Perception of Alcoholic Intoxication in Speech

Florian Schiel

Bavarian Archive for Speech Signals, Institute for Phonetics and Speech Processing,
Ludwig-Maximilians-Universität, München, Germany

schiel@bas.uni-muenchen.de

Abstract

The ALC sub-challenge of the Interspeech Speaker State Challenge (ISSC) aims at the automatic classification of speech signals into intoxicated and sober speech. In this context we conducted a perception experiment on data derived from the same corpus to analyze the human performance on the same task. The results show that human still outperform comparable baseline results of ISSC. Female and male listeners perform on the same level, but there is strong evidence that intoxication in female voices is easier to be recognized than in male voices. Prosodic features contribute to the decision of human listeners but seem not to be dominant. In analogy to Doddington's zoo of speaker verification we find some evidence for the existence of lambs and goats but no wolves.

Index Terms: alcoholic intoxication, speech perception, forced choice, intonation, Alcohol Language Corpus

1. Introduction

Alcoholic intoxication (AI) is known to influence a broad range of mental and motoric abilities in humans. It is therefore reasonable to expect that AI also changes the speech production in certain typical ways.

In most legal systems the amount of AI measured in absolute blood alcohol concentration (BAC) plays an important role in court decisions as well as prosecution procedures. For example in Germany a $BAC > 0.0005$ is considered as a traffic infraction while a $BAC > 0.0011$ is treated as a felony. Forensic phoneticians have therefore in several studies investigated the influence of AI on the speaker's voice mainly with the aim to provide evidence in court cases ([1, 2, 3, 5, 6, 7, 8, 9, 10, 14]). Unfortunately, until recently no public available speech corpus with speech of intoxicated speakers has been available to compare and replicate published results.

End of 2010 the Bavarian Archive for Speech Signals (BAS)[1] released the Alcohol Language Corpus (ALC) for unrestricted scientific usage. It contains a large collection of alcoholic intoxicated speech materials produced by male and female speakers (a detailed description of ALC can be found in [12]). ALC is one of the two benchmark corpora for the Interspeech 2011 Speaker State Challenge (ISSC) which aims at the automatic detection of AI and sleepiness based on the speech signal ([13]).

In this context the questions arises what performance is to expected as a gold standard in the task of AI detection. One straightforward way is to use the measured BAC which is provided for each speaker of the ALC. On the other hand it is to be expected that some speakers are able to mask their AI perfectly and therefore produce speech signals indistinguishable from normal (sober) speech. Hence, it would be interesting to see how human listeners perform on the same task, since for many recognition tasks concerning speech humans are considered to perform better than machines.

If the performance of human listeners is significantly above chance, the questions arise which features they use for their (successful) decisions and whether there is a difference between female and male listeners. Another question is the speaker dependency on the AI detection task. More specifically: are there distinctive speaker groups that

- reveal their AI more easily,
- mask their AI better than others, or
- appear to be under AI although being sober?

In analogy to Doddington's 'zoo of speaker verification' ([4]) I will refer to these three groups as *lambs, wolves* and *goats* in the remaining paper.

Based on common experience most listeners claim that they can reliably recognize AI in the speech uttered by intoxicated persons. This is probably not entirely true, because in real life situations the listener uses several other perceptual sources, such as facial expression, gesture, posture, gait and the situational context as input for her/his decision.

In a number of earlier studies results of identification tests on laboratory speech of intoxicated speakers have been reported. In [10] 44 male subjects performed a forced choice test on 192 sentences read by 8 (male) speakers resulting in an identification rate of 62.5%. Another study [7] reported a recognition rate of 54.0% on 30sec of read text spoken by 11 male speakers intoxicated with < 0.001 BAC and judged by 12 listeners; recognition rates increased to a maximum of 82.0% when the BAC was above 0.001. In [9] 33 male speakers produced read and semi-spontaneous speech under varying intoxication levels. 10-12sec long stimuli derived from these recording were used in an identification task performed by 30 listeners yielding an average recognition rate of 66.8%; recognition rates increased linear from 50.0 to 96.0% with increasing breath alcohol concentration over a range of $0.0004 - 0.002$.

In this study we investigate the human identification performance of AI based on the speech of male and female speakers derived from the ALC. At the same time the results should be comparable to the recognition rates achieved by participants of the ISSC. More specifically we want to answer the following questions:

1. What is the performance of human listeners on the AI detection task (binary decision)? Is that performance gender specific? Does it depend on the speech style?
2. What role play prosodic features of the speech signal, such as intonation contour, speech rate, pauses and rhythm?

[1] http://www.bas.uni-muenchen.de.de/Bas

3. Are human listeners able to predict the amount of AI, i.e. the blood alcohol concentration?
4. Are there any lambs, wolves and goats in the context of AI detection?

