Comparative Study on Feature Selection and Fusion Schemes for Emotion Recognition from Speech

Santiago Planet and Ignasi Iriondo
La Salle – Universitat Ramon Llull

***Abstract* — The automatic analysis of speech to detect affective states may improve the way users interact with electronic devices. However, the analysis only at the acoustic level could be not enough to determine the emotion of a user in a realistic scenario. In this paper we analyzed the spontaneous speech recordings of the FAU Aibo Corpus at the acoustic and linguistic levels to extract two sets of features. The acoustic set was reduced by a greedy procedure selecting the most relevant features to optimize the learning stage. We compared two versions of this greedy selection algorithm by performing the search of the relevant features forwards and backwards. We experimented with three classification approaches: Naïve-Bayes, a support vector machine and a logistic model tree, and two fusion schemes: decision-level fusion, merging the hard-decisions of the acoustic and linguistic classifiers by means of a decision tree; and feature-level fusion, concatenating both sets of features before the learning stage. Despite the low performance achieved by the linguistic data, a dramatic improvement was achieved after its combination with the acoustic information, improving the results achieved by this second modality on its own. The results achieved by the classifiers using the parameters merged at feature level outperformed the classification results of the decision-level fusion scheme, despite the simplicity of the scheme. Moreover, the extremely reduced set of acoustic features obtained by the greedy forward search selection algorithm improved the results provided by the full set.**

***Keywords* — Acoustic and linguistic features, decision-level and future-level fusion, emotion recognition, spontaneous speech**

I. Introduction

ONE of the goals of human-computer interaction (HCI) is the improvement of the user experience, trying to make this interaction closer to human-human communication. Inclusion of speech recognition was one of the key points to include "perception" to multimedia devices. This improved their user interfaces [1]. However, the analysis of affective states by the study of the implicit channel of communication (i.e. the recognition of not only what is said but also how it is said) may improve HCI making these applications more usable and friendly. This is because, in general, inclusion of skills of emotional intelligence to machine intelligence makes HCI more similar to human-human interaction [2]. There is a wide range of contexts where the analysis of speech and emotion in the input of the systems –and also the synthesis of emotional speech at the output– can be applied to, including automatic generation of audio-visual content, virtual meetings, automatic dialogue systems, tutoring, entertainment or serious games.

There are many studies related to emotion recognition based on different approaches. However, a big amount of these works are based on corpora consisting of utterances recorded by actors under supervised conditions. Nowadays this is not the current trend because of the lack of realism of these data [3].

The first study where authors attempted to work with a corpus of spontaneous speech seems to be [4], collecting utterances from infant directed speech. Many other works tried to deal with realistic data, such as [5] and [6]. Nevertheless, it is difficult to compare the results of these approaches when they are using different data and different evaluation methods. A framework to generalise the research on this topic was proposed by [7]. This framework was based on a corpus of spontaneous speech where two different subsets were defined in order to allow speaker-independence during the analysis. Speech was non-acted and, for this reason, utterances were characterised by being non-prototypical and having low emotional intensity. Results obtained within this framework [8] give an idea of the complexity of the task. The combination of 7 classification approaches considering different sets of features achieved 44.00% of unweighted average recall (UAR). We worked under the same naturalistic conditions in this article.

The task of emotion recognition from speech can be tackled from different perspectives [3]. We considered the analysis of two modalities: the acoustic (referred to the implicit message) and the linguistic (referred to the explicit message), extracting acoustic parameters from the speech signal and linguistic features from the transcriptions of the utterances of the corpus. Because in a realistic scenario the analysis of acoustic information could be not enough to carry out the task of emotion recognition from speech [9] the linguistic modality could improve an only-acoustic study. In this article, both modalities were combined at the decision level and at the feature level to compare the performance of different classification approaches using both procedures. To improve

DOI: 10.9781/ijimai.2012.166

the performance of the classifiers and optimize the experiment we reduced the acoustic set of features (the largest one) by selecting the most relevant parameters by a greedy algorithm before starting the learning stage. Also, for this feature selection stage, we compared two search methods (forwards and backwards) through the space of feature subsets.

This paper is structured as follows: Section II describes the corpus and details its acoustic and linguistic parameterization. Section III defines the methodology of the experiment, describes the feature selection algorithms used to optimize the acoustic set of data and details the two fusion schemes proposed. Section IV summarises the results. Conclusions are detailed in Section V.

II. CORPUS

This work was based on the FAU Aibo Corpus [10] as it was defined in [7]. In this Section we describe this corpus and its acoustic and linguistic parameterization.

