The influence of alcoholic intoxication on the short-time energy function of speech

Christian Heinrich[a] and Florian Schiel

Institute of Phonetics and Speech Processing, Ludwig-Maximilians-Universität, Schellingstrasse 3, 80799, Munich, Germany

(Received 26 April 2013; revised 19 March 2014; accepted 27 March 2014)

This study investigates rhythmic features based on the short-time energy function of speech signals with the aim of finding robust, speaker-independent features that indicate speaker intoxication. Data from the German Alcohol Language Corpus, which comprises *read*, *spontaneous*, and *command&control* speech uttered by 162 speakers of both genders and various age groups when sober and intoxicated, were analyzed. Energy contours are compared directly (Root Mean Squared Error, statistical correlation, or the Euclidean distance in the spectral space of the contour) and by parameterization of the contour using the Discrete Cosine Transform (DCT) and the first and second moments of the lower DCT spectrum. Contours are also analyzed by Principal Components Analysis aiming at fundamental "eigen contour" changes that might encode intoxication. Energy contours differ significantly with intoxication in terms of distance measures, the second and fourth DCT coefficients, and the first and second moments of the lower DCT spectrum. Principal Components Analysis did not yield interpretable "eigen contours" that could be used in distinguishing intoxicated from sober contours. © *2014 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.4870705]

PACS number(s): 43.70.Dn, 43.72.Ar [CHS] Pages: 2942–2951

I. INTRODUCTION

Intoxicated speech is assumed to be prone to a variety of feature changes compared to sober speech (for an overview see Chin and Pisoni, 1997). Statistically valid knowledge about the nature of these feature changes under the influence are of interest to forensic phoneticians and might form the basis for automatic detection of intoxication in the speech of a driver voice-controlling his vehicle. The latter approach would require a combination of robust features that are as follows:

(1) they are either speaker-independent or speaker-dependent based solely on sober speech material of the speaker (see our discussion in Schiel, 2011), since usually there is plenty of opportunity to record the speaker's voice in a sober but not in an intoxicated state;
(2) they are language independent;
(3) they can be extracted automatically (i.e., requires no manual segmentation or labeling) from the speech signal recorded at a distance from the speaker's lips, as would occur for a microphone mounted on the dashboard of a vehicle;
(4) they are not sensitive to other speaker states such as pathological, emotional, or caused by stress; and
(5) they are not sensitive to background noise.

Previous studies about different linguistic and phonetic features regarding intoxication often lack enough speakers or deal with male speakers only and, therefore, are not likely to yield statistically robust results (e.g., Aldermann *et al.*, 1995; Behne *et al.*, 1991; Braun, 1991; Chin and Pisoni, 1997; Cooney *et al.*, 1998; Cummings *et al.*, 1995; Hollien *et al.*, 2001; Klingholz *et al.*, 1988; Künzel and Braun, 2003; Levit *et al.*, 2001; Martin and Yuchtman, 1986; Pisoni *et al.*, 1985; Sigmund and Zelinka, 2011; Sobell *et al.*, 1982; Trojan and Kryspin-Exner, 1968). Furthermore, the data on which these studies are based are not available for other researchers, so results cannot be replicated. Recently new studies based on the publicly available German Alcohol Language Corpus (ALC) (see Sec. III) provide statistically firm findings regarding phonetic standard features (Baumeister *et al.*, 2012; Heinrich and Schiel, 2011; Schiel, 2011; Schiel and Heinrich, 2009; Schiel *et al.*, 2010): in short, the most prominent phonetic features that change under the influence are a decreased speaking rate and raised fundamental frequency (Heinrich and Schiel, 2011; Baumeister *et al.*, 2012). Unfortunately, the same features are also prone to changes under stress (e.g., Hansen and Patil, 2007), by the Lombard effect (e.g., Folk and Schiel, 2011), and when the speaker's emotional state is one of anger, joy, or sadness (e.g., Mathon and de Abreu, 2007; Yildirim *et al.*, 2004). It is therefore interesting to investigate other features for changes caused by intoxication and possibly in the long term combine multiple potential features for a more robust and speaker-independent classification system for intoxication.

The present study focuses on rhythmic properties based on the short-time energy function of intoxicated and sober speech, in order to clarify if there are measurable differences between the two. This goal is pursued using three different approaches:

(1) direct distance measures between two energy contours;
(2) interpretable parameterization of contour shapes; and

[a] Author to whom correspondence should be addressed. Electronic mail: heinrich@phonetik.uni-muenchen.de

(3) principal components analysis (PCA) of a set of energy contours to identify dominant prosodic "eigen shapes" whose PCA scores can then be treated as features.

It is not the aim of this study to predict the blood alcohol concentration (BAC) from a single feature or a combination of rhythmic features but rather to clarify which automatic feature extraction techniques, i.e., without any manual annotation of data, yield promising correlations to the (binary) intoxication state of the speaker.

The paper is organized as follows: The following section discusses existing studies regarding rhythmic features and intoxication and presents the motivation to analyze energy contours. Section III briefly describes the ALC database on which the present study is based. Section IV describes the methodology and results of the experiments regarding RMS contours. Finally, in Sec. V, the findings are summarized and discussed.

II. RHYTHM AND INTOXICATION

Schiel and Heinrich (2009) investigated rhythmic properties of intoxicated speech using rhythm metrics (established for the classification of languages into stress-timed, syllable-timed, and mora-timed languages, see e.g., Ramus *et al.*, 1999; Grabe and Low, 2002; Dellwo, 2006; Wagner and Dellwo, 2004). These metrics allow the analysis of the time patterns of a speech signal based on a given segmentation of the speech into vocalic, consonantal, and silence intervals. When read speech is used, almost all metrics show significant differences between intoxicated and sober speech. This holds for metrics that do not normalize for speech rate, like the vocalic/consonantal delta metric (ΔV, ΔC) of Ramus *et al.* (1999) or the raw Pairwise Variability Index (PVI) metrics (Grabe and Low, 2002), and also holds for metrics taking the varying speech rate into account, such as the variation coefficients of the delta metrics (e.g., $Varco\Delta V$, $Varco\Delta C$; Dellwo, 2006), the normalized PVI metrics (e.g., *nPVI-V*; Grabe and Low, 2002), or other PVI based metrics (e.g., *YARD*; Wagner and Dellwo, 2004).

