ENHANCED SEMI-SUPERVISED LEARNING FOR MULTIMODAL EMOTION RECOGNITION

Zixing Zhang[1], *Fabien Ringeval*[1], *Bin Dong*[1], *Eduardo Coutinho*[2], *Erik Marchi*[3], *Björn Schuller*[1,2]

[1]Chair of Complex & Intelligent Systems, University of Passau, Germany
[2]Department of Computing, Imperial College London, UK
[3]Machine Intelligence & Signal Processing group, MMK, Technische Universität München, Germany
zixing.zhang@uni-passau.de, bjoern.schuller@imperial.ac.uk

ABSTRACT

Semi-Supervised Learning (SSL) techniques have found many applications where labeled data is scarce and/or expensive to obtain. However, SSL suffers from various inherent limitations that limit its performance in practical applications. A central problem is that the low performance that a classifier can deliver on challenging recognition tasks reduces the trustability of the automatically labeled data. Another related issue is the noise accumulation problem – instances that are misclassified by the system are still used to train it in future iterations. In this paper, we propose to address both issues in the context of emotion recognition. Initially, we exploit the complementarity between audio-visual features to improve the performance of the classifier during the supervised phase. Then, we iteratively re-evaluate the automatically labeled instances to correct possibly mislabeled data and this enhances the overall confidence of the system's predictions. Experimental results performed on the RECOLA database demonstrate that our methodology delivers a strong performance in the classification of high/low emotional arousal (UAR = 76.5%), and significantly outperforms traditional SSL methods by at least 5.0% (absolute gain).

Index Terms— Multimodal emotion recognition, enhanced semi-supervised learning

1. INTRODUCTION

In the field of automatic emotion recognition, an increasing number of researchers and developers are trying to apply research achievements to real-life applications, such as, video games [1], service robots [2], or health care systems [3]. However, a major challenge for these applications is the limited amount of labeled data that are yet necessary to develop robust Machine Learning systems. Indeed, the great majority of emotional databases that are publicly available at present have only a few hours of annotated instances, and even less for specific applications, such as autism [1, 2] or depression in elderly [4], which is by far not comparable with the datasets available to train automatic speech recognition systems [5].

One simple way to deal with this issue of data scarcity is to agglomerate multiple databases and train an emotion recognition system on the agglomerated dataset [6]. Such procedure makes however the recognition task even more complex because the variability over the different corpora (e.g., microphone, room impulse response) is hard to compensate [6, 7]. Other techniques that have gained a strong momentum in the last few years focus instead on unlabelled data. The main reason is that, unlike labeled databases, unlabelled instances are broadly available. One of the most attractive techniques is based on Semi-Supervised Learning (SSL) [8, 9], as it aims to use these data without the intervention of human annotators.

Many studies have shown the benefits of SSL for emotion recognition [10, 11, 12, 13]. However, most of these studies have focused on a single modality – either on audio [11], video [10, 8], or physiology [9]. Nowadays, *multimodality* has been increasingly and widely implemented for emotion recognition [14, 15, 16]. The main reasons are not only the broad availability of cameras and microphones, but more importantly the combination of various modalities can boost the emotion recognition accuracy [15, 17, 18], since each modality can provide complementary information. For SSL, these information is simply ignored in the context of emotion recognition.

Another long-standing issue of SSL is the performance degradation as the learning process evolves over time [19, 20]. This is because the selected data are sometimes misclassified by the system and then accumulated in the training set. As a consequence, the model becomes less precise, and the noise accumulation leads to a negative vicious circle [21, 19]. We therefore propose in this paper a novel SSL approach that: (i) exploits the complementarity of audio-visual data to perform robust emotion recognition, and (ii) sequentially re-evaluates previously selected data to tackle the issue of noise accumulation.

The remainder of this paper is organized as follows. The proposed method is described in detail in Section 3. This method is then evaluated by an emotion recognition task in Section 4. Finally, conclusions and future work are given in Section 5.

