EMOVO Corpus: an Italian Emotional Speech Database

Giovanni Costantini[1,2], Iacopo Iadarola[3], Andrea Paoloni[3], Massimiliano Todisco[1,3]
[1]Department of Electronic Engineering, University of Rome "Tor Vergata", Rome, Italy
[2]Institute of Acoustics "O. M. Corbino", Rome, Italy
[3]Fondazione "Ugo Bordoni", Rome, Italy
E-mail: costantini@uniroma2.it, iacopo.iadarola@gmail.com, pao@fub.it, massimiliano.todisco@uniroma2.it

Abstract

This article describes the first emotional corpus, named EMOVO, applicable to Italian language,. It is a database built from the voices of up to 6 actors who played 14 sentences simulating 6 emotional states (disgust, fear, anger, joy, surprise, sadness) plus the neutral state. These emotions are the well-known Big Six found in most of the literature related to emotional speech. The recordings were made with professional equipment in the Fondazione Ugo Bordoni laboratories. The paper also describes a subjective validation test of the corpus, based on emotion-discrimination of two sentences carried out by two different groups of 24 listeners. The test was successful because it yielded an overall recognition accuracy of 80%. It is observed that emotions less easy to recognize are joy and disgust, whereas the most easy to detect are anger, sadness and the neutral state.

Keywords: emotional corpus, speech corpus, speech database

1. Introduction

Linguistic information is all that can be carried by text, but it is only a small part of the spoken message. As humans, when listening to speech, we are sensitive to extra-linguistic information about the identity and the state of the speaker, as well as to paralinguistic information about the speaker's intentions underlying the utterance.

An important role in the communication is given by the emotion (Guerrero et al., 1998). For example, a simple text dictation that doesn't reveal any emotion, it does not covey adequately the semantics of the text.

Speech emotion recognition (SER) systems can be used by actors for emotion speech consistency, by disabled people for communication as well as for interactive TV, for constructing virtual teachers, in the study of human brain malfunctions, and the advanced design of Text To Speech systems. To achieve such ambitious goals, the collection of emotional speech databases is a prerequisite.

Studies on emotional speech have verified that there are strong correlates between speech and emotion. Major findings in this area are described in key review articles of (Davitz, 1964), (Scherer, 1986), (Murray and Arnott, 1993), (Cowie et al., 2001) and (Ramakrishnan, 2012). The most commonly studied emotions in related studies are chosen from the "basic emotions" such as happiness, sadness, fear, disgust, anger, and surprise (Murray and Arnott, 1993), (Giovannella et al., 2009, 2012).

Different types of corpora have been used for the study of emotions in speech. Some groups have employed spontaneous emotional speech, trying to get the greatest authenticity in the emotions (Douglas-Cowie et al, 2003), (Cresti, 2005), (Burkhardt et al, 2005). Others have worked with induced emotions, putting the speaker into situations to rouse a specific emotion (Kuroda, 1976). A third option has been to use a speaker with acting skills to simulate emotions. Though this latter technique can exaggerate the emotions, the fact is that they are recognized, so that practical modelling can be derived from them. In this work we selected acted speech corpus because it is the only one that allows a complete control over the recorded text.

2. Emotional speech database

Prominent example of acted DB are the EMO (Berlin emotional speech), the DES (Danish Emotional speech corpus), Polzin in English and Groningen in Dutch. Induced corpora are Amir (Russian), Tolkmitt and Scherer in German Fernandez and al. in English. As concerning the real word corpus Recola in French, Belfast natural Database in English, Smartkom in German. A good list of emotional corpora is available in <http://emotion-research.net/wiki/Databases> . It is based on the Humaine deliverable D5c.

EMOVO (Iadarola, 2009) is the first database of emotional speech for the Italian language. Six actors were summoned, three males and three females with proven expertise, and have made them perform fourteen sentences (assertive, interrogative, lists) based on six basic emotional states (disgust, fear, anger, joy, surprise, sadness) plus the neutral state.

