Word-Level Emotion Recognition Using High-Level Features

Johanna D. Moore, Leimin Tian, and Catherine Lai

University of Edinburgh
School of Informatics
Informatics Forum, EH8 9AB Edinburgh, UK
J.Moore@ed.ac.uk, s1219694@sms.ed.ac.uk, clai@inf.ed.ac.uk

Abstract. In this paper, we investigate the use of high-level features for recognizing human emotions at the word-level in natural conversations with virtual agents. Experiments were carried out on the 2012 Audio/Visual Emotion Challenge (AVEC2012) database, where emotions are defined as vectors in the Arousal-Expectancy-Power-Valence emotional space. Our model using 6 novel disfluency features yields significant improvements compared to those using large number of low-level spectral and prosodic features, and the overall performance difference between it and the best model of the AVEC2012 Word-Level Sub-Challenge is not significant. Our visual model using the Active Shape Model visual features also yields significant improvements compared to models using the low-level Local Binary Patterns visual features. We built a bimodal model By combining our disfluency and visual feature sets and applying Correlation-based Feature-subset Selection. Considering overall performance on all emotion dimensions, our bimodal model outperforms the second best model of the challenge, and comes close to the best model. It also gives the best result when predicting Expectancy values.

1 Introduction

Affective Computing, the study of recognizing, understanding, and synthesising human emotions using computational technologies, has shown great potential both in academic studies of human behaviour as well as industrial applications. For example, by detecting affective states, such as boredom, an Intelligent Tutoring System can improve student learning and increase user satisfaction [1]. Multimodal emotion recognition has recently become a focus of affective computing. However, this task remains challenging, especially with respect to spontaneous spoken dialogue. Much of the early work on this topic was based on acted expressions of emotions [2], leading to models with good performance when the test and training data are similar, but which perform poorly when applied to a system working in a more natural environment. Moreover, differences in data collection and annotation style make it difficult to compare results across studies.

To address these issues, recent studies have focused on recognizing emotions in more realistic dialogues while shared tasks such as the annual Audio/Visual

A. Gelbukh (Ed.): CICLing 2014, Part II, LNCS 8404, pp. 17–31, 2014.

Emotion Challenge (AVEC) have been held with the goal of comparing different approaches on common datasets of spontaneous speech. Despite these steps forward, the performance of existing multimodal emotion recognition models leaves much room for improvement. Predicted values from the top competitors in AVEC2012 [3], for example, exhibit relatively weak correlations for both the frame and word level subchallenges. As a regression task, the average correlation-coefficients over all test sessions for the best Fully-Continuous Sub-Challenge (FCSC) model [4] and the best Word-Level Sub-Challenge (WLSC) model [5] are 0.456 and 0.280 respectively, i.e., weak to moderate correlations. A possible reason for this poor performance is that the lexical, acoustic, and visual features often examined in these tasks are too low-level to predict emotions.

In this paper, we investigate the predictiveness of high-level features in the word-level emotion recognition task. These features include six disfluency features and locations of facial landmarks. We extracted our features from the AVEC2012 database and compared the predictiveness of our high-level features with that of the conventional lower-level audio and visual features used by the AVEC2012 WLSC baseline models. These include spectral and prosodic (SP) features and Local Binary Patterns (LBP) [6]. We compare our models to the corresponding AVEC2012 WLSC baseline models, as well as the three best performing models from the AVEC2012 WLSC. We find that our high-level features are more predictive than the low-level features, and the performance of our best bimodal model is competitive with the highest scoring models from the AVEC challenge, while using at most 22 features.

1.1 Background

Previous approaches to emotion prediction based on the AVEC data work with a high dimensional space of low-level features (1842 SP features and 5908 LBP features in the baseline model). However, the results from the top performing WLSC model [5] show that significant gains can be made by including lexical features. In this paper, we investigate whether other higher level features can be used to reduce feature space dimensionality and improve performance for this task.

Studies of both human cognition [7] and natural language processing [8] suggest that disfluencies are powerful clues for recognizing the emotional states of a speaker. Thus, disfluency features may have a stronger relationship with emotions than SP features or more general lexical features extracted from content words, and may contain less noise. Therefore, we conjecture that a unimodal emotion recognition model using disfluency features will outperform models using SP features or more general lexical features, and may contain less noise. Therefore, we conjecture that a unimodal emotion recognition model using disfluency features will outperform models using SP features or more general lexical features.

