Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors

Daniel Bone[1], *Matthew P. Black*[1], *Ming Li*[1], *Angeliki Metallinou*[1], *Sungbok Lee*[1,2], *Shrikanth S. Narayanan*[1,2]

[1] Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA
[2]Department of Linguistics, University of Southern California, Los Angeles, CA, USA

{dbone@,matthepb@,mingli@,metallin@,sungbokl@,shri@sipi.}usc.edu

Abstract

Speaker state recognition is a challenging problem due to speaker and context variability. Intoxication detection is an important area of paralinguistic speech research with potential real-world applications. In this work, we build upon a base set of various static acoustic features by proposing the combination of several different methods for this learning task. The methods include extracting hierarchical acoustic features, performing iterative speaker normalization, and using a set of GMM supervectors. We obtain an optimal unweighted recall for intoxication recognition using score-level fusion of these subsystems. Unweighted average recall performance is 70.54% on the test set, an improvement of 4.64% absolute (7.04% relative) over the baseline model accuracy of 65.9%.

Index Terms: intoxication detection, speaker state, hierarchical features, speaker normalization, GMM supervectors

1. Introduction

Paralinguistic study of speech often includes characterizing speaking styles, mental states of cognition and socio-emotions, and individual attributes such as age, and gender. Extracting this information automatically from a speech signal has various applications in commerce, education and healthcare. Features and methods developed in a particular paralinguistic domain can often be applied in other distinct but overlapping domains. This paper focuses on identifying an intoxicated speaker state from a single utterance. One potential application of intoxication detection is identifying impaired vehicle operators [1]. It is possible that such a speech system could be used as a stand-alone device or combined with chemical-based systems.

In this paper, we present an intoxicated speaker state recognition study using speech collected from 154 German speakers comprising the Alcohol Language Corpus (ALC) database [2]. The two speaker states of interest are intoxicated (indicated by a blood-alcohol content above 0.5mg/L) or sober.

The automatic classification of the intoxicated speaker state from this dataset is confounded by at least three factors. First, there are many speakers in the database. The speakers have diverse vocal tract and glottis physiologies, as well as differing ranges of vocal expressivity (e.g., activation level during communication). Second, the database consists of three styles of speech: read speech, spontaneous speech, and command and control. Vocal characteristics of each style differ greatly. For example, speech rate is expected to be much slower and less variable in read or command speech, as compared to spontaneous speech for a given speaker. Third, the utterance durations range from 0.5 to over 60 seconds, indicating drastically different amounts and types of information. Taken together, different speakers, speech styles, and utterance durations provide a complex dataset that makes feature generalization challenging.

The difficulty of identifying intoxicated speech may be easily understood from the results of a perceptual listening experiment conducted by Pisoni and Martin [3]. Raters achieved only 73.8% accuracy in identifying which of two lexically identical read sentences from a single speaker was recorded during intoxication and only 64.7% accuracy when given an arbitrary sentence. These findings underscore the inherent variability in intoxicated speech production and the need for an effective speaker normalization scheme.

Research conducted on the effect of alcohol on articulation concluded that articulation became more difficult as level of intoxication increased. Pisoni showed that intoxicated speakers had difficulty controlling abrupt opening and closing of the vocal tract [3]. The effects were more pronounced during complex speech production, such as vocalizations that involved coordination with laryngeal actions. We expect spectral features to capture inhibited articulatory ability. We also expect jitter and shimmer (peak to peak variations in pitch and energy, respectively) to quantify stuttering and/or "quivering" voice that is associated with intoxicated speech [3,4].

Although Mel-frequency cepstral coefficients (MFCCs) are generally found to outperform formant features across many speech-based classification tasks, formants were shown to be the top feature set in a related paralinguistic classification, sleepiness detection [5]. As a result, we also include the first three formants and their bandwidths here.

Reduced speech rate on the same read stimuli has been noted for intoxicated speakers [3,6]. Whether due to increased difficulty of articulation and/or slower cognitive processing, the average duration for speaking a sentence in these studies was found to increase with alcohol consumption. Motivated by this knowledge, we analyzed speech rate features obtained after forced alignment of phonemes following data transcription.

