Noname manuscript No.
(will be inserted by the editor)

DEMoS – An Italian Emotional Speech Corpus

Elicitation methods, machine learning, and perception

Emilia Parada-Cabaleiro · Giovanni Costantini · Anton Batliner · Maximilian Schmitt · Björn W. Schuller

Received: date / Accepted: date

Abstract We present DEMoS (Database of Elicited Mood in Speech), a new, large database with Italian emotional speech: 68 speakers, some 9 k speech samples. As Italian is under-represented in speech emotion research, for a comparison with the state-of-the-art, we model the ‘big 6 emotions’ and guilt. Besides making available this database for research, our contribution is three-fold: First, we employ a variety of Mood Induction Procedures (MIPs), whose combinations are especially tailored for specific emotions. Second, we use combinations of selection procedures such as an alexithymia test and self- and external assessment, obtaining 1,5 k (proto-) typical samples; these were used in a perception test (86 native Italian subjects, categorical identification and dimensional rating). Third, Machine Learning (ML) techniques—based on standardised brute-forced openSMILE ComParE features and Support Vector Machine (SVM) classifiers—were applied to assess how emotional typicality and sample size might impact machine learning efficiency. Our results are three-fold as well: First, we show that appropriate induction techniques ensure the collection of valid samples, whereas the type of self-assessment employed turned out not to be a meaningful measurement. Second, emotional typicality—which shows up in an acoustic analysis of prosodic main features—in contrast to sample size is not an essential feature for successfully training machine learning models. Third, the perceptual findings demonstrate that the confusion patterns mostly relate to cultural rules and to ambiguous emotions.

Keywords Emotional Speech · Italian Corpus · Elicitation · Prototype · Mood Induction Procedures · Machine Learning

Emilia Parada-Cabaleiro
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing
University of Augsburg, Germany
Tel.: +49 (0) 821 598 - 2908
E-mail: emilia.parada-cabaleiro@informatik.uni-augsburg.de

1 Introduction

Since emotional information is essential to build up congruent and efficient human relationships (Bennett, 1979), systems for human-machine interaction are often based on speech emotion recognition technology (El Ayadi et al., 2011). Yet, for the implementation of artificial intelligent systems with real life applications, reliable datasets mirroring everyday emotional speech are essential. Despite this, corpora of natural emotional speech are usually characterised by high level of background noise and their collection is limited by privacy issues. Furthermore, natural emotions do not always conform to the categorical model, which is predominant in the speech emotion literature, as e. g., the 'big six' (Ekman, 1984). On the other hand, datasets of acted emotional speech present high audio quality but are limited by the lack of authenticity. Corpora of emotional speech elicited by *Mood Induction Procedures* (MIPs) are a reliable compromise between acoustic quality, emotional taxonomy correspondence, and naturalness (Douglas-Cowie et al., 2000). Despite this, such corpora are mostly restricted to specific languages (Ververidis and Kotropoulos, 2006), and even though the combination of multiple MIPs (Martin, 1990; Westermann et al., 1996) has shown to be more effective (Westermann et al., 1996), data were mostly collected by applying MIPs individually (Klasmeyer et al., 2000).

We present `DEMoS` (Database of Elicited Mood in Speech), a corpus of induced emotional speech in Italian, a language underrepresented in speech emotion recognition (Costantini et al., 2014; Mencattini et al., 2014; Parada-Cabaleiro et al., 2018). `DEMoS` encompasses 9,365 emotional and 332 neutral samples produced by 68 native speakers (23 females, 45 males) in seven emotional states: the 'big six' anger, sadness, happiness, fear, surprise, disgust (Ekman, 1984), and the secondary emotion guilt. We employ these big six for a better comparison of these Italian data with the state of the art; to get more realistic productions, we do not employ acted speech but speech elicited by combinations of induction procedures. Guilt, according to previous research (Keltner, 1996) a secondary emotion, is also taken into account, in order to evaluate an ambiguous emotion typical of real life. Three elicitation methods are presented, made up by the combination of at least three MIPs, and considering six different MIPs in total. To select samples 'typical' of each emotion, evaluation strategies based on self- and external assessment are applied. We evaluate the reliability of the considered elicitation and selection methods; Machine Learning (ML) experiments are carried out to assess the extent to which emotional typicality and sample size influences their performance. In addition, the selected part of the corpus, which encompasses 1,564 samples produced by 59 speakers (21 females, 38 male), is evaluated by 86 native Italian listeners through a perceptual test based on the categorical and dimensional models of emotion[1].

[1] The corpus is available upon request through a personalised download link.

The rest of the manuscript is laid out as follows: related work is presented in Section 2; in Section 3 and 4, the induction program and the corpus are described; in Section 5, the selection procedures are presented; in Section 6, both elicitation methods and selection strategies are evaluated; Sections 7 and 8 introduce the ML approach and discuss the acoustic findings; Section 9 analyses the perceptual outcomes; finally, in Sections 10 and 11, the limitations and conclusions of our work, as well as future goals, are presented.

2 Related work

2.1 Acted vs natural emotional speech

Most of the available corpora of emotional speech have been collected by considering 'acted speech' (Bänziger et al., 2006), i. e., simulated emotional speech expressed by actors, and 'natural speech' (Devillers et al., 2005b), i. e., real emotional speech spontaneously expressed and collected in the wild. The two available Italian corpora are both acted: EMOVO (Costantini et al., 2014) and EmoFilm (Parada-Cabaleiro et al., 2018). When considering acted speech, high quality audio samples are collected, and specific emotional states can be chosen beforehand to be acted. Yet, the validity of acted emotions has been extensively criticised, since they are considered to be more exaggerated (Batliner et al., 2000; Douglas-Cowie et al., 2003). Indeed, even though semi-professional actors and naïve speakers have been taken into account in order to reduce the artificiality typically linked to professional actors' performance (El Ayadi et al., 2011), natural speech is still considered a better option. Nevertheless, natural speech often lacks in acoustic quality, due to real world environmental noise and speaker overlap. Moreover, naturally occurring emotions are often ambiguous and/or mixed (Devillers et al., 2005a)—this cannot easily be modelled with a simple categorical approach; other restrictions are based on the 'Observer's Paradox' (Labov, 1972)—speakers do not behave fully naturally when they are being observed—or on very specific scenarios such as broadcasting recordings (scripted reality shows or political discussions, just to mention a few). Yet, not informing a person of being recorded (in order to minimise such a condition) would violate their personal privacy. Although natural speech is the ultimate goal, for systematic investigations, properly induced speech is chosen as a compromise between naturalness, acoustic quality, and emotional taxonomy correspondence allowing the study of specific emotions chosen a priori.

2.2 Emotional speech elicited by MIPs

The application of MIPs allows to collect speech produced in specific emotional states (Martin, 1990; Westermann et al., 1996), offering at the same time a compromise between acoustic quality and naturalness. Yet, in order not to affect a subject's psychological stability, such mechanisms should be restricted to the elicitation of transitory emotions (Douglas-Cowie et al., 2011). In this regard, the utilisation of some MIPs, as e. g., hypnosis, the use of drugs, or sleep deprivation (Zou et al., 2011), is highly arguable (Martin, 1990). Furthermore, not all the MIPs show the same reliability, e. g., the emotions elicited by reading an emotionally connoted text (*Empathy MIP*) have been considered as similar to those produced by actors (Schröder, 2004). In addition, the accuracy of the MIPs will also depend on the intended emotional state to be induced, e. g., the elicitation of anger has shown to be particularly challenging (Gross and Levenson, 1995), since it relates to the frustration of specific individual expectations, thus varying considerably amongst different subjects. In this regard, the combination of several MIPs has shown to be more reliable (Westermann et al., 1996), due to the increment in the effectiveness given by the complementarity that might be created between different methods; e. g., listening to music may intensify the effect of reading an emotional text. Despite this, for recording corpora of elicited emotional speech available for research purpose, the utilisation of MIPs has been considered individually, rather than in combination (Klasmeyer et al., 2000; Douglas-Cowie et al., 2007).

