INCREMENTAL ADAPTATION USING ACTIVE LEARNING FOR ACOUSTIC EMOTION RECOGNITION

Mohammed Abdelwahab and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

mxa129730@utdallas.edu, busso@utdallas.edu

ABSTRACT

The performance of speech emotion classifiers greatly degrade when the training conditions do not match the testing conditions. This problem is observed in cross-corpora evaluations, even when the corpora are similar. The lack of generalization is particularly problematic when the emotion classifiers are used in real applications. This study addresses this problem by combining *active learning* (AL) and supervised *domain adaptation* (DA) using an elegant approach for *support vector machine* (SVM). Active learning selects samples in the new domain that are used to adapt the speech classification models using domain adaptation. This paper demonstrates that we can increase the performance of the speech recognition system by incrementally adapting the models using carefully selected samples available after active learning. We propose a novel iterative fast converging incremental adaptation algorithm that only uses correctly classified samples at each iteration. This conservative framework creates sequences of smooth changes in the decision hyperplane, resulting in statistically significant improvements over conventional schemes that adapt the models at once using all the available data.

Index Terms— Emotion recognition, active learning, SVM adaptation

1. INTRODUCTION

Recognizing emotions from speech is an important problem with clear applications in many domains. In order for current speech emotion classifiers to perform well, the data used for training and testing the models should be similar, ideally coming from the same domain. The performance degrades heavily when there is a mismatch [1], so it is important to develop strategies that can mitigate the drop in performance in the presence of a new domain. We formulate this problem by having a *source* domain with emotional labels, which is used to train the model, and a *target* domain with unlabeled data, which is used to test the model. While the most obvious solution is to annotate enough data in the target domain to achieve good performance, generating labels for new data is both expensive and time consuming [2]. It commonly requires multiple raters to evaluate the data, whose annotations are later fused to generate gold standard labels. It is important that the proposed method (1) requires limited labeled data from the target domain, and (2) efficiently use these labeled data.

A popular approach for building a classifier that performs well in a new domain is *active learning* (AL), where the task is to identify the most informative samples in the target domain that can be used to improve the classifier [3]. Since not all the samples are equally beneficial to the classifier [4], different strategies are used to select the most useful samples, which are later annotated by human raters. The new samples are commonly added to the source domain to train the classifier. Another common approach is to adapt one or more classifiers trained on domains that are different, but close to the new domain [5]. *Domain adaptation* (DA) can be used with small set of labeled data (supervised) or with labels automatically generated by the classifiers (unsupervised). DA uses classifiers trained on source domains, correcting the hyperplanes by leveraging data from the target domain. Combining AL and DA can be an appealing framework to reduce the required annotations, reducing the mismatch between train and test conditions.

This paper explores the use of active learning, along with supervised domain adaptation for *support vector machines* (SVMs) in speech emotion recognition. The key contribution of the paper is the proposed data selection framework used to adapt the classifiers, which efficiently uses the new labeled data to improve the performance of the classifiers. We use AL to identify a limited set of samples from the target domain. Instead of using all the samples for DA, we propose an iterative algorithm where we consider only the samples where the predicted labels at a given step match the annotations of the new labeled data. The process is repeated multiple times, where we evaluate different stopping criteria. This is a conservative approach to adapt the SVM classifier, avoiding large changes in the hyperplane caused by including samples in the wrong side of the current decision boundary. The promising results show the importance of data selection in adapting classifiers to a new domain, providing an ideal framework to reduce the amount of data and time needed to generate a robust classifier.

2. RELATED WORK

The key challenge in speech emotion recognition is to build classifiers that perform well under various conditions. The cross-corpora evaluation in Shami and Verhelst [6] demonstrated the drop in classification performance observed when training on one emotional corpus and testing on another. Several approaches have been proposed to solve this problem.

Shami and Verhelst [6] proposed to include more variability in the training data by merging emotional databases. They demonstrated that it is possible to achieve classification performance comparable to within-corpus results. The main approach to attenuate the mismatch between train and test conditions is to minimize the differences between both domains. Hassan et al. [7] used *kernel mean matching* (KMM), *Kullback-Leibler importance estimation procedure* (KLEIP), and *unconstrained least-squares importance fitting* (uLSIF) to increase the weight of the training data that matches the test data distribution.

