SUPERVISED DOMAIN ADAPTATION FOR EMOTION RECOGNITION FROM SPEECH

Mohammed Abdelwahab and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
Email: mxa129730@utdallas.edu, busso@utdallas.edu

ABSTRACT

One of the main barriers in the deployment of speech emotion recognition systems in real applications is the lack of generalization of the emotion classifiers. The recognition performance achieved in controlled recordings drops when the models are tested with different speakers, channels, environments and domain conditions. This paper explores supervised model adaptation, which can improve the performance of systems evaluated with mismatched training and testing conditions. We address the following key questions in the context of supervised adaptation for speech emotion recognition: (a) how much labeled data is needed for adaptation to achieve good performance? (b) how important is speaker diversity in the labeled set? (c) can spontaneous acted data provide similar performance than naturalistic non-acted recordings? and (d) what is the best approach to adapt the models (domain adaptation versus incremental/online training)? We address these problems by using a multi-corpus framework where the models are trained and tested with different databases. The results indicate that even small portion of data used for adaptation can significantly improve the performance. Increasing the speaker diversity in the labeled data used for adaptation does not provide significant gain in performance. Also, we observe similar performance when the classifiers are trained with naturalistic non-acted data and spontaneous acted data.

Index Terms: emotion recognition, supervised domain adaptation

1. INTRODUCTION

Emotion recognition from speech is an emerging research area due to the clear benefits of emotionally aware interfaces in several applications, including *intelligent tutoring system* (ITS), entertainment, call center, and instrumental tools for the diagnostic and prognostic of mental health conditions. The performance of speech emotion recognition systems strongly depends on the differences between training and testing settings. Small differences in background noise, microphone settings, dialects, different languages and speaker variations reduce the classification performance. Robustness and generalization of emotion classifiers are key open challenges in the area of affective computing [1].

Previous studies have proposed various approaches to increase the robustness of speech emotion classifiers: collecting naturalistic databases [2, 3], speaker normalization [4, 5, 6, 7, 8], robust feature selection [9]. In addition to these research directions, an appealing solution is model adaptation where the classifiers are modified to reduce the gap between training and testing settings. To classify new data with good accuracy, a classifier needs to be trained with enough labeled data, resembling the distribution of the target data. Instead of collecting and annotating extra data when the target domain changes, many machine learning studies have proposed transfer learning schemes that use limited labeled data (supervised) or unlabeled data (unsupervised) from the new task, to improve the performance of the classifiers. These frameworks offer interesting solutions for speech emotion recognition. This paper explores the benefits of using classifiers pre-trained with available labeled data that are adapted using supervised methods. We consider three emotional databases, where two of them (IEMOCAP [10] and SEMAINE [2]) are used for training, and one for testing (RECOLA [11]).

While previous studies on speech emotion recognition have started to consider model adaptation [12, 13, 14, 15], this paper addresses key questions for supervised adaptation that remain open. (a) How much labeled data is needed for adaptation to achieve good performance? We address this question by comparing the classification performance in terms of the amount of data used for adaptation. We observe over 10% increase in performance even with a small data set used for adaptation. (b) How important is speaker diversity in the labeled set? We address this question by comparing two conditions: using data from a small number of subjects, and using data from multiple subjects. When we consider the same amount of data for adaptation, we do not observe improvement in classification performance by increasing the speaker variability. (c) Can spontaneous acted data provide similar performance than naturalistic non-acted recordings? We address this question by creating two classifiers trained with acted (IEMOCAP [10]) and natural (SEMAINE [2]) spontaneous interactions. Both of these classifiers are adapted with the same set. We observe very few cases where models originally trained with natural recordings provide better performance, suggesting that spontaneous acted recordings are valuable resources. (d) What is the best approach to adapt the models? We address this question by comparing domain adaptation for *support vector machine* (SVM), and incremental/online adaptation of SVM. Both approaches provide similar classification performance.

2. BACKGROUND AND RESOURCES

2.1. Relation to Prior Work

An important research direction in speech emotion recognition is improving the generalization of the classifiers arising from having different training and testing conditions. These differences include mismatches in noise level, microphone settings, language, and speaker variations. Schuller et al. [16] studied the detrimental effect of background noise and reverberated speech on emotion recognition performance. They showed that optimizing the selected feature set is an effective way to adapt to different noise conditions. Shami and Verhelst [17] conducted a multi-corpus study on emotion speech recognition. They showed that cross-corpus testing resulted in drop in performance. They proposed to merge the databases to train a classifier, achieving performance comparable to within-corpus results.

