Wrapper-Based Acoustic Group Feature Selection for Noise-Robust Automatic Sleepiness Classification

Dara Pir[1], *Theodore Brown*[1,2], *Jarek Krajewski*[3,4]

[1]Dept. of Computer Science, The Graduate Center, City University of New York, New York, USA
[2]Dept. of Computer Science, Queens College, City University of New York, New York, USA
[3]Institute for Safety Technology, University of Wuppertal, Wuppertal, Germany
[4]Engineering Psychology, Rhenish University of Applied Science, Cologne, Germany

dpir@gradcenter.cuny.edu, tbrown@gc.cuny.edu, krajewsk@uni-wuppertal.de

Abstract

This paper presents a noise-robust Wrapper-based acoustic Group Feature Selection (W-GFS) method and its large noise Optimized (OW-GFS) version for automatic sleepiness classification and compares their performances with Correlation-based Feature Selection (C-FS) and Pearson Correlation Coefficient Feature Selection (CC-FS) filters. We use Interspeech 2011 Speaker State Challenge's "Sleepy Language Corpus" and baseline feature set. Group Feature Selection (GFS) considers the feature space in Low Level Descriptor groups rather than individually. Reduced time-complexity and potential generalization power of GFS are discussed. A model to predict on test data with changing Signal-to-Noise Ratio (SNR) is presented based on results from artificially corrupted development data with 10 dB SNR white-noise. Using Support Vector Machine, W-GFS achieves 2.6%, 4.2%, and 1.9% relative Unweighted Average Recall (UAR) improvement over the C-FS, CC-FS, and baseline feature set systems, respectively, on white-noise corrupted test data with randomly changing SNR within a broad range. The corresponding improvements for OW-GFS, using Voted Perception, are 4.8%, 9.8%, and 2.2% relative UAR on strongly white-noise corrupted test data with randomly changing SNR between -5 and +5 dB. Finally, we discuss consistent results obtained using everyday environment noises.

Index Terms: robust paralinguistics, computational paralinguistics, noise-robust feature selection, wrapper method, filter method

1. Introduction

The prevalence of sleep related accidents [1, 2, 3] and the imperative to prevent them highlights the importance of sleepiness detection systems. In situations where the use of certain types of detection methods, e.g., a spontaneous eye-blink detection system [4] requiring the use of intrusive sensors, is not optimal, speech can offer a unique advantage [5, 6, 7]. Moreover, the widespread nature of the sleep phenomenon is indicative of the abundance of applications concerned with its detection.

Computational paralinguistics tasks like sleepiness classification deal with the manner in which something is said rather than the content of what is said [8]. The binary task of Sleepiness Sub-Challenge was presented as part of the Interspeech 2011 Speaker State Challenge and employed the "Sleepy Language Corpus" (SLC) [9]. The 4368 acoustic baseline features generated using the openSMILE software [10] include those deemed relevant to sleepiness state [11] and result in a Sub-Challenge baseline score of 70.3% Unweighted Average Recall (UAR). The findings of the Sub-Challenge demonstrate that using larger feature sets result in superior performances. Furthermore, in the presence of various types and levels of noise, larger feature sets provide a larger pool for subsequent feature selection operations to choose from, in a data-driven fashion [12]. Using domain knowledge to design relevant features for classification in noisy environments is an alternative feature-based approach [13].

The two main types of feature selection methods are filters and wrappers [14]. The filter evaluates feature subsets based on statistical properties of data whereas the wrapper uses a classifier's performance score for the evaluation. The wrapper searches the feature space and evaluates feature subsets for selection. Wrapper-based Group Feature Selection (W-GFS) [15] uses a linear method, a fast variant of Best Incremental Ranked Subset (BIRS) [16], for feature space search and WEKA toolkit's [17] Support Vector Machine (SVM) [18] implementation, Sequential Minimal Optimization (SMO) [19] with linear Kernel, for feature subset evaluation. W-GFS modifies the basic wrapper by considering features in groups defined by Low Level Descriptor (LLD) partitions [20] rather than individually. Group Feature Selection (GFS) approach is motivated by two factors. First, GFS improves the tractability of the computationally intensive wrapper method by reducing the time complexity of the subset search component [15]. Second, an LLD-based GFS could potentially improve the generalization power of the classification algorithm by avoiding overfitting that may result from using a detailed individual feature search. Optimized Wrapper-based Group Feature Selection (OW-GFS) operates identically to W-GFS but its more restrictive selection criteria does not consider groups with evaluation scores of less than 55% UAR for selection.

