Contemporary Stochastic Feature Selection Algorithms for Speech-based Emotion Recognition

Maxim Sidorov[1], *Christina Brester*[2], *Alexander Schmitt*[1]

[1]Ulm University, Ulm Germany
[2]Siberian State Aerospace University, Krasnoyarsk Russia

maxim.sidorov@uni-ulm.de, christina.brester@sibsau.ru, alexander.schmitt@uni-ulm.de

Abstract

In this study a class of Multi-Objective Genetic Algorithms (MOGAs) is proposed to select the most relevant features for the problem of speech-based emotion recognition. The employed evolutionary algorithms are the *Strength Pareto Evolutionary Algorithm (or SPEA)*, the *Preference-Inspired CoEvolutionary Algorithm with goal vectors (or PICEA)*, and the *Nondominated Sorting Genetic Algorithm II (or NSGA-II)*. Performances of the proposed algorithms were compared against conventional feature selection methods on a number of emotional speech corpora. The study revealed that for some of the corpora the proposed approach significantly outperforms the baseline feature selection methods up to 5.4% of relative difference.

Index Terms: Speech-based emotion recognition, human-computer interaction, computational paralinguistics

1. Introduction

A system which can deal with human emotions is highly desirable in such domains as human-computer interaction, human-robot interaction, biological and entertainment spheres. Nevertheless, the performance of the existing emotion recognizers is not sufficient for real-world applications yet. The most essential components of emotion recognizers are the used speech-based features, the modelling algorithm and Feature Selection (FS) techniques. Depending on the selected features and algorithms academic groups from all over the world obtain different Emotion Recognition (ER) performances. Moreover, the optimal feature set and the algorithms used are still under examinations. This means, that in each particular case different approaches and features should be investigated in order to select the quasi-optimal ones.

Based on previous successful applications of MOGAs in the field of multi-objective unconstrained optimization [1, 2] we applied these algorithms for the FS in speech-based ER. As it was mentioned, the performance of ER highly depends on the Classification Algorithm (CA), therefore we utilized several of the most used modelling algorithms in order to set-up the baseline results, but also to investigate the appropriateness of using MOGAs in the ER domain.

The rest of the paper is organized as follows: Significant related work is presented in Section 2. Section 3 describes the applied corpora and outlines their differences. State-of-the-art approaches and their results are shown in Section 4. Our approach on automated speech-based ER using MOGAs-based FS techniques is presented in Section 5. Conclusion and future work are described in Section 6.

2. Previous research

The main concept of MOGAs is inspired by the ideas of evolution and natural selection in Darwinism [3]. A survey of related papers revealed the most powerful ones, in terms of function optimization efficiency. These algorithms are SPEA [4], PICEA [5] and NSGA-II [6]. The majority of the published papers describe the results of their theoretical applications - mostly the optimization of a pre-defined function set [7, 8].

Since the problem of FS can be formulated as an multi-objective optimization problem [9] (i.e. minimization of the intra-class and maximization of inter-class distances or maximization of F_1 measure [10] and minimization of the feature number) we proposed using MOGAs for FS in the ER problem.

Regarding the ER procedure itself by analysing the related papers it was figured out that the most frequently used modelling algorithms are Multi Layer Perceptron (MLP) [11, 12], Support Vector Machine (SVM) [11, 13, 14, 15], k-Nearest Neighbours (kNN) [11, 16], and linear Logistic classifier [17]. Concerning the baseline FS methods the most frequently used are Information Gain Ratio (IGR) [18], and Principal Component Analysis (PCA) [19]. Additionally we suggest using the Chi- and GA-based [20] FS methods as a baseline.

As a baseline for acoustic features we consider the 384-dimensional feature vector which was used within InterSpeech 2009 Emotion Challenge [21, 22, 23].

3. Corpora description

All evaluations were conducted using several audio emotional databases. Here are their brief description and statistical characteristics.

The AVEC-2014 database was used for the fourth Audio-Visual Emotion Challenge and Workshop 2014 [24]. In order to obtain the level of depression, participants have been asked to fill in a standard self-assessed depression questionnaire (the Beck Depression Inventory-II) consisting of 21 questions. Each affect dimension (Arousal, Dominance, and Valence) has been annotated separately by a minimum of three and a maximum of five raters.

