Drink and Speak: On the automatic classification of alcohol intoxication by acoustic, prosodic and text-based features

Tobias Bocklet, Korbinian Riedhammer, Elmar Nöth

Chair of Pattern Recognition, University of Erlangen-Nuremberg

tb@speech.informatik.uni-erlangen.de

Abstract

This paper focuses on the automatic detection of a person's blood level alcohol based on automatic speech processing approaches. We compare 5 different feature types with different ways of modeling. Experiments are based on the ALC corpus of *IS2011 Speaker State Challenge*. The classification task is restricted to the detection of a blood alcohol level above 0.5 ‰. Three feature sets are based on spectral observations: MFCCs, PLPs, TRAPS. These are modeled by GMMs. Classification is either done by a Gaussian classifier or by SVMs. In the later case classification is based on GMM-based supervectors, i.e. concatenation of GMM mean vectors. A prosodic system extracts a 292-dimensional feature vector based on a voiced-unvoiced decision. A transcription-based system makes use of text transcriptions related to phoneme durations and textual structure. We compare the stand-alone performances of these systems and combine them on score level by logistic regression. The best stand-alone performance is the transcription-based system which outperforms the baseline by 4.8 % on the development set. A Combination on score level gave a huge boost when the spectral-based systems were added (73.6 %). This is a relative improvement of 12.7 % to the baseline. On the test-set we achieved an UA of 68.6 % which is a significant improvement of 4.1 % to the baseline system.

Index Terms: GMM, alcohol intoxication, system fusion

1. Introduction

Besides linguistic information spoken language contains also non-verbal information about speaker related characteristics. Examples of such characteristics contained in spoken language can be divided into two groups: characteristics that vary never or slowly (like identity of a speaker, gender, age) or characteristics which can change abruptly or within a short time. Examples for the latter are fatigue [1] and sleepiness [2], stress [3], emotion [4] or alcohol intoxication [5].

This work addresses the automatic detection of alcohol intoxication in speech. This has been a topic at the INTERSPEECH 2011 Speaker State Challenge (IS2011-SS) [6] where the Alcohol Language Corpus (ALC) is provided. The task of this challenge is to detect whether a speaker is alcoholized (blood alcohol concentration (BAC)> 0.5‰) or not alcoholized (BAC ≤ 0.5‰).

The question became quite popular in the late 80ies during the Exxon Valdez oil spill where the captain of an oil tank ship was suspected of being intoxicated. As part of the investigation the captain's speech from marine radio communication was examined for alcohol-related effects [7]. Different studies investigated acoustic [8], prosodic [9] and word-based cues [10] and revealed that there are measurable acoustic, prosodic and lexical differences in alcoholized speech. However, the studies dealt with read speech, which is quite different from the ALC corpus that contains read speech, isolated words, and spontaneous speech.

We pick up some of the mentioned ideas and come up with different automatic systems for the detection of alcohol intoxication. The goal is to compare their difference and to reveal if a combination of these systems could achieve a significant improvement. The first set of systems is of purely acoustic nature and models utterances by Gaussian Mixture Models (GMMs) of spectral features. The classification is either performed by Gaussian classifiers or by Support Vector Machines (SVMs). The second kind of system models each utterance by a prosodic feature vector based on a voiced-unvoiced (VUV) decision. These systems have been applied (and combined) successfully to the task of age or gender recognition ([11, 12]). The third kind of system focuses on the exploitation of the available (word and phoneme) transcriptions of the ALC corpus. We are following the impression, that durations of phones might be different when speakers are alcoholized and that textual cues like rate of speech, irregularities, word abortions or hesitations are useful features for our task. Finally, we used the provided openSMILE features as additional system. Note that this system uses features similar to our acoustic and prosodic systems but with a different way of modeling.

In Sec. 2 we will present the 6 different acoustic systems with 3 different kinds of features, the prosodic system, and the different transcription-based systems. Their stand-alone results on the development set are described in Sec. 3.1. Combination results on development and test set are presented in Sec. 3.2. The paper will be finished by a conclusion and outlook (Sec. 4).