Both the best [4] and the second best [9] performing FCSC models of AVEC2012 chose high-level visual features that describe the facial expressions of the speaker using positions of facial landmark points, instead of the LBP features that describe the orientation of pixels. Their results suggest that high-level visual features may also improve performance when recognizing emotions at word level.

Studies in cognitive science [10] and affective computing [11] show that the audio and visual modalities have different strengths and weaknesses when predicting different emotion dimensions. Therefore, combining the modalities should lead to improved performance, at least compared to the lower performing modality. We also expect to see better performance by combining our high-level features as opposed to combining low-level features.

However, differences in performance may arise depending on how the modalities are combined. For example, the bimodal model of Savran et al. [5] uses Decision-Level (DL) fusion, in which unimodal models are built separately and their individual predictions are then combined. This bimodal model outperforms both unimodal models. However, the WLSC baseline model [3] uses Feature-Level (FL) fusion, in which the audio and visual features are concatenated and a single classifier is built from this combined feature set. This bimodal model only outperforms the worse performing unimodal model (the visual model). This suggests that applying feature engineering methods (e.g., Principal Components Analysis (PCA) and Correlation-based Feature-subset Selection (CFS)), to the concatenated feature set may improve the performance of the bimodal model by reducing the drawbacks of the less predictive features and increasing the benefits given by the more predictive features.

To sum up, in this work, we test the following three hypotheses:

1. *Using high-level features will improve the performance of emotion recognition compared to using low-level features.*
2. *Fusing modalities by concatenating feature sets will give better results compared to unimodal models.*
3. *Applying feature engineering to the concatenated feature set will improve performance of the bimodal model.*

The rest of this paper is organised as follows: In Section 2, we introduce the AVEC2012 database and our experimental setup. Section 3 presents the results of regression experiments using different feature sets. Section 4 provides general discussion and future directions for our work. Section 5 contains the conclusions.

2 Method

2.1 The AVEC2012 Challenge

The Database and Its Definition of Emotions

We use the AVEC2012 database [3] in the following experiments. It includes audio-visual recordings and manual transcriptions with word timings of 24 subjects conversing with 4 virtual agents, which were collected as part of the SEMAINE corpus [12]. Each agent is designed with a different personality, namely even-tempered Prudence, happy Poppy, angry Spike, and depressive Obadiah. Conversations were conducted in a Wizard-of-OZ setup. Topics of conversation varied from daily life to political issues. The 24x4 recordings are divided into training set, development set, and test set, each of which contains 32 dialogue sessions. In the AVEC database, subjects in the test set are different people from those in the training and development sets. For the WLSC, each word spoken by a subject is a data instance. The number of instances contained in the training set, development set, and test set are 20169, 16300, and 13405, respectively.

The AVEC2012 database uses real-value vectors in the Arousal-Expectancy-Power-Valence (AEPV) space to represent emotions. Arousal represents the activeness of the subject; Expectancy represents the predictiveness the subject feels towards the conversation; Power represents the degree of dominance the subject feels over the conversation; Valence represents the positiveness the subject feels towards the conversation [13]. For example, using this representation, we may describe the emotional state of someone who has just been informed that she has won the best paper award as $\boldsymbol{a}$ = (0.6, -0.3, -0.1, 0.9), which means she is excited (A = 0.6) about this great (V = 0.9) news, and cannot stop herself (P = -0.1) from jumping up and down at this surprise (E = -0.3). The AEPV emotional space is capable of describing most of our everyday emotions [13]. The original emotion annotations in the AVEC2012 database had different value ranges for the four dimensions. In our work, we rescale all the AEPV values into the range [-1, 1].

Baseline Audio and Visual Features Provided by the Challenge

The AVEC2012 baseline SP feature set provides 1842 audio features, including pitch, energy, voicing, spectral related low-level descriptors, and voiced/unvoiced duration features, which were extracted from the users' speech using OpenSMILE [14] over words.

The AVEC2012 baseline LBP feature set provides 5908 features related to the size and position of the facial regions, as well as LBP descriptors. Faces in frames are detected using OpenCV's Viola and Jones face detector [15], then aligned by eye locations.

The Best Performing WLSC Model

The model that won the 2012 WLSC, proposed by Savran et al. [5], uses a subset of the baseline audio and visual features, together with lexical features they extracted from the transcripts provided.