The remaining paper is organized as follows: the next section describes the selection of stimuli from ALC for the perception experiment. Section 3 gives details about listeners and the experimental setup. Section 4 presents the results of the perception experiments as well as some comments on the research questions given above.

2. Speech Stimuli

It is not possible to replicate the benchmark test of the ISSC by means of a perception experiment exactly, because the number of test samples is too high (50 speakers x 60 recordings). For our experiment we selected 16 speakers (8f+8m) of the same age group (24-30, average 26.5 years) and dialect region (southern German). The BAC is constant for each speaker and equally distributed from 0.0005 to 0.00142 (average 0.000945) across speakers. From each speaker we selected 12 stimuli of 7-10sec length according to Table 1. Read speech stimuli were

Table 1: *Selected stimuli per speaker*

	normal speech		inverse filtered	
	read	spont	read	spont
alc	2	2	1	1
sober	2	2	1	1

taken from identical read sentences spoken under sober and intoxicated conditions, while spontaneous speech was extracted from dialogues where the dialog partner was not audible and no longer silence intervals occurred. The inverse filtered speech was produced using praat's[2] 'hum' function, that is the excitation signal derived from the autocorrelation of the speech signal is fed into a neutral vocal tract model producing a hum that only reflects the intonation contour and rhythm pattern. All stimuli were intensity normalized to avoid different loudness which may be caused by varying mouth-microphone distances. In total 192 stimuli (128 normal / 64 filtered) were presented to the listeners.

3. Method

3.1. Listener Subjects

47 listeners (30f+17m) of age 20-39 (average 25.1) were recruited for the experiment. They all have the same dialectal background as the speakers and claimed not to have more than average contact to intoxicated persons (such as policemen, doctors etc.). The subjects were informed about the aim of the listening test, namely to identify alcoholic intoxication and that they will hear normal speech as well as humming speech, but nevertheless should concentrate on the intoxicated/sober decision in those cases.

3.2. Experimental Setup

The 192 stimuli were presented in random order and using the same technical equipment (headphone, sound-card) by means of a simple Web-Interface. Subjects could listen to each stimulus as often as they wished, and had to decide in a forced choice whether the speaker was intoxicated or not. Additionally subjects were asked to evaluate the reliability of each decision on a 7-step scale from 'very reliable' to 'very unreliable'.

[2]http://www.praat.org/

Table 2: *Recognition scores in % for normal and inversely filtered speech, total and separately for the factors* voice gender, listener gender *and* speech style*. Below: the top result of the ISSC ([13]).*

	normal speech		inverse filtered	
total	71.65		56.85	
voice gender	female	male	female	male
	77.07	66.23	56.78	56.91
listener gender	female	male	female	male
	71.54	71.76	56.04	58.27
speech style	read	spont	read	spont
	75.73	68.08	-	-
ISSC	67.56		-	
voice gender	female	male	-	-
	66.53	68.60	-	-

4. Results and Discussion

Table 2 shows the identification rates of the human listeners for normal and inversely filtered stimuli and three additional factors *voice gender, listener gender* and *speech style*. Comparable baseline recognition rates of the ISSC are given below[3].

4.1. General Performance, Gender & Speech Style

The total identification rate of 71.65% for unfiltered speech is well beyond chance ($p < 0.0001$)[3]. The rates based on female (77.07%) and male (66.23%) voice stimuli differ significantly ($p < 0.0001$). The performance of female (71.54%) and male listeners (71.76%) do not differ significantly. Looking at different speech styles we find that read speech (75.73%) is significantly better detected than spontaneous speech (68.08%, $p < 0.0001$).

Human listeners outperform the baseline system of the ISSC by absolute 4.09%. The performance in the baseline identification test is not significantly different on female (66.53) and male (68.60) voices ($p = 0.242$).

The answer to our first research question is: yes, listeners are able to detect intoxication from speech only, but with a considerable error rate of 28.35%. There is no gender specific difference among listeners, but it seems that female intoxicated voices are easier to be detected than male voices. The identification rate on male voices 66.23% confirms those reported earlier on intoxicated male speech where the BAC is below 0.001. The fact that female voices are easier to be recognized corresponds to other gender specific findings on long-term fundamental frequency F0 where is was shown that female intoxicated speakers raise their F0 more consistently than male speakers ([11]). Read intoxicated speech is recognized better than spontaneous speech. Probably because intoxicated speakers have no diffi-

[3]All statistical tests in this section are based on a simply χ^2 test on the respective number of judged stimuli.

culties speaking freely, while the additional mental load when reading text may cause them to produce more pronunciation errors or to lower their speaking rate, which then can be detected by the listeners.