The novel aspects of this work, to the best of our knowledge, are the following. First, although W-GFS has been used for another paralinguistics classification task [15], a specialized selection mechanism was employed that removed less than 1% of the features in the best performance. In this work, our two GFS methods remove about 80% and 90% of the features. In this mode, which achieves meaningful dimensionality reduction, the use of W-GFS is novel. Second, implementation of W-GFS in the context of noise-robust paralinguistics is novel. Finally, OW-GFS is a novel method that provides further noise-robustness under high noise conditions.

This paper is organized as follows. Section 2 describes the LLD-based partitioning and the BIRS algorithm for feature space search. Section 3 provides details about the corpus. Noise-robust feature selection and performance evaluation

Table 1: *Results in % UAR of SMO and VP classifications using the four feature selection methods and the baseline (BL) represented by columns of the table on high noise level test data. The best performances for each column are depicted in bold.*

CLS	W-GFS	OW-GFS	C-FS	CC-FS	BL
SMO1	61.5	62.3	59.6	59.9	**65.0**
SMO2	62.5	62.7	61.1	60.4	64.2
SMO3	**64.2**	63.5	61.6	**60.5**	63.2
SMO4	**64.2**	64.2	62.5	60.4	60.8
SMO5	63.4	64.0	62.9	60.2	58.5
SMO6	63.5	64.3	62.7	60.0	57.5
SMO7	62.6	64.1	62.9	60.4	55.9
VP	63.1	**66.4**	**63.4**	59.1	58.9

Table 2: *Results on medium noise level test data.*

CLS	W-GFS	OW-GFS	C-FS	CC-FS	BL
SMO1	64.1	64.3	64.1	61.9	65.8
SMO2	66.1	64.7	64.9	62.9	**66.4**
SMO3	67.2	65.4	65.3	64.6	66.1
SMO4	**67.5**	65.7	65.9	64.7	64.7
SMO5	**67.5**	65.9	65.9	64.9	61.7
SMO6	66.7	65.6	66.0	64.7	61.0
SMO7	65.4	66.1	66.1	**65.9**	59.5
VP	65.1	**67.4**	**66.2**	61.9	62.6

methods are explained in section 4. The experimental results are discussed in section 5 and the paper's conclusions and suggested future work are covered in the last section.

2. Background

2.1. LLD-Based Groups

Acoustic features are generated by chunk level application of functionals like arithmetic mean to LLD contours like RMS energy [21, 9]. The Sleepiness Sub-Challenge uses three sets of LLDs, each having a corresponding set of functionals listed in [9]. Using LLD-partitioned groups is acoustically motivated. If application of a statistical functional to an LLD contour generates a feature relevant to a classification task, it is likely that application of other functionals to the same LLD could be useful for the task as well and vice versa [15].

2.2. BIRS Search

BIRS is a linear forward search algorithm performed in two steps: ranking and feature subset selection. In the ranking step, the features are ranked from highest to lowest based on their evaluation score. In the feature subset selection step, the entire ranked feature set is traversed starting with an empty subset which selects features whose addition results in a subset that is evaluated to a higher UAR value, by a threshold level. Our fast variant of the algorithm used here does not employ cross-validation and t-test in the subset selection step. Wrapper evaluation cycles are used as the time complexity measure. The algorithm performs $2 * N$ evaluations, where N is the number of individual features in the search space. Our LLD-based GFS reduces the algorithm's $N = 4368$ evaluation cycles, in each step, to 118 cycles, i.e., the number of LLDs in the baseline feature set.

Table 3: *Results on low noise level test data. An additional complexity parameter = 0.01 (used by classifier SMO8) is needed to cover the range of interest for CC-FS.*

CLS	W-GFS	OW-GFS	C-FS	CC-FS	BL
SMO1	66.6	66.6	65.9	63.1	66.8
SMO2	68.2	67.3	66.3	64.8	67.1
SMO3	69.0	68.1	67.2	65.6	67.1
SMO4	**69.6**	**68.2**	**67.8**	66.1	**67.2**
SMO5	69.3	68.0	67.4	66.2	66.9
SMO6	68.6	68.0	67.1	66.4	67.0
SMO7	67.8	67.0	66.4	66.8	64.6
SMO8	...	...	...	**67.2**	...
VP	63.7	65.1	63.9	63.5	64.5

