

Genetic Algorithm Based Feature Selection Applied on Predicting Microsleep from Speech

J. Krajewski¹, M. Golz², D. Sommer² and R. Wieland¹

¹ Work and Organizational Psychology, Univ. of Wuppertal, Wuppertal, Germany

² Neuroinformatics and Signal Processing, Univ. of Applied Sciences Schmalkalden, Schmalkalden, Germany

Abstract—Within this study we apply a 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 an usually large feature set (45k) utilizing a broad pool of diverse elementary statistics and spectral descriptors. Several (un-)supervised subset selection methods including genetic algorithm based methods were employed on the feature space in an attempt to prune redundant dimensions. The resulting dimensionality reduced feature space was applied to speech samples gained from a car simulator based sleep deprivation study (N=12; 01.00-08.00 a.m.). Among the tested dimensionality reduction methods a simple correlation filter approach (130 features remaining) reached the best recognition rate (85.1%, SVM) in predicting microsleep endangered sleepiness stages.

Keywords— Acoustic Features, Sleepiness Detection, Feature Selection, Genetic Algorithm.

I. INTRODUCTION

Little empirical research has been done to examine the effect of microsleep endangered sleepiness states [1,2] on acoustic voice characteristics. Most studies have analyzed only single features [3,4] 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 [5-7]. Thus, the aim of this study is to apply a state-of-the-art speech emotion recognition engine [8,9] 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 (LLDs) 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 and 6.

II. BRUTE-FORCE FEATURE EXTRACTION

The trend in speech emotion recognition is towards a thorough (“brute-force”) exploitation of the feature space, resulting in hundreds or even thousands of features used for classification [8,9]. The signal processing, speaker recognition and speech recognition based 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 low level descriptors (LLDs) were often chosen: fundamental frequency, energy (model intensity, based on the amplitude in different intervals), harmonics-to-noise ratio, formant 1-6 (represent spectral maxima, and are known to model spoken content and speaker characteristics), MFCCs (have been proven beneficial in speech emotion recognition, and speech recognition tasks; homomorphic transform with equidistant band-pass-filters on the Mel-scale), LFCCs (emphasise changes or periodicity in the spectrum, while being relatively robust against noise), duration of voiced/unvoiced speech segments (model temporal speech rhythm aspects), spectral features as band-energies, roll-off, centroid or flux, wavelets based features and long term average spectrum (LTAS; averages out formant information, giving general spectral trends).

The next processing step projects the values of the univariate time series (LLD), onto a scalar feature x , which captures temporal information of the acoustic contour (LLD). An important advantage of this sequential approach is the improved ability to model the contribution of smaller units (word) and larger units (chunks) within the prosodic structure of an utterance. 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: a minimum of two functionals has to be applied in either a sequential way (e.g. max of regression error) or combinatorial way (e.g. ratio of mean of two different LLD) [10].

This computationally demanding features extraction procedure results usually in a huge number of features (<1k) and a comparatively small number of samples. This problem is well known as curse of dimensionality and can impair the reliable classification. Thus, the optimization of high dimensional feature spaces seems a must in view of performance and real-time-capability. Optimization can be performed by (un-)supervised feature subset selection (e.g. correlation filter based or genetic algorithm based wrapper selection) and (un-)supervised feature transformation methods (e.g. Principal Component Analysis, Single Value Decomposition, Linear Discriminant Analysis). As genetic algorithm based optimization methods are promising in terms of performance and real-time-capability, they are explained in detail.

Genetic algorithm based feature subset selection. Genetic algorithms (GA) are a method for solving optimization inspired by biological processes of mutation, natural selection and genetic crossover. The GA is a powerful feature selection tool, especially when the dimensions of the original feature set are large. Starting from an initial candidate population of chromosomes (or sets of parameters to be optimized), operators mimicking the biological ones of crossover and mutation to select and reproduce fittest solutions, which is given by a scoring function. Basically, mutation enables the algorithm to explore new regions of the search space by randomly altering some genes (components) of some chromosomes in the population. On the other hand, crossover reinforces prior successes by recombining parent-chromosomes so as to produce fittest offsprings. GA can play an important role in selecting successful feature subsets and even generating new features by combining old features with elementary operators (e.g. +, -, /, *, log, x^{-1}) using again the GA operators of selection, crossover, mutation, generation, and evaluation. In sum, GA does not try to provide an exact match but an approximation of the optimal solution within an acceptable tolerance, which improve their effectiveness [11].