The Emo-DB emotional database [25] was recorded at the Technical University of Berlin and consists of labelled emotional German utterances which were spoken by 10 actors (5 females).

The EmotiW-2014 (or AFEW) audio-visual emotion corpus [26] was used for the first [27] and the second [28] "Emotion Recognition in the Wild Challenges".

The LEGO emotional database [29, 30] comprises non-acted English (American) utterances which were extracted from the SDS-based bus-stop navigational system [31].

Database	Language	Full length (min.)	File level duration		Paralinguistic Labels (Type)
			Mean (sec.)	Std. (sec.)	
AVEC-2014	German	164.08	65.63	46.22	Valence, arousal, dominance (Dimensions)
Emo-DB	German	24.7	2.7	1.02	Anger, boredom, disgust, anxiety, happiness, sadness, neutral (Categories)
EmotiW-2014	English	55.38	2.43	1.03	Angry, disgust, fear, happy, neutral, sad, surprise (Categories)
LEGO	English	118.2	1.6	1.4	Angry, slightly angry, very angry, neutral, friendly, non-speech (Categories)
RadioS	German	278.5	6.26	5.17	Neutral, happy, sad, angry (Categories)
SAVEE	English	30.7	3.8	1.07	Anger, disgust, fear, happiness, sadness, surprise, neutral (Categories)
UUDB	Japanese	113.4	1.4	1.7	Pleasantness, arousal, dominance, credibility, positivity (Dimensions)
VAM	German	47.8	3.02	2.1	Valence, Activation, Dominance (Dimensions)

Table 1: Databases description.

The RadioS database consists of recordings from a popular German radio talk-show. Within this corpus, 69 native German speakers talked about their personal troubles.

The SAVEE (Surrey Audio-Visual Expressed Emotion) corpus [32] was recorded as a part of an investigation into audio-visual emotion classification, from four native English male speakers.

The UUDB (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database [33] consists of spontaneous Japanese speech through task-oriented dialogue which was produced by 7 pairs of speakers (12 females), 4,737 utterances in total.

The VAM [34] dataset was created at Karlsruhe University and consists of utterances extracted from the popular German talk-show "Vera am Mittag" (Vera in the afternoon).

There is a statistical description of the used corpora in Table 1.

4. Baseline experiments

As a baseline approach we consider a number of machine learning algorithms, namely k-NN, MLP, SVM trained by SMO [35], and linear Logistic classifier as modelling algorithms. Further, a number of FS techniques were considered as a baseline, namely PCA, IGR, conventional Genetic Algorithm (GA) in wrapper mode [36], and Chi-based FS method. In order to enhance the statistical reliability we performed the same 6-fold emotion-stratified Cross-Validation (CV) for each emotional corpus, CA, and FS methods. For each classifier 5 types of FS methods were applied (without any FS, PCA, IGR, GA, and Chi). Thus, we performed 160 (8 databases x 4 classifiers x 5 FS methods) CV-based experiments as a baseline approach. The F_1 measure was selected as a main classification performance criterion.

More precisely, in case of experiments without any FS method on every iteration of CV procedure all the training portions were used to train the model and testing portions to test the emotional model. Finally, one average (over 6 folds of CV) value of F_1 measure is used as a performance. A similar procedure was repeated for all of the CAs (kNN, MLP, SVM, and Logistic). In the end, the result of the CA with the highest mean of the CV-based experiment was selected for each database and presented in Figure 1 with the label *"without"*.

Further, in case of conventional GA-based experiments boolean *true* means that the corresponding feature is essential and boolean *false* means it is an unessential one so that GA's individual is a 384-dimensional vector of ones and zeros. Thus, on each iteration of the GA the corresponding training partition is used to form a new instance of the explored database with the selected features. After that, the resulting training portion of the database with the selected features is used to run one more inner 6-fold CS in order to assess the current individual. An outcome of the inner CV is the quasi-optimal set of features for the current iteration of the outer CV. Then, the corresponding training and testing portions of the outer CV were transformed based on the optimal feature set and used to obtain the F_1 measure. This procedure was repeated 6 times within the outer CV for each CA. Again, the result of the CA with the highest mean was selected for each database and presented in Figure 1 with the label *"GA"*.

