Published in final edited form as:
Comput Speech Lang. 2015 January 1; 29(1): 203–217. doi:10.1016/j.csl.2014.04.002.

Acoustic and Lexical Representations for Affect Prediction in Spontaneous Conversations

Houwei Cao[a,*], **Arman Savran**[a], **Ragini Verma**[a], and **Ani Nenkova**[b]

[a]Department of Radiology, Section of Biomedical Image Analysis, University of Pennsylvania, 3600 Market street, Suite 380, Philadelpia, PA 19104

[b]Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut street, Philadelphia, PA 19104

Abstract

In this article we investigate what representations of acoustics and word usage are most suitable for predicting dimensions of affect|AROUSAL, VALANCE, POWER and EXPECTANCY|in spontaneous interactions. Our experiments are based on the AVEC 2012 challenge dataset. For lexical representations, we compare corpus-independent features based on psychological word norms of emotional dimensions, as well as corpus-dependent representations. We find that corpus-dependent bag of words approach with mutual information between word and emotion dimensions is by far the best representation. For the analysis of acoustics, we zero in on the question of granularity. We confirm on our corpus that utterance-level features are more predictive than word-level features. Further, we study more detailed representations in which the utterance is divided into regions of interest (ROI), each with separate representation. We introduce two ROI representations, which significantly outperform less informed approaches. In addition we show that acoustic models of emotion can be improved considerably by taking into account annotator agreement and training the model on smaller but reliable dataset. Finally we discuss the potential for improving prediction by combining the lexical and acoustic modalities. Simple fusion methods do not lead to consistent improvements over lexical classifiers alone but improve over acoustic models.

Keywords

emotion; affect; spontaneous speech; lexical features; acoustics

1. Introduction

In this article we set out to compare a variety of representations of lexical usage and acoustics for dimensional affect analysis in spontaneous speech. The goal of our work is to establish which of these representations are most suitable for predicting individual affective dimensions and to what extent they can combine to further improve prediction.

*Corresponding author: Houwei.Cao@uphs.upenn.edu (Houwei Cao).

Our experiments are carried out on the AVEC (audio-visual emotion recognition grand challenge) 2012 shared task (Schuller et al., 2012) for continuous prediction of multi-dimensional affective states from non-segmented spontaneous speech. The task is to recognize the affect dimensions of (AROUSAL, EXPECTNCY, POWER, VALENCE) continuously on the multimodal SEMAINE database (McKeown et al., 2012) of naturalistic video and audio of human-agent interactions, in terms of audio analysis, video analysis, or combination of them.

In this article we describe in detail and further expand the work on lexical and acoustic analysis which was part of our official submission to the AVEC 2012 challenge, achieving the best results on the word-level prediction competition (Savran et al., 2012). We use the same training and test corpus definitions as the AVEC 2012 challenge; the results are therefore comparable with the challenge benchmarks and with results published by other particpants.

For lexical analysis, we rely on manual speech transcripts to extract features. We experiment with a domain-independent resource, the ANEW (Affective Norms of English Words) dictionary Bradley and Lang (2010), which provides human ratings of words for the AROUSAL, VALENCE, and DOMINANCE dimensions. We also test domain-dependent techniques which exploit the mutual information between words and the opposing ends of dimensional scales. Such information has been successfully used in prior work on natural emotional speech, but in non-sparse representations. We demonstrate that sparse representation using mutual information between words and dimension as weights is more powerful for the task at hand.

For acoustic analysis, we use forced alignment between the audio and the manual transcripts to introduce a novel representation which proves to be highly advantageous for capturing affect-related cues in the voice. We predict emotion on the utterance level, but introduce the notion of regions of interest and compute the long-term statistics of acoustic features of frames that fall in different regions separately. The main motivation is to capture the changes that occur in the regions of speech corresponding to lexical stress. These representations better capture the intuition that affective cues can be expressed to a greater extent in some regions than others, and, thus, increase the discriminating power of acoustic features. In addition, we address a question related to how affect should be modeled which has never been studied before. We show that by training classifiers on acoustic information only from the regions for which annotators agree on the affect label results in a much better model. Prior work has shown that training and testing on data for which annotators agree leads to higher performance (Litman and Forbes-Riley, 2006). We in contrast show that we can improve the results on the full test set, as originally labeled either reliably or unreliably, by using only reliably annotated data in training. Our results contribute strong evidence that automatic prediction of emotion benefits from carefully motivated representations.

Finally, we present experiments with simple decision-level fusion techniques to combine the two modalities. Many studies have discussed the advantage of exploiting jointly acoustic and lexical information, particularly in the prediction of the VALENCE dimension, i.e. classification of negative versus non-negative utterances (Lee and Narayanan, 2005a),

discriminating anger utterance from neutral (Batliner et al., 2000) or other emotional speech (Schuller et al., 2004; Polzehl et al., 2011). We show that while decision-level fusion works better than feature-level fusion, simple fusion techniques do not lead to consistent improvements over the stronger lexical modality.

2. Affect Database

We used the AVEC 2012 Grand Challenge dataset, which is a subset of free publicly available SEMAINE corpus (McKeown et al., 2012). It contains multi-modal recordings and provides video, audio, and manual speech transcripts of dyadic communications between people and virtual characters role-played by human operators. Each of the four virtual characters has a preselected emotionally stereotyped behavior: *Prudence* is even-tempered and sensible, *Poppy* is happy and outgoing, *Spike* is angry and confrontational, and *Obadiah* is sad and depressive. In our work we do not use the identity of the virtual agents for prediction of the other interlocutor's affective state.[1]

The challenge dataset contains 95 sessions of data from 24 recordings. Most recordings consist of four sessions of conversations of the same person with all four virtual characters. These are split into three non-overlapping partitions of training, development and testing subsets. Each subset consists of about 8 recordings from 8 different users. The summary of the dataset is given in Table 1.

All sessions in the AVEC dataset were annotated by two to eight annotators (with the majority annotated by six raters) for four dimensions of affect: AROUSAL, EXPECTANCY, POWER, and VALENCE. These dimensions were identified as the most important ones for distinguishing among a range of emotional states (Fontaine et al., 2007). The dimension of AROUSAL can be described as overall inclination of activation ranging from calming to exciting; the dimension of EXPECTANCY, also known as anticipation, ranges from unpredictable to familiar; the dimension of POWER can be interpreted in terms of potency-control related to the feeling of power or weakness; and the dimension of VALENCE refers to the pleasantness which ranges from highly positive to highly negative.