Our approach is motivated by several previous works [7-10]. Schuller et al. [7] and Black et al. [8] showed that hierarchical features, i.e., computing functionals-of-functionals across windowed regions of an utterance helped to smooth-out noise and provide improved results in classification with large datasets, as compared to only computing global functionals of feature streams. Busso et al. demonstrated a technique for iterative speaker normalization to minimize inter-speaker differences while still preserving emotional discrimination [9]. The method estimates the neutral emotion samples for each speaker and normalizes based on the hypothesized neutral-class statistics, rather than the conventional global speaker normalization approach. They demonstrated accuracy only 2.5% lower than the optimum (oracle) and 9.7% higher than without normalization. Li et al. incorporated a fusion method based on systems using a set of Gaussian Mixture Model (GMM) supervectors at the acoustic level for automatic speaker age and gender recognition [10,11]. In this work, we also propose the GMM latent factor analysis (LFA) [12] based Eigenchannel factors as a new kind of GMM supervector for intoxication detection. We present a classification study based on these three previously unexplored methods in intoxicated speech detection.

In section 2, methodology and approach are explained. Experimental results and discussion are presented in section 3, and conclusions and future work are discussed in section 4.

2. Methodology and Approach

In Section 2.1 we detail data pre-processing. In section 2.2 we discuss the competition scoring metric. Sections 2.3-2.6 detail openSMILE features, Praat features, hierarchical features, and rate features respectively. In Section 2.7 we describe iterative feature normalization. We present the GMM supervector systems in Section 2.8 and the classifiers used in Section 2.9.

2.1. Pre-processing

Before feature extraction, all pauses and noise marked in the transcription files were removed in an effort to extract features only during speech segments from the speaker of interest.

2.2. Competition Scoring Metric

Our train and development (devel) data subsets are biased roughly 70%/30% towards sober utterances. Weighted average recall would be a misleading statistic since a simple 'Zero-Rule' classifier could beat the baseline unweighted accuracy (65.3%). Unweighted average recall is a measure that does not weight the class accuracies by the number of samples from each class, but gives classes equal weights. It is a metric that attempts to simultaneously maximize the performance in each class, and therefore it is used as the performance metric for this study.

2.3. Base Features

The openSMILE base feature set contains a number of common acoustic low-level descriptors (LLDs) [13]. It includes MFCCs, log magnitude of Mel-frequency bands (MFBs), fundamental frequency (f_0), energy, jitter, and shimmer, among other features. The final base feature set is produced by computing 'global' static functionals (e.g., mean, standard deviation) across each of these LLD streams. We refer to these as 'global' features, since the functionals are computed across the entire utterance in this case. Further description of the chosen base feature set may be found in [2].

2.4. Praat Features

While many LLDs are extracted in the base feature set, we felt complementary acoustic information could be extracted using Praat [14]. We extracted eight feature contours with Praat using a 25ms window and 10ms period. f_0 was extracted using the autocorrelation function method, with a minimum and maximum f_0 of 75Hz and 500Hz, respectively. We normalized the pitch on a logarithmic scale, $\log_2(f_0 / f_{0\mu})$, where $f_{0\mu}$ is the mean pitch for the utterance. Energy was normalized as E / E_μ, where E_μ is the mean energy for the utterance. Formants 1-3 and their bandwidths were also estimated using Praat, motivated by the findings in [5].

2.5. Hierarchical Features

Counting both the base and Praat feature subsets, there are 130 LLDs (Table 1). Utterance durations for this corpus range from 0.5 seconds to over 60 seconds, so computing functionals across the entire utterance may result in features that are not comparable for widely varying utterance durations.

The motivation behind extracting hierarchical features is two-fold: 1) because of the windowing technique used, we hope the hierarchical features will be more comparable for varying utterance durations, and 2) we hope these features better capture moment-to-moment changes in an utterance, compared to the global features.

We calculated the hierarchical features by first windowing each LLD at two temporal granularities: 0.1s and 0.5s. Then, the 15 functionals shown in Table 1 are extracted for each windowed segment of the LLD; this will produce a contour for each LLD and functional combination. We generate the final hierarchical features by computing the 'core' functionals (Table 1) across each of these resulting contours; we only use the 'core' functionals to prevent an even larger feature set from this combinatoric framework.

