An Enroll-to-Verify Approach for Cross-Task Unseen Emotion Class Recognition

Jeng-Lin Li, *Student Member, IEEE,* and Chi-Chun Lee, *Senior Member, IEEE*

Abstract—Most speech emotion recognition studies often focus on recognizing pre-set emotion classes. However, the task definition may change due to a shift in focus to a previously unseen class in real-world applications. This cross-task modeling has not been addressed previously. Lengthy data re-collection, model retraining, and the traditional adaptation and transfer learning approaches are not applicable to this cross-task setting. This study proposes an enroll-to-verify framework to avoid model retraining and rapidly perform a new task prediction using only a handful of enrolled samples. Specifically, we use negative angular margin prototypical loss in a pretrained multiclass network as an emotion encoder. Then, we enroll a few samples corresponding to emotion classes in the new task definition and simply compare the encoded embedding distance to perform recognition. In the experiments on the IEMOCAP dataset, given a four-class pretrained emotion encoder, we achieved a 71.9% unweighted average recall in the frustration (unseen) recognition task. The MELD dataset was used where the unseen class was surprise, fear, or disgust. The results revealed that enrolling only 20 samples without retraining was comparable to supervised training using the complete dataset. Further analyses were conducted to demonstrate the working mechanism of our proposed enroll-to-verify approach.

Index Terms—negative margin, prototypical loss, unseen class, cross-task modeling

1 Introduction

Speech emotion recognition (SER) has advanced substantially for modern intelligent applications, such as smart assistant, customer service, and human-robot interaction. The complexity of emotion causes SER to be a challenging modeling task. A broad range of factors, such as social, cognitive, and contextual elements, make emotion categories inherently difficult to define precisely. Most SER studies to-date focus on the task of basic emotion categories that is defined based on the consensus of primitive emotions: anger, happiness, sadness, disgust, fear, and surprise [1]. Further, due to the natural frequency of occurrence in daily life, the current studies tend to simplify the recognition goal as a four-class recognition task (anger, happiness, neutral, and sadness) [2], [3], [4]. Much effort has focused on improving the general-purpose models for these basic emotions and tend to ignore other emotions, such as frustration, amusement, and guilt. However, a recent study has shown that up to 27 emotion categories can be specified by social, cognitive, and contextual circumstances [5]. They may be rare in the collected datasets but common in our daily lives [6]. Researchers often discard these rarely occurring emotion classes to mitigate data imbalance and model learning difficulties. However, this SER development trend limits the applications for deploying the models in real-life scenarios.

Conventionally, an emotion database is collected based on prior knowledge or a pre-defined task, often resulting in an initial collection of pre-set emotion samples (often primitive categories). However, the task definition can change over time in real life due to different targeted application scenarios. Although the recording environment stays the same, the unseen emotion class in the new task hinders the use of the originally trained models. For example, a customer call dataset initially contains basic four emotion categories, but the task may change to frustration detection. If the frustration class is unseen in the original training dataset, using the originally trained models based on the basic emotions is no longer feasible; re-collecting the frustration data and retraining the model based on the new task lead to additional costs. Furthermore, current transfer learning frameworks are not directly applicable in this case because they only match source to target domain of the same emotion categories. We term this situation as *cross-task modeling*. The dataset has N categories of basic emotions $E = \{e_1, e_2, ..., e_N\}$ labeled initially, but the target task is to recognize emotion classes $E' = \{e_i, e_j\}$ where e_i and e_j can be unseen or seen emotions in the initially collected dataset. We consider the situations with at least one unseen emotion. The possible situations are $e_i \in E, e_j \notin E$ or $e_i, e_j \notin E$.

This study introduces an enroll-to-verify approach for the cross-task problem inspired by the speaker verification process [7], [8] that can rapidly handle verification of unseen speakers. Following the definition of *cross-task modeling*, given adequate samples of primitive emotions (e.g., $E = \{angry, happy, neutral, sad\}$), we aim to enable a new classification task $E' = \{e_i, e_j\}$ by enrolling data collected in the same database, with minimal manual effort. Although many prior studies have achieved high performances on the multiclass basic emotion recognition task [9], [10], none can handle a new task E' without retraining. To address this cross-task problem elegantly, we avoid the meta multitask learning [11], few-shot learning [12] and zero-shot learning [13] approaches which require either complex model retraining for each new task or additional auxiliary knowledge on the data distribution of the unseen categories. In this work,

J.-L. L. and C.-C. L., are with Department of Electrical Engineering, National Tsing Hua University, Taiwan (e-mail: cllee@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw).

We propose using a negative angular margin prototypical (NAMP) loss to pretrain a multiclass (E) network as a prototypical emotion encoder. At the cross-task inference stage, given a new task E', a few utterances of the specified classes are *enrolled* by computing the encoded embedding using the pretrained prototypical emotion encoder, and treated as emotion prototypes. We could then *verify* whether each query utterance is in the e_i or e_j class using the shortest distance to these emotion prototypes.

We evaluated this approach for the cross-task setting on two databases, IEMOCAP and MELD, specifies the predefined task as $E = \{angry, happy, neutral, sad\}$ and verified different new tasks with various combinations of e_i and e_j. In our results, for IEMOCAP, in the newly defined binary frustration recognition task ($E' = neutral, frustration$), by enrolling only 20 samples (ten neutral and ten frustration samples), we achieved performance comparable to using the model trained on all available neutral and frustration samples. A similar setup was carried out on the MELD where the new task was defined as binary classification between the neutral and an unseen class (surprise, fear, or disgust). Furthermore, we compared different loss functions and model architectures used in the pretrained encoder network, changing various new task definitions as different pairs of emotion classes. In our analysis, the visualization of embedding and the change in pretrained emotion classes revealed the working mechanism of the encoder loss and the pretrained emotion space used as the emotion prototype encoder. Another analysis is designed to examine the effects of the enrolled sample number for the verification results.

We highlight our contribution in this study as follows:

- This study is the first to introduce an enroll-to-verify approach to address the cold-start cross-task SER modeling task.
- We conduct a comprehensive evaluation of the enroll-to-verify approach on two large emotion databases.
- We perform analyses to illustrate the different designs of the emotion prototype encoder and enroll-to-verify procedure.

The rest of the paper is organized as follows. Section 3 introduces the enroll-to-verify steps and details of the proposed NAMP loss. Next, Section 4 demonstrates the experimental setups and comparison results for two databases, and Section 5 presents the analysis results. Finally, Section 6 concludes the work and describes future works.

2 Related Works

Current developments in SER research are broadly divided into *within-* and *cross*-database settings. We summarize the past SER studies in Table 1 based on the major modeling issue, targeted topic, and specified task in their respective studies. The table can be used to categorize our proposed study and differentiate it from the past studies. Although a wide range of studies has been carried out to address different issues, researchers usually select a pre-defined set of labels. Two common emotion labels are used: dimensional and categorical, where dimensional attributes include activation, valence, and dominance attributes, and categorical labels often include basic emotion categories.

The *within*-database studies focus on optimizing context (database) specific models which have achieved promising accuracy. One direction of work involves feature space learning for SER, it comprise studies addressing temporal dynamics [14] and in-the-wide data [15], [16] problem to improve recognition performances. Another major line of work has considered various contextual factors, such as speaker variability [9], [17], conversational interlocutor behavior [18], [19], and emotion under particular expressive conditions (monologue speech under stress [20] or dialog when using bus services [32]), to improve SER performances. Issues regarding emotion labels including uncertainty [21], inter-category relationship [22], and inter-task relationship [11] are emerging topics to be addressed. For example, Ma et al. attempted to improve the targeted emotion category classification leveraging category-wise correlations [22]. Cai et al. leveraged dimensional emotion attributes as an auxiliary task to facilitate the categorical classification task [11]. The studies in addressing the inter-category relationship [22] and inter-task relationship [11] problems have leveraged the transfer learning techniques originally developed for the cross-database setting. As these topics are recently identified, they have not been clearly defined in the past transfer learning research paradigm [33], [34].

