CROSS-CORPUS ACOUSTIC EMOTION RECOGNITION FROM SINGING AND SPEAKING: A MULTI-TASK LEARNING APPROACH

Biqiao Zhang, Emily Mower Provost, Georg Essl

University of Michigan, Ann Arbor
Computer Science and Engineering
{didizbq, emilykmp, gessl}@umich.edu

ABSTRACT

Emotion is expressed over both speech and song. Previous works have found that although spoken and sung emotion recognition are different tasks, they are related. Classifiers that explicitly utilize this relatedness can achieve better performance than classifiers that do not. Further, research in speech emotion recognition has demonstrated that emotion is more accurately modeled when gender is taken into account. However, it is not yet clear how domain (speech or song) and gender can be jointly leveraged in emotion recognition systems nor how systems leveraging this information can perform in cross-corpus settings. In this paper, we explore a multi-task emotion recognition framework and compare the performance across different classification models and output selection/fusion methods using cross-corpus evaluation. Our results show the classification accuracy is the highest when information is shared only between closely related tasks and when the output of disparate models are fused.

Index Terms— emotion recognition, multi-task learning, cross-corpus, speech emotion, sung emotion

1. INTRODUCTION

Emotion can be expressed over many vocal communication domains, including song and speech. Automatic emotion recognition has received increasing attention in recent years because of its various potential applications, such as interactive robots or personal assistants, emotion-aware agents in cars or call centers, computer-enhanced learning, computer games, and information retrieval systems [1]. Since humans can express and recognize emotion across various domains, emotion recognition systems that can perform similarly will have broad application, enabling more natural cross-domain human-computer interaction.

Previous research has focused on both speech and music emotion recognition [2–6]. In recent years, attention has been paid to the commonalities and differences between spoken and sung emotional communication [7, 8]. Our previous work [9] explored ways of building a shared emotion recognition model using both single-task and multi-task approaches, with song and speech as the two tasks. We found that emotion classification systems can benefit from multi-task learning, suggesting that spoken and sung emotion recognition tasks are different, but related, and can be considered together. Finally, previous studies have shown that gender-dependent emotion recognizers outperform gender-independent ones [10–12]. However, the combined influence of domain and gender on emotion recognition systems has not been analyzed yet.

In addition, there has been growing interest in studying the cross-corpus generalizability of speech emotion recognition to face the challenges brought by the differences in speakers and recording conditions in real-life applications [13–17]. Schuller et al. found that speaker-dependent normalization works better than corpus-level normalization [13]. Lefter et al. [14] and Schuller et al. [15] found that cross-corpus performance could be improved by combining databases and fusing classifiers. Schuller et al. also found that cross-corpus performance improves when systems select prototypical data, defined as datasets with large distances between class centers and as instances close to the corresponding class center [16]. Peng et al. [17] found that transferring feature representations from one corpus to another using Maximum Mean Discrepancy Embedding (MMDE) optimization and dimensionality reduction is beneficial to cross-corpus classification accuracy. However, although joint emotion recognition from speech and song introduces additional sources of variation, analysis on the cross-corpus generalizability of multi-domain emotion classification models is still missing.

In this work, we conduct cross-corpus evaluation on two datasets that contain sung and spoken emotion expressions. We build emotion recognition models that work for both domains (speech and song) and genders (female and male). We define a task as emotion recognition in a domain-gender pair, for example, female-speech and male-song. We recognize emotion from acoustic singing and speaking using four different models: (1) a simple model, where a single classifier is built using data from all four tasks; (2) a single-task (ST) model, where a separate classifier is built for each task; (3) a multi-task feature selection/learning (MTFS/MTFL) model, where all the four tasks are considered to be related; (4) a group multi-task feature selection/learning model (GMTFS/GMTFL), where tasks grouping is also learned, and only tasks within the same group share information with each other. The terms "feature selection" and "feature learning" in the model names specify whether the regularizer enforcing sparsity is imposed on the original feature space or transformed feature space, respectively. Excepting the simple model, all models learn T weight vectors, one for each task (e.g., female-song), and output T predicted labels for a given test instance for a specific classification problem. We propose five methods to fuse the T labels into a final predicted label: (1) oracle, where we assume that the task associated with the test data is known; (2) decision tree (DT), where we predict the task of the testing data; (3) majority vote (MV), where we perform a majority vote over the T labels; (4) weighted majority vote (WMV), where the voting is weighted by a measure of confidence; (5) maximum distance (MD), where the most confident decision is selected. No prior knowledge about the task of test data is needed aside from the baseline "oracle" method. We solve the multi-class classification problem as the combination of one-against-one binary classification problems (e.g. angry vs. happy and angry vs. sad). Fig. 1 illustrated our proposed method for one pair of emotion. The

