Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language xxx (2012) xxx–xxx

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors☆

Daniel Bone[a,*], Ming Li[a], Matthew P. Black[a], Shrikanth S. Narayanan[a,b]

[a] *Signal Analysis & Interpretation Laboratory (SAIL), University of Southern California, 3710 McClintock Ave., Los Angeles, CA 90089, USA*[1]

[b] *Department of Linguistics, University of Southern California (USC), 3620 McClintock Ave., Los Angeles, CA 90089, USA*

Received 13 March 2012; received in revised form 8 August 2012; accepted 18 September 2012

Abstract

Segmental and suprasegmental speech signal modulations offer information about paralinguistic content such as affect, age and gender, pathology, and speaker state. Speaker state encompasses medium-term, temporary physiological phenomena influenced by internal or external bio-chemical actions (e.g., sleepiness, alcohol intoxication). Perceptual and computational research indicates that detecting speaker state from speech is a challenging task. In this paper, we present a system constructed with multiple representations of prosodic and spectral features that provided the best result at the Intoxication Subchallenge of Interspeech 2011 on the Alcohol Language Corpus. We discuss the details of each classifier and show that fusion improves performance. We additionally address the question of how best to construct a speaker state detection system in terms of robust and practical marginalization of associated variability such as through modeling speakers, utterance type, gender, and utterance length. As is the case in human perception, speaker normalization provides significant improvements to our system. We show that a held-out set of baseline (sober) data can be used to achieve comparable gains to other speaker normalization techniques. Our fused frame-level statistic-functional systems, fused GMM systems, and final combined system achieve unweighted average recalls (UARs) of 69.7%, 65.1%, and 68.8%, respectively, on the test set. More consistent numbers compared to development set results occur with matched-prompt training, where the UARs are 70.4%, 66.2%, and 71.4%, respectively. The combined system improves over the Challenge baseline by 5.5% absolute (8.4% relative), also improving upon our previously best result.

Keywords: GMM supervectors; Speaker normalization; Hierarchical features; Intoxication detection; Speaker state; Cognitive and motor load

1. Introduction

Understanding and modeling variations in prosody and articulation are at the core of spoken language study. Increasingly, researchers are interested in computing the information carried in a person's speech beyond the linguistic content. Notably, the study of paralinguistic aspects of speech has recently included modeling and estimating

[1] The SAIL homepage is http://sail.usc.edu.

☆ This paper has been recommended for acceptance by Björn Schuller, Ph.D.

[*] Corresponding author. Tel.: +1 213 740 3477; fax: +1 213 740 4651.
E-mail address: dbone@usc.edu (D. Bone).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.09.004

pathological speaking styles, mental states of cognition and socio-emotions, and individual attributes such as age and gender. Such information can serve to provide a more complete description of human behavior and social interaction. With this inspiration, the field of behavioral signal processing aims to quantify human behavior states in a variety of application fields such as education and health (Black et al., 2011a,b; Lee et al., 2011).

Intoxicated speaker state recognition is a topic that provides important opportunities to learn about paralinguistic speech production and perception. A unique aspect of intoxication detection is that the data utilizes a (direct) chemical signal for the reference labels, as opposed to human annotation (of inferred labels) based on observed behavioral cues. Features and methods developed in a single paralinguistic domain can often transfer to, or inform, another. Furthermore, it is interesting to consider how intoxicated speech cues interact with cues from other paralinguistic domains. These are some of the motivations that led to the collection of the first large-scale corpus of intoxicated speech, the Alcohol Language Corpus, or ALC (Schiel et al., 2011). Another motivating application was to determine if the speech commands used in modern automobile technology can dually be used to detect possible intoxication, and prevent road accidents associated with impaired driving.

Alcohol affects cognitive and motor function, leading to perceptible changes in behavior, including communication. Compared to other drugs, alcohol has a medium-to-large negative effect on psycho-motor function (Hindmarch et al., 1991). Alcohol has lasting cognitive effects, impairing information processing even during decreasing blood-alcohol concentration (BAC) (Schweizer et al., 2004). Such impairment has been demonstrated in vision (Abroms and Fillmore, 2004), hand-writing (Galbraith, 1986), and many other motor tasks. Speech is the result of high-level sensory, cognitive, and motor processes (Hollien et al., 2009). Alcohol-induced sensory-motor impairment justifies the presence of traceable cues of alcohol intoxication in speech.

Listening experiments corroborate theorized changes in speech patterns due to alcohol intoxication. Hollien et al. (2009) demonstrated listeners are able to perceive increasing levels of intoxication within a speaker, but have difficulty ascertaining the specific level of BAC. It was not found that professionals (e.g., highway patrolmen) had acute discernment of intoxicated speech. However, there were differences in classes of speech, such that increasing text difficulty correlated positively with perceptibility. Pisoni and Martin (1989) conducted similar listening experiments and reached analogous results. Compared to 65% accuracy for an arbitrary utterance, they demonstrated a 9% absolute increase in binary perception of alcohol intoxication when classifying two utterances from the same speaker, supporting the need for speaker normalization. Schiel (2011) investigated aural coding accuracies on the Alcohol Language Corpus for a subset of read and spontaneous prompts, finding an average (human) decoding accuracy of 71.65%. It was observed that read sentences were more easily identifiable than spontaneous speech. This may have been because the cognitive-motor task was more complicated than the interview-style dialogue found in spontaneous speech. Schiel also reported that female speakers were more easily identifiable as intoxicated than male speakers. Pitch studies on the same corpus corroborate these conclusions since females were found to increase pitch more consistently than their male counterparts (Schiel and Heinrich, 2009).

Analyses of intoxicated speech have sought to understand the pairwise-correlations between intoxication and acoustic features, but also note an apparent speaker dependency in the expression of intoxicated speech. For instance, mean pitch is largely seen to rise with intoxication (Hollien et al., 2001; Baumeister and Schiel, 2010), but previous studies reported the opposite can also be true (Pisoni and Martin, 1989; Schiel and Heinrich, 2009). In Pisoni and Martin's experiment, one of four speakers exhibited a decrease in pitch when inebriated. Increased disfluencies in the form of increased pauses, mispronunciations, elongations, and interruptions have been noted (Barfusser and Schiel, 2010). Correspondingly, speech rate is consistently observed to decline with intoxication (Pisoni and Martin, 1989; Sobell and Sobell, 1972; Behne et al., 1991). Behne et al. (1991) reported reliable increases in durations of read sentences, but not in isolated monosyllabic words. This is possibly due to the low cognitive load required for producing well-practiced, independent syllables. Besides the durational and static aspects of prosody, dynamic aspects can also be affected, such as generating a perceived "quivering" voice quality (Pisoni and Martin, 1989).

Features that have been proposed in order to capture intoxicated speech statistics and dynamics in classification experiments include: rhythmicity features for speaking rate and vowel triangle area indicating articulatory variability (Schiel et al., 2010); jitter and shimmer for prosodic variability and articulatory features including Mel-frequency cepstral coefficients (MFCCs) (Schuller et al., 2008); and prosodic features including duration, isochrony, pairwise variability indices, and global interval proportions (Honig et al., 2011).

Classification results with these systems have demonstrated accuracies on par with human evaluations. Levit presented results on a smaller German database and a cutoff of 0.8 g/L, achieving 69% accuracy. As expected, classification accuracy was worst near the cutoff, and improved with increased intoxication or sobriety (Levit et al., 2001). In our previous work at the Interspeech 2011 Speaker State Intoxication Subchallenge (Schuller et al., 2011), our proposed approach yielded the top unweighted average recall of 70.5% on the Alcohol Language Corpus test set by fusing scores from various feature-representation models, utilizing hierarchical feature extraction, and applying two types of speaker normalization (Bone et al., 2011). Corresponding to speaker normalization benefits, differences at the phoneme level have been discovered in single-vowel experiments, where speaker-specific models for voice-excitation features achieved an accuracy of 70% (Sigmund and Zelinka, 2011).

