Expert Systems With Applications 237 (2024) 121368

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Semi-supervised cross-lingual speech emotion recognition

Mirko Agarla [a], Simone Bianco [a], Luigi Celona [a,*], Paolo Napoletano [a], Alexey Petrovsky [b], Flavio Piccoli [a], Raimondo Schettini [a], Ivan Shanin [c]

[a] *University of Milano - Bicocca, Milano, Italy*
[b] *Independent researcher*
[c] *Institute of Informatics Problems, FRC CSC RAS, Russia*

ARTICLE INFO

Keywords:
Semi-supervised domain adaptation
Speech emotion recognition
Cross-lingual
Semi-supervised learning

ABSTRACT

Performance in Speech Emotion Recognition (SER) on a single language has increased greatly in the last few years thanks to the use of deep learning techniques. However, cross-lingual SER remains a challenge in real-world applications due to two main factors: the first is the big gap among the source and the target domain distributions; the second factor is the major availability of unlabeled utterances in contrast to the labeled ones for the new language. Taking into account previous aspects, we propose a Semi-Supervised Learning (SSL) method for cross-lingual emotion recognition when only few labeled examples in the target domain (i.e. the new language) are available. Our method is based on a Transformer and it adapts to the new domain by exploiting a pseudo-labeling strategy on the unlabeled utterances. In particular, the use of a hard and soft pseudo-labels approach is investigated. We thoroughly evaluate the performance of the proposed method in a speaker-independent setup on both the source and the new language and show its robustness across five languages belonging to different linguistic strains. The experimental findings indicate that the unweighted accuracy is increased by an average of 40% compared to state-of-the-art methods.

1. Introduction

SER is a fundamental aspect of computational paralinguistics as it concerns the analysis of the non-verbal elements of speech (Schuller & Batliner, 2013). SER, which aims to infer the emotional state of a speaker (El Ayadi, Kamel, & Karray, 2011), could support a wide range of domains, including human–computer interaction (Schuller & Batliner, 2013), healthcare (Tumanova, Woods, & Wang, 2020), and public safety (Lefter & Jonker, 2017). For instance, SER systems could be employed in interactive dialogue systems to make them empathetic (Bertero et al., 2016), in healthcare systems for the diagnosis of disorders and diseases (Hansen et al., 2022), and in commercial applications for detecting customer satisfaction in call-centers and by employment agencies to find suitable candidates (Perez-Toro, Vasquez-Correa, Bocklet, Noth, & Orozco-Arroyave, 2021). Existing SER models have achieved satisfactory results for valence/arousal estimation (Xiao, Wu, Zhang, & Tao, 2016) and emotion classification (Scheidwasser-Clow, Kegler, Beckmann, & Cernak, 2022) when the training and test data are from the same corpus. However, the performance of these models degrades when applied to new corpora of same/different languages due to domain shift (Feraru, Schuller, et al., 2015). This problem occurs mainly in real-world scenarios where the people using a given SER system may differ or speak languages other than those used to train the system.

Over the years, several methodologies have been developed to speed up the adaptation of a pre-trained system to new people or a new language by leveraging semi-supervised/incremental learning (Zhang et al., 2016) and transfer learning (Feraru et al., 2015). Numerous approaches have been proposed to reduce the domain shift problem for cross-corpus or cross-lingual SER, namely eliminate or reduce the difference between the source and target data distribution (Cai et al., 2021; Tamulevičius et al., 2020). Most of these approaches are based on deep learning techniques as they generally prove to be more effective than traditional machine learning techniques also for SER (Tamulevičius et al., 2020). Supervised Domain Adaptation (SDA) methods for SER exploit labeled utterances of the target corpus to adapt the recognition model to work properly on the new set of data (Neumann et al., 2018; Tamulevičius et al., 2020; Zhou & Chen, 2019). However, these methods require the new language utterances to be labeled, which may not be possible as their collection is expensive. Therefore, a more

[*] Corresponding author.
E-mail addresses: m.agarla@campus.unimib.it (M. Agarla), simone.bianco@unimib.it (S. Bianco), luigi.celona@unimib.it (L. Celona), paolo.napoletano@unimib.it (P. Napoletano), alexey@petrovsky.eu (A. Petrovsky), flavio.piccoli@unimib.it (F. Piccoli), raimondo.schettini@unimib.it (R. Schettini), ivan.shanin@gmail.com (I. Shanin).

https://doi.org/10.1016/j.eswa.2023.121368
Received 13 October 2022; Received in revised form 14 July 2023; Accepted 27 August 2023
Available online 3 September 2023

practical solution is Unsupervised Domain Adaptation (UDA) which only demands unlabeled utterances from the new language. Many UDA methods try to reduce the distribution shift between the source and target languages (Cai et al., 2021; Latif, Qadir, & Bilal, 2019; Li, Yan, & Wang, 2021; Ocquaye, Mao, Xue, & Song, 2021).

In this paper we formulate the cross-lingual SER as a SSL problem. This scenario assumes that for the new language there are few labeled and many unlabeled utterances. We first train a deep learning based SER model on the source language dataset in which all utterances are annotated with the emotion label (see Fig. 1(a)). The SER model is then adapted to a new language for which the emotion of most training utterances is unknown. The labeled data of the first language are available (see Fig. 1(b)). Pseudo-labeling is adopted to generate labels for the unlabeled utterances and guide the learning process. Unlike most cross-lingual SER methods which focus on the binary classification of valence, our approach deals with the prediction of five emotion categories. In our experiments we consider English as the source language since it is the most widespread language in the world (Berlitz, 2021).

The proposed method for cross-lingual SER based on pseudo-labeling is suitable for use in all-day consumer technologies, such as smartphones, smartmirrors, and smartwatches. These devices collect massive amounts of unlabeled data, making traditional supervised learning methods difficult to implement. The proposed method overcomes this challenge by requiring only small amounts of labeled data and large amounts of unlabeled data, lowering manual annotation costs and shortening data preparation time. As a result, the method can be used in consumer technologies at a much lower cost than traditional supervised learning-based methods, making it a more practical and accessible solution. By implementing this method, consumer technologies can accurately recognize and respond to emotions expressed in different languages, improving communication and user experience. For example, a smartmirror with cross-lingual SER could provide personalized recommendations based on a user's emotional state, or could adjust lighting and temperature to create a more comfortable environment based on the user's emotions.

Apart from a method for cross-lingual SER, this work provides an analysis of different models as utterance encoder. In particular, it is demonstrated that a Transformer-based utterance encoder trained to build meaningful representations of speech boosts the performance compared to state of the art methods. Furthermore, in the adaptation procedure it is verified that balancing pseudo-labeled vs. labeled utterances helps to improve the generalization capabilities of the learned model.

To summarize, the main contributions of this paper are:

- A cross-lingual SER framework spanning five languages.
- A SSL based cross-lingual SER method for emotion categorization.
- The experimentation of several utterance encoders, i.e. a CNN for speech emotion classification, a CNN and a Transformer trained for speech representation learning.
- Two different approaches for generating pseudo-labels are investigated.
- An utterance rebalancing strategy to reduce the cardinality gap between the labeled utterances available for the source language and the labeled or pseudo-labeled utterances for the new language.
- A thorough analysis of how the variation in the number of labeled utterances for the new language impacts performance.

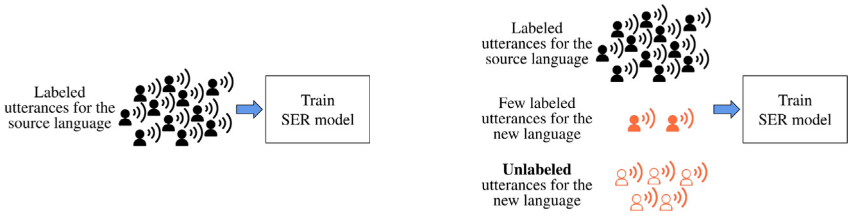
The rest of the paper is organized as follows. Section 2 introduces some previous works on cross-lingual emotion recognition. In Section 3, SSL based cross-lingual SER is formalized and then the proposed method is described. Experimental setup and result analysis are presented in Sections 4 and 5, respectively. Finally, we conclude in Section 6.