A. Corpus Description

The FAU Aibo Corpus consisted of 8.9 hours of audio recordings of German speech from the interaction of children from two schools playing with the Sony's Aibo robot in a *Wizard of Oz* (WOZ) scenario. These audio recordings were divided into 18,216 chunks. A chunk is each one of the segmentations of the audio recordings of the corpus into syntactically and semantically meaningful small parts. These parts were defined manually following syntactic and prosodic criteria [10]. The chunks of the two schools were divided into two independent folds (fold 1 and fold 2) to guarantee speaker-independence. Thus, each fold contained speech recordings from different children. Each chunk, after parameterization, was considered an instance of the datasets used to train and test the classification schemes. The number of resulting instances was 9,959 for the fold 1 and 8,257 instances for the fold 2. The emotions considered to label the corpus were defined by these five category labels: Anger (A), including angry (annoyed), touchy (irritated as a previous step of anger) and reprimanding (reproachful); Emphatic (E) (accentuated and often hyper-articulated speech but without sentiment); Neutral (N); Positive (P), which included motherese (similar to infant-directed speech but from the child to the robot) and joyful states; and Rest (R), a garbage class collecting three affective states: surprise (in a positive sense), boredom (with a lack of interest in the interaction with the robot) and helpless (doubtful, speaking using disfluencies and pauses).

Because of the use of a WOZ scenario to record the affective states of the children, the corpus collected spontaneous utterances of naturalistic emotional speech in a real application environment. For this reason, it included non-prototypical emotions of low intensity. Moreover, the distribution of the emotion labels was very unbalanced. For example, the majority class (N) consists of 10,967 utterances (60.21% of the whole corpus) while the minority class (P) consists of only 889 utterances (4.88% of the whole corpus). For a full description of this corpus cf. [7].

B. Acoustic Parameterization

The acoustic analysis of the corpus consisted on calculating 16 low-level descriptors (LLDs). These LLDs were: the zero-crossing rate (ZCR) analysed in the time signal, the root mean square (RMS) frame energy, the fundamental frequency (F0) normalised to 500 Hz, the harmonics-to-noise ratio (HNR) and 12 mel-frequency cepstral coefficients (MFCC). We also computed the derivative of these LLDs.

We calculated 12 functionals from these LLDs and, also, from their derivatives. These functionals were: the mean, the standard deviation, the kurtosis and the skewness, the value and range and position of the extremes, and the range and two linear regression coefficients with their mean square errors (MSE).

To perform this parameterization we used the openSMILE software included in the openEAR toolkit release [11], obtaining $16 \times 2 \times 12 = 384$ features per instance.

C. Linguistic Parameterization

The linguistic parameterization was based on the transcriptions of the corpus. These transcriptions defined the words that children used to communicate with the robot Aibo. We used the concept of emotional salience proposed by [12] to translate the words of a chunk into 5 emotion-related features. Assuming independence between the words of a chunk, the salience of a word is defined as the mutual information between a specific word and an emotion class. Therefore, an emotionally salient word is a word that appears more often in that emotion than in the other categories. Considering this definition, let $W = \{v_1, v_2, ..., v_n\}$ be the n words of a chunk and let $E = \{e_1, e_2, ..., e_k\}$ be the emotional space defined by a set of k emotion classes. Mutual information between the word v_m and the emotion class e_j is defined by (1).

$$i(v_m, e_j) = \log \frac{P(e_j \mid v_m)}{P(e_j)} \tag{1}$$

where $P(e_j|v_m)$ is the posterior probability that a chunk containing the word v_m implies the emotion class e_j and $P(e_j)$ is the a priori probability of the emotion e_j.

The emotional salience of the word v_m related to the emotional space E is defined by (2).

$$sal(v_m) = \sum_{j=1}^{k} P(e_j \mid v_m) i(v_m, e_j) \tag{2}$$

We calculated the emotional salience of all the words of the training dataset retaining only those with a value greater than a threshold empirically chosen at 0.3. This resulted in a list of emotionally salient words. Next, we calculated 5 linguistic features for each chunk. These features, called activations and

denoted by a_j, were calculated following (3) [12].

$$a_j = \sum_{m=1}^{n} I_m i(v_m, e_j) + \log P(e_j) \quad (3)$$

where I_m is 1 if the word matches the list of salient words or 0 otherwise.

To guarantee the independence of the two folds during the parameterization stage, the list of emotionally salient words was created considering only the fold used for training. Next, we calculated the activation features for both folds but using only the emotional salience values and the a priori probabilities from the training fold. By following this procedure the test data remained unseen during the analysis of the training data to extract the information about the emotional salience of the words of the corpus.

