Laying the Foundation for In-car Alcohol Detection by Speech

Florian Schiel, Christian Heinrich

Bavarian Archive for Speech Signals (BAS), Ludwig-Maximilians-Universität München, Germany

schiel@bas.uni-muenchen.de, heinrich@bas.uni-muenchen.de

Abstract

The fact that an increasing number of functions in the automobile are and will be controlled by speech of the driver rises the question whether this speech input may be used to detect a possible alcoholic intoxication of the driver. For that matter a large part of the new Alcohol Language Corpus (ALC) edited by the Bavarian Archive of Speech Signals (BAS) will be used for a broad statistical investigation of possible feature candidates for classification. In this contribution we present the motivation and the design of the ALC corpus as well as first results from fundamental frequency and rhythm analysis. Our analysis by comparing sober and alcoholized speech of the same individuals suggests that there are in fact promising features that can automatically be derived from the speech signal during the speech recognition process and will indicate intoxication for most speakers.
Index Terms: alcohol detection, speaker characteristics, speech corpus, Alcohol Language Corpus, BAS

1. Introduction

Alcoholic intoxication (AI) has always been and still is one of the major causes for traffic accidents. AI can be measured by (ordered by descending reliability): taking blood samples (BAC), breath alcohol detectors (BRAC) and a variety of psychological tests (mainly reaction time and motor control). All these tests can only be applied either in random checks or post-accidentally, that is after an accident has already happened. Currently there are no known practical methods to routinely check on the AI of a driver pre-emptively.

The debate whether AI can reliably be detected from the speech signal has been going on for quite a while now. When the oil tanker Exxon Valdez stranded in Alaska in 1989, the captain of the ship was suspected for being under alcoholic influence during the time of the crisis. Forensic analysis of the recorded air traffic indicated that the spectra of the phone /s/ were skewed in the direction of an /S/ sound[1] which was considered as an indicator for drunkenness ([1]). Aside from this case study a number of – mostly forensic – studies have investigated phonetic and linguistic parameters of laboratory speech to find reliable indicators for AI (e.g. [2],[3],[4],[5],[6],[7]). Unfortunately most of these studies dealt with read speech in an acoustically clean environment, male speakers, less than 40 test persons and the AI was not measured reliably by BAC.

Nowadays automobiles are equipped with a growing number of functions controlled by speech input. Prominent examples are entertainment (radio, CD), control of the hands-free telephone and input to the navigation system. The type of speech applied here is typical command & control consisting of a limited number of pre-determined commands (often only 3-7 words) and issued to the car system after pressing a button on the steering wheel via a built-in microphone in the roof of the cabin. However, it is to be expected in the near future that more sophisticated voice input in the form of keyword activation and free speech (longer sentences) – as already being demonstrated in prototype systems these days – will be incorporated into standard car systems.

This leads to the interesting question whether a pre-emptive test of alcoholic intoxication using speech input might be feasible in the automotive environment: Since the driver of an automobile will increasingly use his or her voice to communicate with the car system, would it be possible for the car system to automatically retrieve indicators for AI and react accordingly, for instance by warning the driver about her or his condition?

To test this proposition a database of speech samples is required which contains speech of drivers being sober and under the influence of AI. The database needs to contain a large number of female and male speakers and enough material to train and test algorithms of statistically based pattern recognition (e.g. [8]). Furthermore, a broad and statistically sound analysis of phonetic, linguistic and para-linguistic parameters is needed to come up with a set of features that can be applied successfully for this purpose.

In 2007 the Bavarian Archive for Speech Signals (BAS) based at the Ludwig-Maximilians-Universität München, Germany started to collect speech samples for the Alcohol Language Corpus (ALC) project. In close cooperation with the Institute of Legal Medicine, München, speech of a variety of volunteers were recorded under the influence of AI ([9]). Although the main objective of the ALC project is to find features indicating AI in a broad range of speech styles, a considerable proportion of the data were designed for the aim to yield realistic command & control speech from the automobile environment.

The aim of this paper is two-fold: Firstly we present the available data from the ALC project which are suitable to test the above formulated proposition and invite other researchers to use this data for their own investigations towards the goal of developing alcohol detectors based on speech. Secondly we will present some first results about features that already have been statistically tested as indicators for AI based on the current release of the ALC corpus. The paper is organized as follows: Section 2 relates some background information about AI and its relation to traffic accidents. Section 3 gives a very short overview about the ALC project while section 4 concentrates on the command & control speech within the ALC corpus. The remaining two sections report first results on fundamental frequency and rhythm features derived from a set of 82 recorded speakers.