The SER studies using the *cross*-database setting concerns cross-domain model generalization. The technical developments are broadly covered under the term, "transfer learning" which defines transductive and inductive transfer learning based on the availability of target domain data [33], [34]. The transductive transfer learning focuses on a scenario with unavailable labels in the target domain and usually uses domain adaptation techniques. This line of works concentrates on developing algorithms to align the cross datasets feature distribution. Exemplary studies involve common space learning, which improve feature transferability [23], [24], [25]. The inductive transfer learning involves a scenario with available target domain labels and even includes contextualized factors modeling that align the domains by considering gender and language factors [26], [27], [35]. The domain shift problem caused by different distributions in the same emotion category between domains are partially mitigated by domain adversarial learning [28], [29], [30], [31]. While major effort has been carried out in this domain, the application setting of these algorithms is also constrained. One obvious limitation is that although every context realistically includes nonoverlapping emotion categories, these *cross*-database algorithms are designed to match the same emotion categories between the source and target domains. This constraint, while partially alleviated by working in the universal dimensional representation of emotion (activation and valence) to avoid different label categories in the target domain [27], [31], must be relaxed to be flexible in a real world setting. Also, the dimensional attributes are usually derived using a rough mapping rule from the categorical labels, which introduces unnecessary artifacts and uncertainty to the task.

In summary, most if not all of these prior works in settings of either *within*-database or *cross*-database confronted a similar problem in which they do not address the issue of efficiently perform emotion recognition for unseen or rarely occurring emotion categories.

TABLE 1
A summary of speech emotion recognition studies.

Within Database			
Issue	feature space learning	contextualized factors	labeling
Topic	temporal dynamics [14] in-the-wide data [15], [16]	speaker variability [9], [17] conversation [18], [19] conditioned situations [20]	uncertainty [21] inter-category relationship [22] task transfer [11] cross-task verification (Ours)
Task	dimension [16] category [14], [15]	dimension [18], [20] category [9], [17], [19]	dimension [21] category [22] dimension → category [11] category → new category (Ours)
Cross Database			
Issue	feature space learning	contextualized factors	domain shift
Topic	common space learning [23], [24], [25]	multitask learning [26] cross language [12], [27]	domain adversarial [28], [29] [30], [31]
Task	dimension [23] category [24], [25]	dimension [27] category [12], [26]	dimension [28], [30], [31] category [29]

TABLE 2
A summary of two databases emotion distribution

	IEMOCAP	MELD		
Emotion	cross-validation	train	validation	test
angry	1103	1109	153	345
happy	1636	1743	163	402
neutral	1708	4710	470	1256
sad	1084	683	111	208
frustration	1849	-	-	-
disgust	2	271	22	68
fear	40	268	40	50
surprise	107	1205	150	281

3 Methods

3.1 Emotion Databases

We used the IEMOCAP and MELD databases, and the emotion distributions of the datasets are listed in Table 2.

3.1.1 IEMOCAP

The IEMOCAP is a benchmark SER dataset [36] which includes around 12 hours of dyadic spoken interaction sessions from ten speakers. The resulting 10039 utterances are labeled for emotion categories by at least three annotators. We include the anger, happiness, neutral, and sadness categories widely-used in the previous studies [9], [37] as the pre-set (basic) emotions for model pre-training. The excitement category is merged into the happiness class. In addition, the frustration class is a non-primitive yet common emotion class and has enough samples. We use it as the unseen emotion class for our cross-task modeling.

3.1.2 Multimodal EmotionLines Dataset

The MELD is a multimodal multi-party database collected from the TV series "Friends" [38]. It contains seven emotion categories (anger, disgust, fear, happiness, sadness, surprise, and neutral). The dataset is split into training, validation, and testing sets released for research. For cross-task experiments, we used the same four emotion classes (anger, happiness, neutral, sadness) in the pretrained model, and used disgust, fear, and surprise as the unseen classes for the targeted recognition task.

3.2 Features

In this study, we extracted vq-wav2vec as the acoustic features [39] trained using a self-supervised context prediction task with contrastive loss to embed information from raw audios. The pretrained audio encoder could directly project the raw waveform into the latent space. The vq-wav2vec extended wav2vec by conducting vector quantization using the Gumbel-Softmax approach. Latent features were organized as different groups mapped to codebook vectors. Discrete tokens were used in BERT training, optimized with random masks. We used the pretrained model on the Librispeech corpus [40] containing 960 hours of audio. The 512 dimensional latent vq-wav2vec vectors were extracted using the fairseq tool [41].

3.3 Enroll-to-Verify Framework

The overall framework is illustrated in Fig. 1. The three primary steps include: the backbone prototype encoder network pretraining, targeted task enrollment, and unseen emotion verification. These steps are similar to the speaker verification procedure [42]. As the cross-task modeling has not been clearly defined, and the enroll-to-verify steps have not been used in the SER domain, we introduce the procedures in detail in the following subsections. We describe the pretrained prototype encoder network architecture in section 3.3.1, and the NAMP loss in Subsection 3.3.2. The targeted task enrollment and verification procedure are described in section 3.3.3. We also organize the enroll-to-verify approach as a step-by-step procedure shown in Algorithm 1.

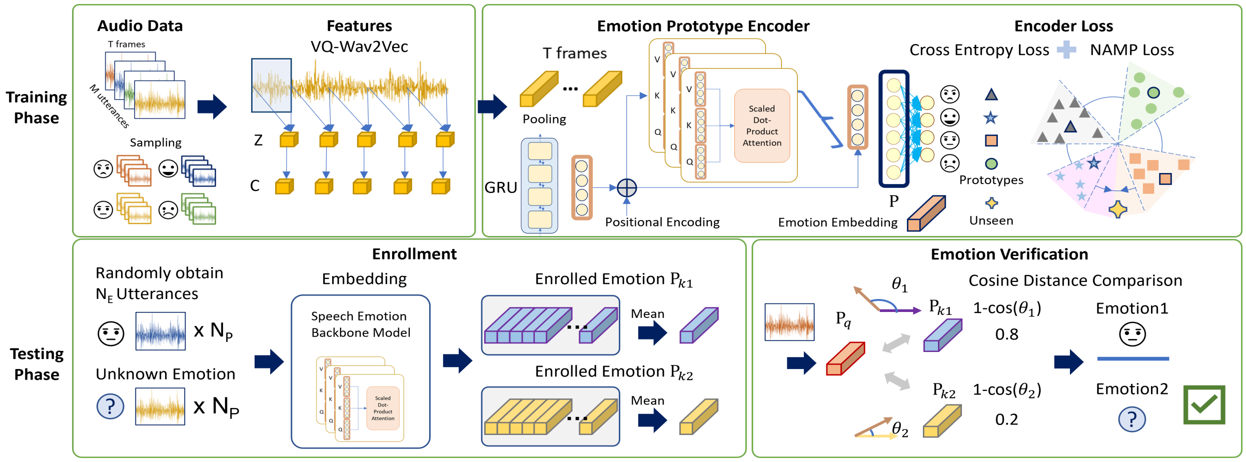


Fig. 1. *This is the overall enroll-to-verify approach for cross-task emotion recognition. The emotion prototype encoder (GRU-transformer) uses vq-wav2vec features to optimize with both multi-class cross entropy loss and negative angular margin prototypical (NAMP) loss for further emotion embedding extraction. The enrollment includes N_P samples for each of the targeted emotion classes. Finally, by computing the distance between the averaged enrolled emotion embeddings and the emotion embedding in query, the class with smaller distance is assigned as the final prediction.*

3.3.1 Emotion Prototype Encoder

The network architecture of the emotion prototype encoder is a GRU-transformer. We performed two-step temporal pooling on the vq-wav2vec features before the GRU layer. Then, we used a transformer stacked with multi-head self-attention layers [43] to capture different temporal components. The acoustic features $x_t \in \mathbb{R}^D$ at a timestep t are transformed to the key and query space by the scaled dot-product self-attention:

$$\alpha_t\tau = \frac{exp(\beta x_t^T W_q^T W_k x_\tau)}{\sum_{\tau'} exp(\beta x_t^T W_q^T W_k x'_\tau)}, \tag{1}$$

where $W_q, W_k \in \mathbb{R}^{D\times D}$ and $\beta = \frac{1}{\sqrt{D}}$ is a scaling factor. We transformed x_t into the value space and calculated the weighted sum with scaled dot product attention to obtain the output embedding P. Empirically, the representations are more robust when including the GRU layer before the transformer for multiclass recognition. After performing mean pooling to obtain an embedding vector of the latent transformer output, the final fully connected layer maps the mean embedding vector to the multiclass emotion space. Thus, the embedding vector is also used for loss calculation and verification steps.

