Cogn Process (2012) 13 (Suppl 2):S541–S550
DOI 10.1007/s10339-012-0516-2

REVIEW

On the recognition of emotional vocal expressions: motivations for a holistic approach

Anna Esposito · Antonietta M. Esposito

Received: 7 May 2012 / Accepted: 11 July 2012 / Published online: 8 August 2012

Abstract Human beings seem to be able to recognize emotions from speech very well and information communication technology aims to implement machines and agents that can do the same. However, to be able to automatically recognize affective states from speech signals, it is necessary to solve two main technological problems. The former concerns the identification of effective and efficient processing algorithms capable of capturing emotional acoustic features from speech sentences. The latter focuses on finding computational models able to classify, with an approximation as good as human listeners, a given set of emotional states. This paper will survey these topics and provide some insights for a holistic approach to the automatic analysis, recognition and synthesis of affective states.

Keywords Emotional vocal expressions · Processing algorithms · Computational models

This article is part of the Supplement Issue on 'Social Signals. From Theory to Applications', guest-edited by Isabella Poggi, Francesca D'Errico, and Alessandro Vinciarelli.

A. Esposito
Department of Psychology, Second University of Naples,
Caserta, Italy

A. Esposito (✉)
International Institute for Advanced Scientific Studies (IIASS),
Vietri sul Mare, Italy
e-mail: iiass.annaesp@tin.it

A. M. Esposito
Sezione di Napoli Osservatorio Vesuviano,
Istituto Nazionale di Geofisica e Vulcanologia, Naples, Italy
e-mail: aesposito@ov.ingv.it

Introduction

In a daily body-to-body interaction, emotional expressions play a vital role in creating social linkages, producing cultural exchanges, influencing relationships and communicating experiences. Emotional information is transmitted and perceived simultaneously through verbal (the semantic content of a message) and non-verbal (facial expressions, vocal expressions, gestures, paralinguistic information) communicative tools and contacts and interactions are highly affected by the way this information is communicated/perceived to/from the addresser/addressee. Research devoted to the understanding of the relationship between verbal and non-verbal communication modes, and to investigate the perceptual and cognitive processes involved in the recognition/perception of emotional states (as well as their mathematical modeling and algorithmic implementation) is particularly relevant in the field of Human–Human and Human–Computer Interaction both for building up and strengthening human relationships and developing friendly and emotionally colored assistive technologies.

In the present paper, considerations are set on emotional vocal expressions and their automatic synthesis and recognition. Demand for and delivery to date of sophisticated and functional computational instruments able to recognize, process and store these relevant emotional cues, as well as to interact with people, displaying reactions that show abilities of appropriately sensing and understanding emotional vocal changes (under conditions of limited time or other resource) and producing suitable, autonomous and adaptable responses to the various emotional displays has produced great expectations in the information communication technology (ICT) domain. It is unmistakable that the same utterance may be employed for teasing, challenging, stressing, supporting, or as expressing an authentic doubt.

The appropriate continuance of the interaction depends on detecting the addresser's mood. A machine interface unable to comprehend the affective differences will have difficulty in managing the interaction. Progress toward the understanding and modeling of such interactional facets is crucial for implementing a friendly human–computer interaction that exploits synthetic agents and sophisticated human-like interfaces and will simplify user access to future and profitable remote social services. The application of such techniques could be very useful, for example, in monitoring psycho-physical conditions of subjects engaged in high responsibility tasks, researching new means for socio-behavioral investigations, clinical studies, media retrieval, call centers and remote applications where information about the caller's emotional state might provide data about her/his contentment and/or her/his health state (Jones and Deeming 2008; Petrushin 1999).[1]

Nowadays, we do have emotionally colored systems, but far from human ability. The achievement of a human level machine for emotional behavior (and in general of human level automaton intelligence) raises the need for more accurate solutions to the following challenges:

(a) Identify a set of processing algorithms able to capture emotional invariant features from multimodal social signals and in particular from speech;

(b) Implement simple and fast computational models trained to classify, as well as humans, emotional acoustic features for the maintenance of sentences hierarchically structured, time dependent and reciprocally connected through complex relations, such as a set of multifaceted emotional feelings.

Another problem to be dealt with when researching into affective vocal expressions is the lack of adequate recordings of genuine emotions. Indeed, most of the studies take advantage of the work of actors (not always professional) who are required to portray collections of phrasal groups with specifically required emotional intonations. Since it is not obvious whether actors reproduce a genuine emotion or generate a stylized idealization of it, it is questionable whether their emotional vocal expressions authentically represent the characteristics of speech used by ordinary people when they spontaneously experience similar affective states.

The commonly applied approach in creating automatic emotional speech analysis systems is to start with a database of emotional speech that has been annotated with emotional tags by a panel of listeners (generally a limited number of expert judges or a group of naïve ones). The next step is to perform an acoustic analysis of these data and correlate statistics of certain acoustic features, mainly related to fundamental frequency, with the emotion tags. In the last step, the obtained parameters are verified and adapted by assessing the system performance through human interaction. Most approaches focus on six basic emotions—happiness, sadness, fear, anger, surprise, and disgust—supposed to be universally shared since reliably associated with basic survival problems such as nurturing offspring, earning food, competing for resource, avoiding and/or facing dangers (Ekman 1992; Izard 1992; Plutchik 1993). Few attempts have been made to cover a wider range of emotions.

