Proceedings of the 19th International Conference on Pattern Recognition
(ICPR 2008), Tampa, 2008. 4 pages, no pagination

Multiple Classifier Applied on Predicting Microsleep from Speech

Jarek Krajewski [1], Anton Batliner [2], Rainer Wieland [1]
[1] *Work and Organizational Psychology, Univ. of Wuppertal, Germany*
[2] *Lehrstuhl fuer Mustererkennung, Univ. of Erlangen-Nuremberg, Germany*
{krajewsk, wieland}@uni-wuppertal.de, batliner@informatik.uni-erlangen.de

Abstract

The aim of this study is to apply a state-of-the-art speech emotion recognition engine on the detection of microsleep endangered sleepiness states. Current approaches in speech emotion recognition use low-level descriptors and functionals to compute brute-force feature sets. This paper describes a further enrichment of the temporal information, aggregating functionals and utilizing a broad pool of diverse elementary statistics and spectral descriptors. The resulting 45,088 features were applied to speech samples gained from a car simulator based sleep deprivation study. After a correlation-filter based feature subset selection, which was employed on the feature space in an attempt to maximize relevance, several classification models were trained. The best model (Support Vector Machine, dot kernel) achieved 86.1% recognition rate in predicting microsleep endangered sleepiness stages

1. Introduction

Little empirical research has been done to examine the effect of microsleep endangered sleepiness states [13] on acoustic voice characteristics. Most studies have analyzed only single features [7,16] or small feature sets containing only perceptual acoustic features, whereas signal processing based speech and speaker recognition features (e.g. MFCCs) have received little attention [8,9,11]. Thus, the aim of this study is to apply a state-of-the-art speech emotion recognition engine [2,3,10,14] on the detection of critical sleepiness states. Attention is drawn particularly on the computation of a 45k feature set using low-level descriptors (LLDs) and their temporal information aggregating functionals.

The rest of this paper is organized as follows: In Section 2 the procedure of computing low-level descriptors and functionals are explained. Section 3 describes the design of the sleep deprivation study used for building a sleepy speaker database. After the results of the sleepiness detection are provided in Section 4, the paper closes with a conclusion and a discussion of the future work in Section 5.

2. Brute-force feature extraction

The acoustic features (low-level descriptors, LLDs) can be computed for each single speech signal frame, and connected to raw contours. This procedure results in speech feature contours as e.g. the fundamental frequency contour or the bandwidth of formant 4 contour. In detail, the following LLDs are often chosen: fundamental frequency, intensity, harmonics-to-noise ratio, formant 1-6 (amplitude, position and bandwidth), MFCCs, LFCCs, duration of voiced/unvoiced speech segments, spectral features as band-energies, roll-off, centroid or flux, wavelets based features and long term average spectrum (LTAS). The next processing step captures temporal information on the acoustic contours (LLDs) by computing functionals.

Frequently used functionals are percentiles (quartiles, quartile ranges, and other percentiles), extremes (min/max value, min/max position, range), distributional functions (number of segments/ intervals/reversal points), spectral functionals (DCT coefficients), regression functions (intercept, error, regression coefficients), higher statistical moments (standard deviance, skewness, kurtosis, length, and zerocrossing-rate), means (arithmetic mean and centroid), and sequential and combinatorial functionals: a minimum of two functionals has to be applied in either a sequential way (e.g. max of regression error) or combinational way (e.g. ratio of mean of two different LLD).

