CO-TRAINING SUCCEEDS IN COMPUTATIONAL PARALINGUISTICS

Zixing Zhang, Jun Deng*, and Björn Schuller*

Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
zixing.zhang@tum.de

ABSTRACT

Data sparsity is one of the major bottlenecks in the field of Computational Paralinguistics. Partially supervised learning approaches can help leverage this problem without the need of cost-intensive human labelling efforts. We thus investigate the feasibility of co-training for exemplary paralinguistic speech analysis tasks spanning along the time-continuum: from short-term-related emotion to mid-term-related sleepiness and finally to long-term trait of gender. By dividing the acoustic feature space with two views as independent and sufficient as possible, the semi-supervised learning approach of co-training selects instances with high confidence scores in each view, and agglomerates them along with their predictions into initial training sets per iteration. Our experimental results on official Interspeech Computational Paralinguistics Challenge tasks effectively demonstrate co-training's superiority over the baseline formed by single-view self-training, especially for the short- and medium-term tasks emotion and sleepiness recognition.

Index Terms— Computational Paralinguistics, Co-Training, Semi-supervised Learning, Emotion, Sleepiness, Gender

1. INTRODUCTION

According to Abercrombie, "the conversational use of spoken language cannot be properly understood unless paralinguistic elements are take into account" [1]. Nowadays, Computational Paralinguistics attract more and more interest in the field of speech and language processing due to its potential broad application. For example, speakers' intention interpretation, conversation analysis and mediation, health detection, serving quality management, multimedia retrieval, more sensitively and naturally communicative robotics, can indeed benefit from paralinguistic information [2, 3].

There is, however, a crucial bottleneck in this field – data scarcity [2] –, which limits Computational Paralingustics in real-world application. Many approaches are trying to deal with this issue. Passive learning (PL), where data instances are annotated by human expects manually, is extremely time- and money-consuming. Thus, active learning (AL) and semi-supervised learning (SSL) seem to be promising alternatives. AL, aiming to select the 'most informative' instances from a large amount of unlabelled data, and mark them manually, not only can actually save a lot of time and labour, but enhance the performance in emotion recognition [4, 5]. In comparison with AL, SSL annotates massive unlabelled data by a system trained on small amount of labelled data automatically. This way without requiring more efforts from experienced human annotators, attracts much attention to enhance the robustness of existing classifiers [6]. Several ways of SSL haven been investigated, of which the most common way is 'single-view' self-training. This method is well evaluated in sound event [7], speech recognition [8, 9], and further tasks. Practical studies by Google Voice Search show that SSL has in fact already turned into common practice. In the related paralinguistics area, it was shown to give a better result in cross-corpus emotion classification [10]. Another approach of SSL – co-training –, as proposed in [11], has drawn considerable attention recently [12, 6]. Co-training assumes that the initial data can be divided into two (or more) disjoint sets of features or 'views' at the problem at hand. Then, the instances which are classified with high confidence score per view are added with the predicted labels to the training set over repeated iterations.

The main purpose of this paper is to investigate whether additional performance improvements can be obtained by applying co-training in a large scale and realistic paralinguistic classification task. According to [2, 3], a most intuitive taxonomy for paralinguistic phenomena is along the time axis, from short-term states, like emotion, confidence, stress, over medium-term phenomena, e. g., speaker states like intoxication, sleepiness, health state, to long-term traits, such as age, social status, personality, race, or gender. Albeit one cannot expect this paper could cover all of these cases, we select three common representative tasks to span the time continuum which were officially studied in INTERSPEECH Challenges from 2009 – 2011: short-term-related *emotion* [13], mid-term-related *sleepiness* [14], and long-term-related *gender* [15] of speakers.

In the following, we firstly introduce the three databases used for evaluation (Section 2); then, we describe the algorithms used for co-training and separation of features in Sections 3 and 4; further, we investigate the performance of co-training in three paralinguistic tasks (Section 5); finally, in Section 6 we conclude.

2. DATABASES

To investigate effectiveness of co-training in the paralinguistic field, we selected three frequently used publicly available databases for our experiments: the FAU Aibo Emotion Corpus, the Sleepy Language Corpus, and the Agender database. The main tasks of the three corpora cover different time-relations of paralinguistic groups from the short-term state of emotion, over the medium-term phenomena of sleepiness, to the long-term trait of gender. Speaker-independent partitioning of instances is shown in Table 1. In the following, we briefly introduce these three databases.

