The influence of alcoholic intoxication on the fundamental frequency of female and male speakers

Barbara Baumeister, Christian Heinrich, and Florian Schiel[a)]
Institute of Phonetics and Speech Processing, Ludwig-Maximilians-Universität München, 80539 Munich, Germany

(Received 11 November 2011; revised 11 April 2012; accepted 11 May 2012)

This study investigates long-term features and utterance contours of fundamental frequency (f0) derived from the German Alcohol Language Corpus. The corpus comprises *read*, *spontaneous*, and *command&control* speech uttered by 148 speakers of both genders and various age groups when sober and intoxicated. f0 median, f0 range, and f0 contours are analyzed for intoxication and interactions with gender and age. Contours are compared both 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 discrete cosine transform and the first and second moment of the lower contour spectrum. Results partly confirm earlier findings, i.e., f0 average and range are mostly raised with intoxication, and also suggest that the majority of speakers do not follow a general trend, but show idiosyncratic alterations to f0. f0 contours differ significantly with intoxication, but a more detailed analysis could not assign these changes to specific general form changes like decline or curvature. The results suggest that it is not possible to predict intoxication from f0 in a single model across different speakers. Instead a speaker-dependent model to account for the individual speaker behavior is proposed. © *2012 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4726017]

PACS number(s): 43.70.Dn, 43.72.Ar, 43.71.Bp [AL] Pages: 442–451

I. INTRODUCTION

This study is concerned with the influence of alcoholic intoxication (AI) on the fundamental frequency of the speaker (f0). f0 plays a major role in speaker identification/verification techniques and prosodic analysis (e.g., the disambiguation of sentence types, the location of sentence focus, etc.). For a researcher dealing with f0 features or developer of speech applications it is therefore useful to know about the expected influence of specific speaker states such as sleepiness, stress, drug related and alcoholic intoxication on f0, as well as the interaction of this influence with other basic factors, such as speaking style, speaker gender, and age. On the other hand, f0 might act as a robust feature for the detection of speaker states [a good overview is given in Müller (2007)], applied, for instance, in automatic background checks within speech dialogue systems operated by humans in high responsibility work places, such as bus/train driver, pilot, medical supervision, etc. (see, e.g., Schiel and Heinrich 2009).

In several earlier studies on intoxicated speech, f0 has been analyzed and often found to change significantly with intoxication (see list of references in Table I). However, the findings are partly inconclusive: Section II gives a concise summary about the results of earlier studies regarding f0 in intoxicated speech.

The present study on f0 in intoxicated speech is based on a large number (148) of male and female speakers, three different speaking styles, and nonlaboratory speech. Further, the database Alcohol Language Corpus (ALC) used in this study (see Sec. III) is publicly available so that interested readers may replicate our analysis or conduct alternative investigations on the same data.

Fundamental frequency is intrinsically related to speech effort: If a speaker raises the intensity of his or her voice, there is a high probability that fundamental frequency will increase as well [e.g. Gramming (1991) reports 0.3–0.5 semitone/dB]. It is therefore possible that the observed changes in f0 in different speaker states are simply a derivate of the change in effort and, hence, intensity or loudness should be analyzed instead. However, there is a technical problem with analyzing absolute loudness from recordings outside the lab, namely that the mouth–microphone distance—which influences the sound pressure in a cubic function—cannot be controlled satisfactorily in field recordings. To avoid this problem and to facilitate comparison with earlier work, this study is concerned with measured f0 only [see, for instance, Schiel *et al.* (2010) for a study of intensity dynamics in intoxicated speech].

The remaining paper is organized as follows: After summarizing and reviewing earlier studies, Sec. III briefly describes the database on which the present study is based. Sections IV and V present the methodology and results of two experiments regarding long-term f0 features and f0 contours, respectively. Finally, in Sec. VI the findings are summarized and discussed.

II. EARLIER STUDIES REGARDING f0

The majority of earlier studies measured alcoholic intoxication of the speakers by estimating the blood alcohol concentration (BAC) (i.e., the true value) from the breath

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

 0001-4966/2012/132(1)/442/10/$30.00

TABLE I. Summary of findings of earlier studies regarding f0 in intoxicated speech.[a]

Study	Subjects	AI measure[b] %		$f0_{average}$[c]	$f0_{range}$[d]	$f0_{contour}$[e]
Sobell *et al.* (1982)	16 m	BrAC	0.016–0.117	–	•	•
Pisoni *et al.* (1985)	4 m	BrAC	0.100–0.170	–	↑	•
Klingholz *et al.* (1988)	11 m	BAC	0.067–0.159	⇑	⇑	•
Behne and Rivera (1990)	6 m	BrAC	0.085–0.170	⇑	⇑	•
Künzel *et al.* (1992), pp. 41–44	33 m	BrAC	0.015–0.212	⇓⇑	⇑	•
Chin and Pisoni (1997)	2 m/2f	BrAC	0.035–0.130	–	⇑	•
Watanabe *et al.* (1994)	37 m/11f	BAC	•	↓	⇑	•
Aldermann *et al.* (1995)	•	BrAC	0.120	↓	•	•
Cummings *et al.* (1995)	4 m	BrAC	0.100–0.170	⇑	⇑	–
Chin *et al.* (1996–1997)	9 m	BrAC	0.075–0.190	–	⇑	•
Cooney *et al.* (1998)	12•	•	•	–	•	•
Hollien *et al.* (1999)	19 m	BrAC	0.040–0.113	⇑	↑	•
Hollien *et al.* (2001)	19 m/16f	BrAC	0.040–0.113	⇑	↑	•
Künzel and Braun (2003)	33 m	BrAC	< 0.040–> 0.200	⇓⇑	⇑	•

[a]⇑/⇓ = sign. rising or falling of the feature (↑/↓ = n.s.); – = no change; • = not reported.
[b]Shows either the direct measurement of blood alcohol concentration (BAC) or the estimated BAC on the basis of the breath alcohol concentration (BrAC) in volume percent.
[c]Denotes long-term average of f0.
[d]Denotes range of f0.
[e]Regards investigations of the f0 contours across words/phrases.

alcohol concentration (BrAC). In an earlier study the authors of the present paper estimated the Pearson correlation between BrAC and BAC based on samples taken from 152 intoxicated persons to be $r = 0.89$. From the observed distribution of differences the chance for a deviation between BrAC and BAC of more than 0.0001 can be estimated to at least 29% (Schiel *et al.*, 2012). In the present study, therefore, only BAC measurements are used as a reference.