Some of the most common MIPs employed for the elicitation of emotional speech are: *Autobiographical Recall MIP*, based on the recall of emotional personal memories (Amir et al., 2000); *Self-statement MIP* (Velten, 1968), based on the repetition of emotional sentences (Barkhuysen et al., 2010); *Empathy MIP*, based on the creation of an empathic reaction by reading text with an emotional content (Chiţu et al., 2008; Douglas-Cowie et al., 2003; Grichkovtsova et al., 2012; Iida et al., 2003; Sobin and Alpert, 1999); and *Social Feedback MIP*, based on a simulated social task such as the Wizard-of-Oz paradigm, specially successful in the collection of children's emotional speech (Batliner et al., 2004; Zhang et al., 2004) but used also with adults (Aubergé et al., 2003; Tato et al., 2002; Türk, 2001). A similar procedure to Social Feedback MIP is *Game Feedback MIP*, based on the elicitation of emotions by presenting cooperative (Cullen et al., 2006) and challenging tasks (Fernandez and Picard, 2003; Tolkmitt and Scherer, 1986), often based on manipulated feedbacks (Johnstone and Scherer, 1999; Johnstone et al., 2005; Truong et al., 2012). Finally, even though induced corpora of emotional speech have been created in a variety of languages, as e. g., English (Douglas-Cowie et al., 2003; Sobin and Alpert, 1999), Dutch (Chiţu et al., 2008), Japanese (Iida et al., 2003), French (Grichkovtsova et al., 2012; Aubergé et al., 2003), German (Barkhuysen et al., 2010), or Hebrew (Amir et al., 2000), Italian has never been considered so far.

3 Induction Program

Emotion induction was performed in one session per subject, lasting around 70 minutes. Before starting, all participants signed the consent agreement required for personal data collection and utilisation with research purposes[2]. In

[2] The consent agreement was designed by *Santa Lucia Foundation* (Research and Health Care Institute).

the induction program, six different MIPs were considered: Music MIP, Autobiographical Recall MIP, Film MIP, Picture MIP, Self-statement MIP, and Empathy MIP (cf. subsection 3.2). The emotions of the corpus were induced through an arousal-valence progression which takes into account two degrees of valence (positive and negative), and three degrees of arousal (low, medium, and high). The progression was created by the specific stimulus selected in the MIPs considered for the elicitation of each emotion, from positive to negative valence and from low to high and again to low arousal: happiness, surprise, fear, anger, disgust, guilt, and sadness, i. e., happiness and surprise (positive and low arousal), fear (negative and medium arousal), anger and disgust (negative and high arousal), guilt (negative and medium arousal), sadness (negative and low arousal). Such a progression relates to the dimensional value encoded in the texts of the Empathy MIP. Note that the use of MIPs should never alter the emotional stability of the participants in any extreme way (Douglas-Cowie et al., 2011); thus, high aroused elicited emotions are not as intense as high aroused real emotions. Three 'Elicitation Methods' (cf. subsection 3.3) were designed by combining at least three different MIPs, chosen as the most suitable for the induction of each specific emotion. To make the subjects familiar with the elicitation procedure, as well as to record neutral speech, induction sessions began with the subject reading a text of neutral content aloud (Ciceri and Anolli, 2000); then, each emotion was induced. The influence of the experimenter's presence was minimised by leading the induction sessions through a computer-based interface (operated by the participants themselves), which presented each elicitation method one after the other. Surrounding distractions were minimised by performing the sessions in a semi-dark and quiet environment, with the workstation used for recording being hidden to the participants. To select the samples more representative of each emotion, self- and external assessment were considered (cf. subsection 3.4).

3.1 MIPs & Stimulus description

(i) *Music MIP* (emotional elicitation by listening to music): Each song was chosen considering harmonic and rhythmic aspects (Husain et al., 2002) according to the emotional content encoded in the texts used for *Empathy MIP* (cf. below, paragraph vi). Major key was considered for positive emotions (happiness and surprise), minor key for negative (guilt and sadness), static rhythm for low aroused emotions (happiness, surprise, sadness), and 'ostinato', i. e., repetitive rhythmic patterns, for an emotion with medium arousal (guilt). *Spiegel im spiegel* (A. Pärt) was chosen for the induction of happiness and surprise (both positive and low aroused), *To the edge of the earth* (M. Nyman) for the induction of guilt (negative and mildly aroused), and *Sotto vento* (L. Einaudi) for the induction of sadness (negative and low aroused). Since *Music MIP* was mainly taken into account to create an acoustic surrounding background which would encourage the effectiveness of *Autobiographical Recall MIP*, any linguistic bias (Singhi and Brown, 2014) was avoided by considering only instrumental music.

(ii) *Autobiographical Recall MIP* (emotional elicitation by recalling personal memories): Four short 'passages' of implicit guidelines—the intended emotions were not expressively indicated—were written in order to be performed as spoken instructions for leading the subjects into their own memories, thereby eliciting the intended emotional states. Such instructions were designed in a figurative language, based on the use of metaphors and sensorial concepts such as *warm* (for positive concepts) or *empty* (for negative), which follows a practice typical of the *Guided Affective Imagery* (Bonny, 2002; Utay and Miller, 2006), a psychotherapeutic intervention based on the patients evocation of mental images through spoken instructions given by the practitioner. For instance, the introductory sentence of the instructions which intended to lead the participants into a happy emotional state, was the following: "Let the music lead your thoughts, towards that pleasant sensation". A semi-professional speaker (a student of voice acting—dubbing), recited the texts being recorded; subsequently, each pre-recorded sample of the spoken guidelines was mixed with each song.

(iii) *Film MIP* (emotional elicitation by watching movies): The topics of public humiliation (a scene of bulling) and danger of bodily integrity (a killer chase scene) were chosen, according to previous work (Gross and Levenson, 1995; Cavanagh et al., 2011), to elicit anger and fear, respectively. The two scenes were extracted from the films *Ben X* (N. Balthazar), for the induction of anger, and *High tension* (A. Aja), for the induction of fear. To encourage the subjects' empathic identification with the victim (i. e., the protagonist of each scene), the victim had a similar age to our subjects in both cases; each scene lasted around five minutes—a length suitable for elicitation purposes (Cavanagh et al., 2011).

(iv) *Picture MIP* (emotional elicitation by watching pictures): Images that 'typically' affect human sensibility, such as insects or blood, were selected to induce disgust. Natural science images, e. g., bugs and spiders, were taken from the *Geneva Affective Picture Database*—GAPED (Dan-Glauser and Scherer, 2011). Human physiology images, e. g., internal organs, were taken from the freely available image database *Pathology Education Informational Resource* (PEIR)[3].