2.1. Emotion: FAU Aibo Emotion Corpus

Emotion recognition is evaluated on the FAU Aibo Emotion Corpus (AEC) [16], the official corpus of the INTERSPEECH 2009 Emotion Challenge (EC) [13]. It deals with recordings of children interacting with Sony's pet robot Aibo via German speech. The Wizard-of-Oz controlled Aibo robot sometimes disobeyed children's commands thus provoking various emotional reactions. The recording was executed at two different schools – MONT and OHM –, and feature 51

*The author acknowledges funding from the Chinese Research Council.

Table 1. Databases: Number of speakers (spk.) and instances per partition (Train, Develop or Test) for three paralinguistic corpora – AEC, SLC, and Agender. NEG: negative; IDL: idle; (N)SL: (non-)sleepy; C: children; M/m: male; F/f:female. No development set is defined on the FAU Aibo Emotion Corpus. Agender test labels are not freely available.

	AEC				**SLC**				**Agender**				
	# spk.	# instances			# spk.	# instances			# spk.	# instances			
		NEG	**IDL**	Σ		**NSL**	**SL**	Σ		**C**	**M**	**F**	Σ
Train	13m/13f	3 358	6 601	9 959	16m/20f	2 215	1 241	3 366	471	4 406	13 985	14 135	32 526
Develop					13m/17f	1 836	1 079	2 915	299	2 396	8 508	9 644	20 548
Test	8m/17f	2 465	5 792	8 257	14m/19f	1 957	851	2 808					
Σ	21m/30f	5 823	12 393	18 216	43m/56f	5 918	3 171	9 089	770	6 802	22 493	23 779	53 074

children with 21 males and 30 females, with ages ranging from 10 to 13 years. For our experiments, we use the whole corpus consisting of 18 216 chunks, and the 2-class labelling: **NEG**ative (subsuming $angry$, $touchy$, $reprimanding$, and $emphatic$) and **IDL**e (consisting of all other states). AEC labels came from five annotators on the word level.

2.2. Sleepiness: Sleepy Language Corpus

The Sleepy Language Corpus (SLC) [17], the official corpus of the Sleepiness Sub-Challenge of the INTERSPEECH 2011 Speaker State Challenge (SSC) [14], is employed for sleepiness recognition. To build this corpus, 99 participants with an age range of 20–52 years took part in six partial sleep deprivation studies. The recording took place in a realistic car environment or in lecture-rooms, including read and spontaneous German speech as detailed in [14]. To annotate the value of sleepiness, the Karolinska Sleepiness Scale (KSS) was used by the subjects and two raters. Scores ranging from 1–10 are given from *extremely alert* (1) to *cannot stay awake* (10). For training and classification purpose, the recordings (mean = 5.9, standard deviation = 2.2) were binarised into two classes: not sleepy ('**NSL**') and sleepy ('**SL**') with the threshold of 7.5 on the KSS.

2.3. Gender: Agender Database

For gender recognition, we choose the Agender database [18], the official corpus of the INTERSPEECH 2010 Paralinguistic Challenge (PC) Gender Sub-Challenge [15]. This database was collected by an external company aiming to identify possible speakers of the targeted age and gender groups. The participants were asked six times to call an automated Interactive Voice Response system to repeat given German utterances or produce free content. Each subject's six calls had to be done with a (mobile) phone in various recording environments and alternative days in order to ensure more variation of the voices. In the Challenge task, gender classification is treated as a three-class task – **C**hildren, **M**ale and **F**emale.

3. CO-TRAINING

Co-training is a paradigm of SSL. In comparison with single-view self-training, which regards the whole data as 'single view', co-training is considered as a multi-view learning algorithm. It presumes the features in the training data can be naturally separated into two sets [11], or more [19]. Firstly, a small set of labelled data serves as initial training material, and a large amount of unlabelled data are candidates to be exploited. Then, the initial training set is divided into two 'views', to build two classifiers based on these. After that,

Given:
- A small amount of labelled data L
- A large amount of unlabelled data U
- A learning domain with features X

Repeat:
- Split the domain features X into two views: X_1, X_2.
- Use X_1 to train classifier h_1, then classify U, after that choose N_1 examples with the most confident prediction
- Use X_2 to train classifier h_2, then classify U, after that choose N_2 examples with the most confident prediction
- Remove $N_1 \cup N_2$ from set U
- Add $N_1 \cup N_2$ to the labelled data L

Fig. 1. Algorithm of co-training.

each classifier recognises the whole unlabelled data and selects the instances that are predicted with high confidence score. Finally, these instances together with their predicted labels are added to the training set. Likewise, in one iteration an instance is either discarded, added once, added twice with the same label, or added twice with different labels. The whole process repeats until a stopping criterion is met. Figure 1 summarises the algorithm. Overall, thus, two views 'learn' mutually with additional informative instances, boosting the robustness of the final hypothesis.