Table 1. *A list of acoustic low-level descriptors (LLDs) and static functionals; the six 'core' functionals are starred (*).*

LLD	120 OpenSmile, 10 Praat
Functional	Mean*, median*, standard deviation*, 0.01/0.99 quantiles*, 0.01/0.99 quantile range*, skewness, kurtosis, min/max positions, upper/lower quartiles, interquartile range, linear approximation slope coeff., linear approximation MSE

2.6. Speech Rate Features

A total of 103 speech rate features are computed globally per utterance. Phoneme durations are not computed per utterance for specific phonemes since this could lead to over-fitting.

The phoneme sequence and the phoneme durations are extracted after forced alignment with a manual transcription (included as part of the corpus). Each phoneme duration in an utterance is z-normalized based on phoneme durations of all identical phonemes in the training data. This provides a contour of normalized phoneme durations on which functionals can be computed. Two types of normalization are used, z-normalization and quartile-normalization (subtracting the median and dividing by the inter-quartile range). Features are normalized by both the sober statistics and the intoxicated statistics. Some functionals are also computed on the delta contour of normalized phoneme durations for an utterance. Only phonemes with at least 1000 instances and only phonemes that occur in both the intoxicated and the sober training data are considered; we ignore consequences to functionals computed on the delta contour. Consonant-vowel duration ratios and intra-pause duration to voiced speech durations were also chosen as features.

2.7. Iterative Speaker Normalization

Iterative speaker normalization is a technique in which features are normalized repeatedly in an effort to normalize by class statistics even when the classes are unknown, such as in the case of unlabeled 'test' data. Once an initial estimate of the class labels is obtained, classification is performed. Features are re-normalized by the new class labels and repetition of this process continues until convergence.

The motivation for such a method is demonstrated in so-called 'oracle' experiments using the devel set. We find that if we know the class labels, we are able to achieve much greater performance from normalizing each speaker by the sober class statistics than by normalizing each speaker with global statistics for all utterances by that speaker.

In our work we initialize the class labels to the result of global z-normalization. Iterative normalization can be sensitive to the classifier parameter settings and initial conditions (class estimates), and convergence to a higher performance is not guaranteed.

When the class label distribution is changed drastically, it is intuitive that global z-normalization may fail. Alternatively, a potential strength (or weakness) of the iterative method is that its performance depends on the ability to seek out the true class labels. By repeatedly finding new estimates of class labels, iterative speaker normalization may still succeed when class label distribution changes cause global speaker z-normalized classifiers to fail. Four iterations were seen to be enough to provide convergence for the iterative speaker normalization method, based on empirical analysis.

2.8. GMM supervectors system

A 512 component GMM was trained upon the 39 dimensional MFCC features from the training dataset. Then, MAP adaptation and Universal Background Model (UBM) scoring were performed for every utterance in the entire dataset. The GMM mean supervectors were generated by concatenating the mean vectors of all components from the adapted GMM. The Tandem posterior probability (TPP) supervector [10,11] is the maximum likelihood (ML) estimate of the mean of a multinomial distribution. Linear kernel and Bhattacharyya probability product kernel were used for the SVM modeling for mean and TPP supervectors, respectively. Further information can be found in [10,11]. Furthermore, in the GMM factor analysis framework for speaker verification [12], we can consider the intoxicated speech as normal speech being corrupted by channel variability. Let us denote $M_{s,c}$ as the speaker and channel dependent mean supervector. Then $M_{s,c}$ can be decomposed into speaker dependent mean supervector plus the channel variability Ux, where U is the low rank Eigenchannel matrix learned from the pooled within speaker covariance matrix.

$$M_{s,c} = M_s + Ux.$$

U is a factor loading matrix and the components of x are channel factors [12]. Rather than reducing the intoxicated variability for speaker verification, we directly adopt the low dimensional Eigenchannel factors x as our Eigenchannel factor supervector for SVM modeling. In this LFA approach, the GMM size and Eigenchannel matrix rank are 256 and 4, respectively. The Eigenchannel matrix was trained on both the train set and the devel set.

2.9. Classifier Type

A support vector machine (SVM) with linear kernel, L2 regularization, and L2 loss is used. Although SMOTE was used on the baseline model, we instead choose to exploit knowledge of class bias to adjust the decision threshold of the SVM model.

SVM training was performed using LIBLINEAR. The cost parameter, *C*, prevents over-fitting. We can obtain a more balanced recall between the two classes by setting the class weight parameters and *C* appropriately; these parameters were optimized on the devel set.