(v) *Self-statement MIP* (emotional elicitation by pronouncing emotionally connoted sentences): Seven emotional sentences (one for each emotional state) and one neutral were considered. The sentences to elicit happiness and sadness were chosen from those proposed by Velten (1968), while the remaining ones were similarly generated. The originals in Italian and their English translations are as follows: *Sento che oggi sarà la mia giornata*—'This is just one of those days when I'm ready to go!' (happiness); *Non me ne va bene una*—'I have too many bad things in my life' (sadness); *Lasciami in pace! Ti odio!*—'Leave me alone! I hate you!' (anger); *Che schifo! Non voglio più vedere*—'It is disgusting! I do not want to look any more!' (disgust); *Veramente? Non me lo aspettavo proprio!*—'Really? I did not expect it!' (surprise); *Cosa volete farmi?*

[3] `http://peir.path.uab.edu/library/`

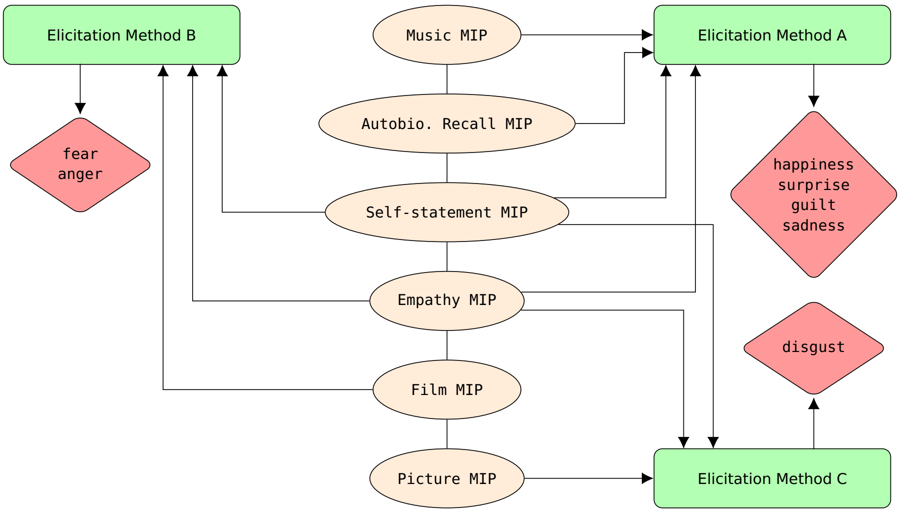


Fig. 1 Flowchart to summarise the Induction Program. The relationship between the six MIPs (Music MIP, Autobiographical Recall MIP, Self-statement MIP, Empathy MIP, Film MIP, Picture MIP) and the three Elicitation Methods (A, B, C) is indicated. The seven emotions were induced in the following order: happiness, surprise, fear, anger, disgust, guilt, and sadness; Empathy MIP and Self-statement MIP were presented as the last MIPs for each Elicitation Method.

No, no!—'What do you want to do to me? No, no!' (fear); *È tutta colpa mia, se li avessi dato retta*—'Everything is my fault, if I would have listened' (guilt); *Parigi è la capitale della Francia*—'Paris is the capital of France' (neutral).

(vi) *Empathy MIP* (emotional elicitation by reading emotionally connoted texts): Five texts, expressively written for the induction of emotional speech (Ciceri and Anolli, 2000) were considered to elicit surprise, fear, anger, guilt, and sadness. For the induction of happiness, a text in line with the previous one was written by the experimenters whereas for the elicitation of disgust, a text was taken from the novel *Perfume: The Story of a Murderer* (P. Süskind). The texts start with an introduction in third person that initially leads the participant into the emotional induction.

For a review of the presented and other Mood Induction Procedures see (Martin, 1990; Gerrards-Hesse et al., 1994; Westermann et al., 1996).

3.2 Elicitation methods

Self-statement MIP and *Empathy MIP* were considered in all three elicitation methods in order to consistently collect emotional speech. The other MIPs were chosen, according to previous research, as those more appropriate for the induction of specific emotions (cf. Figure 1). Considering the single MIPs' durations together, each elicitation method lasted around seven minutes in total, a length optimal for the induction of emotional states (Västfjäll, 2001)—a shorter one would be insufficient to reach the emotional climax, a longer one might impair its maintenance.

Elicitation Method A: made up by the combination of *Music MIP*, *Autobiographic Recall MIP*, *Self-statement MIP*, and *Empathy MIP*; considered for the elicitation of happiness, surprise, guilt, and sadness. Controversial outcomes have been presented on whether the emotions elicited by listening to music would depend on the listeners' musical preferences or not (McCraty et al., 1998; Västfjäll, 2001); still, it has been shown that music can create a surrounding atmosphere which in any case might increase the effectiveness of other MIPs (Mayer et al., 1995). With this as a foundation, we considered *Music MIP* together with *Autobiographic Recall MIP*, a technique that has shown to be effective in the induction of both negative (Van der Does, 2002) and positive (Konečni et al., 2008) emotional states. As a way to obtain reading aloud, immediately after the *Music & Autobiographical Recall MIPs*, we considered two MIPs typically used in eliciting emotional speech, i. e., *Self-statement MIP* (Barkhuysen et al., 2010) and *Empathy MIP* (Grichkovtsova et al., 2012).

Elicitation Method B: made up by the combination of *Film MIP*, *Self-statement MIP*, and *Empathy MIP*; considered for the elicitation of anger and fear. Even though the elicitation through films might not be the most effective method to induce anger and fear (Gross and Levenson, 1995), we discarded more efficient MIPs, such as the encouragement of conflictive situations or the use of hypnosis, as ethically arguable (Martin, 1990). Again, *Self-statement MIP* and *Empathy MIP* were concatenated immediately after *Film MIP* in order to obtain read aloud utterances.

Elicitation Method C: made up by the combination of *Picture MIP*, *Self-statement MIP*, and *Empathy MIP*; considered for the elicitation of disgust. Previous research has shown that the induction through pictures is a successful method to induce disgust (Schienle et al., 2005). According to this, images of several typologies, including natural science (spiders, bugs, and insects in general) and human physiology (blood, skin illness, and internal organs) were chosen to affect the sensibility of a variety of subjects. Again, *Self-statement MIP* and *Empathy MIP* were considered.

4 Data collection

We recorded 68 subjects (23 females, 45 males). The corpus—comprising 9,365 samples in seven emotional states (cf. Table 1) and 332 neutral samples—was recorded in PCM-wave mono format and 48kHz/16-bit sampling rate/bit depth. Subsequently, the emotional speech was manually segmented into samples (mean length 2.9 sec, std 1.1 sec). The manual segmentation was performed in syntactic chunks (*S-Chunks*), i. e., by considering syntax and punctuation; yet, when the participants' prosody deviates, e. g., by phrasing the sentences in an unnatural way, a subject's individual prosody was prioritised in order to avoid an unnatural segmentation that might lead to sudden cuts between words; thus, the resulting prosodic chunks (*P-Chunks*) do not always correspond to the *S-Chunks*. For the data collection, two workstations

Table 1 Distribution of the emotional samples produced by the 68 participants in the seven emotions. *Chunks* relates to the samples extracted by the segmentation of the texts used in the *Empathy MIP*, and can be of two types: *S-Chunks* (Syntactic Chunks) gives the number of utterances grammatically defined in the texts; *P-Chunks* (Prosodic Chunks) gives the number of samples segmented by prioritising the prosody of the subjects. *Sentences* indicates the pre-defined utterances (neutral and emotional) which relate to the *Self-statement MIP*. *Utterances* relate to grammatically defined linguistic units, which might—but not necessarily—coincide with the segmented samples; they can be conceived as types. For each emotion, the number of tokens (collected samples) per emotion are given for females and males separately (*P-Chunks* and *Sentences*) and combined (#), as well as the sum of both ($\sum$).

Emotion		Utterances (types)		Female (tokens)			Male (tokens)			$\sum$
Category	Label	*S-Chunks*	*Sentences*	*P-Chunks*	*Sentences*	#	*P-Chunks*	*Sentences*	#	
Anger	ANG	20	2	461	55	516	898	63	961	1,477
Sadness	SAD	19	2	437	95	532	837	161	998	1,530
Happiness	HAP	17	2	398	126	524	747	124	871	1,395
Fear	FEA	11	2	328	87	415	672	69	741	1,156
Disgust	DIS	17	2	493	103	596	963	119	1,082	1,678
Guilt	GUI	9	2	318	82	400	629	100	729	1,129
Surprise	SUR	10	2	310	39	349	616	35	651	1,000
#		103	14	2,745	587	3,332	5,362	671	6,033	9,365

(one for the recordings and another for the induction program), one hypercardioid close-talk microphone, headphones, and a professional sound card were utilised. To avoid influencing the participants' natural speech production, in each induction session, the subjects were explicitly instructed to wear headphones only for the procedures that required audio, e. g., *Music MIP* and *Film MIP*; this was indicated via the computer-based interface.

