

Automatic Emotion Recognition from DEMoS Corpus by Machine Learning Analysis of Selected Vocal Features

Giovanni Costantini¹^a, E. Parada-Cabaleiro² and Daniele Casali¹^b
¹*Department of Electronic Engineering, University of Rome Tor Vergata, 00133 Rome, Italy*
²*Institute of Computational Perception, Johannes Kepler University Linz, Austria*

Keywords: Speech Emotional Recognition, Italian Corpus, Mood Induction, Natural Speech, Acoustic Features.

Abstract: Although Speech Emotion Recognition (SER) has become a major area of research in affective computing, the automatic identification of emotions in some specific languages, such as Italian, is still under-investigated. In this regard, we assess how different machine learning methods for SER can be applied in the identification of emotions in Italian language. In agreement with studies that criticize the use of acted emotions in SER, we considered DEMoS, a new database in Italian built through mood induction procedures. The corpus consists of 9365 spoken utterances produced by 68 Italian native speakers (23 females, 45 males) in a variety of emotional states. Experiments were carried out for female and male separately, considering for each a specific feature set. The two feature sets were selected by applying Correlation-based Feature Selection from the INTERSPEECH 2013 ComParE Challenge feature set. For the classification process, we used Support Vector Machine. Confirming previous work, our research outcomes show that the basic emotions anger and sadness are the best identified, while others more ambiguous, such as surprise, are worse. Our work shows that traditional machine learning methods for SER can be also applied in the recognition of an under-investigating language, such as Italian, obtaining competitive results.


1 INTRODUCTION


Since everyday life is closely related to sophisticated machines, to improve the human-machine interaction has become more important than ever before. The primary communication channel between humans is speech, i.e., a natural, fast but also complex signal that offers information about the speaker and the message. This information arrives mainly through two channels: the verbal channel, which transmits explicit messages; and the non-verbal channel, which transmits implicit messages. In order to correctly understand the meaning of the message, both channels are essential. Cowie et al. (2001) consider that the non-verbal channel tells people how to interpret what is transmitted through the verbal channel. Indeed, the same words can acquire different meanings if they are used as a joke or a genuine question that seeks an answer. This source of information plays a vital role in an interactive communication process, because the speaker's affective state, not only enriches the human

communication but help us to predict possible speaker feedback too. For this reason, the understanding of the speaker's emotional state is crucial in the development of natural and efficient human-machine interactions.

The goal of Speech Emotion Recognition (SER) is to automatically identify the human emotional state analyzing the speech signal. The typical pattern of automatic recognition systems contains four modules: speech input, feature extraction and selection (which contains emotional information from the speaker's voice), classification, and emotion output (Joshi and Zalte, 2013).

SER is a growing field in developing friendly human-machine interaction systems with wide use in telecommunication services such as call center applications and mobile communication (Chateau et al., 2004; Lee and Narayanan, 2003), multimedia devices, such as video and computer games (Costantini et al., 2014; Cullen et al., 2008; Ocquaye et al., 2021; Parada-Cabaleiro et al., 2018a), diagnostic medical tools (Alessandrini et al., 2017;

^a <https://orcid.org/0000-0001-8675-5532>

^b <https://orcid.org/0000-0001-8800-728X>

France et al., 2000; Saggio et al., 2020; Suppa et al., 2020) and security services, such as surveillance systems or lie detectors (Clavel et al., 2006a, 2006b).

The Ekman’s classification of universal basic emotions, known as “the big six” (anger, disgust, fear, happiness, sadness and surprise), is usually utilized in SER research (Cowie and Cornelius, 2003; Cowie et al., 2005). Ekman (1972, 1977, 1984; Ekman and Friesen, 1971) supports his universal classification on different cross-cultural studies about the facial expression of emotions. These studies demonstrate that different cultures have the same emotion perception due to the natural use of the same basic facial expressions. In our study five of the “big six” emotions are considered: surprise, fear, happiness, anger and sadness. Although disgust is also contained in the DEMoS dataset (Parada-Cabaleiro et al., 2020), this was discarded since the induction of this emotion did not yield reliable results. In addition to the five basic emotions, the secondary emotion guilt was also taken into account (Parada-Cabaleiro et al., 2018b; Parada-Cabaleiro et al., 2020).

