Vaaras, Einari; Ahlqvist-Björkroth, Sari; Drossos, Konstantinos; Lehtonen, Liisa; Räsänen, Okko

Development of a speech emotion recognizer for large-scale child-centered audio recordings from a hospital environment

Published in:
Speech Communication

DOI:
10.1016/j.specom.2023.02.001

Published: 01/03/2023

Document Version
Publisher's PDF, also known as Version of record

Speech Communication 148 (2023) 9–22

Contents lists available at ScienceDirect

Speech Communication

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

Development of a speech emotion recognizer for large-scale child-centered audio recordings from a hospital environment

Einari Vaaras [a,*], Sari Ahlqvist-Björkroth [b], Konstantinos Drossos [a], Liisa Lehtonen [b,c], Okko Räsänen [a,d]

[a] Unit of Computing Sciences, Tampere University, Finland
[b] Department of Clinical Medicine, University of Turku, Finland
[c] Department of Pediatrics and Adolescent Medicine, Turku University Hospital, Finland
[d] Department of Signal Processing and Acoustics, Aalto University, Finland

ARTICLE INFO

Keywords:
Speech emotion recognition
Speech analysis
Real-world audio
Daylong audio
LENA recorder

ABSTRACT

In order to study how early emotional experiences shape infant development, one approach is to analyze the emotional content of speech heard by infants, as captured by child-centered daylong recordings, and as analyzed by automatic speech emotion recognition (SER) systems. However, since large-scale daylong audio is initially unannotated and differs from typical speech corpora from controlled environments, there are no existing in-domain SER systems for the task. Based on existing literature, it is also unclear what is the best approach to deploy a SER system for a new domain. Consequently, in this study, we investigated alternative strategies for deploying a SER system for large-scale child-centered audio recordings from a neonatal hospital environment, comparing cross-corpus generalization, active learning (AL), and domain adaptation (DA) methods in the process. We first conducted simulations with existing emotion-labeled speech corpora to find the best strategy for SER system deployment. We then tested how the findings generalize to our new initially unannotated dataset. As a result, we found that the studied AL method provided overall the most consistent results, being less dependent on the specifics of the training corpora or speech features compared to the alternative methods. However, in situations without the possibility to annotate data, unsupervised DA proved to be the best approach. We also observed that deployment of a SER system for real-world daylong child-centered audio recordings achieved a SER performance level comparable to those reported in literature, and that the amount of human effort required for the system deployment was overall relatively modest.

1. Introduction

Speech contains a vast amount of information other than the linguistic content of speech, such as the speaker's health state, attitude, emotions, and personality (Batliner and Schuller, 2013). In speech emotion recognition (SER), the aim is to recognize the emotional state of the speaker from a speech signal (Batliner et al., 2010). Determining the emotional content of speech is particularly interesting in the study of infants' auditory environments, where the early affective and social experiences of infants can impact their later cognitive and socio-emotional development.

Preterm infants are commonly deprived from normal vocal communication with their parents during their first months when they require hospital care. Hospitalization for a premature infant can last from a few weeks up to several months. During hospital care, both the quantity and quality of vocal communication is likely to be different compared to home environment. Preterm infants have an increased risk for abnormal cognitive development including language development (Nyman et al., 2017), as well as for emotional problems such as depression (Upadhyaya et al., 2021). These problems are partly caused by early parent-infant separation and a lack of parents' participation in the care of their infant. In this context, parents' positive vocal expressions in a neonatal intensive care unit (NICU) environment have been shown to be linked to the interaction with their infant (Filippa et al., 2019). Furthermore, it has been shown that the expressed emotions during this interaction can be detected from the parents' speech patterns (Filippa et al., 2019). However, the associations between parental emotional speech and subsequent development of a preterm infant have not been studied.

To better understand the effect of parental proximity and communication on the long-term development of preterm infants, a large

* Corresponding author.
E-mail address: einari.vaaras@tuni.fi (E. Vaaras).

https://doi.org/10.1016/j.specom.2023.02.001
Received 6 May 2022; Received in revised form 7 February 2023; Accepted 13 February 2023
Available online 15 February 2023

audio corpus of hundreds of hours of child-centered audio recordings from two NICUs in Finland and Estonia has recently been collected as a part of a so-called APPLE study (Ståhlberg-Forsen et al., 2021). To analyze the emotional characteristics of speech in the dataset, a SER system is essentially required. However, there are no existing in-domain SER systems for child-centered audio data nor for Finnish or Estonian real-world audio. In the absence of speech emotion labels, traditional supervised learning methods cannot be applied for this dataset without human labeling efforts. This raises the question of how to most efficiently deploy a SER system for this new domain. Despite the advances in techniques such as *domain adaptation* (DA; e.g., Ben-David et al. (2009)) and *active learning* (AL; e.g., Settles (2012)), the existing literature is not clear on what is the most practical approach to deploy a SER system for data from a new domain, especially when the amount of required manual effort is weighed against the obtained SER performance.

To this end, the aim of this paper is to explore the applicability of *cross-corpus generalization* (CCG) and state-of-the-art AL and DA methods as alternative approaches for developing a SER system for real-world unannotated child-centered audio. We compare these methods in out-of-domain system deployment by conducting simulation experiments using four already existing labeled SER corpora, followed by application of the best identified practices and model settings to the Finnish subset of the NICU audio data (from now on "NICU-A"). As a result, we obtain a functional SER system for the NICU-A data, enabling future research (and potential interventions; see Swain (2017)) on infant emotional environments. To the best of our knowledge, this is the first functional SER system for child-centered LENA data, and the first one for real-life daylong audio. In addition, the simulations and the case study with NICU-A provide new knowledge on how to approach SER system deployment in the future for unannotated data with unknown statistical properties, and how well an effective strategy could be identified using prior simulations on already labeled datasets.

This study is continuation to our previous work in Vaaras et al. (2021), where we briefly introduced NICU-A and reported SER system performance on the final NICU-A test set. Here we expand on that work by thoroughly describing our simulation-based investigations and discussing their findings for out-of-domain system deployment. We also introduce the NICU-A dataset, its properties and its annotation process in a more complete manner. Taken together, the study aims to provide new knowledge on best strategies for developing a novel SER system for a new data in general, and for our present NICU-A dataset in particular.

2. Related work

Perhaps the most straightforward strategy to develop a novel SER system is CCG, which has been examined in multiple SER studies (e.g. Schuller et al. (2010), Schuller et al. (2011b), Zhang et al. (2011), and Zhang et al. 2019). CCG values simplicity over complexity by relying on the naive assumption that the statistical properties of one or multiple labeled training corpora are close enough to those of an unlabeled test corpus; i.e., it is possible to train a well-performing classifier that generalizes from the labeled corpora to the test data. Schuller et al. (2010) performed extensive intra- and inter-corpus CCG experiments using six frequently-used SER corpora of various languages. Their study highlighted many issues with cross-domain SER, such as CCG-based SER being feasible only with certain corpora and emotional classes, even with corpora of similar cultural backgrounds. This same domain mismatch issue has been perhaps the most commonly occurring problem with CCG-based SER methods, and has been denoted earlier (Schuller et al., 2011b; Zhang et al., 2011). In order to counteract domain mismatch, Zhang et al. (2019) proposed a novel loss function to preserve label similarities in a learned feature space. They showed that by combining their loss function with a traditional cross-entropy loss, they were able to significantly outperform a reference method which only utilized a cross-entropy loss.

For more advanced strategies to tackle the problem of domain mismatch, various DA methods have been developed for SER. For example, Sagha et al. (2016) studied a multi-language cross-corpus SER setting using four corpora. They presented a novel DA method that tries to find a common representation space for the source and target language. Through extensive testing, they showed that their method improved the average classification performance compared to the state-of-the-art DA method for SER at the time. Deng et al. (2017) proposed adding a Universum loss to the reconstruction loss of an autoencoder-based classifier for unsupervised DA. Abdelwahab and Busso (2018) proposed a deep neural network (DNN)-based unsupervised adversarial DA method for SER. Their method learns a domain-invariant feature space between labeled data from a source domain and unlabeled data from a target domain, while simultaneously maintaining good performance on the primary SER task. As pointed out in Kouw and Loog (2021), the success of DA can be very sensitive both to the chosen algorithm and its hyperparameter configuration on a particular dataset.

Another common strategy that has been successfully used in SER is AL. In AL, human annotation effort is reduced by using automatic algorithms to identify a subset of the most informative samples that the human should annotate for machine learning system deployment. Zhang and Schuller (2012) presented two iterative AL methods which efficiently reduced the required number of annotations in their experiments. The first method selects instances for human annotation which the method predicts as a sparse class, and the second method chooses the instances for which the method predicts a medium confidence score for annotation. Zhao and Ma (2013) proposed an iterative AL algorithm which utilizes conditional random fields to quantify the uncertainty on each unlabeled sample, and the most uncertain samples were then selected for human annotation. In most cases, their method outperformed random sampling for sample selection. Abdelwahab and Busso (2019) examine various AL methods that are based on uncertainty and maximizing the diversity in the training set to simulate limited annotated data in DNN classifiers. Their experiments reveal that the tested AL methods outperform random sampling-based methods when selecting samples for annotation.

Only a handful of studies have been conducted on large-scale SER datasets. Jia et al. (2019) studied SER with a vast 7-million-utterance internet voice corpus. They pretrained their two novel DNN-based models with 90,000 unlabeled utterances and then fine-tuned and evaluated their models on 3000 randomly selected utterances from the same dataset. Their experiments revealed that both proposed methods outperformed traditional SER models. Fan et al. (2021) presented a large-scale SER dataset with a little over 147,000 utterances from 820 test subjects with a total duration of over 200 h. They proposed a novel SER model containing pyramid convolutions which outperformed other models that were tested on the dataset. Furthermore, they showed that existing models are prone to overfit to small-scale datasets which limits the ability of these models to generalize for real-life data. As far as we know, no systematic work has been conducted on performing SER for child-centered audio recordings in general, nor for speech recordings collected from the vicinity of preterm infants in a hospital environment.

