Interaction Design and Architecture(s) Journal - IxD&A, N. 15, 2012, pp. 102-111

An exploration on possible correlations among perception and physical characteristics of EMOVO emotional portrayals.

Carlo Giovannella[1,2] , Daniele Floris[1], Andrea Paoloni[3]

[1] ISIM Garage Dept. of Education Science and Technology and [2] Scuola IaD
University of Rome Tor Vergata
via della ricerca scientifica 1, 00133 Rome, Italy
[2] Fondazione Ugo Bordoni, Rome, Italy
giovannella@scuolaiad.it, info@mifav.uniroma2.it

Abstract. This article presents the first attempt to investigate the existence of possible quantitative correlations among the physical characteristics of emotional portrayals and the emotions perceived by humans during their listening. Our aim was: a) to design and develop a new investigation protocol; b) to obtain information useful to recognition and synthesis of emotions conveyed by the human voice. Our results, obtained on a subset of the emotional portrayals contained in the corpus EMOVO, show that, apart from well known qualitative indications, it is also possible to observe clear quantitative trends for some couples of "emotion-signal characteristics" as function of the recognition rate of the emotion.

Keywords: voice emotional portrayals, perception of emotion, emotion recognition and synthesis, EMOVO

1 Introduction

The expectations arising from the today's impressive technological developments are such that one day, not far by this time, the interaction among humans and technology enhanced environments could develop in a very natural way, using gestures and voice [1]. Unavoidably the naturalness of the interaction, and thus of the experience, will depend on the ability of the designer to consider the characteristics of either the contexts and the individuals, including all levels of the human interaction [2]. Among the latter, extremely relevant is the emotional level [3,4] that on the other hand, is still very difficult to monitor and synthesize, also, but not exclusively, because of the lack of a universally shared model of emotions. At the present, in fact, many models compete among them; just think of the two major categories of emotional models: dimensional models [5-7] (the most known of which is the bi-dimensional one, based on valence and arousal [8]) and finite-state models (e.g. Eckman [9], Plutchik [10], etc., and many others summarized in the review paper by Ortony and Turner [11]), to which we can add the effort to put in relation affective states and dimensional representations [12], and more complex representations such as the Parrot's taxonomy [13], the model of Ortony, Clore and Collins [14], the four-dimensional model of Sherer and coworker [15]. It maybe worthwhile to stress that relatively little is known on the relationships among such models.

At present, moreover, very little is known also on the quantitative modification of the voice induced by emotions and how the perception of emotions is influenced by such variations. It is not by chance, in fact, that the present physical-mathematical models of vocal synthesis, by themselves, are not able to reproduce in a convincing manner emotionally colored human voices. Such difficulty, however, is not specific to the human voice but extends to all areas of the synthesis of the reality: for example, at the present, the best results on the rendering of illuminated spaces [16], on the modeling of the human gestures [17], and on the synthesis of speech signals [18] are obtained when the combination of sophisticated parallel physical-mathematical models is fed by complementary information otained by sampling the reality.

In this context, we think that a useful contribution to improve our understanding may be given from studies aimed at investigating possible quantitative correlations among physical characteristics of the stimuli and results of perceptive tests. As far as we know such studies have not been carried on up to now. Usually, in fact, researchers limit themselves to: i) validate the speakers (not the listeners); ii) to compare recognition levels of man and machine; iii) provide qualitative data on changes induced by the emotion on physical parameters of the voices [26, 4].

The study we describe in the following is intended to contribute to fill the gap and was conducted using the corpus EMOVO created by the Fondazione Ugo Bordoni. It may be worthwhile to stress that EMOVO represents the first attempt to build an Italian corpus of emotionally colored portrayals, although the present study is not, as far as the Italian language is concerned, the first one that has been devoted to the vocal expression of emotions [19]. EMOVO is a database containing 588 records: 14 Italian sentences each colored with 7 different emotional states by 6 professional actors. The emotional colors were chosen making reference to the Ekman model [9] and were: disgust (physical rather than moral), joy, fear, anger ("hot" as defined by Klaus Scherer [20]), surprise (used for the whole duration of the sentence), sadness; for comparison, the neutral color was also added. More details on the corpus can be found in [21]. Here below, after a brief summary of the results obtained in the past [21-23] - i.e. the background of this study - we shall describe the outcomes of the physical characterization of the EMOVO portrayals and the strategy that we have adopted to identify relevant physical quantities that can be correlated with the results of the perceptive tests. A brief discussion on future perspectives of this work will follow.