2. Alcoholic Intoxication in Traffic

Alcoholic intoxication adversely affects the ability of drivers to safely control their vehicles. Since the beginning of mo-

[1] Phonetic symbols in SAM-PA.

bility at the start of the 19th century this has caused a significant proportion of road accidents, many of them with fatal consequences. For example in 1977 54000 accidents caused by alcoholic intoxication have been registered in West Germany with 75000 injured people (14% of all cases) and 3800 fatalities (22%). Since then the total number of fatalities in traffic has fortunately dropped to about one sixth in 2007 and with it the number of alcohol induced accidents ([10]). However this is mainly caused by better security standards and techniques in modern cars rather than by a more responsible behavior of the drivers.
In 2007 the proportion of fatalities in alcohol induced accidents is with 2.7% significantly higher than the corresponding number in all traffic accidents (1.5%). 61% of all persons involved in an alcohol induced accident are of age 18 to 44 (the majority of 26% in the segment 18-26). 33% of these show a BAC level between 0.05% and 0.109%, the remainder 67% show at least 0.11% and above (numbers of 2007, Germany, [11])

3. Alcohol Language Corpus

A brief description of the ALC project is given in this section. For a detailed description of the corpus please refer to [9].

3.1. Recordings

Speakers are recruited within the age range 21 - 75, equally distributed for both genders and in 5 different locations in Germany. Non-native speakers, speakers with a strong dialect as well as non-cooperative speakers were excluded from participation. For the final ALC corpus 150 speakers are envisaged.
Speakers voluntarily undergo a systematic intoxication test supervised by the staff of the Institute of Legal Medicine. Before the test each speaker chooses the blood alcohol concentration she wants to reach during the intoxication test. The possible range is between 0.05% and 0.20%. Using both Watson- and Widmark formula the amount of required alcohol for each person is estimated and handed to the subject. After consumption the speaker waits another 20 minutes before undergoing a breath alcohol concentration test and a blood sample test. Immediately after the tests, the speaker is asked to perform the ALC speech test which will last no longer than 15 minutes to avoid significant changes caused by fatigue or saturation/decomposition of the measured blood alcohol level.
At least two weeks later the speaker is required to undergo a second recording in sober condition, which takes about 30 minutes. Both tests take place in the same acoustic environment and are supervised by the same member of the ALC staff, who also acts as the conversational partner for the dialogue recordings.
The speech signal is recorded with two different microphones: one headset Beyerdynamic Opus 54.16/3 and one AKG Q400 mouse microphone, frequently used for in-car voice input, located in the middle of the front ceiling of the automobile. Both microphones are connected to an MAUDIO MobilePre USB audio interface where the analog signal is converted to digital and transferred to a laptop. The recording platform is SpeechRecorder [12], the sampling rate 44,1kHz, 16 bit, PCM.
The following meta data are associated with each recording: date and time, speaker ID, age, gender, weight, height, profession, smoker/non-smoker, drinking habits, the region in which the speaker attended elementary school, the environment in which the recording took place, BRAC, BAC, self-judged emotional state of the day in general and during the speech test.

3.2. Content

The set of sober recordings (B) contains twice as many recordings as in the intoxicated case (A): $|B| = 2|A| = 60$; set A is a true subset of B: $A \subset B$ (except for one address). All speakers are prompted with the same material.
Three different speech styles are part of each ALC recording: read speech, spontaneous speech and command & control (see section 4 for details). The read speech consists of numbers, addresses, spelling and tongue twisters. When designing the read speech prompts many combinations of sounds were included that have been reported as being prone to error under alcoholic intoxication (e.g. [2], [1]), such as the alveolar voiceless fricative alternating with the post-alveolar voiceless fricative, the alveolar voiceless plosive alternating with the velar voiceless plosive as well as all voiceless plosives alternating with their voiced counterparts. Spontaneous speech is covered by 3 (6 sober) monologues and 2 (4 sober) dialogues with the recording supervisor, which are initiated by pictures and questions; the length of the monologues and dialogues is restricted to a maximum of 60 sec each. Particularly the monologues and dialogues evoke rather spontaneous speech that comes fairly close to real-life-situations.

3.3. Quality control and Annotation

All recordings are reviewed by staff of the BAS to ensure that they fulfill the required quality standards. The begin and end of the recorded speech is manually marked on the time scale and the spoken input is transcribed using an extended SpeechDat transcription format (for details about the annotation please refer to [9]). If there exist repeated recordings for the same prompt, one is selected for further analysis.
Based on the transcript an automatic segmentation into phonemic categories is done using the MAUS system ([13]). A similar segmentation could be achieved by backtracking the results of a phoneme based ASR system.

