Relevant Acoustic Group Features for Automatic Sleepiness Recognition

Dara Pir[1] and Jarek Krajewski[2,3]
[1]*Information Technology Program, Guttman Community College, City University of New York, New York, U.S.A.*
[2]*Institute for Safety Technology, University of Wuppertal, Wuppertal, Germany*
[3]*Engineering Psychology, Rhenish University of Applied Science Cologne, Cologne, Germany*

Keywords: Automatic Sleepiness Recognition, Acoustic Group Features, Computational Paralinguistics, Human-Computer Interaction.

Abstract: This paper compares the discriminating powers of various acoustic group features for the task of automatic sleepiness recognition using three different classifiers: Voted Perceptron, Simple Logistic, and Random Forest. Interspeech 2011 Sleepiness Sub-Challenge's "Sleepy Language Corpus" (SLC) is used to generate the 4368 acoustic features of the official baseline feature set. The feature space is divided into Low-Level Descriptor (LLD) partitions. We consider the resulting feature space in groups rather than individually. A group feature corresponds to a set of one or more LLD partitions. The relevance of various group features to sleepiness state is then evaluated using the mentioned classifiers. Employing larger feature sets has been shown to increase the classification accuracy in sleepiness classification. Our results, however, demonstrate that a much smaller subset of the baseline feature set outperforms the official Sub-Challenge baseline on the SLC test data.

1 INTRODUCTION

Sleep is a widespread phenomenon and there is great interest in detecting it. An important area of interest is the prevalent sleep related road accidents (Pack et al., 1995; McCartt et al., 1996; Vanlaar et al., 2008) where sleep detection systems may play a critical role in preventing them. Another area of great interest is in the emerging fields of Ubiquitous Computing, Intelligent Companion, and Robots for Eldercare, where both the naturalness and efficiency of Human-Computer Interaction can be enhanced by knowing the speaker's various states such as fatigue and sleepiness. The system may provide feedback about the user's state to appear more emphatic and may adapt its output to the user's state to render the communication more intelligible (Krajewski et al., 2008).

Employing the speech mode in recognition applications offers advantages: 1) over modes that use intrusive or inconvenient sensors which require being attached to the subject or 2) under conditions that degrade performance of alternative modes, e.g., low-light environment for the visual mode (Krajewski and Kröger, 2007; Krajewski et al., 2008; Hönig et al., 2014a; Hönig et al., 2014b; Pir et al., 2017).

The binary task of sleepiness classification, a computational paralinguistics (CP) task, was presented at Interspeech 2011 Sleepiness Sub-Challenge (Schuller et al., 2011). Whereas Automatic Speech Recognition (ASR) tries to determine which words are spoken, CP attempts to discover how those words are spoken and thereby gain knowledge about the various aspects and conditions of the speakers, e.g., age, gender, sleepiness, friendliness, etc. (Schuller and Batliner, 2014; Hantke et al., 2016).

The Sleepiness Sub-Challenge employs the openSMILE toolkit (Eyben et al., 2010) to generate the 4368 baseline acoustic features from the "Sleepy Language Corpus" (SLC) (Schuller et al., 2011). A score of 70.3% Unweighted Average Recall (UAR) is obtained by the Sub-Challenge baseline feature set which is considered to be a collection of relevant features for the task of sleepiness (Schuller et al., 2011; Dhupati et al., 2010). The UAR measure takes the imbalance between class instances into account (Schuller et al., 2011)

We divide the feature space into Low-Level Descriptor (LLD) partitions and consider the resulting feature space in groups rather than individually (Pir and Brown, 2015). A group feature corresponds to a set of one or more LLD partitions. The Sub-Challenge findings demonstrate that using larger fea-

Pir, D. and Krajewski, J.
Relevant Acoustic Group Features for Automatic Sleepiness Recognition.
DOI: 10.5220/0006779802090214
In *Proceedings of the 4th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2018)*, pages 209-214
ISBN: 978-989-758-299-8

Table 1: Type: Type of features included in LLD set. LLD-Set: Names of LLDs contained in the set. nLLD: Number of LLDs in the LLD set.

