Rhythm and Formant Features for Automatic Alcohol Detection

Florian Schiel, Christian Heinrich, Veronika Neumeyer

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

schiel|heinrich|vroni@bas.uni-muenchen.de

Abstract

Two speech feature sets, RMS rhythmicity and formant frequencies F1-F4, are analyzed for their ability to distinguish alcoholized from sober speech. We describe the statistical framework based on the Alcohol Language Corpus (ALC), including other factors such as gender, age and speaking style, and its application to our case. Rhythm features are calculated using a new method based solely on the short-time energy function; formant features are derived using the standard formant tracker SNACK. Our findings indicate that 3 rhythm and 3 formant features have a high potential to detect intoxication within the speech data of a subject. We also tested the hypothesis that vowels are more centralized in the F1/F2 space for alcoholized speech, but found that, on the contrary, subjects tend to hyperarticulate when being tested for intoxication.

Index Terms: alcoholized speech, rhythm, formants, Alcohol Language Corpus, BAS, vowel centralization

1. Introduction

Can alcoholic intoxication be reliably detected from the speech signal alone?
This question has been debated in the field of forensic phonetics for a long time. The answers of earlier studies are inconclusive (e.g. [1],[2],[3],[4],[5],[6],[7]): a number of speech features have been investigated with regard to alcoholic intoxication (AI) and with differing, sometimes contradicting results, most likely depending on the empirical data analyzed. As for technical applications, to our knowledge nobody has yet presented an algorithm which automatically derives an estimate of the blood alcohol concentration (BAC) from the speech signal.

Were such an algorithm to be found, it would potentially help not only forensic sciences but could also allow pre-emptive alcohol testing in the automotive environment, as discussed in [9].

In the years 2007 to 2010 the Bavarian Archive for Speech Signals (BAS) located at the Ludwig-Maximilians-Universität, München collected alcoholized and sober speech of 77 female and 84 male speakers. The Alcohol Language Corpus (ALC) is publicly available so that other interested researchers may replicate our findings or perform their own studies on alcoholized speech[1]. ALC comprises different speech styles: read speech, list style, semi-spontaneous (picture task), spontaneous speech (dialogue) and situational prompted commands from the automotive environment (see [11] for details on Situational Prompting). The speech content covers simple digit strings (telephone/credit-card numbers), word lists, addresses, tongue twisters, picture descriptions, interview style answering and free dialogue about casual topics. Each speaker provided roughly 6min of speech in AI and 12min in sober condition. In addition, 20 speakers were recorded a third time in the exact same situation as in the AI recordings but sober to provide a control group for statistical reference analysis. The speech signals (close and distant microphone) were recorded in the same car environments (two different models) for the sober and intoxicated condition. All recordings were manually transcribed and tagged for para-linguistic events. An automatic phonemic segmentation and labeling into the German SAM-PA phonetic symbol set is provided for all recordings[2]. Meta information about speaker characteristics (age, dialectal origin, height, weight etc.) and recording conditions are provided to allow statistical testing for influencing factors other than alcoholization. For a more detailed description of ALC see [10].

Based on these empirical data we investigated whether two sets of features could distinguish sober from alcoholized speech. One feature set concerns the rhythm of speech, that is features that are derived from the dynamics of the sound pressure energy (the alternation of relatively loud and quiet speech parts). The other set of parameters concerns the resonance of the vocal tract, that is the location of the formants and their respective relation to long-time averaged centroids in the two-dimensional F1-F2 space (also known as the 'vowel space').

The paper is structured as follows: the next section discusses some general issues when dealing with individual speaker features and the statistical framework applied. Sections 3 and 4 describe two different feature sets under investigation, how they were obtained automatically from the speech signal and the statistical results.

2. Statistical Framework

When dealing with slowly changing speaker conditions like tiredness, stress and also alcoholization we hardly ever observe general 'laws' valid for all speakers on how the linguistic and phonetic features of the speech signal behave. We rather find that speaker groups or even individual speakers have their own idiosyncratic way of expressing these states. For example, if we measure the average fundamental frequency f_0 over an utterance of alcoholized speech and compare this to sober speech of the same speaker, we observe an overall tendency of the female speakers to rise their f_0 significantly independent of the measured AI, but we also find a small group of female speakers who behave the other way round. Male speakers behave differently and are more inconsistent than female speakers, but nearly all of them increase or decrease f_0 when being intoxicated ([9]).

