Human perception of alcoholic intoxication in speech

Barbara Baumeister, Florian Schiel

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

bba | schiel@phonetik.uni-muenchen.de

Abstract

In this paper we describe a perception experiment on intoxicated and sober speech of 161 speakers recorded in the German Alcohol Language Corpus. 72 listeners achieved an average discrimination rate of 63.1% when asked to choose from pairs of stimuli which one sounded intoxicated. Perception results were not gender-dependent and no hidden effects were found in a control group test. Since earlier studies reported higher fundamental frequency for intoxicated speakers, the influence of fundamental frequency as a potential acoustic cue in human perception of intoxication was analyzed. Results show a significantly higher detection rate for speakers who produce a higher fundamental frequency when being intoxicated, and a higher success rate for listeners who show a general preference for choosing the stimulus with higher fundamental frequency. However, human listeners do not consistently exploit this acoustic cue, since a simple algorithm which always classifies the stimulus with higher fundamental frequency as intoxicated would lead to a better performance of 82% discrimination rate.

Index Terms: speech perception, Alcohol Language Corpus, alcoholic intoxication

1. Introduction

Alcohol consumption has various effects on the drinker, for example impaired balance, coordination problems, and slow reaction time. Another well-known effect is so-called "slurred speech", hence speech might act as a modality through which intoxication could be detected. In contrast to other test methods for intoxication, speech has the advantage that it can be observed without any obtrusive interaction with the potentially intoxicated person. Since voice controlled applications already exist in the automotive environment, the possibility of automatic detection of alcoholic intoxication by speech is of interest. If it was known which acoustic parameters change under the influence of alcohol, it might be possible to automatically detect intoxication in a built-in vehicle computer to prevent driving under influence.

A number of previous studies examined the effect of alcohol on the acoustic properties of the speech signal, including fundamental frequency (f0), but for f0 findings are inconsistent. They vary from a significant increase (e.g. [7], [6]) to a decrease (e.g. [16], [1]) to a change dependent on the breath alcohol concentration (BRAC) [8] or even to no change at all (e.g. [15], [10], [5]). Some of these contradictory findings might be due to varying experimental setups and the low number of participants: the number of speakers range from 4 to 35 and most studies were conducted with male participants only. In an earlier study [3] we conducted an f0 analysis of intoxicated speech based on the German Alcohol Language Corpus (ALC) [12] which was designed to provide a publicly available, large and statistically sound corpus for speech recorded in an automotive environment. The majority (79.1%) of speakers within the ALC corpus increase their median f0 when being intoxicated.[1]

During the INTERSPEECH 2011 Speaker State Challenge (ISSC) [13] a gender balanced set of 154 speakers from the ALC was provided as a benchmark set, and researchers were invited to test their automatic method of recognizing intoxication by means of a simple identification task. The best result was reached by Bone et al. [4] who reported an identification rate of 70.5%. For a summary of all results of the ISSC see [14].

There is also the question of the performance of humans in detecting alcoholic intoxication solely by speech, and what strategies they may use to fulfill the task. Here again previous findings vary. A relatively high discrimination rate (82%) was reported in [7], but only for speakers whose BAC was above 0.1%. For those below, intoxication was only discriminated in 54.2% of the cases. In a forced choice identification task [9] reports an accuracy rate of 61.5%; the speech material of eight male speakers was judged by 44 listeners. In [11] a discrimination test was conducted on the data of 16 speakers (8f, 8m, BAC of 0.05% - 0.142%) of the ISSC benchmark[2]. The discrimination test revealed a high detection rate (47 listeners reached an average accuracy of 71.65%), but the small number of speakers envolved suggests that this result might be statistically unreliable.

The aim of the present study is to test whether these results hold for a larger population of speakers and listeners, how speakers with lower BAC than 0.05% behave, whether f0 is a major cue for the detection of intoxication for humans, and finally if there are possible other (hidden) factors that might influence the outcome of this type of perception experiment.

The outline of the paper is as follows: Section 2 and 3 describe the experimental setup and the speech data. Section 4.1 presents the most prominent results regarding the performance of the listeners and the varying perceptibility across different speakers and speech styles. In section 4.2 the possibility of f0 as a major perceptual cue is tested, and all results are discussed in Section 5.