Studies have explored feature transformation to reduce mismatches between train and test conditions. Zhang et al. [8] showed that by separately normalizing the features of each corpus, it is possible to minimize cross-corpus variability. Deng et al. [9] trained a sparse autoencoder on the target data and used it to reconstruct the source data. This approach used feature transformation in a way that

This work was funded by NSF CAREER award IIS-1453781.

exploits the underlying structure in emotional speech learned from the target data. Deng et al. [10] used two denoising autoencoders. The first autoencoder is trained on the target data and the second autoencoder is trained on the source data, but it is constrained to be close to the first autoencoder. The second autoencoder is then used to reconstruct both source and target domain data.

Deng et al. [11] used autoencoders to find common feature representation across the domains. They trained the autoencoder such that it minimizes the reconstruction error on both domains. Motivated by the work of Deng et al. [11], Sagha et al. [12] used *principal component analysis* (PCA) along with *Kernel canonical correlation analysis* (KCCA) to find views with the highest correlation between the source and target corpora. First, they used PCA to represent the feature space of the source and target data. Then, the features for source and target domains are projected using the PCA in both domains. Finally, they used KCCA to select the top N dimensions that maximize the correlation between the views.

Zhang et al. [11] proposed an enhanced versions of Self-Training and Co-Training, where they kept track of the changes in the labels assigned to sentences added to the training set. This approach allowed them to detect and correct noise in the labels used for training. Zhang et al. [12] combined semi-supervised and uncertainty based active learning. The proposed algorithm outperformed methods that separately used either active learning or semi-supervised learning. They were able to maintain the classification performance by using only 25% of the data. Also using active learning, Zhang et al. [13, 14] studied different querying criteria and showed that sparse instance selection boosts performance and reduces the amount of data annotation needed in the case of unbalanced classes.

The contribution of this paper is an appealing framework to combine *active learning* (AL) and *domain adaptation* (DA). The closest paper related to this work is the study by Zhang et al. [12]. The key difference is the data selection used after annotating the data from the target domain. This paper proposes an iterative domain adaptation algorithm that considers the prediction of the classifiers on the labeled data in the target domain. By adapting only with the samples that are correctly recognized, we create a conservative adaptation scheme that increases the performance of the system.

3. DATABASES

We evaluate our experiments in a cross corpus setting, with the USC-IEMOCAP database [15] as our source domain (training) and the MSP-IMPROV database [16] as our target domain (testing). Table 1 shows the turn distribution across the emotions for both databases.

3.1. USC-IEMOCAP corpus

The USC-IEMOCAP database is an audiovisual corpus recorded from ten actors during dyadic interaction [15]. It has approximately 12 hours of recordings with detailed motion capture information carefully synchronized with audio (this study only uses the audio). The goal of the data collection was to elicit natural emotions within a controlled setting. This goal was achieved with two elicitation frameworks: emotional scripts, and improvisation of hypothetical scenarios. These approaches allowed the actors to express spontaneous emotional behaviors driven by the context, as opposed to read speech displaying prototypical emotions. Several dyadic interactions were recorded and manually segmented into turns. Each turn was emotionally annotated by three evaluators into categorical emotions, where individual annotations are later merged using majority vote rule. For this study, we combine samples labeled as excited and happiness, creating a four class problem: anger, happiness, sadness and neutral speech.

Table 1. Distribution of turns per class.

Databases	# turns per class				
	A	H	S	N	$\sum$
USC-IEMOCAP	1103	1636	1084	1708	5531
MSP-IMPROV	788	2624	886	3454	7752

[A - Anger; H - Happiness; S - Sadness; N - Neutrality]