When evaluating linguistic or phonetic features derived from the speech signal concerning their potential to distinguish between sober and alcoholized speech, we therefore look for

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

[2] Provided by the MAUS system ([8]).

relative changes within the data of an individual speaker rather than the whole population. For the application of these findings in forensic sciences or preemptive alcohol detection this is not a problem since usually there is plenty of sober speech available for the subject under consideration.

In the following sections we discuss our test statistics, the calculation of correlations and pattern recognition frameworks.[3]

2.1. Test Statistic

Let the feature $f[s(t)]$ be derived from the speech signal $s(t)$. Then we call $f[s(t)]$ the dependent test variable and the following factors the independent test variables which are further sub-categorized into *within speaker factors*, alcoholization (alc), speech style (sty), vowel class (vow), and *between speaker factors*, age (age) and gender (sex):

alc	alcoholized (a) / non-alcoholized (na)	within
age	young (21-36) / old (36-65)	between
sex	female (F) / male (M)	between
sty	read (r) / spontaneous (s) / command (c)	within
vow	/a:/, /i:/, /u:/	within

To test the dependent variable we apply either a repeated measures ANOVA (RM-ANOVA) or a mixed model (MM) with the speaker ID as the *random variable* to filter out between speaker variability.

2.2. Correlation

Since the ALC corpus does not contain speech data of different BAC levels *produced by the same speaker*, a test for linear or non-linear correlation between the feature $f[s(t)]$ and BAC within the data of one speaker is not possible. To test for linear dependencies across a group of speakers the measured features have to be normalized accordingly. Thus individual speaker behavior, in the sense of different linear models, cannot be captured within our experimental setting[4].

2.3. Model Fitting and Prognosis

A first step to test potential features for alcohol detection is to fit a logistic regression model for each individual speaker and estimate the prognosis with regard to alcoholization (on the same data). Based on 121 speakers from ALC and using MFCC as speech features, such a model yields an average prognosis rate of 77% ([14]). However, the prognosis rates across speakers vary from 64-95%, indicating that alcohol detection is indeed a highly speaker dependent task. Interestingly we found no dependencies to gender, age or the actual blood alcohol concentration in this model.

2.4. Pattern Recognition

The classical approach to divide the data set into two disjunctive sets of training and test speakers (using for instance a leave-one-out schema) does not work in this case for the same reasons discussed above. The solution is to divide the corpus into two sets of development and test speakers first and then divide the data sets of each individual test speaker into four parts:

sober train	sober test
alcoholized train	alcoholized test

The method and heuristics of the feature extraction are then calculated using the data of the development speakers only. Finally the derived method is applied to the training data sets of each individual test speaker to achieve a speaker dependent model, which then in turn can be tested using the test data sets of the speaker. By this we yield individual scores and failures for each speaker which can then be added up to calculate overall performance scores.

3. Rhythm Features

The short-time sound pressure energy of speech (RMS) represents a sequence of alternating relatively loud and quiet parts and thus can be used to describe rhythmicity. Also RMS has the advantage that it can be easily derived from the speech signal without any linguistic information such as the CVC sequence.

We calculate RMS values for the ALC data using a Blackman window with 200 ms size and a 20 ms shift. All RMS data are then normalized to the total mean within each utterance to yield comparable measurements. From the obtained normalized RMS we retrieve a sequence of data points in the RMS trajectory at local minima and maxima, whose RMS values are below or above the mean RMS of the total recording. This sequence naturally contains clusters of minima and maxima as well as single minima and maxima. To obtain a continuous alternating sequence in the form *min-max-min-max...* we derive exactly one representative minimum/maximum from each cluster, where the representative minimum/maximum is simply defined as the one with the lowest/highest value within the cluster. The resulting *min-max* sequence forms the base material for a set of measures describing what we call *RMS rhythmicity*.