2. Speech data

The data used in the perception test are taken from the Alcohol Language Corpus (ALC) which comprises recordings of intoxicated and sober speech of 162 German speakers in

[1] In this study only speakers with a blood alcohol concentration (BAC) higher than 0.05% were analysed, resulting in a total of 148 speakers.

[2] The perception test had a slightly different design. In the ISSC speakers with BAC $< 0.05\%$ were treated as sober whereas in the perception test only speakers with BAC = 0.0% were treated as sober.

three different speech styles: read speech (numbers, addresses and tongue twisters), spontaneous speech (image descriptions (monologues), question answering, e.g. "Which was the best gift you've ever received?"), and command and control (C&C) speech (typically used with a vehicle navigation and edutainment system). The read and C&C items are the same for all speakers. Each speaker was recorded sober and with one self-selected BAC level varying from 0.023% to 0.175% (median is 0.089%). For further information about the recordings and the ALC see [12].

For the perception experiment, the same 16 stimuli pairs (8 each of read and C&C speech style) were selected per speaker. Another 8 stimuli pairs in spontaneous speech style were manually cut (average length 5s) and matched according to content across intoxicated and sober speech. Laughter or slips of the tongue were excluded from the spontaneous stimuli pairs as far as possible. This procedure results in 24 discrimination pairs per speaker; the mean duration of one pair of stimuli varies from 0.8s to 15.8s (median is 3.9s). One speaker of the ALC had to be excluded, because it was not possible to extract long enough stimuli of spontaneous speech without laughter. Also part of the perception test are the recordings of a control group of 20 speakers which were recorded in the same experimental setup as the main corpus, but were sober in both conditions. In total this results in $(161 + 20) * 24 = 4344$ stimuli pairs.

3. Perception test

In the forced-choice discrimination test one pair of stimuli of the same speaker was presented in random order (sober and intoxicated, sober and sober for the control group respectively). The listeners had to decide in which of the recordings the speaker was intoxicated. They were allowed to listen up to five times to each stimulus. Because it is not possible for each listener to judge 4344 pairs of stimuli, an experimental design was chosen in which each speaker was heard once by each listener, hence each listener heard 181 pairs of stimuli. 24 different stimuli sets with balanced speech styles and one pair of stimuli per speaker were automatically generated. Each set was presented to three listeners resulting in a total number of 72 (36 female and 36 male) listeners, and each pair of stimuli was judged three times by three different listeners.

Listeners were native German speakers aged between 20 and 36 (median is 23).

4. Results

4.1. Discrimination and detection rates

Figure 1 illustrates the performance of the listeners. The average discrimination rate is 63.1% and varies among the listeners from 52.8% to 76.4%. There was no significant difference between the performance of male and female listeners and we found no significant difference between the ability of female listeners to judge male speakers compared to female speakers and vice versa. Table 1 shows the individual cross-gender results.

The detection rate across speakers varies even more from 40.3% to 87.5% (Figure 2). The intoxicated speech of one speaker was only recognized in about 40% of the trials; in this case even the sober speech of the speaker was more likely to be judged intoxicated than the intoxicated speech.

Presumably the distribution of the discrimination rate of the listeners (Figure 1) is more homogeneous because all listeners