4.2. The Influence of Prosody

The total recognition rate for the 'humming' stimuli (Table 2, inverse filtered), where listeners have to rely on prosodic features only, is still highly significant above chance ($p < 0.0001$) but the performance dropped by an absolute of 14.8% to 56.85%. There are no significant gender differences; speech styles have not been tested for filtered speech.

So, there is evidence that listeners exploit some of the prosodic information still contained in the filtered speech signal to detect intoxication; the dramatic drop in recognition performance could be either attributed to the filtering (which is unfamiliar to listeners) or possibly indicates that prosodic features are not the predominant features for this task.

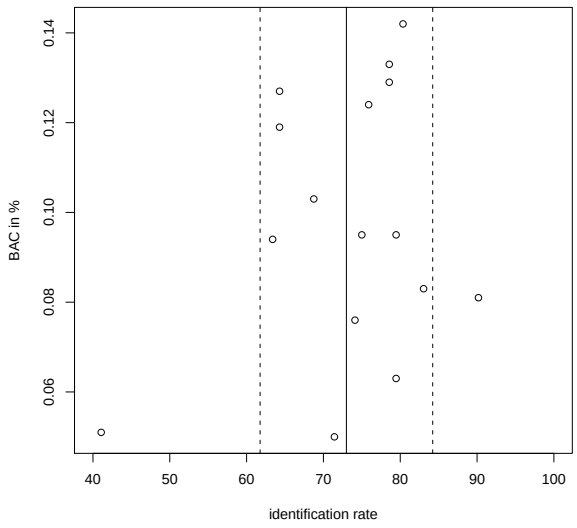

Figure 1: *Scatter plot of BAC values (ordinate, in %) vs. identification rates (x-axis, in %) per speaker.*

4.3. Predictability of BAC

Each speaker is associated with a measured BAC value which is valid for her/his stimuli uttered under alcoholic intoxication. Therefore we can calculate recognition rates for each speaker separately and correlate these to the BAC values to answer the question whether listener can predict the amount of intoxication based on their recognition performance. Figure 1 shows the scatter plot for the 16 speakers of our perception experiment. The Pearson correlation yields only 0.19, meaning that there is no relevant correlation between performance and BAC across different speakers. Using the averaged reliability scores per speaker instead of the performance rates leads to an increased correlation of 0.33 to BAC, but that still is not a reliable linear prediction. Scatter plots of both correlations did not indicate any non-linear behavior.

The answer to the third research question is therefore: no, there is no evidence that listeners can perceive the amount of intoxication from the speech signal across different speakers.

This contradicts earlier findings [7, 9] where identification rates were found to be nicely correlated with the amount of intoxication. One possible explanation is that those results were obtained from multiple recordings of the same speakers with varying BACs, while in this study each speaker is recorded with only one fixed BAC. It seems that inter-speaker differences in intoxicated speech are much larger than within speaker differences and therefore a linear increase of detection rate along with BAC is not generally true across different speakers.

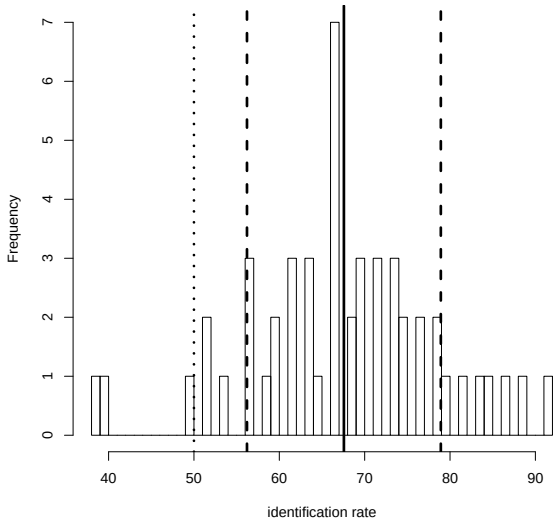

Figure 2: *Histogram of identification rates of speakers in the Interspeech 2011 Speaker State Challenge test set.*

4.4. Lambs, Wolves and Goats?

Since only 16 speakers were involved in our experiment, we cannot give definite answers to the question of lambs, wolves and goats in AI detection. However, looking at the recognition results of the individual speakers (Figure 1) we can see that rates scatter up to 90.18% (female speaker, $BAC = 0.00081$) with a standard deviation of 11.2. Hence, speakers differ considerable in the way that their intoxication can be detected which can be seen as evidence for a lamb group. If we (arbitrarily) define a potential lamb as being above the standard deviation range, we find one lamb candidate among the 16 speakers.

None of our 16 speakers showed recognition results at or around 50% chance; not even below the standard deviation range. The lowest recognition rate per speaker is 63.39% (male speaker, $BAC = 0.00094$) which is still significantly above chance ($p = 0.00035$). So, at least among the 16 speakers of this experiment no evidence for any wolves can be found.

