Conflict Intensity Estimation from Speech Using Greedy Forward-Backward Feature Selection

Gábor Gosztolya

MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
ggabor@inf.u-szeged.hu

Abstract

In the recent years extracting non-trivial information from audio sources has become possible. The resulting data has induced a new area in speech technology known as computational paralinguistics. A task in this area was presented at the ComParE 2013 Challenge (using the SSPNet Conflict Corpus), where the task was to determine the intensity of conflicts arising in speech recordings, based only on the audio information. Most authors approached this task by following standard paralinguistic practice, where we extract a huge number of potential features and perform the actual classification or regression process in the hope that the machine learning method applied is able to completely ignore irrelevant features. Although current state-of-the-art methods can indeed handle an overcomplete feature set, studies show that they can still be aided by feature selection. We opted for a simple greedy feature selection algorithm, by which we were able to outperform all previous scores on the SSPNet Conflict dataset, achieving a UAR score of 85.6%.

Index Terms: computational paralinguistics, conflict detection, feature selection, SVM

1. Introduction

In the past, within the field of speech technology, most of the researchers' efforts were devoted to speech recognition. But in recent years they have turned their attention to other areas as well like emotion detection [1, 2], speaker verification [3], speaker age estimation [4], detecting social signals like laughter and filler events [5, 2, 6] and estimating the amount of physical or cognitive load during speaking [7, 8, 9]. What these tasks have in common is that what is considered noise in speech recognition (i.e. non-verbal audio information) becomes important, whereas what was relevant in speech recognition (i.e. what the speaker actually said) becomes irrelevant.

One such task is to determine the level of conflict in speech. Conflicts influence the everyday lives of people to a significant extent, either in their public or personal lives, and they are one of the main causes of stress [10]. With the rise of socially intelligent technologies, the automatic detection of conflicts can be the first step of handling them properly. Furthermore, conflict detection has straightforward applications such as monitoring incoming calls in call centers, where an important feedback of the employees is how they can handle conflicted situations.

In this study, we focus on the automatic estimation of the level of conflict in televised political debates [11, 12, 13]. This is mainly a regression task [14], as the intensity of conflict should be estimated, which are represented by real values. Of course, regarding actual applications, a categorical approach can be more practical, where the question is whether there is a conflict present or not; then the task is turned into a (binary) classification one.

In computational paralinguistics the standard approach is to extract a huge variety of features from each utterance, and perform classification or regression in this high-dimensional space. This approach clearly contradicts the ideal set-up of machine learning: ideally the feature set consists of only those features (possibly violated by some noise) that are relevant to the actual learning task. Features which capture only noise might even deteriorate the performance of the learning model because it becomes more prone to overfitting (i.e. it learns the noise).

Another problem with irrelevant features is that, due to the curse of dimensionality [15], models trained in a high-dimensional space require more examples. In contrast, more restricted feature sets usually lead to more compact models, which provide better generalization. An additional, obvious advantage of a smaller feature set is that it is also computationally cheaper to extract, store and process.

For these reasons, some researchers already applied feature selection in various paralinguistic tasks such as detecting emotion [16, 17, 18], depression [19] and autism [20, 21]. A number of such studies addressed the same problem as this paper does, i.e. feature selection for conflict intensity estimation [21, 22]. (For a detailed list, see [22].) Perhaps due to the special aspects of the area (few training vectors along with a huge feature set containing both redundant and irrelevant features), most of the efforts focused on adapting existing algorithms to paralinguistic tasks, or developing new ones.

2. The SSPNet Conflict Corpus

The SSPNet Conflict Corpus [11] contains recordings of Swiss French political debates taken from the TV channel "Canal9". It consists of 1430 recordings, 30 seconds each, making a total of 11 hours and 55 minutes. The ground truth level of conflicts was determined by manual annotation performed by volunteers not understanding French (French-speaking people were excluded from the list of annotators). Each 30-second long clip was tagged by 10 annotators, and in the end each recording was assigned a score in the range [-10, 10], 10 meaning a high level of conflict and -10 meaning no conflict at all.