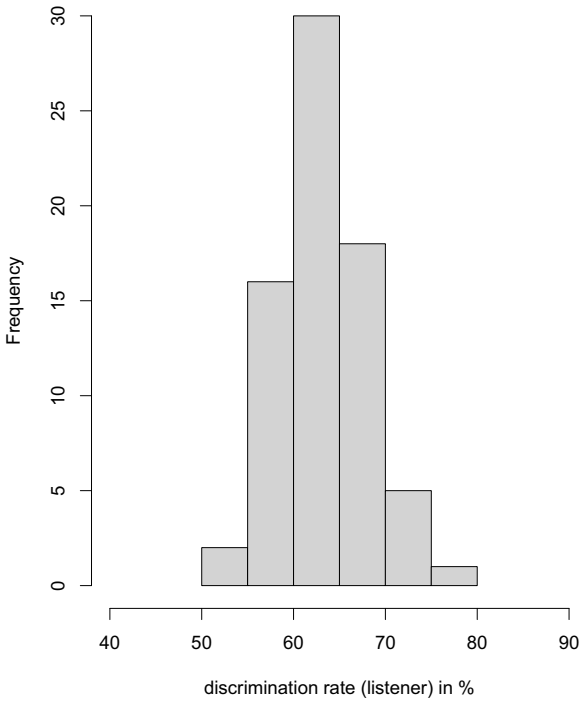

Figure 1: *Histogram of **listeners** discrimination rates*

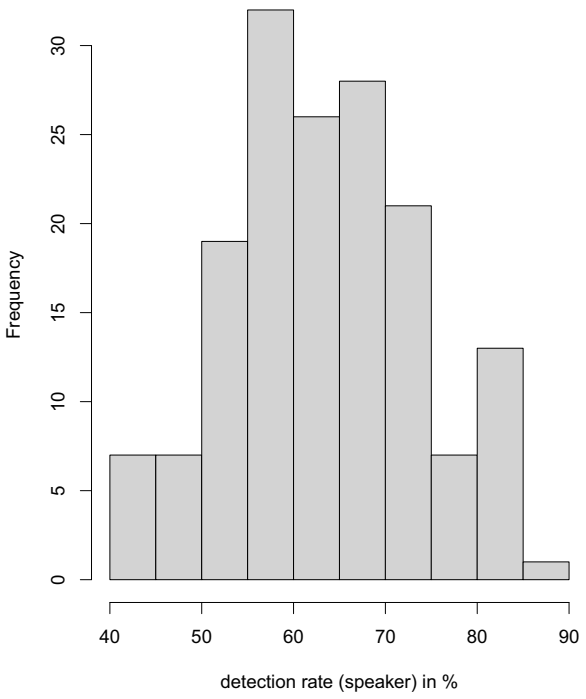

Figure 2: *Histogram of **speakers** detection rates*

Table 1: *Cross-gender discrimination rates (differences not significant)*

MF ← M	MF ← F	M ← MF	F ← MF
63%	63.3%	63.7%	62.5%
M ← M	**M ← F**	**F ← M**	**F ← F**
63.4%	64%	62.5%	62.5%

M = male, F = female speakers/listeners, ←= are judged by

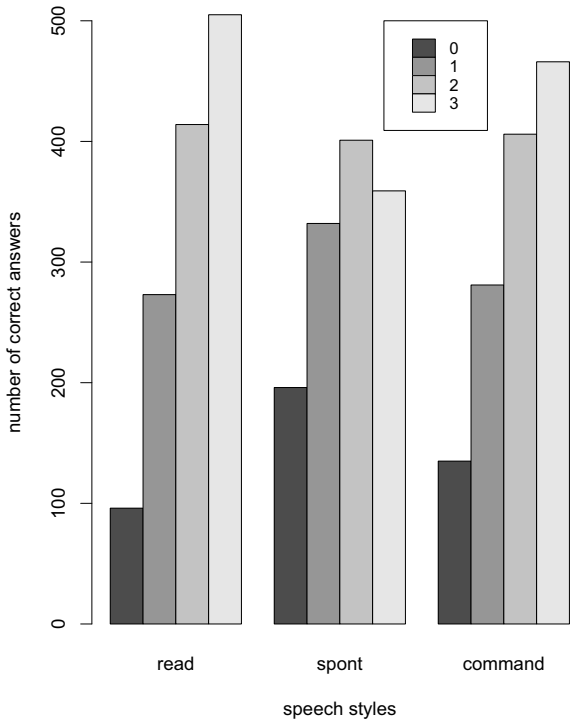

Figure 3: *Performance of discrimination: number of correct answers for pairs of stimuli across speech styles*

performed under the same conditions, whereas speaker conditions varied because of their different degree of intoxication.

The results of the control group (with both stimuli containing sober speech recorded in two experimental setups), showed that the listeners chose randomly between the two recordings (discrimination rate: 49.2%). It follows that there are no hidden factors in the different recording setups that bias listener judgements.