3. Methods

The main goal of our methodology was to deploy a SER system for the automatic analysis of infants' auditory environments for the NICU-A data. Since there was an absence of labeled target domain data, alternative machine learning-based approaches, namely CCG, AL, and DA, were compared in the present experiments. The present CCG approach acted as our "naive" baseline approach, while the AL method, medoid-based active learning (MAL) (Zhao et al., 2017), and the DA method, Wasserstein distance-based domain adaptation (WDA) (Drossos et al., 2019), of the present experiments were selected based on their state-of-the-art performance in their respective audio-related tasks. Fig. 1

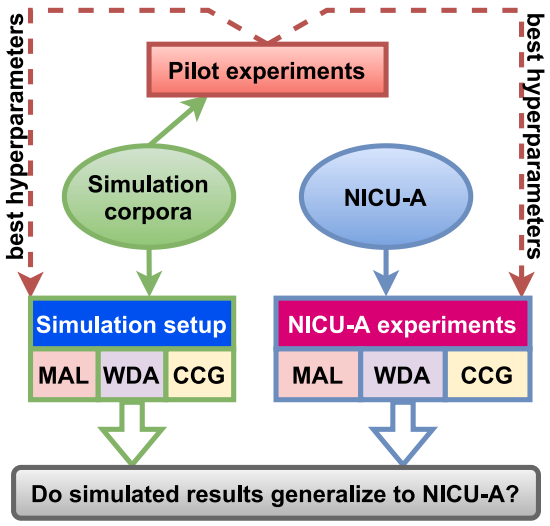

Fig. 1. A block diagram of the present experimental setup in which MAL, WDA, and CCG methods are compared in both the simulation setup and the NICU-A experiments.

depicts a block diagram of the present experimental setup. First, already existing SER corpora, referred to as the *simulation corpora*, were thoroughly experimented with in pilot experiments to find suitable hyperparameters for the tested methods in the SER task. Then, simulations were carried out using these hyperparameters with the simulation corpora in order to compare the different methods, and to estimate how much data needs to be annotated for SER when using MAL. Next, NICU-A was partially annotated based on these findings. Finally, experiments similar to the simulations were carried with NICU-A data to create a SER system for NICU-A. The results of these NICU-A experiments were also compared with the results of the simulations to test the generalizability of the simulated results.

3.1. Medoid-based active learning

For situations when there is a limited number of labels that can be manually assigned, i.e. a limited *labeling budget*, and when the annotations add up to only a small portion of the data, Zhao et al. (2017) proposed MAL as an AL method for sound event classification. Since this is also the premise of the annotation process of the present study, MAL serves as the foundation for the AL method used in the present experiments. The MAL algorithm can be divided into three consecutive stages: (1) obtain an affinity matrix that contains the pairwise similarities between each sample in a dataset, (2) perform k-medoids clustering using this affinity matrix, and (3) starting from the largest cluster, query for human annotations for the medoids in a descending cluster size order.

In the first stage, the similarity metric used in the present experiments was selected based on pilot experiments with MAL using existing SER corpora. To obtain the affinity matrix, A, each sample in a dataset is first represented as a 600-dimensional utterance-level log-mel feature representation (see Section 4.1). Then, a 32-dimensional latent representation of the log-mel features is obtained using a DNN-based autoencoder with six layers (see Section 4.3.3). Finally, A is defined by computing Pearson distances (Immink and Weber, 2014), d_P, for the bottleneck features across all the samples.

In the second stage, k-medoids clustering is applied to the data. First, a random sample is selected as a member of a set of medoids, S, after which $k-1$ additional samples are added to S using the *farthest-first traversal* algorithm. Here, the distance from a sample, $\boldsymbol{a}$, to the set S is defined as

$$d_P(\boldsymbol{a}, S) = \min_{\boldsymbol{b} \in S} d_P(\boldsymbol{a}, \boldsymbol{b}) \ . \tag{1}$$

Then, the samples in S are used as the initial medoids for a k-medoids clustering algorithm (see e.g. Park and Jun (2009) for a detailed description) to assign each sample into one of the clusters.

In the final stage, the clusters are sorted in a descending order based on the number of elements in each cluster. The cluster medoids are then presented to human annotators for labeling. In the present experiments, we examined two different strategies for using these labels: (i) assigning the annotated medoid label for all samples in a cluster (as in Zhao et al. (2017); referred to as "cluster labels"), or (ii) only using the annotated medoid samples for classifier training ("medoid labels"; a condition not studied in the original MAL paper (Zhao et al., 2017)). Based on pilot experiments, a suitable value for k was found to be $\frac{N}{3}$, where N is the number of samples in a corpus.

3.2. Wasserstein distance-based domain adaptation

The DA method of the present experiments is based on the WDA method proposed by Drossos et al. (2019), originally designed for acoustic scene classification. In WDA, a neural network classifier is adapted to a target corpus, D_T, by using labeled data from a source corpus/corpora, D_S. This classifier, aka the *source model* M, consists of two parts: a feature extractor, F, and a label classifier, C_L. The adaptation process of WDA involves two steps, which are depicted in Fig. 2.

The first step (Fig. 2, top) consists of training M using D_S samples, X_S, and their respective labels, Y_S, to obtain the initial trained feature extractor, F_S. This is achieved by using binary cross-entropy (Drossos et al., 2019) as the loss, defined as:

$$L_M(\boldsymbol{x}, \boldsymbol{y}) = - \sum_{(\boldsymbol{x},\boldsymbol{y}) \in (X_S, Y_S)} \boldsymbol{y}^T log_{10}(C_L(F(\boldsymbol{x}))). \tag{2}$$

In the second step (Fig. 2, bottom), F_S is adapted into D_T to obtain the adapted feature extractor, F_T. This is done by minimizing the Wasserstein-1 distance, W_d, between the distributions of D_S and D_T using an adversarial training formulation, namely a WGAN framework (Arjovsky et al., 2017). In the process, F_S is adapted into F_T by finding a common feature representation for D_S and D_T by iteratively minimizing the two loss functions:

$$L_{C_D}(\boldsymbol{x}, \boldsymbol{z}) = \sum_{\boldsymbol{x} \in X_S} C_D(F_S(\boldsymbol{x})) - \sum_{\boldsymbol{z} \in X_T} C_D(F_T(\boldsymbol{z})) \tag{3}$$

$$L_{F_T}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{\boldsymbol{z} \in X_T} C_D(F_T(\boldsymbol{z})) + L_M(\boldsymbol{x}, \boldsymbol{y}) \ , \tag{4}$$

where C_D is the domain discriminator and X_T are the target corpus samples. The parameters for C_D and F_T are updated in turns, where Eqs. (3) and (4) are the losses for updating the parameters of C_D and F_T, respectively. The output feature representation of F_T acts as the input features for C_D. In addition, the parameters of F_S are the initial parameters of F_T. As shown in Drossos et al. (2019), the minimization of L_{C_D} and L_{F_T} is shown to minimize W_d between the distributions of D_S and D_T. In order to account for the performance degradation of C_L during the adaptation, the authors of Drossos et al. (2019) added L_M into Eq. (4). For a detailed formulation of the WDA algorithm, see Algorithm 1 in Drossos et al. (2019).

In addition to unsupervised WDA, a semi-supervised variant which was not part of the original WDA paper (Drossos et al., 2019) was examined. This version utilizes a small subset of target corpus labels, Y_T, to compute the label classification accuracy after each training iteration of the adaptation process. The model with the highest accuracy on this subset of Y_T is then selected as the final adapted model.

3.3. Cross-corpus generalization

In our CCG baseline approach, n labeled source corpora are merged into one training set S. Next, a classifier is trained with S using supervised learning. Finally, the trained classifier is tested on an unlabeled target corpus T, where $T \notin S$.

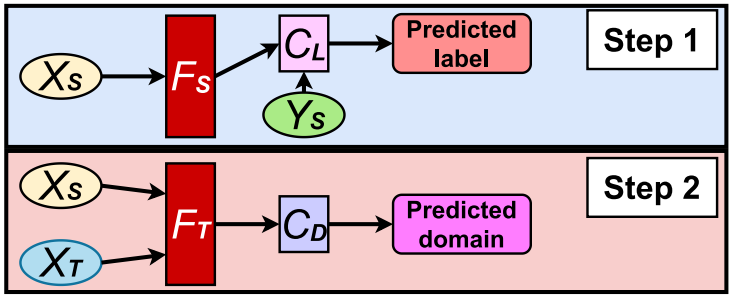

Fig. 2. The two-step adaptation process of WDA. In the first step, the source model M consisting of F_S and C_L is trained using source corpus samples X_S and their respective labels Y_S to classify X_S into emotion categories. In the second step, F_S is adapted into F_T with X_S and target corpus samples X_T using a domain discriminator C_D in an adversarial training process.

4. Simulation setup

Before any NICU-A data had been annotated, simulations with four already existing SER corpora were carried out with the CCG, AL, and DA methods of the present study. The aim of these simulations was to simulate and compare different strategies for deploying a SER system on a new unannotated corpus of potentially different language, speaking style, and recording context, and to estimate the number of samples that needs to be annotated for the SER task when using MAL. All model hyperparameters were based on extensive pilot experiments with the four simulation corpora. For all experiments, the unweighted average recall (UAR %) is used as the primary evaluation measure.

4.1. Features

Log-mel, GeMAPS, and eGeMAPS features (Eyben et al., 2016) were used in all experiments with the exception of DA experiments, where only log-mel features were used based on their superior performance in pilot experiments. The GeMAPS and eGeMAPS are minimalistic features proposed by Eyben et al. (2016) as an attempt to unify features in affective computing, including SER, and have since been used in many SER studies (e.g. Latif et al. (2019), Trigeorgis et al. (2016), Cummins et al. (2017)). For the log-mel features, 40 mel filters were used with a Hann window using a 30-ms window size and 10-ms shifts. To get constant-dimensional utterance feature representations, seven functionals (the first four moments, min, max, and range) were taken from the time series of the log-mel features. In addition, four functionals (the first four moments) were applied to the first and second order delta features, which resulted in a 600-dimensional feature representation for the log-mel features. The 62- and 88-dimensional GeMAPS and eGeMAPS features were extracted using the openSMILE toolkit (Eyben et al., 2013). The features for each corpus were normalized using z-score normalization in order to have zero mean and unit variance for each of the features at corpus level.