The SER system is based on the INTERSPEECH 2013 ComParE Challenge feature set, extracted with openSMILE (Eyben et al., 2013) and on the use of Support Vector Machine (SVM) classifier, implemented on Weka (Hall et al., 2009).

For our experiments, we considered the DEMoS database (Parada-Cabaleiro et al., 2020), i.e., an induced-based emotional speech corpus in Italian language.

The main goals of this work are: (i) to examine how the induced speech is recognized with the proposed methods; (ii) to identify the confusion patterns typical of each emotion; (iii) to compare the performance of different classifiers and feature processing methods.

The rest of this paper is organized as follows: in Section 2, our methodology is presented, giving an overview of the DEMoS corpus as well as a brief description of the machine learning set-up. In Section 3, results are discussed. Finally, in Section 4, the general conclusions are presented.

2 MATERIALS AND METHODS

2.1 DEMoS: An Italian Corpus for SER

In this work, we used the DEMoS copus (Parada-Cabaleiro et al. 2020). The corpus consists of 9365 audio utterances pronounced by 68 native speakers (23 females, 45 males, mean age 23.7 years, std 4.3 years).

Following different validation procedures, the dataset presents a reduced sub-set of selected utterances considered prototypical, i.e., clearly representative of each given emotion.

The sub-set of the corpus encompasses 1564 samples produced by 59 speakers (21 females, 38 males): 422 express sadness, 246 anger, 177 fear, 203 surprise, 167 happiness, and 209 guilt.

The number of sentences is variable for each emotion and for each speaker, because the induction techniques have different level of effectiveness for different emotions and people. This happens because emotions are not always induced with the same success. For instance, the induction of anger showed to be much more difficult than the induction of sadness.

In order to assess to which extent, the emotional instances were representative of the given emotions, these were evaluated in a perceptual study by 86 listeners. In table 1 and 2, confusion matrices showing the perceptual results for the listeners’ evaluation of female and male voices are given, respectively

Table 1: Confusion matrix for the listeners’ perception of emotional speech produced by female speakers in: happiness (hap), guilt (gui), fear (fea), anger (ang), surprise (sur), and sadness (sad).

%	hap	gui	fea	ang	sur	sad
hap	74.3	0.6	1.7	1.5	11.9	9.9
gui	7.2	34.5	4.7	7.5	10.0	36.0
fea	3.2	4.7	66.0	6.9	12.3	6.9
ang	1.0	2.3	6.1	81.2	7.0	2.4
sur	26.6	1.9	6.1	12	46.5	6.9
sad	4.39	6.9	9.6	6.8	9.9	62.3

Table 2: Confusion matrix for the listeners’ perception of emotional speech produced by male speakers in: happiness (hap), guilt (gui), fear (fea), anger (ang), surprise (sur), and sadness (sad).

%	hap	gui	fea	ang	sur	sad
hap	74.2	0.7	2.3	2.3	16.4	2.1
gui	16.9	23.2	6.7	6.6	8.1	38.5
fea	8.5	3.2	39.5	13.4	13.4	22.0
ang	1.0	3.0	9.4	77.0	7.1	2.5
sur	33.8	1.3	3.1	5.6	52.1	4.0
sad	4.6	5.6	8.2	5.4	4.2	72.0

Each row gives the “reference”, each column “identifies as”, for each of the six considered emotion. The results show that guilt was particularly difficult to recognize for both, female and male speakers, showing a prominent confusion pattern towards sadness (cf. gui vs sad for female and male, in Table 1 and 2, respectively).