III. EXPERIMENTATION

In this Section we explain the methodology of the experiment, the feature selection algorithms used to reduce the acoustic set of features and the two procedures to fusion the acoustic and linguistic modalities.

A. Methodology

The acoustic feature vector contained a big amount of information (384 features), being much larger than the vector of linguistic parameters (5 features). The inclusion of irrelevant features in the space of parameters could deteriorate the performance of the classifiers used in the learning stage [13]. Moreover, if these data were merged with the linguistic features without any previous processing then the resulting vectors would be very unbalanced because they would contain many more features related to the acoustic information than features related to the linguistic information.

Feature selection techniques are designed to create subsets of features without redundant data by discarding irrelevant input variables with little predictive information. These reduced subsets could improve the performance of the classifiers and obtain a more generalizable classification model [14]. We used a wrapper method [15] to evaluate the candidate subsets created by a search algorithm and two ways of searching the feature space to create these subsets, as it is explained in detail in Section III.B.

In the classification stage, we considered two procedures to fusion the acoustic and the linguistic data. On the one hand, we performed a decision-level fusion of these modalities classifying the acoustic and the linguistic data independently and merging the classification results by a third classifier. On the other hand, we used a feature-level fusion procedure merging the acoustic and the linguistic parameters before the classification stage. These procedures are detailed in Section III.D and Section III.E, respectively.

We evaluated the classifier schemes in a 2-fold cross-validation manner. We used one fold for training and the other fold for testing and vice versa. This allowed us to guarantee speaker-independence in the experiment. The mean value of the performances of both folds was also calculated.

We considered three learning algorithms in this experiment using the implementations provided by the WEKA data mining toolkit [13]. The first learning algorithm was a Naïve-Bayes (NB) classifier. This algorithm was found to be the most relevant in [16] despite its simplicity. For this reason it was used as the baseline in this experiment. To improve the performance of this classifier we applied, prior to the training stage, a supervised discretisation process based on the Fayyad and Irani's Minimum Description Length (MDL) method [17]. The second classification approach was a support vector machine (SVM) classifier. For this work, we chose a SVM with a linear kernel using sequential minimal optimisation learning [18]. To allow the algorithm to deal with a problem of five classes we used pairwise multi-class discrimination [19]. Finally, the third classifier was a logistic model tree as described in [20]. This is a model tree using logistic regression at the leaves instead of linear regression. This is named Simple Logistic (SL) in WEKA.

We used the UAR measure to compare the performances of the classification approaches because the distribution of the classes in the FAU Aibo Corpus was very unbalanced. Comparing the UAR of the classifiers, instead of the weighted-average recall (WAR) measure, the most even class-wise performance was intended. Thus, the same importance was given to the majority and the minority classes of the corpus because we considered the detection of the interactions with emotional content as important as the detection of the neutral interactions. However, in most of other studies of emotion recognition the WAR measure was used because the distribution of the classes of their corpora was usually quite balanced. Equation (4) shows that the recall for one class *c* is calculated as the proportion of correctly classified cases (True Positives) with respect to the corresponding number of instances (True Positives and False Negatives) of this class. Equation (5) shows the computation of UAR performance of a classifier considering the recalls of each class *c*.

$$recall_c = \frac{TP_c}{TP_c + FN_c} \quad (4)$$

$$UAR = \frac{\sum_{c=1}^{|C|} recall_c}{|C|} \quad (5)$$

where *TP* stands for True Positives, *FN* stands for False Negatives and |*C*| represents the number of classes.

B. Feature Selection Process

To reduce the set of acoustic features we chose a wrapper method. A wrapper method uses a learning algorithm to evaluate the subsets created by a search algorithm. These subsets are the candidates to be the optimal ones. We considered the Naïve-Bayes classifier to assess the goodness-

of-fit of the candidate subsets. We searched the space of features by means of two greedy procedures to automatically create these subsets:

- Greedy forward (FW) search. This algorithm carried out an iterative exhaustive search through the feature space creating subsets starting with no features and adding one parameter at each iteration.
- Greedy backward (BW) search. In this case, the iterative exhaustive search consisted on creating subsets starting with all the features and discarding one at each iteration.

Before starting the feature selection stage we resampled the fold 1 reducing it by half to speed up the process and biased it to a uniform distribution. To guarantee independence between both datasets, we used only the fold 1 to select the candidate subsets of features and evaluated them on all the instances of this fold.