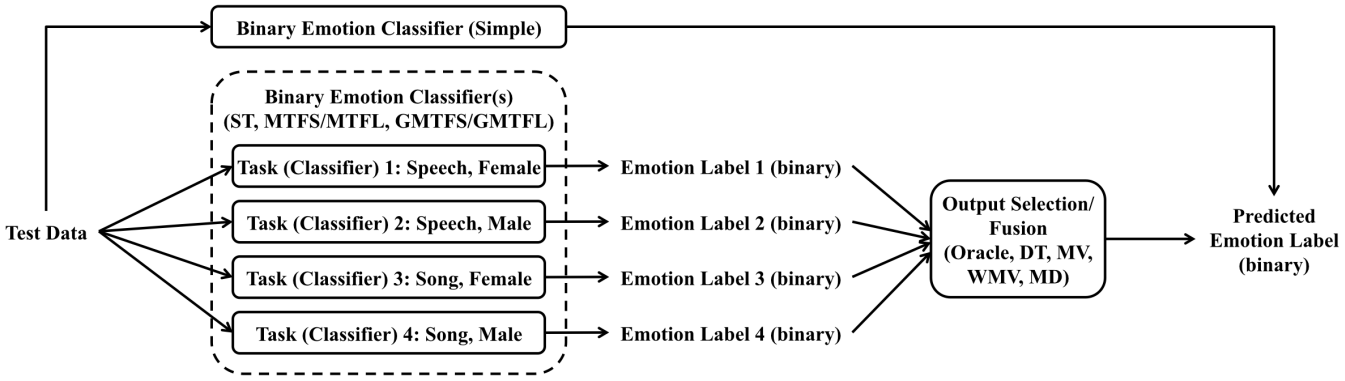

Fig. 1: System diagram of the proposed classification models and output selection/fusion methods. In the simple model, only one binary classifier is trained for each emotion pair. In the other models, either four binary classifiers are trained (ST) or one binary classifier with four classification tasks is trained (MTFS/MTFL and GMTFS/GMTFL) for each emotion pair. The latter approach outputs four labels for a test case, and the predicted binary label is the result of selection (oracle, DT and MD) or fusion (MV and WMV). ST: single-task model, MTFS/MTFL: multi-task feature selection/learning model, GMTFS/GMTFL: group multi-task feature selection/learning model. DT: decision tree, MV: majority vote, WMV: weighted majority vote, MD: maximum distance.

final multi-class label is generated from the binary predictions by majority voting.

The results show that the GMTFS model works the best among all models, which suggests that the tasks (domain-gender pairs) are generally related and that some of the tasks have closer relationships than others. The weighted majority vote achieves the highest performance among all output selection/fusion methods, including oracle. This indicates that a task (e.g. emotion recognition for female-song) in one corpus is not identical to the same task in another corpus. The novelty of this paper includes: (1) an investigation into the influence of communication domain and gender on emotion recognition; (2) an analysis of the generalizability of cross-corpus and cross-domain emotion recognition; (3) an exploration into output selection/fusion methods when the the task of the test data is not known.

2. DATASETS

In this paper, we use the University of Michigan Song and Speech Emotion Dataset (UMSSED) [18] and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [19]. For both datasets, only the audio recordings were used.

2.1. The UMSSED Dataset

The UMSSED Dataset contains audio-visual recordings of three performers (1 female, 2 male) singing and speaking seven sentences with angry, happy, neutral and sad emotions. Seven melodies were composed for the singing performances, one each for the seven sentences. This results in 168 utterances. The target emotion of the performers was used as the ground truth. The utterances were evenly distributed across the four emotions. See [18] for additional details.

