Detecting Levels of Interest from Spoken Dialog with Multistream Prediction Feedback and Similarity Based Hierarchical Fusion Learning

William Yang Wang
Department of Computer Science
Columbia University
New York, NY 10027
ywu2347@columbia.edu

Julia Hirschberg
Department of Computer Science
Columbia University
New York, NY 10027
julia@cs.columbia.edu

Abstract

Detecting **levels of interest** from speakers is a new problem in Spoken Dialog Understanding with significant impact on real world business applications. Previous work has focused on the analysis of traditional acoustic signals and shallow lexical features. In this paper, we present a novel hierarchical fusion learning model that takes feedback from previous multistream predictions of prominent seed samples into account and uses a mean cosine similarity measure to learn rules that improve reclassification. Our method is domain-independent and can be adapted to other speech and language processing areas where domain adaptation is expensive to perform. Incorporating Discriminative Term Frequency and Inverse Document Frequency (D-TFIDF), lexical affect scoring, and low and high level prosodic and acoustic features, our experiments outperform the published results of all systems participating in the 2010 Interspeech Paralinguistic Affect Subchallenge.

1 Introduction

In recent years, there has been growing interest in identifying speakers' emotional state from speech (Devillers and Vidrascu, 2006; Ai et al., 2006; Liscombe et al., 2005). For Spoken Dialog Systems (SDS), the motivation has been to provide users with improved over-the-phone services by recognizing emotions such as anger and frustration and directing users to a human attendant. Other forms of *paralinguistic* information which researchers have attempted to detect automatically include other classic emotions, charismatic speech (Biadsy et al., 2008), and deceptive speech (Hirschberg et al., 2005). More recently, the 2010 Interspeech Paralinguisic Affect Subchallenge sparked interest in detecting a speaker's **level of interest** (LOI), including both the speaker's interest in the topic and his/her willingness to participating in the dialog (Schuller et al., 2010). Sensing users' LOI in SDS should be useful in sales domains, political polling, or service subscription.

In this paper, we present a similarity-based hierarchical regression approach to predicting speakers' LOI. The system has been developed based on the hierarchical fusion learning of lexical and acoustic cues from speech. We investigate the contribution of a novel source of information, **Discriminative TFIDF**; lexical affect scoring; and prosodic event features. Inspired by the successful use of Pseudo Relevance Feedback (Tao and Zhai, 2006) techniques in Information Retrieval and the cosine similarity measure (Salton, 1989) in Data Mining, we design a novel learning model which takes the multistream prediction feedback that is initially returned from seed samples [1] and uses a mean cosine similarity measure to calculate the distance between the new instance and prominent seed data points in the Euclidean Space. We then add this similarity measure as a new feature to perform a reclassification. Our main contributions in this paper are: (1) *the novel Discriminative TFIDF approach for lexical modeling and keywords spotting;* (2) *using lexical affect scoring and language modeling techniques to augment lexical modeling;* (3) *combin-*

[1] Seed samples are from a random small subset in the test set.

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 152–161,
Portland, Oregon, June 17-18, 2011. ©2011 Association for Computational Linguistics

ing (1) and (2) with additional low-level prosodic features together with voice quality and high-level prosodic event features; and (4) *introducing a multistream prediction feedback and mean cosine similarity based fusion learning approach.*

We outline related work in Section 2. The corpus, system features, and machine learning approaches are described in Section 3. We describe our experimental results in Section 4 and conclude in Section 5.

2 Related Work

Schuller et al. (2006) were among the first to study LOI from conversational speech. They framed this task as either a three-way or binary classification, extracting standard acoustic features and building a bag-of-words vector space model for lexical analysis. By linearly combining lexical features with acoustic features, they achieved high F-measures when using Support Vector Machine (SVM). Since a bag-of-words model is a naive model, there may be more valuable lexical information that it cannot capture. Moreover, as lexical and acoustic features are extracted from different domains, a single layer linear combination may not yield the optimal results.

