Automatic classification of emotions via global and local prosodic features on a multilingual emotional database

Antonio Origlia[1], *Vincenzo Galatà*[2], *Bogdan Ludusan*[1]

[1] Natural Language Processing Group, Department of Physical Sciences, "Federico II" University, Naples, Italy
[2] Laboratory of Phonetics, Department of Linguistics, University of Calabria, Italy
antori@gmail.com, vgalata@libero.it, ludusan@na.infn.it

Abstract

In this paper we introduce the €motion database, a multilingual emotional database consisting of emotional sentences elicited in four European languages: Italian, French, English and German. Along with this, a new set of features, containing both global and local prosodic features, for automatic classification of emotions is presented and their appropriateness for this task tested. The results obtained using these features on a monolingual emotional database are comparable with the state of the art results previously obtained.

Index Terms: automatic emotions classification, multilingual emotional database, prosodic features

1. Introduction

Searching among the 64 speech resources reviewed in [1] for a multilingual emotional speech database, it is surprising to find out that such resources are very rare, especially when dealing with European languages. According to it, there are very few resources containing speech material in two languages (English and German in [2, 3], English and Slovene in [4], English and Spanish in [5]), and no database containing material in Italian language. Due to this lack of consideration for the Italian language in multilingual emotional speech resources, €motion was realized.

Previous research on speech emotion recognition systems (SER) followed different directions to provide a set of features that would be descriptive for emotions:

- using fixed sets of features.
- applying feature selection algorithms to reduce large sets of features before performing classification.

However, both approaches have weaknesses: using a fixed set of features exposes to the risk of missing something important, while feature selection causes systems to use a different set of features each time a new data set is analyzed, thus giving less insight into the emotions phenomenon and increases the complexity of the classifier.

Features used for emotion recognition are usually global statistic measures on the voice signal. These measures are related to prosodic analysis, so F0 and intensity are mainly taken into account. Even though it has been stated that global features are more suitable to be employed in SER systems [6], we present here a method to locally extract features in areas of the signal we deemed relevant for emotions conveying. We followed the indications in [7] regarding the relations between the prosodic and the segmental level of a speech signal and used a syllable based segmentation for extracting the local features.

Emotional classification is usually performed using a wide range of classifiers: Neural Networks (NN), Support Vector Machines (SVM), as well as methods adapted from other fields. In this paper we present the results obtained using a Neural Network and a Support Vector Machines trained with our set of features achieved on the Berlin database of emotional speech (EMO-DB) [8] and on the €motion corpus. The Berlin corpus has been used to perform a monolingual test while the €motion corpus has been used to perform multilingual tests and to check the performance difference when classifying actor speakers against naive speakers.

2. €motion: a multilingual emotional database

€motion represents a first attempt to experimentally collect an emotional speech database in four European languages (Italian, French, English and German). Its original purpose [9] was referred to a research on vocal emotions in a cross-linguistic and cross-cultural study described extensively in [10]. The objective of this survey was, among others, to verify people's ability to decode vocal emotions in European languages: the main idea is that there are, perhaps, more elements in common between closer cultures than in cultures being poles apart (e.g. American vs. Japanese or Chinese).

2.1. General observations

Research on vocal emotions has followed research on facial expression, but as [11] underlines "Contrary to research on the face, researchers interested in the voice cannot present a valid 'snapshot' which represents the vocal attributes of an emotion." This difficulty has been challenged in different ways causing emotional speech resources to differ among them for their characteristics. These include primarily choice and number of encoders (with reference to gender and type: actor vs. naive speakers); recorded speech units (such as vowels or syllables, words, non-sense utterances, interjections, and so on); kind of speech (spontaneous vs. laboratory speech, read or acted); emotion elicitation procedures (photographs, scenarios, verbal labels, etc.); last but not least, the number of emotions considered differs greatly, too.

Because of the great differences between emotional speech databases, a comparison among them is often not possible. This issue suggested the creation of the €motion database.

2.2. Description of the €motion database

The €motion database has been collected bearing in mind to administer a perceptual test to listeners to verify their ability to decode emotions in a foreign language as well as to verify how much of the differences in the recognition accuracy rates were due to language differences.