4.1 Corpus description

The 68 participants (23 females, 45 males) were all students from an engineering faculty (mean age 23.7 years, std 4.3 years), who obtained academic credits for their voluntary participation. The corpus encompasses 9,697 samples: 3,444 produced by females (3,332 with an emotional content and 112 neutral); 6,253 produced by males (6,033 with an emotional content and 220 neutral). In Table 1, the distribution of the samples is given. As mentioned above, due to prosodic variations between different speakers, the number of tokens per *P-Chunks* (cf. columns 'Female' and 'Male' in Table 1), i. e., the samples based on segmenting the texts according to the speakers' prosody, do not always coincide with that expected by multiplying the number of subjects by the number of tokens per *S-Chunks* (cf. the section 'Utterances' in Table 1), i. e., the number of utterances grammatically defined by the syntax and punctuation of each text. Similarly, since some of the participants repeated more than once the pre-defined utterances, the number of tokens per *Sentences*, i. e., the samples based on *Self-statement MIP* (cf. *Sentences* column for 'Female' and 'Male' in Table 1), is not the same for the different emotions.[4] Note that

[4] To give an example: For anger produced by females, we would expect 460 samples (20 S-Chunks x 23 participants) based on *Empathy MIP* and 46 samples (2 Sentences x 23 participants) based on *Self-statement MIP*. Yet, we end up with 461 samples based on

from now on, the emotional labels indicated in Table 1 will be used throughout the article.

5 Corpus Selection

To guarantee the reliability of the induction mechanisms, in emotional elicitation studies, it is common to evaluate how the participants perceive their own emotions (Bradley and Lang, 2000; Mikula et al., 1998). Nevertheless, factors such as subjective cognitive appraisal or self-defense mechanisms might influence the responses of individual subjects who may not be able to fully identify their own emotions (Scherer, 2013). Differently, in emotional speech research, the selection of samples that faithfully represent an emotion, i. e., prototypes[5] (Batliner et al., 2005; Russell, 1991), is mainly made by annotators, often experts (Burkhardt et al., 2005). Yet, such an external evaluation cannot guarantee the collection of a reliable 'ground truth' (Schuller et al., 2011), since the subjectivity inherent of perception inevitably biases listeners' responses. To get a reliable 'gold standard', samples of the corpus more representative of each emotion were selected by considering both self-assessment (performed by the participants) and external assessment (performed by experts). The participants' and experts' ability to reliably identify their own and others' emotions was evaluated by an alexithymia test (Roedema and Simons, 1999). The samples formulated in third person at the beginning of the texts used in the *Empathy MIP*—whose goal was to initially lead the participant into the induction—were excluded, since they might be less likely to express emotion. Also those produced by a participant who had acting experience were discarded, since they might be more artificial.

We would like to emphasise that the goal of the selection procedure is to identify a subset of 'prototypical' samples for a further evaluation of the role of sample size and typicality in ML approaches (cf. Section 7). Given the difficulty to collect 'prototypical' emotions in a non-acted setting—non-acted emotions are often characterised by a certain degree of ambiguity, as shown by mixed motions (Mower et al., 2009)—the understanding of whether typicality or sample size is more relevant in the performance of ML systems is a crucial topic. Still, this selection does not imply that the non selected samples should be discarded, as indeed will be shown in Section 7, given the importance of sample size.

5.1 Selection criteria

(i) The alexithymia test is an instrument to assess a subject's ability to correctly identify and describe their own and others' emotional states. To guaran-

Empathy MIP (cf. *P-Chunks* in Table 1) and 55 samples based on *Self-statement MIP* (cf. *Sentences* in Table 1). *P-Chunks* can integrate across syntactic boundaries but more often, they partition *S-Chunks* into smaller units.

[5] A'prototype' is a central, natural category (Rosch, 1973) with a unique representation, not composed by a combination of simpler ones.

tee that both, participants and experts, had an emotional awareness adequate to correctly identify emotions, the *Scala Alessitimica Romana* (SAR) was performed. The SAR (Baiocco et al., 2005) evaluates five areas: somatic expression of emotion, emotional identification, emotional communication, emotional thought, and empathy. It is structured in 27 statements expressed both positively and negatively (e. g., 'When I feel an emotion, I understand why'), and each must be rated with one of the following options: never, sometimes, often, or always. Even though alexithymia tests are a common practice in psychological studies, they have, to the best of our knowledge, never been considered in affective computing research. Still, since alexithymic subjects might not display an accurate perception of own and others' emotions (Roedema and Simons, 1999), to guarantee the reliability of self- and external assessment, a measurement strategy like this should be employed. The three experts successfully passed the alexithymia test SAR whereas eight participant did not (cf. subsection 6.4).

(ii) Self-assessment, i. e., the evaluation of the emotional states produced by the participants themselves, was performed to evaluate the effectiveness of the induction program by verifying that the emotions self-perceived by the subjects coincide with those intended to be elicited. The *Circumflex Model* of emotions (Russell, 1980), where each emotion is identified with a unique position in a bi-dimensional space—arousal (activation) and valence (pleasure)—was employed as a reference. This model is typically used in external assessment (Cowie et al., 2000; Schröder, 2004), whereas the categorical model is more common for self-assessment procedures (Scherer, 2013). Despite this, it is not clear yet which of both models (the dimensional or the categorical) might be more suitable to evaluate emotional self-perception, as shown by a succinct literature that proposes arguments against and in favor of each of them (Philippot, 1993; Scherer, 2005). In this regard, self-assessment methods such as the *Geneva Emotion Wheel* – GEW (Scherer et al., 2013) integrate both models in a unique procedure.

Considering this, an emotional diagram comprising the arousal dimension and categorical labels was designed and presented to each participant immediately after each elicitation method (cf. Figure 2). In addition to some of the emotional labels indicated in the circumflex model (Russell, 1980), the emotional categories disgust, surprise, and guilt, since considered in the elicitation but not present in Russel's circumflex, were also included in the diagram, placed in a suitable position of the dimensional space—e. g., surprise was placed adjacent to fear (Schlosberg, 1954). All in all, the diagram presents the emotions intended to be elicited (surprise, happiness, sadness, disgust, guilt, anger, and fear), as well as other, not induced emotional categories typical of the circumflex model (excitement, elation, gladness, serenity, calm, sleepiness, tiredness, boredom, and anxiety) that were considered as so called *distractor labels* (Murray and Arnott, 1995), i. e., emotional labels displayed to encourage a task based on identification (by increasing the number of possible responses) rather than on discrimination, thus ensuring accurate responses. When the induction procedure for eliciting each emotion was finished, each participant

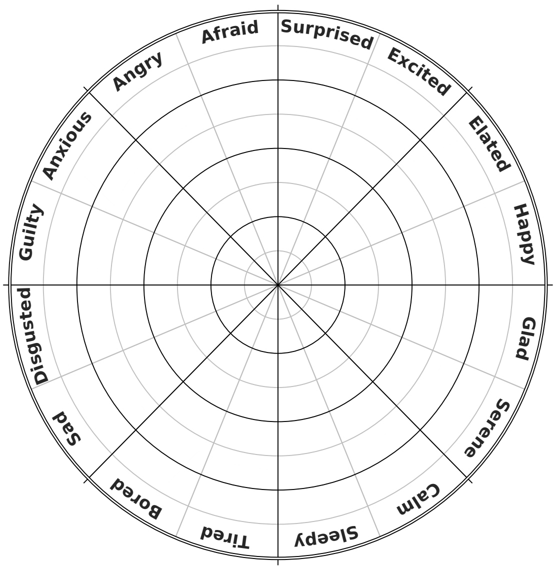


Fig. 2 Diagram for the self-assessment. Both the emotions intended to be elicited (surprised, happy, sad, disgusted, guilty, angry, afraid) and the distractors, i. e., emotional labels displayed to ensure accurate responses but not referring to the induced emotions (excited, elated, glad, serene, calm, sleepy, tired, bored, anxious), are indicated; the closer to the center, the lower the arousal. The emotional labels—possible answers to the questions: 'How did you feel during the activity?'—were given as adjectives.