However, the calculation of these rhythm metrics depends on a given phonetic segmentation, which is usually not available in practical applications (see also Levit *et al.*, 2001, p. 1). Preferably analysis should be based on features that can be derived directly from the speech signal without human intervention to allow the development of fully automatic detection algorithms. Since rhythmic events are expressed by changes of loudness (among other features such as fundamental frequency), the energy function is a promising basis for such features. Levit *et al.* (2001) pursued this approach using very basic prosodic features (positions, values, and regressions of fundamental frequency and energy function) in a classification system to detect alcoholic intoxication. Feature vectors were calculated from automatically segmented "phrasal units" (Levit *et al.*, 2001, p. 2), which basically resemble intonation phrases, although no attempt was made to model the specific contour form. In Schiel *et al.* (2010), the RMS signal was converted into a sequence of local RMS minima and maxima, which provide a basis for the calculation of RMS "rhythmicity" features. Although showing significant differences between intoxicated and sober speech, these features—like the rhythm metrics described earlier—reduce the prosodic information inherent in a speech signal (or the corresponding RMS signal) to only a few parameters per recording, which cannot explain those changes morphologically beyond a certain degree. For instance, a parameter influenced by syllable rate will not change, if syllable rate decreases in the first part of the sentence and increases in the second half. On the other hand prosodic features often are expressed in macroscopic movements, e.g., the decline of fundamental frequency in a declarative sentence.

For this study, the complete energy contour of declarative sentences was analyzed instead. The hypothesis is that changes in the rhythmic structure of speech caused by alcoholic intoxication are reflected in changes in the characteristic form of the RMS contours. To test this hypothesis, three general approaches are pursued: Direct comparison of contours to test several distance measures, the parameterization of contour forms based on discrete cosine transform, and principal components analysis (PCA) to learn more about basic contour form changes caused by intoxication.

III. SPEECH DATA USED IN EXPERIMENTS

For all analyses presented in this study, speech material from the Alcohol Language Corpus (ALC) was used. The ALC is a collection of sober and intoxicated speech of 162 German speakers, 77 female and 85 male. The data has been collected in southern parts of Germany and subsequently annotated between 2007 and 2010 by trained phoneticians at the Bavarian Archive for Speech Signals.

Volunteers for a controlled intoxication experiment were asked to drink up to a self-selected intoxication level between 0.03 and 0.15 vol.% blood alcohol concentration (BAC). After a waiting period of 30 min to stabilize the BAC as well as the breath alcohol concentration (BrAC), samples were taken by medical staff of the Institute of Legal Medicine, LMU Munich, to monitor the actual intoxication level (BAC and BrAC). Immediately after monitoring, participants were asked to deliver a speech sample, which took about 10–12 min. Intoxication levels are assumed to remain constant during that recording session. After a period of at least 14 days, the speaker delivered a second speech sample in the same environment and with the same dialogue partner, this time being sober. To cross-check for hidden factors that could have an influence on measurements, 20 (10 female, 10 male) of 162 speakers were recorded a third time, again being sober but otherwise under the same conditions as in the intoxicated recording. In the present study, however, this set of data functioned as reference points for relative distance measures (see Sec. IV). BAC levels for these 20 speakers in the intoxicated recording all exceeded 0.05%.

The corpus comprises different speaking styles: Read speech, spontaneous speech, and command and control speech as typically used in an automotive environment. The speech content covers simple digit strings (telephone/credit card numbers), word lists, addresses, tongue twisters, picture descriptions, read and elicited commands, interview style answers (mostly monologue), and free dialogue. Every speaker delivered roughly 6 min of intoxicated and 12 min of sober speech. Actual measured BAC levels (constant for every speaker and intoxicated recording session) across speakers ranged from 0.023% to 0.175%.

Recordings in the automotive environment were made using close and distant microphones, at distances of approximately 4 and 40 cm from the speaker's lips, respectively, in two different car types. Car type and other meta data (age group, dialectal origin, height, weight, BrAC, BAC, weather, etc.) were documented for every recording session to allow for statistical testing of influencing factors other than intoxication. For a more detailed description of ALC, see Schiel *et al.* (2012). A subset of ALC was used for the INTERSPEECH 2011 Speaker State Challenge (Schuller *et al.*, 2012).

For the present study 150 speakers (68 female, 82 male) were selected from the ALC. Their BAC levels while intoxicated exceeded 0.05%, which is the legal limit for driving in Germany. Every speaker read 19 utterances in both the intoxicated and sober states. Of the 150 speakers, 20 read the 19 sentences a third time, when sober (control recordings). The utterances were prompted in the same way for each recording. A total of $6080 \text{ utterances} = (150 \times 2 + 20) \times 19$ form the empirical data for this investigation.

IV. RMS CONTOURS—METHODOLOGY AND RESULTS

The following section describes how the discrete-time sampled raw RMS contour data were processed to form comparable data sets. Then three approaches to the analysis of RMS contours are presented: The direct comparison (distance measures), the analysis of RMS contour parameters as features (interpretable parameterization), and the analysis of PCA scores of dominant prosodic "eigen shapes."

A. RMS contour as a feature

The short-time RMS of a speech signal describes the dynamics of the sound pressure energy which can be seen as a sequence of relatively loud and quiet portions of the signal. This energy function is not only a (smoothed) estimate of the sound pressure level but also can be used to investigate rhythm parameters like speaking rate (Morgan and Fosler-Lussier, 1998; Pfau and Ruske, 1998; Dekens *et al.*, 2007; Heinrich and Schiel, 2011), syllable position (Xie and Niyogi, 2006), or so-called Rhythmicity Parameters to distinguish rhythmically different speech samples (Schiel *et al.*, 2010).

In this study, RMS values were calculated using the standard RMS algorithm of the ASSP tool kit within the Emu Speech Database System (Cassidy and Harrington, 2001). To ensure the capture of normal syllable rates (maximum ten syllables per second) while also smoothing out fine grained suprasegmental energy movements (like bursts in obstruents), signals were sampled using a Blackman window of 100 ms length and a window shift of 20 ms. Before any further processing, the logarithmically scaled RMS contours were all normalized by subtracting the mean RMS level from a given RMS contour. This results in an energy function alternating around 0 dB.

B. Distances measures

1. Distance measures—Method

Various distance measures can be calculated between two RMS contours; the measures described below have already been applied to f0 contours in Baumeister *et al.* (2012).

The first two distance measures, root mean squared error and correlation distance, described below, both require that the contours to be compared have the same number of samples. The non-sober or sober control contour was re-sampled to the number of samples in the sober contour, respectively (the third distance measure, Euclidean Distance of Discrete Cosine Transform, was applied to the non-time-normalized contours). Linear normalization was applied instead of nonlinear normalization techniques like dynamic-time-warping since the timing of the inherent dynamics should not be distorted. According to the hypothesis this timing is assumed to carry information about the speaker's intoxication state and should therefore not be altered before the contour analysis.

Figure 1 illustrates the Euclidean distance between two time-normalized RMS contours. Distances are hypothesized to be larger between intoxicated and sober contours (henceforth referred to as "intoxicated distances") than between sober and sober control contours (henceforth referred to as "sober distances"):

$$D(\text{intoxicated}) > D(\text{sober}). \qquad (1)$$

It follows that direct distance measures cannot be applied in a speaker-independent classification system but require the same sentences as the recording in question of the same speaker recorded in a sober state. For instance, a voice-controlled navigation system might store recurring control statements of the speaker over a longer time-span and compare these with the current recorded input.

