Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 168–189

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification

William Yang Wang [a,*], Fadi Biadsy [a], Andrew Rosenberg [b], Julia Hirschberg [a]

[a] *Department of Computer Science, Columbia University, United States*
[b] *Computer Science Department, Queens College (CUNY), United States*

Received 2 May 2011; received in revised form 22 March 2012; accepted 23 March 2012
Available online 3 April 2012

Abstract

Traditional studies of speaker state focus primarily upon one-stage classification techniques using standard acoustic features. In this article, we investigate multiple novel features and approaches to two recent tasks in speaker state detection: level-of-interest (LOI) detection and intoxication detection. In the task of LOI prediction, we propose a novel Discriminative TFIDF feature to capture important lexical information and a novel Prosodic Event detection approach using AuToBI; we combine these with acoustic features for this task using a new multilevel multistream prediction feedback and similarity-based hierarchical fusion learning approach. Our experimental results outperform published results of all systems in the 2010 Interspeech Paralinguistic Challenge – Affect Subchallenge. In the intoxication detection task, we evaluate the performance of Prosodic Event-based, phone duration-based, phonotactic, and phonetic-spectral based approaches, finding that a combination of the phonotactic and phonetic-spectral approaches achieve significant improvement over the 2011 Interspeech Speaker State Challenge – Intoxication Subchallenge baseline. We discuss our results using these new features and approaches and their implications for future research.

Keywords: Emotional speech; Paralinguistic; Speaker state

1. Introduction

Although the automatic detection of speaker state has attracted considerable interest in recent years, most studies have focused on the analysis of anger, frustration, and other classic emotions (Litman and Forbes-Riley, 2004; Liscombe et al., 2005; Devillers and Vidrascu, 2006; Ai et al., 2006; Grimm et al., 2007; Gupta and Nitendra., 2007). This focus is motivated primarily by Spoken Dialogue System (SDS) applications, such as call centers and tutoring systems, for which it would be useful to recognize a speaker state such as anger or uncertainty in order to improve the user experience as well as task performance by automatically adapting the system-controlled conversation in real time (Bhatt et al., 2004; Gupta and Nitendra., 2007). The benefit of adapting SDS to the speaker's state is shown by recent work (Forbes-Riley and Litman, 2011) that demonstrates successful deployment of a speaker state classifier in a tutoring system. However, emotional state is not the only important speaker state to recognize. More recently, there have been studies of speaker states that do not map directly to the classic or even derived emotions: studies of charismatic speech

* Corresponding author. Tel.: +1 347 226 1057.
E-mail addresses: yww@cs.cmu.edu (W.Y. Wang), biadsy@google.com (F. Biadsy), andrew@cs.qc.cuny.edu (A. Rosenberg), julia@cs.columbia.edu (J. Hirschberg).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.03.004

(Biadsy et al., 2008), of deceptive speech (Hirschberg et al., 2005), and of medical conditions such as depression or autistic disfunction (Hirschberg et al., 2010) broaden the scope of paralinguistic analysis considerably.

In this paper, we present novel approaches to two of these speaker states: level-of-interest (LOI) and degree of speaker intoxication, both topics in the Interspeech Paralinguistic Challenges. In 2010, the Interspeech Paralinguistic Challenge launched a sub-challenge to detect speaker's LOI (Schuller et al., 2010; Wang and Hirschberg, 2011). Detecting LOI in a topic, product, or person is an important task in many domains. By automatically detecting users' interest in a product or service, for example, it should be easier for sales representatives to identify potential customers. In the political domain, the automatic detection of interest could augment traditional polling activities. Also, understanding the speaker's interest in a conversation might have a significant influence on improving customer service behavior. The 2011 Interspeech Speaker State Challenge launched another sub-challenge: intoxication detection (Schuller et al., 2011; Biadsy et al., 2011), a still more critical task from the point of view of public safety in countries like the United States, where hundreds of thousands of people are the victims of drunk driving every year. A system to detect a person's level of intoxication via minimally invasive means would be able to significantly aid in the enforcement of drunk driving laws, and ultimately to save lives.

We describe our analyses of both data sets from both these Paralinguistic Challenges and compare our features and their performance on each below. In Section 2, we review previous work. In Section 3, we describe our studies of LOI, including the corpus, features and methods we employ. In Section 4, we describe the corpus, features and methods we use for the intoxicated speech studies. We then compare our results on both data sets to understand why different features and methods are better used on different types of data.

2. Related work

2.1. Level-of-interest (LOI)

Schuller et al. (2006) were among the first to study automatic LOI detection from conversational speech. They designed their task as a multiclass classification task, extracting standard acoustic features, such as Mel-Frequency-Cepstral-Coefficients (MFCC), and building a bag-of-words (BoW) vector space model for lexical modeling. When concatenating the bag-of-words feature vector with the acoustic feature vector into a single vector, they achieved good F-measures using a Support Vector Machine (SVM). However, a bag-of-words approach clearly fails to capture contextual information in utterances. For example, the BoW model might not be able to capture negation (e.g. "This product is not bad at all."). In addition, since lexical and acoustic-spectral features are extracted from different domains, a single stage linear combination may not yield optimal results. The 2010 Interspeech Paralinguistic Challenge (Schuller et al., 2010) included an LOI subchallenge, encouraging researchers from many groups to propose new features and methodologies. Each team was given the same conversational speech corpus with annotated LOI, baseline acoustic features, and two baseline results, which were obtained using one a single layer classification. The evaluation metric used for the challenge was primarily the cross correlation (CC) measure (Grimm et al., 2008), with mean linear error (MLE) also taken into consideration. The baseline was built only from acoustic features with Random-Sub-Space meta-learning using unpruned REPTrees, and the CC and MLE for training vs. development sets were 0.604 and 0.118. For the test data, CC and MLE scores of 0.421 and 0.146 were observed.

Participants in this subchallenge included Gajšek et al. (2010), who based their system on the Gaussian Mixture Models as Universal Background Model (GMM-UBM) approach, with relevance MAP (Maximum A-Posteriori) estimation for the acoustic data motivated by the success of GMM-UBM modeling in speaker identification (Reynolds et al., 2000). They achieved CC and MLE of 0.630 and 0.123 in the training vs. development condition, but CC and MLE of only 0.390 and 0.143 in test. This performance difference may have been due to the fact that different subsets of the corpus include different speakers: acoustic features alone may not be robust enough to capture the speaker variation.

Jeon et al. (2010) won the 2010 Subchallenge by including lexical and subjectivity information in the form of term frequency and a subjectivity dictionary. In addition to a linear combination of all lexical and acoustic features, they designed a hierarchical regression framework with multiple levels of combinations. Its first two combiners combine hypotheses from different acoustic classifiers and then use a final stage SVM classifier to combine the overall acoustic posteriors with lexical posteriors to form the final output. They report a result of 0.622 for CC and 0.115 for MLE. On

the test set, they report CC and MLE of 0.428 and 0.146, respectively. Jeon et al. (2010) have also hypothesized that the drop in testing performance might be related to the robustness issue of traditional acoustic features (e.g. Mel-frequency cepstral coefficients) when modeling different groups of speakers.

2.2. Intoxication detection

Pisoni and Martin (1989) conducted early experiments on human perception of intoxicated speech, testing naive subjects and police judgments on the task of distinguishing intoxicated from sober speech. Their subjects could easily distinguish between utterances produced under sober and intoxicated conditions. Results of their acoustical analyses revealed consistent and well-defined changes in speech articulation between sober and intoxicated speech. A decade later, Hollien et al. (2001) investigated the relationship between prosodic characteristics and level of intoxication. Subjects were required to produce four types of utterances when sober and at four strictly controlled levels of intoxication: three ascending and one descending. Prosodic cues examined included fundamental frequency, intensity, speaking rate and disfluencies. These researchers found statistically significant changes for increasing intoxication, including increases in F0, in task duration and in number of disfluencies. They concluded however, that although certain changes in speech suprasegmentals, appear to occur in some speakers as a function of increasing intoxication, these patterns could not be viewed as universal, since approximately 20% of their subjects exhibited no or little change. In the same year, Levit et al. (2001) began to study machine perception and automatic detection of intoxication from speech. Their goal was to detect extreme intoxication, with blood alcohol level greater than 0.8 per mille, and cast this as a binary classification problem, using acoustic and prosodic features, including shimmer, jitter, duration of voiced and unvoiced segments. They used an intoxication speech corpus collected from the Police Academy of Hessen, Germany, which contains 120 readings (approx. 87 min) of the German version of the "The Sun and the Northern Wind" story, produced by 33 male speakers at different intoxication levels, with alcohol blood level varying between 0 and 2.4 per mille. The utterances were divided in intoxicated and sober speech with a boundary value of 0.8 per mille. They achieved almost 69% accuracy.

More recently, Schiel and Heinrich (2009) attempted to detect in-car intoxicated speech, collecting a corpus of read and spontaneous speech from multiple domains. Their first analysis of F0 showed that most speakers raise F0 under intoxication, although this was not consistent across genders. They also found that rhythmic features showed significant changes under alcohol.

3. Automatic detection of level-of-interest from speech

In this section, we first briefly describe the corpus, and analyze the distributions of three datasets in the corpus. Then, we list a comprehensive description of feature streams used in our investigation, and propose a two-tier hierarchical fusion approach, which is based on a novel feedback technique: multistream prediction feedback. Finally, we perform three major experiments, demonstrating the effectiveness of our approach to this task.