This work was funded by NSF (IIS-1217104, IIS-1329659) and Samsung Research America.

One approach to reduce variability between training and testing domains is feature normalization. Schuller et al. [4] used a simple corpus-dependent normalization scheme to reduce domain variability. They presented cross-corpora experiments, where normalizing each corpus separately was effective. Speaker normalization is a common approach to improve performance by attenuating speaking variations from the feature set [1, 5]. Busso et al. [6, 7] proposed the *iterative feature normalization* (IFN) algorithm that estimates the normalization parameters by detecting neutral speech. The normalization parameters are estimated over this neutral set, such that the first and second order statistics of the acoustic features for neutral speech are similar across speakers. The parameters are then applied to the entire set. This approach was effectively used by Rahman and Busso [8] to personalize an emotion classifier to a target user.

Model adaptation is another appealing solution to generalize an emotion recognition system. Supervised learning relies on limited labeled data from the new domain. Unsupervised learning does not require extra labeled data. Zhang et al. [15] studied the effect of using unlabeled data in a multi-corpus experiment. Working with six databases, they merged three databases to build a baseline classifier. They used the models to predict the emotional content of two different corpora. Then, they re-trained the classifiers with the five corpora, where the predictions were used as actual labels. The models were evaluated on a different corpus, where this unsupervised learning scheme achieved better performance than using the baseline classifier only trained with the three original corpora. Using neural networks, Deng et al. [12] presented a sparse autoencoder for feature transfer learning that exploited the underlying structure in emotional speech learnt from the labeled target data to reconstruct the source data accordingly. Maeireizo et al. [13] used co-training to automatically label spoken dialog data for an emotion detection problem, showing that co-training is highly effective when combined with a good set of features. Liu et al. [14] proposed an enhanced co-training algorithm, where the classifiers represented two conditionally independent attribute views. The approach showed promising results for speech emotion recognition. This study focuses on transfer learning under supervised domain adaptation, where labeled data is available.

2.2. Adaptation Schemes

There are several methods for supervised adaptation that can be used based on the framework used for classification. For example, for *Gaussian mixture models* (GMMs) the adaptation schemes based on *maximum a posteriori* (MAP) adaptation [18] and *Maximum Likelihood Linear Regression* (MLLR) [19] are effective and widely used in speaker, language, and speech recognition. Recent studies have proposed various solutions for model adaptation for *support vector machine* (SVM). While SVM has been widely used in emotion recognition, we are not aware of any of those adaptation methods employed in the field of speech emotion recognition. This study explores two alternative approaches to adapt a SVM using limited labeled data: an adaptive SVM algorithm [20] and incremental SVM training method [21].

Adaptive SVM: This work uses the adaptive SVM algorithm proposed by Yang et al. [20] in an attempt to transform existing SVM classifiers into a new effective SVM classifier that would work on a new dataset with limited number of labeled data. The approach aims to minimize both the classification error over the training examples, and the discrepancy between the originals and adapted classifier. The new optimization problem seeks a decision boundary close to that of the classifier trained from the source domain, while managing to separate the new labeled data from the target domain.

Incremental SVM: Incremental SVM classifiers were introduced to reduce batch SVM memory and computational requirements, especially for very large data sets. One of the useful features of incremental learning is the ability to add more training data. These approaches employ incremental learning techniques, where only a subset of the data is considered at each step of the learning process, discarding old data while maintaining the support vectors learned in previous steps [22, 23]. Shalev et al. [24] proposed and analyzed a simple and effective stochastic sub-gradient descent algorithm for solving the optimization problem imposed by SVMs. Each iteration of the algorithm operates on a single training example selected at random. By selecting the training examples at random, the authors demonstrated that the solution converges in probability regardless of the data used in the classification problem.

