Cross-Corpus Acoustic Emotion Recognition with Multi-Task Learning: Seeking Common Ground While Preserving Differences

Biqiao Zhang, *Student Member, IEEE*, Emily Mower Provost, *Member, IEEE*, and Georg Essl, *Member, IEEE*

Abstract—There is growing interest in emotion recognition due to its potential in many applications. However, a pervasive challenge is the presence of data variability caused by factors such as differences across corpora, speaker's gender, and the "domain" of expression (e.g., whether the expression is spoken or sung). Prior work has addressed this challenge by combining data across corpora and/or genders, or by explicitly controlling for these factors. In this work, we investigate the influence of corpus, domain, and gender on the cross-corpus generalizability of emotion recognition systems. We use a multi-task learning approach, where we define the tasks according to these factors. We find that incorporating variability caused by corpus, domain, and gender through multi-task learning outperforms approaches that treat the tasks as either identical or independent. Domain is a larger differentiating factor than gender for multi-domain data. When considering only the speech domain, gender and corpus are similarly influential. Defining tasks by gender is more beneficial than by either corpus or corpus and gender for valence, while the opposite holds for activation. On average, cross-corpus performance increases with the number of training corpora. The results demonstrate that effective cross-corpus modeling requires that we understand how emotion expression patterns change as a function of non-emotional factors.

Index Terms—Emotion recognition, cross-corpus, multi-task learning

1 Introduction

Emotion plays an important role in our perception, attention, memory and decision-making processes [1]. Emotion is not only crucial in human-to-human communication, but also vital in human-computer interaction [2]. For example, Reeves and Nass found that people tend to treat computers as if they are intelligent and emotion-aware [3]. This demonstrates a growing need for agents imbued with proper affective behavior and affective understanding in areas such as interactive robots, story telling agents, computational medical assistants and computer games [4], [5]. One challenge that arises in these real use cases is the presence of variations in emotion expression that occur naturally in the wild, caused by factors including speaker characteristics, languages, lexical content, noise level and recording conditions. Researchers have approximated this challenge by performing cross-corpus analyses [6], [7], [8], [9], [10] and have demonstrated the efficacy of using multiple training corpora for enhancing cross-corpora robustness [6], [8]. However, it is not yet known how to best take advantage of the variability introduced by these training corpora.

- *B. Zhang and E. Mower Provost are with the University of Michigan, Ann Arbor, MI 48109. E-mail: {didizbq, emilykmp}@umich.edu.*
- *G. Essl is with the University of Wisconsin-Milwaukee, Milwaukee, WI 53211. E-mail: essl@uwm.edu.*

Manuscript received 18 Apr. 2016; revised 16 Oct. 2016; accepted 20 Feb. 2017. Date of publication 19 Mar. 2017; date of current version 7 Mar. 2019. Recommended for acceptance by J. Epps.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2017.2684799

There are additional sources of variation that emotion recognition systems need to handle, such as "domain" (e.g., whether the expression is spoken or sung) and gender. Emotion recognition from song and speech are often considered separately. However, our previous work found that one can achieve higher accuracy when training classifiers that allow for information sharing between domains [11]. On the other hand, while most works in emotion recognition use gender-independent systems [12], [13], [14], previous studies have shown that gender-dependent models outperform those that are gender-independent [15], [16], [17]. This suggests that there exist similarities in emotion expression across domains [11] and genders [12], [13], [14], and that the performance of systems increases when controlling for the pervasive differences across the two factors [11], [15], [16], [17].

Most of the previous work in audio emotion recognition addresses the variations caused by corpus, domain and gender differences in two ways: (a) increasing the variations in the training data, such as merging multiple corpora during training [6], [8]; (b) controlling for particular sources of variation in the training data, such as training gender-dependent models [17] or training multiple corpus-specific classifiers and performing late fusion [8]. While multi-task learning has been demonstrated to be useful in affect recognition from visual input [18], [19], [20], its effectiveness on audio emotion recognition is under-explored. In this work, we investigate the influence of corpus, domain, and gender on emotion recognition by combining (a) and (b) using multi-task learning. We hypothesize that we will obtain a

1949-3045 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

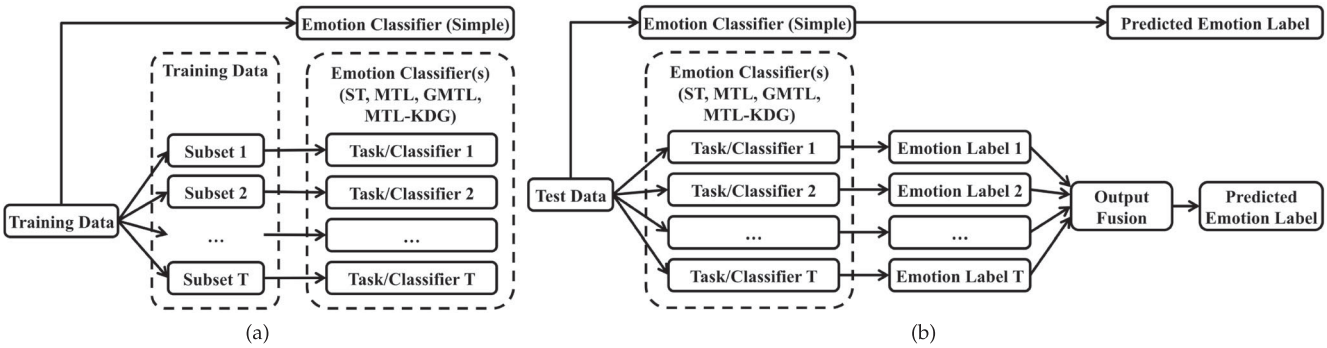


Fig. 1. System diagram for the proposed classification framework, including the: (a) training phase and (b) testing phase. In the simple model, only one classifier is built using all the training data and only one label is generated. In all other models, either T classifiers are trained (ST) or one classifier with T classification tasks is trained (MTL, GMTL and MTL-KDG). T labels are output for each test case and are fused to determine the final label. ST: separate-task model, MTL: multi-task learning model, GMTL: group multi-task learning model, MTL-KDG: multi-task learning with knowledge-driven grouping.

more accurate emotion recognition system, compared to (a) and (b), by seeking common ground across different factors, while preserving the differences in the learned emotion patterns associated with a specific corpus, domain, gender, or their combination. In this paper, multi-task learning refers to jointly training multiple tasks, which contain non-overlapping sets of instances that share a same set of labels.

We explore five models to test our hypothesis: (1) a simple model, where we train a single classifier using all the data; (2) a separate-task (ST) model, where we train task-specific classifiers individually; (3) a multi-task learning (MTL) model, where all the tasks are considered related; (4) a group multi-task learning model (GMTL), where only the intra-group relatedness is assumed and the task grouping is learned with task-specific weights; (5) a multi-task learning with knowledge-driven grouping (MTL-KDG) model, where the group is pre-defined based on knowledge instead of learned as in (4). The first two models are our baselines because they have been shown to be useful for cross-corpus emotion recognition that considers corpora as tasks [6], [8]. This paper extends our previous work that investigated the impact of domain and gender on cross-corpus emotion recognition, using the simple, ST, MTL and GMTL models [21]. In this paper, different experimental settings, additional models and experiments are used. Fig. 1 illustrates the training and testing phase of the proposed methods.

We present two sets of experiments: Experiment 1 investigates the influence of domain and gender using two cross-domain emotion datasets and Experiment 2 explores the influence of corpus and gender using four speech emotion datasets. We do not conduct an experiment with all three factors because the data are insufficient for decoupling domain and corpus in experiments using cross-corpus evaluation. In our experiments, the training data are separated into subsets according to corpus, gender, or domain, where each subset is treated as a task. We perform weighted majority voting to fuse the test labels output by the tasks, where the votes are weighted by a measure of confidence, as in [21].

We find that variations in corpus, domain, and gender all influence emotion recognition. In general, models using multi-task learning methods outperform models that treat the tasks as identical or independent. Data-driven grouping is better than or comparable to knowledge-driven grouping. Domain is a larger differentiating factor than gender when multiple domains are involved, while gender is as important as corpus for single domain data (i.e., speech). Defining tasks by gender is more beneficial than by corpus or both corpus and gender for valence, while the opposite holds for activation. On average, the system performance increases with the number of training corpora. The novelty of this paper includes: (1) an analysis of the benefits of multi-task learning in cross-corpus emotion recognition, with tasks defined by corpus, domain and/or gender; (2) an exploration of effective ways to define tasks for valence and activation; (3) an examination of the influence of sparsity on different feature spaces; (4) a comparison of knowledge-driven and data-driven task grouping.

2 Related Works

2.1 Cross-Corpus Emotion Recognition

Applications of emotion recognition face many challenges, including differences in the acoustic properties of speech due to variations across individuals and recording conditions, among others [6]. Growing attention has been paid to cross-corpus generalizability in speech emotion recognition.

Shami et al. [22] evaluated the generalizability of a segment-based speech emotion recognition method across two corpora, using three settings: within-corpus, cross-corpus, and integrated-corpus (i.e., merging corpora for training and testing). They found that cross-corpus performed the worst, but integrated-corpus was more accurate than within-corpus. Lefter et al. [7] found that cross-corpus performance was higher than within-corpus performance when the intra-corpus training set was very limited, and integrating multiple corpora during training was beneficial. These findings suggest that there are differences between corpora, but that common ground also exists.

Schuller et al. [6] assessed the cross-corpus performance of emotion classification using four normalization methods (i.e., speaker-level, corpus-level, speaker-and-corpus-level, and no normalization) and found that speaker-level normalization performed the best. They also found that cross-corpus performance could be improved by selecting datasets that have large distances between class centers, or selecting instances that are close to class centers [9]. Lefter et al. [23] found that in cross-corpus evaluation, corpus-level normalization was better than normalizing based on

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

the neutral instances of each corpus and that upsampling the sparse class had a positive effect. Vlasenko et al. [24] proposed a phoneme-based emotion classification system and achieved the state-of-the-art cross-corpus performance on two German emotion datasets.

