ENSEMBLE FEATURE SELECTION FOR DOMAIN ADAPTATION IN SPEECH 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

When emotion recognition systems are used in new domains, the classification performance usually drops due to mismatches between training and testing conditions. Annotations of new data in the new domain is expensive and time demanding. Therefore, it is important to design strategies that efficiently use limited amount of new data to improve the robustness of the classification system. The use of ensembles is an attractive solution, since they can be built to perform well across different mismatches. The key challenge is to create ensembles that are diverse. This paper proposes the use of active learning along with feature selection to build a diverse ensemble that performs well in the new domain. The diversity and accuracy of the ensemble are achieved by (1) training emotional classifiers with bias toward specific emotions, (2) eliminating overlap in the feature sets of the ensemble, and (3) conducting feature selection by maximizing the performance over the new labeled data. We study various data selection criteria, and different sample sizes to determine the best approach toward building a stable diverse ensemble that generalize well on new domains.

Index Terms— Emotion recognition, ensemble, active learning, machine learning.

1. INTRODUCTION

An important challenge in speech emotion recognition is to build a robust classifier that performs well regardless of the application [1]. To account for the environmental, emotional and idiosyncratic variabilities, the classifier would need enough labeled data which is expensive and not easy to collect. It is important to identify alternatives in machine learning that effectively use limited labeled data from a new domain to improve the classification performance. An appealing framework is the use of ensembles.

This paper explores the use of ensembles in emotion recognition, which effectively use limited data from a new domain. The idea behind ensembles is to train multiple classifiers which are later fused, improving the overall performance of the system. Ensembles have multiple benefits [2, 3]. With abundant data, the classifiers can be trained in parallel using partitions of the dataset. With limited data, the ensemble can use bootstrapping algorithms [4, 5]. Another benefit of ensembles is that they are more resilient to performance degradation due to mismatched conditions. While some classifiers in the ensemble may be affected, the diversity of the classifiers can attenuate the drop in performance by using the ensemble in a different domain. For these reasons, ensembles have been used in emotion recognition from speech [6, 6–8] and even for multimodal emotion recognition [9–12].

The challenge in building ensembles is to increase the diversity between classifiers [2]. If all the classifiers in the ensemble are similar, their fusion will not provide additional benefits. This paper proposes a feature selection algorithm that maximizes the performance of an ensemble on the target domain and ensures the diversity of its speech emotion classifiers. The algorithm makes use of *active learning* (AL) to minimize the amount of labeled data needed for feature selection. The diversity of the ensemble is achieved by (1) training emotional classifiers with bias toward specific emotions, (2) eliminating overlap in the feature sets of the ensemble, and (3) conducting feature selection by maximizing the performance over the labeled data from the target domain. We study how the size and criteria of selecting samples used in feature selection affect the ensemble performance. The proposed approach improves performance by increasing the robustness of the ensemble.

2. RELATION TO PRIOR WORK

Previous studies have shown the advantages of using ensembles in emotion recognition. Lee and Narayanan [13] used an ensemble of three classifiers to detect negative emotion in spoken dialog, where each classifier was separately trained with acoustic, lexical and discourse features, respectively. They showed that the fusion of the classifiers' decisions outperform that of a single classifier. They also showed the importance of the ensemble diversity, where improvements can only be achieved with complementary classifiers.

There are various approaches to build an ensemble that would perform well in a new domain. Schuller *et al.* [14] showed that it was possible to outperform the performance of a classifier trained on a single database, by adding different databases during training either by pooling them into a single training set or by training multiple classifiers which are then fused. Adding databases adds diversity in the training data in terms of language, recording settings, and noise, which helps building more robust classifiers. Duan *et al.* [15] proposed to combine classifiers trained on different databases by weighting them based on the distance in the distribution between training and testing datasets. The algorithm uses the fusion of the classifiers to optimize a new *support vector machine* (SVM) classifier over a small amount of labeled data from the test domain, regularized by the unlabeled data.