performed the self-assessment, once for each emotion. The participants were invited to mark the emotion 'self-perceived' (i. e., the emotion felt during the induction procedure) with an 'X' at a unique position in the diagram according to the arousal level (the closer to the center of the circle, the lower the activation) and according to the categorical label (encoded in the circumflex perimeter). We considered the responses encoded within the categorical section which related to the emotion intended to be elicited as valid; otherwise, they were excluded. An exception to this were the emotions elation and gladness, which were also accepted as a self-perception of happiness since on one side, the three emotion categories (i. e., elation, happiness, and gladness) can be interpreted as a different arousal representation of the same emotion; on the other side, considering that seven emotional categories with a positive valence were distractors, this would penalise too much the participants' performance in comparison to the other emotions. Even though the presented selection criteria may considerably reduce the size of the collected samples, our intention with this was to guarantee a selection of samples that truly represent genuine emotions.

(iii) External assessment, i. e., the evaluation of the emotional states made by another subject, was carried out by three experts in the field of affective computing. *Empathy MIP* and *Self-statement MIP* presuppose suitable reading aloud skill of the listeners; thus, through the external assessment not only samples lacking in emotional expressivity, but also those void of reading flu-

Table 2 Distribution of non-selected samples per emotion is given for females and males separately (#) and combined ($\sum$). Samples discarded by the alexithymia test (SAR), those produced by the actress (Act), thus excluded as unnatural, those expressed in third person (3^{rd}p), thus unlikely to express emotion, and those rejected by self- (self-A) and external assessment (ext-A), are indicated.

Label	Female						Male					$\sum$
	SAR	Act	3^{rd}p	self-A	ext-A	#	SAR	3^{rd}p	self-A	ext-A	#	
ANG	24	21	147	163	88	443	143	266	237	142	788	1,231
SAD	25	20	62	158	185	450	151	114	210	183	658	1,108
HAP	25	20	42	192	157	436	104	76	384	228	792	1,228
FEA	19	16	105	104	110	354	105	190	238	92	625	979
DIS	28	24	0	178	295	525	155	0	647	211	1,013	1,538
GUI	19	17	0	139	145	320	111	0	253	236	600	920
SUR	16	14	84	109	73	296	96	152	116	137	501	797
#	156	132	440	1043	1053	2,824	865	798	2,085	1,229	4,977	7,801

ency (which might be 'unnatural'), were rejected. Samples discarded by two out of the three experts were excluded in the selected version of the corpus. In Table 2, the distribution of the samples excluded after the selection process is given.

5.2 Selected corpus description

Since the successful performance of the 'alexithymia test' is a requirement for the reliability of self- and external assessment, this was the first procedure to be performed; thus, only the participants who successfully passed it were considered for the self- and external assessment. The self-assessment was performed by all the participants and the positive results obtained from this, i. e., productions which related to the emotions identified by the participants as those intended to be elicited, were taken into account in the external assessment. Out of the 68 participants, nine were excluded from the selected version of the corpus: eight did not successfully pass the 'alexithymia test' (one female and seven male, who produced 1,021 samples in total); one had professional experience as an actress (cf. Table 2); thus, her speech might be considered as more artificial (132 samples in total). In addition, the samples of each emotion for which the participants did not successfully pass the self-assessment (3,128 in total) were excluded, as well as those produced in third person in the *Empathy MIP* (1,238 in total). Finally, the samples considered by the external assessment as lacking in reading fluency and void of emotional expressiveness (2,282 in total) were also discarded. After performing the alexithymia test, both assessments, and rejecting the instances expressed in third person, a total of 7,801 samples were discarded. In Table 3, the distribution of the samples after the selection process is given, i. e., 1,564 samples produced by 59 speakers: 508 by 21 female and 1,056 by 38 male. Even if the drop-out is big, we expect the remaining items to be good examples (prototypes) of the considered emotions.

Table 3 Distribution of the selected samples produced by 59 participants in the seven emotions. *Chunks* relate to the *Empathy MIP*; *Sentences* to the *Self-statement MIP* (cf. caption of Table 1). The number of tokens (collected samples) per emotion are given for females and males individually (*P-Chunks*—Prosodic Chunks—and *Sentences*) and combined (#), as well as the sum of both ($\sum$) and the difference (*diff*) between the full and the selected corpus (gray cells). Notice that for $\sum$, only the selected cases are considered; in the column *S-Chunks* (Syntactic Chunks), the utterances expressed in third person are not included.

Label	Utterances (types)		Female (tokens)				Male (tokens)				$\sum$
	S-Chunks	*Sentences*	*P-Chunks*	*Sentences*	#	*diff*	*P-Chunks*	*Sentences*	#	*diff*	
ANG	13	2	66	7	73	443	161	12	173	788	246
SAD	16	2	59	23	82	450	278	62	340	658	422
HAP	15	2	69	19	88	436	54	25	79	792	167
FEA	6	2	47	14	61	354	83	33	116	625	177
DIS	17	2	42	29	71	525	32	37	69	1,013	140
GUI	9	2	53	27	80	320	84	45	129	600	209
SUR	6	2	44	9	53	296	131	19	150	501	203
#	82	14	380	128	508	2,824	823	233	1,056	4,977	1,564

6 ML Approach: Corpus Evaluation

In order to evaluate the effectiveness of the elicitation methods (cf. subsection 3.3), we employed ML techniques to classify the collected samples in an automatic way, considering the emotional speech produced by each speaker individually, i. e., experiments were carried out 68 times, and each time only the samples produced by one participant were considered. Unweighed Average Recall (UAR), i. e., the average of the recalls per class (emotion), was used as a measure of comparison for the classification performance. Using a rank-based approach, the 68 participants were ordered by the UAR; then, the recall for each emotion was compared across speakers in order to evaluate the classification performance for each emotion. We assume that a higher recall per class relates to emotional speech that is more representative of each emotion, i. e., prototypical (Batliner et al., 2005; Russell, 1991), which would present high intra-class homogeneity (samples within each class would be similar to each other) and high inter-class diversity (samples of each class would be dissimilar to those of another class). Emotional speech 'typicality', i. e., the emotional speech characteristic of each emotion, is considered as an indicator that the elicitation methods were effective. Yet, lower recall per class would not necessarily mean that the elicitation methods were not efficient, but the emotional state might be particularly ambiguous—also known as mixed motions (Mower et al., 2009), thus, not 'typical' of an emotion, but related to more than one. In order to guarantee a comparable evaluation across speakers, we propose a novel but promising automatic method, as a plausible alternative to the subjectivity, time constraints, and effort linked to human annotation of big datasets.