2 Background

In the recent past, using 42 portrayals of the corpus EMOVO, those based on the nonsense sentence "La casa forte vuole col pane" ("The house wants strong with the bread"), we have developed a test to measure the emotional perception induced by the portrayals, a test that has been experienced with many different categories of subjects (e.g.: a sample of standard Italian population, music composers, children, etc.). The software application specially developed to allow for an easy design of such kind of tests and, as well, to record and analyze data have been described in details in other previous papers [21-23].

Table 1. Emotion-Actor Recognition-Matrix for a representative sample of standard Italian population: the first percentage is calculated by integrating data from all the nuances of a given petal of the Plutchik's flower (see fig.1); the second one is calculated only by summing those points localized in the portion of the petal related to the exact emotion that the actor was required to convey (1/3 of the Plutchick's petal: the central area). Av.Tot: average calculated on 17 reliable listeners; Av*: average calculated subtracting the M3 contribution (because unreliable speaker) [21].

actor	anger	disgust	sadness	surprise	fear	joy	neutral
M1	65%-30%	6%-6%	44%-25%	41%-23%	65%-29%	6%-6%	29%
M2	82%-71%	23%-12%	6%-0%	12%-0%	60%-29%	18%-6%	18%
M3	6%-0%	53%-12%	23%-0%	12%-0%	18%-6%	23%-0%	29%
F1	47%-18%	47%-35%	70%-44%	12%-6%	6%-0%	53%-23%	18%
F2	53%-12%	6%-0%	23%-6%	29%-23%	100%-53%	71%-59%	23%
F3	41%-23%	6%-6%	53%-23%	35%-18%	44%-6%	41%-23%	37%
Av.Tot	49%-25%	23%-12%	36%-16%	23%-12%	48%-21%	35%-20%	26%
Av.*	58%-31%	18%-12%	39%-20%	26%-14%	55%-24%	38%-24%	25%
Av.M*	73%-50%	15%-9%	25%-12%	26%-12%	62%-29%	12%-6%	23%
Av. F	47%-18%	20%-14%	48%-24%	25%-16%	50%-20%	55%-35%	26%

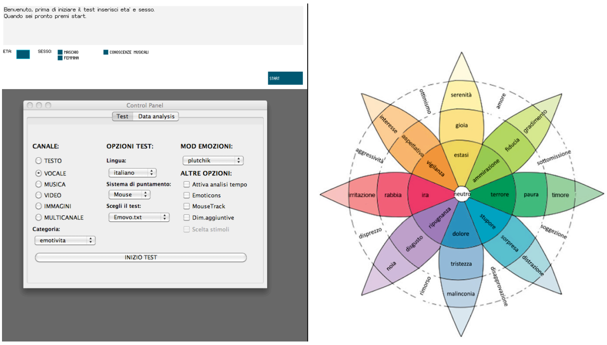
Table 2. Variation of the physical parameters of the voice induced by the emotions conveyed: qualitative trends. Left data from Juslin and Scherer (2005)[26], right our findings. M=male; F=female.

activation	*high*	*low*	*low*	*high*	*high*	*high*
	anger	disgust	sadness	surprise	fear	joy
f0 mean	> | >	< | =	< | =	> | >	> | >	> | >
f0 SD	> | >	> | =	< | =(F<)	> | >	< | ?	> | >
intensity	> | >	> | ?	< | <	> |?(M>)	= | ?	> | >
f1 mean	> | >	> |?(M>)	< | ?	< | ?	< | >	> |? (F>)
art. time	< | >	> | >	< | >	< |?(M<)	< | <	< | >
pause	< | <	> | ?	> | >	< | >	< |?(M<)	< | <
jitter	< | ?	< | ?	= | >	> | ?	> | >	= | ?

Here we limit ourselves to remind that the subjects were required to listen to all 42 portrayals that have been presented in random sequence. For each portrayal the subject was required to indicate, by means of a mouse click, the emotion perceived, among those contained in a graphical representation of a given model of emotions, for example the Plutchik model [10], see fig.1.

Through a careful examination of the data collected it was possible to apply adequate filters to select high quality speakers and, as well, reliable listeners. At the end, we obtained the results reported in Table 1 showing the recognition rates we obtained as function of the speaker and the emotion.

