Published in final edited form as:

Proc ACM Int Conf Multimodal Interact. 2012 ; 2012: 485–492. doi:10.1145/2388676.2388781.

Combining Video, Audio and Lexical Indicators of Affect in Spontaneous Conversation via Particle Filtering

Arman Savran[♮], **Houwei Cao**[♮], **Miraj Shah**[♮], **Ani Nenkova**[♯], and **Ragini Verma**[♮]

[♮]Department of Radiology, Information Science University of Pennsylvania Philadelphia, PA 19104

[♯]Department of Computer & Information Science University of Pennsylvania Philadelphia, PA 19104

Abstract

We present experiments on fusing facial video, audio and lexical indicators for affect estimation during dyadic conversations. We use temporal statistics of texture descriptors extracted from facial video, a combination of various acoustic features, and lexical features to create regression based affect estimators for each modality. The single modality regressors are then combined using particle filtering, by treating these independent regression outputs as measurements of the affect states in a Bayesian filtering framework, where previous observations provide prediction about the current state by means of learned affect dynamics. Tested on the Audio-visual Emotion Recognition Challenge dataset, our single modality estimators achieve substantially higher scores than the official baseline method for every dimension of affect. Our filtering-based multi-modality fusion achieves correlation performance of 0.344 (baseline: 0.136) and 0.280 (baseline: 0.096) for the fully continuous and word level sub challenges, respectively.

Keywords

emotion recognition; affective computing; particle filtering; emotion dynamics; multi-modality fusion; local binary patterns; class-spectral features; lexical analysis; adaboost; svm

1. INTRODUCTION

The affective state of a speaker is conveyed in several communication modalities: facial expressions, voice characteristics and intonation and the words the speaker says all contribute useful cues for interpreting the affective state. Even dealing with a single modality is challenging and the complexity of processing increases considerably when

arman.savran@uphs.upenn.edu
houwei.cao@uphs.upenn.edu
mirajs@seas.upenn.edu
nenkova@seas.upenn.edu
ragini.verma@uphs.upenn.edu

Categories and Subject Descriptors I.5.4 [**Computing Methodologies**]: Pattern Recognition—Applications: Signal processing; Computer vision; Waveform analysis; Text processing

multiple modalities are used in automatic affect processing. In this paper, we study facial video, audio and lexical indicators for affect estimation and present results for individual classes of features, as well as combination of all classes for the AVEC2012 challenge [15]. The AVEC2012 challenge follows the dimensional affect modelling theory. The task is to estimate the four affect dimensions—AROUSAL, EXPECTANCY, POWER and VALENCE—of speaker affective states during spontaneous dyadic conversations.

There is a rich literature on automatic emotion recognition independently done with video and speech modalities. For video analysis, almost all prior work [5] focuses on categorical emotion descriptions or detection of facial action units [3] that can be used as intermediate tools for emotion analysis. As far as we know, the only previous video-based work that aims at continuous estimation of the affect states during dyadic conversations uses 20 facial feature points and employs output-associative Relevance Vector Machines [10].

Previous work on acoustic analysis focused on identifying representative features and developing sophisticated classification paradigms[18]. In addition, linguistic and lexical features have been also incorporated in recent years [14].

In contrast, multi-modal affect recognition is limited compared to the richness of single modality studies, partly because of the necessity of interdisciplinary expertise and insufficient number of available databases. A review about previous multi-modal studies is available in [20].

The AVEC2011 challenge [16], where binary classification of affective dimension was performed, was organized in an attempt to bring multidisciplinary researchers together. In AVEC 2012, the objective is the continuous estimation of the affect dimensions rather than binary classification of each dimension. Most of the research teams in AVEC2011 had not considered fusion of audio and video modalities. The only two available fusion studies [2, 6] do not show considerable improvements when audio and video are fused, and even some degradations were observed for several affect dimensions. These prior findings clearly demonstrate the need for improved techniques for fusion.

In our work, we extract rich video, audio and lexical features. Video features comprise the best temporal features amongst the temporal statistics of texture descriptor components (Section 4). Audio features include prosodic, spectral and voice quality features (Section 2). Lexical features are based on the mutual information between a word and high activation in a specific dimension of affect (Section 3). We then fuse the regressions from the different classes of indicators using a Bayesian filtering approach (Section 5); the algorithm is based on Particle Filtering and makes use of learned dynamics of affect (Section 5.2) in the fusion process. We obtain improved estimations for all three classes of modalities, and our fusion is able to combine these independent estimators effectively, as demonstrated in Section 6.