The present work focuses on two questions. Firstly, we investigate speaker state classification with our general model, focusing on improving performance on the Intoxication Subchallenge test set. Our approach is detailed, and the benefits of individual components are thoroughly evaluated. Since previous studies have noted perceptual differences in prosodic expression between genders during intoxicated speech, we also investigate gender-dependent models.

Our complete fusion model is comprised of: (i) hierarchical contour functionals; (ii) alternative feature representation with Gaussian mixture model (GMM) mean supervectors, Universal Background Model (UBM) weight posterior probability (UWPP) supervectors, and latent factor analysis (LFA) supervectors; and (iii) global and iterative speaker normalization on the feature vectors generated by each of these representations.

Secondly, we focus our intoxicated speech detection analysis toward a system that can marginalize certain sources of variability, in turn gaining understanding about the perceptibility of intoxicated speech under different conditions. Experimental results are presented on the development dataset for prompt-type specific models, experiments using matched prompts that occur in both the sober and intoxicated condition, and various speaker normalization techniques. Prompt-type specific models may indicate that certain types of prompts are more or less difficult to model due to increasing cognitive load. Matched-prompt data reduces the tendency of a system to model lexical content. One of the proposed speaker normalizations that is suitable for live applications is referred to as *background normalization* in this paper. This type of normalization makes a practical assumption that each speaker is enrolled with sober intoxication data. Having such held-out enrollment data from unseen prompts leads to accuracies comparable to other speaker normalization techniques that may be dependent on the speaker's class-label distribution.

The rest of this article is organized as follows: Section 2 describes the ALC corpus and experimental setup; Section 3 details the feature extraction, modeling, and classification methods utilized; general speaker state experimental results are discussed in Section 4; Section 5 presents results of the final, fused model; error analysis is conducted in Section 6; and conclusions and ideas for future research are presented in Section 7.

2. Database and experimental setup

All experiments were conducted on a large, controlled-environment collection of alcoholized and non-alcoholized speech named the Alcohol Language Corpus. The following subsections detail the corpus, the corresponding matched-prompt subset, the Interspeech Speaker State Challenge and classification metric, and pre-processing.

2.1. Alcohol Language Corpus

The Alcohol Language Corpus (ALC) is the first publicly available speech corpus of intoxicated individuals, encompassing an impressive number of 162 female and male speakers of German (Schiel et al., 2011; Schuller et al., 2011). The corpus was collected with the intent of determining whether advances in automated speech processing and machine learning could capture the changes in a person's speech patterns due to intoxication.

The collection and the data elicitation were diverse in an effort to make results generalizable, valid, and complete. Speech was acquired across a broad range of speaker ages (21–75 years, mean = 31 years) at 5 different locations in Germany and balanced for gender (78 female, 84 male). Data were recorded in an automotive environment to approximate real-world application scenarios. Three types of speech styles were obtained: read, spontaneous, and command and control. The prompts included: read digit strings, tongue twisters, commands, addresses, and spelling; spontaneous monologues of picture description, question answering, and commands; and dialogues with a researcher.[2]

[2] Detailed description is provided by the corpus authors, Schiel et al. (2011).

Table 1
Groupings of ALC recordings and their respective numbers within the official and matched experiment designs.

Item type	Official		Matched		
	Sober	Intoxicated	Sober	Intoxicated	Remaining sober
Address	11	6	2	2	9
Command and control	19	9	8	8	11
Dialogue	5	2	1	1	4
Monologue	5	3	3	3	2
Digit string	10	5	5	5	5
Tongue twister	10	5	5	5	5
All	60	30	24	24	36

To begin, subjects drank a desired amount of alcohol. The subject BAC – measured by blood test – was between 0.28 and 1.75 g/L during the alcoholized speech samples. Next, the subject was asked to respond to 30 prompts in the alcoholized condition (not necessarily above the Challenge BAC limit). The ‘alcoholized’ speech sample was collected in less than 15 min. Within two weeks the speakers recorded 60 sober prompts in the same acoustic environment.

Audio recorded from a headset microphone is downsampled to 16 kHz for the Challenge. Gender is provided for the Challenge training and development sets, but not for the Challenge test set. All Challenge sets come with speaker labels, supporting speaker normalization.

2.2. *Matched prompt sets*

While the paper largely focuses on classification using the framework of the Intoxication Subchallenge, we also conducted additional experiments with a subset of the data to gain greater insights into speaker state system design. It came to our attention that a significant degradation in performance was noted when training on the train and development sets and testing on the test set, which had its subset of prompts. It seems reasonable that prompt-specific features were being captured by the classifiers implicitly. As further evidence, when classification was performed by training on the set of 36 sober prompts that did not appear identically in the intoxicated set (i.e., those read and spontaneous prompts that would have surely led to unique lexical content) and testing on the entire intoxicated set, unweighted accuracy reached 95%. In short, the classifier was performing prompt identification rather than intoxication detection. One remedy in real-world applications would be to have a large, random set of prompts for either condition; and the ALC allows us a step toward that by providing 24 identical prompts from either condition – designated our “matched set”. The matched set contains samples of all three speech styles. Even if a ‘matched’ prompt requires a spontaneous response, we will be modeling intricacies specific to intoxication. Item types from the non-matched and matched sets, written as they will be reported in Results, are shown in Table 1. It should be noted that the one ‘spelling’ prompt was grouped with the 10 read ‘address’ prompts for analysis. Another inherent benefit was the opportunity to train models with sober background data (from the 36 remaining sober prompts) as previously suggested in Schiel et al. (2010).

2.3. *Interspeech Speaker State Challenge and classification metric*

The Interspeech 2011 Speaker State Challenge consisted of two Subchallenges – Intoxication and Sleepiness (Schuller et al., 2011). The Intoxication Subchallenge used the Alcohol Language Corpus (ALC). The classification was posed as a two-way, high- and low-intoxication-level division set at Germany’s legal limit for ‘intoxicated’ driving, 0.5 g/L. This led to some speakers not meeting the cutoff for intoxication and having all utterances designated as sober. The train and development data encompasses all 90 utterances, 60 sober and 30 alcoholized, per speaker while the test set contained only 60 utterances per speaker – uncertain during the Challenge, 30 are sober and 30 are alcoholized per speaker. 154 speakers were randomly chosen in order to obtain a balanced gender set (77 female, 77 male). The speakers were further randomly divided into three groups: train (60 speakers), development (44 speakers), and test (50 speakers). The official classification metric was unweighted average recall – which is the average of per-class recall percentages.

2.4. Corpus pre-processing

We focus our analysis on the speech of the subject, ignoring all speech defects or pauses and all speech from the research conductor. Manual lexical transcriptions which include tagging of speech defects are supplied for the subject's speech. No such transcript is provided for the rare background speech from the research conductor. The database also provides associated phonetic alignments based on the manual transcriptions. We remove any frames aligned as silence, garbage, defects, or background noise from the original wave files prior to any feature extraction (this is a standard approach used in GMM supervector modeling in speech processing).

3. Methods

Our framework consists of five subsystems. Two are built on static- and hierarchical-functionals and are different only in the adopted speaker normalization method – global versus iterative. The remaining three are GMM supervector constructs that incorporate global speaker normalization. The static/hierarchical functional systems are described in the first subsection, followed by details of the GMM systems. Next, speaker normalization techniques are discussed, followed by description of a feature selection technique with which we have experimented. Lastly, score-level fusion is described. All classification is performed using LIBLINEAR (Fan et al., 2008; Chang and Lin, 2011), except for RBF-kernel SVM in association with feature selection which uses LIBSVM (Chang and Lin, 2011).

