The Influence of Alcohol Intoxication on Silent Pause Duration in Spontaneous Speech

Hong Zhang[1*]

[1]*Department of Linguistics, University of Pennsylvania, Philadelphia, PA, USA*
*zhangho@sas.upenn.edu

This study looks at the effect of alcohol intoxication on the distribution of silent pause durations in spontaneous monologue and dialogues using speech from the German Alcohol Language Corpus. Distribution of silent pause durations for each speaker is represented as the joint density function of silent pause duration against the following speech duration. Density estimations are then projected onto lower dimensional space through Singular Value Decomposition (SVD). Group difference between sober and intoxicated speakers can be effectively visualized using the first 3 dimensions in the derived space. Good classification results are achieved using Support Vector Machines (SVM) with gaussian kernel in the binary intoxication classification task. These results suggest that alcohol intoxication has global effect on the distribution of silent pause duration relative to the following speech utterance durations. The derived features can provide a representation for tasks such as alcohol intoxication detection.

INTRODUCTION

Alcohol intoxication can cause deterioration in various aspects of cognitive processing, which may not only lead to problems in the motor control of speech production, but also result in deficits in speech planning [1]. Previous research has shown that speakers under the influence of alcohol intoxication tend to produce higher overall fundamental frequency (F0) [2], increased rate of disfluencies [3] and changed short-term energy function and F0 contour [2,4]. Practically speaking, successful detection based on altered speech signal caused by alcohol intoxication can be helpful in the prevention of alcohol related health issues, such as drunk and drive. To facilitate the development of systems that improve the efficiency of alcohol intoxication detection, Alcohol Language Corpus (ALC) [5] has been developed and used for a speaker state detection challenge [6]. In the challenge, a common set of acoustic features were used to train systems on utterance level classification with a baseline test accuracy (Unweighted Average Recall, UAR) of 65.9%. The best system [7] following the paradigm of this challenge achieved a UAR score of 71.4%. Here we ask the question of how the distribution of silent pause durations changes when the speaker is alcohol intoxicated.

In this study, we take the same ALC and ask if the distribution of pause durations changes at individual level when the speaker is intoxicated. As suggested in [2,4], the effect of alcohol intoxication on speech is highly speaker dependent, meaning that the same effect may surface in the opposite direction on the same acoustic measures for different individuals. This property of intoxicated speech may partly explain the relatively poor performance of utterance level classifiers, even if trained using state-of-the-art neural network architecture with rich acoustic representation [8]. Therefore we take a global perspective, with the goal of exploring the feature space that can efficiently represent the change induced by alcohol intoxication.

METHOD

The speech data

ALC is a collection of speech from a total of 162 German speakers (85 males, 77 females) produced in two conditions: sober and alcohol intoxication at a self-chosen intoxication level. The actual blood alcohol concentration (BAC) level was measured immediately before recording. Speech tasks used in the corpus include read speech, monologue (such as picture description, commands and instructions) and short conversations. Speech from the picture description task and short dialogues with the interviewer is chosen for the current study. The speech is recorded with a sample rate of 44.1 kHz with 16 bit rate. Verbatim transcriptions at phoneme level are available and the recordings are aligned.

Feature generation

The distribution of pause durations is represented as the joint density function of silent pause duration against the speech segment duration immediately following the silence. For each individual in each intoxication state, the joint distribution function is derived from all the selected speech in that condition. All silent pauses longer than 50ms are considered in the calculation. A 100x100 grid is used to sample from the 2D density function. Therefore the joint distribution is represented by a 100x100 matrix per speaker condition.

To reduce the sparsity of this representation and achieve a compact representation of the distributions, each 100x100 matrix is flattened as a 1x10,000 vector. SVD is then performed on the full 162x10,000 matrix stacked from all the individual feature vectors in each intoxica-

tion condition. The left singular matrix (dimension 162x162) is used as the final feature representation of all the speakers in each state, where each row corresponds to an individual in the given condition.

Feature evaluation

The derived feature vector is first evaluated by visualizing individual speakers in two conditions using the first three dimensions in the derived 162-dimensional feature space. A binary classification task is then performed using a simple SVM with the full feature vector to classify each individual as sober or intoxicated.

RESULTS

Fig 1 illustrates the difference in the joint distribution of silent pause duration and the following speech segment duration for a single speaker in intoxicated (left) and sober (right) conditions. A clear distinction between the two joint distributions can be observed. Silent pauses produced in intoxicated condition appear to be shorter, and the overall distribution is multi-modal compared to the sober condition.

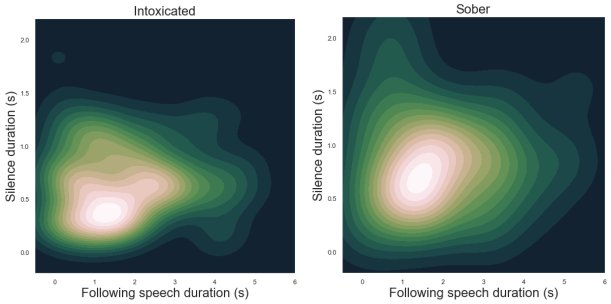

Fig. 1. 2D density plot of the joint distribution of silent pause duration (y-axis) against following speech segment duration (x-axis) for a single speaker in intoxicated (left) and sober (right) conditions.