The following distance measures were investigated:

(1) The root mean squared error (RMSE), which is equal to the Euclidean distance between two vectors of the same

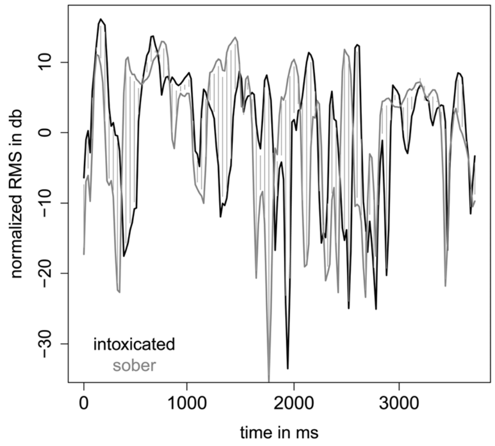

FIG. 1. Example for two time- and RMS-normalized contours and the raw distance between them.

length. Here it reflects the physical distance between two time-normalized contours x and y along the time line. A larger value indicates a greater distance between the contours and a smaller value a smaller distance,

$$D_{\mathrm{rmse}} = \sqrt{\frac{\sum_{t=1}^{N} (x(t) - y(t))^2}{N}}. \tag{2}$$

(2) The second measure is based on the correlation coefficient which here describes the synchronicity of up and down movements of the two contours. The correlation distance is calculated as 1 minus the correlation coefficient, where x and y are time-normalized contours, $\bar{x}$ and $\bar{y}$ their mean values, and sd_x and sd_y their standard deviations,

$$D_{\mathrm{corr}} = 1 - \left(\frac{1}{N-1} \sum_{t=1}^{N} \left(\frac{x(t) - \bar{x}}{sd_x} \right) \left(\frac{y(t) - \bar{y}}{sd_y} \right) \right). \tag{3}$$

(3) The third measure is the distance in a low-dimensional spectral parameterization space. Both contours are transformed into a fixed-dimensional spectral space, and then, the Euclidean distance between the contours within this space is calculated.

The Discrete Cosine Transform (DCT) decomposes a waveform x into factors of its inherent cosine waves $\psi_x(\nu)$ (referred to as "DCT coefficients" henceforth, e.g., Harrington, 2010). DCT coefficients $\psi_x(\nu)$ with lower indices ν represent low-frequency movements (ripples) in the transformed waveform, while coefficients with higher indices represent ripples of high frequency. The first coefficient of the DCT, $\psi_x(\nu = 1)$, is the same for all contours due to the preceding normalization and is therefore not considered in the analysis (Baumeister *et al.*, 2012).

Varying the number of lower DCT coefficients showed that ripple frequency indices 2 to 7 yielded the best distinction between intoxicated and sober RMS contours,

$$D_{\mathrm{dct}} = \sqrt{\sum_{i=2}^{7} (\Psi_x(i) - \Psi_y(i))^2}. \tag{4}$$

2. Distance measures—Results

Sober and intoxicated distances D_{rmse}, D_{corr}, and D_{dct} were calculated for contour pairs for the 20 speakers with control recordings. Results were tested by Mixed Effect Model analysis (MEM; Baayen, 2008) with intoxication, gender as fixed factors and speaker and utterance as random factors. The binary factor intoxication here refers to intoxicated and sober distance. MEM analysis compared to traditional test methods (e.g., MANOVA) has the advantage that data do not have to be averaged across a single strata before analysis. So, in this case the subject and the utterance can be used as a random factor to avoid test errors caused by the statistical dependencies of data points within a subject or sentence category.

Figure 2 shows boxplots of the sober and intoxicated contour distances for all utterances across the 20 speakers. There is a tendency toward larger distances for intoxicated than sober in all three measures D_{rmse}, D_{corr}, and D_{dct}.

MEM analysis only reports F statistics, and the classical estimation of levels of significance using the number of samples (760) as the degree of freedom would lead to unrealistic low values. Therefore, following Reubold *et al.* (2010) throughout this study, p-levels are estimated conservatively from F values using a fixed value of 60 for the degree of freedom because that is roughly the point in the F statistics where the gain in p-level becomes flat.

For D_{rmse} the MEM reports a highly significant increase of the intoxicated against the sober distance ($F = 27.3$, $p < 0.001$). Therefore, the hypothesis is confirmed, that contours of intoxicated speech physically differ more from those of sober speech than do contours from the sober control recording.

The correlation distance D_{corr} also exhibits a highly significant increase ($F = 14.6$, $p < 0.001$) for intoxication. Hence, the movements of sober and corresponding sober control contours are more synchronous than those of sober and corresponding intoxicated contours.

The Euclidean Distance in the 6-dimensional DCT space, D_{dct}, also shows a highly significant increase ($F = 12.4$, $p < 0.001$) for intoxication.

There is no significant interaction with speaker gender for any of the three distance measures.

C. Parameterization of RMS contours

In contrast to distance measures, the direct parameterization allows calculating potential features for each contour, testing the features for the factor speaker intoxication, and correlating to the measured BAC values. In other words, parameterizations are not relative like distances, do not

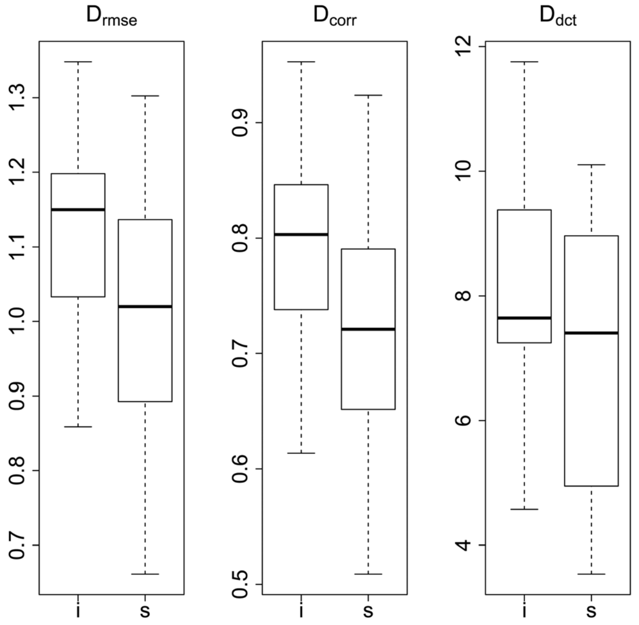

FIG. 2. Sober (s) and intoxicated (i) contour distances across 20 speakers and all utterances

require control recordings as a reference measure, and can be treated as speaker-independent features in classification. Hence, the following analysis can be carried out not only for the 20 speakers with control recordings but for all 150 speakers of the selected sub-corpus of ALC.