One male speaker showed a reverse behavior as we would expect from a goat. For some reason this speaker was judged to be intoxicated when in fact being sober and vice versa. His recognition rate was 41.07% and therefore below chance with a weak significance ($p = 0.016$). So, we can at least confirm a small possibility of goats in AI from our perception data (see Table 3).

We also analyzed the baseline identification results of the

ISSC derived from the 50 speakers of the official test set ([13]). Figure 2 shows a histogram of the individual identification rates per speaker (based on 60 balanced trials per speaker). The mean and SD interval are plotted as solid and dashed lines respectively; chance of 50% is marked with a dotted line.

If we again consider speakers above the SD range to be potential lambs, we find 7 out of 50 speakers (14%). There seem to be two candidates for goats with 40.0% and 38.3% identification rate. But due to the low number of trials per speaker these values are still within the 95% confidence interval around chance. In contrast to the speakers in our perception experiment, we find 4 possible candidates for wolves (identification rates around chance), but again these results are not significant due to the large confidence interval.

Table 3: *Candidates for lambs, goats and wolves in analogy to Doddington's zoo ([4]) among the 16 tested speakers in the perception experiment and among the 50 speakers of the ISSC.*

task	total speakers	lambs	goats	wolves
perception	16	2	1	0
ISSC	50	7	2 (?)	4 (?)

5. Conclusion

A small class-balanced sample drawn from 16 speakers of the ALC was used in a simple forced choice perception experiment to quantify the ability of human listeners to detect alcoholic intoxication. The average accuracy of 47 listeners is with 71.65% significantly higher than the top baseline recognition result reported for the Interspeech 2011 Speaker State Challenge ([13]) (67.5%, unweighted accuracy, *training + development vs. test set* with 50 speakers)[4]. Hence, there is still some room for improvement in classification techniques.

Human listeners are more successful in detecting intoxication in female voices than in male voices, and in read rather than in spontaneous speech. On the other hand, female and male listeners show the same detection performance. Prosodic information can be exploited by human listeners for the decision process but probably not as much as other types of features.

There is some evidence that AI detection is strongly influenced by the individual behavior of speakers. More specifically, some speakers are easier recognized than the average (lambs), while some speakers are even judged to be intoxicated when in fact being sober and vice versa (goats). We found no significant indication for speakers that can mask their intoxication perfectly (wolves).

6. Acknowledgments

I would like to thank my student Ulrike Aulich for performing the perception experiment and my colleagues Felix Weninger and Björn Schuller for providing me with the baseline results on the Interspeech 2011 ALC Speaker Challenge and Anton Batliner for valuable advice.

[4]The difference to the 65.9% reported in [13], Table 5 is caused by a slightly different design: while the IS2011 Challenge classifies recordings of speakers with a $BAC < 0.0005$ as non-intoxicated, in our experiment only speakers with $BAC = 0.0000$ are treated as non-intoxicated. Therefore we re-calculated the rates based on the classification results of the baseline experiment to be as close as possible to our perception experiment.

7. References

[1] 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.

[2] Braun A (1991): Speaking while intoxicated: Phonetic and forensic aspects. Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, pp. 146-149.

[3] 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.

[4] Doddington G R (1998) Sheep, Goats, Lambs and Wolves - An Analysis of Individual Differences in Speaker Recognition Performance. In: Proc. of the ICSLP 1998, Sidney, Australia.

[5] 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.

[6] 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.

[7] 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.

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

[9] Künzel H J, Braun A, Eysholdt U (1992): Einfluß von Alkohol auf Sprache und Stimme. Kriminalistik Verlag Heidelberg.

[10] Martin C S, Yuchtman M (1986): Using speech as an Index of Alcohol-Intoxication. Research on Speech Perception, No. 12, pp. 413-426.

[11] Schiel F, Heinrich Chr (2009): Laying the Foundation for In-car Alcohol Detection by Speech. Proc. of the INTERSPEECH 2009, Brighton, UK, pp. 983-986.

[12] Schiel F, Heinrich Chr, Barfüßer S (2011): Alcohol Language Corpus. Language Resources and Evaluation, Springer, Berlin, New York, in print.

[13] Schuller B, Steidl S, Batliner A, Schiel F, Krajevski J (2011) The Interspeech 2011 Speaker State Challenge. In: Proc. of the Interspeech 2011, Florence, Italy, to appear.

[14] Sobell L C, Sobell M B, Coleman R F (1982): Alcohol-Inducted Disfluency in Non-alcoholics. In: Folia Phoniatrica, No. 34, pp. 316-323. Trojan F, Kryspin-Exner K (1968): The Decay of Articulation under the Influence of Alcohol and Paraldehyde. Folia Phoniatrica, No. 20, pp. 217-238.