2. Related work

Cross-lingual SER methods involve the use of two languages, the source language for which the emotion information is available for all the samples and the target language, for which only few labeled samples are available. The aim of cross-lingual SER is then to learn from the source language and extend the learned knowledge on the unlabeled samples of the target language. This propagation is an active process that involves the use of the few samples available and, if present, of auxiliary information available for both languages. At each learning episode, unlabeled samples of the target language which are believed to belong to an emotional class, are labeled. This process is called pseudo-labeling and in turn will support future learning episodes for evaluating the remaining samples. The choice of the acoustic features is quite important and must be kept into account (Tamulevičius et al., 2020).

The primary taxonomy of cross-lingual SER methods is given by the strategy used to convey knowledge on the new language. Two are the main schemes used in the state of the art. The first one is the use of auxiliary information that is available on both languages to learn a shared feature space. In this context, the emotion recognition task and the side task are performed simultaneously through a joint-training. Cai et al. (Cai et al., 2021) proposed a neural network with two branches, leading respectively to emotion and language classification. During training, the first branch is trained only with the source language corpus with emotion labels while the second by both languages. This training schema allows to exploit all existing information of the first and the second language to create a shared feature space. Gradient Reversal Layer (GRL) is adopted to force the features to be meaningful for the primary task (emotion classification) and at the same time to be indistinguishable for the auxiliary task (language classification). Performance is measured in terms of valence and arousal on Urdu, Estonian, IEMOCAP, Persian, and German (4 emotions). Li et al. (Li et al., 2021) proposed an SSDA memory-based system called Neural Network with Pseudo Multilabel (NNPM). In first place, they use a siamese network with self attention for projecting source and target utterances in a learned feature space. Then, the source-domain features are dynamically stored in the dynamic external memory. Emotion similarity is gained through cosine distance between features in the memory and of the target utterance. Pseudo-labels are given to the target domain utterances based on the similarity score. Hard negative sample mining strategy is used to improve the learning whereas the features result less representative. Performance is measured with weighted and unweighted accuracy. Ocquaye et al. (Ocquaye et al., 2021) exploited joint training to perform SSDA. A neural network with one common branch and a set of task-specific branches is proposed. Two branches perform emotion recognition respectively on the source samples and the pseudo-labeled target samples. The evaluation has been conducted on SAVEE, IEMOCAP, EMO-DB, FAU-AIBO (German), and EMOVO for valence classification.

The second strategy widely used in the context of cross-lingual SER is the use of adversarial training. This technology showed great ability in domain transfer and thus is very effective in this scenario. Latif et al. (Latif et al., 2019) proposed a method for learning a language-independent emotion recognition feature vector in the context of UDA. This system is based on Generative Adversarial Networks (GANs) as this technology has shown great potential in learning the underlying data distribution. Specifically, the proposed method has two generators to project respectively the source and the target utterance in a common feature space that is later evaluated through a critic. The feature space is then constrained to carry emotion information through a classification loss. In other words, the adversarial loss makes the feature space homogeneous both for the source and the target languages while the classification loss makes the features meaningful for the task of emotion recognition. Performance assessed on EMOVO, SAVEE, Urdu, and EMO-DB for valence classification.


Labeled utterances for the source language
Train SER model
Labeled utterances for the source language
Few labeled utterances for the new language
Unlabeled utterances for the new language
Train SER model

(a) Single-language SER

(b) Train on a new language

Fig. 1. Pipeline of the proposed cross-lingual SER method. (a) The model is first trained on the source language and (b) it is then adapted to the new language for which a few labeled utterances are available.

Domain Adversarial Neural Network (DANN) is a method that generates domain invariant feature representations. This allows to reduce the gap among source and target domain features (Abdelwahab & Busso, 2018).

However, the effectiveness of domain adversarial training strongly depends on the distribution of the two databases: in fact, adversarial attacks and instabilities may occur in the training phase if the data points are significantly different from each other. Aggregate multi-task Learning (AL) is another technique that has been used to improve the generalization of the trained model by incorporating information of gender and naturalness (Kim, Englebienne, Truong, & Evers, 2017).

Extending the work of Sung et al. (Sung et al., 2018), Ahn et al. (Ahn, Lee, & Shin, 2021) presented Few-shot Learning and Unsupervised Domain Adaptation (FLUDA), which aims to train an embedding and a metric module that respectively project the utterances in a meaningful shared feature space and learn the differences among classes. The embedding and metric module are optimized to predict class similarity for each episode by exploiting few samples composing the support set and pseudo-labels assigned in the previous episode. During training, an auxiliary module is used to determine whether the labeled sample is real or pseudo-labeled. The proposed method estimates four categorical emotions (neutral, happy, sad and angry) and uses IEMOCAP and CREMA-D as source corpora while MSP-IMPROV, EMO-DB or KME were used as target corpus. However, the samples in few-shot learning significantly depend on the choice of the support set, that can make its application challenging to a practical setup. Furthermore, the strong assumption that the support set is uniformly sampled from a single distribution, leads to the selection of an unstable number of samples for each class during training. With the aim of solving the previous challenges, Zhou et al. (Zhou & Chen, 2019) used adversarial network to perform SSDA. Specifically, a GAN is modified such that the generator projects the utterance in a feature space carrying emotion information and the critic has to determine which emotion class belongs the input feature and the used language. This stage is trained with the source language for which the labels are known. In second place the critic is frozen and the encoder is adapted to the new language. This second step forces the encoder to adapt and generate compatible features with the ones obtained in the first step. The method has been benchmarked on EMO-DB and Aibo in terms of positive and negative emotions.

Recently, Das et al. (Das, Lønfeldt, Pagsberg, & Clemmensen, 2022) presented a Variational AutoEncoder (VAE) for learning a latent space able to discriminate emotions and to generalize on different languages simultaneously. They achieved this goal by (i) exploiting a Kullback–Leibler (KL) loss annealing using cyclic scheduling to improve emotion discrimination, (ii) employing semi-supervised training of the VAE by incorporating a clustering loss in the learning function. Experimental results have been collected for IEMOCAP, SAVEE, EMO-DB, CaFE, and AESD in terms of four emotions. Kshirsagar et al. (Kshirsagar & Falk, 2022) explored the combined use of Bag-of-Words (BoW) methodology, domain adaptation and data augmentation as strategies to counter the damaging effects of cross-lingual SER. The authors also proposed a new method called N-CORAL in which all languages are mapped to a common distribution. Experiments with the German, Hungarian, Chinese, and French languages show the advantages of the proposed N-CORAL method, combined with data augmentation and BoW for valence-arousal estimation.

The related works can be thus summarized as follows:

- Several studies in the literature demonstrate that it is possible to perform cross-language SER without labeled utterances for the new language using auxiliary information, generative methods, or adversarial and few-shot learning.
- Domain adaptation approaches based on adversarial neural networks are widely used for cross-lingual SER; however, there is still room for performance improvement.
- The use of the newer Transformer architectures for utterance encoding has not been explored and used for cross-lingual SER.

3. Method

The problem of cross-lingual SER is first formulated in Section 3.1 and then the proposed SSL method for cross-lingual SER is described in Section 3.2.

3.1. Problem formulation

We represent sets with special Latin characters (e.g., $\mathcal{S}$). Lower or uppercase normal fonts, e.g., K denote scalars. Matrices are in uppercase bold letters (e.g., $\mathbf{M}$), while lowercase bold letters represent vectors as in $\mathbf{v}$. We use lowercase Latin letters to represent indices (e.g., i).

We formulate our cross-lingual SER as the following domain adaptation task. We have a source language corpus with N_s labeled utterances as source domain, $\mathcal{D}_s = \{(\mathbf{X}_i^s, y_i^s)\}_{i=1}^{N_s}$, and a new language corpus, $\mathcal{D}_t$, as target domain. The new language corpus, $\mathcal{D}_t = \{\mathcal{U}_t \cup \mathcal{K}_t\}$, consists of a set N_u of unlabeled utterances $\mathcal{U}_t = \{\mathbf{X}_i^t\}_{i=1}^{N_u}$, and a set N_k of labeled utterances $\mathcal{K}_t = \{\mathbf{X}_i^t, y_i^t\}_{i=1}^{N_k}$. The number of utterances of the source language N_s is much higher than the number of labeled utterances for the new language N_k, i.e. $N_s \gg N_k$. Utterances $\mathbf{X}_i^s$ and $\mathbf{X}_i^t$ are elements of $\mathbb{R}^{F\times T}$, where F and T are the number of frequency bins and the number of time frames, respectively. The utterance labels y_i^s and y_i^t are scalar values such that $y \in \mathbb{Z} : 1 \leq y \leq C$ where C is the number of emotion categories within the corpora. We consider that the source and target corpora contain the same number C of emotion categories.