3.1. The LOI corpus

The corpus we use in our LOI experiments is the 2010 Paralinguistic Challenge Affect Subchallenge corpus Technische Universität München Audiovisual Interest Corpus (TUM AVIC), provided by Schuller et al. (2010). The corpus includes 10 h of audio–visual recordings of interviews in which an interviewer provides commercial presentations of various products to a subject. The subject and interviewer discuss the product, and the subject comments on his/her interest in it. Subjects were instructed to relax and not to worry about politeness in the conversation. 21 subjects participated (11 male, 10 female), including three Asians and the rest of European background. All interviews were conducted in English; while none of the subjects were native speakers, all were said to be fluent. 11 subjects were younger than 30 and 7 were between 30 and 40, and 3 were over 40. The subject portions of the recordings were segmented into speaker turns (continuous speech by one speaker with backchannels by the interviewer ignored). These were further segmented into sub-speaker turns at grammatical phrase boundaries such that each segment is shorter than 2 s. These smaller segments were annotated by four male undergraduate psychology students for subject LOI, using a 5-point scale as follows: (−2) *Disinterest* (subject is totally tired of discussing this topic and totally passive); (−1) *Indifference* (subject is passive and does not want to give feedback); (0) *Neutrality* (subject follows and participates

in the dialog, but it is not recognized if she/he is interested in the topic); (1) *Interest* (subject wants to talk about the topic, follows the interviewer and asks questions); (2) *Curiosity* (subject is strongly interest in the topic and wants to learn more). A normalized mean LOI is then derived from mean LOI/2, to map the scores into $[-1, +1]$. (Note that no negative scores occur for this corpus.) In our experiments, we consider the normalized mean LOI score as the label for each sub-speaker turn segment; we refer to this as "mean LOI" below. The corpus was divided for the Subchallenge into training, development, and test corpora; we use these divisions in our experiments.

To verify our earlier hypothesis on the different distributions of provided acoustic features among train, development, and test datasets, we first calculate the means $\mu_{i\in\{tr,de,te\}}$ and variances $\sigma^2_{i\in\{tr,de,te\}}$ of the distributions of the 1582 acoustic features (see Section 3.2 for details about the features). Then, we assume that the training distribution $\Phi_{(tr)}(\mu_{(tr)}, \sigma^2_{(tr)})$ is the background distribution $\Phi_{(m)}$ of the corpus and take the mean of the absolute differences beween development set distribution $\Phi_{de}(\mu_{de}, \sigma^2_{de})$, test set distribution $\Phi_{te}(\mu_{te}, \sigma^2_{te})$, and the background $\Phi_{(m)}$. Thus, the mean absolute difference in means $\delta_{(\mu)}$ can be calculated by using

$$\delta_{(\mu)} = \left| \frac{\sum_{i=1}^{D}((\sum_{j=1}^{N} X_{i,j}^{(t)}/|N|) - (\sum_{k=1}^{M} X_{i,k}^{(m)}/|M|))}{|D|} \right|$$

where $|N|$ is the number of instances of the target distribution, $|M|$ is the number of instances of the background distribution, and $|D|$ is the dimension of the feature space. Similarly, we can calculate the mean absolute difference in variances $\delta_{(\sigma^2)}$

$$\delta_{(\sigma^2)} = \left| \frac{\sum_{j=1}^{D} var(X_j^{(t)}) - var(X_j^{(m)})}{|D|} \right|$$

where $var(\cdot)$ denotes the variance array for all components in vector X. By looking at the values of $\delta_{(\mu)}$ and $\delta_{(\sigma^2)}$ from development vs. training, and test vs. training, we can have a general idea about the differences in distributions.

We computed the $\delta_{(\mu)}$ between the development vs. the training sets to be only 1.798, but the $\delta_{(\mu)}$ between test vs. training sets is 248.401, which implies a large difference of the mean of the absolute differences in means. When looking at the $\delta_{(\sigma^2)}$ between development vs. training sets, the result is 2.66×10^5 whereas test vs. training sets has a $\delta_{(\sigma^2)}$ of 9.55×10^{10}, which does not have the same magnitude as the former. Clearly, this again shows the different distributions of features among train, development, and test sets. This also entails that training and development sets are much more similar than training vs. test sets, and explains why previous work had unexpected drops of performances in the testing scenarios.

3.2. Features

In the task of LOI detection, we design a comprehensive set of features, spanning lexical, prosodic, and acoustic streams. In the lexical stream, our primary features are Discriminative TFIDF, Lexical Affect Scoring, and Language Modeling features. In the prosodic stream, we not only extract low level energy, duration, F0 and Voice Quality features, but we also include high level prosodic events. For the acoustic stream, we use a large set of traditional acoustic features including MFCC and other spectral cues, provided by the 2010 Interspeech Paralinguistic Challenge organizers. Table 1 provides an overview of the feature sets in the LOI experiments.

3.2.1. Discriminative TFIDF

As found in the official 2010 Interspeech Paralinguistic Challenge, the system (Gajšek et al., 2010) using acoustic features alone might be insufficient to capture LOI from speech. A possible explanation for this is that acoustic features used in the study are not adapted to speaker variations, and the test set is drawn from a completely different set of speakers, so acoustic cues alone might not be enough to capture LOI. In contrast, the winning system (Jeon et al., 2010) included some lexical features, and their results were more robust in the testing condition. As a result, in this study, we investigated lexical cues extensively.

Table 1
Feature sets.

Feature sets	Features
Discriminative TFIDF	Sum of word-level Discriminative TFIDF scores
Lexical Affect Scoring	Sum of word-level lexical affect scores
Language Modeling	Trigram language model
	log-likelihood and perplexity
Acoustic features	1582 acoustic features
	Detail see Schuller et al. (2010)
Prosodic and VQ	# Pulses, # periods, mean periods, SDev period
	Voicing fraction, # voice breaks, degree,
	Voiced2total frames
	Jitter local, local (absolute), RAP, PPQ5
	Shimmer local, local (dB), APQ3, APQ5, APQ11
	Harmonicity mean autocorrelation,
	Harmonicity mean NHR, mean NHR (dB)
	Duration seconds
	F0 min, max, mean, median, SDev, MAS
	Energy min, max, mean, SDev
Prosodic Events	Pitch accents, intermediate phrase, and intonational boundaries

VQ: Voice Quality; SDev: standard deviation; RAP: relative average perturbation; PPQ5: five-point period perturbation quotient; APQn: n-point amplitude perturbation quotient; NHR: noise-to-harmonics ratio; MAS: mean absolute slope.

In the standard vector space model, each term is associated with its Term Frequency (TF) in the utterance. The Inverse Document Frequency (IDF) provides information on how rare the term is over all utterances. The standard TFIDF vector of a term t in an utterance u is represented as $\mathbf{V}(t, u)$:

$$\mathbf{V}(t,u) = TF * IDF = \frac{C(t,u)}{C(v,u)} * \log \frac{|U|}{\sum_{t=1}^{n} u(t)}$$

TF is calculated by dividing the number of occurrences of term t in the utterance u by the total number of terms v in the utterance u. IDF is the log of the total number of utterances $|U|$ in the training set, divided by the number of utterances n in the training set in which the term t appears. $u(t)$ can be viewed as a simple function: if t appears in utterance u, then it returns 1, otherwise 0. In Discriminative TFIDF, we add additional information to the TFIDF metrics. When calculating IDF, we weight each term by the distribution of its labels in the training set. This helps us to weight terms by the LOI of the utterances they are uttered in. An intuitive example is this: although the terms "chaos" and "Audi" both appear once in the corpus, the occurrence of "Audi" is in an utterance with a Mean LOI score of 0.9, while "chaos" appears in an utterance with a label of 0.1. A standard TFIDF approach will give these two terms the same score.

We define our Discriminative TFIDF (DTFIDF) measure as follows – particularly to distinguish between such cases:

$$\mathbf{V}'(t,u) = \frac{C(t,u)}{C(v,u)} * \log \frac{|U|}{\sum_{t=1}^{n} u(t) * (1 - |MeanLOI(t,u)|)}$$

Here, the Mean LOI score ranging from (0, 1) is the label of each utterance.[1] Instead of summing the binary outputs of $u(t)$ directly, we now assign a weight to each utterance. The weight for term t in a particular utterance u is $(1 - |MeanLOI(t, u)|)$ in our task. The overall IDF score of terms important to identifying the LOI of an utterance will thus be boosted, as the denominator of the IDF metric decreases compared to the standard TFIDF. The Discriminative TFIDF measure is similar to the idea of Delta TFIDF (Martineau and Finin, 2009), but it is tailored for regression problems. Wang and McKeown (2010) show that adding Part-of-Speech (POS) information to a text can be helpful in similar classification tasks. So we have used the Stanford POS tagger (Toutanova et al., 2003) to tag these transcripts before calculating the Discriminative TFIDF score.

[1] In this dataset, there are no negative LOI scores among the annotations. However, in other data, if negative LOI instances are present, we could apply a linear function $y' = (y + 1)/2$ to project the original LOI score y to y' that is within the range of (0, 1).

3.2.2. Lexical Affect Scoring

Whissell's Dictionary of Affect in Language (DAL) (Whissell, 1989) attempts to quantify emotional language by combining results of human raters judging 8742 words collected from various sources including college essays, interviews, and teenagers' descriptions of their own emotional state. The DAL *pleasantness* (EE) score indicates the negative or positive valence of a word, rated on a scale from 1 to 3. For example, "abandon" scores 1.0, implying a fairly low level of pleasantness.[2] We note that the DAL is able to catch even some very slightly different semantics. For example, "successes" has a higher pleasantness score than "success" in the DAL. To calculate an utterance's overall pleasantness score, we first remove stopwords and then sum up the EE score of each word in the utterance.

3.2.3. Statistical Language Modeling

In order to capture contextual information, we also train a statistical language model[3] to augment the Discriminative TFIDF and lexical affect scores. We train trigram language models on the training set using the SRI Language Modeling Tookit (Stolcke, 2002). We apply the Witten-Bell smoothing (Bell et al., 1990) technique to smooth the trigram distribution. In the testing stage, the log likelihood and perplexity scores are used as Language Modeling features.

3.2.4. Acoustic, prosodic and Voice Quality features

As noted above, the TUM AVIC corpus is distributed with acoustic features (Schuller et al., 2010) for all of the data sets. These include 1582 acoustic features: PCM loudness, MFCC[0-14], log Mel Frequency Band[0-7], Line Spectral Pairs Frequency [0-7], F0 by Sub-Harmonic Sum., F0 Envelope, Voicing probability, Jitter local, Jitter consecutive frame pairs, and Shimmer local. In addition to these acoustic features, we have extracted 32 standard prosodic and Voice Quality features to augment these, including Glottal Pulses, Voicing, Jitter, Shimmer, Harmonicity, Duration, Fundamental Frequency, and Energy (see Table 1).