In contrast to parameter-free GA-based FS method for the rest of the baseline FS procedures (PCA, IGR, and Chi) another important parameter should be selected, namely the number of selected features. For the PCA-based approach this parameter corresponds to the number of principal components, whereas for the IGR- and Chi-based approaches this parameter is the number of included features with the highest ranking. In order to deal with the optimization of the parameter we introduce to use conventional GA, where genotype is encoded number of features (from 1 to 384) and the fitness value is F_1 measure obtained with the corresponding selected features again based on 6-fold inner CV. For this reason, in each iteration of outer CV the corresponding train portion is used to perform FS only once (since these FS methods do not require the inner CA), then the number of selected features was optimized with the GA approach. As early the outcome of the inner CV is the number of top-ranked features obtained by GA. Further, this optimal number of features was used to form the new instance of the outer CV's portions to train and test the CA. This procedure was repeated for each fold of outer CV and the averaged F_1 measure over 6 folds is used as a performance. Finally, only results of the best CA in terms of the highest average F_1 measure were included in Figure 1 with the corresponding captures.

It should be noted that the inner CVs were performed *only* with the train portion of outer CV to obtain the quasi-optimal solution and form the new instance of outer CV portions. The obtained portions were used to train and test the CA on each

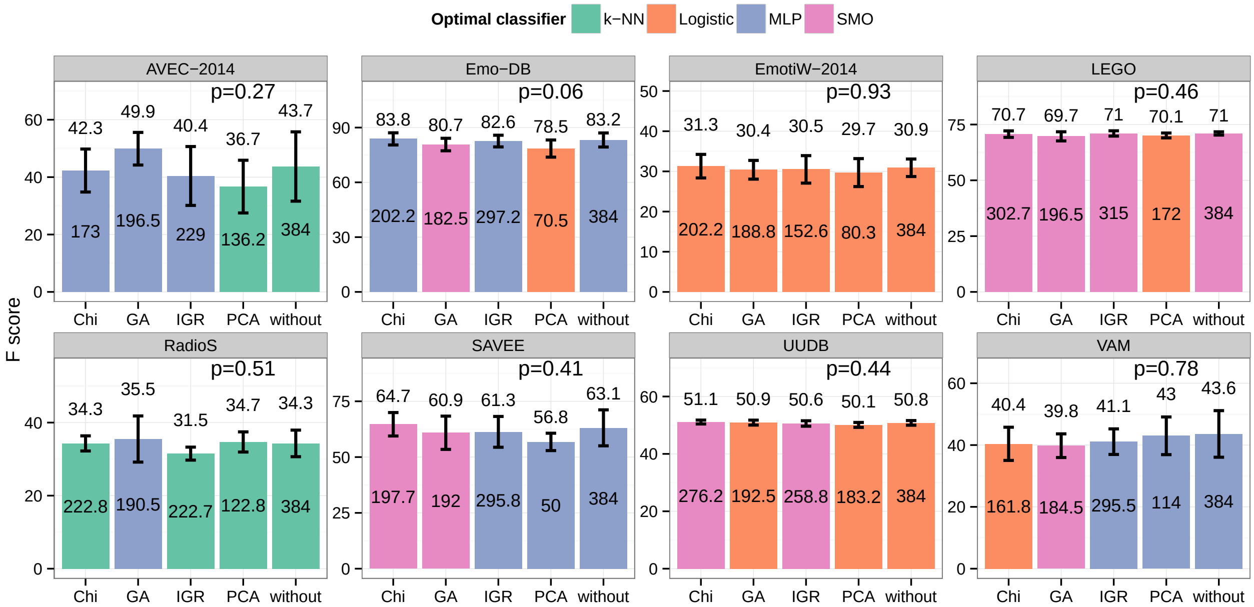

Figure 1: F_1 measure of speech-based emotion recognition with baseline feature selection methods and without any dimensionality reduction (without). Colours show the optimal classifiers, whereas the average number of features is shown within the bars. Result of single factor ANOVA test is above for all of the databases. All the experiments are 6-fold cross-validation emotion-stratified. Error bars demonstrate the population-based standard deviation of F_1 within the cross-validation.

iteration of outer CV.