It is worth to mention that the discrete categorization of emotions as reported above is just one of the many and varied theories and models developed over the years that attempt to explain emotions either from a holistic point of view or as atomic components of individuals' emotional experience (Oatley and Jenkins 2006). The discrete theory is widely used since it best suits the needs of an information processing approach that can produce immediate market applications. More sophisticated approaches, as in the affective computing (AC) field (Picard 2000), are dedicated to specific facets of emotion synthesis motivated by the attempt to develop emotionally capable artificial intelligences able to emulate human abilities such as flexibility, decision making, creativity and reasoning, exploiting limited memory and bounded information. To date, an emotionally complete computer architecture is yet to be developed even though the AC literature has provided several more or less sophisticated attempts (Blumberg et al. 1996; de Byl and Toleman 2005; El-Nasr 1998; Kaehms 1999; Penrose 1989; Sloman 2001; Velasquez 1999).

The encoding issue

Automatic recognition of emotion from speech (as well as automatic speech recognition) has been revealed to be a computationally hard problem due to the fact that emotional voices appear to be affected at various degrees by many sources of variability that cause distortions and modifications in the original signal, thus modifying the acoustic features useful for its recognition. Such sources of variability are coarsely clustered into 4 groups: phonetic variability (i.e., the acoustic realizations of phonemes are highly dependent on the context in which they appear), within-speaker variability (as a result of changes in the speaker's physical and emotional state, speaking rate, voice quality), across-speaker variability (due to differences in the socio-linguistic background, gender, dialect, size and shape of the vocal tract), and acoustic variability (as a result of changes in the environment as well as the position and characteristics of the transducer).

[1] Sony AIBO Europe, Sony entertainment. www.sonydigital-link.com/AIBO/.

In order to overcome the limitations in the system performances elicited by the above sources it becomes necessary to know, at any stage of the recognition process, which would be the most appropriate encoding and computational approach.

There is no doubt that emotions produce changes in respiration, phonation and articulation, which in turn affect vocalizations and the acoustic parameters of the corresponding signal (Bachorowski 1999; Banse and Scherer 1996; Friend 2000; Scherer et al. 2001). Moreover, the acoustic realization of specific emotions is to a large extent speaker dependent.

Acoustic features of emotional speech are derived from perceptual cues of loudness, pitch and timing, which in turn are expressed in changes of acoustic variables such as amplitude (which quantifies sound pressures), sound intensity (which is representative of sound energy at different frequencies), signal fundamental frequency F0 (in Hz), some F0-derived measures, speaking rate, utterance and syllable lengths and distribution of empty and filled pauses in the utterance.

The values of the above acoustic features are taken over long-lasting speech utterances (supra-segmental), since it is expected that emotional feelings are more durable than single phonetic segments. The most common acoustic attributes supposed to encode information useful to detect emotions in vocal expressions are: F0 contour, F0 maximum and minimum excursion, F0 jitter (i.e., random fluctuations on F0 values), spectral tilting, vocal envelope—defined as the time interval for a signal to reach the maximum amplitude and decay to zero amplitude—long term average spectrum (LTAS), energy values in different frequency bands, inverse filtering measurements (Banse and Scherer 1996; Breitenstein et al. 2001; Hozjan and Kacic 2006; Hozjan and Kacic 2003; Klasmeyer and Sendlmeier 1995; Navas et al. 2006; Nushikyan 1995). In addition, some authors also propose Mel Frequency Cepstral Coefficients (MFCC) (Hu et al. 2007), perceptual critical band features (PCBF) (Esposito and Aversano 2005), Mel Bank Filtering (MELB) (Busso et al. 2007; Esposito and Aversano 2005), erceptual linear predictive coefficients (PLP) (Hermansky 1990) as well as other comparable encodings (see El Ayadi et al. 2011; Fragopanagos and Taylor 2005 for a review) together with their first (Δ) and second derivatives (ΔΔ).

These measurements are generally considered by the current literature as the acoustic correlates of the small set of discrete emotional states referred to as basic emotions (Russell 1980; Scherer 1989, 2003). Yet, so far, there is little systematic knowledge about the details of the decoding process, that is, the precise acoustic cues the listeners use for inferring the speaker's emotional state. It is evident from the above that the acoustic attributes which seem to play a role in signaling emotions, are the same acoustic attributes which are modified by the phonetic context, the inter- and intra-speaker variability, as well as the environmental noise. Therefore, the quest in the search for reliable algorithms able to encode emotional speech features is strictly related to the quest in searching invariant features for speech recognition.

In the following section, we will report some results on our experience using two different databases and different encoding algorithms.

The first experiment was based on a database of 504 utterances of infant-directed speech (BabyEars). The recordings were made at the Interval Research Corporation (IRC), California, US, by Slaney and McRoberts (Slaney and McRoberts 2003) and consisted of sentences spoken by 12 parents (six males and six females) talking to their infants (from 10 to 18 months old). The sentences were divided into three emotional categories, *approval* (212 sentences), *attention* (149 sentences) and *prohibition* (148 sentences).

It can be objected that infant-directed speech cannot be properly included in the set of emotional vocal expressions since it has generally a social and educative intent. However, parents are really pleased, or worried, or concerned when producing their sentences. Therefore, it seems appropriate to relate their approving, prohibitive and attentional affective acoustic productions to the emotional categories of happiness (for the infant is doing something good or new showing learning and progress), fear (for the infant is putting herself/himself in a dangerous situation) and distress (for the infant is attention demanding) respectively. In addition, these types of affective vocalizations are of great interest both from a psychological and an information communication technology (ICT) point of view, since their prosodic contours seem to be universally recognized (Bryant and Barrett 2007) and can be used to facilitate a robot teaching process in realistic affective computing applications (Breazeal and Aryananda 2002).