For this study we concentrate on measures based on the RMS maxima only. For all recordings belonging to the same factor combination (see section 2.1) we calculate the mean (*A*) and standard deviation[5] (*B*) of the *time delay between successive maxima*, the median and quarter quantile distance of the *differences of the RMS values between successive maxima* (*C*,*D*), the same for *absolute differences* (*E*, *F*), and finally the quarter quantile distance of the *relative distance of maxima and minima to the normalized mean RMS* (*G* and *H*). We also adopted the *normalized Pairwise Variability Index* (*nPVI*) as introduced by Grabe and Low [15] to describe the average durational difference between successive periods of time between maxima (*I*). All those measures were only applied to data points within utterance boundaries.

Measures *A* and *B* reflect the timing and regularity of the occurrence of maxima: a higher mean would be an indicator for a lower speech-rate and a larger SD could indicate an increased irregularity in the speech rhythm. *C*,*D*,*E* and *F* reflect the change in the energy dynamics of the speech signal whereas *G* and *H* reflect the energy dynamic itself. Measure *I* represents changes in the sequential structure of the speech signal based on the RMS maxima.

Table 1 shows the differences of the measures between sober and alcoholized speech, giving the direction of change and the speech style for which differences are significant (other factors did not interact). All measures rise with alcoholization and differ significantly between sober and alcoholized speech (RM-ANOVA, p<0.001). The statistical analysis yielded no dependency to the factor *sex* but significant interactions with

[3] Note that this contribution deals with test statistics only; nevertheless we discuss correlation and pattern recognition for the benefit of other researchers using the ALC speech corpus.

[4] It is however possible to compensate for simple biases in the data, e.g. the fundamental frequency of female and male speakers.

[5] For *A* and *B* we apply the less robust mean and SD since the measurements are time-discrete with a 20ms interval; the remaining (continuous) measures are estimated by the more robust median and quarter quantile distance.

Table 1: *Significant differences for measures A to I based on the data of 128 speakers for the three speech styles.* $\Uparrow$ *: rises with alcoholization.*

	A	B	C	D, E
alc	$\Uparrow$	$\Uparrow$	$\Uparrow$	$\Uparrow$
score	p<0.0001	p<0.0001	p<0.001	p<0.0001
sty	r,s	r,s	c	r,s,c

	F	G	H	I
alc	$\Uparrow$	$\Uparrow$	$\Uparrow$	$\Uparrow$
score	p<0.0001	p<0.0001	p<0.0001	p<0.0001
sty	r,s,c	r,s	r,s	r

the speech style (*sty*)[6]: the last line in Table 1 indicates those speech styles for which the feature was found to be significantly different.

As an example the box-plots in Figure 1 show the difference between alcoholized and sober speech, averaged for both genders and the three styles of speech for the measure *E* (median of the absolute value of the differences of the RMS values between maxima). The tendency of *E* to increase from sober to alcoholized speech for all conditions is clearly visible.[7]

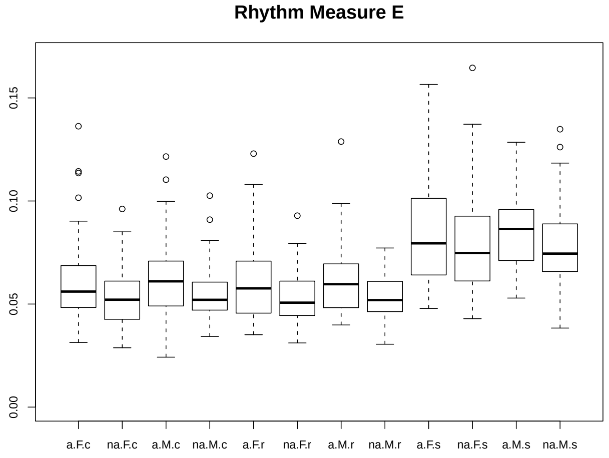

Figure 1: *Box-plots of measure E for female (F) and male (M) speakers, the speech styles (c,r,s) and alcoholized (a) vs. sober (na) speech.*

4. Formant Features

Formants are the primary resonances of the vocal tract caused by different geometric configurations of the articulating organs. They are used by listeners to distinguish phonemic minimal pairs, for instance vowel classes and transitions from vowels into different consonants. A formant is defined by its frequency (or position), amplitude and bandwidth; for this study we concentrate on the frequencies of the first four formants $F1 - F4$.

