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

M. Brendel, R. Zaccarelli, L. Devillers

LIMSI-CNRS; France
{brendel, zaccarelli, 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 of 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 for improving emotion detection performances. After feature/selection we obtain good results even with our strict speaker independent testing method.

1. Introduction

The modelling of realistic emotional behaviour is needed for various applications, like embodied agents, robots and dialog systems in call centres. 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. 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 et al., 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, 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 48 speakers and the second, JEMO is obtained by an emotion detection game with 33 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 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 and finally the union of both. Feature selection will be studied and also the use of separate models for male and female. 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. 48 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., ACII 2009, Schuller et al., submission to LREC 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	SAT	SAD	ANG	NEU
Male	112	197	140	149
Female	29	129	159	96

Table 1: CINEMO sub-corpus

As it can be seen in Table 1, satisfaction, a positive emotion is underrepresented in CINEMO.

2.2. The JEMO corpus

The JEMO corpus features 1062 instances after segmentation of speech recorded from 33 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, SATisfaction 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).

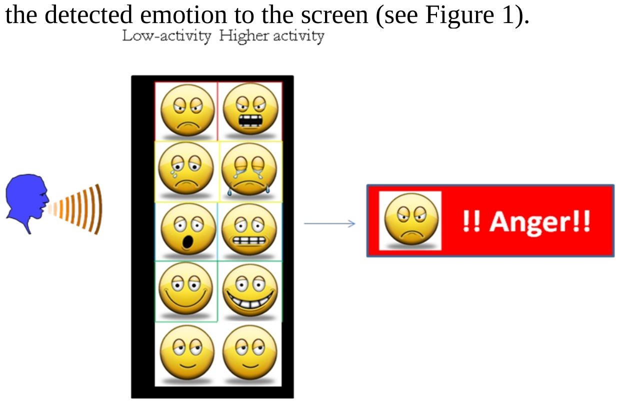

Figure 1: The emotion detection game used for recording JEMO

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. 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.

Table 2 is a description of the JEMO sub-corpus:

JEMO	SAT	SAD	ANG	NEU
Male	110	99	106	183
Female	177	97	90	199

Table 2: Class-statistics of corpus JEMO

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

3. Evaluation protocol

There is a consistent and significant difference between speaker dependent (SpD) and speaker independent (SpI) cross-validation. The SpD cross/validation (like the one in Weka) gives better results and is generally too optimistic. In our opinion SpI cross/validation shall be used as the standard evaluation/protocol (Brendel et al., submitted to ICASSP 2010). In our experiments we only use SpI cross-validation, except when we have 2 corpora, when speaker independence is guaranteed.

Our Speaker Independent (SpI) testing method is a 10-fold cross-validation made on the set of speakers instead of the instances. This way the separation of speakers in the train and test set is ensured, with a cost of unbalancedness, which is sufficiently enough corrected by the cross-validation.

Evaluation between the two corpora is made by simple train and test. SpI cross-validation is used if a corpus is tested on itself. In this case cross-validation is done 10 times and the accuracy values are averaged not only for the folds, but for 10 random runs.

4. Features

For all corpora each segment is then passed through an executable which performs spectral (16 MFCCs) and prosodic analysis (pitch, zero-crossing and energy). Eventually, each segment corresponds to a text file storing the time series describing pitch/energy contours and MFCCs. The time series were passed into a feature extractor executor programmed in Java. The feature extractor calculates basic statistical features on voiced parts: min, max, mean, standard deviation, range, median quartile, third quartile, min and max intra range (of each voiced segment), min and max inter range (between voiced segments). We got 237 features: 13 for pitch, 32 for energy (12 RMS, 12 energy and 8 zero-crossings) and 192 for MFCC1-16. Finally, we trained the data set using Weka (Witten et al., 1999) with a SVM with a RBF-kernel by means of Sequential minimal optimization algorithm (Platt, 1999).

5. Results on both corpora independently

First, we tested the two corpora on separate and on each other to see, what errors each corpus causes.