Our goal is to learn a reliable emotion classifier on $\mathcal{D}_s$, $\mathcal{U}_t$, and $\mathcal{K}_t$, which preserves performance on source language $\mathcal{D}_s$ and generalizes well on $\mathcal{D}_t$.

3.2. SSL for cross-lingual SER

The proposed semi-supervised cross-lingual SER is a deep learning model f parameterized with θ that maps an input utterance $\mathbf{X}$ into a basic emotion y, $y = f(\mathbf{X}, \theta)$.

As depicted in Fig. 2, our method consists of two modules, namely the *SER recognition backbone* and the *adaptation module*. The SER recognition backbone (f_θ) classifies emotions for both the source and new

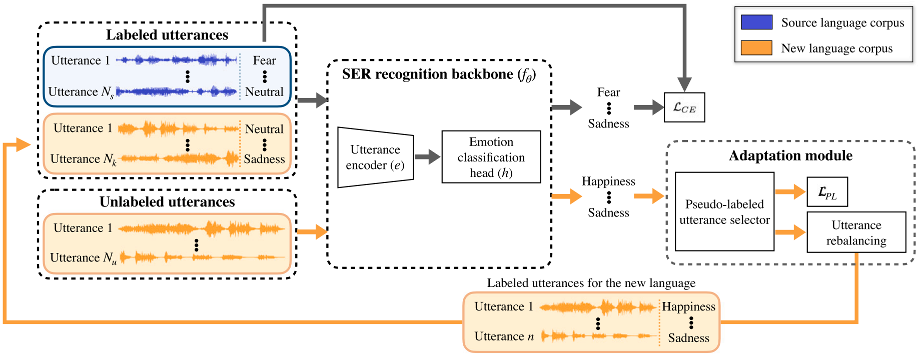


Fig. 2. The pipeline of our method.

language utterances. During the training phase, we introduce the adaptation module to improve the discriminative power and generalization ability of f on the new language. The adaptation module relies on a pseudo-labeling strategy to allow model training on the unlabeled utterances of the new language. Furthermore, we include an utterance rebalancing mechanism to avoid that the model is biased on the source language due to the higher cardinality of utterances compared to those of the new language.

In the next sections we detail the previously introduced modules.

3.2.1. SER recognition backbone

It is the core of our method that deals with utterance classification. It consists of two modules, i.e., the utterance encoder e and the classification head h. The utterance encoder e is a deep architecture like a CNN or a Transformer that takes a raw waveform or an audio representation as input, $\mathbf{X}$, and returns a d-dimensional feature vector $\mathbf{x} \in \mathbb{R}^D$. The choice of the utterance encoder is important for the performance of the proposed method. Therefore, three different deep architectures are considered, namely EmotionCNN (Tamulevičius et al., 2020), Bootstrapping Your Own Latent for Speech (BYOL-S) (Scheidwasser-Clow et al., 2022), and Hidden-unit BERT (HuBERT) (Hsu et al., 2021). **EmotionCNN** is a CNN architecture for cross-lingual speech emotion recognition. It is composed of three convolutional layers. Each convolutional layer is followed by a ReLU, a batch normalization layer, and 3 × 3 max pooling, respectively. The model is fed with a cochleagram-based representation of the raw waveform and outputs a 128-dimensional feature vector. **BYOL-S** is a CNN model for audio representation inspired by the Bootstrapping Your Own Latent (BYOL) model initially proposed for self-supervised image classification (Grill et al., 2020). It is trained on the speech utterances of the AudioSet (Gemmeke et al., 2017) dataset. BYOL-S is currently the state of the art in SER (Scheidwasser-Clow et al., 2022). The model accepts input utterances of variable length and returns a single 1024-dimensional feature vector per input utterance. All utterances are converted to a log-scaled Mel spectrogram with a window size of 64 ms, hop size of 10 ms, and mel-spaced frequency bins $F = 64$ in the range 60–7800 Hz. Each spectrogram is normalized by subtracting the mean and dividing by the estimated standard deviation for the frames of the spectrogram. **HuBERT** is a Transformer-based approach for self-supervised speech representation learning. It consists of a convolutional waveform encoder, a BERT-Base encoder (Devlin, Chang, Lee, & Toutanova, 2019), a projection layer, and a code embedding layer. It is trained on the Librispeech 960h dataset (Panayotov, Chen, Povey, & Khudanpur, 2015) to classify randomly masked frames to pseudo-labels. The labels are generated by running K-Means clustering with 100 clusters on 39-dimensional MFCC features. The model accepts raw waveforms of variable length as input and returns a single 768-dimensional feature vector per input utterance. The main characteristics of the three architectures are summarized in Table 1.

The feature vector obtained from one of the previously described utterance encoders is processed by the classification head h to predict $\mathbf{p} \in \mathbb{R}^C$, i.e., the probability distribution over the C emotion categories. The classification head consists of a linear layer followed by a softmax:

$$\mathbf{p} = \mathrm{Softmax}(h(\mathbf{x}, \theta_h)), \tag{1}$$

where θ_h is the set of weights $\mathbf{W} \in \mathbb{R}^{D\times C}$ and bias $\mathbf{b} \in \mathbb{R}^C$.

3.2.2. Adaptation module

This module aims to exploit the unlabeled utterances of the new language in the model training. To this purpose, an SSL approach is used to generate pseudo-labels $\tilde{y}$ for the unlabeled utterances of the new language dataset, $\mathcal{D}_t$. This operation is repeated at each step to take advantage of the knowledge learned in previous steps and results in the creation of a hybrid dataset $\tilde{\mathcal{D}}$ composed by real samples $(\mathbf{X}_i, y_i)$ and samples with generated ground truth $(\mathbf{X}_i, \tilde{y}_i)$.

A key decision is how to generate the pseudo-labels $\tilde{y}$ for the N_u unlabeled utterances. In this paper, we experiment with the use of hard pseudo-labels and soft pseudo-labels.

Hard pseudo-labels. In this approach hard pseudo-labels are directly obtained from network predictions. Let $\mathbf{p}_i$ be the probability outputs of our trained h_θ model for the utterance. Using the probability vector, the pseudo-label for the utterance $\mathbf{X}_i$ corresponds to $\tilde{y}_i = \arg\max(\mathbf{p}_i)$. We select the subset of pseudo-labels which are less noisy to limit the confirmation bias, i.e. the overfitting of incorrect pseudo-labels predicted by the model. In particular, we select only the pseudo-labels corresponding to high-confidence predictions. Let $\mathbf{g} = \{g_b\}_{b=1}^B$ be a binary vector representing the selected pseudo-labels in a mini-batch B. This vector is obtained as follows:

$$g_b = \mathbb{1}\left[\max(p_b) \geq \tau\right], \tag{2}$$

where $\tau \in (0, 1)$ is a confidence threshold.

Soft pseudo-labels. We investigate the use of soft pseudo-label inspired by Arazo, Ortego, Albert, O'Connor, and McGuinness (2020) since it has demonstrated in some cases to perform better than hard pseudo-labels (Tanaka, Ikami, Yamasaki, & Aizawa, 2018).

Let $\mathbf{p}_i$ be the probability outputs of our trained h_θ model for the utterance $\mathbf{X}_i$. Two regularization terms are used to improve convergence. The first regularization term discourages the model to assign all samples to a single class by adding:

$$R_A = \sum_{c=1}^{C} \mathbf{p}_c \log\left(\frac{\mathbf{p}_c}{\bar{\mathbf{h}}_c}\right), \tag{3}$$

where $\mathbf{p}_c$ is the prior probability distribution for class c assumed as a uniform distribution $\mathbf{p}_c = 1/C$ and $\bar{\mathbf{h}}_c$ denotes the mean softmax probability of the model for class c across batch utterances.

