Building a system for emotions detection from speech to control an affective avatar

Mátyás Brendel, Riccardo Zaccarelli, Laurence Devillers

LIMSI-CNRS, France

{mbrendel,riccardo,devil}@limsi.fr

Abstract

In this paper we describe a corpus set together from two sub-corpora. The CINEMO corpus contains acted emotional expression obtained by playing dubbing exercises. This new protocol is a way to collect mood-induced data in large amount which show several complex and shaded emotions. JEMO is a corpus collected with an emotion-detection game and contains more prototypical emotions than CINEMO. We show how the two sub-corpora balance and enrich each other and result in a better performance. We built male and female emotion models and use Sequential Fast Forward Feature Selection to improve detection performances. After feature-selection we obtain good results even with our strict speaker independent testing method. The global corpus contains 88 speakers (38 females, 50 males). This study has been done within the scope of the ANR (National Research Agency) Affective Avatar project which deals with building a system of emotions detection for monitoring an Artificial Agent by voice.

1. Introduction

The modelling of realistic emotional behavior is needed for various applications, like embodied agents, robots and dialog systems in call centers. Recognition of emotions in speech is a complex task due to the fact that there is no unambiguous answer to what emotion is for a given speech. The term ”emotion” has been so far used for the affective state including the emotions, moods, interpersonal stances, etc. Results reported on emotional material collected in real-world context are sparse in the literature (Devillers and Vidrascu, 2009) in spite of the fact that this topic of research is becoming a key technology for next generation human-machine interaction.

This study comes within the scope of the ANR (National Research Agency) Affective Avatar project which deals with building a system of emotions detection for monitoring an Artificial Agent by voice. The chosen application is Skype where the speaker is depicted by his/her avatar. In this application, the speaker gender will be given by the user in the interface. The avatar should show the expressive behavior (e.g. anger) corresponding to the emotion detected (e.g. irritation). This application has two main challenges: speaker-independent emotion detection and real-time emotion detection. This paper focuses on the first one. The main point is to find an appropriate corpus with sufficient number of speakers for training the emotion detection system and a large variability of emotional expressions.

The choice of appropriate corpora for training computational models is fundamental. The training data must be as close as possible to the behaviors observed in the real application but also large enough, with sufficient variability of emotional expressions, including complex, mixed and shaded emotions. Likewise, expressions of emotion should be collected as they occur in everyday action and interaction rather than as idealized archetypes. Spontaneous emotions are hard to collect, to annotate, and to distribute due to privacy problems. The available corpora in the community are mainly acted, without any application in sight. Moreover, they are small, including few speakers and little variations in the expression of emotions. The emotional corpora already existing at LIMSI have been mainly collected in call centers (bank, emergency or stock exchange call centers). These corpora overcome many of the previous limitations: they contain spontaneous manifestation of emotions, complex emotions and a large diversity of speakers (more than 700 in a call center corpus named CEMO (Devillers et al., 2005), (Devillers and Vidrascu, 2007) but they are telephonic data with mainly negative emotions.

There has not been any accessible corpus of everyday talk present for training the emotional model for our Skype-application, neither any software whereby we would be able to collect data fitting into our framework. Thus, we have selected emotional classes and have built protocols to collect data in everyday talks. In order to obtain a wide range of emotional expressions from speakers with various acoustic features and a large number of speakers, we used two kinds of corpus, the first named CINEMO (Rollet et al., 2009) is speech acted in context by 50 speakers and the second, JEMO is obtained by an emotion detection game with 39 speakers.

Section 2 will describe both our corpora and annotations. We will focus on 4 emotional classes (which are the most represented in our corpora): positive (including satisfaction, amusement, joy and all positive behaviors), sadness (including different levels of sadness such as disappointment), anger (irritation) and neutral (non emotional manifestations). In Section 3 the LIMSI Affective Avatar features used are described. In section 4 the protocol of evaluation is given. Then we will provide results for the 4 emotional macro-classes detection (POS, SAD, ANG and NEU) using the first, then the second corpus in section 5. We provide results for the united corpus in section 6, and for balanced corpora in section 7. The use of separate models for male and female (Section 8) will be studied and also feature selection (Section 9). Our conclusion will be on the possibility to mix different kind of corpus for training more

efficient classifiers.