Some works have focused on cross-corpus adaptation. Shah et al. [25] proposed two cross-corpus adaptation methods: (1) removing training instances that are classified incorrectly according to the development set in the test corpus; (2) penalizing the distance between the weights learned on the training corpus and on the development set in the test corpus. They found that both methods increased cross-corpus performance. Abdelwahab and Busso [26] investigated two variants of Support Vector Machines (SVM) for domain adaptation: adaptive SVM and online SVM. They found that for both methods, a significant performance gain could be achieved using only a small portion of the data from the target corpus for adaptation. Similar findings were made in [27] using a domain adaptation method based on the idea of sharing priors between related classes of the source and the target corpora. Song et al. proposed transfer learning variants of two feature learning algorithms: Maximum Mean Discrepancy Embedding [10], and Non-negative Matrix Factorization [28]. They demonstrated the effectiveness of their proposed methods for cross-corpus evaluation on three speech emotion datasets.

Previous works have also investigated methods of enhancing the cross-corpus performance of speech emotion recognition using multiple training corpora. Schuller et al. [8] proposed two methods: (1) merging multiple corpora for training; (2) training one classifier on each of the available training corpora, and fusing the results using majority vote. They showed that both methods improved cross-corpus generalizability, although the preferred method varied across test corpora. In the contrast, Lefter et al. [23] found that for the recognition of negative interaction, training on two merged datasets produced a slightly lower performance than the best performance of training on each dataset separately. Zhang et al. [29] found that adding unlabeled data to merged multi-corpus training data increased the performance of cross-corpus emotion recognition. However, the increase was only approximately 50 percent of the increase brought by adding labeled data.

2.2 Variations in Domain

Although speech and music emotion recognition have traditionally been investigated separately [4], [30], [31], [32], research has demonstrated that there are similarities between music and speech emotion perception and expression. Juslin and Laukka [33] conducted a meta-analysis and found that there are similar patterns in some acoustic features of music and speech emotion expressions. For example, tempo and voice intensity often decrease in sadness and tenderness. Ilie and Thompson [34], [35] found that altering certain acoustic features in music and speech led to similar emotion perception. For example, fast tempo was associated with greater energy in both music and speech. Scherer et al. [36] found a high degree of similarity in the patterns of sung and spoken expressions of emotion. Livingstone et al. [37] found that emotion was expressed similarly in many acoustic cues across the two domains, but that there were also differences in acoustic signals such as vocal loudness, spectral properties and vocal quality [37].

More recently, there has been work recognizing emotion across domains. Coutinho et al. [38] predicted the emotion of music and speech across domains by transferring the feature space. Their results suggested that models trained on one domain could be adapted to the other domain, with a decrease in performance. Our previous work [39] predicted the emotion perception associated with sung and spoken emotion using within- and cross-domain settings. We found that activation was perceived more similarly across domains, compared to valence, and that visual features could better predict emotion perception across domains compared to acoustic features. We explored methods of recognizing emotion from speech and song using a shared model [11]. We showed that multi-task learning, with song and speech as the two tasks, brought benefits to emotion classification for both domains. This suggested that emotion recognition from speech and song can be considered together.

2.3 Variations in Gender

Most research in audio emotion recognition has focused on gender-independent models [12], [13], [14]. However, researchers have analyzed gender variations in emotion recognition. Brendel et al. [40] measured similarity between emotional corpora or sub-corpora of different genders using four similarity measures: recognition rate, correlation, groups of features and feature-ranks. They found that the data were less similar across genders than across corpora when using recognition rate and correlation as measures, yet the opposite held when the latter two measures were used. This suggested that the differences between genders could be as large as the differences between corpora. Alghowinem et al. [41] found that the best features for detecting depression from speech were different for females and males. For example, log energy and shimmer were the most important for females, while loudness was the best feature for males. Vlasenko et al. [42] applied context dependent vowel-level analysis based on gender-dependent features to emotion classification. They showed that the system could detect high-arousal emotions accurately.

Ververidis and Kotropoulos [15] selected relevant features for each gender separately and trained gender-dependent classifiers. Their results showed that the classification accuracy of gender-dependent classifiers was higher than that of a gender-independent classifier. Vogt and André [17] combined gender detection and gender-dependent emotion recognition into a two-stage system. They found that their system increased the emotion recognition rate by 2-4 percent, compared to gender-independent emotion recognition system. Similar observations were made in [43].

2.4 Multi-Task Learning in Affective Computing

Researchers have investigated the effectiveness of multi-task learning for facial and body gestural emotion recognition. Romera-Paredes et al. [18] predicted pain level from facial expression and muscle activity from body gestures by applying multi-task learning in a transfer learning setting (MTL-TL), with subjects as tasks to account for idiosyncrasy. They showed that the proposed model outperformed models without MTL-TL. Another paper [44] proposed a multilinear

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
Dataset Details of UMSSED and RAVDESS

Dataset	Language	Lexical Content	Type	Performer-level			Utterance-level								
				# All	# F	# M	# All	# A	# H	# N	# S	# F	# M	# So	# Sp
UMSSED	English	fixed	acted	3	1	2	168	42	42	42	42	56	112	84	84
RAVDESS	English	fixed	acted	23	11	12	1,288	368	368	184	368	616	672	644	644

F: Female; M: Male; A: Angry; H: Happy; N: Neutral; S: Sad; So: Song; Sp: Speech

multi-task learning method, and showed its effectiveness on synthetic and real data, including recognizing the intensity of facial action units (AUs) associated with pain, with subjects and AUs as tasks in a tensor structure. With the same task definition, Almaev et al. [19] proposed a MTL-TL framework, and showed that it performed well even only with limited labeled data for the target tasks. Shields et al. [20] added multi-task component to the Conditional Restricted Boltzmann Machines (CRBM). They showed that jointly recognizing action, affect, and gender using their proposed model improved the performance of each task, compared to traditional CRBM and other baseline methods.

Related work has indicated that there exist both differences and similarities across corpora, domains, and genders for emotion recognition. In addition, multi-task learning methods have been demonstrated effective in visual affective computing. However, most works in audio emotion recognition either concentrated on increasing data variability (e.g., merging of multiple corpora as the training set), or focused on controlling for variability (e.g., separate classifiers for each available training corpus, gender-dependent classifiers). The design of classification approaches that leverage common ground across different corpora, domains, and genders while preserving the inherent differences is still under-explored.

3 Datasets

The number of publicly available emotion datasets has continued to grow along with the popularity of the field. Early datasets in emotion recognition, such as the Berlin Emotional Speech-Database (EmoDB) [45] and the Danish Emotional Speech Corpus [46], were recorded in laboratory environments with fixed lexical content and acted emotions [6]. The field has recognized the importance of modeling natural behaviors and have introduced new datasets that capture natural displays of affect, including human-robot interaction (e.g., FAU Aibo [47]), or recordings taken from public media (e.g., VAM [48]). Researchers have also focused on emotion induction as a technique to elicit emotional behaviors (e.g., SEMAINE [49], [50]). Recent acted datasets have included altered elicitation protocols to increase naturalness, for example, using improvisation (e.g., IEMOCAP [51]), increasing the diversity of speakers' cultural backgrounds (e.g., eNTERFACE [52]), or including additional domains (e.g., RAVDESS [53]).

We select six datasets covering different types of emotion (acted and spontaneous), languages (English and German) and domains (speech and song) to investigate the cross-corpus generalizability of our proposed methods. We conduct two sets of experiments concentrating on the variations caused by: (1) domain and gender, and (2) training corpus and gender. In (1), we use the University of Michigan Song and Speech Dataset (UMSSED) [39] and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [53], as they are the only available emotion corpora that contain both speech and song. These experiments use categorical emotion labels. In (2), we use EmoDB and the eNTERFACE dataset to represent acted emotion in German and English, respectively, and the Vera am Mittag German Audio-Visual Emotional Speech Database (VAM) and the AVEC (2011) corpus [54], [55] to represent spontaneous emotion, again in German and English, respectively. These experiments use binary labels of valence (negative versus positive) and activation (calm versus excited). The meta information about the datasets can be found in Tables 1 and 3.

UMSSED contains audio-visual recordings of song and speech with fixed sentences, produced by three performers. During recording, each sentence was embedded into four passages that were intended to evoke anger, happiness, neutrality, and sadness. For the song recordings, each sentence is accompanied by a unique melody that is the same across the four emotional variations. The final dataset contains 168 utterances. We use the emotion target provided to the actors to match the labels available in RAVDESS. See [39] for additional details.

RAVDESS contains emotional audio-visual recordings of song and speech with fixed lexical content, collected from 24 performers. The six emotions for song are neutrality, calmness, happiness, sadness, anger, and fear. There are two additional emotions, disgust and surprise, for speech. The data were collected at two emotional intensities (except for the class of neutrality). Three melodies were composed for positive, neutral, and negative emotions for the singing performances. We only use utterances with anger, happiness, neutrality, and sadness to match UMSSED. We drop one performer with missing data. This results in 1,288 utterances. We use the target emotion labels because the perception evaluations had not been fully released at the time of this experiment. More details can be found in [37], [53], [56].

EmoDB consists of audio recordings of 10 German speakers reading lexically neutral sentences in seven emotions: anger, boredom, disgust, fear, joy, sadness and neutrality.

TABLE 2
Mapping from Categorical Emotions to Binary Valence and Activation

Emotion	Appearance	Valence	Activation
Anger	EmoDB, eNTERFACE	−	+
Happiness	EmoDB, eNTERFACE	+	+
Neutrality	EmoDB	+	−
Sadness	EmoDB, eNTERFACE	−	−
Fear	EmoDB, eNTERFACE	−	+
Disgust	EmoDB, eNTERFACE	−	−
Surprise	eNTERFACE	+	+
Boredom	EmoDB	−	−

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

TABLE 3
Dataset Details of EmoDB, eNTERFACE, VAM and AVEC

Dataset	Language	Lexical Content	Type	Speaker-level			Utterance-level						
				# All	# F	# M	# All	# V (+)	# V (−)	# A (+)	# A (−)	# F	# M
EmoDB	German	fixed	acted	10	5	5	493	142	351	246	247	286	207
eNTERFACE	English	fixed	acted	43	9	34	1,287	427	860	857	430	270	1,017
VAM	German	natural	spontaneous	47	36	11	947	72	875	445	502	751	196
AVEC	English	natural	spontaneous	16	10	6	2,368	1,534	834	1,280	1,088	1,620	748

F: Female; M: Male; V: Valence; A: Activation

The utterances were labeled using the target emotion. Utterances with a recognition rate higher than 80 percent and naturalness higher than 60 percent during human evaluation were kept. This results in 493 utterances. We map the seven categorical emotions to binary valence and activation labels (see Table 2), following [6], [8], in order to match the labels in VAM and AVEC. More details can be found in [45].