In 2010, Interspeech launched a Paralinguistic Challenge (Schuller et al., 2010) that included the task of detecting LOI from speech as a subchallenge. Competitors were given conversational speech corpora with annotated LOI, baseline acoustic features, and two baseline results. The evaluation metric used for the challenge was primarily the cross correlation [2] (CC) measure (Grimm et al., 2008), with mean linear error [3] (MLE) also taken into consideration. The baseline was built only on acoustic features, 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.

Gajsek et al. (2010) participated in this challenge and proposed the use of Gaussian Mixture Models as Universal Background Model (GMM-UBM) with relevance MAP estimation for the acoustic data. This is based on the success of GMM-UBM modeling in the speaker identification tasks (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 the testing condition. This performance 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) approach won the 2010 Subchallenge for this task. In addition to the baseline acoustic features provided, they used term frequency and a subjectivity dictionary to mine the lexical information. In addition to a linear combination of all lexical and acoustic features, they designed a hierarchical regression framework with multiple level of combinations. Its first two combiners tackle the prediction problems from different acoustic classifiers and then uses a final stage SVM classifier to combine the overall acoustic predictions with lexical predictions 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.

3 Our System

Unlike previous approaches, we emphasize lexical modeling, to counter problems of speaker variation in acoustic features (Jeon et al., 2010). We propose an improved version of standard TFIDF (Spärck Jones, 1972) — Discriminative TFIDF — which computes the IDF score of the target word by discriminating its different mean LOI score tags during training to produce more informative keyword spotting in testing.

In addition to Discriminative TFIDF, we utilize the Dictionary of Affect in Language (DAL) (Whissell, 1989) to detect lexical affect and compute an utterance-level affect score. To maximize the coverage of lexical cues, we also train trigram language models on the training data to capture contextual information and use the test output log likelihoods and perplexities as features. Besides these lexical features and the 1582 baseline acoustic features from the Interspeech Paralinguistic Challenge, we extract 32 additional prosodic and voice quality features using Praat (Boersma, 2001). In order to model sentence-level prosodic events, we use Au-

[2] Pearson product-moment correlation coefficient is a measure of the linear dependence that is widely used in regression settings.

[3] MLE is a regression performance measure for the mean absolute error between an estimator and the true value.

ToBI (Rosenberg, 2010) to extract pitch accent and phrase-based features. These features are described in detail in Section 3.2.

The simplest approach to classification is to include all features in a single classifier. However, different features streams include different number of features, extracted and represented in different domains. The Sum Rule approach (Kittler et al., 1998) is an early solution to this classifier combination problem. Instead, we train 1st-tier classifiers for each of the feature streams and then train a 2nd-tier classifier to weight the posterior predictions of the 1st-tier classifiers. We further improve this method by integrating a novel model which considers the 1st-tier multistream prediction feedback from the seed samples and uses a mean cosine similarity method to measure the distance between a new instance and prominent seed samples. We use this similarity measure to improve classification.

3.1 Corpus

The corpus we use in our experiments is the 2010 Paralinguistic Challenge Affect Subchallenge corpus Technische Universtät Munchën Audiovisual Interest Corpus (TUM AVIC), provided by Schuller (2010). The corpus consists of 10 hours of audiovisual 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 fluent. 11 subjects were younger than 30; 7 were between 30-40; 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 2sec.

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 experimens.

3.2 Feature Sets

Table 1 provides an overview of the feature sets in our system.

Discriminative TFIDF

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

$$V(t,u) = TF * IDF = \frac{C(t,u)}{C(v,u)} * \log \frac{|U|}{\sum u(t)}$$

TF is calculated by dividing the number of occurrences of term *t* in the utterance *u* by the total number of tokens *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 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 word by the distribution of its labels in the training set. This helps us to weight words by the LOI of the utterances they are uttered in. An intuitive example is this: Although the words "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

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)
Pulses	# 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, Mean NHR, Mean NHR (dB)
Duration	Seconds
Fundamental Frequency	Min, Max, Mean, Median, SDev, MAS
Energy	Min, Max, Mean, SDev
Prosodic Events	Pitch accents, intermediate phrase, and intonational boundaries.