Intoxication in read speech was recognized best (67.7%), spontaneous speech showed the worst accuracy rate of 57.2%, C&C speech reached a total of 64.5% (all differences significant with $p < 0.005$). As mentioned above, each pair of stimuli was judged by three different listeners, so it was possible to be judged correct from zero to three times. To illustrate the discrimination performance across speech styles Figure 3 shows the number of correct answers for each single pair of stimuli, separated by speech style. For the spontaneous speech style we see a significantly lower number of pairs of stimuli which were judged correctly by all three listeners.

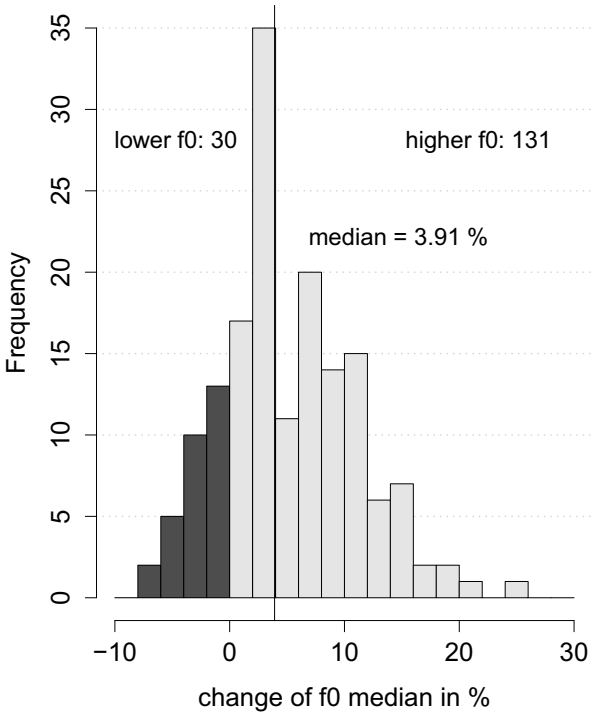

Figure 4: *Histogram of relative changes of the median f0 between sober and intoxicated across speakers; a positive value on the x-axis denotes a rise of f0 with intoxication*

4.2. Fundamental frequency (f0) and perception

To test the hypothesis that f0 is an important cue for the perception of intoxication we calculated the relative difference of f0 medians between sober and intoxicated stimuli for each speaker (Figure 4) which was raised on average by approximately 4%. F0 medians were higher for intoxicated speech for 81.4% of the speakers.

Figure 5 shows the correlation between these relative f0 median changes (Figure 4) and the speaker specific detection rate (Figure 2). A tendency for better detection rates for speakers who show a bigger change in f0 can be seen, though the correlation is weak (r = 0.23). A mixed effect model analysis [2] showed a significantly higher probability to choose the intoxicated stimulus for pairs of stimuli in which the f0 in the intoxicated stimulus is higher.

We also calculated the general preference of each listener to choose the stimulus with higher f0 as being intoxicated (a speaker with 100% preference would always choose the stimulus with the higher f0 as being intoxicated). Most listeners (66 out of 72) more or less followed this strategy (preference above chance ranging from 51 to 68%). In Figure 6 the correlation between the discrimination rates and these individual preferences is given. Although the correlation is weak (r = 0.42), listeners who have a higher preference to choose the stimulus with higher f0 tend to have more success.

5. Discussion and conclusion

The result of the perception experiment, an average discrimination rate of 63.1%, shows that the discrimination rate of 71.65%