The speech signal was processed using two different algorithms: the perceptual linear predictive (PLP) coding (Hermansky and Morgan 1994), and the well known linear predictive coding (LPC) (Makhoul 1975). The computational model employed for the sentence classification was a simple recurrent neural (SRNN) network (Elman 1991). The SRNN architecture consisted of 50 hidden nodes (and respectively 50 context units) and 3 output nodes. The training set included 242 examples and the validation and testing set, comprised each 132 examples. The classification results with the SRNN gave a high percentage of correct classification on the training set (100 % of correct classification), but the performance was poor on the validation and testing set. In particular, the total percentage of correct classification was 59 and 62 % using LPC and PLP features

Table 1 SRNN percentage of correct classification on the testing set using the LPC encoding schema

%	Approval	Attention	Prohibition
Approval	62	20	18
Attention	18	60	22
Prohibition	19	34	47

Table 2 SRNN percentage of correct classification on the testing set using the PLP encoding schema

%	Approval	Attention	Prohibition
Approval	59	21	20
Attention	22	63	15
Prohibition	14	17	69

respectively (details are reported in Tables 1 and 2), with PLP features providing a slightly better performance.

A second attempt was made using the data collected at the School of Psychology Queen's University Belfast, headed by Cowie and Douglas-Cowie (see www.image.ece.ntua.gr/physta for more details), in the context of the European project Principled Hybrid Systems: Theory and Applications (PHYSTA). The database consisted of video clip extracts from television programs where subjects were invited to speak about their own life and interact with an interlocutor in a way that was considered to be essentially genuine. Associated with each video clip there was also an audio file containing only the speech of the main speaker (the interlocutor's voice was removed) and a file describing the emotional state that three observers (expert judges) were attributing to the subject using an automatic system called Feeltrace (Douglas-Cowie et al. 2000). The data were produced by 100 subjects, 77 females and 23 males, each showing at least two emotional states, one always labeled as *neutral*, and one or more marked emotions among the 16 under examination. From these data, after a qualitative and quantitative evaluation, 122 audio files were selected, containing several utterances of different lengths, associated with 4 emotional states: neutral (N), angry (A), happy (H), and sad (S)—details in (Esposito 2002).

These waves were encoded as PLP, RASTA-PLP (RelAtive SpecTrAl), and PCBF (perceptual critical band feature) coefficients (Aversano et al. 2001; Hermansky and Morgan 1994; Hermansky 1990) and classified using a time delay recurrent neural network (TDRNN) with 10 input units, 50 hidden units and 4 output units (Ström 1997). The net performance on the test sets is reported in Table 3 for each encoding procedure.

The results in Table 3 revealed that the PCBF encoding schema better captured the emotional content of the data. However, the net performance was far from being acceptable and this was mostly due to the low signal-to-noise ratio of the recorded waveforms (recordings were made during talk shows). However, also the naturalness of the emotional sentences may have played a role since the waveforms being produced by ordinary people when they spontaneously experience emotions, may contain speech characteristics not accounted in the encoding process.

Given the difficulty to decipher which encoding schema is appropriate for a given emotional database, the most recent approaches compute a high number of acoustic attributes using different encodings and then apply a feature selection algorithm in order to reduce the dimensions of the final input vectors as well as to select the most appropriate features for a given classification task. For example, in (Atassi and Esposito 2008) it was shown that the best mean classification rate of 63 % (obtained for the anger, fear, happiness, boredom, sadness, disgust and neutral emotions) on the Berlin Database of Emotional Speech (BDES) (Burkhardt et al. 2005) was accomplished using a Gaussian mixture model (GMM) as classifier and feature vectors obtained combining [through the sequential floating forward selection (SFFS) algorithm (Pudil et al. 1994)] PLP, ΔPLP, PCBF, ΔΔMELB (Mel Bank) coefficients.

Interestingly, using a feature vector of only PLP, PCBF, MELB, and MFCC coefficients or combining them with their first and second derivatives, the performance was 10 % lower, if not worse, than 63 %, suggesting that relevant emotional acoustic attributes are only partially encoded by each processing schema.

Table 3 TDRNN percentage of correct classification on the test set, using PLP, RASTA, and PCBF encoding schemes

	PLP coding + energy + F0				RASTA coding + energy				PCBF coding			
	N	A	H	S	N	A	H	S	N	A	H	S
N	**54.8**	14.4	11.2	19.6	**15.1**	21.1	40.7	23.1	**39.6**	18.9	23	18.5
A	31.4	**44.2**	10.8	13.6	8.8	**32.2**	32.9	26.1	18.5	**51.7**	17.3	12.4
H	51.2	6.7	**25.6**	16.6	11.8	19.5	**48.8**	19.9	23.3	12.4	**50.1**	14.2
S	31.7	19.6	9.6	**39.2**	12.8	25.8	34.5	**26.9**	13.6	16	19	**51.4**

Bold values indicate the percentage of correct identification of each emotion category, neutral (N), angry (A), happy (H), and sad (S), while the other values in each row indicate the percentage of samples confused with the other emotion categories

The Berlin database of emotional speech (BDES) (Burkhardt et al. 2005) like most of the existing emotional audio databases, consists of audio stimuli produced by professional actors and recorded under studio conditions. Such stimuli, being acted, are unlikely to possess either naturalistic and/or genuine emotional characteristics. To this aim the above combined encoding schema was also tested on the Italian COST 2102 database of emotional sentences (http://cost2102.cs.stir.ac.uk) that has a certain degree of spontaneity since the actors/actresses producing the sentences were acting according to a movie script and their acted emotional states were judged as appropriate to the required emotional context by the movie director (supposed to be an expert). In addition, as audio records extracted from movies, the emotional sentences are either noisy or slightly degraded by interruptions, defining a more realistic environmental context (Atassi et al. 2010; Esposito and Riviello 2010; Esposito et al. 2009a). The classification accuracy for this setup was extremely poor with a mean recognition rate of 40 % accuracy on the six basic emotions of happiness, disgust, fear, sadness, surprise, and anger.