2.2 Extracted Features

The feature set used in this work is the INTERSPEECH 2013 ComParE Challenge feature set, extracted with openSMILE feature extractor (Eyben et al., 2013). Created in the scope of the European EU-FP7 research project SEMAINE (<http://www.semaine-project.eu>), openSMILE (The Munich open Speech and Music Interpretation by Large Space Extraction) is a feature extractor for signal processing and machine learning applications used in the field of speech recognition, affective computing, and music information retrieval. The open-source version of openSMILE is available for research and educational use but not for commercial aims. The set includes energy, spectral and voicing related low-level descriptors (LLDs), that contains all together 6373 features (Schuller et al., 2010). The data format RIFF-WAVE (PCM), on which the database is made, can be read by openSMILE, which generates the output data in a Weka ARFF file, subsequently considered for the feature selection and classification. Since the speech signal is not stationary, in speech processing, it is usual to fragment the signal into small segments known as frames and considered as stationary. Global features may surpass local features in accuracy, classification time, and computational cost; yet, as these features do not have temporal information about signals, to use classifiers such as the hidden Markov model or the SVM is unreliable. Furthermore, some studies have shown that global features are not efficient in distinguishing between emotions with similar arousal (El Ayadi et al., 2011). In order to take advantage of both types of features some authors decided to use global and local features combined (Vlasenko et al., 2007; Li e Zhao, 1998).

Another still open question regards the most suitable features for SER. These have to be efficient in the characterization of emotions but also independent of both the speaker and the linguistic content. For a long time, the most used features in SER were the prosodic features, also called supra-segmental acoustic features (Lee e Narayanan, 2005; Busso et al., 2012) or continuous features (El Ayadi et al., 2011). These features regard mainly three perceptive aspects: pitch, related to fundamental frequency (F0); loudness, related to energy; and timing, related to speech and pause duration. However, in the last years the spectral features, also called segmental acoustic features, which relate to the distribution of spectral energy (Lee e Narayanan, 2005; Busso et al., 2012; Krothapalli e Koolagudi, 2012), along to phonetic features (Nwe et al., 2003)

and system features (Krothapalli e Koolagudi, 2012), have become much more used (Schuller et al., 2010). In addition to the previous, voice quality features, i.e., those describing the excitation glottal properties, and also known as intrasegmental acoustic features (Busso et al., 2012), qualitative features (El Ayadi et al., 2011) or excitation source features (Krothapalli e Koolagudi, 2012), are becoming much more used as well (Schuller et al., 2010). Although having an effective features set is essential to create an efficient SER system, so far there is no agreement about which features are more suitable for the SER tasks. Since considering many features suppose a big computational cost, while considering too few risks overlooking fundamental aspects, in order to reduce the computational cost without risk to omit essential information, we performed features selection, by this eliminating the redundant ones (El Ayadi et al., 2011). In Table 3 and 4, a description of the different feature groups considered in this work is presented.

2.3 Feature Selection and Processing

In this work features were also discretized (Fayyad, 1993) and subsequently selected by means of the Correlation-base Feature Selection (CFS) method. The CFS algorithm, chooses only those features that have higher correlation with the class and lower correlation among themselves. According to this algorithm, the following formula is adopted to measure the “merit” of a feature subset S containing k features:

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k+(k-1)\overline{r_{ff}}}} \quad (1)$$

where $\overline{r_{cf}}$ is the mean feature-class correlation ($f \in S$) and $\overline{r_{ff}}$ is the average feature-feature inter-correlation (Asci et al., 2020).

We selected two feature sets: one for male voices and another one for female voices. Most selected features were related to magnitude spectrum, voicing features (pitch, energy, etc.), Mel-Frequency Cepstral Coefficients (MFCC), and RelAtive SpecTrAl (RASTA) coefficients (Hermansky & Morgan, 1994).

In our experimental results we will compare the results obtained with and without features’ discretization.

Table 3: Selected feature for males. LLD: Low Level Descriptor; MFCC: Mel Frequency Cepstral Coefficient. The suffix “de” indicates that the current feature is a 1st order delta coefficient (differential) of the smoothed low-level descriptor (delta regression coefficients computed from the feature). “iq” means inter-quartile and “p” means percentile, “q” means quartile.

