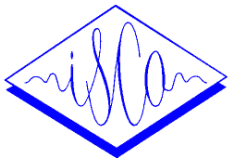
ISCA Archive
http://www.isca-speech.org/archive



ITRW on Prosody in
Speech Recognition and Understanding
Molly Pitcher Inn, Red Bank, NJ, USA
October 22-24, 2001

Use of prosodic speech characteristics for automated detection of alcohol intoxication

Michael Levit†, *Richard Huber*‡, *Anton Batliner*†, *Elmar Noeth*†

†Chair for Pattern Recognition, University of Erlangen, Germany
{levit,batliner,noeth}@immd5.informatik.uni-erlangen.de
‡ Sympalog Speech Technologies AG, Germany
huber@sympalog.com

Abstract

In this paper we describe our methodology for automatic detection of speaker alcoholization. Our task is restricted to detection of considerable alcoholization (alcohol blood level $\geq$ 0.8 per mille), so that a two-class classification problem is to be solved. In particular, our attention is focused on the influence of the alcohol intoxication on the prosodical aspect of the spoken language. A new kind of signal intervals underlying the extraction of prosodic features (*phrasal units*) is proposed along with a method for their localization, which makes it possible to avoid the word segmentation of the speech signal as a preceding stage of the classification process. We also assess the utility of various prosodic features computed on such intervals for the task specified above. In our experiments on unseen data, we achieved classification rates of almost 69% when discriminating between alcoholized vs. not alcoholized speech.

1. Introduction

It is a well known fact that diverse qualities of spoken language can be influenced by factors such as stress experienced by the speaker [13], his emotions [11], or impairment of his physiological functionality caused by drug or alcohol intoxication [10, 8, 3].

There have been several efforts to identify and classify stress and emotions using spoken language [12, 14, 1]. Both acoustic features (e.g. cepstral coefficients) and prosodic features (e.g. evolution of fundamental frequency F_0) have been used in these experiments. To the knowledge of the authors, no experiments on automated detection of alcohol intoxication via spoken language have been conducted as yet. Nonetheless, we deem these problems to be closely related to each other. Indeed, in [4] we find a definition of stress as *"... a psycho-physiological state characterized by subjective strain, dysfunctional physiological activity, and deterioration of performance"*. On the other hand, similar effects can be attributed to alcohol intoxication as well. Thus, we expect alcohol intoxication to affect several prosodic characteristics of the spoken language.

One possibility to attack this classification problem by means of *structural prosodic features* is to calculate one vector of prosodic features for each signal interval corresponding to a lexical unit of speech (e.g. word) occurred in the signal [7, 9]. Thus, a speech recognition engine must be run prior to the actual classification. This can be a major bottleneck, since the word recognition rates of a system drop significantly when speech abnormalities (for instance emotions [13]) are present.

In this paper we propose a new approach for the determination of signal intervals which underly extraction of prosodic features. The strategy suggested in section 2 allows to avoid the employment of Automated Speech Recognition at the preceding stages of the classification process, by relating the prosodic structural features to the signal intervals localized by means of *basic prosodic features* (e.g. zero-crossing rate, fundamental frequency and energy) only. We call such intervals *phrasal units*.

Another aspect of a classification problem is the set of the employed prosodic features itself. In section 3 we present four different groups of features which we use in our classification experiments. In section 4 an experimental evaluation of phrasal units as signal intervals underlying the feature extraction and of the features themselves is given for the task, in which alcohol intoxication is to be detected. The conclusion given in section 5 summarizes the content of the paper.

2. Phrasal units

When looking for indicators of alcohol intoxication in speech, we assume that these indicators are stable, i.e. they persist throughout the entire speech signal. This observation allows for three strategies to choose intervals (*prosodic units*) for which prosodic features will be computed:

- micro-intervals (e.g. 10-msec-frames);
- entire signal as a single interval;
- moderate number of macro-intervals (typical duration of a few seconds).

From the alternatives above only the last one appears to result in a time resolution which is fine enough to produce feature vectors accounting for local changes in the prosodic characteristics of the utterances (such as regression coefficients for F_0) but still allow their meaningful and reliable estimation.