Table 1
Utterance encoder architectures considered within the proposed method.

	EmotionCNN (Tamulevičius et al., 2020)	BYOL-S (Scheidwasser-Clow et al., 2022)	HuBERT (Hsu et al., 2021)
Input	Cochleagrams	MFCC	Waveform
Architecture	CNN	CNN	Transformer
Feature vector dim.	128	1024	768
Pretraining	–	AudioSet	Librispeech 960h
ML paradigm	–	self-supervised	self-supervised
Num. of params	35,584	1.6M	95M

The second regularization is the average per-sample entropy (R_H stands for entropy regularization) that forces the probability distribution to peak on a single class:

$$R_H = -\sum_{i=1}^{B}\sum_{c=1}^{C} h_\theta^c(\mathbf{X}_i)\log(h_\theta^c(\mathbf{X}_i)), \tag{4}$$

where $h_\theta^c(\mathbf{X}_i)$ denotes the c class value of the softmax output $h_\theta(\mathbf{X}_i)$ and it is estimated on a mini-batch B. The total loss is the following:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_A R_A + \lambda_H R_H, \tag{5}$$

where λ_A and λ_H control the contribution of each regularization term.

To limit the confirmation bias problem, we exploit the mixup data augmentation technique proposed in Zhang, Cisse, Dauphin, and Lopez-Paz (2018). It combines data augmentation with label smoothing to reduce the confidence of the model on its predictions. Mixup trains on convex combinations of sample pairs ($\mathbf{X}_p$ and $\mathbf{X}_q$) and corresponding labels (y_p and y_q):

$$x = \alpha\mathbf{X}_p + (1-\alpha)\mathbf{X}_q, \tag{6}$$

$$y = \alpha y_p + (1-\alpha)y_q, \tag{7}$$

where α is randomly sampled in the range $(0, 1)$.

3.2.3. Utterance rebalancing

As stated in Section 3.1, the number of labeled utterances for the new language N_k is much lower than the number of labeled utterances for the source language N_s. The use of pseudo-labeling to adapt the model to the new language can only partially reduce the imbalance between source and new language corpus. The imbalanced ratio γ_l between the number of source utterances, N_s, and the number of labeled utterances for the new language, N_k, is defined as N_k/N_s and a γ_l far from 1 indicates more severe utterance imbalance.

To tackle the utterance imbalance and get $\gamma_l = 1$ we exploit random oversampling. Specifically, the utterances of the new language are randomly replicated to match the number of utterances of the source language.

The rebalancing algorithm is run at the end of each pseudo-labeling iteration during the adaptation procedure. Furthermore, it is performed at the beginning of the training procedure if the number of labeled utterances for the new language is non-zero.

3.2.4. Training procedure

The proposed model $f_\theta(\mathbf{X})$ is trained end-to-end for E epochs using a set of N training utterances belonging to the joint domain $\mathcal{D} = \{\mathcal{D}_s \cup \mathcal{K}_t\}$. $\mathcal{D}$ contains a set of $N_s + N_k$ labeled utterances $\{\mathcal{D}_s, \mathcal{K}_t\}$ coming from both the source and the target language. Every W epochs, the model adaptation procedure is performed, in which supervised learning is accompanied by pseudo-labeling on the set $\mathcal{U}_t$ of unlabeled utterances for the new language. The set $\mathcal{D}$ is then expanded with the generated pseudo-labels $\mathcal{K}'_t\}$. The complete training procedure is presented in Algorithm 1.

The parameters θ of the model are optimized using categorical cross-entropy:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{B} y_i \log(f_\theta(\mathbf{X}_i)), \tag{8}$$

where $f_\theta(x)$ are the softmax probabilities predicted by the model, $\log(\cdot)$ is applied element-wise and y_i can be a real or pseudo label, and B is the number of batch utterances.

To mitigate the risk of overfitting, early stopping is implemented by selecting the model weights from the epoch that achieves the best performance on the validation set for both the source and new languages.

Algorithm 1 Our training procedure.

```
Input: Total training epochs $E$, interval of epochs for pseudo-labeling $W$, source language corpus $\mathcal{D}_s$, labeled utterances for the new language $\mathcal{K}_t$, unlabeled utterances for the new language $\mathcal{U}_t$.
Initialize the model $f_\theta$.
Initialize labeled corpus $\mathcal{D} = \mathcal{D}_s \cup \mathcal{K}_t$.
for e = 1 to $E$ do
    Train and update $f_\theta$ on $\mathcal{D}$.
    if mod(e, $W$) = 0 then
        Generate pseudo-labels for $\mathcal{U}_t$ using $f_\theta$.
        Form $\mathcal{K}'_t$ by applying pseudo-label policy on $\mathcal{U}_t$.
        Expand labeled set by $\mathcal{D} = \mathcal{D} \cup \mathcal{K}'_t$.
    end if
end for
Return: $f_\theta$
```

4. Experiments

In this section the datasets considered for experiments and the experimental setup are presented.

4.1. Datasets

A summary of the datasets used for our experiments is presented in Table 2. We consider seven speech emotion classification datasets in five languages: three in English (RAVDESS, SAVEE and TESS), one in French (CaFE), German (EmoDB), Italian (EMOVO), and Persian (ShEMO). In each dataset, speech samples have three attributes: audio data (i.e., the raw waveform, in mono), speaker identifier, and emotion label (e.g., angry, happy, sad). The datasets comprise scripted and acted utterances and vary in size (i.e., number of utterances), number of speakers, sample rate, and number of classes. All of them comprise utterances in which the speaker acts a specific emotion.

The considered datasets share the same five primary emotions, which are anger (A), fear (F), happiness (H), neutral (N), and sadness (S). Following Tamulevičius et al. (2020), in this study we consider only utterances annotated with one of the previous five emotions and discard the remaining utterances. Thus, the number of emotion categories is $C = 5$. TESS, SAVEE, and RAVDESS datasets are merged to obtain a large English language dataset. A summary of the distributions of utterances by emotion for each language is shown in the Table 3.

4.2. Experimental setup

Each dataset was split into training, validation, and testing sets to respectively train, optimize and evaluate task-specific emotion speech

Table 2
List of considered datasets for speaker emotion recognition.

Name	Spkrs	Emot.	SR (Hz)	Utter.	Lang.	Avg. dur. (s)	Tot. dur. (h)
CaFE (Gournay, Lahaie, & Lefebvre, 2018)	12	7	48,000	864	French	4.5	1.1
EMO-DB (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005)	10	7	16,000	535	German	2.8	0.4
EMOVO (Costantini, Iaderola, Paoloni, & Todisco, 2014)	6	7	48,000	588	Italian	3.1	0.5
RAVDESS (Livingstone & Russo, 2018)	24	8	48,000	1,440	English	3.7	1.5
SAVEE (Wang, 2010)	4	7	44,100	480	English	3.8	0.5
ShEMO (Nezami, Lou, & Karami, 2019)	87	6	44,100	3,000	Persian	4.0	3.3
TESS (Pichora-Fuller & Dupuis, 2020)	2	7	24,414	2,800	English	2.0	1.6

Table 3
Utterance distribution of the selected datasets across emotion classes.

	A	F	H	N	S	Total
English (RAVDESS, SAVEE, TESS)	652	652	652	652	616	3224
French (CaFE)	144	144	144	144	72	648
German (EMO-DB)	127	69	71	79	62	408
Italian (EMOVO)	84	84	84	84	84	420
Persian (ShEMO)	1059	38	201	1028	449	2775

classifiers. Following SERAB (Scheidwasser-Clow et al., 2022), each dataset is split into 60% training, 20% validation, and 20% testing sets. Each data partition is speaker-independent, i.e., the sets of speakers included in each part are mutually disjoint.

Importantly, the utterances of the various datasets have different sampling rates, for this reason they were all resampled to 16 kHz before any processing. During training a sequence of 2 seconds is randomly sampled from the whole utterance for augmentation, while the whole utterance is used at testing time. Voice Activity Detection (VAD) is not used in training and testing. The linear layer for emotion speech classification is randomly initialized.