Selected features for males		
Families of LLDs	LLDs	Functionals
RASTA coefficients	L1 norm	Lpc3, mean rising slope, q 3, Min. pos., rise time, p, max pos., up level time 50, iq 1-2, skewness
	Coefficient of band 0, 2 (de), 5 (de), 6 (de), 10, 12 (de), 16, 21 (de), 22 (de), 23, 23, 24 (de)	
Magnitude Spectrum	Roll off 25.0, 75,90,0	Kurtosis, rise time Rise time iq 2-3 Max pos, uplevel time 25, min pos, p 1, range, skewness
	Spectral flux	
	Spectral entropy	
	Spectral slope	
	Spectral sharpness	
	Harmonicity	
Voicing Related	Fundamental Frequency flatness	Mean, max pos, iq 2-3, max pos, min pos
MFCC	Mel Coefficient 1,2,3,5,6,11,12,14	Linear regression coffefficient 1, mean, mean falling slop, peak mean abs

Table 4: Selected feature for females. LLD: Low Level Descriptor; MFCC: Mel Frequency Cepstral Coefficient. The suffix “de” indicates that the current feature is a 1st order delta coefficient (differential) of the smoothed low-level descriptor (delta regression coefficients computed from the feature). “iq” means inter-quartile and “p” means percentile, “q” means quartile.

Selected features for females		
Families of LLDs	LLDs	Functionals
RASTA coefficients	Coefficients of band 2,5, 17	Lpc3, Mean rising slope
	Coefficients of band 2, 10 (de)	
Magnitude Spectrum	250 - 650	P 1, kurtosis, rise time, standard deviation, q 2, Inter-quartile 1-2, uplevel 75
	1000-22000	
	Roll off 25	
	Spectral flux	
	Harmonicity	
	roll off 25, 50 (de)	
Voicing Related	100-2000 (de)	Mean, quartile 1, iq1-2, p 99, skewness
	Fundamental Frequency flatness	
MFCC	Mel Coeff. 1,2,3,7,9,12,14	Quartile 3, p 1, p 99, quartile 2, lpc1 q 3, iq 1-2, il 1-2

2.4 Classification

Currently, there is not agreement about which classifier is the most accurate in SER, since all of them present some advantages and limitations.

In this work, the software chosen for the classification is Weka (Hall et al., 2009), a free software of machine learning written in Java and developed at the University of Waikato, New Zealand. Weka contains a collection of algorithms for data analysis and predictive modeling such as classification or feature selection. We used the SVM classifier, which ensures high performance, even with large datasets and with audio data, as is also confirmed by previous research (Costantini et al., 2010a; Costantini et al., 2010b; Saggio et al., 2011).

3 RESULTS

In Table 5(a) and 6(a), the accuracy percentage obtained by the automatic recognition of all the

emotions together, without feature discretization, is displayed for male voices and female voices respectively. Our results show an accuracy of 69% for the male voices and 58% for the female voices. While basic emotions like anger and sadness present high level of recognition (near 80% in the male voices), emotions like fear, guilt, and surprise (for female voices) are the worst recognized. A comparison with Tables 1 and 2 shows that performances of man and machine in recognition are quite similar for most emotions, with few exceptions, especially for guilt.

These results are in line with previous work on the perception of emotional speech (Parada-Cabaleiro et al., 2018a), that showed that some emotions, as anger and sadness, present standard expressions across different cultures, whereas others, as e.g., surprise or guilt, defined as ambiguous labels (Scherer, 1986), are expressed very differently between cultures and individuals. Similarly, comparable outcomes have been presented in the SER domain (Oudeyer, 2003; Jiang and Cai, 2004; Borchert and Dusterhoft, 2005; Austermann et al., 2005; Lugger and Yang, 2007; Mencattini et al. 2014; Wu et al., 2009).

Table 5: Accuracy and confusion matrix for male voices, without (a, c, e) and with (b, d, e) feature discretization.

%	hap	gui	fea	ang	sur	sad
hap	41.8	15.2	5.1	6.3	13.9	17.7
gui	4.7	45.0	4.7	1.6	1.6	42.6
fea	5.2	6.0	67.2	4.3	9.5	7.8
ang	2.89	1.2	3.5	83.2	5.2	4.1
sur	6.7	4.0	9.3	8.7	65.3	6.0
sad	12.1	3.2	2.1	1.8	1.8	79.4

a) accuracy = 69%

%	hap	fea	ang	sur	sad
hap	65.8	2.5	1.3	10.1	20.3
fea	2.6	71.6	4.3	10.3	11.2
ang	0.0	5.8	82.1	6.9	5.2
sur	7.3	9.3	11.3	66.0	6.0
sad	4.1	4.1	1.8	2.7	87.4