4.2. Simulation corpora

Four already available speech corpora with emotion labels were used in the simulations:

(1) *The Berlin Emotional Speech Database* (EMO-DB) (Burkhardt et al., 2005) is a well-known and perhaps the most widely used SER corpus, containing 535 spoken utterances in German from 10 actors (five male). The actors read sentences with predefined emotions in seven emotional labels: anger, boredom, disgust, feat, joy, neutral, and sadness.

(2) *eNTERFACE* (Martin et al., 2006) is an audiovisual database consisting of 1287 video samples in English from 42 test subjects (eight female) from 14 nationalities. Each test subject listened to six successive short stories, each of them evoking an emotion from six categories: anger, disgust, fear, joy, sadness, and surprise. Only the audio tracks were used in the present study.

(3) *The Finnish Emotional Speech Corpus* (FESC) (Airas and Alku, 2006) consists of 450 spoken 83-word passages of Finnish prose from nine Finnish professional actors (five male) portraying emotions of five categories: neutral, sadness, joy, anger, and tenderness. These passages were further split into 4254 utterances based on long silences as defined by an energy threshold (Vaaras, 2021).

(4) *The Ryerson Audio–Visual Database of Emotional Speech and Song* (RAVDESS) (Livingstone and Russo, 2018) is a multimodal database, including a total of 7356 recordings in English from 24 professional actors (12 male). Only the 1440 speech-only recordings were used in the present experiments, covering eight emotional labels: neutral, calm, happy, sad, angry, fearful, surprise, and disgust.

These corpora were selected with the aim of incorporating corpora of different languages and sizes to the present experiments. Furthermore, excluding FESC, we wanted to select publicly available corpora that have been popularly used in SER literature as a means to be able to compare the achieved performance level to other SER studies. FESC was selected in order to study if it is easier to adapt a matching Finnish-language corpus to NICU-A using CCG or DA, and also since FESC was the only Finnish-language SER corpus available.

The emotional labels of the simulation corpora were mapped into the quarters of the valence-arousal plane following the mapping of Schuller et al. (2010) (Fig. 3), which has also been used in multiple SER studies (e.g. Schuller et al. (2011b), Zhang et al. (2011), Sagha et al. (2016), Latif et al. (2019), Deng et al. (2014), Mao et al. (2017)). This mapping was made for two reasons: First, in order to simplify the task of annotating NICU-A samples (see Section 6.1) and since the researchers in the APPLE study were interested in the quantity of positive valence in NICU-A, the annotation of NICU-A was carried out in terms of valence and arousal. Hence, the emotional labels of the simulation corpora were also decided to be mapped to the valence-arousal plane in order to harmonize the differences between the emotional labels both across the simulation corpora and also between NICU-A and the simulation corpora. Second, the mapping was made to simplify the classification task into two binary classifications, valence (positive/negative) and arousal (high/low), since reliable emotion classification is easier to perform on a binary scale rather than using fine-grained emotion categories. Table 1 shows the class distributions of the simulation corpora.

4.3. Simulation experiments

Four types of simulation experiments were conducted using the simulation corpora:

4.3.1. Within-corpus experiments

Each corpus was examined individually to get an estimate of the accuracy that is achievable, if annotations for the entire dataset were available for classifier training and evaluation. Each corpus was randomly split into a training and test set in a 85:15 ratio, followed by

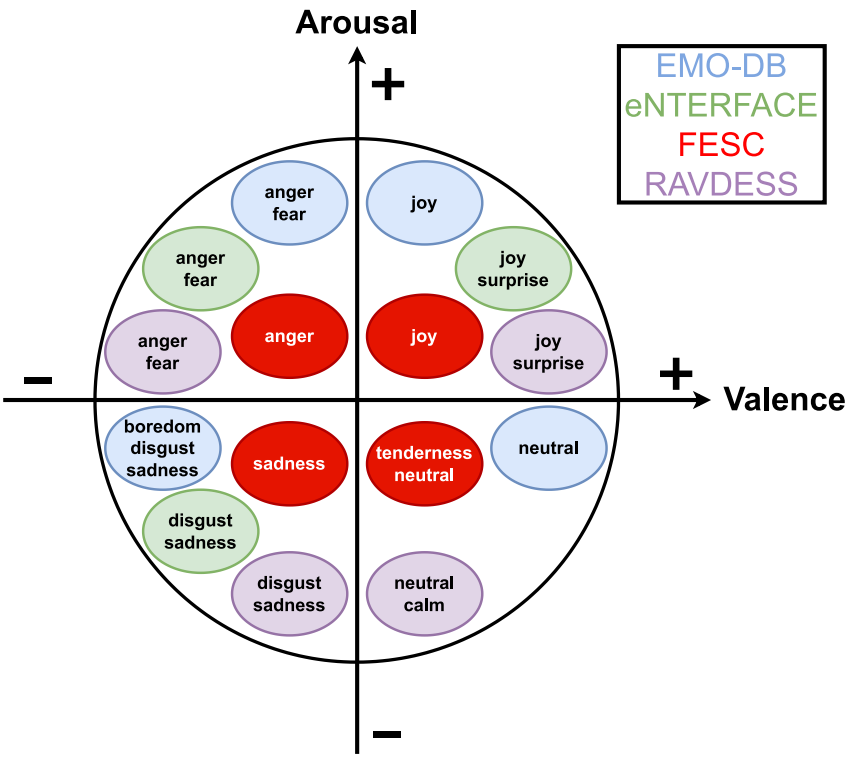

Fig. 3. Mapping of emotion categories of the utilized SER corpora into the valence-arousal plane.
Source: Figure adapted from Vaaras (2021).

Table 1
The class distributions of the corpora used in the simulations.

	Valence			Arousal	
	positive	*neutral*	*negative*	*high*	*low*
EMO-DB	71	79	385	267	268
eNTERFACE	427	–	860	857	430
FESC	1755	1082	1417	1969	2285
RAVDESS	576	96	768	768	672

training a support vector machine (SVM) classifier with a radial basis function (RBF) kernel. Imbalances in the training data class distributions were countered by weighting each sample inversely proportional to its class frequency. Optimal SVM hyperparameters (box constraint C and kernel scale parameter γ) were selected for each feature type and for both valence and arousal individually based on a grid search using 5-fold cross-validation for the training set. Then, the trained SVM was evaluated on the test set.

4.3.2. Cross-corpus generalization

The CCG simulation experiments consisted of two settings: 1-to-1 and 3-to-1 CCG. In the 1-to-1 settings, all possible combinations were tested in which one of the simulation corpora was used as the training set and another corpus as the test set for an SVM with an RBF kernel. In this setting, the optimal values for C and γ were taken from the within-corpus experiments. In the 3-to-1 setting, three simulation corpora were used for SVM training and the fourth corpus was used for testing. In a similar manner as in the within-corpus experiments, optimal values for C and γ were determined based on a grid search on training/development data split.

4.3.3. Active learning

The AL simulation experiments were conducted in a within-corpus manner using MAL, in which each simulation corpus was randomly split into a training, validation, and test set in a 70:15:15 ratio (the same test set as in the within-corpus experiments). An autoencoder was used to compress the log-mel features into a latent representation. The encoder network consisted of three dense layers of 512, 512, and 32 units with exponential linear unit (ELU) nonlinearities, followed by two 512-unit dense layers with ELU nonlinearities and a linear reconstruction layer. For the first two layers, a dropout of 10% was used. The autoencoder was trained with a batch size of 1024, mean squared error loss, and early stopping with a patience of 300 based on validation loss. Adam optimizer (Kingma and Ba, 2015) was used with a learning rate (lr) of 10^{-4}.

After obtaining a latent representation of the log-mel features and after clustering the data according to the MAL algorithm, the annotation process for MAL was simulated for both valence and arousal using labeling budgets of 3%, 6%, and 10% of the total samples in each simulation corpus. The same set of samples as in the training set of the within-corpus experiments was used for the simulated annotations. Experiments with and without assigning the label of the medoid as the label of all the cluster members were carried out. Then, optimal values for C and γ were taken from the within-corpus experiments, and the annotated samples were used to train an SVM classifier with all three features for both valence and arousal. The performance of the classifier was then tested on the test set.

4.3.4. Domain adaptation

In a similar manner as in the CCG simulation experiments, the DA simulations were conducted in 1-to-1 and 3-to-1 settings, and for both