2.2. The RAVDESS Dataset

The RAVDESS dataset consists of the audio-visual recordings of 24 performers (12 female, 12 male) singing and speaking two sentences with six and eight emotions respectively, each with two repetitions. The song recordings consist of neutral, calm, happy, sad, angry, fearful emotions. The speech recordings have two additional emotions of disgust and surprise. All emotions except for neutrality were performed at two emotional intensities. Three melodies that differ in two notes were used for the singing performances, one each for positively valenced, neutral, and negatively valenced emotions.

We decreased the size of the datasets by only selecting angry, happy, neutral and sad emotions to match the UMSSED dataset. One performer with missing data was dropped. This results in 1288 utterances. The percentages of angry, happy, neutral and sad utterances are 28.6%, 28.6%, 14.3% and 28.6%. See [19] for additional details.

3. METHODOLOGY

3.1. Feature Extraction

We extracted the 65 frame-level acoustic low level descriptors (LLDs) described in the Interspeech 2013 Computational Paralinguistics Evaluation (ComParE) feature set [20], using openSMILE [21]. We extracted statistics, including mean, standard deviation, max, min, range, interquartile range, mean absolute deviation, skewness and kurtosis from the non-silence part of the LLDs and delta LLDs, to generate the 1170 utterance-level acoustic features. We applied performer-dependent z-normalization to the utterance-level features, such that in both datasets, each feature of each performer has zero mean and standard deviation of one. This method was demonstrated to outperform other normalization methods in [13].

3.2. Classification Models

We solve the multi-class emotion classification problem using the one-against-one strategy. Six binary classifiers were trained, one for each pair of emotions. The final label of each binary classifier gives one vote to the winning class and the data was labeled with the most voted class. In the case of ties, the class with the smallest index is selected. We present four models for the binary classification problem of each emotion pair: the simple model, the single task (ST) model, the multi-task feature selection/learning (MTFS/MTFL) model, and the group multi-task feature selection/learning (GMTFS/GMTFL) model.

The simple model is the same as the method used in [13] for cross-corpus acoustic emotion recognition of speech. It builds a sin-

gle emotion classifier using a support vector machine (SVM) with a linear kernel using all utterances in RAVDESS as training data. In the ST model, the MTFS/MTFL model and the GMTFS/GMTFL model, the utterances of the training data were split into four different tasks, one for each domain-gender pair. Emotion classification tasks were generated for female performers singing, male performers singing, female performers speaking and male performers speaking. In the ST model, a separate SVM emotion classifier with linear kernel was created for each task. The MTFS/MTFL and the GMTFS/GMTFL models are described in the following sections.

3.2.1. The MTFS/MTFL Model

The MTFS/MTFL model is based on the method introduced in [22, 23]. It considers the tasks as related problems, which contrasts with the ST model's assumption that the tasks are independent. It learns a common representation across tasks using a $L_{2,1}$-norm regularizer, which both couples the tasks and enforces sparsity of the learned weights for the features.

The multi-task learning algorithm in [22, 23] has two settings: (a) feature selection and (b) feature learning. In our paper, MTFS and MTFL correspond to setting (a) and (b), respectively. In (a), regularization is imposed directly on the weight matrix of the features, and the objective function for the multi-task learning problem is given by Eq. (1). The first term of Eq. (1) is the summation of the loss, L, across all T tasks, where m_t is the amount of training data in task t, y_{ti} is the output label of the ith training data of task t, and $\langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle$ is the inner product of $\mathbf{w_t}$ (weight vector of task t) and $\mathbf{x_{ti}}$ (ith training data of task t, with dimensionality d). The second term is a regularization term, where γ is the regularization parameter, W is a $d \times T$ matrix with column $\mathbf{w_t}$, and the $\|W\|_{2,1}^2$ is the $L_{2,1}$-norm.

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

In (b), it is assumed that $W = UA$, where U is a $d \times d$ matrix with orthogonal columns that can transform the original feature space into a new space with an orthogonal basis, and A is the $d \times T$ weight matrix for the transformed feature space. The $L_{2,1}$-norm is imposed on A instead of W. Therefore, the corresponding objective function of the multi-task classification problem becomes