c) accuracy = 78.4%

%	hap	fea	ang	sad
hap	64.6	6.3	6.6	22.8
fea	6.0	74.1	5.2	14.7
ang	3.5	5.2	85.6	5.8
sad	3.8	2.4	2.3	90.9

e) accuracy = 83.9%

%	hap	gui	fea	ang	sur	sad
hap	70.9	8.9	2.5	2.5	6.3	8.9
gui	3.9	60.4	3.9	1.6	0.8	29.4
fea	1.7	6.9	82.8	2.6	2.6	3.4
ang	1.1	0.6	2.3	89.0	4.6	2.3
sur	2.7	2.7	2.0	1.3	87.3	4.0
sad	0.6	10.0	1.2	1.5	1.2	85.6

b) accuracy = 81.7%

%	hap	fea	ang	sur	sad
hap	79.8	3.8	1.3	8.9	6.3
fea	0.0	87.1	1.7	3.5	7.8
ang	1.2	2.3	88.4	5.8	2.3
sur	2.0	4.7	6.7	82.7	4.0
sad	1.2	1.2	0.6	0.3	96.8

d) accuracy = 89.7%

%	hap	fea	ang	sad
hap	83.6	5.1	2.5	8.9
fea	0.9	86.2	2.6	10.3
ang	1.7	2.9	91.9	3.5
sad	2.4	0.9	0.3	96.5

f) accuracy = 92.2%

Table 6: Accuracy (a) and confusion matrix for female voices, without (a, c, e) and with (b, d, e) feature discretization.

%	hap	gui	fea	ang	sur	sad
hap	73.3	11.6	5.8	2.3	2.3	4.7
gui	9.6	56.6	6.0	4.8	4.8	18.1
fea	5.17	20.7	39.7	12.5	15.5	3.4
ang	14.8	14.8	22.2	18.5	18.5	11.1
sur	13.0	4.3	13.0	34.8	34.8	0.0
sad	6.2	30.0	2.5	1.3	1.3	58.8

a) accuracy = 58.4%

%	hap	fea	ang	sur	sad
hap	78.4	4.6	3.41	5.7	8.0
fea	17.7	48.4	12.9	9.7	11.3
ang	4.1	8.2	78.1	5.5	4.1
sur	15.1	13.2	7.6	56.7	8.0
sad	14.6	6.1	2.4	1.2	75.7

c) accuracy = 69.3%

%	hap	fea	ang	sad
hap	81.8	4.6	4.6	9.1
fea	11.5	60.7	18.0	9.8
ang	11.0	12.3	74.0	2.7
sad	14.6	6.1	1.2	78.1

e) accuracy = 74.7%

%	hap	gui	fea	ang	sur	sad
hap	84.1	4.6	1.1	3.4	3.4	3.4
gui	8.8	76.3	0.0	1.3	2.5	11.2
fea	3.3	4.9	80.3	3.3	3.3	4.9
ang	4.1	4.1	2.7	84.9	4.1	0.0
sur	13.2	5.7	5.7	3.8	69.8	2.0
sad	4.9	12.2	1.2	3.7	1.2	76.8

b) accuracy = 79.2%

%	hap	fea	ang	sur	sad
hap	96.6	0.0	0.0	2.3	1.1
fea	8.1	71.0	8.1	6.5	6.5
ang	4.11	6.9	84.9	1.4	2.7
sur	11.3	7.6	9.4	69.8	1.9
sad	4.9	4.9	0.0	0.0	90.2

d) accuracy = 84.4%

%	hap	fea	ang	sad
hap	89.8	4.6	1.1	4.6
fea	4.9	75.4	13.1	6.7
ang	4.1	9.6	86.3	0.0
sad	6.1	7.3	0.0	86.6

f) accuracy = 85.2%

In Table 5(b) and 6(b), results for the feature discretization are given, procedure which yielded significantly better results, with an overall performance of 81.7% for male voices and 84.1%, for female voices. Despite the overall improvement in the performance, guilt, fear and surprise are still the emotions worse recognized, especially for female

voices. This confirms, once again, the outcomes above presented, i.e., more ambiguous emotions are worse identified even in optimal conditions, i.e., with a more efficient feature set.