2. Corpora

2.1. The CINEMO Corpus

The CINEMO corpus used in this paper consists of 1012 instances after segmentation of emotional French speech amounting to a total net playtime of 2:13:59 hours. 50 speakers (15 to 60 years) dubbed 27 scenes of 12 movies. For some scenes, the two roles have been played by different persons, making a total of 31 different linguistic scripts. Each linguistic script contains one to twelve speaker turns. Each scene was repeated around 1.67 times in average. This corpus is described in details in ((Rollet et al., 2009) (Schuller et al., 2010)). A subset of the more consensual segments was chosen for training models for detection of 4 classes (POS, SAD, ANG and NEU). The rich annotation of CINEMO was used to build these 4 macro-classes; for example the class NEU contains segments annotated as neutral plus low-level intensity and activation for positive, sadness and stress emotions. We have not considered mixtures of emotions for training our models in that experiment. Table 1 is a description of the CINEMO sub-corpus:

CINEMO	POS	SAD	ANG	NEU
# segments	313	364	344	510

Table 1: CINEMO sub-corpus, number of segments for 50 speakers.

As it can be seen in table 1, positive emotion is underrepresented in CINEMO.

2.2. The JEMO Corpus

The JEMO corpus features 1062 instances after segmentation of speech recorded from 39 speakers (of 18 to 60 years old). JEMO is a corpus collected with an emotion-detection game. This game used a segmentation tool based on silenced pauses and used a first system of 5-emotions detection (ANGer, FEAr, SADness, POSitive and NEUtral) and a system of activation detection (low/high) built on CINEMO data. The linguistic content is free. The system detects the emotion (among the 5 classes) and the activity (low or high) from the audio signal and sends an emoticon of the detected emotion to the screen (see Figure 1).

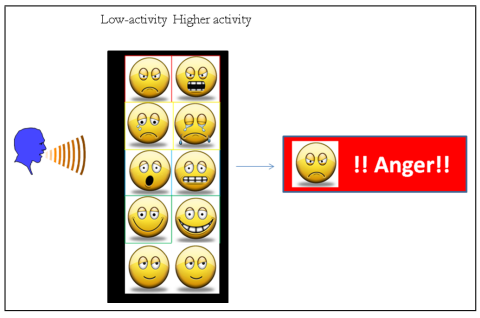

Figure 1: The emotion detection game used for recording JEMO.

This game is a first prototype of a real time detection system with an error recognition rate still significantly high. Fear emotion is for example very badly recognized. Therefore we omitted this class. Anger and Sadness are the best recognized emotions. However, this aspect led to a more challenging game for the players, whose reactions were more spontaneous and differentiated (e.g., several negative reactions due to an emotion often not recognized and a positive reaction when the emotion was finally recognized). Thus, speakers generated spontaneous sentences with higher level of expressivity than in CINEMO.
The JEMO corpus has been annotated by two coders with major and minor emotions. These data were more prototypical than in the corpus CINEMO because very few mixtures of emotions were annotated.

JEMO	POS	SAD	ANG	NEU
# segments	316	223	179	416

Table 2: JEMO sub-corpus, number of segments for 39 speakers.

In table 2 it can be seen we have here much more examples in the POS class especially with women. Furthermore the POS class of the JEMO corpus contains more prototypical expressions of Joy than in CINEMO which contains generally speaking more complex and shaded emotions.