6.1 Methodology

In the ML experiments, we employ the `ComParE` feature set (Schuller et al., 2013), comprising 6,373 acoustic features (Eyben et al., 2015) computed by applying statistical functions to 65 Low-Level Descriptors (LLDs), extracted by the openSMILE feature extractor (Eyben et al., 2010), and a Support Vector Machine (SVM) classifier with a linear kernel from the open-source toolkit Liblinear (Fan et al., 2008). Even though Deep Neural Networks (DNNs) are prevalent nowadays for ML tasks, in affective computing research their performance is not yet superior to rather classic ML procedures such as SVMs. This can be seen in the series of Interspeech Challenges, from Schuller et al. (2013) to Schuller et al. (2018) and might simply be due to the sparse data problem: DNNs need very large databases; such databases do not exist for emotion modelling. Therefore, we chose an SVM classifier as it has only few hyperparameters, compared to recent deep learning approaches, and thus gives more reliable results in terms of robustness during training; our approach is more focused on understanding and less on optimising classification. The experiments were carried out by dividing the samples into three folds, i. e., a training set (used for model training), a development set (used to evaluate training hyperparameters), and a test set (used for final evaluation). The split was done in such a way that the samples per emotion for each speaker are balanced over the three folds. Since the goal of this procedure is to compare the performance achieved for each subject individually, a speaker dependent classification was performed, i. e., the models are adapted to the speakers present in the corpus and the evaluation shows how good the emotions can be recognised again for each speaker.

In the training phase, the SVM model was learnt using the training set. Subsequently, in the development phase, the development set was considered as a ‘preliminary’ test set in which the complexity (the most important SVM hyperparameter) was optimised by considering 30 different levels, from 2^{30} to 2^{0}. The complexity level which yielded the maximum UAR on the development set was considered to set up the SVM for the final training phase in which the samples of the training and the development sets were combined and used as a final training set. The final test was then performed on the test set. The experiments were done in a cross-validation setup, considering all six possible permutations of the folds, i. e., considering each fold as either training, development, or test set; the results were averaged, reporting the mean UAR and the average of the recall per class over all permutations. Thus, we conducted 68 experiments, considering each time for training, development, and test the samples produced by only one participant; in the following, we give ‘mean, std’ of samples for these constellations: speakers (142.6, 22.7); anger (21.7, 2.2); sadness (22.5, 3.6); happiness (20.5, 5.1); fear (17, 3.0); disgust (24.7, 4.6); guilt (16.6, 2.8); surprise (14.7, 1.4); neutrality (4.9, 3.1). The dimensionality of the feature space (6,373) is much higher than the number of instances ($<$ 143 for all emotions), which might lead to overfitting. However, we avoid this by using separate development and test partitions, tuning the classifier only on the development partition and evaluating the performance on the unseen test samples. From our experience (Schuller et al., 2013), the ComParE feature set in combination with a linear SVM classifier is generally robust against overfitting.

Table 4 Full corpus automatic classification. Mean (ALL_{mean}) and std (ALL_{std}) are given for the results of the 68 participants considering: recall per class for each emotion (cf. caption of Table 1) and neutrality (neu); Unweighted Average Recall (UAR) for the seven classes; frequency of evaluated cases (#). Absolute results, mean, and mean differences (diff_{mean}) for the participants in the first (UAR > 60%) and last (UAR < 40%) positions of the classification rank are given considering: recall per class, UAR, and frequency of cases (#). Values higher than 50% are highlighted in bold. Experiments were carried out individually for each participant.

%		ANG	SAD	HAP	FEA	DIS	GUI	SUR	NEU	UAR	#
ALL_{mean}		**57.1**	**63.5**	46.9	42.8	**55.8**	45.1	41.2	33.3	47.9	142.6
ALL_{std}		16.6	15.1	17.8	15.2	16.3	20.9	18.0	25.7	8.3	22.7
First Rank Pos.	1	**88.0**	**85.7**	**71.0**	41.6	**52.0**	**78.3**	**89.1**	16.6	**65.3**	127
	2	47.9	**80.9**	**68.3**	42.8	43.3	**80.1**	**61.1**	83.3	**63.6**	164
	3	**76.7**	**79.1**	**60.7**	**61.1**	**67.5**	32.7	**60.0**	58.3	**62.0**	146
	4	**77.7**	**63.8**	43.8	43.3	**78.5**	41.6	**64.1**	83.3	**62.0**	117
	5	**78.8**	**56.8**	**61.1**	39.6	49.8	**84.1**	**57.9**	66.6	**61.8**	195
	6	**84.2**	48.2	**61.3**	44.4	**70.1**	42.7	40.0	100	**61.3**	149
mean		**75.6**	**69.1**	**61.0**	45.5	**60.2**	**59.9**	**62.0**	68.0	**62.7**	149.7
Last Rank Pos.	63	40.4	**68.6**	21.1	20.0	36.9	31.6	28.3	50.0	37.1	138
	64	20.6	**65.8**	40.0	**52.7**	**50.0**	23.3	43.3	0.0	36.9	129
	65	29.1	**65.7**	31.2	39.2	20.3	33.3	**59.1**	8.3	35.8	149
	66	34.9	**64.2**	32.9	15.0	27.0	45.0	15.0	33.3	33.4	128
	67	41.0	**72.6**	26.7	20.5	31.4	33.3	13.3	16.6	31.9	149
	68	33.7	**51.1**	16.6	23.3	49.6	3.3	30.8	16.6	28.1	121
mean		33.3	**64.7**	28.0	28.4	35.9	28.3	31.6	20.8	33.9	134.7
diff_{mean}		42.3	4.4	33.0	17.1	24.3	31.6	30.4	47.2	28.8	15

6.2 Evaluation of elicitation methods

From this speaker-dependent 8-class classification problem—the seven emotions and neutrality were considered as a recognition target—the emotions best classified were sadness (recall > 60%), anger, and disgust (recall > 55% each); cf. row ALL_{mean} in Table 4. This suggests that the elicitation methods A, B, and C (cf. subsection 3.3) were mostly successful in the induction of sadness, anger, and disgust, respectively. The classification of happiness, guilt, fear, and surprise achieved a lower recall, which might not only be due to the elicitation methods' inefficiency but also because such emotions are harder to induce. By evaluating the first and last positions of the classification rank, i. e., those with an UAR above 60% and below 40%, it is confirmed that the elicitation method A (*Music MIP + Autobiographic Recall MIP*) is adequate to induce sadness (Van der Does, 2002), as shown by a similar recall for the first and last rank position (cf. the low mean difference = 4.4% between both). Still, to some extent, this might also relate to sadness being an emotion characterised by low pitch, tone, and energy, thus, closer to an 'undefined' category, and therefore similarly classified in selected and non-selected samples. Differently, and confirming the inaccuracy of *Film MIP* to elicit fear (Gross and Levenson, 1995), the recall for this emotion was particularly low even for the first rank positions. The induction of anger, happiness, disgust, guilt, and surprise yielded substantially lower recall in the last rank positions, which shows that the elicitation of these emotions was successful for some participants but not for all. Furthermore, these emotions commonly display diverse representations which can compromise their recognition—unlike sadness, which has a more standardised expression. Disgust and surprise are ambiguous emotions (Ortony and Turner, 1990), guilt is a secondary emotion, and anger and happiness are typically represented by two arousal levels, i. e., cold anger vs hot anger and amusement vs elation.