Their lexical features were computed using Pointwise Mutual Information (PMI) values, which are measurements of the correlations between words and binarized emotion dimensions. They extracted two types of lexical features, i.e., sparse PMI features using a 1000-dimension bag-of-words approach, and non-sparse PMI features using word counts. These were the only lexical features used in the AVEC2012. Their experiments showed that the sparse PMI features gave significantly better results than the non-sparse PMI features, and were the most predictive features. Since our disfluency features are also extracted from the transcripts, performance of our disfluency feature model will be compared to their lexical model.

2.2 Disfluency Features

In our work, we extract 6 novel disfluency features from the manual transcripts and word-timings provided by the challenge. Each of our disfluency features describes one type of disfluency as described in the following list:

1. **Filled pauses:** non-lexical sounds people make when speaking. For example, "Hmm" in the utterance "Hmm... Maybe we should try another road." The three most common filled pause words in the AVEC database are "em", "eh", and "oh".
2. **Fillers:** phrases used by speakers when they pause to think but they still want to hold the turn. For example, "you know" in the utterance "I just want to, you know, get a drink and forget all about it.". The three most common fillers in the AVEC database are "well", "you know", and "I mean".
3. **Stutters:** words or parts of words the speaker involuntarily repeats when speaking. For example, "Sa" in the utterance "Sa Saturday will be fine.", or the first "I didn't" in the utterance "I didn't, I didn't mean it."
4. **Laughter:** sounds labelled as ⟨LAUGH⟩ in the transcripts provided with the AVEC challenge.
5. **Breath:** sounds labelled as ⟨BREATH⟩ in the transcripts.
6. **Sigh:** sounds labelled as ⟨SIGH⟩ in the transcripts.

We note that this is only a subset of the types of disfluencies that are studied in natural language. For example, content based repairs ("I went hiking on Saturday...no, Sunday.") were not annotated. We choose to use these 6 types because they are the most common disfluencies occurring in the corpus and they are relatively easy to detect from transcripts alone.

Filled pauses, fillers, and stutters were detected semi-automatically: we first ran a keyword search for known disfluency words on the transcripts, then manually checked the annotation results to reduce mistakes such as annotating the "well" in "It works well" as a filler. For laughter, breath, and sighs, we use the annotations provided by the challenge, which were manually labelled. In this paper, we use the manually corrected "gold standard" disfluency features to establish whether our disfluency features are good predictors for emotion recognition. If so, we plan to develop methods to detect disfluencies fully automatically.

To calculate the disfluency features, we used a moving window, which contains the target word and its 14 preceding words, as shown in Figure 1. Our window works on dialogue sessions, and it slides word by word, until it reaches the end of a session. The window for w_n contains w_{n-14} to w_n when $n > 15$. When $n \leq 15$, a window that contains w_1 to w_{15} is used. We chose a window length of 15 words because this is the average length of a speaker turn, and the emotional states of the words within a speaker turn are often highly correlated.

We compute our disfluency features using equation (1):

$$D_i = \frac{t_d}{T_i} \tag{1}$$

Here, D_i is the disfluency feature D of the i-th word; t_d is the total duration of disfluency type D within the window of this word; T_i is the total utterance length

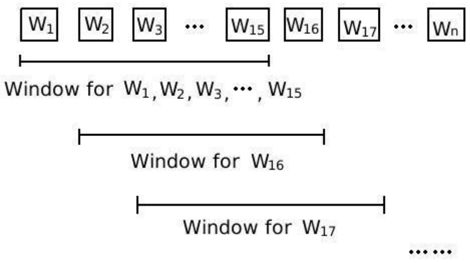

Fig. 1. The moving window

of all 15 words within this window. The reason we use durations of disfluencies instead of their counts is that duration of disfluency features also contains clues for emotions, e.g., saying “Hmm…” (longer duration) may indicate that the speaker feels more uncertain than saying “Hmm” (shorter duration).

One issue with our disfluency features is that some types of disfluency are very infrequent in the data. The percentages of non-zero values of our disfluency features are shown in Table 1 (before speaker normalization is applied). We can see that some types of disfluency are very sparse, with non-zero values occurring less than 10% of the time. However, these infrequent disfluency features may also contribute to emotion recognition by indicating where the also infrequent periods of strong emotions are located. This may help to predict the emotional values of time periods that are highly emotional, as well as those of time periods that are calm and neutral.