Table 1: **Feature Sets**. *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.*

will give these two terms the same score. To differentiate the importance of these two words, we define our Discriminative TFIDF measure as follow:

$$V(t,u) = \frac{C(t,u)}{C(v,u)} * \log \frac{|U|}{\sum u(t) * (1 - |MeanLOI|)}$$

Here, the Mean LOI score ranging from (0,1) is the label of each utterance. Instead of summing the *u(t)* scores directly, we now assign a weight to each utterance. The weight is $(1 - |MeanLOI|)$ in our task. The overall IDF score of words 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. Discriminative TFIDF can be viewed as a generalized version of Delta TFIDF (Martineau and Finin, 2009) that can be used in various regression settings.

Wang and McKeown (2010) show that adding Part-of-Speech (POS) information to text can be helpful in similar classification tasks. So we have also used the Stanford POS tagger (Toutanova and Manning, 2000) to tag these transcripts before calculating the Discriminative TFIDF score.

Lexical Affect Scoring

Whissell's Dictionary of Affect in Language (DAL) (Whissell, 1989) attempts to quantify emotional language by asking raters to judge 8742 words collected from various sources including college essays, interviews, and teenagers descriptions of their own emotional state. Its 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. A previous study (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. To calculate an utterance's lexical affect score, we first remove the stopwords and then sum up [4] the EE score of each word in the utterance.

Statistical Language Modeling

In order to capture the contextual information and maximize the use of lexical information, we also train a statistical language model 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). In the testing stage, the log likelihood and perplexity scores are used as language modeling features. Due to the data sparsity issue, we are not able to train language models on subsets of training data that correspond to different LOI scores.

[4] We have experimented with Min, Max and Mean scores, but the results were poor.

Acoustic, Prosodic and Voice Quality Features

As noted above, the TUM AVIC corpus includes acoustic features (Schuller et al., 2010) for all of the data sets. These include: 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 Difference of Difference of Periods, and Shimmer local. We have extracted an additional 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).

Prosodic Event Features

To examine the contribution of higher-level prosodic events, we have also experimented with AuToBI (Rosenberg, 2010) to automatically detect pitch accents, word boundaries, intermediate phrase boundaries, and intonational boundaries in utterances. AuToBI requires annotated word boundary information; since we do not have hand-annotated boundaries, we use the Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008) to align each utterance with its transcription. We use AuToBI's models, which were trained on the spontaneous speech Boston Directions Corpus (BDC) (Hirschberg and Nakatani, 1996), to identify prosodic events in our corpus.

3.3 Fusion Learning Approaches

Assuming that our various lexical, acoustic and prosodic feature streams are informative to some extent when tested separately, we want to combine information from the streams in different domains to improve prediction. We experimented with several approaches, including Bag-of-Features, Sum Rule combination, Hierarchical Fusion, and a new approach. We 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. A potential problem with this method is that, when combining 1582 acoustic features with 10 lexical features, the classifier will treat them equally, so potentially more useful lexical features will not be evaluated properly. A second problem is that our features are extracted from different domains using different methods, and normalization across domains is not possible in a bag-of-features classification/regression approach. Another possible approach is the Sum Rule Combiner, which uses product or sum rules to combine the predictions from 1st-tier classifiers. Kittler et al. (1998) show that the Sum Rule approach outperforms the product rule, max rule and mean rule approaches when combining classifiers. Their sensitivity analysis shows that this approach is most resilient to estimation errors.