Previous findings concerning f0 under the influence of alcohol are inconsistent (see Table I). Some studies reported an increase of the long-term feature (LTF) f0 in intoxicated condition (see, e.g. Klingholz *et al.*, 1988; Hollien *et al.*, 2001), some a decrease (see, e.g., Watanabe *et al.*, 1994; Aldermann *et al.*, 1995). Künzel and Braun (2003) observed a lower LTF f0 when the subjects had a BrAC lower than 0.08%, and a higher f0 at higher BrAC levels. Finally, others found no significant change at all (see, e.g., Sobell *et al.*, 1982; Pisoni *et al.*, 1985;[1] Cooney *et al.*, 1998). As to the LTF variation of f0, the results are more consistent and show a significant increase of LTF f0 variation while being intoxicated (see column $f0_{range}$ in Table I). The contradictory findings concerning f0 might be due to the different experimental designs. The number of speakers for example, varied from 4 (Pisoni *et al.*, 1985; Chin and Pisoni, 1997; Cummings *et al.*, 1995) to 35 (Hollien *et al.*, 2001) and the speech style recorded was mostly read speech.

To our knowledge, only one study considered f0 contours, i.e., the explicit movement of f0, as a feature for intoxicated speech. In Cummings *et al.* (1995) f0 contours over words were investigated, but no quantifiable results were presented.

Most of the previous studies solely investigated the speech of male speakers, only in the studies of Chin and Pisoni (1997), Watanabe *et al.* (1994), and Hollien *et al.* (2001) female speakers were considered.

III. SPEECH DATA USED IN EXPERIMENTS

All analyses presented in this study are based on the ALC (Schiel *et al.*, 2012). Between 2007 and 2010 the Bavarian Archive for Speech Signals (BAS) collected and annotated intoxicated and sober speech of 77 female and 85 male speakers. Each speaker provided a set of recordings in a sober condition and in an intoxicated condition with one fixed BAC level. BAC levels ranged across speakers from 0.023% to 0.175%.

The ALC comprises different speech styles: *Read speech* (including list style), *semi-spontaneous* (picture task), and *spontaneous* speech (dialogue), as well as read and situational prompted *command&control* speech from the automotive domain (for details on situational prompting, see Mögele *et al.*, 2006). The speech content covers simple digit strings (telephone/credit-card numbers), word lists, addresses, tongue twisters, picture descriptions, interview style answering, and free dialog about casual topics. Each speaker provided, on average, 6 min of speech in intoxicated and 12 min in sober conditions.

In addition, 20 speakers were recorded a third time under the exact same circumstances as in the AI recordings but sober to provide a control group for statistical reference analysis (see Sec. V). All 20 speakers of the control group exhibit a blood alcohol concentration above 0.05% in their intoxicated recordings.

The speech (close and distant microphone) was recorded in the same car environments (two different vehicle types) for the sober and intoxicated conditions. All recordings were manually transcribed and tagged for paralinguistic events. Automatic phonemic segmentation and labeling into the German SAM-PA phonetic symbol set (Wells, 1997) are available for all recordings. Meta data about speaker characteristics (age group, dialectal origin, height, weight, etc.) and

recording conditions (BrAC, BAC, car type, weather) allow statistical testing for influencing factors other than intoxication. The resulting speech corpus can be obtained via the BAS so that other interested researchers may replicate our findings or perform their own studies on intoxicated speech. For a more detailed description of ALC see Schiel *et al.* (2012).

For this study 148 speakers have been selected from ALC. Their BAC level exceeded 0.05%, which is the legal limit for driving in Germany.

IV. LTF OF f0—METHODOLOGY AND RESULTS

In this study we present results from two basic f0 long-term features, the median and the interquartile range (IQR). To test these features for a significant effect on the intoxication we apply a repeated measures design, as it is to be expected that f0 features are speaker-idiosyncratic. The motivation for looking at only simple long-term features (and not, for instance, a parametric model of the distributions) and for not combining them into a higher dimensional feature space are the following:

(i) For the correct application of an analysis of variance (ANOVA) in a repeated measures design all measured data of the same speaker–factor combination must be averaged in a proper manner (e.g., Baayen, 2008).
(ii) Median and IQR are independent from the underlying distribution form, i.e., we do not have to assume that the measured values are Gaussian (or mixture Gaussian) distributed.
(iii) Median and IQR can be more or less directly compared to the results of earlier studies (see Table I).
(iv) Median and IQR are robust against outliers and noise, which is a problem when using out-of-the-laboratory speech and automatic f0 detection.
(v) By including only one LTF value per speaker–factor combination for the final analysis, each speaker has the same contribution to the total result, no matter how much speech material is available for that particular speaker and speaking style (the recording of spontaneous speech necessarily leads to different amounts of speech per speaker).

F0 was calculated for every speech signal from the close talk microphone every 5 ms using the Schaefer-Vincent algorithm (Schaefer-Vincent, 1983) with different search ranges for female (100–500 Hz) and male speakers (50–250 Hz).

The Schaefer-Vincent algorithm operates solely on the digital signal and requires no spectral analysis. As a first step the signal is reduced to a series of extremal values. Then this series is searched for potential "twin periods" by inspecting all possible triples of extremals. Finally a search algorithm finds the best consecutive chain of twin periods that forms a consistent f0 trajectory under certain constraints. In a cross-evaluation of different f0 algorithms we found that the Schaefer-Vincent algorithm is robust against noise, and provides a rather consistent classification of the speech signal into voiced and unvoiced parts.

After discarding all frames that were judged as unvoiced by the f0 detector LTF median and interquartile ranges were calculated over all data frames belonging to one of 888 data partitions defined by speaker (148), intoxication (2), and speaking style (3).

A. Long-term feature f0 median

Figure 1 shows the distribution of the speakers' LTF f0 medians according to intoxicated/sober speech, gender, and speaking style.

It is apparent that the distribution of the medians in intoxicated speech (a) (see Fig. 1) shows higher f0 values than that of sober speech (na) for both female (F) and male (M) speakers and all speaking styles. This confirms the findings on laboratory speech of the majority of earlier studies (7 out of 9) listed in Table I. Significance for this effect was confirmed using a repeated measures ANOVA with the speaker as random factor [$F(1, 146) = 99.2$, $p < 0.001$]; no significant interactions were found for gender, age group (below/above 40 years of age) or speaking style.

To take into account the individual pitch of speakers and to see how each speaker is influenced by intoxication, we calculated the difference of medians between intoxicated and sober speech of each individual speaker. The relative increase/decrease of the LTF f0 medians of each speaker is plotted in the histogram in Fig. 2. Male and female speakers are considered together, because no significant interaction on gender was found in the root mean (RM)-ANOVA and the natural differences in the height of f0 between men and women can be left aside due to the normalization.

The histogram shows that a majority (79.1%) of speakers produce speech with a higher f0 while being intoxicated. On the other hand, the remaining 20.9% of speakers lower their fundamental frequency.