3. Features

Each speech segment is passed through spectral (16 MFCCs) and prosodic analysis (pitch, zero-crossing and energy). The feature extractor next calculates basic statistical features on voiced parts: min, max, mean, standard deviation, range, median quartile, third quartile, min and max intra and intra range, and the mean and standard deviation of the coefficients of least square fitting regression (of each voiced segment); min and max inter range (between voiced segments). Overall, 458 features are thus obtained: 23 for pitch, 51 for energy (from these 22 root mean square energy), 18 zero-crossings and 366 for MFCC1-16
Table 3 shows the low level descriptors and functionals used in generating the LIMSI Affective Avatar features for these experiments.

4. Evaluation Protocol

We call methods speaker independent (SpI) if they ensure that the same speaker is not present in the training and test set. Speaker dependent (SpD) testing denotes the opposite: when the same speaker may occur in the test set and the train set. First we compared simple SpI and SpD train-test evaluation, but as a function of the ratio of the train and test sets. We can call this method Leave-n-Speaker-Out testing: n speakers were randomly taken as the test set and the remainder speakers were taken as the train set. Thus we obtained the SpI version of testing. The ratio of the train and test set was computed and a partition with the same ratio was created with Weka (Witten and Frank, 2005) to obtain the SpD version of the train and test set. Such train-test turns were done for n=1,2..., N and 30 turns were done

LLD	Functionals
Energy	*moments(2)*
RMS Energy	absolute mean, max
F0	*extremes(2)*
Zero-Crossing-Rate	2 x values, range
MFCC 1-16	*linear regression(2)*
	MSE, slope
	quartiles(2)
	quartile, tquartile

Table 3: *LIMSI features: low-level descriptors and functionals. Abbreviations: root mean square (RMS), Mel Frequency Cepstral Coefficients (MFCC), Mean Absolute/Square Error (MAE/MSE). Note that we do not have all the combinations.*

for each n. N was chosen to be approximately the half of all speakers, but we deleted cases, where the train set was smaller than 50% of the corpus. Figure 2 shows the Recognition Rate (RR) for the united CINEMO JEMO corpus.

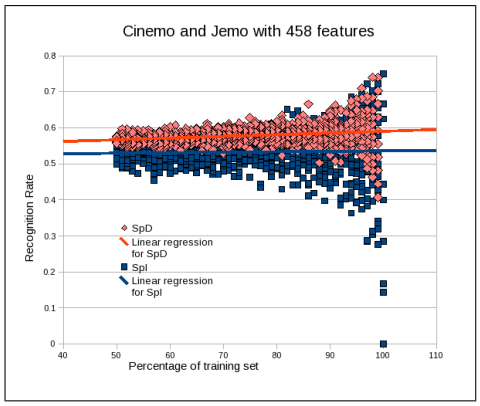

Figure 2: Speaker dependent (SpD) and Speaker independent (SpI) testing. Recognition Rate (RR) as a function of train-test ratio. Note that points of SpD mask a lot of point of SpI.

As it can be seen in figure 2, SpD results are systematically better than SpI results, which means that SpD is over-optimistic. Another conclusion is that performance depends on ratio of train and test sets. Interestingly, this dependence fits well to the linear regression, with an increasing deviation as the test-set size decreases to zero.
This means that a common SpI testing technique, leave-one-speaker-out (LOSO) used for example in (Steidl, 2009) is not comparable to a 50 or 60% split or 10-fold Cross-validation (CV). Moreover, it can be seen that close to the 100% of train set ratio, where the LOSO test cases are, the variance is huge. This means that there migh be some consern when using LOSO. We have to develop instead a SpI 10-fold CV.
Note that in Weka - the most commonly used tool for machine learning - there is only SpD CV present as a built-in tool. However, as we have shown, this SpD CV is too optimistic. Especially considering our application, where the end-user is not included in the training set. Therefore a SpI CV shall be used as the standard evaluation/protocol.