A third method of combining features is the Hierarchical Fusion approach of fusing multistream information, which involves multiple classifiers and performs classification/regression in multiple stages. This can be implemented by first training 1st-tier classifiers for each single stream of features, collecting the predictions from these classifiers, and training a 2nd-tier supervector classifier to weight the importance of predictions from the different streams and make a final prediction. The rationale behind this approach is to solve the cross-domain issue by letting the 2nd-tier classifier weight the streams, as the predictions from 1st-tier classifiers will be in a unified/normalized form (e.g. 0 to 1 in this task).

The Multistream Prediction Feedback and Mean Cosine Similarity based Hierarchical Fusion

Our Multistream Prediction Feedback and Mean Cosine Similarity based Hierarchical Fusion approach combines a similarity based two-stage approach with a multistream feedback approach. Figure 1 shows the architecture of this system. It is based on the intuition that, if we can identify the prominent samples (e.g. the samples that all 1st-tier classifiers assign high average prediction scores), then we can measure the average distance between a new sample and all these prominent samples in the Euclidean Space. Furthermore, we can use this average distance (average similarity) as a new feature to improve the 2nd-tier classifier's final prediction.

To implement this process, we first train five 1st-tier Additive Logistic Regression (Friedman et al., 2000) classifiers and a Random Subspace meta learning (Ho, 1998) 1st-tier classifier (for the acoustic stream), using six different feature streams in our training data. In the testing stage, we use a random subset of the test set as seed samples. Next, we run the seed samples for each of these 1st-tier classifiers

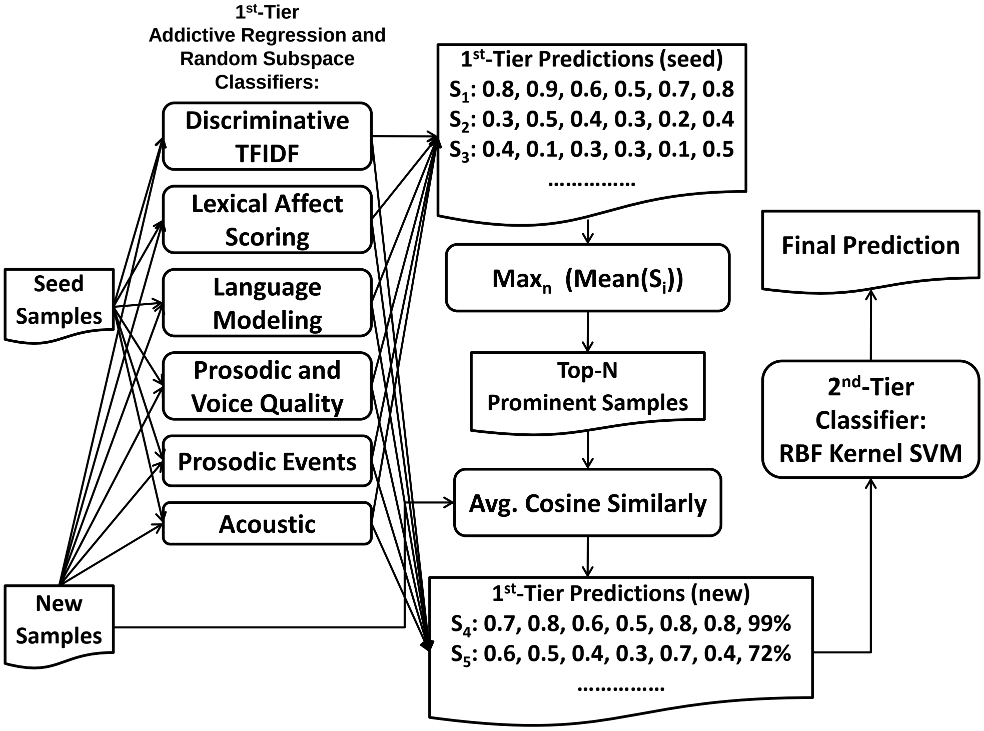

Figure 1: **The Overview of Multistream Prediction Feedback and Mean Cosine Similarity based Hierarchical Fusion Learning**

to obtain prediction scores ranging from 0 to 1. Now, we take the mean of these predicted scores for each sample, and use the following method to select the top *n* samples from the seed samples *S* as "prominent samples":