3.4. Availability

The ALC corpus is available for unrestricted scientific and commercial usage (pre-releases are available; the final release is expected end of 2009). Interested parties may obtain copies of the corpus at BAS[2]. Please contact *bas@bas.uni-muenchen.de* or refer directly to the BAS catalogue at *www.bas.uni-muenchen.de/Bas*.

4. Command & Control Speech

All ALC recordings take place in an automobile, to ensure the same acoustic environment for the different recording locations and for the two main conditions *intoxicated/sober*. The engine is switched off except for the command speech where the running engine creates a realistic ambience for control commands. For security reasons no recordings are performed in the moving car.
About 2/3 of the recorded items in the ALC project can be considered as being potential speech used as command & control within the mobile environment. These are (numbers per recording in brackets for the sober case):

- 3 telephone numbers (6) – engine silent
- 1 credit card number (2) – engine silent
- 5 addresses (10) – engine silent

[2]BAS distribution fees apply.

Table 1: Examples of recorded commands by situational prompting. In the top line the prompt text followed by the recorded commands (translated into English)

You're listening to the radio, but would like to switch to CD. Ask your car system to do that!
Change to CD!
Change from radio to CD!
Radio off, CD on!
Play CD!
Please change to CD.
Switch from radio to CD, please.
Car radio off, CD on!
Car radio: please activate CD player!
Entertainment CD!
Please start CD!
Radio: source CD!
Hi, no more radio, I like a CD, please switch to CD.
Hey, fatso: switch off that jingle-jangle and put in the Johnny Cash CD! Hop to it!

- 1 spelled city name (1) – engine running
- 4 read commands (9) – engine running
- 5 spontaneous commands (10) – engine running

The read commands were randomly taken from a real prototype system for automobile speech control. They comprise 3.9 words (range 2-7) and 8.3 syllables (range 4-12) in average and cover each vowel and diphthong of the German vowel system at least two times.
The spontaneous commands are elicited using the 'situational prompting' technique developed by Mögele et al ([14]): The speaker is prompted by a description about a driving situation, in which she is using a fully functional speech interface in the running car. The speaker is then asked to perform certain control actions by addressing the car system in her own words, that is the actual commands are *not* part of the prompted text. After thinking about an appropriate command phrase the speaker hits a push-to-talk button and issues the command. The average word count per situational prompted command in ALC is 4.98 (based on 927 recorded commands).

To illustrate the outcome of this technique table 1 shows some of the issued commands of several sober speakers after being asked to switch the entertainment system from radio to CD.

To summarize, the descriptive factors for the command & control subset of the ALC corpus are:

Factor	Values
alcoholization	intoxicated, sober
speech content	number, address, command, spelling
speech style	read, spontaneous
background	engine running, engine silent
microphone	close, mid range

The following two sections present results of fundamental frequency and rhythm analysis on the close microphone signals of the command & control set recorded of 82 speakers (45 f + 37 m). Beside the intoxication we tested the type of speech, read or spontaneous, as well as the gender of the speakers as additional factors in the statistical analysis.

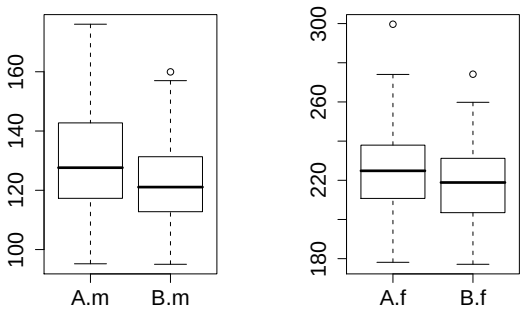

Figure 1: *F0 medians for alcoholized (A) and non-alcoholized (B) speech separated for male (37) and female (45) speakers.*

5. Fundamental Frequency F0

Fundamental frequency has been the principally investigated feature in earlier studies of laboratory alcohol speech (e.g. [1], [2], [3], [4], [6]). In most cases F0 is reported to rise with alcoholization; mostly male speakers were tested.
We calculated the fundamental frequency over the total utterances using the robust pitch detection algorithm of Vincent-Schaefer ([15]) and deleted all frames that were classified as unvoiced. From the remaining frames the median f_m and the quarter-quantile distance Δf_{qq} for each speaker were calculated. We prefer the median and quarter-quantiles over mean and standard deviation because they are more robust against outliers which are inevitable in automatic pitch detection. Figure 1 shows the distribution of f_m for alcoholized and non-alcoholized speech separately for female and male speakers. Testing the differences for alcoholic and non-alcoholic speech within each speaker using repeated measures ANOVA (RM-ANOVA) yields significant differences for both features f_m ($p < 0.001$) and Δf_{qq} ($p < 0.001$), although a post hoc test reveals that the latter is only significant for spontaneous commands ($p < 0.001$) but not for read speech ($p = 0.046$). Gender of the speakers does not have any significant effect.

