Relevant Acoustic Group Features for Automatic Sleepiness Recognition

Dara Pir\textsuperscript{1} and Jarek Krajewski\textsuperscript{2,3}

\textsuperscript{1}Information Technology Program, Guttman Community College, City University of New York, New York, U.S.A.
\textsuperscript{2}Institute for Safety Technology, University of Wuppertal, Wuppertal, Germany
\textsuperscript{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; Vantaa 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; Höning et al., 2014a; Höning 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).

The Sleepiness Sub-Challenge findings demonstrate that using larger 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-
tecture 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 partitioning is acoustically moti-

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.

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

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 through 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.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Basic</th>
<th>Delta</th>
<th>Comb</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>65.0</td>
<td>62.3</td>
<td>65.7</td>
</tr>
<tr>
<td>SL</td>
<td>64.7</td>
<td>61.2</td>
<td>65.1</td>
</tr>
<tr>
<td>VP</td>
<td>62.9</td>
<td>59.2</td>
<td>62.6</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Basic</th>
<th>Delta</th>
<th>Comb</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>35%</td>
<td>8%</td>
<td>57%</td>
</tr>
</tbody>
</table>

### 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.

### Table 5: Three top performing group features using any classifier.

<table>
<thead>
<tr>
<th>Group</th>
<th>Cat</th>
<th>Cls</th>
<th>nLLD</th>
<th>% UAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC3</td>
<td>Comb</td>
<td>RF</td>
<td>6</td>
<td>70.7</td>
</tr>
<tr>
<td>MFCC2</td>
<td>Comb</td>
<td>RF</td>
<td>4</td>
<td>70.4</td>
</tr>
<tr>
<td>MFCC4</td>
<td>Basic</td>
<td>RF</td>
<td>4</td>
<td>70.2</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Group</th>
<th>Cat</th>
<th>nLLD</th>
<th>% UAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC2</td>
<td>Basic</td>
<td>2</td>
<td>68.0</td>
</tr>
<tr>
<td>MFCC2</td>
<td>Comb</td>
<td>4</td>
<td>67.8</td>
</tr>
<tr>
<td>MFCC3</td>
<td>Comb</td>
<td>6</td>
<td>67.1</td>
</tr>
</tbody>
</table>

### 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


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