Proceedings of the 19th International Conference on Pattern Recognition
(ICPR 2008), Tampa, 2008. 4 pages, no pagination

3. Experimental Method

3.1. Database

Twelve students, recruited from the University of Applied Sciences, Schmalkalden, Germany, volunteered in taking part in this study. Initial screening excluded those having severe sleep disorders or sleep difficulties. The participants were instructed to maintain their normal sleep pattern and behaviour. Due to recording and communication problems, the data of 2 participants could partly not be analyzed (4 speech samples). We conducted a within-subject sleep deprivation design (01.00 - 08.00 a.m). During the night of sleep deprivation a well established, standardised self-report sleepiness measure, the Karolinska Sleepiness Scale (KSS) [1], was used by the subjects and the two experimental assistants almost every hour just before the speech recordings. In the version used in the present study, scores range from 1 to 10 (extremely alert =1; sleepy, but no effort to stay awake =7; sleepy, but some effort to stay awake =8; very sleepy, great effort to stay awake =9; extremely sleepy, can't stay awake =10). Given the verbal descriptions, scores of 8 and higher appear to be most relevant from a practical perspective as they describe a state in which the subject feels unable to stay awake. During the night, the subjects were confined to the laboratory, conducting a driving simulator task and were supervised throughout the whole period.

The recording took place in a laboratory room with dampened acoustics using a high-quality, clip-on microphone (sampling rate: 44.1 kHz, 16 bit). The input level of the sound recording was kept constant throughout the recordings. Furthermore the subjects were given sufficient prior practice so that they were not uncomfortable with this procedure. The verbal material was taken from formulaic pilot-air traffic controller communication: "Cessna nine three four five lima, county tower, runway two four in use, enter traffic pattern, report left base, wind calm, altimeter three zero point zero eight". The participants recorded other verbal material at the same session, but in this article we focus on the material described above. For training and classification purposes, the records were further divided into two classes: alert (A) and microsleep endangered sleepy (MS) with the microsleep validated boundary value KSS $\geq$ 7.5 (8 samples per subject; total number of speech samples: 94 samples; 34 samples A, 60 samples MS; KSS:= mean of the three KSS-Ratings; M= 7.22; SD= 2.87). As described above, the Acoustic Sleepiness Analysis follows a speech adapted pattern recognition approach: (a) recording speech, (b) preprocessing, (c) feature extraction, (d) dimensionality reduction, (e) classification, and (f) validation.

3.2. Feature extraction

All acoustic measurements were taken utterance-wise using the Praat speech analysis software for computing the LLDs [4]. As mentioned above we estimated the following 58 LLDs: fundamental frequency, fundamental frequency peak process, intensity, harmonics-to-noise ratio, formant position and bandwidth (F1-F6), 15 LPCs, 12 MFCCs, 12 LFCCs, duration of voiced, duration of unvoiced speech segments and long term average spectrum (LTAS). These 58 LLDs are joined by their first and second derivates (velocity and acceleration contours). Furthermore these 174 speech feature contours are modeled in average by 129.56 functionals in time and frequency domain feature space.

(i) functionals from elementary statistics *(time domain)*: min, max, range, mean, median, trimmed mean 10%, trimmed mean 25%, 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentile, interquartil range, mean average deviation, standard deviation, skewness, kurtosis, robust regression coefficients, intercept, frequency of values beyond different threshold (median +/- 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0*median), min and max position, relative min and max position; entropy, number of peaks, mean standard deviation, min and max of peak position, peak amplitude value, delta peak position, and delta peak amplitude.

(ii) functionals from the *spectral domain:* spectral envelope (regression coefficient, intercept), power spectral density of 5 frequency bands, relative power, maximum within 5 frequency bands.

This procedure of combining LLDs and functionals results in 22,544 raw features. To take individual response patterns into account, we added the same amount of speaker normalized features (differences between raw feature vectors and the speaker specific mean of this feature vector). In sum, we computed a total amount of 45,088 features per speech sample.

3.3. Feature selection and classification

The purpose of feature selection is to reduce the dimensionality, which can otherwise hurt the performance of the pattern classifiers. The small amount of data also suggested that longer vectors would not be advantageous due to overlearning of data.

In this study, we used a rather relevance maximizing then redundancy minimizing correlation filter approach (pearson correlation >.40) [17].