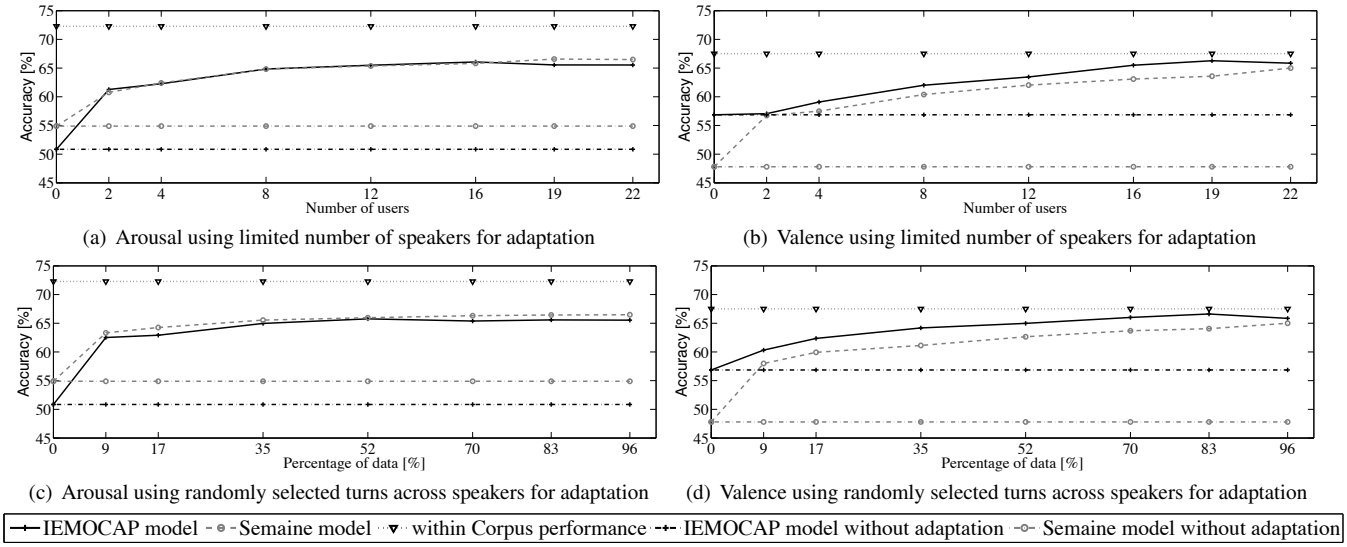
2.3. Databases

The study relies on a multi-corpus framework. We assume that two English emotional databases are available to train emotional classifiers: the IEMOCAP [10] and SEMAINE [2] database. For testing, we use the RECOLA database [11], which was recorded in French. We briefly describe these corpora.

IEMOCAP: The USC *IEMOCAP* database contains approximately 12 hours of audio-visual data recorded from five male and five female actors [10]. It contains detailed motion captured information of the recordings that are carefully synchronized with the audio (this study uses only the audio). The goal of the data collection was to elicit natural emotions within a controlled setting. This goal was achieved with two elicitation framework: 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 of approximately five minutes were recorded, which were 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 relies on the dimensional scores which takes values between one and five. We consider 6829 turns recorded by the 10 actors.

SEMAINE: The SEMAINE database is an audiovisual database with natural emotional displays [2]. The corpus includes sessions recorded from two individuals, an *operator* and a *user*, interacting through teleprompter screens from two different rooms. The emotions were elicited with the *sensitive artificial listener* (SAL) framework, where the operator assumes four personalities aiming to elicit positive and negative emotional reactions from the user. The sessions were emotionally annotated by 6-8 raters. Instead of assigning global labels to the speaking turns, the evaluators provided time-continuous emotional traces using the FEELTRACE toolkit [25]. As the evaluators watch the recordings, they move the mouse cursor over a *graphical user interface* (GUI), where the axes represent specific emotional attributes. The interface records the position of the cursor, providing a continuous profile, or trace, for that emotional dimension. Among other descriptors, the perceptual evaluation considered the dimensions activation (calm versus active), valence (negative versus positive), control (weak versus strong) and expectation (predictable versus unexpected) [26]. This study focuses on activation and valence, which are the most commonly used emotional dimensions. This study uses 2315 turns from 10 speakers (users) interacting with the operators.

