Knowledge-Based Systems 63 (2014) 68–81

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure

Arianna Mencattini [a,*], Eugenio Martinelli [a], Giovanni Costantini [a,d], Massimiliano Todisco [a], Barbara Basile [b,c], Marco Bozzali [b], Corrado Di Natale [a,d]

[a] Dept. of Electronic Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy
[b] Neuroimaging Laboratory, Santa Lucia Foundation, Rome, Italy
[c] School of Cognitive Psychotherapy, Rome, Italy
[d] Institute of Acoustics and Sensors "Orso Mario Corbino", Via del Fosso del Cavaliere, 100 Rome, Italy

ARTICLE INFO

Article history:
Received 19 December 2013
Received in revised form 12 March 2014
Accepted 22 March 2014
Available online 2 April 2014

Keywords:
Speech emotion recognition (SER)
Circumplex model of emotions
Partial least square (PLS) regression
Pearson correlation coefficient
Pitch contour characterization
Audio signal modulation

ABSTRACT

Speech emotion recognition (SER) is a challenging framework in demanding human machine interaction systems. Standard approaches based on the categorical model of emotions reach low performance, probably due to the modelization of emotions as distinct and independent affective states. Starting from the recently investigated assumption on the dimensional circumplex model of emotions, SER systems are structured as the prediction of valence and arousal on a continuous scale in a two-dimensional domain. In this study, we propose the use of a PLS regression model, optimized according to specific features selection procedures and trained on the Italian speech corpus EMOVO, suggesting a way to automatically label the corpus in terms of arousal and valence. New speech features related to the speech amplitude modulation, caused by the slowly-varying articulatory motion, and standard features extracted from the pitch contour, have been included in the regression model. An average value for the coefficient of determination R^2 of 0.72 (maximum value of 0.95 for fear and minimum of 0.60 for sadness) is obtained for the female model and a value for R^2 of 0.81 (maximum value of 0.89 for anger and minimum value of 0.71 for joy) is obtained for the male model, over the seven primary emotions (including the neutral state).

1. Introduction

In the community of Human Computer Interface (HCI) researchers have been working for several years in trying to emulate a human communication system, using innovative technologies and methodologies, based on the emotion recognition in facial expressions and speech [1–4]. Speech analysis techniques find applications in the area of legal field-law for the assessment of an individual's psychological integrity, in security and surveillance, in computer tutorial applications, in car board systems, in automatic translations systems, in call center applications, and in diagnostic tools for therapists in clinical studies, psychosis monitoring and diagnosis of neuropsychological disorders [5,6]. Audio communication is a mixture of verbal and non-verbal coding that compose the so called *vocal communicative act* [7]. Hence, the communicative effect of a word, pronounced in a certain utterance, may assume different meanings according to the emotions that accompany the utterance. In various applications, speech emotion recognition (SER) seems the optimal way to decode emotions in a subject, superior to other communication channels: speech acquisition is simple and does not face problems due to subject movements (as in facial analysis); speech analysis is not influenced by visual artifacts (occlusions due to glasses, mustaches, beard, etc.); speech can be easily recorded in any environment and does not require large memory storage support or particular acquisition devices or procedures (i.e., webcam, stereo vision); most of the subjects refused to be filmed for shame, for protection of privacy, or simply for nuisance, especially among patients; due to privacy preserving and anonimization of sensible data, it is very difficult to implement a study on a large population of "naif" subjects which is based on facial analysis. To be challenging, a SER system should be robust and effectively solve some relevant problems: (i) the large set of features available for speech analysis and the absence of a stable evidence of the most relevant ones [8]; (ii) the huge variability of utterances that can be pronounced and no evidence on the best choices (sense vs. non-sense utterances); (iii) an intrinsic inter-speakers variance of the speech features

* Corresponding author. Tel.: +39 0672597368.
E-mail address: mencattini@ing.uniroma2.it (A. Mencattini).

http://dx.doi.org/10.1016/j.knosys.2014.03.019
0950-7051/

and no evidence on the optimal way to normalize them to reduce it; (iv) low emotion recognition rate obtained by the SER systems described in the literature; and (v) cross-linguistic testing.

