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

[1]Department of Electrical Engineering, University of Texas at Dallas
{txr100020, sxm096221, sxk095920, gxl083000, 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 nonintrusive 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 ($\sim$ 40%), development ($\sim$ 30%) and testing ($\sim$ 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 (λ_B)

A baseline SVM classifier is trained as reference, following a similar approach proposed by Schuller *et al.* [12]. A set of com-

28 – 31 August 2011, Florence, Italy

Table 1: Sleepiness detection on the development set. λ_B – Baseline SVM classifier (Sec.3); λ_L – SVM classifier trained with likelihoods of reference models (Sec. 4.1); λ_F – SVM classifier trained with functionals across voiced segment statistics (Sec.4.2); λ_P – GMM classifier trained with PMVDR+SDC features (Sec.4.3); λ_M – GMM classifier trained with MFCCs (Sec.4.4).

Classifier	%WA	%UA	%Recall	%Precision	Class
λ_B	70.70	67.45	80.10	75.10	NSL
			54.80	58.10	SL
λ_L	66.60	63.95	74.00	73.20	NSL
			53.90	54.90	SL
λ_F	50.60	57.50	31.10	76.60	NSL
			83.90	41.70	SL
λ_P	59.45	59.28	67.10	70.32	NSL
			51.81	48.07	SL
λ_M	61.44	57.32	64.87	68.72	NSL
			49.77	45.43	SL

monly used acoustic and prosodic features in various speech processing tasks are extracted using openSMILE. This package is the backend of Emotion and Affect Recognition (openEar) toolkit [13]. The feature set includes 59 *Low-level descriptors* (LLDs) related to energy (4), spectral features (50) and voiced related features (5). 33 base functionals and 6 F0 functionals are estimated from the LLDs, producing 4,368 sentence level features. 431 features had constant values across sentences, so they were removed from the set. The readers are referred to Schuller *et al.* [12] for more details about the features.

A linear kernel *Support Vector Machine* (SVM) with *Sequential Minimal Optimization* (SMO) is used as classifier. The SVM is trained and tested with the WEKA data mining toolkit [14], using all the features. The *synthetic minority oversampling technique* (SMOTE) is employed to compensate unbalanced classes in the training set. This baseline classifier is referred to here as λ_B. The *complexity* parameter of the classifier, c, is optimized on the development set, by maximizing the *Unweighted Accuracy* (UA) (i.e., the unweighted average recall). For $c = 0.02$, the SVM classifier achieves the highest UA. Table 1 gives its performance in terms of UA, *weighted accuracy* (WA), precision and recall.

4. Systems trained with proposed features

4.1. Contrasting speech with neutral reference models (λ_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

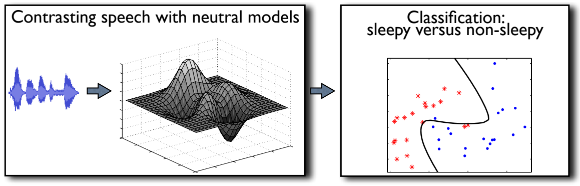

Figure 1: Neutral model based likelihood features [9, 15].

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 Section 3. This classifier is referred to here as λ_L. The classification results on the development set are given in Table 1. Although the performance is lower than the baseline λ_B, Section 5 indicates that they provide complementary information.

4.2. Statistics of F0 contour across voiced segments (λ_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 λ_F. The results on the development set are provided in Table 1.

4.3. GMM trained with PMVDR+SDC (λ_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 i^{th} order *linear prediction* (LP) analysis via Levinson-Durbin recursion using perceptual autocor-

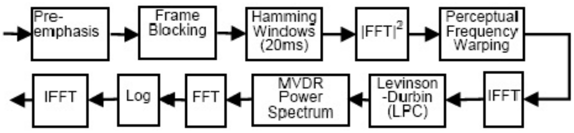

Figure 2: Block diagram to extract PMVDR features.

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

	λ_L	λ_B	λ_F	λ_P	λ_M
λ_L	1	0.8689	0.2451	0.3705	0.7911
λ_B	-	1	0.1854	0.4904	0.8272
λ_F	-	-	1	0.0322	0.4754
λ_P	-	-	-	1	0.4360
λ_M	-	-	-	-	1

relation lags, 4) calculate the i^{th} order MVDR spectrum from the LP coefficients, and 5) obtain the final cepstrum coefficients using the straightforward FFT-based approach.

The study uses a 10-dimensional PMVDR feature vector. The analysis window is set to 25 ms with 15 ms of overlap. Cepstral mean normalization (CMN) is applied to the final feature vector. Previous studies have showed that the SDC operation can incorporate additional temporal information into the feature vector to produce better performance [10]. This study employs the same strategy. A GMM is trained to process the data frame by frame. The normalized sum of the likelihoods is used to classify SL versus NSL at the sentence level. This classifier is referred to here as λ_P. The results for the development set are shown in Table 1.

4.4. GMM trained with MFCCs (λ_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 λ_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 λ_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 λ_F, which is higher than the recall achieved by the baseline λ_B (55%). The implication is that λ_F is more likely to correctly recognize SL samples than λ_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 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%).