2. PREVIOUS WORK

Two main SSL approaches have been proposed for emotion recognition in the literature: *Self-Training* (ST) [22, 12, 8] and *Co-Training* (CT) [13, 9, 12]. The work in [22] applied ST to multiple emotional corpora. This was further extended by combining Active Learning [12], which more efficiently reduced the human annotation work and improved the learning performance. Because the ST procedure requires little annotation work from human, it has been considered as a useful option to enhance the robustness of an existing emotion classifier [22, 12, 8]. Whereas in CT, the mutual agreement between two distinct 'views' (i.e., classifiers) of an unlabeled instance is used to consider its inclusion in the training set [23]. The work done by Zhang et al. [12] and Liu et al. [13] has shown the capability of CT for retrieving the emotion information on unlabeled data via separating the feature sets into two 'views' in the speech domain. The work in this paper is a further step in continuation of the authors' previous work on exploiting unlabeled data for emotion recognition by exploiting multiple modalities and a refined SSL algorithm.

Algorithm 1: Enhanced multimodal Self-Training (emmST).

Initialize: Number of additional selected data per learning iteration n, and predefined iteration times I

for $i = 1, ..., I$ **do**
- Tandem features $\mathbf{x} = [\mathbf{x}_a, \mathbf{x}_v]$; *% one view*
- (Optional) $\mathcal{L}_s^i \leftarrow upsample(\mathcal{L}^i)$;
- Train classifier $h^i := f(\mathcal{L}_s^i(\mathbf{x}, y)|\mathcal{L}^i(\mathbf{x}, y))$;
- Classification $(y'_{\mathbf{x}}, C(y'_{\mathbf{x}})) \leftarrow h^i(\forall \mathbf{x} \in \mathcal{U})$; *% re-evaluate the whole original unlabeled set*
- Set $n^i = i \times n$;
- Copy $\mathcal{S}^i$ from $\mathcal{U}$, $size(\mathcal{S}^i) = n^i$, and satisfy $\underset{\forall \mathbf{x} \in \mathcal{S}^i}{C(y'_{\mathbf{x}})} \geq \underset{\forall \mathbf{x}' \in (\mathcal{U} \setminus \mathcal{S}^i)}{C(y'_{\mathbf{x}'})}$;
- $\mathcal{L}^{i+1} = \mathcal{L}^0 \cup \mathcal{S}^i$;

end

3. ENHANCED MULTIMODAL SEMI-SUPERVISED LEARNING

Let us assume a small set of labeled audio-visual data $\mathcal{L}^0 = \{(\mathbf{x}_{ai}, \mathbf{x}_{vi}, y_i), i = 1, \ldots, N_L\}$, and a large set of unlabeled audio-visual data $\mathcal{U} = \{(\mathbf{x}_{ai}, \mathbf{x}_{vi}), i = 1, \ldots, N_U\}$, where $\mathbf{x}_a \in \mathcal{X}_a$ and $\mathbf{x}_v \in \mathcal{X}_v$ denote the feature vectors in the audio and visual domains, respectively; $y \in \mathcal{Y}$ is the domain category; and N_L and N_U indicate the number of labeled and unlabeled instances, respectively. It should be noted that, N_L is much smaller than N_U ($N_L \ll N_U$) due to the well-known limited availability of labeled data in the field of emotion recognition.

3.1. Self-Training and Co-Training

As mentioned in Section 2, ST and CT are two frequently used SSL approaches. For ST, a classifier is firstly trained with the original human-labeled data set $\mathcal{L}$. After that, the classifier is used to recognize the unlabeled data set $\mathcal{U}$. Typically, the unlabeled data $\mathcal{S}$ that are recognized with high confidence $C(\mathbf{x})$, together with their predicted labels, are added to the original training set ($\mathcal{L} \cup \mathcal{S}$), and removed from the unlabeled data set ($\mathcal{U} \setminus \mathcal{S}$). The classifier is then retrained with the updated training set and this process is repeated several times until a predefined stopping criterion is met.

To cease the learning process, several criteria can be implemented, for example, (i) no performance improvement is shown on the evaluation set, (ii) a predefined repeating times is matched or (iii) no target data remains in the unlabeled data set. Note that, in this paper, the second stopping criterion is chosen through all of the experiments to ease performance comparison.

