Intoxication Detection using Phonetic, Phonotactic and Prosodic Cues

Fadi Biadsy[1]*, William Yang Wang*[1]*, Andrew Rosenberg*[2]*,Julia Hirschberg*[1]

[1]Computer Science Department, Columbia University, New York, USA
[2]Computer Science Department, Queens College (CUNY), New York, USA
{fadi,julia}@cs.columbia.edu, yw2347@columbia.edu, andrew@cs.qc.cuny.edu

Abstract

In this paper, we investigate multiple approaches for automatically detecting intoxicated speakers given samples of their speech. Intoxicated speech in a given language can be viewed simply as a different accent of this language; therefore we adopt our recent approach to dialect and accent recognition to detect intoxication. The system models phonetic structural differences across sober and intoxicated speakers. This approach employs SVM with a kernel function that computes similarities between adapted phone GMMs which summarize speakers' phonetic characteristics in their utterances. We also investigate additional cues, such as prosodic events, phonotactics and phonetic durations under intoxicated and sober conditions. We find that our phonetic-based system when combined with phonotactic features provides us with our best result on the official development set, an accuracy of 73% and an equal error rate of 26.3%, significantly higher than the official baseline.

1. Introduction

The impaired judgement and slowed response-times associated with intoxication can have dramatic consequences. In the United States, hundreds of thousands of people are injured from drunk driving accidents every year. Legal requirements of sobriety while driving, flying and in other sensitive positions have been in place for many years throughout the world. A system to detect that a person is intoxicated through minimally invasive means would be able to significantly aid in the enforcement of these laws, and ultimately save lives. Additionally, there are medical implications in detecting the degree to which a person is intoxicated both for the application of care and avoiding medication interactions with alcohol and other intoxicants. Exploration of techniques for the recognition of alcohol-induced intoxicated speech is the focus of the Intoxication Sub-Challenge of the Interspeech 2011 Speaker State Challenge [1].

From the point of view of speech research and technology, Intoxication can be viewed as a state that has a temporary effect on discernible features of a person's speech. We can thus consider intoxicated speech to be similar to speech resulting from the temporary experience of other kinds of speaker state, whichhave been well studied in the literature, including the classic emotions such as anger, happiness, and sadness. Similar to the effect of emotion on speech, it has been found that there are aspects of speech that are temporarily modified through intoxication, while the broad physiological properties of a person's speech production system – e.g. vocal tract length, mouth and nasal cavity size and shape – remain unchanged. The task in studying intoxicated vs. non-intoxicated speech is thus to identify which speech properties are modified by alcohol-based intoxication and how.

Alcohol is a neurological depressant; the release of potassium to the bloodstream through metabolization of ethanol slows neuron firing intervals. The broad effect of metabolized alcohol is an overall depression of the central nervous system. This slowing of neural responses can lead to asynchronicity in the timing of speech articulators, including the lips, tongue, jaw, vocal chords. This may have a profound impact on the realization of phones and their durations.

In this work, we investigate three speech qualities that may be impacted by intoxication. We first explore symbolic prosodic qualities of intoxicated and sober speech. The hypothesis here is that intoxicated speakers may use prosody in predictable ways, realized through changes in phrasing and accenting behavior. Next, we investigate the variation of phone durations and phonotactic constraints under intoxicated and sober conditions. We hypothesize that articulator mistiming, and modified speech rhythm in intoxicated speech may be observable through the relative duration of phone units. Third, we investigate the use of phone-sensitive acoustic modeling to detect intoxication. This approach is based on the idea that the acoustic features corresponding to specific phones may systematically vary in intoxicated and sober states. This system has been successfully applied to the identification of spoken dialects and accents with state-of-the-art performance on variety of dialects and accents of multiple languages. The motivation here is that intoxicated speech in a given language (e.g., German here) can be viewed simply as a different accent of this language.

We discuss the materials we use in this work in the next Section. In Section 3, we describe our prosodic modeling approach and its results. Similarly, we describe our phone duration and phonotactic modeling in Sections 4 and 5, respectively. Then, we describe our novel approach for this task using our phonetic-based SVM kernel approach along with our experimental results. Finally, in Section 7, we conclude and propose directions for future work.

2. Materials

For our experiments, we use the Interspeech 2011 Speaker State Challenge German Alcohol Language corpus [1]. The official training set contains 3750 sober utterances and 1650 intoxicated utterances, giving a majority class for this data set of 69.4%. The official development set contains 2790 sober utterances and 1170 intoxicated utterances, for a majority class of this development data set of 70.5%. The official "weighted accuracy" of the baseline system on the development set is 69.2% [1].