3.2.5. Prosodic Event features

To examine the relationship between Prosodic Events and LOI, we generate hypothesized ToBI (Tones and Break Indices) (Silverman et al., 1992) annotations using the AuToBI Toolkit (Rosenberg, 2010). The ToBI standard for the annotation of prosody describes linguistically significant prosodic variation in terms of phrasing and intonational prominence. ToBI defines 5 levels of juncture in its *break index tier*; a *tones tier* in which *pitch accents*, *phrase accents*, and *boundary tones* are defined; an *orthographic tier* in which words are aligned with the waveform; and a *miscellaneous tier* in which other phenomena maybe be annotated. Intonationally prominent words in English and German are marked by a pitch accent. ToBI describes the types of pitch accents in both languages using High (H) and Low (L) tones. In Standard American English (SAE), pitch accents may be simple (one tone) or complex (two tones). The inventory of pitch accent types in SAE includes H*, L*, L+H*, L*+H, H+!H*. When High tones are produced in a compressed pitch range, a "downstepped" indicator (!) is added to the tone, leading to three additional pitch accent types, !H*, L+!H*, and L*+!H. Prosodic phrasing in ToBI is hierarchical, such that each *intermediate phrase* contains one or more accented words plus a phrase accent, and each intonational phrase contains one or more intermediate phrases plus a boundary tone. There are three types of phrase accent, H-, L-, and !H-. Phrase accents, which end intermediate phrases, describe the pitch between the last pitch accent and the end of the phrase. Boundary tones, which end intonational phrases, describing the pitch at the boundary itself. There are two types of boundary tones, H% and L%. G-ToBI(S) (Mayer, 1995) is a version of ToBI defined for German. G-ToBI contains a similar tone inventory to the SAE version of ToBI. The pitch accents are H*L (fall), L*H (rise), HH*L (early-peak), L*HL (rise-fall), H*M (stylized contour), H* (high target), L* (low-target) and a down stepped fall (!H*L). The phrase accents and boundary tones are identical in G-ToBI and SAE ToBI. Moreover, there may be substantial differences in the acoustic realizations of these tones between the two languages. Moreover, the relative distributions of tone types are substantially different across the two languages. For example, in SAE ToBI, the most frequent pitch accent types are H* and !H*. In the Boston University Radio News Corpus (BURNC) (Ostendorf et al., 1995) of read English Broadcast News speech,

[2] Agarwal et al. (2009) notes that one of the advantages of this dictionary is that it has different scores for various forms of a root word. For example, the words "affect" and "affection" have very different meanings; if they were given the same score, the lexical affect quantification might not be discriminative.

[3] Due to the data sparsity issue, we did not train multiple language models with different mean LOI scores, though it might be possible to obtain benefits using the delta perplexity scores from language models trained with different groups of mean LOI scores were it not for this issue.

H* makes up 58% of all pitch accents with !H* making up an additional 25%. On the other hand in the MS corpus (Schweitzer, 2011) of read speech in German, the most frequent pitch accent type is L*H – characterized by a rise, rather than a high tone target – which makes up 51% of all pitch accents; the H* accent comprises only 12% of accents (Schweitzer, 2011).

AuToBI (Rosenberg, 2010)[4] is an open-source toolkit that automatically predicts ToBI annotations aligned to a word-segmentation. AuToBI first detects pitch accents and phrase boundaries (intermediate and intonational), and then classifies these based on the inventory described in the ToBI standard, as described above. It classes each type of Prosodic Event in a speech file aligned with an orthographic transcript – pitch accents, intermediate and intonational phrase boundaries – into categorical types based on the speech waveform aligned with each orthographic word. Since manually aligned transcripts are unavailable for this corpus, we align each utterance to its transcription using the Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008). Hypothesized prosodic events are then generated using AuToBI models trained on the spontaneous portion of the Boston Directions Corpus (BDC) (Hirschberg and Nakatani, 1996). For our experiments, we use frequency-based Prosodic Event features representing the rate of each Prosodic Event type.

3.3. Fusion learning approaches

If one finds that feature streams are informative when tested separately, it is useful to combine information from the streams from different domains to improve prediction. We have experimented with several approaches to feature combination in this work, including bag-of-features, Sum Rule combination, Hierarchical Fusion, and a new approach and present here results of each on our LOI prediction task. In the bag-of-features approach, a simple classification method includes all features in a single classifier. However, a potential problem for this method is that, when one combines 1582 acoustic features with 10 lexical features, some classifiers (e.g. naive Bayes, classification and regression trees), especially unregularized classifiers, will treat them equally; thus, potentially useful lexical features will not be evaluated properly. A second problem is that, when features are extracted from different streams using different methods, the scaling (e.g. normalization and/or standardization) of feature vector components can be challenging.

Another approach we examined was the Sum Rule Combiner, which uses product or sum rules to combine predictions from first-tier classifiers. Kittler et al. (1998) show that this approach outperforms the product rule, max rule and mean rule approaches when combining classifiers. Their sensitivity analysis shows that the Sum Rule Combiner is more resilient to estimation errors. However, although Sum Rule Combination can be useful in combining posteriors, it nonetheless assigns the same weight to each stream. In practice, this may not yield optimal results. A third feature combination method is the Hierarchical Fusion approach in which multistream information is fused using multiple classifiers and performing classification/regression in multiple stages. This approach can be implemented by first training first-tier classifiers for each single stream of features, collecting predictions, and training a second-tier supervector classifier to weight the utility of predictions from the different streams and to make a final prediction. This approach solves the scaling issue by letting the second-tier classifier weight the streams, as the predictions from the first-tier classifiers will be in a unified/normalized form (e.g. 0–1 in this task).

Another issue in the generalization performance of spoken language processing classifiers comes from speaker differences. Like many spoken language understanding tasks, in LOI detection, if we have a different set of speakers with different genders, ages, and speaker styles, the overall feature distribution for lexical, prosodic, and acoustic cues in the test set can be very different from the training set. Traditional speaker adaptation techniques typically focus only on the acoustic stream and may be very expensive to perform. None of the above fusion methods address the issue of speaker idiosyncrasies. We define a new method, improving over the Hierarchical Fusion approach, by extracting more knowledge about the lexical, prosodic, and acoustic features' distributions.

3.3.1. The multistream prediction feedback approach

Our Multistream Prediction Feedback and Mean Cosine Similarity based Hierarchical Fusion approach combines a hierarchical fusion approach with a multistream feedback approach. Fig. 1 shows the architecture of this system. Our proposed approach is influenced by the idea of Pseudo Relevance Feedback (PRF) (Yu et al., 2003) in Information

[4] http://eniac.cs.qc.cuny.edu/andrew/autobi/.

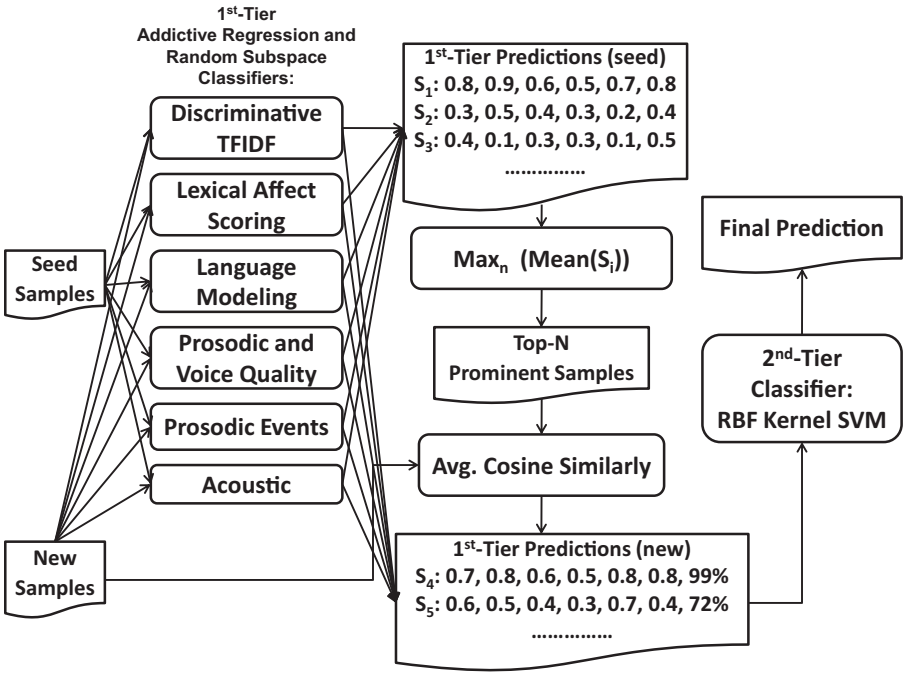

Fig. 1. Overview of multistream prediction feedback and mean cosine similarity based hierarchical fusion approach.

Retrieval (IR), where the PRF approach has the following two-staged strategies: (1) it performs an initial search of a given query, and assumes the top-k research snippets are (pseudo) relevant results; (2) it uses the results from step one to retrain or adapt the IR model and performs a final search of the same query using the updated model.

In Spoken Language Processing, there has been some recent work using PRF for meeting summarization (Chen et al., 2011), spoken term detection (Chen et al., 2011) and spoken document retrieval (Tu et al., 2011). These authors reported that using PRF could significantly improve the performances in all of the above spoken language processing tasks. However, PRF does not take cues from multiple feature streams into account and does not evaluate the similarities and relations among positive instances in the training stage, pseudo relevance instances (prominent samples), and new samples from the testing stage.

Our own fusion approach is based on the intuition that, if we can identify the *Prominent Samples* (e.g. the samples that all first-tier classifiers assign high average prediction scores),[5] then we can calculate the average distance between any new sample and these prominent samples in the Euclidean Space. We can also use this average distance as a new feature to improve the second-tier classifier's final prediction. By using the average distance, the gain here is that when running the trained model in the testing scenario, even the feature space might be shifted due to speaker differences, we explicitly incorporate a new similarity feature that captures these differences.