In the case of the FW search, the acoustic dataset was reduced from 384 features to 28 features: 21 related to the MFCC parameters, 3 related to the RMS frame energy, 2 related to the F0, 1 related to the HNR and 1 related to the ZCR. The BW search was a more conservative approach and created a set of 305 features.

A comparison of the performances of the classifiers using the full set of acoustic features and the reduced sets is shown in Fig. 1. For each algorithm we show three results: the Fold 1 column indicates the results obtained when training the classifiers with the fold 1 and testing with the fold 2, the Fold 2 column is the opposite and the Mean column is the mean of the previous results. As it can be observed, focusing on the mean values of the Fold 1 and Fold 2 experiments and except the case of the Naïve-Bayes classifier, UAR values were slightly better for the reduced sets than for the full set of features. In the case of the Naïve-Bayes classifier, the dataset created by the FW search degraded dramatically the performance of this classifier. Nevertheless, the performance was slightly improved using the dataset created by the BW search. Thus, we chose the reduced sets for this experiment decreasing the computational cost of the classification algorithms.

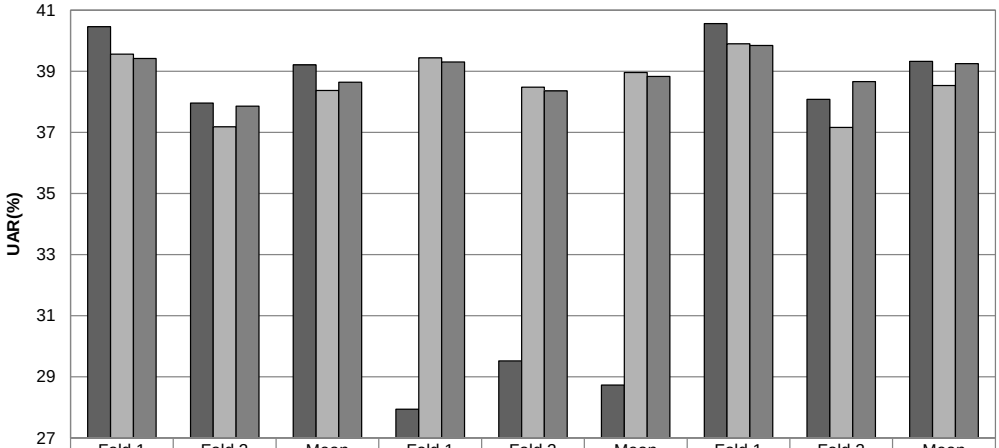
C. Dataset Pre-processing

To optimize the performance of the classifiers we pre-processed the datasets used to train them. Datasets were biased to a uniform class distribution by means of a resampling with replacement technique and duplicating the total number of instances. We did not bias the distribution of classes in the case of the Naïve-Bayes algorithm because this process degraded its performance. In the case of the SVM, data was also normalised by the Euclidean norm.

D. Decision-Level Fusion

Decision-level fusion is based on the processing of the classification results of prior classification stages. The main goal of this procedure is to take advantage of the redundancy of a set of independent classifiers to achieve higher robustness by combining their results [21].

In this experiment, decision-level fusion was performed by combining hard decisions from the classifiers that were trained and tested by the acoustic and linguistic features independently. Although soft decisions could also be used, hard decision classifiers provide the least amount of information to make their combinations [22]. We followed the stacked generalization strategy introduced by [23] and used a decision tree to merge the classifications obtained by the two classifiers. This stacking approach proved to be useful in the field of emotion recognition in previous works like those by [24] and [25]. The decision tree used to merge the hard decisions of the classifiers was a J4.8 classifier. This is the WEKA implementation of the C4.5 Revision 8 algorithm [13], a slightly improved version of the C4.5, based on entropy



	All acoustic features (384)			FW Selected acoustic features (28)			BW Selected acoustic features (305)		
	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean
NB	40.46	37.96	39.21	27.94	29.52	28.73	40.56	38.08	39.32
SVM	39.56	37.18	38.37	39.44	38.48	38.96	39.90	37.16	38.53
SL	39.42	37.86	38.64	39.30	38.36	38.83	39.84	38.66	39.25

Fig. 1. Unweighted average recall of the classifiers using the full dataset of 384 acoustic features and using the reduced sets of the 28 and 305 acoustic features selected by the greedy forward search and greedy backward search selection algorithms, respectively. NB stands for Naïve-Bayes, SVM stands for Suport Vector Machine and SL stands for Simple Logistic. FW and BW stands for the greedy forward search and greedy backward search selection algorithms, respectively.

information [26].