Dai *et al.* [16] proposed a modified version of Adaboost, where they divided the training data into two partitions: one that agrees with the test data distribution, and another that follows a different distribution. They limited the error caused by training and testing mismatches by increasing the weight on misclassified instances from the similar distribution and by reducing the weight on instances from the dissimilar distribution. Xia *et al.* [17] used a similar approach for sentiment classification, where they used PCA to select samples from the source domain that follow the same distribution of the target domain. After training an ensemble of classifiers with different feature sets, they fused them using weights learned from their performance on a small labeled set from the target domain. Instead of having global weights for the ensemble classifiers, Gao *et al.* [18] proposed a locally weighted ensemble framework. The proposed framework mitigates the uncertainty caused by different models assigning different labels to test instances.

The contribution of this paper is on the feature selection step in training the classifiers in the ensemble. The novel algorithm min-

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

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

Table 1: Distribution of turns per emotional class (A: Anger; H: Happiness; S: Sadness; N: Neutrality).

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

imizes the mismatches between source and target domains, maximizes performance in the target domain using limited labeled data, and increases the diversity in the classifiers.

3. DATABASES AND ACOUSTIC FEATURES

We evaluate our experiments in a cross corpus evaluation, where we use the USC-IEMOCAP corpus as our source domain (i.e., training) and the MSP-IMPROV databases as our target domain (i.e., testing). Table 1 shows the speaker turns across emotions for both databases.

The USC-IEMOCAP database has about 12 hours of audiovisual data recorded from ten actors [19]. The recordings were collected with scripts, and improvisation of hypothetical scenarios, which allowed the actors to express spontaneous emotional behaviors driven by the context. Several dyadic interactions were recorded and manually segmented into turns. Each turn was annotated into ten categorical emotions (e.g., anger, happiness, or neutrality – three evaluators per turn), as well as dimensional scores (valence, activation, dominance – two evaluators per turn). This study focus on categorical emotions. We consider only four emotional categories: anger, happiness, sadness and neutrality. The happy set also includes speaking turns labeled as excited.

The MSP-IMPROV corpus is a multimodal emotional database designed to promote natural emotional behaviors while maintaining control over lexical and emotional content in the recordings [20]. The corpus relied on a novel elicitation scheme, where two actors improvise scenarios that lead one of them to utter target sentences. For each target sentence, emotion-dependent scenarios are created so that the renditions convey the target emotions expressed as dictated by the context. The corpus also includes the rest of the speaking turns during the improvisation and the natural interaction between actors during the breaks. The corpus consists of 8,438 turns of emotional sentences recorded from 12 actors (over 9 hours). The turns are manually segmented into speaking turns and emotionally annotated into categorical emotions (anger, happiness, sadness, neutrality and other). The annotation was conducted with crowdsourcing, where we estimate in real time the performance of the raters, stopping the evaluation when their performance dropped below an acceptable level [21]. Each speaking turn was evaluated by at least five raters, where we assign consensus using majority vote rule.

The study uses the feature set proposed for the Interspeech 2013 Computational Paralinguistics Challenge [22], extracted with OpenSMILE [23]. This feature set includes statistics of prosodic, spectral and voice quality features, resulting in 6,373 features. Some of the features are highly correlated, therefore, we reduce the set with *correlation feature selection* (CFS). CFS selects features that correlate with the class label, but that are not correlated with previously selected features. The set of features consider for the evaluation consists of the first 3000 features selected by CFS using the USC-IEMOCAP corpus (source domain).