Figure 2 shows the change of median $f_m(A) - f_m(B)$ for each tested speaker grouped by female and male speakers. The female speakers rather uniformly increase their average fundamental frequency while the male speakers behave in both directions. Note that only 16% speakers keep their register constant in both conditions. A reasonable assumption is therefore that in most cases intoxicated female speakers increase while male speakers either increase or decrease their pitch.

6. Rhythm Features

In the context of this investigation rhythm refers to the time patterns of voiced and unvoiced segments within continuous speech. All of the following rhythm features are based on a segmentation in consonantal, vocalic and silence segments which were automatically derived from the phonemic segmentation by grouping consonants and vowels into clusters. The position of the syllable nucleus is assumed to be in the middle of the vowel cluster; nucleus distances of more than 500msec are discarded.
We investigated 17 different rhythm features so far but due to limited space we'll concentrate on the following 5 features: *standard deviation (SD) of duration of vowel clusters* (deltaV.sd, [15]), *SD of distances between syllable nuclei* (deltaSN.sd), *average durational difference of consecutive*

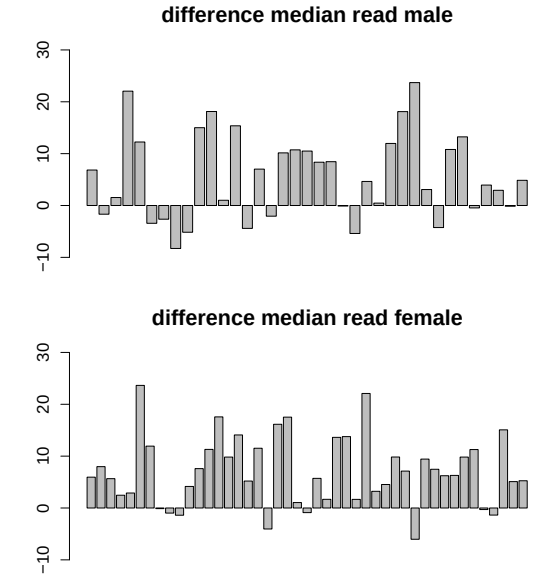

Figure 2: *Differences of median F0 for male (top) and female speakers.*

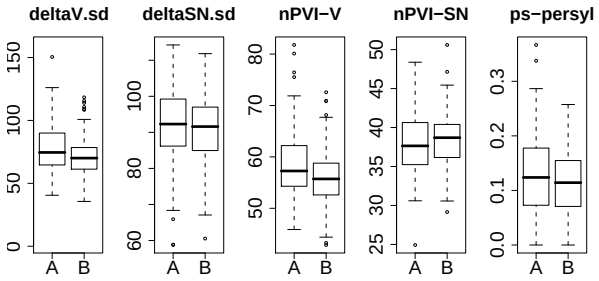

Figure 3: *Box-plots of the 5 rhythm features (see text) for alcoholized (A) and non-alcoholized (B) speech.*

vowel clusters (nPVI-V, [16]), *average difference of consecutive syllable nuclei distances* (nPVI-SN) and *short pause* ($< 1sec$) *rate per syllable* (ps-persyl). The first four features are supposed to show increased values on speech that contains varying durational patterns of voiced, unvoiced and silence intervals; the fifth feature simply reflects the increased usage of un-filled hesitations. While the features deltaV.sd, deltaSN.sd and ps-persyl are independent of the CVC order, the two features nPVI-V and nPVI-SN reflect the intrinsic structure of segmental durations.