From a first analysis of the physical characteristics of the 42 portrayals we used it has been possible to compile Table 2 showing qualitative information on the modification of the physical parameters induced by each emotion. The agreement with the literature is quite good. It is better for emotions characterized by a high level of activation and worsens in the case of emotions that have low rates of recognition such as the disgust. The discrepancies are almost all on the *articulation time* and on the *duration of pauses*, i.e. quantities that, more than others, are related to the specificity of a given language.



Benvenuto, prima di iniziare il test inserisci eta' e sesso.
Quando sei pronto premi start.

ETA:
SESSO: ■ MASCHIO ■ FEMMINA
■ CONOSCENZE MUSICALI

START

Control Panel

Test | Data analysis

CANALE:
○ TESTO
● VOCALE
○ MUSICA
○ VIDEO
○ IMMAGINI
○ MULTICANALE

Categoria: emotivita

OPZIONI TEST:
Lingua: italiano
Sistema di puntamento: Mouse
Scegli il test: Emovo.txt

MOD EMOZIONI: plutchik

ALTRE OPZIONI:
□ Attiva analisi tempo
□ Emoticons
□ MouseTrack
□ Dim.aggiuntive
□ Scelta stimoli

INIZIO TEST

Fig. 1. Screenshot of the interface of the test module of our application. Right: emotional palette (flower) representing the Plutchik model.

A first attempt to correlate the quantitative variations of the physical characteristics of the stimuli with the perceptual responses enabled us to provide preliminary indications on threshold that such variations should take to ensure a high probability of recognition of the emotions conveyed by the portrayals. The purpose of this article is to deepen such investigation and to work out possible quantitative functional dependencies, if any.

3 Analysis of the EMOVO portrayals

By means of PRAAT [24] we have analyzed the entire corpus EMOVO: 588 portrayals. Having obtained for each portrayal the value of all physical quantities that can be measured using PRAAT, we calculated for such physical quantities the variation induced by the emotions conveyed into the voice, i.e. the difference between the values measured for a given emotional portrayal and the corresponding one measured for the neutral sentence.

For most of the physical quantities we observed a too large variability within the subsets of portrayals played by the same actor. Taking as an example the *fear*, the quantity (*f0max-f0min*) shows, as might be expected, a strong variability with sentences and does not provide useful indications to identify a given actors, see fig. 2 top. The situation changes when one considers the variation of *f0mean* that, despite few inevitably jumps, on the whole, shows a higher stability and, therefore, can be considered a good descriptor of the peculiar style used by a given speaker to convey emotions, see fig. 2 bottom. Physical parameters deserving such quasi-stable trends are good candidates to explore possible correlation among modulation of the vocal signal and perception of emotion induced in the audience.

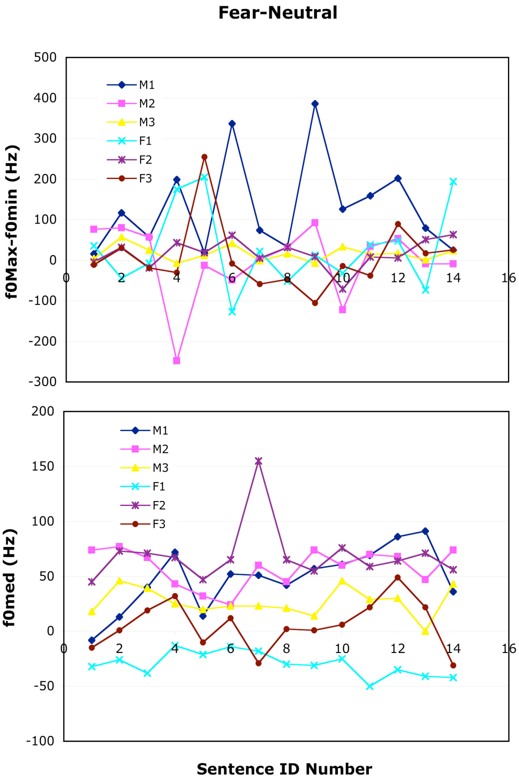


Fig. 2. Fear portrayals: variation with respect to the neutral portrayals of (f0max-fmin) and (f0mean).

Fig. 3 shows that in the case of *fear* the correlation between the variation of *f0mean* and the percentage of recognition of the emotion (that can be taken as a measure of the probability to induce the perception of a given emotion) is quite evident and robust. In fact, it is detectable either if we use the variation of *f0mean* measured for the no-sense sentence used during the tests and, as well, if we use the variation of *f0mean* averaged over all the 14 sentences interpreted by a given actor, fig 3a. The correlation persists, albeit with a different functional dependence, also when on the x-axis we report the percentage of listeners' responses integrated on all the nuances of the petal that includes also the fear, fig. 3b (we remind that in the Plutchik model the fear corresponds to the central portion of the green petal of fig. 1). The observation of this correlation is particularly relevant considering that it was obtained using both male and female voices.