The database contains both audio and video recordings, and the annotators could rely on both sources. In the recent experiments, however, attention was focused only on the audio information for a number of reasons. Firstly, the annotators judged the level of conflict in a similar way based on the two sources: the correlation of the scores was 0.95 [11]. Secondly, in a television political debate, audio can be a more reliable indicator: the subjects can hear all the participants, but they can only see the one that the cameraman of the debate has chosen, which is not the one speaking in many cases (especially in the heat of a debate when several persons may be speaking at the same time).

The audio clips of this dataset were also used in the Conflict sub-challenge of the Interspeech 2013 ComParE Challenge [12]. Besides completely discarding video information, the organizers of the Challenge made other steps to standardize the work on this dataset, and their setup has since been adopted by most researchers. Firstly, separate training, development and test sets were defined instead of relying on cross-validation as was done in [11]; secondly, a standard feature set was defined and extracted from the utterances by utilizing the tool openSMILE [23]. The set includes energy, spectral, cepstral (MFCC) and voicing-related low-level descriptors (LLDs) as well as logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness, leading to a total of 6373 features calculated for each utterance.

The evaluation metrics used for this dataset were also defined. Although it was admitted that this was mainly a regression task, and cross-correlation (CC) was used to measure performance, this task was also converted into a two-class classification one, defining the classes *low* and *high* based on the signum of the score. Classification accuracy was measured by using the Unweighted Average Recall (UAR) score [12]. This score was used to evaluate contributions submitted to the Challenge (e.g. [6, 21]), and it has been used in research papers since then (e.g. [22]). As this is clearly more like a regression task than a classification one, we find CC a more appropriate metric and we will primarily rely on it, but also give the UAR scores.

3. Feature Selection for Conflict Intensity Estimation

Feature selection seeks to reduce the list of features extracted for a given task. A large number of methods just consider the correlation of the target score and each feature, such as Correlation-based Feature Selection [24]. We can find several such applications in computational paralinguistics as well; for example, Kaya et al. proposed a feature selection method based on Canonical Correlation Analysis (CCA) and applied it for depression recognition [9]. Later, an improved version of this method was proposed, which randomly selected feature subsets and calculated the ranking of features based on the CCA-based score got on these subsets [22].

The weakness of performing feature selection only on the basis of correlation is, however, that after feature selection, some kind of machine learning method (SVM, ANN etc.) is trained using the restricted feature set, but the special aspects of this method are ignored during feature selection. This can be improved if we incorporate the machine learning method in the feature selection process [24]. This approach has the advantage that we will more likely pick those features which are relevant *for the given machine learning method*. Of course, this approach could lead to higher computational times than before; still, as in computational paralinguistics we do not have so many examples, this can be a good choice as long as we can keep the number of features low.

Such methods are usually divided into two types: forward and backward feature selection methods. Forward methods start with an empty (or very restricted) feature set, and expand it step-by-step [25]. Perhaps the most well-known forward method is the Sequential Forward Selection algorithm [26]: this, for each step, tries to add each feature to the set, and keep the one which resulted in the highest improvement in accuracy. Unfortunately, for such a large set of available features as in this task, this would be a very expensive method, as choosing each feature would require several thousands of model trainings.

Algorithm 1 Greedy Forward Feature Selection

Require: $features$: an ordered list of features
Require: $trainingData$: the training dataset
Require: $developmentData$: the development dataset
  $selectedFeatures := \emptyset$
  $bestCC := 0$
  for i := 1 $\rightarrow$ len(features) **do**
    $testFeatures := selectedFeatures \cup \{ features(i) \}$
    $model := trainML(trainingData(testFeatures))$
    $actScores := evaluateML(model, \ldots$
        $developmentData(testFeatures))$
    $actCC := CC(developmentData, actScores)$
    if $actCC \geq bestCC + \epsilon$ **then**
      $selectedFeatures := testFeatures$
      $bestCC := actCC$
    end if
  end for
  return $selectedFeatures$

Backward methods work in the opposite way: they start with the full feature set, which is reduced over time [25]. This approach also has a drawback in our case: due to the large number of features, we have to train our classifiers by using a lot of features (at least in the first part of the selection process), making it rather time-consuming as well.