4. ACOUSTIC FEATURES AND SEPARATION

In order to keep in line with the INTERSPEECH Challenge 2009–2011 conditions, we employ the same feature sets per task in our experiments as in the respective original Challenge. Thus, for emotion recognition, 384 features by brute-forcing based on 31 low-level descriptors (LLDs) / 42 functionals are implemented; for sleepiness detection, 4 368 features comprising 59 LLDs and 39 functionals are used, and for gender classification, 450 features composed by 38 LLDs and 21 functionals. As in the Challenge baselines, the features are extracted by our toolkit openSMILE [20]. For more details on the LLDs and functionals, please refer to [13, 14, 15].

Co-training, as multi-view learning, relies on two assumptions: *compatibility* and *independence* [11, 21]. Compatibility requires that each view is sufficient to train a good classifier. The assumption of independence demands that the two sets are conditionally independent. Aiming to approach the requested independence in a straight-forward choice, we split the whole LLDs into three partitions: energy-related,

Table 2. Feature separation based on LLDs. The symbols $\dagger, *, \sharp$ indicate the feature group which view-1 of co-training bases on for emotion, sleepiness, and gender recognition, respectively.

Group	Features in Group
Energy-related*	Sum of energy in auditory bands (loudness) Sum of RASTA-style filtered auditory spectral band energies RMS Energy RASTA-style filtered auditory spectral bands 1–26 (0–8 kHz) Spectral energy 25–650 Hz, 1 k–4 kHz
Spectral	Zero-Crossing Rate Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90 Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Slope F0, Probability of voicing Jitter (local, delta), Shimmer (local)
Cepstral$^{\dagger,\sharp}$	MFCC 1–12

spectral, and cepstral. Taking the (largest) feature set for sleepiness recognition as an example, Table 2 depicts this feature splitting. In comparison to the LLD groups shown in [14], the feature separation described in Table 2 differs – the one chosen here proved more suitable for the assumption of independence. For the other two tasks of emotion and gender recognition, the feature separation rule is the same. After dividing the whole feature domain into three partitions, the distribution of features is rather unbalanced across partitions. To solve this problem and satisfy the first assumption of sufficiency for each view, we re-arrange the three groups into two views. That is, the two groups including less features are agglomerated as one view. In Table 2, the symbols of $\dagger, *, \sharp$ mark the feature group which view-1 bases on for emotion, sleepiness, and gender recognition, respectively (remember that, the feature sets differ from task to task). Thus, the remaining two feature groups together each form view-2. Eventually, attribution ratios of view-1/view-2 are obtained as 288/96, 2 294/2 074, and 240/210 for the three tasks, respectively.

5. EXPERIMENTS AND RESULTS

5.1. Experimental Setup

For classification, we employ Support Vector Machines (SVMs) trained by the Sequential Minimal Optimization (SMO) algorithm as implemented in the Weka toolkit [22], keeping in line with the 2009 EC [13], 2010 PC[15], and 2011 SSC[14]. Further, we exactly follow the feature sets and classifier set-ups of the three Challenges: SVMs with linear kernel and a complexity constant optimised on development data of 0.05, 0.02, 0.1 for emotion, sleepiness and gender recognition, respectively. To evaluate the co-training algorithm in paralinguistics against a baseline formed by single-view self-training, AEC, SLC, and Agender, as described in Section 2, serve as data for emotion, sleepiness, and gender recognition. We randomly select 500 instances as initial human labelled training set from AEC and SLC, and 4 000 instances from Agender due to its larger size, which resembles approximately 3 %, 6 %, 8 % of each database. At each new iteration, 100 instances are selected by each view of co-training. Thus, for the baseline experiment of single-view self-training, 200 instances are chosen per iteration to provide a fair comparison. Finally, 25, 20, and 20 rounds of SSL iterations for AEC, SLC, and Agender are executed. Furthermore, to reduce the influence of 'lucky' or 'unlucky' selection for the initial training set, we repeat five times with different random generator initialisations ('seeds'), leading to five rounds of the whole iteration process executed. In addition, to deal with class imbalance, instance upsampling is used per iteration for emotion and sleepiness recognition. Details of the three experimental set-ups are given in Table 3.