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

In both settings, the convex loss function L can be freely chosen. In this paper, we use hinge loss, such that $L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) = \max(0, 1 - y_{ti} \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle)$. The convex equivalencies of Eq. (1) and Eq. (2) can be solved by iteratively performing the supervised step for task-specific optimization and the unsupervised step for learning the common sparse representations across the tasks. The former step is given by

$$\mathbf{w_t} = \operatorname{argmin} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \langle \mathbf{w_t}, D^{-1} \mathbf{w_t} \rangle \tag{3}$$

where D is initialized as $D = \frac{I}{d}$. Solving (3) is equivalent to solving the linear SVM with a variable transformation trick. The latter step is updating D by Eq. (4) (setting (a)) or Eq. (5) (setting (b)). The $\mathbf{w^i}$ in Eq. (4) refers to the ith row of W, and the ϵ in (5) is a small perturbation parameter used to ensure the convergence of the problem. Note that D is diagonal in Eq. (4), but not in in Eq. (5). This is because in (a), the features are only "chosen" by sparsity, not transformed to a new feature space, as in (b).

$$D = Diag(\lambda), \text{where } \lambda^i = \frac{\|\mathbf{w^i}\|_2}{\|W\|_{2,1}} \tag{4}$$

$$D = \frac{(WW^T + \epsilon I)^{\frac{1}{2}}}{trace(WW^T + \epsilon I)^{\frac{1}{2}}} \tag{5}$$

3.2.2. The GMTFS/GMTFL Model

While the MTFS/MTFL models consider all tasks to be related, the GMTFS/ GMTFL model assumes that the tasks can be clustered into groups, and only the tasks in the same group share information. This model uses an algorithm proposed in [24], which is formed as a mixed integer programming problem. The problem can be solved by iteratively performing two steps: (a) solving the multi-task learning problem as in Eq. (1) or Eq. (2) for each group of tasks, and (b) identifying the optimal group assignments. The optimal number of groups G is not known a priori and is treated as a hyper-parameter.

In this paper, we use both the feature selection setting and the feature learning setting for step (a), which results in two corresponding versions of the group multi-task learning: the GMTFS model and the GMTFL model. As step (b) is not a convex problem, the final solution could be a local optimum. We address this problem by training multiple times. We determine the final output labels using a suite of methods, discussed in detail in section 3.3.

3.3. Output Selection/Fusion Methods

All models, other than the simple model, learn multiple sets of weight vectors for a specific classification problem (i.e., a pair of emotion). Although it is common in the literature to assume that the tasks of the test data are known [22–24], only our oracle method (discussed below) makes this assumption. The decision tree (DT), majority vote (MV), weighted majority vote (WMV), and maximum distance (MD) methods make no such assumption. The selection/fusion methods are defined as follows:

- Oracle assumes that the domain and gender information are known. It outputs the estimated emotion class associated with the known task. The GMTFS/GMTFL identifies a final label by performing a majority vote over the five runs. This approach is used as the baseline.
- DT trains a domain classifier and a gender classifier using SVM with a Radial Basis Function (RBF) kernel. First, the domain and gender of the test data are identified. Second, the emotion is identified using the task-appropriate model. Again, the GMTFS/GMTFL identifies a final label by performing a majority vote over the five runs.
- MV performs majority vote over the output of all tasks and selects the most voted label. In the case of a tie, the label with the smaller index was returned.
- WMV is similar to MV, but each vote is given a weight defined as the distance to the decision hyperplane.
- MD adopts the output label associated with the largest distance to the hyperplane over all tasks and runs.

4. RESULTS AND DISCUSSION

We use RAVDESS as the training data and UMSSED as the test data, because the amount of data in the latter one is insufficient for training when split into four tasks. The hyper-parameters of the models were

Table 1: 4-class emotion classification accuracy of different models and labeling methods (%). For each model, the best result is italicized. The best result over all models and methods is bolded.

	Within-corpus	Cross-corpus					
		Simple	ST	MTFS	MTFL	GMTFS	GMTFL
Oracle	54.76	45.24	44.05	49.40	44.64	45.24	47.02
DT			42.86	45.83	44.64	45.24	45.83
MV			50.60	50.00	47.62	51.19	51.79
WMV			*54.17*	*52.38*	*53.57*	***57.14***	51.19
MD			52.98	*52.38*	52.38	52.38	*54.17*

Table 2: Per-performer 4-class emotion classification accuracy of the best performance of each model (%). The highest accuracy for each performer is bolded.

