Detecting sleepiness by fusing classifiers trained with novel acoustic features

Tauhidur Rahman, Soroosh Mariooryad, Shalini Keshavamurthy, Gang Liu, John H.L. Hansen, and Carlos Busso

1Department of Electrical Engineering, University of Texas at Dallas
{txr100020, sxm096221, sxk095920, gx1083000, John.Hansen, busso}@utdallas.edu

Abstract

Automatic sleepiness detection is a challenging task that can lead to advances in various domains including traffic safety, medicine and human-machine interaction. This paper analyzes the discriminative power of different acoustic features to detect sleepiness. The study uses the sleepy language corpus (SLC). Along with standard acoustic features, novel features are proposed including functionals across voiced segment statistics in the F0 contour, likelihoods of reference models used to contrast non-neutral speech, and a set of robust to noise spectral features. These feature sets, which have performed well in other paralinguistic tasks such as emotion recognition, are used to train classifiers that are combined at the feature and decision levels. The best unweighted accuracy (UA) is obtained by combining the classifiers at the decision level under a maximum likelihood framework (UA = 70.97%). This performance is higher than the best results reported in the corpus.

Index Terms: Speaker State Recognition, Paralinguistics, Affective Computing, Sleepiness

1. Introduction

Sleepiness impairs cognitive abilities, reducing the efficiency of individuals to perform operationally relevant tasks. Long distance journey causes fatigue, strain and drowsiness even to professional drivers. According to the National Highway Traffic Safety Administration (NHTSA), over 22-24% of car accidents occur due to sleepy drivers [1]. Detecting sleepiness is also an important problem for many other domains including the study of sleep disorders, and the design of human-machine interfaces. This paper aims to detect sleepiness from acoustic features.

Several studies have reported progress toward developing sleepiness/fatigue detection system, using eye blinks [2, 3], vision based features [4, 5], and speech [6, 7, 8]. This paper is particularly interested in detecting sleepiness from speech, which can be captured from noninvasive sensors. Krajewski and Kröger used standard set of prosodic and spectral features to train artificial neural networks (ANNs) and linear discriminant analysis (LDA) [6]. They reported accuracies of 88.2% in a database collected during a sleep deprivation study. They extend their analysis by considering other features and other classifiers [7, 8]. The best performance was achieved by support vector machine (SVM) [7].

The contribution of this paper is the use of novel acoustic features for detecting sleepiness. These features were introduced in our previous work in the context of emotion recognition [9, 10]. The first set of features is estimated by contrasting speech with reference neutral models trained with Gaussian mixture models (GMMs). The likelihood scores of the models are used as features. These features perform well even with language mismatch in the training and testing sets. The second set of features corresponds to novel statistics derived from F0 contour. Speech is segmented into voiced and unvoiced segments. Local statistics from the F0 contour are estimated for each voiced segment, which are then used to derive sentence level statistics (e.g. the maximum of the pitch slopes derived from voiced segments). The third set of features corresponds to perceptual minimum variance distortionless response (PMVDR) [11] and shifted delta cepstrum (SDC) [10]. These features are robust to noisy environment [10]. For each of these feature sets, separate classifiers are trained. In addition, a baseline SVM classifier is trained with the standard acoustic features described by Schuller et al. [12]. The individual classifiers are combined at the feature and decision levels. For feature level fusion, the study compares the performance when the dimension of the feature set is reduced to different values using a chi-squared feature selection. For decision level fusion, the classifiers are combined using hard and soft labels under a maximum likelihood framework. The study uses the sleepy language corpus (SLC) [6, 8]. The best unweighted accuracies are achieved by fusing the classifiers at the decision level using soft labels (68.68% - development set, 70.97% - testing set). These accuracies outperform the best results reported for this corpus [12].

The paper is organized as follows. Section 2 introduces the database. Section 3 describes the baseline SVM system trained with standard speech features. Section 4 presents the classifiers trained with the proposed acoustic features. Section 5 describes the fusion techniques to combine different features and classifiers and their corresponding results. Section 6 concludes the paper with discussion and future directions.

2. Database

The study uses the sleepy language corpus (SLC). It consists of 21 hours of speech from 99 participants that were recorded either in a realistic car-environment or in a lecture room (9,089 turns). The speech data includes isolated vowels, read speech, commands/requests and spontaneous speech. The corpus was annotated using the Karolinska sleepiness scale by the participants (self assessment) and by two trained evaluators. The raters assigned a number between 1-extremely alert and 10-extremely sleepy. If the value was above 7.5, the sample was labeled as sleepy (SL). Otherwise, it was labeled as non-sleepy (NSL). The corpus is divided into three groups: training (≈ 40%), development (≈ 30%) and testing (≈ 30%). The datasets have similar gender proportion (57% female, 43% male). The samples for each speaker are exclusively contained in only one of the sets (speaker independent partitions). Further details about the corpus are given in references [6, 8, 12].