For this reason the €motion database was collected for four European languages, namely Italian, French, English and

German, considering the emotions defined in [12] as *basic emotions* (happiness, anger, fear, sadness, disgust, surprise). A neutral version was also considered.

All the recordings have been collected in a soundproof room in a 44.1kHz - 16 bit - mono *.wav format using a *Sennheiser e835* microphone and an *Edirol R-09* digital recorder.

What distinguishes €motion from other emotional speech resources is given by the attempt to collect emotional speech using both actors and naive speakers following an experimental emotion elicitation protocol.

2.2.1. *Choosing a carrier sentence*

The choice of a carrier sentence was highly sought-after following some *a-priori* established *desiderata*, according to which the sentence had to be:

- Semantically neutral if taken out of a specific context: every sentence spoken during a speech act carries an amount of information that cannot be excluded *a-priori*. Since the interpretation and emotional value is in the mind of the listener and not in the sentence itself, the semantically neutral refers here to the possibility of inserting the sentence in any given scenario of a specific emotion;
- Consistent with the general construct of any of the situations presented;
- Correct according to the general rules of each language examined avoiding confusion or distraction firstly in the encoders, and subsequently in the listeners;
- Be easy to analyze, meaning that the sentence onset should be easily identified for analysis: sentence onset with unvoiced consonants was for this reason avoided.

Following previous studies (such as [13, 14]) the resulting sentence choice was:

- For Italian: *Non è possibile. Non ci posso credere.*
- For French: *Oh là là. C'est incroyable.*
- For English: *It can't be. I cannot believe it.*
- For German: *Das ist nicht möglich. Ich kann es nicht glauben*.

2.2.2. *Emotion elicitation protocol*

Different methodologies and emotion eliciting procedures have been proposed to collect emotional speech material so far (see [15] for a full review).

However, when dealing with more languages, robust and reproducible procedures are not only suggested but absolutely needed if comparable emotional speech material is desired.

Subjects (mainly actors) involved in the recordings, are usually asked to speak out the text "x" with the emotion "y". This instruction can cause great variability and confusion in cross-language studies: in different languages different emotional labels can be used to refer to a given emotion, and a specific emotional label can refer to more affective states.

To avoid the use of verbal labels, an adapted version of the scenario approach used in [15] was adopted by inserting the carrier sentence in a scenario for each of the six emotions.

Since a certain degree of spontaneity was desired, specific instructions were given to the subjects: these instructions represent an experimental protocol to test the influence on the emotional productions caused by the amount of information made available to encoders.

The material was therefore recorded according to the above mentioned 4 instructions, resulting in 4 recording modes:

- Mode A: after a private reading, read again the six scenarios with sense and in a natural and spontaneous way;
- Mode B: read the text once more with sense and in a natural and spontaneous way considering the desired emotion letting yourself personally get involved in the story proposed in the text;
- Mode C: repeat the carrier sentence according to the requested emotion and to the scenario proposed in each text;
- Neutral mode: simply read a list of sentences containing the carrier sentence.

No information about the experiment was given to the subjects involved until the end of the recording procedure.

2.2.3. *Collecting the speech material*

Once the experimental setting has been established a total of 12 male actors and 27 male naive speakers were recruited and recorded (for further information about the encoders please refer to Table 1).

Following the elicitation procedure above described, for each of the subjects the following material was recorded:

- Mode A: 6 productions (1 per emotion);
- Mode B: 6 productions (1 per emotion);
- Mode C: 24 productions (4 per emotion);
- Neutral mode: 4 neutral productions.

Table 1. *Summary for the recorded subjects.*

Language	Encoder	N°	Mean age (sd)
Italian	actor	5	35,8 (5,2)
	naive	6	25,8 (1,6)
French	actor	3	43,7 (11,0)
	naive	6	29,5 (12,8)
English	actor	2	46,0 (8,5)
	naive	6	45,0 (10,9)
German	actor	2	50,5 (2,1)
	naive	9	26,0 (3,9)

The collecting procedure produced a total of 1560 files as reported in Table 2.