3.1. Static- and hierarchical-functionals

In this subsection, details are given of the static and hierarchical functionals computed on the baseline and Praat low-level descriptors (LLDs). In total, 150 LLDs and 33,618 functionals (features) are computed.

3.1.1. openSMILE baseline

The openSMILE baseline is constructed on numerous, commonly used acoustic LLDs (Eyben et al., 2010). The computed LLDs include relative spectral (RASTA) MFCCs, spectral energies, spectral roll-offs, spectral flux and entropy, zero-crossing rate, loudness, probability of voicing, fundamental frequency (f_0), energy, jitter, and shimmer, among others. In total, 120 LLDs (60 LLDs and corresponding first-order deltas) are extracted using openSMILE. The baseline system for the Interspeech Speaker State Intoxication Subchallenge is composed of 33 base, utterance-level static functionals such as mean, standard deviation, quartiles, mean value of peaks, linear prediction (LP) coefficients 1–5, and duration of signal above 95% quantile. An additional set of 6 functionals is extracted on the f_0 contour. In total, there are 4368 baseline features. Further description of the baseline features can be found in the Challenge Description Paper (Schuller et al., 2011).

3.1.2. Praat prosody and formants

We expected that a complementary set of acoustic LLDs could be extracted using Praat (Boersma, 2001). Eight feature contours were computed using a 25 ms window with 10 ms period: f_0, intensity, and the first three formants and their bandwidths. The formants and formant bandwidths were included based on their success in sleepiness detection (Krajewski et al., 2009). Pitch was extracted using the autocorrelation method with a minimum of 75 Hz and maximum of 500 Hz. Normalized versions of pitch (Eq. (1)) and intensity (Eq. (2)) were additionally computed per-speaker. First- and second-order deltas are calculated, totaling 30 Praat LLDs.

$$f_{0,norm}(t) = log\left(\frac{f_0(t)}{\overline{f_0}}\right) \tag{1}$$

$$int_{norm}(t) = \frac{int(t)}{\overline{int}} \tag{2}$$

3.1.3. Hierarchical functionals

Hierarchical feature extraction uses multi-level windowed statistics (functionals-of-functionals) to extract features. Hierarchical features have proven more effective than utterance-level functionals in emotion, sleepy speaker state, and couples' therapy machine learning tasks (Schuller et al., 2008; Krajewski et al., 2009; Black et al., 2011b). Although

Table 2
A list of the 15 static functionals computed at the frame-level on LLD contours; the six ‘core’, hierarchical functionals are starred (*).

LLDs	150 total – 120 openSmile, 30 Praat
Functionals	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

this combinatoric representation generates very large feature sets, linear support vector machine (SVM) has proven to provide high classification accuracy in all of these domains.

There are two rationales behind extracting hierarchical features – to find features that are robust to utterance duration and to better capture moment-to-moment changes in an utterance compared with utterance-level static functionals. The utterance durations in the ALC range from 0.5 s to over 60 s. Wider variations will occur for functionals computed on shorter-duration utterances, even if those utterances are being produced by the same generating distribution. We expect that using hierarchical feature extraction will produce features that are robust to utterance duration. Because we have shown in our previous work that hierarchical features increase accuracy when using global speaker normalization, we do not consider the baseline features alone (although we do investigate feature selection) (Bone et al., 2011). Dynamic variations may also be better captured because functionals are now computed within smaller windows throughout the utterance, supplying a finer granularity to utterance modeling.

First, primary-level windowing is performed on each LLD at two temporal granularities: 0.1 s and 0.5 s. Next, the 15 functionals shown in Table 2 are computed within each windowed segment of each LLD. This is a new contour of functionals, upon which we compute a second set of six ‘core’ functionals designated by an asterisk, ‘*’. A reduced set of core functionals is used to avoid an even larger feature set resulting from this multiplicative framework. Hierarchical extraction is depicted in Fig. 1, where $f_1(n)$ is one of 15 frame-level functionals and $f_2(n)$ is one of six core, hierarchical functionals. For each of the 150 LLDs, 15 utterance-level functionals are extracted and 90 (15×6) hierarchical functionals are calculated at the two temporal granularities. In total 29,250 functionals (150×195) are computed and added to the 4368 baseline features, completing the final 33,618 functional feature-vector. Some functionals may be co-linear or even identical, but linear SVM is expected to be robust in such cases. In the rare case that an utterance had a feature that could not be computed because the utterance was too short, that feature was set to 0. This implicitly incorporated utterance length into our model. Features which had very small variance compared to the corresponding mean absolute value were removed.

3.2. GMM supervectors

GMM methods enable alternative feature representations that provide various opportunities to account for orthogonal variances due to speaker and ‘channel’ effects representing the multiple sources of potential variability. Our GMM supervector systems consist of established, state of the art methods drawn from areas such as age and gender identification (Li et al., 2013), speaker and language recognition (Castaldo et al., 2007), and paralinguistic classification (Li et al., 2012). Three GMM-based subsystems are implemented in our experiments. It is generally shown that larger GMMs produce higher accuracies. We also would like to keep the mean supervector dimensionality low enough that

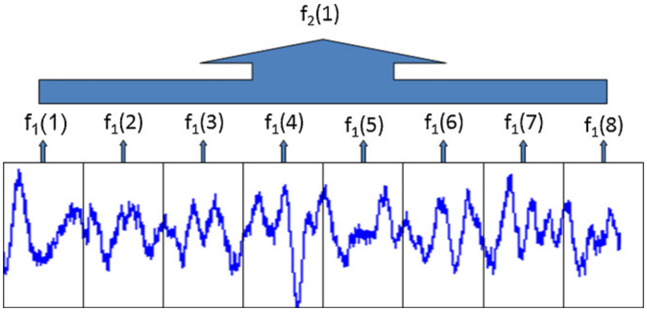

Fig. 1. Pictorial description of hierarchical feature computation.

computation is feasible. We choose a 512-dimensional GMM constructed with 39-dimensional MFCC feature vectors $(13+\Delta+\Delta\Delta)$ for the mean-supervector and the Universal Background Model (UBM) weight posterior probability (UWPP) supervector. We construct a 256-dimensional GMM for latent factor analysis, as was done in our previous work (Li et al., 2012). Standard GMM scoring was not included because we already are considering three related GMM systems, and have also previously found the GMM baseline system to have lower performance than the latent factor analysis (LFA) subsystem for intoxication detection.

3.2.1. *GMM mean-supervector*

The 512-dimensional Universal Background Model (UBM) is trained on all 39-dimensional MFCC feature vectors from 10 ms speech frames in the training set (all frames are speech frames due to pre-processing). Separate GMMs are trained for the official, complete prompt set of 5400 utterances and the reduced, matched-prompt set of 2880 utterances. Then, the means of the UBM are MAP-adapted for each training and evaluation utterance. Each supervector is constructed by concatenating the 39-dimensional means of all 512 Gaussians in the GMM, plus the bias (19,968 + 1 features). Lastly, each supervector feature is normalized by the corresponding standard deviation and GMM-component weight to fit a supervector kernel that is a bounded approximation of the Kullback–Leibler divergence (Campbell et al., 2006).

3.2.2. *GMM UWPP supervector*

For each utterance in the training and evaluation sets, unweighted posterior probability (UWPP) feature extraction is performed on the appropriate 512-dimensional UBM (official or matched set). The UWPP supervector consists of the average posterior probability associated with each GMM-component, and is therefore a 512-dimensional vector. The computation is conducted as shown in Eq. (3), where b_i is the ith UWPP-supervector-component, λ_i is the ith GMM-component, $\mathbf{o}_t$ is the tth window's MFCC vector, and T is the number of frames in the utterance.

$$b_i = \frac{1}{T}\sum_{t=1}^{T} p(\lambda_i|\mathbf{o}_t) = \frac{1}{T}\sum_{t=1}^{T}\gamma_i(t) \tag{3}$$

Lastly, the vector is modeled using the Bhattacharyya probability product kernel as has been previously proposed for age and gender identification (Li et al., 2013).