To train the J4.8 algorithm we trained and tested each one of the three classifiers with the full training sets, both the acoustic and the linguistic. Next, we created a dataset merging the hard decisions of each classifier for both sets of features. This dataset was used to train the J4.8 learning scheme after biasing it to a uniform distribution and duplicating the number of instances. Once more, and as in other stages of this experiment, test data remained unseen during the training process. When the J4.8 classifier was trained, we evaluated the hard decisions of the classifiers tested with the test data, measuring the performance of the full scheme at the end.

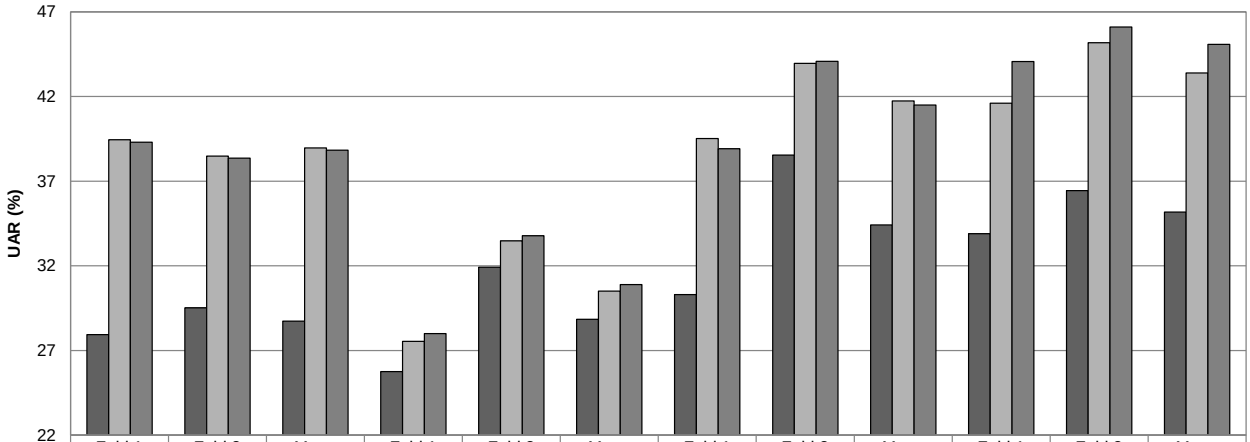
E. Feature-Level Fusion

A feature-level fusion scheme integrates unimodal features before learning concepts, as it is described in [27]. The main advantage of a feature-level fusion scheme is the use of only one learning stage. Moreover, this fusion scheme allows taking advantage of mutual information from data. We used concatenation of the reduced set of acoustic features and the linguistic set to create a multimodal representation of each instance. Thus, the amount of features for the merged dataset was of 33 elements per instance.

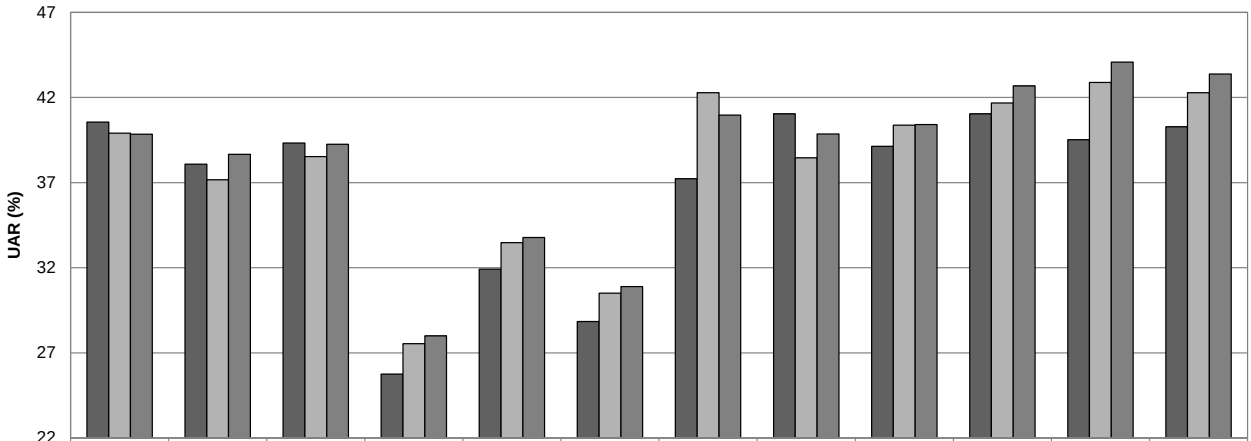
IV. RESULTS

Results of this experiment are shown in Fig. 2. Like in Fig. 1, for each algorithm we show three results: the Fold 1 column indicates the results obtained when training the classifiers with the fold 1 and testing with the fold 2, the Fold 2 column is the opposite and the third result is the mean of the previous results. The results obtained by the dataset created by the FW search procedure are shown at the top and the results achieved by the BW search dataset are shown at the bottom.