3. Baseline SVM system $\lambda_B$

A baseline SVM classifier is trained as reference, following a similar approach proposed by Schuller et al. [12]. A set of com-
Table 1: Sleepiness detection on the development set. $\lambda_B$ – Baseline SVM classifier (Sec.3); $\lambda_L$ – SVM classifier trained with likelihoods of reference models (Sec. 4.1); $\lambda_F$ – SVM classifier trained with functionals across voiced segment statistics (Sec.4.2); $\lambda_P$ – GMM classifier trained with PMVDR+SDC features (Sec.4.3); $\lambda_M$ – GMM classifier trained with MFCCs (Sec.4.4).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>%WA</th>
<th>%UA</th>
<th>%Recall</th>
<th>%Precision</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_B$</td>
<td>70.70</td>
<td>67.45</td>
<td>80.10</td>
<td>75.10</td>
<td>NSL</td>
</tr>
<tr>
<td>$\lambda_L$</td>
<td>66.60</td>
<td>63.95</td>
<td>74.00</td>
<td>73.20</td>
<td>NSL</td>
</tr>
<tr>
<td>$\lambda_F$</td>
<td>50.50</td>
<td>57.50</td>
<td>31.10</td>
<td>67.60</td>
<td>SL</td>
</tr>
<tr>
<td>$\lambda_P$</td>
<td>59.45</td>
<td>59.28</td>
<td>67.10</td>
<td>70.32</td>
<td>NSL</td>
</tr>
<tr>
<td>$\lambda_M$</td>
<td>61.44</td>
<td>57.32</td>
<td>64.87</td>
<td>68.72</td>
<td>NSL</td>
</tr>
</tbody>
</table>

Contrasting speech with neutral models

4.1. Contrasting speech with neutral reference models ($\lambda_L$)

We have proposed the use of neutral reference models to contrast emotional speech [9, 15]. Fig. 1 describes the general framework of the two-step approach. First, a neutral corpus is used to build robust speech models (e.g., GMM and HMM). Then, the likelihood scores are used as feature to discriminate between neutral and non-neutral speech. The implicit assumption is that acoustic features derived from non-neutral speech – sleepy speech – deviate from the patterns observed in the ones from neutral speech. Since the reference models will not properly fit non-neutral speech, it is expected that the likelihood scores will be lower. One advantage of the approach is that robust, speaker independent reference model can be built, since there are several emotionally neutral databases available. Also, the approach can capture paralinguistic information conveyed in the testing set, even when they are not properly represented in the training set, as long as they differ in any aspect from neutral speech properties. Our previous studies have shown that this approach achieves better performance than classifiers directly trained with speech features. It also generalizes better, even in the presence of language mismatch in the training and testing sets [9]. Here, we explore the benefits of using this approach in the context of sleepiness detection.

For each of the features contained in the baseline set, a neutral model is implemented with univariate GMM (4 mixtures). These reference GMMs are trained with the spontaneous sentences from the Wall Street Journal-based Continuous Speech Recognition Corpus Phase II (WSJ). Notice that there is a language mismatch between the neutral corpus (English) and the SLC (German). The likelihoods of the models are used to train SVM classifiers following the same procedure described in Sec. 3. This classifier is referred to here as $\lambda_L$. The classification results on the development set are given in Table 1. Although the performance is lower than the baseline $\lambda_B$, Section 5 indicates that they provide complementary information.

4.2. Statistics of F0 contour across voiced segments ($\lambda_F$)

In Busso et al., we proposed sentence-level F0 features derived from the statistics of the voiced regions’ patterns (see Table III in [9]). These features are estimated as follow. First, speech is segmented into voiced and unvoiced regions. Then, basic functionals such as range, maximum, quartiles, slope, curvatures and inflections are estimated from the F0 contour for each voiced segment. These values describe local statistics conveyed in the F0 contour. Then, we compute the mean, maximum and standard deviation across the functionals estimated over the voiced segments (e.g., the mean of the pitch range estimated across voiced segments). These statistics provide insights about the local dynamics of the pitch contour. For example, while the pitch range at the sentence-level gives the distance between the extreme values, the mean of the pitch range across voiced regions will indicate whether the pitch in voiced regions have flat or inflected shapes.