3.2.3. *GMM LFA supervector*

We have adopted the GMM latent factor analysis (LFA) framework that was recently implemented for speaker state detection (Li et al., 2012). LFA is often used to remove channel effects for speaker verification and diarization (Kenny et al., 2010), but we repose the problem to model the 'channel' factors which we anticipate to be resultant from alcoholized speech. If we let $\mathbf{M}_{s,c}$ denote the speaker- and channel-dependent mean supervector and $\mathbf{M}_s$ denote the 'clean' mean supervector of a speaker, then we can define the speaker and channel dependent mean supervector in terms of the 'clean' mean supervector and a low-rank Eigenchannel projection, **Ux**, which represents channel effects (Eq. (4)).

$$\mathrm{M}_{\mathrm{s,c}} = \mathrm{M}_{\mathrm{s}} + \mathbf{Ux} \tag{4}$$

The Eigenchannel matrix is computed with Principal Component Analysis (PCA) on the pooled within speaker covariance matrix. In order to train the Eigenchannel matrix with the 256-dimensional GMM-UBM, we need to use data from multiple speakers, with each speaker having data from each class. We therefore use only the 96 speakers in the training and development sets who reached the intoxication limit. We combine all utterances per speaker and per state into single vectors for adaptation based on previous success (Li et al., 2012). We then construct a within speaker variability matrix **S** by concatenating all of the mean-subtracted supervectors (vectors from MAP adaptation). The Eigenchannel matrix **U** is given by the R PCA eigenvectors of the within speaker covariance matrix, $(1/J)\mathbf{SS}^t$, with the largest eigenvalues (J is the number of supervectors).

The speaker state factor **x** of the LFA framework is calculated as in Eqs. (5) and (6) (Castaldo et al., 2007).

$$\mathbf{x} = (\mathbf{A}+\mathbf{E}^{-1})^{-1}\sum_{i=1}^{N}\mathbf{U}_i^{\mathbf{t}}\sum_{t=1}^{T}\gamma_i(t)\frac{\mathbf{o}_t-\boldsymbol{\mu}_i}{\Sigma_i} \tag{5}$$

$$\mathbf{A} = \sum_{i=1}^{N} \frac{\mathbf{U}_i^{\mathbf{t}}\mathbf{U}_i}{\Sigma_i} \sum_{t=1}^{T} \gamma_i(t) \qquad (6)$$

$\mathbf{U}_i$ is the ith Gaussian component's sub-matrix of $\mathbf{U}$, $\mathbf{o}_t$ is the tth feature vector, and $\gamma_i(t)$ is the occupancy probability of the ith Gaussian component for the tth feature. $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and the diagonal covariance of the UBM, respectively. The diagonal covariance matrix $\mathbf{E}$ is constructed from the R eigenvalues of the Eigenchannel matrix $\mathbf{U}$.[3] The Eigenchannel factor $\mathbf{x}$ is computed for all utterances. Linear SVM is then used for modeling on the channel factors.

3.3. *Speaker normalization*

Speech features are well-known to be corpus and person specific. Accommodation through model adaptation or speaker normalization is extremely important in automatic speech recognition, language identification, and emotion recognition. Speaker normalization is investigated in three variants – global, iterative, and background – each carrying assumptions about the data, but all showing marked improvements over using raw feature vectors. Global speaker normalization computes normalizing statistics on all data per speaker, regardless of class-label distribution. In contrast, iterative and background speaker normalization techniques normalize data by an approximation of the baseline-class (sober) statistics. Iterative speaker normalization seeks to automatically determine the baseline samples by a recursive-classification algorithm. Background speaker normalization utilizes a set of 'control' data that is often of distinct lexical content.

We choose Z-normalization over only mean subtraction or median and interquartile normalization because we initially found that it led to higher accuracies. In the per-prompt experiments, normalization is only performed using sentences of the same prompt type.

3.3.1. *Global normalization*

Global speaker normalization consists of computing the mean and standard deviation of each feature across all utterances from a speaker, indifferent to the class-label distribution. An inherent disadvantage is that the expected location for the optimal hyperplane should change if the class-label distribution changes. In our previous work, we attempted to fit the expected class-label distribution and saw a small gain (Bone et al., 2011). However, our current work investigates the robustness of our subsystems to the observed change in the class-label distribution on the test set – specifically, the training data that is distributed approximately 2-1 (sober-alcoholized), but the test data is distributed approximately 1-1. An additional benefit of global speaker normalization is the simplicity of computation.

3.3.2. *Iterative normalization*

Iterative speaker normalization recursively uses hypothesized sober labels for prediction, until convergence. The method is motivated by 'oracle' experiments in which the class-labels are actually known. If Z-statistics for speaker normalization are calculated on the true sober samples, the highest accuracies reported on the ALC are obtained (Bone et al., 2011). Iterative speaker normalization seeks a local optimum, and hence it will be dependent upon a proper initialization (Busso et al., 2011). To provide an appropriate seed-point, the hypothesized labels already obtained from global speaker normalization are used. Although the first iteration will likely provide a similar result to the global-normalization seed-point, the new hypothesis need not be identical since the hyperplane is being drawn in a space that has undergone different linear transformations between speakers. Besides the global cost parameter for SVM, the number of iterations can also be tuned. The iterative algorithm was run for three iterations based on quick convergence of the algorithm and observed development set performance.

3.3.3. *Background normalization*

Speaker normalization is a common approach for increasing performance in speech processing applications, but not all types of speaker normalization are practical in real-world applications. Take for example speaker state detection, where it is likely that we will not have representative data for a person in a particular speaker state. However, we likely can gain access to neutral/baseline speech from the speaker in similar recording conditions. This approach is similar to

[3] Further details can be obtained from Burget et al. (2007) and Castaldo et al. (2007).

iterative speaker normalization in that it approximates a speaker's neutral state statistics for a given utterance, except that instead of computing those sober sample-statistics through iterative classification on the target prompt-responses, we are estimating those features with sober utterances with comparable qualities.

Background speaker normalization requires a set of additional reference utterances for each speaker. Instead of performing another layer of cross-validation on the development set, we evaluate background normalization when a full held-out set of utterances is available as in the matched-prompt experiments. Out of the original 60 non-alcoholized prompts, 24 are repeated identically in the test set, leaving 36 to be used as background data on which to compute *Z*-statistics. For prompt-specific experiments, we perform speaker normalization regardless of the number of background utterances of that prompt-type available for each speaker (given in Table 1). Monologue prompts only have 2 utterances per speaker available as background, thus that specific result should be analyzed with care.

3.4. Feature selection

We have previously shown that the inclusion of hierarchical features increases unweighted average recall (UAR) over the baseline openSMILE functionals in the case of speaker-normalized features (Bone et al., 2011), but feature selection may lead to further improvements. Feature selection is a common machine learning method to improve modeling by reducing the 'noisy' features or the co-linearity between 'important' features. However, feature selection can be computationally expensive and may not provide gain in recall. This is especially true for very high-dimensional feature spaces where some standard techniques are computationally infeasible, and practical implementations may not succeed – e.g., Black et al. (2010) showed that using LDA with forward-feature selection on hierarchical features did not improve recall in dyadic interaction behavior modeling.