Type	LLD-Set	nLLD
Energy	Sum of auditory spectrum	1
	Sum of RASTA-style filtered auditory spectrum	1
	RMS Energy	1
	Zero-Crossing Rate	1
Spectral	RASTA-style filt. auditory spectrum	26
	Spectral energy	2
	Spectral Roll Off	4
	Spctral Flux, Entropy, Variance, Skewness, Kurtosis, Slope	6
	MFCC	12
Voice	F0	1
	Probability of voicing	1
	Jitter (local)	1
	Jitter (delta)	1
	Shimmer	1

ture sets improves the classification performance. By evaluating the various constituent group features of the baseline feature set we attempt to discover those that are more relevant for the task. This information, in turn, may be helpful in designing feature sets with superior performance.

The novel aspect of this paper in the context of sleepiness recognition, to the best of our knowledge, is the use of multiple classifiers in evaluating the discriminating power of the various group features that comprise the openSMILE generated baseline feature set without employing any feature selection operation.

This paper is organized as follows. Section 2 describes group features, LLDs, and the Mel-Frequency Cepstral Coefficients (MFCCs). Section 3 describes the corpus. Section 4 covers the group features considered in experimental evaluation, the over-sampling step, and the classifiers employed. Section 5 presents the experimental results and the paper's conclusions and suggested future work are discussed in Section 6.

2 FEATURES

Acoustic features are generated, on the chunk level, by application of functionals like statistical moments or quartiles to LLD contours such as Fundamental Frequency or Zero-Crossing Rate (Schuller et al., 2009; Weninger et al., 2013; Schuller et al., 2011).

2.1 Group Features

Acoustic group features are comprised of LLD partitions. LLD-based portioning is acoustically motivated since the features within an LLD are suprasegmental information on the same single LLD and therefore related (Schuller et al., 2011; Pir and Brown, 2015; Pir et al., 2016; Pir et al., 2017)

2.2 LLDs

The list of basic LLDs used by the openSMILE toolkit to generate the baseline acoustic features are shown in Table 1. The LLDs are divided by type into: energy-related, spectral, and voice-related sets. Details about the full set of functionals applied to these LLDs can be found in (Schuller et al., 2011). For each basic LLD shown, there is a corresponding delta LLD. Delta is defined as the first order difference function of the related LLD (Eyben, 2016). The total number of LLDs, i.e., basic and delta combined, are therefore twice the numbers shown in Table 1.

2.3 MFCCs

The MFCCs (Davis and Mermelstein, 1980) are among the most popular features for ASR and have also been successful in many other audio processing tasks such as speaker identification, music signal processing, and CP (Eyben, 2016; Lerch, 2012). We therefore chose to investigate the performance of four of their smaller subsets in addition to the full set which is included in the openSMILE feature set.

3 CORPUS

The 21 hours of SLC speech recordings were made from 99 subjects. The recordings have a sampling rate

Table 2: Classification results in % UAR on test data using three classifiers: VP, SL, and RF. Type: Type of features included in the group feature. Group: Abbreviation for the group feature. Basic: Results for basic group features. Delta: Results for delta group features. Comb: Results for the combined basic and delta group features. The best performance for each row is depicted in bold.