In general, there is no agreement on which encoding schema better encodes emotional features and the same encoding schemes may give different performance on different databases (see El Ayadi et al. 2011; Fragopanagos and Taylor 2005 for a survey). It will be shown in the following, that this problem also affects the automatic recognition of emotions and produces difficulties in identifying a powerful computational model for this task.

The computational issue

The possibility of extracting features from speech which can be used for the detection of the speaker's emotional states finds large interest in the automatic speech recognition and speech synthesis framework and to a large extent in the field of human–machine interaction. Since the exploited computational models are based on informational contents extracted from processed data, special care must be applied both to the collected data and the encoded feature vectors utilized for training such computational devices. This is not to say that the computational model does not play a role on the final system achievement. In the following it will be shown that different models produce different performance on the same processed data. Nevertheless, it is fair to assume that an appropriate encoding combined with an appropriate computational model would provide results that encompass solutions where only one of these two facets has been accounted for. For example, it has been shown in Tables 1 and 2 that the performance of an SRNN on the BabyEars emotional database depends on the encoding schema (with the PLP more appropriate than the LPC schema). Using as computational model a time delay recurrent neural network (TDRNN) and a PLP coding (Apolloni et al. 2000; Esposito 2000; Ström 1997) the mean correct classification rate on this same emotional database, was 95 % on the training set, 75 and 77 % on the testing and validation sets respectively (see Tables 4 and 5 where the TDRNN confusion matrices obtained on the test set with both the LPC and PLP processing are reported).

Table 4 TDRNN percentage of correct classification on the test set using the LPC encoding schema

%	Approval	Attention	Prohibition
Approval	64	18	18
Attention	25	55	20
Prohibition	12	22	66

Table 5 TDRNN percentage of correct classification on the test set using the PLP encoding schema

%	Approval	Attention	Prohibition
Approval	76	17	7
Attention	13	74	13
Prohibition	13	15	72

It is worth to note that both the simple recurrent neural (SRNN) network (Elman 1991) and the time delay recurrent neural (TDRNN) network (Ström 1997) are computational models able to capture temporal and spatial information in the data, which are important information in speech, mostly when emotional features should be evaluated. However, the SRNN with static inputs was not able to follow the temporal evolution of an emotional sentence. Therefore, even though both the models learned very well from the training examples, the SRNN did not generalize, and its performance on the validation and testing sets were poorer than those obtained on the training set. The TDRNN model, instead, adopted a more general internal representation of the data resulting in a performance on the testing and validation sets as good as that reached by non-native human listeners asked to give an emotional label to the same audio waves (Apolloni et al. 2000). Nevertheless, the same model did not generalize at all on the PHYSTA database, independently from the coding schema, as it is shown in Table 3.

These contradictory results are not a flaw of such particular computational models. For example, in Atassi and Esposito (Atassi and Esposito 2008) a much more complex procedure (known in literature as multiple classifier systems) was settled up in order to overcome the processing and computational drawbacks discussed above for the automatic recognition of emotions. The idea was to split the recognition task into two steps. The first step used as encoding schema a combination of PLP, ΔPLP, PCBF and ΔΔMELB features that were given as input vector to a GMM classifier for classifying six basic emotional states and identifying among them the couple with the highest likelihood. The GMM parameters were initialized through the K-means procedure and estimated using the expectation maximization (EM) algorithm (Duda et al. 2003). The second steps re-encoded, through a set of prosodic and voice quality measures, the two emotional states that obtained the highest likelihood scores in the first step, and again used a GMM classifier for selecting the winner. The obtained average classification rate (80.7 %) was an improvement to that (74.5 %) reported by Lugger andYang (2007) on the same emotional database (BDES) (Burkhardt et al. 2005). However, this same approach was unsuccessful on the COST 2102 Italian database (Atassi et al. 2010) for which it was necessary to use a hybrid classification model in order to improve the mean classification rate from 40 % to 61 %. Notice that all the data discussed above refer to a speaker-independent approach to the recognition of emotional vocal expressions. Speaker-dependent recognition methods always give, under the same conditions, better classification results.

In addition to artificial neural networks (Apolloni et al. 2004; Apolloni et al. 2000), multiple classifier systems (Atassi et al. 2010; Atassi and Esposito 2008; Lugger and Yang 2007) and Gaussian mixture model (GMM) (Slaney and McRoberts 2003), several other computational models have been proposed for automatic recognition of emotional vocal expressions such as k-NN classifiers (Schuller et al. 2004), fuzzy classifiers (Razak et al. 2005), decision trees (Pierre-Yves 2003), linear discriminant analysis (Fragopanagos and Taylor 2005), support vector machine (Schuller et al. 2004), hidden Markov model (Nwe et al. 2003). A detailed survey on both the processing and computational issues can be found in (El Ayadi et al. 2011; Fragopanagos and Taylor 2005).

It is worth to note that independently from their computational complexity all the proposed models showed the same drawbacks, that is, their performance appeared to be strongly dependent on the database and the data feature representation, suggesting that for the achievement of a human level machine emotional behavior (and in general of human level automaton intelligence) there is a need for a strong shift into the computational approaches applied up to now.