The eNTERFACE dataset consists of speech with fixed lexical content in six emotions: angry, happy, fearful, sad, disgust, and surprise. The emotions of speakers were elicited by short stories during the recording. The released dataset contains audio-visual recordings of 44 speakers from 14 different nations that were assessed as emotionally unambiguous by two experts. In this paper, we drop the data of speaker number 6 because the recordings are not segmented. This results in 1,287 utterances from 43 speakers. Again, the categorical emotions are mapped to binary valence and activation (Table 2). See [52] for more details.

VAM consists of spontaneous emotional speech from a German TV talk-show. We use the VAM-Audio portion of the corpus, which contains audio recordings from 47 speakers that were evaluated as "very good" or "good" by human evaluators in terms of the usability for emotion analysis. The recordings were provided as utterances, which are mostly complete sentences, but also include some exclamations, affect bursts, and incomplete sentences, because of the spontaneous nature of the data. This results in 947 utterances. The utterances were continuously labeled by human evaluators (17 for speaker 1-19, 6 for speaker 20-47) on valence, activation and dominance (weak versus strong). In this paper, we use only the valence and activation evaluations. We take the sign of the mean valence and mean activation of each utterance as the binary labels, following the process in [8]. See [48] for more details about the VAM corpus.

AVEC (2011) was created from the Solid-SAL partition of SEMAINE [49], [50]. It contains interactions between users and four emotionally stereotyped characters played by human operators. In this paper, we use the training and development set of AVEC. This results in 63 sessions, where each session is the interaction between a user and a character. Each interaction was fully transcribed, and was annotated by at least two raters along the dimensions of valence, activation, expectation (expecting versus being taken unaware) and power (weak versus strong). Binary word-level labels are provided for each dimension. In this work, we use the valence and activation dimensions. We segment the recordings data into turns and generate the turn-level emotion labels from the word-level labels using majority vote. Additional information about the AVEC and SEMAINE datasets can be found in [49], [50], [54], [55].

4 Classification Models

We present five classification models: the simple model, separate task (ST) model, multi-task learning (MTL) model [57], [58], group multi-task learning (GMTL) model [59], and multi-task learning with knowledge-driven grouping (MTL-KDG) model [57], [58]. We define a task as emotion recognition using data from a specific factor (e.g., a corpus), or a specific combination of two factors (e.g., a corpus-gender pair). The five classification models correspond to five different assumptions about the tasks. The simple model assumes that the tasks are identical, and merges data from all the tasks for training. The ST model sees the tasks as independent, and trains a separate classifier for each task. The simple and ST models are similar to the "pooling" and "voting" strategies in [8], respectively, if we consider each corpus as a task. Therefore, we use simple and ST as baselines in our experiments. The MTL model assumes that the tasks are related and share a common sparse feature representation. The GMTL model assumes that the tasks can be clustered into groups, and only intra-group information sharing is allowed. Finally, the MTL-KDG model assumes that information is shared within a group, but it predefines groups based on knowledge such as domain, gender, or corpus, instead of learning the groups from data.

We use linear Support Vector Machine (SVM) in the simple and ST models, as in previous cross-corpus emotion recognition works [6], [8], [9]. We adopt two types of regularization: L_2-regularization ([6], [8], [9]), and L_1-regularization, which assumes sparsity of the features. The MTL model and each group of tasks in the MTL-KDG model use the multi-task feature learning algorithm [57], [58]. The GMTL model uses the group multi-task learning algorithm [59]. These two algorithms are used in our prior work [21]. Details on the two algorithms are provided in the following sections.

4.1 Multi-Task Feature Learning

The multi-task feature learning algorithm [57], [58] learns a common feature representation across tasks using the $L_{2,1}$-norm regularization, which enforces sparsity of the features across tasks. There are two settings of this algorithm: (a) feature learning (FL) and (b) feature selection (FS). The major difference between them is that in (a), the $L_{2,1}$-norm regularization is imposed on a transformed feature space, while in (b) the regularization is imposed directly on the original feature space.

The objective function of setting (a) is given by Eq. (1). It is assumed that the weight matrix, W, whose column vectors are the weights $\mathbf{w_t}$ of individual tasks, can be rewritten into $W = UA$, where $U^T U = I$ (identity matrix) and A is the weight matrix for a transformed feature space.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

$$\min_{U,A} \sum_{t=1}^{T} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{a_t}, U^T \mathbf{x_{ti}} \rangle) + \gamma \|A\|_{2,1}^2. \quad (1)$$

Eq. (1) contains two terms: the loss term (first) and the regularization term (second). The loss term is the summation of the loss, $L(.)$, across T tasks. Here, m_t is the number of training instances in task t, $y_{ti} \in \{-1, 1\}$ is the label of the ith instance in task t, $\mathbf{a_t}$ is tth column of A, $\mathbf{x_{ti}}$ is the ith training instance of task t, and $< >$ stands for inner product. The regularization term is the product of the regularization parameter γ and the squared $L_{2,1}$-norm of A. $L_{2,1}$-norm is defined as the L_1-norm of the vector produced by taking the L_2-norm of each row of A.

Setting (b) is a special case of (a). In (a), U and A are learned together from the data, while in (b), we force $U = I$. In this way, the "feature learning" in (a) reduces to the special case of "feature selection" in (b) [57], [58].

The problem given by Eq. (1) is non-convex. However, [57], [58] proved that it has an equivalent convex form that can be solved by iteratively minimizing over W (Eq. (2)) and a $d \times d$ matrix D, where d is the dimensionality of the input features. Specifically, we first initialize D to $\frac{I}{d}$, and then iteratively perform two steps:

- Fix D, solve the task-specific optimization by Eq. (3).
- Fix W, update D using Eq. (4) for setting (a) or Eq. (5) for setting (b). The ϵ in Eq. (4) is a perturbation parameter used to ensure the convergence of the problem. The $\mathbf{w^i}$ in Eq. (5) denotes the ith row of W.

$$\min_{W} \sum_{t=1}^{T} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \sum_{t=1}^{T} \langle \mathbf{w_t}, D^{-1} \mathbf{w_t} \rangle. \quad (2)$$

$$\mathbf{w_t} = \arg\min \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \langle \mathbf{w_t}, D^{-1} \mathbf{w_t} \rangle. \quad (3)$$

$$D = \frac{(WW^T + \epsilon I)^{\frac{1}{2}}}{trace(WW^T + \epsilon I)^{\frac{1}{2}}}. \quad (4)$$

$$D = Diag(\lambda), \text{where } \lambda_i = \frac{\|\mathbf{w^i}\|_2}{\|W\|_{2,1}}. \quad (5)$$

Eq. (3) holds for any convex loss function. In this paper, we choose the hinge loss (Eq. (6)) to match the linear SVM used in the simple model and ST model, as in our prior work [21]. Note that Eq. (3) with hinge loss is equivalent to linear SVM with a variable transformation trick.

$$L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) = \max(0, 1 - y_{ti} \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle). \quad (6)$$

4.2 Group Multi-Task Learning

Group multi-task learning [59] assumes that the tasks belong to several groups that can be learned together with task-specific weights. Only the tasks that are grouped together share information. This method was built directly on the multi-task feature learning algorithm above. In [58], it was proved that the optimization problem given by Eq. (1) is equivalent to Eq. (7), where $\|W\|_{tr}^2 = trace(WW^T)$

$$\min_{W} \sum_{t=1}^{T} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \|W\|_{tr}^2. \quad (7)$$

Analogous to Eq. (7), the objective function of group multi-task learning becomes Eq. (8) given the group assignments. Here, G is the number of groups. Q_g is a diagonal matrix with diagonal entries being the binary group assignment values for group g, and $\sum_g Q_g = I$. The optimal G is not known a priori and is treated as a hyperparameter. When $G = T$, group multi-task learning is equivalent to solving each task individually, and when $G = 1$, it is the same as the multi-task feature learning.

$$\min_{W,Q} \sum_{t=1}^{T} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \sum_{g=1}^{G} \|WQ_g\|_{tr}^2. \quad (8)$$

Eq. (8) is a mixed integer programming problem. It can be solved by iteratively performing two steps:

- Fix Q, solve group-specific optimization given by Eq. (9). $W_g = WQ_g$, and q_{gt} is the tth diagonal entry of Q_g.
- Fix W and solve for Q. See details in [59].

$$\min_{W_g} \sum_{t:q_{gt}=1} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \|W_g\|_{tr}^2. \quad (9)$$

The second step is non-convex, and the solution could become stuck in a local optimum. We address this problem by training multiple times and fusing the labels.

All other models, except for the simple model, learn a different weight vector for each task. Therefore, there could be T predicted labels for a given test instance. Although it is common in the multi-task learning literature to assume knowledge about the tasks of the test data [57], [58], [59], we do not make this assumption. This is because: (1) it requires additional information about the test data, which may not be available in real applications; (2) the test data may not strictly belong to any of the tasks (e.g., test data is from an unseen corpus when using each training corpus as a task). In this paper, we generate the final output label by weighted majority vote, where each task gives a vote to the label it outputs, weighted by the distance to the decision hyperplane. This method was demonstrated to outperform other output selection or fusion methods in [21].