A typical choice for such intervals are linguistically based units such as words or syllables. However, this presupposes word or syllable recognition as the preceding step, and in section 1 we have pointed out that this is highly error prone. An alternative to linguistically motivated units are prosodically motivated units. In our methodology, the corresponding speech intervals are delimited by silence intervals which, on their part, are determined by means of frame-wise calculated basic prosodic features such as:

- fundamental frequency (where possible);
- zero-crossing rate;
- energy.

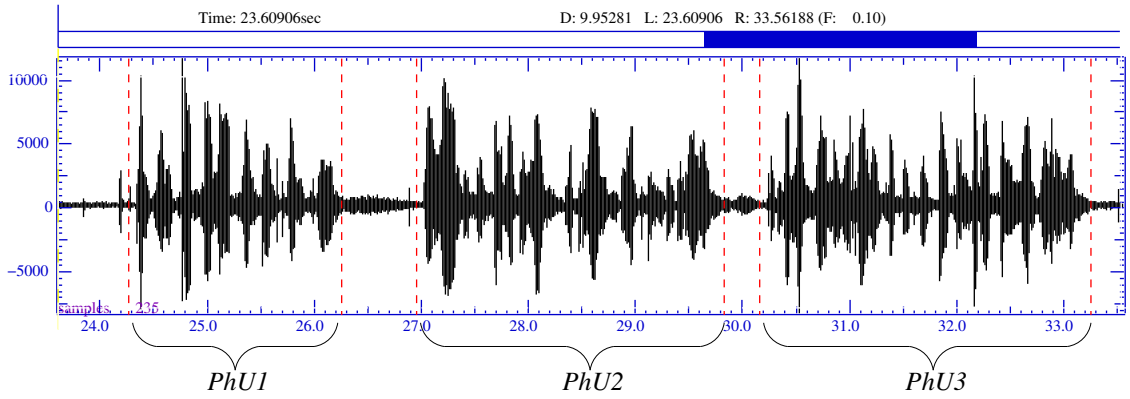


Figure 1: *An example of syntactically justified phrasal units.*

We call the prosodic units that are generated this way *phrasal units (PhU)*. It is remarkable that the boundaries between phrasal units correlate strongly with syntactic boundaries in the underlying text. Indeed, the detected boundaries coincide most of the time (precision 85%) with *full intonational boundaries* which, according to [2], can take over the role of punctuation marks in spoken language. Picture 1 shows an example where the speech signal based upon the German text

> *"Endlich gab der Nordwind den Kampf auf. Nun wärmte die Sonne die Luft mit ihren freundlichen Strahlen und schon nach wenigen Augenblicken zog der Wanderer seinen Mantel aus."*

has been split in three phrasal units which correspond to the first sentence and to the main and subordinate clauses of the second sentence. Also, intervals of silence have been localized.

The average length of phrasal units depends strongly on the speech tempo in the corpus but typically lies between one and three seconds.

3. Features for classification

For each prosodic unit we calculate one vector of features representing this signal interval. At this place, we introduce four groups of features which will be evaluated in our experiments.

The first group *PM21* consists of prosodic features which describe macro-tendencies in the fundamental frequency and energy. Calculation of these features is a standard component of the *prosody module* of the VERBMOBIL-project [9]. In particular, we consider regression coefficients and positions and values of maxima and minima for both energy and fundamental frequency, on- and offset positions and values for fundamental frequency, durations of pauses around prosodic units and some other features [7]. Altogether, this group contains 21 features.

Durational characteristics of voiced and unvoiced intervals within prosodic unit compose the second group *VU11*. Those are 11 features, reflecting absolute lengths and numbers of voiced and unvoiced intervals in the speech signal as well as their proportions [6].

The only group of non-prosodic features that we use for our experiments is the set of long-term cepstral coefficients *LTM24*. This group is formed by 12 mel-cepstral coefficients along with their time differences. To calculate these coefficients, we average the frame-wise computed mel-cepstral coefficients and 12 time differences over the entire signal and assume that all of its prosodic units possess these average values. We expect these coefficients to account for specific properties of the alcoholized speech such as increased nasalization.