After all, the ambiguous emotions, which are more complex than the so-called universal emotions, present many different expressive characteristics,

both in the same culture and in different cultures, reason why their recognition is more difficult. This becomes clear if we think about fear, which can be both active and passive, as well as surprise, which can be both positive and negative. Indeed, surprise is not considered as a primary emotion by some authors because it does not present a clear valence, an aspect instead essential in the others primary emotions (Ortony and Turner, 1990). Concerning guilt, the ambiguity becomes even more prominent, since is a secondary emotion. On the other hand, the universal emotions, which present more standardized expressions among different cultures and individuals, such as anger and sadness, are easily recognized. Interestingly, the confusion patterns shown by the SER system are also confirmed by the perceptual study, as shown in Tables 1 and 2.

Since guilt was the only secondary emotion, we considered this might have increased considerably the level of confusion. Due to this, the experiments were performed again without considering this emotion. Experiment on the five emotions yielded to an overall improvement: for male voices the SER system achieved 78.4% accuracy; for female, 69.2%; cf. results in Table 5(d) and 6(d), respectively. Fear was one of the emotions that most increased in accuracy level, passing from 39.7% to 48.4% in the female voices; cf. fea in Table 6(a) and 6(c), respectively. The recognition of sadness was enhanced as well: 58.8% to 75.7% (cf. Table 6(a) and (c), respectively), something however expected, since sadness was the emotion mainly misclassified as guilt (30% of the instances of sadness were wrongly identified as guilt); cf. sad vs gui in Table 6(a). The recognition of surprise showed an unexpected behavior, in male voices: from 65.3% in the classification with seven emotions, to 66% in the classification with six; cf. Table 5(a,c). This might be given by the fact that sadness, although considered a basic emotion for some authors, is in any case more ambiguous than happiness, sadness, fear and anger; thus, presenting confusion patterns even in optimized conditions. Probably the confusion of happiness is related to the fact that usually this emotion is similar to neutral expressions when arousal is low. This was demonstrated in previous research (Parada-Cabaleiro et al., 2018a), where happiness was mainly confused with a neutral label. Indeed, happiness in male voices was confused mainly with sadness, usually characterized by low level arousal, and so, similar to neutral. The experiments were also carried out with feature discretization, with results shown in Table

5(d) and 6(d). We can observe differences between experiments performed with and without discretization: for example, fear, for female voices, didn't increase so much, because its accuracy was already considerably high with six emotions. Differently, the accuracy of sadness increases particularly (from 85.6% to 96.8% in male voices, Table 5(b, d) and from 76.8% to 90.2% in female voice, Table 6(b, d), is still noticeable. Since surprise is considered a secondary emotion by some authors (Ortony and Turner, 1990), we decided to carry out the experiments again without considering it, i.e., evaluating only the four basic emotions happiness, fear, anger, and sadness. In Table 5(e) and Table 6(e), the results for this setting are given. The accuracy of the system was considerably better, both for the female and male voices (achieving 83.6% and 74.7% of accuracy, respectively). This time the until now best recognized emotions, i.e., anger and sadness, presented levels of accuracy very similar to the other two emotions, i.e., happiness and fear. The accuracy levels for fear increased particularly for the female voice: from 48.4% considering five emotions, to 60.7% considering four; cf. Tables 6 (c) and (e), respectively. The results for male voices confirmed again that the more standardized emotions were anger and sadness, although the results of this last test demonstrated that more ambiguous emotions, such as fear and happiness, can improve notoriously their recognition level if the test is performed in optimized conditions, such as considering a lower number of emotions. When doing feature discretization (Tables 5(f) and 6(f)), we found that accuracy improvement is not so strong, but still we can observe more uniformity among the different emotions.