5 Experimental Design

We designed two sets of experiments, both of which use cross-corpus evaluation. Experiment 1 investigates the influence of domain and gender on emotion recognition using multi-domain data. Experiment 2 investigates the influence of corpus and gender on emotion recognition using speech data. We hypothesize that we can achieve better performance by splitting data into tasks, and controlling for the degree of information sharing between the tasks using multi-task learning. We solve the linear SVMs using Liblinear [60]. We use a fixed number of iterations as the stop criteria for multi-task learning, as in [61], [62]. For multi-task feature learning, we fix the number of iterations to 20, according to [58]. For group multi-task learning, we fix the outer-iteration to five as in the example code from the author of [59], and the group-specific inner-iteration to 20. The detailed experimental settings are described below.

5.1 Experiment 1: Domain and Gender

We analyze the influence of domain and gender in experiment 1 using four categorical emotion labels: angry, happy,

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

neutral, and sad. We use RAVDESS for training and UMSSED for testing, because UMSSED has very limited number of instances and may not be sufficient for training.

We conduct three sub-experiments to understand the impact of domain (1d), of gender (1g), and of the combination of domain and gender (1dg). Experiment 1d treats expressions from speech and song as two tasks. We train classifiers using the simple, ST and MTL models, which posit three different relationships between the emotion expression across domains: identical (simple), different (ST), or related (MTL). Similarly, experiment 1g analyzes the impact of gender, using female and male as tasks. Experiment 1dg focuses on the joint influence of domain and gender. The four tasks in experiment 1dg correspond to the four domain-gender pairs. In addition to simple, ST and MTL, we train classifiers with the GMTL and MTL-KDG models. Both models assume that closely related tasks can be grouped together and can share information. The difference is that the grouping in GMTL is data-driven, while the grouping in MTL-KDG is knowledge driven. In the MTL-KDG model, we group the tasks by domain (denoted as MTL-GD) and gender (denoted as MTL-GG), respectively. Experiment 1dg is similar to our previous work [21], but has additional classification models and different parameter tuning process. Note that the simple model is the same across sub-experiments.

We solve the multi-class classification problem as the combination of binary classifications using the one-against-one strategy. For each pair of emotions, a binary label is predicted following the process in Fig. 1b. The final multi-class label is generated from the six binary predictions of pairwise emotions using majority vote. We select the class with the smallest index in the occurrence of ties.

We extract the 65 frame-level low-level descriptors (LLDs) in the ComParE feature set [63] (see Table A.1 in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2017.2684799), using openSMILE [64]. We remove unvoiced portions (where $F0 = 0$) of the data, and calculate nine statistics over the LLDs and ΔLLDs, including mean, standard deviation, max, min, range, interquartile range, mean absolute deviation, skewness and kurtosis. This results in 1,170 utterance-level acoustic features. We apply speaker-dependent z-normalization to the features, because of its effectiveness in cross-corpus experiments [6].

We tune the hyper-parameters by maximizing the leave-one-speaker-out (LOSO) cross-validation accuracy of the training set, in the ranges below:

- Regularization parameter γ (in simple, ST, MTL, GMTL and MTL-KDG): $\{10^{-4}, 10^{-3}, \ldots, 10^{3}\}$. Note that γ is equivalent to the cost parameter C for the error term in linear SVM (simple and ST), where $C = 1/(2 \times \gamma)$.
- Permutation parameter ϵ (in FL setting of MTL, GMTL and MTL-KDG): $\{10^{-8}, 10^{-7}, \ldots, 10^{0}\}$.
- Number of Groups G (in GMTL): $\{1, 2, \ldots, T\}$.

5.2 Experiment 2: Corpus and Gender

In experiment 2, we analyze the influence of corpus and gender on the binary (positive versus negative) classification of valence and activation, using four speech emotion datasets.

We conduct three sets of sub-experiments, focusing on the impact of corpus (2c), of gender (2g) and of the combination of corpus and gender (2cg). In experiments 2c and 2g, we train the simple, ST, and MTL models using corpora and genders as tasks, respectively. Experiment 2cg investigates the joint impact of corpus and gender, with each corpus-gender pair as a task. In 2cg, we also train GMTL and MTL-KDG in addition to simple, ST and MTL. We group the tasks by corpus and gender for MTL-KDG, denoted as MTL-GC and MTL-GG below. In each sub-experiment, we use one, two or three corpora for training, and test on each of the remaining corpora separately. Note that the simple model is the same across all sub-experiments when the same training corpora are used. When there is only one training corpus, all models in 2c are identical to the simple model and experiment 2cg is not performed.

In each sub-experiment, we compare the performance of different models to test the underlying assumptions. We compare the performance of ST and MTL across the sub-experiments, to investigate the three ways of defining the tasks (corpus, gender, or corpus-gender pair). In addition, we compare the performance on the same test corpus when using different numbers of corpora for training to investigate the impact of adding additional training corpus.

As we have more data in experiment 2, we use a larger feature set. We extract the "emo_large" feature set from openSMILE, as in [8]. It consists of 6,669 features, generated from 57 acoustic LLDs (listed in Table A.2 in the Appendix, available online) by calculating 39 statistics over the LLDs, ΔLLDs, and $\Delta\Delta$LLDs. We apply speaker-dependent z-normalization to the utterance-level features.

We use the same parameter ranges discussed in experiment 1. We use 5-fold cross-validation, where the folds are divided at speaker-level for each task to avoid overfitting to known speakers. In the cross-validation process, we use average UAR of the tasks as the performance measure if the model contains more than one task, because the data are not evenly distributed across the tasks.

6 Results

6.1 Experiment 1: Domain and Gender

Table 4 shows the results in experiment 1, including unweighted average recall (UAR) of the cross-corpus four-class emotion classification, and the LOSO within-corpus UAR.

6.1.1 Comparing Different Versions of the Models

There are two versions of each model: the L_1-regularization (L_1) and L_2-regularization (L_2) of the simple and ST models, and the feature selection (FS) and feature learning (FL) settings of the MTL, GMTL and MTL-KDG models. The UAR of the two versions are shown in the first and second rows of Table 4, respectively. For the simple and ST models, L_1 assumes sparsity on the feature space, while L_2 does not. For the MTL, GMTL and MTL-KDG models, FS assumes feature sparsity on the original space, while FL assumes that there is a transformed feature space where the features are sparse. We compare L_1 versus L_2 and FS versus FL to investigate whether a sparse representation on the original feature space that generalize well can be found in multi-domain data.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

TABLE 4
Experiment 1. UAR of the Four-Class Clasification Task (%)

Version	Cross-corpus										Within-corpus
	Simple	Task: Domain		Task: Gender		Task: Domain-gender Pair					
		ST	MTL	ST	MTL	ST	MTL	GMTL	MTL-GD	MTL-GG	
L_1/FS	45.8	**58.9**	52.4	50.0	**51.8**	57.1	57.1	**60.1**	56.6	54.8	48.8
L_2/FL	41.1	53.0	51.2	46.4	47.6	54.8	53.0	52.4	53.0	54.2	**54.8**

The chance performance is 25%. The overall best result is underlined. The best result within each sub-experiment is bolded. L_1: L_1-regularization; L_2: L_2-regularization; FS: feature selection setting; FL: feature learning setting.

We find that in the cross-corpus results of experiment 1, L_1 consistently outperforms L_2, and FS consistently outperforms FL. These results support the assumption of feature sparsity on the original feature space. The higher performance of L_1 and FS may stem from the within-corpus nature of the tasks. The training tasks share the same performers (in exp. 1d), lexical content, and recording conditions. In addition, both datasets contain acted emotion in the same language. This may increase the transferability of the sparse feature representation learned from one corpus to the other. The rest of the analyses on experiment 1 (Section 6.1) use L_1 (simple and ST) and FS (MTL, GMTL and MTL-KDG).

6.1.2 The Influence of Model and Task Definition

When domain is used as the task (experiment 1d), ST achieves the highest UAR. When gender is used (experiment 1g), MTL achieves the highest UAR. This may indicate that emotion is expressed more similarly across genders than across domains. In other words, domain is a stronger differentiating factor than gender. This is supported by experiment 1dg, which shows that grouping the tasks by domain outperforms grouping the tasks by gender.

The best performance is achieved in experiment 1dg with GMTL. This suggests that it is beneficial to consider variations in both domain and gender. GMTL outperforms both ST and MTL, indicating that the tasks separated by domain and gender are partially related, and the close relationship between the tasks only happens within group. Interestingly, the data-driven grouping (GMTL) has higher UAR than the knowledge-driven grouping (MTL-GD, MTL-GG). This may suggest that closeness between the tasks does not exactly correspond to the obvious grouping factors. Another possible explanation is that GMTL has access to all training instances, while in MTL-KDG, each classifier only has access to the instances within a group (i.e., a single domain for MTL-GD, or a single gender for MTL-GG).

6.1.3 Cross-Corpus versus Within-Corpus

The best cross-corpus UAR, 60.1 percent, is higher than the best within-corpus UAR of UMSSED, which is 54.8 percent. This performance difference is not significant when tested using paired t-test on the per-performer UAR. This may be due to the small number of performers in UMSSED (three performers). However, using GMTL-FS with domain-gender pairs as tasks improves the UAR of two out of three performers by more than 8 percent. This indicates that the proposed cross-corpus approach can outperform, or achieve comparable results to, models trained using the within-corpus setting with limited data, despite the corpus variations in performer, lexical and melodic content, recording conditions and noise level.

6.2 Experiments 2: Corpus and Gender

In experiment 2, we analyze the binary classification results of valence and activation on four speech emotion datasets. We compare the performance between: (1) different assumptions on feature sparsity, (2) different training-testing combinations, (3) different models while controlling for task definition (e.g., corpus as the task), (4) different task definitions while controlling for model, (5) different number of training corpora while controlling for model and task definition, and (6) cross-corpus and within-corpus.

We use a repeated measure model (denoted as RM) with mixed factors for the comparisons. We treat the test corpus (e.g., EmoDB) as the between-subject factor because there are multiple experiments run on each test corpus. Thus, the overall set of results has underlying dependencies. The within-subject factors (denoted as WSF) include: version (e.g., L_1-regularization), model (e.g., ST), task definition (e.g., gender), and number of training corpora.