4. PROPOSED ALGORITHM

Discriminative features vary from one domain to another. Ideally we want to select discriminative features that are insensitive to changes between domains. However, this is rarely possible. An alternative is to use *active learning* (AL) to annotate a limited subset from the target domain. After annotating new data in the target domain, it is important to implement algorithms that efficiently use this set. The straightforward approach to leverage limited data from the target domain is to include them in the training set by either retraining or adapting the models. However, the features used for the ensemble are not modified, relying only on information from the source domain. We hypothesize that the information from the target domain may be effective to identify better discriminative features for the new domain. We propose a feature selection algorithm that optimizes the ensemble's performance over the selected data with AL, while making sure the ensemble is diverse. We increase the diversity by introducing different class specific bias across the classifiers and by reducing the overlap between feature sets across the classifiers in the ensemble.

4.1. Dataset Selection for Active Learning

Active learning identifies samples in the target domain, which are then labeled. This section defines various criteria used to select the samples (denoted D^t). This dataset will be used in our proposed feature selection scheme to train the ensemble.

Vote Entropy (VE) selects samples in the target domain with the highest levels of disagreement among different classifiers. Assuming that we have k classifiers, vote entropy is defined as [24]:

$$D(x) = -\sum_{c} \frac{V(c,x)}{k} log \frac{V(c,x)}{k} \quad (1)$$

where c ranges over all possible labels and $V(c, x)$ is the number of ensemble members assigning a class c to the input sample x. The aim of vote entropy is to resolve confusion between different classifiers. We implement this method with an ensemble with 40 linear *Support Vector Machine* (SVM) trained on the source domain. To increase diversity, each classifier in the ensemble is trained with 40 unique features that are not used by any other classifier. We use an adapted version of *forward feature selection* (FFS), where one feature is sequentially added to each classifier in order. After the feature is added, it is removed from the set of feature available to the rest of the classifiers. The ensemble is used to evaluate the sentences from the target domain, selecting samples according to Equation 1.

Uncertainty sampling (US) identifies samples that a classifier is less confident. We implement this framework with a single linear SVM trained on the source domain. We reduce the feature set from 3000 to 300 using FFS. The SVM is then evaluated on the sentences from the target domain, selecting sentences that are closest to the hyperplane.

Random sampling (RS) randomly selects samples from the target domain until reaching the required number of samples.

4.2. Proposed Feature Selection for Ensemble

The goal of the proposed feature selection algorithm is to minimize the mismatch between train and test conditions, while preserving the diversity of the ensemble. Algorithm 1 presents the pseudo code for the feature selection, which has three main aspects.

First, the classifiers are biased toward specific emotions. This approach selects features that are discriminative of specific emotional classes, covering different parts of the feature space. This leads the ensemble to be more diverse. If we have L classes, it partitions the k classifiers in the ensemble into L groups. The m^{th} classifier is biased towards the $m \bmod L$ class. We bias each group toward a specific class by maximizing the F_2-score for that class. The F_β-score is defined as:

$$F_\beta = (1+\beta^2)\frac{precision * recall}{\beta^2 * precision + recall} \quad (2)$$

Second, it eliminates the overlap in the feature sets of the ensemble by modifying the FFS algorithm. Each classifier starts with an empty feature set, adding at each stage the feature that maximizes its performance. The difference here is that the feature selection of all classifiers are done together in order. We sequentially add feature to

Algorithm 1 Proposed Feature Selection

```
Input:
    $k$ : Number of classifiers in the ensemble
    $N_{fs}$ : Number of features per classifier
    $L$ : Number of classes
    $D^s$ : Source training data
    $D^t$ : Target labeled data
    $F_{av}$ : Set containing the features in the available feature space
Output: $F_{sel}$: Set containing the features selected for the ensemble
define: $H_j$ is classifier j, $f^j_{sel}$ features selected for $H_j$
define: $j = 1, i = 1, l = 1$ and $F_{sel} = \phi$
for $i <= N_{fs}$ do
    for $j <= k$ do
        for $l <= |F_{av}|$ do
            $f_e = F_{av}(l)$
            Train $H_j$ Using $D^s$ and $f^j_{sel} \cup f_e$
            Evaluate $H_j$ on $D^t$ and $f^j_{sel} \cup f_e$
        $m = j \bmod L$
        $f_s$:= feature that maximizes $H_j$'s $F_2$-score over class $m$
        $f^j_{sel} = f^j_{sel} \cup f_s$
        $F_{av} = F_{av} - f_s$
    $F_{sel} = f^1_{sel}, \ldots, f^{N_{cl}}_{sel}$
```