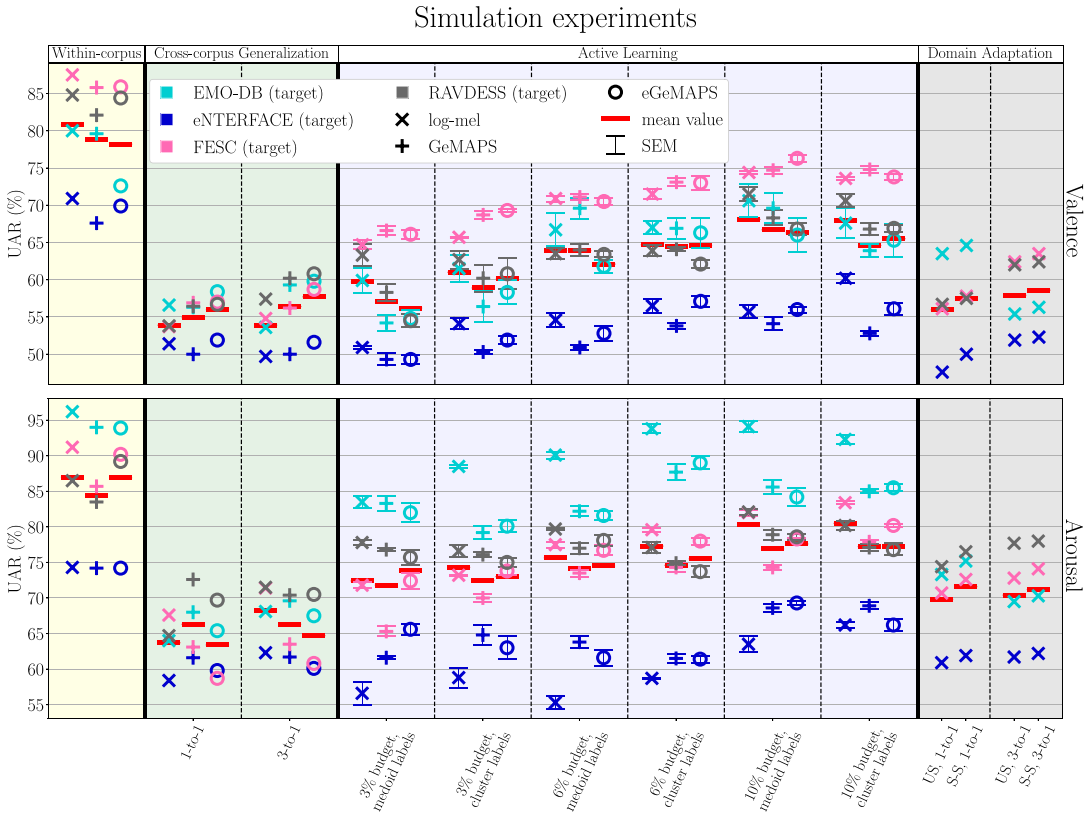

Fig. 4. The unweighted average recall (UAR) performance scores of the simulation experiments for valence (top) and arousal (bottom). The result of each individual experiment is reported for each test corpus and for all three features, with the exception of only using log-mel features in the DA experiments. In the 1-to-1 CCG and DA results, the reported value of each simulation corpus is the mean value of the results of the three separate 1-to-1 settings where the given simulation corpus is used as the test set/target corpus. For AL experiments, the results are shown with three different labeling budgets (3%, 6%, and 10% of the total number of samples in each simulation corpus) and with either medoid labels or cluster labels being used. Since the medoids are initialized at random in MAL, the average of the classification accuracies of five consecutive experiments is reported together with the standard error of the mean (SEM). In the DA experiments, the results for the unsupervised (US) and semi-supervised (S-S) variant of WDA are given. For all experiments, the mean value of the results of the four simulation corpora is given with a red dash. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

valence and arousal tasks separately. The data for each simulation corpus was split as in the within-corpus experiments. In the 1-to-1 settings, one simulation corpus acted as the source corpus and another corpus acted as the target corpus for WDA. Similarly, in the 3-to-1 settings, one corpus acted as the target corpus, and the training and test set for the source data was the combination of the three remaining source corpora. For the first step of the adaptation process, the training set of the source data was used to train M using the Adam optimizer ($lr = 10^{-4}$), batch size of 1024, and early stopping with a patience of 100 based on test set accuracy. The log-mel features were the input for F, which consisted of three dense layers of 512, 512, and 256 units. Each layer was followed by batch normalization, and the first two layers had leaky rectified linear unit (ReLU) (Maas et al., 2013) nonlinearities and a dropout of 40%. C_L consisted of three dense layers of 256, 256, and 2 units, the first two layers having leaky ReLU nonlinearities and a dropout of 30%, and the last layer was followed by a softmax function. A unique M was trained for both valence and arousal and for each distinct set of source data. For the second step of the adaptation process, the entire unlabeled source data and the unlabeled training set of the target corpus were used for training. Following Drossos et al. (2019), the unsupervised variant of WDA was trained until the saturation of the first term in Eq. (4). For the semi-supervised variant, the labels of a random subset of the target corpus training set (5% of corpus size) were utilized. C_D consisted of four dense layers of 512, 512, 256, and 1 units, the first three layers having ReLU nonlinearities. The parameters of C_D and F_T were updated with RMSProp (Tieleman and Hinton, 2012) and Adam optimizers, respectively. For F_T, $lr = 5 \cdot 10^{-5}$ was used in the 1-to-1 setting, and $lr = 3 \cdot 10^{-5}$. For both settings, $lr = 5 \cdot 10^{-5}$ was used for C_D. Finally, the performance of the adapted model was evaluated on the test set of the target corpus.

5. Results on the simulation setup

Fig. 4 presents the results of the simulation setup. The achieved performance level in both the within-corpus and CCG experiments is on par with those of earlier literature. For example, in the within-corpus experiments for EMO-DB, the 96.2% UAR accuracy for arousal with log-mel features was similar to the results reported for the same corpus in Eyben et al. (2016) (97.8% UAR) and Schuller et al. (2010) (97% UAR). As an example of the CCG experiments, eNTERFACE with log-mel features obtained an accuracy of 62.3% UAR for arousal. This is close to the accuracies of 64% UAR by Schuller et al. (2010) and 58.2% UAR by Zhang et al. (2011). However, it should be noted that the data split and the used features were not identical between the studies, as well as the training corpora in the CCG experiments. In all experiments, it is notable that the classification performance with eNTERFACE as the target corpus for both valence and arousal tasks is distinctly lower than that of other simulation corpora. This may be explained due to the different nature of eNTERFACE compared to the other corpora. The emotional expressions of eNTERFACE were evoked from non-actor test subjects of multiple different nationalities, the majority having a different native language than the one used in the corpus, whereas the emotions in other corpora were expressed by actors using speakers' native languages.

When comparing the tested methods in Fig. 5, it can be observed that there is a notable drop in classification performance without

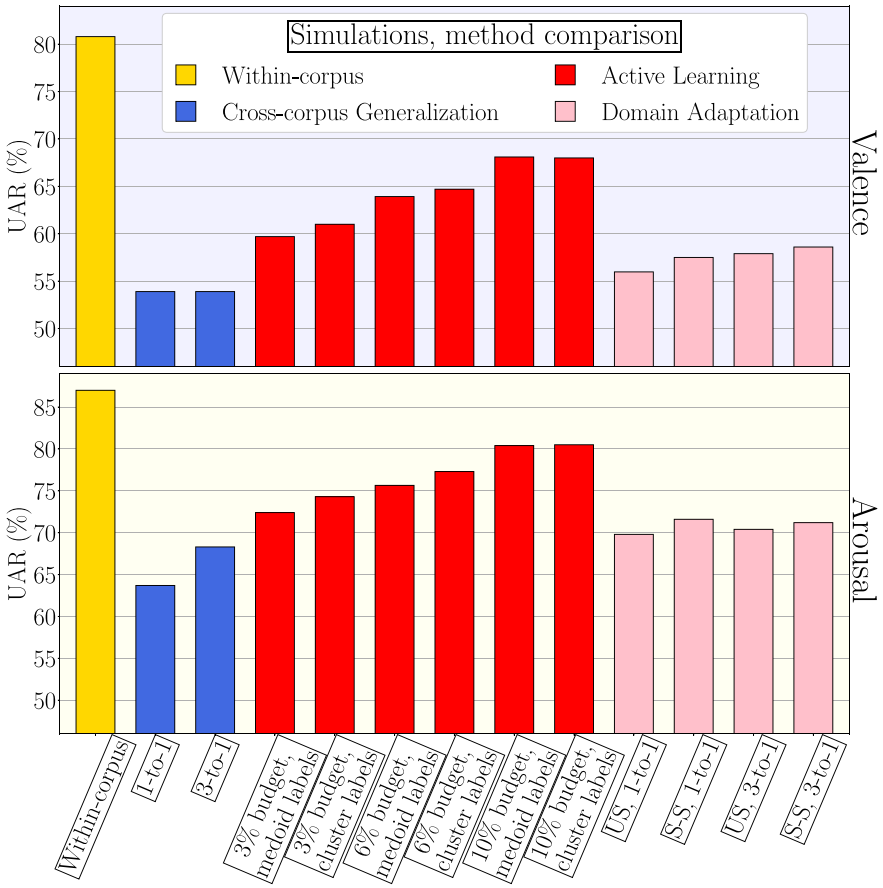

Fig. 5. The comparison of the tested methods in the simulation setup for valence (top) and arousal (bottom) for the log-mel features, which were (on average) the best-performing features, and the only common features for all tested methods. For each method, the mean classification accuracy of the individual results of the four simulation corpora is reported in unweighted average recall (UAR). For the DA experiments, the unsupervised (US) and semi-supervised (S-S) variant of WDA are compared.

labeled target corpus data available when comparing the within-corpus and CCG results. When comparing the 1-to-1 and 3-to-1 CCG results, the latter achieves better results overall. This indicates that a training set consisting of multiple corpora is better for CCG than a training set consisting of a single randomly selected corpus. The 1-to-1 CCG experiments also suggest that having training and testing data from a matching language family results in a better CCG performance. This is since, in most cases, the best training corpus for a Germanic-language corpus was also a Germanic-language corpus (see Table 5.2 of Vaaras (2021) for detailed results). However, in general, there is clearly a need for more advanced methods for new-domain system deployment, as the 3-to-1 CCG performance is rather low compared to the within-corpus performance. From the 1-to-1 and 3-to-1 CCG results in Fig. 4, one can also note that the mean classification accuracies are close to chance level with valence, particularly with the log-mel features.

The results of the AL simulations in Fig. 5 show the effect of the increase in the number of annotations. There is a major performance increase when the labeling budget is increased from 3% to 6%, and another major increase in performance when increasing the number of annotations from 6% to 10%, albeit this increase is not as large as when comparing labeling budgets of 3% and 6%. By further observing the AL simulation results in Fig. 4, there is plenty of variation in the results corpus- and feature-wise. However, on average, the use of cluster labels in MAL leads to accuracy gains. This increase in performance was most noticeable with smaller labeling budgets, which suggests that the use of cluster labels provides accuracy gains most likely in situations when the labeling budget is small. Furthermore, what stands out in the AL experiments as unintuitive is that, in some cases, the classification performance does not monotonically increase and the standard error of the mean (SEM) does not decrease as the number of annotations grows. This may be explained due to the selected distance measure and the properties of the MAL method. The distance measure was selected based on average performance on the simulation corpora and not the corpus-wise performance. The distance measure, in turn, affects the medoid initialization process in MAL.

Regarding the results of the DA simulations in Fig. 4, the 3-to-1 adaptation setting performed better than the 1-to-1 counterpart on average, indicating that a set of multiple source corpora is better for WDA than a single randomly chosen source corpus. The semi-supervised variant of WDA consistently outperformed the unsupervised variant in all experiments. This implies that the use of a small amount of labeled target data helps in selecting the best adapted model in the adaptation process. However, by comparing the AL and DA results in Fig. 5, AL with a labeling budget of 3% outperformed semi-supervised WDA, which utilized a labeling budget of 5%. This suggests that it is more beneficial to annotate data using MAL and use these annotations directly for training a model instead of annotating data for DA purposes.