Each rater annotated the above four dimensions continuously, with continuous value on every frame (20 ms) by using the FeelTrace tool (Cowie et al., 2000). The ground-truth of the frame-level challenge labels were obtained by taking the average of the annotations over all raters. In addition, word-level labels are also provided by calculating the mean value of the frame-level labels on the entire word. The experiments we present in this paper are evaluated on the word-level affect prediction task.

In our work we aim to develop accurate models of vocal expressions of emotional states. We hypothesize that a way to achieve this goal is to rely only on data labeled with high agreement, thus reliably linking audio features with perceptually recognizable changes in expressed affect. For this purpose here we further analyze the inter-rater agreement of multiple labelers using the measurement of Cronbach's alpha (George and Mallery, 2003). Table 2 gives the distribution of inter-rater agreement on the training and development

[1]Other studies have productively exploited this information (Soladié et al., 2012).

partitions of the challenge data. We group these according to different alpha levels of unacceptable (alpha < 0.5), acceptable (alpha < 0.7), and good (alpha > 0.8). Alpha depends on the number of raters, so we list this information as well. We do not give information about the test partition in this analysis, since the individual labels of different raters on that part of the data are still not publicly available. As we will show later in Section 5.2, using only the reliably labelled data in training markedly improves performance on the full test set with original annotations of varying reliability.

In general, raters have higher agreement on POWER and VALENCE, lower on AROUSAL and EXPECTANCY. In addition, the ground-truth labels tends to be more reliable as the number of raters increases.

3. Lexical Representation

In this section we analyze lexical content, as given by the manual transcripts and its association with levels of affect dimensions AROUSAL, EXPECTATION, POWER and VALENCE. We study a variety of representations, including domain-specific statistics and domain-independent psychological norms for three of the affect dimensions.

3.1. PMI features

We first consider pointwise mutual information (PMI) (Fano, 1961), which gives an insight about affective connotations of different words. Specifically, we calculate the PMI between a word and a given affect dimension. In prior work PMI has been successfully applied for categorical emotion recognition between anger and non-anger utterances (Lee and Narayanan, 2005b). Later work showed that PMI representations of lexical content does not have much advantage compared to bag of word approaches in three different corpora (Polzehl et al., 2011). However, both of these studies derive a single feature from PMI, corresponding to a weighted sum for all words in the utterance. In contrast in most semantic processing applications, PMI is used for feature selection and as weights for individual words in sparse bag of words representations (Turney and Pantel, 2010). Here we compare the non-sparse representations used in prior work on emotion recognition with the sparse, PMI-weighted representation, and show that the latter is considerably more powerful.

Computing PMI between a word w and a dimension of affect ε is straightforward when affect is coded as binary presence or absence of a given property. In the AVEC 2012 data, however, AROUSAL, EXPECTATION, POWER and VALENCE are coded as continuous variables. We transform these into binary labels in order to compute PMI. To do this, we computed the average value of each affect dimension over all words in the dataset. Then the binary labels are assigned depending on whether the continuous label for a word is above or below the overall average. Class 1 is assigned if the original value is above the sample average and class 0 is assigned if the original value is below the average. We calculate mutual information on the combined training and development set. Having a larger set for computing such co-occurrance statistics is important, especially for lexical analysis, in order to mitigate to some extent the expected issues of unseen (out of vocabulary) words and low counts.

Now the PMI between a word and a binary emotion dimension can be calculated as

$$PMI(\varepsilon, w) = \log \frac{P(\varepsilon, w)}{P(\varepsilon)P(w)} = \log \frac{P(\varepsilon|w)}{P(\varepsilon)} \quad (1)$$

$P(\varepsilon)$ is the prior probability of an affect dimension and $P(\varepsilon|w)$ is the conditional probability of the affect dimension given the word w. Both probabilities are computed directly from counts on the combined training and development data.

For each word w in the training set and for each affect dimension (AROUSAL, EXPECTATION, POWER and VALENCE) we compute two PMI values: one for association between the word and class 0, which corresponds to low values for the affect dimension, and class 1, which corresponds to high values. Table 3 lists examples of the ten words with the highest PMI that are associated with low (class 0) and high (class 1) values of the affect dimensions. For instance, “bloody”, “anxious”, and “fighting” carry strong negative connotation, while “interesting”, “holiday”, and “love” are typical positive words, based on the analysis of PMI on the dimension of VALENCE on the AVEC dataset. The words associated with the other dimensions are not that easily interpretable. Some of the words associated with low arousal, such as “anxious”, “inside” and “rain” can intuitively be associated with low inclination for physical activity but others such as “language” and “three” are rather opaque. Regardless of the lack of intuitive interpretation for the PMI features, we will see in later experiments that they lead to the best prediction on the test data.

PMI-based features can be used for either word-level or turn-level represenations:

WL PMI: Word level PMI. This representation consist of two components, each equal to the PMI that the word has with the low (0) and high (1) classes of the dimension we wish to predict.

TL PMI: Turn level PMI. Similarly to the word level PMI, the turn-level PMI representation also consists of only two features. They are computed as the average of word-level PMI of all words in the turn. In testing, the values for each turn are calculated by taking the average only for words that appeared in training, ignoring any other word.

For the word-level representation, any unseen words will be smoothed by the feature vector of their preceding words. For the turn-level representation, if a turn in testing consist entirely of words that did not appear in training, we take the values from the preceding turn as features. Compared with word level PMI, the turn level representation may help produce more robust prediction and partially overcome the out-of-vocabulary (OOV) problems caused by the appearance of words in the test data that did not occur in the training data.

3.2. Sparse lexical features

The turn level sparse lexical features are inspired by conventional bag of words representations in which texts are represented as sparse vectors of occurrence counts of words from a predefined vocabulary. Here, we investigate four lexical representations with various sparse feature spaces with different vocabulary.

Sparse BOW: Bag-of-word features. The feature space is defined by a selected number of words. The representation of utterance has value zero for words not in the utterance, 1 for words that do appear in it.

Sparse PMI: Sparse lexical representation with PMI selected words. This representation is similar to the BOW one, however the value of the component corresponding to a particular word is equal to the PMI between the word and the affect dimension that is being predicted.

Sparse ANEW: Sparse lexical representation with words in the ANEW dictionary. The values are the norms from the dictionary, equal to the average rating of multiple people, indicating the extent to which they associate the work with a given dimension.

TF.IDF: Term Frequency–Inverse Document Frequency features. The feature space is defined by all words in the training data. The values of components are determined by the term frequency–normalized word counts. The inverse document frequency for words is determined by the number of all conversations and the number of conversations that contain the particular word.