B. Long-term feature f0 range

Figure 3 shows that interquartile ranges tend to be higher for intoxicated speakers (a) than for sober speakers

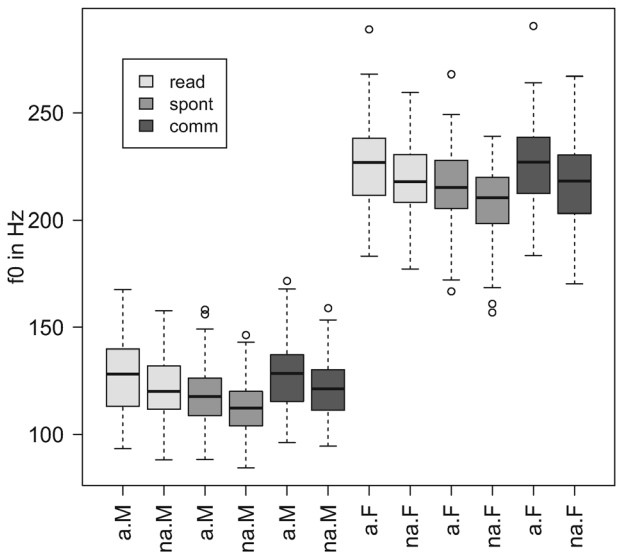

FIG. 1. LTF f0 medians of 148 speakers partitioned in intoxicated (a) and sober (na), female (F) and male (M) speakers, and three speaking styles *read*, *spontaneous*, and *command&control* speech.

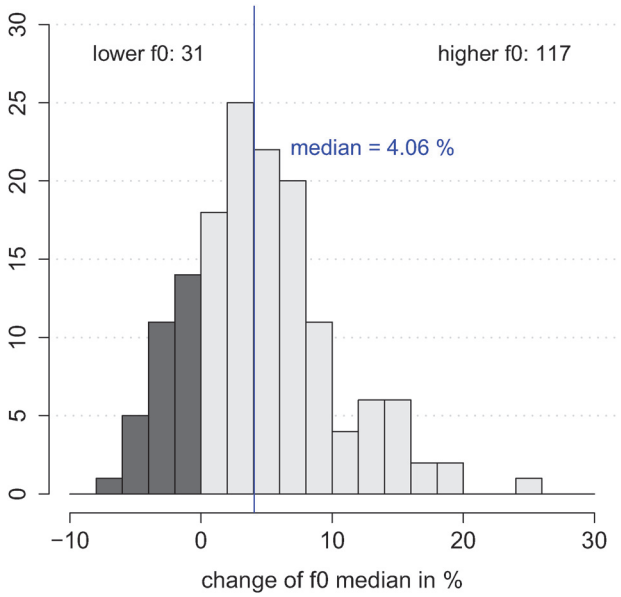

FIG. 2. (Color online) Histogram of relative increase/decrease of LTF f0 medians over 148 speakers, both genders, and all speaking styles.

(na) for both genders and all three speaking styles, which means that speakers, on average, use a wider intonation range when intoxicated. Statistical analysis with RM-ANOVA yielded a highly significant effect for intoxication [$F(1, 146) = 35.21$, $p < 0.001$], thus confirming the unanimous results of earlier findings (see Table I). No interactions were found for gender or age group. In contrast to the LTF f0 median, a significant interaction was found for speaking style [$F(2, 145) = 7.42$, $p < 0.001$]. A *post hoc* test with Bonferroni correction showed that the difference between intoxicated and sober speech in *spontaneous* and *command&control* speech is more significant ($p < 0.001$) than in *read* speech ($p < 0.05$). Figure 3 also shows a large variance over speaker-individual f0 ranges, which is caused by

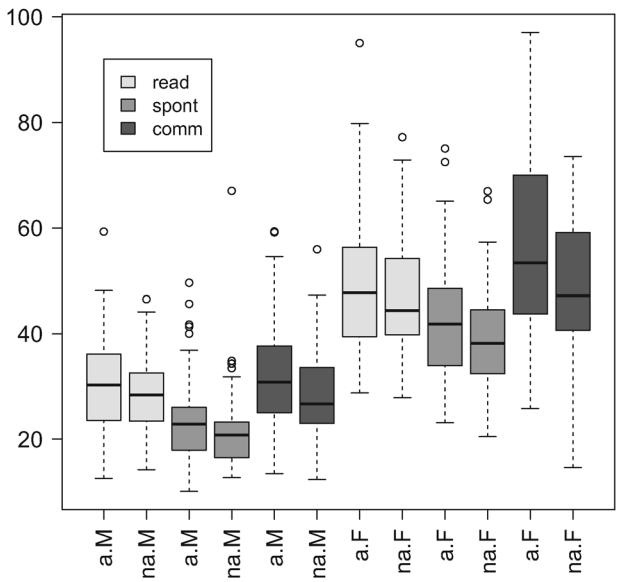

FIG. 3. LTF f0 interquartile ranges of 148 speakers partitioned in intoxicated (a) and sober (na), female (F) and male (M) speakers, and three speaking styles *read*, *spontaneous*, and *command&control* speech.

the different pitch of speakers. Hence, speaker-individual LTF f0 range differences were calculated in the same way as the individual f0 median differences described previously, but separately for the three speaking styles (see histograms in Fig. 4). It is noticeable that again the individual differences vary considerably, and the trend for speakers to use a wider frequency range is not as clear as in the f0 medians (*read* speech: 59.5%, *spontaneous*: 62.8%, *command&control*: 68.9% of all speakers).

C. Correlation of LTF f0 vs BAC level

Linear regression models were fitted for BAC depending on LTF f0 features for all 162 speakers of the ALC and also separately for the three different speaking styles. It turns out that regression coefficients for all features never exceeded 0.21. Therefore, the BAC level cannot be predicted from LTF f0 features across different speakers by a linear model. The scatter plots did not show any evidence for a nonlinear correlation. However, Künzel *et al.* (1992) reported reliable linear regressions, but only when fitting a model on features of the same (male) speaker at different BrAC levels above 0.08%.

V. f0 CONTOURS—METHODOLOGY AND RESULTS

A fundamental frequency or pitch contour is the (sometimes interpolated) f0 track for a specific speech or musical recording. As the f0 contour is closely related to rhythm and subjects questioned about features of intoxicated speech often refer to "rhythm change" (Schiel, 2011), the question arises whether such "changes" may manifest as measurable features in the pitch contour.

The following sections discuss known methods of contour distance calculation from other scientific areas and then present the approach to process raw f0 tracks into comparable data sets used in this study. Then two approaches are described: The direct comparison of f0 contours (distance measures) and the analysis of f0 contour parameters as features (parameterization).

A. Evaluation of f0 contours

There have been a number of investigations into pitch contours in music, many of them dealing with the similarity between a pitch contour query and a given pitch contour derived from a piece of music (Francu and Nevill-Manning, 2000; Lu *et al.*, 2001; Zhu and Kankanhalli, 2002; Shmulevich, 2004). The proposed methods mostly feature different kinds of dynamic alignment algorithms, e.g., dynamic time warping (DTW) (Sakoe and Chiba, 1978), to match a given pitch contour sample against another contour considering both time and tone pitch, yielding a minimized distance between the aligned contours.