RECOLA: The RECOLA corpus consists of dyadic interactions where the participants engaged in video conference while completing a task requiring collaboration [11]. Each participant received a questionnaire to evaluate his or her current emotional state by using Self-Assessment Manikin (SAM) [27]. The participants of a team



(a) Arousal using limited number of speakers for adaptation

(b) Valence using limited number of speakers for adaptation

(c) Arousal using randomly selected turns across speakers for adaptation

(d) Valence using randomly selected turns across speakers for adaptation

Fig. 1. Classification performance using supervised adaptation of SVM proposed by Yang et al. [20].

were then introduced to each other. They were separated into different rooms, where they were told that the experiments was to understand communication between people by using computer-supported tools. The data comprises different multimodal data (i.e., audio, video, ECG and EDA) that were continuously and synchronously recorded. 46 participants took part in the data collection. All the subjects are native French speaking. According to the self-reports filled by the subjects, only 20% of the participants knew well their teammate. Only the first five minutes of each interaction was kept to ease the emotion annotation process. Six annotators continuously evaluated the emotional content using two dimensions: arousal and valence. Other social behavior labels were also collected. This study uses annotated emotional data from 23 speakers interacting with their teammate. In total, we consider 899 turns.

The SEMAINE and RECOLA databases have time-continuous emotional traces. To derive a single score for each speaking turn, we assign the average value across the duration of the turn and across evaluators. We repeat this approach for arousal and valence.

3. EXPERIMENTS

We describe the proposed framework to explore the optimum size, and variety of the labeled data used for adaptation, the nature of the training corpus, and best approach to incorporate new labeled data.

3.1. Classification Problem and Experimental Settings

The evaluation consists in detecting low and high level of arousal and valence. For these attributes, we separately normalized the values, per corpora, using z-normalization. We create the negative class (low arousal or valence) with turns where the normalized values are less than $z_l = -0.3$. We define the positive class (high arousal or valence) with turns when the normalized values are higher than $z_h = 0.3$. The samples lying between these thresholds are discarded to create good separation between classes. With these thresholds, the number of turns considered for activation/valence are 5,266/5,073 for IEMOCAP, 1,734/1,680 for SEMAINE, and 700/677 for RECOLA.

Size of the labeled data: The RECOLA database is used to evaluate the classification performance. We split this corpus into two speaker-independent partitions, one for adaptation and one for testing (all the data for one subject is either in the adaptation or testing sets). The partition is implemented with a *leave-one-speaker-out* (LOSO) cross-validation approach, where, in each fold, only one out of 23 subjects is used for testing. The results are reported by averaging the classification performance across the 23 folds. We sequentially increase the size of the labeled data to understand the optimum size needed for adaptation (see the x-labels in Figures 1 and 2). Since data from 22 subjects is potentially available for adaptation, we evaluate 20 different partitions per experiment (i.e., if we adapt the models with data from two subjects, we randomly select 20 different pairs within the 22 remaining subjects, without repetitions).

Speaker variety of the labeled data: For a given size of the adaptation set, we evaluate two conditions to assess the benefits of increasing the speaker diversity in the adaptation set (1) we create the adaptation set from a limited number of speakers, (2) we create the adaptation set at random from all the subjects in the RECOLA database that are not in the testing set. For example, consider Figures 1(a) and 1(c). Using eight subjects (Fig. 1(a)) can be directly compared with using 35% of the corpus for adaptation (Fig. 1(c)).