III. MATERIALS AND METHODS

A. Procedure, subject and speech material

Twelve students took part in this study voluntarily. 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) [12], 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, neither alert nor sleepy =5, 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). Furthermore the subjects were given sufficient prior practice so that they were not uncomfortable with this procedure. The verbal material consisted of a simulated 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; $M = 7.22$; $SD = 2.87$).

B. Feature extraction

All acoustic measurements were taken utterance-wise using the Praat speech analysis software for computing the LLDs [13]. 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 derivatives (velocity and acceleration contours). Furthermore these 174

speech feature contours are described in average by 129 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%, 10th, 25th, 75th, and 90th percentile, interquartil range, mean average deviation, standard deviation, skewness, kurtosis, robust regression coefficients, intercept, frequency of values beyond different threshold, 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 *spectral domain*: spectral envelope (regression coefficient, intercept), power spectral density of 5 frequency bands, relative power, max 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

C. Dimensionality reduction and classification

The purpose of dimensionality reduction is to reduce the decorrelate the feature space, 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 combined filterbased subset selection method consisting of high classcorrelation (maximizing relevance; filter criteria: Pearson correlation $>.40$ with KSS) and low inter-feature correlation (minimizing redundancy; filter criteria: Pearson correlation $<.95$ with other single features) [14]. This low computational effort demanding technique leads to a compact representation of the feature space. Furthermore wrapper based supervised subset selections were employed to optimize predicting accuracy using search strategies as forward selection, backward elimiantion, and genetic algorithm based methods. In addition we employed feature space transformation techniques for dimensionality reduction as Principal Component Analysis, Self Organizing Map, and Single Value Decomposition.

Referring to the classifier choice one could consider the use of SVM here, as they have proven in many works to best model static acoustic feature vectors [8]. Nevertheless we applied several often used classifiers as well. Thus, we used for the classification 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 Neighbor (KNN; $k = 1, 2, \text{ or } 3$), 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. 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.

IV. RESULTS

In order to determine the multivariate prediction performance, different classifiers were applied on the 130 features remaining after the correlation-filter procedure. For all configurations, we trained the classifier and applied them on the test sets. The averaged recognition rates (RR = ratio correctly classified samples through 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), Naive 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 reached the highest sum of RR and CL, and was therefore applied to further detailed dimensionality reduction. The results are depicted in Table 1.

Table 1 Recognition rates and class-wise averaged classification rate for different dimensionality reduction techniques using SVM as classifier and leave-one-sample-out validation strategy (# = number of remaining components)

Feature Transformation	#	Feature Selection	#	RR	CL
PCA (90%)	36	-	-	77.6	73.6
PCA (99%)	61	-	-	76.6	72.7
SVD	2	-	-	72.3	66.2
SVD	3	-	-	79.8	74.6
SVD	5	-	-	77.6	72.3
SOM	2	-	-	63.8	50.0
-	-	Forward Select.	14	74.5	71.1
-	-	Backward Elimin.	124	76.6	70.2
-	-	Genetic Algorithm	37	83.0	77.7
PCA (90%)	36	Genetic Algorithm	21	78.7	71.8
SVD	3	Genetic Algorithm	2	72.3	64.3
-	-	<i>Reference:</i> <i>Correlation Filter</i>	130	85.1	83.9

As shown in Table 1 the most successful dimensionality reduction approach is the simple correlation filter based feature subset selection with 85.1% recognition rate and 83.9 class-wise averaged classification rate.