1. DCT and moments of DCT—Method

DCT coefficients $\psi_x(\nu)$ with $\nu = 2\ldots7$ were calculated as features for every sober and intoxicated RMS contour as described in Sec. IV B.

As a further parameterization of the DCT spectrum the first and second moments of the DCT coefficients were calculated (Baumeister *et al.*, 2012). The first two moments encode basic properties of the DCT spectral shape, i.e., the center of gravity (m_1) and the variance across the DCT frequency range (m_2). Considering the absolute value of the DCT spectrum $|\psi(\nu)|$ at the ripple frequency ν as an analog to the probability that the contour contains this specific ripple with frequency index ν, the statistical moments $m_{1,2}$ on this probability distribution can be calculated as (e.g., Harrington, 2010, p. 298):

$$m_k = \frac{\sum_{\nu} |\Psi(\nu - m_{k-1})|^k}{\sum_{\nu} \Psi(\nu)}, \tag{5}$$

with $m_0 = 0$ and $k = 1, 2$.

A low value in the first moment m_1, which corresponds to a low center of gravity, is expected for contours with dominant long-term and fewer short-term movements, i.e., a more flattened contour. A more dynamic contour should result in higher m_1 values.

The second moment m_2 is determined by the variance within the DCT spectrum: RMS contours exhibit a low variance in DCT across ripple frequencies, if they are of a regular form, e.g., a uniform sequence of RMS peaks. In contrast, irregular or random contours should have a higher m_2 (Baumeister *et al.*, 2012).

To calculate $m_{1,2}$, the DCT spectrum over ripple frequency indices $\nu = 2\ldots51$ was used; the smallest wavelength of RMS movement considered is therefore

$$\frac{4}{51}L = 0.078L, \tag{6}$$

where L is the total length of the recording in seconds. Since the average syllable number of the test sentences was 12.3 (which equals a wave length of $0.081\,L$), the DCT range $\nu = 2\ldots51$ should roughly cover all RMS movements up to the syllable rate (Baumeister *et al.*, 2012).

2. DCT and moments of DCT—Results

Tests of significance were performed on DCT coefficients 2 to 7 and first and second DCT moments applying MEM analysis (Baayen, 2008) with intoxication and gender as fixed factors, and utterance and speaker as random factors. Here intoxication refers to the (binary) intoxication state of the speaker.

The DCT coefficients $\psi_x(\nu)$ with $\nu = 2$ and $\nu = 4$ are lowered significantly with intoxication ($F = 9.1/25.0$, $p < 0.01/0.001$); all other DCT coefficients yield no significant effect for intoxication. The ascertained significant global distance shift in the 6-dimensional DCT space [see Eq. (4)] can therefore be attributed mainly to a decrease of the energy of the ripple frequencies 0.5 cosine waves (often associated with "slope", e.g., Harrington, 2010, pp. 305) and 1.5 cosine waves within the energy contours (sometimes associated with "skewness" of the contour).

Figures 3 and 4 show the changes (intoxicated minus sober) of the mean DCT coefficients 2 and 4 (of 19 utterances) from sober to intoxicated speech sorted across the 150 speakers.

The MEM tests also reveal a weak but significant decrease of the first DCT moment m_1 ($F = 8.8$, $p < 0.05$) for intoxication, and a significant increase ($F = 9.7$, $p < 0.01$) for the second DCT moment m_2.

Figures 5 and 6 show the changes (intoxicated minus sober) of the mean DCT moments (of 19 utterances) from sober to intoxicated speech sorted across the 150 speakers.

A decrease in the first DCT moment, i.e., a shift in the center of gravity to smaller values, indicates slower movements in the intoxicated than in the sober RMS contours. This seems to be a reasonable result since it has been reported that most individuals reduce their rate of speech when intoxicated (Heinrich and Schiel, 2011).

An increase of the second DCT moment, i.e., a larger variance, indicates that energy contours of intoxicated speech are more irregular than those of sober speech. This result is in line with the previous results about the correlation distance.

Table I shows the Pearson's product-moment correlations between the four significant features across speakers. The only significant correlation found is between the DCT coefficients 2 and 4 ($t = 5.3$, $Df = 158$, $p < 0.001$). DCT moments $m_{1/2}$ and either DCT coefficient 2 or 4 can therefore be considered as prospective independent feature candidates for intoxication detection.

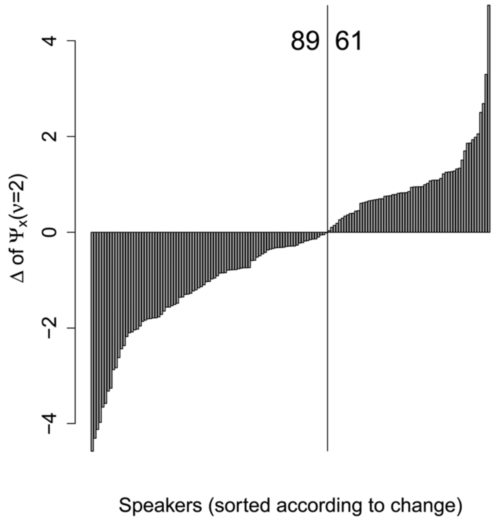

FIG. 3. Change of DCT coefficient 2 from sober to intoxicated sorted across 150 speakers.

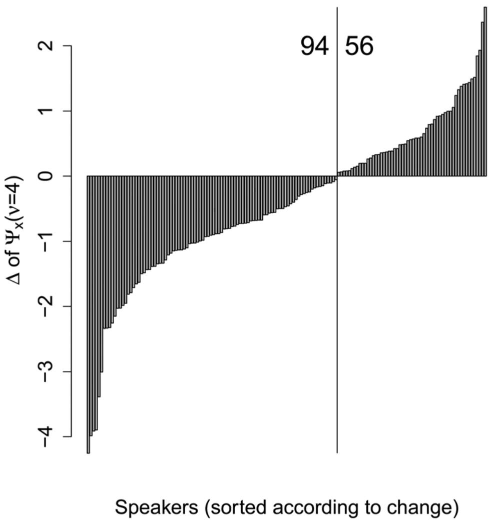

FIG. 4. Change of DCT coefficient 4 from sober to intoxicated sorted across 150 speakers