After fitting the results into an RM, we perform the repeated-measure ANOVA (denoted as RANOVA) for each dimension (i.e., valence and activation). If the WSF is significant, we perform the Tukey's honest significant difference test (denoted as Tukey test), which is a pairwise comparison between different values of the WSF using the model statistics of RANOVA.

6.2.1 Comparing Different Versions of the Models

We first investigate if a sparse representation on the original feature space can be found across corpora and genders for speech emotion data. We compare the UARs as a function of regularization (L_1 versus L_2) for the single-task methods (simple and ST), and feature handling (FS versus FL) for the multi-task methods (MTL, GMTL and MTL-KDG).

We use two RMs to compare: (1) L_1 versus L_2 using all the experimental results of simple and ST, and (2) FS versus FL using all the experimental results of MTL, GMTL and MTL-KDG. We use the version of the model as the WSF. For L_1 versus L_2, the influence of regularization is significant for valence (RANOVA, $F(1, 84) = 4.3$, $p = 0.042$), but not for activation. The Tukey test (Fig. 2) shows that L_2 is significantly better than L_1 for valence ($p = 0.042$). For FS versus FL, the influence of feature handling is significant for both valence ($F(1, 104) = 19.2$, $p = 2.8e\text{-}05$) and activation ($F(1, 104) = 12.8$, $p = 5.3e\text{-}04$). FL significantly outperforms FS for valence

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

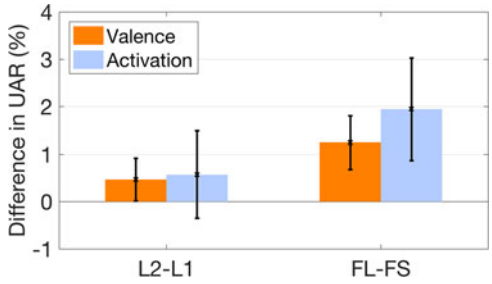


Fig. 2. Experiment 2. The left bars show the difference in UAR as a function of regularization (L_2-regularization minus L_1-regularization) in the simple and ST models. The right bars show the same difference as a function of feature handling (feature learning minus feature selection) in MTL, GMTL and MTL-KDG. The black lines represent the 95% confidence interval of the Tukey's honest significant difference test.

(Tukey test, p = 2.8e-05) and activation (p = 5.3e-04), as shown in Fig. 2.

These results indicate that we cannot find a sparse feature representation on the original feature space that transfers well across corpora, which is different compared to the results from experiment 1. This may be because nearly six times more features are used in experiment 2. The orthogonal projection decouples the original features by "collapsing" similar information onto the same dimension. Therefore, the sparse representation on the new feature space might be able to keep more emotion-related information. In addition, there are higher variations in languages, lexical content, speakers and recording conditions within each individual task, and across tasks. As a result, the emotion-related patterns on the original feature space may be further masked. Therefore, we may need to learn a feature space where a sparse representation that generalizes well from the data.

For simplicity, we only present results of the L_2-regularization (simple and ST) and the feature learning setting (MTL, GMTL and MTL-KDG) for the rest of the analyses on experiment 2. We present the UAR of simple, ST and MTL with a single training corpus in Table 5 (only experiment 2g has multiple tasks). We present the UAR of all models when using multiple training corpora, with corpora, genders and corpus-gender pairs as tasks in Table 6.

6.2.2 Different Corpus as Training Set

While we treat corpus as a single factor in this experiment, it includes variations in language, type of emotion, in addition to recording condition. We compare the cross-corpus UAR of classifiers trained on each dataset, and each combination of two datasets, to get some insights on the impact of language and type of emotion (i.e., acted and spontaneous).

When a single corpus is used for training, we find that the models trained on VAM achieve the highest UAR (bolded in Table 5) on EmoDB, eNTERFACE and AVEC for activation. We also find when testing on eNTERFACE, models trained on VAM outperform cross-corpus models from the literature (e.g., [29] achieved a maximal UAR of 63.9 percent, Table 7). This is surprising because eNTERFACE is in a different language and with a different type of emotion. There may be several reasons. First, VAM contains recordings about personal and very emotional topics (e.g., paternity questions or affairs) [65], which makes the content more emotionally expressive. Second, VAM only contains the speakers evaluated as "very good" and "good", which also results in data that are more emotionally expressive. In addition, the speakers in eNTERFACE are from 14 different nations. This may reduce the advantage of training on corpus with the same language, due to the presence of different accents in the testing data.

TABLE 5
Experiment 2. UAR (%) of Valence and Activation Using a Single Training Corpus

Test on	Train on	Valence			Activation		
		Simple	Task: Gender		Simple	Task: Gender	
			ST	MTL		ST	MTL
EmoDB	eNT	56.1	**61.0**	60.7	72.0	78.7	75.5
EmoDB	VAM	48.1	46.3	47.1	86.4	**87.8**	81.1
EmoDB	AVEC	52.3	52.6	52.5	51.5	58.6	54.0
eNT	EmoDB	49.6	48.0	48.9	63.2	65.5	63.7
eNT	VAM	49.3	47.1	47.9	66.2	66.7	**69.8**
eNT	AVEC	54.5	53.3	**56.0**	54.3	62.1	69.5
VAM	EmoDB	50.6	49.4	50.0	67.7	68.0	68.2
VAM	eNT	**59.3**	56.8	56.7	61.7	61.1	58.6
VAM	AVEC	51.4	56.0	53.4	53.4	64.2	**72.2**
AVEC	EmoDB	54.1	52.9	54.0	55.1	55.7	55.2
AVEC	eNT	53.5	53.7	**54.3**	55.9	57.2	56.1
AVEC	VAM	49.9	51.2	50.0	58.8	60.2	**60.7**
Total Avg.		52.4	52.4	52.6	62.2	65.5	65.4

The best result for each test corpus and dimension is bolded. eNT: eNTERFACE.

However, the models trained on VAM do not achieve the highest UAR for valence, even when the test corpus is of the same language (EmoDB) or same type of emotion (AVEC). This may be because it has very unbalanced data for valence, as shown in Table 3. These findings suggest that the cross-corpus performance is not only related to the connection in language or type of emotion between training and testing datasets, but is also influenced by other aspects, such as data distribution and quality.

We compare different combinations of training corpora for activation, where the data are more balanced compared to valence. We noticed that the highest UAR of each test corpus (italicized in Table 6) is achieved by training on two corpora. Interestingly, for EmoDB, VAM and AVEC, the best training combinations consist of the two corpora that have an aspect in common with the test corpus (i.e., a corpus with same language, and a corpus with the same type of emotion), but do not share common factors between them. The only exception is eNTERFACE, in which the performances of different training combinations are similar. This may be because: (1) we are able to combine knowledge related to language and type of emotion by training on corpora that each share a different common factor with the test corpus; (2) when the training corpora are more dissimilar in language and type of emotion, the common ground between them have higher possibility to be emotion-related. However, when the training corpora have the same language or type of emotion, we may be overfitting to this common factor. Therefore, the classifiers do not generalize well when the factor is different in the test corpus.

6.2.3 The Influence of Model

We hypothesize that the influence of model is significant, when task definition is controlled. Specifically, multi-task learning models are better than the simple model and the

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

TABLE 6
Experiment 2. UAR (%) of Valence and Activation Using Multiple Training Corpora

Dim	Test on	Train on				Simple	Task: Corpus		Task: Gender		Task: Corpus-gender Pair				
		EmoDB	eNT	VAM	AVEC		ST	MTL	ST	MTL	ST	MTL	GMTL	MTL-GC	MTL-GG
V	EmoDB		✓	✓		52.9	57.8	58.4	53.8	55.7	56.8	56.8	59.6	56.0	56.3
	EmoDB		✓		✓	59.9	58.3	57.9	58.9	59.9	*60.8*	56.6	56.6	55.0	60.4
	EmoDB			✓	✓	54.1	51.8	50.8	53.3	57.2	51.4	50.9	53.0	50.0	49.0
	EmoDB		✓	✓	✓	57.0	58.1	58.8	59.5	57.1	57.1	57.9	59.1	56.3	59.8
	eNT	✓		✓		49.7	48.4	49.9	50.4	51.3	46.3	48.3	47.4	47.5	48.1
	eNT	✓			✓	54.2	50.5	49.9	52.9	54.0	48.4	50.5	50.5	54.0	50.5
	eNT			✓	✓	49.0	52.7	52.4	50.8	*60.0*	49.1	56.5	56.7	55.4	55.8
	eNT	✓		✓	✓	50.5	50.5	51.3	52.1	57.7	47.0	50.7	49.0	54.3	49.8
	VAM	✓	✓			48.2	54.0	51.1	49.6	51.1	52.6	51.5	53.8	52.7	52.1
	VAM	✓			✓	48.9	51.0	48.9	54.9	55.3	53.0	49.3	49.3	52.9	49.3
	VAM		✓		✓	51.7	56.4	*58.1*	53.2	55.8	56.9	56.9	56.9	55.9	57.5
	VAM	✓	✓		✓	51.6	54.2	52.0	53.5	57.3	51.0	52.2	52.4	54.5	52.2
	AVEC	✓	✓			54.9	*55.9*	55.2	54.7	55.0	54.9	55.1	55.4	55.3	55.1
	AVEC	✓		✓		51.0	53.4	53.9	53.1	54.6	52.9	52.2	53.0	54.0	53.8
	AVEC		✓	✓		52.6	53.8	53.3	53.3	54.3	53.1	53.8	53.8	53.4	54.0
	AVEC	✓	✓	✓		53.3	54.9	55.7	53.9	54.4	54.7	55.6	55.0	55.2	54.7
	Avg. of Valence					52.5	**53.9**	53.6	53.6 ∗	**55.7** ∗ †	52.9	53.4	**53.9**	**53.9**	53.7
A	EmoDB		✓	✓		74.7	86.2	87.6	84.2	86.0	*89.1*	88.9	87.6	81.7	88.0
	EmoDB		✓		✓	56.8	70.8	73.0	58.2	62.9	77.1	75.3	75.3	56.8	70.4
	EmoDB			✓	✓	74.7	83.0	79.1	76.9	70.0	85.8	75.7	75.7	70.4	75.7
	EmoDB		✓	✓	✓	69.0	85.2	85.6	75.5	79.1	87.2	85.4	85.6	71.2	80.3
	eNT	✓		✓		65.7	67.0	68.5	66.4	67.7	67.2	68.2	66.8	69.6	67.2
	eNT	✓			✓	59.3	63.4	67.8	62.7	67.1	65.9	68.2	68.1	69.7	67.8
	eNT			✓	✓	64.5	66.9	69.9	64.5	69.0	67.8	*70.5*	*70.5*	70.2	*70.5*
	eNT	✓		✓	✓	62.8	66.9	69.2	64.0	68.0	67.5	68.3	68.3	70.2	68.6
	VAM	✓	✓			62.1	68.2	68.0	65.7	65.2	67.8	67.2	64.6	67.1	67.7
	VAM	✓			✓	63.8	68.1	73.1	68.3	74.5	69.3	*74.8*	74.2	73.8	74.6
	VAM		✓		✓	60.1	61.0	65.9	62.1	69.6	65.5	68.6	68.6	72.1	73.4
	VAM	✓	✓		✓	63.1	68.9	70.7	65.6	72.6	68.9	71.1	71.0	73.6	73.8
	AVEC	✓	✓			56.2	56.1	55.7	58.3	56.7	56.8	56.7	55.7	56.4	57.2
	AVEC	✓		✓		59.0	56.9	59.2	58.6	59.1	57.8	59.4	58.5	60.5	60.1
	AVEC		✓	✓		58.8	59.3	59.9	59.7	60.1	60.2	60.6	61.2	60.6	*61.7*
	AVEC	✓	✓	✓		58.8	57.2	58.5	59.1	59.1	58.5	58.2	58.5	60.7	59.5
	Avg. of Activation					63.1	67.8 ∗	**69.5** ∗ †	65.6 ∗	**67.9** ∗ †	69.5 ∗	**69.8** ∗	69.4 ∗	67.8 ∗	**69.8** ∗