Compared with ST, where the classifier uses its own prediction to teach itself, CT tries to exploit the mutual information between two models ('views' or feature domains) – $\mathcal{X}_1$ and $\mathcal{X}_2$, each of which uses its predictions to teach not only itself but also the other one. Specifically, each 'view' is used to create two 'good' classifiers h_1 and h_2, and each classifier is tested on the unlabeled data set $\mathcal{U}$. The unlabeled data ($\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$) predicted with high confidence values $C(\mathbf{x})$ are then added (together with the new label) to the training set ($\mathcal{L} \cup \mathcal{S}$) and removed from the unlabeled data set ($\mathcal{U} \setminus \mathcal{S}$). Afterwards, the two classifiers are retrained from the new training set based on the corresponding feature sets, and the process is repeated until the stopping criterion is met.

CT relies on two assumptions [23]: (a) sufficiency – Each 'view' is sufficient for classification on its own. That is, the two hypotheses $f_1 : \mathcal{X}_1 \mapsto \mathcal{Y}$ and $f_2 : \mathcal{X}_2 \mapsto \mathcal{Y}$ are good enough for recognition; (b) conditional independence – The 'views' are conditionally independent given the class label [23], that is, $p(y_i|\mathbf{x}) \leftarrow p(y_i|\mathbf{x}_1)p(y_i|\mathbf{x}_2)$.

Algorithm 2: Enhanced multimodal Co-Training (emmCT).

Initialize: Number of additional selected data per learning iteration n, and predefined iteration times I

for $i = 1, ..., I$ **do**
- **for** $\mathbf{x} = \mathbf{x}_a, \mathbf{x}_v$ **do** *% two views*
  - (Optional) $\mathcal{L}_s^i \leftarrow upsample(\mathcal{L}^i)$;
  - Train classifier $h^i := f(\mathcal{L}_s^i(\mathbf{x}, y)|\mathcal{L}^i(\mathbf{x}, y))$;
  - Classification $(y'_{\mathbf{x}}, C(y'_{\mathbf{x}})) \leftarrow h^i(\forall \mathbf{x} \in \mathcal{U})$; *% re-evaluate the whole original unlabeled set*
  - Set $n^i = i \times \lfloor n/2 \rfloor$;
  - Copy $\mathcal{S}$ from $\mathcal{U}$, $size(\mathcal{S}) = n^i$, and satisfy $\underset{\forall \mathbf{x} \in \mathcal{S}}{C(y'_{\mathbf{x}})} \geq \underset{\forall \mathbf{x}' \in (\mathcal{U} \setminus \mathcal{S})}{C(y'_{\mathbf{x}'})}$;
  - $\mathcal{S}^i = \bigcup \mathcal{S}$
- **end**
- $\mathcal{L}^{i+1} = \mathcal{L}^0 \cup \mathcal{S}^i$;

end

3.2. Multimodal Semi-Supervised Learning

In the case of CT, there are two 'views' employed to train different models. In the field of emotion recognition, however, the two 'views' normally belong to the same domain/model (e.g., speech) [12, 13]. To refine the unimodal SSL algorithms as discussed in Section 3.1, multiple modalities (e.g., audio and video) can be used together for both ST and CT.

To do this, audio and video feature sets are joined (*early fusion*) as one set for ST, i.e., $\mathbf{x} = [\mathbf{x}_a, \mathbf{x}_v]$. After that, the learning process proceeds as typical ST algorithms. In this paper, we will refer to this method as *multimodal Self-Training* (mmST). However, for CT, both audio and video feature sets can be served as different 'views', i.e., $\mathcal{X}_1 = \mathcal{X}_a$, and $\mathcal{X}_2 = \mathcal{X}_v$ compared with the work in [24]. This method is called *multimodal Co-Training* (mmCT) in the paper.

3.3. Enhanced Semi-Supervised Learning

As mentioned in Section 1, one main drawback of SSL is noise accumulation. For traditional SSL, the data selected by machine oracle are fully trusted and pooled into the training data set. However, some of these data are mislabeled actually. As the learning process continues, more and more mislabeled data (noise) might be accumulated in the training set, eventually leading to a vicious circle of erroneous learning [21, 19, 25].