An interesting approach was proposed by Räsänen et al. [21]: for each iteration they selected a random subset of features, and a classifier model was trained on this feature subset. The usefulness of each feature was calculated from the average classification accuracy using the given feature. (This is similar to applying a Multi-Armed Bandit setup during the training of AdaBoost.MH. [27, 28]) The drawback of this approach is, however, that it requires a really huge number of iterations to provide reliable accuracy estimates: the authors suggested $300,000$ iterations for the case of the SSPNet Conflict corpus. This may be an option only if we apply a simple machine learning method (they used KNN), but then the final accuracy score can be expected to be suboptimal after feature selection as well, due to the simplistic nature of the classifier.

3.1. Greedy Forward Feature Selection Method

For the above reasons, we opted for a greedy forward feature selection heuristic. In the proposed method we first set up an ordering of the possible features. Then we examine the features in this order; each feature is examined only once: if using the actual feature improves the quality of regression (or classification), we permanently add this feature to our set of selected features, otherwise we permanently discard it. For the pseudocode of this method, see Algorithm 1.

In this algorithm the order of the features is quite important because after choosing a feature we cannot discard it any more (and vice versa, after rejecting a feature, we cannot add it to the selected subset any more). Hence, instead of setting up a random order, we opted for a more sophisticated solution: first we measured the correlation of each feature with the target score. Then the features were sorted according to the absolute value of their correlation score in descending order, and this order was used during feature selection.

Besides this order of features (which is calculated automatically from the training set), we have one other parameter of the method. As in the CC score very small improvements are

Algorithm 2 Greedy Backward Feature Elimination

Require: $features$: an ordered list of features
Require: $trainingData$: the training dataset
Require: $developmentData$: the development dataset
  $selectedFeatures := features$
  $model := trainML(trainingData(selectedFeatures))$
  $actScores := evaluateML(model, \ldots$
    $developmentData(selectedFeatures))$
  $baseCC := CC(developmentData, actScores)$
  for i := len(features) $\rightarrow 1$ **do**
    $testFeatures := selectedFeatures \setminus \{ features(i) \}$
    $model := trainML(trainingData(testFeatures))$
    $actScores := evaluateML(model, \ldots$
      $developmentData(testFeatures))$
    $actCC := CC(developmentData, actScores)$
    if $actCC \geq baseCC$ **then**
      $selectedFeatures := testFeatures$
    end if
  end for
  return $selectedFeatures$

possible, the set of selected features can grow quite large by features which aid the regression only vaguely; therefore we required the improvement to be at least ϵ (a parameter of our method), thereby limiting the number of features selected.

3.2. Greedy Backward Feature Elimination Method

As was pointed out, forward methods tend to select a redundant feature set, since once a feature is picked, there is no way to get rid of it. For this purpose, it is not uncommon (e.g. [29, 30]) to use a backward feature elimination pass on the output of a forward pass to prune the feature set. If we apply the backward feature elimination step on the output of a forward pass, we also avoid the high running times of backward methods mentioned earlier, as the number of features examined can be expected to be much smaller than the total number of input attributes.

In our proposed Greedy Backward FS method, we process the feature set in the opposite order as we did in the forward pass (i.e. features less correlated to the target score are checked first). As the reason for this pass is not to improve the accuracy, but to exclude the redundant features, for each step we check if, by using the restricted feature set, we can achieve at least the same cross-correlation value as that with the input feature set. If we are able to do so, we can freely discard the given feature; otherwise we have to keep it. (See Algorithm 2.)

4. Experiments and Results

4.1. Experimental Setup

We generally followed the set-up of the ComParE challenge [12] regarding feature set and training / development / test set split, and applied both the CC and UAR metrics. Before feature selection every set was standardized independently, i.e. they were converted so as to have a zero mean and unit variance. We chose this strategy based on preliminary tests, and as it was standard practice in earlier works on this dataset (e.g. [21, 22]). We utilized the LibSVM [31] library, using the nu-SVR method with linear kernel; the value of C was tested in the range $10^{\{-5,\ldots,1\}}$ (even during feature selection), while ν was left at its default value of 0.5. ϵ for the forward feature selection pass was set to 10^{-4}.