Table 4: *Results on unknown noise level test data.*

CLS	W-GFS	OW-GFS	C-FS	CC-FS	BL
SMO1	64.1	64.4	63.2	61.6	**65.9**
SMO2	65.6	64.9	64.1	62.7	**65.9**
SMO3	66.8	65.7	64.7	63.6	65.5
SMO4	**67.1**	66.1	**65.4**	63.7	64.2
SMO5	66.7	66.0	**65.4**	63.8	62.4
SMO6	66.2	66.0	65.3	63.7	61.8
SMO7	65.3	65.7	65.1	**64.4**	60.0
VP	64.0	**66.3**	64.5	61.5	62.2

3. Corpus

The SLC used in our classification contains speech recordings of 99 subjects made in realistic car and lecture-room settings and has a duration of 21 hours. The original 44.1 kHz recordings made with a microphone-to-mouth distance of 0.3 m are down-sampled to 16 kHz and use 16 bit quantization [9]. The levels of sleepiness 1 through 10 are reported according to the Karolinska Sleepiness Scale (KSS) [22] which is shown to be valid in certain studies [23]. A level equal or below 7.5 is classified as non-sleepy and one above 7.5 as sleepy.

4. Method

We first explain how our two W-GFS and OW-GFS methods and the two filters, Correlation-based Feature Selection (C-FS) [24] and Pearson Correlation Coefficient Feature Selection (CC-FS) [25] (implemented by WEKA's CfsSubsetEval and CorrelationAttributeEval, respectively), are used in the development phase to obtain the four noise-robust feature sets which are subsequently used in the evaluation phase. For the GFS methods, in the development phase, we train on the training set and predict on the development set. For the two filters, we train on the combined training set (training plus development sets combined). Next, we describe our evaluation of the four noise-robust selected feature sets using test data sets with changing noise levels. For evaluation, we train on the combined training set and report predictions on the test set. Features are standardized to standard normal and WEKA's Synthetic Minority Oversampling Technique (SMOTE) implementation [26] is used to balance the number of the classes in the development sets.

4.1. Noise-Robust Feature Selection on Development Data

In the absence of knowledge about the nature of the background noise, we model our feature selection systems using additive white Gaussian noise. First, in matched manner, we use devel-

Table 5: *Best performance results (bold entries) from Tables 1, 2, 3, and 4. The highest value of W-GFS and OW-GFS methods is displayed under "Best GFS" column.*

Noise	Best GFS	C-FS	CC-FS	BL
High	66.4	63.4	60.5	65.0
Med	67.5	66.5	65.9	66.4
Low	69.6	67.8	67.2	67.2
Unknown	67.1	65.4	64.4	65.9

Table 6: *% Improvement in relative UAR of the best performing model (Best Pair) over the best C-FS, CC-FS, and baseline models on each noise level test data.*

Noise	Best Pair	↑ C-FS	↑ CC-FS	↑ BL
High	OW-GFS, VP	4.8	9.8	2.2
Med	W-GFS, SMO4	1.6	2.5	1.8
Low	W-GFS, SMO4	2.6	3.5	3.6
Unknown	W-GFS, SMO4	2.6	4.2	1.9

opment data with additive white-noise of 10 dB Signal-to-Noise Ratio (SNR) level (generated by MATLAB's Communications System Toolbox function *awgn* [27]) to find the optimum linear kernel SMO complexity parameter. We model our feature selection methods for noise-robustness based on results from this mid-range SNR level. Second, using the obtained complexity, we perform W-GFS to obtain the noise-robust feature set which we will use to evaluate W-GFS on test data. Third, to add more robustness under high noise levels, the selected feature set obtained by W-GFS is reduced by removing groups with less than 55% UAR scores. The resultant feature set will be used to evaluate OW-GFS on test data. Finally, we obtain the two other feature sets using the C-FS and CC-FS filters. For CC-FS, we use the top 400 features as in [28]. The four noise-robust selected feature sets, in the mentioned order, are of sizes 935, 407, 138, and 400, respectively.