2. System Description

2.1. Acoustic Systems

The subsystems in this section model acoustic features by *Gaussian Mixture Models* (GMMs). Two features use short-time spectral analysis, namely Mel Frequency Cepstrum Coefficients (MFCCs) and Perceptual Linear Prediction (PLPs). The third system is based on TempoRAl Patterns (TRAPS), which observe the speech signal within a longer temporal context. These features are either modeled by a standard GMM with Universal Background Model (GMM-UBM) [13] or by GMM-Support Vector Machine (GMM-SVM) modeling [11].

2.1.1. Mel Frequency Cepstrum Coefficients

A Hamming window with a size of 25 ms and a time shift of 10 ms is applied to the speech signal. Afterwards the Mel-spectrum with 26 triangular filters is calculated and processed by Discrete Cosine Transform (DCT). We take the first 13 Mel-

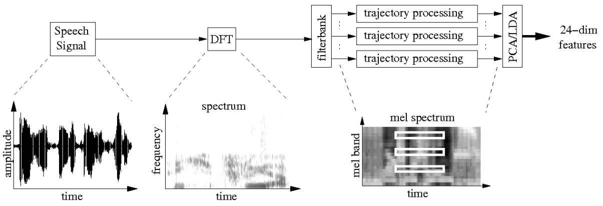

Figure 1: Adapted feature extraction for Temporal Patterns

frequency cepstral coefficients including C_0. Cepstral mean subtraction (CMS) is applied and first- and second order derivatives of these features are calculated over a context of 5 and 9 consecutive frames. In the end a 39-dimensional feature vector is created. Non-speech frames are discarded within each utterance, so that only speech frames are considered for further modeling. The speech/non-speech detection is based on BUT's Hungarian phoneme recognizer [14].

2.1.2. Perceptual Linear Prediction

The second set of short-time spectral features are Perceptual Linear Prediction (PLP) features [15]. Parameter settings are similar to our MFFCs. The differences lie in the use of a Bark filter bank and the use of linear prediction (LP) before Cepstrum computation. PLPs are said to be more robust in noisy conditions. Non-speech frames are also discarded for this feature set.

2.1.3. Temporal Patterns

Our *TempoRAl PatternS* (TRAPS) in this work are quite similar to the original approach of Hermansky [16]. The main difference of our approach is the different processing within the time trajectories. A detailed explanation can be found in [17].
Figure 1 illustrates the processing step of our TRAPS: The time trajectories consider a long temporal context of 31 coefficient (310 ms) in 18 bands. These bands are generated by a Mel filter bank. Each trajectory is smoothed by a Hamming window and transformed into frequency domain In a fusion step we concatenate the 31 coefficients of each band together to a high-dimensional feature vector ($D = 558$). This vector is then transformed by a *Linear Discriminant Analysis* (LDA) to a 24-dimensional vector. The LDA is trained on a 578-speaker subset of the German Verbmobil database [18]. 46 German phonetic classes serve as labels for this transformation.

2.1.4. GMM-UBM modeling

After extraction of the spectral features an UBM, i.e., a class-independent GMM with 256 Gaussians, is trained on the whole training set by Expectation-Maximization (EM) algorithm. Note that we use full covariance matrices for this way of modeling. Weights, means and variances of the UBM are adapted by relevance *Maximum A Posteriori* (MAP) adaptation in order to get two different class GMMs (not alcoholized and alcoholized). Scoring is based on the arithmetic mean of frame-wise likelihood calculation.

2.1.5. GMM-SVM modeling

In GMM-SVM modeling GMMs are created for each utterance in the same manner as described in Sec. 2.1.4. Instead of using 256 fully-equipped Gaussians, we use 512 Gaussians and only adapt the means in the adaptation process. Finally, for each utterance a single GMM is trained. The mean vectors of each Gaussian are extracted and concatenated to a big vector ($D * 512$), where D is the dimension of the acoustic features (39 for MFCC and PLP, 24 for TRAPS). For each utterance (for both the training and the test set) one of these so-called GMM-supervectors is created and then used for classification with *Support Vector Machines* (SVM). This approach is common in the field of speaker identification and has been applied to age recognition in [11]. The C-values and kernel method have been optimized on the development set and resulted in $C = 0.01$ using the linear kernel. Since training examples for the two classes are not balanced, we applied a random resampling technique in order to bias the class distribution toward a uniform distribution. The resampling technique was used for all systems that use SVMs as classifier.