2. AUDIO REPRESENTATION

Utterances/speaker turns are the unit we use to create the regressors in our experiments with audio. 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. In the AVEC challenge data, speaker turns are short, 2–3 seconds on average, so it is reasonable to expect that they have consistent emotion profiles.

Features are calculated to characterize the emotion content in turns and affect dimensions are estimated for each turn. To evaluate word-level or frame-level estimation, we simply assign the estimation obtained for the turn as the value for all constituent words and frames.

We use the openSMILE toolkit [4] to obtain a comprehensive set of acoustic features. In Table 1 we list all of the 26 prosodic and spectral low–level descriptor (LLD) contours used. For each of these, the 19 functional statistics listed in the second column are also computed, as well as the first order delta coefficients.

Along with the conventional turn-level acoustic analysis, we also consider class-level spectral features. In the 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 summarized only for the frames that correspond to a particular phoneme class. In order to do this, we perform Viterbi-based forced alignment between the manual transcript and the audio, in order to identify which segments of audio correspond to which phoneme class [19]. In this way we 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. After that, for each spoken turn, we group phonemes into three distinct phoneme classes—stressed vowels, unstressed vowels and consonants. Class-level spectral features were extracted by computing statistics of spectral measurements from parts of the spoken turns, corresponding to these classes.

In prior work on categorical emotion estimation, class-level spectral features have proven to be complementary to turn-based features, with superior performance as a standalone class [1][8]. In the work presented here, we analyze the effectiveness of class-based spectral features for the estimation of continuous emotion dimensions. We compute turn-level prosodic, turn-level spectral and class-level spectral features. Later, we will discuss the performance of the following combinations:

TL acoustic

Turn prosodic and spectral

CL acoustic

Turn prosodic and class spectral

3. LEXICAL REPRESENTATION

Many words have strong positive or negative connotations and often what people say (the words they use) carries rich information about the affective state of the speaker. Much work in text processing has shown that subjectivity, opinion and emotion can be successfully estimated simply on the basis of lexical features [11].

In our work we calculate the pointwise mutual information (PMI) between a word and a given affect dimension. PMI has been successfully applied for categorical emotion estimation [9] and is widely used as a measure of association in a range of semantic processing applications [17].

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 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.

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

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

P(ε) is the prior probability of an affect dimension and *P*(ε|*w*) is the conditional probability of the affect dimension given the word *w*. Both probabilities are computed directly from counts on the 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 2 lists examples of the five words with the highest PMI, that are associated with low (class 0) and high (class 1) values of the affect dimensions.

As with audio, the unit of representation and estimation for the lexical models is a speaker turn. We experimented with two turn representations based on the PMI between the words in the turn and affect dimensions.

TL PMI

Turn level PMI

TL Sparse Lex

Turn level sparse lexical representation

The turn-level PMI representation consists of only two features. They are computed as the average of word-level PMI of all the 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 turns that consist entirely of words that did not appear in training, we take the values from the preceding turn as 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. 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 utterances. Speaker turn in the AVEC dataset consists of 15 words on average, so these representations are indeed sparse.

4. VIDEO REPRESENTATION

In video, unlike in audio, regression is performed directly on the frame level. An immediate complication is that now the size of the training set is prohibitively large, containing more than 500,000 frames. Moreover, neighbouring frames carry redundant information due to the high frame rate (50 fps). To address these problems, we resample the training data. For each emotion dimension, different set of training examples are selected by sampling from frames for which affect value is greater or smaller than half of the standard deviation estimated over the entire training set. Resampling is done randomly by rejection from this set. We empirically determined the rate of the resampling as 40 per minute and rejection neighbourhood size as one second. If a sampled frame falls in the neighbourhood of an already selected frame, it is rejected.

We choose to work with low level image features. However, we employ temporal features as well as spatial features in order to capture temporal cues that can have information about one’s affective state, such as fast head and face movements or stable moments. Duration of the visual signals may also correlate with the emotional states. For this reason, we consider temporal features with different durations as well.

Our temporal features are based on still image texture descriptors known as Local Binary Patterns (LBP). We employed the LBP features that are provided within the challenge [15]. These features are extracted after a registration stage that involves face and eye detection and normalization onto a 200×200 pixel image by aligning the eyes. Registered facial images are divided into 10×10 blocks and a histogram with 59 bins is calculated for every block, resulting in 5900 bins to be used as features. Our feature set also involves face and eye detection outputs, resulting in 5908 features of the static face.

