INTERSPEECH 2009 BRIGHTON

GTM-URL Contribution to the INTERSPEECH 2009 Emotion Challenge

Santiago Planet, Ignasi Iriondo, Joan-Claudi Socoró, Carlos Monzo, Jordi Adell

GTM – Grup de Recerca en Tecnologies Mèdia
La Salle – Universitat Ramon Llull
Quatre Camins 2, 08022 Barcelona (Spain)
{splanet,iriondo,jclaudi,cmonzo,adell}@salle.url.edu

Abstract

This paper describes our participation in the INTERSPEECH 2009 Emotion Challenge [1]. Starting from our previous experience in the use of automatic classification for the validation of an expressive corpus, we have tackled the difficult task of emotion recognition from speech with real-life data. Our main contribution to this work is related to the Classifier Sub-Challenge, for which we tested several classification strategies. On the whole, the results were slightly worse than or similar to the baseline, but we found some configurations that could be considered in future implementations.

Index Terms: emotion, recognition, speech parameters, challenge, classification

10.21437/Interspeech.2009-104

1. Introduction

There is a growing trend towards the use of speech in human-machine interaction. In this field, the inclusion of automatic emotion recognition or expressive speech synthesis can improve communication by making it sound more natural. It could be used in the automatic generation of audiovisual content (e.g., for entertainment purposes), for virtual meetings, or even in automatic dialogues to adapt the system to the user's emotional state. Our research group has interest in both technologies. In a previous study [2], we developed an expressive speech database from acted speech. Based on [3], we applied techniques of emotion recognition to validate its expressive content and to refine the recorded speech database. The refinement was conducted by an automatic system. It was designed to emulate humans' subjective criteria for the identification of emotions in speech. The system was trained with the results of a subjective evaluation, which was carried out on a small part of the corpus using [4].

The speech utterances of this corpus were completely different from the data provided for this Emotion Challenge. The speech material supplied in the Emotion Challenge differed from our previous experience because it was highly spontaneous and was not recorded in studio. On the other hand, the corpus developed in [2] was recorded in a studio by a professional speaker and the styles were balanced.

Our work deals with two of the proposed sub-challenges, the Classifier Sub-Challenge and the Feature Sub-Challenge, in which we considered five non-prototypical emotional classes.

In Section 2 we propose four approaches for the Classifier Sub-Challenge in which we consider various learning schemas, describing them and comparing the results. In Section 3 we face the Feature Sub-Challenge by adding new features to the provided dataset and using a feature-selection algorithm based on mutual information to select a smaller subset. Sections 4 and 5 contain discussion and conclusions, respectively.

2. Classifier Sub-Challenge

The Classifier Sub-Challenge is described in detail in [1]. In this study we faced the task of classifying the provided parameterised corpus considering five emotional labels. We found that the main problem is that the corpus is unbalanced, and this makes it difficult to tackle the classification task with a standard classification algorithm. For this reason we considered four different approaches, as described below.

2.1. Description of classifiers

Because the datasets were very unbalanced, two measures had to be taken into account. The former was the weighted average (WA) recall, which defines the percentage of correctly classified examples. However, when the number of examples corresponding to one class is much larger than the others, this measure can provide a mistaken idea of the classifier accuracy; e.g., a majority class predictor will get a high WA recall, but all other examples of other classes will be misclassified. For this reason unweighted average (UA) recall, defined as the arithmetic mean of the recalls of each class, must be considered.

In the first stage we studied various learning algorithms that have been tested successfully in previous studies [5], leaving the corpus unmodified, and evaluated the algorithms with a 10-fold cross-validation.

In the second stage we considered a two-level classifier scheme. In this approach, at the first level of the classifier structure examples are classified into binary splits according to whether they belong to the class N or any other different (NoN) class. At the second level, a classifier is trained with the whole corpus resampled to get a uniform distribution of classes. The idea is to classify N-labelled examples at the first level by means of a specialised two-class classifier and, later, to classify the rest of the examples at the second level.