2.2. Prosodic System

The prosodic system is not based on any speech recognition output or forced time alignments. The prosodic features are calculated whenever a voiced speech segment is found. The voiced-unvoiced (VUV) decision is based on the zero crossing rate, the normalized energy of the signal and the maximum energy.
Prosodic base features are calculated on the whole utterance. These are fundamental frequency (F_0), energy, VUV segments and pitch periods. Structured prosodic features are calculated on the voiced segments. Segments which are shorter than 50 ms are deleted, i.e., the neighboring segments are merged. the corresponding F_0 contour is interpolated to make the segmentation more robust. Context segments, that merge two adjacent segments together, are used additionally. All in all, 73 features are calculated for each segment modeling F_0, energy, duration, pauses, jitter and shimmer. A detailed description of the whole feature set is given in [19]. Finally, we compute mean, minimum, maximum and standard deviation of these 73 segments features. This forms our 292-dimensional prosodic feature vector. Classification is performed by SVM with a resampling of the training instances as mentioned in Sec. 2.1.5, with $C = 1$ for the linear kernel (optimal on development set).

2.3. openSMILE System

Additionally, we used the openSMILE feature set of the baseline system described in the challenge paper [20]. We also used SVM classification with $C = 0.01$ and a linear kernel. We used a different resampling technique (see Sec. 2.1.5) in order to achieve a balanced number of training instances for the two classes.

2.4. Transcription-based Systems

The transcription-based features are motivated by the impression, that durations of phones might be different when speakers are alcoholized and that textual cues like rate of speech, irregularities, word abortions or hesitations are useful features for our task. Similar ideas are mentioned in [10] and [5].

2.4.1. Phoneme Duration System

From the phoneme alignments, we extracted duration statistics for pauses (excluding initial and final), schwas, vowels and diphthongs. Mean and standard deviation were computed both for the individual phones as well as their group (open, mid, ...) to obtain better statistics.

2.4.2. Textual System

Following the general impression that intoxicated people usually tend to speak differently, we computed the following textual features from the rich transcriptions.

- duration of the turn (in milliseconds)
- number of false, dialectical, unintelligible words
- number of restarts, interrupts, irregularities, hesitations, words
- approximate rate of speech (characters / duration)

Furthermore, we extracted a lexicality feature motivated by the fact that an ideal (sober) utterance is free of repetitions, and an intoxicated speaker repeats words or sequences more often.

$$\text{lex}_{\text{obs}} = -\sum_{w} p(w) \cdot \log p(w) \quad (1)$$

where $p(w) = \text{count}(w)/N$, N is number of words. The Kullback-Leibler divergence between ideal (i.e. $q(w) = 1/N$) and observed lexicality can be computed as

$$\text{lex}_{\text{div}} = \sum_{w} p(w) \cdot \log \frac{p(w)}{q(w)} = \sum_{w} p(w) \cdot \log(Np(w)) \quad (2)$$

2.5. System Fusion

Fusion of the different systems is performed on score level and is either based on linear logistic regression (LLR) as it is implemented in the *FoCal* toolkit [21] or on a simple majority voting. In the former case, the system combination is achieved by a (calibrated) weighted sum α_i, $i = 1,..,K$ of the K different system scores with an accounting for a possible class offset $\vec{\beta}$. α_i and $\vec{\beta}$ are optimized in a training step where $\vec{\beta}$ can remain zero. The weights α_i can also be selected equally. This allows 4 different ways of combination:

- logistic regression with optimized α_i and $\vec{\beta}$
- logistic regression with optimized α_i and $\vec{\beta} = \vec{0}$
- logistic regression with equal weights α_i
- majority voting

Note that α_i and/or $\vec{\beta}$ are optimized on the development set, which leads to optimal combination results on the development set.

3. Experiments and Results

We first show the results of our stand-alone systems introduced in Section 2 achieved on the development set. After that we summarize our results achieved on the development set with the different fusion approaches. This section also contains the results of the five possible submissions on the test set.