each classifier, moving to the second feature once all the classifiers have one feature. After adding a feature to a classifier, this feature is not longer available for the remaining classifiers. This process repeats until reaching the desired number of features. This approach maximizes performance of each classifier, avoiding overlap in the feature set, and increasing the diversity across classifiers.

Third, we conduct the feature selection by maximizing the performance over the new labeled data from the target domain. The classifiers are trained with the source domain, but they are evaluated in the target domain using D^t. This is a novel way to leverage D^t, minimizing the mismatch between source and target domains.

5. EXPERIMENTAL SETTINGS

We evaluate the approach using an ensemble with 40 SVM classifiers for a four class problem (happiness, anger, sadness and neutrality). We use the *LibLinear* toolkit [25] to train all the SVM classifiers. For simplicity, we use linear kernel with the cost factor C set to 1. To ensure balanced classes in both train and test sets, we used random sub-sampling. After sub-sampling, the number of instances for training and testing are 4336 and 3152, respectively. The evaluation is repeated 10 times to ensure consistent results across the random sub-sampling iterations. For each iteration, we (1) select samples from the target domain for annotation using all the data selection techniques presented in Section 4.1, (2) run the proposed feature selection algorithm on the selected samples, (3) train the ensemble on source domain data with the selected features, and (4) evaluate the performance on the target domain, excluding the selected samples (i.e., D^t). For each sample in the test domain, every classifier outputs the probability of that sample belonging to each of the four classes. During the fusion of the ensemble predictions, we calculate the probabilities of the classes by equally weighting the classifiers in each class-dependent group, dividing by the sum across all the groups. The reported results are the average of the performance across all sub-sampling iterations.

The baseline for the evaluation is an ensemble containing 40 linear SVM classifiers, where each of them is trained on the source domain with 40 features. The key differences are (1) we do not bias the classifiers to emotional classes, and (2) we do not optimize performance over the target domain (we do not use AL for the baseline). However, we do limit the overlap between the feature sets across the classifiers to increase the diversity of the ensemble.

We study the effect of the data used for feature selection. First, we consider the size of the set, where we consider three sample sizes 200, 400, and 600. Notice that we report the performance on the remaining data in the target domain, so the testing set across evaluations are not necessary the same. Second, we consider the selection criteria used in sampling the data: vote entropy, uncertainty sampling, and random sampling. We evaluate statistical significance between classification results using one tail matched pair population mean t-test, asserting significance at p-value$= 0.05$.

6. RESULTS

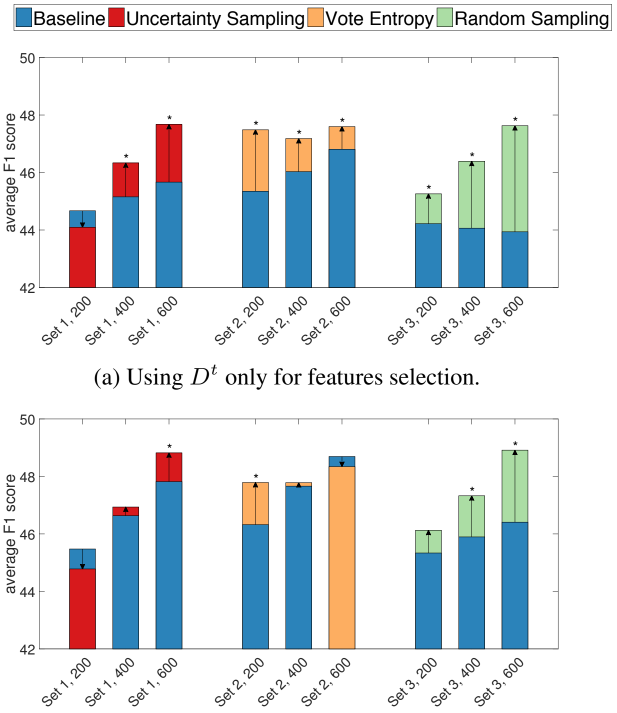
6.1. Proposed Feature Selection Method versus Baseline