All experiments are run three times with different random seeds, and the unweighted accuracy is chosen as our evaluation criterion.

4.2.1. Hyperparameters

In our experiments, we train the model for a total of 100 epochs using the Adam optimizer with an initial learning rate equal to 1×10^{-3} which decays by a factor of 0.95 every 10 epochs, a batch of 100 utterances, and exponential decay rates β_1 and β_2 equal to 0.9 and 0.999. The pseudo-labeling procedure is executed every 30 epochs, i.e. $W = 30$. We experiment with different values of τ (see Section 5.6.2 for a study of this hyperparameter) which lead to the choice of $\tau = 0.50$, but do not attempt a careful adjustment of the regularization weights λ_A and λ_H and simply set them to 0.8 and 0.4 as done in Tanaka et al. (2018).

5. Results

In this section the results achieved for different configurations are described. In all the experiments we consider English as the first language while the other languages were chosen one at a time as the second language.

5.1. Cross-lingual results

In this section the performance obtainable by our method on a totally unknown new language is measured. These results give an idea of the worst accuracy, defined as the lower bound, achievable for cross-lingual SER. To this end, experiments using only the training set of the source language (i.e. English) and testing on the new language data are performed. Table 4 reports accuracy results achieved by the three utterance encoders for the four new languages. As it is possible to see, HuBERT achieves the best accuracy for all the languages, while the EmotionCNN obtains the worst performance. Regarding HuBERT, the highest accuracy equal to 78.87% is achieved for the German language while the lowest accuracy of 33.24% is obtained for Persian. The same gap is registered for all utterance encoders and depends on the fact that as stated by several linguistic distances (Chiswick & Miller, 2005; Gamallo, Pichel, & Alegria, 2017; Petroni & Serva, 2008) Persian belongs to a linguistic strain very different from the English one. Therefore, it is conceivable that learning on utterance in Persian will produce an important gain in performance.

Table 4
SER accuracy on new languages by training only on the source language (English). Results are reported for three different versions of the proposed model in which a different utterance encoder is exploited. Best result for each utterance encoder is in **bold**.

	Accuracy (%)		
	EmotionCNN	BYOL-S	HuBERT
French	29.63	54.32	**58.02**
German	29.58	60.56	**78.87**
Italian	26.43	49.28	**52.14**
Persian	15.15	18.28	**33.24**

5.2. Multi-lingual results

In this section, multi-lingual SER experiments are performed following a supervised training that uses all data of the source and the new language. The results obtained give the upper bound for cross-lingual SER, i.e. the best accuracy obtainable having all the labels of the new language available. Since we want to evaluate whether the SER classifier generalizes on the new language but also if it preserves the performance on the source language. Table 5 shows the accuracy on the source language (i.e. first column of the table) and the accuracy on the new language (i.e. second column of the table) achieved by the three utterance encoders.

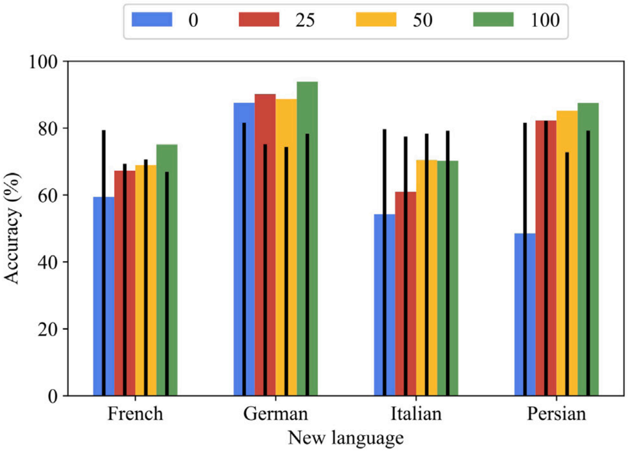
From the results it is possible to make various considerations. First, as expected, there is an increase in performance on the new language by using the training data of both the source language and the new one. This result is particularly evident for Persian, where the performance increases of about 50% for all the utterance encoders. Second, EmotionCNN achieves significantly lower performance than the other two encoders, i.e. about 20% less than BYOL-S and 30% less than HuBERT. Third, the performance on English, which is the source language, does not degrade by adding the training data of the new language. For HuBERT, which confirms itself as the best encoder, starting from the 81.50% of accuracy obtained by training only on English, we obtain a loss of about 3% for both French and Persian. On the other hand, for Italian and German, the variation in accuracy is less than 1%.

5.3. Pseudo-labeling results

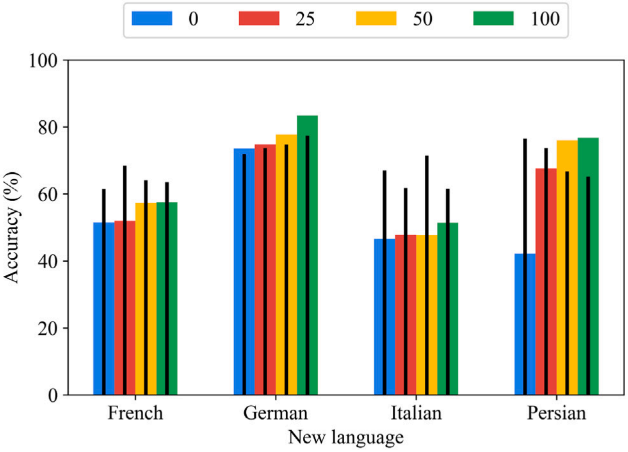
In this subsection, the results of using the proposed method with HuBERT for cross-lingual SER are presented. Performance is reported while changing the number of labeled utterances for the new language. The numbers of labeled utterances considered for the new language are 0, 25, 50, and 100, while all the utterances for the source language are labeled (i.e. 1682 utterances). From the previous numbers we obtain the imbalanced ratios, γ_l, equal to 0, 0.01, 0.03 and 0.06. Fig. 3 shows the performance achieved by using hard pseudo-labels (see Fig. 3(a))

Table 5
SER accuracy on new languages by training on the combination of source and new language training sets. Results are reported for three different versions of the proposed model in which a different utterance encoder is exploited. Best result for each utterance encoder is in **bold**.

	Accuracy on the source language (%)			Accuracy on the new language (%)		
	EmotionCNN	BYOL-S	HuBERT	EmotionCNN	BYOL-S	HuBERT
English	58.50	79.13	**81.50**	–	–	–
English & French	57.32	60.08	**78.19**	39.63	58.02	**77.78**
English & German	52.68	75.04	**82.44**	54.93	73.24	**94.37**
English & Italian	53.31	73.78	**81.02**	32.86	59.28	**74.29**
English & Persian	51.57	71.26	**78.50**	62.02	86.43	**92.24**



(a) Results for hard pseudo-labels



(b) Results for soft pseudo-labels

Fig. 3. SER accuracy by varying the number of labeled utterances for the new language and using (a) hard pseudo-labels and (b) soft pseudo-labels for the unlabeled utterances. Black bars indicate the accuracy on the source language (i.e. English).

and soft pseudo-labels (see Fig. 3(b)). Each colored bar represents the accuracy of a given pseudo-label approach with a given number of available labels for the new language. Black narrow bars represent the accuracy on the source language.

Several considerations can be made. First, overall hard pseudo-labels obtain better performance than soft pseudo-labels. Second, soft pseudo-labels not only result in worse performance than hard pseudo-labels, but also cause significant performance degradation on the source language. This behavior can be due to the effect of mixup augmentation which results in too noisy pseudo-labels that do not allow the model to converge properly. Third, for both pseudo-label approaches and for all languages, having more available labels for the new language allows to achieve higher accuracy. The pseudo-label approaches cannot reach the upper bound in any language but in any case they manage to improve the performance of the lower bound (for the Persian language it is even possible to have an increase of about 60%).