These emotions are the well-known big six (Cowie & Cornelius, 2003) found in much of the literature relating to emotional speech. The recordings were made, with appropriate professional tools, at the laboratories of the Fondazione Ugo Bordoni,

In addition, EMOVO has been validated with a test of discrimination of emotions on two phrases: the house loud wants with the bread and the cat is flowing in pear conducted in parallel and separately. Both tests had a sample of 12 subjects recognizers, which had in turn indicate, choosing between two possible answers, the emotional state of the sentences heard. The test was successful because it was an overall accuracy of the recognition of 80%. We also assessed the scale of recognition of emotions, and there has been a substantial agreement with the scale of recognition inferred from the

literature in this regard.

3. Choice of emotion and actors

The emotional performance expressed in the database concern to the following:

1. disgust
2. joy
3. fear
4. anger
5. surprise
6. sadness
7. neutral

These are the emotions that have received more attention from the Italian scientific community and therefore it was decided to stay within this range of basic emotions rather than analyse others.
To avoid too different interpretations of these emotions, we are given the appropriate explanation for the emotions with more problematic definition. In particular, we explained to each actor that anger to express is identified by Scherer Scherer K. R. (1979) as hot. While, for what concerns to disgust, we pointed out that it meant a physical rather than moral disgust. Regarding the surprise we invited the actors to maintain this emotional state for the entire duration of the sentence, and not to use for this purpose a questioning tone. For the rest of the emotions we have not considered it necessary to provide further explanation. Finally, we asked the actors not to deliver an explicit emotional indicators such as laughter, tears etc., which would distort the recognition test.
We chose 6 professional actors, 3 males and 3 females, the age is from 23 to 30 years old and we care to check their professional background, especially their knowledge of the famous Stanislavsky method, consisting to self induced emotional states recalling the situations in their own life where he/she felt intensely the same emotion. The actors were classified as M1, M2, M3, F1, F2, F3.

4. Linguistic material

This is the list of phrases to elicit:

1. Gli operai si alzano presto.
2. I vigili sono muniti di pistola.
3. La cascata fa molto rumore.
4. L'autunno prossimo Tony partirà per la Spagna nella prima metà di ottobre.
5. Ora prendo la felpa di là ed esco per fare una passeggiata.
6. Un attimo dopo s'è incamminato ... ed è inciampato.
7. Vorrei il numero telefonico del Signor Piatti.
8. La casa forte vuole col pane.
9. La forza trova il passo e l'aglio rosso.
10. Il gatto sta scorrendo nella pera
11. Insalata pastasciutta coscia d'agnello limoncello.
12. Uno quarantatré dieci mille cinquantasette venti.
13. Sabato sera cosa farà?
14. Porti con te quella cosa?

In English:

1. Workers get up early.
2. Firefighters are equipped with a gun.
3. The waterfall makes a lot of noise.
4. The next fall Tony will leave for Spain in the first half of October.
5. Now I take the sweatshirt and go for a walk.
6. A moment later he hath walked ... and stumbled.
7. I would like the telephone number of Mr. Piatti.
8. The strong house wants with bread.
9. The force is up and red garlic.
10. The cat is flowing in pear.
11. Pasta salad leg of lamb limoncello.
12. One forty-three ten thousand fifty-seven twenty.
13. Saturday night, what will?
14. Bring with you that thing?

Regarding the content of the sentences the necessary condition is that the semantic value of this content be emotionally neutral in order to use this sentences also in subjective tests where the text meaning could influence the results. To this end it is possible to use phrases that do not relate specifically to a given emotional expression, and phrases "nonsense". If the former can pose a challenge to the actor to place them in the right emotional state, the latter involve the risk of being recited in a stereotypical manner, as the actor may not be able to "hear" as natural. Aware of these problems, we decided to use both categories (sentences 1-7, 13, 14 semantically neutral, sentences 8-10 nonsense), adding two sentences which are located in an intermediate position with respect

The phrases are semantically neutral and are formed by short versions (1-3) and long (4-7).
For the purpose of spectral analysis have been satisfied, the following basic conditions:

- presence in the sentences of all the phonemes of the Italian language
- presence in every sentence of a fair balance between voiced and unvoiced consonant

5. Recording and storing data

The performances of the actors were recorded in a room of the laboratories of the Fondazione Ugo Bordoni in Rome. Two professional microphones SHURE SM58LC model and a digital recorder Marantz PMD670 model were used. The recordings were performed with a sampling frequency of 48 kHz, 16 bit stereo, wav format.
The actors were able to move freely and this has obviously affected the data on absolute signal intensity, depending on the distance of the mouth from the microphone. In other cases it is due to provide manual lowering of the volume of registration, in order to avoid the saturation of the signal in the cases of emotions that generate high levels of energy, such as anger.