In the third stage a two-level classifier was also considered. In this approach, at the first level examples are classified into binary splits according to whether they belong to the class N or any other (NoN) class. At the second level, however, two classifiers are trained using the training dataset after being classified by the first-level classifier. The former is trained with those N-examples that, at the first level classifier, have been correctly predicted as N (True Positives), and with those NoN-examples that have been incorrectly classified as N (False Positives). The latter is trained with those N-examples incorrectly classified as NoN (False Negatives) and those NoN-examples that have been correctly classified as NoN (True Negatives). With this strategy we could get better balanced datasets to train the classifiers of the second level.

Finally, we studied another approach based on five classifiers following a waterfall structure. The first classifier is trained with the whole training dataset to distinguish the majority class from the others, i.e. N-examples from NoN-

examples. In the second level, a classifier is trained to carry out the same task but, in this case, distinguishing the second most populated class from the others, i.e. E-examples from NoE-examples. Only training examples classified as NoN by the first-level classifier are used to train this second classifier. This structure is repeated to the fifth classifier, distinguishing P-examples from NoP-examples. In this case, NoP-examples are classified as N-examples by default.

2.2. Results

For the first experiment, we used the original dataset provided for the Classifier Sub-Challenge, in which we considered five classes of emotion and studied six learning schemes using Weka [6]: J4.8 (a decision tree based on C4.5), IBk-A (Instance Based learner (1 solution) using Euclidean distance), IBk-B (like IBk-A but using Manhattan distance), Naïve-Bayes (using previous discretisation), a decision table, and linear SMO (Weka implementation of SVM). Table 1 shows the results of the learning schemes when using either a 10-fold cross-validation (10-FCV) or the provided test set. Thereby, differences between the 10-FCV and the test set evaluation could be observed. In general, results evaluating the test set were worse in terms of UA recall than those obtained using the 10-FCV. However, the Naïve-Bayes algorithm is not particularly susceptible to evaluation with a 10-FCV or by means of the supplied test set, although it gets low WA recall. Moreover, Naïve-Bayes outperforms the best baseline indicated in [1], corresponding to a linear SMO with a dataset standardised and resampled by SMOTE. Table 2 shows the confusion matrix for this schema of evaluating the test set. The best WA recall was achieved by the linear SMO in both the 10-FCV and the test set, and its UA recall was the second best.

Table 1. *Results of WA recall and UA recall (%) with a 10-fold cross-validation (10-FCV) for evaluating the supplied test set.*

	10-FCV		Test set	
Learning scheme	WA	UA	WA	UA
J4.8	48.26	33.31	48.23	26.82
IBk-A	50.83	33.55	48.92	26.14
IBk-B	54.82	39.37	50.05	27.35
Naïve-Bayes	41.67	41.72	39.14	41.16
Decision table	59.89	28.33	64.71	24.53
Linear SMO	63.71	37.54	65.62	28.94

Table 2. *Confusion matrix for the Naïve-Bayes algorithm for evaluating the test set. The first row represents the predicted class, and the first column the actual class.*

	A	E	N	P	R
A	310	129	84	60	28
E	359	712	317	74	46
N	1230	1030	2027	861	229
P	14	7	52	130	12
R	122	67	178	126	53

For the second experiment, we chose two linear SMO classifiers for the first-level binary classifier and the second-level 5-class classifier. In this case, two datasets were selected: the original dataset and its standardised version. Table 3 shows the results evaluated in the supplied test (second and third rows), and compared to the baseline provided in [1] (first row).

Table 3. *Results of WA recall and UA recall (%) for two different approaches: one linear SMO and a two-level classifier.*

Dataset	Algorithm	WA	UA
Original	Linear SMO	65.62	28.94
Original	Two level SMO	60.78	30.82
Original stand.	Two level SMO	58.71	32.30

Because the standardised dataset improves the UA recall of the experiment, we chose it for the rest of the experiments in this section.

The third experiment added a classifier to the one described above. In this approach, examples classified as N by the first-level classifier (1stC) were reclassified into the five possible classes in a second-level classifier (2ndC-1). Also, examples labelled as NoN in the first level were processed by a classifier in the second level (2ndC-2).