6.3 Evaluation of selection strategies

In order to evaluate the effectiveness of the qualitative strategies of control (cf. subsection 3.4), the results of these tests will be evaluated by focusing on the upper and lower half of the classification ranks (i. e., the participants in the 34 first and the 34 last positions). The efficiency of the induction methods seems not to relate to the speakers' gender, since a balanced distribution of males and females are included in both halves of the rank (12 females in the upper, 11 females in the lower; 22 males in the upper, 23 males in the lower). By performing a one-way ANOVA, the difference between the means for the UAR achieved by female and male speakers turned out not to be statistically significant: F(1,68) = .066, p = .797.[6]

(i) The alexithymia test. Eight participants out of the 68 (one female) did not successfully pass the 'alexithymia measurement test' SAR (Baiocco et al., 2005). The subjects who did not have the capacity to correctly identify their own emotions (those who did not pass the test) were also those for whom in general the induction procedures were less successful, as given by their classification results which mostly yielded a UAR lower than 48.5%, i. e., the threshold between the higher and lower half ranks (cf. Figure 3). Five out of the seven males who did not pass the test are in the lower half rank, whereas the other two are borderline, i. e., in the two last positions of the upper half rank (UAR = 48.9 and UAR = 49.3). Yet, the female who did not pass the test occupies the 22^{nd} position (out of 34) of the upper half rank, which indicates that even though this test can be generally considered as an indicator of reliable data, exceptions might be taken into account. Indeed, the alexithymia is a condition in which individuals are not able to identify their own and others' emotional states—which might not mean that they do not have the capacity to feel and therefore express the elicited emotional states. Since the goal of performing the alexithymia test is to guarantee the reliability of the qualitative strategies of control, i. e., self- and external assessment, the samples produced by the eight subjects who did not pass the alexithymia test

[6] Null-Hypothesis-Testing with p-values as decisive criterion has been critised repeatedly from its beginning; we refer to the statement of the American Statistical Association in Wasserstein and Lazar (2016). Throughout this article, we will thus report p-values not as criteria for a binary yes-no decision 'significant/not significant' but rather as a descriptive device; note that we do not correct for repeated measurements.

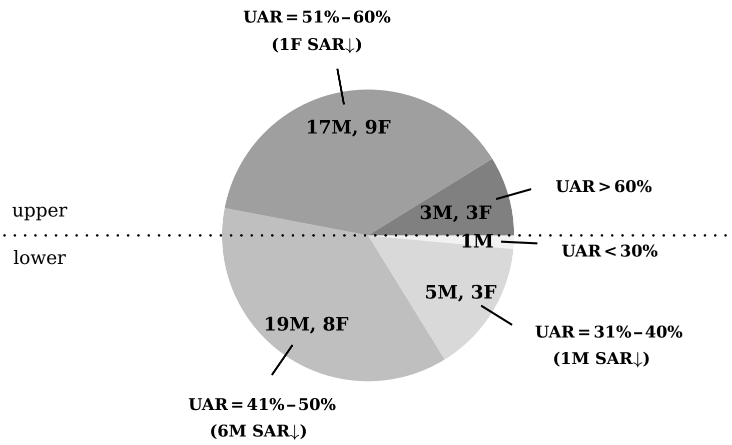


Fig. 3 Number of participants in the upper half rank (upper side of the pie chart, i. e., UAR > 48.5%) and in the lower half rank (lower side of the pie chart, i. e., UAR < 48.5%). Classification rank position for females (F) and males (M), and number of participants who did not successfully pass the alexithymia measurement test (SAR), e. g., 1F SAR↓ in the upper half rank, are indicated. The darker the shadowing, the higher the UAR.

were not considered in the selected corpus—the self-assessment performed by them may not be trustworthy.

(ii) Self-assessment. After discarding the eight participants who did not pass the alexithymia test, and the actress, responses of 59 participants were considered for the self-assessment evaluation. The efficiency of the induction methods, evaluated according to the UAR in the classification task—the higher the UAR, the more 'prototypical' the degree of the samples, thus the more efficient the induction—was not corroborated by the self-assessment, which displays high similarity between the participants' responses in the upper and lower half rank (cf. Table 5). By performing Pearson Chi-square, for the difference between emotions correctly identified and those misidentified in the upper and lower half rank, we got $p = .29$ for females and $p = .15$ for males. This confirms the idea that self-assessment might not be fully reliable (Schutte et al., 1998); to a certain extent, there always might be a subjective bias (Scherer and Ceschi, 1997), as shown by the high number of cases wrongly identified in the upper half rank. Sadness was one of the emotions better identified in the self-assessment, which confirms a successful induction; still, also other emotions with a lower UAR, such as fear or surprise, were accurately identified by the participants. This suggests that the induction of fear and surprise might easily encourage the *demand effect*, i. e., the condition in which a participant is aware of the emotion intended to be elicited (Vaughan, 2011). Yet, the low UAR contrasting with the high self-assessment accuracy may also display that despite a successful induction, these two emotions are scarcely prototypical, thus classified with difficulty.

(iii) External assessment. Since human annotation might be highly time-consuming, mechanisms as, e. g., crowdsourcing have been developed and suc-

Table 5 Results of the self-assessment for the upper and lower half rank considering the emotions individually and combined ($\sum$) for both females (F) and males (M). When the self-assessment coincided with the induced emotion, ✓ is given, otherwise ✗. Answers for 58 participants are encoded: in the upper half 31 (11F, 20M); in the lower half 28 (10F, 18M).

Upper Half Rank (UAR $>$ 48.5 %)																	
		anger		sadness		happiness		fear		disgust		guilt		surprise		$\sum$	
✓		21		23		13		22		15		18		22		134	
✗		10		8		18		9		16		13		9		83	
	#	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
F	77	6	5	8	3	5	6	7	4	8	3	5	6	5	6	44	33
M	140	15	5	15	5	8	12	15	5	7	13	13	7	17	3	90	50
Lower Half Rank (UAR $<$ 48.5 %)																	
		anger		sadness		happiness		fear		disgust		guilt		surprise		$\sum$	
✓		19		18		15		15		10		17		22		116	
✗		9		10		13		13		18		11		6		80	
	#	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
F	70	7	3	5	5	6	4	7	3	6	4	7	3	8	2	46	24
M	126	12	6	13	5	9	9	8	10	4	14	10	8	14	4	70	56

Table 6 Results of the external assessment for the upper and lower half rank considering the emotions individually and combined ($\sum$) for both females (F) and males (M). Emotions for which more than 50% of the samples were approved are indicated with ✓, otherwise with ✗. Results are given only for the emotions correctly identified in the self-assessment. For the number of evaluated cases (#), cf. ✓ in Table 5.

Upper Half Rank (UAR $>$ 48.5 %)																	
		anger		sadness		happiness		fear		disgust		guilt		surprise		$\sum$	
✓		12		12		5		16		6		11		15		77	
✗		9		11		8		6		9		7		7		57	
	#	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
F	44	3	3	2	6	3	2	4	3	3	5	4	1	5	0	24	20
M	90	9	6	10	5	2	6	12	3	3	4	7	6	10	7	53	37
Lower Half Rank (UAR $<$ 48.5 %)																	
		anger		sadness		happiness		fear		disgust		guilt		surprise		$\sum$	
✓		9		8		3		10		1		7		9		47	
✗		10		10		12		5		9		10		13		69	
	#	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
F	46	2	5	1	4	1	5	5	2	0	6	2	5	2	6	13	33
M	70	7	5	7	6	2	7	5	3	1	3	5	5	7	7	34	36

cessfully applied in listeners' evaluation of big data, with this collective external assessment minimising the individual effort. Nevertheless, the reliability of such annotations is hardly comparable to that achieved by an expert evaluation. Considering this, and since the main goal of the selection process is to identify the samples of the corpus more representative of each emotion, i. e., prototypes, only the samples selected through the self-assessment procedure were evaluated by the three experts. The number of samples discarded by the experts in the lower half rank (69) is higher than in the upper (57), unlike those accepted, whose number is higher in the upper half rank (77) than in the lower (47); cf. Table 6 (p = .007 in Pearson's Chi-square). Yet, also from

Table 7 Binary classification of non-selected (n-sel) and selected (sel) samples considering actual size (act-s), down-sampling (down-s), and up-sampling (up-s) for training (train), development (dev), and test.