The best average performance for each dimension in each experiment is bolded. The overall best performance of each dimension is underlined. The ∗ (†) indicates that the difference in the mean UAR between the marked model and the simple (ST with same task definition) model is statistically significant when tested using the Tukey's honest significant difference test at 95% confidence level. V: valence; A: activation; GC/GG: group tasks by corpus/gender; eNT: eNTERFACE.

ST model. We test this hypothesis when corpus, gender or the corpus-gender pair is used to define tasks, respectively, using RMs with model as the WSF.

When corpus is used as the task, the influence of model is significant for activation (RANOVA, $F(2, 24) = 54.5$, $p = 1.2e\text{-}09$), but not for valence. A pairwise comparison for activation shows that MTL significantly outperforms both simple and ST (Tukey test, $p = 1.0e\text{-}05$ and 0.016, respectively). This supports the notion that different corpora should be treated as related tasks for the prediction of activation.

When gender is used as the task, we test the influence of model on the results from training on single datasets (from Table 5) and on multiple datasets (from Table 6). This is to be consistent with other task definitions (i.e., corpus and corpus-gender pair), where only multiple training datasets results can be compared. We find that the influence of model is significant for both valence (RANOVA, $F(2, 24) = 14.3$, $p = 8.2e\text{-}05$) and activation ($F(2, 24) = 16.8$, $p = 2.8e\text{-}05$) when training on multiple corpora, but not when training on a single corpus. This may be because we are not capturing the full range of gender variability with only one training corpus. In addition, splitting data by gender for a single corpus may result in insufficient training data per task. The Tukey tests show that when we use multiple training corpora, MTL significantly outperforms both simple and ST for valence ($p = 0.0025$ and 0.018, respectively) and activation ($p = 0.0016$ and 0.043, respectively). This reinforces the importance of separating data from different genders, yet still considering the relatedness between them.

When corpus-gender pairs are used as tasks, we find that the influence of model is significant for activation (RANOVA, $F(5, 60) = 24.3$, $p = 2.8e\text{-}13$), but not for valence. The Tukey test for activation shows that all models that explicitly consider the variations in corpus and gender (ST, MTL, GMTL, MTL-GC and MTL-GG) significantly outperform the simple model ($p = 1.7e\text{-}06$, 5.0e-04, 0.0015, 7.3e-04 and 1.3e-04 for ST, MTL, GMTL, MTL-GC and MTL-GG versus simple, respectively). Interestingly, there is no significant difference between

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

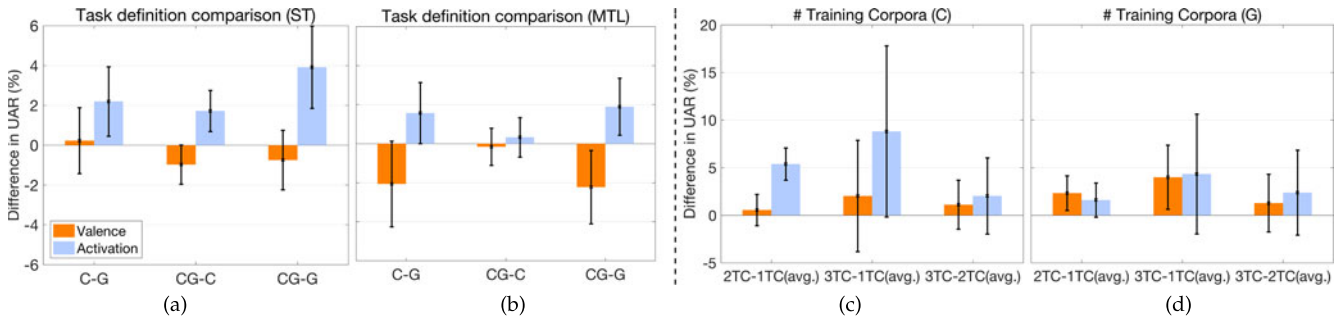


Fig. 3. Experiment 2. Difference in UAR between different experimental conditions (e.g., C-G is the difference between defining the tasks by corpus and gender), along with the 95% confidence interval of the Tukey test, between: different task definitions for (a) ST and (b) MTL; different numbers of training corpora when (c) defining corpus as the task (using MTL) and (d) defining gender as the task (using MTL). Note that for (c) and (d), results with fewer training corpora are averaged across each corpus (1TC) or each combination of training corpora (2TC). MTL with only one training corpus in (c) is the same as simple. C: corpus as the task; G: gender as the task; CG: corpus-gender pairs as the tasks; TC: training corpora.

ST and other multi-task learning models. This may be because we fuse the results by weighted majority vote over all the tasks. Therefore, we are not only considering the differences in corpus and gender by training task-dependent classifiers, but also utilizing knowledge learned from all the tasks instead of just one. Comparing the multi-task learning methods, we can see that on average, GMTL and MTL-GC perform the best for valence, and MTL and MTL-GG perform the best for activation. However, the differences between the multi-task learning models are very small and not statistically significant, except for between MTL-GC and MTL-GG for activation ($p = 0.019$). We notice that grouping the tasks by corpus or gender generates the highest UAR on several classification tasks (e.g., MTL-GG for valence of VAM when using eNTERFACE and AVEC for training, and for activation of AVEC when using eNTERFACE and VAM for training, MTL-GC for activation of eNTERFACE and AVEC when training on three corpora), but their performances are not stable. For example, the UARs of MTL-GC on the activation of EmoDB are the lowest for all the training corpora combinations, compared to all other models except for the simple model. This may indicate that the closeness between the tasks may be related to the common factors between them, but that the relationship is not guaranteed.

These results support the notion that variations in training corpus, gender, and their interactions all modulate the data. It is beneficial to control for these sources of variation by defining tasks and allowing the tasks to share information using multi-task learning. Improvement in valence is harder to achieve, compared to activation, as found in [9].

6.2.4 The Influence of Task Definition

We hypothesize that the way we define the tasks significantly influences the performance of a model. We test this hypothesis using RMs for ST and MTL, respectively, across experiment 2c, 2g and 2cg, with task definition as the WSF.

For ST, the effect of task definition (i.e., corpus, gender, corpus-gender) is significant for activation (RANOVA, $F(2, 24) = 19.7$, $p = 8.8e\text{-}06$), but not for valence. The pairwise Tukey test for ST (Fig. 3a) suggests that for activation, using either corpora or corpus-gender pairs as tasks is significantly better than using genders as tasks ($p = 0.015$ and 7.6e-04, respectively) and that the corpus-gender pairs significantly outperform corpora as tasks. ($p = 0.0021$).

For MTL, the impact of task definition is significant for both valence (RANOVA, $F(2, 24) = 7.1$, $p = 0.0038$) and activation ($F(2, 24) = 7.8$, $p = 0.0025$). The pairwise comparison for MTL is shown in Fig. 3b. For valence, gender is a significantly better task-separator than corpus-gender pair (Tukey test, $p = 0.021$), while for activation, the result is the opposite ($p = 0.012$). In addition, the advantage of gender over corpus is approaching significance for valence ($p = 0.066$), and the advantage of corpus over gender as the task is approaching significance for activation ($p = 0.05$).

The results indicate that defining tasks by gender is the best for valence, while defining a task as a corpus-gender pair is the most beneficial for activation. Interestingly, the benefits of using corpus-gender pairs as tasks in activation is consistent for ST and MTL, but the advantage of using gender as the task in valence only shows in MTL. This suggests that information sharing between genders is important for learning a more robust pattern associated with valence.

6.2.5 Number of Training Corpora

We hypothesize that the number of training corpora (denoted as TC) significantly influences the system performance, when both task definition and model are controlled. Specifically, we hypothesize that adding additional TC is helpful. We test this hypothesis by comparing the performance as the number of TC changes. The model is MTL and the task is either corpus or gender. We build RMs with the number of TC as the WSF for three settings: (a) 2TC versus 1TC, (b) 3TC versus 1TC, and (c) 3TC versus 2TC. The challenge is that each TC size is associated with a different number of results. We compare by averaging over relevant subsets. For example, in the 3TC setting, where we are testing on VAM, the training corpora include EmoDB, eNTERFACE, and AVEC. We compare this result to the 2TC results, still with VAM as a testing corpus. In this case, we take the average performance of systems trained on EmoDB and eNTERFACE, EmoDB and AVEC, and eNTERFACE and AVEC. When comparing to 1TC, we calculate the average obtained by training systems on each of the training corpora, individually. We repeat this over all test corpora. The same comparison applies to 2TC versus 1TC. Thus, in (a) there are 12 results for each dimension, in (b) and (c) there are four results for each dimension. The comparisons between different numbers of TC are shown in Figs. 3c (corpus as task) and 3d (gender as task).