3. Experimental Results and Discussion

3.1. Non-GMM Features

First, we divided our non-GMM features into 4 sets and ran a grid search to optimize performance of each. The four feature sets are the original 4368 base features, global functional and hierarchical features of the Praat LLDs, all hierarchical features, and rate features. It became apparent that speaker z-normalization performed better on the openSMILE base set than speaker mean normalization, and we used only this normalized set for our classifiers in order to reduce the final set of features.

From error analysis, we noticed empirically that the iterative speaker-normalization method seemed to be optimizing the unweighted accuracy by performing better at classifying instances of sobriety, whereas the global speaker-normalization method, when optimized, performed better at classifying instances of intoxication. We implemented a basic score-level fusion classifier that tries to exploit the possible complementarity of these methods by summing up the class-weighted confidence of each instance ("naive fusion").

The results are presented in table 2. The table contains columns for the non-speaker-normalized features (None), speaker z-normalization without regard to class labels (Global), iterative sober-class z-normalization (Iter.), naive fusion of global and iterative classifiers (Naive Fusion), and oracle sober-class z-normalization (Oracle).

Table 2. *Unweighted classification results on devel set. The baseline is 65.3% [2].*

	Speaker Normalization				
Set	None	Global	Iter.	Naive Fusion	Oracle
All Feats	0.6287	0.7104	0.7176	**0.7258**	0.7831
Only: Original	*0.6439*	0.7079	0.6838	0.7033	0.7471
Praat	0.5950	0.6617	0.6323	0.6414	0.7114
Hierarchical	0.6225	0.7083	0.7052	0.7145	0.7770
Rate	0.5732	0.5791	0.5292	0.5429	0.5764

Speaker normalization by sober class z-statistics with oracle knowledge achieves the best performance in all cases, showing the potential power of this speaker normalization scheme. The best performance on the development set without oracle knowledge is from the naive fusion method when using all features.

The hierarchical feature set has the highest unweighted average recall of the individual feature sets. It is followed closely in the four speaker normalization columns by the original base feature set. The Praat set has much lower classification accuracy. This is likely because the Praat feature set is constructed from functionals and hierarchical features of 10 prosody and formant-related LLDs, whereas more informative features may be contained in utterance-level functionals of the 120 LLDs comprising the original base set.

Rate features appear to not generalize well. This is understandable given the diversity of the utterance type and duration. Preliminary tests show potential improvement for the speaking rate features if three separate models are trained based on utterance length, which roughly divides speech styles. This method is one effort to concurrently address two of the issues corrupting speaking rate feature performance, speech style and length of utterance. This is an area of future research.

The top model from the train/devel classifiers was chosen to classify the test set. The best model was "naive fusion" using all systems. While classification on the devel set shows

72.58% accuracy (7.28% absolute gain), test set classification provided only 66.85% accuracy (0.95% absolute gain).

The reduced performance on the test set may be due to a number of factors. We suspected that the major corruption in our model came from the difference in class distribution between the train and devel sets and the effect it has on the absolute feature values resulting from speaker normalization.

In order to better match the test set class distribution per speaker that reaches intoxication (there are 1620 and 1380 sober and intoxicated utterances respectively [2]), we duplicated the training instances marked as alcoholized for each speaker such that each speaker in the training set had 60 alcoholized instances and 60 non-alcoholized instances. All speakers have 30 alcoholized utterances, indicating that they drank alcohol, but were not necessarily intoxicated. We empirically found this to perform better than when using SMOTE. Next, speaker normalization was conducted the same as before. This time, we optimized the fusion weights between the confidence scores generated by the two speaker normalization methods. Our unweighted accuracy improved to 68.14% (a 2.24% absolute improvement).

3.2. Inclusion of GMM Features

Fusion results on the devel set are shown in Table 3. Optimal weights were found for certain feature set sub-groups, and further optimal weights were found for the combination of the fused sub-groups. Among the three GMM supervectors, the Eigenchannel factor supervector achieved the best performance of 70.39% on the devel set. In addition, by fusing with the GMM mean and TPP supervector systems, the accuracy is further improved to 71.43%. Finally, by fusing GMM supervector based systems with the hierarchical features based subsystems, the performance is enhanced to 76.96%. The results of the global speaker normalization method, the iterative method, and their fusion are slightly different than in Table 2 because of the revised, class-unbiased speaker normalization method described at the end of Section 3.1.