To overcome this noise accumulation problem, we propose to not always trust the machine labeled data. We call this method *enhanced SSL* (eSSL). The core principle of this extension is to maintain the previously selected data in the original unlabeled data set at all learning iterations. By doing this, the previously selected data will be re-evaluated by the following enhanced model. Therefore, it is possible to correct mislabeled data in future iterations with an improved model. Naturally, the previously selected instances may not be selected again in the following learning process, i.e., $\mathcal{S}^i \not\subset \mathcal{S}^j$, $i < j$. The advantage of this method is that it guarantees that the machine oracle will perform better when selecting the unlabeled instances for automatic annotation. The pseudocode describing the

Table 1. Distribution of speakers and instances per partition of the RECOLA [27]. spks: speakers, POS: positive, NEG: negative.

	# spks	*# arousal*		
		POS	NEG	$\sum$
pool	23	623	344	967
eval.	11	366	149	515

algorithms for both enhanced multimodal Self-Training (emmST) and enhanced multimodal Co-Training (emmCT) are shown in Algorithm 1 and 2, respectively.

4. EMPIRICAL EXPERIMENTS AND RESULTS

In the following, we firstly describe the selected database and acoustic/visual feature sets. Then, we focus on evaluating the performance of the proposed multimodal SSL and its enhanced extension.

4.1. Selected Database

For the purpose of evaluating the different SSL approaches, we chose the RECOLA database [26, 27]. It includes spontaneous and natural affective behaviours collected from 46 French speaking participants while solving a task in dyads and remotely; 27 females, 19 males, mean age is 22 years and standard deviation is 3 years. The database includes 9.5 h of continuous and synchronous multimodal recordings, i.e., audio, video, electrocardiogram, and electro-dermal activity. Due to consent of the participants to share their data, the data is reduced to a subset of 34 participants with an overall duration of 7 hours. Rating of emotion was performed by 6 French-speaking assistants (3 male, 3 female) using the ANNEMO web-based annotation toolkit [26]. Emotional dimensions (arousal and valence) were rated time-continuously for the first 5 minutes of each recording by all raters.

For the purpose of this study, these continuous ratings are further discretized into a binary category – POSitive and NEGative. To do this, the audiovisual time series are firstly split into sequential short segments (instances) according to the voice activity and face detection [27], i.e., an instance is defined when both voice activity and face are detected simultaneously. Then we assigned POS or NEG to each of these instances if the average rating value is above or under zero; the average of the ratings is normalised to zero-mean for each recording. The audiovisual instances are finally divided into speaker independent pool (unlabeled data set) and evaluation sets. Details on the speakers and the distribution of instances used in this paper are shown in Table 1. Note that we only used the arousal dimension as some issues were found on the valence, which we suspect to be due to the normalisation procedure.

4.2. Feature Set

As acoustic features, we chose the same set of Low-Level Descriptors (LLDs) as in the past three INTERSPEECH Computational Paralinguistic ChallengEs (ComParE 2013-2015) [28]. It contains 4 energy related LLDs (loudness, RASTA spectrum, RMS energy and zero-crossing rate), 55 spectral related LLDs (e.g., spectrum bands, MFCC 1-14, spectral energy, spectral flux/centroid/entropy/slope, psychoacoustic sharpness, harmonicity, spectral variance/skewness/kurtosis), and 6 voicing related LLDs (pitch, probability of voicing, logHNR, jitter, shimmer). These 65 LLDs of speech with their first order derivate leads to 130 LLDs in total. Functionals (min, max, range, mean, variance) are then computed on the LLDs over an instance, which thus provides 650 acoustic features in total per segment.

As visual features, we extracted 20 LLDs and their first order derivate (40 LLDs in total) for each frame in the video recordings. The 20 LLDs contain 15 facial actions units, head-pose in three dimensions, and the mean and standard deviation of the optical flow in the region around the head – the computation of those features is described in details in [27]. Similar to the acoustic features, the same 5 functionals are applied per segment after extracting the frame-based LLDs, which provides 200 visual features per segment in total.