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

TABLE 7
Within-Corpus UAR ("Within") Using the Simple Model and the Best Cross-Corpus UAR ("Cross") in Experiment 2 (%) from Our Models, and the Within-Corpus and Cross-Corpus UAR from Literature

Dim	Setting	From	EmoDB	eNT	VAM	AVEC
V	Within	Our Model	84.5	83.4	53.2	53.6
		[66]	87.0	78.7	49.2	-
	Cross	Our Model	61.0	60.0	59.3	55.9
		Literature	-	58.4 [29]	58.6 [8]	-
A	Within	Our Model	95.9	84.0	76.1	56.5
		[66]	96.8	78.1	76.5	-
	Cross	Our Model	89.1	70.5	74.8	61.7
		Literature	-	63.9 [29]	71.9 [24]	-

eNT: eNTERFACE; Dim: dimension; V: valence; A: activation.

We find that when corpus is used to define the tasks, the influence of the number of TC is significant for activation (RANOVA for 2TC versus 1TC, $F(1,8) = 53.7$, $p = 8.2e\text{-}05$), but not for valence. The Tukey test demonstrates that 2TC is significantly better than 1TC ($p = 8.2e\text{-}05$). The improvements of 3TC over 1TC and 2TC are not significant. However, there are only four results to be compared in these two tests.

When gender is used to define the tasks, the influence of the number of TC is significant for valence (RANOVA, $F(1,8) = 8.7$, $p = 0.018$ for 2TC versus 1TC, $F(1,3) = 14.3$, $p = 0.033$ for 3TC versus 1TC), but not for activation. Both 2TC and 3TC are significantly better than 1TC for valence ($p = 0.018$ and 0.033, respectively). The performance gain of adding a third training corpus to a set already composed of two is not statistically significant.

These findings suggest that the addition of training corpora is helpful, especially given limited variability in the data (e.g., single training corpus). The results also support our earlier findings that gender is a better task-separator than corpus for valence, while corpus is a better task-separator than gender for activation.

6.2.6 Cross-Corpus versus Within-Corpus

We present our best cross-corpus UAR and within-corpus LOSO UAR using the simple model in Table 7. We compare these results to both the benchmark within-corpus LOSO UAR from [66] and the state-of-the-art cross-validation UAR from the literature (see Table 7). We are not able to compare to [6], [9] because the UARs of the individual test datasets are not provided. Note that the number of instances in this paper is off by 1 for EmoDB and VAM, and off by 10 for eNTERFACE, compared to [8], [29], [66]. We do not compare the results of EmoDB to [24] because the label matching method is different.

We find that the advantage of within-corpus classification is dominant for datasets with acted emotion (EmoDB and eNTERFACE). A possible explanation is that these acted datasets use fixed lexical content, making emotion recognition much easier. However, We find that cross-corpus classification is effective for datasets with spontaneous emotion. The performance of cross-corpus classification is higher for VAM valence and for AVEC valence and activation. It is slightly lower for VAM activation. Direct comparison between our model and the literature is not possible due to the small differences in data described above and the differences in training datasets. We note that our models achieve comparable results to the state of the art.

7 Discussion

In this paper, we explore the influence of domain, corpus and gender in emotion recognition by conducting two sets of experiments. We propose a multi-task learning approach to recognize emotion across corpora, with data from multiple domains or datasets as the training set. We present five different models: the simple model, the separate-task model, the multi-task learning model, the group multi-task learning model, and the multi-task learning model with knowledge-driven grouping. These models correspond to five assumptions about the relationship between the tasks: identical, independent, related, partially related and can be grouped based on data similarity, and partially related and can be grouped based on knowledge.

Our results show that a generalizable sparse feature representation on the original space can be found across two acted corpora with both spoken and sung data (experiment 1). However, we find that for the speech domain (experiment 2), a common sparse feature representation on a transformed feature space is more beneficial, compared to on the original feature space. We assume that this may be due to the higher dimensionality of the features and larger variability in languages, types of emotion, lexical content, speakers and recording conditions.

In experiment 2, the best cross-corpus performance with a single training corpus is not always achieved by training on a corpus that shares common language or type of emotion with the test corpus. This may indicate that the quality of the training corpus, in terms of cross-corpus generalizability, is not only related to its similarity with the test corpus, but is also influenced by factors such as class imbalance and the quality of the emotion content. This is inline with Schuller et al. [9]. They found that models trained on VAM produced the best cross-corpus performance on various testing datasets for activation, and that the supreme performance of VAM could be related to the large distance between the positive and negative classes. In experiment 2, we also observed that training on corpora that each share common factors with the test corpus, but not with each other, improves activation recognition, in most cases.

Our results support that variations in corpus, domain and gender all influence emotion recognition. Overall, separating tasks by these factors and allowing for information sharing between tasks using multi-task learning methods is advantageous. When a single factor is considered, the best performances happen predominantly in cases where we treat the tasks as related, instead of identical or independent. When multiple factors are considered (domain and gender, corpus and gender), group multi-task learning either achieves the highest performance or is comparable to the best performance generated by other multi-task learning models. This suggests that when defining the tasks by more than one factor, some tasks are more closely related than others. Although we are not able to get a stable grouping since group multi-task learning is non-convex, we find that data-driven grouping works better than knowledge-driven grouping for domain and gender, and is comparable to knowledge-driven grouping for corpus and gender.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

Comparing different factors, we find that domain is a larger differentiating factor than gender for multi-domain data. This explains why researchers often consider speech emotion recognition and music (or song) emotion recognition as separate fields of research. One might expect that corpus is a more dominant differentiating factor for speech emotion recognition, compared to gender. However, we find that when using multiple datasets for training, separating data based on either corpus or gender, and training emotion classifiers with multi-task learning generates better results, compare to merging all the data together or training independent classifiers. This is inline with the findings in [40] that differences between genders can be as large as the differences between datasets. More specifically, we find that gender is a better task-separator for valence, compared to corpus or corpus-gender pair, while corpus and corpus-gender pair are better task-separators for activation, compared to gender.

The best cross-corpus performance in our experiments is better than or comparable to the within-corpus performance using the baseline method in two situations: (1) when the test corpus has limited data (experiment 1); (2) when the test corpus contains spontaneous emotion (VAM and AVEC in experiment 2). The first situation supports the findings of Lefter et al. [7] that cross-corpus performance could be higher than within-corpus performance when the intra-corpus training set suffers from data scarcity. In the second situation, our findings may be influenced by the high degree of variability within the spontaneous dataset, which may have reduced the advantage of within-corpus testing.

8 Conclusion

In this paper, we investigate methods of increasing the generalizability of audio emotion recognition systems, by controlling for three sources of variation: corpus, domain, and gender. These factors define our tasks. We use multi-task learning to enable the information sharing across tasks.

In general, defining the tasks by domain, corpus and/or gender, and allowing for information sharing across tasks is beneficial. For multi-domain data, domain is a stronger differentiating factor than gender. For speech domain, defining tasks by corpus or both corpus and gender is better than by gender for activation predictions, while gender is the best task-separator for valence predictions. When multiple factors are used to define the tasks, data-driven grouping performs at least comparably to knowledge-driven grouping. On average, the system performance increases with the number of training corpora.

In the future, we plan to continue this work in the following directions. First, inspired by our observation from Section 6.2.2, we are interested in investigating (1) how the cross-corpus performance of the same pair of training-testing datasets changes as a function of emotion expressiveness, class distribution, or noise level of the training dataset; and (2) how cross-corpus performance changes as a function of the similarity between training corpora, or the similarity between training and testing corpora, where similarity is defined by either language or type of emotion. Further, we will investigate how these findings change when information about the corpus-level similarity is not known.

Second, we would like to explore finer-grain tasks, such as separating the data by speaker identity. We plan to jointly train speaker-dependent multi-task emotion classifiers, by learning latent representative tasks, and treating the known tasks as combinations of the latent tasks. We will use this approach and unsupervised transfer learning to achieve both the advantage of sufficient training data, and the benefit of speaker-dependent emotion classification.

Finally, we will explore feature modeling with deep learning methods that are effective and robust for cross-corpus emotion recognition.

References

[1] T. Brosch, K. R. Scherer, D. Grandjean, and D. Sander, "The impact of emotion on perception, attention, memory, and decision-making," *Swiss Med Wkly*, vol. 143, 2013, Art. no. w13786.

[2] R. Beale and C. Peter, "The role of affect and emotion in HCI," in *Affect and Emotion in Human-Computer Interaction*. Berlin, Germany: Springer, 2008, pp. 1–11.

[3] B. Reeves and C. Nass, *How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, U.K.: CSLI Publications and Cambridge University Press, 1996.

[4] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012.

[5] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation," in *Affect and Emotion in Human-Computer Interaction*. Berlin, Germany: Springer, 2008, pp. 75–91.

[6] B. Schuller, et al., "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Trans. Affective Comput.*, vol. 1, no. 2, pp. 119–131, Jul.–Dec. 2010.

[7] I. Lefter, L. J. Rothkrantz, P. Wiggers, and D. A. Van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," in *Proc. Int. Conf. Text, Speech Dialogue*, 2010, pp. 353–360.

[8] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" in *Proc. INTERSPEECH*, 2011, pp. 1553–1556.

[9] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting training data for cross-corpus speech emotion recognition: Prototypicality versus generalization," in *Proc. Afeka-AVIOS Speech Process. Conf.*, 2011, p. 4.

[10] P. Song, Y. Jin, L. Zhao, and M. Xin, "Speech emotion recognition using transfer learning," *IEICE Trans. Inf. Syst.*, vol. 97, no. 9, pp. 2530–2532, 2014.