	Within-corpus	Cross-corpus					
		Simple	ST	MTFS	MTFL	GMTFS	GMTFL
P1	62.50	55.36	60.71	53.57	62.50	**64.29**	62.50
P2	**51.79**	37.50	42.86	44.64	41.07	44.64	41.07
P3	50.00	42.86	58.93	58.93	57.14	**62.50**	58.93

selected on the training data using a grid search by optimizing the 5-fold cross-validation accuracy. The linear SVMs were solved using Liblinear [25]. As a reference, we also performed evaluation using a within-corpus setting on the UMSSED in a leave-one-performer-out cross-validation manner, using the simple model. Other models that involve multiple tasks were not tested because the number of training data would be very limited for each task.

Table 1 shows the 4-class emotion classification accuracy of different models and labeling methods, along with the within-corpus 4-class emotion classification accuracy. The domain classification and gender classification used for DT have accuracies of 95.2% and 70.8%, respectively. The GMTFS model achieved the highest accuracy of 57.14%, which is higher than the within-corpus performance. The performance difference is not significant (per-performer accuracy, paired t-test). This is because there are only three performers in the UMSSED, making it difficult to assess significance. However, we show the per-performer 4-class emotion classification accuracy of the best label generating method for each model in Table 2. It can be seen that the GMTFS model improves the performance of two out of three performers. This suggests that despite the differences in recording conditions, noise overlay, lexical and melodic content of the two datasets, our proposed cross-corpus approach can still achieve comparable results to within-corpus training and testing.

Our results show that GMTFS either outperforms or achieves comparable results to the other models for all performers. That is, the best performance was achieved when information sharing only happens within groups. This may indicate that the tasks were neither completely coupled nor completely uncoupled, and that some tasks have closer relationships than others. This is in line with our previous finding that models that explicitly control for task relatedness outperform those that do not for cross-domain emotion recognition [9]. We also notice that the best accuracy of all models with four tasks is higher than that of the simple model. This suggests that the performance of emotion recognition systems can be improved by explicitly considering communication domain and gender.

The GMTFS model, which enforces task relatedness and sparsity directly on the original feature space, works better than GMTFL, which first transforms features onto a new space with orthogonal basis. We hypothesize that this occurs because the lexical information in speech, rather than the emotion information, is dominant. Consequently, the transformation process obscures the information of interest. Past work [26] has also found that PCA leads to loss of information in emotion recognition.

We found that weighted majority vote was the most effective output selection/fusion strategy, followed by maximum distance. The unweighted version of the majority vote has a lower accuracy than the weighted version. Oracle has a slightly higher performance than DT, indicating that knowledge of the "correct" task is advantageous. However, it is interesting to see that we can achieve an average performance gain of 4.17%, 7.62% and 6.79% by using MV, WMV and MD, respectively, compared to oracle. This may suggest that the tasks in the training set are not guaranteed to have a one-to-one correspondence to tasks in the testing set, due to the differences in recording conditions, performers, lexical content and melodic content. Another possibility is that the variability of the training data is not sufficient to make the learned classifiers good representatives of the tasks. Interestingly, while ST with oracle or DT has the lowest accuracy among all methods, ST with MWV achieves the second best result. The reason might be that by fusing the outputs, the prediction is not only based on a single task, but also takes the knowledge from other tasks into account.

5. CONCLUSION AND FUTURE WORKS

In this paper, we present a multi-task learning approach to recognize emotion from song and from speech. We consider four different models: the simple model, the single-task model, the multi-task feature selection/learning model, and the group multi-task feature selection/learning model. These models correspond to the assumptions that the tasks are identical, independent, related, or partially related, respectively. We propose five different output selection/fusion methods, including oracle, decision tree, majority vote, weighted majority vote and maximum distance. We performed experiments in a cross-corpus setting, using RAVDESS as training data and UMSSED as test data to study the generalizability of the models and methods.

Among all models, group multi-task feature selection achieves the highest accuracy. This suggests that some tasks are more closely related than others, and sharing information only among closely related tasks is beneficial. A limitation of this work is that we are not able to get a static grouping for the tasks, because the GMTFS/GMTFL is non-convex and has unstable outputs. We are interested in measuring task relatedness explicitly in future work. In addition, it would also be interesting to explore the situation when only domain or gender is used as task-separator, and the performance of knowledge-based grouping (e.g. grouping by gender or domain) compared to data-based grouping as in the GMTFS/GMTFL model.