3.2. MSP-IMPROV corpus

MSP-IMPROV is a multimodal emotional database recorded from actors interacting in dyadic sessions [16]. The recording were carefully designed to promote natural emotional behaviors, while maintaining control over lexical and emotional content. The corpus relied on a novel elicitation scheme, where two actors improvise scenarios that lead one of them to utter target sentences. For each of these target sentences, four emotional scenarios were created to contextualize the sentence to elicit happy, angry, sad and neutral reactions, respectively. The approach allows the actor to express emotions elicited by the scenarios, avoiding prototypical reactions that are characteristic of other acted emotional corpus. Busso et al. [16] shows that the target sentences occurring within these improvised dyadic interactions were perceived more natural than read renditions of the same sentences. The MSP-IMPROV corpus not only includes the target sentences, but also other sentences during the improvisation and natural interactions between actors during the breaks. The corpus consists of 8,438 turns (over 9 hours) of emotional sentences recorded from 12 actors. The turns are manually segmented into speaking turns, which are emotionally annotated with perceptual evaluations using crowdsourcing [17]. The labels include four categorical emotions (anger, happiness, sadness, or neutrality) as well as dimensional attributes scores (arousal, valence, and dominance). The label assigned to each turn is decided by the majority vote rule, where we only consider turns that reach an agreement.

4. METHODOLOGY

The section presents the proposed algorithm that combines both AL and DA to build a classifier that improves performance in new domains. This work is motivated by Adapt-SVM [18, 19], however, this can be applied to other model parameter adaptation algorithms. We first describe the AL criteria used in this study (Sec. 4.1) and the details of the DA approach (Sec. 4.2).

We define the following notation used in the paper: $\boldsymbol{x}_i$ is a d-dimensional feature vector, y_i is the class label for feature vector i, $\mathcal{D}^s = ([\boldsymbol{x}_1^s, y_1^s], \ldots, [\boldsymbol{x}_M^s, y_M^s])$ denotes the labeled source domain data, $\mathcal{D}^t = (\boldsymbol{x}_1^t, \ldots, \boldsymbol{x}_{N_u}^t)$ is the target domain data, which is assumed to be unlabeled. L is the subset of sentences in the target domain that we want to annotate. After collecting the annotations, this set becomes $\mathcal{D}_l^t = ([\boldsymbol{x}_1^t, y_1^t], \ldots, [\boldsymbol{x}_{N_l}^t, y_{N_l}^t])$ (i.e., the target domain data labeled by AL).

4.1. Active Learning (AL)

The task in AL is to identify sentences in the target domain that can improve the classification performance when added to the training set. The most common AL strategies are: *uncertainty sampling*, selecting samples where the classifier is less confidence; *committee based query*, selecting samples where multiple classifiers disagree; *estimated expected error*, selecting samples that if added to the labeled set would minimize the expected error in the future; *variance reduction*, selecting samples that maximize future variance reduction derived from the estimated distribution of the model's outputs; and, *density weighted strategies*, selecting samples according to the

Algorithm 1 Adaptation using all the samples -Baseline

1: Train a classifier $\mathcal{H}$ using $\mathcal{D}^s$, and then classify $\mathcal{D}^t$
2: Rank the classified labels based on classification confidence
3: Select a subset L with the lowest confidence
4: Submit L to be annotated (i.e., AL step)
5: Remove L from testing data $\mathcal{D}^t$
6: Adapt classifier $\mathcal{H}$ using subset $\mathcal{D}_l^t$. eq. (2)
7: Evaluate the adapted classifier on the testing data $\mathcal{D}^t$

underlying distribution, avoiding selecting outliers. This paper uses uncertainty sampling querying criteria, since it works well for max-margin algorithms such as SVM. The most informative samples for SVM classifiers are the support vectors, which are the samples that lie between the hyperplane and the margins. These samples are also the ones the classifier has the least confidence in their labels, since they are closer to the hyperplane.

4.2. Domain Adaptation (DA)

Adapt-SVM was proposed by Yang et al. [18]. It aims to learn a new classifier decision function from the adaptation data. This is done by learning a delta function Δf:

$$f(\boldsymbol{x}) = f^s(\boldsymbol{x}) + \Delta f(\boldsymbol{x}) = f^s(\boldsymbol{x}) + \Delta \boldsymbol{w}^T \phi(\boldsymbol{x}) \tag{1}$$

where $f^s(\boldsymbol{x})$ is the decision function learned from the training data from source domain and $\boldsymbol{w}$ are the parameters to be estimated from the labeled data in the new domain. The problem is formulated such that the new decision function minimizes both the error on the adaptation data and the deviation from the old decision function. They defined the objective function as:

$$\begin{aligned} \min_{\boldsymbol{w}} \ & \frac{1}{2}\|\Delta \boldsymbol{w}\| + C\sum_{i=1}^{N} \xi_i \\ \boldsymbol{s.t.} \ & \xi_i \geq 0, \\ & y_i(f^s(\boldsymbol{x}_i) + \Delta \boldsymbol{w}^T \phi(\boldsymbol{x}_i)) \geq 1 - \xi_i \ \ \forall (\boldsymbol{x}_i, y_i) \in \mathcal{D}_l^t \end{aligned} \tag{2}$$

where $\phi(\boldsymbol{x}_i)$ is a kernel mapping function. The cost factor C controls the contribution of the loss function on adaptation data.