Focusing on the mean value of the two folds, it can be observed that the performance of the classifiers that only used



	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean
	FW Selected acoustic features (28)			Linguistic features (5)			Decision-level fusion (FW)			Feature-level fusion (FW)		
NB	27.94	29.52	28.73	25.76	31.92	28.84	30.30	38.54	34.42	33.90	36.44	35.17
SVM	39.44	38.48	38.96	27.54	33.48	30.51	39.52	43.96	41.74	41.60	45.18	43.39
SL	39.30	38.36	38.83	28.00	33.78	30.89	38.92	44.08	41.50	44.06	46.10	45.08



	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean
	BW Selected acoustic features (305)			Linguistic features (5)			Decision-level fusion (BW)			Feature-level fusion (BW)		
NB	40.56	38.08	39.32	25.76	31.92	28.84	37.22	41.04	39.13	41.04	39.52	40.28
SVM	39.90	37.16	38.53	27.54	33.48	30.51	42.28	38.46	40.37	41.68	42.88	42.28
SL	39.84	38.66	39.25	28.00	33.78	30.89	40.96	39.86	40.41	42.68	44.08	43.38

Fig. 2. Unweighted average recall of the classifiers using the selected set of acoustic features (28 features selected by the greedy forward search selection algorithm (top) and 305 features selected by the greedy backward search selection algorithm (bottom)), the set of 5 linguistic features, the decision-level fusion scheme and the feature-level fusion scheme.

the 28 acoustic features selected by the FW search was better, in general, than the performance of the classifiers that only used the 5 linguistic parameters. In the case of the SVM classifier, the use of the acoustic features improved the performance of the linguistic parameters by 8.45% absolute (27.70% relative). In the case of the Simple Logistic performance was improved by 7.94% absolute (25.70% relative). Only the Naïve-Bayes got its performance improved (by only 0.11% absolute, 0.38% relative) using the linguistic features instead of the acoustic parameters. In the case of the set of features selected by the BW search, the performance of the classifiers using the 305 features was better, in all cases, than using the 5 linguistic parameters. The improvement in the case of the Naïve-Bayes, the SVM and the Simple Logistic classifiers was 10.48% absolute (36.34% relative), 8.02% absolute (26.29% relative) and 8.36% absolute (27.06% relative), respectively.

However, the combination of the linguistic and the acoustic features at the decision and at the feature levels improved the performance of the classifiers that considered both modalities independently. For the FW search, the decision-level fusion results improved the mean of the performances achieved by the acoustic and the linguistic sets in the case of the Naïve-Bayes, the SVM and the Simple Logistic classifiers by 5.63% absolute (19.56% relative), 7.00% absolute (20.15% relative) and 6.64% absolute (19.05% relative), respectively. The improvement in the case of the feature-level fusion scheme was 6.38% absolute (22.16% relative), 8.65% absolute (24.90% relative) and 10.22% absolute (29.32% relative), respectively. Considering the BW search, the decision-level fusion results improved the mean of the performances achieved by the acoustic and the linguistic sets in the case of the Naïve-Bayes, the SVM and the Simple Logistic classifiers by 5.85% absolute (16.95% relative) and 5.34% absolute (15.23% relative), respectively. In the case of the Naïve-Bayes classifier, performance was slightly degraded. The improvement in the case of the feature-level fusion scheme was 6.20% absolute (18.19% relative) for the Naïve-Bayes, 7.76% absolute (22.48% relative) for the SVM and 8.31% absolute (23.70% relative) for the Simple Logistic classifier. As it can be observed, the improvement achieved by the fusion of the acoustic and the linguistic parameters (regardless the classifier considered) is more significant in the case of the acoustic FW search selected features than in the case of the acoustic BW search selected features.

In all the cases, the fusion of both modalities at the feature level outperformed the results of the fusion at the decision level. Considering the FW search selected features, for the Naïve-Bayes, the SVM and the Simple Logistic classifiers, the feature-level fusion scheme improved the performance of the decision-level scheme by 0.75% absolute (2.18% relative), 1.65% absolute (3.95% relative) and 3.58% absolute (8.63% relative), respectively. In the case of the BW search selected features, the feature-level fusion scheme considering the Naïve-Bayes, the SVM and the Simple Logistic improved the performance of the decision-level scheme by 1.15% absolute (2.94% relative), 1.91% absolute (4.73% relative) and 2.97% absolute (7.35% relative), respectively.