Finally, we consider jitter and shimmer, short-term fluctuations in energy and fundamental frequency. Most definitions of jitter and shimmer are rather descriptive than constructive. For instance, in [5, p.79] we read:

> *"Jitter means (stochastic) small frequency changes and modulations of a signal..."*

Several calculation rules have been proposed for jitter and shimmer (see for instance [12]). For our experiments we make use of a linear filtering mechanism. In order to obtain sequence of jitter values J_i, we consider the sequence of F_0-values f_i computed using an ANN-driven period-synchronous method [7, pp.145–150]. This sequence is highpass-filtered via convolution and normalized by f_i:

$$J_i = \frac{\sum_{\mu=0}^{m} f_{i-\mu} g_\mu}{f_i}. \tag{1}$$

The employed filter has impulse response $g=\{g_0, \ldots, g_m\}$ set as a sequence of binomial coefficients. For example, such a filter of size $m = 2$ has the impulse response $g=\{-1, 1\}$ and the filter of size $m = 4$ the impulse response $g=\frac{1}{4}\{-1, 3, -3, 1\}$. The higher the filter size, the larger is the preserved portion of high frequencies in the spectrum of the sequence of the consequent F_0-values. Which filter size is optimal for a classification is one of the subjects of the discussion in section 4.2. Finally, the average and variance of the values J_i within the prosodic unit are computed for the last group of features *JS4*:

$$J_{avg} = \frac{\sum_{J_i<J_\theta} J_i}{\#_{J_i<J_\theta} J_i}; \tag{2}$$

$$J_{dev} = \sqrt{\frac{\sum_{J_i<J_\theta} (J_i - J_{avg})^2}{\#_{J_i<J_\theta} J_i}}, \tag{3}$$

where $\#_{J_i<J_\theta} J_i$ is the number of periods of fundamental frequency detected in the prosodic unit, such that the corresponding jitter value does not exceed the given threshold J_θ (thus, obvious outliers are not taken into account). Calculation rules for the shimmer-based features S_{avg} and S_{dev} are analogous but based upon the energy values of the localized F_0-periods.

Alcohol Blood Level	0.0	< 0.4	< 0.8	< 1.2	< 1.6	< 2.0	< 2.4
Recordings	32	20	20	18	20	7	3

Table 1: *Distribution of recordings in the corpus over alcohol blood level.*

4. Experiment

4.1. Database

For our experiments we used a collection of alcoholized speech samples assembled at the Police Academy of Hessen, Germany. It contains 120 readings (approx. 87 minutes) of the German version of the fable "The Sun and the Northern Wind", produced by 33 male speakers in different alcoholization conditions with alcohol blood level varying between 0 and 2.4 per mille. The phrasal units obtained from this corpus as described in section 2 have the following characteristics: average duration 2.3 sec, average speech tempo 20.8 PhU/min.

The distribution of the collected recordings over alcohol blood level is shown in table 1. For training and classification purposes the records were further divided in two classes: alcoholized (AL) and not alcoholized (NAL) with the boundary value 0.8 per mille.

4.2. Results

We employ Artificial Neuronal Networks (ANN) as classifier, whereby several *Multi-Layer Perceptron (MLP)*-topologies are trained and tested independently[1]. Two criteria are used to assess the achieved classification success:

1. RR_{BEST} — the highest recognition rate over different MLP-topologies obtained for the given set of classification features;
2. RR_{AVG} — the average recognition rate over all investigated MLP-topologies.

In both cases recognition rates are defined as the ratio of the number of correctly classified phrasal units to the total number of phrasal units in the set.

Due to the data sparsity we start by splitting the corpus in training and validation set, the latter acting as a test set at the same time. The neuronal networks are trained with the training set and the validation set is used to first determine the best MLP-topology for each combination of features, and then the best set of features. Both RR_{BEST} and RR_{AVG} are taken into account when deciding which set of features is the most useful for classification. However, since we test on the validation set, we give RR_{AVG} more influence on this decision, arguing that more useful features are in general capable of better classification, no matter how the topology of the neuronal network to be trained is specified.

First of all, we compared the strategy which uses phrasal units as prosodic units against the strategy which acquires prosodic units by splitting the speech signal in equal time intervals (with a length equal to the average length of phrasal units in the corpus). Using group *PM21* we found that the decrease in performance in the latter case amounted to almost 3 percentage points for both RR_{BEST} and RR_{AVG}, which proved the meaningfulness of phrasal units.