3.3.2 Encoder Loss

We proposed using the NAMP loss to train the emotion prototype encoder network described in section 3.3.1. This section first states the formulation of the angular margin prototypical loss to explain why the loss design can be used in the enroll-to-verify process. Then, we elaborate a generalization issue on the loss design if we directly apply the loss to the SER domain. Therefore, we describe the design of NAMP loss to mitigate the issue for cross-task SER application scenarios.

The idea to train the emotion prototype encoder with designed loss functions originated from the speaker verification domain. The designed losses enhanced the ability to represent new data by extracting embeddings from a learned encoder and allowed the enroll-to-verify procedure by a simple distance measurement. With N utterances sampled from K emotion classes, the i^{th} sample embedding from class $k \in K$ is denoted as $P_{k,i}$, where $i \in N$. This sampling technique from metric learning implemented the same procedure of the verification steps that enrolled a few utterances as prototypes. This N number is usually chosen by the same number for enrollment and thus the training can match the same process of enroll-to-verify in the testing stage. Similar to the typical prototypical network [44], we deemed the average of the episodic mini-batch embeddings as the centroid representation c_k:

$$c_k = \frac{1}{N-1}\sum_{i=1}^{N-1} P_{k,i} \tag{2}$$

We computed the distance between the embedding $P_{j,i}$ from class $j \in K$ to each centroid c_k using cosine similarity:

$$S_{j,k} = w \cdot cos(P_{j,i}, c_k) + b \tag{3}$$

where $w > 0$ was a learnable scaling factor. This cosine similarity measurement projected the distance metric onto a normalized hypersphere. In the two conditions, $j = k$ and $j \neq k$, the loss could be designed to reduce the distance between representations for the $j = k$ condition, and enlarged the distance for the $j \neq k$ condition. Most previous speaker and face verification studies had also followed this loss design for discriminative learning of the network [7], [45], [46], [47]. The latent embedding extracted from the discriminative network could thus properly represent each acoustic sample in the enroll-to-verify procedure.

Combined with the angular margin loss [7], we could introduce a margin $m \in \mathbb{R}$ to the similarity function to adjust the inter-class and intra-class emotion distribution on the hypersphere.

$$S_{j,k} = \begin{cases} w \cdot (cos(P_{j,i}, c_k) + m) & j = k \\ w \cdot cos(P_{j,i}, c_k) & j \neq k \end{cases}. \tag{4}$$

However, Liu et al. have found a generalization issue in a few-shot learning study that the discriminative loss overemphasized on large inter-class variance and small intraclass variance of the pretrained classes and led to large intra-class variance of the unseen class [48]. In the case of the enroll-to-verify for cross-task SER, the pretrained emotion prototype encoder cannot be updated as the few-shot learning does. If the pretrained emotion prototype encoder cannot enlarge the intraclass variance of the unseen class, it would deteriorate testing performance. In addition, the naturally existing emotion ambiguity is another reason making the generalization issue serious. The small intra-class variance of the pretrained prototype emotion classes reduces the representation capability on the overlapped pre-trained emotion space where the unseen emotion might be located. Hence, we constrained the margin $m < 0$ learned as a negative value, to penalize the over-discriminated decision boundaries between different basic emotion classes. The similarity $S_{j,k}$ was transformed via a negative log likelihood function as the NAMP loss. Finally, we added the cross-entropy loss L_{ce} to balance the discriminability and the emotion ambiguity of the NAMP loss:

$$L = L_{ce} - \frac{1}{K}\sum_{j=1}^{K} \log \frac{e^{S_{j,j}}}{\sum_{k=1}^{K} e^{S_{j,k}}}. \tag{5}$$

The cross-entropy loss also maintains the robustness of model training, and alleviates convergence failure on the basic emotion classification task.

Algorithm 1 Enroll-to-verify

```
Input Acoustic features $X = \{x_i | 1 <= i <= N\}$ and each $x_i \in \mathbb{R}^{T \times D}$ has time step $T$ and feature dimension $D$
Training Randomly initialize parameters $\theta$ in an emotion prototype network $f$ with an encoder $enc$ and a classifier $clf$. The training will run for $N_{iter}$ iteration.
for $i = 1$ to $N_{iter}$ do
    Calculate embedding $P_i = enc(x_i)$
    Calculate the centroid $c_k$ of the sample's corresponding to label class $k$
    Generate prediction by the embedding, $y_i = clf(P_i)$
    Calculate the Loss $L$ with the prediction $y$ for softmax loss and the embedding $P_i$ along with the centroid $c_k$ for the NAMP loss
    Optimize parameters $\theta$ with the calculated loss
Enroll Acoustic features $X_{k1}^{val} = \{x_{k1,1}^{val} | 1 <= i <= N_p\}, X_{k2}^{val} = \{x_{k1,2}^{val} | 1 <= i <= N_p\}$ for class $k1$ and $k2$ sampled from the validation set with $N_p$ samples for each class
$\overline{P}_{k1} = \frac{1}{N_p}\sum_{i=1}^{N_p} enc(x_{k1,i}^{val})$
$\overline{P}_{k2} = \frac{1}{N_p}\sum_{i=1}^{N_p} enc(x_{k2,i}^{val})$
Verify Acoustic features of a query sample $x_q$.
Calculate embedding $P_q = enc(x_q)$
Compute cosine distance $D_{q,k1}$ between $P_q$ and $\overline{P}_{k1}$
Compute cosine distance $D_{q,k2}$ between $P_q$ and $\overline{P}_{k2}$
Output class $k$ by $\operatorname*{argmin}_{k}\{D_{q,k1}, D_{q,k2}\}$
```

3.3.3 *Enrollment and Verification Procedure*

With the trained prototype multiclass network encoder based on NAMP loss, we extracted the transformer-encoded output P as the emotion embedding for each utterance. In the testing phase, we enrolled N_P embeddings from utterances for each emotion class in a newly defined task which includes unseen emotion class. Then, we computed the average of these N_P embeddings of the emotion class $k \in K$ as an enrolled emotion representation $\overline{P}_k$. This emotion representation $\overline{P}_k$ is an emotion prototype that characterizes the emotion class in the learned deep network embedding space. Finally, for each query embedding P_q for testing, we calculated the cosine distance $D_{q,k}$ (one minus cosine similarity) between this embedding P_q and each enrolled embedding $\overline{P}_k$. The utterance is predicted as emotion class k by choosing the smallest $D_{q,k}$, finding the most similar enrolled emotion prototype using $\min_{k'} D_{q,k'}$.

4 Experiments

4.1 Experimental Setup

We performed leave-one-dyad-out cross-validation on the IEMOCAP dataset, and split the training, validation, and testing sets on the MELD, following the common settings for these two databases. On the IEMOCAP, we left one speaker for testing and another speaker from the same dyad for validation in each fold. We reported the unweighted average recall (UAR) values and weighted F1 scores in Table 4 for a comprehensive evaluation. We conducted the following three experiments to compare different state-of-the-art models and validate the network parameters.