In addition to evaluating our system on this official development set (while training on the official training set), we also decided to test our system using a (nearly) balanced data for training and testing. We attempted to balance both number of speakers and number of utterances at the same time. To do that, we first combine the training and development sets and then randomly select 20% of the speakers (from the grouped data) from each class as the new development set and 80% for training. Next we attempt to equalize the number of utterances in both classes in training and testing, by downsampling. The details of this selection are presented in Table 1. We denote this selection as the *balanced set*. For this new division, the majority class

of the development set is now 52%; the majority class of the training set is 53.7%.

Class	# Train Spk.	# Train Utt.	# Test Spk.	# Test Utt.
Intoxicated	74	2220	20	600
Sober	83	2573	21	651

Table 1: Number of speakers and utterances in our balanced set

3. Prosodic Variation

We first hypothesize that intoxicated speakers may use prosodic contours differently from sober speakers. For example, energetic intoxicated speakers may systematically more emphasize than sober speakers, whereas depressed intoxicated speakers might use less emphasis. Phrasing is thought to be influenced by sentence planning [2, 3], if a speakers ability to plan future constituents is impaired by alcohol, they may include more disfluencies and intonational phrase boundaries.

We use the AuToBI toolkit [4] to identify prosodic events on the IS11 Speaker State Challenge (IS11-SSC) material automatically. AuToBI is an open-source toolkit that automatically predicts ToBI (Tones and Break Indices) [5] annotations aligned to a word-segmentation. AuToBI first detects pitch accents and phrase boundaries, and then classifies these based on the inventory described in the ToBI standard. The ToBI standard encodes three types of prosodic events, pitch accents, intermediate and intonational phrase boundaries. Each of these are classified into categorical types based on pitch contours coincident with their realization. These types are defined based on high (H) and low (L) tones which identify pitch accents, phrase accents, and boundary tones in Standard American English (SAE) speech. The AuToBI models are trained on SAE speech from the Boston Directions Corpus material [6]. AuToBI uses word boundaries to determine regions of analysis, and align predictions, but does not use the lexical identity of the words to make predictions. While there are some similarities between SAE and German intonation, AuToBI is likely to generate noisy hypotheses the IS11-SSC material due to differences between SAE and German intonation; however, AuToBIs hypothesized tones may still contain some discriminative information with respect to intoxicated vs. sober speech.

To represent an utterances prosodic contour, we use a feature representation capturing the n-gram frequencies of the prosodic event sequence without explicitly constructing a Markov chain model. For each value of n, we calculate the rate of occurrence of each n-gram in the observation sequence. To approximate the function of a backoff language model we include these n-gram features at $n = \{1, 2, 3\}$. We construct these features using the full set of ToBI tones and collapsing some similar tones. We also explore an n-gram representation in which deaccented words are represented. In addition, we include features such as the relative frequency of pitch accent, phrase accent and boundary tone types, and the overall accenting and phrasing rates and the number of tones in the sequence. Using these features, we train a logistic regression classifier with L_1-regularization.

We first perform 10-fold cross validation on the official training material and observe 69.8% accuracy. Evaluating on the official development set, the prosodic modeling fails to significantly outperform the majority class baseline, with 69.6% accuracy and an f-measure of 0.032. Note that the skewed distribution toward sober speakers could have a profound impact on the classifier performance. Although cross-validation does not explain the task difficulty, since the same speakers can be in both the test and train folds, reporting this accuracy throughout this paper can still be valuable, especially in cases where there are (training) samples from a speaker to be tested.

Evaluating the model trained on the balanced training set on the balanced development data, the accuracy remains at baseline, 53.3%, yet the f-measure is 0.457. This corresponds to 0.53 precision and 0.40 recall, suggesting that there is some discriminative information in the prosodic signal. There are a few possible explanations for our poor performance on this material. First, the AuToBI hypotheses are generated using models trained on English speech; the errors from applying these models to German material may be too great to yield a useful representation of prosody. Second, prosodic analysis is most effective on longer utterances, the short productions that make up much of the material may also limit the efficacy of this approach. Finally, while it is likely that a single speaker's prosody is modified when intoxicated, the differences across speakers may not be consistent enough to be detected using this approach.

4. Phone Duration Variation

We hypothesize that intoxication may lead to changes in certain phone durations. We make use of the phones and temporal alignment provided for the training and development data to extract phone duration statistics for each phone type in each utterance. For each utterance, we extract the following features for each phone type: minimum, maximum, mean and standard deviation of durations of all phone instances of this phone type in the utterance. We also include global phone duration statistics at the utterance level. Specifically, we extract additional four duration features: minimum, maximum, mean and standard deviation of the durations of all phone instances from *all* types.