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

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

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

where "·" indicates 'dot product'. Now, given our hypothesized prominent samples, for each of these samples and new samples, we choose the original Discriminative TFIDF, Lexical Affect Scoring, Language Modeling, Prosodic and Voice Quality, and Prosodic Event features as *k* vectors to represent all the samples in Euclidean Space. The reason we drop the acoustic features from the vector space model is because of the dimensionality issue — 1582 acoustic features. We substitute our 32 standard 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:

$$Sim(U_n, U_S) = Mean\left(\frac{\sum_{i=1}^{k} V_n * V_s}{\sqrt{\sum_{i=1}^{k} V_n^2} * \sqrt{\sum_{i=1}^{k} V_s^2}}\right)$$

In the next step, we add this mean cosine similarity measure as a new feature and include it in the 2nd-tier classifier for reclassification. Now, in the reclassification stage, all 1st-tier feature stream predictions will be re-weighted by the new 2nd-tier classifier that incorporated with Multistream Feedback information.

The reason why the Multistream Prediction Feedback is useful in this task is that, 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. So, by extracting more knowledge about the lexical, prosodic, and acoustic features distributions in test set using our novel approach, we will have a better understanding about the skewed distributions in the test set. In addition, our approach is inexpensive and does not require extra unlabeled data.

4 Experiments and Results

We conduct our experiments in three parts. First, we examine how well the Discriminative TFIDF feature performs, compared with standard TFIDF feature. Secondly, we look at how different feature sets influence our results. For the first two parts, we evaluate our features using the Subchallenge training vs. development sets only. Finally, we compare our similarity based multistream fusion feedback approach to other feature-combining approaches. We examine our final system first comparing training vs. development performance, and then combined training and development sets vs. the test set. WEKA (Witten and Frank, 2005) and LIBSVM (Chang and Lin, 2001) are used for regression.

4.1 TFIDF v.s. Discriminative TFIDF

Method	CC	MLE
TFIDF	0.296	0.142
D-TFIDF	0.368	0.140
S-D-TFIDF	**0.381**	**0.136**

Table 2: **Single TFIDF Feature Stream Single Regression Results** *(Train vs. Develop, Additive Logistic Regression). D-TFIDF: Discriminative TFIDF. S-D-TFIDF: the POS tagged version of D-TFIDF. CC: Cross Correlation. MLE: Mean Linear Error.*

When working with the training and development sets, we are able to access the label and transcriptions of each set to calculate the Discriminative TFIDF scores. For the testing scenario discussed in in Section 4.3, we do not have these annotations. So, we redefine the task as a keyword spotting task, where we can use the identified keywords in the training and development sets as keyword features in testing. We also sum up the word-level TFIDF scores and use the sentence-level TFIDF as a single feature for the classification experiment. The regression algorithm we use is Additive Logistic Regression with 50 iterations. Table 2 shows how different approaches perform in the experiment. 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. In contrast, our Discriminative TFIDF has stronger discriminative power and picks a total number of **59** truly rare terms as top-1 level keywords.

4.2 Regression with Different Feature Streams

Table 3 shows performance using different feature streams in our system. We see that the acoustic and prosodic features are the dominating features in this task. The Prosodic Events feature stream also emerges as a new informative high-level prosodic feature in this task.

Feature Streams	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**

Table 3: **Comparing Contributions of Different Feature Streams in the 2nd-tier Classifier** *(Training vs. Developmen, Random Subspace for the 1st-tier classifier of Acoustic Stream, and Additive Logistic Regression for other 1st-tier classifiers. Radial Basis Function (RBF) Kernel SVM as 2nd-tier Classifier.) S-D-TFIDF: the POS tagged version of D-TFIDF. VQ: Voice Quality. n: Top-n Feedback. CC: Cross Correlation. MLE: Mean Linear Error.*

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. 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 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.