Table 1: *Cross-correlation (CC) and UAR scores obtained in the literature for the SSPNet Conflict Corpus, following the ComParE 2013 setup. "—" means that the given score was not provided.*

Method	CC	UAR
Challenge baseline ([12])	0.816	80.8%
Speaker overlap (Grèzes, [6])	—	83.1%
Random Subset FS (Räsänen, [21])	0.826	83.9%
SLCCA FS (Kaya, [22])	—	84.6%
Greedy Forward FS (CC)	0.835	**85.6%**
Greedy Forward + Backward FS (CC)	**0.842**	85.1%
Greedy Forward FS (UAR)	0.838	83.9%
Greedy Forward + Backward FS (UAR)	0.836	84.3%

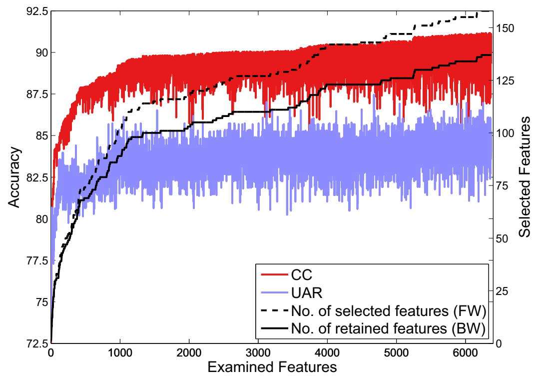

Figure 1: *Cross-correlation (CC) and AUC scores during the feature selection process. The dashed lines indicate the number of features chosen in the forward pass (FW) and kept in the backward pass (BW).*

During feature selection, we always trained our regression model on the training set, and calculated the correlation and UAR scores based on the predictions made on the development set. For each step, all possible values of C were tested. Finally, we trained our regression model on the training and development sets united, using the optimal feature set and SVR parameter, and evaluated it on the test set.

4.2. Results

The results achieved here and in the preceding studies can be seen in Table 1. As the UAR score was the de facto standard evaluation metric on this dataset and setting so far, we listed the UAR score of all previous studies; however, we also give the cross-correlation scores where available. It is clear that we exceeded the performance of all previous studies achieved on this dataset regarding UAR, and the CC scores are much higher than those in the previous studies as well. The backward pass improved cross-correlation even further, meaning that achieving the same CC score on the development set using fewer features improves the generalization capabilities of a machine learning method, and make the results more robust. (The backward pass somewhat reduced the UAR value, though.)

Note that our scores even exceeded the results presented in [11], where correlation scores around 0.75 were reported. These values cannot be directly compared, though, as the authors of [11] used 5-fold cross validation and a more restricted feature set (90 attributes overall).

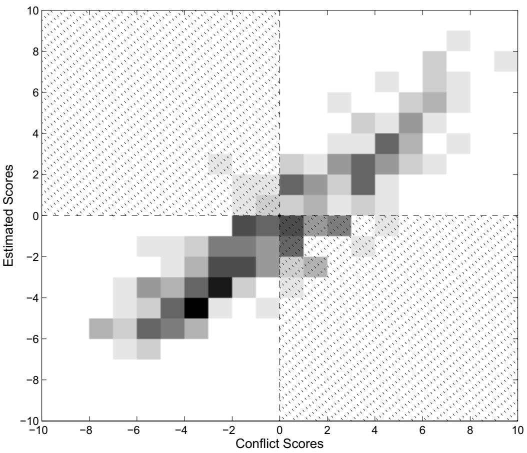

Figure 2: *Density scatter plot of the predictions on the dev set, after FW+BW feature selection optimizing for CC; these estimates had a CC score of 0.912 and a UAR score of 84.4.*

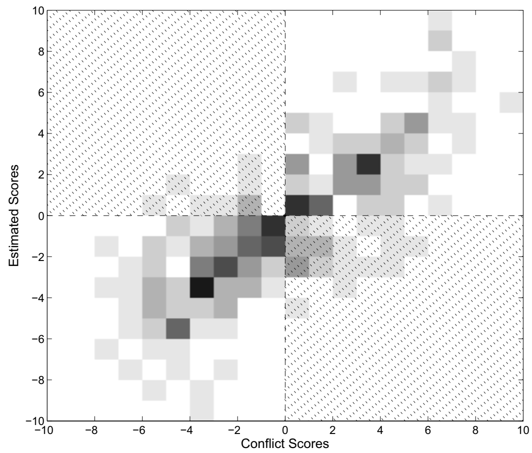

Figure 3: *Density scatter plot of the predictions on the dev set, after FW+BW feature selection optimizing for UAR; these estimates had a CC score of 0.796 and a UAR score of 85.4.*