4 CONCLUSIONS

Confirming previous work (Parada-Cabaleiro et al., 2018a), our study shows that confusion patterns typical of perceptual studies (Cowie and Cornelius, 2003), i.e., more standardized emotional expressions, such as anger and sadness, are better identified than those more ambiguous, such as surprise and guilt, are also shown in SER of Italian language. This is confirmed by the fact that, when many ambiguous emotions are recognized together, the overall accuracy of the SER system decreases considerably, while the standardized emotions (anger and sadness) still maintain a good performance. As expected, the recognition of ambiguous emotions improves when

few of these are recognized together. This is due to the fact that ambiguous emotions share common characteristics, i.e., there is no a prototypical representation of these emotions, reason why they cannot be easily differentiated when recognized together.

REFERENCES

- Alessandrini, M., Micarelli, A., Viziano, A., Pavone, I., Costantini, G., Casali, D., Paolizzo, F., Saggio, G. (2017). Body-worn triaxial accelerometer coherence and reliability related to static posturography in unilateral vestibular failure. *Acta Otorhinolaryngologica Italica*, 37(3).
- Asci, F., Costantini, G., Di Leo, P., Zampogna, A., Ruoppolo, G., Berardelli, A., Saggio, G., & Suppa, A. (2020). Machine-Learning Analysis of Voice Samples Recorded through Smartphones: The Combined Effect of Ageing and Gender. *Sensors*, 20(18), 5022.
- Austermann, A., Esau, N., Kleinjohann, L., Kleinjohann, B. (2005). Prosody based emotion recognition for MEXI. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1138–1144.
- Borchert, M., Dusterhoft, A. (2005). Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. *IEEE Proceedings of NLPKE*, 147–151.
- Busso, C., Bulut, M., Narayanan, S. S., (2012). Toward effective automatic recognition systems of emotion in speech. In: J. Gratch - S. Marsella (Eds.), *Social emotions in nature and artifact: emotions in human and human-computer interaction*, 110–127.
- Chateau, N., Maffiolo, V., Blouin, C., (2004). Analysis of emotional speech in voice mail messages: The influence of speakers' gender. *International Conference on Speech and Language Processing (ICSLP)*.
- Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T., Richard, G., Vasilescu, I., Devillers, L., Ehrette, T., Richard, G., (2006a). Fear-type emotions of the safe corpus: annotation issues. *Proc. 5th Int. Conf. on Language Resources and Evaluation (LREC)*.
- Clavel, C., Vasilescu, I., Richard, G., Devillers, L., (2006b). Voiced and Unvoiced content of fear-type emotions in the SAFE Corpus. *Speech Prosody, Dresden*—to be published.
- Costantini, G., Todisco, M., Perfetti, R., Basili, R., & Casali, D. (2010a). SVM Based Transcription System with Short-Term Memory Oriented to Polyphonic Piano Music. *Mediterranean Electrotechnical Conference (MELECON) 201*
- Costantini, Giovanni, Casali, D., & Todisco, M. (2010b). An SVM based classification method for EEG signals. *Proceedings of the 14th WSEAS international conference on Circuits*, 107–109.
- Costantini, G., Iaderola, I., Paoloni, A., Todisco, M. (2014). EMOVO Corpus: An Italian emotional speech database. *LREC*.
- Cowie, R., Douglas-Cowie, E., Tsatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G., (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18 (1), 32–80.
- Cowie, R., Cornelius, R. R., (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40 (1–2).
- Cowie, R., Douglas-Cowie, E., Cox, C., (2005). Beyond emotion archetypes: databases for emotion modelling using neural networks. *Neural Networks*, 18, 371–388.
- Cullen, C., Vaughan, B., Kousidis, S., (2008). Emotional speech corpus construction, annotation and distribution. *Proceedings of the 6th International Conference on Language Resources and Evaluation*.
- Ekman, P., (1972). Universals and cultural differences in facial expressions of emotion. J. Cole (Ed.), *Nebraska Symposium on Motivation*, 19, Lincoln University of Nebraska Press, 207–282.
- Ekman, P., (1977). Biological and cultural contributions to body and facial movement. J. Blacking (Ed.), *The anthropology of the body*, Academic Press, London, 39–84.
- Ekman, P., (1984). Expression and the nature of emotion. *Approaches to Emotion*, 1–25.
- Ekman, P., Friesen, W. V., (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17 (2), 124–129.
- El Ayadi, M., Kamel, M. S., Karray, F., (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44.
- Eyben, F., Wenginger, F., Gross, F., Schuller, B., (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. *Proceedings of the 21st ACM International Conference on Multimedia*, Barcelona, Spain, 835–838.
- Fayyad, U. and K. Irani (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *IJCAI*.
- France, D., Shiavi, R., Silverman, S., Silverman, M., Wilkes, D., (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47 (7).
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, 11 (1).
- Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2.
- Jiang, D. N., Cai, L. H., (2004). Speech emotion classification with the combination of statistic features and temporal features. *IEEE International Conference on Multimedia and Expo (ICME)*, 3, 1967–1970.
- Joshi, D., Zalte, B. M., (2013). Speech emotion recognition: A review. *IOSR Journal of Electronics & Communication Engineering (JECE)*, 4 (4), 34–37.