4.2. Evaluation System on Test Data

We use WEKA's linear kernel SMO and VotedPerceptron (VP) [29] implementations in the evaluation phase. If the type of noise is known, evaluation can be performed in a matched manner. In the absence of knowledge about the nature of the everyday environment noise, our four prediction models are trained in a partially matched fashion, i.e., using the clean combined training set reduced by the four noise-robust feature sets obtained using additive white-noise in the development phase. Predictions are made on noisy test data. Since the degree of similarity between white-noise and the particular everyday environment noise is unknown, prediction in a fully matched manner could produce unpredictable outcome. Noisy test data is produced as described below.

We generate three test sets with high, medium, and low levels of additive white-noise, respectively. To generate the high level noise test data, following a uniform distribution, we randomly add white-noise to the test data using an SNR level between -5 and +5 dB. This generation process allows for evaluation under changing noise levels. The medium noise level test data is generated in a similar manner except that the SNR range is between +5 and +15 dB. In order to include clean data as part of our test sets, the low noise level test data is generated similarly to the other levels using the +15 to +25 dB range but only with a 50% chance following a uniform distribution. The remaining 50% of data is clean.

Table 7: *Results on everyday environment noise test data (counterpart of Table 6's last row). The "Best Pair" obtains 63.1 % UAR.*

Noise	Best Pair	↑ C-FS	↑ CC-FS	↑ BL
Unknown	OW-GFS, SMO7	2.1	3.1	1.4

In practice, hyperparameters tuned in the development phase are used for prediction in the test and evaluation phases. However, using W-GFS for tuning the SMO complexity parameter gives the model an unfair advantage over others. To fairly compare our four feature selection and baseline models using the SMO classifier, therefore, we need to evaluate their performances using several SMO complexity parameters spanning the range of interest. The seven values of interest range from 0.00005 to 0.005 in approximately double increments, i.e., 0.00005, 0.0001, 0.0002, ..., 0.005. The corresponding classifiers are named SMO1, SMO2, SMO3, ..., SMO7, respectively. For the VP classifier, WEKA's default settings are used.

5. Experimental Results

Table 1 depicts results obtained on high noise level test data. The highest value in this table represents the model (feature selection method and classifier pair) that achieves best performance on high noise level test data. Tables 2 and 3 are generated similarly for the medium and low level noise test data. Table 4 is the average of the high, medium, and low noise level test data tables and represents the unknown noise level. The highest value in this table represents the model that achieves best performance under changing and unknown noise levels.

To facilitate comparison of results obtained by the four feature selection methods and the baseline we generate Tables 5 and 6. Table 5 displays the best performance results (bold entries) from Tables 1, 2, 3, and 4. Results from these tables demonstrate that our two GFS methods obtain the top two performances for each noise level. The highest value obtained by the W-GFS and OW-GFS methods is displayed under the common "Best GFS" column. Table 6 is constructed in the following manner. Column 1 displays the noise level. Column 2 identifies the model (method and classifier pair) that attains best performance on each noise level test data. Column 3 (↑ C-FS) depicts, for each level, the percent improvement in relative UAR of the best model over the best C-FS model. Similarly, columns 4 (↑ CC-FS) and 5 (↑ BL) show improvements of the best model over the best CC-FS and best baseline models. These results demonstrate that the best GFS method consistently outperforms the C-FS, CC-FS, and baseline models on all four noise level test data. Specifically, for high noise, the OW-GFS and VP pair outperforms the best C-FS, CC-FS, and baseline models by 4.8%, 9.8%, and 2.2% relative UAR, respectively. The overall best performing model under changing and unknown noise level, the W-GFS and SMO4 (SMO with complexity = 0.0005) pair, outperforms the best C-FS, CC-FS, and baseline models by 2.6%, 4.2%, and 1.9% relative UAR, respectively.

Finally, we evaluated the four feature selection methods and the baseline on test data with additive everyday environment noises. Recording of nature plus driving car sounds was undergone SNR level changes according to the same distributions that was used in generating the unknown noise level test data for additive white-noise. The resultant test data was generated directly rather than through the averaging process used for the white-noise case. The results are displayed in Table 7. The performance improvement pattern is similar to that of the white-noise case (last row of Table 6) although the best performance value of 63.1% UAR using everyday environment noise (not shown in the table) is expectedly lower than the 67.1% obtained by the white-noise counterpart.

6. Conclusions and Future Work

In the absence of specific knowledge about the type and number of noise sources, we used additive Gaussian white-noise to model the background noise. This noise model was employed by four feature selection methods to obtain four reduced feature sets. Systems based on these reduced feature sets performed sleepiness classification on the SLC test data with additive white and everyday environment noises whose SNR levels are changed dynamically following a uniform distribution. In a partially matched design, our best GFS systems showed performance improvement over the two alternative filter systems and the baseline. For further real-world noise-robustness, our GFS systems could be trained on models that incorporate actual everyday environment noises and subsequent predictions could be made in a matched manner.