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

Features #	%WA	%UA
7812(all feature)	66.20	65.25
5000	68.20	66.90
3000	68.00	67.40
1000	63.30	64.00

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, λ_i, $i \in \{1, \ldots, n\}$, the goal is to infer the true class label $\omega \in \{SL, NSL\}$. If the classifier λ_i predicts the class ω_{λ_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.

$$\begin{aligned}\hat{\omega} &= \underset{\omega_\theta}{\operatorname{argmax}}\ P(\omega_{\lambda_1}, \omega_{\lambda_2}, \ldots, \omega_{\lambda_n} | \omega = \omega_\theta) \\ &= \underset{\omega_\theta}{\operatorname{argmax}}\ \frac{P(\omega_{\lambda_1}, \omega_{\lambda_2}, \ldots, \omega_{\lambda_n}, \omega = \omega_\theta)}{P(\omega = \omega_\theta)} \qquad (1)\end{aligned}$$

5.2.1. Fusion with hard decision labels

With hard decision, $\omega_{\lambda_i}, \omega_\theta \in \{SL, NSL\}$. 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 (ω_{λ_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 $P(\omega_{\lambda_1}, \omega_{\lambda_2}, \ldots, \omega_{\lambda_n} | \omega = \omega_\theta)$. 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.

Classifiers	Decision Fusion			
	Hard		Soft	
	%WA	%UA	%WA	%UA
λ_B,λ_L	69.33	67.53	65.80	64.06
λ_B,λ_F	70.70	67.42	68.30	66.53
λ_B,λ_M	70.70	67.42	70.77	68.30
λ_B,λ_P	70.70	67.42	69.67	67.86
λ_L,λ_F	66.11	63.29	68.06	67.33
λ_L,λ_M	61.06	60.97	66.96	64.58
λ_L,λ_P	66.55	63.95	67.82	65.48
λ_F,λ_M	64.77	63.80	61.65	64.09
λ_F,λ_P	64.19	60.37	55.71	59.65
λ_M,λ_P	62.13	62.69	62.92	64.05
λ_B,λ_L,λ_F	68.89	68.23	67.07	66.56
λ_B,λ_L,λ_M	69.47	68.14	69.06	67.77
λ_B,λ_L,λ_P	70.36	68.27	68.92	68.22
λ_B,λ_F,λ_M	68.92	66.67	69.26	67.17
λ_B,λ_F,λ_P	68.82	66.57	68.27	66.77
λ_B,λ_M,λ_P	69.47	67.03	70.63	68.34
λ_L,λ_F,λ_M	64.70	65.54	66.04	64.54
λ_L,λ_F,λ_P	63.91	64.01	66.79	65.10
λ_L,λ_M,λ_P	66.21	65.09	67.41	65.05
λ_F,λ_M,λ_P	64.77	64.43	62.50	63.62
λ_B,λ_L,λ_F,λ_M	68.30	67.19	68.40	67.43
λ_B,λ_L,λ_F,λ_P	70.05	68.07	67.86	67.24
λ_B,λ_L,λ_M,λ_P	70.12	68.22	**70.15**	**68.68**
λ_B,λ_F,λ_M,λ_P	68.85	66.67	69.06	67.43
λ_L,λ_F,λ_M,λ_P	64.08	64.93	66.07	65.12
λ_B,λ_L,λ_F,λ_M,λ_P	68.82	67.07	68.92	67.72

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

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 λ_B, λ_L, λ_M and λ_P 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 (68.68% – 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 λ_B 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.

Table 5: Sleepiness classification with decision level fusion with soft labels for the test set.

$Classifier$	$\%WA$	$\%UA$
Schuller, et al.[12]	73.00	70.30
λ_B,λ_L,λ_M,λ_P	74.07	70.97

7. References

[1] S. Klauer, T. Dingus, V. Neale, J. Sudweeks, and D. Ramsey, "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," National Highway Traffic Safety Administration, Tech. Rep., 2006.

[2] N. Galley, R. Schleicher, and L. Galley, *Blink parameters as indicators of driver's sleepiness-possibilities and limitations*. Elsevier, Amsterdam, 2004, vol. X, pp. 189–196.

[3] P. Caffier, U. Erdmann, and P. Ullsperger, "Experimental evaluation of eye-blink parameters as a drowsiness measure," *European Journal of Applied Physiology*, vol. 89, pp. 319–325, 2003.

[4] I. Garcia, S. Bronte, L. Bergasa, N. Hernandez, B. Delgado, and M. Sevillano, "Vision-based drowsiness detector for a realistic driving simulator," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, Sept 2010, pp. 887–894.

[5] D. Sandberg, T. Akerstedt, A. Anund, G. Kecklund, and M. Wahde, "Detecting driver sleepiness using optimized nonlinear combinations of sleepiness indicators," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 1, pp. 97–108, March 2011.

[6] J. Krajewski and B. Kröger, "Using prosodic and spectral characteristics for sleepiness detection," in *INTERSPEECH*, 2007, pp. 1841–1844.

[7] J. Krajewski, A. Batliner, and R. Wieland, "Multiple classifier applied on predicting microsleep from speech," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, Dec 2008, pp. 1–4.

[8] 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.

[9] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[10] G. Liu, Y. Lei, and H. Hansen, "A novel feature extraction strategy for multi-stream robust emotion identification," in *Interspeech 2010*, Makuhari, Chiba, Japan, September 2010, pp. 482–485.

[11] U. H. Yapanel and J. Hansen, "A new perceptually motivated mvdr-based acoustic front-end(pmvdr) for robust automatic speech recognition," *Speech Communication*, vol. 50, pp. 142–152, 2008.

[12] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Interspeech 2011*. Florence, Italy: ISCA, August 2011.

[13] F. Eyben, M.Wöllmer, and B. Schuller, "openEAR-introducing the munich open-source emotion and affect recognition toolkit," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009, pp. 576–581.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, November 2009.

[15] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.

[16] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, pp. 181–207, May 2003.