For each feature component, we represent temporal information with the mean and standard deviation over fixed length temporal windows. The mean captures the dominant value for each component in a temporal window and smooths out variation, while the standard deviation conveys information about the variability. For each time point, the statistics are computed for five different temporal windows with widths of $\left\{2^k\right\}_{k=-1}^{3}$ seconds immediately preceding the time point. The window range is from 25 to 400 frames, given the frame rate of 50 selected for the cuprous. We also include the descriptor bin values of the current frames. Thus, the size of the final feature vector is (5 × 2 + 1) × 5908 = 64988.

The number of features is too large to include in regression directly, so we perform feature selection. We apply classification AdaBoost, as done in prior work on facial action unit intensity estimation [13], to select compact feature sets (200 features) according to their success at discriminating between high and low values, such that each feature is a binary weak classifier (nearest mean classifier). The high and low values are defined as the values above and below the mean, for each affect dimension.

AdaBoost feature selection is performed on the resampled training set. The selected features are used in epsilon Support Vector Machine regressors (SVR) with linear kernel. Training is also performed on the resampled training set.

5. BAYESIAN FILTERING FOR FUSION

Regression does not incorporate any sequence information. However, prediction of the current state can greatly benefit from the valuable information contained in past estimates from the sequence because emotion sequences are continuous and in most cases emotion states are stable for short intervals of time. In order to take advantage of these temporal dependencies between emotional states, we perform Bayesian filtering via particle filters that adaptively process data to fuse different modalities. The filter operates on the regression outputs of each modality and incorporates prior knowledge about the dynamics of the emotional states.

5.1 Particle Filtering

Particle filtering [12] is an effective Bayesian framework to handle non-linear models and non-Gaussian probability density propagation processes. In our case, an important motivation for using particle filtering is the fusion, because outputs of different regressors can cause multi-modal posterior distributions if they do not agree on the affect state. By propagating the posterior of the states, particle filters can achieve near-optimal estimations. Moreover, the observation model is not completely linear since it is possible that at a given time not all observations are available, as explained in Section 5.3. Estimations with other Bayesian filters that impose strong assumptions, such as linearity and Gaussianity as in Kalman filtering, would be far from optimal.

To define the filtering process, let x_t be our state variable corresponding to an affect dimension at discrete time t. It evolves according to a transition function at t

$$x_t = f_t\left(x_{t-1}, v_t\right) \quad (2)$$

where v_t is an i.i.d. process noise sequence. The objective is to estimate x_t recursively using the measurements

$$\mathbf{z_t} = \mathbf{h_t}\left(x_t, \mathbf{n_t}\right) \quad (3)$$

where n_t is an i.i.d. measurement noise sequence. Here, the components of the measurement vector $\mathbf{z_t}$ are the regression outputs of different modalities at discrete time t.

Particle filtering estimates the posterior distribution of the state given past observations, $p(x_t|\mathbf{z_{0:t}})$, by sequential Monte Carlo method [12]. The posterior is represented by weighted N particles, $\left\{x_t^i, w_t^i\right\}_{i=1}^N$ where weights w_t^i are updated at each time step. We use Sampling Importance Resampling (SIR) filter [12], also known as Bayesian Bootstrap filter. The SIR filter uses the prior density as the importance density to draw the particles from. Thus the filter first performs N predictions according to the transition density,

$$x_t^i \sim p\left(x_t|x_{t-1}^i\right), \quad (4)$$

and then each prediction is weighted according to the observations in the update step by the likelihood function

$$w_t^i = p\left(\mathbf{z_t}|x_t^i\right). \quad (5)$$

Particles are re-sampled according to their weights at each iteration to attain equal weights. This is achieved by resampling the same particle multiple times depending on the weight. For inference, we apply minimum mean square error estimation (MMSE) using 50 particles.

5.2 Auto-regressive Emotion Dynamics

During filtering, the current state is predicted according to the transition function f_t (equation 2). A transition model which better approximates the dynamics of the states means we have more realistic prior (prediction) distribution, thus leading to better estimations. Proper predictions can also reduce the number of necessary particles since they may be placed close to the observation likelihood peaks more frequently, i.e., fewer wasted particles.

We find proper models of the emotion dimension dynamics by fitting auto-regressive (AR) models with order K

$$x_t = a_0 + \sum_{k=1}^{K} a_k x_{t-k} + v_t \quad (6)$$

where a_k are the model coefficients and $v_t \sim N(0, \sigma^2)$. AR models are estimated by the Burg method [7]. However, it is important to note that if the model parameters are estimated over the concatenated sequences, spurious signals will appear due to the jumps at the concatenation site. To circumvent this issue, we skip the samples at the concatenation points so that samples from the consecutive sequences are never used in the same iteration, while estimating the autocorrelations.