Unfortunately, such clear correlation was not observed for all physical quantities measured by mean of PRAAT. Fig. 4a, for example, shows that, apparently, the jitter does not deserve any correlation with the fear recognition rate, although its increase seems to facilitate the perception/recognition of sadness, see figure 4b.

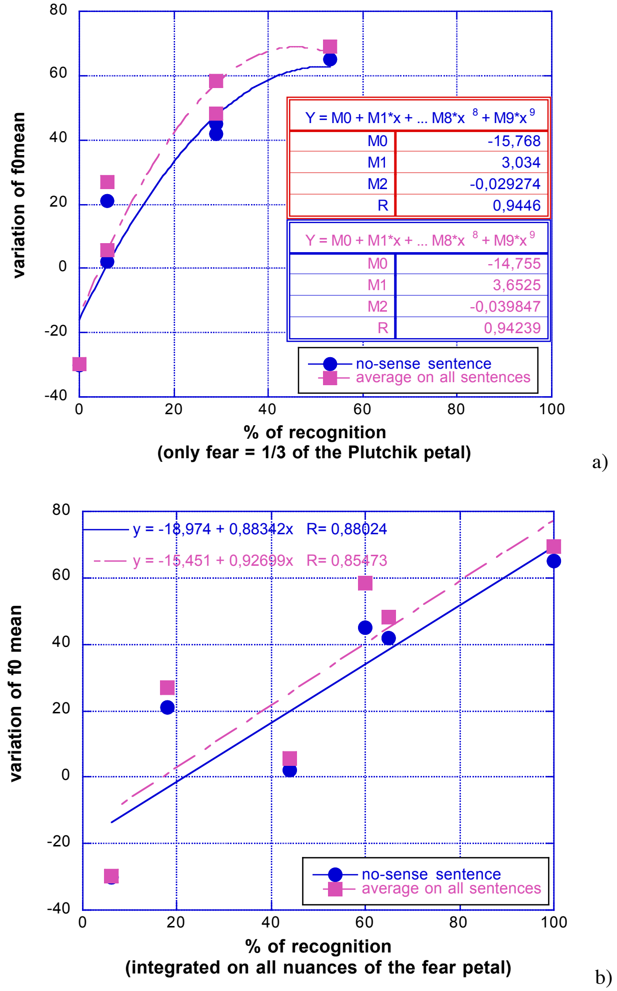


Fig. 3. Fear portrayals: variation of f0mean with respect to the neutral portrayals versus the recognition rate of fear (a) or fear-family emotions (b) of the Plutchik's flower (see fig.1).

Moreover the fact that the variation of a given physical quantity shows a strong correlation with the recognition rate of a given emotion does not assure that it will show similar correlations also for all other emotions taken in consideration. As an example fig. 5 shows the variation of f0mean plotted against the recognition rate for all other five emotions considered here. It is quite clear that the recognition rate of surprise correlates as good as for fear with the increase of f0mean. Also in the case of anger there is a strong tendency toward an increase of f0mean with the recognition rate; a similar trend is observed also for the joy, but in none of these cases the correlation is so clear and robust as in the case of fear. For sadness one could hazard a tendency of f0mean to decrease with the recognition rate. Finally for disgust a very weak tendency to increase have been observed.