Although the Naïve-Bayes classifier performed well in a prior study [16], in the case of the FW search selected features its performance was below the other two classifiers. The main reason can be found in the fact that the feature selection algorithm used in Section III.B was not designed to avoid dependencies among the chosen parameters, being independence of features one of the requirements of this classification algorithm [28]. This degradation was not observed analysing the features selected by the BW search because it contains a larger number of parameters.

Only the Fold 1 columns of Fig. 2 must be taken into account to compare these results with the experiments carried out by other authors in the same scenario. This column shows the performance of the classification algorithms when using fold 1 for training and fold 2 for testing, i.e. the two different schools independently, as detailed in [7]. Reference [8] compiled a list of results achieved by several authors working in the same conditions and their fusion by a majority voting scheme. The fusion of the best 7 results achieved a performance of 44.00% UAR, considering different learning schemes and datasets. The best result obtained in this paper by means of the Simple Logistic classifier and the feature-level fusion scheme considering the acoustic FW search selected features (i.e. using 33 features) improved this result by 0.06% absolute (0.14% relative). Although both results were quite similar, it is noteworthy that the number of features involved in our study was dramatically lower and also the complexity of the learning scheme.

V. Conclusion

In this paper we presented a comparison between decision-level and feature-level fusion to merge the acoustic and the linguistic modalities in a real-life non-prototypical emotion recognition from speech scenario. Also, we compared two procedures to select the most relevant features from the large set of acoustic parameters.

We parameterized the audio recordings of a naturalistic speech corpus obtaining 384 acoustic and 5 linguistic features. To reduce the amount of acoustic features we compared two greedy search procedures for feature selection analysing the full set of features forwards and backwards, obtaining 28 and 305 relevant parameters, respectively. The performance of the classifiers with these reduced datasets was, except for the case of the Naïve-Bayes algorithm with the FW search selected features, slightly better than using the full dataset. Using fewer features we were able to speed up the emotion recognition process because we simplified the parameterization stage and the small datasets reduced the computational cost of the classification stage.

Linguistic information, by themselves, did not create a good dataset for the classifiers of this experiment and their performance was even below the performance achieved by

using only the acoustic dataset. However, the combination of these modalities by means of any of the two fusion procedures outperformed the results achieved by both modalities on their own. It is remarkable, then, the importance of analysing the acoustic modality (how things are said) and the linguistic modality (what things are said) to achieve the best results in an automatic emotion recognition experiment, in a similar way as we do in the human communication. This outperformance is more significant in the case of the fusion of the linguistic parameters and the acoustic FW search selected features than in the case of the fusion of the linguistic parameters and the acoustic BW search selected features. Moreover, in general, results from the FW scheme are better than in the BW scheme, except for the case of the Naïve-Bayes algorithm.

Feature-level fusion revealed as the best scheme to merge the acoustic and the linguistic information. Moreover, this kind of fusion is simpler than decision-level fusion, which reduces the complexity of the analysis of the speech recordings. In this feature-level fusion scheme we used only one classifier to analyse a reduced set of acoustic and linguistic parameters merged by simple concatenation of vectors. The performance of this scheme was better than the decision-level scheme consisting of three classifiers: two for each modality and one to merge their results.

The best classifier in this experiment was the Simple Logistic algorithm. Although the Naïve-Bayes is a simple classifier able to achieve good results, its performance was degraded when working with the smallest set of acoustic features (those selected by the FW search procedure). One of the requirements of this classifier is the use of independent parameters but our feature selection procedure was not intended to achieve it. For this reason, in future work, we will experiment with other methods to select relevant feature subsets but also eliminating the redundancy of the data, like [29].

Future work will be related to the enhancement of the linguistic parameterization by considering not only individual words but also groups of them in the form of n-grams. With these n-grams we will be able to study the relation of more complex linguistic structures and the relations between words. Also, we will include an automatic speech recogniser module to work in a more real scenario.

REFERENCES

[1] J. Canny, "The future of human-computer interaction," Queue, vol. 4, no. 6, pp. 24–32, July 2006.

[2] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1175–1191, October 2001.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39–58, January 2009.

[4] M. Slaney and G. McRoberts, "Baby Ears: a recognition system for affective vocalizations," in Proceedings of 1998 International Conference on Acoustics, Speech and Signal Processing, vol. 2, Seattle, WA, pp. 985–988, 1998.