		RECOGNIZED EMOTION						
		NEUTRAL	DISGUST	JOY	FEAR	ANGER	SURPRISE	SADNESS
ELICITED EMOTION	NEUTRAL	93%	1%	0%	0%	4%	0%	2%
	DISGUST	3%	67%	2%	6%	10%	6%	6%
	JOY	2%	4%	65%	7%	7%	10%	4%
	FEAR	2%	7%	2%	74%	3%	3%	9%
	ANGER	1%	1%	1%	3%	92%	1%	1%
	SURPRISE	1%	3%	4%	1%	1%	81%	9%
	SADNESS	2%	2%	1%	3%	0%	0%	92%

Table I: EMOVO validation test.

Each recording session is about an hour. Subsequently recorded phrases were stored in the database in 6 different folders corresponding to the 6 actors. Each file has been labelled as follows:

neu-m1-b1.wav

The first part of the name indicates the emotional state (neutral, disgust, joy, fear, anger, surprise, sadness).
The second part of the name indicates the actor/actress (m1, m2, m3, f1, f2, f3) that reads the sentence.
The third part of the name indicates the type of sentence:

- the sentences 1-3 are ranked respectively b1, b2, b3 (short)
- the sentences 4-7 are ranked respectively l1, l2, l3, l4 (long)
- the sentences 8-12 are ranked respectively n1, n2, n3, n4, n5 (nonsense)
- the sentences 13,14 are ranked respectively d1, d2 (questions).

Finally, all the sentences have been recorded into a database in Microsoft Access, in which we can sort the records according to the emotion, the text of the sentence, the sentence type, and we also can listen the sentence clicking on the proper link.
In total 588 records have been archived: for each of the 6 actors 98 sentences were recorded, corresponding to 14 sentences spoken in 6 emotional states plus the neutral state.
The total number of recorded material, excluding breaks, is approximately 10 minutes per actor, then an hour for the entire database. EMOVO is currently available on a single CD-ROM and will be distributed to a request.

6. Applications of Emovo

A corpus of emotional voice is essential to build systems of emotion recognition and systems of text to speech synthesis from text with emotional voice. The introduction of the appearance of emotion in human-machine communication finds its most important application in man-machine interface and then in call canter, but also in many other areas. Among these we will quote the Lie detector (X13-VSA PRO) a system capable of judging what is he lies, the use in the automotive industry with important implications in the field of safety, the use in video games, in communication robot man, and in many other application fields such as e-learning.

7. Validation test

In order to evaluate the performances of emotional actors, or if the sentences produced contain the information needed to recognize the emotion that had been acted, we have built a subjective test of emotion recognition. After listening to a pre-test consisting of 6 sentences, one for each emotional state, subjects were asked to choose between two possible emotions (ex: anger / joy) then the listeners was subjected to full validation testing. For each actor were chosen 2 nonsense sentences to avoid that the semantic content could influence the choice between 6 for every emotion possible. The subject then had to choose between two possible emotions. A total of 42 items were proposed to subject choices (6 choices for 7 emotional states).
Each subject listened the signals of a male and a female actor for a total of 84 tests. The accuracy was calculated according to the formula:

Accuracy = (ca - wa) / nr

where *ca* are the correct answers, *wa* are the wrong answers and *nr* are the number of responses.