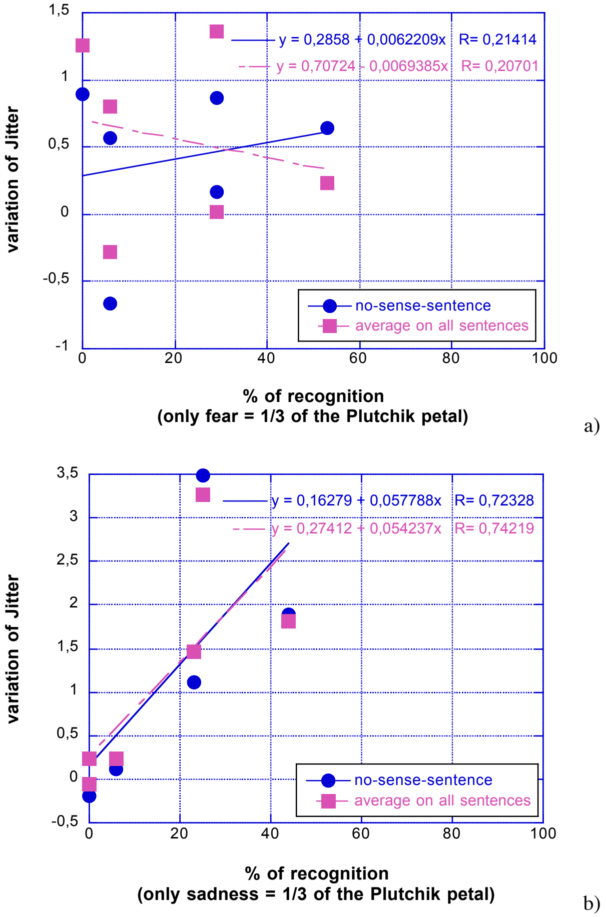


Fig. 4. Fear and Sadness portrayals: same as fig. 3 but for the physical quantity Jitter.

4 Discussion and conclusions

The results discussed in the previous section show fairly clear correlations between the variation of certain physical parameters of the portrayals and the recognition rate of the emotion conveyed by the human voice. It is also very clear that, like for the emotions expressed by the face, also for the voice not all physical characteristics of the signal are used with the same intensity to convey emotions. The main difference is that whereas in the case of the face expressions one can make reference to quite sophisticated models, e.g. the FACS [25] - which allow to catalog all effects generated by the movement of all muscles of the face - in the case of the voice a mapping between physical characteristics of the signals and the emotions they convey seems still far to be available. This depends certainly on the lack of studies on the perception of emotions but also on the scarce information we have on the physical quantities that may play a role in the identification of the emotion conveyed by the voice. These are very unexplored aspects of this still open domain of research.

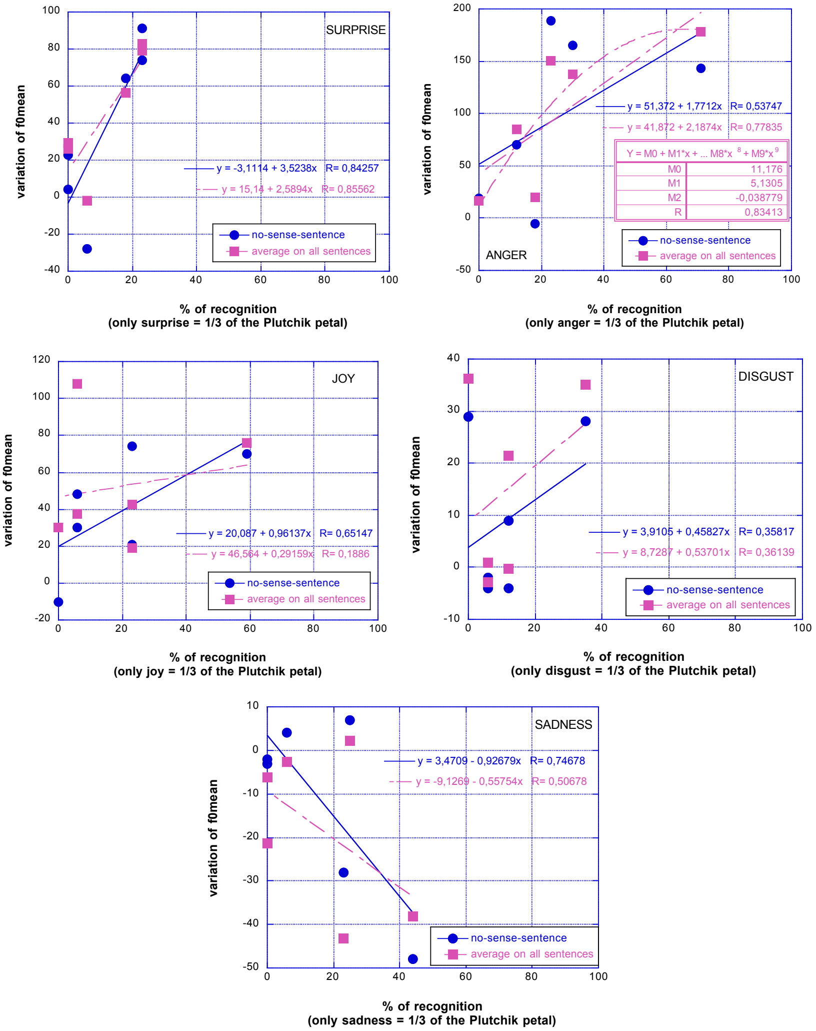


Fig. 5. As for fig. 3a but for the remaining 5 emotion featured by the actors.