Among all labeling methods, the weighted majority vote works the best. The fact that this method is advantageous, compared to the oracle method, reveals that the one-to-one correspondence in the training and testing tasks is not guaranteed, due to dataset differences or insufficient variability in the training data. It will be interesting to continue to explore this problem by testing over additional datasets, treating dataset as an additional task separator.

Finally, cross-corpus classification is closely related to domain adaptation and transductive transfer learning [27]. There have been many works studying ways to improve performance in this setting, such as importance sampling and re-weighting methods like kernel-mean matching [28] and feature representation transferring with denoising autoencoder [29, 30]. Our future work will involve combining these methods with multi-task learning.

6. REFERENCES

[1] Thurid Vogt, Elisabeth André, and Johannes 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*, pp. 75–91. 2008.

[2] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] Klaus R Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1, pp. 227–256, 2003.

[4] Emily Mower, Maja J Mataric, and Shrikanth Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

[5] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull, "Music emotion recognition: A state of the art review," in *Proceedings of International Society for Music Information Retrieval*, 2010, pp. 255–266.

[6] Yi-Hsuan Yang and Homer H Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, pp. 40, 2012.

[7] Klaus R. Scherer, Johan Sundberg, Lucas Tamarit, and Gláucia L. Salomão, "Comparing the acoustic expression of emotion in the speaking and the singing voice," *Computer Speech & Language*, 2013.

[8] Steven R Livingstone, Katlyn Peck, and Frank A Russo, "Acoustic differences in the speaking and singing voice," in *Proceedings of Meetings on Acoustics*, 2013, vol. 19.

[9] Biqiao Zhang, Georg Essl, and Emily Mower Provost, "Recognizing emotion from singing and speaking using shared models," in *Proceedings of Affective Computing and Intelligent Interaction*, 2015.

[10] Dimitrios Ververidis and Constantine Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proceedings of the Signal Processing Conference*, 2004, pp. 341–344.

[11] Chul Min Lee and Shrikanth S Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[12] Thurid Vogt and Elisabeth André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proceedings of the Language Resources and Evaluation Conference*, 2006.

[13] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[14] Iulia Lefter, Leon JM Rothkrantz, Pascal Wiggers, and David A Van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," in *Proceedings of Text, Speech and Dialogue*, 2010, pp. 353–360.

[15] Björn Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?," in *Proceedings of INTERSPEECH*, 2011, pp. 1553–1556.

[16] Björn Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll, "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization," in *Proceedings of the Afeka-AVIOS Speech Processing Conference*, 2011.

[17] Song Peng, Jin Yun, Zhao Li, and Xin Minghai, "Speech emotion recognition using transfer learning," *IEICE Transactions on Information and Systems*, vol. 97, no. 9, pp. 2530–2532, 2014.

[18] Biqiao Zhang, Emily Mower Provost, Robert Swedberg, and Georg Essl, "Predicting emotion perception across domains: A study of singing and speaking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

[19] Steven R Livingstone, Katlyn Peck, and Frank A Russo, "Ravdess: The ryerson audio-visual database of emotional speech and song," in *Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science*, 2012.

[20] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH*, 2013.

[21] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, 2010, pp. 1459–1462.

[22] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil, "Multi-task feature learning," in *Advances in Neural Information Processing Systems*, 2007, vol. 19.

[23] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[24] Zhuoliang Kang, Kristen Grauman, and Fei Sha, "Learning with whom to share in multi-task feature learning," in *Proceedings of the International Conference on Machine Learning*, 2011, pp. 521–528.

[25] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[26] Yongjin Wang and Ling Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[27] Andrew Arnold, Ramesh Nallapati, and William W Cohen, "A comparative study of methods for transductive transfer learning," in *Data Mining Workshops, IEEE International Conference on Data Mining*, 2007, pp. 77–82.

[28] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola, "Correcting sample selection bias by unlabeled data," in *Advances in neural information processing systems*, 2006, pp. 601–608.

[29] Jun Deng, Zixing Zhang, Florian Eyben, and Bjorn Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[30] Jun Deng, Zixing Zhang, and Björn Schuller, "Linked source and target domain subspace feature transfer learning–exemplified by speech emotion recognition," in *Proceedings of the International Conference on Pattern Recognition*, 2014, pp. 761–766.