Figure 3 shows box-plots of the alcoholized and non-alcoholized sets for the 5 rhythm features. RM-ANOVA yields significant differences for deltaV.sd ($p = 0.0014$) and nPVI-V ($p < 0.001$); ps-persyl shows only $p = 0.049$. A post-hoc test on feature nPVI-V reveals an interaction with the type of speech (read or spontaneous command): only command speech is significant with $p < 0.001$ while read speech shows no significant changes ($p = 0.79$). The opposite is true for the feature ps-persec: only read speech has significant changes ($p < 0.001$) while command speech does not ($p = 0.99$).
Both features based on the syllable nuclei, deltaSN.sd and nPVI-SN, yield no significant changes. There was no evidence for any gender effects.

7. Conclusion and Future Work

We presented a new speech corpus of alcoholized speech recorded in the automotive environment which could eventually lead to the development of automatic alcohol detection in the car. A first analysis of the fundamental frequency confirms that most speakers rise F0 under intoxication; we also found that this is not consistently the case across genders. Rhythm parameters also show significant changes under alcohol although here the influence of the speech style (read vs. spontaneous) is significant. Future work will include a broader analysis of phonetic features (RMS, formants, spectral tilt) and prosodic contours (F0, energy) that finally will lead to a statistical classificator (e.g. using a SVM) based on a combination of significant features.

8. References

[1] Johnson K, Pisoni D B, Bernacki R H (1990) Do voice Recordings Reveal whether a Person is Intoxicated? A Case Study. In: Phonetica, vol. 41, pp. 215-237.

[2] Künzel H J, Braun A (2003): The effect of Alcohol on Speech Prosody. In: Proc. of the ICPhS. Barcelona, pp. 2645-2648.

[3] Hollien H, De Jong G, Martin C A, Schwartz R, Liljegren K (2001): Effects of ethanol intoxication on speech suprasegmentals. In: The Journal of the Acoustical Society of America, pp. 3198-3206.

[4] Cooney O M, McGuigan K, Murphy P, Conroy R (1998): Acoustic analysis of the effects of alcohol on the human voice. In: The Journal of the Acoustical Society of America, p. 2895.

[5] Behne D M, Rivera S M, Pisoni D B (1991): Effects of Alcohol on Speech: Durations of Isolated Words, Sentences and Passages. In: Research on Speech Perception, No 17, pp. 285-301.

[6] Klingholz F, Penning R, Liebhardt E (1988): Recognition of low-level alcohol intoxication from speech signal. In: Journal of the Acoustical Society of America, vol. 84, 1988, pp. 929-935.

[7] Sobell L C, Sobell M B, Coleman R F (1982): Alcohol-Inducted Dysfluency in Nonalcoholics. In: Folia Phoniatrica, No. 34, pp. 316-323.

[8] Levit M, Huber R, Batliner A, Nöth E (2001): Use of prosodic speech characteristics for automated detection of alcohol intoxication. In: Bacchiani M, Hirschberg, J, Litman D, Ostendorf M (Eds.): Proc. of the Workshop on Prosody and Speech Recognition 2001, Red Bank, NJ, pp. 103-106.

[9] Schiel F, Heinrich Chr, Barfüsser S, Gilg Th (2008). ALC - Alcohol Language Corpus. In: Proc. of LREC 2008, Marrakesch, Marokko, paper 419.

[10] Bund gegen Alkohol und Drogen im Strassenverkehr. www.bads.de/Alkohol/statistik.htm. Cited 2009-03-23.

[11] Statistisches Bundesamt, Wiesbaden, Germany (2007) Alkoholunfälle 2007. e.g. www.bads.de/Statistikdaten/Alkohol/AlkVU%202007.pdf. Cited 2009-03-23.

[12] Draxler Chr, Jänsch K (2004) SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software. In: Proc. of the LREC. Lisbon, Portugal.

[13] Schiel F (1999) Automatic Phonetic Transcription of Non-Prompted Speech. In: Proc. of the ICPhS. San Francisco, August 1999. pp. 607-610.

[14] Mögele H, Kaiser M, Schiel F (2006) SmartWeb UMTS Speech Data Collection: The SmartWeb Handheld Corpus. In: Proc. of the LREC 2006, Genova, Italy, pp. 2106-2111.

[15] Schäfer-Vincent K (1983) Pitch period detection and chaining: method and evaluation. Phonetica 1983, Vol 40, pp. 177-202.

[15] Ramus F, Nespor M, Mehler J (1999) Correlates of linguistic rhythm in the speech signal. Cognition, Volume 73, Number 3, pp. 265-292, Elsevier.

[16] Grabe E, Low E L (2004) Durational Variability in Speech and the Rhythm Class Hypothesis. In: Gussenhoven C, Warner N (eds) Papers in Laboratory Phonology 7, Berlin, New York: Mouton de Gruyter.