We chose to use a second order model by analysing the reflection coefficients. We learned AR models for each of the emotion dimensions separately and here was a sharp drop for coefficients of higher order for all emotion dimensions, which suggested that the acceleration model is the most appropriate one. Thus the samples are drawn from the distribution

$$x_t^i \sim N\left(x_t;a_0+a_1x_{t-1}+a_2x_{t-2},\sigma^2\right) \quad (7)$$

5.3 Measurement Model for Fusion

We combine measurements coming from each modality by

$$\mathbf{z_t}=\mathbf{1}.\mathbf{x_t}+\mathbf{n_t} \quad (8)$$

where 1 is vector of ones, and $\mathbf{n_t}$ is multivariate Gaussian measurement noise with covariance **R**, $\mathbf{n_t} \sim N(\mathbf{0}, \mathbf{R})$. $\mathbf{z_t}$ is composed of regression outputs of video, audio and lexical indicators. Thus if the variance of error is smaller for one modality, in filtering it will have higher influence than the other modalities. Also, overall higher measurement variance increases the role of predictions for the estimation since posterior distribution becomes more similar to prediction distribution. We estimate full covariance matrices over the error sequences in order to take the correlations between the modalities into account as well.

An issue during fusion is that features for each modality cannot be extracted at every point in time. For instance, audio measurements only exist when the subject is speaking; video measurements are not available if the face is not detected. At times it is even possible that none of the measurements are available. Therefore, for each point in time we use only the available measurements to compute the likelihood of the state. Thus for each combination of measurements we have different likelihood models. In case there is no observation at all, all the particles have the same likelihood, i.e., estimations are based only on the predictions. However, as explained and discussed in Section 6.3 and Section 6.4, we also experiment with a method where we fill missing measurements with values from preceding turns. This can be interpreted as artificial likelihood values over non-speech regions which prevent large deviations from the preceding over-word estimations.

Let the nominal variable c_t represent a measurement combination at time t which indicates a mapping of $\mathbf{z_t}$ to a new observation vector $\mathbf{z_t^{c_t}} \in R^{n_{c_t}}$ where n_{c_t} is the number of available measurements. Then the algorithm is:

- Given: $\left\{x_{t-1}^i, w_{t-1}^i\right\}_{i=1}^N, \mathbf{z_t}, \mathbf{c_t}$
- Predictions: $x_t^i \sim N\left(x_t;a_0+a_1x_{t-1}+a_2x_{t-2},\sigma^2\right)$
- Update if any measurement exists:
  - $w_t^i=N\left(\mathbf{z_t^{c_t}}-\mathbf{1}.\mathbf{x_t};\mathbf{0},\mathbf{R^{c_t}}\right)$
  - $w_t^i \leftarrow w_t^i/\Sigma_{j=1}^N w_t^j$
  - Resample ($\left\{x_t^i, w_t^i\right\}_{i=1}^N$)
- Posterior: $p\left(x_t|\mathbf{z_{0:t}}\right) \approx \frac{1}{N}\Sigma_{i=1}^N \delta\left(x_t-x_t^i\right)$

- Apply MMSE estimation

6. EXPERIMENTAL RESULTS

6.1 Challenge Datasets

The challenge dataset [15] is composed of 95 sequences of video (upper body video), audio and manual speech transcripts of dyadic interactions between human subjects and virtual agents operated by a human. All frames are annotated for the four dimensions of affect: AROUSAL, EXPECTANCY, POWER and VALENCE. There are four different virtual agents who have unique moods, like angry and cheerful. The subjects have conversations with all of the agents, thus there are four sequences per subject which usually exhibit different combinations of affect states. The total duration of the footage is about 7.5 hours with more than 50 thousands words. The dataset is divided into training, development and test subsets with 31, 32 and 32 sequences respectively. Every subset consist of eight participants, each interacting with all four virtual agents.

There are no subjects that appear both in the training and testing set, however some subjects appear in both training and development data.

For individual classes of features, we use Support Vector Regression (SVR) with linear kernel for regression. SVM parameters are optimized on the development set and the SVR regressors are trained on the training set. For fusion with particle filtering, the training set is used to learn the affect dynamics, and measurement noises are estimated over the regression errors on the development set.

6.2 Analysis of Speech Related Features

Here we discuss the performance and robustness of the audio and lexical features. Table 3 shows the average correlation coefficients of the word-based sub-challenge. Models are trained on the training data and the development sets is used for testing. We analyze the robustness of regression and list both speaker-dependent (SD) and speaker-independent (SI) results in the table. For SD regression, we report results only for the 12 samples from speakers in the development set that also appeared in the training set. For SI regression, we only report on the 20 samples from speakers that did not appear in training.