To implement this method, we first train five first-tier Additive Logistic Regression (Friedman et al., 2000) classifiers and a Random Subspace meta-learning (Ho, 1998) first-tier classifier (for the acoustic stream), resulting in six different feature streams in our training procedure. In testing, we use a random subset of the test set as seed samples. We classify these seed samples using each of our first-tier classifiers to obtain prediction scores ranging from 0 to 1 and calculate the mean predicted score for each sample. We then select the top n samples from the seed samples S to serve as our Prominent Samples by simply taking the top-n samples from the ranked list of the mean predicted LOI scores $Mean(S)$:

$$Prominent(S, n) = Max_n(Mean(S))$$

[5] In this work, we calculate Prominent Samples based on the predicted LOI scores from first-tier classifiers, but it is also possible to substitute the posteriors (confidence) scores from each first-tier classifier for these predicted scores.

Recall that the cosine similarity (Salton, 1989) of two utterances U_i, U_j in the vector-space model can be represented as:

$$\cos(U_i, U_j) = \frac{U_i \cdot U_j}{||U_i||_2 * ||U_j||_2}$$

where "·" indicates 'dot product'. Now, given our hypothesized Prominent Samples, we can estimate the distance from each new sample as follows: for each prominent sample and each new sample, we choose the original Discriminative TFIDF, Lexical Affect Scoring, Language Modeling, Prosodic and Voice Quality, and Prosodic Event features as a k-dimensional vector to represent the samples in Euclidean Space. When calculating the cosine similarity between positive examples in the training set and prominent samples in the test stage, we drop the 1582 acoustic features from the vector space model here because of the efficiency issue and the robustness issue of MFCC-style features, and substitute our 32 standard normalized prosodic features instead.

Now we use the mean cosine similarity score to represent how far a new sample U_n is from the Prominent Samples U_S in the space:

$$Sim(U_n, U_S) = \frac{1}{|S|}\sum_{s=1}^{|S|}\left(\frac{\sum_{i=1}^{k} \mathrm{V}_{(\mathrm{n,i})} * \mathrm{V}_{(\mathrm{s,i})}}{\sqrt{\sum_{i=1}^{k} \mathrm{V}_{(\mathrm{n,i})}{}^2} * \sqrt{\sum_{i=1}^{k} \mathrm{V}_{(\mathrm{s,i})}{}^2}}\right)$$

In the above equation, k is the total number of feature components in each feature vector $\mathbf{V}$.

We first sum up the cosine similarities of all possible combinations between the new sample and prominent samples, then normalize by the size of prominent samples $|S|$ to derive the mean cosine similarity between this new sample U_n and our prominent samples U_S. In the next step, we provide this mean cosine similarity measure as a new feature to the second-tier classifier, to use in reclassification. Like domain adaptation techniques, our approach also estimates the distribution of our features in the test set,[6] but it is inexpensive and does not require extra unlabeled data.

3.4. Evaluation

We conduct our experiments in three parts. First, we examine how well the Discriminative TFIDF feature performs, compared with the standard TFIDF feature. Secondly, we look at how our different feature streams affect the results. For these first two sets of experiments, we evaluate our features using the Subchallenge training vs. development sets only. Finally, in our third set of experiments, we compare our Multistream Prediction Feedback and Mean Cosine Similarity based Hierarchical Fusion approach to other feature-combining approaches. In these experiments, we first compare training vs. development performance, and then compare combined training and development sets vs. the test set. WEKA (Witten and Frank, 2005) and LIBSVM (Chang and Lin, 2001) are used for regression.

3.4.1. TFIDF vs. Discriminative TFIDF

Note that, when working with the training and development sets, we can access the label and transcriptions of each set to calculate the Discriminative TFIDF scores. For the testing scenario discussed in Section 3.1, we do not have these annotations. So, we redefine the task as a keyword spotting task, where we can use the keywords identified in the training and development sets as features in testing. We also sum up the word-level TFIDF scores and use the sentence-level TFIDF as a feature in the classification experiment. The regression algorithm we use is Additive Logistic Regression with 50 iterations. In this experiment, we directly evaluate the performances of TFIDF features in a one-stage classification setting. Table 2 shows how different approaches perform in the experiment.

We first show the experimental results from two simple baselines. For the Baseline1 in Table 2, we calculate the mean μ and the standard deviation σ from training set, and randomly generate a Gaussian distribution to predict all testing instances in the development set. The CC and MLE from Baseline1 are only −0.0067 and 0.222, respectively, suggesting a very pool predictive power. In Baseline2, when we simply assign the mean LOI score from the training set

[6] In real applications, instead of using a randomly chosen subset in the test set, we could incrementally select incoming prominent test samples, and iteratively calculate the cosine similarity features between prominent samples and the training data.

Table 2
Single TFIDF feature stream single regression results (Training vs. Develop, Additive Logistic Regression).

Method	CC	MLE
Baseline1 (Gaussian w. learned mean μ and std σ*)	−0.067	0.222
Baseline2 (mean LOI score**)	–	0.152
TFIDF	0.296	0.142
D-TFIDF	0.368	0.140
S-D-TFIDF	**0.381**	**0.136**

*Baseline1: using randomly generated Gaussian distribution with mean μ and standard deviation σ calculated from training set to predict all testing instances. **Baseline2: using the average score of all mean LOI scores in the training set to predict all testing instances. D-TFIDF: Discriminative TFIDF; S-D-TFIDF: the POS tagged version of D-TFIDF; CC: cross correlation; MLE: mean linear error.

Table 3
Comparing contributions of different feature streams in the second-tier classifier (Training vs. Development, Random Subspace for the First-tier Classifier of Acoustic Stream, and Additive Logistic Regression for other first-tier classifiers. Radial Basis Function (RBF) Kernel SVM as second-tier classifier).

Feature stream	CC	MLE
S-D-TFIDF	0.394	0.132
Language Modeling	0.404	0.141
Prosodic Events	0.458	0.133
Lexical Affect Scoring	0.459	0.132
Standard prosody + VQ*	0.591	0.122
Acoustic**	**0.607**	**0.118**
Multistream feedback ($n = 3$)	0.234	0.150
Multistream feedback ($n = 10$)	0.262	0.149
Multistream feedback ($n = 20$)	**0.290**	**0.146**

S-D-TFIDF: the POS tagged version of D-TFIDF; VQ: Voice Quality; *n*: top-*n* feedback; CC: cross correlation; MLE: mean linear error. * refers to the 32 prosodic and Voice Quality features we described in Section 3.2.4. ** refers to the 1582 acoustic features we described in Section 3.2.4.

to the development set, we obtain a baseline MLE of 0.152,[7] which is much better than the randomly generated Gaussian baseline. We see that the Syntactic Discriminative TFIDF approach is much more informative than the standard TFIDF approach. Note that, after calculating the global IDF score, the standard TFIDF approach selects **732** terms as top-1 level keywords.[8] In contrast, our Discriminative TFIDF has stronger discriminative power and picks a total number of **59** truly rare terms as top-1 level keywords.

3.4.2. Regression with different feature streams

In this experiment, we evaluate the contributions of different feature streams by utilizing the predicted LOI score of each feature stream from the output of its corresponding first-tier classifier as a single feature, and incorporating it into the second-tier classifier. Table 3 compares the performance of different feature streams. The second half of the table shows the contributions of multistream prediction feedback approach, when only using cosine similarity scores as the feature in the second-tier classifier. We see that the acoustic and prosodic features dominate in this task. The Prosodic Events feature stream emerges as an informative high-level prosodic feature. When testing the multistream feedback information as a single feature stream, we see in the bottom half of Table 3 that CC and MLE are improved when we increase the number of prominent samples. In the experiments of Section 3.4.3, the actual seed samples we take into account is 200, and the chosen feedback parameter is 30. Discriminative TFIDF and Language Modeling are also important, as seen from these results, but the Lexical Affect Scoring feature performs best among the lexical features in this task. We suspect that the reason may be a data sparsity issue, as we do not have a large amount of data

[7] Note that it is impossible to calculate the Pearson CC for this baseline method, due to the uniform distribution of the predicted LOI scores.

[8] Top-1 level keywords represent a set of words that have the same highest IDF scores in the corpus. We use the top-1 level keyword here to demonstrate the discriminative power of S-D-TFIDF, but in our experiments we utilize all levels of keywords in our training vocabulary.

Table 4
Comparing different systems.

Method	CC	MLE
Schuller et al. (2010)	0.604	0.118
Jeon et al. (2010)	0.622	0.115
Gajšek et al. (2010)	0.630	0.123
Bag-of-features fusion	0.602	0.118
Sum rule combination	0.617	0.117
SVM hierarchical fusion	0.628	0.115
Feedback + hierarchical fusion	**0.640**	**0.113**
Gajšek et al. (2010)	0.390	0.143
Schuller et al. (2010)	0.421	0.146
Jeon et al. (2010)	0.428	0.146
Bag-of-features fusion	0.420	0.145
Sum rule combination	0.422	0.138
SVM hierarchical fusion	0.450	0.131
Feedback + hierarchical fusion	**0.480**	**0.131**

Above: Training vs. Development; bottom: Combined Training +Development vs. Test; CC: cross correlation; MLE: mean linear error.

for training robust global Discriminative IDF scores, language models, and the feedback stream. In contrast, the DAL is trained on much larger amounts of data.

3.4.3. Comparing our multistream feedback based hierarchical fusion learning approach with state-of-the-art learning systems

Table 4 compares our Multistream Feedback based Hierarchical Fusion Learning approach to alternative learning approaches. The first half of this table reports results on training vs. development sets, and the second half compares combined training and development vs. test set results. Note that, in order to transcribe the test data, we have trained a 20-Gaussian-per-state 39-MFCC Hidden Markov Model speech recognizer with HTK, using the training and development sets together with TIMIT (Fisher et al., 1986), the Boston Directions Corpus (BDC) (Hirschberg and Nakatani, 1996), and the Columbia Games Corpus (Hirschberg et al., 2005). The word error rate (WER) is 29% on the development set.