Table 2. Statistical performance comparison between the *multimodal* (audio + video) and the *unimodal* (audio or video) SSL, the enhanced (e) and the non-enhanced SSL, on the means of Self-Training (ST) and Co-Training (CT). *initial*, *last*, *max.*, and *mean* denote the initial, last, maximum and mean unweighted average recalls (UARs) over the 40 learning iterations

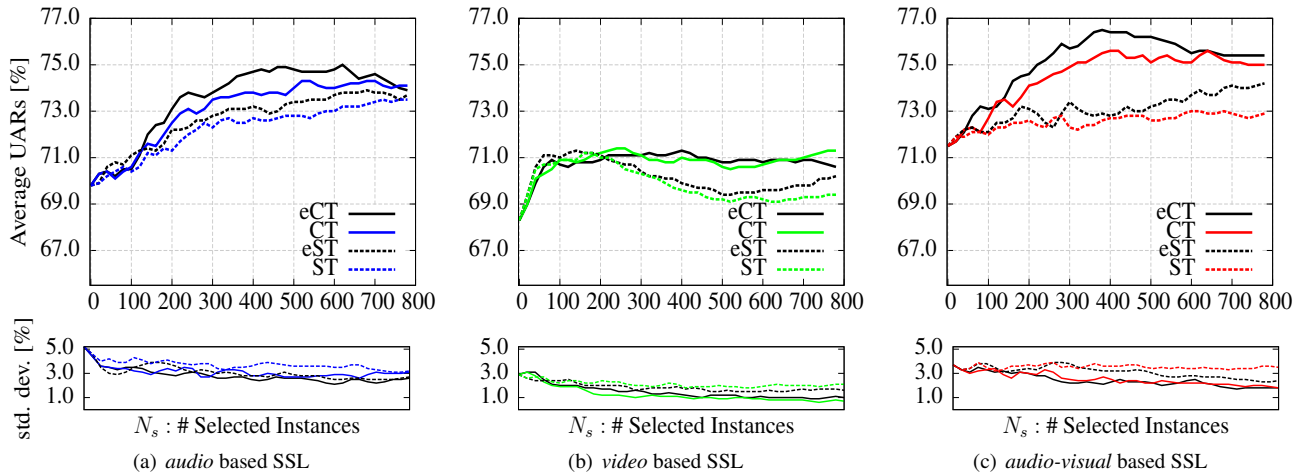
	average UARs			
[%]	*initial*	*last*	*max.*	*mean*
audio, ST	69.8	73.5	73.5	72.2
video, ST	68.3	69.4	71.2	69.9
audio+video, ST	71.5	72.9	73.0	**72.6**
audio, eST	69.8	73.7	73.9	72.6
video, eST	68.3	70.2	71.3	70.1
audio+video, eST	71.5	74.2	74.2	**73.1**
audio, CT	69.8	74.1	74.3	73.0
video, CT	68.3	71.3	71.4	70.8
audio+video, CT	71.5	75.0	75.6	**74.5**
audio, eCT	69.8	73.9	75.0	73.5
video, eCT	68.3	70.6	71.3	70.8
audio+video, eCT	71.5	75.4	76.5	**75.1**

4.3. Performance Evaluation

In this paper, we focus on the automatic recognition of the arousal. As classifier, we opted for linear Support Vector Machines (SVMs), as were used in the series of INTERSPEECH ComParEs [28], with a fixed complexity of 0.05. In terms of performance evaluation, we used the unweighted average recall (UAR). It equals the sum of the recalls per class divided by the number of classes, and better reflects the overall accuracy in the presence of imbalanced classes.

Before the SSL process, we randomly selected $N_L = 50$ instances from the pool set with the annotation by human oracle for initial training, which corresponds to approximately 5% of the whole pool set. The remaining instances in the pool set are considered as the unlabeled ones. At each machine-supervised learning iteration, we selected 20 additional instances for both ST and CT. Specifically, for the CT each 'view' chose an equal number of instances, i.e., each 'view' selected 10 instances. Note that, the stopping criterion is defined for an iteration time of $I = 40$, to ease performance comparison, and the whole learning process is conducted 30 independent times through all the following experiments.