In automatic speech synthesis the intonation—and thus the f0 contour—also plays an important role. To evaluate different synthesis systems and to compare their output to natural speech samples, pitch contour differences are often used to quantify their overall similarity. Latsch and Netto (2011) recently proposed a matching method that combines

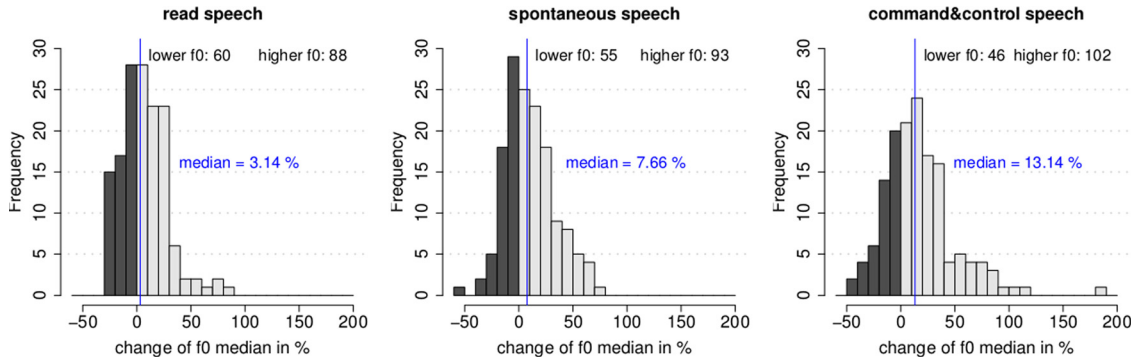

FIG. 4. (Color online) Histogram of relative increase/decrease of LTF f0 interquartile ranges over 148 speakers shown for three different speaking styles.

the pitch-adjusting mechanism [*time domain pitch synchronous overlap and add* (Moulines and Charpentier, 1990)] with the time alignment of speech using DTW. Clark and Dusterhoff (1999) calculated differences between contours using time alignment techniques and compared them to the results of a perception experiment to evaluate synthetic intonation. Hermes (1998a,b) conducted a similar study: Results of a perception experiment, where phoneticians rated the dissimilarity of contours both visually and auditorily, were compared to automatic ratings obtained by different automatically derived distance measures. In his Ph.D. thesis on German intonation in speech synthesis, Moehler (1998) specified numeric and perceptive evaluation methods for contour similarity. Moehler's numeric evaluation method features two of the distance measures described in Sec. V C. Barlow and Wagner (1988) calculated distances between energy, fundamental frequency, voicing, and linear prediction error contours using DTW for a speaker identification experiment.

Many of the contour distance measurements listed previously distort the time axis to minimize the overall distance. This might eliminate subtle rhythmic contour differences, which are of interest in the context of this study. Therefore, instead of using DTW, this investigation compares either linearly time normalized contours directly (Moehler, 1998) or parameterizes contours as described in the following sections.

B. Comparable f0 contours

As stated earlier, f0 tracks were calculated for all speech recordings in the corpus. As the direct comparison of contours requires utterances of equal content, only 19 recordings of *read* speech per speaker with equal content in sober and intoxicated conditions are considered.

As a first step, the f0 tracks are linearly interpolated in voiceless regions (indicated by the Schaefer-Vincent algorithm as $f0 = 0$). Figure 5 shows the f0 contour of an example recording as given by the original f0 calculation (gray) and the interpolated f0 contour (black).

In order to make contours of different lengths comparable, or rather the distances between them measurable, the intoxicated and the control f0 contour were resampled to the same sample number as in the sober f0 contour. This gave 19×3 contours of equal length for 20 speakers as the base material for the distance measurements (Sec. V C). The analysis of contour parameterization does not require time normalization, as the parameters used here are independent of utterance length (Sec. V D).

C. Distance measures between contours

1. Distance measures—method

Figure 6 illustrates the difference between two interpolated and time normalized f0 contours. Such a difference can be quantified by a distance measure. Control recordings are essential for the evaluation of a distance measure. More specifically the distance between the sober and the sober control contour serves as a baseline to which the distance between intoxicated and sober contour can be compared (for convenience the distance between sober and sober control contour is hence referred to as the *sober distance* and the distance between the intoxicated and the sober contour as *intoxicated distance*). If f0 contours differ in the speech of intoxicated speakers, the authors expect that intoxicated distances are significantly larger than sober distances.

In this study distance was determined in three different ways. First, the root mean squared error (RMSE), which is frequently used to describe the differences between time

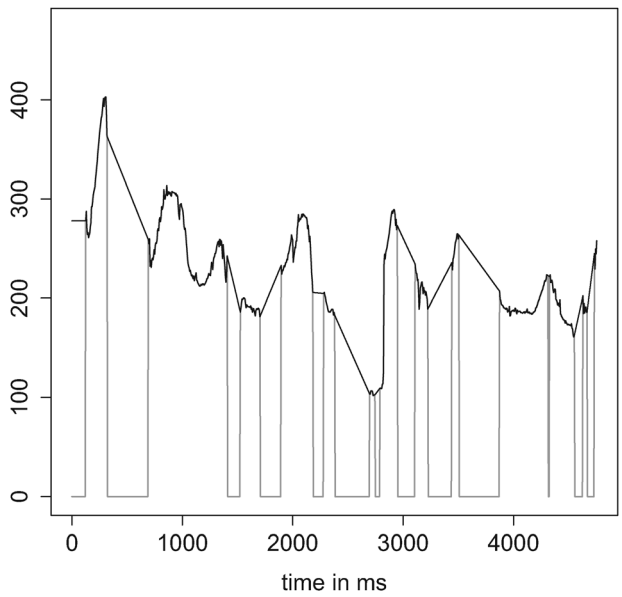

FIG. 5. Original f0 contour and interpolated f0 contour.