The optimal "weighted fusion" model was re-trained on the combined train and devel sets and used to classify the test set. The unweighted accuracy on the test set is 70.54%, a 4.64% absolute (7.04% relative) improvement over the baseline. Table 4 contains the confusion matrices from the devel and test classifications. On the development set, the same method achieved a 11.66% absolute improvement over the baseline. The drop in performance may be due to overfitting or various potential differences between the devel and test sets.

Table 3. *Unweighted classification results on devel and test sets. The baselines are 65.3% and 65.9% respectively [2].*

Set	UW Acc	Set Fusion	UW Acc
Global (1)	71.29	(1)+(2)	71.45
Iterative (2)	69.28	(3)+(4)	69.06
TPP (3)	63.93	(3)+(4)+(5)	71.43
LFA (4)	70.39	All (1-5)	76.96
GMM (5)	65.05	All [test set]	70.54

Table 4. *Confusion matrices of the weighted fusion system on the devel and test sets.*

Data	Ref. | Pred.	Sober	Intoxicated	Sum
devel	Sober	2130	630	2760
	Intoxicated	297	903	1200
test	Sober	1127	493	1620
	Intoxicated	393	987	1380

4. Conclusions and Future Work

Intoxication detection is an important, but challenging area of paralinguistic speech research with potential real-world applications. We tested the efficacy of several acoustic-based methods with the potential to provide benefits to this paralinguistic machine learning task. The methods include hierarchical acoustic features, iterative speaker normalization, GMM supervectors, an Eigencannel supervector, and score-level fusion. We obtained a balanced recall for intoxication recognition using score-level fusion of these subsystems. Unweighted average recall performance on the test set was 70.54%, an improvement of 4.64% absolute (7.04% relative) over the baseline model accuracy of 65.9%.

Since speaker normalization can be highly susceptible to changes in the class-label distributions between the test and development sets, the robustness of the two speaker normalization techniques to these distribution changes should be investigated as part of our future work.

Another area of future work should be automatic identification of speech style. We expect less feature variability within speech styles, leading to performance gains.

5. Acknowledgements

This work was supported in part by funds from NSF and NIH.

6. References

[1] F. Schiel, C. Heinrich, V. Neumeyer, "Rhythm and Formant Features for Automatic Alcohol Det.," in *Proc. Interspech*, 2010.

[2] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. Interspeech*, 2011.

[3] D. Pisoni, C. Martin, "Effects of Alcohol on the Acoustic-Phonetic Properties of Speech: Perceptual and Acoustic Analyses," *Alcoholism: Clin. and Exp. Research*, Jul./Aug. 1989.

[4] Y. Horii, P. Ramig, "Pause and utterance duration and fundamental frequency characteristics of repeated oral readings by stutterers and non-stutterers," *Journal of Fluency Disorders*, vol. 12, pp. 257-270, issue 4, Aug. 1987.

[5] J. Krajewski, A. Batliner, M. Golz, "Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach," *Behavior Research Methods*, 2009.

[6] L. Sobell, M. Sobell, "Effects of Alcohol on the Speech of Alcoholics," *Journal of Speech and Hearing Research*, vol. 15, pp. 861-868, December, 1972.

[7] B. Schuller, M. Wimmer, L. Mosenlechner, C. Kern, D. Arsic, G. Rigoll, "Brute-forcing Hierarchical Functional for Para-linguistics: A Waste of Feature Space?" in *Proc. ICASSP*, 2008.

[8] M. Black, P. Georgiou, A. Katsamanis, S. Narayanan, "'You made me do it:' Classification of Blame in Married Couples' Interactions by Fusing Automatically Derived Speech and Language Information," submitted to *Proc. Interspeech*, 2011.

[9] C. Busso, A. Metallinou, S. Narayanan, "Iterative Feature Normalization for Emotional Speech Detection," in *Proc. ICASSP*, 2011.

[10] M. Li, C.-S. Jung, K. Han, "Combining Five Acoustic Level Modeling Methods for Automatic Speaker Age and Gender Recognition", in *Proc. Interspeech*, 2010.

[11] M. Li, K. Han, S. Narayanan, "Automatic Speaker Age and Gender Rec. Using Acoustic and Prosodic Level Information Fusion," submitted to *Computer speech and language*, 2011.

[12] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007

[13] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR-Introducing the Munich open-source emotion and affect recognition toolkit," *Proc. IEEE ACII*, 2009.

[14] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341-345, 2001.