Does it Groove or Does it Stumble - Automatic Classification of Alcoholic Intoxication Using Prosodic Features

Florian Hönig, Anton Batliner, Elmar Nöth

Pattern Recognition Lab, University of Erlangen-Nuremberg

{hoenig,batliner,noeth}@informatik.uni-erlangen.de

Abstract

This paper studies how prosodic features can help in the automatic detection of alcoholic intoxication. We compute features that have recently been proposed to model speech rhythm such as the pair-wise variability index for consonantal and vocalic segments (PVI) and study their aptness for the task. Further, we use a large prosodic feature vector modelling the usual candidates – pitch, intensity, and duration – and apply it onto different units such as words, syllables and stressed syllables to create generalizations of the rhythm features mentioned. The results show that the prosodic features computed are suitable for detecting alcoholic intoxication and add complementary information to state-of-the-art features. The database is the intoxication database provided by the organizers of the 2011 Interspeech Speaker State Challenge.

Index Terms: prosody, rhythm, alcohol intoxication

1. Introduction

To improve classification performance for phenomena that – hopefully – manifest themselves within the speech signal, be this, e. g. age, gender, emotion, or intoxication, we can try to optimise with regard to several aspects, e. g., classifiers, feature reduction or selection, and, of course, the types of features themselves. The acoustic feature vector provided by the organizers of the Interspeech 2011 Speaker State Challenge [1] is fairly complete, which makes it even more challenging trying to improve its performance, by adding information obtained with other types of features. In this vein, we use a large prosodic feature vector for the challenge task that has been proven to be efficient for the automatic assessment of non-native speech [2]. The obvious idea is that intoxication is just a variety of speech (register), the same way as any other idiosyncrasy or non-native trait is; as for considerations along similar lines, cf. [3].

The authors of the present paper are all affiliated with and partly identical with the organizers of the Challenge. Thus, they do not contribute in the Challenge. To ensure comparability of results, we strictly follow the procedures defined within [1], and we do not use any data or information that were not available to all competitors.

2. Database

The database used is the ALCOHOL LANGUAGE CORPUS (ALC) [3] which has been provided by the organizers of the Interspeech 2011 Speaker State Challenge. In accordance with the etiquette of the Challenge, we do not describe further details and refer to [1]. The data is divided into three partitions and into NAL/AL (non-alcoholised with BAC $\leq$ 0.5 per mill vs. alcoholised with BAC $>$ 0.5 per mill); in parentheses, number of cases which we will call (segmented) *chunks* is given: Train (3750/1650), Develop (2790/1170), and Test (1620/1380). All 162 speakers were recorded in three different speech registers: read speech, spontaneous speech, and command & control. The speech register is not given in the Challenge distribution.

3. Features

To start with, we use the extended set of 4368 openSMILE features provided by the organizers of the challenge and described in [1]; it is fairly representative for state-of-the-art acoustic features. In addition, we compute a large number of features modelling different prosodic traits. All processing is done fully automatic; however, we use the phonemic transcription of the spoken words which is provided with the database.

3.1. Identification of Stressed Syllables

For some of the prosodic features that will be described below, it is necessary to know which syllables are stressed. A syllable is considered as stressed if

(a) it is a mono-syllabic word bearing a primary or secondary phrase accent, or if

(b) it is part of a multi-syllabic word, having either secondary or primary word accent.

The provided distribution of ALC includes word stresses and a classification into (non-)function-words for Train and Develop. However, no phrase accents are marked, and the categorization into (non-)function-words can only be used as an approximate substitute. Most importantly, both word stress and function word categorization are missing for Test.