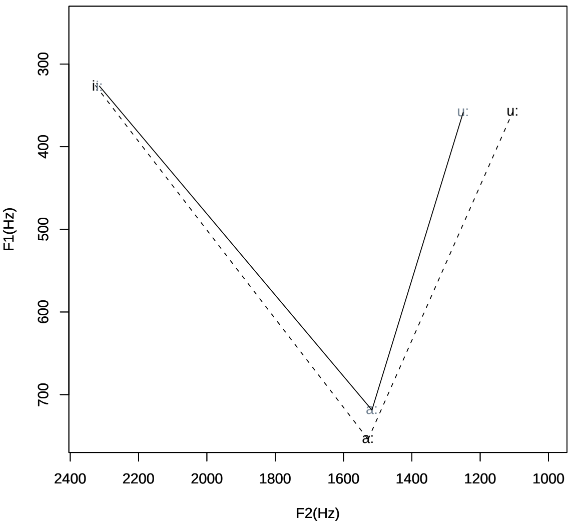

Figure 2: *Vowel triangle in the F1/F2 plane. The solid line combines vowel centroids of sober speech, the dashed line those of alcoholized speech.*

Formant measurements usually require phonetic segmentation and labeling (to determine the location of the vowel centers) and a manual correction of the automatically detected formant trajectories. Since we are looking for an automatic method to distinguish sober from AI speech, we first apply the automatic segmentation tool MAUS ([8]) and then the formant tracker of the SNACK package[8] without any subsequent manual correction to the speech recordings of 131 speakers (67 f + 64 m). The formant tracker is configured to track the first 4 formants and uses a nominal $F1$ frequency of 531Hz for male and 595Hz for female speakers[9]. We then derive the median F_m and quarter-quantile distance F_{qq} of the formant frequencies $F1 - F4$ from the 30% mid-section of each vowel segment, average these 8 values over all vowels within one factor combination (*speaker(sex), alc, vow, sty*, see section 2.1) and test them for significant differences using repeated measure ANOVA (random factor is the speaker).

$F1_m$ increases significantly in alcoholized vs. sober speech (F=16.5, p<0.0001) but post-hoc Tukey tests show that this is only significant for the vowel /a:/ (although the tendency for /i:/ and /u:/ is the same). $F2_m$, $F3_m$ and $F4_m$ do not differ significantly. $F1_{qq}$ shows a weak significant increase (F=4.05, p=0.046) but only for the vowel /u:/. $F3_{qq}$ has a non-significant tendency (F=3.84, p=0.051) and finally $F4_{qq}$ (F=8.61, p=0.0040) increases significantly for all factor combinations.

To produce clearly distinguishable vowels the articulators (lips, tongue, jaw) have to move into the correct target positions, which requires time and effort. It is therefore reasonable to predict that fast, spontaneous or blurred speech will show

[6] Factors *age* and *vow* have not been tested here.

[7] Note that the box-plots show data averaged over a large number of speakers and are therefore not as distinguishable as in the data of a single speaker.

[8] http://www.speech.kth.se/snack/

[9] Based of average vocal tracts lengths of 16cm for male and 14,3cm for female speakers given by [13].

vowel formants that are not in the ideal target positions (see for instance Lindblom's *hypo-hyper speech continuum* [12]). A common way to demonstrate the separability of vowel classes is to plot measured formant positions into the two-dimensional F1-F2 vowel space which is roughly spanned between the Cardinal Vowels /i/, /a/ and /u/ (vowel triangle).

Based on these assumptions we hypothesize that the frequencies of $F1$ and $F2$ will be *less distinguishable (less apart)* for the tense, lexically accented vowels /a:/, /i:/ and /u:/ in alcoholized speech than in sober speech.