When using an SVM classifier, the SVM may also offer the most successful feature selection. However, we cannot utilize exhaustive methods and instead choose a recently proposed method in biological learning. This feature selection method based on recursion is named recursive SVM (R-SVM) (Zhang et al., 2006). The intent is to find the dimension(s), j, for which the difference between class-average distances from the hyperplane and that component's weight is properly maximized for each class (standardized features are assumed). We generally violate the algorithm's assumption (and initial inspiration) that the classes can be separated in the feature space; however, this assumption does not appear to affect the algorithm's motivation and derivation and we still adopt this technique in consideration of feature selection constraints for high-dimensional feature spaces and previous failure of standard techniques. The method is explained through Eqs. (7) and (8). In these equations, n_1 and n_2 are the number of data points in classes 1 (+) and 2 (−), $f(\mathbf{x})$ is the distance from the hyperplane, m_j^+ and m_j^- are the means of feature j for each class, d is the total number of features, and w_j is the jth component of the weight vector, $\mathbf{x}$. Our goal is to select features, j, corresponding to maximal values of s_j.

$$S = \frac{1}{n_1}\sum_{\mathbf{x}^+ \epsilon class1} f(\mathbf{x}^+) - \frac{1}{n_2}\sum_{\mathbf{x}^- \epsilon class2} f(\mathbf{x}^-) = \sum_{j=1}^{d} w_j m_j^+ - \sum_{j=1}^{d} w_j m_j^- = \sum_{j=1}^{d} w_j (m_j^+ - m_j^-) \tag{7}$$

$$s_j = w_j(m_j^+ - m_j^-) \tag{8}$$

To begin, the final feature-dimensionality and step size, which determines the number of 'new' features that are considered at each iteration, are chosen. We opted to select feature-dimensionalities of 4000 and 200 with the feature step size set to be half the final feature size. At each iteration, 6000 or 300 randomly chosen features are used for classification in the 4000 or 200 feature-set cases, respectively. The top N features are kept at each iteration. This process is iterated until all features have been considered. For each selected feature set, we again perform classification with linear SVM. For small feature sets, kernel-SVM using a radial basis function (RBF) kernel may provide an increased, modeling capability. Thus, we also utilize RBF-SVM for the case of 200 selected features.

3.5. Score-level fusion

Late-fusion is chosen to combine all five subsystems by their 'scores' – i.e., distances from the SVM decision-hyperplane. Each score is standardized by its median and inter-quartile ratio, calculated across all utterances. When deciding when to fuse each subsystem, the most-related and weakest subsystems should be fused first. We group the

Table 3
UAR for no speaker normalization (*Raw*) and *global* speaker normalization on the official and matched development sets. Results of feature selection for *All*, 4000, and 200 features are presented. Linear (LIN) and radial basis function (RBF) kernels are used.

	Raw functionals				Global speaker normalization			
	All LIN	4000LIN	200LIN	200RBF	All LIN	4000LIN	200LIN	200RBF
Official development set								
All	**0.641**	0.632	0.590	0.536	**0.719**	0.712	0.640	0.595
Address	0.737	0.736	0.693	0.585	0.761	0.749	0.719	0.681
Command and control	0.660	0.634	0.603	0.523	0.702	0.709	0.628	0.582
Dialogue	0.604	0.637	0.645	0.500	0.765	0.711	0.689	0.567
Monologue	0.638	0.640	0.574	0.506	0.768	0.766	0.675	0.656
Digit string	0.654	0.613	0.538	0.499	0.747	0.688	0.656	0.600
Tongue twister	0.711	0.713	0.654	0.514	0.783	0.763	0.731	0.664
All-divided	**0.676**	0.664	0.618	0.526	**0.745**	0.727	0.676	0.623
Matched development set								
All	**0.588**	0.591	0.575	0.533	**0.697**	0.674	0.617	0.599
Address	0.665	0.606	0.569	0.526	0.682	0.763	0.608	0.610
Command and control	0.564	0.564	0.555	0.505	0.679	0.676	0.615	0.635
Dialogue	0.590	0.546	0.581	0.500	0.752	0.775	0.638	0.627
Monologue	0.632	0.627	0.598	0.572	0.751	0.765	0.668	0.744
Digit string	0.583	0.566	0.534	0.537	0.701	0.668	0.600	0.611
Tongue twister	0.591	0.609	0.592	0.560	0.727	0.702	0.680	0.670
All-divided	**0.592**	0.584	0.566	0.533	**0.706**	0.702	0.632	0.649

Bold indicates points of focus for analysis.

two static-functional subsystems and all GMM subsystems for fusion because those groups are expected to be closely related. We fuse the GMM mean supervector with the UWPP supervector and then add the LFA supervector. The final score is then thresholded at zero and used for classification.

4. Speaker state experiments and discussion

This section has two major goals: (i) to experiment with speaker normalization methods, gender-dependent models, and feature selection in order to obtain a system that provides the best accuracy; and (ii) to analyze the results both per prompt-type and in relation to non-matched and matched sets, discussing generalizability and commenting on underlying cognitive-motor coordination in intoxicated speech. Such investigation is broadly informative to speaker state research.

4.1. Feature selection on functionals

In this subsection, classification performance is evaluated in terms of the following: feature selection, official or prompt-matched development sets, global speaker normalization, and prompt-type specific models. Speaker-independent cross-validation was performed for parameter selection. The results are displayed in Table 3.

A relatively consistent decline in performance is observed when anything less than the full feature set is modeled, whether using linear-kernel or RBF-kernel SVM. Feature selection is non-beneficial in all four combinations of development-set type and speaker normalization type. The matched development set with global speaker normalization may be the most important experiment to examine because it is unaffected by bias in modeling prompt type and provides higher performance than without speaker normalization. In this region, feature reduction to 4000 features while using linear SVM increases UAR for the address prompts, but does not notably improve classification for the remaining prompt-types or remaining feature dimension. It was determined through experiment repetition with random feature selection that the benefits were coming from increased modeling ability due to reduced feature size, not from feature selection. The results indicate that a subset of robust features may exist, although they may be prompt specific and are difficult to select. We therefore have empirical support to avoid feature selection on the test set.

Table 4
Unweighted average recall on the matched prompts and the non-matched prompts in reference to matched and non-matched prompt training. Global speaker normalization is used.

	Tested on:	
	Non-matched	Matched
Trained on:		
Official	**0.771**	0.658
Matched	N/A	**0.697**

Bold indicates points of focus for analysis.

Although large discrepancies in unweighted accuracy exist between the official and matched development sets, the matched-prompt results may be more robust. The matched-prompt experiments are motivated by the observation that lexical content that only occurs in a single class is being implicitly modeled (prompt identification, Section 2.2), affecting the potential of the system to model the true underlying speaker state. To further analyze this claim, results using global speaker normalization on both development sets are presented in Table 4. The matched model appears more successful in identifying cues of alcoholized speech since it has a higher accuracy on the matched-prompt evaluation set. Utterance identification is evident in the 11% absolute UAR increase between the matched (24 alcoholized/non-alcoholized prompts) and non-matched (36 alcoholized and 6 non-alcoholized prompts) test sets – together the 'matched' and 'non-matched' prompts are the official set. We did not evaluate matched-prompt model performance on non-matched data since the test would have been on unseen data (we do this later for the test set). It is clear that using matched prompts should produce the least biased view of intoxication detection and we may focus our discussion on this set.

The success of global speaker normalization is demonstrated in the right-half of Table 3. Regardless of other variables, using only the global speaker-normalized features (without raw features) is seen to always improve unweighted averaged recall compared to the raw features. For instance, UAR increases from 64.1% to 71.9% on the official development set and from 58.8% to 69.7% on the matched-prompt development set when using all features and combining all prompt types.

Lastly, we analyze the outcome of classification for different types of prompts and the potential gain from modeling each prompt type individually. Focusing on the lower-right quadrant of Table 3, we observe that modeling precision is consistently low for command-and-control utterances which contain both read and spontaneous responses. While these utterances tend to be shorter than certain other prompt-types (the median lengths for command & control and for tongue twister utterances are 3.2 s and 4.9 s, respectively, on the development set), they are also of lower cognitive requirement. They utilize familiar words and phrases, and in the case of spontaneous commands, potentially require less complex planning in sentence construction. Therefore, both the motor and cognitive loads on the participants are low. In comparison, we consider read tongue-twisters for which the motor load is prominently increased because the phrases and phonetic-pairings are unusual for natural speech. These findings are in accordance with perceptual studies (Hollien et al., 2009).