- Exp I: comparison of different encoder loss functions
- Exp II: comparison of different input features and network architectures of the emotion prototype encoder
- Exp III: results from varying definitions of new tasks E' of different emotion classes

The pretrained task E included four basic classes (angry, happy, neutral, and sad), and a new task E' was defined with an unseen emotion e_5 and an observed class using the same database. In Exp I and Exp II, the cross-task experiment was set up to pretrain a four-class emotion network encoder for one task and recognize $E' = \{neutral, e_5\}$ for another task. We chose frustration in the IEMOCAP and disgust, fear, and surprise in the MELD as the unseen emotion, e_5 for the targeted task. In Exp III, we changed the new task definition to $E' = \{e_i, e_5\}$ where e_i can be angry, happy, or sad. In the enroll-to-verify procedure, we enrolled 10 neutral samples and 10 targeted class samples from the validation set and performed the binary emotion verification task on the testing set. The setting examined the situation in which we could not retrain the model and none of the prior information for the targeted emotion was accessible. We detailed the compared loss functions of Exp I in Section 4.1.1, and the compared features and architectures of Exp II in Section 4.1.2.

4.1.1 *Compared Loss Functions*

Exp I was conducted to compare the encoder loss described in Section 3.3.2 with other state-of-the-art loss functions for verification. We showed the four-class results on the pretrained task, which was comparable to recent studies, and focused on the performance comparison of the newly defined task with the unseen emotion class. In Table 4, *Prior* denotes the best performance from the prior SER works on the four-class prediction task using the IEMOCAP [9], [10] and MELD [19], [37]. This baseline is reported to demonstrate the performance our pretrained encoder network. *All* denotes the model trained with the complete two-class data from the entire dataset for the newly defined binary classification task. For example, we used 1,849 frustration samples and 1,708 neutral samples to conduct cross-validation for this binary classification experiment. This strong baseline indicates the scenario when we re-collect labeled data and re-train a model for the new task. For the other compared models, we followed the setting described in Section 4.1, which only used 10 neutral samples and 10 targeted class samples from the validation set for the final binary task. Considering that no previous study has this cross-task problem explicitly, we compared several state-of-the-art speaker verification approaches for this cross-task experiment, as follows:

- **SM**: Basic multiclass cross entropy loss.
- **AM**: Angular margin softmax loss [7], [45].
- **AAM**: Additive angular margin softmax loss [7].
- **Tri**: Triplet loss, minimizing intraclass distances and maximizing interclass distances [49], [50].
- **Proto**: Prototypical loss, computing the squared Euclidean distance against centroids for classification [44].
- **AProto**: Angular prototypical loss, using the cosine similarity function with a linear transformation [8].
- **AMProto**: Constraint for the margin to be learned as positive values for *AProto*.
- **NAMP**: Our proposed approach in Section 3.3.2.

The *AM* and *AAM* are two angular margin approaches designed to enlarge interclass variance and minimize intraclass variance. The *Tri*, *Proto*, and *AProto* approaches are metric learning approaches that directly optimizes the distance metric for small intraclass and large interclass distances. *Tri* assigns a positive and a negative sample to an anchor while *Proto* and *AProto* estimates the centroid to compare the intraclass and interclass distance. The *AMProto* and *NAMP* combine the angular margin with the prototypical network and use positive and negative margins, respectively. Besides the different loss functions, we also reported margin m as a hyper-parameter to show the effects of angular margin prototypical loss. We changed margin m in Eq. (4) from -0.4 to 0.3 and presented the results in Fig. 2.

4.1.2 Compared Features and Encoder Architectures

In Exp II, we changed the acoustic feature set (Section 3.2) and the emotion prototype encoder architecture (Section 3.3.1) to other state-of-the-art features and networks. These networks were all trained using the *NAMP* loss described in Section 3.3.2. In Table 5, we compared the following features and networks:

- **NAMP**: Our proposed approach in Section 4.1.1
- **Emobase**: Using the emobase feature set extracted from the opensmile toolbox [51] to replace the vq-wav2vec in Section 3.2 for the GRU-Transfomer
- **Transformer**: NAMP without the GRU layer before the transformer
- **BLSTM**: The Bidirectional Long Short Term Memory network using the vq-wav2vec features described in Section 3.2
- **BLSTM+NetVLAD**: BLSTM with NetVLAD layer [52] using vq-wav2vec features
- **CNN+NetVLAD**: Convolutional Neural Network with NetVLAD layer [53] using spectrogram as input

TABLE 3
The hyperparameters used in our experiment. N_{layer} and N_{head} denote number of layers and heads in the transformer network. The notation N, K, and N_p are minibatch sampled size, emotion classes, and number of enrolled samples which are consistent to the notation in Section 3.3.2.

Network Architecture				
Dataset	**GRU**	**transformer**	$\mathbf{N_{layer}}$	$\mathbf{N_{head}}$
IEMOCAP	256	512	1	2
MELD	256	128	1	2
Training Parameters	**epoch**	**learning rate**	**decay step**	**decay rate**
	100	0.001	20	10
	N	**optimizer**	**K**	$\mathbf{N_p}$
	10	Adamax	4	10

The *emobase* has been used in several SER studies [9], [54], and contains the pitch (fundamental frequency), intensity (energy), loudness, cepstral (12 MFCC), voicing probability, fundamental frequency envelope, eight line spectral frequencies, zero-crossing rate, and delta regression coefficients of these low-level descriptors. *BLSTM* has been widely used in previous works [9], [17] , and *BLSTM+NetVLAD* and *CNN+NetVLAD* include the NetVLAD encoding layer has been widely used in the speaker recognition [55]. In addition, *CNN+NetVLAD* was fine-tuned from the speaker recognition pretrained model on the VoxCeleb database [56].

4.1.3 Network Configuration

The hyperparameters were obtained through the best validation performance in a grid search. The 512-dimensional vq-wav2vec features were input into a single-layer GRU with 256 nodes, followed by {1, 2} transformer layers with {64, 128, 256, 512} nodes. The number of attention heads was selected from the set {1, 2, 4}. We trained the network with 100 epochs and a 0.001 learning rate, and we divided the learning rate by 10 every 20 epochs. The mini-batch size was ten for each emotion class to match the enrolled number in the verification stage and the optimizer was Adamax. The searching range of the hyperparameters covers the parameters used in the previous four-class emotion recognition studies [9], [19]. Table 4 shows a summary of the hyperparameters. The models are trained using a NVIDIA GTX 1080 GPU with 11GB memory. The parameters N and K indicate that we sample N samples from the K classes in a batch for training the encoder described in Section 3.3.2.

4.2 Exp I: Comparison of Encoder Losses

For the four-class pretraining task, as listed in Table 4, the proposed *NAMP* achieved 62.3% and 40.3% UAR on the IEMOCAP and the MELD, respectively. The results were

TABLE 4
The results of different cross-task modeling approaches. Details of the abbreviations (Prior: Prior Works, All: re-training with the whole dataset, SM: softmax, AM: angular margin, AAM: additive angular margin, Tri: triplet, Proto: prototypical, Aproto: angular prototypical, AMProto: angular margin prototypical, NAMP: negative angular margin prototypical) refer to session 4.1.1.