		VP			SL			RF		
Type	Group	Basic	Delta	Comb	Basic	Delta	Comb	Basic	Delta	Comb
Energy	ENER	61.3	60.8	62.4	64.0	64.5	**65.7**	58.4	65.3	62.9
Spectral	RFilt	66.0	61.6	64.9	**68.3**	65.5	67.7	64.9	63.6	65.5
	SpEn	42.7	45.8	46.1	54.2	57.7	**58.7**	51.0	53.0	51.0
	SpRo	62.0	56.4	58.2	61.9	62.2	62.0	63.7	62.6	**64.9**
	Mom+	63.3	62.4	**64.4**	63.2	64.1	63.8	60.7	61.1	61.2
	MFCC12	63.7	59.4	62.1	64.2	59.0	63.3	**67.9**	61.7	66.0
Voice	VOICE	66.7	63.3	65.3	66.6	65.8	66.7	**67.8**	62.6	67.6
All	ALL	67.1	65.3	67.9	**70.4**	66.5	68.9	68.7	69.0	69.8
MFCC	MFCC1	66.4	55.9	63.5	66.6	54.0	67.3	66.2	61.5	**67.8**
	MFCC2	66.4	60.8	65.8	67.4	59.4	67.1	70.1	62.5	**70.4**
	MFCC3	65.3	59.1	65.5	64.9	56.9	65.2	70.1	62.3	**70.7**
	MFCC4	64.4	60.0	64.5	64.8	59.0	64.4	**70.2**	62.4	**70.2**

of 16 kHz and are quantized at 16 bits (Schuller et al., 2011).

SLC uses the Karolinska Sleepiness Scale (Shahid et al., 2012), which defines levels of sleepiness 1 though 10. A level greater than 7.5 represents the sleepy state and one equal to or less than 7.5 indicates a non-sleepy state.

4 METHOD

4.1 Group Features Considered

For each LLD, three different feature sets are considered for evaluation. The first set consists of the basic features. The second set is comprised of the delta features of the corresponding basic features of the first set. The third set is the basic and delta sets combined.

The number of energy-related LLD sets is small and hence they are combined into a single group feature for performance evaluation. The voice-related LLD sets are also combined for the same reason. Each of the five spectral LLD sets is considered as a group feature and evaluated separately. In addition, four MFCC group features are considered where the first group includes the first coefficients, the second group includes the first and second coefficients, and so forth.

4.2 Synthetic Minority Over-sampling Technique

WEKA's Synthetic Minority Over-sampling Technique (SMOTE) implementation (Chawla et al., 2002) is used to balance the number of class instances in the development set.

4.3 Classifiers

Our three classifiers use WEKA's (Hall et al., 2009) implementations: VotedPerceptron (VP) (Freund and Schapire, 1999), SimpleLogistic (SL) (Sumner et al., 2005), and RandomForest (RF) (Breiman, 2001). The training is done on training and development data combined and the evaluation is performed on test data using the Sub-Challenge defined data partitions.

5 EXPERIMENTAL RESULTS

5.1 Relevance

The classification results on test data are shown in Table 2 for the basic, delta, and combined group features using three classifiers: VP, SL, and RF. The bold entries in each row of the table depict the best performance for the associated group feature. The results demonstrate that relevance depends on the classifier used, e.g., Mom+ is most relevant for VP, RFilt for SL, and SpRo for RF. Alternatively, average relevance could be defined as the classification accuracies' average for all classifiers.

5.2 Comparison Among Basic, Delta, and Combined Groups Features

Table 3 shows the classification accuracy averages for the basic, delta, and combined group features using each classifier. The basic groups' averages outperform those of the delta for every classifier and the combined groups' averages outperform those of the basic for the RF and SL classifiers.

Table 3: Classification accuracy averages in % UAR for the basic, delta, and combined group features using each classifier. Classifier: Classifier used. The rest of the columns are described as in Table 2.

Classifier	Basic	Delta	Comb
RF	65.0	62.3	65.7
SL	64.7	61.2	65.1
VP	62.9	59.2	62.6

In addition, Table 4 shows the percentage of cases where each of the basic, delta, and combined groups achieves top performance for every classifier. The combined groups are associated with about 57% of the cases followed by the basic groups representing around 35% of the cases. The delta groups achieve top results in only about 8% of the cases.

Table 4: The percentages of best performances for every classifier. Classifier: Classifier used. Basic, Delta, and Comb: The percentage of times the corresponding category of group features achieves top performance for a particular classifier and group feature.