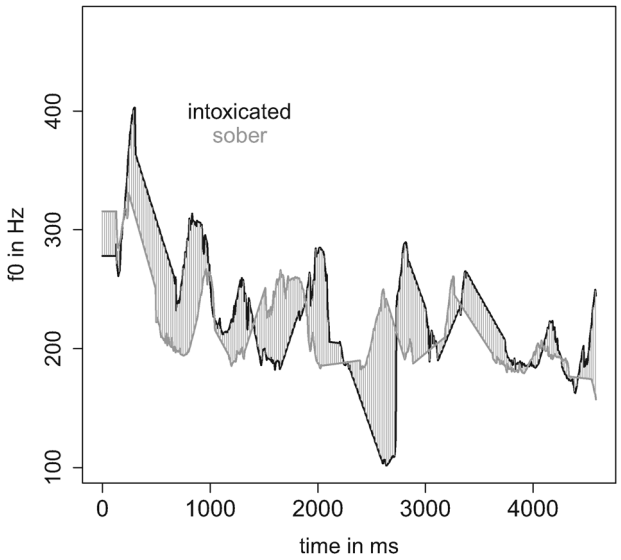

FIG. 6. Physical distance between two pitch contours.

sequences of values. It reflects the physical distance between contours x and y along a timeline. Higher values indicate a larger distance and lower values a smaller distance between contours (Moehler, 1998),

$$D_{\mathrm{RMSE}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}. \tag{1}$$

The second distance measure between pairs of contours is calculated as one minus the correlation coefficient (Klabbers and van Santen, 2004). This measure reflects the consistency of up and down movements within the two contours. A higher correlation coefficient (resulting in a smaller distance) is expected for contours with similar directions in their movements whereas a lower correlation coefficient (or a larger distance) indicates different or less similar directions. If x and y are time normalized contours, $\bar{x}$ and $\bar{y}$ their mean values, and sd_x and sd_y their standard deviations, the correlation distance is defined as

$$D_{\mathrm{CORR}} = 1 - \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{sd_x}\right)\left(\frac{y_i - \bar{y}}{sd_y}\right)\right]. \tag{2}$$

In contrast to D_{RMSE} this measure is bias-independent, i.e., a change in the average f0 does not influence D_{CORR}.

Another approach rather than comparing the contours directly is to transform the contours into a fixed-dimensional spectral space and then calculate the Euclidean distance between contours within this space. The discrete cosine transform (DCT) decomposes a positive waveform x into factors of its inherent cosine waves $\Psi_x(\nu)$ (called "DCT coefficients" henceforth) (see, e.g. Harrington, 2010). DCT coefficients $\Psi_x(\nu)$ with small indices ν represent low-frequency movements (ripples) in the transformed waveform, whereas coefficients with higher indices represent ripples of high frequency. The first coefficient of the DCT $\Psi_x(\nu = 1)$ reflects the mean of the underlying waveform x and is therefore left out of our analysis.

As a third distance measure the Euclidean distance between DCT coefficients 2–7 of two contours x and y was applied as follows:

$$D_{\mathrm{DCT}} = \sqrt{\sum_{i=2}^{7}[\Psi_x(i) - \Psi_y(i)]^2}. \tag{3}$$

Technically the six DCT coefficients 2–7 represent the smoothed and unbiased f0 contour up to a wavelength of $\frac{4}{7}L$, where L is the total length of the recording. In other words, only the most general features of the phrase contour, such as tilt, curvature etc., can be captured by this distance measure.

2. Distance measures—results

Sober and intoxicated distances D_{RMSE}, D_{CORR}, and D_{DCT} were calculated for 19 recordings from 20 speakers. The boxplot in Fig. 7 shows the distribution of distance measures for the same read utterance spoken by 20 speakers (10 female, 10 male). It can be seen that the intoxicated distances (a) tend to be larger than the sober distances (na).

Statistical analysis across all sentences was carried out for each distance measure separately using mixed effect model analysis (Baayen, 2008) to allow for a pairwise comparison of the 19 sentences for every speaker without averaging across speakers. Intoxication, gender, and sentence were treated as fixed factors and the speaker as a random factor. Intoxication here refers to the intoxicated and the sober distance as described previously. As the mixed effect model analysis only reports F statistics the authors applied a rather conservative method to estimate p-levels as given in Reubold *et al.* (2010): Instead of using the number of samples (760) as a degree of freedom, a fixed value of $Df = 60$ is applied to estimate the p-level, because that is roughly the point in the F statistics were the gain in p-level becomes flattened.

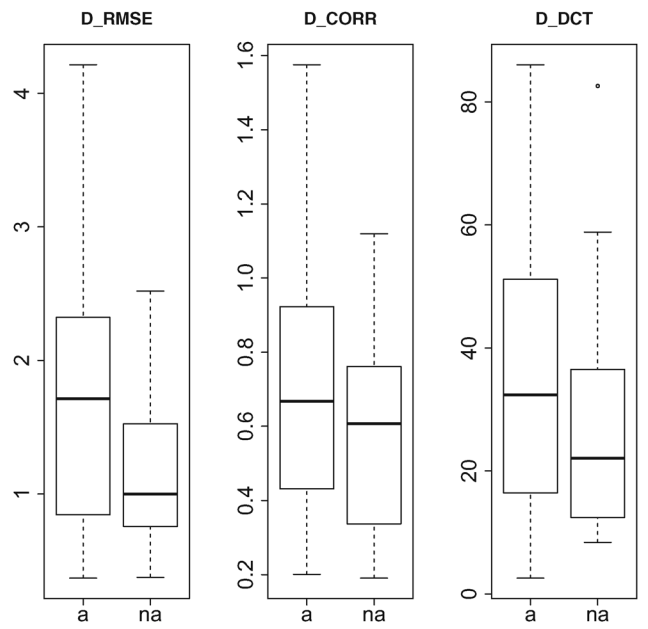

FIG. 7. Boxplot of sober (na) and intoxicated (a) contour distances for one example sentence across 20 speakers.

For D_{RMSE} the mixed model analysis reports a highly significant increase of the intoxicated against the sober distance ($F = 41.983$, $p < 0.001$), which confirms our hypothesis that f0 contours of intoxicated speech differ from those of sober speech. This increase is partly due to the increase of LTF f0 median as shown in Sec. IV A, as a difference in average f0 affects D_{RMSE}.

The correlation distance D_{CORR} also shows a significant increase ($F = 9.3299$, $p < 0.01$) for intoxication. This indicates that the direction of movements of the sober and sober control contours are more similar than those of the sober and intoxicated contours, as expected. Finally, the Euclidean distance between the 6 DCT coefficients D_{DCT} exhibits a highly significant increase ($F = 23.199$, $p < 0.001$) for intoxication as well.

There is no significant interaction with speaker gender in the three distance measures. Female speakers have a (non-significant) tendency to show more distinctive distance differences for D_{RMSE} than male speakers.

The results suggest that f0 contours of intoxicated speech differ globally from those of sober speech, even when changes in the average f0 between the two are not considered. This leads to the question of whether these global differences can be attributed to elementary form changes like phrase decline or contours of f0 peaks or valleys as discussed in the following.

D. Parameterization of f0 contours

1. DCT coefficients and moments—method

As described in Sec. V C the DCT coefficients $\Psi_x(\nu)$ were calculated from the interpolated f0 contour x. Harrington (2010) associates the second to fourth DCT coefficient (in Harrington's nomenclature DCT-1 to DCT-3) with tilt, curvature, and skewness of the contour. DCT coefficients with higher indices might be influenced by syllable rate and/or peak contours. To address the question about specific form changes in f0 contours of intoxicated speech the DCT coefficients for $\nu = 2, \ldots, 7$ were analyzed as independent features regarding intoxication.