4.3. Combining Active Learning and Domain Adaptation

Algorithm 1 describes the straightforward approach to combine AL and DA. After training the SVM with the source domain, AL selects samples according to the uncertainty sampling querying criteria (samples close to the hyperplane). After the annotation process of $\mathcal{D}_l^t$, all the labeled data are used to adapt the hyperplane of the SVM with Equation (2). This approach is used as our baseline to compare our proposed approach.

4.3.1. Motivation of the Proposed Approach

The data used for adaptation ($\mathcal{D}_l^t$) have a significant role in the performance of the adapted classifier. Samples from $\mathcal{D}_l^t$ in the wrong side of the hyperplane (i.e., the classifier incorrectly predicts its label) can have an important influence on the decision boundary. As the classifier adapt its decision boundary to correct these cases, the influence of the original model trained with the source domain decreases, unlearning information that may be important. This is particularly relevant in emotion recognition problems, when the labels are annotated by human subjects. The perception of emotions is intrinsically listener-dependent. Therefore, the emotional labels derived from perceptual evaluations are noisy with low inter-evaluator agreement [1]. Even if the label is correctly annotated, the sample from $\mathcal{D}_l^t$ that disagrees with the classifier can greatly affect the performance of the system as it will require a larger shift in the hyperplane compared to a sample that agrees with the classifier. Since all the labels in $\mathcal{D}_l^t$ are close to the hyperplane, many of these samples will become support vectors. It is crucial that the adapted model does not unlearn useful information from the original model while trying to adapt the boundary to account for these samples. We hypothesize that a conservative adaptation approach can prevent these problems. We propose to consider whether the current model correctly or incorrectly classifies samples in $\mathcal{D}_l^t$. We derive an elegant iterative algorithm that at each step evaluates whether the current model agrees with available labeled data, adapting the models with only the samples in the correct side of the current hyperplane.

4.3.2. Proposed Algorithm

Algorithm 2 shows the proposed approach, where we incrementally adapt only with samples from $\mathcal{D}_l^t$ that the classifier has correctly predicted at a given iteration. Using the dataset selected by AL ($\mathcal{D}_l^t$), we evaluate our current classifier. We identify the subset N_a from $\mathcal{D}_l^t$ where the labels are correctly predicted. We update the models with Equation (2), using only samples in N_a. This process continues until meeting the stopping criterion, which is discussed in Section 4.3.3. It is important to emphasis that Algorithm 1 (baseline) and Algorithm 2 (proposed approach) operate with the exact same data selected by AL. Our proposed algorithm does not require to annotate new samples during each iteration. The key difference between both algorithms is that Algorithm 1 uses all of the available data in one iteration, while Algorithm 2 uses a subset of $\mathcal{D}_l^t$ in each iteration.

4.3.3. Stopping Criteria

As described in the experimental evaluation, we implement this approach as a multi-class problem for four basic emotions: anger, happiness, sadness, and neutrality. Based on the multi-class aspect of the problem, various stopping criteria are considered for the incremental adaptation algorithm. *Criterion 1* stops the algorithm when the number of emotional classes represented in N_a is less than three. *Criterion 2* stops the algorithm when the number of emotional classes represented in N_a is less than two. *Criterion 3* adds an extra iteration to Criterion 2, where we add all the remaining samples, including the samples in N_a that are incorrectly classified.

We implement this multi-class SVM problem with the "one-against-all" method, where a binary SVM classifier is built for each class. If at a certain iteration the correctly classified data only belong to two classes, then only the classifiers related to those two classes are adapted; the rest of the classifiers remain unchanged.