V. DISCUSSION

To cover possible prosodic, speech quality and articulatory changes in sleepy speech an uncommonly large 45k feature-space was reduced with several dimensionality reduction methods and then fed into diverse classifier. The best classifier results were achieved for a SVM classifier. Overall results of 85.1% recognition rate of microsleep endangered sleepiness stages highly emphasize the benefits of the brute-force feature computation and a simple correlation filter based feature selection. The application of any additional dimensionality reduction methods including genetic algorithm based techniques reduced the sleepiness recognition rates. These results are surprising as genetic algorithm based selection and generation claimed showed superior prediction accuracies [11]. Nevertheless the overall prediction results of about 80% for the 2-class sleepiness prediction problem were largely as could be expected [5-9].

Nevertheless our results are limited by several facts. The present results are preliminary and need to be replicated using a natural speech environment. Furthermore, it would seem advisable that future studies address the following issues: (a) the computation of signal processing features derived from state space domains as e.g. average angle or length of embedded space vectors [1,10], Lyapunov exponents, correlation dimensions, time resolved densities, fractal dimensions, multiscale entropies, and recurrence quantification analyses [15] should be computed. In addition, different normalization procedures could be applied as, e.g. computing speaker specific baseline corrections not on functional level features but on duration adapted LLD level. (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). Another method for reducing the dimensionality of the feature space are unsupervised feature transformations methods (e.g. PCA Network, Nonlinear Autoassociative Network, Multidimensional Scaling, Sammon Map, Enhanced Lipschitz Embedding) or supervised feature transformation methods (e.g. LDA)

VI. CONCLUSIONS

Overall results of predicting microsleep endangered sleepiness stages from speech highly emphasize the benefits of the brute-force feature computation in combination with simple dimensionality reduction techniques as correlation filter.

REFERENCES

1. Sommer D, Chen M, Golz M, Trutschel U, Mandic D (2005) Fusion of State Space and Frequency Domain Features for Improved Microsleep Detection. In W Dutch et al. (Eds.) *Int Conf Artificial Neural Networks (ICANN 2005)*, pp 753-759. Springer: Berlin
2. Golz M, Sommer D, Chen M, Trutschel U, Mandic D (2007) Feature Fusion for the Detection of Microsleep Events. *J VLSI Signal Proc Syst*, 49:329-342
3. Harrison Y, Horne JA (1997) Sleep deprivation affects speech. *Sleep* 20:871-877
4. Whitmore J, Fisher S (1996) Speech during sustained operations. *Speech Communication* 20:55-70
5. Krajewski J, Kröger B (2007) Using prosodic and spectral characteristics for sleepiness detection. *Interspeech Proc.*, Antwerp, Belgium, 2003, pp 1841-1844
6. Krajewski J, Wieland R, Batliner A (in press) An acoustic framework for detecting fatigue in human-computer interaction. *ICCHP Proc.*, Linz, Austria, 2008
7. Nwe TL, Li H, Dong M (2006) Analysis and Detection of Speech under Sleep Deprivation. *Interspeech Proc.*, Pittsburgh, USA, 2006, pp 17-21
8. Vlasenk B, Schuller B, Wendemuth A, Rigoll G (2007) Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. *Interspeech Proc.*, Antwerp, Belgium, pp 2249-2252
9. Batliner A, Steidl S, Schuller B, Seppi D, Laskowski K, Vogt T, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V (2006) Combining Efforts for Improving Automatic Classification of Emotional User States. In Erjavec T & Gros JZ (Eds.): *Language Technologies, IS-LTC 2006*, Ljubljana, Slovenia, pp 240-245
10. Mierswa I, Morik K (2005) Automatic feature extraction for classifying audio data". *Machine Learning Journal* 58:127-148
11. Schuller B, Reiter S, Rigoll G (2006) Evolutionary Feature Generation in Speech Emotion Recognition. *ICME 2006*, Toronto, Canada, 2006, pp 5-8
12. Åkerstedt T, Gillberg M (1990) Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience* 52:29-37
13. Boersma P (2001) PRAAT, a system for doing phonetics by computer. *Glott International* 5:341-345
14. Witten IH, Frank E (2005) *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco
15. Webber C L, Zbilut J P (1994) Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology* 76: 965-973