The results of speech-based ER using the baseline approaches are demonstrated in Figure 1. There, for all the databases and feature selection methods only the results of the best classifier in terms of the higher F_1 measure achieved with CV experiments is provided. Further, in order to figure out significance difference throughout the baseline approaches a single factor ANOVA [37, 38] was conducted. The corresponding P values are also listed in Figure 1. Thus, in case of the EmotiW-2014 the corresponding value is equal to 0.93 which means that one may use whatever feature selection without significance difference in terms of F_1 measure. In contrast, in case of the Emo-DB with the corresponding value $p = 0.06$ the difference between PCA and Chi-based approaches is rather significant.

5. Proposed approach

The proposed heuristic feature selection scheme is based on estimating statistical metrics. We introduce the two-criteria model, specifically, the Intra-class distance (IA) and the Inter-class distance (IE) as optimization criteria:

$$IA = \frac{1}{n}\sum_{r=1}^{k}\sum_{j=1}^{n_r} d({p_j}^r, p_r) \rightarrow min, \tag{1}$$

$$IE = \frac{1}{n}\sum_{r=1}^{k} n_r d(p_r, p) \rightarrow max, \tag{2}$$

where ${p_j}^r$ is the j-th example from the r-th class, p is the central example of the data set, $d(.,.)$ denotes the Euclidian distance, p_r and n_r represent the central example and the number of examples in the r-th class.

As a feature selection technique we propose to use MOGAs operating with binary strings, where unit and zero correspond to a relative attribute and an irrelevant one respectively.

The common scheme of any MOGA includes the same steps as any conventional one-criterion GA:

```
Generate the initial population
Evaluate criteria values
while stop-criterion!=true do
    Estimate fitness-values;
    Choose the most appropriate individuals with the
    mating selection operator based on their
    fitness-values;
    Produce new candidate solutions through
    recombination;
    Modify the obtained individuals through mutation;
    Compose the new population (environmental
    selection);
end
```

Algorithm 1: General scheme of GA.

In contrast to one-criterion GAs, the outcome of MOGAs is the set of non-dominated points which form the Pareto set approximation. Designing a MOGA, researchers are faced with some issues which are referred to fitness assignment strategies, diversity preservation techniques, and ways of elitism implementation. Therefore, in this study we investigate the effectiveness of MOGAs, which are based on various heuristic mechanisms, from the perspective of the feature selection procedure. NSGA-II, PICEA with goal vectors, and SPEA were used as tools to optimize the introduced criteria (1), (2). Table 2 demonstrates the main characteristics of the used MOGAs.