A further parameterization of the DCT spectrum is achieved by calculating the first and second moment of the DCT coefficients, which encode basic properties of the DCT spectral shape. If we consider the absolute value of the DCT spectrum $|\Psi(\nu)|$ at the ripple frequency ν as an analogon to the probability that the contour contains this specific ripple with frequency ν, we can calculate the statistical moments $m_{1,2}$ on this probability distribution as (e.g., Harrington, 2010, p. 298):

$$m_k = \frac{\sum_{\nu} |\Psi(\nu - m_{k-1})|^k}{\sum_{\nu} \Psi(\nu)}, \tag{4}$$

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

The motivation for this parameterization is that a lower center of gravity in the DCT spectrum (i.e., a lower first moment) is caused by contours with dominant long-term and fewer short-term movements, i.e., a more flattened intonation, whereas a higher center of gravity is caused by a more dynamic f0 contour (see Fig. 8 for a schematic sketch). The second moment is determined by the variance within the DCT spectrum: f0 contours exhibit a low variance in DCT across frequencies, if they are of a regular form, e.g., a uniform sequence of f0 peaks. In contrast, irregular or random contours should have a higher second moment of DCT spectrum.

To exclude the influence of average f0 encoded in the first DCT coefficient (which is equal to LTF f0 mean) only the DCT spectrum over $\nu = 2, \ldots, 51$ was used in the calculation of moments. This means that the smallest wavelength of f0 movement considered is

$$\frac{4}{51} L = 0.078\, L, \tag{5}$$

where L is the total length of the recording in seconds. As the average syllable number of the test sentences was 12.3 (which equals a wavelength of $0.081L$), this DCT range should roughly cover all f0 movements down to the syllable rate.

Note that in contrast to the previous section no reference distance is needed for single DCT coefficients and DCT moments, as a DCT coefficient or moment can be calculated from a sober or intoxicated contour independently like the LTF features described in Sec. IV. Hence DCT coefficients 2–7 and first and second moments were calculated for all 148 speakers with a BAC above 0.05% and all 19 matching *read* items.

2. DCT coefficients and moments: Results

Significance was tested on DCT coefficients 2–7 and moments using a mixed effect model analysis (Baayen, 2008) with intoxication, gender, and sentence as fixed factors and the speaker as a random factor (repeated measures design). Intoxication levels are again intoxicated and sober, but here they directly relate to the intoxication state of the speaker.

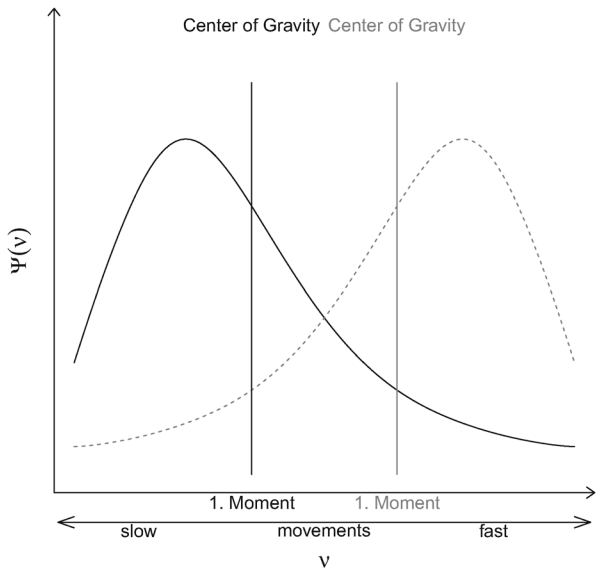

FIG. 8. Schematic diagram showing the effect of a change of the center of gravity (first moment) in the DCT spectrum on contour movements.

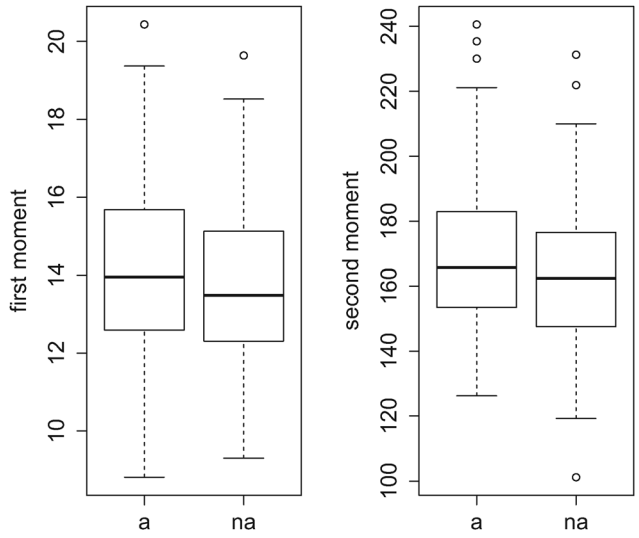

FIG. 9. Boxplot of sober (na) and intoxicated (a) DCT moments for one example sentence across 148 speakers.

The model yields no significant effect between intoxicated and sober speech for all DCT coefficients 2–7. Hence, the significant global distance shift in the six-dimensional DCT space [see Eq. (3)] cannot be attributed to specific form changes in the contours.

On the other hand, our expectation for a shift of the center of gravity in the DCT spectrum was confirmed: The first moment exhibits a highly significant increase ($F = 21.7046$, $p < 0.001$) for intoxication. See, e.g., Fig. 9 for the distribution of the first moment summarized for 148 speakers and one *read* sentence. For the second moment there is only a weak effect ($F = 4.5704$, $p < 0.05$) for intoxication. In both moments there was no evidence for either a gender specific effect or any interaction with age groups. Hence f0 contours of intoxicated *read* speech seem to contain significantly more fast f0 movements than contours from sober speech

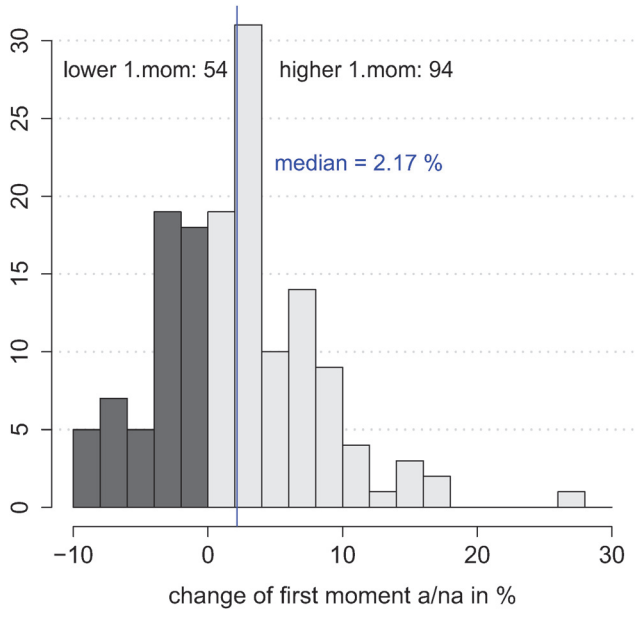

FIG. 10. (Color online) Histogram of relative decrease/increase of the first moment in the DCT spectrum (index 2-51) over 148 speakers (see text).