Bag-of-words (BOW) is one of the commonly used feature in information retrieval and text classification tasks. For the feature space in this representation we consider only the words which appear at least three times in the training data; there are 1,048 such words. Words occurring fewer times are discarded. Each turn is represented by a vector of length 1,048, each component corresponding to one of the words. The value of the component is 1 if the corresponding word occurs in the speaker turn and 0 for if the word does not appear in the utterance.

In **Sparse PMI**, we use a different set of 1,000 words for each affect dimension. These are the 500 words with highest PMI for class 0 and class 1 respectively. Each turn is represented by 1,000 features with the corresponding PMI values for words that occur in the speaker turn and zeros for words which do not appear in the utterance. Table 4 compares the pre-defined vocabularies in **Sparse BOW** and **Sparse PMI** representations on the four affective dimensions. It is clear that more than 50% of the words are different in these representations, in all affective dimension except for AROUSAL.

Sparse ANEW employs the vocabulary of the Affective Norms for English Words (ANEW) (Bradley and Lang, 2010), which provides a set of normative emotional ratings for a large number of words in English, in the affective dimensions of AROUSAL, POWER, and VALENCE. In ANEW, the level of affective connotation of about 2,500 words were defined in terms of perceptual test in a psychology study. As Table 5 reveals, the association between words and affect dimensions are much more easily interpretable in the ANEW dictionary than in the dictionary derived using PMI. In that table we list the words with highest and lowest average human rating for the three dimensions covered in ANEW.

However only 564 of the words in normed in the ANEW dictionary appeared in the AVEC training data set. Consequently, in the **Sparse ANEW** representation, each turn is represented by 564 features with the normative emotional ratings values for words that occur in the speaker turn and zeros for words which do not appear in the utterances.

TF.IDF is another conventional representation used in information retrieval tasks. In this study, we build the feature space by considering all 2,992 words that appeared in the training dataset; each turn is represented by a vector of length 2,992.

4. Audio Representation

In the present study, the super-segmental approach is applied to extract fixed-length feature vectors for each speech chunk. We use the openSMILE toolkit (Eyben et al., 2010) to obtain a comprehensive set of acoustic features in terms of low–level descriptors (LLD) and their various statistics over the turn. In Table 6 we list the 26 prosodic and spectral LLDs we considered in this study. For each of these, the 19 functional statistics listed in the second column are also computed, as well as the first order delta coefficients.

We study the predictive power of two types of acoustic features, one from standard representations at different granularity and one from regions of interest (ROI) that have different potentials for expressive prosody. The first type corresponds to traditional super-segmental acoustic features composed of various statistics of acoustic parameters (low–level descriptors) computed over the entire speech chunks. The second type are ROI acoustic features, for which we first divide the whole speech chunk into sub-segments in terms of distinct phonemes or word classes. Then all chunks corresponding to a region category are concatenated together and acoustic features are extracted from that new concatenated acoustic sequence. The feature vectors for all ROIs are concatenated to obtain the representation of the entire utterance.

4.1. Traditional super-segmental acoustic features

We first consider the conventional super-segmental acoustic analysis on speech chunks with different granularity: word-level and turn-level. In word-level analysis, we extract acoustic features from the speech signal descriptors of each word and the prediction is also performed on every single word.

In the turn-level analysis, features are calculated to characterize the emotion content in turns and affect dimensions are estimated for each turn. This decision is motivated by the assumption that any changes in audio features within the same turn creates an effect so that the utterance is perceived as conveying particular affective states, but it is the same mix of emotions conveyed in the entire utterance.[2] In the AVEC challenge data, speaker turns are short, 4–5 seconds on average, so it is reasonable to expect that they have consistent emotion profiles. To evaluate word-level estimation, we simply assign the estimation obtained for the turn as the value for all constituent words.

4.2. Class-based acoustic features

We analyze the behavior of two ROI acoustic representations which we describe in this section.

[2]Work on anger recognition has also shown that individual words can be perceived as angry when heard by listeners in isolation, but the full utterance sounds neutral, indicating the need for utterance-level context even for people (Cauldwell, 2000).

Phoneme ROI: acoustic features captured on distinct phoneme classes. Here features are computed separately for regions of the utterances corresponding to consonants, stressed vowels and unstressed vowels.

Accent ROI: acoustic features captured on distinct classes of words. Here we characterize words according to whether they are typically accented or deaccented in conversational speech.

In the traditional turn-level representation, features are extracted from 20ms frames and values for each feature are summarized as functional statistics of all frames in the utterance. In class-level representation, features are computes over different type of phone/word classes that correspond to a particular ROI. In the ideal situation, we will have precise manual segmentation of smaller acoustic units, i.e., phonemes, words, which respect to every particular ROI. In this study, automatic forced alignment is applied to segment this large portion of speech data into words and phonemes.

In the **Phoneme ROI** analysis, we first perform Viterbi-based forced alignment between the manual transcript and the audio to detect the start time and end time of each phoneme, as well as the presence of lexical stress for each vowel in the speech data (Young et al., 2006). After that, for each spoken turn, we divide the entire speaker turn into three new regions associated with the three distinct phoneme classes of stressed vowels, unstressed vowels and consonants, by concatenating phonemes into these three phoneme classes accordingly. Finally, **Phoneme ROI** features can be extracted by computing statistics of spectral measurements from these new acoustic signals of concatenated phonemes corresponding to different phoneme classes.

In prior work on categorical emotion estimation, Phone ROI spectral features have proven to be complementary to turn-based features, with superior performance as a standalone class (Bitouk et al., 2010; Lee et al., 2004). In the work presented here, we analyze the effectiveness of ROI spectral features for the estimation of continuous emotion dimensions. Our **Phoneme ROI** features are the concatenated features of turn-level prosodic and class-level spectral features. Their performance will be discussed in Session 5.2.

Accent ROI features have not been studied in emotion prediction before our work. In the **Phoneme ROI** representation we identify through forced alignment if a vowel is stressed or unstressed, as realized in the specific utterance. In the **Accent ROI** instead we use information about entire words, i.e. if the words tend to be realized with a pitch accent in spoken utterances or not. Specifically we use the words that tend to be deaccented in conversational speech as one class and all other words fall in the second class. The list of words that are deaccented more often then expected by chance in a sample of conversational speech (Godfrey et al., 1992) are taken from the accent ratio dictionary (Nenkova et al., 2007) created for pitch accent prediction. Words that tend to be deaccented include words from different parts of speech such as “would”, “me”, “then”, “them”, “though”, “make”, “just”, “said”, “much”, “has”, “way”, “when”, “where”, “take” etc. In expressive and affective speech the realization of such words may change. For example words that are normally deaccented may be accented. In sad or fearful speech the opposite phenomenon

may occur. By separating the two ROIs we remove the noise in acoustic variation related to linguistic phrase realization and are more likely to capture affect-related changes.