A drawback of prompt-specific modeling is reduced training data, but a potential gain may come from modeling intricacies of each speech type. In all four cases presented in Table 3, modeling each prompt individually increases the overall performance. However, we have not developed an appropriate automated prompt-identification system, and will not use such a method for the test set experiments. Likely, the delivered prompt would be known in real-world systems, and the system could presumably leverage this knowledge.

4.2. Speaker normalization methods

Three speaker normalization techniques are implemented in this work. The first is global speaker normalization, in which all samples from a speaker are used to normalize each feature. The intuitive assumption is that the class-distribution is relatively constant from speaker-to-speaker; we will examine the robustness of this assumption when classifying on the test set. Secondly, iterative speaker normalization attempts to automatically find the sober samples through recursion. Lastly, background speaker normalization assumes the system has access to samples of a speaker's baseline speech. The following subsections show the development set performance of these methods.

Table 5
Unweighted average recall for iterative and background speaker normalization when compared to no speaker normalization (Raw) and global speaker normalization on the official and matched development sets.

	Official development set			Matched development set			
	Raw	Global	Iterative	Raw	Global	Iterative	Background
All	0.641	**0.719**	0.684	0.588	**0.697**	0.676	**0.685**
Address	0.737	0.761	0.760	0.665	0.682	0.695	0.690
Command and control	0.660	0.702	0.677	0.564	0.679	0.663	0.671
Dialogue	0.604	0.765	0.768	0.590	0.752	0.729	*0.713*
Monologue	0.638	0.768	0.759	0.632	0.751	0.644	*0.512*
Digit string	0.654	0.747	0.729	0.583	0.701	0.648	0.654
Tongue twister	0.711	0.783	0.752	0.591	0.727	0.681	0.682
All-divided	0.671	**0.744**	0.724	0.592	**0.706**	0.667	**0.672**

Bold indicates points of focus for analysis.

4.2.1. Global versus iterative normalization

Results are displayed collectively and per prompt-type on the official and matched development sets in Table 5. Global speaker normalization is reliably superior to iterative speaker normalization. Iterative speaker normalization is shown to consistently improve upon no speaker normalization. We will consider potential benefits of iterative speaker normalization during fusion in the next section.

4.2.2. Background normalization

Background normalization has been suggested as a viable adaptation mechanism for real-world systems, given that neutral/baseline speech should be easier to obtain, whereas procuring training data for particular states of interest may be impractical (Schiel et al., 2010). With this in mind, we evaluate performance using background normalization on the matched development set, having tuned parameters on the matched training set with speaker-independent cross-validation (Table 5). We do not evaluate the performance on the test set, and we analyze these results only in order to provide reference numbers for future research on intoxicated speech detection and for the general relevance to speaker state research. This valuable technique should be largely independent of the test set distribution of class-labels.[4]

Speaker normalization using statistics from a speaker's background data provides higher unweighted average recall than non-normalized features, but lower accuracy than global speaker normalization. In fact, the UARs for background normalization are nearly identical to the UARs for iterative normalization. This is intuitive as the techniques approximate the same knowledge, a speaker's would-be neutral state characteristics for any given utterance. These experiments show the utility of background data.

4.3. Gender-dependent modeling

Gender-dependent models have demonstrated improved accuracy across a variety of speech based classification tasks. Pitch studies performed on the ALC by Schiel and Heinrich (2009) found that females increased their pitch more consistently than their male counterparts. But it is unknown whether these trends will apply across the vast feature set employed, or if the benefits of extra training data will outweigh the gains from modeling these gender-specific feature tendencies. The results of gender-dependent modeling on the official and matched development sets in the cases of no speaker normalization and global speaker normalization are shown in Table 6.

Gender-combined modeling outperforms gender-dependent modeling in all of the situations considered. As previously stated, it may be more informative to focus analysis on the matched-prompt set of experiments. For the raw feature case, there is similar performance whether the model is trained separately or jointly. Also, males are more accurately modeled than females – a surprising finding. This may be because males have lower variance in median pitch, and possibly other features, across speakers than females. In the speaker-normalized feature case, gender-combined accuracy is higher than both gender-dependent accuracies. Female accuracy is comparable to male accuracy, possibly

[4] Some evidence for this proposition is found through experiments in Section 6.

Table 6
UAR for gender-dependent models using all features for no speaker normalization (Raw) and for global speaker normalization on the official and matched development sets.

Data set	Raw				Global			
	Baseline	Female	Male	Average	Baseline	Female	Male	Average
Official development set	**0.641**	0.605	0.648	0.626	**0.719**	0.711	0.702	0.707
Matched development set	**0.588**	0.556	0.605	0.581	**0.697**	0.681	0.673	0.678

Bold indicates points of focus for analysis.

because the variance in female neutral state features is now comparable to the variance in male neutral state features due to speaker normalization. While it is still possible that other modeling techniques and features may make use of gender-specific information, we have shown that gender-dependent models are not necessary for our feature set with linear SVM modeling.

5. Final system construction: late fusion of functionals and GMM-SVM systems

In this section we present our final system constructed by late-fusion of five classifiers – two static functional systems, and three GMM-supervector systems. We show results for each subsystem on the development and test sets (Table 7) and compare with previous results on the Speaker State Intoxication Subchallenge test set.

5.1. Subsystem results

The modeling capabilities of all five subsystems should be examined in order to determine which subsystems are best overall. The global and iterative speaker normalization subsystems, which operate on static and hierarchical functionals, produce the highest single-system accuracies on the matched development set and the official development set, with the exception of latent factor analysis (LFA). However, LFA's performance drops notably from the official to the matched-prompt case. The GMM mean-supervector (MGMM) has the fourth highest unweighted accuracy, followed only by the lower dimensionality (512) unweighted posterior probability supervector (UWPP).

In terms of robustness, the closely related global and iterative functional-feature subsystems and the GMM mean-supervector system, all of which are very high-dimensional, show consistency between development and test set when trained on matched-prompt data. LFA and MGMM obtain considerably higher performance on the official development set than the matched development set, suggesting that the GMM methods built on MFCC features were exploiting utterance-specific lexical information to improve classification performance. The test set is identical in the official and matched experiments, only the training datasets differ.

The LFA rank R (the size of the $\mathbf{U}$ matrix) and the sub-rank r (the ordered-subset of the R factors used for SVM modeling) were tuned on the development set, with R having a maximum value of 55. For the non-matched set, cross-validation showed only $r = 10$ factors were necessary for peak performance when chosen from the LFA matrix of rank $R = 55$. For the matched set, $R = 40$ with $r = 20$ provided optimal performance. The additional factors included in the matched-prompt set may have provided higher robustness based on its slightly higher performance on the test set,

Table 7
Unweighted average recall for all five subsystems on the development (D) and test (T) datasets, official and matched, and the accuracies obtained from fusion of the subsystems. Results submitted for the Challenge are listed as "Prev., Test Set Optimized". All results on the GMM subsystems are for global speaker normalization.

Train data	Test data	Global-1	Iterative-2	[1,2]	MGMM-3	UWPP-4	[3,4]	LFA-5	[3–5]	[1–5]
Official	D,Official	0.719	0.684	**0.724**	0.677	0.659	0.704	0.700	**0.711**	**0.751**
Matched	D,Matched	0.697	0.676	**0.699**	0.626	0.639	0.658	0.656	**0.668**	**0.717**
Official	T,Official	0.699	0.690	**0.697**	0.632	0.615	0.656	0.619	**0.651**	**0.688**
Matched	T,Official	0.699	0.687	**0.704**	0.641	0.629	0.654	0.637	**0.662**	**0.714**
Prev., Test Set Optimized				**0.681**						**0.705**

Bold indicates points of focus for analysis.