Table 2. *Total number of productions recorded per language and type of encoder.*

	Naive	Actor	Total	N° files
Italian	6	5	11	*440*
French	6	3	9	*360*
English	6	2	8	*320*
German	9	2	11	*440*
Total	*27*	*12*	*39*	*1560*

2.2.4. *Database validation and stimuli selection*

All the collected files were validated using the same encoders recruited for the recordings. Each encoder, this time considered as judge, listened to his productions and to those of other two encoders. Each set of stimuli for each encoder has been arranged to be judged by three judges/encoders.

The validation procedure was carried out by means of a forced choice test with 7 options available (happiness, anger, fear, sadness, disgust, surprise, and neutral) using the PRAAT software [16]. Judges were asked to carefully listen to the randomly presented files and to indicate which of the choices available they recognized in the presented files. Judges were not allowed to go back to previously presented stimuli.

All those stimuli correctly recognized by at least 2 out of 3 judges were retained valid. See Table 3 for a more detailed account of the stimuli available per language and emotion.

Table 3. *Summary of the €motion stimuli selected for the present study.*

Language	Emotion							Tot
	ang	dis	fea	joy	neu	sad	sur	
Italian	40	28	26	26	33	43	24	220
French	27	8	15	17	32	37	27	163
English	24	9	19	8	23	34	23	140
German	42	6	25	21	38	44	22	198

3. The classification task

3.1. Feature Extraction

The first step to perform automatic emotion recognition is to extract features from the audio files in the data set being analyzed. Due to the fact that emotions are vocally conveyed through intonational patterns, prosodic analysis is the primary key to perform classification.

To perform Feature Extraction, we used the speech analysis software PRAAT. The global features we extracted are mostly classic measures like mean, standard deviation, max and min values for F0 and intensity. The most relevant global features other than the traditional ones are statistical measures on harmonicity and on the spectral centroid.

Harmonicity [17] refers to the degree of acoustic periodicity within the signal or Harmonic-to-Noise Ratio (HNR) and helps in finding emotionally relevant areas being at the same time a feature itself.

The idea of using local features along with global ones is motivated by the fact humans tend to use localized intonational strategies to convey support data, like their own emotional state [18] or the interrogative/affirmative intent of the sentence [8], along with a generic voice modification. To extract local features, we developed a method to identify areas of the signal carrying more emotional information. These emotionally descriptive region of the signal correspond to sustained syllable nuclei and the algorithm used for their identification is presented in the next paragraph.

To perform syllable nuclei search we used an energy based approach like the one presented in [19]. After filtering and enhancing the signal in the frequency domain, we filtered it in the time domain to remove a number of artifacts and we took as syllable nuclei energy prominences complying with a rules set. To recognize syllable nuclei carrying emotional data we considered the length of the harmonic region in which the nucleus falls: if this length was higher than a fixed threshold, local features were extracted in the harmonic region. In Figure 1 we show the energy profile of a speech signal along with its harmonicity and pitch curves.

The local features we extracted were statistical measures regarding the length of sustained nuclei, their energy and the voice quality of the sustained vowel in terms of jitter and shimmer. The use of these features has been indicated in previous studies to provide aid in emotion recognition, for example in [20]. The total number of features employed, both global and local, is 31.

Syllable nuclei search has also been used to evaluate the speaker's speech rate by considering the mean occurring frequency of syllable nuclei.

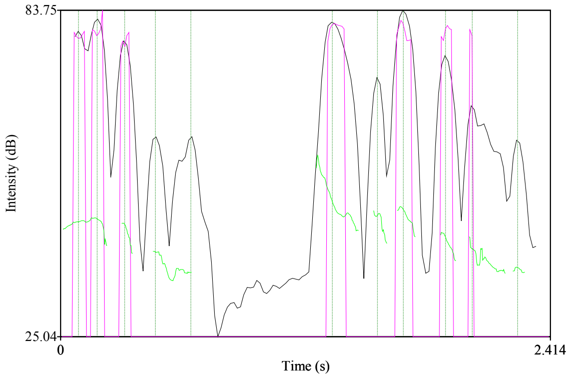

Figure 1: *An energy profile (black) of a speech signal along with its pitch (green) and harmonicity (magenta) curves. Syllable nuclei are marked with dashed vertical lines.*

3.2. Classification experiments

The feature extraction algorithm was applied both on the €motion and on the Berlin corpus using two classifiers: Multi Layer Perceptron (MLP) and a Support Vector Machine. The test protocol was the stratified 10-fold cross-validation. MLP experiments were conducted using the WEKA [21] environment while SVM experiments were conducted using the libSVM [22] package.