Using these features in a logistic regression classifier, we obtain an accuracy of 69.6% with 10-fold cross validation using the official training data. Testing on the official development set, we obtain an accuracy of 70.5%. It is interesting that, with such relatively simple features, we obtain an accuracy higher than the 69.2% obtained by the baseline system. Although our accuracy is not higher than the majority class, when we look at the confusion matrix, we see that our classifier does not always choose the majority class. Training and testing this classifier on our balanced sets, we obtain an accuracy of 62.5%, which is significantly better than the majority class baseline (52%). From these results it appears that phone duration statistics to be valuable in distinguishing intoxicated vs. sober speakers.

5. Phonotactic Variation

Phonotactic modeling has been quite successful for language and dialect identification [7]. Here, we hypothesize that intoxication may cause speakers to pronounce words differently, choosing certain pronunciation variants more frequently than others, and may even choose certain words more frequently, affecting the phonotactic patterns in each class. We take a vector-space based phonotactic modeling approach. We first collect the set of all triphones in the training data.[1] We construct a feature vector for each utterance, where each element in this vector corresponds to a single triphone in our set. The value of this element is the frequency of this triphone in this utterance. To compensate for utterance duration differences, we normalize this vector by its Euclidian norm.

We use these feature vectors to train an SVM classifier with linear kernel. The 10-fold cross-validation on the official training data is 70.1%. Training and testing on the official training and development data, respectively, we obtain an accuracy of 71.1%, which is significantly higher than the official baseline system (69.2%). Also, this accuracy is higher than the majority

[1] We add "start" and "end" symbols to the borders of each utterance.

class baseline (70.5%), although the difference is not significant. If we train an SVM classifier using this approach on our balanced training data and test it on the balanced test set, we obtain an accuracy of 71.1%, which is significantly higher than the majority class baseline (52%). These results suggest that the phonotactic distributions across the two classes are significantly different.

Since the task of this challenge is detection, we believe it is useful to report the Detection Error Tradeoff (DET) curve, which plots false alarm vs. miss probabilities (of missing intoxicated speakers), as is standard in speaker verification. The DET curve allows us to determine the detection threshold of interest. The DET curve has also an advantage over accuracy here due to the skewness of the official development set.[2] As shown in Figure 1, the Equal Error Rate (EER) of the phonotactic approach on the official development set is 33.5%, significantly better than chance.[3] We obtain a slightly better EER when employing our balanced set, of 30.8%, also significantly better than chance (see Figure 2).

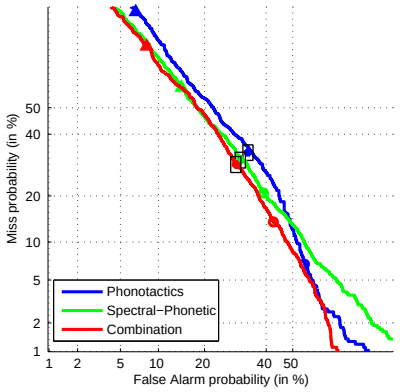

Figure 1: DET curve for the official development set (training on the official training set)

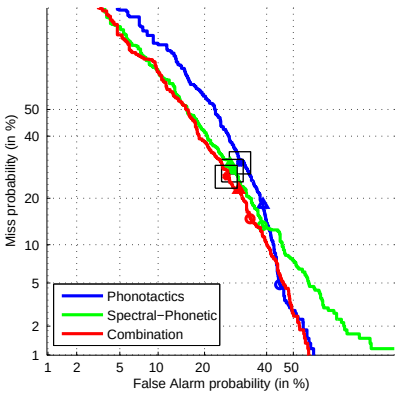

Figure 2: DET curve for our balanced set (training on our balanced training set)

6. Spectral-Phonetic Variation

In this section, we rely on the hypothesis that intoxicated speakers realize certain phones differently than sober speakers. To model phonetic structural differences across these classes (sober and intoxicated), we adopt our recent and successful approach to dialect and accent recognition [8], treating intoxicated speech in a given language simply as a different accent of this language. Again, we make use of the phones and alignments provided for this task, although one could also use a high-quality phone recognizer to obtain such information.