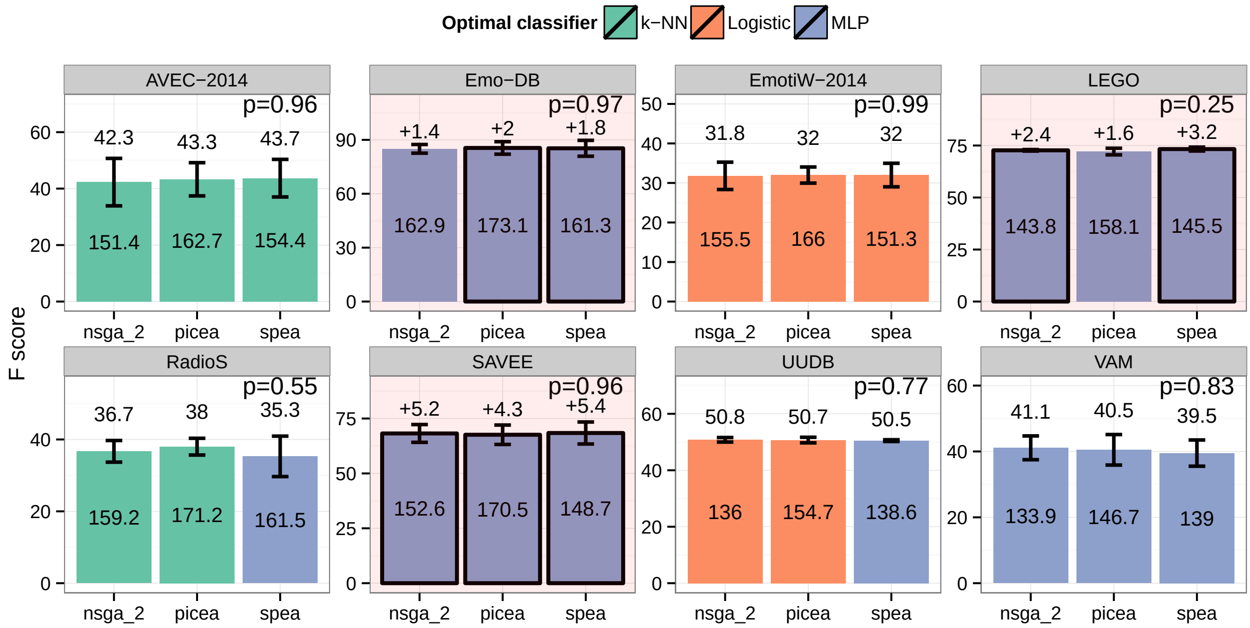

Figure 2: F_1 measure of speech-based emotion recognition with MOGA-based feature selection methods. Colours show the optimal classifiers, whereas the average number of features is shown within the bars. The result of a single factor ANOVA test is provided above for all of the databases. All the experiments are 6-fold cross-validation emotion-stratified, used in baseline experiments. Error bars demonstrate the population-based standard deviation of F_1 within the cross-validation. Coloured background shows the cases where ANOVA-based significant differences with baseline approaches have been achieved, moreover a relative improvement over the best baseline result is shown with the ”+” sign. Bars with bold frames indicate Z-test-based significant improvement against the best values from the baseline approaches (see the corresponding bars in Figure 1).

MOGA	Fitness Assignment	Diversity Preservation	Elitism
NSGA-II	Pareto-dominance (niching mechanism) and diversity estimation (crowding distance)	Crowding distance	Previous population and offspring
PICEA-g	Pareto-dominance (with generating goal vectors)	Nearest neighbour	Archive set, previous population and offspring
SPEA	Pareto-dominance (dominance rate)	Clustering technique	Archive set

Table 2: The main features of the used MOGAs.

As we have noticed, MOGAs return the set of candidate-solutions which cannot be preferred to each other. Taking into account this fact, we have proposed a way to derive the final solution based on the set of non-dominated points. It is assumed that the outcome of the MOGA is N binary strings (the set of non-dominated solutions). Each chromosome should be decoded to the reduced database, according the rule: if a gene is equal to ’0’ then eliminate the corresponding attribute, and if a gene is equal to ’1’ then include the respective feature in the database reduced. In short, we obtain N different sets of features and train N various classifiers based on these data. For each test example the engaged models vote for different classes according to their own predictions. The final decision is defined as a collective choice based on the majority rule.

Taking into consideration predictions of several classifiers is a good alternative to choosing one particular solution from the set of non-dominated points. In fact, candidates, which demonstrate high effectiveness on the training data, might often be the worst on the test data. Therefore, to avoid such cases, we use the scheme described.

The results of speech-based emotion classification with MOGAs are listed in Table 2.

6. Conclusion and future work

The achieved results show that the MOGA-based feature selection methods perform significantly better than the baseline methods for some of the used corpora, whereas in case of the rest of the corpora the proposed methods show the same performances. Moreover, in fact the proposed approach results in fewer features than state-of-the-art methods.

It can be concluded, that a class of algorithms such as MOGA-based FS can be wider used in the field of speaker state recognition and dialogue analysis.

One possible direction for future work is the examination of cooperative schemes of the mentioned algorithms, where several CAs and FS methods can be incorporated in order to use features of each other.