Issue (i) can be addressed by searching for new speech features and implementing novel feature selection and reduction procedures accounting for the maximum relevance of the features with the target and the minimum redundancy among them. Concerning the kind of utterances, the use of both sense and non-sense sentences could be a choice (issue ii)) also to implement cross-linguistic testing (issue v)). To reduce inter-speakers variance, features are preliminarily mean centered and divided by each standard deviation (autoscale). Features normalization can have a severe impact on the performance of many machine learning algorithms. We believe that other kinds of normalization could reduce the capability to recognize emotions in speech and the generalizing properties of the implemented SER system (issue iii)). To prove this assumption, we compare standard autoscale of each feature with the use of min–max normalization (each feature is subtracted of the minimum value and then scaled by the difference of the maximum and the minimum value thus belonging to the range $[0, 1]$) and with only the rescaling by the standard deviation normalization. The very different range of the features extracted requires feature normalization at least for range equalization. Section 5 will illustrate results of this comparison. At this moment, the low recognition rate reported in the literature reduces the appealing of using SER systems as the key strategy in novel HCI systems. We believe that one of the limitations is that standard SER approaches are based on a categorical emotional model [9,10] in which emotions are categorized into 6 primary emotions of disgust, joy, fear, anger, surprise, and sadness, with the addition of the neutral state that identifies the absence of emotions. Following this model, emotions are assumed to belong to independent and distinct categories. However, this taxonomy does not allow to explain the frequent co-morbility among different psychological diseases, and does not provide a way to relate neurophysiological substrate to emotions. During last 10 years, the categorical model has been partially substituted by a *dimensional* approach of emotions. This model has been supported by several studies and statistical analyses through the observation of the affective states from verbal description, facial expressions, subjective experience and memories, self-reporting, etc. in different cultural frameworks [11–16] and of neuroimaging findings, showing involvement of specific neuronal substrates. The so called *circumplex model* [15,16] of affects assumes that the affective state can be represented on a continuous of two dimensions: valence (from unpleasantness to pleasantness) and arousal (from low activation to high activation) as illustrated in Fig. 1. A linear combination of these two quantities permits a two-dimensional representation of all emotions. As an example, joy is an affective state with positive valence and a moderate arousal level while sadness might be characterized by negative valence and low arousal. The circumplex model assumes that emotions are not discrete and specific affective states, but they are fuzzy and ambiguous experiences, often interacting, and simultaneously present in a subject [17]. In analogy to colors, it seems more correct to speak of an halo of emotions when, for example, a subject feels simultaneously joyful, surprised, and excited. In a practical scenario in which "naif" subject express their emotions in a verbal, subjective, and natural way, this model seems to be the correct framework of representation of spontaneous affective states. Although promising, automatic dimensional emotion recognition is still in its pioneering stage. First attempts in this field have addressed similar or associated sub-problems. Mainly, due to the intrinsic complexity of regression vs. classification approaches, emotion recognition from utterances is addressed by solving the binary problem of negative vs. positive valence [18] or the 4-quadrants problem of positive-active, positive-passive, negative-active, and negative-passive emotions [19,20]. Only the work in [17] addressed the problem of valence vs. arousal using a regression strategy applied on utterances.

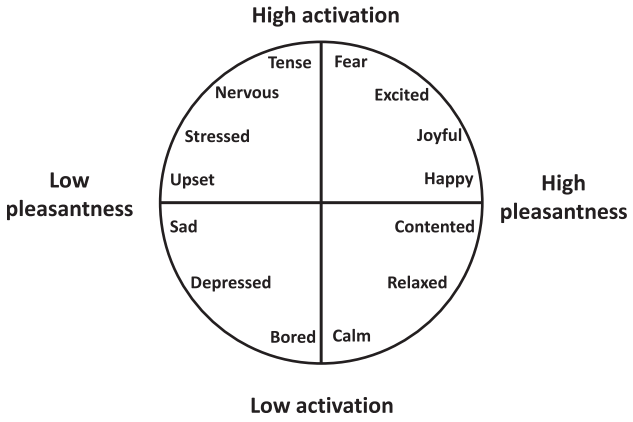

Fig. 1. The circumplex model of emotions.

The rest of the paper is organized as follows: Section 2 contains a description of literature review. Section 3 contains the speech corpus used for validation and testing. Section 4 contains the description of the methodology. Section 5 reports results and comparisons. Section 6 contains discussions and Section 7 reports final conclusions.

2. Related works and contributions

The importance of the proposed theme is proved by an increasing number of related works that can be found in the literature. Basically the works can be grouped according to two relevant aspects: speech emotion classification vs. regression ($2D$–$3D$ domain) and features extraction and selection techniques (filter methods vs. wrappers and embedded approaches) [21–29].

The innovation of this study concerns: the presentation of a reliable and speaker-independent automatic gender identification procedure; the use of a large set and partially innovative features (suprasegmental and amplitude modulation) for the estimation of valence and arousal as a flexible and powerful alternative to the categorical emotion recognition; the implementation of a novel dual-stage features selection procedure to minimize inter-correlation among features and select the most significant characteristics, using a selection method that is totally independent from the recognition step; the development of an easy and robust regression technique based on PLS for the valence/arousal estimation; the representation of the result on the circumplex diagram for a visual and simultaneous understanding of the results either in terms of emotional categories and by their dimensional localization; the testing of the system on the novel Italian database EMOVO with the inclusion of nonsense utterances.

The methods can serve as a valid instrument in conducting special studies to identify psychological-motivated relationships between features and emotion dimensions [30], in identifying relevant utterances and/or representative actors/persons for an intra-corpus characterization [31], in performing cross-linguistic studies due to the absence of semantic features [3]. Moreover, the approach proposes a solution to solve the problem of observer's rating variation in dimensional labelling [32,33].

3. Materials

One of the first Italian attempt to build a database of emotional speech has been performed by Fondazione Ugo Bordoni (FUB) in Rome, with EMOVO [34]. The EMOVO database is a speech corpus