A comparison of the used features is demonstrated in Fig. 6. All in all, the log-mel features had the best average performance, although, on many occasions, the lower-dimensional GeMAPS and eGeMAPS features come relatively close in performance. In the within-corpus experiments, the log-mel features stood out particularly well. Only when classifying arousal with RAVDESS using a RAVDESS-trained model, the eGeMAPS features outperformed the log-mel features. In CCG, in the GeMAPS and eGeMAPS features worked generally better than log-mel features, except for arousal in the 3-to-1 arousal experiments. Although the reason for this is unclear, it is likely that the lower dimensionality of the GeMAPS and eGeMAPS features and their tailored focus on paralinguistic aspects of speech might alleviate the mismatch between feature

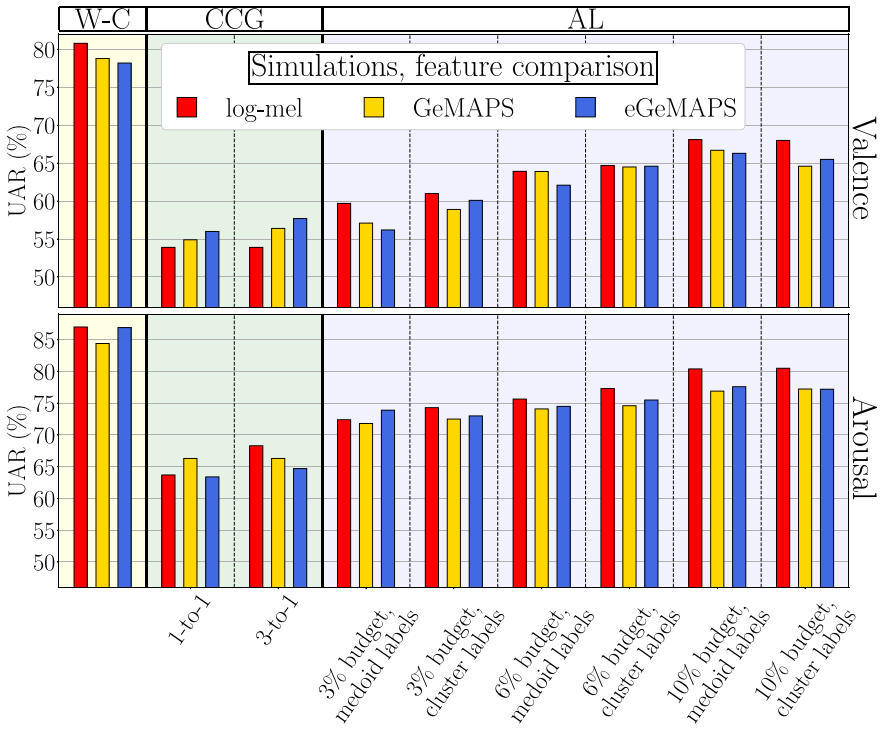

Fig. 6. The comparison of the tested features in the simulation setup for both valence (top) and arousal (bottom) in the within-corpus (W-C), CCG, and AL experiments. For each reported accuracy, the mean result of the four simulation corpora is given in unweighted average recall (UAR). The results for the DA experiments are left out due to the WDA method using only log-mel features.

distributions of training and testing corpora. Otherwise, the log-mel features were superior compared to other features in the within-corpus and AL experiments. These results are also in line with the findings of Eyben et al. (2016), who found that the GeMAPS and eGeMAPS features nearly matched or surpassed the performance level of multiple high-dimensional feature sets in SER.

What can be concluded from the simulation experiments is that for cases when there are no labeled data available, unsupervised DA is superior to CCG in both the 1-to-1 and 3-to-1 adaptation settings (Fig. 5). However, when one can afford to label a small portion of the data (approx. 3% of all samples in a corpus), the present AL method stood out as the best approach. Out of the features, the high-dimensional log-mel features performed the best on average (Fig. 6). Still, even though some general trends regarding the tested SER corpora and methods can be identified from the simulation experiments, there are practically always exceptions to these trends. For example, even though the log-mel features stood out as the best classification features for AL, they did not fare as well in valence classification of FESC or arousal classification in eNTERFACE. This implies that although some methods or features can be identified as being better than others in the simulations, it is not possible to determine in advance what will ultimately be the best performing approach on a completely new dataset without trying them out and validating them on at least some amount of annotated data.

5.1. Class distribution analysis

Since both CCG and DA are affected by properties of the source corpus, one question is whether a mismatch in class distributions between source and target corpora adversely impacts system performance. To study this, we quantified the degree of *distribution mismatch* in class frequencies as

$$d_{\text{mismatch}}(source, target) = |p_{source} - p_{target}|, \tag{5}$$

where p_{source} and $p_{\text{target}} \in [0, 1]$ refer to the relative proportion of positive and neutral valence samples compared to all samples (or proportion of high arousal samples in case of arousal classification) in the source and target corpora, respectively. If $d_{\text{mismatch}} \in [0, 1]$ is 0, both corpora have the same class proportions in the two classes in our present binary classification tasks whereas > 0 indicates a mismatch. The mismatch score was calculated for each possible pair of source and target corpora (12 in total; refer to Table 1) and for both valence and arousal tasks, resulting in a total of 24 mismatch scores. These scores were then compared against the CCG and DA performance in each of the corresponding source-target combinations. The performance scores were averaged across the three compared features for CCG and using unsupervised performance scores for DA. Also, instead of comparing against the original UAR scores, we normalized each of the scores by dividing it by the topline UAR for the given target corpus (the "within-corpus" results in Figs. 4 and 5), as this helps to normalize for irrelevant corpus-dependent variability in the scores. Finally, rank correlation was calculated between the mismatch scores and normalized performance ratings with pooled valence and arousal data points (for improved statistical robustness), thereby testing if lower performance scores were associated with higher mismatches in class distributions.

For AL, the process was somewhat different, as AL performance is only (potentially) affected by the class distribution of the target corpus. Therefore, we parametrized class *imbalance* of each target corpus as

$$d_{\text{imbalance}}(target) = |0.5 - p_{target}| \tag{6}$$

with an analogical definition of p_{target} to Eq. (5). We then compared the imbalance scores to the 8 normalized performance scores of AL for the four target corpora with rank correlation, again pooling valence and arousal scores, and by using the average of cluster and medoid label performance at 10% labeling budget as the performance score of interest.

Fig. 7 shows the results for the class mismatch/balance tests. As can be observed, both CCG and DA performance decrease with increasing class distribution mismatch between the source and target corpus ($r = -0.52$ for CCG, $r = -0.56$ for DA). For AL, there is also a negative performance trend as a function of class imbalance, but the correlation is not

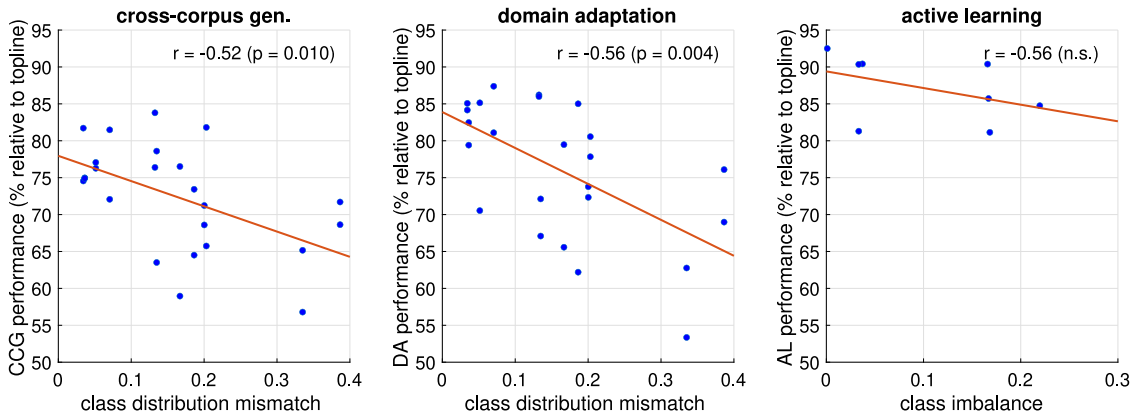

Fig. 7. Dependency of SER performance on the class distribution mismatch between source and target corpora for CCG (left) and DA (middle), or dependency of performance on class balance of target corpus for AL (right). For CCG and DA, each dot corresponds to a unique combination of source corpus, target corpus, and valence/arousal classification. For AL, each dot corresponds to one target corpus with either valence or arousal classification. Correlation r is reported for Spearman rank correlation with a significance criterion of $p < 0.05$ (with n.s. standing for non-significant correlation).

significant, likely due to a limited number of data points. Overall, the analysis shows that class distribution mismatch does affect performance detrimentally. This also happens in the case of DA, which is supposed to adapt to the statistics of the target corpus. In the present case, the mismatch explains a non-negligible part of the SER performance variance across different source-target corpus pairs, and should therefore be taken into account whenever possible. However, a major challenge is that the class distribution of a completely new emotional speech dataset is not generally known, although some prior information on speaking style and context might help to infer some above-chance priors for different emotion categories of interest. Unfortunately, there was not enough data to reliably estimate distributional effects for valence and arousal separately.

6. Experimental setup with NICU-A

6.1. NICU-A child-centered audio dataset

The FinEst NICU Audioset (NICU-A) is a large audio corpus consisting of hundreds of hours of child-centered audio recordings. It was recorded in a collaborative project between Turku University Hospital, Finland, and Tallinn Children's Hospital, Estonia, called the *Auditory environment by Parents of Preterm infant; Language development and Eye-movements* (APPLE) study (Ståhlberg-Forsen et al., 2021). A subset of NICU-A consisting of only Finnish speakers is used as the primary audio material in the present study.

The subset includes families in which both parents had Finnish as their reported mother tongue. It was recorded at the NICU of Turku University Hospital using LENA recorders (https://www.lena.org/) to capture the sound environment of preterm infants in an intensive care unit. LENA consists of both software and a recording device, and is considered as the standard for measuring vocal interactions with children up to three years in age (Xu et al., 2008). The recorder was set next to the child, and parents and nurses were instructed to keep the recorder near the infant in all situations. The recordings were carried out in relatively calm rooms of the NICU, where were only one or two infants with their parents (primary talkers), and, occasionally, nurses and doctors carrying out healthcare routines were present.