Table 8
Increased accuracy of final prediction in relation to utterance length for the test set when trained with ALC matched-prompt data. Utterance length groupings are determined by quantile.

Utterance length (quantile)	[0,25]	(25,75]	(75,100]	[0,100]
Unweighted average recall	0.67	0.71	0.76	0.71

although it is universally observed that training with matched-data provides more consistent results when transitioning to the test set.

5.2. Fusion results

Fusion was carried out at the score-level in order to increase robustness to unexpected failures of specific sub-systems (results in Table 7). In parallel, the functional-feature subsystem scores are merged as are the GMM subsystems. Then, both composites are united to form the final system.

Performance of the static-functional systems ([1,2]) is comparable to the GMM systems ([3,4,5]) on the official development set, but not when evaluated on the official test set because the GMM systems overfit the non-matched lexical content. Additionally, the LFA vector receives a large fusion weight (0.79 versus 0.69) with the other two fused GMM classifiers (MGMM and UWPP) due to high development set performance, but it notably declines in recall on the test set, explaining the reduced performance in the official prompt case. In contrast, the accuracies of the matched-prompt models are consistent in both evaluation sets (devel and test).

The final accuracies on the test set are 68.8% for modeling the official prompts and 71.4% when modeling the matched-prompts. The matched-prompt-training results are higher than our best reported result in the Interspeech Speaker State Intoxication Subchallenge, where we achieved the maximal recall of 70.5%, but the official-prompt-training results are lower. The largest factor that explains this discrepancy is that we assumed a 1-1 distribution of the test set when performing speaker normalization during the Challenge, but here we relied on tuning of the SVM weight and cost parameters, among others, to produce the most generalizable results. Our matched-prompt-type accuracy of 71.4% is a 5.5% absolute improvement and 8.4% relative improvement of the openSmile baseline accuracy of 65.9% unweighted average recall, and the best result reported thus far.

6. Error analysis

Analysis of errors can improve future system performance and give insight into the limitations of the modeling task. With increasing utterance length, we can expect statistical variance to decrease, and our modeling accuracy to increase. Unweighted average recall (UAR) is displayed in relation to utterance length when evaluating on the test set with models obtained from our most-accurate system (Table 8). Unweighted average recall is 4% lower for the shorter utterances (first quartile) than the overall UAR. This serves as one explanation for differences in accuracy between prompts, although the more complex cognitive processes may even induce longer utterance durations.

Next, we examine performance of our system near the ‘intoxication’ line of 0.5 g/L. The top subplot in Fig. 2 demonstrates that sober recall is nearly unaffected on either side of the threshold, while intoxicated recall is by default 0 on the ‘sober’ side. Furthermore, the bottom subplot indicates that the percentage of samples classified as sober or intoxicated remains close to 50/50 regardless of the side of the ‘intoxication’ line it is on. Our system finds it difficult to tell ‘intoxicated’ from ‘sober’ speech, but can detect ‘non-alcoholized’ versus ‘alcoholized’ speech to a reasonable degree. Additionally, when considering only UAR on samples above the ‘intoxication’ threshold (because the ones below are much lower), there is a weak, significant correlation between BAC and UAR – Spearman’s rank-correlation coefficient $\rho = 0.35$ ($p < 0.05$). This correlation aligns with listening experiments conducted by Hollien et al. (2009). We also note that the recalls are similar between both classes for all speakers with no significant dependence between UAR and BAC ($p = 0.49$).

Interestingly, background normalization (performed on the matched-prompt development set) does not demonstrate independence between the predicted labels and blood-alcohol concentration. As BAC increases, the percentage of samples predicted as ‘intoxicated’ increases, $\rho = 0.35$ ($p < 0.05$). This is relevant because a person should be expected to produce a higher number of “drunk” sounding utterances if they are more intoxicated, and the system should pick

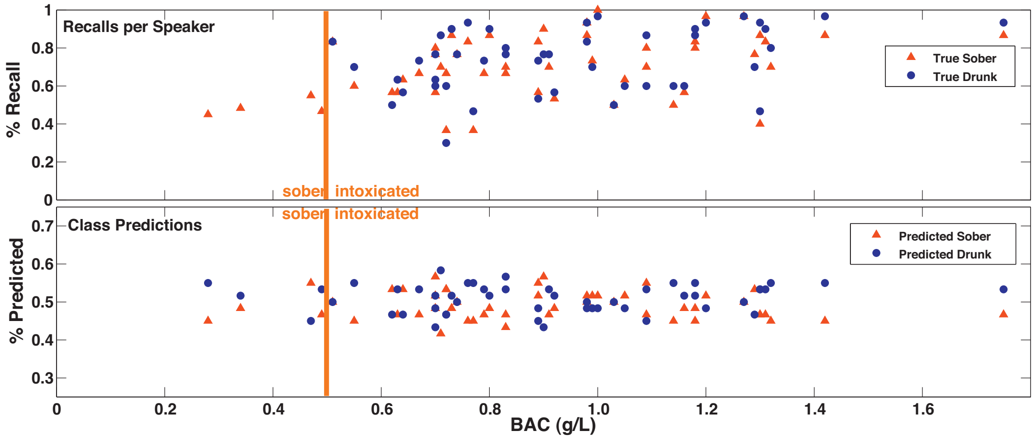

Fig. 2. Recall and prediction per speaker plotted against each speaker's BAC for final model – trained on matched-prompt train&devel sets and evaluated on the official test set.

up on that. We again find a weak, significant correlation between BAC and UAR when using background speaker normalization, $\rho = 0.30$ ($p < 0.10$); however, we still cannot explicitly detect the 'intoxication' cutoff since at least 30% of samples are classified as intoxicated per speaker. Background speaker normalization is a practical technique which provides comparable performance to other speaker normalization methods, but with lesser implicit assumptions about the distribution of evaluation set class labels built into the model.

7. Conclusion and future work

Speaker states are medium-term phenomena that are reflected, at least in part, in the speech signal patterns. We have presented a fused system composed of five classification subsystems that provides robustness and improvements when evaluated on the test set. The final unweighted accuracy on the test set was 71.4% when training with only matched-prompts (approximately matched lexical content) and 68.8% when modeling with the official prompt set. If optimized to improve UWA directly for the test set (assuming a 1-1 reference label distribution), we were able to achieve 70.5% performance (Bone et al., 2011). While such a result from models trained using the non-matched prompts was better than our results without assuming the test set label distribution (68.8%), training with matched prompts avoided modeling of extraneous lexical content and improved our final prediction (71.4%). We demonstrated that global and iterative speaker normalizations are both useful tools, although global normalization generally obtained higher results. In a potentially more realistic experiment, we have shown that with background speaker normalization, in which only baseline data is required for each speaker, we were able to achieve higher results than with no speaker normalization, and very similar numbers to what the related iterative speaker normalization produced. Importantly, we performed all classifications in this paper without making assumptions about the test set label distribution, showing some robustness in our system design.

Individual prompt-types were classified and the results were analyzed – the observations were congruous with psychological theory. It was demonstrated that modeling each prompt separately may provide higher recalls, but we did not explore this on the test set since we did not have a prompt recognition system and this information may be known in live systems. We corroborated perceptual studies on increasing intoxication perception given either amplified cognitive load or higher BAC. We demonstrated that longer utterances increased detection performance. It was also shown that gender-dependent models did not improve upon combined-gender models in our hierarchical functional subsystem.