Discussion

In the past decade, there has been a big effort in implementing automatic systems that can be used in most environments and are able to decrease human work and therefore, human errors. Most of these systems are devoted to applications, such as emotion recognition and synthesis, where the computational approach cannot be modeled through a deterministic Turing machine since, the computational complexity required to perform one or more of the necessary algorithmic steps is prohibitive. These are *NP-complete* and *NP-hard* problems in computer science, where NP indicates that the automatic procedure has a *Non-Polynomial* solution either in terms of computational time or in terms of memory occupancy, or both. To allow the computational tractability of these problems, some powerful and new research fields such as artificial intelligence and cognitive systems have been developed with the aim to propose computational and mathematical models, such as neural networks and expert systems, able to infer and gain the required knowledge from a set of contextual examples or rules. However, both the approaches showed drawbacks and limitations: The neural paradigm was unable to justify and explain the solutions obtained by the proposed models, whereas the artificial intelligence paradigm was unable to define the exact rules for describing algorithmically the required knowledge that the proposed expert systems must exhibit.

To overcome these difficulties, it was suggested to combine the two paradigms and infer an understandable solution to these problems directly from the data identifying features from them that uniquely describe some of their properties. However, due to several sources of variability affecting the data, the search for such invariant features was impracticable. Thus, the computational standstill moved from the identification of an appropriate computational model to that of an appropriate data representation. In addition, the general idea driving this search was that there are rules (or features) governing speech variability and that such rules can be learned and applied in practical situations. This point of view is not generally accepted by all experts, since it is related to the classical problem of reconciling the physical and linguistic description of speech, that is, the invariance issue (see Lindblom 1990).

The importance of the real data from which computational models must extract knowledge is highly stressed in the field of automatic recognition and synthesis of affective states. It is clear that special care must be put in collecting the data for training such intelligent devices, since the classification accuracy of both human subjects and speech emotion recognizers strongly depend on the data. However, in order to move forward from the current impasse, it is important that the machines to be developed should be equipped with a human level automaton intelligence, where dynamical and contextual issues are also considered. This will require some considerations on the problem at the hand.

Emotions, as well as any facet of human interaction, are not entities arriving vacuum-packed and amenable to study in pristine isolation. They are communicated through a gestalt of actions which involve much more than the speech production system. Facial expressions, head, body and arm movements (grouped under the name of non-verbal signals) all potentially provide emotional information, transmitting (through different channels) the speaker's psychological state, attitude, and feeling. There is a link of converging and interweaving cognitive processes that cannot be totally untangled. The definition and comprehension of this link can be understood only by identifying some *macro-entities* involving mental processes more complex than those devoted to the simple peripheral pre-processing of the received signals. To understand how humans exploit information which, even arriving from several channels, all potentially contributes to the semantic and pragmatic meaning of the interaction, it is necessary to gather multimodal dynamic data (comprising verbal and non-verbal) and analyze them across a spectrum of disciplines. This information can be fundamental for depicting the structure of such *macro-entities* and may enable the development of new mathematical models that favor the implementation of intelligent emotional interfaces.

The concept of macro-entity promotes a comprehensive view of the verbal/non-verbal packaging that is critical for disambiguating among them. As an example, let us reconsider the idea that the same utterance may be employed for teasing, challenging, stressing, supporting, or as expressing an authentic doubt. In this case, a challenge will be likely accompanied by emblematic hand and shoulder motions and by head and eye movements dynamically distinct and with a different temporal alignment with the speech produced for teasing or expressing a genuine question. The recognition of an emotional state can be then captured through the appropriate phasing (temporal order) of these verbal and non-verbal signals. This phasing is critical for identifying the macro-entity that assembles the real meaning of the conveyed feeling and will solve possible ambiguities.

Other crucial aspects that have not been investigated, during emotional speech, are some sets of non-lexical expressions carrying specific emotional and communicative values such as turn-taking and feedback mechanisms regulating the interaction, or empty and filled pauses and other hesitation phenomena, vocalizations and nasalizations signaling positive or negative reactions, and the so called "speech repairs" which convey information on the speaker's cognitive state and the planning and re-planning strategies she/he is typically using in a discourse (Butterworth and Beattie 1978; Chafe 1987; Esposito 2008; Esposito and Marinaro 2007). These phenomena have never been accounted for in synthesizing or recognizing emotions.

In addition, recent results in social psychology have shown that social information processing involves embodiment, intended here as the mutual influence of the physical environment and the human activities that unfold within it. The underlying idea is that embodiment emerges from the interaction between our sensory-motor systems and the inhabited environment (that includes people as well as objects) and dynamically affects/enhances our reactions/actions, or our social perception. Several experimental data seem to support this idea. For example, Schubert (2004) showed that the act of making a fist influenced both men's and women's automatic processing of words related to the concept of strength. Similar effects in different contexts have been described by several authors (see Bargh et al. 1996; Stepper and Strack 1993) suggesting that the body and the context rule the individual's social conduct as a practical ability to render the world sensible and interpretable in the course of everyday activities. Context interaction, therefore—the organizational, cultural, and physical context—plays a critical role in shaping social conduct providing a means to interpret and understand individuals' choices, perception, actions and emotions. Previous cognitive theories had not accounted for such findings. The popular metaphor about the mind is that cognitive processes (such as inference, categorization and memory) are independent from their physical instantiations. As a consequence, mental operations are based on amodal representations performed by a central processing unit that exploits the sensory (input) and motor (output) subsystems for collecting and identifying representations of the external world and execute commands respectively (Block 1995; Dennett 1969; Fodor 1983; Newell and Simon 1972; Pylyshyn 1984). Only recently, new cognitive models have been proposed, which account for embodied knowledge acquisition and embodied knowledge use (Barsalou et al. 2003; Smit and Semin 2004). In order to bring further support to these theories, it is necessary to set-up a series of perceptual experiments that show how perception and action are affected by the communicative context (internal and external) and how a successful interaction is a function of the user's correct interpretation of the contextual communicative instance.