Table 4 shows the results for the experiments conducted using the third approach. In this case, three configurations were tested, introducing the RBF SMO (SMO with a Gaussian radial basis function kernel). Thus, an improvement of the WA recall and UA recall was seen in the third configuration (two linear SMOs and a Naïve-Bayes classifier for examples classified as NoN), compared to the two-level SMO described above. Using an RBF SMO improves WA recall but worsens the UA recall. The linear SMO structure presents an intermediate result.

Table 4. *Results of WA recall and UA recall (%) for the third experiment.*

1stC	2ndC-1	2ndC-2	WA	UA
Linear SMO	Linear SMO	Linear SMO	63.04	30.20
RBF SMO	RBF SMO	RBF SMO	64.28	27.97
Linear SMO	Linear SMO	Naïve-Bayes	60.20	32.63

In the fourth experiment we tried to generate more-balanced datasets at each level of the waterfall structure by discarding those examples that were being classified at each stage. In this case, working with linear SMOs at each level, WA recall was 65.53%, while UA was 28.03%. Making a small modification to this approach, we changed the order of the final two classifiers. We did this because the R-examples are a set of cases that are not well defined (Rest), and we believed they could be difficult to characterise. With this modification WA recall was 65.33% (almost the same value compared with the previous version) and UA recall was 28.12%. Compared with the other experiments, this approach was better in terms of WA recall, but worse in terms of UA recall. However, training was faster because the datasets used in each step were smaller than in the other cases. Table 5 shows the confusion matrix of this approach when we evaluated the test set. We found that good classification of N-examples could be made using this approach.

Table 5. *Confusion matrix for the waterfall strategy evaluation of the test set.*

	A	E	N	P	R
A	92	133	382	4	0
E	17	399	1082	10	0
N	80	346	4904	47	0
P	5	0	193	16	1
R	18	20	487	21	0

3. Feature Sub-Challenge

To select the most characterising features for this corpus we used a feature-selection algorithm. The goal was to evaluate a large set of attributes consisting of the standard acoustic features provided for the Classifier Sub-Challenge and several others extracted from the audio files. A detailed explanation of the corpus and the provided features can be found in [1].

3.1. Source corpus

To the acoustic features specified in [1] several other parameters, referring to Voice Quality (VoQ), have been added. These parameters have proved to be useful in other emotion-recognition experiments [2][7]. Specifically, these parameters are:

- Jitter and Shimmer. These parameters compute the cycle-to-cycle variations of the fundamental period and waveform amplitude, respectively; i.e., they describe frequency and amplitude modulation noise. These parameters were adapted to be used in expressive speech applications [8].
- Hammarberg Index (hammarberg). This index is the difference between the maximum energy in the 0-2000 Hz and 2000-5000 Hz frequency bands.
- Spectral Flatness Measure (SFM). This parameter is computed as the ratio of the geometric to the arithmetic mean of the spectral energy distribution.
- Drop-off of spectral energy above 1000Hz (do1000). This is a linear approximation of spectral tilt above 1000 Hz. This measure has been standardised.
- Relative amount of energy in the high- (above 1000Hz) versus low-frequency range of the voice spectrum (pe1000).

For all previous parameterisations, only voiced parts of speech were taken into account. This information, extracted from pitch marks (only voice areas were marked), was used as a reference for the voiced speech analysis. Parameters were extracted from a 40 ms window size and a 20 ms window rate, for each audio file. If a parameter could not be calculated, a value of -200 was returned as an error code.

For each file and parameter, eleven statistics were calculated: mean, standard deviation, median, maximum value, minimum value, range, first and third quartile, inter-quartile range, *skewness* and *kurtosis*. The total number of features was 66.

In addition, two rhythm-related parameters were considered:

- Accent group duration. This parameter was estimated by the distance between local maxima in a pitch contour. This contour was obtained by subtracting a base contour from the original pitch contour. The base contour consisted of the original contour low-pass filtered at 5Hz.
- Syllable duration. The duration was estimated by the distance between intensity peaks of the intensity contour.

Both the pitch and the intensity contours were obtained using the Praat tool [9].

For each of both parameters and file, two statistics were measured: mean and standard deviation. Considering the VoQ parameters, 70 parameters were added to the original dataset. The resulting dataset had 454 features per example, excluding the label of the emotion. The training dataset consisted of 9959 examples.