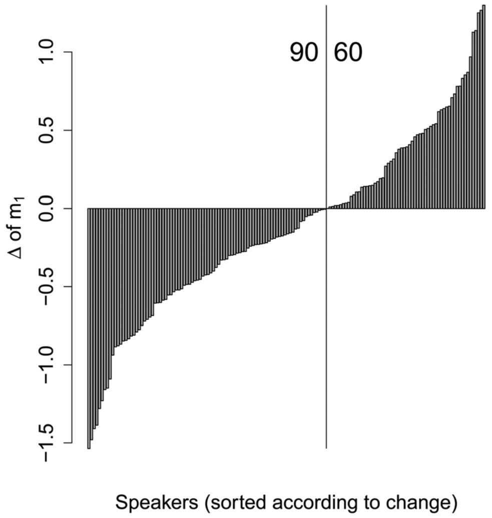

FIG. 5. Change of first DCT moment from sober to intoxicated sorted across 150 speakers

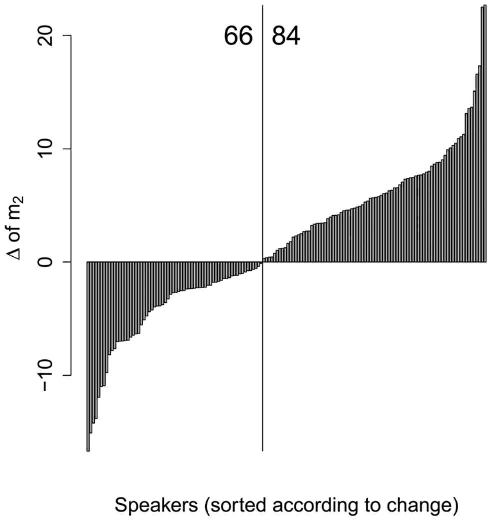

FIG. 6. Change of second DCT moment from sober to intoxicated sorted across 150 speakers.

TABLE I. Pearson's correlation coefficients between mean changes of significant features across 150 speakers.

	Ψ_x ($\nu = 4$)	Δm_1	Δm_2
$\Psi_x(\nu = 2)$	$r = 0.40$	$r = 0.06$	$r = 0.10$
$\Psi_x(\nu = 4)$	-	$r = 0.18$	$r = 0.02$
Δm_1	-	-	$r = 0.08$

3. *Correlation to BAC level*

Linear regression models were fitted for BAC depending on the mean change of the significant DCT coefficients 2 and 4, and first and second DCT moments for all speakers of the ALC (including those with a BAC lower than 0.05%). Figure 7 shows the corresponding scatterplots (top left: first DCT moment, top right: second DCT moment, bottom left: DCT coefficient 2, bottom right: DCT coefficient 4). No significant linear dependencies were found for the significant features. Also, there is no indication of any non-linear behavior. The best regression coefficient is −0.15 (first DCT moment). Therefore, the BAC level of a speaker cannot be predicted from the average change of a single feature by a linear model.

D. Principal components analysis of RMS contours

A further step in the analysis of the RMS energy function is to consider the contour as a whole rather than calculating parameters to represent the contour or a distance between contours.

All 19 sentences investigated for the present study are simple declarative sentences without subordinate clauses. The voice onset and offset of each sentence can therefore be assumed to correlate with a major prosodic boundary. The energy contour between these major boundaries is expected to consist of varying rhythmic forms caused by the (varying) syllable structure and accent pattern but also by varying global forms that cover the complete utterance such as the (physiologically caused) energy decline toward the final boundary which can be steep or flat depending on the sentence type. Based on this, two hypotheses can be formulated:

(1) basic varying energy contour forms common to all declarative sentences in German exist; and

(2) if such basic common energy contour forms exist, their individual contribution to the overall contour varies with intoxication.

If the assumed basic varying energy contours are linearly independent, i.e., encode independent information, then one possible way to falsify these hypotheses is the application of a principal components analysis (PCA; Pearson, 1901) to a large set of energy contours.

1. *Principal components analysis—Method*

The PCA decomposes a value of a multidimensional data set with size *N* into weighted sums of principal components (of the same dimensionality as the value). The *N* principal components (PCs) derived from the data set are under certain assumptions linearly independent and ranked

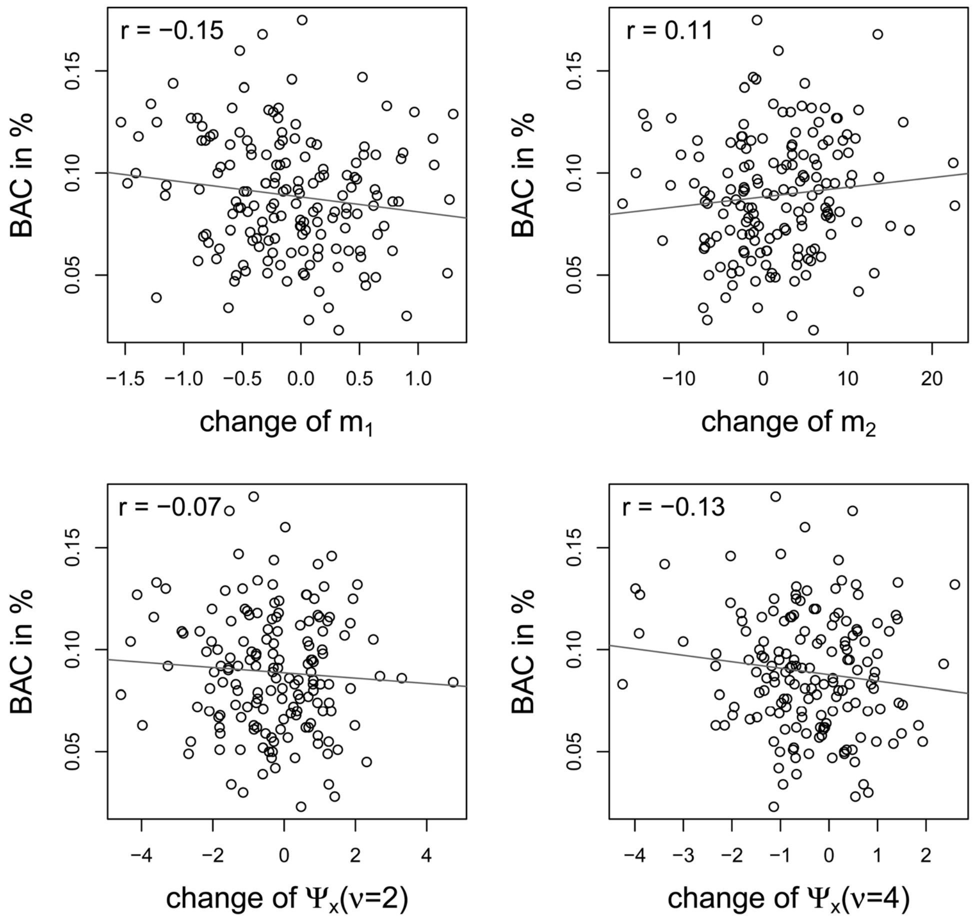

FIG. 7. Correlation between BAC level and the mean change per speaker of (top left) first DCT moment, (top right) second DCT moment, (bottom left) DCT coefficient 2, and (bottom right)DCT coefficient 4.