Table 1. Frequency of non-zero values of our disfluency features

Data set	Filled Pause	Filler	Stutter	Laughter	Breath	Sigh
train(%)	45.1	16.0	12.2	8.2	2.9	0.6
devel(%)	37.3	20.7	12.0	10.7	1.5	0.5
tests(%)	34.2	20.9	12.4	9.6	3.0	0.1

We examined two additional types of features: silent pause features, which are calculated by finding silent gaps between word timings, and window-based PMI features. However, our experiments showed that these features are not as predictive as our disfluency features. Therefore, these features and the experiments related to them will not be further discussed in this paper.

2.3 The ASM Visual Features

In our work, we use the horizontal and vertical positions of 77 facial landmarks as our visual features. The face detection and eye alignment procedure is similar to that employed in the AVEC2012 baseline visual feature extraction. To locate the facial landmarks on detected facial regions, the Active Shape Model (ASM) [16] is used. In ASM, a model for the shape of an object is first constructed from the training samples based on the geometric features calculated using PCA. This model is then applied to the test sample and iteratively fit to it. The reason we chose ASM instead of the Active Appearance Model (AAM) used by the best [4] and the second best [9] FCSC models, is that, in general, ASM works better than AAM when the test subjects are different from the training subjects [17]. In our case, we used an existing face model that is trained on the MUCT Face Database [18], so ASM is a more reasonable choice here.

The STASM (ASM with SIFT descriptors) tool [19] was used to automatically locate facial landmarks. Horizontal and vertical locations of these points are shown in Figure 2. This gives us a 154-dimension vector representing the facial expressions in each frame. We then compute a mean representation over all the frames within the duration of a word. The same moving window used in the disfluency feature extraction is applied again and the mean representation of each word within the window is concatenated. This leads to our ASM visual feature set of 2310 (154x15) features.

2.4 Speaker Normalization

We applied z-score speaker normalization to all our features to reduce the influence of speaker variance, as follows.

$$V_a^{'} = \frac{V_a - \bar{V_a}}{Std_a} \tag{2}$$

$V_a^{'}$ is the normalized value of an attribute a; V_a is the original value of attribute a; $\bar{V_a}$ is the mean value of attribute a over all the samples extracted from the speaker; Std_a is the standard deviation of values of attribute a. Speaker normalization is applied after grouping the data by speaker.

2.5 Modality Fusion and Feature Engineering

In our work, we applied a FL modality fusion method and concatenated our disfluency and ASM visual feature sets into one set. Simple concatenation without any further feature engineering is referred to as Basic-FL in the following. The feature set used by our Basic-FL model contains 2316 features.

We also study the influence of two feature engineering methods on the Basic-FL model. The first method is PCA, which maps the original features to a lower dimensional space, thus reducing the size and redundancy of the feature set. After reserving 99% of the total variance, the new feature set contains 59 transformed features.

The second method is CFS [21], which ranks features based on their predictiveness and correlation with other features. The predictiveness of features is evaluated by building single feature classifiers. Features with the best performance and low redundancy are iteratively added to our starting set containing

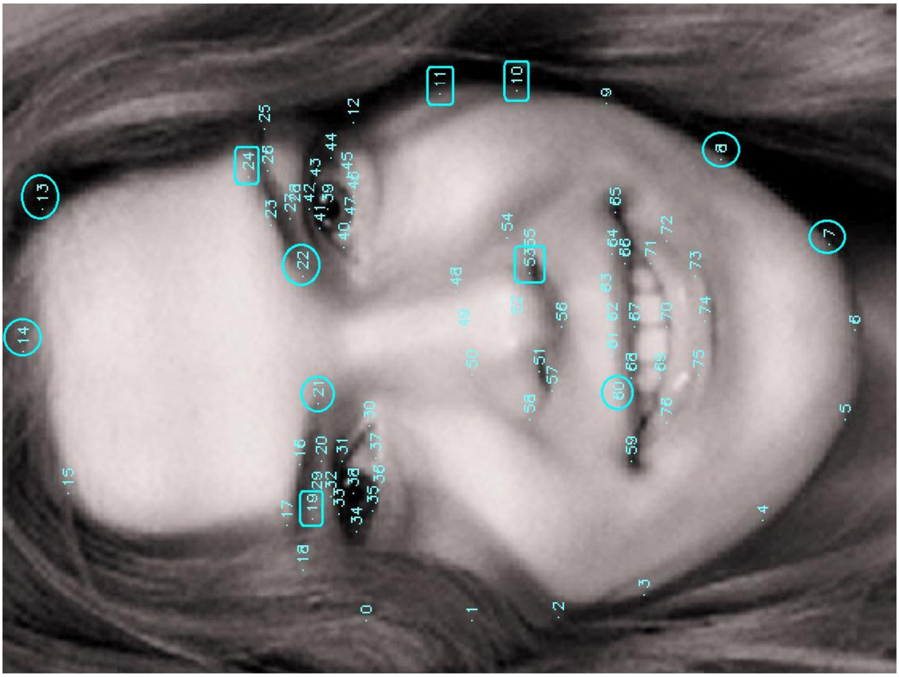

Fig. 2. Locations of 77 facial landmarks [20]