For the unimodal SSL, the selected instances at each learning iteration are removed from the unlabeled data pool. The performance of the audio or the video based ST (dash) and CT (solid) is illustrated in Fig. 1 (a) and (b). Specifically, its corresponding statistical performance is indicated in Table 2. It can be seen that the audio-based SSL performs better than the video-based SSL for arousal emotion recognition, which is consistent with the literature



(a) *audio* based SSL (b) *video* based SSL (c) *audio-visual* based SSL

Fig. 1. Comparison between enhanced and non-enhanced Self-Training ((e)ST) and Co-Training ((e)CT) based on the audio (a), video (b), and audio-visual (c). The charts show the average unweighted average recalls (UARs) across 30 independent runs (and respective standard deviations) vs. number of selected instances (N_s).

[15, 16, 17, 18, 27, 29]. These improvements show not only in the initial learning process (69.8% vs. 68.3% of average UAR at $N_s = 0$), but also in the following consecutive learning process. The absolute gain is 3.7% and 4.5% for audio-based ST and CT over the whole 40 learning iterations, respectively. These gains are higher than those obtained for video – 1.9% (ST) and 2.1% (CT). These results can be attributed to the better initial performance of the audio based model, which is able to correctly label more instances in the learning process.

The results of the multimodal experiments are depicted in Fig. 1 (c). The initial performance of the model achieved an UAR of 71.5% (see Table 2), which is higher than the performance obtained by either audio (69.8%) or video (68.3%), showing thus the complementarity of audio-visual features for emotion recognition [15, 16, 17, 18, 27, 29]. During the CT process, the gain steadily increases as in the mono-modal experiments, and achieves a top performance of 75.6%. Through the whole 40 learning iterations, the mmCT has a statistically significant performance improvement compared with either audio or video based CT ($p < .001$ in Student's t-test). Similarly, mmST also outperforms the audio or video based ST at the significance level of .01 and .001, respectively.

Unlike the afore investigated SSL methods, the eSSL method selected $20 \times i$ instances at the i-th learning iteration, since the instances selected at any iteration were put back into the unlabeled pool set and considered as equal as the others for the next data selection. Fig. 1 shows the average performance of the enhanced Self-Training (eST) and Co-Training (eCT) based on the audio (a), video (b), and audio-visual (c) features averaged over 30 independent runs. The best performance is achieved by emmCT with an average UAR of 76.5%, which improved by 8.2% (one-side z-test, $p < .002$), 6.7% ($p < .01$), and 5.0% ($p < .05$) the initial video, audio, and audio-visual classifiers, cf. Table 2. A statistical comparison of the performance in the various experiments indicates that the eSSL performs significantly better ($p < .001$ in Student's t-test) than the conventional SSL in five out of six cases (except the video based eCT). This suggests that the enhanced learning procedure improves the quality of the selected instances at each learning iteration.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed the applications of multimodal Semi-Supervised Learning (SSL) in the context of emotion recognition. Unlike the conventional SSL for emotion recognition, we combined the audio and video modalities, which are known to provide complementary views of affective behaviours [15, 16, 17, 18, 27, 29]. Our hypothesis was that a performance improvement to the initially trained models could result in a more efficient SSL process by reducing the amount of wrongly labeled data at each iteration. Furthermore, we proposed an enhanced SSL algorithm that allows to correct wrongly labeled data with subsequent version of the enhanced model. In our experiments we compared unimodal (audio or video) and multimodal (audio and video) SSL using both Self-Training (ST) and Co-Training (CT) strategies.

Our experiments clearly demonstrated that the multimodal SSL outperforms the traditional unimodal SSL for arousal classification. For example, the multimodal CT averagely surpasses the audio and video based CT with about 1.5% and 3.7% of absolute UARs at the whole learning iterations, respectively. Furthermore, we have shown that our enhanced SSL model performs significantly better than the traditional SSL algorithm in most cases. In this case we achieved the best performance of all experiments by reaching a classification accuracy of 76.5% (UAR).

In future work, we plan to investigate valence recognition, as well as other multimodal databases. We also plan to extend the algorithm to process physiological data alongside audio-visual data. Furthermore, inspired by the work in [12], a cooperative learning that tries to efficiently share the annotation work between human and machine oracles will be further considered.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and the European Union's Horizon 2020 Programme through the Research Innovation Actions #645378 (ARIA-VALUSPA), #645094 (SEWA), and #644632 (MixedEmotions).