(increase of first DCT moment), but there is only weak evidence for a decrease in regularity (increase of second DCT moment).

The histogram in Fig. 10 shows the relative change of the mean first DCT moment of the 19 recordings for 148 speakers. It can be seen that for most speakers (63.5%) the first DCT moment shifts to higher values when intoxicated. About one-third of the speakers exhibit a shift in the opposite direction.

E. Correlation of DCT moments vs BAC level

Analogous to Sec. IV C linear regression models were fitted for BAC depending on the first and second DCT moment for all 162 speakers. As for LTF f0 features (see Sec. IV C) no linear dependencies were found for DCT moments; the highest regression coefficient is 0.11. It follows that the BAC level cannot be predicted from the DCT moments by a linear model. Again there was no indication from the scatter plots for nonlinear behavior.

VI. CONCLUSION

Long-term fundamental frequency features and f0 contours over complete utterances were investigated for 148 intoxicated speakers of both gender and in three different speaking styles.

The findings on LTF f0 medians confirm that the majority of speakers (79%) increase their fundamental frequency when intoxicated not only for *read* speech, but also for *spontaneous* and *command&control* speech. However, 21% consistently decrease their f0. There was no evidence that these two speaker groups (increase vs decrease) are caused by an exceptional high or low BAC level of the individual speakers, or any statistical interaction with gender or age. It therefore remains unclear what causes a speaker to lower or increase her/his fundamental frequency.

The majority of the speakers also increase their LTF f0 interquartile range, but this effect is not as clear as the effect on f0 medians and, further, is dependent on speaking style: f0 variation increases significantly more in intoxicated *spontaneous* and *command&control* speech than in *read* speech. These two findings somewhat contradict earlier findings on laboratory speech, where the increase in f0 range was reported to be more consistent than the increase of the average f0.

f0 contours spanning over complete read utterances, in general, differ significantly between sober and intoxicated read speech. This was confirmed for direct contour distance measures based on RMSE, on statistical correlation, and in the six-dimensional DCT space. All three effects seem to be robust across different sentences and independent of speaker age and gender. There was no indication that these global differences can be attributed to changes in certain form features of the f0 contour (DCT coefficients 2–7). A majority of speakers (63.5%) shifted the first moment in the DCT spectrum to higher values for intoxicated speech, which can be interpreted as a higher proportion of fast f0 movements. The second moment of the DCT spectrum does not show any significant trend, which indicates that the regularity of f0

movements does not change with intoxication across the speaker population.

Neither LTF f0 features nor f0 contour parameters correlated with BAC levels of speakers. To this end it is unlikely that a statistical model across different speakers will be able to reliably predict the intoxication level based on f0 features.

LTF f0 features and the first DCT moment were also tested for correlations against each other. A high correlation would indicate that one feature is dependent on the other and therefore redundant. Table II shows the pairwise correlation coefficients. Although LTF median and LTF range are slightly correlated ($r = 0.55$), both the LTF median and LTF range do not correlate with the first DCT moment ($r = 0.05$, $r = -0.03$). Hence, LTF features and the first DCT moment can be considered as statistically complementary features for intoxication.

In both experiments speakers tend to follow an individual trend rather than a global trend, which is valid for all speakers. For example in Fig. 11 we can see the distribution of the LTF f0 median calculated for all sober (na) and intoxicated (a) utterances of two different male speakers (utterance medians). These two speakers follow opposite trends: Whereas the speaker on the left rather consistently increases his utterance medians, the speaker on the right decreases when intoxicated. To verify this effect, the authors conducted *t*-tests on the utterance medians for each individual speaker and found that 60% (89 out of 148) of speakers followed their individual trend, i.e., increasing or decreasing their utterance medians, on a significance level of $p = 0.05$.

This speaker-dependent behavior is indirectly confirmed by the findings of 10 studies of independent research groups that participated in the Interspeech Speaker State Challenge (Schuller *et al.*, 2011). The task there was to predict intoxication from the speech signal in a simple binary classification problem (sober <0.05% vs intoxicated >0.05%). The best speaker-independent classification model developed by Bone *et al.* (2011) yielded ~70% correct unweighted detection rate (chance being 50%). In Schiel (2011) a perception study on the same data set resulted in a human detection performance of 72%. All these findings together with the results of this study suggest that BAC levels or even a simple binary decision about intoxication cannot reliably be predicted by a general statistical model for unknown speakers. Our findings about individual speaker behavior indicate that speaker-dependent models trained on sober speech of individual speakers might be more appropriate. This hypothesis and also the influence of other speaker states such as sleepiness, fatigue, or emotions on f0 will be subject to future investigations.

TABLE II. Correlation coefficients between feature changes across 148 speakers.

	LTF f0 range	DCT first moment
LTF f0 median	$r = 0.55$	$r = 0.05$
LTF f0 range	—	$r = -0.03$

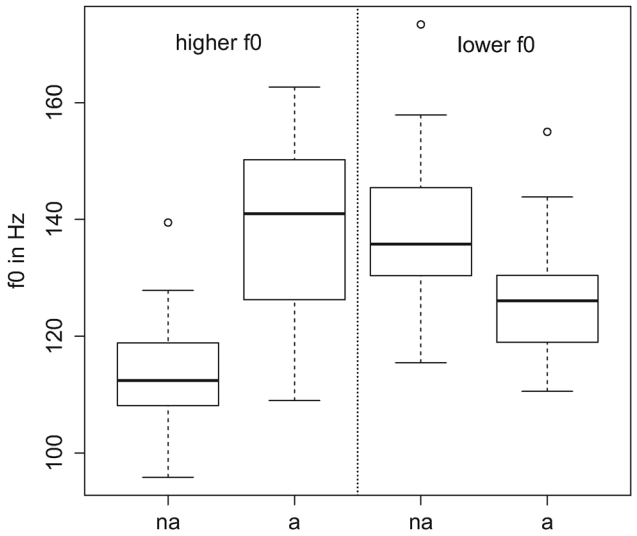

FIG. 11. Distribution of LTF f0 medians for two male speakers with opposite trends.

ACKNOWLEDGMENTS

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, by the Deutsche Forschungsgemeinschaft (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 within the BAS and the Institute of Legal Medicine for their help and support.