6.1. Phone GMM-UBM

The first step in our approach is to build an acoustic model for each German phone type. In particular, we first extract acoustic features temporarily aligned to each phone instance in the training data from both classes. The acoustic features we use here are 13 RASTA-PLP features (including energy) plus delta and delta-delta, resulting in a 39D feature vector from each frame. Afterwards, using the frames aligned to the same phone type (in all training utterances), we train a Gaussian Mixture Model (GMM), with 60 Gaussian components with diagonal covariance matrices, for this phone type, employing the EM algorithm. We have observed that some phone types occur infrequently in the training data; therefore, we build only a single GMM for each of the most frequent 45 phone types. Each phone GMM can be viewed as a GMM-Universal Background Model (GMM-UBM) for that phone type, since it models the general realization of that phone in both classes [9]. We term these GMMs as phone GMM-UBMs.

6.2. Phonetic Representation

For our approach, we need a representation that captures the acoustic-phonetic features for each phone type in a given utterance (U). We adopt the GMM representation [10] but at the level of phone types rather than the entire utterance. Specifically, we first obtain the acoustic frames aligned to every phone instance of the same phone type in U. We then use these frames to MAP (Maximum A-Posteriori) adapt the corresponding phone GMM-UBM. We adapt only the means of the Gaussians using a relevance factor of $r = 0.1$. We denote the resulting GMM of phone type ϕ as the *adapted phone-GMM* (f_ϕ). The intuition here is that f_ϕ 'summarize' the variable number of acoustic frames of all the phone instances of a phone-type ϕ in a new distribution specific to ϕ in U.

6.2.1. A Phone-Type-Based SVM Kernel

We represent each utterance U as a set S_U of adapted phone-GMMs, each of which corresponding to one phone type. Therefore, the size of S_U is at most the size of the phone inventory ($|\Phi|$). Let $S_{U_a} = \{f_\phi\}_{\phi\in\Phi}$ and $S_{U_b} = \{g_\phi\}_{\phi\in\Phi}$ be the adapted phone-GMM sets of utterances U_a and U_b, respectively. Now we design a kernel function to compute the 'similarity' between pairs of utterances given their adapted phone-GMM sets. In this work, we compare the Kullback–Leibler (KL) divergence between the two adapted phone-GMMs, following [10, 11]. Unfortunately, the KL-divergence is not symmetric and does not satisfy the Mercer condition, and thus does not meet the requirements for use as the kernel function for an SVM. However, Campbell et al. [10] have proposed a kernel function between GMMs based on an upper bound for their KL-divergence proposed by Do [12]. This function assumes that only the means of the GMMs are adapted, which is true in our case.

Using this KL-divergence-based kernel between two adapted phone-GMMs modeling phone ϕ, we obtain the kernel function:

$$K_\phi(f_\phi, g_\phi) = \sum_i (\sqrt{\omega_{\phi,i}}\Sigma_{\phi,i}^{-\frac{1}{2}}\mu_i^f)^T(\sqrt{\omega_{\phi,i}}\Sigma_{\phi,i}^{-\frac{1}{2}}\mu_i^g) \quad (1)$$

[2] Now chance is the line that goes through (50,50) with a slope of -1.

[3] We use the NIST scoring software developed for LRE07: www.itl.nist.gov/iad/mig/tests/lre/2007

where $\omega_{\phi,i}$ and $\Sigma_{\phi,i}$ respectively are the weight and diagonal covariance matrix of Gaussian i of the phone GMM-UBM of phone-type ϕ; μ_i^f and μ_i^g are the mean vectors of Gaussian i of the *adapted* phone-GMMs f_ϕ and g_ϕ, respectively. We define our kernel function between a pair of utterances:

$$K(S_{U_a}, S_{U_b}) = \sum_{\phi \in \Phi} K_\phi(f'_\phi, g'_\phi) \quad (2)$$

where f'_ϕ is the same as f_ϕ but we subtract from its Gaussian mean vectors the corresponding Gaussian mean vectors of the phone GMM-UBM (of phone type ϕ). g'_ϕ is obtained similarly from g_ϕ. The subtraction allows zero contributions from Gaussians that are not affected by the MAP adaptation.[4]

It is interesting to note that, for (2), when K_ϕ is a linear kernel, such as the one in (1), we can represent each utterance S_{U_x} as a single vector. This vector, say W_x, is formed by stacking the mean vectors of the adapted phone-GMM (after scaling by $\sqrt{\omega_\phi}\Sigma_\phi^{-\frac{1}{2}}$ and subtracting the corresponding $\vec{\mu}_\phi$) in some (arbitrary) fixed order, and zero mean vectors for phone types not in U_x. This representation allows the kernel in (2) to be written as in (3). This vector representation can be viewed as the 'phonetic finger print' of the speaker. It should be noted that, in this vector, the phones constrain which Gaussians can be affected by the MAP adaptation (allowing comparison under linguistic constraints), whereas in the GMM-supervector approach [13], in theory, any Gaussian can be affected by any frame of any phone.