Table 3. Experimental set-ups for AEC, SLC, and Agender. R: round number of whole processing; L: number of initial human labelled training instances; N_1+N_2: number of instances selected by view-1 and view-2 per iteration; I: iteration times per round.

#	R	L	N_1+N_2	I
AEC	5	500	100+100	25
SLC	5	500	100+100	20
Agender	5	4 000	100+100	20

5.2. Performance Evaluation

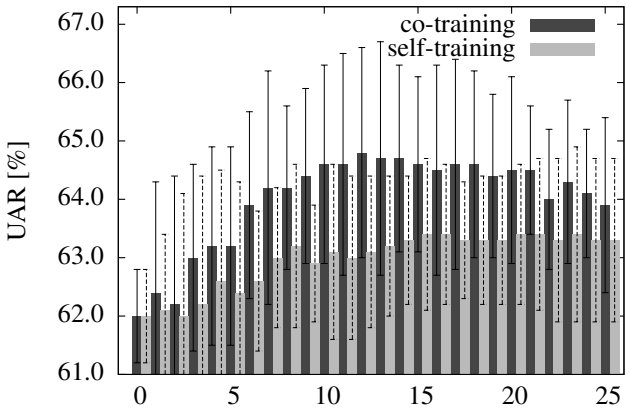
For performance evaluation, we use unweighted average recall (UAR), the sum of the recalls per class divided by the number of classes, which is the official competition measure of the 2009 EC, 2010 PC, and 2011 SSC. The chance level of UAR is 50.0 % for the binary emotion and sleepiness classification, and 33.3 % for the three-class gender classification.

Figure 2 displays the comparison of average performance and standard deviations between co-training (dark grey histograms with solid error lines) and single-view self-training (light grey histograms with dotted error lines) in five independent rounds for the three experiments based on the AEC, SLC, and Agender databases.

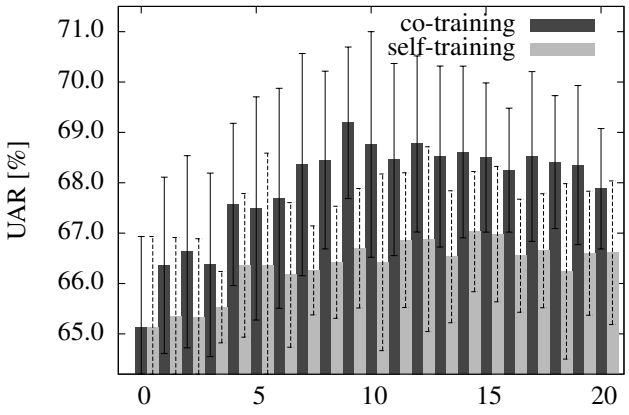
For the emotion recognition based on AEC, as seen in Figure 2 (a), the best mean UAR obtained by co-training with two-view learning based on feature partition in five independent rounds is 64.8 % UAR at the 12th iteration (24 k instances combined by co-training). This value boosts the initial mean UAR of 62.0 % UAR without any SSL iteration at the .001 significance level in a one-side z-test, and even greatly higher than the best mean UAR of 63.4 % achieved by single-view self-training at the 15th iteration at the .05 significance level (cf. Table 4). This improvement means that, the two-view SSL of co-training incorporates more additional information than single-view self-training. Further, one can also notice that the performance degrades quicker than in single-view self-training after the highest UAR gain. This phenomenon can probably be attributed to falsely labelled data by both views when the instances with increasingly lower confidence score are selected as iteration goes on, leading to doubling or accelerating error numbers added per iteration.

Figure 2 (b) depicts the UAR for recognition of sleepiness based on SLC. The gain obtained by co-training is also notable with a boost in mean UAR of almost 4.1 %, and 2.2 % absolute in comparison to the initial results (UAR of 65.1 %) and the best mean UAR achieved by single-view self-training (UAR of 67.0 %), respectively. Both improvements are significant at the .001 and .05 level (one-side z-test, cf. Table 4).