3.2. Feature selection

A corpus such as that described above is usually too large to be processed efficiently by a classification algorithm. In many situations, a large number of attributes are irrelevant and could be discarded without degrading the classification rate, and in fact could improve it. Many techniques can be considered to reduce the dimensionality of a dataset by selecting the most characterising attributes [6]. These techniques can follow two different approaches [10]:

- Filter methods: Attributes are selected before the learning process starts. Selection is independent of the posterior learning scheme.
- Wrapper methods: Candidate subsets of attributes are evaluated by means of the learning scheme. In this case computational complexity is higher than in the filter approach, and the final subset is fitted to a specific algorithm.

We chose a subset of attributes by means of a filter approach. The attributes were selected following the minimal-redundancy-maximal-relevance criterion (mRMR) [11]. The goal of this method is to select those mutually exclusive attributes whose relevance is closest to their class.

We used a MATLAB implementation of the mRMR feature-selection algorithm. The training dataset had been previously processed. First, missing values of each attribute were replaced with their means. In addition, due to the large size of the dataset, it was sub-sampled at 50% while maintaining the class distribution. Finally, attributes were discretised: each attribute was standardised, multiplied by a factor k and rounded to the nearest integer. In this case, the value k=10 was empirically chosen.

$$attribute_i = round(k*zscore(attribute_i)) \quad (1)$$

The algorithm was configured to obtain 200 features, but we chose only 100 according to the specifications of the Feature Sub-Challenge [1]. Figure 1 shows a simulation of the selected features with a linear SMO learning schema. It shows the WA recall and UA recall when evaluating a linear SMO by means of a 10-FCV strategy. For 100 features the selected dataset still had not reached the best performance. A linear SMO trained with it and tested with the provided test set resulted in a WA recall of 64.95% while UA recall was 21.32% (variation of -0.67% and -7.62%, respectively, compared to the original dataset, which was 284% larger).

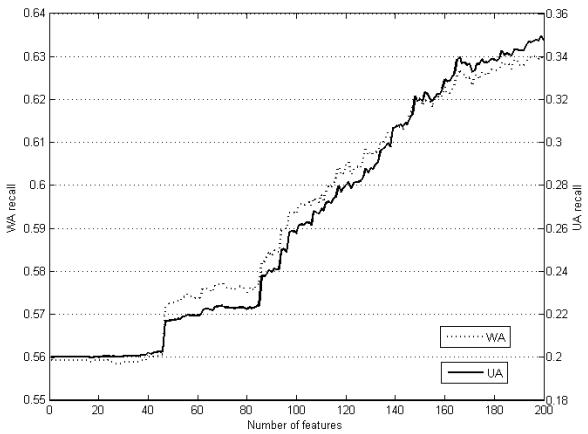

Figure 1: *WA recall and UA recall of a linear SMO evaluated with a 10-FCV, choosing from 1 to 200 features previously selected by the mRMR algorithm.*

4. Discussion

Addressing the task of classifying a corpus such as that proposed in this challenge has revealed important differences from other emotion-recognition studies where more selected corpora are chosen. In this sense, it can be observed that it is difficult to get good results (high WA recall and UA recall) with a simple learning schema, because the dataset is heavily unbalanced. Generally, improving one measure means the degradation of the other. Furthermore, important differences have been observed when evaluating the models with a 10-FCV strategy or when using a different test set for this task. Nevertheless, the Naïve-Bayes algorithm shows a similar behaviour between the evaluation strategies and gets better results in terms of UA recall than even more complex algorithms like SMO, despite its relative simplicity. As seen in Table 2, the confusion matrix of the algorithm reveals a high confusion rate of class N compared with the other classes. This explains why WA recall is so low. However, this could be offset by classifying N-examples first and the remainder of the examples later. This strategy was used in the third classification approach described in this paper, but UA recall became degraded. We propose to adapt the fourth approach to take advantage of this algorithm. In this study a waterfall schema was presented using linear SMO classifiers at each level of the structure. Despite its low UA recall, the WA recall for the waterfall schema was higher than for other schemas. Also, as can be seen in Table 5, N-examples were quite well classified. For this reason, it could be interesting to use a linear SMO for the first level of the waterfall structure and Naïve-Bayes classifiers for the remainder.