Parents of preterm infants were invited in the study if the infant was born below 32 weeks of gestation and did not have acutely life-threatening diseases (not likely to survive until the age of 12 months) or major congenital anomalies or syndrome of clinical significance. The recording was performed when the infant had reached postmenstrual age of 32 weeks.

The data consisted of continuous 16-hour recordings from 43 different participating families (a total of 688 h of audio) with a 16-kHz sampling frequency. By utilizing the broad-class diarization (male/female /key child/other child) of LENA software (Xu et al., 2008), the 16-h recordings were split into utterances. Based on the validity study for the same data reported in Siirilä (2019), adult speech from both 'near' and 'far' categories were included in the analyses to capture caregiver speech, as the data were clean enough to also support the processing on far-field talkers (see Cristia et al. (2020) for general guidelines for the usage of LENA 'far' data). Utterances shorter than 600 ms were discarded from further analysis. This resulted in a total of 129,007 utterances with an average length of 1.57 s (approx. 56 h of speech, corresponding to approx. 8% of the recordings).

Two professionals familiar with the research project carefully selected 35 families for the training set and eight families for the test set, referred to as the gold standard (GS) set. Table 2 showcases the demographics of NICU-A and its training and GS sets. The table includes the number of infants using non-invasive breathing support (CPAP, NIV-NAVA, or nasal high-flow cannula) and the number of infants using invasive breathing support. Additionally, Table 2 includes the information whether the infant was a singleton (i.e. non-twin) or a twin. The sound environment of twins is different compared to singletons, since twins are usually kept together, sometimes even in the same bed. The criterion for selecting families for the training and GS sets was to maximize the representativeness of both data sets in terms of covariates presented in Table 2. After pre-processing the data of NICU-A, both the training set and GS set were partially annotated.

For the training data (101,813 samples), samples were selected for annotation using MAL, similar to the procedure described in Section 4.3.3. The training and validation data for the autoencoder were based on a random split of the unlabeled training set into two sets with a ratio of 80:20 utterances. Two annotators performed labeling for distinct subsets of the data, except for the first 200 samples that were annotated by both in order to measure the inter-rater reliability of the annotations. Each sample was annotated in terms of valence (negative/neutral/positive) and arousal (high/low) in a random order, and using a keyboard-controlled text interface on a computer screen. The utterance was played before each separate task and the user was able to replay utterances without any restrictions. Also, the annotation could be stopped and resumed at any time. A sample could also be labeled as erroneous, if the samples were corrupted by noise, had overlapping speakers, had very short speech fragments, or did not contain any speech at all. Furthermore, the annotator was able to go to the previous utterance at any given time. For the training data, the annotation process took approx. 19 s for each sample (approx. six working days in total).

Similarly, GS annotations were created by three speech or clinical experts for a randomly selected subset of samples for the test data (27,194 samples). Each sample was independently annotated by all

Table 2
The demographics of NICU-A and its training and gold standard (GS) sets at the time of birth (top) and on the recording day (bottom). For each given statistic, either the mean (± standard deviation) is reported, or the number of infants (N) and their proportion (%) of the given set is reported.

		NICU-A (N=43)	Training set (N=35)	GS set (N=8)
Birth	Birth weight (g)	1172 (±400)	1153 (±382)	1153 (±493)
	Gestational week	28.9 (±2.2)	28.9 (±2.1)	29.2 (±2.7)
	Gender, girl	25 (58.1%)	20 (57.1%)	5 (62.5%)
	Singleton (non-twin)	27 (62.8%)	21 (60.0%)	6 (75.0%)
Recording day	Postmenstrual age (weeks)	33.3 (±0.5)	33.3 (±0.6)	33.4 (±0.3)
	Infant was in single family room	23 (53.5%)	18 (51.4%)	5 (62.5%)
	Breathing support:			
	None	18 (41.9%)	15 (42.9%)	3 (37.5%)
	None-invasive	20 (46.5%)	16 (45.7%)	4 (50.0%)
	Invasive ventilation	5 (11.6%)	4 (11.4%)	1 (12.5%)

Table 3
The class distributions of the annotated training and gold standard sets of NICU-A.

	Valence			Arousal	
	positive	*neutral*	*negative*	*high*	*low*
Training set	1509	3391	298	3165	2033
Gold standard set	120	214	11	89	256

three annotators, and the final labels were determined by performing majority voting for the three labels. The samples for which a majority agreement could not be determined were removed from the GS set. For the GS annotations, the annotators had access to 10 s of audio preceding the sample-to-be-annotated in order to obtain a contextual understanding of the communicative situation. However, the annotators were instructed to assign labels based on only the utterance following this 10-second context. It was possible to replay utterances with and without the context with no restrictions. For the GS set, the annotation process took approx. 40 s for each sample, on average (approx. three working days in total for all annotators combined).

After removing all samples which were tagged as erroneous, the sizes of the labeled training and GS sets were 5198 and 345 samples, respectively, corresponding to approx. 4.0% and 0.3% of all samples in NICU-A. By labeling all samples belonging to a cluster based on the cluster's medoid label, the size of the labeled training set was increased to 33,979 samples for the AL experiments involving cluster labels. Table 3 shows the class distributions of the labeled training and GS sets. Although the distributions for valence labels are approximately the same for the training and GS sets, it is evident that the label distribution for arousal is notably different between the two sets. This difference can be explained with MAL as the algorithm selects training samples based on acoustic dissimilarity. Here, the high arousal samples are clearly more acoustically distinct based on the MAL dissimilarity criterion as the algorithm prefers selecting high arousal samples for annotation. Furthermore, considering that the researchers in the APPLE study were interested in the proportion of positive valence over other types of valence, the 'neutral' and 'negative' classes for valence were merged into 'non-positive' for the experiments regarding NICU-A.

The training data inter-annotator agreement rate in terms of kappa score was 0.78 for valence and 0.64 for arousal. For the GS set, kappa scores were 0.48 and 0.28 for valence and arousal, respectively. The kappa score was 0.77 and 0.51 for a binary decision whether a sample was erroneous or not. These scores showcase the inherent difficulty in annotating the emotional content of random samples of real-world speech, even for binary or ternary emotion categories. Overall, based on the given kappa scores, verbal expression of valence can be regarded as more transparent in NICU-A than that of arousal. The difference between the agreement rates in the training and GS set may be explained due to the use of MAL when selecting training samples. With the MAL algorithm, the first 200 samples annotated by both annotators were also the most acoustically distinct samples in the training set, according to the MAL dissimilarity criterion. This can also be observed by investigating the kappa scores of the first 40 samples (0.95) and the last 40 samples (0.59) of the 200 mutual samples for arousal. With a larger number of mutual annotated samples for the training data, the kappa scores of the training set would most likely be closer to those of the GS set.

The data collection was conducted with an ethical permission from the Hospital District of Southwest Finland with decisions no. TO8/027/16 and TO8/049/17. The APPLE study is registered at ClinicalTrials.gov (identifier NCT04826978).

6.2. Experiments with NICU-A

Once the annotations for NICU-A had been acquired, similar experiments as in Section 4 were conducted with NICU-A in order to deploy a SER system for NICU-A and to test the generalizability of the simulated strategies. To better correspond to the labels of NICU-A, the emotional label mapping for the simulation corpora was modified for valence so that 'neutral' and 'negative' were merged into 'non-positive'. All experiments were conducted on the NICU-A GS set.

6.2.1. Cross-corpus generalization experiments with NICU-A

For CCG with NICU-A, experiments corresponding to those in Section 4.3.2 were explored with NICU-A in two settings: 1-to-1 and 4-to-1 CCG. In the 1-to-1 setting, each of the simulation corpora was used individually as the training set, and in the 4-to-1 setting all simulation corpora were used as the training set.

6.2.2. Active learning experiments with NICU-A

In a similar manner as described in Section 4.3.3, an autoencoder was trained on the unlabeled training set of NICU-A and the MAL algorithm was performed for each of the 35 training set families to obtain annotations for the data (Section 6.1). Then, the labeled training samples were used to fine-tune and train an SVM classifier with an RBF kernel, and the trained model was then tested on the GS set. This was done separately for both cluster labels and medoid labels.

6.2.3. Domain adaptation experiments with NICU-A

With similar specifications as in the DA simulations in Section 4.3.4, 1-to-1 and 4-to-1 DA settings were conducted on NICU-A, the former number referring to the number of simulation corpora used as source corpora for WDA. After training M with the source data, the full unlabeled data from the source corpus/corpora and the 96,615 unlabeled

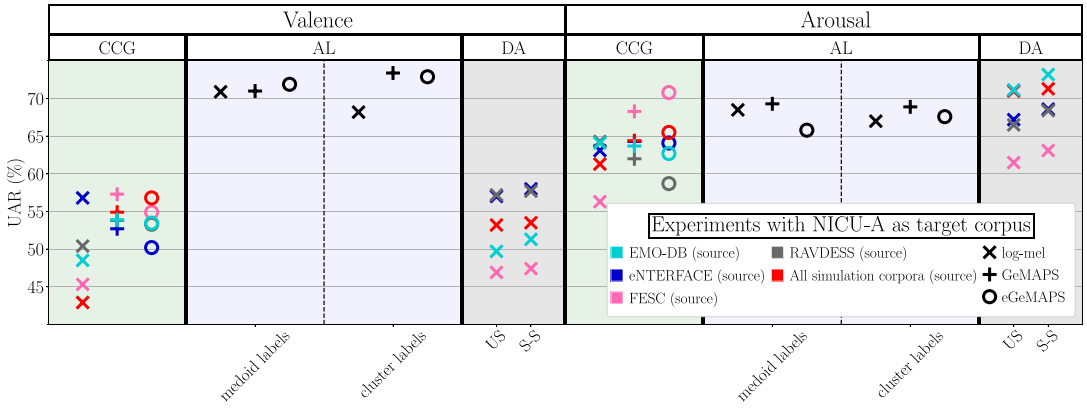

Fig. 8. The unweighted average recall (UAR) performance scores of the experiments conducted on NICU-A for both valence (left) and arousal (right). For each reported result, the NICU-A GS set was used as the test set. For CCG and DA experiments, the results are reported for each variant of training/source data (one/all simulation corpora). For CCG and AL, the results are given for all three features. For AL, medoid and cluster labels are compared, and for DA, the unsupervised (US) and semi-supervised (S-S) variants of WDA are compared.

training samples of NICU-A were used for the second step of the adaptation process. The unsupervised and semi-supervised variants of WDA were trained according to the procedure described in Section 4.3.4, with the accuracy on the 5198 labeled training samples being used as the model selection criterion for the semi-supervised variant. In the 1-to-1 settings, $lr = 5 \cdot 10^{-5}$ was used, except with FESC for valence and with RAVDESS for arousal, where $lr = 7 \cdot 10^{-5}$. For the 4-to-1 settings, $lr = 7 \cdot 10^{-5}$ was used for valence and $lr = 6 \cdot 10^{-5}$ for arousal.