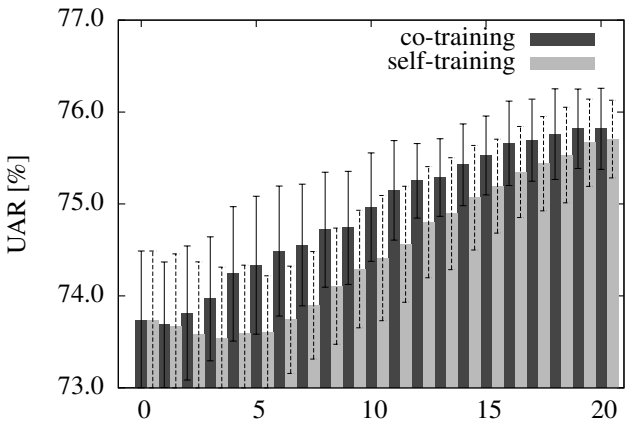
Finally, Figure 2 (c) shows the performance for recognition of gender based on the Agender database. It can been seen that, both co-training and single-view self-training significantly increase the initial UAR from 73.7 % to 75.8 %, and 75.7 %, at the significance level of .001 and .001 in a one-side z-test (cf. Table 4), respectively.



(a) *Emotion*: FAU Aibo Emotion Corpus



(b) *Sleepiness*: Sleepy Language Corpus



(c) *Gender*: Agender Corpus

Iteration

Fig. 2. Unweighted average racall (UAR) vs. number of iterations. Comparison between single-view semi-supervised learning and co-training in five independent rounds for three paralinguistic corpora – AEC, SLC, and Agender.

Overall, in terms of UAR for emotion, sleepiness, and gender recognition, the gain achieved by co-training based on feature multi-view is highly significant in comparison with the initial results of supervised learning for all three tasks. This holds even when compared to the baseline SSL approach. Details of performance improvement are given in Table 4.

Table 4. Classification evaluation comparison of co-training and single-view self-training in five independent rounds for three corpora of AEC, SLC, Agender. UAR: unweighted average recall; initial: initial supervised learning result; delta: absolute improvement of co-training over single-view self-training.

Mean of UAR[%]	**initial**	**self-training**	**co-training**	**delta**
AEC	62.0	63.4 ●○	**64.8** ●●	1.4 ●○
SLC	65.1	67.0 ○○	**69.2** ●●	2.2 ●○
Agender	73.7	75.7 ●●	**75.8** ●●	0.1 ○○

Significance levels [23]: ○○ not significant ●○ 0.05 ●● 0.001

6. CONCLUSIONS AND FUTURE WORK

In this paper, we evaluated the suitability of co-training in a large-scale study of paralinguistic tasks classification, by investigating the three representative cases of personal affect, speaker state, and speaker trait recognition as various temporal aspects. The results indicate that adding unlabelled data with a co-training algorithm can significantly enhance the performance of initial supervised learning – here by 2.8 %, 4.1 %, and 2.1 % UAR absolutely for emotion, sleepiness, and gender classification –, and even impressively improve over the performance of commonly used single-view self-training for the former two cases with a mean UAR of 1.4 % and 2.2 % absolutely (one-side z-test, $p<.05$). This renders co-training beneficial in real-world applications of Computational Paralinguistics analysis, where labelled data is sparse, but unlabelled data can be easily collected. of our previous paralinguistic tasks recognition model.

Future work may focus on the question of how to best partition features to fulfil sufficiency and independence of the views. This may lead to partitioning also according to functionals rather than LLDs. An obvious partitioning could also be by acoustic and linguistic feature information, once the latter should be taken into account.

7. RELATION TO PRIOR WORK

Previous work of co-training has been tested on a broader range of pattern recognition tasks such as web page classification: Co-training takes the text on the page as one view and the anchor text of the hyperlinks as the other view in [11]. Further, a similar idea in the domain of co-training for semi-supervised Expectation Maximisation is called co-EM [24] and exploited for the same task. For human action recognition, in [25] a boosted co-training algorithm is proposed, where inter-view and intra-view confidence addresses the view-sufficiency and dependence issues in co-training. In addition, co-training is also researched in on-line biometrics [26], music mood [27], vehicle [28], and handwritten word [29] classification.