Lastly, we showed through recall analysis that our system could detect 'alcoholized' versus 'non-alcoholized' speech, but could not do so explicitly at the threshold of 0.5 g/L. An important finding was that when using background speaker normalization on the matched-prompt development set, the number of predicted drunk utterance increases with BAC, $\rho = 0.35$ ($p < 0.05$) – an indication that this alternative speaker normalization framework will make even less implicit assumptions about the distribution of evaluation set class-labels. Unweighted average recall improved for higher BAC and for longer utterance lengths. These results should inform design of real-world speaker state detection systems.

Our model is general, and may be readily applied to other speaker state investigations, such as sleepiness detection, or it can be fused with additional, orthogonal subsystems. Future research into intoxicated speech detection may focus on further feature optimization, such that more complex modeling may be successful on a reduced set of features. Phonetic-based prosodic systems appear promising and intuitive (Honig et al., 2011). Dynamical modeling of prosody may also prove beneficial. Since we have demonstrated that recall increases with increasing BAC, ordinal techniques to create soft-labels may be useful for improved modeling.

Acknowledgments

This research was supported in part by the National Science Foundation and the National Institute of Health. The authors are very grateful to the organizers of the Interspeech Speaker State Challenge and to Dr. Florian Schiel for making the ALC database accessible to the Challenge participants.

References

Abroms, B., Fillmore, M., 2004. Alcohol induced impairment of inhibitory mechanisms involved in visual search. Experimental and Clinical Psychopharmacology 12, 243–250.

Barfusser, S., Schiel, F., 2010. Disfluencies in alcoholized speech. In: IAFPA Annual Conference.

Baumeister, B., Schiel, F., 2010. On the effect of alcoholisation on fundamental frequency. In: IAFPA Annual Conference.

Behne, D.M., Rivera, S.M., Pisoni, D.B., 1991. Effects of alcohol on speech: durations of isolated words, sentences, and passages. Research on Speech Perception 17, 285–301.

Black, M., Tepperman, J., Narayanan, S.S., 2011a. Automatic prediction of children's reading ability for high-level literacy assessment. IEEE Transactions on Acoustics, Speech and Language Processing 19, 1015–1028.

Black, M.P., Katsamanis, A., Baucom, B.R., Lee, C.C., Lammert, A.C., Christensen, A., Georgiou, P.G., Narayanan, S.S., 2011b. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. Speech Communication, http://dx.doi.org/10.1016/j.specom.2011.12.003.

Black, M.P., Katsamanis, A., Lee, C.C., Lammert, A.C., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.S., 2010. Automatic classification of married couples' behavior using audio features. In: Proceedings of Interspeech, pp. 2030–2033.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot International 5, 341–345.

Bone, D., Black, M.P., Li, M., Metallinou, A., Lee, S., Narayanan, S.S., 2011. Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors. In: Proceedings of Interspeech, pp. 3217–3220.

Burget, L., Matejka, P., Schwarz, P., Glembek, O., Cernocky, J., 2007. Analysis of feature extraction and channel compensation in a GMM speaker recognition system. IEEE Transactions on Acoustics, Speech and Language Processing 15, 1979–1986.

Busso, C., Metallinou, A., Narayanan, S., 2011. Iterative feature normalization for emotional speech detection. In: Proceedings of ICASSP, pp. 5692–5695.

Campbell, W., Sturim, D., Reynolds, D., Solomono, A., 2006. SVM-based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proceedings of ICASSP.

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., Vair, C., 2007. Compensation of nuisance factors for speaker and language recognition. IEEE Transactions on Audio, Speech, and Language Processing 15, 1969–1978.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 1–27.

Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE – the Munich versatile and fast open-source audio feature extractor. In: ACM Multimedia, Firenze, Italy, pp. 1459–1462.

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. LIBLINEAR: a library for large linear classification. Journal of Machine Learning Research 9, 1871–1874.

Galbraith, N., 1986. Alcohol: its effect on handwriting. Journal of Forensic Science 31, 580–588.

Hindmarch, I., Kerr, J., Sherwood, N., 1991. The effects of alcohol and other drugs on psychomotor performance and cognitive function. Alcohol and Alcoholism 26.

Hollien, H., DeJong, G., Martin, C.A., Schwartz, R., Liljegren, K., 2001. Effects of ethanol intoxication on speech suprasegmentals. Journal of the Acoustical Society of America 110, 3198–3206.

Hollien, H., Harnsberger, J.D., Martin, C.A., Hill, R., Alderman, G.A., 2009. Perceiving the effects of ethanol intoxication on voice. Journal of Voice 23, 552–559.

Honig, F., Batliner, A., Noth, E., 2011. Does it groove or does it stumble – automatic classification of alcoholic intoxication using prosodic features. In: Proceedings of Interspeech, pp. 3225–3228.

Kenny, P., Reynolds, D., Castaldo, F., 2010. Diarization of telephone conversations using factor analysis. IEEE Journal of Selected Topics in Signal Processing 4, 1059–1070.

Krajewski, J., Batliner, A., Golz, M., 2009. Acoustic sleepiness detection: framework and validation of a speech-adapted pattern recognition approach. Behavior Research Materials 41, 795–804.

Lee, C.C., Katsamanis, A., Black, M., Baucom, B., Georgiou, P., Narayanan, S., 2011. An analysis of PCA-based vocal entrainment measures in married couples' affective spoken interactions. In: Proceedings of Interspeech, pp. 3101–3104.

Levit, M., Huber, R., Batliner, A., North, E., 2001. Use of prosodic speech characteristics for automated detection of alcohol intoxication. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding.

Li, M., Han, K.J., Narayanan, S., 2013. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. Computer Speech and Language 27 (1), 151–167.

Li, M., Metallinou, A., Bone, D., Narayanan, S., 2012. Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling. In: Proceedings of ICASSP.

Pisoni, D.B., Martin, C.S., 1989. Effects of alcohol on the acoustic–phonetic properties of speech: perceptual and acoustic analyses. Alcoholism: Clinical and Experimental Research 13, 577–587.

Schiel, F., 2011. Perception of alcoholic intoxication in speech. In: Proceedings of Interspeech, pp. 3281–3284.

Schiel, F., Heinrich, C., 2009. Laying the foundation for in-car alcohol detection by speech. In: Proceedings of Interspeech, pp. 983–986.

Schiel, F., Heinrich, C., Barfüsser, S., 2011. Alcohol Language Corpus – the first public corpus of alcoholized German speech. Language Resources and Evaluation.

Schiel, F., Heinrich, C., Neumeyer, V., 2010. Rhythm and formant features for automatic alcohol detection. In: Proceedings of Interspeech, pp. 458–461.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., 2011. The INTERSPEECH 2011 Speaker State Challenge. In: Proceedings of Interspeech, pp. 3201–3204.

Schuller, B., Wimmer, M., Mosenlechner, L., Kern, C., Arsic, D., Rigoll, G., 2008. Brute-forcing hierarchical functionals for paralinguistics: a waste of feature space? In: Proceedings of ICASSP, pp. 4501–4504.

Schweizer, T.A., Jolicoeur, P., Vogel-Sprott, M., Dixon, M.J., 2004. Fast, but error-prone, responses during acute alcohol intoxication: effects of stimulus-response mapping complexity. Alcoholism: Clinical and Experimental Research 28, 643–649.

Sigmund, Zelinka, 2011. Analysis of voiced speech excitation due to alcohol intoxication. Information Technology and Control 40, 145–150.

Sobell, L.C., Sobell, M.B., 1972. Effects of alcohol on the speech of alcoholics. Journal of Speech and Hearing Research 15, 861–868.

Zhang, X., Lu, X., Shi, Q., Qin Xu, X., Chiu E. Leung, H., Harris, L.N., Iglehart, J.D., Miron, A., Liu, J.S., Wong, W.H., 2006. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinformatics 7, 197–209.

Please cite this article in press as: Bone, D., et al., Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors. Comput. Speech Lang. (2012), http://dx.doi.org/10.1016/j.csl.2012.09.004