We note that a bag-of-features model using Random Subspace with REP Trees gives worse results than the official baseline, which uses only the acoustic stream. When we use Sum Rule combination with Bagging to combine different feature streams, we obtain a CC score of 0.422. Although this improvement may seem small, it is quite significant (2-tailed paired t-test, $p < 0.0001$), comparing to the bag-of-features model. When using the SVM as the second-tier supervector classifier to weight different prediction streams, we achieve 0.628 CC and 0.115 MLE in training vs. development data, and 0.450 CC and 0.131 MLE on the test set; this result is significantly different from the bag-of-features baseline (paired t-test, $p < 0.0001$), but it is not significantly different from the Sum Rule Combination approach. Finally, augmenting the SVM hierarchical fusion learning approach with multistream feedback in our hierarchical approach, we obtain a final CC of 0.480 and MLE of 0.131 in the test mode, which is significantly different from the bag-of-features approach (paired t-test, $p < 0.0001$), but does not differ significantly from the SVM hierarchical fusion approach.

Thus, we can see that our Multistream Prediction Feedback approach does appear to represent a significant improvement over other methods of classifier combination, at least when using data sets in which there is considerable variation between the training and test sets in terms of the distribution of feature values.

4. Automatic detection of intoxicated speech

In this section, we explore approaches designed to detect a type of speaker state rather different from LOI – intoxication detection. Due to the nature of the phenomenon and the available data, we explore a somewhat distinct feature set and different methods for this task. The task is different from the LOI task because the corpus contains a fair amount of read speech across the training and testing splits, meaning that any training on lexical features will

Table 5
Number of speakers and utterances in our balanced set.

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

bias the results. As a result, the fusion of acoustic and lexical streams might not be the primary concentration of this task. In contrast, our approaches in this intoxication detection task concentrates on the acoustical and prosodic analysis that are motivated from the tasks of speaker and dialect identification. We believe that the direct representation of utterance-level lexical and acoustic features cannot capture the subtle durational, phonetic, and prosodic pattern changes between intoxicated and sober speech.

Our approach to investigating intoxicated speech is based upon intuitions about the features that should change significantly between speech of individuals when they are intoxicated, compared with the speech of the same individuals when they are sober. We hypothesize three possible qualities that may be impacted by intoxication. Our first hypothesis is that intoxicated speakers use prosody in different, but predictable ways when they are intoxicated compared to when they are sober, and that these differences may be realized through changes in phrasing and accenting behavior. Our second hypothesis is that a speaker's phone durations and phonotactic behavior differ under intoxicated and sober conditions. That is, articulator timing and speech rhythm may be modified in intoxicated speech, and this modification may be observable through the relative duration of phone units and the sequencing of such units. Our third hypothesis is that the quality of a speaker's phones – the acoustic characteristics of phones – may vary between intoxicated and sober states. To test this, we investigate the use of phone-sensitive acoustic modeling to detect intoxication, using a system which has been successfully applied to the identification of spoken dialects and accents. Under this third hypothesis, we view intoxicated speech in a given language (e.g. German) as simply a different accent of this language. We investigate these three hypotheses on the 2011 Interspeech Speaker State Challenge Intoxication Detection sub-challenge using the features and approaches described below.

4.1. The intoxication corpus

The Interspeech 2011 Speaker State Challenge German Alcohol Language Corpus (ALC) (Schuller et al., 2011) consists of 162 speakers (84 male, 78 female) within the age range of 21–75 (mean age 31.0 years and standard deviation 9.5 years) from 5 different locations in Germany. To acquire a gender balanced set, 77 male and 77 female speakers were selected randomly from the ALC corpus for the Challenge. All conversations are in German. The collection was divided into two parts. In the first, each subject was asked to choose the Blood Alcohol Concentration (BAC) level he/she wanted to reach, and was given the required amount of alcohol to do so, evaluated by standard medical formulas. After alcohol consumption, each speaker waited for 20 min to test his/her BAC level. The recorded range was from 0.28 to 1.75 per mille. Immediately after the BAC measurement, subject speech was recorded in three conditions: read, spontaneous and command-and-control; these recordings lasted less than 15 min, to prevent a drop in BAC level. Two weeks after the first experiment, the same subjects were asked to record 30 min of sober speech, in the same environment and supervised by the same staff. The official training set includes 3750 sober utterances and 1650 intoxicated utterances, such that a majority class accuracy for this data set is 69.4%. The official development set contains 2790 sober utterances and 1170 intoxicated utterances, for a majority class of 70.5%. The official baseline system on the development set achieves 65.3% accuracy(Schuller et al., 2011). For the purposes of our experiences, we also decided to create an approximately balanced data set for training and testing, in which we attempted to balance both number of speakers and number of utterances simultaneously, by first combining the training and development sets and then randomly selecting 20% of the speakers (from the grouped data) from each class as the new development set and 80% for training. We then attempted to equalize the number of utterances in both classes in training and testing by downsampling. The results of this selection are presented in Table 5. Below we term this the *balanced set* for the intoxication detection task. For this new division, the majority class of the development set is 52% and the majority class of the training set is 53.7%.

4.2. Prosodic Event modeling

Our investigation of the importance of prosodic changes between sober and intoxicated speech arises from scientific evidence of the effect of alcohol on the human body. Alcohol intoxication can resemble a stimulant or a depressant with respect to the mood and energy of a speaker. On the one hand, energetic intoxicated speakers may use more emphasis than sober speakers, leading to a higher rate of accenting, and use of accents associated with greater emphasis (L+H*, or L*+H). On the other hand, depressed intoxicated speakers may use less emphasis, realized as fewer accented words, or a greater rate of L* accents. Sentence planning has also been hypothesized as a major factor in prosodic phrasing (Krivokapic, 2010; Breen, 2011). Due to an impairment of an intoxicated speaker's ability to plan future lexical content, intoxicated speech may also include a greater rate of disfluencies and or of intonational phrase boundaries.

4.2.1. Features

As in Section 3.2.5, we use the AuToBI toolkit (Rosenberg, 2010) to identify Prosodic Events in the ToBI framework. When hypothesizing Prosodic Events on the IS11 Speaker State Challenge (IS11-SSC) material, we use the AuToBI models trained on all of the SAE speech from the Boston Directions Corpus material (Nakatani et al., 1995), both spontaneous and read material. While AuToBI uses word boundaries as the regions of analysis on which hypotheses are aligned, it does not use any other lexical information in its classification process. Therefore the lexical differences between the SAE training data and German evaluation data will only impact the hypothesis process due to durational differences between SAE and German words. Other influences are due to prosodic differences between the two languages. While SAE and German have significant similarities regarding their intonation, the G-ToBI(S) description of german intonation and the ToBI description of English point to some significant differences in the distribution of types of accenting and phrasing behavior. Moreover, the acoustic realizations of these similar tone descriptions of Prosodic Events may demonstrate significant differences across languages.

AuToBI is likely to generate errorful hypotheses on the IS11-SSC material. However, despite this noise, the hypothesized tones may still capture discriminative information concerning intoxicated vs. sober speech.

4.2.2. Approach

For the modeling of Prosodic Events for the detection of intoxicated speech, we use a representation of n-gram frequencies of prosodic events without constructing a Markov chain model. This feature representation decision allows us to discriminatively train a model that captures the same information as a traditional generative language model, while taking class information into consideration. For each value of n, we calculate the rate of occurrence of each n-gram in the observation sequence. To incorporate the backoff function of a standard language model, we include n-gram features for $n = \{1, 2, 3\}$. We construct these features in three ways: (1) using the full inventory of ToBI tones, (2) collapsing high tones (H) with downstepped high tones (!H) in pitch accents and phrase accents and (3) including a DEACCENTED tone to represent words that have no pitch accent. We also include distributional features such as the relative frequency of pitch accent, phrase accent, and boundary tone types, the overall accenting and phrasing rates, and the number of tones in the sequence.

4.2.3. Experiments

With this feature vector, we train a logistic regression classifier with L_1-regularization. Using 10-fold cross validation on the IS11-SSC training material, we observe 69.8% accuracy. Performing cross-validation on the training data makes use of speech material from the same speaker in training and testing folds. These cross-validation folds were generated by random fold assignment, and were not included as part of the distribution of the original corpus data. On the official development set, the prosodic modeling fails to significantly outperform the majority class baseline, with 69.6% accuracy and an F-measure of 0.032. It seems clear that the unbalanced distribution toward sober speakers has a major impact on this classification performance.

To maintain speaker independence while using a less skewed distribution of evaluation points, we evaluate this approach using class-balanced training and development sets. These data sets are proper subsets of the official training and development sets, constructed to have approximately equal amounts of intoxicated and sober material. Evaluating the Prosodic Event models trained on this balanced training set on the balanced development data, the accuracy remains at baseline, 53.3%, while the F-measure rises to 0.457 ($p = 0.53$, $r = 0.40$). This indicates that there is some discriminative information in the prosodic signal, despite the low performance on the official development data.

There are at least two explanations for the poor performance of this model on this material. First of all, there are differences between English and German intonation. The AuToBI hypotheses are generated for German speech using models trained on English material.

It is worthwhile to recall, that the G-ToBI description of German intonation contains a similar but distinct inventory of tones to describe Prosodic Events as the American English ToBI standard. However, the distribution of these tones differ radically in the two languages.

While we anticipated some additional noise due to the differences in the acoustic correlates to the ToBI-described Prosodic Event types across the two languages, this noise may be too great to yield a meaningful representation of prosody. At the time of writing, there are no AuToBI models trained on German speech. This explanation could be tested if such models were available.

The prosodic qualities that indicate that a speaker is intoxicated or sober may be dependent on the "way" in which he or she reacts to alcohol. As indicated above, sometimes intoxicated speakers are energetic, and other times they are sullen. This difference may impact the consistency of prosodic variation in the intoxicated condition, making classification difficult. Moreover, two speakers may have prosodic changes to their speech when intoxicated that vary wildly. The changes may not be consistent enough to be detected using this approach. Finally, prosodic analysis tends to be more effective when analyzing longer utterances. Much of the IS11-SSC material is quite short, sometimes only a few words. This too may limit the efficacy of this approach.