Consequently, we developed an SpI CV. Here the folds are built over the set of speakers: they are divided randomly in 10 folds. From the 10 folds of the speakers the 10 fold of instances is built by gathering the corresponding instances together. This way the separation of speakers in the train and test set is ensured. Note that the partition of speakers instead of instances restricts the possibilities and it may result in an additional unbalancedness, which is sufficiently enough corrected by the averaging of the folds. The recognition rate is accumulated through the folds, resulting in the recognition rate of the entire corpus. Beside of this, unweighted average recall of the classes is also accumulated through the folds and computed as the simple average of the recall of the classes in the end. UAR is important, because it takes minority classes in the test set more into account.
We trained the data set using Weka (Witten and Frank, 2005) with a SVM with a polynomial-kernel by means of Sequential Minimal Optimization algorithm (Platt, 1999).

RR/UAR	SpD CV	SpI CV
CINEMO	0.5715/0.5668	0.5003/ 0.4807
JEMO	0.6340/0.5948	0.6014/0.5514
C. and J.	0.5816/0.5595	0.5414/0.5077

Table 4: Comparing SpI and SpD 10-fold CV. RR=Recognition Rate, UAR=Unweighted Average Recall, C.=CINEMO, J.=JEMO.

Table 4 shows that there is a consistent and significant difference between speaker dependent (SpD) and speaker independent (SpI) cross-validation both in RR and UAR. UAR is important, because it rules out the case that SpI is worse because of a possibly unbalanced test-set.Thsu we can conlude that the higher performance of SpD is an artefact: speaker dependent relations are learned, which increases performance, but which can not be exploited when the application is used with different speakers. Note, that the performance of JEMO is better than of CINEMO, since it contains more prototypical emotions.
Also note, that there is a more significant difference in JEMO between RR and UAR, which is because JEMO is more unbalanced than CINEMO. However, the above mentioned two trends are still obviously valid.
Consequently, in our further experiments we only use SpI CV.

5. Results on both corpora independently

First, we tested the two corpora on each other to see, what errors each corpus causes. To do this, we adapted our SpI CV to cross-corporal testing: the 10 train folds of CINEMO were tested with the test-folds of JEMO and vice versa. This way the train and test sets are the same for inter and intra-corpus testing.
As it can be seen in Table 5 training on CINEMO and testing on JEMO performs better than vice-versa. This is due to that JEMO contains more prototypical emotions. Obviously, it is better to train on a wider set and test on a narrower than the other way. Surprisingly, training on CINEMO then testing on JEMO gives a slightly better performance than testing on CINEMO itself. However, we can

RR/UAR	Test C.	test J.
Train C..	0.5003/ 0.4807	0.5123/0.4805
Train J.	0.4272/0.3899	0.6014/0.5514

Table 5: Cross-corpora results. C.=CINEMO=C., J.=JEMO.

explain this in the same way: JEMO contains more prototypical emotions, which perform better, when testing even when training was carried out on CINEMO. Training and testing on JEMO is way the best, which is not surprising at all.

6. Results on the union of the corpora

The union of CINEMO and JEMO was made by uniting all the corresponding classes of CINEMO and JEMO. The united corpus is more balanced and contains a larger variability of emotional expressions (acted from JEMO and more shaded and complex emotions from CINEMO), so we tested how this will be reflected in the results (0.5414/0.5077). This is worse than JEMO on itself (0.6014/0.5514), but much better than any other value. The united corpus performs better than the average of JEMO on both corpora: 0.513/0.4691, computed from the tables 5 and 4.
This means that the unification of the corpora improved the results. We could not be better than JEMO, but it is obvious that the good result of JEMO on itself is because it is a small corpus with prototypical emotions only, and it has no good generalization power, see table 5: training on JEMO, testing on CINEMO.

7. Results on balanced corpora

Several factors could be the reason why the united corpus is better than the two sub-corpus. The first is that it is more balanced, the second that it richer, and third is the sheer amount of instances: the united corpus is bigger. To test this, we made experiment with corpora which were balanced and of the same size. Since the class anger in JEMO is the smallest (179), we did take the same number of instances for all the classes in all the corpora. This way we created a balanced JEMO, CINEMO and a mixed corpus with 716 instances. Note that these balanced corpora are much smaller, therefore we expect lower performance.