Similarly to what we performed in the **Phoneme ROI** analysis, we group words into two classes accordingly. **Accent ROI** features are extracted by computing statistics of spectral and prosody measurements from the concatenated parts of the spoken turns corresponding to the two word classes.

The performance of forced alignment is important for the two ROI acoustic representations we introduced. If the alignment is not accurate, it may generate odd artifact on the phoneme/word boundaries. However, for most of the case, we believe that the alignment is reliable, and it would not seriously affect the accuracy of the following frame-level feature extraction (i.e., extraction of MFCC) on the new speech signal with concatenate phonemes/words. The other issue, even with perfect segmentation is the discontinuity of representation at the concatenation points. For some features, like pitch and intensity, calculating summary statistics and shape over the concatenated region is informative. For example changes of pitch only on the segments that bear pitch accent actually abstracts away the variation of change in pitch that is required for the realization of the utterance. For other features some of the shape-related feature (slope, curvature) would be less meaningful. So in the experiments presented here we calculate only spectral features for the ROIs and compute prosodic features on the entire turn.[3]

5. Continuous Affect Recognition Experiments

Two sub-challenges are addressed in the AVEC 2012 grand challenge: the fully continuous sub-challenge (FCSC) and the word-level sub-challenge (WLSC). The FCSC involves fully continuous affect recognition, where the level of affect has to be predicted at frame-level, for every moment of the recording no matter if the user is speaking or listening. The WLSC focuses on the prediction of the level of affective dimensions at word level, only when the user is speaking. In the present study, since we are interested in how levels of affective states are associated with the speech-related features, we only consider the task of the WLSC in the voice-based scenario.

We address the task of automatic continuous affect recognition as a regression problem. Table 7 summarizes different acoustic and lexical representations we introduced in the previous section. For each type of features, we use Support Vector Regression (SVR) with linear kernel for regression. In order to increase the generality of the regression models, we make use of all available data by combining the training and development partitions together in the training phrase. For each set of regression models, we first optimized the SVR parameters in terms of cross-validation with Leave-One-Subject-Out (LOSO) paradigm. Then the SVR regressors are trained on the combined set with the optimized parameters. The performance of the regression models are evaluated in terms of the cross-correlation between the ground-truth labels and the predictions, on the same testing partition as defined in the AVEC challenge.

[3]In-depth feature selection and analysis studies can be performed to better understand the behavior of the ROI representations but these are outside the scope of the current article.

5.1. Affect perdition via lexical representation

Table 8 compares the correlation performances of various lexical indicators on the test partition. We can see that the correlation scores on the **WL PMI** is the lowest among all lexical models, for every affective dimension. The correlation score is extremely low on all dimensions except for VALENCE. Single word information has low predictive power for affective states. This result is consistent with what we observed in Table 3, where the example word do not show interpretable association with affect dimensions except for VALENCE.

We can obtain remarkable improvement with the **TL PMI** models which take into account turn-level mutual information contributed by several words in the spoken turn. We further observe substantial improvements from the sparse lexical models, except for the **Sparse ANEW**. Compared with averaged correlation coefficient of 0.091 of the **TL PMI**, the high dimensional **Sparse BOW**, **Sparse PMI**, and **TF.IDF** obtain about double absolute improvement for almost all affect dimensions. These three representations achieve average scores of 0.182, 0.214, and 0.177 respectively cross all dimensions.

The **Sparse ANEW** representation performs worst among the four sparse lexical models, due to the mismatch of the lexicon defined in the ANEW dictionary and the vocabulary of the AVEC dataset. Only about 25% of the words in the AVEC data had been given normative ratings on the dimension of AROUSAL, POWER, and VALENCE in the ANEW dictionary. As a results, 46.2% of all speaker turns in the AVEC dataset consist entirely of out of vocabulary words which are not in the ANEW dictionary. In these cases, in both training and testing, we take the affect prediction from the preceding turn as the prediction for the current turn, assuming continuity of the emotional state when no evidence for chance has been presented. Consequently, the relatively poor performance of **Sparse ANEW** model may be partially related to the high percentage of turns consisting entirely of out of vocabulary words. Moreover, we notice that the **Sparse ANEW** model achieves much higher correlation scores on POWER and VALENCE than on AROUSAL. It obtains scores of 0.125 and 0.129 on POWER and VALENCE respectively, which is comparable with that of **TL PMI**, despite the huge sparsity issues with ANEW. This suggests that the level of affective connotation of specific words is more robustly related to the affective states of POWER and VALENCE across different corpora, while the level of activation of target words may vary considerably.

Similarly, we also analyze the occurrence of the complete OOV turns in the **Sparse BOW**, **Sparse PMI** and **TF.IDF** representations. The corresponding percentage of the OOV turns in the test partition is shown in Table 8. Since we consider all word occurrences in the training and development partitions for **TF.IDF**, we notice only 1.1% OOV turns in the test partition. Surprisingly, although we only consider the frequent words which appear at least three times for the **Sparse BOW**, we observe similar level of 1.5% OOV turns in the test partition, which is actually lower than we might have expected. On the other hand, we observe noticeably higher rate of OOV turns in the **Sparse PMI** representations. The percentage of OOV turns of **Sparse PMI** also markedly vary for different affective dimensions, from 4.8% of AROUSAL to 23.1% of POWER. This happens because we

selected different words based on the PMI values to construct the lexical feature space for different affective states.

In addition, our results indicate that the **Sparse PMI** features obtain the best continuous recognition performance among all lexical representations, although we observe relatively high OOV rate in this representation. They yield noticeably higher affect recognition accuracy compared to **Sparse BOW** and **TF.IDF** features for most of the affective states, i.e., AROUSAL, EXPECTANCY and POWER. The only exception is the recognition of the affective dimension of VALENCE, where **Sparse BOW** and **TF.IDF** perform better. We believe that this is due to the higher OOV rate of 12.2% observed in **Sparse PMI**.