Our next goal was to determine the optimal size of the highpass-filter used to calculate the four jitter- and shimmer-based classification features. For this purpose we conducted a series of experiments using only these four features for training and classification. Table 2 shows the achieved values of RR_{BEST} for filter sizes $m = 2, 3, 5$. We see that the simple definition of jitter and shimmer as the normalized difference of two neighbor F_0- and energy-values, respectively, yields the best classification rates. In the following, we therefore keep these definitions.

[1] We employ MLP with no hidden layers, one hidden layer with 3 or 5 nodes and two hidden layers with respectively 5 and 3 nodes.

	$n = 2$	$n = 3$	$n = 5$
RR_{BEST}	**70.8%**	68.6%	68.0%

Table 2: *Influence of the filter size in the calculations of jitter and shimmer on the classification performance.*

PM21	JS4	VU11	LTM24	RR_{AVG} %	RR_{BEST} %
+	–	–	–	69.9	72.8
–	+	–	–	67.6	70.8
–	–	+	–	59.8	62.7
–	–	–	+	67.1	**84.7**
+	+	–	–	**72.2**	74.8
+	–	+	–	69.0	72.5
+	–	–	+	65.3	73.0
+	+	+	–	70.8	73.7
+	+	+	+	64.6	77.6

Table 3: *Recognition rates on the validation data set using different classification features; used groups are marked as "+", unused as "–".*

Next, we focused our attention on the optimal choice of the set of features for our task domain. We simplified the problem by looking for the optimal combination of the feature groups as they were defined in section 3, instead of ascertaining the usability of each feature for itself. Table 3 shows classification results for various combinations of feature groups.

We see that the combination of groups *PM21* and *JS4* on average leads to the best recognition rates, even though there was one MLP-topology which resulted in the absolutely best recognition rates when using classification features from *LTM24*.

For practical applications, it suffices to provide the entire speech signal (record) with only one label identifying the alcoholization of the speaker. Since the alcoholization is a stable speaker characteristic, the straightforward *majority-voting* solution can be applied, labeling the whole speech signal as alcoholized (AL) if the majority of the phrasal units it consists of are labeled so, and not alcoholized (NAL) otherwise. With this method we obtained recognition rates $RR_{BEST} = 80.2\%$, using the combination of feature groups *PM21* and *JS4*. Table 4 presents the confusion matrices for this combination, based on the phrasal units as well as on the entire records.

The purpose of our next experiment was to obtain classification results on unseen data using the features from the groups *PM21* and *JS4*. Because of the small size of the corpus we followed a *Leave-One-Out* strategy. The corpus was split into 5 equal parts, and 5 independent tests with the MLP-topology corresponding to RR_{BEST} were conducted. The resulting recog-

Phrasal units		
Reference	Hypothesis	
	NAL	AL
NAL	155 (79.1%)	41 (20.9%)
AL	42 (29.6%)	100 (70.4%)

Entire records		
Reference	Hypothesis	
	NAL	AL
NAL	15 (93.8%)	1 (6.2%)
AL	3 (33.6%)	6 (66.7%)

Table 4: *Confusion matrices for phrasal units and entire records; groups* PM21 *and* JS4.

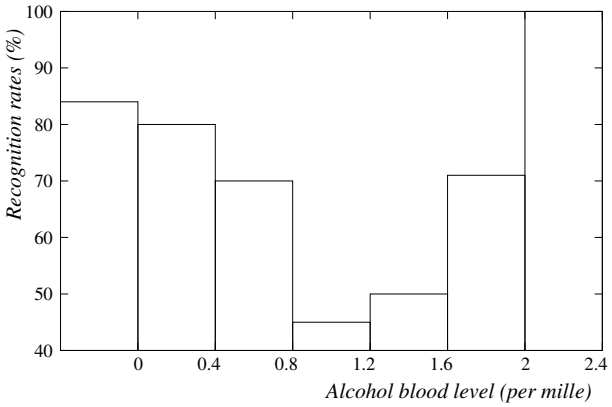


Figure 2: *Distribution of record-based recognition rates over alcohol blood level.*

nition rates amounted to 61.7% when considering each phrasal unit for itself, and to 68.8% when classifying record-wise.