4.3 Comparing with State-of-the-Art Systems

Table 4 compares our 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 developemen vs. test set result.

Method	CC	MLE
Shuller et al.,(2010)	0.604	0.118
Jeon et al., (2010)	0.622	0.115
Gajsek 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**
Gajsek et al. (2010)	0.390	0.143
Shuller 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**

Table 4: **Comparing Different Systems.** *Above: Training vs. Development. Bottom: Combined Training+ Development vs. Test. CC: Cross Correlation. MLE: Mean Linear Error.*

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 Game Corpus (Hirschberg et al., 2005). The word error rate (WER) is 29% on the development set.

Note that a Bag-of-Features approach combining all features results in poorer performance than the use of acoustic features alone. The Sum Rule approach improves over this method by achieving CC score of 0.422. Although the improvement of CC seems small, it is extremely statistically significant (Paired t-test with two-tailed P-value less than 0.0001), comparing to the Bag-of-features model. However, when using the SVM as the 2nd-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.

Augmenting the SVM hierarchical fusion learning approach with multistream feedback, we observe a significant improvement over all other systems and methods. We obtain a final CC of 0.480 and MLE of 0.131 in the test mode, which is sigificantly different from the Bag-of-features approach (paired t-test p < 0.0001), but does not differ significantly from the SVM hierarchical fusion approach.

5 Conclusion

Detecting **levels of interest** from speakers is an important problem for Spoken Dialog Understanding. While earlier work, done in the 2010 Interspeech Paralinguistic Affect Subchallenge, employing traditional acoustic features and shallow lexical features, achieved good results, our new features — Discriminative TFIDF, lexical affect scoring, language modeling, prosodic event — when used with standard prosodic features and our new Multistream Prediction Feedback and Mean Cosine Similarity heuristic-based Hierarchical Learning method improves over all published results on the LOI corpus. Our method is domain-independent and can be adapted to other speech and language processing areas where domain adaptation is expensive to perform. In the future, we would like to experiment with different distributional similarity measures and bootstrapping strategies.

Acknowledgments

The first author was funded by Kathleen McKeown while conducting the research. We would also like to thank Andrew Rosenberg and three anonymous reviewers for their useful comments.

References

Agarwal, Apoorv and Biadsy, Fadi and Mckeown, Kathleen R. 2009. Contextual Phrase-Level Polarity Analysis Using Lexical Affect Scoring And Syntactic N-Grams. in *EACL 2009*.

Ai, Hua and Litman, Diane J. and Forbes-Riley, Kate and Rotaru, Mihai and Tetreault, Joel and Purandare, Amruta 2006. Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs. in *INTERSPEECH 2006*.

Biadsy, Fadi and Rosenberg, Andrew and Carlson, Rolf and Hirschberg, Julia and Strangert, Eva. 2008. A Cross-Cultural Comparison of American, Palestinian, and Swedish Perception of Charismatic Speech. in *Proceedings of the Speech Prosody 2008*.

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. in *Glot International*.

Chang,Chih-Chung and Lin, Chih-Jen. 2001. LIBSVM: a library for support vector machines. Software available at *www.csie.ntu.edu.tw/˜ cjlin/libsvm*.

Devillers, Laurence and Vidrascu, Laurence 2006. Real-life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs. in *INTERSPEECH 2006*.

Fisher, William M. and Doddington, George R. and Goudie-Marshall, Kathleen M. 1986. The DARPA Speech Recognition Research Database: Specifications and Status. in *DARPA Workshop on Speech Recognition*.

Friedman, Jerome and Hastie, Trevor and Tibshirani, Robert. 2000. Additive logistic regression: a statistical view of boosting. in *Ann. Statist.*.

Gajšek, Rok and Žibert, Janez and Justin, Tadej and Štruc, Vitomir and Vesnicer, Boštjan and Mihelič, France. 2010. Gender and Affect Recognition Based on GMM and GMMUBM modeling with relevance MAP estimation. in *INTERSPEECH 2010*.