All our acoustic and lexical models outperform the AVEC12 audio baseline, which we also listed for comparison. The baseline acoustic system relies on features extracted on the word level, unlike ours which works on the turn level. Acoustic features are more sensitive to unseen speakers and performance degrades considerably for all affect dimensions, especially for AROUSAL and VALENCE. The turn level sparse lexical representation is unique because with it, results are higher for speaker independent experiments than for speaker dependent.

In both acoustic and lexical features, different representations lead to best performance on each of the four affect dimensions. The sparse lexical representations are better than the average PMI for AROUSAL and VALENCE, but not for EXPECTANCY and POWER. We observe the same effect for acoustic features. In particular, class-based spectral features will give better regressions on arousal and power than turn-level features. These two dimensions are more sensitive to sub-band energy and therefore the more focused class-based spectral-features yield improvements.

Figure 1 gives a comparison of different acoustic and lexical features for the word-level sub-challenge (WLSC) on the test set. The audio baseline results are again shown in the figure for comparison. As in the development set, our audio and lexical features outperform the acoustic baseline by a large margin. Class-level acoustic feature and sparse lexical features achieve correlation coefficients of 0.114 and 0.162 respectively with the gold-standard annotations. The correlation for the audio baseline is only 0.081. On the test set, sparse lexical features prove to be the best individual representation, with 0.2 cross-correlation on all dimensions except arousal. Class-level acoustic features perform notably well for estimating power, where they emerge as the best representation that yields more than double improvement over the baseline.

Lexical features appear to be more robust than acoustic features, as performance for them degrades only slightly compared to the results we obtained on the development set. There is a larger performance gap between development and test set for the acoustic features.

6.3 Evaluation of Single Modality Regression Estimations

Here we report both frame-level (FCSC) and word-level (WLSC) correlations with the gold-standard affect annotation. Table 4 shows the averaged correlation coefficient scores (average of correlation coefficients evaluated over each sequence) for both sub-challenges (frame level FCSC and word level WLSC) on development and test sets. We see that on the development set for fully continuous sub-challenge (Table 4.a), our video-only results are above the baseline for all affect dimensions, with 0.283 correlation score on average (baseline video: 0.146). We compute word level video estimates by taking the average values of the frames over the corresponding words, and obtain 0.272 WLSC score (baseline video: 0.124) (Table 4.b).

Our audio (TL acoustic) and lexical indicators (TL Sparse Lex) scores are 0.228 and 0.218 for WLSC (baseline audio: 0.074) (Table 4.b), respectively. To convert speech modality regressions to the frame level, we assume that all frames in a turn have the affect value estimated for the entire turn. Frames that correspond to regions during which the virtual agent speaks (the subject is silent) are assigned the affect value of the preceding turn of the subject. This extrapolation is reasonable if the gaps between the turns are short or if there are still cues, i.e., visual signals, that conform with the current affective state while the subject is listening. Otherwise, extrapolations can cause misleading predictions.

As can be seen in Table 4.a, we obtain 0.295 and 0.263 FCSC correlation scores for audio and lexical indicators, respectively. The extrapolation improves the frame-level power estimations dramatically compared to those on word-level, for instance for audio from 0.147 to 0.455. This shows that the power of the preceding turn is a great indicator of the affective state of the subject while the subject is listening to their interlocutor in interactive conversations. Valence also improves slightly by the extrapolation, but there is slight drop in performance for arousal and expectancy.

On the test set (Table 4.c and Table 4.d), we observe a marked degradation of results for all modalities (our systems and baselines) compared to those on the development set. This may be because all the test subjects are different from the subjects of the training set, while some training subjects are the same in the development set. The mismatch in performance is largest for POWER in the video modality, AROUSAL and EXPECTANCY for audio and arousal for lexical indicators. The reason may be the lack of generalizability on the test set. However, our results are above the baseline: frame level video of 0.178 vs. 0.104, and 0.091 audio and 0.162 lexical vs. baseline audio of 0.081.

6.4 Evaluation of Particle Filtering Based modality Fusion

Correlation scores for several Particle Filtering based fusion types on the development set are shown in Table 4.a. All the fusions are performed at frame level but we convert frame level estimates to word level by taking average over the words. We compare two fusions for the speech only modality, with and without non-speech region extrapolation as explained in Section 6.3. Here, by means of extrapolation, we fill the missing values of the non-speech regions before filtering, thus the particle filter uses the extrapolated values as if they are the current measurements (regression outputs). Actually, both types of fusion over the non-speech regions perform only prediction from the past regression outputs and the learned models of the emotion dimension dynamics (see Section 5.2). However, this extrapolation acts as a mechanism in the filtering process which forces the affect states to become much closer to the preceding speech region estimation.