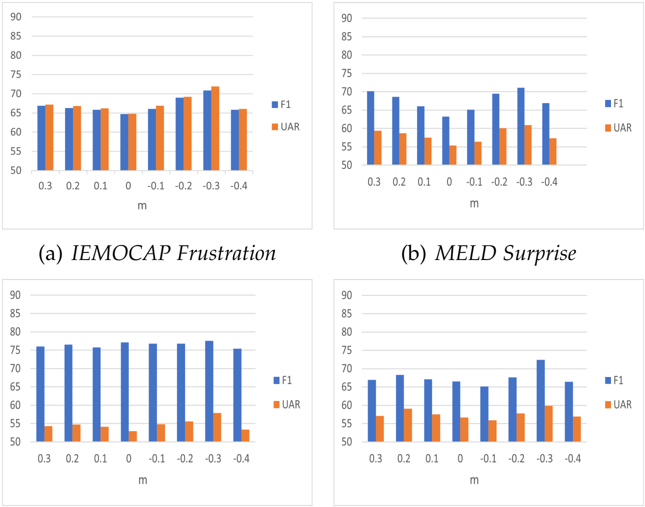
IEMOCAP	**4-class**	**Prior**	**SM**	**AM**	**AAM**	**Tri**	**Proto**	**AProto**	**AMProto**	**NAMP**
Pretrain	UAR	63.1	58.2	62.3	57.9	49.3	58.3	61.1	59.8	62.3
Verify	**2-class**	**All**	**SM**	**AM**	**AAM**	**Tri**	**Proto**	**AProto**	**AMProto**	**NAMP**
frustration	F1	69.2	62.2	64.4	64.1	64.1	62.4	58.3	**66.9**	**70.9**
	UAR	70.0	63.1	65.2	65.1	64.5	62.8	61.1	**67.2**	**71.9**

MELD	**4-class**	**Prior**	**SM**	**AM**	**AAM**	**Tri**	**Proto**	**AProto**	**AMProto**	**NAMP**
Pretrain	UAR	40.3	40.0	40.3	39.6	34.9	36.4	39.7	39.2	40.4
Verify	**2-class**	**All**	**SM**	**AM**	**AAM**	**Tri**	**Proto**	**AProto**	**AMProto**	**NAMP**
surprise	F1	75.6	57.8	**72.0**	66.8	64.8	59.5	61.3	70.2	**71.1**
	UAR	69.6	56.8	59.7	56.9	**60.4**	55.7	55.6	59.4	**60.9**
fear	F1	79.0	70.8	**86.7**	77.5	77.0	65.5	75.6	76.0	**77.6**
	UAR	56.4	**59.3**	54.9	56.2	51.0	50.2	57.8	54.3	**57.9**
disgust	F1	76.3	**71.8**	56.6	68.4	63.5	59.1	68.6	66.9	**72.4**
	UAR	63.2	54.9	56.6	57.1	51.9	49.4	**58.0**	57.1	**59.8**

comparable to the previous works on the IEMOCAP (57.1% to 63.1%) [9], [10] and the MELD (39.4% to 40.3%) [19], [37]. Although [10] achieved 63.1% and 55.5% UAR on the IEMOCAP using a transformer with and without ensemble learning, respectively, the proposed framework alleviated the heavy computational burden of the ensemble approach and still obtained a comparable UAR.

For the newly defined binary verification task (cross-task modeling), our proposed *NAMP* (F1 = 70.9% and UAR = 71.9%) performed slightly better than *All* (F1 = 69.2% and UAR = 70.0%) and significantly better (p-value < 0.05 in McNemar's test [57]) than all the other methods in the frustration versus neutral classification on the IEMOCAP dataset. We achieved this comparable result by enrolling only 20 utterances (10 frustration utterances and 10 neutral utterances) rather than training with over 2000 samples. This result indicates the capability of the well pretrained network to perform cross-task recognition without explicit retraining. Other margin-based algorithms such as *AM* and *AAM* outperformed *SM* by 2.2% for the F1 score and 2.1% for UAR and by 1.9% for the F1 score and 2.0% for UAR, respectively, which indicates the necessity of using added margin in this cross-task modeling. In addition, *AMProto* had a 66.9% F1 score and a 67.2% UAR, demonstrating that the margin was also beneficial to prototypical learning. Although this positive margin approach worked, the negative margin (*NAMP*) performed even better. This result implies that regularizing the basic emotion pretraining with the negative margin is more favorable than overly enlarging inter-class distances when the learned model derives representations for unseen emotions.

On the MELD, *NAMP* attained the highest UAR in recognizing surprise and disgust versus neutral, and achieved the second highest UAR for the fear recognition task. The results for the new task on surprise were superior to those of *All* by 4.5% for the F1 score and 8.7% for UAR, and its results for disgust were superior by 3.9% F1 and 3.4% UAR, whereas the fear task had a minor difference (-1.4% for F1 and +1.5% for UAR). The baseline method, *SM*, failed to obtain comparable results on every metric which also occurred on *AM* and *Tri*. Generally, *AProto* and *AMProto* performed better than *Tri* and *Proto*. The positive margin improved only the surprise task results (8.9% for F1 and 3.8% for UAR). However, with its negative margin, *NAMP* surpassed *AProto* in recognizing surprise (9.8% for F1 and 5.3% for UAR), fear (2.0% for F1 and 0.1% for UAR), and disgust (3.8% for F1 and 1.8% for UAR). In addition, *NAMP* outperformed the angular loss methods *AM* and *AAM*. Here, *NAMP* maintained the advantages from the prototypical loss that mimics enrollment and verification via episodic sampling and the discriminative space from the margin approaches.



(a) *IEMOCAP Frustration*

(b) *MELD Surprise*

(c) *MELD Fear*

(d) *MELD Disgust*

Fig. 2. *The results using different margin in the loss.*

In comparing different margins in the angular margin prototypical loss, we set the margin m in Eq. (4) as a hyper-parameter to examine its corresponding F1 and UAR values in the cross-task modeling. Fig. 2 presents the results of frustration on the IEMOCAP dataset, and surprise, fear, dis-

TABLE 5
Results compared with different network architectures and parameters.

IEMOCAP	**4-class**	**NAMP**	**Emobase**	**Transformer**	**BLSTM**	**BLSTM+NetVLAD**	**CNN+NetVLAD**
Pretrain	UAR	**62.3**	59.4	60.8	56.5	57.9	55.1
Verify	**2-class**	**NAMP**	**Emobase**	**Transformer**	**BLSTM**	**BLSTM+NetVLAD**	**CNN+NetVLAD**
frustration	F1	**70.9**	65.5	67.8	62.8	65.8	62.3
	UAR	**71.9**	65.4	68.9	62.9	65.9	63.0
MELD	**4-class**	**NAMP**	**Emobase**	**Transformer**	**BLSTM**	**BLSTM+NetVLAD**	**CNN+NetVLAD**
Pretrain	UAR	**40.3**	38.4	39.1	35.4	36.3	34.9
Verify	**2-class**	**NAMP**	**Emobase**	**Transformer**	**BLSTM**	**BLSTM+NetVLAD**	**CNN+NetVLAD**
surprise	F1	**71.1**	64.3	70.3	58.8	65.0	59.1
	UAR	**60.9**	57.2	59.8	56.6	59.6	55.9
fear	F1	**77.6**	75.8	77.2	74.1	77.4	73.2
	UAR	57.9	55.4	56.9	51.8	**58.5**	52.3
disgust	F1	**72.4**	70.1	70.8	70.7	70.6	69.3
	UAR	**59.8**	56.6	58.6	52.3	58.5	52.2

gust on the MELD dataset. The most effective margin values were usually -0.3 and 0.3, while the lowest F1 and UAR often occurred when the margin was equal to zero. The difference in the positive margin values could attain the best results using $m = 0.2$ or $m = 0.3$, and no significant performance difference was found for $m = -0.2$ and $m = -0.3$. However, the results deteriorated when the margin value equals -0.4. A large negative margin might overly encourage overlap between the pretrained feature space and degrade the basic discriminative power on emotion. Therefore, properly selecting the margin is an important issue because a tradeoff exists between the discriminative capability and generalizability. The large positive margin would overemphasize the discriminative capability and reduce the generalizability for the cross-task scenario while the large negative margin can directly damage the classification ability on the pretrained network. We can regard the negative margin as a regularization mechanism, and select a proper negative margin value which still result in competitive performance on the pretrained four-class recognition task.