according to their explained variance in the analyzed data set. Without loss of generality a time series $x_i(t)$ of fixed length $t = 1\ldots T$ can be treated as a data point of dimensionality T, and can thus be approximated by the sum of C PCs $\phi_c(t)$ weighted by their corresponding PC scores $\delta_{c,i}$ with $C <= N$ where $\mu(t)$ is the mean of all time series,

$$x_i(t) \approx \mu(t) + \sum_{c=1}^{C} \left(\phi_c(t)\, \delta_{c,i}\right). \qquad (7)$$

Applied to energy contours this means that the PCs resemble orthogonal "eigen contours" so that the first-ranking PCs represent basic contour forms that vary most in the analyzed data set. The aforementioned weights or "scores" of the PCs then express for each individual contour $x_i(t)$ how much (and with which sign) the PC's "eigen contour" $\phi_c(t)$ contributes to this specific energy contour $x_i(t)$, and can therefore be thought of as features regarding basic contour forms in the same sense as the DCT coefficients in Sec. IV C represent the proportion of basic cosine functions within the contour. If the analyzed data set contains data points, i.e., energy contours from different sources (here sober and intoxicated speakers), and if the differences of these sources manifest themselves in varying basic forms, then the PCA should yield PCs that depict these basic contour forms, and the percentage of explained variance in the analyzed data for the first PCs should be quite large.

The intoxicated and sober contours of the 19 read utterances of 150 speakers were re-sampled to $T = 200$ samples each, since the PCA expects all data points to have the same dimensionality. This normalization across time is justified since the hypothesis assumes the existence of basic contour forms within the prosodic structure of a declarative sentence; therefore, the absolute length of the utterance is not relevant. Because of the fact that principal component scores can be treated like features, the control recordings were not considered here. The complete set of $N = 5700$ contours (sober and intoxicated) were then analyzed using the R function FPCAdecomp provided by Fabian Scheipl of the Computational Statistics group at the University of Munich. The analysis yields N PCs and for each input contour the set of N PC scores. PC scores associated with the first 9 PCs were then analyzed for their correlation to intoxication.

2. Principal components analysis—Results

Figure 8 depicts the first nine PCs based on $N = 5700$ input contours (50% sober, 50% intoxicated). The low ratios of explained variance $\hat{\Phi}_c = 0.06, 0.05, 0.05, 0.04, 0.04, 0.04, 0.04, 0.03, 0.03$ suggest input data that cannot be clustered easily. A visual inspection of the contour forms does not reveal any typical prosodic processes such as a decline of energy with time etc.

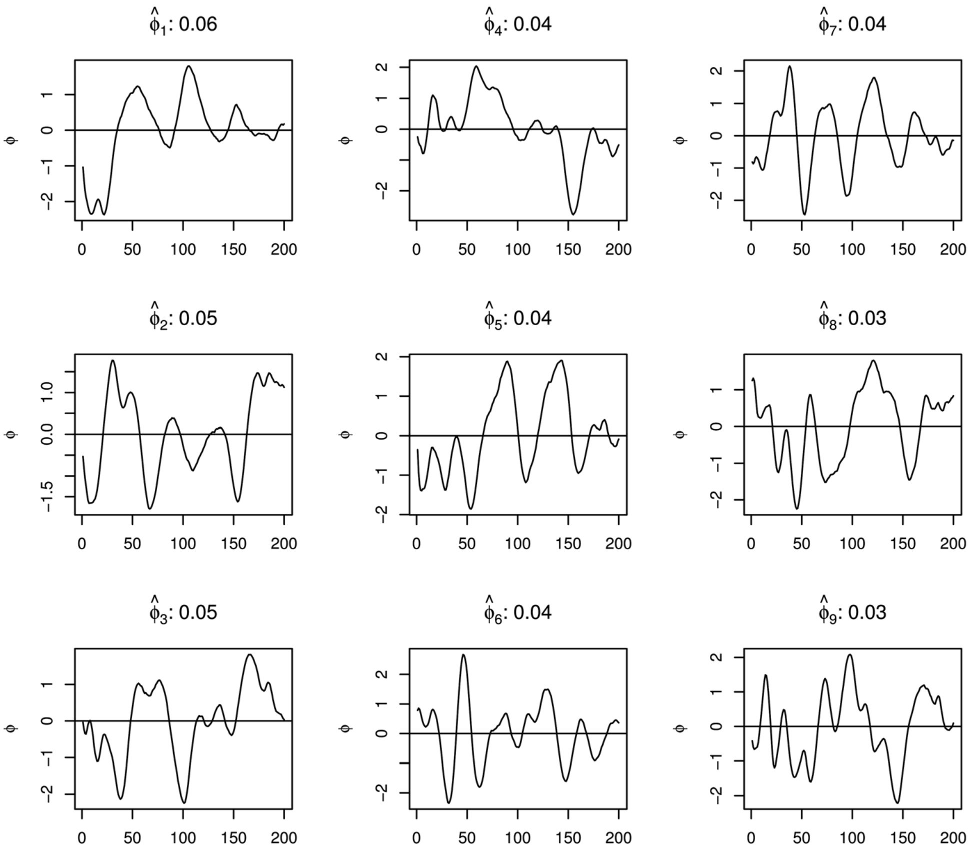

FIG. 8. Principal components 1–9 based on 5700 sober and intoxicated contours; $\hat{\Phi}_c$ is the ratio of explained variance in the data set.

Hypothesis 1, that fundamental contour patterns can be determined by the PCA which explain larger parts of the variance caused by intoxication, has therefore to be rejected.

Tests of significance were performed on PC scores $\delta_c, c = 1\ldots 9$ applying MEM analysis as described in Sec. IV B 2 with intoxication and gender as fixed factors, and utterance and speaker as random factors. Here intoxication refers to the (binary) intoxication state of the speaker.

Only the score of the first PC δ_1 is significantly increased with intoxication ($F = 19.9$, $p < 0.001$); the remaining PC scores $\delta_c, c = 2\ldots 9$ yield no significant effect for intoxication ($p > 0.01$). As in previous analyses, there is no significant interaction with the speaker's gender. Figure 9 shows a boxplot of the first PC score for sober and intoxicated speech; in Fig. 10 the sorted mean score difference between sober and intoxicated speech across the 150 speakers is plotted. For 88 speakers, the score δ_1 increases with intoxication, while for 62 speaker, the score decreases.