the 6 disfluency features, until the performance decreases. The CFS-FL model uses different feature sets when predicting different emotion dimensions, and there are at most 22 features (for Valence) in these subsets. The variance in the number and features contained in the subsets for predicting different emotion dimension also highlights the varying relationship between features and these dimensions.

Taking a closer look at the feature set selected and referring to the annotations of facial landmarks shown in Figure 2, we also find that the small number of visual features selected by the CFS method often represents facial expression changes within the moving window. For example, the 8 visual features selected when predicting Arousal values are w_1y_{21}, w_3y_{22}, w_5y_{21}, w_7y_{21}, w_9x_{24}, $w_{10}y_{22}$, $w_{12}y_{22}$, $w_{15}y_{21}$. Here, w_i represents the i-th word in the window. We use w_ix_j and w_iy_j to represent the averaged horizontal and vertical positions of the facial landmark numbered at j during word w_i. In Figure 2, we can see that the facial landmarks No.21 and No.22 are the inner corners of eyebrows. This subset of visual features mainly describes the vertical movements of these two key facial points within our window. Other important facial landmarks are also labelled in Figure 2. Those marked by circles indicate that the vertical movement of this point is used, and those marked by squares indicate that the horizontal movement is used.

2.6 Regression and Evaluation Metric

We use Support Vector Regression (SVR [22]) as our regression method for comparability with the AVEC2012 baseline and top performing models. Following the settings of the best WLSC model [5], we implemented epsilon-SVR with a linear kernel using the LibSVM [23] toolbox on the WEKA [24] interface. Before regression begins, all features are normalized to range [0, 1] in the regression models. The AEPV values are predicted independently. We use training, development, and test sets as set out by the AVEC guidelines.

In the AVEC2012 challenge, Cross-Correlation Score (CCS), which is the average of correlation-coefficients of all 32 test sessions, was defined as the evaluation metric. The value range of CCS is from 0 to 1, with higher scores representing better performance. In this paper, we evaluate significance of differences between CCS scores using a two-tailed z-test after Fisher's r-to-z transformation.

3 Experiments and Results

Results of our experiments are shown in Table 2. In Table 2, DF is our disfluency feature model; PMI is the sparse PMI feature model used by Savran et al. [5]; SP is the AVEC2012 WLSC baseline audio model using SP features; ASM is our ASM visual feature model; R-LBP is the dimensionality-reduced LBP model used by Savran et al. [5]; LBP is the AVEC2012 WLSC baseline visual model using LBP features; B-FL, P-FL, C-FL are our Basic-FL, PCA-FL, CFS-FL bimodal models; AV-B, AV-1, AV-2, AV-3 are the AVEC2012 WLSC baseline, the best [5], the second best [25], and the third best [26] audio-visual (AV) models.

3.1 The Disfluency Feature Model

As shown in Table 2, compared to all the other unimodal models, our disfluency feature model has the best performance when predicting all four emotion dimensions. Its overall performance is not significantly different from the best result of WLSC ($p = 0.424$).

As shown in Figure 3, our disfluency feature model outperforms the baseline model using SP features. Comparing our disfluency features with the PMI-based lexical features used by Savran et al. [5], our features look at data from a higher level, and give significantly better performance when predicting all emotion dimensions. This is consistent with our hypothesis that using high-level features will improve model performance.

Note that we only have 6 disfluency features, while there are 1842 SP features and 1000 PMI features. This huge difference in dimensionality will influence the efficiency of the emotion recognizer greatly, and lower dimensional features are often preferred, especially in real-time interactive systems. Therefore, our high-level disfluency features are more predictive and more efficient.