4.3 Exp II: Comparison of Features and Architectures

The pretrained emotion prototype encoder plays a key role in providing a proper embedding space for the enroll-to-verify process. Therefore, we compared different input features and several architectures to examine the effects of pretrained encoder networks on the emotion verification results in the cross-task setting.

In Table 5, using the emobase feature set achieved a 59.4% UAR for the four-class recognition task and a 65.5% F1 score and a 65.4% UAR for the frustration classification task on the IEMOCAP dataset. The inferior performance on both tasks suggests that vq-wav2vec features are more representative of affective information than the emobase feature set. A similar tendency was observed on the MELD.

The *transformer*, when removing the GRU layer, negatively affects the pretrained and verification tasks on two datasets. However, it still outperformed other features and network architectures in Table 5. The *BLSTM*, *BLSTM+NetVLAD*, and *CNN+NetVLAD* have recently been used in the SER works. The *BLSTM+NetVLAD* generally achieved better performance than *BLSTM* and *CNN+NetVLAD*. For the frustration recognition task, the verification performance of *BLSTM+NetVLAD* (F1 = 65.8% and UAR = 65.9%) was higher than that for *BLSTM* (F1 = 62.8% and UAR = 62.9%) with 3.0% improvement in the F1 score and UAR. Similar results were also observed on the MELD, where *BLSTM+NetVLAD* attained margins for the F1 score of 6.2%, 3.3%, and -0.1%, and for the UAR of 3.0%, 6.7%, and 6.2% for surprise, fear, and disgust verification tasks, respectively compared to the *BLSTM*. Although both models obtained similar F1 scores on the disgust verification task, *BLSTM+NetVLAD* still had a higher UAR value in recognizing the disgust class. The NetVLAD layer improved the BLSTM network for the four-class recognition and the verification cross-task results.

The proposed *NAMP* using a GRU layer and transformer outperformed *BLSTM+NetVLAD* on the frustration verification task on the IEMOCAP dataset (5.1% F1 and 5.0% UAR) and the surprise verification on the MELD (6.1% F1 and 1.3% UAR). In addition, *CNN+NetVLAD* was designed for speaker recognition [55] although it did not improve SER performance. Directly applying the network architecture from other similar enroll-to-verify recognition tasks (such as speaker verification) can be ineffective for SER.

In this experiment, we see that the performance of the pretraining task and new verification task is mostly positively correlated. The pretrained feature space of the emotion representations influences the verification results. Therefore, a discriminative network architecture is essential for robust cross-task verification.

4.4 Exp III: Different Emotion Verification Tasks

In this section, we performed new verification tasks for different emotion classes versus frustration, (i.e., angry versus frustration, happy versus frustration, and sadness versus frustration). The results are listed in Table 6. In the IEMOCAP dataset, the binary verification results to differentiate sad and frustration achieved an 80.1% F1 score and an 82.0% UAR value, higher than results of other emotion classes. The sad class obtained 64.9% recall in the pretrained four-class result which was inferior to the angry class. The pretrained results did not directly indicate the performance of a new task, and the enroll-to-verify approach was effective to be

TABLE 6
The results different verification tasks E', choosing any of the four emotion categories (neutral, angry, happy, sad) versus an unseen emotion class on the IEMOCAP dataset and the MELD dataset.

IEMOCAP	**Pretrain**	**Verify**		**MELD**	**Pretrain**	**Verify**					
	4-class	Frustration			4-class	Surprise		Fear		Disgust	
Emotion	UAR	F1	UAR	**Emotion**	UAR	F1	UAR	F1	UAR	F1	UAR
Neutral	61.4	70.9	71.9	**Neutral**	36.1	71.1	60.9	77.6	57.9	72.4	59.8
Angry	69.1	64.4	65.1	**Angry**	44.7	57.2	56.9	60.4	55.4	61.4	55.0
Happy	54.0	67.6	65.8	**Happy**	41.0	53.2	52.6	66.2	54.2	60.9	66.5
Sad	64.9	80.1	82.0	**Sad**	38.6	57.5	57.0	63.5	54.1	65.3	68.6

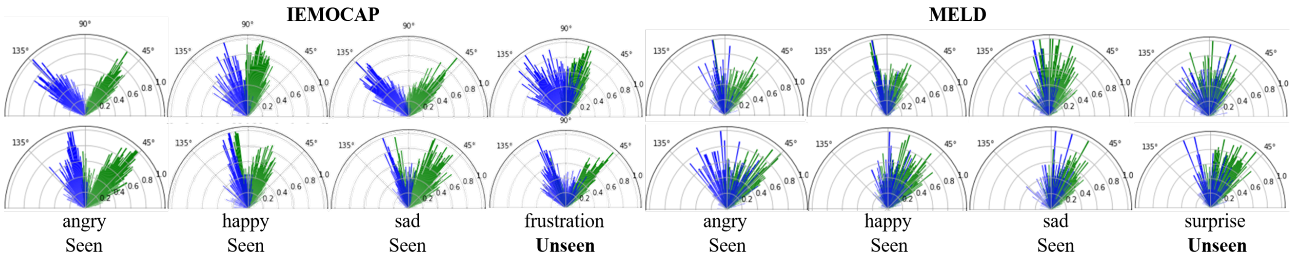


Fig. 3. *The cosine distance visualized histogram of the each emotion versus neutral. The first row are figures using positive margin and the second row are figures using negative margin.*

applied to different task definitions. In the MELD, the tasks, differentiating surprise, fear, or disgust classes from the neutral class generally outperformed the tasks differentiating these classes from other classes (angry, happy, and sad). The surprise, fear, and disgust classes are all basic emotions, whose acoustic samples might not be fully represented by the four classes used in the pretrained model.

The performance on the new verification task reflects the relationship between the two emotion classes. The sad class versus surprise class obtained a 57.5% F1 score and a 57.0% UAR value, whereas the happy class attained the lowest F1 score and UAR value. The acoustic samples of the surprise class are embedded as a representation much closer to the samples of the happy class than the sad class in the latent embedding space. The low UAR results in the tasks of the fear or disgust class were due to an imbalanced data distribution. The UAR value of the disgust versus angry (55.0%) task was relatively lower than the disgust versus happy or sad class. These results imply similar acoustic characteristics between the angry and disgust classes.

TABLE 7
The verification task results using three-class and four-class pretrained model.

IEMOCAP	**5-class**		**4-class**		**3-class**	
Pretrain	F1	UAR	F1	UAR	F1	UAR
	56.7	52.1	61.3	62.3	71.7	73.2
Verify	F1	UAR	F1	UAR	F1	UAR
Frustration	66.4	66.4	70.9	71.9	68.9	69.2
MELD	**5-class**		**4-class**		**3-class**	
Pretrain	F1	UAR	F1	UAR	F1	UAR
	24.5	32.8	43.3	40.4	41.5	50.7
Verify	F1	UAR	F1	UAR	F1	UAR
Surprise	-	-	71.1	60.9	68.3	59.1
Fear	64.9	54.7	77.6	57.9	75.9	56.9
Disgust	62.8	55.8	72.4	59.8	71.2	58.3

5 Analyses

5.1 Embedding Distance Visualization

In this section, we use visualized distance polar bars to demonstrate the effects of the negative margin in the framework. We compute the enrolled neutral representation ($\overline{P}_{neu}$) and enrolled emotion representation ($\overline{P}_k$) for every emotion class by averaging all emotion embeddings in that class. For each encoded query sample E_q, we compute the cosine distance to $\overline{P}_{neu}$ and $\overline{P}_k$ as D_{neu} and D_k, and the polar angle using $180 * \frac{D_{neu}}{D_{neu}+D_k}$. We set neutral at 0° and the target emotion at 180° to examine the intra-class distance to the emotion centroids and interclass data variance. Fig. 3 reveals the accumulated cosine distance histogram of emotion classes with *AMProto* in the first row (positive margin) and *NAMP* in the second row (negative margin). By examining angry, happy, and sad classes relative to the neutral class, we observe that the positive margin results in a greater distance between the target emotion group in blue and neutral group in green compared to the negative margin. This phenomenon is more evident in the IEMOCAP than the MELD due to the better multiclass pretraining performance. However, the unseen frustration samples in the IEMOCAP and the surprise samples in the MELD have more overlapping distributions; therefore, they are less discriminative on the unseen classes with a positive margin. This conclusion corroborates the cross-task results in Table 4 revealing that a properly overlapping emotion distribution in the latent space of the basic emotion model can enable more representative space for unseen classes, especially in verifying a case that requires no further supervised training.