Classifier	Basic	Delta	Comb
All	35%	8%	57%

5.3 Most Relevant Group Features Overall

The three most relevant, i.e., top performing, group features using any classifier are shown in Table 5. All the top three performances are achieved by the RF classifier on some MFCC group. The best result, 70.7% UAR, achieved by the combined MFCC3 group feature outperforms the official Sub-Challenge baseline of 70.3% while being comprised of only 6 LLDs.

We have not included in the table the 70.4% UAR result obtained by the SL classifier on the entire baseline feature set as our goal is to find particular group features that are relevant to the sleepiness state.

5.4 Most Relevant Group Features on Average

The three most relevant group features on average are shown in Table 6. The MFCC2 basic group feature, comprised of only 2 LLDs, achieves the highest average performance of 68.0% UAR. We have not included performance results on the entire baseline feature set for the reason described above.

Table 5: Three top performing group features using any classifier. Group: Abbreviation for the group feature. Cat: Group feature category. Cls: Classifier. nLLD: Number of LLDs in the group feature. % UAR: Classification result in % UAR. Performance results that are superior to the Sub-Challenge baseline are depicted in bold.

Group	Cat	Cls	nLLD	% UAR
MFCC3	Comb	RF	6	**70.7**
MFCC2	Comb	RF	4	**70.4**
MFCC4	Basic	RF	4	70.2

Table 6: Columns are described as in Table 5.

Group	Cat	nLLD	% UAR
MFCC2	Basic	2	68.0
MFCC2	Comb	4	67.8
MFCC3	Comb	6	67.1

5.5 Irrelevant Group Feature

Our analysis has allowed us to identify a group feature, the SpEn (Spectral Energy) group, that performs worse than chance for all three categories as shown in Table 7. We note that the group is not irrelevant when using the SL classifier and the performance of the RF classifier is near chance.

Table 7: Classification results of the VP classifier for SpEn (spectral energy) group feature. Columns are described as in Table 2.

	VP		
Group	Basic	Delta	Comb
SpEn	42.7	45.8	46.1

5.6 Comparison with Previous Results

Of the six accepted papers in the Interspeech 2011 Sleepiness Sub-Challenge only three surpassed the highly competitive baseline (Schuller et al., 2014). The best performing system achieved a UAR of 71.7% which is not a significant improvement over the baseline at an $\alpha = 0.05$ level (Schuller et al., 2014). The mentioned system employed two other standard feature sets in addition that of the Sleepiness Sub-Challenge. Classification results were obtained using the authors' proposed Asymmetric Simple Partial Least Squares method, SVM, and fusions (Huang et al., 2011; Schuller et al., 2014). Our best performance is achieved, however, using only about 5% of the features (6 out of 118 LLDs) in the baseline feature set. This reduction is significant in two important ways. First, it renders the training phase of computationally intensive classifiers more tractable. Second, it provides knowledge to domain experts by identifying those features that are better suited to the task.

Although (Hönig et al., 2014a) reports a state-of-the-art result of 71.9% UAR, the dataset used is smaller and a direct performance comparison cannot be made.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we compared the accuracy performances of LLD-based group features that comprise the Sleepiness Sub-Challenge's baseline feature set using three different classifiers. Our analysis has revealed the relative discriminating powers of various group features for a specific classifier as well as averaged over all classifiers. Our top performance, which achieved improvement over the official baseline, was obtained using the Random Forest classifier and the MFCC group feature containing the first three coefficients. The mentioned MFCC group feature includes only 6 LLDs out of the 118 that comprise the baseline feature set.

Future work includes extending the current framework for evaluating relevance for group features in the context of other paralinguistics tasks as well as developing feature selection methods that incorporate the knowledge obtained about group feature relevance in this paper.

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

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

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

Dhupati, L. S., Kar, S., Rajaguru, A., and Routray, A. (2010). 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*, pages 917–921. IEEE.

Eyben, F. (2016). *Real-time speech and music classification by large audio feature space extraction*. Springer.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM.