We also compared the predictiveness of different types of disfluency using the rank generated by the CFS method (see Section 2.5). Results are shown in

Table 2. Experimental results

Models	A	E	P	V	Mean
DF	0.250	0.313	0.288	0.235	0.271
PMI	0.131	0.285	0.254	0.188	0.214
SP	0.014	0.038	0.016	0.040	0.027
ASM	0.205	0.246	0.172	0.219	0.211
R-LBP	0.184	0.156	0.146	0.226	0.178
LBP	0.005	0.012	0.018	0.005	0.011
B-FL	0.205	0.274	0.223	0.207	0.227
P-FL	0.214	0.268	0.269	0.225	0.244
C-FL	0.274	0.258	0.266	0.215	0.253
AV-B	0.021	0.028	0.009	0.004	0.015
AV-1	0.302	0.194	0.293	0.331	0.280
AV-2	0.210	0.240	0.289	0.208	0.237
AV-3	0.267	0.241	0.223	0.138	0.192

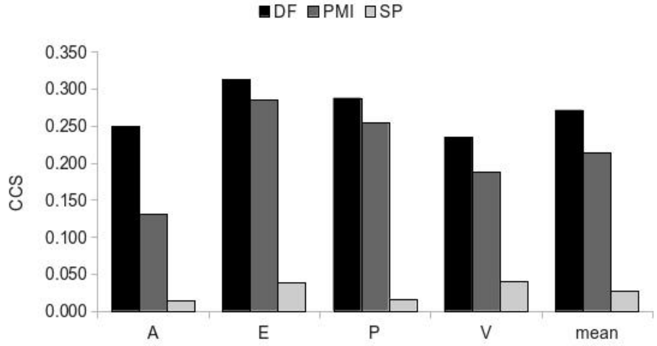

Fig. 3. The disfluency feature model

Table 3. The rank of a disfluency feature when predicting a particular emotion dimension is from 1.0 to 6.0, with 1.0 representing the highest predictiveness. The results indicate that filled pauses and laughter are the most predictive disfluency features in this task.

3.2 The ASM Visual Feature Model

As seen in Figure 4, for all four emotion dimensions our high-level ASM visual features are more predictive than the pixel-level LBP features extracted from the whole facial region. Our model also outperformed the feature-selected LBP model on most of the emotion dimensions. On the Valence dimension, our model has slightly lower CCS, but the difference is not significant ($p = 0.549$). These results verified our hypothesis that our high-level ASM features are more predictive then the low-level LBP features.

Table 3. Predictiveness rank of different disfluency features

Disfluency	A	E	P	V	mean
Filled pause	1.0	2.0	1.0	2.0	1.5
Filler	6.0	5.0	5.0	6.0	5.5
Stutter	5.0	6.0	6.0	3.0	5.0
Laughter	2.0	1.0	2.0	1.0	1.5
Breath	4.0	3.0	4.0	5.0	4.0
Sigh	3.0	4.0	3.0	4.0	3.5

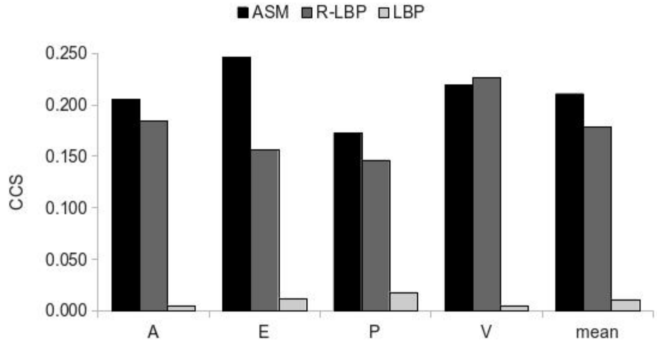

Fig. 4. The ASM visual feature model

3.3 The Bimodal Models

The performance of our unimodal models (DF and ASM) and our bimodal models (B-FL, P-FL, and C-FL) is shown in Figure 5. As we can see, our disfluency feature model outperforms our ASM visual model on all emotion dimensions. After simple concatenation of the feature sets, the increase on mean CCS of the lower-performing visual modality is not significant ($p = 0.168$). Recall that there are only 6 disfluency features, while there are 2310 ASM visual features. This suggests that the large visual feature set is dominating and introduces noise into the model. We can see that the PCA-FL model and the CFS-FL model perform better than the Basic-FL model in general, thus applying feature engineering to the concatenated feature set helps to reduce the influence of noisy visual features. These two feature-engineered bimodal models both give significant improvements on mean CCS compared to the lower-performing visual model. However, compared to the Basic-FL model, only CFS gives significant improvement on mean CCS ($p = 0.024$).

