10.12871/clicit2014220

EVALITA 2014: Emotion Recognition Task (ERT)

Antonio Origlia
University “Federico II”, Napoli, Italy
antonio.origlia@unina.it

Vincenzo Galatà
ISTC-CNR, UOS Padova, Italy
Free University of Bozen-Bolzano, Italy
vincenzo.galata@pd.istc.cnr.it

Abstract

English. In this report, we describe the EVALITA 2014 Emotion Recognition Task (ERT). Specifically, we describe the datasets, the evaluation procedure and we summarize the results obtained by the proposed systems. On this basis we provide our view on the current state of emotion recognition systems for Italian, whose development appears to be severely slowed down by the type of data available nowadays.

Italiano. *In questo report, descriviamo il task EVALITA 2014 dedicato al riconoscimento di emozioni (ERT). In particolare, descriviamo i set di dati utilizzati, la procedura di valutazione e riassumiamo i risultati ottenuti dai sistemi proposti. Su questa base, descriveremo la nostra posizione sullo stato attuale dei sistemi per il riconoscimento di emozioni per l’Italiano, il cui sviluppo sembra essere fortemente rallentato dal tipo di dati disponibili attualmente.*

1 Introduction

After the Interspeech 2009 Emotion Challenge (Schuller et al., 2009) and the Interspeech 2010 Paralinguistics Challenge (Schuller et al., 2010), the EVALITA Emotion Recognition task (ERT) represents the first evaluation campaign specifically dedicated to Italian Emotional speech. Unlike the two Interspeech challenges, we move here the first steps for Italian by using acted emotional speech collected according to Ekman’s classification model (Ekman, 1992) as this is, so far, the only type of speech material we have knowledge. In this task, we aimed at evaluating the performance of automatic emotion recognition systems and to investigate two main topics, covered by two different subtasks:

- cross language, open database task
- Italian only, closed database task

First of all, we wanted to estimate the performance that could be obtained on Italian using emotional speech corpora in other languages. We also wanted to verify to what extent it would have been possible to build a model for emotional speech starting from a single, professional, speaker portraying the discrete set of emotions defined by Ekman (1992) (anger, disgust, fear, joy, sadness, surprise, and neutral).

In this first evaluation of emotional speech recognition systems on Italian, the material we use is composed of acted speech elicited by means of a narrative task. The material is extracted from two emotional speech corpora containing similar material and sharing basic characteristics:

- the E-Carini corpus
- the €motion corpus

Concerning the second subtask, the goal of the evaluation was to establish how much information could be extracted from material coming from a single, professional source of information whose explicit task is to portray emotions and obtain models capable of generalizing to unseen subjects.

2 Datasets

For both development and training sets, *.wav files were provided along with their Praat *.TextGrid file containing a word level (wrd) annotation carried out by means of forced alignment. Pauses in the *.TextGrid file are labelled as “.pau”. The material consists of PCM encoded WAV files (16000Hz).

2.1 Development set: the €motion corpus

Participants were provided with a development set taken from the yet unpublished €motion corpus (Galatà, 2010) to obtain reference results for the test material during the system preparation time. The material extracted from €motion consists of the Italian carrier sentence "Non è possibile. Non ci posso credere." (*It can't be. I can't believe it.*), recorded by one professional actor according to 4 instructions (or recording modes) as follows:

- 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 himself 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).

Following the above described elicitation procedure, the 40 sentences were provided as development set:

- 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.

The file name structure for this data set provides information on the way the sentence has been collected as well as the discrete emotion label assigned and intended for its production. Given the file name it_ang_a_mt_c1 as example, the file name provides the following information:

- Language: it;
- Intended emotion: 6+1 discrete emotion labels (eg. ang, sur, joy, fea, sad, dis, neu);
- Type of subject: a (actor);
- Subjects name: mt;
- The recording mode: a, b or c (for the neutral mode this slot is left out);
- Occurrence number: 1, 2, 3 or 4.