For the classification we used a Support Vector Machine (SVM; dot kernel function), a Multilayer Perceptron (MLP; feedforward net, backpropagation, 2 hidden sigmoid layer, 5 nodes each), a k-Nearest Neighbour (KNN; k = 1, 2, or 3), a Decision Tree, a Random Forest, a Naive Bayes, a Basic Rule Learner, a Radial Basis Function (RBF), a Logistic Base, a Fuzzy Lattice Reasoning and a Logistic Regression. Specifically SVM have proven to best model static acoustic feature vectors [14] and were therefore chosen and computed with Matlab software. Due to data sparcity, a speaker-dependent approach has been chosen, a leave-one-sample-out cross-validation, i.e. in turn, one case was used as test set and all other as train. The final classification errors were calculated averaging over all classifications.

4. Results

In order to determine the multivariate prediction performance, different classifiers were applied on the 230 features remaining after the correlation-filter procedure (among the selected features we found e.g. the plausible results of reduced fundamental frequency and reduced formant 1 values for sleepy speaker; r =-.42. resp. r =-.35). For all configurations, we trained the classifier and applied them on the test sets.

The averaged recognition rates (RR = ratio correctly classified samples divided by all samples, and CL = class-wise averaged classification rate) of the different classifiers for the two class prediction problems are: SVM (86.1/82.8), MLP (80.9/79.3), 1-NN (73.4/70.3), 2-NN (62.8/69.5), 3-NN (76.6/72.1), DT (75.5/70.6), Random Forest (68,1/62.9), Naïve Bayes (73.4/70.9), Basic Rule Learner (71.3/71.7), RBF (72.3/68.2), Logistic Base (86.1/82.4), Fuzzy Lattice Reasoning (75.5/75.1) and Logistic Regression (86.2/82.4). The SVM prediction achieved the highest class-wise averaged classification rate, which reached significance compared to a pure chance based classification (χ^2 =45.5; df =1; p <.001), and was therefore applied for further detailed LLD based analyses. The results are depicted in Table 1.

5. Discussion

The most important LLD feature classes for this prediction were according to (a) the sum of features remaining the correlation-filter: LFCCs, LPCs, and duration of voiced/unvoiced; according to (b) the prediction accuracy of the single LLD feature class: Formants, F0, and LFCCs. Using all LLDs we achieved on this two-class classification problem a recognition rate of over 86% on unseen but speaker dependent data with a Support Vector Machine classifier. Our classification performance is in the same range as has been obtained for comparable tasks, e.g. for emotional user state classification, cf. [2,3,8,9,11,12,14].

Table 1: *Recognition rates (RR) and class-wise averaged classification rate (CL) (in %) on the test set using different LLDs feature sets (raw and speaker normalized features surviving the correlation-filter; # = number of features) on the SVM classifier.*

LLDs	**Raw**			**Raw & Normalized**		
	#	**RR**	**CL**	**#**	**RR**	**CL**
Formants	2	71.3	65.4	8	86.2	82.8
F0	2	72.3	68.1	3	78.7	75.7
LFCCs	18	73.4	70.3	72	77.7	72.9
MFCCs	5	72.3	67.5	19	74.5	69.2
LPCs	14	74.5	71.1	67	70.2	65.8
HNR/ Int	11	70.2	65.8	20	66.0	60.0
Duration	1	64.9	57.8	39	64.9	56.0
LTAS	0	-	-	2	67.0	54.4
All LLDs	*53*	*70.2*	*65.8*	*230*	*86.1*	*82.8*