[5] M. Chetouani, A. Mahdhaoui, and F. Ringeval, "Time-scale feature extractions for emotional speech characterization," Cognitive Computation, vol. 1, no. 2, pp. 194–201, 2009.

[6] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in 10th Annual Conference of the International Speech Communication Association, pp. 1595–1598, 2009.

[7] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009). Brighton, UK, pp. 312–315, September 2009.

[8] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," Speech Communication, Special Issue: Sensing Emotion and Affect – Facing Realism in Speech Processing, vol. 53, no. 9-10, pp. 1062–1087, November 2011.

[9] A. Batliner, K. Fischer, R. Hubera, J. Spilkera J, and E. Noth, "How to find trouble in communication," Speech Communication, vol. 40, pp. 117–143, 2003.

[10] S. Steidl, Automatic classification of emotion-related user states in spontaneous children's speech. Logos Verlag, 2009.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - Introducing the Munich open-source emotion and affect recognition toolkit," in Proceedings of th 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction, pp. 576–581, 2009.

[12] C. M. Lee, S. S. Narayanan, "Towards detecting emotions in spoken dialogs," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 2, pp. 293–303, March 2005.

[13] I. H. Witten and E. Frank, Data mining: Practical machine learning tools and techniques. 2nd Edition. San Francisco, CA: Morgan Kaufmann, June 2005.

[14] Y. Kim, N. Street, and F. Menczer, "Feature selection in data mining," in J. Wang (ed.) Data mining: Opportunities and challenges, pp. 80–105. Idea Group Publishing, 2003.

[15] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol.3, pp. 1157–1182, 2003.

[16] S. Planet, I. Iriondo, J. C. Socoró, C. Monzo, and J. Adell, "GTM-URL contribution to the Interspeech 2009 Emotion Challenge," in Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009). Brighton, UK, pp. 316–319, September 2009.

[17] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in Proceedings of the 13th International Joint Conference on Artificial Intelligence. pp. 1022–1029, 1993.

[18] J. Platt, "Machines using sequential minimal optimization," in B. Schoelkopf, C. Burges, and A. Smola (eds.) Advances in Kernel Methods: Support Vector Learning, MIT Press, 1998.

[19] T. Hastie and R. Tibshirani, "Classification by pairwise coupling", Annals of Statistics, vol. 26, no. 2, pp. 451–471, 1998.

[20] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees". Machine Learning, vol. 59, no. 1–2, pp. 161–205, 2005.

[21] J. C. Bezdek, J. M. Keller, R. Krishnapuram, and N. R. Pal, Fuzzy models and algorithms for pattern recognition and image processing. Norwell, MA: Kluwer Academic Publishers, 1999.

[22] D. Ruta and B. Gabrys, "An overview of classifier fusion methods," in Computing and Information Systems, vol. 7, no. 1, pp. 1–10, 2000.

[23] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, pp. 241–259, 1992.

[24] D. Morrison, R. Wang, and L. C. D. Silva, "Ensemble methods for spoken emotion recognition in call-centres," Speech Communication vol. 49, no. 2, pp. 98–112, 2007.

[25] I. Iriondo, S. Planet, J.-C. Socoró, E. Martínez, F. Alías, and C. Monzo, "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification," Speech Communication, vol. 51, no. 9, pp. 744–758, 2009.

[26] J. R. Quinlan, C4.5: Programs for machine learning. 1st Edition. Morgan Kaufmann, January 1993.

[27] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in 13th Annual ACM International Conference on Multimedia, pp. 399–402, 2005.

[28] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, pp. 338–345, 1995.

[29] M. A. Hall, Correlation-based feature subset selection for machine learning. Hamilton, New Zealand, 1998.

Santiago Planet received his M.Sc. degree from the Ramon Llull University (URL), Spain, in 2007, while he was carrying out researching and teaching tasks in La Salle School of Engineering and Architecture (URL). Currently, he is a PhD student at the same university. He has published several articles and has made contributions to national and international scientific conferences. Also, he has contributed to several research and development projects. His main research topic is affective recognition by means of multimodal processing.

Ignasi Iriondo received the PhD degree in Electrical Engineering from the Ramon Llull University (URL), Spain, in 2008. He worked in a company dedicated to Speech Technologies oriented to people with disabilities. Nowadays he is lecturer on Digital and Speech Processing and coordinator of the degree of Audiovisual Systems Engineering in La Salle School of Engineering (URL). His major research interests are audiovisual speech synthesis and acoustic modelling of emotional expression.