Finally, in order to evaluate the statistical significance of the difference between the best lexical representations (**Sparse PMI**) and the alternative lexical features that we studied, we perform a non-parametric two sided Wilcoxon sign-rank test (Wilcoxon, 1945). For each of the four affect dimensions, we take the 32 individual correlation scores for each test conversation and we perform a paired test to compare the change of correlation per individual conversation. In addition, we also perform an overall test for significance, considering all affect dimensions simultaneously, combining correlation scores of different affect dimensions for the test. We show the corresponding p-values in Table 9.

It is clear that the **Sparse PMI** features are significantly better than the non-sparse ones, e.g., **WL PMI**, on all affect dimensions. Among all sparse features, the **Sparse PMI** features show different levels of significance in improvements on various affective states. For example, they consistently outperform all other sparse lexical features on POWER while the difference on VALENCE is not significant. Overall, the best **Sparse PMI** features systematically outperform the **Sparse ANEW** and **TF.IDF** with p-value 0.005 and 0.02, while its improvements over **Sparse BOW** is not significant (p-value 0.14).

5.2. Affect perdition via acoustic representation

In Table 10 we compare the affect recognition performance on the test partition of various acoustic features introduced in Session 4. A single prediction is obtained for each utterance, then each word in the utterance is assigned the same affective value. The prediction is evaluated as the average word-level correlation across all sessions.

Consistent with what we observed in the lexical representations, the traditional **Turn-level** super-segmental acoustic features significantly outperform the **Word-level** features: correlations for the **Turn-level** features are more than five times greater than these for **Word level** prediction. Prior work on emotion classification tasks has documented that larger units of analysis are beneficial for emotion classification (Vlasenko et al., 2008). Our results of word-level and turn-level acoustic and lexical features further prove the superiority of the larger units for the tasks of continuous affect recognition. Our experiments with ROI-based features however reveal that specific regions of the turn may still carry more reliable information about the expression of affect.

ROI-based acoustic features provide substantial performance improvements over the conventional **Turn-level** features in most of the cases, with the exception of the recognition

of the affective dimension of AROUSAL with **Accent ROI** acoustic features. For example, the performance gain is as high as 144% for the recognition of the affective dimension POWER and 61% for recognition of affective state EXPECTANCY with **Accent ROI** features respectively in terms of correlation coefficients.

In addition, although we observe slightly higher average correlation scores of 0.136 for **Accent ROI** than the 0.126 achieved by **Phoneme ROI** features, our results also indicate that the **Accent ROI** features do not provide consistent improvement over **Phoneme Class** features on different affective dimensions. Different ROI-based features lead to best performance on each of the four affect dimensions. The **Accent ROI** features are much better than **Phoneme ROI** features for EXPECTANCY, while the **Phoneme ROI** features outperform **Accent ROI** features for AROUSAL. The two types of ROI-based features exhibit relatively similar performance for POWER and VALENCE.

One of the difficult aspects for emotion and affect recognition on spontaneous speech is the uncertainty inherent in the labeling of the data. However, to the best of our knowledge, effects of the uncertainty of the labeling for continuous affect recognition on spontaneous speech have not been explored for the prediction of affect.

We hypothesize that if the human raters are uncertain about the annotations, the model of affect recognition will be affected by the noise and will overfit values that do not correspond to any realistic perceptual correlates of affect. In other words, lower inter-rater agreement in the training set would also lead to lower recognition performance in the testing.

As we discussed in Session 2, many conversation sequences in the AVEC dataset have relatively low inter-rater agreement, particularly in the affective dimension of AROUSAL and EXPECTANCY. In order to investigate how continuous affective states recognition performance is affected by the reliability of labeling, we differentiate reliable data from unreliable ones with uncertain human labels and perform robust training with relatively small portion of reliable data with high inter-rater agreement.

Specifically, for each affect dimension we first analyze the reliability of annotations in terms of the inter-rater agreement on every session. If any session has inter-rater agreement score lower than the threshold of acceptable level ($a < 0.7$), this session will be classified as unreliable data and therefore removed from the training set. If any session obtains inter-rater agreement score higher than the threshold of good ($a > 0.8$), the whole session will be classified as reliable data and therefore kept in the training phrase. Furthermore, for the session with the inter-rater agreement between the level of acceptable and good ($0.7 < a < 0.8$), we further examine the pair-wise correlations of any two individual raters. If most of the pair-wise correlations are on the lower side, the session will be treated as unreliable and excluded from robust training, otherwise it will be included as part of the training data.

For ease of comparison, we also give the affect recognition performance on exactly the same testing partitions, for the two best acoustic representations of **Phoneme ROI** and **Accent ROI** with reliable training, in Table 10. As we expected, compared with the results obtained with the full set of data, we achieve noticeable improvement on both **Phoneme ROI** and

Accent ROI features based on robust training with reliable data. This is a remarkable improvement given that the size of the reliable training data is much smaller than that of the whole set without selection. The largest improvement is in the prediction of the AROUSAL dimesion.

Finally, similarly to what we did for lexical representations, we also perform two sided Wilcoxon sign-rank test for statistical significance between the best performing **Accent ROI reliable** and the other acoustic representations. The p-values are listed in Table 11.

For the traditional super-segmental acoustic features, these results further confirm that turn-level acoustic features are systematically and consistently better than the word-level ones, with highly significant differences between them and our best **Accent ROI reliable** for all affect dimensions. In addition, in overall comparison between turn-level and **Accent ROI reliable** acoustic features proves the advantage of our proposed ROI-based features, with p-value of 0.003. Compared to the standard turn-level representations, the best features perfrom significantly better on EXPECTANCY and POWER. The differences in individual dimensions between the different ROI representations is not significant.

6. Combination of different indications

Now we turn to discuss the combination of acoustic and lexical modalities for continuous affect recognition. The performance of the combined multimodal affect recognition system is likely to be highly associated with the performance of the uni-modal systems, the redundancy among modalities and the effectiveness of the applied fusion methods. We study these issues in this section.

The extent to which different modalities agree with each other or differ in their prediction can be measured by computing the correlation between predictions from different models. Fusing models that differ in their prediction is more likely to lead to improvements in joint multi-modal prediction. In general we expect that different affect recognition models will give similar predictions if they are trained with features that capture similar sources of information. In such case, impressive improvements would not be expected even from powerful fusion techniques over single modalities, for example in combining all the lexical representations, since the prediction of these will be highly correlated.

In order to investigate the relationship between different acoustic and lexical indicators, we compute the cross-correlation between the output predictions obtained from different indicators. Table 12 gives averaged correlation scores (over four affective dimensions) of different lexical and acoustic indicators. Here, we only consider various turn-level acoustic features and sparse lexical representations, since they are the most likely candidates for fusion because they have been proven much better than the word-level acoustic features and two dimensional PMI features respectively. In addition, we do not consider the predictions from **TF.IDF** representations, because we established they are very similar to the **Sparse BOW** results.