[1]The following references may be downloaded from http://www.iu.edu/~srlweb/publications (Last viewed 4/10/12): Behne and Rivera (1990), Behne, Rivera, and Pisoni (1991), Martin and Yuchtman (1986), and Pisoni, Hahtaway, and Yuchtman (1985).

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), pp. 263–328.

Barlow, M. G., and Wagner, M. (**1988**). "Prosody as a basis for determining speaker characteristics," in *Proceedings of the Second Australian International Conference on Speech Science and Technology*, Sydney, Australia, pp. 80–85.

Behne, D. M., and Rivera, S. M. (**1990**). "Effects of alcohol on speech: Acoustic analyses of spondees," Research on Speech Perception Progress Report No. 16 (Indiana University, Bloomington, IN), pp. 263–291.

Behne, D. M., Rivera, S. M., and Pisoni, D. B. (**1991**). "Effects of alcohol on speech: Durations of isolated words, sentences and passages," Research on Speech Perception Report No. 17 (Indiana University, Bloomington, IN), pp. 285–301.

Bone, D., Black, M. P., Li, M., Metallinou, A., Lee, S., and Narayanan, S. S. (**2011**). "Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors," in *Proceedings of the Interspeech2011*, Florence, Italy, pp. 3217–3220.

Chin, S. B. Large, N. R., and Pisoni, D. B. (**1997**). "Effects of alcohol on the production of words in context: A first report," Res. Spoken Lang. Process. Progr. Rep. **21**, 403–420.

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

Clark, R. A. J., and Dusterhoff, K. E. (**1999**). "Objective methods for evaluating synthetic intonation," in *Proceedings of the 99′ Eurospeech*, Budapest, Hungary, Vol. **4**, pp. 1623–1626 (1999).
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 Lang. Process. Progr. Rep. **20**, 359–386.
Francu, C., and Nevill-Manning, C. G. (**2000**). "Distance metrics and indexing strategies for a digital library of popular music," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Vol. 2, pp. 889–892.
Gramming, P. (**1991**). "Vocal loudness and frequency capabilities of the voice," J. Voice **5**, 144–157.
Harrington, J. (**2010**). *Phonetic Analysis of Speech Corpora* (Wiley–Blackwell, Chichester, UK), pp. 297–316.
Hermes, D. J. (**1998a**). "Auditory and visual similarity of pitch contours," J. Speech Lang. Hear. Res. **41**, 63–72.
Hermes, D. J. (**1998b**). "Measuring the perceptual similarity of pitch contours," J. Speech Lang. Hear. Res. **41**, 73–82.
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.
Hollien, H., Liljegren, K., Martin, C. A., and DeJong, G. (**1999**). "Prediction of intoxication levels by speech analysis," in *Advances in Phonetics*, edited by A. Braun (Steiner, Stuttgart, Germany), Vol. 106, pp. 40–50.
Klabbers, E., and van Santen, J. P. H. (**2004**). "Clustering of foot-based pitch contours in expressive speech," in *ISCA Speech Synthesis Workshop*, Vol. 5, 73–78.
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 ICPhS2003*, Barcelona, Spain, pp. 2645–2648.
Künzel, H. J., Braun, A., and Eysholdt, U. (**1992**). *Einfluß von Alkohol auf Sprache und Stimme (Influence of Alcohol on Speech and Voice)* (Kriminalistik-Verlag, Heidelberg, Germany), pp. 1–117.
Latsch, V. L., and Netto, S. L. (**2011**). "Pitch-synchronous time alignment of speech signals for prosody transplantation," in *International Symposium on Circuits and Systems (ISCAS)* (IEEE, New York), pp. 2405–2408.
Lu, L., You, H., and Zhang, H.-J. (**2001**). "A new approach to query by humming in music retrieval," in *Proceedings of IEEE International Conference on Multimedia and Expo* (IEEE, New York), pp. 595–598.
Martin, C. S., and Yuchtman, M. (**1986**). "Using speech as an index of alcohol-intoxication," Research on Speech Perception Report No. 12 (Indiana University, Bloomington, IN), pp. 413–426.
Moehler, G. (**1998**). "Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese (Theory-based modelling of German intonation for speech synthesis)," Ph.D. thesis, Universität Stuttgart, pp. 96–114.
Mögele, H., Kaiser, M., and Schiel, F. (**2006**). "SmartWeb UMTS Speech Data Collection: The SmartWeb Handheld Corpus," in *Proceedings of the LREC2006*, Genova, Italy, pp. 2106–2111.
Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.
Müller, C., ed. (**2007**). *Speaker Classification I*, LNAI 434 ed. (Springer, Berlin), pp. 1–353.
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," Research on Speech Perception Progress Report No. 11 (Indiana University, Bloomington, IN), pp. 109–171.
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.
Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust. Speech Signal Process. **26**, 43–49.
Schaefer-Vincent, K. (**1983**). "Pitch period detection and chaining: Method and evaluation," Phonetica **40**, 177–202.
Schiel, F. (**2011**). "Perception of alcoholic intoxication in speech," in *Proceedings of the Interspeech2011*, 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*, Brighton, UK, pp. 983–986.
Schiel, F., Heinrich, C., and Barfüsser, S. (**2012**). "Alcohol language corpus: The first public corpus of alcoholized German speech," *Language Resources and Evaluation* (in press).
Schiel, F., Heinrich, C., and Neumeyer, V. (**2010**). "Rhythm and formant features for automatic alcohol detection," in *Proceedings of the Interspeech2010*, Chiba, Japan, pp. 458–461.
Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajevski, J. (**2011**). "The Interspeech 2011 speaker state challenge," in *Proceedings of the Interspeech2011*, Florence, Italy, pp. 3201–3204.
Shmulevich, I. (**2004**). "A note on the pitch contour similarity index," J. New Music Res. **33**, 17–18.
Sobell, L. C. Sobell, M. B., and Coleman, R. F. (**1982**). "Alcohol-induced dysfluency in nonalcoholics," Folia Phoniatr. **34**, 316–323.
Watanabe, H. Shin, T. Matsuo, H. Okuno, F. Tsuji, T. Matsuoka, M. Fakaura, J., and Matsunaga, H. (**1994**). "Studies on vocal fold injection and changes in pitch associated with alcohol intake," J. Voice **8**, 340–346.
Wells, J. C. (**1997**). "SAMPA computer readable phonetic alphabet," in *Handbook of Standards and Resources for Spoken Language Systems*, edited by D. Gibbon, R. Moore, and R. Winski, Part IV, Sec. B (Mouton de Gruyter, Berlin).
Zhu, Y., and Kankanhalli, M. (**2002**). "Similarity matching of continuous melody contours for humming querying of melody databases," in *Proceedings of IEEE Workshop on Multimedia Signal Processing* (IEEE, New York), pp. 249–252.