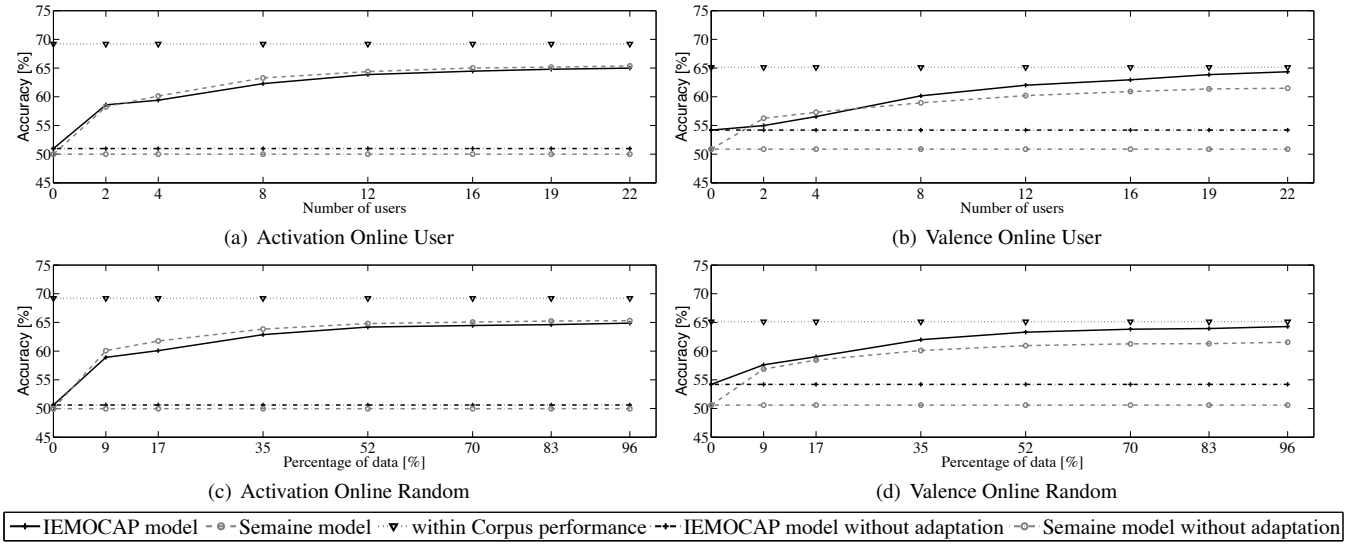
Acted versus natural database: Across conditions, we train two different classifiers using either the IEMOCAP (i.e., acted) or the SEMAINE (i.e., natural) corpus. These classifiers are then adapted using the RECOLA database.

Adaptation scheme: We evaluate the adaptive SVM (Fig. 1) and online SVM (Figs. 2) frameworks. For the adaptive SVM framework, we train a linear kernel SVM classifier implemented using LIBSVM [28]. This classifier is later modified with the adaptive SVM framework proposed by Yang et al. [20]. The online SVM classifier was implemented using the VLFeat library [21], which provides the stochastic sub-gradient descent algorithm described in Section 2.2.

We provide two baseline classifiers. First, we estimate the within-corpus performance on the RECOLA database. For consistency, we use LOSO cross-validation approach (i.e., 22 training, 1 testing), where the results are the average values across the 23 folds. The second baseline corresponds to training with either the IEMOCAP or SEMAINE corpus and testing with the RECOLA database without adaptation (see straight lines in Figs. 1 and 2).

3.2. Acoustic Features

The evaluation considers the feature set proposed for the Speaker State Challenge in Interspeech 2011 [29]. The set includes 4,368 features derived from spectral, prosodic and voice quality features. For each database, we normalize the features using z-normalization.



(a) Activation Online User (b) Valence Online User

(c) Activation Online Random (d) Valence Online Random

Fig. 2. Classification performance using incremental/online training implemented with the VLFeat library [21]

Given the dimension of the feature vector, we reduce the set using a two-layer feature selection approach. In the first layer, we use *correlation attribute evaluation* using a ranker search method to limit the set to 500 features. The approach selects the features with the highest correlation with the class label. This approach efficiently selects candidate features. The second layer reduces further this set to 50 features by using *correlation feature selection* (CFS). This feature selection approach is implemented with a greedy stepwise search method, where new features are added only if they are not highly correlated with previously selected features. This approach is efficient and general, since it does not depend on the performance of any classifier. We separately implement this approach for the SEMAINE and IEMOCAP databases for arousal and valence. The only setting where the RECOLA turns are used for feature selection is for the within-corpus classifier (upper bound performance).

4. RESULTS

Figures 1 and 2 give the classification results across settings. This section discusses our research questions.

4.1. How much labeled data is needed for adaptation?

The baseline classifiers without adaptation are significantly lower than the within-corpus classification performance. For arousal, Figures 1(a) and 1(c), and Figures 2(a) and 2(c) show significant improvements even when we only use data from two subjects for adaptation ($\sim$9% of the data). The one-tailed population proportion test indicates that these differences are significant (p-value < 0.05). For valence, the improvement in performance is more gradual as we increase the adaptation set. When using only two subjects, we only observe significant improvements over the model without adaptation for the SEMAINE database. For the IEMOCAP corpus, we need to consider more data in the adaptation set.