The subjects were organized into two groups, both of 12 elements, a group active in the laboratory that had acquired the corpus in Rome, and another in university laboratory in Cosenza. The results of the two laboratories are statistically compatible and are therefore summarized in a single table (see Table I) where the columns represent the emotion recognized and the lines that elicited. It is observed that the emotions less easily recognized are the joy and the disgust while the most easily recognized are the neutral, the anger and sadness. Validation test has been done to ensure the goodness of emotional performances of the actors, for this purpose we have built a subjective test of emotion recognition.

Each listener heard a couple of sentences with different emotions, and then was asked to choose between two possible emotions such as disgust or joy.

Listeners were organized into two groups, both of 12 elements, a group at the laboratory that had acquired the corpus in Rome, and another in another laboratory in Cosenza. The results of the two laboratories are sufficiently compatible and are therefore summarized in a single table (see Table I), where the columns represent the emotion recognized and rows the one elicited. It is observed that emotions less recognized are the joy and disgust, while the overall accuracy of the recognition is about 80%. It is certainly true that good actor can generate speech that listeners classify reliably.

8. Conclusion

In the study of speech signal the emotional aspect plays an increasing role because only through emotional connotations become artificial voices more natural and because the emotion recognition is an important parameter in many applications. In order to characterize the emotional voices it is essential to have a reference corpus. Thus, the realization of EMOVO has made available to the scientific community the first corpus of emotional voice for Italian.

This paper has presented a detailed description of the corpus and also some significant results of its subjective evaluation. Based on indications obtained in the tests, we believe that EMOVO is suitable for use in studies of emotion. However, a data base which records sentences where emotion has been simulated by actors might be questioned for use in a system for automated recognition of emotions in speech. To override this potential problem, we aim at building an alternate corpus made up of sentences uttered by persons caught in a real emotional state

The EMOVO corpus can be downloaded at:

http://voice.fub.it/EMOVO

9. References

Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W. & Weiss B. (2005), *A database of german emotional speech*. In Interspeech 2005, Lisbona, pp. 1517- 1520,

Cowie, R., Cornelius, R.R. (2003). *Describing the Emotional States that Are Expressed in Speech*. Speech Communication, 40(1,2), 2-32.

Cresti E. & Moneglia M. (2005), *C-ORAL-ROM Integrated reference corpora for spoken romance languages*. Benjamins, Amsterdam-Philadelphia.

Davitz, J.R., 1964. *A review of research concerned with facial and vocal expressions of emotion*. In Davitz, J.R. (Ed.), The Communication of Emotional Meaning. McGraw-Hill, NY, pp. 13–29.

Douglas-Cowie E., Campbell N., Cowie R. & Roach P. (2003), *Emotional speech: towards a new generation of databases*, "Speech communication" 40, pp. 33-60.

Giovannella C., Floris D., Paoloni A., *Transmission of vocal emotion: Do we have to care about the listener? The case of the Italian speech corpus EMOVO*, 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009.

Giovannella C., Floris D., Paoloni A., *An exploration on possible correlations among perception and physical characteristics of of EMOVO's emotional portrayals*. IxD&A Journal, N. 15, 2012, pp. 102-111.

Guerrero, L.K., Andersen, P.A., Trost, M.R., 1998. *Communication and emotion: basic concepts and approaches.* Speech Communication 40 (2003) 161–187.

Iadarola, I. (2009), *EMOVO, database di parlato emotivo per l'italiano*, Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Arcavacata di Rende (CS), 3-5 dicembre 2007.

Kuroda I., Fujiwara O., Okamura N. & Utsuki N. (1976), *Method for determinino pilot stress through analysis of voice communication*. In "Aviation, space, and environmental medicine" 47, pp. 528-533.

Scherer, K.R., 1986. *Vocal affect expression: a review and a model for future research*. Psychological Bulletin 99 (2), 143–165.

Murray, I.R., Arnott, J.L., 1993. *Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion*. Journal of Acoustic Society of America 93 (2), 1097–1108.

Ramakrishnan, 2012. *Recognition of Emotion from Speech: A Review*. In Speech Enhancement, Modeling and Recognition - Algorithms and Applications,

Scherer K. R. (1979), *Nonlinguistic vocal indicators of emotion and psychopathology*. In Izard C. E., Emotion and personality in psychopathology, PlenumPress, NY.