	Tested on CINEMO	Tested on JEMO
Trained on CINEMO	0.4833	0.4958
Trained on JEMO	0.3778	0.5803

Table 3: Training and testing on the 2 corpora

As it can be seen in Table 3 JEMO performs better than CINEMO, this is mainly because it contains prototypical emotions. The same reason can be given why the model trained on CINEMO performs as well, moreover, slightly better on JEMO.

[EXTENSIVE ANALYSES PROVIDED IN THE FULL VERSION]

6. Results on the union of both corpora

The union of both corpora 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. Average recognition rate is 0.56, which is a little bit worse than JEMO on itself, but much better than any other value. This means that the unification of the corpora improved the results obtained on 4 classes. We could not be better than JEMO on itself, 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 itself.

[EXTENSIVE ANALYSES PROVIDED IN THE FULL VERSION]

6. 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).

The downside is that the divided corpus is smaller. The result is 0.54 for the male model and 0.63 for the female one. It seems that with this corpus we can only benefit from the female model.

[EXTENSIVE RESULTS PROVIDED IN THE FULL VERSION]

7. Feature selection

We chose Sequential Fast Forward Selection method (Somol and Pudil, 2009) for feature selection as this is currently well established and used widely.

Male feature selection was running until 50 features. Best accuracy is 0.56 with 38 features.

Female model was running until 50 features, best accuracy is 0.64 with 37 features.

For Male and Female models, features selection allows to obtain better results. The number of the features is similar, but an important difference is that no energy-feature is used in the female model, while it is used extensively in the male one.

[EXTENSIVE RESULTS and ANALYSES PROVIDED IN THE FULL VERSION]

8. Conclusion

Our conclusion will be on the possibility to mix different kind of corpus for training more efficient classifiers.

9. 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.

7. References

Brendel, M., Zaccarelli, R. Devillers L., (2010), Towards a speaker independent Evaluation protocol for emotion-detection from speech, submitted to ICASSP 2010.

Devillers, L., Vidrascu, L., Layachi, O., (2009) Automatic detection of emotion from vocal expression, in Scherer, K.R., Bänziger, T., & Roesch, E. (Eds.) A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing. Oxford: Oxford University Press (under press).

Devillers, L. Vidrascu L. & Lamel, L. (2005). “Challenges in real-life emotion annotation and machine learning based detection”, Neural Networks 18, pp. 407-422.

Devillers, L., Vidrascu, L. (2007), Emotion recognition, «Speaker characterization », Christian Müller, Susanne Schötz (eds.), Springer-Verlag.

Devillers, L., Cowie, R., Martin, J.-C., Douglas-Cowie, E., Abrilian, **S.**, McRorie, M. (2006) Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. (LREC 2006), Genoa, Italy, 24-27 may.

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

Platt, J. (1999): Fast training of Support Vector Machines using Sequential Minimal Optimization.

In: Advances in Kernel Methods – Support Vector Learning. MIT Press (1999)

Rollet, N., Delaborde, A., & Devillers, L. (2009) “Protocol CINEMO: The use of fiction for collecting emotional data in naturalistic controlled oriented context,” in Proc. ACII, Amsterdam, The Netherlands, 2009, IEEE.

Schuller, B., Zaccarelli, R., Rollet, N., Devillers L. (2010) CINEMO – A French Spoken Language Resource for Complex Emotions: Facts and Baselines, submitted to LREC 2010.

Somol, P. Pudil P. (2002): “Feature selection toolbox”, Pattern Recognition, 35(12): 2749-2759, 2002.

Ververidis D. and Kotropoulos C. (2004). Automatic speech classification to five emotional states based on gender information. In *Proc. 12th European Signal Processing Conference*, pages pp. 341–344, Vienna, September.

Witten, I.H., Franck, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S.J., (2009) “Weka: Practical machine learning tools and techniques with Java implementations”, Proc ANNES’99 International Workshop: Emerging Engineering and Connectionnist-Based Information Systems, 1999, p 192-196.