Fusion results are better than those of the individual modalities being fused, but several other interesting results can be observed in Table 4.a and Table 4.b. When missing values are extrapolated before filtering, we obtain dramatical improvements on POWER estimations compared to when they are not filled, from 0.283 to 0.556. VALENCE fusion is also considerably better when values are filled than when they are not. The same effect is present for word-level fusion. A reason for this peculiarity might be that our assumption that affect states are stable over short periods of time is more applicable to some dimensions than to others. From our results, it appears that AROUSAL and EXPECTANCY are more stable for a person over time, while POWER and VALENCE are more variable, especially depending on whether the subject is speaking or listening. Recall that when values are not extrapolated, only the emotion dynamics leads the estimation. For noisy measurements of highly variable emotion dimensions, the dynamics will cause more deviations over non-speech regions. Therefore, extrapolation improves the correlation performances by reducing these deviations.

We also perform fusion of all three modalities for both sub-challenges, with and without extrapolation for speech indicators. Results are shown in Table 4.a and Table 4.b. For all dimensions, combining video with speech related features leads to results better than those for video only prediction. Again, extrapolation shows similar effects on the scores.

We can see from the results on the development set that each affect dimension should be predicted with a different fusion method. The best results for each dimension are shown in bold in the tables and the last line in the table gives the performance when the best combination is chosen for each dimension (mixed fusion). Motivated by these results on the development data, we also perform mixed fusion on the test data: for AROUSAL and EXPECTANCY video and speech modalities are fused without extrapolation; for VALENCE estimation we fuse video and speech-related features with extrapolation; for POWER we fuse only speech-related features with extrapolation. The mixed fusion gives the best performance on the test set, supporting the observation that different combinations of modalities are appropriate for different affect dimensions.

All of our particle filtering fusion results have higher correlation scores than the baseline results on the development set. In the baseline, feature level fusion is applied, simply combining all features in a single regression. The benefits from particle filtering are clear and can be attributed to the filtering process which is able to propagate multi-modal posterior distribution in time via the learned dynamics.

Finally, we assess our fusion results on the test set (Table 4.c and Table 4.d). The success of our fusion method is validated by FCSC correlation of 0.344 (baseline: 0.136) and WLSC correlation of 0.280 (baseline: 0.096). The contribution of our fusion is much higher. For instance, for WLSC, while we obtain fusion score of 0.280 by combining single modalities with scores 0.178 (video), 0.091 (audio) and 0.162 (lexical), the baseline single modality fusion score is 0.096 after combining single modalities with scores 0.117 (video) and 0.081 (audio).

7. CONCLUSIONS

In this paper, we study estimation of the four dimensions of affect, namely AROUSAL, EXPECTANCY, power and VALENCE during dyadic conversations on the AVEC 2012 challenge dataset. We first develop single modality regression-based estimators with video, audio and lexical indicators, respectively. For video, we extract temporal statistics of LBP texture descriptor components and consider different time widths to better capture visual clues happening in different time scales better. This makes the number of the possible features impractically large. Also, the total duration of training video is close to three hours. To cope with the excessive amount of video data, we apply automatic video frame selection for training, and discriminative feature selection. Our video regressor obtains 0.178 correlation score compared to the baseline score of 0.104, which highlights the contribution of the temporal features and the effective training frame selection procedure. Video scores of all of the dimensions are above the baseline and are particularly good for the AROUSAL and VALENCE dimensions.

For the speech-related modalities, we exploit both acoustic and lexical information. For the audio modality, we experiment with class-based spectral features for three broad phoneme classes in addition to the conventional turn-based spectral and prosodic features. For the lexical analysis, we computed the word-level mutual-information between words and high and low values for affect dimensions. Words with strongest association are used as the basis for feature space representation of individual turns. Our audio and lexical regressors achieve 0.114 and 0.162 averaged correlation coefficients respectively on test set for the word-level sub-challenge, outperforming the audio baseline of 0.081 by a large margin.

We consider each single modality regression output as an affect measurement in a Bayesian filtering framework. Predictions are done based on past observations and the learned dynamics of the affect states. This is realized by Particle filtering which propagates the multi-modal posterior distributions in time according to the auto-regressive processes.

Our filtering based fusion dramatically improves the scores for single modalities and outperforms baseline fusion which is done on the feature level. This constitutes a major advance compared to the results from the AVEC 2011 challenge where fusion did not appear beneficial [2, 6]. We report fusion correlation scores of 0.344 (baseline: 0.136) and 0.280 (baseline: 0.096) for the FCSC and WLSC, respectively.