As a further step into the investigation of multimodal aspects of emotions, multisensory integration of auditory and visual stimuli must be investigated. It has been proved that the human brain has the ability to merge information from different sensory systems thus offering a more accurate and faster ability to operate in response to environmental stimuli (Frens et al. 1995; Hughes et al. 1994). This ability to integrate different signals in a unique percept is especially appreciated in noisy environments with corrupted and degraded signals (Benoit et al. 1994; Perrott et al. 1991). Research in neuroscience had proved that audiovisual, visual-tactile and audio-somatic sensory inputs are constantly synchronized and combined into a reasoned percept (Callan et al. 2003; Macaluso et al. 2004; Schulz et al. 2003; Stein et al. 2001). In speech, the effects of vision influence on the auditory perception is proved by the well known McGurk effect at the phoneme level (McGurk and MacDonald 1976) as well as by recent results on emotional labeling of combined audio and video stimuli (Esposito and Riviello 2011; Esposito et al. 2009b; Esposito 2007, 2009).

Multisensory integration of audio and visual emotional stimuli must be investigated to take into account the amount of information conveyed by the auditory and visual channels and how this information integrates for the identification of emotional states expressed in dynamic emotional vocal and facial expressions.

Finally, taken singularly, signal processing, pattern recognition and machine learning strategies are not sufficient for succeeding in the algorithmic modeling of emotional vocal expressions. There is need to take account of the physical, social and organizational context, as well as to provide the system of an *a priori* knowledge that dynamically changes according to the system experience. This motivates the holistic approach.

Conclusions

This paper does not provide a set of rules on how to implement intelligent emotional interfaces. Instead, it presents a personal account on how to identify a theoretical framework to extract rules from multimodal emotional data. We first emphasized the role of the data encoding process, that is, the way knowledge can be extracted from the data and encoded in a unique representation. This is a very delicate stage since a device which uses this representation can only rely on the information that such a representation is able to encode. Then, we underlined the importance of the context, the way the data are collected and the amount of data available for training the proposed computational models. We then considered the computational models and pointed out that a new level of automaton machine intelligence approach would be necessary to solve the ditches related to the synthesis and recognition of affective states. Finally, we ended with a discussion on all the modalities that humans exploit during interaction to gather emotional information. From the section above, it can be seen that there are several sources of emotional information that have not been taken into account for the automation of the recognition process. Together with these sources, also the ability of the automaton model was always very limited, with very few, if none, prospects to gather tools that enable learning from experience and associations as well as the ability to build up a personal representation of the external and internal world. In summary, what is missed is a holistic approach to the computational treatment of affective states. Neither the signal alone, the feature processing alone, nor the computational model alone can solve the computational handling of affective states in speech, but the three aspects combined together and restricted to a given contextual application may boost up, both from a theoretical and a practical point of view, the research in speech and affective computing.

Acknowledgments This work has been supported by the European projects: COST 2102 "Cross Modal Analysis of Verbal and Nonverbal Communication" (cost2102.cs.stir.ac.uk/) and by COST TD0904 "TIMELY: Time in MEntaL activity" (www.timely-cost.eu/). Acknowledgements go to three unknown reviewers, to Isabella Poggi and to Maria Teresa Riviello for their useful comments and suggestions. Miss Tina Marcella Nappi is acknowledged for her editorial help.

References

Apolloni B, Aversano G, Esposito A (2000) Preprocessing and classification of emotional features in speech sentences. In: Kosarev Y (ed) Proceedings of international workshop on speech and computer. SPIIRAS, pp 49–52

Apolloni B, Esposito A, Malchiodi D, Orovas C, Palmas G, Taylor JG (2004) A general framework for learning rules from data. IEEE Trans Neural Networks 15(6):1333–1350

Atassi H, Esposito A (2008) Speaker independent approach to the classification of emotional vocal expressions. In: Proceedings of IEEE conference on tools with artificial intelligence (ICTAI), vol 1. Dayton, OH, USA, pp 487–494

Atassi H, Riviello MT, Smékal Z, Hussain A, Esposito A (2010) Emotional vocal expressions recognition using the COST 2102 Italian database of emotional speech. In: Esposito A et al (eds) LNCS, vol 5967. Springer, Berlin, pp 406–422

Aversano G, Esposito A, Esposito AM, Marinaro M (2001) A new text-independent method for phoneme segmentation. In: Ewing RL et al (eds) Proceedings of the IEEE international workshop on circuits and systems, vol 2, pp 516–519

Bachorowski JA (1999) Vocal expression and perception of emotion. Curr Dir Psychol Sci 8:53–57

Banse R, Scherer K (1996) Acoustic profiles in vocal emotion expression. J Pers Soc Psychol 70(3):614–636

Bargh JA, Chen M, Burrows L (1996) Automaticity of social behavior: direct effects of trait construct and stereotype activation on action. J Pers Soc Psychol 71:230–244

Barsalou LW, Niedenthal PM, Barbey AK, Ruppert JA (2003) Social embodiment. In: Ross BH (ed) The psychology of learning and motivation, vol 43. Academic Press, San Diego, pp 43–92

Benoit C, Mohamadi T, Kandel S (1994) Effects of phonetic context on audio-visual intelligibility of French. J Speech Hear Res 37: 1195–1203

Block N (1995) The mind as the software of the brain. In: Smith EE, Osherson DN (eds) Thinking. MIT Press, Cambridge, pp 377–425