[11] B. Zhang, G. Essl, and E. Mower Provost, "Recognizing emotion from singing and speaking using shared models," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2015, pp. 139–145.

[12] Y. Kim and E. Mower Provost, "Say cheese versus smile: Reducing speech-related variability for facial emotion recognition," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 27–36.

[13] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Proc. 18th Int. Conf. Pattern Recog.*, 2006, pp. 1136–1139.

[14] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[15] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proc. 12th Eur. Signal Process. Conf.*, 2004, pp. 341–344.

[16] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[17] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. Language Resources Eval. Conf.*, 2006, pp. 1123–1126.

[18] B. Romera-Paredes, M. S. Aung, M. Pontil, N. Bianchi-Berthouze, A. C. d. C. Williams, and P. Watson, "Transfer learning to account for idiosyncrasy in face and body expressions," in *Proc. 10th IEEE Int. Conf. Workshops Automatic Face Gesture Recog.*, 2013, pp. 1–6.

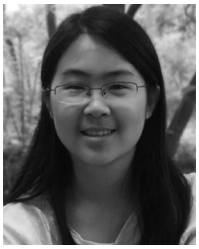
[19] T. Almaev, B. Martinez, and M. Valstar, "Learning to transfer: Transferring latent task structures and its application to person-specific facial action unit detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3774–3782.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

[20] T. J. Shields, M. R. Amer, M. Ehrlich, and A. Tamrakar, "Action-affect classification and morphing using multi-task representation learning," *arXiv preprint arXiv:1603.06554*, 2016.
[21] B. Zhang, E. Mower Provost, and G. Essl, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2016, pp. 5805–5809.
[22] M. Shami and W. Verhelst, "Automatic classification of emotions in speech using multi-corpora approaches," in *Proc. 2nd Annu. IEEE Benelux/DSP Valley Signal Process. Symp.*, 2006, pp. 3–6.
[23] I. Lefter, H. T. Nefs, C. M. Jonker, and L. J. Rothkrantz, "Cross-corpus analysis for acoustic recognition of negative interactions," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2015, pp. 132–138.
[24] B. Vlasenko, D. Prylipko, R. Böck, and A. Wendemuth, "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Comput. Speech Language*, vol. 28, no. 2, pp. 483–500, 2014.
[25] M. Shah, C. Chakrabarti, and A. Spanias, "Within and cross-corpus speech emotion recognition using latent topic model-based features," *EURASIP J. Audio Speech Music Process.*, vol. 2015, no. 1, pp. 1–17, 2015.
[26] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2015, pp. 5058–5062.
[27] Q. Mao, W. Xue, Q. Rao, F. Zhang, and Y. Zhan, "Domain adaptation for speech emotion recognition by sharing priors between related source and target classes," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2016, pp. 2608–2612.
[28] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Commun.*, vol. 83, pp. 34–41, 2016.
[29] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. IEEE Workshop Automatic Speech Recog. Understanding*, 2011, pp. 523–528.
[30] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, no. 9, pp. 1062–1087, 2011.
[31] Y. E. Kim, et al., "Music emotion recognition: A state of the art review," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 255–266.
[32] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, 2012, Art. no. 40.
[33] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814, 2003.
[34] G. Ilie and W. F. Thompson, "A comparison of acoustic cues in music and speech for three dimensions of affect," *Music Perception: An Interdisciplinary J.*, vol. 23, no. 4, pp. 319–329, 2006.
[35] G. Ilie and W. F. Thompson, "Experiential and cognitive changes following seven minutes exposure to music and speech," *Music Perception: An Interdisciplinary J.*, vol. 28, no. 3, pp. 247–264, 2011.
[36] K. R. Scherer, J. Sundberg, L. Tamarit, and G. L. Salomão, "Comparing the acoustic expression of emotion in the speaking and the singing voice," *Comput. Speech Language*, vol. 29, pp. 218–235, 2013.
[37] S. R. Livingstone, K. Peck, and F. A. Russo, "Acoustic differences in the speaking and singing voice," in *Proc. Meetings Acoustics*, 2013, Art. no. 035080.
[38] E. Coutinho, J. Deng, and B. Schuller, "Transfer learning emotion manifestation across music and speech," in *Proc. Int. Joint Conf. Neural Netw.*, 2014, pp. 3592–3598.
[39] B. Zhang, E. Mower Provost, R. Swedberg, and G. Essl, "Predicting emotion perception across domains: A study of singing and speaking," in *Proc. 29th AAAI Conf. Arti. Intell.*, 2015, pp. 1328–1334.
[40] M. Brendel, R. Zaccarelli, B. Schuller, and L. Devillers, "Towards measuring similarity between emotional corpora," in *Proc. Satellite Workshop of LREC*, 2010, pp. 58–64.
[41] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "From joyous to clinically depressed: Mood detection using spontaneous speech," in *Proc. 25th Int. Florida Artif. Intell. Res. Soc. Conf.*, 2012, pp. 141–146.
[42] B. Vlasenko, D. Philippou-Hübner, D. Prylipko, R. Böck, I. Siegert, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal emotions," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2011, pp. 1–6.
[43] I. M. A. Shahin, "Gender-dependent emotion recognition based on HMMs and SPHMMs," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 133–141, 2013.
[44] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *Proc. 30th Int. Conf. Int. Conf. Machine Learn.*, 2013, pp. 1444–1452.
[45] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.
[46] I. S. Engberg and A. V. Hansen, "Documentation of the Danish emotional speech database des," Internal AAU report, Center for Person Kommunikation, Aalborg Øst, Denmark, 1996.
[47] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the Fau Aibo emotion corpus," in *Proc. Satellite Workshop LREC*, 2008, pp. 28–31.
[48] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2008, pp. 865–868.
[49] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *IEEE Int. Conf. Multimedia Expo*, 2010, pp. 1079–1084.
[50] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 5–17, Jan.–Mar. 2012.
[51] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proc. IEEE*, vol. 101, no. 5, pp. 1203–1233, May 2013.
[52] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops*, 2006, p. 8.
[53] S. R. Livingstone, K. Peck, and F. A. Russo, "Ravdess: The Ryerson audio-visual database of emotional speech and song," in *Proc. Annu. Meet. Canadian Soc. Brain, Behaviour Cognitive Sci.*, 2012.
[54] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011–the first international audio/visual emotion challenge," in *Proc. 4th Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 415–424.
[55] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: The continuous audio/visual emotion challenge," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2012, pp. 449–456.
[56] S. R. Livingstone, W. F. Thompson, M. M. Wanderley, and C. Palmer, "Common cues to emotion in the dynamic facial expressions of speech and song," *Quarterly J. Exp. Psychology*, vol. 68, no. 5, pp. 952–970, 2015.
[57] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 41–48.
[58] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.
[59] Z. Kang, K. Grauman, and F. Sha, "Learning with whom to share in multi-task feature learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 521–528.
[60] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.
[61] A. Agarwal, S. Gerber, and H. Daume, "Learning multiple tasks using manifold regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 46–54.
[62] K. Crammer and Y. Mansour, "Learning multiple tasks using shared hypotheses," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1475–1483.
[63] B. Schuller, et al., "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc.*, 2013,pp. 148–152.
[64] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
[65] F. B. Pokorny, F. Graf, F. Pernkopf, and B. W. Schuller, "Detection of negative emotions in speech signals using bags-of-audio-words," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2015, pp. 879–884.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.

[66] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2009, pp. 552–557.



Biqiao Zhang (S'16) received the BS degree in information management and information system, the BA degree in economics, in 2011, and the MS degree in information science, in 2013, all from Peking University, Beijing, China. She is working toward the PhD degree working with professor Emily Mower Provost and professor Georg Essl in computer science and engineering with the University of Michigan. She is a student member of the IEEE. Her research interests are in human-centered computing and affective computing using acoustic signal processing and machine learning methods. In particular, she is interested developing methods that can address factors that negatively impact the generalizability of automatic emotion recognition systems.

Emily Mower Provost (S'07-M'11) received the BS degree in electrical engineering (summa cum laude and with thesis honors) from Tufts University, Boston, Massachusetts, in 2004 and the MS and PhD degrees in electrical engineering from the University of Southern California (USC), Los Angeles, California, in 2007 and 2010, respectively. She is an assistant professor in computer science and engineering with the University of Michigan. She is a member of the Tau-Beta-Pi, Eta-Kappa-Nu, and a member of the IEEE and the ISCA. She has been awarded the National Science Foundation Graduate Research Fellowship (2004-2007), the Herbert Kunzel Engineering Fellowship from USC (2007-2008, 2010-2011), the Intel Research Fellowship (2008-2010), the Achievement Rewards For College Scientists (ARCS) Award (2009 2010), and the Oscar Stern Award for Depression Research (2015). She is a co-author on the paper, "Say Cheese versus Smile: Reducing Speech-Related Variability for Facial Emotion Recognition," winner of Best Student Paper at ACM Multimedia, 2014. She is also a co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. Her research interests are in human-centered speech and video processing, multimodal interfaces design, and speech-based assistive technology. The goals of her research are motivated by the complexities of human emotion generation and perception.

Georg Essl (S'93-M'02) received the undergraduate degree in telematics from Graz University of Technology in 1996. He received the PhD degree from Princeton University, in 2002 working with Perry Cook on real-time sound synthesis method for solid objects. He is a visiting research professor in the College of Letters & Science at the University of Wisconsin-Milwaukee. He received his PhD from Princeton University in 2002 working with Perry Cook on real-time sound synthesis method for solid objects. He has been on the faculty of the University of Michigan, the University of Floria, worked at MIT Media Lab Europe and the Deutsche Telekom Laboratories at the Technical University of Berlin. He is a member of the IEEE, the ASA, the ACM, the AMS, the ICMA, and serves on the advisory board of NIME. His research interests are computer music, mobile HCI and mobile music making, human-computer interfaces, real-time physical simulation of audio, tactile feedback and foundations of numerical methods for interactive applications.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR. Downloaded on May 29,2022 at 06:12:07 UTC from IEEE Xplore. Restrictions apply.