The MLP we used is a feed-forward neural network, having a single hidden layer with 76 nodes. For obtaining the number of nodes of the hidden layer that balances the NN complexity and its accuracy, the following formula was used:

$$N=2(n_c+n_f) \quad (1)$$

where N is the number of nodes of the hidden layer, n_c is the number of classes and n_f is the number of features. The network was trained using the Back-Propagation method.

A test was conducted to find the most suited kernel for the SVM classification: the RBF kernel was found to be the one giving the best results. Thus, for our SVM classifier we used the RBF kernel.

First, we tested our feature set against the MLP reported in [23], which is the best MLP we found in literature, and against the new algorithm presented in the same work, which is TGI+.2. The data set we used was the one obtained from the Berlin corpus analysis. As reported in Table 4, our MLP and our SVM achieved accuracy levels identical to the reference MLP, while being inferior to TGI+.2. This, however, is sufficient to check the predictive power of a data set including both globally and locally extracted features.

Table 4. *Accuracy levels achieved by each algorithm on the Berlin corpus*

Algorithm	Accuracy
Reference MLP	73,9%
TGI+.2	78,58%
MLP	73,83%
SVM	73,64%

Since we do not use feature selection by preferring a fixed feature set, we wanted to be sure that it would not perform well only on the Berlin corpus because, as stated in [24], we could have been just "lucky to guess the right set". We also wanted to test if the fact of emotions being cross-cultural could be checked in speech as it has been done in facial expressions [25] and, finally, we wanted to check the accuracy difference between actor and naive speakers. The results obtained on the €motion corpus, using MLP as well as SVM, are reported in Table 5.

Table 5. *MLP and SVM accuracies on the €motion corpus and subsets of it*

Dataset	MLP	SVM
€motions - full	58,31%	56,03%
€motions – actors	70,3%	64,33%
€motions – naive	60,1%	54,40%

As we expected, naive speakers, even not being really spontaneous, are far more difficult to recognize than actors are as it has been shown in [26]. From the test on the actors subset, we understand that emotions are cross-lingual because the accuracy level we achieved is similar to the one obtained on the monolingual test. We use this subset of the €motion corpus for the comparison because the Berlin corpus contains actor speakers only.

4. Conclusions and future work

The accuracies obtained on the two databases are very promising and seem to suggest that locally extracted features do improve speech emotion recognition. Results analysis has shown that local features improved the accuracy of an MLP using global features only of about 2% on the full €motion set, 7% on the actors subset and 5% on the naive subset respectively. This suggests that global and local features bring complementary information to the automatic classification of emotions task. Another fact that encourages us is that the same set of features made the classification algorithms perform quite well on two data sets showing a very high degree of dissimilarity. In particular, we obtained good results on a database, €motion, considerably more difficult to deal with than EMO-DB.

As a future development of this work we would like to obtain a representation capable of capturing the sequence in which local prosodic events appear in the signal. We will also investigate the performance drop occurring when mixing actors and naive speakers: at present time, we believe that trying to build a unique model for two kind of speakers that are proven to convey emotions differently is the key.

Acknowledgments

Bogdan Ludusan's work was supported by the EU FP6 Marie Curie Research Training Network "Sound to Sense".

5. References

[1] Ververidis, D. and Kotropulos, C., "Emotional speech recognition: Resources, features, and methods", Speech Communication, 48: 1162-1181, 2006.

[2] Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russel, M., and Wong, M., " 'You stupid tin box' - children interacting with the AIBO robot: a cross-linguistic emotional speech corpus", in: Proceedings of the 4th International Conference of Language Resources and Evaluation (LREC '04), pp. 171-174, 2004.