Acknowledgments

This work was supported by Grant Number R01-MH-073174 from the NIH.

9. REFERENCES

[1]. Bitouk D, Nenkova A, Verma R. Class-level spectral features for emotion recognition. Speech communication. 2010:613–625. [PubMed: 23794771]

[2]. de Melo, CM.; Carnevale, P.; Antos, D.; Gratch, J. ACII 2011. Memphis, TN: Oct.. 2011 A computer model of the interpersonal effect of emotion displayed in a social dilemma.

[3]. Ekman P, Friesen WV, Hager JC. Facial Action Coding System. 2002

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

[5]. Fasel B, Luettin J. Automatic Facial Expression Analysis: A Survey. Pattern Recognition. 2003; 36(1):259–275.

[6]. Glodek M, Tschechne S, Layher G, Schels M, Brosch T, Scherer S, KÃd'chele M, Schmidt M, Neumann H, Palm G, Schwenker F. Multiple classifier systems for the classification of audio-visual emotional states. IJATEM'11. 2011:359–368.

[7]. Kay, S. Prentice Hall; Englewood Cliffs, NJ: 1988.

[8]. 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.

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

[10]. Nicolaou M, Gunes H, Pantic M. Output-associative rvm regression for dimensional and continuous emotion prediction. FG 2011. 2011

[11]. Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2007; 2(1-2):1–135.

[12]. Pitt, MK.; Shephard, N. Sequential Monte Carlo Methods in Practice. de Freitas, N.; Doucet, A.; Gordon, NJ., editors. Springer-Verlag; New York: 2001.

[13]. Savran A, Sankur B, Bilge MT. Regression-based intensity estimation of facial action units. Image and Vision Computing. 2011 in press.

[14]. Schuller B, Müller R, Lang MK, Rigoll G. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. INTERSPEECH. 2005:805–808.

[15]. Schuller B, Valstar M, Eyben F, Cowie R, Pantic M. Avec 2012, the continuous audio/visual emotion challenge. AVEC 2012 Grand Challenge.

[16]. Schuller B, Valstar M, Eyben F, McKeown G, Cowie R, Pantic M. Avec 2011, the audio/visual emotion challenge. AVEC 2011 Grand Challenge.

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

[18]. Ververidis D, Kotropoulos C. Emotional speech recognition: Resources, features, and methods. Speech communication. 2006:1162–1181.

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

[20]. Zeng Z, Pantic M, Roisman GI, Huang TS. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE TPAMI. Jan; 2009 31(1):39–58.

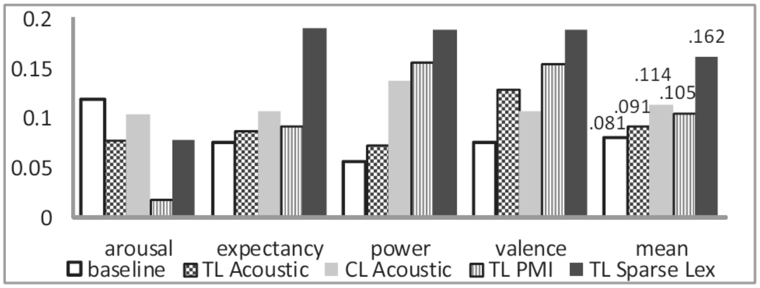

Figure 1.

Comparison between the audio baseline and different acoustic and lexical features on the test set. Performance is measured in cross-correlation average over all sequences on the WLSC

Table 1

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	max, min, mean, standard deviation, liner regression: offset, slope, linear, quadratic error extremes: value, range
Spectral features: MFCC 1–12, LSF 1–8	relative position, skewness, kurtosis, quartile 1–3, 3 inter-quartile ranges

Table 2

Examples of the words with the highest PMI for 4 affective dimensions on binary classes

	words associated with class 0	**words associated with class 1**
AROUSAL	three, plan, alone, change, depressing	car, shopping, around, she, afternoon
EXPECTANCY	plan, their, around, mind, goes	fair, children, question, told, aright
POWER	face, fighting, matter, room, true	afternoon, along, cathy, four, phone
VALENCE	depressing, fighting, room, unnecessarily, angry	afternoon, cathy, four, meet, photograph

Table 3

Correlation performances (average of correlation coefficients evaluated over each sequence) of different audio and lexical features in terms of speaker-dependent (SD) and speaker-independent (SI) sequence, on the development set