4.3. Phone duration and phonotactic modeling

Our second hypothesis is that intoxication may lead to changes in a speaker's phone durations and in their phonotactic patterns. Phone durations may be affected by the slurred speech characteristic of some intoxicated speakers. Phonotactic modeling has been quite successful for language and dialect identification (Zissman, 1996). Here, we hypothesize that intoxication may cause speakers to pronounce words differently, choosing some pronunciation variants more frequently than others, and even choosing certain words more frequently; each type of behavior would affect the phonotactic patterns in intoxicated speech.

4.3.1. Features

To investigate this hypothesis, we make use of the phones and temporal alignment provided for the training and development data in the sub-challenge to extract phone duration statistics for each phone type in each utterance. For each utterance, and each phone type, we extract the following features: minimum, maximum, mean and standard deviation of durations of all phone instances of this phone in the utterance. We also include global phone duration statistics at the utterance level. Specifically, we extract four additional duration features: minimum, maximum, mean and standard deviation of the durations of all phone instances from *all* types. For our phonotactic experiments we simply use the phonetic transcriptions provided in the corpus.

4.3.2. Approaches and experiments

To examine whether there are reliable phone duration differences between intoxicated and sober speech, we use the features described above in a logistic regression classifier. We obtain an accuracy on the official training set of 69.6%, using 10-fold cross validation. Testing on the official development set, our accuracy is 70.5%. It is interesting that, with such relatively simple features, we obtain an accuracy higher than the 65.3% obtained by the baseline system. Although our accuracy is not higher than the majority class, we see that our classifier does not always choose the majority class. Training and testing this classifier on our balanced sets, we obtain an accuracy of 62.5%, which is significantly better than the majority class baseline (52%). From these results it appears that phone duration statistics to be valuable in distinguishing intoxicated vs. sober speakers.

We then explore a vector-space based phonotactic modeling approach. We first collect the set of all triphones in the training data.[9] We then construct a feature vector for each utterance, where each element in this vector corresponds to a single triphone in our set. The value of this element is the frequency of this triphone in this utterance. To compensate for utterance duration differences, we normalize this vector by its Euclidian norm. We use these feature vectors to

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

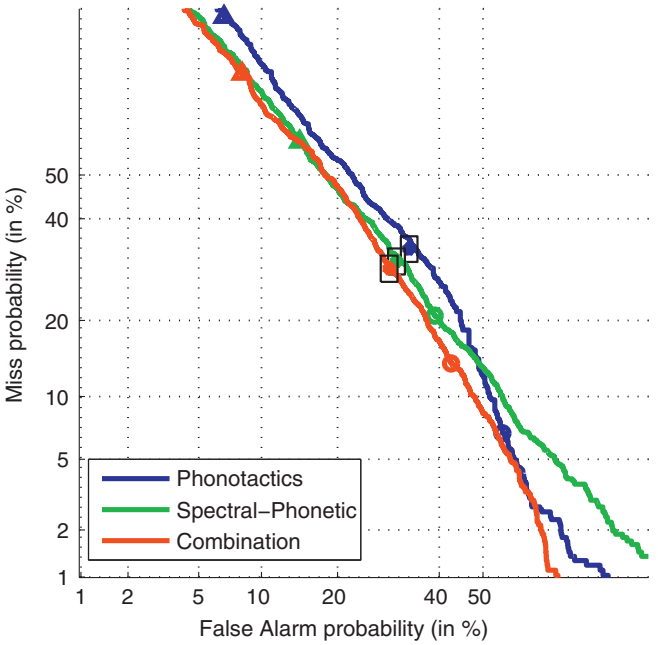

Fig. 2. DET curve for the official ALC development set (training on the official training set).

train an SVM classifier with linear kernel. The 10-fold cross-validation on the official training data is 70.1%. Training and testing on the official training and development data, respectively, we obtain an accuracy of 71.1%, which is significantly higher than the official baseline system (65.3%). Also, this accuracy is higher than the majority class baseline (70.5%), although the difference is not significant. If we train an SVM classifier using this approach on our balanced training data and test it on the balanced test set, we obtain an accuracy of 71.1%, which is significantly higher than the majority class baseline (52%). These results suggest that the phonotactic distributions across the two classes are significantly different.

We next examine the Detection Error Tradeoff (DET) curve, which plots false alarm vs. miss probabilities (of missing intoxicated speakers), as is standard in speaker verification (Martin et al., 1997). The DET curve allows us to determine the detection threshold of interest, and has also an advantage over accuracy for this data due to the skewness of the official development set.[10] As shown in Fig. 2,[11] the Equal Error Rate (EER) of the phonotactic approach on the official development set is 33.5%, significantly better than chance.[12] We obtain a slightly better EER when employing our balanced set, of 30.8%, also significantly better than chance (see Fig. 3).

4.4. Spectral-phonetic modeling

In this section, we test the hypothesis that intoxicated speakers realize certain phones differently than sober speakers. To model phonetic structural differences across these classes (sober and intoxicated), we adopt our recent and successful approach to dialect and accent recognition (Biadsy et al., 2011), treating intoxicated speech as a different accent of the speaker's native language.

4.4.1. Features

Again, we make use of the phones and alignments provided for this task, although it would also be possible to use a high-quality phone recognizer to obtain such information. The first step in creating our features is to build an acoustic

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

[11] Note that we are unable to plot the DET curve for the challenge baseline, because the posteriors from the baseline using WEKA SMO classifier are always zeros and ones.

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

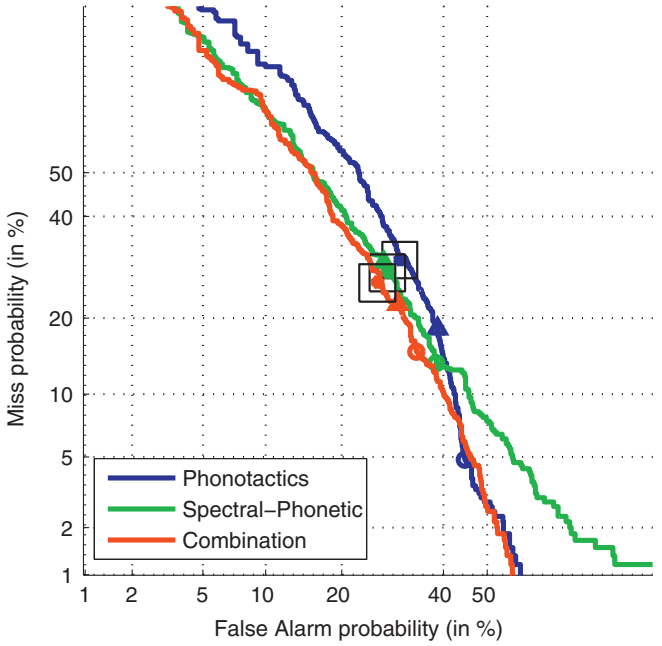

Fig. 3. DET curve for our balanced set (training on our balanced training set).

model for each phone type. We first extract acoustic features temporarily aligned to each phone instance in the training data from both classes, intoxicated and sober. We extract 13 RASTA-PLP features (including energy) plus delta and delta-delta, resulting in a 39D feature vector from each frame. Using the frames aligned to the same phone type (in all training utterances), we next train a Gaussian Mixture Model (GMM), with 60 Gaussian components with diagonal covariance matrices, for this phone type, employing the EM algorithm. Since some phone types occur infrequently in the training data, we build only a single GMM for each of the most frequent 45 phone types. Each phone GMM can be viewed as a GMM-Universal Background Model (GMM-UBM) for that phone type, since it models the general realization of that phone in both classes (Reynolds et al., 2000). We term these GMMs phone GMM-UBMs.

4.4.2. Approach

For our approach, we need a representation that captures the acoustic-phonetic features for each phone type in a given utterance (U). We adopt the GMM representation of (Campbell et al., 2006), but at the level of phone types, rather than the entire utterance. Specifically, we first obtain the acoustic frames aligned to every phone instance of the same phone type in U. We then use these frames to MAP adapt the corresponding phone GMM-UBM. We adapt only the means of the Gaussians using a relevance factor of $r = 0.1$. The resulting GMM of phone type ϕ we denote as the *adapted phone-GMM* (f_ϕ). The intuition here is that f_ϕ 'summarize' the variable number of acoustic frames of all the phone instances of a phone-type ϕ in a new distribution specific to ϕ in U.

We represent each utterance U as a set S_U of adapted phone-GMMs, each of which corresponding to a single phone type. Therefore, the size of S_U is, at most, the size of the phone inventory ($|\Phi|$). Let $S_{U_a} = \{f_\phi\}_{\phi\in\Phi}$ and $S_{U_b} = \{g_\phi\}_{\phi\in\Phi}$ be the adapted phone-GMM sets of utterances U_a and U_b, respectively. Next we design a kernel function to compute the 'similarity' between pairs of utterances, given their adapted phone-GMM sets. We compare the Kullback–Leibler (KL) divergence between the two adapted phone-GMMs, following Moreno et al. (2004) and Campbell et al. (2006).[13] The KL-divergence is not symmetric and does not satisfy the Mercer condition; thus it does not meet the requirements for use as the kernel function for an SVM. However, Campbell et al. (2006) proposed a kernel function between GMMs, based on an upper bound for their KL-divergence proposed by Do (2003). This function assumes that only the means

[13] Note that in these previous works, the entire utterance is represented by a single adapted GMM, independent of the linguistic labels and their segmentation.

of the GMMs are adapted, which is true in our case. Using this KL-divergence-based kernel between two adapted phone-GMMs modeling phone ϕ, we obtain the kernel function:

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

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

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

where f'_ϕ is the same as f_ϕ except that we subtract from its Gaussian mean vectors the corresponding Gaussian mean vectors of the phone GMM-UBM (of phone type ϕ). g'_ϕ is obtained similarly from g_ϕ. The subtraction allows zero contributions from Gaussians that are *not* affected by the MAP adaptation; this subtraction slightly improves accuracy in our dialect recognition work (Biadsy et al., 2011). For (2), when K_ϕ is a linear kernel, such as the one in (1), we can represent each utterance S_{U_x} as a single vector. This vector, say W_x, is formed by stacking the mean vectors of the adapted phone-GMM (after scaling by $\sqrt{\omega_\phi}\Sigma_\phi^{-1/2}$ and subtracting the corresponding $\vec{\mu}_\phi$) in some (arbitrary) fixed order, and zero mean vectors for phone types not in U_x. This representation allows the kernel in (2) to be written as in (3). This vector representation forms a kind of 'phonetic finger print' of the utterance's speaker. We noted that, in this vector, the phones constrain which Gaussians can be affected by the MAP adaptation (allowing comparison under linguistic constraints), whereas in the GMM-supervector approach (Campbell et al., 2006), in theory, any Gaussian can be affected by any frame of any phone.