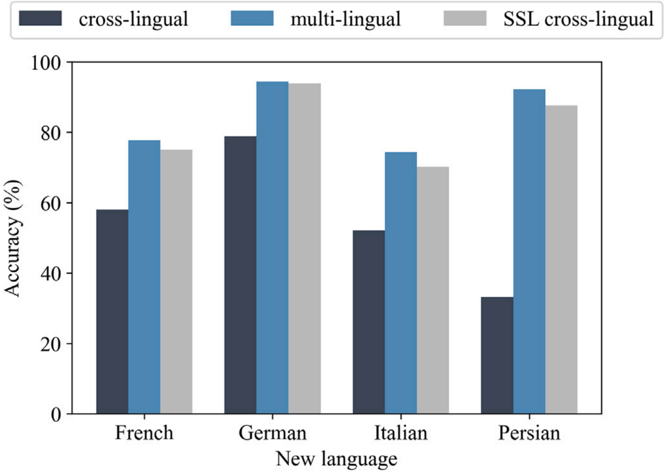


Fig. 4. Comparison of the results on new languages between the cross-lingual, multi-lingual and the best SSL (hard pseudo-labels considering 100 labeled utterances).

Fig. 5 illustrates the accuracy trends of the SER model with HuBERT and 100 labeled utterances across training epochs, depicting the performance of the training and validation sets. Each plot showcases accuracy curves for both the source and new languages. Despite having disjoint sets of speakers, the accuracy remains relatively consistent across the train and validation sets. However, the French and Italian languages, identified as the weakest performers, exhibit a tendency to overfit the model.

5.4. Discussion

Fig. 4 summarizes the results achieved by our best method with HuBERT in the different configurations evaluated for the recognition of emotions on new languages. The aim is to highlight the gap between training the method with all the data labeled or with the use of pseudo-labeled samples. In the chart we compare the performance for (i) the cross-lingual experiment (results collected from the Table 4), (ii) the multi-lingual experiment (accuracy reported in the Table 5), and (iii) the SSL cross-lingual experiment based on the use of hard pseudo-labels having 100 labeled utterances for the new language (see Fig. 3(a)). As expected, multi-lingual SER with all the utterances labeled results in a noticeable increase in performance compared to the cross-lingual SER. This increase is particularly significant for Persian (+60%), a language that has very different linguistic traits from those of English and for this reason the adaptation of the model is very important. Cross-lingual SSL based on hard pseudo-labels improves the performance of the cross-lingual configuration for all the considered languages (the increase is 50% for the Persian language).

An analysis of the feature space before and after the adaptation to the new language is also provided. The analysis aims to verify

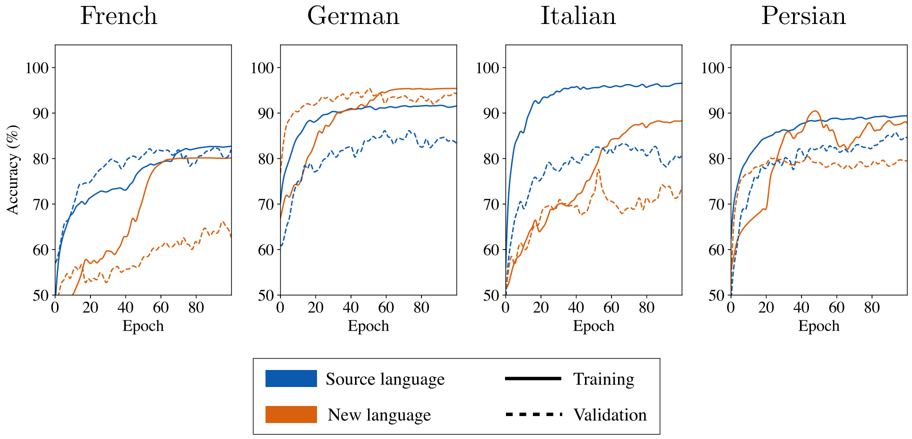


Fig. 5. Accuracy curves for training and validation sets over the training of HuBERT with 100 labeled utterances for the new language. Each plot reports the accuracy on both the source language, i.e. English, and on each of the languages considered *(better see in color and magnified)*.

whether the pseudo-label approach can effectively reduce the variations between different languages while retaining information related to emotions. To illustrate this, we use Principal Component Analysis (PCA) to project the learned feature representation, i.e., the output of the utterance encoder into 3D space. Furthermore, the silhouette score (Rousseeuw, 1987) is exploited to estimate the ability of the learned representation to discriminate emotions independently of language. The silhouette's best score value is 1 and the worst value is −1. Values close to 0 indicate overlapping clusters.

The first row of Fig. 6 displays the test utterances for both English and the new language encoded using HuBERT trained solely on English. Conversely, the second row shows the test utterances for both English and the new language encoded using HuBERT adapted on the new language via hard pseudo-labeling, and without any labeled utterances for the new language. The results indicate that the representation learned solely on English is unable to capture emotions for the new language, resulting in a very low silhouette score of approximately 0.07. In contrast, the adaptation procedure produces well-defined clusters based on emotion, independent of language, as evidenced by an average increase in the silhouette index of 0.36.

5.5. Comparison with state-of-the-art

In this section, we compare the performance of the proposed method with four recent state-of-the-art methods, AL (Kim et al., 2017), DANN (Abdelwahab & Busso, 2018), FLUDA (Ahn et al., 2021), and NNPM (Li et al., 2021). The above methods are trained for discriminating 4 emotion categories namely anger, happiness, neutral, and sadness. Implementations of the methods are not publicly available so we first report their performance on the German language from the original documents or reimplementations. For a more in-depth comparative analysis we reimplemented the considered methods in order to be able to estimate the performance also on the other three languages, namely French, Italian and Persian. We only report the results obtained for DANN and NNPM, for which the performance are in line with those declared. For a fair comparison with previous methods, we exclude from the all corpus the utterances labeled with emotion of fear and train the methods.

Our method is retrained without using labeled utterances of the new language and taking advantage of hard pseudo-labels. Table 6 shows the comparison between two versions of the proposed method, i.e. with the backbone based on BYOL-S and HuBERT, and other methods. The results achieved for our implementations are respectively "DANN (our reimplementation)" and "NNPM (our reimplementation)". Results show that both versions of the proposed method perform better than recent state-of-the-art methods. More specifically, our HuBERT-based method outperforms the second best method, which is also based on pseudo-labeling, namely NNPM, with an improvement in relative accuracy of 30%. We get better performance than the multi-task learning method, i.e. AL, with 57% better accuracy, and the unsupervised cross-corpus SER model based on few-shot learning (i.e., FLUDA) of an increase of 64%.

Table 6
Comparison with other state-of-the-art methods on new languages. Best result for each language is in **bold**.

Method	French	German	Italian	Persian
DANN (Ahn et al., 2021)	–	28.5	–	–
DANN (our reimplementation)	31.5	32.6	25.9	30.7
FLUDA (Ahn et al., 2021)	–	34.9	–	–
AL (Ahn et al., 2021)	–	42.5	–	–
NNPM (Li et al., 2021)	–	50.6	–	–
NNPM (our reimplementation)	39.7	52.	40.2	56.0
Our (BYOL-S)	52.8	66.7	48.4	68.4
Our (HuBERT)	**70.7**	**89.0**	**66.9**	**79.6**

This high gain in performance with respect to previous methods is due to three main aspects. First, the state-of-the-art methods consist of very simple architectures compared to that of HuBERT. Second, the model training procedure is profoundly different between the previous methods and that used for HuBERT. While the purpose of previous methods is to directly learn a specialized mapping of a speech signal into an emotion category, HuBERT and BYOL-S are trained to learn a general-purpose and robust representation of a speech signal. This last aspect allows the obtained representation to be more effective for the different tasks, including the recognition of emotions. Ultimately, the difference between HuBERT and BYOL-S is due to both the architectural aspect of the model and the cardinality of the dataset used for the pre-training. In fact, HuBERT is trained on a much larger and challenging dataset than the one used for BYOL-S.

5.6. Ablation study

This section presents an ablation study of the main design choices that led to the definition of the final method. The adaptability to the new language of CNN vs. Transformer based utterance encoders is evaluated. The effect of different values for the hard pseudo-labeling τ parameter is investigated. Finally, the impact of utterance rebalancing on the performance is estimated.