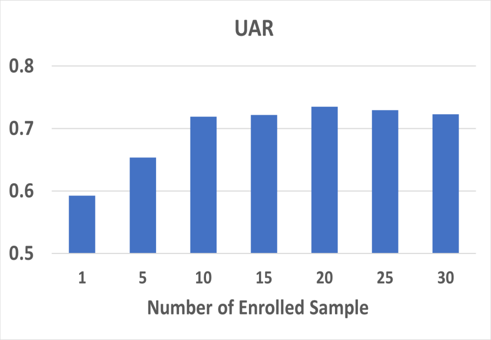


Fig. 4. *The cross-task verification results using a different number of enrolled samples. We used the number of samples on the x-axis for enrollment and performed verification to derive the resulting UAR on the IEMOCAP dataset.*

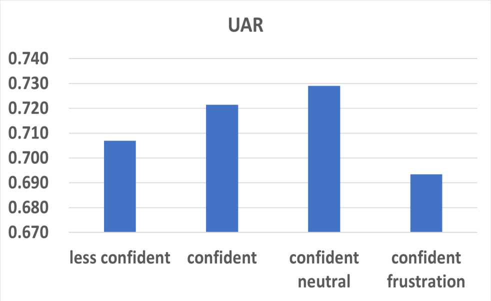


Fig. 5. *The cross-task verification results using enrolled samples with different label confidence. We used the different confidence groups of samples on the x-axis for enrollment and performed verification to derive the resulting UAR on the IEMOCAP dataset.*

5.2 Pretrained Emotion Classes for Prototype Encoder

The previous emotion verification experiments were conducted by pretraining the commonly-used four-class emotion recognition model. This analysis experiment used a five-class pretrained model by adding the surprise class and a three-class pretrained model by removing the happy class from the four-class. In Table 7, the three-class pretrained models usually obtained lower F1 and UAR than four-class models in different tasks and the five-class pretrained models performed worse than both the three-class and four-class models. The results demonstrated the importance of the pretrained emotional feature space. Increasing diversity of the pretraining data from the three emotion classes to four classes enhances the ability to properly represent new samples in a new task, and thus improves the verification performance. When removing data for the happy class from the pretraining in the IEMOCAP, the differentiation task for the frustration and neutral classes, the F1 score declined from 70.9% to 68.9%, and the UAR declined from 71.9% UAR to 69.2%. The added surprise class in the five-class pretrained models had much fewer samples (107 samples) and was not easy to be discriminated which might negatively affect the discriminative power of the learned emotional feature space. In the MELD, the pretrained five-class, four-class, and three-class tasks obtained 32.8%, 40.4% and 50.7% UAR values, respectively. New tasks for surprise, fear, and disgust verification performed slightly better using the four-class than the three-class prototype encoder. The difference in F1 score and UAR values on the three tasks between these two pretrained prototype encoders was less than 2%. The reduction in the pretrained emotion classes degrading the verification performances suggests the importance of constructing a pretrained emotion space using more emotion classes. However, the five-class pretrained model with only 24.5% F1 score resulted in much lower cross-task results than pretrained models using 4 basic emotion classes. The recall of the surprise class was only 26.7%. The results led to a conclusion that the pretrained emotional feature space would be poor if the discriminatory boundary between pretrained emotion classed was not learned effectively.

5.3 Effects on the Number of Enrolled Samples

In this section, we explored the effects on the number of enrolled samples. We chose the IEMOCAP dataset in this experiment because it performed well when given a sufficient number of validation and testing samples in each emotion class. In the previous setting described in Section 4.1, we enrolled 10 utterances for each emotion class in a new task, and these utterances were randomly selected from the validation set. In Fig. 4, we presented the UAR value of the binary frustration verification task using different enrolled utterances for each emotion class.

Enrolling more utterances should result in more accurate embedding descriptions for the prototype of the unseen frustration class. However, by increasing the number of enrolled utterances, it would bring about more cost and inconvenience in real-world applications. This experiment showed that enrolling around 10 utterances for each emotion class can achieve comparable performance with more utterances. The estimated total utterance length of the 10 enrolled utterances for each emotion class are around 25.5 seconds in IEMOCAP. If only enrolling, one utterance, the frustration verification result could still attain a 59.2% UAR. However, using very few utterances for enrollment is likely to cause significant variation. This analysis shows that representing an emotion class with a pre-trained prototype created by 10 enrolled utterances is adequate for robust cross-task emotion verification.

5.4 Effects on Label Confidence of Enrolled Samples

We investigated the verification results using enrolled samples with different label confidence. In the IEMOCAP dataset, each sample was labeled by at least 3 annotators and thus we could identify samples consistently annotated as the same labels to examine the effect on the verification results. We calculated a consistent ratio for each sample by the number of annotators labeling the same as the final label divided by the number of total annotators. The final label was determined by the majority vote and thus the possible ratio values for the most samples labeled by 3 or 4 annotators were 0.5, 0.67, 0.75, and 1. The samples with confident ratio smaller than 0.5 were not included in the training and evaluation. We set a strict rule to determine a confident group as the totally consistent samples (consistent ratio equals to 1) and a less confident group with other samples. We also examined the cases only enrolling confident neutral samples but less confident frustration samples (denoted as confident neutral in Fig. 5) and the cases enrolling confident frustration samples but less confident neutral samples (denoted as confident frustration in Fig. 5).

Fig. 5 shows the verification results using different label confidence group in the enrollment step. We observed that the less confident group obtained lower UAR (70.7%) than the confident group (72.1%). The confident neutral group achieved superior UAR (72.9%) than the confident group. In contrast, the confident frustration group even degraded than the less confident group. The results reflected that enrolling samples consistently labeled as neutral was more beneficial which could provide a powerful baseline prototype for the verification step. The frustration class was naturally easy to co-occurred with other emotions such as anger or sadness. The use of confident frustration but less confident neutral samples might result in confusion for the complex frustration samples. The neutral space enrolled by less confident samples also brought about negative effects.

6 Conclusion

This work proposes a novel and simple emotion verification approach to perform cross-task emotion recognition task with only a few enrollment samples. We demonstrated the promising accuracy on two different larget scale datasets, and these results are comparable to the scenario of recollecting and labeling the unseen emotion class. The negative margin also reflects a pivotal strategy to construct the emotion space for unseen classes. Several experimental results indicate that the pretrained encoder using a discriminative architecture and properly selected negative margin can benefit the cross-task performance. The verification can be applied to different new task definitions, and the relationships between emotion classes, such as the frustration class in the IEMOCAP dataset are more similar to the manifestation of the anger class than sadness class. In the analyses, we showed the effects of different modules in the framework, such as the loss margin, pretrained emotion classes, and enrollment samples.

Several future directions can be advanced to further generalize SER. Extending cross-task to cross-database scenarios allows the model to recognize unseen emotion categories collected in different data recording conditions. A variety of datasets can be explored for diverse applications on various emotion classes. The finding that pretrained feature space is vital for verification leads to a future direction to pretrain using different emotion classes. Learning a pretrained space using additional contextual factors and different advanced deep networks will increase model robustness and generalizability. Investigating multimodal signals using the enroll-to-verify approach is also a direction to improve recognition performance. We hope this work initiates a new line of research in the SER domain.

References

[1] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[2] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6474–6478.