Figure 1(a) shows average F1 score of ensembles trained using the proposed feature selection algorithm for different selection criteria, and different size for D^t. The first three bars give the results for uncertainty sampling (set 1) when the size for D^t is 200, 400 and 600, respectively. The next three bars give the results for vote entropy (set 2), and the last three bars give the results for random sampling (set 3). The arrows indicate whether the performance increases or decreases with respect to the baseline performance (blue bars). For example, we observe the largest improvement (about 3.7% absolute) in Figure 1(a) for random sampling with 600 samples. We represent statistical significant improvements of the proposed feature selection algorithm over the baseline with an asterisk (*) above the bar. Since samples in D^t are removed from the testing set, we cannot compare the results across sample selection criteria, since they are evaluated over different sets. We address pairwise comparison between methods in Section 6.2.

Figure 1(a) shows that the proposed feature selection approach gives statistical significant improvements over the baseline in all cases, except for uncertainty sampling when the selected samples from the target domain is just 200. As we increase the sample size, our feature selection approach with uncertainty sampling significantly outperforms the baseline improving the performance in 2% (absolute) with 600 samples. Both uncertainty sampling and vote entropy criteria select samples that are harder to classify. Since these samples are removed from the test set, the evaluation become easier. As a result, the performance for the baseline performance increases. This is why the performance of the baseline changes depending on the sample selection criteria, especially when the number of selected samples increase. For random sampling, removing these samples does not make the problem easier so the baseline F1-score remains somewhat constant for different test sets.

For vote entropy, Figure 1(a) significantly outperforms the baseline, regardless of the sample size. The performance gap between the proposed feature selection approach and the baseline decreases as the selected sample size increases. While the baseline performance increases from 45.4% (200 samples) to 46.8% (600 samples), the performance of the proposed method remains relatively constant around 47.4%. For random sampling, Figure 1(a) shows that the performance gap increases from 1.0% for 200 samples, to 3.7% for 600 samples.

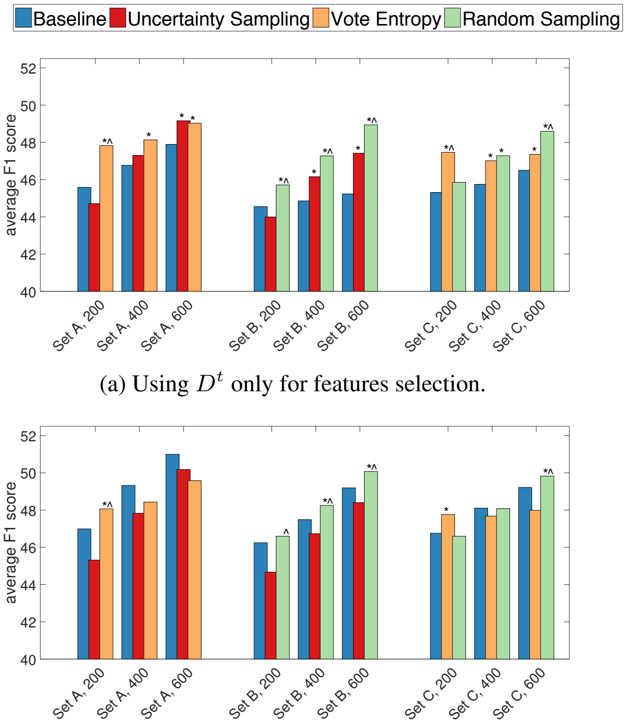
Similar to Figure 1(a), Figure 1(b) gives the performance when the selected samples for AL are not only used for the proposed feature selection method, but also included in the training set. This is a variation of the proposed framework. Since the baseline classifiers are trained with the selected samples from the target domain, their performance increases from the one shown in Figure 1(a). For this case, the improvement in performance is statistically significant over the baseline in four out of nine cases. Interesting, the trends are very similar to the ones observed in Figure 1(a). With uncertainty sampling, our ensemble performs worse than the baseline for small sample size (not statistically significant), but outperforms it for larger