Let $\vec{C}_f$ and $\vec{C}_m$ be the centroids of all formant frequencies $F1_m$ and $F2_m$ of female and male speakers in the two-dimensional F1-F2 space. Then for a given sample of speech (= vowels within a factor combination) the amount of 'being apart' ED can be measured by the sum of the Euclidean distances from the vowel centroids in this sample $\vec{C}_a$, $\vec{C}_i$ and $\vec{C}_u$ to the overall centroid $\vec{C}_{f|m}$.

$$ED = \sum_{x=a,i,u} \sqrt{\sum_{k=F_1,F_2} (C_{x,k} - C_{f|m,k})^2}$$

Figure 2 shows the vowel triangles of female speakers for sober (solid) and alcoholized speech (dashed). We clearly see a significant change in ED (F=20.75, p<0.0001 for both genders), which contradicts our hypothesis, because sober speech appears to be more centralized than alcoholized speech. Post-hoc Tukey tests show that only female speakers and mainly the vowel /u:/ contribute to this significant change.

One possible explanation for this surprising result is that female speakers try to compensate for their impaired state by using hyper-articulated speech. This makes sense since the recorded subjects knew that they are being tested and tried to act as naturally as possible to 'pass the test'. Assuming that the rhythm measure A (section 3) is an inverse representation of the overall speaking rate (which is lowered for alcoholization) another explanation could be that the speakers simply have more time to reach the articulatory targets.

Table 2 summarizes the formant measurements that distinguish alcoholized from sober speech with regard to direction of change and factors for which the change is significant. Only one feature, the quarter-quantile distance of the forth formant $F4_{qq}$, shows a highly significant within-speaker increase for alcoholized vs. sober speech in both genders, for all vowel classes and speaking styles. Other significant features are either dependent on gender, vowel class or both.

Table 2: *Significantly different formant measurements based on the data of 131 speakers.* ⇑ *: rises with alcoholization.*

	$F1_m$	$F1_{qq}$	$F4_{qq}$	ED
alc	⇑	⇑	⇑	⇑
score	p<0.0001	p=0.046	p=0.004	p<0.0001
sex	F,M	F,M	F,M	F
vow	/a:/	/u:/	/a: i: u:/	/u:/
sty	r,s,c	r,s,c	r,s,c	r,s,c

5. Conclusion

We presented a statistical framework to investigate the ability of features derived from the speech signal to distinguish alcoholized from sober speech. Based on a subset of 128/131 speakers we analyzed 9 rhythm and 9 formant features automatically derived from the speech signal. The rhythm features (RMS rhythmicity) were based on a simple RMS measurement and peak picking algorithm, while formant frequencies were calculated using SNACK. Three formant and three rhythm features show highly significant differences independent of gender and speaking style. Contrary to our expectation vowel formants did not centralize in the F1/F2 space for alcoholized speech but rather decentralize. We explain this by the increased compensatory hyper-articulation of alcoholized subjects when being tested for intoxication. Other experiments based on MFCC features ([14]) using the same database indicated an average prognosis rate of 77%. Whether these baseline results may be improved by incorporating the feature sets described here will be the subject of future work.

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

[9] Schiel F, Heinrich Chr (2009): Laying the foundation for In-cat Alcohol Detection by Speech. In: Proceedings of the Interspeech 2009, Brighton, United Kingdom, pp. 9983-986.

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

[11] 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.

[12] Lindblom B (1990) Explaining phonetic variation: A sketch of the H&H theory. In: Hardcastle W J, Marshal A (eds) Speech Production and Speech Modelling. Amsterdam:Kluwer, pp. 403-439.

[13] Menard L, Schwartz J-L, Boe L-J, Aubin J (2007) Articulatory-acoustic relations during vocal tract growth for French vowels: analysis of real data and simulations with an articulatory model. Journal of Phonetics 35 (2007), pp. 1-19

[14] Kobayashi C, Schober L, Straub A, Stuckart C (2009) Klassifikation von Sprachsignalen in alkoholisiert vs. nüchtern. Project Report Computational Statistics, LMU München.

[15] 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.