[3] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct Modelling of Speech Emotion from Raw Speech," in *Proc. Interspeech 2019*, 2019, pp. 3920–3924.

[4] J.-L. Li, T.-Y. Huang, C.-M. Chang, and C.-C. Lee, "A waveform-feature dual branch acoustic embedding network for emotion recognition," *Frontiers in Computer Science*, vol. 2, p. 13, 2020.

[5] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[6] T. Dalgleish and M. Power, *Handbook of cognition and emotion*. John Wiley & Sons, 2000.

[7] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1652–1656.

[8] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In Defence of Metric Learning for Speaker Recognition," in *Proc. Interspeech 2020*, 2020, pp. 2977–2981.

[9] J.-L. Li and C.-C. Lee, "Attentive to individual: A multimodal emotion recognition network with personalized attention profile." in *Interspeech*, 2019, pp. 211–215.

[10] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7189–7193.

[11] R. Cai, K. Guo, B. Xu, X. Yang, and Z. Zhang, "Meta Multi-Task Learning for Speech Emotion Recognition," in *Proc. Interspeech 2020*, 2020, pp. 3336–3340.

[12] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Processing Letters*, vol. 28, pp. 1190–1194, 2021.

[13] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. W. Schuller, "Autonomous Emotion Learning in Speech: A View of Zero-Shot Speech Emotion Recognition," in *Proc. Interspeech 2019*, 2019, pp. 949–953.

[14] M. A. Jalal, R. K. Moore, and T. Hain, "Spatio-temporal context modelling for speech emotion classification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 853–859.

[15] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and 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)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246. [Online]. Available: https://aclanthology.org/P18-1208

[16] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–7.

[17] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7144–7148.

[18] J.-L. Li and C.-C. Lee, "Encoding Individual Acoustic Features Using Dyad-Augmented Deep Variational Representations for Dialog-level Emotion Recognition," in *Proc. Interspeech 2018*, 2018, pp. 3102–3106.

[19] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6685–6689.

[20] M. Jaiswal, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost, "MuSE: a multimodal dataset of stressed emotion," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1499–1510. [Online]. Available: https://aclanthology.org/2020.lrec-1.187

[21] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8384–8388.

[22] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Speech Emotion Recognition with Emotion-Pair Based Framework Considering Emotion Distribution Information in Dimensional Emotion Space," in *Proc. Interspeech 2017*, 2017, pp. 1238–1242.

[23] Z. Huang, W. Xue, Q. Mao, and Y. Zhan, "Unsupervised domain adaptation for speech emotion recognition using pcanet," *Multimedia Tools Appl.*, vol. 76, no. 5, p. 6785–6799, mar 2017. [Online]. Available: https://doi.org/10.1007/s11042-016-3354-x

[24] A. Marczewski, A. Veloso, and N. Ziviani, "Learning transferable features for speech emotion recognition," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, ser. Thematic Workshops '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 529–536. [Online]. Available: https://doi.org/10.1145/3126686.3126735

[25] p. song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 265–275, 2019.

[26] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards Speech Emotion Recognition "in the Wild" Using Aggregated Corpora and Deep Multi-Task Learning," in *Proc. Interspeech 2017*, 2017, pp. 1113–1117.

[27] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer Learning for Improving Speech Emotion Classification Accuracy," in *Proc. Interspeech 2018*, 2018, pp. 257–261.

[28] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[29] Y. Gao, J. Liu, L. Wang, and J. Dang, "Domain-adversarial autoencoder with attention based feature level fusion for speech emotion recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6314–6318.

[30] H. Zhou and K. Chen, "Transferable positive/negative speech emotion recognition via class-wise adversarial domain adaptation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3732–3736.

[31] Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, "Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition," *Speech Communication*, vol. 93, pp. 1–10, 2017.

[32] A. Schmitt, S. Ultes, and W. Minker, "A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 3369–3373.

[33] W. M. Kouw and M. Loog, "An introduction to domain adaptation and transfer learning," *arXiv preprint arXiv:1812.11806*, 2018.

[34] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[35] A. Naman and L. Mancini, "Fixed-maml for few shot classification in multilingual speech emotion recognition," *arXiv preprint arXiv:2101.01356*, 2021.

[36] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[37] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "A dialogical emotion decoder for speech emotion recognition in spoken dialog," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6479–6483.

[38] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[39] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2019.

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

[41] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

[42] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, pp. 1–22, 2004.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[44] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-based deep metric learning for speaker recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3652–3656.

[45] Y. Liu, L. He, and J. Liu, "Large Margin Softmax Loss for Speaker Verification," in *Proc. Interspeech 2019*, 2019, pp. 2873–2877.

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

[47] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[48] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, and H. Hu, "Negative margin matters: Understanding margin in few-shot classification," in *European Conference on Computer Vision*. Springer, 2020, pp. 438–455.

[49] C. Zhang, K. Koishida, and J. H. L. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, 2018.

[50] Z. Ren, Z. Chen, and S. Xu, "Triplet based embedding distance and similarity learning for text-independent speaker verification," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 558–562.

[51] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[52] J. Huang, J. Tao, B. Liu, and Z. Lian, "Learning Utterance-Level Representations with Label Smoothing for Speech Emotion Recognition," in *Proc. Interspeech 2020*, 2020, pp. 4079–4083. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1391

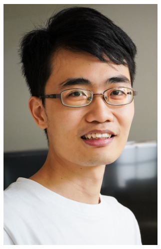
[53] Y.-L. Huang, B.-H. Su, Y.-W. P. Hong, and C.-C. Lee, "An Attribute-Aligned Strategy for Learning Speech Representation," in *Proc. Interspeech 2021*, 2021, pp. 1179–1183.

[54] S. T. Rajamani, K. T. Rajamani, A. Mallol-Ragolta, S. Liu, and B. Schuller, "A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6294–6298.

[55] W. Xie, A. Nagrani, J. S. Chung, and 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)*, 2019, pp. 5791–5795.

[56] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[57] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.



Jeng-Lin Li is currently a doctoral student. He received the B.S. degree in the Department of Electrical Engineering at National Tsing Hua University, Taiwan in 2016, and is directly pursuing PhD degree. He was awarded with NTHU Principal Outstanding Student Scholarship (2017 - 2020), Garmin Scholarship (2018), Yahoo Scholarship (2019), EMBC student travel grants(2019), Interspeech student travel grants(2019), and Novatek PhD Scholarship (2020, 2021). His research interests are behavior signal processing (BSP), affective computing, and health analytics. He is also a student member of ISCA and AAAC.

Chi-Chun Lee (M'13, SM'20) is an Associate Professor at the Department of Electrical Engineering of the National Tsing Hua University (NTHU), Taiwan. He received his B.S. and Ph.D. degree both in Electrical Engineering from the University of Southern California, USA in 2007 and 2012. His research interests are in speech and language, affective computing, health analytics, and behavioral signal processing. He is an associate editor for the IEEE Transaction on Affective Computing (2020-), the IEEE Transaction on Multimedia (2019-2020), the Journal of Computer Speech and Language (2021-), the APSIPA Transactions on Signal and Information Processing and a TPC member for APSIPA IVM and MLDA committee. He serves as an area chair for Interspeech 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity chair for ACM ICMI 2018, late breaking result chair for ACM ICMI 2023, sponsorship and special session chair for ISCSLP 2018, 2020. He is the recipient of the Foundation of Outstanding Scholar's Young Innovator Award (2020), the CIEE Outstanding Young Electrical Engineer Award (2020), the IICM K. T. Li Young Researcher Award (2020), the NTHU Industry Collaboration Excellence Award (2021), and the MOST Futuretek Breakthrough Award (2018, 2019). He led a team to the 1st place in Emotion Challenge in Interspeech 2009, and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in Interspeech 2019. He is a coauthor on the best paper award/finalist in Interspeech 2008, Interspeech 2010, IEEE EMBC 2018, Interspeech 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. He is also an ACM and ISCA member.