V. DISCUSSION

Based on recordings of read speech from the ALC, the findings suggest that there is a significant difference in the short-time RMS energy function between the intoxicated and the sober speech signal of a speaker. In a speaker-

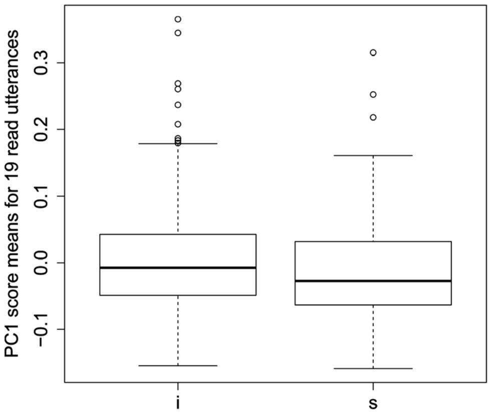

FIG. 9. Sober (s) and intoxicated (i) scores of PC 1 across 150 speakers and all utterances.

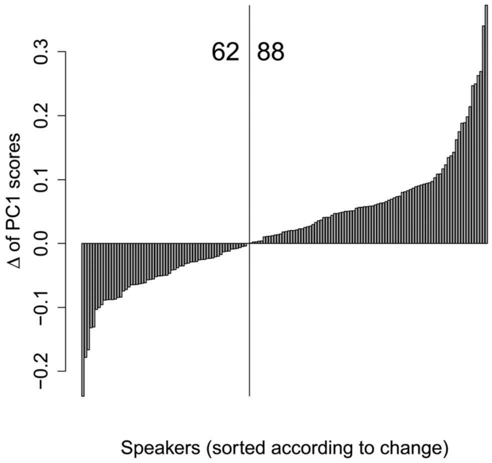

FIG. 10. Change of scores of PC 1 from sober to intoxicated sorted across 150 speakers. 22

dependent classification system, the energy function, which is robust against noise and easily provided, could be used among others as a feature to detect intoxication.

Global measures like the Euclidean distance between pairs of contours, the correlation based distance and the distance in the DCT parameter space are significantly larger between sober and intoxicated contours than between sober and sober control contours. This confirms the hypothesis that RMS contours of intoxicated speech deviate in some way from those of sober speech. However, it does not answer the question in which way these differences are expressed.

For instance, there may be more or longer pauses and therefore a shifting of words, a faster decline of the energy function across the phrase, or a prolongation or shortening of linguistic units. If the main factors responsible could be isolated, the intoxication classification solely based on these factors might be improved.

In order to address this question, contours were parameterized into interpretable DCT coefficients as well as the first and second moments of the lower DCT spectrum. The second and fourth DCT coefficients were lowered significantly with intoxication, which can be interpreted as a weaker decline and weaker skewness within the energy contours. The first and second DCT moments were found to be significantly lowered and raised, respectively, which indicates slower movements and more variation in the energy contours of intoxicated speech compared to sober speech. This confirms earlier findings that intoxication is often correlated with a reduction of speech rate (Heinrich and Schiel, 2011).

However, DCT coefficients have the disadvantage that the base functions of the transform are fixed cosinoidals and not data driven per se. To determine sentence-independent prosodic contour forms and thus examine the differences between intoxicated and sober contours, PCA was applied to the contour data. But the resulting principal components or "eigen contours" do not have the desired effect of reflecting any such differences. The explained variance of the first "eigen contour" did not exceed 6% which indicates that the input data across 19 different sentences are too noisy for PCA. A PCA applied to the data comprising only one kind of sentence would possibly yield better "eigen contours"; but this design would not be useful in real-world applications (because then the analyzed speaker would be required to speak a predefined sentence) and was therefore not pursued in this study. Nevertheless, the PCA score of the first principal component yielded a significant correlation with intoxication, comparable to the DCT moments.

The rhythm-based features found to be significantly different for intoxication in this study can be considered as potential features in a classification system. However, there are some caveats.

(1) Simple distance measures (Sec. IV B) require a fair amount of sober reference data, which would only be available in speaker-dependent classification. Also, it is still to be tested how strong the correlations are between the three distance features analyzed (the data for 20 speakers in this study are not sufficient to calculate reliable correlations).

(2) The first principal component of the PCA (Sec. IV D) is data-dependent. It remains to be tested whether the number of speakers in the ALC is sufficient to estimate "eigen contours" that can be used in a speaker independent classification system.

(3) All analyzed features exhibit a considerable proportion of speakers that "behave in the opposite direction" (see Figs. 3, 4, 5, 6, and 10). Although some of these features are decorrelated (see Sec. IV C 2), the general impression is that the effects on speech under the influence of alcohol are highly speaker-dependent and, therefore, require a speaker-dependent classification system.

ACKNOWLEDGMENTS

Fabian Scheipl of the Computational Statistics group at the Ludwig-Maximilians-Universität München, Germany, provided the code of the functional Principal Components Analysis (R package FPCAdecomp). The work presented in this paper as well as the speech corpus collection ALC was partly funded by the Bavarian Archive for Speech Signals (BAS) at the Ludwig-Maximilians-Universität München, Germany, the Deutsche Forschungsge-meinschaft (SCHI 1117/1-1), the Bund gegen Alkohol und Drogen im Straßenverkehr e.V. (B.A.D.S., League Against Drug and Alcohol Abuse in Traffic), Hamburg Germany, and the BAS Services Schiel, Munich, Germany. The authors thank all colleagues at the BAS and the Institute of Legal Medicine for their help and support.

Aldermann, G. A., Hollien, H., Martin, C., and DeJong, G. (**1995**). "Shifts in fundamental frequency and articulation resulting from intoxication," J. Acoust. Soc. Am. **97**, 3363–3364.

Baayen, R. H. (**2008**). *Analysing Linguistic Data: A Practical Introduction to Statistics Using R* (Cambridge University Press, Cambridge, UK), pp. 263–328.

Baumeister, B., Heinrich, C., and Schiel, F. (**2012**). "The influence of alcoholic intoxication on the fundamental frequency of female and male speakers," J. Acoust. Soc. Am. **132**, 442–451.

Behne, D. M., Rivera, S. M., and Pisoni, D. B. (**1991**). "Effects of alcohol on speech: Durations of isolated words, sentences and passages," Res. Speech Percept. **17**, 285–301.

Braun, A. (**1991**). "Speaking while intoxicated: Phonetic and forensic aspects," in *Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence* (ICPhS Organizing Committee, Aix-en-Provence, France), pp. 146–149.

Cassidy, S., and Harrington, J. (**2001**). "Multi-level annotation in the EMU speech database management system," Speech Commun. **33**(1–2), 61–77.

Chin, S. B., and Pisoni, D. B. (**1997**). *Alcohol and Speech* (Academic Press, San Diego, CA), pp. 258–269.

Cooney, O. M., McGuigan, K. G., and Murphy, P. J. P. (**1998**). "Acoustic analysis of the effects of alcohol on the human voice," J. Acoust. Soc. Am. **103**, 2895–2895.