When we compare the classification performances using an adaptation set with either eight subjects ($\sim$35% of the data) or 22 subjects, we do not observe significant differences for activation and valence (one-tailed population proportion test, asserting significance at p-value=0.05). Using 35% of the data is enough to adapt models.

4.2. How important is speaker diversity in the labeled set?

To address this question, we estimate the hypothesis test for matched pairs, where the matched condition is the size of the adaptation set. We compare the classification results shown in Figures 1(a) and 1(c); Figures 1(b) and 1(d); Figures 2(a) and 2(c); and Figure 2(b) and 2(d). In each of these cases, we do not observe any significant difference (asserting significance at p-value=0.05). Interestingly, speaker variety is not a dominant factor in selecting the adaptation set.

4.3. Can spontaneous acted data provide competitive results?

Without adaptation, the SEMAINE model provides better performance for arousal, while the IEMOCAP model provides better performance for valence. However, the classification performances are very similar when these models are adapted using the RECOLA database. The hypothesis test for matched pairs shows that there are no significant difference in performance for arousal and valence (the matched condition is the training data). With model adaptation, this result suggests that a classifier built with spontaneous acted data can perform as well as a classifier built with natural emotional databases.

4.4. What is the best approach to adapt the models?

Finally, we compare the performance of the adaptation schemes (Fig. 1 versus Fig. 2). We estimate the hypothesis test for matched pairs across conditions where the matched condition was the adaptation scheme. The test reveals that both methods provide similar performance (asserting significance at p-value=0.05).

5. CONCLUSIONS

This work explores important open questions about supervised adaptation in speech emotion recognition. We observe significant improvement in classification performance, even when only 9% of the data is used for adaptation. The study demonstrated that speaker diversity in the adaptation set is not a dominant factor. Furthermore, classifier models built with spontaneous acted data is a viable option when the models are adapted. Finally, using online training yields similar performance to a model adaptation approach.

The study demonstrates the importance of model adaptation. We are currently exploring the use of unsupervised model adaptation coupled with feature normalization. We expect that combining both approaches will lead to emotion recognition systems with better generalization against train and test mismatches.

6. 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] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[3] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[4] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, July-Dec 2010.

[5] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker normalisation for speech based emotion detection," in *15th International Conference on Digital Signal Processing (DSP 2007)*, Cardiff, Wales, UK, July 2007, pp. 611–614.

[6] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, October-December 2013.

[7] C. Busso, A. Metallinou, and S. Narayanan, "Iterative feature normalization for emotional speech detection," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 5692–5695.

[8] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

[9] C. Busso, S. Lee, and S.S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[10] 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.

[11] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013.

[12] 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.

[13] B. Maeireizo, D. Litman, and R. Hwa, "Co-training for predicting emotions with spoken dialogue data," in *Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, July 2004, p. 203206.

[14] J. Liu, C. Chen, J. Bu, M. You, and J. Tao, "Speech emotion recognition using an enhanced co-training algorithm," in *IEEE International Conference on Multimedia and Expo (ICME 2007)*, Beijing, China, July 2007, pp. 999–1002.

[15] 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.

[16] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, HI, USA, April 2007, vol. 4, pp. 941–944.

[17] 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.

[18] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[19] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, April 1995.

[20] 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.

[21] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *ACM international conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1469–1472.

[22] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *Advances in Neural Information Processing Systems 13*, T.K. Leen, T.G. Dietterich, and V. Tresp, Eds., Neural Information Processing, pp. 409–415. A Bradford Book, The MIT Press, Cambridge, MA, USA, April 2001.

[23] N.A. Syed, S. Huan, L. Kah, and K. Sung, "Incremental learning with support vector machines," in *Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence (IJCAI 1999)*, Stockholm, Sweden, August 1999.

[24] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for SVM," *Mathematical programming*, vol. 127, no. 1, pp. 3–30, March 2011.

[25] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, ISCA, pp. 19–24.

[26] J.R.J. Fontaine, K.R. Scherer, E.B. Roesch, and P.C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, December 2007.

[27] M.M. Bradley and P.J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, March 1994.

[28] 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.

[29] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, August 2011, pp. 3201–3204.