We also compared our best bimodal model, the CFS-FL model, with the baseline and the best three models of AVEC2012 WLSC. As shown in Figure 6, performance of our bimodal model is significantly better than the AVEC2012 WLSC baseline model when predicting all four dimensions of emotions, as expected.

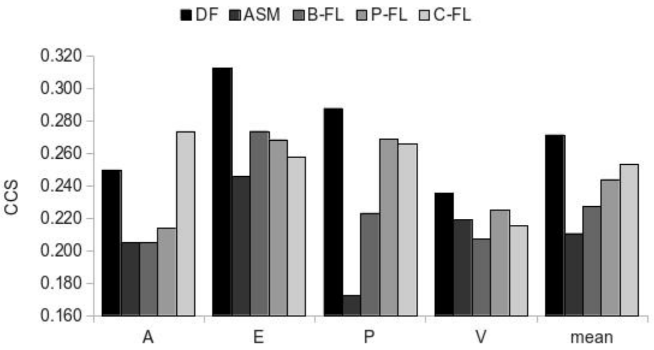

Fig. 5. Modality fusion and feature engineering

Comparing the overall performance (mean CCS), our model is significantly better than the third best WLSC model [26]. Our model also outperformed the second best WLSC model [25], but the difference is not significant ($p = 0.165$). When predicting Expectancy values, our model gives the highest CCS. The reason may be that Expectancy is the easiest dimension to predict for both our disfluency and ASM features, as shown in Figures 3 and 4.

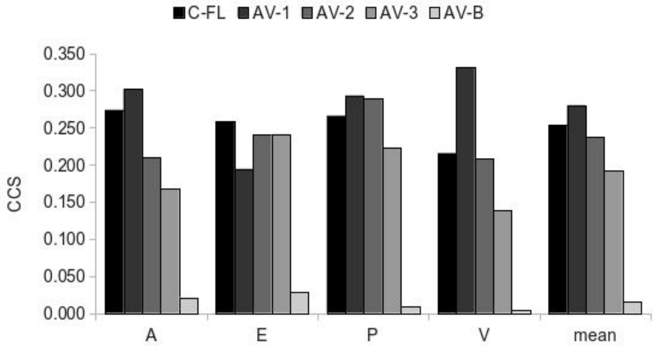

Fig. 6. Comparing our bimodal model with the AVEC2012 WLSC results

4 Discussion

Based on the experimental results, we verified our main hypothesis that using high-level features, namely the disfluency features and the ASM features, improves model performance compared to using low-level features.

The average performance of our disfluency model using only the 6 disfluency features is not significantly different from the best audio-visual model in AVEC2012 WLSC. This indicates that disfluency features are especially powerful in emotion prediction. Our disfluency model also outperformed the PMI model used by Savran et al. [5]. This suggests that learning other high level classes of lexical features may be useful for this task.

In the future, we plan to study whether the disfluency features are still highly predictive of emotion when using other corpora. The utility of disfluencies also depends on how well they can be detected. Further work will investigate performance using disfluencies detected from the output of an automatic speech recognizer, rather than manual transcription. Similarly, integrating work on developing a fully automatic disfluency detection method based on existing studies should be helpful for this task. For example, Liu et al [27] use a Hidden-Markov Model that combines textual and prosodic clues to detect disfluencies. The work of Niewiadomski et al. [28] also highlights the importance of automatic laughter detection.

The visual feature subset selected by the CFS method illustrates a way to compute visual features that also have longer duration. In the future, we will use the position changes of a subset of the 77 facial landmarks over the window as visual features, thus further reducing the dimensionality of the visual feature set.

Our experimental results also verify that fusing modalities can give improvements compared to the lower-performing unimodal model. However, it is difficult for FL fusion models to improve on the better-performing unimodal model. The fact that feature-engineered models do not provide huge gains may be due to a lack of control in feature weighting and, as such, DL fusion may be more appropriate for the task. Compared to FL fusion, DL fusion has the natural advantage of flexibility when weighting different modalities. In the future, we plan to apply DL fusion to our model and study whether or not it improves performance.

Finally, the low CCS of all models in the AVEC2012 may indicate that CCS is not the best evaluation metric for this task. CCS evaluates average performance of the classifier for predicting values of all data in the corpus. However, occurrences of strong emotions are relatively rare in conversations, which makes the values of a large portion of the data unsuitable for classifiers that are designed to predict such emotions. Therefore, a more appropriate evaluation metric is needed. One possible alternative would be to detect emotionally strong events first, using methods such as those previously used for the detection of hot spots in meetings [29], and only evaluate model performance on these segments.