7. Results on NICU-A

Fig. 8 presents the results of the experiments on NICU-A. For the CCG results on valence, the performance level is, on many occasions, close to or below chance level. This indicates that CCG is not a viable solution for NICU-A data for valence classification. The lower-dimensional GeMAPS and eGeMAPS features outperform the log-mel features, and the matching Finnish-language FESC stands out with these features, gaining the best CCG performance for both valence (57.3% UAR, GeMAPS) and arousal (70.8% UAR, eGeMAPS). On the contrary, CCG with FESC performed poorly when using log-mel features. In AL, the GeMAPS and eGeMAPS features achieve the best mean classification accuracy for both valence and arousal. Overall, the use of cluster labels leads to an increase in performance with these features (approx. +1.2%-points UAR), but not with log-mel features. When comparing the results of the matching log-mel features in CCG and DA, the DA method consistently improves from the performance of CCG. In addition, the DA results are on average higher than those of CCG considering all the features, albeit the classification models for FESC and for EMO-DB with valence are clearly not able to adapt properly to NICU-A data. Surprisingly, the matching-language FESC was consistently the worst source corpus for NICU-A in DA, perhaps due to the mismatch between acted emotions and naturalistic affective expressions in an infant caregiving context. Overall, DA was not able to provide a substantial improvement over CCG for valence. For arousal, some of the model adaptations yielded a notable improvement over the CCG and AL results. What stands out is that the best-performing adapted model for arousal (73.2% UAR) was able to outperform all other tested methods by a clear margin. Furthermore, the semi-supervised variant of WDA consistently outperformed the unsupervised variant, similar to the simulation experiments. The confusion matrices for the best-performing models are shown in Fig. 9. All in all, AL performed most consistently with NICU-A and obtained the best performance for valence. However, better results for arousal were obtained by particular configurations of CCG and DA, which was not in line with the results of the simulations, where AL outperformed CCG and DA in both valence and arousal tasks. This can be speculated to be due to the properties of MAL and NICU-A (Section 8).

7.1. Class distribution analysis on NICU-A

Given that we found significant detrimental effects of class distribution mismatch between source and target corpora in our simulation experiments (Section 5.1), we also tested if the obtained performance scores on NICU-A were related to the mismatch in class distributions. To this end, we repeated the CCG and DA analyses described in Section 5.1 for the four possible source corpora with NICU-A as the target corpus. The only difference was that now the scores were normalized with respect to the overall best NICU-A performance obtained for valence and arousal across all the three studied methods (i.e., AL for valence and DA with EMO-DB as the source corpus for arousal). Fig. 10 shows the results of the analysis.

As can be seen from the figure, the performance of CCG or DA does not decrease with an increasing mismatch with the source corpus class distribution. Instead, there appears to be an opposite trend, especially for DA, although correlations are not statistically significant with such a low number of data points. This indicates that the class distribution mismatches do not adversely impact CCG or DA when applying the source models to NICU-A.

As for AL, there is now a higher distribution imbalance in the valence training set ($p_{\text{imbalance}} = 0.21$) and to a somewhat lesser degree in the test set ($p_{\text{imbalance}} = 0.15$), while there is a relatively balanced distribution for arousal in the training set ($p_{\text{imbalance}} = 0.11$) compared to the imbalance in the test set ($p_{\text{imbalance}} = 0.24$). In other words, there is also a larger class distribution mismatch between training and testing data for AL. The AL method appears to find a more balanced set of samples with both low and high arousal for annotation, whereas random sampling of gold standard samples in the test set likely reflects the actual distributional properties of the NICU-A dataset. While this shows that AL is actually operating in a desired manner for sample selection, it may also cause AL-based classifiers to be biased towards erroneous priors on class distributions. On the other hand, we would have expected the same phenomenon to apply to the simulation experiments, but in those AL systematically scored the best in both arousal and valence dimensions. Hence, the reason for superiority of particular configurations of CCG and DA over AL on NICU-A is not completely understood purely in terms of data class distributions.

8. Discussion and conclusions

In this study, we developed a SER system for analyzing the emotional content of speech for unannotated real-life child-centered audio recordings from a NICU. To identify the best approach to deploy a SER system for large-scale unannotated data in terms of required manual effort, three different methods were compared as alternative approaches,

Fig. 9. Normalized confusion matrices for valence (left) and arousal (right) accuracy for the best models on NICU-A. For valence, the best model was an SVM with GeMAPS features and cluster labels from MAL (73.4% UAR). For arousal, the best model was a neural network that was adapted using semi-supervised WDA with EMO-DB as the source corpus (73.2% UAR).

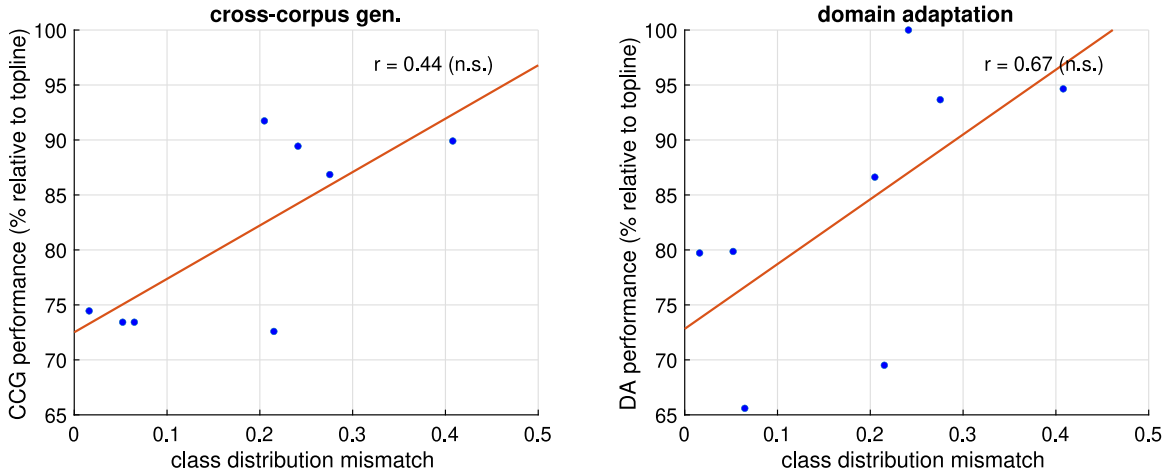

Fig. 10. Dependency of NICU-A SER performance on the class distribution mismatch between different source corpora and NICU-A for CCG (left) and DA (right). Each dot corresponds to a unique combination of a source corpus together with NICU-A as the target corpus. Correlation r is reported for Spearman rank correlation with a significance criterion of $p < 0.05$, and with n.s. standing for non-significant correlation.

namely CCG, AL, and DA. First, simulations were carried out using four already existing SER corpora to fine-tune and compare these three methods for the SER task. Then, these methods were applied to our primary study material, NICU-A, to test how the simulated strategies would work in practice.

The main finding of the simulations was that unsupervised DA worked best for cases without labeled target domain data. However, the present AL method outperformed CCG and DA when even a small proportion of manually annotated data was available (approx. 3% of corpus samples). As a result, we decided to apply AL also for the NICU-A by annotating a subset of its samples as chosen by the AL algorithm, but also comparing CCG and DA with AL on the dataset. As a result, we found that the AL algorithm also performed the best on NICU-A valence classification, hence being in line with the findings from the simulations. However, CCG and DA with specific source corpora combinations outperformed AL in arousal classification, which was not the case for the simulations.

There are at least three potential explanatory factors for the superiority of particular configurations of CCG and DA over AL on NICU-A arousal classification, all related to the properties of MAL and the corpus in question. First, as observed in Zhao et al. (2017), the MAL algorithm benefits from consistent labels, since annotating similar clusters differently might lead to strong confusion between the annotated classes. Therefore, the relatively low inter-rater agreement of NICU-A for arousal (Section 6.1) might have lowered the performance of MAL. Second, the expressed emotions were realistic with NICU-A, whereas they were acted or evoked in the simulation corpora. Emotional categories are more distinguishable from each other in acted emotions than with realistic emotions (e.g. Batliner et al. (2010), Schuller et al. (2011a)), which might make acoustic feature-based grouping of similar emotions into clusters more difficult and hence degrading MAL performance. Third, as analyzed in Section 7.1, the arousal class distributions of the training and test sets were highly dissimilar in NICU-A due to the use of MAL for selecting annotated training samples. This may negatively affect classification performance. However, a similar problem with AL-based sampling bias is applicable to simulation experiments with AL as well, and therefore it is unclear why the sampling bias would specifically have detrimental effects on the NICU-A data but less so on other tested corpora.

Taking together all our experiments with simulations and with NICU-A data, AL performed the best on average, was the most consistent performer, and was not as dependent of specific features or training corpora as CCG and DA. The naive CCG baseline approach performed the worst of the tested methods, and was only successful to some extent with very specific training corpora and feature combinations. The GeMAPS and eGeMAPS features worked best with CCG, although the superiority of one over the other was very case-specific. The DA method outperformed the CCG baseline in both unsupervised and semi-supervised settings, but the method also resulted in large variability in its results across testing conditions. Moreover, it was difficult to know beforehand which source corpus/corpora will result in a good performance, as even the matching Finnish-language FESC was consistently the worst source corpus for NICU-A, and since data class distributions did not seem to explain the performance variability either. Moreover, the success of the present DA method was very dependent on its training hyperparameter configuration. Due to this, we only included the log-mel features in the DA experiments since we could not get the method to work properly with GeMAPS and eGeMAPS features. Overall, these findings suggest that AL should be the primary approach to investigate when developing a novel SER system for similar unannotated large-scale speech data. However, for cases when it is simply not possible to annotate data, unsupervised DA turned out to be the best approach.