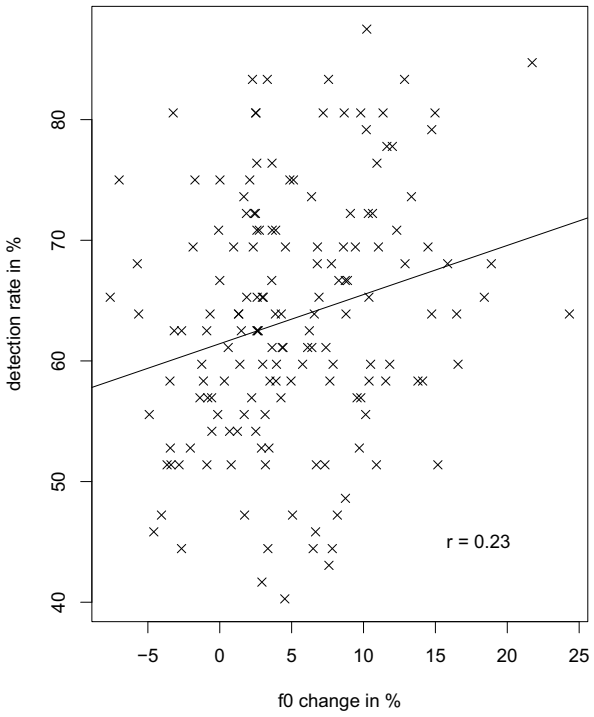

Figure 5: *Correlation between f0 and detection rates (speaker)*

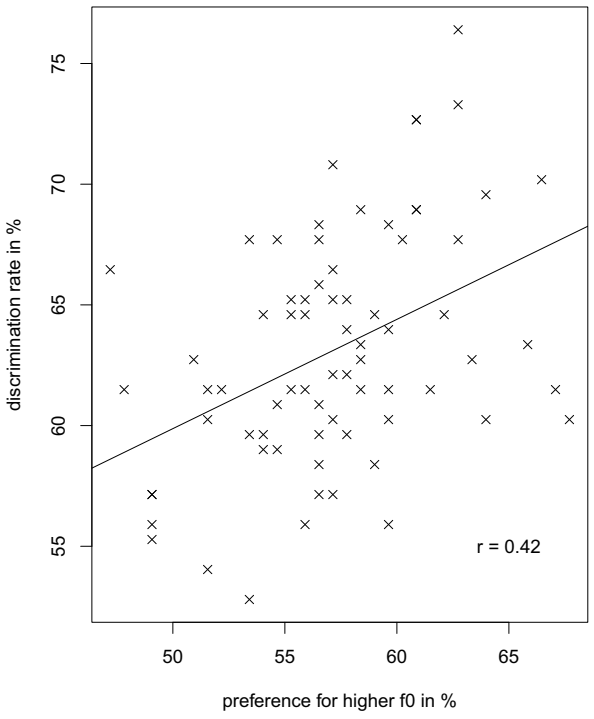

Figure 6: *Correlation between preference for higher f0 and discrimination rates (listener)*

reported in [11] with 16 speakers of the ALC and 47 listeners cannot be replicated with a larger set of data. Aside from neglecting speakers with less than 0.05% BAC in [11] paralinguistic cues such as laughter and speech errors may have been used by the listeners to reach such a high discrimination level. This implies that the detection of intoxication from the speech signal alone – without the possibility of incorporating para-linguistic cues – is harder than assumed. In this sense the achievement of Bone et al. ([4]) in the ISSC, an identification rate of 70.5%, is even more remarkable.

Nevertheless, the overall discrimination rate in the present study is still well above chance and there do not appear to be hidden factors within the different recording situations. The analysis of the speech style revealed that read speech was recognized best (67.7%) while spontaneous speech was recognized worst (57.2%). This may be due to the fact that cognitive demands on reading are higher for the speaker than on spontaneous speech, especially when reading tongue twisters or unknown addresses as was the case in the ALC. This higher cognitive demand may lead to difficulties for speakers in masking intoxication and make it easier for listeners to recognize. The discrimination rate for C&C speech is with 64.5% nearly as good as for read speech. This is a promising result for automatic detection within the car, because C&C speech in the ALC was partly not read from screen but rather prompted through a game situation resulting in very realistic speech samples. The results also demonstrate that the perceptibility of intoxication is both speaker and listener dependent, but the variations can not be derived from gender differences (as reported in [11]).

The (weak) correlation between f0 changes and the speaker specific detection rates shows that a higher f0 in the intoxicated stimulus facilitates the perception of intoxication. In line with this result we also see that the preference of listeners to choose the stimulus with higher f0 correlates positively with their individual performance – but f0 can not be regarded the sole perceptual cue for intoxication. A simple algorithm, which classifies the stimulus with the higher f0 median as the intoxicated one, would achieve a higher discrimination rate in 82% of the cases[3]. It seems that the listeners are not aware of how good a change in f0 as a cue really is.