RR	Test C.	Test J.	Test C. and J.
Train C.	0.4623	0.4092	0.4358
Train J.	0.3713	0.5740	0.4727
Train C. and J.	-	-	0.4553

Table 6: Balanced results. C.=CINEMO, J.=JEMO. Note that since these corpora are balanced RR=UAR.

Table 6 shows the result. Note that the last column is computed by averaging of the first two column at CINEMO and JEMO, while it is a real test for the mixed corpus. As it can be seen, the mixed corpus performs in between JEMO and CINEMO. It seems that performance levels out. There is no extra improvement. We can also conclude that the performance improvement in the previous chapter is mainly due to the large number of instances.

8. Male and Female Models

Differences in acoustic features for male and female speakers are a well-known problem and it is established that gender-dependent emotion recognizers perform better than gender-independent ((Lee and Narayanan, 2005), (Ververidis and Kotropoulos, 2004)).
As it can be seen in table 7 we get two slightly more unbalanced sub-corpora, which are approximately half of the size of the united corpus.

C. & J.	POS	SAD	ANG	NEU
Male	252	262	267	432
Female	377	325	256	494

Table 7: Female and Male sub-corpora of the united corpus, # of segments for 38 female and 50 male speakers.

Figures 3 and 4 show also some qualitative differences. The display of all the features and all the instances would be impossible, therefore we computed speaker-means for two selected feature. The means were computed per speaker and per class. The figure shows some difference between female and male speakers. Although only two important features are displayed we can see that female classes are somewhat different: male speakers sometimes speak with higher energy in anger and pitch than female. Meanwhile some female speakers have higher pitch in the class POS.

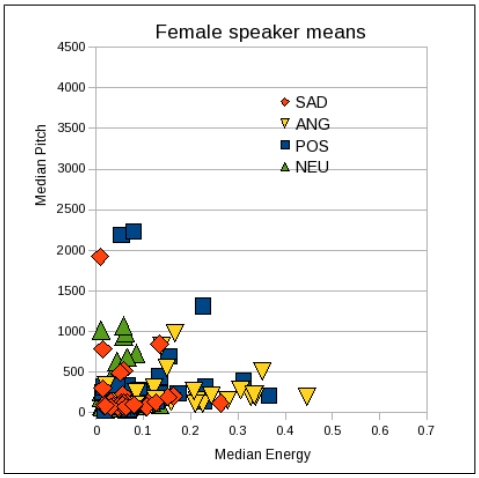

Figure 3: Female speaker means per classes in the space of MedianEnergy and MedianPitch. Note that several values of POS, NEU amd ANG are masked by other classes around the orig.

The reduction in size in the sub-corpora may results in an effect of decreasing performance, on the other hand, the specialization to the different kind of speakers should have an advantage. The balance of the two effects is hardly predictable, so it must be tested.

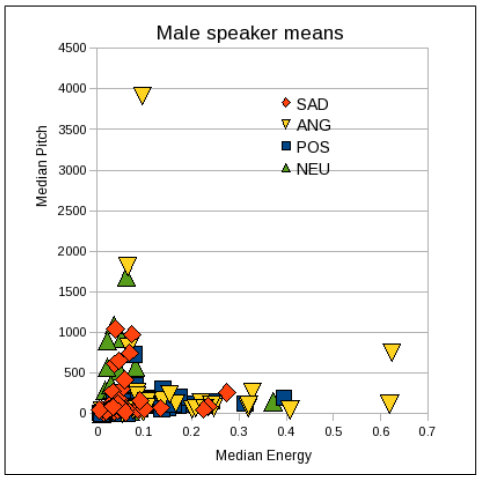

Figure 4: Male speaker means per classes in the space of MedianEnergy and MedianPitch.Note that several values of POS, NEU amd ANG are masked by other classes around the orig.