Freund, Y. and Schapire, R. E. (1999). Large Margin Classification Using the Perceptron Algorithm. *Machine Learning*, 37(3):277–296.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hantke, S., Weninger, F., Kurle, R., Ringeval, F., Batliner, A., Mousa, A. E.-D., and Schuller, B. (2016). I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on asr performance. *PloS one*, 11(5):e.0154486. doi:10.1371/journal.pone.0154486.

Hönig, F., Batliner, A., Bocklet, T., Stemmer, G., Nöth, E., Schnieder, S., and Krajewski, J. (2014a). 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)*, pages 995–999. IEEE.

Hönig, F., Batliner, A., Nöth, E., Schnieder, S., and Krajewski, J. (2014b). Acoustic-prosodic characteristics of sleepy speech – between performance and interpretation. In *Proc. of Speech Prosody*, pages 864–868.

Huang, D.-Y., Ge, S. S., and Zhang, Z. (2011). Speaker State Classification Based on Fusion of Asymmetric SIMPLS and Support Vector Machines. In *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, 2011, Florence, Italy, Proceedings*, pages 3301–3304.

Krajewski, J. and Kröger, B. J. (2007). 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*, pages 1841–1844.

Krajewski, J., Wieland, R., and Batliner, A. (2008). *An Acoustic Framework for Detecting Fatigue in Speech Based Human-Computer-Interaction*, pages 54–61. Springer Berlin Heidelberg, Berlin, Heidelberg.

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons.

McCartt, A. T., Ribner, S. A., Pack, A. I., and Hammer, M. C. (1996). The scope and nature of the drowsy driving problem in new york state. *Accident Analysis & Prevention*, 28(4):511–517.

Pack, A. I., Pack, A. M., Rodgman, E., Cucchiara, A., Dinges, D. F., and Schwab, C. W. (1995). Characteristics of crashes attributed to the driver having fallen asleep. *Accident Analysis & Prevention*, 27(6):769–775.

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

Pir, D., Brown, T., and Krajewski, J. (2016). Wrapper-Based Acoustic Group Feature Selection for Noise-Robust Automatic Sleepiness Classification. In *Proceedings of the 4th International Workshop on Speech*

Processing in Everyday Environments (CHiME 2016), September 13, 2016, San Francisco, CA, USA, pages 78–81.

Pir, D., Brown, T., and Krajewski, J. (2017). Automatic Driver Sleepiness Detection Using Wrapper-Based Acoustic Between-Groups, Within-Groups, and Individual Feature Selection. In *Proceedings of the 3rd International Conference on Vehicle Technology and Intelligent Transport Systems, Edited by: Oleg Gusikhin, Markus Helfert and António Pascoal, ISBN: 978-989-758-242-4, April 22–24, Porto, Portugal*, pages 196–202.

Schuller, B. and Batliner, A. (2014). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Schuller, B., Steidl, S., and Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge. In *INTERSPEECH 2009 – 10th Annual Conference of the International Speech Communication Association, September 6–10, 2009, Brighton, UK, Proceedings*, pages 312–315.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. (2011). The INTERSPEECH 2011 Speaker State Challenge. In *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, 2011, Florence, Italy, Proceedings*, pages 3201–3204.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., Weninger, F., and Eyben, F. (2014). Medium-term speaker statesa review on intoxication, sleepiness and the first challenge. *Computer Speech & Language*, 28(2):346–374.

Shahid, A., Wilkinson, K., Marcu, S., and Shapiro, C. M. (2012). Karolinska sleepiness scale (kss). In *STOP, THAT and One Hundred Other Sleep Scales*, pages 209–210. Springer.

Sumner, M., Frank, E., and Hall, M. (2005). *Speeding Up Logistic Model Tree Induction*, pages 675–683. Springer Berlin Heidelberg, Berlin, Heidelberg.

Vanlaar, W., Simpson, H., Mayhew, D., and Robertson, R. (2008). Fatigued and drowsy driving: A survey of attitudes, opinions and behaviors. *Journal of Safety Research*, 39(3):303–309.

Weninger, F., Eyben, F., Schuller, B., Mortillaro, M., and Scherer, K. (2013). On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in psychology*, 4:227–239.