We therefore resort to an automatic classification system based on [2]. It comprises an acoustic component that estimates the conditional probability for each syllable to be stressed, and an n-gram for modelling a priori probabilities for stress sequences. The acoustic component applies Linear Discrimination Analysis (LDA) to a feature vector (see Section 3.3) computed from the nuclei of the current and ± 2 neighbouring syllables. A syllable stress 4-gram models the a priori probabilities[1]. The acoustic likelihoods and n-gram probabilities have equal weight; an estimate of the most likely sequence of stressed and unstressed syllables is decoded with the help of a Viterbi beam search. The system is trained with two hours of accordingly annotated read speech from four speakers (a German version of

[1] For example, an utterance composed of two syllables has an a priori probability of 62.5% for a singular stress on the first syllable, while a singular stress on the second syllable has 29.2%.

the native data used in [2]); the weighted average recall for the binary classification problem stressed vs. unstressed is 78.8% (unweighted: 78.0%) in a leave-one-speaker-out evaluation.

Prior to the computation of the features, the DC offset is removed from each chunk and the maximal amplitude is normalised. Short-time energy and fundamental frequency (F0) are computed on a frame-by-frame basis (step size 10 msec). For estimation, F0 is assumed to range between 59 and 550 Hz; frame size is 3 times the largest assumed fundamental period, i. e. around 50 msec. F0 is logarithmised and normalised (per chunk), and interpolated during non-voiced segments. Although phoneme alignments are provided by the database, segment boundaries are re-estimated from a forced alignment of the transcribed phonemes using cross-word triphone HMMs. The reason for re-estimating the segmentation is to achieve segmentations as similar as possible to the training material of the syllable stress classification module described above.

In the following, we shortly describe the different types of prosodic features implemented.

3.2. Specialised Prosodic Feature Sets

Duration Measures (Duration): A basic but fundamental property of speech is how fast something is said. We compute the total duration of the chunk, and the average duration of all vocalic segments for each chunk (two features).

Isochrony Features (Iso): In order to capture possible isochrony properties [4], we calculate the distances between the centers of consecutive stressed and consecutive unstressed syllables. The centers are identified as the frames with maximal short-time energy within a nucleus. We compute six chunk-level features: mean distances between stressed and between unstressed syllables, standard deviations of those distances, and the ratio of means and standard deviations.

Variability Indices (PVI): Following [5], we identify vocalic and consonantal segments and calculate the raw Pairwise Variability Index (rPVI) which is defined as the absolute difference in duration of consecutive segments and its normalised version nPVI (rPVI divided by the mean duration of the segments) for vocalic and consonantal segments (four features).

Global Interval Proportions (GPI): Following [6], we compute the percentage of vocalic intervals (of the total duration of vocalic and consonantal segments), and the standard deviation of the duration of vocalic and consonantal segments of a speech segment (three features).

In the experimental evaluation, the feature groups described in this Sub-Section are either analysed individually or pooled into **Rhythm-All** (17 features).

3.3. General-Purpose Prosodic Features

In addition to the specialised features, we apply our comprehensive general-purpose prosody module which has already been successfully applied to as diverse problems as phrase accent and phrase boundary recognition [7], word accent position classification [2], and emotion recognition [8]. The features are based on duration, energy, pitch, and pauses, and can be applied to arbitrary units of speech (here, the units used are words, syllables, and nuclei). Some of the energy and duration based features are normalised versions of a quantity, e. g. the duration of a word divided by the average duration of that specific word. The statistics necessary for these normalization measures have been estimated on the same data that has been used for the syllable stress classification module described above. Trying to be as exhaustive as possible, we use a highly redundant feature set leaving it to data-driven methods to find out the relevant features. However, the procedure is based on knowledge and not on brute force. Features are extracted from a context of one or two units. A varying number of neighbouring units are used to extract features for the current unit: For a *local variant*, the current unit and the context including the current and the following unit are used for feature extraction. A *context-capturing* variant uses contexts up to ± 2 neighbouring units. This process is detailed in Table 1. The features are abbreviated as follows: *duration features 'Dur':* absolute (Abs) and normalised (Norm); *energy features 'En':* regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max) with its position on the time axis (MaxPos), absolute (Abs) and normalised (Norm) values; *F0 features 'F0':* regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max), minimum (Min), onset (On), and offset (Off) values as well as the position of Max (MaxPos), Min (MinPos), On (OnPos), and Off (OffPos) on the time axis; *length of pauses 'Pause':* silent pause before (Pause-before) and after (Pause-after).