(a) Using D^t only for features selection.

(b) Using features selection and adding D^t during training.

Fig. 1: The performance of the ensemble trained with the proposed feature selection scheme using different data selection criteria. We evaluate different size for D^t.



(a) Using D^t only for features selection.

(b) Using features selection and adding D^t during training.

Fig. 2: Direct comparison between data selection criteria for different sizes for D^t. Two criteria are directly compared with the baseline ensemble.

sample size (e.g., 600 samples). With vote entropy, our ensemble only provides statistically significant performance over the baseline when fewer samples are selected (200 samples). For random sampling, the gap in performance increases as we select more samples.

6.2. Pairwise Comparison Between Data Selection Criteria

As mentioned, we cannot directly compare the results in Figure 1 across sample selection criteria, since the testing sets are different. We create pairwise comparisons by defining sets where we use the domain data, excluding the samples selected by two criteria:

- Set A: Target domain minus {uncertainty sampling, vote entropy}
- Set B: Target domain minus {uncertainty & random sampling}
- Set C: Target domain minus {vote entropy, random sampling}

Figure 2(a) shows the direct comparison between sample selection criteria when we used the set D^t only for feature selection. The figure gives the results for different number of selected samples. Each set of results has three bars, one for the baseline (blue bar), and two for the proposed approach implemented with the corresponding criteria. We denote with an asterisk (*) when the proposed approach is statistically better than the baseline, and with a hat ($\wedge$) if one ensemble with one criterion is statistically better than the ensemble with the other criterion.

For Set A, Figure 2(a) shows that vote entropy (47.8%) outperforms the baseline (45.6%) and uncertainty sampling 44.7% when we only select 200 samples. As the sample size increases, the performance of the ensemble with uncertainty sampling (49.2% - 600 samples) provides similar performance than the one for vote entropy (49.0% - 600 samples). Both of them have F1-scores significantly better than the baseline (47.9%). For Set B, Figure 2(a) shows that random sampling consistently outperforms both uncertainty sampling and the baseline across all settings. Ensembles with uncertainty sampling outperforms the baseline when the number of samples is either 400 or 600. When we compare vote entropy and random sampling (i.e., Set C), Figure 2(a) shows that ensemble with vote entropy (47.5%) achieves the best performance with 200 samples, but ensemble with random sampling has the best performance with 600 samples. These results agree with the observations from Figure 1.

Finally, Figure 2(b) gives the results when we also include D^t in the training set. As expected, the baseline results increase when we use data from the target domain. The best sample selection criteria are vote entropy, when the number of samples is small and random sampling when the sample size increases. These results demonstrate the potential of the proposed feature selection approach, which provides statistical significant improvement over the baseline even when D^t is included in the training set.

7. CONCLUSIONS

This paper proposed a feature selection algorithm using active learning for an ensemble, which minimizes the mismatches between train and test conditions, and creates diverse classifiers in the ensemble. The results demonstrated that we can achieve a significant improvement by performing feature selection on a small set from the target domain. We also demonstrated the importance of using the appropriate criterion in selecting samples for annotation, where vote entropy is preferable if the selected sample size is small. When the sample size increases random sampling becomes the best option because it better represents the distribution of the target domain.