5 Conclusions

In this paper, we introduced a new emotion recognition approach that used a small number of human-interpretable high-level features. Our unimodal and bimodal models built using these features have significantly better performance compared to the baseline models, which used a large number of low-level audio, visual, or lexical features. In fact, our models have the best performance for predicting the Expectancy dimension of emotion compared to all AVEC2012 competitors. Using only 6 disfluency features, we built a model with performance not significantly different from the best AVEC2012 bimodal model overall. Previous studies on automatic disfluency detection also give us reason to believe that these features can be computed automatically in real or near-real time with reasonable accuracy. This in turn would allow the development of a fast and accurate emotion classifier which holds promise for future applications in interactive systems.

Acknowledgments. We are grateful to Dr. Ani Nenkova for discussion about her research [5], and to Dr. Michel Valstar for answering questions about the AVEC2012 challenge and database.

References

1. D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., Person, N., Kort, B., el Kaliouby, R., Picard, R., et al.: AutoTutor detects and responds to learners affective and cognitive states. In: Workshop on Emotional and Cognitive Issues at the International Conference on Intelligent Tutoring Systems (2008)
2. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th International Conference on Multimodal Interfaces, pp. 205–211. ACM (2004)
3. Schuller, B., Valster, M., Eyben, F., Cowie, R., Pantic, M.: AVEC 2012: the continuous audio/visual emotion challenge. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 449–456. ACM (2012)
4. Nicolle, J., Rapp, V., Bailly, K., Prevost, L., Chetouani, M.: Robust continuous prediction of human emotions using multiscale dynamic cues. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 501–508. ACM (2012)
5. Savran, A., Cao, H., Shah, M., Nenkova, A., Verma, R.: Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 485–492. ACM (2012)
6. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 971–987 (2002)
7. Scherer, K.R.: Expression of emotion in voice and music. Journal of Voice 9, 235–248 (1995)
8. Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422 (2005)
9. Soladié, C., Salam, H., Pelachaud, C., Stoiber, N., Séguier, R.: A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 493–500. ACM (2012)
10. Silva, L.C., Miyasato, T.: Degree of human perception of facial emotions based on audio and video information. IEICE Technical Report. Image Engineering 96, 9–15 (1996)
11. Chen, L.S., Huang, T.S., Miyasato, T., Nakatsu, R.: Multimodal human emotion/expression recognition. In: Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 366–371. IEEE (1998)

12. McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 1079–1084. IEEE (2010)
13. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.C.: The world of emotions is not two-dimensional. Psychological Science 18, 1050–1057 (2007)
14. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the International Conference on Multimedia, pp. 1459–1462. ACM (2010)
15. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I–511. IEEE (2001)
16. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding 61, 38–59 (1995)
17. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Comparing active shape models with active appearance models. BMVC 99, 173–182 (1999)
18. Milborrow, S., Morkel, J., Nicolls, F.: The MUCT landmarked face database. Pattern Recognition Association of South Africa 201 (2010)
19. Milborrow, S., Nicolls, F.: Active shape models with sift descriptors and mars 1, 5 (2014)
20. Milborrow, S.: Stasm User Manual (2013), `http://www.milbo.users.sonic.net/stasm`
21. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD thesis, University of Waikato, Hamilton, New Zealand (1998)
22. Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. Advances in Neural Information Processing Systems, 155–161 (1997)
23. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2, 27 (2011)
24. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11, 10–18 (2009)
25. Ozkan, D., Scherer, S., Morency, L.P.: Step-wise emotion recognition using concatenated-HMM. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 477–484. ACM (2012)
26. van der Maaten, L.: Audio-visual emotion challenge 2012: a simple approach. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 473–476. ACM (2012)
27. Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M.: Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Transactions on Audio, Speech, and Language Processing 14, 1526–1540 (2006)
28. Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., et al.: et al.: Laugh-aware virtual agent and its impact on user amusement. In: Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems, pp. 619–626. International Foundation for Autonomous Agents and Multiagent Systems (2013)
29. Lai, C., Carletta, J., Renals, S.: Detecting summarization hot spots in meetings using group level involvement and turn-taking features. In: Proceedings of Interspeech 2013, Lyon, France (2013)