5.6.1. Utterance encoder comparison

Cross- and multi-lingual results demonstrate that HuBERT and BYOL-S provide more effective utterance encoding for emotion classification than EmotionCNN (see Sections 5.1 and 5.2). In this section

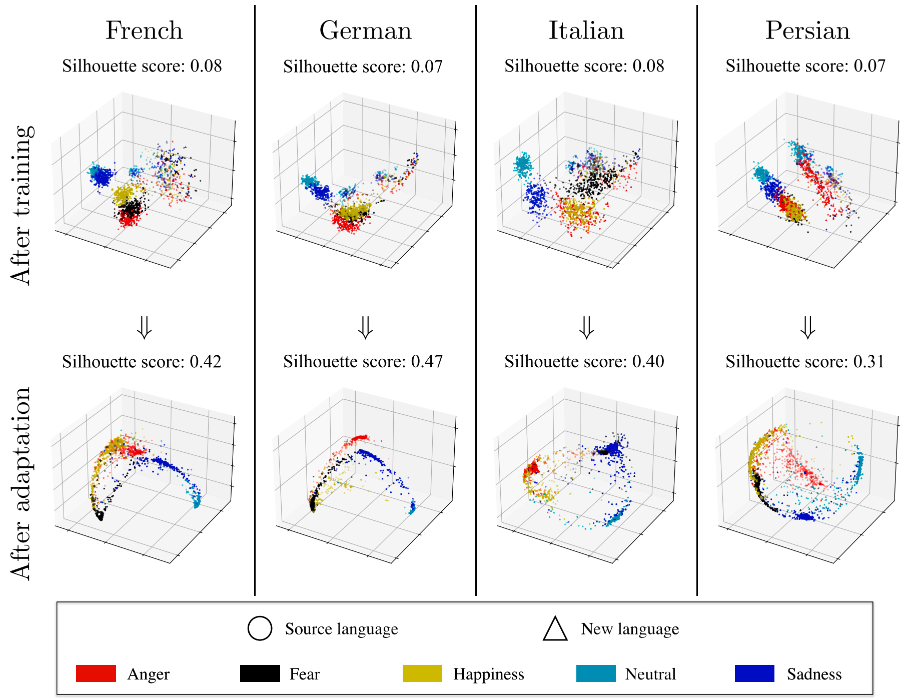


Fig. 6. PCA plot of the learned feature representation with emotion and language labels after training on the English language only (first row) and after adaptation on the new language (second row) *(better see in color and magnified)*.

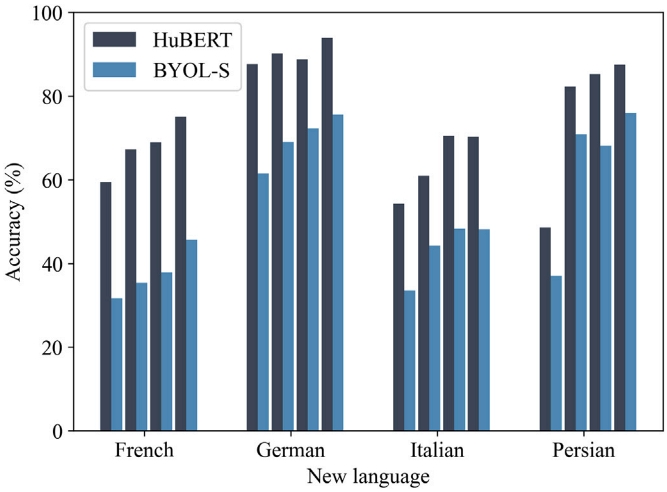


Fig. 7. HuBERT vs. BYOL-S. Results for each new language using one or the other utterance embedding model and hard pseudo-labels. The number of utterances labeled for the new language is varied from 0 to 100.

we perform the comparison for the different languages using hard pseudo-labels. The results for cross-lingual SER by varying the number of utterances labeled for the new language from 0 to 100 are shown in Fig. 7. As it is possible to see, HuBERT outperforms BYOL-S by a large gap (about +20% accuracy). This gap might be motivated by several reasons. First, HuBERT is trained on a larger and more diverse speech corpus (i.e. Librispeech), with both spontaneous and anechoic scripted speech, while BYOL-S is trained on a subset of AudioSet (Elbanna et al., 2022). Second, HuBERT's transformer-based architecture coupled with direct encoding of the raw waveform provides a more robust and powerful representation of speech.

Table 7
SER accuracy obtained using 100 labeled utterances for the new language and varying τ values for hard pseudo-labeling. Best result for each language is in **bold**.

	τ = 0.50	τ = 0.65	τ = 0.85
English	66.98	68.92	**75.91**
French	**75.10**	73.86	74.28
English	**78.35**	76.56	75.87
German	93.95	**94.37**	93.90
English	79.19	**81.89**	77.69
Italian	**74.24**	68.81	73.10
English	79.21	**81.47**	79.13
Persian	**88.53**	88.27	88.18

5.6.2. *Effect of τ*

Among all the hyperparameters, the confidence threshold τ used for hard pseudo-labels (see Section 3.2.2) is the one that needs to be carefully tuned. This subsection studies the effects of τ on our hard pseudo-labeling approach. Table 7 reports the cross-lingual SER results by considering 100 labeled utterance for the new language and varying τ. From the results it is possible to observe that the choice of the best τ does not generalize to all languages. However, for 0.50 the best performance is obtained in most cases and for this reason it was chosen.

5.6.3. *Effect of utterance rebalancing*

Here we quantitatively evaluate the contribution of the utterance rebalancing procedure on the performance of our method (see Section 3.2.3 for details). Table 8 shows the cross-lingual SER results of our method without utterance rebalance "w/o rebalance" and with utterance rebalance "w rebalance". As it is possible to see, the version of the method with utterance rebalancing outperforms the version without utterance rebalancing for all languages. The highest accuracy improvement corresponding to 40% is registered for the Persian language. The lowest gaps of 8% and 15% between the two versions are obtained for French and Italian, respectively.

Table 8
SER accuracy obtained using 100 labeled utterances for the new language and hard pseudo-labeling with $\tau = 0.5$. Performance without and with utterance rebalancing is compared. Best result for each language is in **bold**.

	w/o rebalance	w rebalance
English	67.85	**68.92**
French	65.64	**73.86**
English	74.13	**76.56**
German	78.37	**94.37**
English	73.14	**81.89**
Italian	53.64	**68.81**
English	74.34	**81.47**
Persian	47.42	**88.27**

6. Conclusions

In cross-lingual SER it is common to have many labeled utterances for the English language and a lower availability of labels for other languages. Based on this consideration, an SSL approach for cross-lingual speech emotion recognition is proposed.

The proposed method consists of a transformer able to classify an utterance into an emotion category. For SSL, we experimented with the use of hard and soft pseudo-labels for unlabeled utterances. The proposed method is evaluated using English as source language and four different languages (French, German, Italian and Persian) as new languages. It is revealed that the use of hard over soft pseudo-labels allows for better results on the new language at the expense of a drop in performance on the source language.

Experimental results show that the average accuracy has increased by 40% in comparison with state-of-the-art methods.

The proposed method has some limitations, of course. First, the method assumes that the number of emotions in the source and the new language is the same. Nonetheless, in real-wold applications this constraint could be too stringent. To overcome this limitation, prototypes could be learned from representations for newly-introduced emotion categories in the target language, using a methodology similar to the one described in Bucher, Vu, Cord, and Pérez (2021). Second, for the hard-pseudo labeling approach there is no handling of confirmation bias, i.e. overfitting to incorrect pseudo-labels predicted by the network. In fact, if the model makes several wrong unlabeled predictions, pseudo-labeling can act like a bad feedback loop and deteriorate performance. As future work we plan to handle the confirmation bias issue by averaging the predictions for different views of the unlabeled data as done in Berthelot et al. (2019) or by using reinforcement learning (Latif et al., 2022).