Blumberg BM, Todd PM, Maes P (1996) No bad dogs: ethological lessons for learning in Hamsterdam. In: Proceedings of the 4th international conference on simulation of adaptive behaviour, MIT Press/Bradford Books, Cambridge, pp 295–304

Breazeal C, Aryananda L (2002) Recognition of affective communicative intent in robot-directed speech. Auton Robots 12: 83–104

Breitenstein C, Van Lancker D, Daum I (2001) The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. Cogn Emot 15(1):57–79

Bryant GA, Barrett HC (2007) Recognizing intentions in infant-directed speech. Psychol Sci 18(8):746–751

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: Proceedings of Interspeech, pp 1517–1520

Busso C, Lee S, Narayanan SS (2007) Using neutral speech models for emotional speech analysis. In: Proceedings of Interspeech, Antwerp, Belgium, pp 2225–2228

Butterworth BL, Beattie GW (1978) Gestures and silence as indicator of planning in speech. In: Campbell RN, Smith PT (eds) Recent advances in the psychology of language. Olenum Press, New York, pp 347–360

Callan DE, Jones JA, Munhall K, Callan AM, Kroos C, Vatikiotis-Bateson E (2003) Neural processes underlying perceptual enhancement by visual speech gestures. NeuroReport 14: 2213–2218

Chafe WL (1987) Cognitive constraint on information flow. In: Tomlin R (ed) Coherence and grounding in discourse. John Benjamins, Amsterdam, pp 20–51

de Byl PB, Toleman MA (2005) Engineering emotionally intelligent agents. Encycl Inf Sci Technol II:1052–1056

Dennett DC (1969) Content and consciousness. Humanities Press, Oxford

Douglas-Cowie E, Cowie R, Schroder M (2000) A new emotion database: considerations, source and scope. In: Proceedings of ISCA workshop on speech and emotion. Belfast, Northern Ireland

Duda R, Hart P, Stork D (2003) Pattern classification, 2nd edn. Wiley, New York

Ekman P (1992) An argument for basic emotions. Cogn Emot 6:169–200

El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44:572–587

Elman JL (1991) Distributed representation, simple recurrent neural networks, and grammatical structure. Mach Learn 7:195–225

El-Nasr MS (1998) Modeling emotion dynamics in intelligent agents. M.Sc. dissertation, American University in Cairo

Esposito A (2000) Approaching speech signal problems: an unifying viewpoint for the speech recognition process. In: Memoria of Taller Internacional de Tratamiento del Habla, Procesamiento de Vos y el Language, Suarez Garcia S, Baron Fernandez R (Eds), CIC-IPN Obra Compleata, Memoria. ISBN: 970-18-4936-1

Esposito A (2002) The importance of data for training intelligent devices. In: Apolloni B, Kurfess C (eds) From synapses to rules: discovering symbolic knowledge from neural processed data. Kluwer, Dordrecht, pp 229–250

Esposito A (2007) The amount of information on emotional states conveyed by the verbal and nonverbal channels: some perceptual data. In: Stilianou Y et al (eds) Progress in nonlinear speech processing. LNCS, vol 4391. Springer, Berlin, pp 245–268

Esposito A (2008) Affect in multimodal information. In: Tao J, Tan T (eds) Affective information processing, Springer, Heidelberg, pp 211–234

Esposito A (2009) The perceptual and cognitive role of visual and auditory channels in conveying emotional information. Cogn Comput J 2:268–278

Esposito A, Aversano G (2005) Text independent methods for speech segmentation. In: Chollet G et al (eds) Nonlinear speech modeling and applications, LNCS, vol 3445, pp 261–290

Esposito A, Marinaro M (2007) What pauses can tell us about speech and gesture partnership. In: Esposito A et al (eds) Fundamentals of verbal and nonverbal communication and the biometric issue, vol 18. IOS press, Amsterdam, pp 45–57

Esposito A, Riviello MT (2010) The new Italian audio and video emotional database. In: Esposito A et al (eds) LNCS, vol 5967. Springer, Berlin, pp 406–422

Esposito A, Riviello MT (2011) The cross-modal and cross-cultural processing of affective information. In: Apolloni B et al (eds) Frontiers in artificial intelligence and applications. IOS press, Amsterdam, pp 301–310

Esposito A, Riviello MT, Di Maio G (2009a) The COST 2102 Italian audio and video emotional database. In: Apolloni B et al (eds) WIRN09, vol 204. IOS press, Amsterdam, pp 51–61

Esposito A, Riviello MT, Bourbakis N (2009b) Cultural specific effects on the recognition of basic emotions: a study on Italian subjects. In: Holzinger A, Miesenberger K (eds) USAB 2009, LNCS, vol 5889. Springer, Berlin, pp 135–148

Fodor JA (1983) The modularity of mind. MIT Press, Cambridge

Fragopanagos N, Taylor JG (2005) Emotion recognition in human–computer interaction. Neural Netw 18:389–405

Frens MA, Van Opstal AJ, Van der Willigen RF (1995) Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Percept Psychophys 57:802–816

Friend M (2000) Developmental changes in sensitivity to vocal paralanguage. Dev Sci 3:148–162

Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. JASA 87(4):1738–1752

Hermansky H, Morgan N (1994) RASTA processing of speech. IEEE Trans Speech Audio Process 2(4):578–589

Hozjan V, Kacic Z (2003) Context-independent multilingual emotion recognition from speech signals. Int J Speech Technol 6:311–320

Hozjan V, Kacic Z (2006) A rule-based emotion-dependent feature extraction method for emotion analysis from speech. JASA 119(5):3109–3120