This study uses the same set of 17 features proposed in Busso et al. [9]. A SVM classifier is trained following the same procedure described in Sec. 3. This classifier is referred to as $\lambda_F$. The results on the development set are provided in Table 1.

4.3. GMM trained with PMVDR+SDC ($\lambda_P$)

Our previous work has shown that PMVDR features provide improvements and robustness to classifiers trained to recognize emotion [10]. PMVDR can better model the upper spectral envelope, unveiling perceptually important harmonics [11]. Unlike Mel-frequency cepstrum coefficients (MFCCs), PMVDR features do not require explicit filter-bank analysis. Furthermore, PMVDR coefficients are more robust to noise, which is important since some of the speech files are corrupted by noise. Notice that these features are LLDS.

Fig. 2 gives the block diagram to extract PMVDR features. The algorithm includes the following steps: 1) obtain the perceptually warped FFT power spectrum, 2) Compute “perceptual autocorrelations” by utilizing the IFFT on the warped power spectrum, 3) perform the $\psi^{th}$ order linear prediction (LP) analysis via Levinson-Durbin recursion using perceptual autocor-
Figure 2: Block diagram to extract PMVDR features.

Table 2: Q-statistics for pairwise comparison of similarity between classifiers.

<table>
<thead>
<tr>
<th></th>
<th>$\lambda_L$</th>
<th>$\lambda_B$</th>
<th>$\lambda_P$</th>
<th>$\lambda_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_L$</td>
<td>1</td>
<td>0.8689</td>
<td>0.2451</td>
<td>0.3705</td>
</tr>
<tr>
<td>$\lambda_B$</td>
<td>-</td>
<td>1</td>
<td>0.1854</td>
<td>0.4904</td>
</tr>
<tr>
<td>$\lambda_P$</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.0322</td>
</tr>
<tr>
<td>$\lambda_M$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Sleepiness classification with feature level fusion. Subsets of features are selected with chi-squared feature selection.

<table>
<thead>
<tr>
<th>Features #</th>
<th>% WA</th>
<th>% UA</th>
</tr>
</thead>
<tbody>
<tr>
<td>7812(all feature)</td>
<td>66.20</td>
<td>65.25</td>
</tr>
<tr>
<td>5000</td>
<td>68.20</td>
<td>67.90</td>
</tr>
<tr>
<td>3000</td>
<td>68.00</td>
<td>67.40</td>
</tr>
<tr>
<td>1000</td>
<td>63.30</td>
<td>64.00</td>
</tr>
</tbody>
</table>

4.4. GMM trained with MFCCs ($\lambda_M$)

The study also considers a GMM classifier trained with standard MFCCs (12 coefficients plus their delta and delta-delta values). While the baseline classifier $\lambda_B$ uses functionals derived from MFCCs, this classifier processes the feature vector frame by frame. We used the same window and overlapping rates that are used for PMVDR coefficients. This classifier is referred to here as $\lambda_M$. The results for the development set are shown in Table 1.

5. Fusion and Experiment Results

Table 1 reveals that the proposed classifiers have different confusion matrices. For example, Table 1 indicates that the recall for SL is approximately 84% for $\lambda_F$, which is higher than the recall achieved by the baseline $\lambda_B$ (55%). The implication is that $\lambda_F$ is more likely to correctly recognize SL samples than $\lambda_B$. The difference in performance between the classifiers is also observed in Table 2, which gives the Q-statistics for each pair of classifiers [16]. This statistic gives a measure between -1 and 1 describing the similarity between the outputs of 2 classifiers (the higher the absolute value, the more dependent the classifiers are). In general, the values in Table 2 are low. Given that different classifiers provide and model different, and hopefully, complementary information, we decide to compare feature level fusion and decision level fusion.

5.1. Feature level fusion

To study the performance achieved with feature level fusion, all the sentence-level features are combined to form a large set with 7812 features (baseline features, likelihood feature and 500 F0 statistics). In addition, we compare the performance when the dimension of the feature set is reduced to different values using a chi-squared feature selection technique. In each of these cases, a SVM is built with the training data (complexity $c = 0.005$). Table 3 shows the results on the development set. The best UA is achieved when the feature set is reduced to 3000 features (67.40%). This classifier does not provide any improvement compared to the baseline (67.45%).