[3] Scherer, K. R., Grandjean, D., Johnstone, L. T., G. Klasmeyer, T. B., Acoustic correlates of task load and stress, in: Proceedings of the International Conference on Spoken Language Processing (ICSLP '02), 3: 2017-2020, 2002.

[4] Ambrus, D. C., "Collecting and recording of an emotional speech database", Technical Report, Faculty of Electrical Engineering, Institute of Electronics, Univ. of Maribor, 2000.

[5] Gonzalez, G. M., "Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in Chicanos/Latinos", Technical Report 39, Univ. Michigan, 1999.

[6] Schuller, B., Rigoll, G., Lang, M., "Hidden Markov Model-Based speech Emotion Recognition". Proc. ICASSP 2003, Vol. II, pp. 1-4, Hong Kong, China, 2003.

[7] D'Imperio, M., House, D., "Perception of Questions and Statements in Neapolitan Italian", EUROSPEECH-1997, 251-254.

[8] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., "A database of German emotional speech", in Proc. of Interspeech '05, Lisbon, Portugal, pp. 1517-1520, 2005.

[9] Galatà, V., Romito, L., "Un corpus sperimentale per lo studio cross-linguistico europeo delle emozioni vocali", in: Proceedings of the V AISV Conference - La dimensione Temporale del parlato, University of Zürich, Switzerland, 4-6 february 2009, pp. 603-641, 2010.

[10] Galatà, V., "Production and perception of vocal emotions: a cross-linguistic and cross-cultural study", PhD Thesis, University of Calabria, Italy, (unpublished).

[11] Pell, M. D., Monetta, L., Paulmann, S., Kotz, S. A, "Recognizing emotions in a foreign language", Journal of Nonverbal Behavior, 33: 107-120, 2009.

[12] Ekman, P., "An argument for basic emotions", Cognition and Emotion, 6: 169-200, 1992.

[13] Wallbott, H. G., Scherer, K. R., "Cues and channels in emotion recognition", Journal of personality and social psychology, 51: 690-699, 1986.

[14] Anolli, L., Wang, L., Mantovani, F., De Toni, A., "The Voice of Emotion in Chinese and Italian Young Adults", Journal of Cross-Cultural Psychology, 39: 565-598, 2008.

[15] Coan, J. A., Allen, J. J. B., "The handbook of emotion elicitation and assessment", London: Oxford University Press, 2007.

[16] Boersma, P., Weenink, D., "Praat: doing phonetics by computer" [Computer program], retrieved from http://www.praat.org/.

[17] Boersma, P., "Accurate short term analysisof the fundamental frequency and the Harmonics-to-Noise Ratio of a sampled sound", Institute of Phonetic Sciences, University of Amsterdam, Proceedings 17 (1993), 97-110.

[18] O'Reilly, M., & Ni Chasaide, A., Analysis of intonation contours in portrayed emotions using the Fujisaki model. The Second International Conference on Affective Computing and Intelligent Interaction. Proceedings of the Doctoral Consortium, 2007, pp. 17-24.

[19] M. Petrillo, F. Cutugno, "A Syllable Segmentation Algorithm for English and Italian", EUROSPEECH 2003, 1-4.

[20] Drioli, C., Tisato, G., Cosi, P., Tesser, F., "Emotions and Voice Quality: Experiments with Sinusoidal Modeling", Proceedings of Voqual 2003, 27-29.

[21] Hall, M., Frank, E., Holmes, G. Pfahringer, B. Reutemann, P., Witten, I.H., "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1.

[22] Chih-Chung Chang, Chih-Jen Lin, "LIBSVM : a library for support vector machines", 2001. Software downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm on 20/03/2009.

[23] Sidorova, J., "Speech emotion recognition with TGI+.2 classifier", Proceedings of the EACL 2009 Student Research Workshop, 54–60.

[24] Sidorova, J., DEA report: Speech Emotion Recognition, http://www.glicom.upf.edu/tesis/sidorova.pdf

[25] Ekman, P., Oster, H., "Facial expressions of emotion" Annual Review of Psychology, 527-554.

[26] Vogt, T. Andre, E., "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", Proc. ICME 2005.