2.2 Training set: the E-Carini corpus

The material provided for the E-Carini corpus (Avesani et al., 2004; Tesser et al., 2004; Tesser et al., 2005), consists of a reading by a professional actor of the short story "Il Colombre" by Dino Buzzati. The novel is read and acted according to the different discrete emotion labels provided. The novel is split in 47 paragraphs (from *par01* to *par47* in the file name) and stored in different folder (one for each emotion). This training set provided for the *closed database* task consisted of 1 hour and 17 minutes of speech.

2.3 The test set

All the participants were provided with the test set consisting of emotional productions by 5 actors with the same characteristics as in the *development set* above described. For each emotion, 30 stimuli were included in the test set. In order to allow speaker dependent system training, 4 neutral productions were provided for each speaker in the test set.

All the file names provided for the *test set*, apart from the neutral ones, were masked: the subject ID was, however, available to the participants, while the target emotion was kept hidden. The format given to the files contained the subjects name followed by a three digits random number (eg. as_108). Neutral files followed the format provided with the *development set* files.

3 Evaluation measure

Typically, the objective measure chosen for an emotion classification task would be the F-measure. However, as in this case, the sample accuracy (percentage of correctly classified instances) is used. Since the test set here distributed contains the same number of examples for each class, there is no influence to take into account on the side of data distribution and the sample accuracy results in a better choice.

3.1 Baseline

For the emotion recognition system baseline, we used the features set obtained with the OpenSMILE package (Eyben et al., 2013) in the configuration used for the Interspeech 2010 Paralinguis-

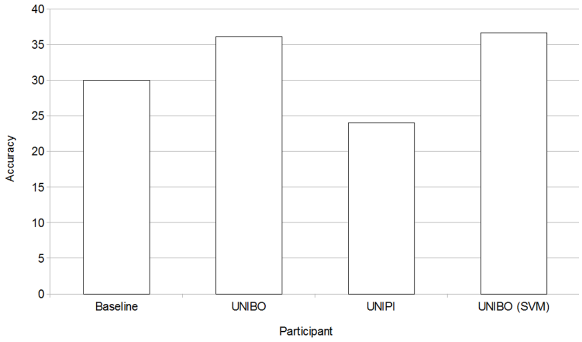

Figure 1: Summary of the submitted results. We also report the experiment provided by UNIBO with an SVM trained with finer parameter optimization than the one used as a baseline.

tics Challenge (Schuller et al., 2010). The LibSVM (Chang and Lin, 2011) implementation of Support Vector Machines (SVM) was trained on this data using an RBF kernel and a basic strategy to optimize the γ and C parameters (grid search between 0 and 2 with 0.4 grid step for both). The obtained classifier reached an accuracy of 30% on the test set.

4 Participation and results

Before receiving the material, all participants were asked to sign an End User License Agreement (EULA). Four participants downloaded the datasets after publication on the EVALITA website.

However, after receiving the test material, only two participants submitted the final test results for the "closed database" subtask and no one for the "open database" subtask. A system from the University of Bologna (UNIBO) and the University of Pisa (UNIPI) were proposed. Results were submitted to the organizers as a two columns *.csv file: the first column containing the file name and the second column the label assigned by the proposed system (eg. as_100, ang; eo_116, fea; etc.).

After the results submission, the participants were provided with a rename table mapping the masked file names on the original ones in order to let them replicate the evaluation results. In the following subsections we summarize the proposed approaches, while in Figure 1 we show the graphical comparison among the approaches with their respective recognition accuracies.

4.1 UNIBO

The system presented by UNIBO performed emotion recognition by means of a Kernel Quantum Classifier, a new general-purpose classifier based on quantum probability theory. The system is trained on the same feature set used for the baseline. The system reached a performance of 36.11% recognition accuracy, which is the highest result obtained in the ERT.

4.2 UNIPI

The system presented by UNIPI used an Echo State Network (Jaeger and Haas, 2004) to perform emotion classification. The system has the peculiarity of receiving, as input, directly the sound waveform, without performing features extraction. Neutral speech productions for each speaker were used to obtain waveform normalization constants for each speaker. Using the proposed approach, a recognition accuracy of 24% was obtained on the test set.