Clearly, the predictions from different types of lexical representations are highly correlated with each other, with correlation of 0.660 between **Sparse PMI** and **Sparse BOW**.

Similarly the predictions from the two ROI-based representations have average correlation of 0.701 and that between turn-level acoustic and ROI representations is over 0.6. Correlations between the predictions from different types of indicators|lexical or acoustic|are low, for example 0.240 between **Sparse PMI** and **Phoneme ROI**.

For various sparse lexical representations, as we expected, **Sparse BOW** and **Sparse PMI** are relatively similar to each other, while **Sparse ANEW** is rather different. On the other hand, the correlation scores obtained between the two ROI-based acoustic representations is slightly better than that with conventional turn-level super-segmental acoustic features. The highest correlation of 0.701 is between two different class-based acoustic features, suggesting that these seemingly different representations in fact covey similar information.

Next, we select for the fusion study one of each acoustic and lexical representations by considering the performance of the single representation and the correlation between the two modalities we wish to combine. Compared with **Sparse BOW**, **Sparse PMI** representations show advantage in both better affect recognition performance and lower correlation with acoustic information. On the other hand, **Phoneme ROI** and **Accent ROI** exhibit comparable performance in both respects. Finally, we pick **Sparse BOW** and **Phoneme ROI** for the fusion experiments we describe next. The correlation between the predictions of these two indicators is 0.240, and this relatively low score shows the potential of improvement from fusion.

In Table 13 we compare the performance of early-stage feature fusion and late-stage fusion. In the first, both types of features are combined to train a regression model. We examine two approaches for late-stage: we combine the prediction obtained from acoustic and linguistic indicators with decision fusion by averaging the individual prediction or use the predictions as features in an SVR. The correlations of predictions for each affective dimension between two individual modalities are also listed in the table.

For early-stage feature fusion, we train SVR with linear kernel with one feature vector concatenating acoustic and linguistic features. We perform speaker-independent LOSO cross-validation on the training and development partitions to determine the hyper-parameters for SVR.

Consistent with the finding in Polzehl et al. (2010), we can see that the early-stage feature fusion shows inferior results, while the fusion performs slightly better than the acoustic models but much worst than the lexical models. The two late-stage fusion systems show comparable performance in general. The SVR combination marginally outperforms the averaging fusion methods on average. Compared with the best single modality of **Sparse PMI**, we obtain insignificant improvements on VALENCE and POWER with SVR fusion. On the other hand, there is no change of performance on EXPECTANCY and performance drops on AROUSAL. This may be because the high correlation of 0.479 and negative correlation of −0.010, between acoustic and linguistic modalities, observed on EXPECTANCY and AROUSAL respectively.

7. Discussions

In this paper, we explored various acoustic and lexical representations for predicting AROUSAL, EXPECTANCY, POWER and VALENCE affect dimensions in spontaneous conversations.

In Lee and Narayanan (2005a), *emotional salience* based on mutual information between words and emotion classes has been found useful to classify negative and non-negative emotions in real-world call center dialog application. On the other hand, Polzehl et al. (2011)'s work indicate the advantage of word-level and bi-gram *emotional salience* over standard bag-of-word sparse representations, in distinguishing anger from non-anger utterances on databases with low word-per-turn rate of 2–3 words. However, in our study, we show the superiority of sparse representations in continuous affect recognition in a corpus with longer spoken turns containing 15 words in average. Moreover, we achieve further improvements by jointly using sparse representations and mutual information.

We also notice that the conversation domain poses a challenge for generic characterization of affective content in lexical usage. The corpus-independent sparse lexical representations show relatively poor performance due to the fact that many words in the testing set were not present in the general corpus-independent ANEW dictionary.

We also address the problem of what unit of acoustic analysis is most suitable to predict dimensions of affect in spontaneous conversations. We first confirm prior findings that the longer utterance-level features are more predictive than word-level features, consistent with the finding in Vlasenko et al. (2008) on emotion classification tasks. Furthermore, we indicate that the affective cues are expressed to a greater extent in some regions of the utterance. We divide the turn into regions of interest for acoustic analysis by considering pitch accent realization in two different ways. Both ROI representations show substantial improvements over less informed utterance-level features. It is also interesting to note that the prediction of the two ROI representations is highly correlated with each other, since they convey similar information in different ways. Last but not least, we find that the performance of acoustic models is influenced by the inter-annotator agreement on the training data. We divide the original training data into two groups of reliable and unreliable partitions according to the agreement among different raters and obtain improvements by using only reliably annotated data in training instead of making use of the larger but noisier training set.

As in any study involving two rather different views of the data|lexical or acoustic modalities in our case|it is tempting to ask which of these broad classes of representation is the better one. Our work however clearly demonstrates that instead we ought to focus on refining the representations within each modality, i.e. how a modality is represented influences performance more than the choice of modality itself.

The combination of different sources of information together for emotion and affect prediction is a challenging task. D'Mello and Kory (2012) reviews 30 studies on both unimodal and multimodal affect detection and concludes that the improvements obtained of multimodal systems over unimodal ones were much lower on natural or seminatural data,

only 4.39% improvement on average. In our study, we achieved modest improvement on VALENCE and POWER by decision fusion with SVR, while this simple fusion method did not show consistent improvements on AROUSAL and EXPECTANCY. However, we observe relatively low correlation between the prediction from acoustic and lexical indicators. This suggests that the proposed acoustic and lexical representations are complementary to each other and the potential for improvement for fusion is high.

Finally, we compared our performance with the AVEC 2012 benchmarks and other participants. The baseline system and most participants applied both video and audio analysis in the challenge, while the lexical information was not included in their analysis. The baseline audio-visual system trained on the union of word-level high-dimensional super-segmental acoustic features and LBP image features obtains a correlation of 0.015 (Schuller et al., 2012), which is much lower than our scores of 0.152 and 0.214 obtained on the best acoustic and lexical representations respectively. Ozkan et al. (2012) applied co-HMM fusion with low-dimensional video (smile, gaze, head tilt), acoustic (energy, articulation rate, F0, Peak slope, Spectral stationarity), and scale time features. Their best performance on word-level prediction is 0.200 which is the second place in the word-level competition. Compared with that we report improvements due to the stronger prediction power of lexical representations. The promising performance suggests the need for including lexical information in prediction of affect dimension in spontaneous conversions in future application, which has received relatively little attention before. Our proposed multimodal affect recognition system with audio, lexical, and video analysis won the word-level prediction competition in AVEC 2012 challenge, obtaining impressive improvement from multimodality via much more advanced fusion methods with particle filtering (Savran et al., 2012).