CRediT authorship contribution statement

Mirko Agarla: Investigation, Conceptualization, Methodology, Software, Formal analysis, Validation, Visualization, Writing – review & editing, Writing – original draft . **Simone Bianco:** Investigation, Conceptualization, Methodology, Formal analysis, Validation, Visualization, Writing – review & editing, Writing – original draft. **Luigi Celona:** Investigation, Conceptualization, Methodology, Software, Formal analysis, Validation, Visualization, Writing – review & editing, Writing – original draft. **Paolo Napoletano:** Investigation, Conceptualization, Methodology, Formal analysis, Validation, Visualization, Writing – review & editing, Writing – original draft. **Alexey Petrovsky:** Investigation, Conceptualization, Methodology, Formal analysis, Validation, Visualization, Writing – review & editing, Writing – original draft. **Flavio Piccoli:** Investigation, Conceptualization, Methodology, Software, Formal analysis, Validation, Visualization, Writing – review & editing, Writing – original draft. **Raimondo Schettini:** Investigation, Conceptualization, Methodology, Formal analysis, Validation, Visualization, Writing – review & editing, Writing – original draft. **Ivan Shanin:** Investigation, Conceptualization, Methodology, Formal analysis, Validation, Visualization, Writing – review & editing, 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.

Data availability

The authors do not have permission to share data

References

Abdelwahab, M., & Busso, C. (2018). Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*, 2423–2435.

Ahn, Y., Lee, S. J., & Shin, J. W. (2021). Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation. *IEEE Signal Processing Letters*, *28*, 1190–1194.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International joint conference on neural networks* (pp. 1–8). IEEE.

Berlitz (2021). The most spoken languages in the world. URL https://www.berlitz.com/blog/most-spoken-languages-world. (Accessed 29 June 2022).

Bertero, D., Siddique, F. B., Wu, C.-S., Wan, Y., Chan, R. H. Y., & Fung, P. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Conference on empirical methods in natural language processing* (pp. 1042–1047).

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, *32*.

Bucher, M., Vu, T.-H., Cord, M., & Pérez, P. (2021). Handling new target classes in semantic segmentation with domain adaptation. *Elsevier Computer Vision and Image Understanding*, *212*, Article 103258.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In *Interspeech* (pp. 1517–1520).

Cai, X., Wu, Z., Zhong, K., Su, B., Dai, D., & Meng, H. (2021). Unsupervised cross-lingual speech emotion recognition using domain adversarial neural network. In *International symposium on Chinese spoken language processing* (pp. 1–5). IEEE.

Chiswick, B. R., & Miller, P. W. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development*, *26*, 1–11.

Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). EMOVO corpus: An Italian emotional speech database. In *International conference on language resources and evaluation* (pp. 3501–3504). European Language Resources Association (ELRA).

Das, S., Lønfeldt, N. N., Pagsberg, A. K., & Clemmensen, L. H. (2022). Towards transferable speech emotion representation: On loss functions for cross-lingual latent representations. In *International conference on acoustics, speech and signal processing* (pp. 6452–6456). IEEE.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. N. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT* (pp. 4171–4186).

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Elsevier Pattern Recognition*, *44*, 572–587.

Elbanna, G., Scheidwasser-Clow, N., Kegler, M., Beckmann, P., Hajal, K. E., & Cernak, M. (2022). BYOL-S: Learning self-supervised speech representations by bootstrapping.

Feraru, S. M., Schuller, D., et al. (2015). Cross-language acoustic emotion recognition: An overview and some tendencies. In *International conference on affective computing and intelligent interaction* (pp. 125–131). IEEE.

Gamallo, P., Pichel, J. R., & Alegria, I. (2017). From language identification to language distance. *Physica A. Statistical Mechanics and its Applications*, *484*, 152–162.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., et al. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *International conference on acoustics, speech and signal processing* (pp. 776–780). IEEE.

Gournay, P., Lahaie, O., & Lefebvre, R. (2018). A Canadian French emotional speech dataset. In *Multimedia systems* (pp. 399–402). ACM.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, *33*, 21271–21284.

Hansen, L., Zhang, Y.-P., Wolf, D., Sechidis, K., Ladegaard, N., & Fusaroli, R. (2022). A generalizable speech emotion recognition model reveals depression and remission. *Acta Psychiatrica Scandinavica*, *145*, 186–199.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460.

Kim, J., Englebienne, G., Truong, K. P., & Evers, V. (2017). Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning. In *Interspeech* (pp. 1113–1117).

Kshirsagar, S., & Falk, T. H. (2022). Cross-language speech emotion recognition using bag-of-word representations, domain adaptation, and data augmentation. *MDPI Sensors*, *22*, 6445.

Latif, S., Cuayáhuitl, H., Pervez, F., Shamshad, F., Ali, H. S., & Cambria, E. (2022). A survey on deep reinforcement learning for audio-based applications. *Springer Artificial Intelligence Review*, 1–48.

Latif, S., Qadir, J., & Bilal, M. (2019). Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. In *International conference on affective computing and intelligent interaction* (pp. 732–737). IEEE.

Lefter, I., & Jonker, C. M. (2017). Aggression recognition using overlapping speech. In *International conference on affective computing and intelligent interaction* (pp. 299–304). IEEE.

Li, J., Yan, N., & Wang, L. (2021). Unsupervised cross-lingual speech emotion recognition using pseudo multilabel. In *Automatic speech recognition and understanding workshop* (pp. 366–373). IEEE.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS One*, *13*, Article e0196391.

Neumann, M., et al. (2018). Cross-lingual and multilingual speech emotion recognition on English and French. In *International conference on acoustics, speech and signal processing* (pp. 5769–5773). IEEE.

Nezami, O. M., Lou, P. J., & Karami, M. (2019). ShEMO: A large-scale validated database for Persian speech emotion detection. *Language Resources and Evaluation*, *53*, 1–16.

Ocquaye, E. N., Mao, Q., Xue, Y., & Song, H. (2021). Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. *International Journal of Intelligent Systems*, *36*, 53–71.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *International conference on acoustics, speech and signal processing* (pp. 5206–5210). IEEE.

Perez-Toro, P. A., Vasquez-Correa, J. C., Bocklet, T., Noth, E., & Orozco-Arroyave, J. R. (2021). User state modeling based on the arousal-valence plane: Applications in customer satisfaction and health-care. *IEEE Transactions on Affective Computing*.

Petroni, F., & Serva, M. (2008). Language distance and tree reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, P08012.

Pichora-Fuller, M. K., & Dupuis, K. (2020). Toronto emotional speech set (TESS). http://dx.doi.org/10.5683/SP2/E8H2MF.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Elsevier Journal of Computational and Applied Mathematics*, *20*, 53–65.

Scheidwasser-Clow, N., Kegler, M., Beckmann, P., & Cernak, M. (2022). SERAB: A multilingual benchmark for speech emotion recognition. In *International conference on acoustics, speech and signal processing* (pp. 7697–7701). IEEE.

Schuller, B., & Batliner, A. (2013). *Computational paralinguistics: Emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Conference on computer vision and pattern recognition* (pp. 1199–1208). IEEE.

Tamulevičius, G., Korvel, G., Yayak, A. B., Treigys, P., Bernatavičienė, J., & Kostek, B. (2020). A study of cross-linguistic speech emotion recognition based on 2D feature spaces. *MDPI Electronics*, *9*, 1725.

Tanaka, D., Ikami, D., Yamasaki, T., & Aizawa, K. (2018). Joint optimization framework for learning with noisy labels. In *Conference on computer vision and pattern recognition* (pp. 5552–5560). IEEE.

Tumanova, V., Woods, C., & Wang, Q. (2020). Effects of physiological arousal on speech motor control and speech motor practice in preschool-age children who do and do not stutter. *Journal of Speech, Language, and Hearing Research*, *63*, 3364–3379.

Wang, W. (2010). *Machine audition: Principles, algorithms and systems: Principles, algorithms and systems*. IGI Global.

Xiao, Z., Wu, D., Zhang, X., & Tao, Z. (2016). Speech emotion recognition cross language families: Mandarin vs. western languages. In *International conference on progress in informatics and computing* (pp. 253–257). IEEE.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In *International conference on learning representations*.

Zhang, Z., Ringeval, F., Dong, B., Coutinho, E., Marchi, E., & Schüller, B. (2016). Enhanced semi-supervised learning for multimodal emotion recognition. In *International conference on acoustics, speech and signal processing* (pp. 5185–5189). IEEE.

Zhou, H., & Chen, K. (2019). Transferable positive/negative speech emotion recognition via class-wise adversarial domain adaptation. In *International conference on acoustics, speech and signal processing* (pp. 3732–3736). IEEE.