Grimm, Michael and Kroschel, Kristian and Narayana, Shrikanth. 2008. The Vera am Mittag German Audio-Visual Emotional Speech Database. in *IEEE ICME*.

Hirschberg, Julia and Nakatani, Christine H. 1996. A prosodic analysis of discourse segments in direction-giving monologues. in *ACL 1996*.

Hirschberg, Julia and Benus, Stefan and Brenier, Jason M. and Enos, Frank and Friedman, Sarah and Gilman, Sarah and Gir, Cynthia and Graciarena, Martin and Kathol, Andreas and Michaelis, Laura. 2005. Distinguishing Deceptive from Non-Deceptive Speech. in *INTERSPEECH 2005*.

Ho, Tin Kam. 1998. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on PAMI*.

Jeon, Je Hun and Xia, Rui and Liu, Yang. 2010. Level of Interest Sensing in Spoken Dialog Using Multi-level Fusion of acoustic and Lexical Evidence. in *INTERSPEECH 2010*.

Kittler, Josef and Hatef, Mohamad and Duin, Robert P. W. and Matas, Jiri. 1998. On combining classifiers. *IEEE Transactions on PAMI*.

Laskowski, Kornel and Burger, Susanne. 2007. Analysis of the Occurrence of Laughter in Meetings. in *INTERSPEECH 2007*.

Liscombe, Jackson and Hirschberg, Julia and Venditti, Jennifer J.. 2005. Detecting Certainness in Spoken Tutorial Dialogues. in *Eurospeech*.

Martineau, Justin and Finin, Tim. 2009. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. in *ICWSM*.

Reynolds, Douglas A. and Quatieri, Thomas F. and Dunn, Robert B. 2000. Speaker verication using adapted gaussian mixture models. in *Digital Signal Processing*.

Rosenberg, Andrew. 2010. AuToBI - A Tool for Automatic ToBI Annotation. in *INTERSPEECH 2010*.

Salto, Gerard 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*.

Schuller, Björn, and Köhler, Niels and Müeller, Ronald and Rigoll, Gerhard. 2006. Recognition of Interest in Human Conversational Speech. in *INTERSPEECH 2006*.

Schuller, Björn, and Steidl, Stefan and Batliner, Anton and Burkhardt, Felix and Devillers, Laurence and Müeller, Christian and Narayanan, Shrikanth. 2010. The INTERSPEECH 2010 Paralinguistic Challenge. in *INTERSPEECH 2010*.

Spärck Jones, Karen. 1972. A statistical interpretation of term specificity and its application in retrieval. in *Journal of Documentation*.

Stolcke, Andreas. 2002. SRILM-an extensible language modeling toolkit. in *ICSLP 2002*.

Toutanova, Kristina and Manning, Christopher D.. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. in *EMNLP/VLC-2000*.

Tao, Tao and Zhai, ChengXiang 2006. Regularized estimation of mixture models for robust pseudo-relevance feedback. in *SIGIR 2006*.

Wang, William Yang and McKeown, Kathleen. 2010. “Got You!”: Automatic Vandalism Detection in Wikipedia with Web-based Shallow Syntactic-Semantic Modeling. in *COLING 2010*.

Wang, Chingning and Zhang, Ping and Choi, Risook and DEredita, Michael. 2002. Understanding consumers attitude toward advertising. in *Eighth Americas conf. on Information System*.

Witten, Ian H. and Frank, Eibe 2005. *Data mining: Practical machine learning tools and techniques, 2nd Edition*. San Francisco: Morgan Kaufmann.

Whissell, Cynthia. 1989. The Dictionary of Affect in Language. in *R. Plutchik and H. Kellerman, Editors, Emotion: Theory Research and Experience*.

Yuan, Jiahong and Liberman, Mark. 2008. Speaker identification on the SCOTUS corpus. in *Proceedings of acoustics ’08*.