7. References

[1] A. I. Pack, A. M. Pack, E. Rodgman, A. Cucchiara, D. F. Dinges, and C. W. Schwab, "Characteristics of crashes attributed to the driver having fallen asleep," *Accident Analysis & Prevention*, vol. 27, no. 6, pp. 769–775, 1995.

[2] A. T. McCartt, S. A. Ribner, A. I. Pack, and M. C. Hammer, "The scope and nature of the drowsy driving problem in new york state," *Accident Analysis & Prevention*, vol. 28, no. 4, pp. 511–517, 1996.

[3] W. Vanlaar, H. Simpson, D. Mayhew, and R. Robertson, "Fatigued and drowsy driving: A survey of attitudes, opinions and behaviors," *Journal of Safety Research*, vol. 39, no. 3, pp. 303–309, 2008.

[4] P. P. Caffier, U. Erdmann, and P. Ullsperger, "Experimental evaluation of eye-blink parameters as a drowsiness measure," *European Journal of Applied Physiology*, vol. 89, no. 3-4, pp. 319–325, 2003.

[5] J. Krajewski and B. Kröger, "Using Prosodic and Spectral Characteristics for Sleepiness Detection," in *INTERSPEECH 2007 – 8th Annual Conference of the International Speech Communication Association, August 27-31, Antwerp, Belgium, Proceedings*, 2007, pp. 1841–1844.

[6] F. Hönig, A. Batliner, T. Bocklet, G. Stemmer, E. Nöth, S. Schnieder, and J. Krajewski, "Are men more sleepy than women or does it only look like – automatic analysis of sleepy speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 995–999.

[7] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Acoustic-Prosodic Characteristics of Sleepy Speech – Between Performance and Interpretation," in *Speech Prosody 2014*, pp. 864–868.

[8] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2014.

[9] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, August 28–31, Florence, Italy, Proceedings*, 2011, pp. 3201–3204.

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

[11] L. S. Dhupati, S. Kar, A. Rajaguru, and A. Routray, "A novel drowsiness detection scheme based on speech analysis with validation using simultaneous eeg recordings," in *Automation Science and Engineering (CASE), 2010 IEEE Conference on*. IEEE, 2010, pp. 917–921.

[12] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion Recognition in the Noise Applying Large Acoustic Feature Sets," in *Speech Prosody 2006*.

[13] V. Mitra, A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5795–5799.

[14] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997.

[15] D. Pir and T. Brown, "Acoustic Group Feature Selection Using Wrapper Method for Automatic Eating Condition Recognition," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015, Proceedings*, 2015, pp. 894–898.

[16] R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," *Pattern Recognition*, vol. 39, no. 12, pp. 2383–2392, 2006.

[17] 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.

[18] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] J. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," Technical Report MSR-TR-98-14, Microsoft Research, April 1998.

[20] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *frontiers in Psychology*, vol. 4, pp. 227–239, 2013.

[21] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *INTERSPEECH 2009 – 10th Annual Conference of the International Speech Communication Association, September 6–10, 2009, Brighton, UK, Proceedings*, 2009, pp. 312–315.

[22] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, "Karolinska sleepiness scale (kss)," in *STOP, THAT and One Hundred Other Sleep Scales*. Springer, 2012, pp. 209–210.

[23] 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, no. 3, pp. 795–804, 2009.

[24] M. A. Hall and L. A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper," in *FLAIRS Conference*, 1999, pp. 235–239.

[25] J. Lee Rodgers and W. A. Nicewander, "Thirteen Ways to Look at the Correlation Coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321–357, 2002.

[27] MATLAB and Communications System Toolbox Release 2014b, The Mathworks, Inc., Natick, Massachusetts, United States.

[28] F. Eyben, F. Weninger, and B. Schuller, "Affect Recognition in Real-Life Acoustic Conditions A New Perspective on Feature Selection," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France, Proceedings*, 2013, pp. 2044–2048.

[29] Y. Freund and R. E. Schapire, "Large Margin Classification Using the Perceptron Algorithm," *Machine Learning*, vol. 37, no. 3, pp. 277–296, 1999.