It has also been difficult to parameterise the audio files because of the uncontrolled conditions of the recordings. This can be a drawback when selecting the best subset of features. Although pre-process techniques can be applied to the datasets, it could be useful pre-process the audio files before their parameterisation.

Finally, in this study we used a 10-FCV strategy in a first stage to try to determine preliminarily the best classification algorithm. Because important differences were observed when using the test set, which is constructed with speech utterances from other subjects, a different approach could be considered including utterances from different users to construct training and test sets, following a speaker-independent cross-validation strategy as exposed in [12].

5. Conclusions

In this paper we have presented our contribution to the INTERSPEECH 2009 Emotion Challenge with the aim of applying our previous knowledge of automatic emotion recognition using realistic data. Two of the sub-challenges were addressed: the Classifier Sub-Challenge and the Feature Sub-Challenge, including five non-prototypical emotion classes.

For the Classifier Sub-Challenge four studies were conducted. First, we tested various algorithms to observe the behaviour of the trained models when evaluating different test sets. To improve the classification results, expressed in terms of WA recall and UA recall, we proposed three approaches that create structures of classifiers to deal with the unbalanced data of the datasets. Results showed that linear SMO is a good classifier to characterise N-labelled examples, and Naïve-Bayes is a simpler algorithm to classify the remainder of the classes.

For the Feature Sub-Challenge the audio files were parameterised to extract VoQ and rhythm features. These features were added to the original dataset to build a more detailed one. To select the more characterising features, an algorithm based on mutual information was chosen. After the selection, an incremental simulation was carried out to study the performance of a linear SMO with a variable number of features. However, other techniques such as forward or backward selection could be considered as well to improve the final dataset.

6. Acknowledgements

This work was partially supported by the European Commission, project SALERO (FP6 IST-4-027122-IP), and the Spanish Government, project SAVE (TEC2006-08043/TCM).

7. References

[1] B. Schuller, S. Steidl and A. Batliner, "The Interspeech 2009 Emotion Challenge", *Interspeech 2009*, ISCA, Brighton, UK, 2009.

[2] 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* (in press), Elsevier, 2008.

[3] P.-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human Computer Interaction*, vol. 59, no. 1-2, pp. 157–183, special issue on Affective Computing, 2003.

[4] S. Planet, I. Iriondo, E. Martínez and J. A. Montero, "TRUE: an online testing platform for multimedia evaluation", in *Proceedings of the Second International Workshop on EMOTION: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation*, Marrakech, Morocco, 2008.

[5] I. Iriondo, S. Planet, J.-C. Socoró and F. Alías, "Objective and subjective evaluation of an expressive speech corpus," in M. Chetouani et al. [Eds.], *Advances in Nonlinear Speech Processing*, LNCS, 4885, pp. 86-94, Springer-Verlag, 2007.

[6] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., San Francisco: Morgan Kaufmann, 2005.

[7] I. Iriondo, S. Planet, F. Alías, J.-C. Socoró, C. Monzo and E. Martínez, "Expressive speech corpus validation by mapping subjective perception to automatic classification based on prosody and voice quality", in *Proceedings of the XVI International Congress of Phonetic Sciences*, Saarbrücken, Germany, 2007.

[8] C. Monzo, I. Iriondo and E. Martínez, "Procedimiento para la medida y la modificación del jitter y del shimmer aplicado a la síntesis del habla expresiva", in *V Jornadas en Tecnología del Habla*, 2008.

[9] P. Boersma and D. Weenink. "Praat: doing phonetics by computer (version 4.3.04)", 2005. [Online]. Available: http://www.praat.org/, 2005.

[10] P. Langley, "Selection of relevant features in machine learning", in *Proceedings of AAAI Fall Symposium on Relevance*, 1994.

[11] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[12] K. Truong and D. Van Leeuwen, "An 'open-set' detection evaluation methodology for automatic emotion recognition in speech", In *ParaLing'07: Workshop on Paralinguistic Speech - between models and data*, pp. 5-10, Saarbrücken, Germany, 2007.