7. REFERENCES

[1] B. Schuller, E. Marchi, S. Baron-Cohen, A. Lassalle, H. O'Reilly *et al.*, "Recent developments and results of ASC-Inclusion: An integrated internet-based environment for social inclusion of children with autism spectrum conditions," in *Proc. of IDGEI*, Atlanta, GA, 2015, no pagination.

[2] E. Marchi, F. Ringeval, and B. Schuller, "Voice-enabled assistive robots for handling autism spectrum conditions: An examination of the role of prosody," in *Speech and Automata in the Health Care*, A. Neustein, Ed. Walter de Gruyter GmbH & Co KG, 2014, pp. 207–236.

[3] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Troster, and C. Haring, "Activity and emotion recognition to support early diagnosis of psychiatric diseases," in *Proc. of Pervasive Health*, Istanbul, Turkey, 2008, pp. 100–102.

[4] M. H. Sanchez, D. Vergyri, L. Ferrer, C. Richey, P. Garcia, B. Knoth, and W. Jarrold, "Using prosodic and spectral features in detecting depression in elderly males." in *Proc. INTERSPEECH*. Florence, Italy: ISCA, 2011, pp. 3001–3004.

[5] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. of INTERSPEECH*. Dresden, Germany: ISCA, 2015, pp. 1–5.

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

[7] 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, 2010.

[8] I. Cohen, N. Sebe, F. G. Cozman, and T. S. Huang, "Semi-supervised learning for facial expression recognition," in *Proc. of ACM SIGMM international workshop on Multimedia information retrieval*, New York, NY, 2003, pp. 17–22.

[9] M. Schels, M. Kächele, M. Glodek, D. Hrabal, S. Walter, and F. Schwenker, "Using unlabeled data to improve classification of emotional states in human computer interaction," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 5–16, 2014.

[10] I. Cohen, N. Sebe, F. Gozman, M. C. Cirelo, and T. S. Huang, "Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data," in *Proc. of CVPR*, vol. 1. Madison, WI: IEEE, 2003, pp. I–595.

[11] A. Mahdhaoui and M. Chetouani, "Emotional speech classification based on multi-view characterization," in *Proc. of ICPR*. Istanbul, Turkey: IEEE, 2010, pp. 4488–4491.

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

[13] J. Liu, C. Chen, J. Bu, M. You, and J. Tao, "Speech emotion recognition using an enhanced co-training algorithm," in *Proc. of ICME*. Beijing, China: IEEE, 2007, pp. 999–1002.

[14] R. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[15] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing,*, vol. 2, no. 2, pp. 92–105, 2011.

[16] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[17] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[18] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2015 – The 5th International Audio/Visual Emotion Challenge and Workshop," in *Proc. of ACM MM*. Brisbane, Australia: ACM, October 2015, pp. 1335–1336.

[19] X. Zhu, "Semi-supervised learning literature survey," Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, Tech. Rep. TR 1530, 2006.

[20] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. of ACL*. Stroudsburg, PA: ACM, 1995, pp. 189–196.

[21] O. Chapelle, B. Schölkopf, A. Zien *et al.*, *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[22] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. of IEEE workshop on Automatic Speech Recognition and Understanding*, Big Island, HY, 2011, pp. 523–528.

[23] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. of COLT*. Madison, WI: ACM, 1998, pp. 92–100.

[24] Z. Zhang, J. Deng, and B. Schuller, "Co-training succeeds in computational paralinguistics," in *Proc. of ICASSP*. Vancouver, Canada: IEEE, 2013, pp. 8505–8509.

[25] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[26] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions," in *Proc. of EmoSPACE 2013*. Shanghai, China: IEEE, April 2013.

[27] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, November 2015.

[28] B. Schuller *et al.*, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. INTERSPEECH*. Lyon, France: ISCA, August 2013, pp. 148–152.

[29] M. Valstar, B. Schuller, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: The 4th international audio/visual emotion challenge and workshop," in *Proc. of ACM MM*, Orlando, FL, 2014, pp. 1243–1244.