5 Discussion

The results obtained in the ERT task highlight an important problem for emotion recognition speech in Italian concerning the available material. While corpora containing Italian acted emotional productions have been successfully used for emotional speech synthesis in the past (this is the case of the E-Carini corpus), it appears it is not straightforward to transfer the model built on one professional actor portraying a set of specific emotions on other subjects, even if they are professional actors too. As a consequence, we believe that the type of emotional speech data available nowadays is inadequate to train emotion recognition systems for Italian. The reason for this inadequacy is mainly due to the difference between the type of data collected so far for Italian and the data that have been collected in other countries (mostly English speaking). For Italian, other than the E-Carini and the €motion corpus, to our knowledge only the EMOVO corpus (Iadarola, 2007; Costantini et al., 2014) is available. This dataset, as the ones here adopted, also contains acted read speech classified using Ekmans schema. Outside Italy, on the contrary, the scientific community appears to be oriented towards more spontaneous speech, mostly elicited through dialogue with artificial agents in a Wizard of Oz setup and annotated with both emotional classes and with continuous measures as done, for example, in the SEMAINE corpus (McKeown et al., 2010). As a matter of fact, the latest international challenges on emotion recognition are evaluated on the capability of automatic systems to track continuous values over the entire utterance (regression), as opposed to recognizing a single class over a full sentence (classification).

In conclusion, the result of the EVALITA 2014 ERT task seems to highlight that the type of data available in Italian emotional speech corpora is outdated at least for the emotion recognition task. Two problems are, in our opinion, important for the Italian community to tackle. First of all, we have observed that it is not straightforward to transfer the knowledge acquired by modelling a single professional source to other professional sources even in the case of read speech in silent conditions with a neutral speech basis available. This indicates that it is necessary for the Italian community working on emotional speech recognition to move away from this kind of data and collect more spontaneous data.

The second problem lies in data annotation. On an international level, automatic classification according to Ekmans basic emotions has been abandoned in favour of dimensional models as proposed, for example, by Mehrabian (1996). We believe it is necessary for the Italian community to move forward in this sense too as the global attention appears to be focused on dimensional annotations.

Acknowledgments

This work has been funded by the European Community and the Italian Ministry of University and Research (MIUR) and EU under the PON OR.C.HE.S.T.R.A. project. Vincenzo Galatà's work has been supported by WIKIMEMO.IT (The Portal of Italian Language and Culture, FIRB-MIUR Project, RBNE078K93).

References

Cinzia Avesani, Piero Cosi, Elisabetta Fauri, Roberto Gretter, Nadia Mana, Silvia Rocchi, Franca Rossi, and Fabio Tesser. 2004. Definizione ed annotazione prosodica di un database di parlato-letto usando il formalismo ToBI. In *Il parlato italiano*, pages 1–14.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. EMOVO corpus: an Italian emotional speech database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838.

Vincenzo Galatà. 2010. *Production and perception of vocal emotions: a cross-linguistic and cross-cultural study*. Ph.D. thesis, University of Calabria - Italy.

Iacopo Iadarola. 2007. EMOVO: database di parlato emotivo per l'italiano. In *Atti del 4 Convegno Nazionale dell'Associazione Italiana Scienze della Voce (AISV)*.

Herbert Jaeger and Harald Haas. 2004. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80.

Gary McKeown, Michel Franois Valstar, Roderick Cowie, and Maja Pantic. 2010. The semaine corpus of emotionally coloured character interactions. In *Proc. of ICME*, pages 1079–1084.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology: Developmental, Learning, Personality, Social*, 14:261–292.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The INTERSPEECH 2009 emotion challenge. In *Proc. of Interspeech*, pages 312–315. ISCA.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan. 2010. The INTERSPEECH 2010 paralinguistic challenge. In *Proc. of Interspeech*, pages 2794–2797.

Fabio Tesser, Piero Cosi, Carlo Drioli, and Graziano Tisato. 2004. Modelli prosodici emotivi per la sintesi dell'italiano. *Proc. of AISV 2004*.

Fabio Tesser, Piero Cosi, Carlo Drioli, and Graziano Tisato. 2005. Emotional Festival - Mbrola TTS synthesis. *Interspeech 2005*, pages 505–508.