References

Batliner A, Fischer K, Huber R, Spilker J, Nöth E. Desperately seeking emotions or: Actors, wizards, and human beings. ISCA Workshop on Speech and Emotion. 2000

Bitouk, D.; Verma, R.; Nenkova, A. Class-level spectral features for emotion recognition; Speech communication. 2010. p. 613-625.

Bradley, MM.; Lang, PJ. Tech rep. The Center for Research in Psychophysiology, University of Florida; 2010. Affective norms for english words (anew): Stimuli, instruction manual and affective ratings.

Cauldwell RT. Where did anger go? the role of context in interperting emotion in speech. ITRW on Speech and Emotion. 2000

Cowie R, Douglascowie E, Savvidou S, Mcmahon E, Sawey M, Schrder M. feeltrace: An instrument for recording perceived emotion in real time. Proceeding of ISCA Workshop on Speech and Emotion. 2000:19–24.

D'Mello SK, Kory J. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. ICMI. 2012:31–38.

Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. ACM Multimedia. 2010:1459–1462.

Fano, RM. Transmission of Information: A Statistical Theory of Communications. MIT Press; Cambridge, MA: 1961.

Fontaine JR, Scherer KR, Roesch EB, Ellsworth PC. The world of emotion is not two-dimensional. Psychological Science. 2007; 18:1050–1057. [PubMed: 18031411]

George, D.; Mallery, P. SPSS for Windows Step by Step: A Simple Guide and Reference. Boston: Allyn & Bacon; 2003.

Godfrey, JJ.; Holliman, EC.; McDaniel, J. ICASSP. 1992. Switchboard: Telephone speech corpus for research and development; p. 517-520.

Lee C, Yildirim S, Bulut M, Kazemzadeh A, Busso C, Deng Z, Lee S, Narayanan S. Emotion recognition based on phoneme classes. INTERSPEECH. 2004:205–211.

Lee CM, Narayanan SS. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing. 2005a; 13 (2):293–303.

Lee CM, Narayanan SS. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing. 2005b; 13:293–303.

Litman DJ, Forbes-Riley K. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Communication. 2006; 48 (5):559–590.

McKeown G, Valstar MF, Cowie R, Pantic M, Schröder M. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. T. Affective Computing. 2012; 3 (1):5–17.

Nenkova A, Brenier JM, Kothari A, Calhoun S, Whitton L, Beaver D, Jurafsky D. To memorize or to predict: Prominence labeling in conversational speech. HLT-NAACL. 2007:9–16.

Ozkan D, Scherer S, Morency L-P. Step-wise emotion recognition using concatenated-hmm. ICMI. 2012:477–484.

Polzehl, T.; Schmitt, A.; Metze, F. Salient Features for Anger Recognition in German and English IVR Portals, Spoken Dialogue Systems Technology and Design Edition. Springer; Boston, USA: 2010. p. 83-105.

Polzehl T, Schmitt A, Metze F, Wagner M. Anger recognition in speech using acoustic and linguistic cues. Speech Communication. 2011; 53 (9–10):1198–1209.

Savran A, Cao H, Shah M, Nenkova A, Verma R. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. ICMI. 2012:485–492. [PubMed: 25300451]

Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. ICASSP. 2004:577–580.

Schuller B, Valstar M, Eyben F, Cowie R, Pantic M. Avec 2012: the continuous audio/visual emotion challenge. ICMI. 2012:449–456.

Soladié C, Salam H, Pelachaud C, Stoiber N, Séguier R. A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. ICMI. 2012:493–500.

Turney PD, Pantel P. From frequency to meaning: vector space models of semantics. J Artif Int Res. 2010; 37(1):141–188.

Vlasenko B, Schuller B, Mengistu KT, Rigoll G, Wendemuth A. Balancing spoken content adaptation and unit length in the recognition of emotion and interest. INTERSPEECH. 2008:805–808.

Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945:80–83.

Young SJ, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P. The HTK Book Version 3.4. 2006

Table 1

Overview of the AVEC Grant Challenge dataset

	Training	Development	Testing
Number of sessions	31	32	32
Duration in total	2.79 h	2.5 h	2.27 h
Number of speakers	7	8	8
Number of turns	1,232	1,145	941
Number of words in total	20,183	16,311	13,856
Average no. of words per turn	16.38	14.25	14.72
Vocabulary size	2,109	1,887	1,759
Average turn length in seconds	5.16	4.65	4.36
Average word durations (s)	0.262	0.276	0.249

Table 2

Distribution of inter-rater agreement, given in ranges of alpha, with different numbers of raters in the training and development partitions of the AVEC challenge dataset.

	AROUSAL	EXPECTANCY	POWER	VALENCE
2 raters				
a < 0.5 (unacceptable)	28.60%	64.30%	28.60%	35.70%
a > 0.7 (acceptable)	28.60%	7.10%	35.70%	35.70%
a > 0.8 (good)	0%	0%	14.30%	28.60%
6 raters				
a < 0.5 (unacceptable)	25.80%	29%	6.50%	9.70%
a > 0.7 (acceptable)	51.60%	32.30%	87.10%	77.40%
a > 0.8 (good)	22.50%	9.70%	38.70%	45.20%
8 raters				
a < 0.5 (unacceptable)	8.30%	8.30%	0	0
a > 0.7 (acceptable)	75%	41.70%	66.70%	83.30%
a > 0.8 (good)	41.70%	33.30%	58.30%	66.70%

Table 3

Words with highest PMI with affective dimensions transformed into binary classes

Class	AROUSAL	EXPECTANCY	POWER	VALENCE
0 - low	anxious	house	depressed	bloody
0 - low	inside	that	hello	tha
0 - low	opinion	deadline	extra	world
0 - low	optimism	goes	inherently	annoy
0 - low	rains	write	rude	insurance
0 - low	three	writing	smiling	reference
0 - low	language	general	travel	annoys
0 - low	change	magical	unfortunate	anxious
0 - low	guess	rest	expect	constructive
0 - low	realise	job	face	fighting
1 - high	afternoon	let’s	write	interesting
1 - high	drawer	anxious	idea	bye
1 - high	funny	balance	lights	it’ll
1 - high	pissed	care	buy	weekend
1 - high	spain	everything’s	goes	tree
1 - high	spanish	face	language	holidays
1 - high	win	funny	calm	phd
1 - high	car	pain	tree	together
1 - high	annoys	supposed	opportunity	she
1 - high	buy	truth	saw	love

Table 4

Comparison of the selected vocabulary in Sparse BOW and Sparse PMI representations on the four affective dimensions.