As far as the work presented here is concerned, it is quite clear that we need to explore further the perceptual dimension to collect a sufficient amount of data to identify more accurately the laws governing the variation of the physical variables that greatly contribute to the identification of emotions.

Although the completion of such "exploration" needs, unavoidably quite a long time, we would like nevertheless to stress that with the present work we have already achieved very relevant results: the definition of a completely new investigation procedure/protocol whose usefulness and effectiveness has been demonstrated by the results reported in par. 3 and that, as far as we know, are the first ones of this kind ever obtained up to now. On the basis of such preliminary results, it is not difficult to imagine, moreover, that, once that the "exploration" will be completed and the goals achieved, the outcomes could be used to get a more natural voice synthesis and recognition.

References

1. Dourish P.: 2004 Where the action is. The MIT Press, Cambridge, MA (2004)
2. Giovannella C., Moggio F.: Toward a general model of the learning experience. In ICALT 2011, IEEE publisher, pp.644--645 (2011)
3. Picard R.W.: Affective Computing. The MIT Press, Cambridge, MA (1997)
4. Calvo R.A., D'Mello S.: Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Application, IEEE Trns. on Affective Computing, 1, 18--37 (2010)
5. Wundt W.: Outlines of psychology. G. E. Stechert Leipzig, New York (1897)
6. Schlosberg H.: Three Dimensions of Emotion. Psychological Review vol. 61, pp. 81--88 (1954)
7. Osgood C., Suci G., Tannenbaum P.H.: The Measurement of Meaning. University of Illinois Press, Urbana (1967)
8. Russell J. A., Feldman-Barrett L.: Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant. Journal of Personality and Social Psychology, vol. 76, pp. 805--819 (1999)
9. Ekman P.: An Argument for Basic Emotions. Cognition and Emotion, 6, 169--200 (1992)
10. Plutchik R.: Emotion: A Psychoevolutionary Synthesis. Harper & Row, New York, (1980)
11. Ortony A., Turner T.: What's Basic about Basic Emotions. Psychological Rev., vol. 97, pp. 315--331 (1990)
12. Russell J.A.: Core Affect and the Psychological Construction of Emotion. Psychological Rev., vol. 110, pp. 145-172 (2003)
13. Parrott W.: Emotions in Social Psychology. Psychology Press, Philadelphia (2001)
14. Ortony A., Clore G.L., Collins A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1998)
15. Fontaine J., Scherer K., Roesch E., Ellsworth P.: The World of Emotions Is Not Two-Dimensional. Psychological Science vol.18, pp. 1050--1057 (2007)
16. Debevec P.: Virtual Cinematography: Relighting through Computation. Computer, 39, 57-65 (2006)

17. Menache A.: Understanding Motion Capture for Computer Animation. Morgan Kaufmann (2010)
18. De Mori R.: Spoken Dialogues with Computers: Signal Processing and Its Applications. Academic Press (1998)
19. Anolli L., Ciceri R.: La voce delle emozioni. Verso una semiosi della comunicazione vocale non verbale delle emozioni. Franco Angeli Press (1997)
20. Banse R., Scherer K.R.: Acustic Profiles in Vocal Emotion Expression, J. of Person. and Soc. Psych. 70, 614--636 (1996)
21. Giovannella C., Santoboni R., Conflitti D., Paoloni A.: Transmission of vocal emotion: do we have to care about the listener?. in ACII09, IEEE publisher, 494--499 (2009)
22. Giovannella C., Carcone S.: A new application to detect 'emotional perception and styles' of children, and their evolution with age. In ICALT 2011, IEEE publisher, pp. 53--55 (2011)
23. Giovannella C., Carcone S.: An application to test the emotion conveyed by vocal and musical signals, in Interspeech, ISCA, 3313--3314 (2011)
24. Praat. http://www.fon.hum.uva.nl/praat/. Accessed 20 December 2012
25. Ekman, P., Friesen, W. V.: The Facial Action Coding System: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto (1978)
26. Juslin P.N., Scherer K.R.: The New Handbook of Methods in Nonverbal Behavior Research. Oxford University Press, Oxford, UK, 65--135 (2005)