Hu H, Xu M, Wu W (2007) GMM supervector based SVM with spectral features for speech emotion recognition. In: Proceedings of ICASSP, vol 4, pp IV 413–IV 416

Hughes HC, Reuter-Lorenz PA, Nozawa G, Fendrich R (1994) Visual auditory interactions in sensorimotor processing: saccades versus manual responses. J Exp Psychol Hum Percept Perform 20:131–153

Izard CE (1992) Basic emotions, relations among emotions, and emotion–cognition relations. Psychol Rev 99:561–565

Jones C, Deeming A (2008) Affective human-robotic interaction. In: Peter C, Beale R (eds) Affect and emotion in HCI, LNCS, vol 4868. Springer, pp 175–185

Kaehms B (1999) Putting a (sometimes) pretty face on the web. WebTechniques, CMP Media. www.newarchitectmag.com/archives/1999/09/newsnotes/

Klasmeyer G, Sendlmeier WF (1995) Objective voice parameters to characterize the emotional content in speech. In: Elenius K, Branderudf P (Eds) Proceedings of ICPhS, Arne Strömbergs Grafiska, vol 1, pp 182–185

Lindblom B (1990) Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle J, Marchal A (eds) Speech production and speech modeling. Kluwer, Dordrecht, pp 403–439

Lugger M, Yang B (2007) The relevance of voice quality features in speaker independent emotion recognition. In: Proceedings of ICASSP, vol 4, pp 17–20

Macaluso E, George N, Dolan R, Spence C, Driver J (2004) Spatial and temporal factors during processing of audiovisual speech: a PET study. NeuroImage 21:725–732

Makhoul J (1975) Linear prediction: a tutorial review. Proc IEEE 63(4):561–580

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748

Navas E, Hernáez I, Luengo I (2006) An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. IEEE Trans Audio Speech Lang Process 14(4):1117–1127

Newell A, Simon HA (1972) Human problem solving. Prentice Hall, Oxford

Nushikyan EA (1995) Intonational universals in texual context. In: Elenius K, Branderudf P (eds) Proceedings of ICPhS 1995, Arne Strömbergs Grafiska, vol 1, pp 258–261

Nwe T, Foo S, De Silva L (2003) Speech emotion recognition using Hidden Markov models. Speech Commun 41:603–623

Oatley K, Jenkins JM (2006) Understanding emotions, 2nd edn. Blackwell, Oxford

Penrose R (1989) The emperor's new mind. Oxford University Press, New York

Perrott DR, Sadralodabai T, Saberi K, Strybel TZ (1991) Aurally aided visual search in the central visual field: effects of visual load and visual enhancement of the target. Hum Factors 33: 389–400

Petrushin V (1999) Emotion in speech: recognition and application to call centers. In: Proceedings of the conference on artificial neural networks in engineering, pp 7–10

Picard R (2000) Toward computers that recognize and respond to user emotion. IBM Syst J 39(3–4):705–719

Pierre-Yves O (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum Comput Stud 59: 157–183

Plutchik R (1993) Emotion and their vicissitudes: emotions and psychopatology. In: Lewis JM, Haviland-Jones M (eds) Handbook of emotion. Guilford Press, New York, pp 53–66

Pudil P, Ferri F, Novovicova J, Kittler J (1994) Floating search method for feature selection with non monotonic criterion functions. Pattern Recogn 2:279–283

Pylyshyn ZW (1984) Computation and cognition: toward a foundation for cognitive science. MIT Press, Cambridge

Razak A, Komiya R, Abidin M (2005) Comparison between fuzzy and nn method for speech emotion recognition. In: Proceedings of 3rd international conference on information technology and applications ICITA, vol 1, pp 297–302

Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39:1161–1178

Scherer KR (1989) Vocal correlates of emotional arousal and affective disturbance. In: Wagner H, Mner H, Manstead A (eds) Handbook of social psychophysiology. Wiley, New York, pp 165–197

Scherer K (2003) Vocal communication of emotion: a review of research paradigms. Speech Commun 40:227–256

Scherer KR, Banse R, Wallbott HG (2001) Emotion inferences from vocal expression correlate across languages and cultures. J Cross Cult Psychol 32:76–92

Schubert TW (2004) The power in your hand: gender differences in bodily feedback from making a fist. Pers Soc Psychol Bull 30:757–769

Schuller B, Rigoll G, Lang M (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proceedings of the ICASSP, vol 1, pp 577–580

Schulz M, Ross B, Pantev C (2003) Evidence for training-induced cross modal reorganization of cortical functions in trumpet players. NeuroReport 14:157–161

Slaney M, McRoberts G (2003) Baby ears: a recognition system for affective vocalizations. Speech Commun 39:367–384

Sloman A (2001) Beyond shallow models of emotion. Cogn Process 2(1):177–198

Smit ER, Semin GR (2004) Socially situated cognition: cognition in its social context. Adv Exp Soc Psychol 36:53–117

Stein BE, Jiang W, Wallace MT, Stanford TR (2001) Nonvisual influences on visual-information processing in the superior colliculus. Prog Brain Res 134:143–156

Stepper S, Strack F (1993) Proprioceptive determinants of emotional and non-emotional feelings. J Pers Soc Psychol 64:211–220

Ström N (1997) Sparse connection and pruning in large dynamic artificial neural networks. In: Proceedings of Eurospeech, vol 5, pp 2807–2810

Velasquez JD (1999) From affect programs to higher cognitive emotions: an emotion-based control approach. In: Proceedings of workshop on emotion-based agent architectures, Seattle, USA, pp 10–15