We implement the algorithm with an ensemble with SVMs. It is interesting to explore whether the benefits observed in the experimental evaluation generalize for other classifiers such as random forest. We are also planning to evaluate other feature selection criteria for AL that consider the data distribution along with the uncertainty (e.g., expected error variance, density weighted strategy).

8. 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] M. Woźniak, Manuel Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, vol. 16, pp. 3–17, March 2014.

[3] R. Polikar, "Ensemble learning," in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds., pp. 1–34. Springer, Boston, MA, USA, February 2012.

[4] G.L. Marcialis and F. Roli, "Fusion of face recognition algorithms for video-based surveillance systems," in *Multisensor Surveillance Systems: The Fusion Perspective*, G.L. Foresti, C.S. Regazzoni, and P.K. Varshney, Eds., pp. 235–249. Springer-Verlag New York, Boston, MA, USA, July 2003.

[5] S. Hashem, "Optimal linear combinations of neural networks," *Neural Networks*, vol. 10, no. 4, pp. 599–614, June 1997.

[6] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *9th European Conference on Speech Communication and Technology (Interspeech 2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 805–808.

[7] S. Scherer, F. Schwenker, and G. Palm, "Emotion recognition from speech using multi-classifier systems and RBF-ensembles," in *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*, B. Prasad and S.R.M. Prasanna, Eds., vol. 83 of *Studies in Computational Intelligence*, pp. 49–70. Springer, November 2010.

[8] D. Morrison, R. Wang, and L.C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, February 2007.

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

[10] A. Kapoor, R. W. Picard, and Y. Ivanov, "Probabilistic combination of multiple modalities to detect interest," in *International Conference on Pattern Recognition (ICPR 2004)*, Cambridge, UK, August 2004, vol. 3, pp. 969–972.

[11] F. Schwenker, S. Scherer, M. Schmidt, M. Schels, and M. Glodek, "Multiple classifier systems for the recogonition of human emotions," in *Multiple Classifier Systems: 9th International Workshop (MCS 2010).*, N El Gayar, J. Kittler, and F. Roli, Eds., vol. 5997 of *Lecture Notes in Computer Science*, pp. 315–324. Springer Berlin Heidelberg, Cairo, Egypt, April 2010.

[12] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction (ACII 2011)*, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., vol. 6975/2011 of *Lecture Notes in Computer Science*, pp. 359–368. Springer Berlin / Heidelberg, Memphis, TN, USA, October 2011.

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

[14] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 3285–3288.

[15] L. Duan, I.W. Tsang, D. Xu, and T.S. Chua, "Domain adaptation from multiple sources via auxiliary classifiers," in *International Conference on Machine Learning (ICML 2009)*, Montreal, QC, Canada, June 2009, pp. 289–296.

[16] W. Dai, Q. Yang, G.R. Xue, and Y. Yu, "Boosting for transfer learning," in *International conference on Machine learning (ICML 2007)*, Corvallis, OR, USA, June 2007, pp. 193–200.

[17] R. Xia, C. Zong, X. Hu, and E. Cambria, "Feature ensemble plus sample selection: Domain adaptation for sentiment classification," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 10–18, May-June 2013.

[18] J. Gao, W. Fan, J. Jiang, and J. Han, "Knowledge transfer via multiple model local structure mapping," in *ACM SIGKDD international conference on Knowledge discovery and data mining*, Las Vegas, NV, USA, August 2008, pp. 283–291.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] I. Dagan and S.P. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Proceedings of the Twelfth International Conference on Machine Learning*, A. Prieditis and S. Russell, Eds., pp. 150–157. Morgan Kaufmann, Tahoe City, CA, USA, July 1995.

[25] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, "LIBLINEAR: a library for large linear classification," *Journal of machine learning research*, vol. 9, pp. 1871–1874, August 2008.