No. of words	AROUSAL	EXPECTANCY	POWER	VALENCE
BOW and PMI C0 shared	326	282	235	179
BOW and PMI C1 shared	343	181	284	311
BOW unique words	379	585	529	558

Table 5

Words with highest/lowest ANEW ratings for each affective dimension

Rating	AROUSAL	POWER	VALENCE
low	boring	rejection	rape
low	relaxed	humiliation	suicide
low	relax	robbery	funeral
low	paper	helplessness	cancer
low	fatigued	insecure	rejection
low	tired	loss	murderer
low	lazy	failure	miserable
low	sign	disability	suffocate
low	sleep	miserable	torture
low	quiet	famine	unhappy
high	rage	me	triumphant
high	orgasm	leader	love
high	rollercoaster	winner	paradise
high	thrill	confident	loved
high	enraged	admired	joy
high	explosion	win	miracle
high	terrified	king	funny
high	killer	complete	humor
high	win	choice	laughter
high	ecstatic	excellence	award

Table 6

Features used for emotion recognition: low-level descriptors (LLD) and functions

LLD (26)	Functionals (19)
Prosody features: intensity, loudness, F0, F0 envelope, probability of voicing, zero-crossing rate ***Spectral features:*** MFCC 1–12, LSF 1–8	max, min, mean, standard deviation, liner regression: offset, slope, linear, quadratic error extremes: value, range relative position, skewness, kurtosis, quartile 1–3, 3 inter-quartile ranges

Table 7

Summary of lexical and acoustic features evaluated in the affect prediction experiments

Lexical representations	Number of features	Acoustic representations	Number of features
WL PMI	2	Word-level	988
TL PMI	2	Turn-level	988
Sparse BOW	1048	Phoneme ROI	2508
Sparse PMI	1000	Accent ROI	1976
Sparse ANEW	548		
TF.IDF	2992		

Table 8

Correlation performances (average of correlation coefficients evaluated over each sequence) of various lexical representations on test set.

	AROUSAL	EXPECTANCY	POWER	VALENCE	mean
PMI features					
WL PMI	0.011	0.008	0.004	0.037	0.015
TL PMI	0.054	0.045	0.123	0.143	0.091
Sparse lexical representations					
Sparse BOW	0.092	0.267	0.168	0.201	0.182
Sparse PMI	0.131	0.285	0.254	0.188	0.214
Sparse ANEW	0.029	X	0.125	0.129	0.094
TF.IDF	0.121	0.217	0.153	0.217	0.177
Percentage of OOV turns					
Sparse BOW	1.5%	1.5%	1.5%	1.5%	1.5%
Sparse PMI	4.8%	11.5%	23.1%	12.2%	12.9%
Sparse ANEW	46.2%	46.2%	46.2%	46.2%	46.2%
TF.IDF	1.1%	1.1%	1.1%	1.1%	1.1%

Comput Speech Lang. Author manuscript; available in PMC 2015 January 01.

Table 9

P-value of sign rank significant test between correlation performances (correlation coefficients evaluated over each sequence of different affect dimensions) between the best **Sparse PMI** feature and other lexical representations.

	WL PMI	TL PMI	Sparse BOW	Sparse ANEW	TF.IDF
AROUSAL	0.01	0.04	0.2	0.14	0.41
EXPECTANCY	<0.001	<0.001	0.59	NA	0.05
POWER	<0.001	0.002	0.14	0.01	0.03
VALENCE	<0.001	0.31	0.68	0.31	0.76
Overall	<0.001	<0.001	0.14	0.005	0.02

Table 10

Correlation performances (average of correlation coefficients evaluated over each test conversation) of various acoustic representations on the test set.

	AROUSAL	EXPECTANCY	POWER	VALENCE	mean
Super-segmental acoustic features over entire speech chunks					
Word-level	0.002	0.009	0.007	0.034	0.013
Turn-level	0.107	0.112	0.054	0.122	0.099
ROI-based acoustic features					
Phoneme ROI	0.112	0.123	0.127	0.144	0.126
Accent ROI	0.093	0.180	0.132	0.138	0.136
ROI-based acoustic features with reliable training					
Phoneme ROI reliable	0.164	0.136	0.134	0.143	0.144
Accent ROI reliable	0.149	0.185	0.139	0.144	0.152

Table 11

P-value of sign rank test between correlation performances (correlation coefficients evaluated over each test conversation) of the best performing **Accent ROI reliable** and the other acoustic features.

	Word-level	Turn-level	Phoneme ROI	Accent ROI	Phoneme ROI reliable
AROUSAL	0.003	0.25	0.33	0.09	0.61
EXPECTANCY	<0.001	0.02	0.02	0.65	0.06
POWER	<0.001	0.04	0.76	0.66	0.75
VALENCE	0.008	0.61	0.60	0.87	0.82
Overall	<0.001	0.003	0.16	0.19	0.36

Table 12

Correlation scores (average of correlation coefficients over four affective dimensions) of various turn-level acoustic and linguistic indicators on the test partition

	PMI	ANEW	TL acoustic	ROI Phoneme	ROI Accent
Sparse BOW	0.660	0.466	0.208	0.311	0.307
Sparse PMI		0.414	0.190	0.240	0.249
Sparse ANEW			0.171	0.256	0.237
TL acoustic				0.603	0.636
ROI Phoneme					0.701

Table 13

Correlation performances (average of correlation coefficients evaluated over each test conversation) of late stage average, SVR, and early stage feature fusion with **Sparse BOW** and **Phoneme ROI** on the test set. The best performance results with respect to correlation score is bold-face for each affective dimension.

	AROUSAL	EXPECTANCY	POWER	VALENCE	mean
Performance of Single modalities					
Corr. of two models	−0.010	0.479	0.213	0.276	0.240
Sparse PMI	**0.131**	0.285	0.254	0.188	0.214
Phoneme ROI	0.112	0.123	0.127	0.144	0.126
Performance of fusion systems					
Average	0.121	0.260	0.245	0.204	0.208
SVR	0.099	**0.287**	**0.265**	**0.210**	0.215
Feature fusion	0.117	0.165	0.140	0.144	0.142

Comput Speech Lang. Author manuscript; available in PMC 2015 January 01.