Table 1: Prosodic features and their context. Filled circles indicate which contexts in columns 2–6 are used for the 31 local features; for the 100 context-capturing features, additionally the contexts indicated by empty circles are used. Curly brackets indicate that all the features displayed in these three rows are computed for all contexts in the three rows in columns 2–6.

features for the actual unit '0'	**context size**				
computed from up to $\pm$ units	-2	-1	0	1	2
Dur: Norm, Abs; En: RegCoeff,		○	●	○	
MseReg, Mean, Abs, Norm;	○		●		
F0: RegCoeff, MseReg, Mean		○		○	
En: Max, MaxPos		○	●	○	
F0: Max, MaxPos, Min, MinPos		○	●	○	
F0: Off, Offpos		○	●		
F0: On, Onpos			●	○	
Pause-before		○	○		
Pause-after			●	○	

Additionally, we compute four features over a larger context of ± 7 units (or less, if the utterance is shorter), representing local estimates of global properties: *RateOfSpeech* (average speech rate), *DurTauLoc* (average duration), *EnTauLoc* (average energy) and *F0MeanGlob* (average fundamental frequency). These are appended to both the local and context-capturing configuration, ending up with 35 and 104 features per unit, respectively. A more detailed overview of the prosodic features is given in [7].

We now use all these prosodic features computed over different units and contexts to construct (again highly redundant) extensions of the *Iso*, *PVI* and *GPI* features (in total 523):

(1) we compute the context-capturing features for all stressed syllables and nuclei of stressed syllables, and the local features for all words, syllables and nuclei, and use the mean values of these features as speaker-level *extended Iso* features (2*104 + 3*35 = 313 features); words, syllables and nuclei at the start and end of the chunks that do not provide enough context units are skipped;

(2) for the *extended PVI* features, the mean absolute difference of the local features of consecutive words, syllables and nuclei (3*35 = 105 features) is computed;

(3) the standard deviations of the local features of all words, syllables and nuclei (3*35 = 105 features) represent the *extended GPI* features.

A last group of 11 global features reflects number and length of voiced and unvoiced segments (as evident from F0 extraction, not from phoneme segmentation), ratios of that numbers, and the standard deviation of F0.

In the experimental evaluation, the features described in this Sub-Section are grouped into

(a) **Pros-Normal**: those 326 features that do not need information about syllable stress,

(b) **Pros-Stressed**: those 208 features that do, and

(c) **Pros-All**: all 534 features.

Similarly, all features described throughout this whole *Section* are grouped into **All-Normal** (335 features) and **All-Stressed** (216 features). The total set of the 551 supra-segmental features described here is called **SUPRA**.

4. Experiments and Results

In order to study the aptness of the proposed feature groups for the task, we evaluate the classification performance on Develop when training with Train. We compare two classifiers: LDA and SVM[2] (linear kernel, $C = 1$). As the performance criterion of the Challenge is unweighted average, we have to account for the unbalanced classes in Train. For LDA, this is trivially done by setting equal a priori class probabilities. For SVM, we achieve a balanced training set by keeping all instances of the more infrequent AL class and randomly sampling the same number of instances from the more frequency NAL class.

The results for the different feature groups are detailed in Table 2. In general, both classifiers seem to be equally suited for this task and these features. It is evident that none of the individual rhythm feature groups Duration, Iso, PVI and GPI achieves much more than random guessing (50%) for both classifiers. However, combining them improves the results, achieving at least 56.2% with LDA. The generic prosody module features achieve somewhat higher classification accuracies, and it is obvious that including information about the syllable stress pays off (e. g. LDA on Pros-All, 62.6% versus Pros-Normal, 61.3%). Finally, the combination of all proposed features results in a further improvement for the SVM, which scores 62.4%.