WISC-development	**arousal**		**expectancy**		**power**		**valence**		**average**	
Baseline Audio	**0.097**		**0.052**		**0.061**		**0.085**		**0.074**	
	SD	**SI**	**SD**	**SI**	**SD**	**SI**	**SD**	**SI**	**SD**	**SI**
Audio features										
TL acoustic	0.340	0.208	0.244	0.235	0.164	0.137	0.340	0.227	0.272	0.202
CL acoustic	0.363	0.216	0.211	0.184	0.170	0.175	0.380	0.203	0.281	0.195
Lexical features										
TL PMI	0.190	0.188	0.260	0.202	0.218	0.171	0.205	0.200	0.218	0.190
TL Sparse Lex	0.268	0.208	0.228	0.204	0.146	0.168	0.216	0.302	0.215	0.221

Table 4

Correlation performances (average of correlation coefficients evaluated over each sequence) of different modalities and their fusions for fully continuous affect dimension state estimation. Results for both sub-challenges (frame level FCSC and word level WLSC) are shown on development and test sets. Extrap. means extrapolation at non-speech regions for the audio and lexical indicators. PF stands for Particle filtering based fusion. Mixed means the use of the best fusion methods for each affect dimension

(a) FCSC - development					
	Arousal	**Expectancy**	**Power**	**Valence**	**Avg.**
Single Modality					
Baseline Video	0.157	0.130	0.115	0.186	0.146
Video	0.306	0.215	0.242	0.370	0.283
Audio (Exterp.)	0.215	0.215	0.455	0.297	0.295
Lexical (Exterp.)	0.171	0.176	0.396	0.308	0.263
Speech Modality Fusions					
Audio+Lex. (PF)	0.276	0.208	0.283	0.373	0.285
Audio+Lex. (Exterp., PF)	0.275	0.205	**0.556**	0.423	0.365
Video and Speech Modality Fusions					
Baseline Video+Audio	0.162	0.162	0.111	0.208	0.157
Video+Audio+Lex. (PF)	**0.383**	**0.266**	0.238	0.408	0.324
Video+Audio+Lex. (Exterp., PF)	0.377	0.210	0.477	0.473	0.384
Video+Audio+Lex. (Mixed, PF)	0.383	0.266	0.556	0.473	**0.420**

(b) WLSC - development					
	Arousal	**Expectancy**	**Power**	**Valence**	**Avg.**
Single Modality					
Baseline Video	0.145	0.111	0.103	0.137	0.124
Video	0.280	0.202	0.223	0.383	0.272
Baseline Audio	0.097	0.052	0.061	0.085	0.074
Audio	0.257	0.239	0.147	0.270	0.228
Lexical	0.230	0.211	0.160	0.270	0.218
Speech Modality Fusions					
Audio+Lex. (PF)	0.323	0.222	0.192	0.358	0.274
Audio+Lex. (Exterp., PF)	0.329	0.169	**0.296**	0.377	0.293
Video and Speech Modality Fusions					
Baseline Video+Audio	0.113	0.090	0.083	0.114	0.100
Video+Audio+Lex. (PF)	**0.350**	**0.263**	0.209	0.437	0.315
Video+Audio+Lex. (Exterp., PF)	0.384	0.182	0.228	**0.458**	0.313
Video+Audio+Lex. (Mixed, PF)	0.350	0.263	0.296	0.458	**0.342**

(c) FCSC - test					
	Arousal	**Expectancy**	**Power**	**Valence**	**Avg.**
Single Modality					
Baseline Video	0.092	0.121	0.064	0.140	0.104
Video	0.251	0.153	0.099	0.210	0.178
Audio (Exterp.)	0.74	0.032	0.418	0.149	0.168
Lexical (Exterp.)	0.046	0.057	0.426	0.228	0.189
Fusion					

(c) FCSC - test					
	Arousal	**Expectancy**	**Power**	**Valence**	**Avg.**
Baseline Video+Audio	0.149	0.110	0.138	0.146	0.136
Video+Audio+Lex. (Mixed, PF)	0.359	0.215	0.477	0.325	**0.344**

(d) WLSC - test					
	Arousal	**Expectancy**	**Power**	**Valence**	**Avg.**
Single Modality					
Baseline Video	0.091	0.114	0.121	0.143	0.117
Video	0.184	0.156	0.146	0.226	0.178
Baseline Audio	0.119	0.075	0.056	0.076	0.081
Audio	0.078	0.087	0.073	0.128	0.091
Lexical	0.078	0.190	0.189	0.189	0.162
Fusion					
Baseline Video+Audio	0.103	0.105	0.066	0.111	0.096
Video+Audio+Lex. (Mixed, PF)	0.302	0.194	0.293	0.331	**0.280**