3.1. Experiments on stand-alone systems

The results on the development set achieved by the different stand-alone systems is summarized in Table 1. Unless stated otherwise, the discussion of the results always refers to unweighted accuracy (UA). The best stand-alone result is achieved by a combination of the two transcription-based systems: text-based and phoneme-based. Combination in this case refers to a simple combination of the features. This system achieves an improvement compared to the baseline of 4.8 %. The short-time spectral GMM-SVM systems achieved the best performances among the acoustic system with an UA of 66.77 % which is an improvement of about 2.3 % compared to the baseline. The acoustic and prosodic systems tend to recognize the non-alcoholized (NAL) trials in higher proportions, transcription-based systems tend to achieve better results on alcoholized (AL) trials. This might be an issue, if miss-classification should not be equally penalized.

system	% UA	% WA	% NAL	% AL
baseline (cf. [20])	65.3	69.2	–	–
MFCC-SVM	66.8	69.1	72.4	61.1
PLP-SVM	66.5	68.5	71.4	61.5
TRAPS-SVM	61.4	62.7	64.6	58.3
MFCC-UBM	62.6	63.7	65.3	59.8
PLP-UBM	64.5	65.1	65.9	63.1
TRAPS-UBM	62.3	67.8	75.7	49.0
PROSODIC	58.8	59.3	59.9	57.7
text	59.1	57.1	62.0	72.2
phoneme	67.6	64.5	60.0	75.2
text+phoneme	68.4	66.9	64.6	72.2
openSMILE (OS)	65.5	65.9	66.5	64.4

Table 1: Stand-alone results of the different systems achieved on the development set. The first line contains the baseline results provided in [20]. The results on the two classes are unknown for this system since we did not reproduce the baseline results.

Since we use a different resampling than the challenge baseline system, we tried our resampling approach on their features. UA results of the different resampling techniques are quite similar, but seem to be more balanced when comparing UA and weighted accuracy (WA).

The prosodic system achieved the lowest stand-alone recognition result of 58.79 %. We are expecting this system to add complementary information for system combination.

3.2. Experiments on System Combination

Combination results with all possible combinations of all different systems with the different fusion mechanisms mentioned in 2.5 have been tried out (~700 combination). A first finding is that the combinations with the best results all contain at least one short-time spectral system, the transcription based system (text+phoneme) and either the prosodic system and/or the openSMILE system. This finding suggest the assumption, that due to the different motivations of the mentioned systems, complementary information is added when these systems are combined. The differences in performances of the top 5 systems are not significant. Nevertheless, we wanted to evaluate how much (complementary) information is added by the different systems.

Table 2 contains combination results for the most interesting system combinations with different ways of combination. $\alpha + \vec{\beta}$ refers to LLR combinations with optimized weights and offset, α refers to optimized weights. The column *equal* contains LLR combination results with equal system weighting, *maj. vot.* refers to majority voting results.

A combination of the three GMM-SVM, PROSODIC (PR), text+phoneme (TP), and openSMILE (OS) systems achieved the most promising results on development set. Whenever a text+phoneme system is accounted in a weighted combination, the weight of this system in general is 2 times higher than the other systems.

Based on these findings we started scoring on the test set.

systems	$\alpha+\vec{\beta}$	α	equal	maj. vot.
SVM	68.5	68.3	68.3	67.4
SVM+OS+PR	69.2	68.9	68.4	68.1
SVM+UBM+OS+PR	70.9	69.2	69.0	66.4
SVM+OS+PR+TP	73.6	72.8	71.6	71.1

Table 2: Combination results (% UA) on the development set with different combination measures

submission	% UA	% WA	% NAL	% AL
sub1	67.0	66.8	64.1	69.9
sub2	64.5	63.6	53.8	75.1
sub3	67.2	67.4	69.9	64.6
sub4	68.6	68.5	66.5	70.7
baseline	65.9	66.4	–	–

Table 3: Combination results (% UA) on test set of the different submissions

All system are retrained on the training and development data. Table 3 contains the results achieved by the five different submissions. Submission 1 (sub1) is an LLR combination of the GMM-SVM, PROSODIC, text+phoneme, and openSMILE systems with equal system weighting. Unfortunately, we did not achieve an improvement as on the development set with this submission. Nevertheless, an improvement of 2.6 % compared to the baseline was achieved. sub2 contained the same systems as sub1, but here we optimized α_i and $\vec{\beta}$ on development set. This combination actually was worse compared to the baseline. Since the weights contained a high proportion for text+phoneme system, we assumed this system to be unsuitable for the test set. We therefore submitted sub3 which is an LLR combination of SVM+PROSODIC+openSmile features with equal weights α_i. Since this submission achieved slightly higher results than sub1 (2,9 % improvement) we did not use the text+phoneme system for further submissions. sub4 contains all systems, except the text+phoneme system. α_i have been optimized on development set for this submission. This system achieved the best results so far and achieves an improvement of 5.1 % compared to the baseline.