4.3. The Selected Features

After the forward pass, our feature selection method retained only 158 features, which is a very small feature set overall: Räsänen et al. had 349 features [21], whereas Kaya et al. chose 500 features from the original 6373-long set [22]. The backward pass further reduced this value to 137. Two thirds of the full feature set consist of spectral-related features (66%), MFCC taking up another 22%; the rest is standing of energy- and voicing related attributes (5% and 7%, respectively). Surprisingly, after the forward pass, only 36% of the retained features belong to the spectral category, which is exceeded by the ratio of energy-related features (38%). The importance of MFCC and voicing-related attributes practically remain unchanged (22% and 4%, respectively), and these ratios are only slightly affected by the backward pass. This suggests that for conflict intensity estimation, mainly energy-related attributes are needed, which is logical if we consider that a high level of conflict is usually accompanied by the participants raising their voice. (We use the categories defined in [32], treating MFCC independently of other spectral features, following [22].)

Figure 1 shows the best CC and UAR scores for each iteration on the development set. CC shows an increasing tendency, which is not surprising as we optimized for it. The tendency of the increasing UAR scores can also be seen, but as it was not our primary metric, the values vary much more in both directions.

The dashed lines in Fig. 1 represent the number of features selected (FW) and kept (BW) after examining the actual feature. It is not surprising that most features are selected from the 1000 most correlated features of the original set, and most of them are kept even after the backward elimination pass. After examining 1000 features, the CC and UAR scores are also close to optimal. Still, valuable features are picked during the later iterations as well, as one-third of the final feature set is chosen in this phase.

5. On the UAR Evaluation Metric

Figure 2 shows the distribution of the output scores of the best-performing regressor model (FW+BW feature selection optimizing for CC) on the development set. The high CC value associated with this prediction is obvious; the UAR metric, however, concentrates only on the ratio of the correctly classified examples after discretization; thus, maximizing UAR practically means minimizing the number of examples falling into the shaded regions of Fig. 2. When compared with Fig. 3, which shows the same plot when optimizing for UAR, we see that although fewer points fall in the shaded region, overall the points are more scattered. The difference is even more obvious in the region around the origin. Although Fig. 3 has a higher UAR score than in Fig. 2, it is hardly better in any other way. And since these figures show the results for the development set, it is not surprising that the latter model appears to be less robust, in the end achieving a worse UAR score on the test set.

Over the past two years since the Challenge, researchers have found that during feature selection, we should focus on correlation instead of UAR [22], which accords with our experience. We also got better results if we first performed regression (i.e. using nu-SVR) and then discretized the predictions, instead of directly focusing on the binary classification task. It turns out that the UAR metric should only be used for final model evaluation; and even in this sub-task, its use can be questioned. For these reasons, we suggest that instead of the UAR score, future studies on this task should be evaluated using only regression metrics like cross-correlation and RMSE.

6. Conclusions

The standard set-up in computational paralinguistics is to extract a highly redundant feature set in the hope that the machine learning method applied can handle the redundancies and irrelevant attributes present. However, recent studies have shown that even current state-of-the-art classification and regression algorithms can be aided by feature selection. We proposed a simple greedy forward-backward feature selection algorithm for conflict intensity estimation; using this method, we were able to significantly outperform all previous scores on this task.

7. Acknowledgements

This publication is supported by the European Union and co-funded by the European Social Fund. Project title: Telemedicine-oriented research activities in the fields of mathematics, informatics and medical sciences. Project number: TÁMOP-4.2.2.A-11/1/KONV-2012-0073

8. References

[1] S. L. Tóth, D. Sztahó, and K. Vicsi, "Speech emotion perception by human and machine," in *Proceedings of COST Action*, Patras, Greece, 2012, pp. 213–224.

[2] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting autism, emotions and social signals using AdaBoost," in *Proceedings of Interspeech*, 2013, pp. 220–224.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[4] G. Dobry, R. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 1975–1985, 2011.

[5] T. Neuberger, A. Beke, and M. Gósy, "Acoustic analysis and automatic detection of laughter in Hungarian spontaneous speech," in *Proceedings of ISSP*, 2014, pp. 281–284.