- Krothapalli, R. S., Koolagudi, S. G., (2012). Emotion recognition using speech features, *Springer Science & Business Media*.
- Lee, C., Narayanan, S., (2003). Emotion recognition using a data-driven fuzzy inference system. *Eurospeech*, Geneva.
- Lee, C., Narayanan, S., (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13 (2), 293–303.
- Li, Y., Zhao, Y., (1998). Recognizing emotions in speech using short-term and long-term features. *International Conference on Speech and Language Processing (ICSLP)*, Sydney, Australia, 2255–2258.
- Lugger, M., Yang, B., (2007). The relevance of voice quality features in speaker independent emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4, 7–20.
- Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M., et al. (2014). Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowledge-Based Systems*, 63, 68–81.
- Nwe, T. L., Foo, S. W., De Silva, L. C., (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, 41 (4), 603–23.
- Ocquaye, E.N.N., Mao, Q., Xue, Y., Song, H. (2021). Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. *International Journal of Intelligent Systems*, 36 (1).
- Ortony, A., Turner, T. J., (1990). What's basic about basic emotions? *Psychological Review*, 97, 315–331.
- Oudeyer, P. Y., (2003). The production and recognition of emotions in speech: Features and algorithms. *International Journal of Human-Computer Studies*, 59 (1–2).
- Parada-Cabaleiro E., Costantini G., Batliner A., Baird A., Schuller B. (2018a). Categorical vs Dimensional Perception of Italian Emotional Speech. *INTERSPEECH 2018*
- Parada-Cabaleiro, E., Schmitt, M., Batliner, A., Hantke, S., Costantini, G., Scherer, K., Schuller, B.W. (2018b). Identifying emotions in opera singing: Implications of adverse acoustic conditions, *ISMIR 2018*.
- Parada-Cabaleiro, E., Costantini, G., Batliner, A., Schmitt, M., Schuller, B.W. (2020). DEMoS: an Italian emotional speech corpus: Elicitation methods, machine learning, and perception, *Language Resources and Evaluation*, 54(2), 341-383
- Saggio, G., & Costantini, G. (2020). Worldwide Healthy Adult Voice Baseline Parameters: A Comprehensive Review. *Journal of Voice*
- Saggio, G., Giannini, F., Todisco, M., & Costantini, G. (2011). *A data glove based sensor interface to expressively control musical processes* (pag. 195). <https://doi.org/10.1109/IWASI.2011.6004715>
- Scherer, K. R., (1986). Voice, stress, and emotion. M. H. Appley, R. Trumbull (Eds.), *Dynamics of stress*, Plenum, New York, 159–181.
- Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G., (2010). Cross-corpus acoustic emotion recognition: variances and strategies. *IEEE Transaction on Affective Computing*, 1 (2).
- Suppa, Antonio, Asci, F., Saggio, G., Marsili, L., Casali, D., Zarezadeh, Z., Ruoppolo, G., Berardelli, A., & Costantini, G. (2020). Voice analysis in adductor spasmodic dysphonia: Objective diagnosis and response to botulinum toxin. *Parkinsonism & Related Disorders*, 73, 23–30.
- Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., (2007). Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. *Affective Computing and Intelligent Interaction (ACII)*, 139–147.
- Wu, S., Falk, T. H., Chan, W. Y., (2009). Automatic recognition of speech emotion using long-term spectro-temporal features. *IEEE the 16th International Conference on Digital Signal Processing*, pp. 1–6.