Cummings, K. E., Chin, S. B., and Pisoni, D. B. (**1995**). "Acoustic and glottal excitation analyses of sober vs. intoxicated speech: A first report," Res. Spoken Language Process. Prog. Rep. **20**, 359–386.

Dekens, T., Demol, M., Verhelst, W., and Verhoeve, P. (**2007**). "A comparative study of speech rate estimation techniques," in *Proceedings of Interspeech 2007* (International Speech Communication Association, Antwerp, Belgium), pp. 510–513.

Dellwo, V. (**2006**). "Rhythm and speech rate: A variation coefficient for DeltaC," in *Language and Language-processing. Proceedings of the 38th linguistic Colloquium*, edited by P. Karnowski, and I. Szigeti (Peter Lang, Frankfurt am Main, Germany), pp. 231–241.

Folk, L., and Schiel, F. (**2011**). "The Lombard Effect in spontaneous dialog speech," in *Proceedings of Interspeech 2011* (International Speech Communication Association, Florence, Italy), pp. 2701–2704.

Grabe, E., and Low, E. L. (**2002**). "Durational variability in speech and the rhythm class hypothesis," in *Papers in Laboratory Phonology*, edited by C. Gussenhoven, and N. Warner (Cambridge University Press, Cambridge, UK), Vol. 7, pp. 515–546.

Hansen, J. H. L., and Patil, S. (**2007**). "Speech under stress: analysis, modeling and recognition," in *LNAI 4343 Speaker Classification I*, edited by C. Mueller (Springer, New York), pp. 108–137.

Harrington, J. (**2010**). *Phonetic Analysis of Speech Corpora* (Wiley-Blackwell, Chichester, UK), pp. 297–316.

Heinrich, C., and Schiel, F. (**2011**). "Estimating speaking rate by means of rhythmicity parameters," in *Proceedings of Interspeech 2011* (International Speech Communication Association, Florence, Italy), pp. 1873–1876.

Hollien, H., DeJong, G., Martin, C. A., Schwartz, R., and Liljegren, K. (**2001**). "Effects of ethanol intoxication on speech suprasegmentals," J. Acoust. Soc. Am. **110**, 3198–3206.

Klingholz, F., Penning, R., and Liebhardt, E. (**1988**). "Recognition of low-level alcohol intoxication from speech signal," J. Acoust. Soc. Am. **84**, 929–935.

Künzel, H. J., and Braun, A. (**2003**). "The effect of alcohol on speech prosody," in *Proceedings of the ICPhS 2003*, Barcelona, Spain, pp. 2645–2648.

Levit, M., Huber, R., Batliner, A., and Noeth, E. (**2001**). "Use of prosodic speech characteristics for automated detection of alcohol intoxication," in *ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, Workshop on Prosody and Speech Recognition 2001*, edited by M. Bacchiani, J. Hirschberg, D. Litman, and M. Ostendorf (Red Bank, NJ), pp. 103–106.

Martin, C. S., and Yuchtman, M. (**1986**). "Using speech as an index of alcohol-intoxication," Res. Speech Percept. **12**, 413–426.

Mathon, S., and de Abreu, S. (**2007**). "Emotion from speakers to listeners: Perception and prosodic characterization of affective speech," in *LNAI 4441 Speaker Classification II*, edited by C. Mueller (Springer, New York), pp. 70–82.

Morgan, N., and Fosler-Lussier, E. (**1998**). "Combining multiple estimators of speaking rate," in *Proceedings of 1998 IEEE International Conference on Acoustics, Speech and Signal Processing* (Institute of Electrical and Electronics Engineers, Seattle, WA), Vol. 2, pp. 729–732.

Pearson, K. (**1901**). "On lines and planes of closest fit to systems of points in space," Philos. Mag. **2**(11), 559–572.

Pfau, T., and Ruske, G. (**1998**). "Estimating the speaking rate by vowel detection," in *Proceedings of 1998 IEEE International Conference on Acoustics, Speech and Signal Processing* (Institute of Electrical and Electronics Engineers, Seattle, WA), Vol. 2, pp. 945–948.

Pisoni, D. B., Hathaway, S. N., and Yuchtman, M. (**1985**). "Effects of alcohol on the acoustic-phonetic properties of speech: Final report to GM research laboratories," Res. Speech Percept. Prog. Rep. **11**, 109–171.

Ramus, F., Nespor, M., and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **73**(3), 265–292.

Reubold, U., Harrington, J., and Kleber, F. (**2010**). "Vocal aging effects on f0 and the first formant: A longitudinal analysis in adult speakers," Speech Commun. **52**, 638–651.

Schiel, F. (**2011**). "Perception of alcoholic intoxication in speech," in *Proceedings of the Interspeech 2011*, Florence, Italy, pp. 3281–3284.

Schiel, F., and Heinrich, C. (**2009**). "Laying the foundation for in-car alcohol detection by speech," in *Proceedings of the Interspeech 2009* (International Speech Communication Association, Brighton, UK), pp. 983–986.

Schiel, F., Heinrich, C., and Barfüsser, S. (**2012**). "Alcohol language corpus: The first public corpus of alcoholized German speech," Lang. Resour. Eval. **46**, 503–521.

Schiel, F., Heinrich, C., and Neumeyer, V. (**2010**). "Rhythm and formant features for automatic alcohol detection," in *Proceedings of the Interspeech 2010* (International Speech Communication Association, Makuhari, Japan), pp. 458–461.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajevski, J., Weniger, F., and Eyben, F. (**2012**). "Medium-term speaker states—A review on intoxication, sleepiness and the first challenge," Comput. Speech Lang. **28**, 346–374.

Sigmund, M., and Zelinka, P. (**2011**). "Analysis of voiced speech excitation due to alcohol intoxication," Inf. Technol. Control **40**, 145–150.

Sobell, L. C., Sobell, M. B., and Coleman, R. F. (**1982**). "Alcohol-induced dysfluency in nonalcoholics," Folia Phoniatr. **34**, 316–323.

Trojan, F., and Kryspin-Exner, K. (**1968**). "The decay of articulation under the influence of alcohol and paraldehyde," Folia Phoniatr. **20**, 217–238.

Wagner, P., and Dellwo, V. (**2004**). "Introducing YARD (yet another rhythm determination) and re-introducing isochrony to rhythm research," in *Proceedings of Speech Prosody 2004* (School of Frontier Sciences, University of Tokyo, Nara, Japan), pp. 227–230.

Xie, Z., and Niyogi, P. (**2006**). "Robust acoustic-based syllable detection," in *Proceedings of Interspeech 2006* (International Speech Communication Association, Pittsburgh, PA), pp. 1571–1574.

Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S. (**2004**). "An acoustic study of emotions expressed in speech," in *Proceedings of International Conference on Speech and Language Processing 2004* (International Speech Communication Association, Jeju Island, Korea), pp. 2193–2196.