[6] F. Grèzes, J. Richards, and A. Rosenberg, "Let me finish: Automatic conflict detection using speaker overlap," in *Proceedings of Interspeech*, 2013, pp. 200–204.

[7] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proceedings of Interspeech*, 2014.

[8] G. Gosztolya, T. Grósz, R. Busa-Fekete, and L. Tóth, "Detecting the intensity of cognitive and physical load using AdaBoost and Deep Rectifier Neural Networks," in *Proceedings of Interspeech*, Singapore, Sep 2014, pp. 452–456.

[9] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, "Canonical Correlation Analysis and Local Fisher Discriminant Analysis based multi-view acoustic feature reduction for physical load prediction," in *Proceedings of Interspeech*, Singapore, Sep 2014, pp. 442–446.

[10] P. Spector and S. Jex, "Development of four self-report measures of job stressors and strain: interpersonal conflict at work scale, organizational constraints scale, quantitative workload inventory, and physical symptoms inventory," *Journal of Occupational Health Psychology*, vol. 3, no. 4, pp. 356–367, 1998.

[11] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli, "Predicting continuous conflict perception with Bayesian Gaussian Processes," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 187–200, 2014.

[12] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social signals, Conflict, Emotion, Autism," in *Proceedings of Interspeech*, 2013.

[13] G. Gosztolya, "Estimating the level of conflict based on audio information using Inverse Distance Weighting," *Acta Universitas Sapientiae, Electrical and Mechanical Engineering*, 2014.

[14] R. Berk, *Statistical Learning from a Regression Perspective*. Springer Verlag, 2008.

[15] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton University Press, 1961.

[16] S. Zhang and Z. Zhao, "Feature selection filtering methods for emotion recognition in Chinese speech signal," in *Proceedings of ICSP*, Beijing, China, Oct 2008, pp. 1699–1702.

[17] T. Kostoulas, T. Ganchev, A. Lazaridis, and N. Fakotakis, "Enhancing emotion recognition from speech through feature selection," in *Proceedings of TSD*, Brno, Czech Republic, Sep 2010, pp. 338–344.

[18] H. P. Espinosa, C. A. R. G. a, and L. V. Pineda, "Bilingual acoustic feature selection for emotion estimation using a 3D continuous model," in *Proceedings of FG*, Santa Barbara, CA, USA, 2011, pp. 786–791.

[19] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," in *Proceedings of ICASSP*, Florence, Italy, 2014, pp. 3757–3761.

[20] K. Kirchhoff, Y. Liu, and J. Bilmes, "Classification of developmental disorders from speech signals using Submodular Feature Selection," in *Proceedings of Interspeech*, Lyon, France, Sep 2013, pp. 187–190.

[21] O. Räsänen and J. Pohjalainen, "Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech," in *Proceedings of Interspeech*, Lyon, France, Sep 2013, pp. 210–214.

[22] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, "Random discriminative projection based feature selection with application to conflict recognition," *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 671–675, 2015.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM Multimedia*, 2010, pp. 1459–1462.

[24] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, University of Waikato, 1999.

[25] L. Molina, "Feature selection algorithms: a survey and experimental evaluation," in *Proceedings of ICDM*, 2002, pp. 306–313.

[26] P. Devijver and J. Kittler, *Pattern Recognition, a Statistical Approach*. Prentice Hall, 1982.

[27] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the Multiarmed Bandit Problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.

[28] R. Busa-Fekete and B. Kégl, "Fast boosting using adversarial bandits," in *Proceedings of ICML*, vol. 27, 2010, pp. 143–150.

[29] J. Liu, J. Ye, and R. Fujimaki, "Forward-backward greedy algorithms for general convex smooth functions over a cardinality constraint," in *Proceedings of ICML*, 2014, pp. 503–511.

[30] M. Brendel, R. Zaccarelli, and L. Devillers, "A quick sequential forward floating feature selection algorithm for emotion detection from speech," in *Proceedings of Interspeech*, Makuhari, Japan, 2010, pp. 1157–1160.

[31] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[32] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-lie acoustic conditions - A new perspective on feature selection," in *Proceedings of Interspeech*, Lyon, France, 2013, pp. 2044–2048.