5.2. Decision level fusion

This section explores fusing the classifiers at the decision level. The proposed approach is based on maximum likelihood criteria. Given $n$ different classifiers, $\lambda_i$, $i \in \{1, \ldots, n\}$, the goal is to infer the true class label $\omega \in \{SL, NSL\}$. If the classifier $\lambda_i$ predicts the class $\omega_i$, the optimal decision ($\hat{\omega}$), based on maximum likelihood criteria is given by Equation 1. These probabilities are estimated using hard and soft decisions.

$$\hat{\omega} = \arg\max_{\omega_i} P(\omega_1, \omega_2, ..., \omega_n | \omega = \omega_i)$$

5.2.1. Fusion with hard decision labels

With hard decision, $\omega_i, \omega_i \in \{SL, NSL\}$. The values of the numerator and denominator in Equation 1 are estimated from the results of the individual classifiers on the development set. The probability in the numerator is estimated by counting the joint frequency of the classifiers’ outputs for each class. The class distribution probability in the denominator is estimated by computing the frequency of each class in the development set. This probability serves as a normalization factor to avoid bias produced by unbalanced classes.

5.2.2. Fusion with soft decision labels

The confidence measure (probability) of each classified sample is potentially more informative than the recognized class labels (binary result). Replacing the recognized labels ($\omega_i$) in Equation 1 with the corresponding probability for one particular class (e.g., SL), results in decision level fusion with soft decision. A Gaussian distribution is built on the probabilities of the classifiers to estimate the conditional distribution $P(\omega_1, \omega_2, ..., \omega_n | \omega = \omega_i)$. During inference, the class with maximum likelihood is selected.

5.2.3. Decision level results on development set

For the results reported in this section, a 10-fold cross-validation approach is implemented to split the development set (30 subjects). Data from 27 subjects is used to estimate the probabilities in Equation 1. Data from the remaining 3 subjects is used for testing the accuracies (speaker independent results). This approach is repeated for each fold. The reported accuracies correspond to the average performance across all subjects.
Table 4: Sleepiness classification with decision level fusion. Reported values are UAs and WAs for the development set. The values are the average across all subjects in 10-fold cross-validation experiments.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Hard</th>
<th>Soft</th>
</tr>
</thead>
<tbody>
<tr>
<td>B, L, M, L'</td>
<td>70.69</td>
<td>68.22</td>
</tr>
<tr>
<td>B, L, L', M</td>
<td>70.40</td>
<td>67.91</td>
</tr>
<tr>
<td>L, M, L', M</td>
<td>70.15</td>
<td>67.66</td>
</tr>
<tr>
<td>L, M, L, M</td>
<td>69.86</td>
<td>67.37</td>
</tr>
<tr>
<td>M, L, M, M</td>
<td>69.57</td>
<td>66.98</td>
</tr>
<tr>
<td>M, L, L, M</td>
<td>69.28</td>
<td>66.69</td>
</tr>
<tr>
<td>L, M, M, M</td>
<td>68.99</td>
<td>66.40</td>
</tr>
<tr>
<td>M, M, M, M</td>
<td>68.70</td>
<td>66.11</td>
</tr>
</tbody>
</table>

Table 4 shows the performance for different combinations (e.g., B, L, M, L’ denotes the combination of B, L, M, and L’). The table shows 11 combinations that achieve better performance than the baseline classifier (67.45%, Table 1). The highest UA is obtained by combining B, L, M, and L’ with soft decisions. This configuration is selected to validate the approach in the testing set (Sec. 5.2.4). Although the classifier L’ is not in this set, Table 4 indicates that incorporating this classifier in some cases improves the overall performance (e.g., B, L, M, L’).

5.2.4. Decision level results on testing set

In this section, the testing set is used to validate the accuracies of the selected classifier (decision level fusion of B, L, M, and L’ with soft decisions). The entire development dataset is used to estimate the probabilities in Equation 1. Table 5 shows the results. The UA is higher than the best result reported for this corpus [12]. Likewise, the proposed classifier provides a 1.07% (absolute) improvement for WA.

6. Conclusions

This paper describes our efforts to detect sleepiness by using novel acoustic features. Different classifiers are trained with these sets of features, which are fused at the feature and decision levels. For decision level fusion, hard and soft decisions from individual classifiers are combined using maximum likelihood criterion. The best performance in term of UA is achieved with decision level fusion using soft decisions (68.86% – development set, 70.97% – testing set). These accuracies outperform the best results previously reported for this corpus.

As part of our future work, we will investigate the benefits of using gender dependent models for the classifiers. During our preliminary experiments, we separately trained the baseline classifier λB for female and male speakers. We noticed significant differences in their UAs (77.70% – male, 62.35% – female). This preliminary result suggests that gender dependent models may improve the overall performance of the system.

7. References