4. Conclusion and Outlook

This work focused on the automatic detection of alcohol intoxication based on automatic speech processing systems. Different systems which employ different feature types have been tested: acoustic, prosodic, and transcription-based. Combination experiments on the development set revealed, that a combination of systems with all types of features leads to a significant improvement over the best stand-alone result. When confirming this on the test set we discovered a problem regarding the transcription-based system. This system lead to lower submission results on the test set. After excluding this system, we achieved a significant improvement of over 5 % compared to the baseline system of IS2011-SS.

5. References

[1] C. Stamoulis, "Effects of human fatigue on speech signals," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2632–2632, 2004.

[2] 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, pp. 795–804, 2009.

[3] H. Steeneken and J. Hansen, "Speech under stress conditions: overview of the effect on speech production and on system performance," in *Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on*, vol. 4, 1999, pp. 2079 –2082 vol.4.

[4] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Commun.*, vol. 40, pp. 117–143, April 2003.

[5] L. C. Sobell and M. B. Sobell, "Effects of alcohol on the speech of alcoholics," *Journal of Speech and Hearing Research*, vol. 15, pp. 861–868, 1972.

[6] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mueller, and C. Narayanan, "The Interspeech 2010 Paralinguistic Challenge," in *Proc. Interspeech (2010)*, 2010, p. no pagination.

[7] M. Brenner and J. Cash, "Speech analysis as an index of alcohol intoxication–the exxon valdez accident," *Aviat Space Environ Med*, vol. 62, pp. 893–898, 1991.

[8] D. B. Pisoni and C. S. Martin, "Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses," *Alcohol Clin Exp Res*, vol. 13, no. 4, pp. 577–587, 1989.

[9] M. Levit, R. Huber, A. Batliner, and E. Nöth, "Use of prosodic speech characteristics for automated detection of alcohol intoxination," in *Proc. of the Workshop on Prosody and Speech Recognition 2001*, M. Bacchiani, J. Hirschberg, D. Litman, and M. Ostendorf, Eds., 2001, pp. 103–106.

[10] D. M. Behne, S. M. Rivera, and D. B. Pisoni, "Effects of alcohol on speech: I. durations of isolated words, sentences, and passages in fluent speech," *The Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 2311–2311, 1991.

[11] T. Bocklet, A. Maier, J. Bauer, F. Burkhardt, and E. Nöth, "Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines," in *Proc. ICASSP 2008*, vol. 1, 2008, pp. 1605–1608.

[12] T. Bocklet, G. Stemmer, V. Zeissler, and E. Nöth, "Age and Gender Recognition Based on Multiple Systems - Early vs. LateFusion," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, ISCA, Ed., 2010.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, pp. 19–41.

[14] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006, p. I.

[15] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustic Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[16] H. Hermansky and S. Sharma, "TRAPS – classifiers of temporal patterns," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998.

[17] T. Bocklet, A. Maier, and E. Nöth, "Text-independent Speaker Identification using Temporal Patterns," in *Text, Speech and Dialogue*, ser. Lecture Notes of Artificial Intelligence, V. Matousek and P. Mautner, Eds., vol. 1, Berlin, 2007, pp. 318–325.

[18] W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*. New York, Berlin: Springer, 2000.

[19] A. Maier, F. Hönig, V. Zeissler, A. Batliner, E. Körner, N. Yamanaka, P. D. Ackermann, and E. Nöth, "A language-independent feature set for the automatic evaluation of prosody," in *Proc. Interspeech 2009*, Brighton, England, 2009, pp. 600–603.

[20] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The Interspeech 2011 Speaker State Challenge," in *Proc. Interspeech (2011)*, 2011, p. no pagination.

[21] N. Brümmer, *FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores*, 2007.