Our best models on the NICU-A data achieve a performance level similar to what was obtained in the simulations, which, in turn, were on a par with those reported in earlier SER literature. In addition, the time

it took to annotate training and test data for these models was overall relatively modest, altogether less than 10 working days. Moreover, to the best of our knowledge, our SER system for NICU-A data is the first functional SER system for real-life daylong audio, and also the first one for child-centered LENA data. Therefore, this automated technology creates a basis for large-scale research on how the exposures of emotional dimensions of language affect child development. This is especially relevant in risk populations, such as preterm infants, which are at particular risk for biased language exposures and abnormal later development. Furthermore, SER technology opens up new possibilities to study parenting in the NICU environment. This is because parental expressions of positive emotions towards the infant may be indicators of the parental bonding and activation of parental brain networks that are related to pleasure circuits (Kim et al., 2010).

It should be noted that the present study only considered one specific method for CCG, AL, and DA each, and only a few variants within these methods. To further compare these three approaches in developing a novel SER system, alternative methods should also be tested with the same data. Furthermore, in addition to the features used in the present experiments, other feature representations such as learning features directly from the data should also be tested. In addition, utilizing pretrained SER models could also be applied to the present experiments. Finally, the different experiments in the simulation setup assumed that the annotator is always correct when performing annotations. To better fine-tune these methods to be more robust to errors in the annotations, different levels of noise could be added to the labels in the simulations.

CRediT authorship contribution statement

Einari Vaaras: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Visualization. **Sari Ahlqvist-Björkroth:** Conceptualization, Data curation, Resources, Writing – review & editing. **Konstantinos Drossos:** Methodology, Writing – review & editing. **Liisa Lehtonen:** Conceptualization, Data curation, Resources, Writing – review & editing. **Okko Räsänen:** Conceptualization, Formal analysis, Writing – review & editing, Supervision, Visualization, Funding acquisition.

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 data that has been used is confidential.

Acknowledgments

This research was funded by Academy of Finland grants no. 314573, 314602, 332962, and 335872, and EU Horizon-2020 grant no. 957337 MARVEL. The authors would like to thank the APPLE consortium for the help in the project.

References

Abdelwahab, M., Busso, C., 2018. Domain adversarial for acoustic emotion recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (12), 2423–2435.
Abdelwahab, M., Busso, C., 2019. Active learning for speech emotion recognition using deep neural network. In: Proc. ACII. pp. 1–7.
Airas, M., Alku, P., 2006. Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient. Phonetica 63, 26–46.
Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (Eds.), Proc. ICML, vol. 70. PMLR, International Convention Centre, Sydney, Australia, pp. 214–223.
Batliner, A., Schuller, B., 2013. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. John Wiley & Sons, Incorporated, New York.
Batliner, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., Vogt, T., Aharonson, V., Amir, N., 2010. The automatic recognition of emotions in speech. In: Emotion-Oriented Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 71–99.
Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2009. A theory of learning from different domains. Mach. Learn. 79 (1–2), 151–175.
Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005. A database of German emotional speech. In: Proc. EUROSPEECH, vol. 5. pp. 1517–1520.
Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., Bergelson, E., 2020. A thorough evaluation of the Language Environment Analysis (LENA) system. Behav. Res. Methods.
Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., Schuller, B.W., 2017. An image-based deep spectrum feature representation for the recognition of emotional speech. In: Proc. ACMMM. Association for Computing Machinery, New York, NY, USA, pp. 478–484.
Deng, J., Xia, R., Zhang, Z., Liu, Y., Schuller, B., 2014. Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition. In: Proc. ICASSP. pp. 4818–4822.
Deng, J., Xu, X., Zhang, Z., Frühholz, S., Schuller, B., 2017. Universum autoencoder-based domain adaptation for speech emotion recognition. IEEE Signal Process. Lett. 24 (4), 500–504.
Drossos, K., Magron, P., Virtanen, T., 2019. Unsupervised adversarial domain adaptation based on the wasserstein distance for acoustic scene classification. In: Proc. WASPAA. pp. 259–263.
Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P., 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7 (2), 190–202.
Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proc. ACMMM. pp. 835–838.
Fan, W., Xu, X., Xing, X., Chen, W., Huang, D., 2021. LSSED: a large-scale dataset and benchmark for speech emotion recognition. ArXiv preprint arXiv:2102.01754.
Filippa, M., Monaci, M.G., Grandjean, D., 2019. Emotion attribution in nonverbal vocal communication directed to preterm infants. J. Nonverbal Behav. 43 (1), 91–104.
Immink, K.A.S., Weber, J.H., 2014. Minimum pearson distance detection for multilevel channels with gain and/or offset mismatch. IEEE Trans. Inform. Theory 60 (10), 5966–5974.
Jia, J., Zhou, S., Yin, Y., Wu, B., Chen, W., Meng, F., Wang, Y., 2019. Inferring emotions from large-scale internet voice data. IEEE Trans. Multimed. 21 (7), 1853–1866.
Kim, P., Leckman, J.F., Mayes, L.C., Feldman, R., Wang, X., Swain, J.E., 2010. The plasticity of human maternal brain: Longitudinal changes in brain anatomy during the early postpartum period. Behav. Neurosci. 124 (5), 695–700.
Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proc. ICLR.
Kouw, W.M., Loog, M., 2021. A review of domain adaptation without target labels. IEEE Trans. Pattern Anal. Mach. Intell. 43 (3), 766–785.
Latif, S., Qadir, J., Bilal, M., 2019. Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. In: Proc. ACII. pp. 732–737.
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 (5), 1–35.
Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: Proc. ICML.
Mao, Q., Xu, G., Xue, W., Gou, J., Zhan, Y., 2017. Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. Speech Commun. 93, 1–10.
Martin, O., Kotsia, I., Macq, B., Pitas, I., 2006. The eNTERFACE' 05 audio-visual emotion database. In: Proc. ICDEW. pp. 1–8.
Nyman, A., Korhonen, T., Munck, P., Parkkola, R., Lehtonen, L., Haataja, L., PIPARI Study Group, 2017. Factors affecting the cognitive profile of 11-year-old children born very preterm. Pediatr. Res. (ISSN: 0031-3998) 82 (2), 324–332.
Park, H.-S., Jun, C.-H., 2009. A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. 36 (2), 3336–3341.
Sagha, H., Deng, J., Gavryukova, M., Han, J., Schuller, B., 2016. Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace. In: Proc. ICASSP. pp. 5800–5804.
Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011a. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun. 53 (9–10), 1062–1087.
Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G., 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Trans. Affect. Comput. 1 (2), 119–131.
Schuller, B., Zhang, Z., Weninger, F., Rigoll, G., 2011b. Using multiple databases for training in emotion recognition: To unite or to vote? In: Proc. INTERSPEECH. pp. 1553–1556.
Settles, B., 2012. Active Learning. Morgan & Claypool Publishers, ISBN: 1608457257.
Siirilä, K., 2019. Language Environment Analysis (LENA) -menetelmän validiteetti keskosvauvojen ääniympäristön arvioinnissa (Master's thesis). University of Turku.

Ståhlberg-Forsen, E., Aija, A., Kaasik, B., Latva, R., Ahlqvist-Björkroth, S., Toome, L., Lehtonen, L., Stolt, S., 2021. The validity of the language environment analysis system in two neonatal intensive care units. Acta Paediatr.

Swain, J.E., 2017. Stress-sensitive parental brain systems regulate emotion response and motivate sensitive child care. In: Filippa, M., Kuhn, P., Westrup, B. (Eds.), Early Vocal Contact and Preterm Infant Brain Development: Bridging the Gaps Between Research and Practice. Springer International Publishing, pp. 241–269.

Tieleman, T., Hinton, G., 2012. Lecture 6.5—RMSProp: Divide the Gradient by a Running Average of Its Recent Magnitude. COURSERA: Neural Networks for Machine Learning.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S., 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: Proc. ICASSP. pp. 5200–5204.

Upadhyaya, S., Sourander, A., Luntamo, T., Matinolli, H.-M., Chudal, R., Hinkka-Yli-Salomäki, S., Filatova, S., Cheslack-Postava, K., Sucksdorff, M., Gissler, M., Brown, A., Lehtonen, L., 2021. Preterm birth is associated with depression from childhood to early adulthood. J. Am. Acad. Child Adolesc. Psychiatr. 60 (9), 1127–1136.

Vaaras, E., 2021. Automatic Emotional Speech Analysis from Daylong Child-Centered Recordings from a Neonatal Intensive Care Unit (Master's thesis). Tampere University, Tampere.

Vaaras, E., Ahlqvist-Björkroth, S., Drossos, K., Räsänen, O., 2021. Automatic analysis of the emotional content of speech in daylong child-centered recordings from a neonatal intensive care unit. In: Proc. INTERSPEECH. pp. 3380–3384.

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., Hansen, J., 2008. Signal processing for young child speech language development. In: Proc. WOCCI.

Zhang, B., Kong, Y., Essl, G., Provost, E.M., 2019. F-similarity preservation loss for soft labels: A demonstration on cross-corpus speech emotion recognition. Proc. AAAI 33 (01), 5725–5732.

Zhang, Z., Schuller, B., 2012. Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In: Proc. INTERSPEECH. pp. 362–365.

Zhang, Z., Weninger, F., Wöllmer, M., Schuller, B., 2011. Unsupervised learning in cross-corpus acoustic emotion recognition. In: Proc. ASRU. pp. 523–528.

Zhao, S., Heittola, T., Virtanen, T., 2017. Active learning for sound event classification by clustering unlabeled data. In: Proc. ICASSP. pp. 751–755.

Zhao, Z., Ma, X., 2013. Active learning for speech emotion recognition using conditional random fields. In: Proc. SNPD. pp. 127–131.