$$K(S_{U_a}, S_{U_b}) = W_a^T W_b \quad (3)$$

Since we found in Section 4 that phone durations are important features, we also include duration statistics for each phone type from U_x in this vector (W_x), including the mean and standard deviation of the log durations of the phone instances of the same type in the utterance. As a result, we include 90 (45x2) new duration features.

Now we test whether our method can capture phonetic differences between sober and intoxicated speakers. For our first experiment, we use the official training data from both classes to train our phone GMM-UBMs. Then, we construct a vector W_x for each utterance in the training data, as described above. Afterwards, employing our kernel function (3), we first compute a kernel matrix for both classes using these vectors. We then train a standard binary SVM classifier using this kernel matrix. Our accuracy on 10-fold cross validation, using all the official training data, is 75.8%. This is significantly better than the majority class which is 69.4% and all our approaches above. Testing our approach on the development set, we obtain a significant improvement in accuracy (72.8%) over both the majority class accuracy (70.5%) and the baseline system's accuracy (69.2%) and better than every other approach above. As shown in Figure 1, the EER of our system using this approach on the official development set is 30.9%, slightly better than the phonotactic system.

To test our system on our balanced data, we train our phone GMM-UBMs, employing our balanced set of training data. We then train an SVM classifier as described above. Evaluating this classifier on our balanced development set, we obtain an accuracy of 71.2%, which is significantly better than majority class (52%). We report the DET curve on our balanced development set in Figure 2; the EER is 28.2%.

We are also interested in testing whether phonotactics and phonetic systems can contribute to the classification task when combined. To plot the combination DET curves, we simply sum the posteriors from the two classifiers. As shown in Figures 1 and 2, we observe that, in fact, the combination of these two approaches improve the EER over using any approach alone for both sets (the official and balanced). We obtain an EER of 29.4% using the official sets, and 26.3% on the balanced sets.

[4] We have observed that this subtraction slightly improves accuracy in our dialect recognition work [8].

7. Conclusions and Future Work

We have conducted a series of experiments designed to automatically detect intoxicated speakers given samples of their speech, as part of 2011 Interspeech Speaker State Challenge: Intoxication Sub-challenge. We have examined classifiers based on prosodic events, phone durations, phonotactic patterns, and spectral phonetic features. We have found that modeling automatically obtained prosodic events does not seem to be effective for this task. However, phone durations do provide an important contribution to the classification task. Moreover, phonotactic features seem to be even more informative features (EER=30.8% on the balanced set). Finally, our novel approach, which relies on the hypothesis that certain phones are realized differently across intoxicated and sober speakers, provides us with better accuracy on this task (EER=28.2%). Yet, the combination of the phonotactic and phonetic approaches gives the best results (EER=26.3%). Both our phonetic-based kernel system (accuracy of 72.8%) and the combination system (73.0%) achieve significant improvements over the majority class and the baseline system (p-value$<$0.001), when trained on the official training data and tested on the official development set as well as on the balanced sets.

We have observed in our dialect and accent recognition work that adding more training data from more speakers substantially improves results. We plan in future work to test this approach when more training data is available. We also plan to incorporate prosodic and phonotactic features directly in our kernel function.

8. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," in *Interspeech*, 2011.

[2] J. Krivokapic, "Speech planning and prosodic phrase length," in *Speech Prosody*, 2010.

[3] M. Breen, "Intonational phrasing is constrained by meaning, not balance," *Language and Cognitive Processes*, 2011.

[4] A. Rosenberg, "Autobi – a tool for automatic tobi annotation," in *Interspeech*, 2010.

[5] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Proc. of the 1992 ICSLP*, 1992.

[6] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: Studies on speech corpora," in *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[7] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processig*, vol. 4, no. 1, 1996.

[8] F. Biadsy, J. Hirschberg, and D. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *Interspeech*, 2011.

[9] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, 2000.

[10] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP variability compensation," in *Proceedings of ICASSP'06*, France, May 2006.

[11] P. Moreno, P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *Advances in Neural Information Processing Systems 16, MIT Press, Cambridge*, 2004.

[12] M. Do, "Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models," *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 115–118, 2003.

[13] W. Campbell, D. Sturim, and D. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.