We carried out SpI CV tests for female and male sub-corpora, which are the female and male speakers of the united corpus.
As you can see in the first column of Table 8, the result of the SpI CV is 0.5229/0.4925 for the male model and 0.5927/0.5560 for the female one. Comparing to the united model (0.513/0.4691) we have a very significant development at the female model and a moderate increase of performance at the male model. The later one has a significant increase in performance considering UAR. In summary this means that haveing gender information the separate male and female model is beneficial.

9. Feature Selection

We chose Sequential Fast Forward Selection method ((Pudil et al., 2002) for feature selection as this is currently well established and widely used. We had to use for the algorithm LIBSVM (Chang and Lin, 2001) instead of Weka to be able to use a PBS cluster. Therefore we have only RR values.
Recognition Rate was also selected as the criterion function of SFFS. Table 8 shows the results.

RR/UAR	All features	SFFS
Female	0.5927/0.5560	0.6505
Male	0.5229/0.4925	0.5523

Table 8: Results for separate female and male models. Note that feature selection was made with RR so we do not have UAR.

The female model was running until 42 features, best performance is 0.6505 with 31 features. The male model was running also until 42 features, best performance is with 38 features. For Male and Female models, features selection allows to obtain better results, which is understandable. In the female and male corpora we have 1214 and 1453 instances. With 458 features, the average number of instances per class is lower than the number of features. Consequently, training with all the features results in some overtraining. Feature selection is not only beneficial in this case because the model is faster and smaller, but actually also better.

10. Conclusion

We have seen a united corpus set together from two, similar, but in some important aspects different kind of corpora. Beside the technical possibility of unification, it has several advantages. First, the number of instances is approximately doubled. Second, the classes are more balanced. And finally, the two corpora enrich each other. We have shown that the increase in performance is mainly due to the first effect, but the two others are also nice features. After the unification we also have shown that splitting the corpus along gender is also beneficial. Here, the number of instances decreases, and the unbalancedness increases. The two sub-corpora represent obviously different kinds of speakers. This splitting is however beneficial: the models trained on the sub-corpora are better. Since gender information is available this may be used in our application. In our paper we took always into account the field of application: an affective avatar. Consequently, our conclusions are directly used in our application.

11. Acknowledgement

This work was partly funded by the ANR Affective Avatar. We thank SME Voxler who provides the tools for segmentation and for extraction of the low-level features.

12. References

Ch.-Ch. Chang and Ch.-J. Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

L. Devillers and L. Vidrascu. 2007. *Emotion recognition, Speaker characterization*. Springer-Verlag.

L. Devillers and O. Vidrascu, L. end Layachi. 2009. Automatic detection of emotion from vocal expression. In *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*. Oxford University Press, Oxford.

L. Devillers, L. Vidrascu, and L. Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Netw.*, 18(4):407–422.

Ch. M. Lee and S. S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293–303.

J. C. Platt. 1999. Fast training of support vector machines using sequential minimal optimization. pages 185–208.

P. Pudil, J. Novovicová, and P. Somol. 2002. Feature selection toolbox software package. *Pattern Recogn. Lett.*, 23(4):487–492.

N. Rollet, A. Delaborde, and L. Devillers. 2009. Protocol cinemo: The use of fiction for collecting emotional data in naturalistic controlled oriented context,. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction, 2009*.

B. Schuller, R. Zaccarelli, N. Rollet, and L. Devillers. 2010. Cinemo a french spoken language resource for complex emotions: Facts and baselines. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

S. Steidl. 2009. Automatic classification of emotion-related user states in spontaneous children's speech. In *Studien zur Mustererkennung*, volume 28. Logos Verlag, Berlin.

D. Ververidis and C. Kotropoulos. 2004. Automatic speech classification to five emotional states based on gender information. In *Proc. 12th European Signal Processing Conference*, page pp. 341344.

I. H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.