Finally, the diagram in figure 2 makes clear that most difficulties occur when dealing with alcohol blood level close to 0.8 per mille, the threshold that we have chosen as a boundary between two classes (AL and NAL). Almost all records exhibiting very high or very low degree of alcoholization were classified correctly, whereas the vast majority of mistakes were made within close proximity of the boundary.

5. Conclusion

We have shown that the problem of automated recognition of alcohol intoxication in human speakers can be tackled using prosodic speech characteristics. We also have demonstrated how to extract prosodic features with good classification abilities from a speech signal without a lexical segmentation of the signal (e.g. word extraction). Indeed, it appears to be sufficient for a successful classification to relate vectors of structural prosodic features to signal intervals localized solely by means of basic prosodic features. Moreover, we have shown that such intervals (called *phrasal units*) often correspond to syntactic structures of the language.

We also determined the set of structural prosodic features capable of the best classification in the domain of automated detection of alcoholization. Along with features which describe macro-tendencies in the fundamental frequency and energy, jitter- and shimmer-based characteristics turned out to be important for successful classification.

In our experiments on a two-class classification problem (alcoholized vs. not alcoholized), we achieved a recognition rate of almost 69% on unseen data, most problems being encountered in the region close to the boundary between the two classes.

6. References

[1] Amir N. and Ron S: *Towards an Automatic Classification of Emotions in Speech*, ICSLP'98, Sydney, v. 3, pp. 555–558.

[2] Batliner A., Kompe R., Kießling A., Mast M., Niemann H. and Nöth E.: *M=Syntax+Prosody: A Syntactic–Prosodic Labelling Scheme for Large Spontaneous Speech Databases*, SpeechComm, North Holland, v. 25, pp. 193–222, 1998.

[3] Brenner M. and Cash J.R.: *Speech Analysis as an Index of Alcohol Intoxication – the Exxon Valdez Accident*; Aviation, Space and Environement Medicine, 1992.

[4] Gaillard A.W.K. and Wientjes C.J.E.: *Mental Load and Work Stress as Two Types of Energy Mobilization*, Work and Stress, pp.141–152, 1994.

[5] Hess W.: *Pitch Determination of Speech Signals*, Springer Series of Information Sciences, Springer–Verlag, Berlin, 1983.

[6] Huber R.: *Prosodisch-linguistische Klassifikation von Emotion*, PhD. dissertation, University of Erlangen–Nuremberg, 2001.

[7] Kießling A.: *Extraktion und Klassifikation Prosodischer Merkmale in der Automatischen Sprachverarbeitung*, Berichte aus der Informatik, Shaker Verlag, Aachen, 1997.

[8] Klingholz F., Penning R. and Liebhardt E.: *Recognition of Low-Level Alcohol Intoxication from Speech Signal*, Journal of the Acoustical Society of America, n. 84, pp. 929–935, 1988.

[9] Kompe R.: *Prosody in Speech Understanding Systems*, Lecture Notes for Artificial Intelligence, Springer–Verlag, Berlin, 1997.

[10] Lester L. and Skousen R.: *The Phonology of Drunkeness*, Papers from the Parasession on Natural Phonology, Bruck, Fox and LaGaly (Eds.), Chicago Linguistic Society, Chicago, 1974.

[11] Scherer K.R.: *Speech and Emotional States*, Darby J.K. (Ed.), Grune and Stratton, New York, 1981.

[12] Slyh R.E.,Nelson W.T. and Hansen E.G.: *Analysis of Mrate, Shimmer, Jitter and F_0 Contour Features across Stress and Speaking Style in the SUSAS Database*, ICASSP'99, Phoenix, Arizona, v. 4, pp. 2091–2094.

[13] Steeneken H.J.M. and Hansen J.H.L.: *Speech under Stress Conditions: Overview of the Effect on Speech Production and on System Performance*, ICASSP'99, Phoenix, Arizona, v. 4, pp. 2079–2082.

[14] Zhou G., Hansen J.H.L. and Kaiser J.F.: *Methods for Stress Classification: Nonlinear Teo and Linear Speech Based Features*, ICASSP'99, Phoenix, Arizona, v. 4, pp. 2087–2090.