The reported classification results are well above chance but clearly below those reported in [1] for the baseline openSMILE feature set (65.3%, for Train vs. Develop). Therefore, we tried to combine our proposed features with the openSMILE feature set to find out whether we can add complementary information.

For the openSMILE features, LDA was well below SVM in terms of classification performance, so we used SVM[3] only for the remaining experiments. To start with, we compared our features, the openSMILE features, and a late fusion[4] of both

[2] $C = 1$ was found optimal among the powers of ten for performance of the SUPRA features on Develop

[3] We chose $C = 0.01$ for the openSMILE SVM system, the optimum among the powers of ten for the performance on Develop.

[4] We fitted logistic models to the output in order to achieve probability estimates. Fusion was done by multiplying the class probabilities, exponentially weighted with weights optimised on Develop.

Table 2: Unweighted average recall in % on Develop when training with Train, for different groups of the proposed supra-segmental features.

Features	LDA	SVM
Duration	52.2	52.3
Iso	53.4	52.4
PVI	51.9	52.4
GPI	51.7	52.4
Rhythm-All	56.2	53.8
Pros-Normal	61.3	61.4
Pros-Stressed	60.2	58.5
Pros-All	**62.6**	62.3
All-Normal	61.2	61.1
All-Stressed	60.3	59.3
SUPRA = All	62.3	**62.4**

SVM systems on Develop when training with Train (see first column with figures in Table 3). For openSMILE, we obtained a slightly lower score of 64.8% than reported in [1], which can be explained by our simpler resampling scheme. The fusion of both systems yields a marked improvement to 67.0% which indeed indicates that our features contribute useful information.

Next, we compared the performances on Test when training with Train+Develop (see second column with figures in Table 3; an overview of all results on Test is given in Table 4). Surprisingly, we observed a pronounced degradation in the performance of our features here: They yield only 57.9% vs. 65.9% in [1], and even hurt in the fusion system (64.2%).

In search for an explanation for this mismatch results, we began inspecting the transliteration of the database. While each speaker has 90 chunks in Train and Develop, there are only 60 chunks contained per speaker in Test. Each of the 90 chunks in Train and Develop has an ID that identifies the prompt used to elicit speech (e. g. `*_A_*_002.wav`, according to the transcripts, was a prompt asking the speaker about his or her last holiday). That ID is obfuscated for Test, but we were able to identify 30 prompts that – according to the transcripts – did not occur at all in Test, i. e. there's a mismatch with respect to the spoken texts between Train and Develop on the one hand, and Test on the other hand. By removing those 30 prompts from Train and Develop, we obtained the versions Train-matched and Develop-matched which are smaller but better matched to Test. Interestingly, all these 30 prompts are from the sober recording sessions (IDs matching `*_N_*.wav`). That means that Train-matched is also better balanced (1650 AL vs. 1950 NAL chunks) than Train (1650 AL vs. 3750 NAL), and that given our resampling scheme, the number of used training instances is the same for Train and Train-matched ($2 \cdot 1650$). It will also be relevant for the individual classification performance on the two classes NAL and AL (end of this Section).

When training with Train-matched and testing with Develop-matched (third column with figures in Table 3), the systems behave similar to Train/Develop, but on a lower level: openSMILE (61.4%) is better than SUPRA (56.8%) but the fusion again is best (62.6%). Employing Train-matched+Develop-matched to score for Test is indeed helpful (last column in Table 3): again, openSMILE (66.3%) is better than SUPRA (60.4%), but the fusion (67.6%) is best. Thus, we

Table 3: Unweighted average recall of SVM in % for the *openSMILE* and the *SUPRA* feature set, and their (late) fusion. (*) is the openSMILE baseline performance as reported in [1].