To summarize, the results of this and previous studies suggest that automatic detection of intoxication solely based on the speech signal seems promising. In contrast, humans seem not to be able to detect alcoholic intoxication from acoustic features alone; humans do not exploit the simple acoustic cue of the rise of f0 but rather seem to rely on para-linguistic cues such as speech errors or laughter.

There is still the issue of those speakers that can mask their intoxication almost perfectly or show acoustic feature changes in the opposite direction (e.g. lowering their f0 instead of raising). Future work could use a different paradigm, in the case where a sufficient amount of sober speech is available for each speaker, which would lead to a speaker-dependent detection schema and would circumvent these problems.

6. Acknowledgements

This work was partly supported by the DFG, contract number SCHI1117/1-1. We would like to thank the ALC team for providing the speech data and the orthographic transcription.

[3]Please note the difference of discrimination and identification rate here: discrimination is an easier task than identification.

7. References

[1] Aldermann, G. A., Hollien, H., Martin, C., DeJong, G. (1995), "Shifts in fundamental frequency and articulation resulting from intoxication". In: Journal of the Acoustical Society of America 97, pp. 3363-3364.

[2] Baayen, R. H. (2008), "Analysing Linguistic Data: A Practical Introduction to Statistics Using R". Cambridge University Press, Cambridge, pp. 263-328.

[3] Baumeister, B., Heinrich, C., Schiel, F. (2012), "The influence of alcoholic intoxication on the fundamental frequency of female and male speakers". In: Journal of the Acoustical Society of America 132, pp. 442-451.

[4] Bone, D., Black, M. P., Li, M., Metallinou, A., Lee, S., Narayanan, S. S. (2011), "Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors". In: Proceedings of the Interspeech 2011, Florence, Italy, pp. 3217-3220.

[5] Cooney, O. M., McGuigan, K. G., Murphy, P. J. P. (1998), "Acoustic analysis of the effects of alcohol on the human voice". In: Journal of the Acoustical Society of America 103, pp. 2895-2895.

[6] Hollien, H., DeJong, G., Martin, C. A., Schwartz, R., Liljegren, K. (2001), "Effects of ethanol intoxication on speech suprasegmentals". In: Journal of the Acoustical Society of America 110, pp. 3198-3206.

[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 84, pp. 929-935.

[8] Künzel, H. J., Braun, A. (2003), "The effect of alcohol on speech prosody". In: Proceedings of the ICPhS2003, Barcelona, Spain, pp. 2645-2648.

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

[10] Pisoni, D. B., Hathaway, S. N., Yuchtman, M. (1985), "Effects of alcohol on the acoustic-phonetic properties of speech: Final report to GM research laboratories". In: Research on Speech Perception Progress Report No. 11 (Indiana University, Bloomington, IN), pp. 109-171.

[11] Schiel, F. (2011), "Perception of Alcoholic Intoxication in Speech". In: Proceedings of the Interspeech 2011, Florence, Italy, pp. 3281-3284.

[12] Schiel, F., Heinrich, C., Barfüßer, S. (2012), "Alcohol Language Corpus: The first public corpus of alcoholized German speech". In: Language Resources and Evaluation 46(3), pp. 503-521.

[13] Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J. (2011), "The INTERSPEECH 2011 Speaker State Challenge". In: Proceedings of the Interspeech 2011, Florence, Italy, pp. 3201-3204.

[14] Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., Weninger, F., Eyben, F. (2012), "Medium-Term Speaker States – A Review on Intoxication, Sleepiness and the First Challenge". In: Computer Speech and Language, in print.

[15] Sobell, L. C., Sobell, M. B., Coleman, R. F. (1982), "Alcohol-induced dysfluency in nonalcoholics". In: Folia Phoniatrica 34, pp. 316-323.

[16] Watanabe, H., Shin, T., Matsuo, H., Okuno, F., Tsuji, T., Matsuoka, M., Fakaura, J., Matsunaga, H. (1994), "Studies on vocal fold injection and changes in pitch associated with alcohol intake". In: Journal of Voice 8, pp. 340-346.