Our results are limited by several facts. The present results are preliminary and need to be replicated using a natural speech environment: it would seem advisable that future studies address the main topics of improving the acoustic sleepiness analysis and finding evidence for its validity in real-world applications. Thus, collecting sleepy speech samples from different types of speakers and real-life speech situations would provide the infrastructural research background that enhances further progress in acoustic sleepiness analysis. Emotion and stress speech databases (e.g. EMO-DB [5]), could serve as a model for this kind of open source speech corpora. For further improvement of the acoustic sleepiness analysis, the following issues have to be addressed: (a) The computation of signal processing features derived from state space domains as e.g. average angle or length of embedded space vectors, and recurrence quantification analyses should

be computed, and feature transformation applied [15]. In addition, different normalization procedures could be applied as, e.g. computing speaker specific baseline corrections not on high-level features but on duration adapted low-level contours. (b) For finding the optimal feature subset, further supervised filter based subset selection methods (e.g. IGA) or supervised wrapper-based subset selection methods, should be applied (e.g. sequential forward floating search). (c) A third class should be added to the classification task serving as a warning stage within a microsleep detection system.

References

[1] T. Åkerstedt & M. Gillberg. Subjective and objective sleepiness in the active individual, International Journal of Neuroscience, 52, 29-37. 1990

[2] A. Batliner, C. Hacker, S. Steidl, E. Noeth, S. D'Arcy, M. Rusell, M. Wong. "You stupid tin box" – Children interacting with the AIBO robot: A crosslinguisitc emotional speech corpus. Proc. LREC 2004, pp 171-174, 2004.

[3] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L., L. Vidrascu, N. Amir, L. Kessous, V. Aharonson. Combining Efforts for Improving Automatic Classification of Emotional User States. In Erjavec, T. & Gros, J.Z. (Eds.): Language Technologies, IS-LTC 2006, Ljubljana, Slovenia, pp 240-245, 2006.

[4] P. Boersma. PRAAT, a system for doing phonetics by computer, Glot International, vol. 5, no. 9/10, pp. 341–345, 2001.

[5] F. Burkhardt, F., A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss. A Database of German Emotional Speech, Proceedings of Interspeech, 1517-1520, 2005

[6] M. Golz, D. Sommer, M. Chen, U. Trutschel, D. Mandic. Feature Fusion for the Detection of Microsleep Events. J VLSI Signal Proc Syst, 49, 329-342, 2007

[7] Y.Harrison, J.A Horne. Sleep deprivation affects speech, Sleep, 20, 871-877, 1997.

[8] J. Krajewski, B. Kröger. Using prosodic and spectral characteristics for sleepiness detection. Proc. Interspeech, 1841-1844, 2007.

[9] J. Krajewski, R. Wieland, A. Batliner. An acoustic framework for detecting fatigue in human-computer interaction. In K. Miesenberger, J. Klaus, W. Zagler, A. Karshmer (Eds.), Computers Helping People with Special Needs (pp. 54-61). Heidelberg: Springer, 2008

[10] I. Mierswa, K. Morik. Automatic feature extraction for classifying audio data. Machine Learning Journal, 58, 127-149, 2005.

[11] T.L. Nwe, H. Li, M. Dong. Analysis and Detection of Speech under Sleep Deprivation, Proc. Interspeech 17-21, 2006.

[12] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, V. Aharonson. The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. Proc. Interspeech, 2253-2256, 2007.

[13] D. Sommer, M. Chen, M. Golz, U. Trutschel, D. Mandic. Fusion of State Space and Frequency Domain Features for Improved Microsleep Detection. In W. Dutch et al. (Eds.) Int Conf Artifical Neural Networks (ICANN 2005), pp. 753-759. Springer: Berlin, 2005.

[14] B. Vlasenko, B. Schuller, A. Wendemuth, G. Rigoll. Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. Proc., Interspeech, 2249-2252, 2007.

[15] C.L. Webber, J.P. Zbilut. Dynamical assessment of physiological systems and states using recurrence plot strategies, Journal of Applied Physiology 76, 1994.

[16] J. Whitmore, S. Fisher. Speech during sustained operations, Speech Communication 20, 55-70, 1996.

[17] I.H. Witten, E. Frank. Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann (Ed.), San Francisco, 133 ff., 2005.