The scattered plots for all speakers in sober and intoxicated conditions in the first three dimensions in the derived space are plotted in Fig 2. In the coordinate defined by the first and second dimension, intoxicated speakers are distributed mainly in the lower right corner, while in the coordinate defined by the second and third dimension, intoxicated speakers are mainly distributed to the left of the vertical line as shown in the figure. Thus the derived feature space is able to represent the group difference in the distribution of silent pause durations as measured by its relation with the following speech duration.

To test the performance of this derived feature space in distinguishing speakers in intoxicated from sober condition, speakers are randomly divided into training and testing set with a 3-to-1 ratio. The training set contains 122 speakers in both intoxicated and sober states, whereas the testing set includes the paired intoxicated and sober states for the rest of the speakers. Therefore the task can be understood as distinguishing between sober and intoxicated states when the speaker is given.

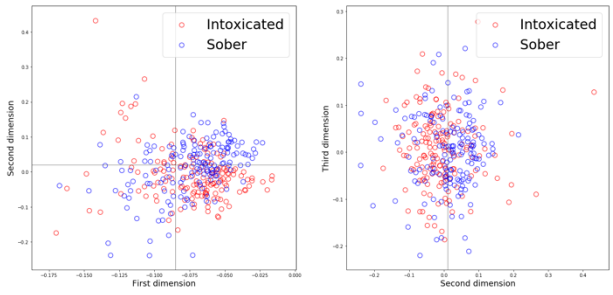

Fig. 2. Scattered plots of individual speakers in the derived space in intoxicated and sober conditions. The left plot shows the distribution along the first (x-axis) and second (y-axis) dimensions, and the right plot shows the distribution along the second (x-axis) and third (y-axis) dimensions.

A simple SVM classifier with gaussian kernel is used for this classification task. Tuning parameters are selected through a grid search with 10-fold cross-validation on the training set. Table 1 reports the simple testing accuracy for models trained on different percentages of the training data.

Results from this simple classification experiment suggest that the derived features are efficient in classifying speaker intoxication states between sober and intoxicated, as 20% of the training data can already achieve above-chance performance. Training on the full training set is able to yield a pretty high testing accuracy (93.75%) on the unseen test examples.

Tab. 1: Testing accuracy for SVMs training on different percentages of the training data

Training data	20%	40%	60%	80%	100%
Accuracy	0.65	0.7125	0.8875	0.7375	0.9375

DISCUSSION

In this study, we addressed the question of how alcohol intoxication affects the distribution of pause durations in spontaneous monologue and conversations at individual speaker level. Using speech produced from the picture description and short conversation tasks in ALC, we demonstrated that the effect of alcohol intoxication on silent pause duration can be effectively represented through the relation between silent pause duration and the following speech segment duration. Dimensionality reduction techniques, such as SVD, are able to offer compact parameterization of the differences observed from the joint distribution. The derived features appear to be highly efficient in intoxication identification.

The good performance of our feature space in the simple SVM setting also shows that although individual variation in particular acoustic dimensions can be problematic in deriving good representations of alcohol intoxicated speech, features derived from rich characterization of joint distributions of related variables can generate robust parameterizations for speaker state detection.

REFERENCES

[1] Peterson, J. B., Rothfleisch, J., Zelazo, P. D., & Pihl, R. O. (1990). Acute alcohol intoxication and cognitive functioning. *Journal of studies on alcohol*, *51*(2), 114-122.

[2] Baumeister, B., Heinrich, C., & Schiel, F. (2012). The influence of alcoholic intoxication on the fundamental frequency of female and male speakers. *The Journal of the Acoustical Society of America*, *132*(1), 442-451.

[3] Schiel, F., & Heinrich, C. (2015). Disfluencies in the speech of intoxicated speakers. *International Journal of Speech, Language & the Law*, *22*(1).

[4] Heinrich, C., & Schiel, F. (2014). The influence of alcoholic intoxication on the short-time energy function of speech. *The Journal of the Acoustical Society of America*, *135*(5), 2942-2951.

[5] Schiel, F., Heinrich, C., & Barfüsser, S. (2012). Alcohol language corpus: the first public corpus of alcoholized German speech. *Language resources and evaluation*, *46*(3), 503-521.

[6] Schuller, B., Steidl, S., Batliner, A., Schiel, F., & Krajewski, J. (2011). The INTERSPEECH 2011 speaker state challenge. In *12th Annual Conference of the International Speech Communication Association*. Florence, Italy, 2011, pp. 3201–3204.

[7] Bone, D., Li, M., Black, M. P., & Narayanan, S. S. (2014). Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors. *Computer speech & language*, *28*(2), 375-391.

[8] Berninger, K., Hoppe, J., & Milde, B., (2016) Classification of Speaker Intoxication Using a Bidirectional Recurrent Neural Network, in: P. Sojka, A. Horák, I. Kopeček, K. Pala (Eds.), *Text, Speech, and Dialogue,* Springer International Publishing, Brno, Czech Republic, pp. 435–442.