7. References

[1] F. Mendoza, J. L. Bernal-Agustin, and J. A. Dominguez-Navarro, "Nsga and spea applied to multiobjective design of power distribution systems," *Power Systems, IEEE Transactions on*, vol. 21, no. 4, pp. 1938–1945, 2006.

[2] D. Buche, P. Stoll, R. Dornberger, and P. Koumoutsakos, "Multiobjective evolutionary algorithm for the optimization of noisy combustion processes," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 32, no. 4, pp. 460–473, 2002.

[3] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[4] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach," *Evolutionary Computation, IEEE Transactions on*, vol. 3, no. 4, pp. 257–271, 1999.

[5] R. Wang, "Preference-inspired co-evolutionary algorithms," Ph.D. dissertation, University of Sheffield, 2013.

[6] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 2, pp. 182–197, 2002.

[7] H. Lu and G. G. Yen, "Rank-density-based multiobjective genetic algorithm and benchmark test function study," *Evolutionary Computation, IEEE Transactions on*, vol. 7, no. 4, pp. 325–343, 2003.

[8] E. Zitzler, K. Deb, and L. Thiele, "Comparison of multiobjective evolutionary algorithms: Empirical results," *Evolutionary computation*, vol. 8, no. 2, pp. 173–195, 2000.

[9] M. Venkatadri and K. S. Rao, "A multiobjective genetic algorithm for feature selection in data mining," *Int. J. Comput. Sci. Inf. Technol*, vol. 1, pp. 443–448, 2010.

[10] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in information retrieval*. Springer, 2005, pp. 345–359.

[11] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–577.

[12] M. Sidorov, S. Ultes, and A. Schmitt, "Comparison of gender-and speaker-adaptive emotion recognition," in *International Conference on Language Resources and Evaluation (LREC)*, 2009, pp. 3476–3480.

[13] B. Schuller, R. Müller, M. K. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles." in *INTERSPEECH*, 2005, pp. 805–808.

[14] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[15] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals." in *INTERSPEECH*, 2003.

[16] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition." in *INTERSPEECH*, 2002.

[17] M. E. Hoque, M. Yeasin, and M. M. Louwerse, "Robust recognition of emotion from speech," in *Intelligent Virtual Agents*. Springer, 2006, pp. 42–53.

[18] T. Polzehl, A. Schmitt, and F. Metze, "Salient features for anger recognition in german and english ivr portals," in *Spoken dialogue systems technology and design*. Springer, 2011, pp. 83–105.

[19] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[20] M. Sidorov, C. Brester, E. Semenkin, and W. Minker, "Speaker state recognition with neural network-based classification and self-adaptive heuristic feature selection," in *Informatics in Control, Automation and Robotics (ICINCO), 2014 11th International Conference on*, vol. 1. IEEE, 2014, pp. 699–703.

[21] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009, 2009, pp. 312–315.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[23] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[24] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

[25] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, 2005, pp. 1517–1520.

[26] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, no. 3, pp. 34–41, 2012.

[27] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 509–516.

[28] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 461–466.

[29] A. Schmitt, S. Ultes, and W. Minker, "A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system." in *LREC*, 2012, pp. 3369–3373.

[30] S. Ultes, M. J. P. Sánchez, A. Schmitt, and W. Minker, "Analysis of an extended interaction quality corpus."

[31] M. Eskenazi, A. W. Black, A. Raux, and B. Langner, "Let's go lab: a platform for evaluation of spoken dialog systems with real world users," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[32] S. Haq and P. Jackson, *Machine Audition: Principles, Algorithms and Systems*. Hershey PA: IGI Global, Aug. 2010, ch. Multimodal Emotion Recognition, pp. 398–423.

[33] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, 2011.

[34] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.

[35] J. Platt *et al.*, "Fast training of support vector machines using sequential minimal optimization," *Advances in kernel methodssupport vector learning*, vol. 3, 1999.

[36] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[37] R. A. Bailey, *Design of comparative experiments*. Cambridge University Press, 2008, vol. 25.

[38] D. C. Montgomery, *Design and analysis of experiments*. John Wiley & Sons, 2008.