Training on:	Train	Train+Dev.	Train-matched	Train+Dev.-matched
Evaluation on:	Develop	Test	Dev.-matched	Test
openSMILE	64.8	**65.9***	61.4	66.3
SUPRA	62.4	57.9	56.8	60.4
openSMILE + SUPRA	**67.0**	64.2	**62.6**	**67.6**

Table 4: Unweighted average recall (UA), weighted average recall (WA), and recall for the classes NAL and AL, of SVM on Test for the different constellations: Condition 'orig' refers to training with the original Train and Develop set; condition 'matched' to using Train-matched and Develop-matched. 'openSMILE + SUPRA' refers to the (late) fusion of the SVM systems using openSMILE and SUPRA, respectively.

Constellation	Condition	**% UA**	% WA	% NAL	% AL
Baseline openSMILE [1]	orig	65.9	66.4	-	-
SUPRA	orig	57.9	56.9	45.6	70.1
openSMILE + SUPRA	orig	64.2	63.7	57.7	70.7
openSMILE	matched	66.3	66.8	72.0	60.6
SUPRA	matched	60.4	60.3	59.6	61.2
openSMILE + SUPRA	matched	**67.6**	**68.0**	72.4	62.8

have shown that our features really add useful information and further improve performance of state-of-the-art features. However, our features seem to be more susceptible to a mismatch between training and test with respect to speech material.

Table 4 details recall for the classes NAL and AL. It is interesting that for the original training set, NAL is recognized considerably worse than AL (last two columns, upper rows in Table 4), while there is a reverse tendency for the matched training set (last two columns, lower rows in Table 4). An explanation for this could be that the NAL class is particularly affected by the train/test mismatch because all of the 30 prompts contained only in Train/Develop are from the sober recording sessions.

5. Conclusion and Outlook

Read and spontaneous speech differ in several respect [9]; as far as prosody is concerned, this relates, e.g., to number of placement of accents and boundaries, speech rate and – most probably in trading relation – number of reduction phenomena (schwa etc.). This might even hold for different narratives. Thus it is possible that a mismatch between train and test set with respect to these phenomena makes it more difficult to compute reliable estimations for our prosody and rhythm based features. In contrast, the openSMILE features used for the baseline are frame-based and computed for larger segments (here: chunks) whose boundaries are obtained automatically [1]; thus they do not take into account the internal phrasal and accentual structuring of speech – this makes them less affected by the mismatch.

We have shown that – given matched conditions with respect to the type of spoken utterances – this vector adds information and classification performance: It grooves. A next step will be a close look at the characteristics of our prosodic-rhythmic features, in order to find out whether intoxicated speech really stumbles somehow, in relation to 'normal' speech, or whether the (prosodic) language model used is simply susceptible to linguistic structures not seen in the training data.

6. Acknowledgements

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the project *C-AuDiT* under Grant 01IS07014B, and by the German Ministry of Economics (*BMWi*) in the framework of the project *AUWL* under grant KF2027104ED0. The responsibility lies with the authors.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The Interspeech 2011 Speaker State Challenge," in *Proc. Interspeech*, 2011, no pagination.

[2] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners," in *Proceedings of SLATE*, Wroxall Abbey, 2009.

[3] F. Schiel, C. Heinrich, and V. Neumeyer, "Rhythm and Formant Features for Automatic Alcohol Detection," in *Proc. Interspeech 2010*, Chiba, Japan, 2010, pp. 458–461.

[4] D. Abercrombie, *Elements of General Phonetics*. Edinburgh: University Press, 1967.

[5] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology VII*, C. Gussenhoven and N. Warner, Eds. Mouton de Gruyter, 2002, pp. 515–546.

[6] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," in *Proc. Speech Prosody*, Aix-en-Provence, 2002, pp. 115–120.

[7] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The Prosody Module," in *Verbmobil: Foundations of Speech-to-Speech Translations*, W. Wahlster, Ed. Springer, 2000, pp. 106–121.

[8] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Tales of Tuning – Prototyping for Automatic Classification of Emotional User States," in *Proc. Interspeech*, Lisbon, 2005, pp. 489–492.

[9] J. Llisteri, "Speaking styles in speech research," in *Proceedings of the ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, Dublin, 1992, 28 pages.