5. EXPERIMENTS AND RESULTS

5.1. Feature Selection

We used OpenSMILE framework to extract the acoustic features [20]. We adopted the feature set released for the INTERSPEECH 2013 Computational Paralinguistic Challenge [21], consisting of 6373 features including turn based statistics of prosodic, spectral and voice quality features. We reduced the dimensionality of the feature vector using a two step approach. First, we use *Correlation Feature Selection* (CFS) to remove redundant features. CFS adds features one at a time that correlate with the labels, but that are not correlated with previously selected features. We reduce the feature set to 3000, reducing the computational complexity of the feature selection process. The second step uses *forward feature selection* (FFS), where the cost function maximizes the performance

Algorithm 2 Incremental Adaptation -Proposed Algorithm

Train a classifier $\mathcal{H}$ using $\mathcal{D}^s$, and then classify $\mathcal{D}^t$
Rank the classified labels based on classification confidence
Select a subset L with the lowest confidence
Submit L to be annotated (i.e., AL step)
Remove L from testing data $\mathcal{D}^t$
Select subset N_a from $\mathcal{D}_l^t$ that the classifier $\mathcal{H}$ predicted correctly
while N_a contains at least two class labels **do**
    Adapt classifier $\mathcal{H}^i$ using subset N_a. eq. (2)
    Remove N_a from $\mathcal{D}_l^t$
    Select subset N_a from $\mathcal{D}_l^t$ that the adapted classifier $\mathcal{H}^{i+1}$ predicted correctly
Evaluate the adapted classifier on the testing data $\mathcal{D}^t$

Table 2. Average F1-Score

Algorithm 1

criteria	# samples	before	After	# iterations
Algorithm 1	200	45.48	46.70	1

Algorithm 2

criteria	# samples	before	After	# iterations
1^{st} step	64.4	45.48	47.78	1
criterion 1	117.8	45.48	48.28	3.71
criterion 2	123.6	45.48	48.13	4.71
criterion 3	200	45.48	45.47	5.71

of the SVMs over the source domain (i.e., train set). The dimension of the feature set for each experiment is 300.

5.2. Experiment

We used LibSVM toolkit to train our SVM classifiers [22]. We implement the multi-class classifier as four "one-against-all" classifiers with a linear kernel and a cost factor C set to 1 (our previous work showed good performance with this setting). We use random subsampling to ensure balanced classes in both train and test conditions. After sampling, the number of instances for training and testing are 4336 and 3152, respectively. We predict the labels for the testing data (target domain) with SVM classifiers, selecting the least confident samples to be annotated with AL. We choose 200 sentences, which are removed from the testing set for fair comparison. We compare Algorithm 1 (baseline) and Algorithm 2 (proposed approach). We repeat the evaluation 20 times to ensure consistent results across the random sub-sampling iterations, reporting the average F1-scores. We assert performance at p-value = 0.05, using matched pair population mean t-test.

5.3. Results

Table 2 shows the F1-score of the classifiers before and after adaptation, along with the average number of samples used for adaptation. The table includes the results for Algorithm 1, and different stopping criterion for Algorithm 2. Algorithm 1 uses all 200 samples to adapt the classifiers trained with the source domain. The adaptation improves the classification performance in 1.23%, which is statistically significant over the F1-score before adaptation (p-value $< 2e^{-4}$).

Algorithm 2 incrementally adapts the classifiers with the subset of samples correctly recognized by the SVMs. Table 2 gives the results after the first iteration, and after each of the stoping criteria. The increase of F1-score just by performing the first iteration is 2.31%, where we only use 64.4 samples on average. The performance increases as we add more iterations. The classification performance using criterion 1 is 48.28%, which represents an absolute improvement of 2.80% over the F1-scores before adaptation, which is statistically significant (p-value $< 2e^{-11}$). This criterion uses 3.71 iteration in average using 117.8 out of the 200 sentences in $\mathcal{D}_l^t$. The classification performance for criterion 2 is 48.13%, which is 2.65% better than the F1-score before adaptation. This result is also statistically significant over the result before adaptation (p-value $< 2e^{-10}$). The last row of Table 2 shows the performance for criterion 3, where the goal is to show the degradation in performance caused by adapting the classifiers with all the available data, including the misclassified samples. In just one iteration, the classification performance drops from 48.13% to 45.47%. This step requires the classifiers to change the hyperplane to accommodate samples in the wrong side of the hyperplane, affecting the decision boundary learnt with the data from the source domain.