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

Since we found in Section 4.3 that phone durations are important features, we also include duration statistics for each phone type from U_x in this vector (W_x), including the mean and standard deviation of the log durations of the phone instances of the same type in the utterance. As a result, we include 90 (45×2) new duration features. Now we test whether our method can capture phonetic differences between sober and intoxicated speakers. For our first experiment we use the official training data from both classes to train our phone GMM-UBMs. We construct a vector W_x for each utterance in the training data, as described above. Afterwards, employing our kernel function (3), we compute a kernel matrix for both classes using these vectors. We then train a standard binary SVM classifier using this kernel matrix. Our accuracy on 10-fold cross validation, using all the official training data, is 75.8%. This is significantly better than the majority class which is 69.4% and all of our approaches above. Testing our approach on the development set, we obtain a significant improvement in accuracy (72.8%) over both the majority class accuracy (70.5%) and the baseline system's accuracy (65.9%) and, again, better than our other approaches described above. As shown in Fig. 2, the EER of our system using this approach on the official development set is 30.9%, slightly better than the phonotactic system.

To test our system on our balanced data, we train our phone GMM-UBMs, on our balanced set of training data. We then train an SVM classifier as described above. Evaluating this classifier on our balanced development set, we obtain an accuracy of 71.2%, which is significantly better than majority class (52%). We report the DET curve on our balanced development set in Fig. 3; the EER is 28.2%.

In Table 6, we summarize the performances of our different approaches in this intoxication detection task.

We are also interested in testing whether phonotactics and phonetic systems can contribute to the classification task when combined. To plot the combination DET curves, we simply sum the posteriors from the two classifiers. As shown in Figs. 2 and 3, we observe that, in fact, the combination of these two approaches improve the EER over using any approach alone for both sets (the official and balanced). We obtain an EER of 29.4% using the official sets, and 26.3% on the balanced sets.

Table 6
Comparing different approaches on 10-fold cross-validation of training set, Official Training vs. Develop, and Balanced Training vs. Develop Set in the Intoxication Detection Task.

Approaches	10-fold X-Valid	Official Dev.	Balanced Dev.
Majority baseline	69.4%	70.5%	52%
Schuller et al. (2011)	–	65.3%	–
Prosodic Events	69.8%	69.6%	53.3%
Phone duration	69.6%	70.5%	62.5%
Phonotactic	70.1%	71.1%	71.1%
Phonetic-spectral	75.8%	72.8%	71.2%

5. Conclusion and further discussion

In this paper, we present a number of novel approaches and analyses to speaker state prediction, testing on two speaker-state detection tasks from the 2010 and 2011 Interspeech Challenges: the 2010 Affect Subchallenge and the 2011 Intoxication Subchallenge.

In the first, we measure Level of Interest (LOI) making use of novel lexical and acoustic/prosodic features, including the Discriminative TFIDF measure and lexical affect scoring using Whissell's Dictionary of Affect (DAL). In addition to direct modeling of acoustic features, we explore the use of hypothesized categorical Prosodic Events, extracted using the AuToBI Toolkit for Standard American English. For this task we also propose a new method for feature combination and feedback, the Multistream Prediction Feedback and Mean Cosine Similarity based Hierarchical Fusion approach, an ensemble of primary classifiers whose decision is combined using a second tier classifier. Results in Sections 3.4.2 and 3.4.3 have been presented in previous our conference paper (Wang and Hirschberg, 2011), and we have updated new results to Section 3.4.1.

In the intoxication detection task, we explore three approaches to the speaker-state classification task. We investigate the use of a state-of-the-art dialect identification technique treating intoxicated speech as a different "accent" of the language in a sober state. We also evaluate the use of phone duration and phonotactic information, as well as hypothesized Prosodic Events in this task. As of the time of writing, the LOI results outperform all systems from 2010 Interspeech Paralinguistic Challenge – Affect Subchallenge. The intoxication detection results, which have been included in our previous conference paper (Biadsy et al., 2011), significantly surpass the baseline result of 2011 Interspeech Speaker State Challenge – Intoxication Subchallenge.

5.1. Novel contributions for the LOI prediction task

5.1.1. The mean absolute difference of means and variances in two distributions

When surveying the previous work of 2010 Interspeech Paralinguistic Challenge, we realize both participating teams in the LOI prediction task suffer from significant downgraded performances between the results on development test set and official test set. We hypothesize that the acoustic features might not be robust enough to model different speaker, and we have proposed the measures of mean absolute difference in means and variances to calculate the difference in distributions of feature space in pair-wise datasets. Our analysis shows that the development set and the training set are much more similar than the test set and the training set (both contains distinct speakers), in terms of feature distributions. This analysis also explains why the previous work show have worse results on the test set.

5.1.2. The Discriminative TFIDF approach

In particular, we found that using Discriminative TFIDF measures improved over simple TFIDF metrics, confirming our hypothesis that using regression labels in the training set can help us better discriminate subtle differences between rare keywords and identify real task-specific keywords that a standard TFIDF approach fails to capture. Discriminative TFIDF also significantly outperforms the baseline where we simply assigning the mean LOI scores from training set to all test utterances, which can be seen as a generative multinomial autoregressive model that predicts the general likelihoods of having LOI in the conversations within a population. Discriminative TFIDF features may also prove

useful in other speech and language tasks. We also discovered that Whissell's Dictionary of Affect in Language proved useful in LOI detection.

5.1.3. The Prosodic Event analysis

Another novel finding in this study is that automatically extracted Prosodic Events can be useful when compared to direct modeling of prosodic information. Results on this task suggest that the use of hypothesized Prosodic Events can prove an effective alternative to the direct modeling of low level acoustic and prosodic features, serving as a kind of intermediate representation. Since Prosodic Event hypotheses are generated using features similar to those used in direct modeling, this symbolic feature representation can be viewed as a linguistically salient dimensionality reduction. Automatically extracted Prosodic Events can serve to distill the information contained in the raw acoustic features into a form that retains discriminative information for classification, while dramatically reducing the domain and number of variables required. This has a particular advantage over empirically defined dimensionality reduction techniques in that there are clear linguistic implications of Prosodic Events.

5.1.4. The multistream feedback and hierarchical fusion approach

In examining the differences between the features that proved most helpful in classifying our corpus, we find some distinctions between features that represent *what* is said vs. *how* a speaker realizes lexical content. Knowing that traditional acoustic features (e.g. MFCC) might not be robust across different datasets, we hypothesized that lexical information should be informative when for investigating speaker states like LOI, where speaker attitude toward particular commercial products has been shown to be identifiable in sentiment analysis studies through lexical cues. Our findings bear out this intuition. However, combining the lexical and the acoustic streams is always a challenging issue. Motivated by Pseudo Relevance Feedback technique in IR, we propose a Multistream Feedback and Hierarchical Fusion Approach, which considers the average similarity between prominent samples in test set, and samples in our training set. Through these experiments, we have validated our hypothesis that a hierarchical combination of features can improve system performance significantly in the LOI prediction task.

5.2. Novel contributions for the intoxication detection task

5.2.1. The phonotactic approach

In the second task, we use an alternative method to capture segmental information: a phonotactic approach. This is based on the assumption that intoxicated speakers pronounce words differently than sober speakers. Since the intoxication corpus includes read as well as spontaneous speech – and these conditions are not marked in the test set – using word-level lexical features cannot be relied upon to discriminate the two classes. However, pronunciation variation may occur even in read speech. Our experiment results clearly show that the phonotatic approach is informative in the intoxication detection task.

5.2.2. The phone durational approach

We found that the phone durational features much more useful than we expected. We discovered that a simple phone duration-based model could significantly improve over the official challenge baseline, which was built on thousands of low-level acoustic features. Our phone duration-based vector space model is motivate by the assumption that the length of individual phonemes will vary between a given speaker's intoxicated and sober states. Speakers may increase or decrease their speaking rate, depending upon their type of intoxication, so that phones may be lengthened or shorted as a consequence.

5.2.3. The phonetic-spectral approach

In the intoxication task, we build our phonetic-spectral model using acoustic frames for phone-level representation, which allows us to capture subtle phonetic differences between intoxicated and sober speech. This is based on the assumption that intoxicated speakers might realize phones differently than sober speakers. We then compute KL-based kernel of SVM supervectors for those phone-type based GMMs over the entire utterance. While the official challenge baseline represents utterance-level acoustic features directly, its performance is much worse than our phone-based acoustic feature representation. Our result is also in line with state-of-the-art results in dialect and language

identification tasks: the results from the phonetic-spectral approach were better than the phonotactic approach and the phone durational approach in our intoxication detection task.

5.3. Similarities and differences of the two tasks

In terms of the similarities between the two tasks, we notice that both tasks are less-researched non-traditional speaker state prediction tasks that have practical applications and potential social impacts. This requires us to think carefully before we start about which existing spoken language processing techniques might be useful for this task, or what useful adaptations might be made to make existing techniques more applicable. The corpora of both tasks also contain a significant portion of spontaneous speech, which always causes robustness issues in automatic speech recognition, as well as other speech application.

One of the obvious differences between the two tasks is the material, which represent distinct languages: English and German. This might require multilingual resources for training to achieve optimal results in testing. Another difference is that the corpus of the second task contains read speech across different partitions of the datasets, which might make the integration of lexical features and acoustic features less interesting.

We have examined approaches on modeling "what was said" in the two tasks. In the LOI task, we perform a comprehensive lexical modeling approach, which includes Discriminative TFIDF, and have obtained a powerful alternative feature stream to augment acoustic features. However, we were not able to examine the utility of lexical features in the intoxication task, since the intoxication corpus includes read speech (from the same material) as well as spontaneous. This is also one of the reasons why we were not able to examine our proposed multistream prediction feedback and fusion approach for the intoxication task.