For paralinguistic tasks recognition, up to now, first studies on co-training focused on emotion [30, 31, 32], which, however, is just one of manifold tasks in Computational Paralinguistics [2, 3]. In this contribution, we provide a large-scale investigation of co-training based on feature multi-view in a broader range of paralinguistics.

8. REFERENCES

[1] D. Abercrombie, *Elements of general phonetics*. Edinburgh University Press, 1967, vol. 203.

[2] B. Schuller, "The computational paralinguistics challenge," *Signal Processing Magazine, IEEE*, vol. 29, no. 4, pp. 97–101, 2012.

[3] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in Speech and Language – State-of-the-Art and the Challenge," *Computer Speech & Language, Special issue on Paralinguistics in Naturalistic Speech and Language*, vol. 27, no. 1, pp. 4–39, 2013.

[4] Z. Zhang and B. Schuller, "Active Learning by Sparse Instance Tracking and Classifier Confidence in Acoustic Emotion Recognition," in *Proc. INTERSPEECH 2012*, Portland, OR, 2012.

[5] D. Wu and T. Parsons, "Active class selection for arousal classification," in *Proc. Affective Computing and Intelligent Interaction (ACII)*, Memphis, TN, 2011, pp. 132–141.

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

[7] Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," in *Proc. ICASSP 2012*, Kyoto, Japan, 2012, pp. 333–336.

[8] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[9] K. Yu, M. Gales, L. Wang, and P. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 7, pp. 652–663, 2010.

[10] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Big Island, HY, 2011, pp. 523–528.

[11] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th annual conference on Computational Learning Theory*, Madison, WI, 1998, pp. 92–100.

[12] O. Chapelle, B. Schölkopf, A. Zien *et al.*, *Semi-supervised learning*. Cambridge, MA: MIT Press, 2006.

[13] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 312–315.

[14] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. INTERSPEECH 2011*, Florence, Italy, 2011, pp. 3201–3204.

[15] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. INTERSPEECH 2010*, Makuhari, Japan, 2010, pp. 2794–2797.

[16] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin: Logos Verlag, 2009.

[17] J. Krajewski, A. Batliner, and M. Golz, "Acoustic sleepiness detection - Framework and validation of a speech adapted pattern recognition approach," *Behavior Research Methods*, vol. 41, pp. 795–804, 2009.

[18] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A Database of Age and Gender Annotated Telephone Speech," in *Proc. 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 2010, pp. 1562–1565.

[19] Z. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529–1541, 2005.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[21] J. Du, C. X. Ling, and Z. Zhou, "When does co-training work in real data?" *IEEE Transactions on Knowledge Discovery and Data Mining*, vol. 23, no. 5, pp. 788–799, 2011.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[23] S. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data mining and knowledge discovery*, vol. 1, no. 3, pp. 317–328, 1997.

[24] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine learning*, vol. 39, no. 2, pp. 103–134, 2000.

[25] C. Liu and P. Yuen, "A boosted co-training algorithm for human action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 9, pp. 1203–1213, 2011.

[26] H. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, A. Noore, and A. Ross, "On co-training online biometric classifiers," in *Proc. 2011 International Joint Conference on Biometrics (IJCB)*, Washington DC, 2011, pp. 1–7.

[27] Y. Zhao, D. Yang, and X. Chen, "Multi-modal music mood classification using co-training," in *Proc. 2010 Computational Intelligence and Software Engineering (CiSE)*, Wuhan, China, 2010, pp. 1–4.

[28] A. Gepperth, "Co-training of context models for real-time vehicle detection," in *Proc. 2012 IEEE Intelligent Vehicles Symposium (IV)*, Alcala de Henares, Spain, 2012, pp. 814–820.

[29] V. Frinken, A. Fischer, H. Bunke, and A. Foornes, "Co-training for handwritten word recognition," in *Proc. 2011 Document Analysis and Recognition (ICDAR)*, Beijing, China, 2011, pp. 314–318.

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

[31] B. Maeireizo, D. Litman, and R. Hwa, "Co-training for predicting emotions with spoken dialogue data," in *Proc. 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Stroudsburg, PA, 2004, pp. 203–206.

[32] A. Mahdhaoui and M. Chetouani, "Emotional speech classification based on multi view characterization," in *Proc. 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, 2010, pp. 4488–4491.