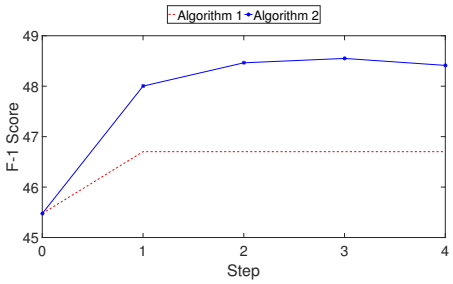

Fig. 1. Average performance of Algorithm 2 across iterations. The F1-score of Algorithm 1 is significantly lower.

When we compare the performance between Algorithm 1 and Algorithm 2, we observe that the proposed approach is significantly better. With the stopping criterion 1, the F1-score is 1.58% better than Algorithm 1 (p-value $< 5e^{-5}$). With the stopping criterion 2, the F1-score is 1.42% better than Algorithm 1 (p-value $< 1e^{-4}$). The improvement in performance is achieved with an average of 3.71 iterations for criterion 1 or 4.71 iterations for criterion 2. The proposed approach does not require high number of iterations.

Figure 1 gives the classification performance for Algorithm 2 per iteration (solid blue line). For comparison, we plot the performance for Algorithm1 (dashed red line). Both curves start with the performance of the classifiers before adaptation. The figure shows the advantage of the proposed algorithm over the standard approach, where even at the first iteration the difference in performance is significant, even though the proposed approach approximately uses only a third of the labeled data. Algorithm 2 converges after few iterations.

6. CONCLUSIONS

This paper proposed an algorithm for incremental supervised SVM domain adaptation. We showed the importance of selecting the data used for adaptation to match the classifiers' predictions, where by adapting with the correct data we can achieve a significant improvement. The approach uses a portion of the labeled dataset, converging to a stable performance after few iterations (between 3 and 5). The evaluation showed that adapting with misclassified data causes a degradation in classification performance.

For future work, we want to modify the optimization function so that we can make use of all of the available data. We can achieve this goal by introducing a variable regularization parameter for each instance. This framework will allow us to reduce the weight assigned to the misclassified samples, instead of ignoring them, as we do in the proposed approach. We will also explore if the proposed algorithm can be used in different supervised domain adaptation approaches (e.g. adapting deep learning frameworks).

7. REFERENCES

[1] C. Busso, M. Bulut, and S.S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds., pp. 110–127. Oxford University Press, New York, NY, USA, November 2013.

[2] D. Braha, *Data Mining for Design and Manufacturing: Methods and Applications*, Kluwer Academic Publishers, Norwell, MA, USA, October 2001.

[3] B. Settles, *Active Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, Long Island, NY, USA, July 2012.

[4] D. Le and E. Mower Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2015)*, Xi'an, China, September 2015, pp. 146–152.

[5] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, April 2015, pp. 5058–5062.

[6] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: A multi-corpus study," in *Speaker Classification II*, C. Müller, Ed., vol. 4441 of *Lecture Notes in Computer Science*, pp. 43–56. Springer-Verlag Berlin Heidelberg, Berlin, Germany, August 2007.

[7] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, July 2013.

[8] Z. Zhang, F. Weninger, M. Wollmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*, Waikoloa, HI, USA, December 2011, pp. 523–528.

[9] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 511–516.

[10] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, September 2014.

[11] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5185–5189.

[12] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 115–126, January 2015.

[13] Z. Zhang, J. Deng, E. Marchi, and B. Schuller, "Active learning by label uncertainty for acoustic emotion recognition," in *Interspeech 2013*, Lyon, France, August 2013, pp. 2856–2860.

[14] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Interspeech 2012*, Portland, Oregon, USA, September 2012, pp. 362–365.

[15] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[16] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. To appear, 2015.

[17] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[18] J. Yang, R. Yan, and A.G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *ACM international conference on Multimedia (MM 2007)*, Augsburg, Germany, September 2007, pp. 188–197.

[19] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *International Conference on Computer Vision (ICCV 2011)*, Barcelona, Spain, November 2011, pp. 2252–2259.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[21] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[22] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27:1–27, April 2011.