While the use of prosodic analysis in the analysis of speaker state and paralinguistic qualities would appear to be a natural fit for both our corpora, and we have performed the Prosodic Event analyses for both tasks, in fact we found that Prosodic Event detection is not useful for the intoxication task. There are several possible explanations for our poor results using Prosodic Events in the intoxication task: (1) the input language is German, while the prosodic analysis models are trained on English; (2) intoxication may lead to disparate changes to a speaker's prosody that are more difficult to classify; and (3) the automatic prosodic event hypotheses may be too noisy to capture relevant prosodic variance for this classification task.

We have seen that acoustic features play a major role in speaker state detection in both corpora. In the LOI task, the acoustic stream dominates all other feature streams. In the intoxication task, we also find that our phonetic-spectral approach outperforms our other sources of information. However, there is a major difference between these two approaches. In the approach for LOI prediction, acoustic features are represented for each utterance directly, while in the intoxication task, we implement a much more complicated model based on a large margin method on the adapted Phone-GMM-UBM based supervectors.

Fusing different feature streams is one of our focused approach of the first task, and we have observed clearly the contributions from our multistream prediction feedback approach. In our second task, even though we have only attempted to combine the two best approaches using simple posteriror combination approach, we wish to perform our multistream prediction feedback approach once we are able to obtain more reliable and predictive feature streams in this task.

5.4. Future work

In the future, we plan to explore more distributional similarity measures (Lee, 1999), such as Euclidean distance, L_1 norm, and Jaccard's coefficient for the LOI task. It will also be interesting to investigate iterative methods for incrementally discovering prominent seed data points, sparse features and seed sample sizes that best generate prominent data points in the test set. After publication of our initial LOI paper (Wang and Hirschberg, 2011), Woellmer et al. (2011) proposed a model that utilizes context information from neighboring utterances, and obtained promising results. Therefore, investigating a joint model that reasons in the intersentential space is also important. Regarding the intoxication detection task, we would like to directly incorporate prosodic features in the kernel of our phonetic-spectral based approach. It would also be interesting to investigate our multi-stream prediction feedback approach for combining the phonotactic, phonetic, and prosodic views of the data.

References

Agarwal, A., Biadsy, F., McKeown, K.R., 2009. Contextual phrase-level polarity analysis using Lexical Affect Scoring and syntactic *n*-grams. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009).

Ai, H., Litman, D.J., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A., 2006. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In: Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006).

Bell, T., Cleary, J., Witten, I., 1990. Text Compression. Prentice Hall, Englewood Cliffs, NJ.

Bhatt, K., Evens, M., Argamon, S., 2004. Hedged responses and expressions of affect in human/human and human/computer tutorial interaction. In: Proceedings of 26th Annual Meeting of the Cognitive Science Society (CogSci 2004).

Biadsy, F., Hirschberg, J., Ellis, D., 2011. Dialect and accent recognition using phonetic-segmentation supervectors. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011).

Biadsy, F., Rosenberg, A., Carlson, R., Hirschberg, J., Strangert, E., 2008. A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech. In: Proceedings of the Fifth Meeting of the Speech Prosody Special Interest Group of the International Speech Communication Association (Speech Prosody 2008).

Biadsy, F., Wang, W.Y., Rosenberg, A., Hirschberg, J., 2011. Intoxication detection using phonetic, phonotactic, and prosodic cues. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011).

Breen, M., 2011. Intonational phrasing is constrained by meaning, not balance. In: Language and Cognitive Processes.

Campbell, W., Sturim, D., Reynolds, D., 2006. Support vector machines using GMM supervectors for speaker verification. In: IEEE Signal Processing Letters.

Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A., 2006. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006).

Chang, C.C., Lin, C.J., 2001. Libsvm: a library for support vector machines.

Chen, Y.N., Chen, C.P., Lee, H.Y., Chan, C.A., Lee, L.S., 2011. Improved spoken term detection with graph-based re-ranking in feature space. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011).

Chen, Y.N., Huang, Y., Yeh, C.F., Lee, L.S., 2011. Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011).

Devillers, L., Vidrascu, L., 2006. Real-life emotions detection with lexical and paralinguistic cues on human–human call center dialogs. In: Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006).

Do, M., 2003. Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models. In: IEEE Signal Processing Letters.

Fisher, W.M., Doddington, G.R., Goudie-Marshall, K.M., 1986. The darpa speech recognition research database: specifications and status. In: Proceedings of the DARPA Workshop on Speech Recognition.

Forbes-Riley, K., Litman, D., 2011. Using performance trajectories to analyze the immediate impact of user state misclassification in an adaptive spoken dialogue system. In: Proceedings of the 12th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2011).

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. The Annals of Statistics.

Gajšek, R., Žibert, J., Justin, T., Štruc, V., Vesnicer, B., Mihelič, F., 2010. Gender and affect recognition based on GMM and GMM-UBM modeling with relevance MAP estimation. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010).

Grimm, M., Kroschel, K., Mower, E., Narayanan, S., 2007. Primitives-based evaluation and estimation of emotions in speech. In: Speech Communication.

Grimm, M., Kroschel, K., Narayana, S., 2008. The vera am mittag German audio–visual emotional speech database. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2008).

Gupta, P., Nitendra, R., 2007. Two-stream emotion recognition for call center monitoring. In: Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007).

Hirschberg, J., Benus, S., Brenier, J.M., Enos, F., Friedman, S., Gilman, S., Gir, C., Graciarena, M., Kathol, A., Michaelis, L., 2005. Distinguishing deceptive from non-deceptive speech. In: Proceedings of the 6th Annual Conference of the International Speech Communication Association (Interspeech 2005).

Hirschberg, J., Hjalmarsson, A., Elhadad, N., 2010. Using computational approaches for modeling speaker state to gauge illness and recovery. In: Speech in Mobile Environments, Call Centers and Clinic.

Hirschberg, J., Nakatani, C.H., 1996. A prosodic analysis of discourse segments in direction-giving monologues. In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL 1996).

Ho, T.K., 1998. The random subspace method for constructing decision forests. The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Hollien, H., DeJong, G., Martin, C.A., Schwartz, R., Liljegren, K., 2001. Effects of ethanol intoxication on speech suprasegmentals. In: Proceedings of the 142nd Meeting of the Acoustical Society of America.

Jeon, J.H., Xia, R., Liu, Y., 2010. Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010).

Kittler, J., Hatef, M., Duin, R.P., Matas, J., 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Krivokapic, J., 2010. Speech planning and prosodic phrase length. In: Proceedings of the Sixth Meeting of the Speech Prosody Special Interest Group of the International Speech Communication Association (Speech Prosody 2010).

Lee, L., 1999. Measures of distributional similarity. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999).

Levit, M., Huber, R., Batliner, A., Nöth, E., 2001. Use of prosodic speech characteristics for automated detection of alcohol intoxication. In: Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding.

Liscombe, J., Hirschberg, J., Venditti, J.J., 2005. Detecting certainness in spoken tutorial dialogues. In: Proceedings of the 6th Annual Conference of the International Speech Communication Association (Interspeech 2005).

Litman, D., Forbes-Riley, K., 2004. Predicting student emotions in computer–human tutoring dialogues. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004).

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The det curve in assessment of detection task performance. In: Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997).

Martineau, J., Finin, T., 2009. Delta tfidf: an improved feature space for sentiment analysis. In: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM 2009).

Mayer, J., 1995. Transcription of German intonation – the Stuttgart system.

Moreno, P., Ho, P., Vasconcelos, N., 2004. A Kullback–Leibler divergence based kernel for SVM classification in multimedia applications. In: Advances in Neural Information Processing Systems, vol. 16.

Nakatani, C., Hirschberg, J., Grosz, B., 1995. Discourse structure in spoken language: studies on speech corpora. In: Proceedings of AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.

Ostendorf, M., Price, P., Shattuck-Hufnagel, S., 1995. The Boston University radio corpus.

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

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. In: Digital Signal Processing.

Rosenberg, A., 2010. Autobi – a tool for automatic tobi annotation. In: The 11th Annual Conference of the International Speech Communication Association (Interspeech 2010).

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.

Schiel, F., Heinrich, C., 2009. Laying the foundation for in-car alcohol detection by speech. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009).

Schuller, B., Köhler, N., Müeller, R., Rigoll, G., 2006. Recognition of interest in human conversational speech. In: Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müeller, C., Narayanan, S., 2010. The interspeech 2010 paralinguistic challenge. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010).

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., 2011. The interspeech 2011 speaker state challenge. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011).

Schweitzer, A., 2011. Production and perception of Prosodic Events – evidence from corpus-based experiments. Doctoral Dissertation, Universitat Stuttgart.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. Tobi: a standard for labeling English prosody. In: Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP 1992).

Stolcke, A., 2002. SRILM – an extensible Language Modeling toolkit. In: Proceedings of the 2002 International Conference on Spoken Language Processing (ICSLP 2002).

Toutanova, K., Klein, D., Manning, C.D., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 4th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL 2003).

Tu, T.w., Lee, H.y., Lee, L.s., 2011. Improved spoken term detection using support vector machines with acoustic and context features from pseudo-relevance feedback. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011).

Wang, W.Y., Hirschberg, J., 2011. Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In: Proceedings of the 12th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2011).

Wang, W.Y., McKeown, K., 2010. "Got you!": automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010).

Whissell, C., 1989. The dictionary of affect in language. In: Emotion: Theory Research and Experience.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed.

Woellmer, M., Weninger, F., Eyben, F., Schuller, B., 2011. Acoustic-linguistic recognition of interest in speech with bottleneck-blstm nets. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011).

Yu, S., Cai, D., Wen, J.R., Ma, W.Y., 2003. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In: Proceedings of the 12th International Conference on World Wide Web (WWW 2003).

Yuan, J., Liberman, M., 2008. Speaker identification on the scotus corpus. In: Proceedings of Acoustics 2008.

Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Transactions on Speech and Audio Processing.