Classification target	**2-class problem** non-selected (n-sel) / Selected (sel)
Experimental set-up	**Speaker independent** 6 permutations of train / dev / test
Partitioning	**Females (average of samples across the 10 partitionings)** up-s: train (1,852) / dev (1,942) / test (1,854) down-S: train (398) / dev (302) / test (316)
	Males (average of samples across the 10 partitionings) up-s: train (3,136) / dev (3,506) / test (3,312) down-s: train (714) / dev (658) / test (740)
Sample size	**Balanced** Up sampling (up-s): n-sel (act-s) / sel (up-s) Down sampling (down-s): n-sel (down-s) / sel (act-s)
# experiments	**28 individual experiments (2 x 7 x 2) x 10 partitionings = 280** Gender (F, M) x Emotion (ANG, SAD, HAP, FEA, DIS, GUI, SUR) x Size (up-s, down-s)

the external assessment (as well as in the self-assessment), a high number of potentially valid samples—produced by speakers in the first positions of the rank—were excluded. Indeed, although sadness was effectively induced, the majority of samples produced by females in this emotion were excluded in both halves of the rank (samples of 10 females discarded, of 3 accepted). An opposite trend is shown for fear with, in the lower half rank, a higher proportion of acceptance (samples of 10 participants accepted, of 5 discarded), even though the induction of fear showed scarce reliability.

7 ML approach: Sample Size vs Typicality

Selection strategies, as those previously evaluated, might be a way to achieve speech 'typical' of an emotion. Yet, such strategies—which might not be fully reliable—may massively reduce the data, as we have seen in our case: The full corpus comprises 9,365 samples, the selected one only 1,564. Moreover, factors such as typicality or sample size might influence the performance of systems for speech emotion recognition. To evaluate the extent to which these two factors affect the performance of ML approaches, we carried out experiments where we systematically varied the sizes of the selected and non-selected samples for training and test. We consider as more 'typical' the 1,564 selected samples (508 produced by females and 1,056 by males), and as 'not typical' the remaining 7,801 non-selected samples (2,824 produced by females and 4,977 by males); note that the 332 neutral samples are not considered in this round.

7.1 Binary Classification: Selected vs Non-Selected

In order to assess whether the selected samples can be recognised when using an automatic approach, a binary classification problem to discriminate between selected and non-selected samples was performed (cf. Classification target in Table 7). Since most of the speakers are the same in the selected and non-selected groups, in order to perform a speaker independent task, both

Table 8 Partitioning of non-selected and selected samples in training (train), development (dev), and test for female and male speakers, computed across the ten 3-fold speaker independent randomly generated partitionings. Mean number of samples (#) per emotion, per fold, and total number of cases ($\sum$) is given.

	Non-Selected (n-sel)							**Selected (sel)**						
#	Female			Male			$\sum$	Female			Male			$\sum$
	train	dev	test	train	dev	test		train	dev	test	train	dev	test	
ANG	144	151	148	246	284	258	1,231	28	24	21	59	55	59	246
SAD	149	153	148	205	230	223	1,108	31	25	26	114	119	107	422
HAP	138	153	145	236	285	271	1,228	46	20	22	35	18	26	167
FEA	119	122	113	200	225	200	979	20	20	21	41	33	42	177
DIS	172	181	172	328	353	332	1,538	27	20	24	14	13	42	140
GUI	108	109	103	201	204	195	920	25	27	28	37	44	48	209
SUR	96	102	98	152	172	177	797	22	15	16	57	47	46	203
$\sum$	926	971	927	1,568	1,753	1,656	7,801	199	151	158	357	329	370	1,564

kinds of samples were split up into three partitions (training, development, and test) by randomly assigning different speakers to each fold. To make a fair comparison between selected and non-selected groups, the random assignment of speakers to each fold was the same for selected and non-selected samples; since some speakers have been excluded from the selected group, these are considered in the non-selected partition only. In order to minimise the different distribution of samples per emotion across speakers in each fold, 10 different 3-fold partitions were automatically generated. The experiments—performed ten times according to the different partitioning—were carried out following the procedure described in Section 6.2, i. e., computing the six possible permutations between training, development, and test sets, and the results across permutations and partitioning were averaged (cf. Experimental set-up and Partitioning in Table 7; Table 8).

Since the unbalanced number of selected and non-selected samples might influence the performance of the classifier, the experiments were carried out twice—balancing both groups. On the one side, the selected speech was up-sampled (for each emotion and speaker in each fold) to match the sample size (actual size) given in the partitioning of the non-selected speech. On the other side, the non-selected speech was down-sampled by randomly deleting the samples (for each emotion and speaker in each fold) that exceed the sample size (actual size) given in the partitioning of the selected speech (cf. Sample size in Table 7). The classification was carried out for each gender and each emotion individually, i. e., 14 experiments considering each time the selected and non-selected samples of only one gender/emotion, and this was performed twice (considering two sample sizes), i. e., 28 experiments in total, all of them with sub-sets of different samples—notice that this process was repeated 10 times according to the different partitionings (cf. # experiments in Table 7). To guarantee the reliability of the down-sampling procedure, the random selection of the non-selected samples was performed ten times, and the experimental results, which were obtained ten times according to ten different randomisations (random seeds), were averaged. As previously, experiments were conducted over all six permutations of the folds for each of the different partitionings. Results are given in Table 9.

Table 9 Mean recall per class (in %) for the speaker independent binary classification of non-selected (n-sel) vs selected (sel) samples achieved in the test phase, averaging over the six permutations. Results computed for each emotion individually are given separately for females (F) and males (M) considering the up-sampling (up-s) of the minority class (sel), and the down-sampling (down-s) of the majority (n-sel). Values higher than 50% are highlighted in bold. The mean UAR values across all emotions is 58% for females and 74.5% for males in up-sampled experiments, and 59.5% for females and 75.3% for males in down-sampled experiments.

up-s	ANG		SAD		HAP		FEA		DIS		GUI		SUR	
	n-sel	sel	n-sel	sel	n-sel	sel	n-sel	sel	n-sel	sel	n-sel	sel	n-sel	sel
F	**79.3**	27.8	**79.4**	47.2	**84.4**	46.4	**84.2**	24.8	**77.4**	44.9	**80.1**	31.3	**81.1**	23.7
M	**87.1**	**74.6**	**73.5**	**69.0**	**89.9**	36.5	**91.3**	**70.4**	**93.5**	**64.4**	**84.3**	**64.8**	**77.9**	**65.1**

down-s	ANG		SAD		HAP		FEA		DIS		GUI		SUR	
	n-sel	sel	n-sel	sel	n-sel	sel	n-sel	sel	n-sel	sel	n-sel	sel	n-sel	sel
F	**61.5**	**51.0**	**67.4**	**59.8**	**67.3**	**63.0**	**63.8**	**50.0**	**63.2**	**61.4**	**69.8**	47.0	**56.0**	**51.2**
M	**79.8**	**79.8**	**73.2**	**71.0**	**77.1**	**53.1**	**84.1**	**79.0**	**81.8**	**77.8**	**79.1**	**72.9**	**72.1**	**73.0**

7.1.1 Results

Contrary to our expectations, the binary classification carried out to identify selected and non-selected emotional speech (performed for each emotion individually) yielded a lower accuracy for the selected samples (cf. Table 9). Still, this difference is only prominent when up-sampling the selected group, which suggests that it might relate to the variability of the samples (notice that when up-sampling the selected group, the considered information is actually the same, just the size is balanced by duplicating samples). Yet, when down-sampling the non-selected group, the recall per class is still comparable or slightly higher for the automatic classification of the non-selected samples, which indicates that such a difference might relate to the experimental task. Indeed, since in binary classification each emotion is considered individually, the non-selected speech, being less ‘typical’ of each emotion, presents high intra-class diversity (samples within each class would be dissimilar to each other) and high inter-class similarity (samples from different classes would be similar to each other). This yields more diversity in training, increasing therefore the robustness of the model for the recognition of non-selected samples and thus encouraging identification in the test phase. On the contrary, the selected samples—since more ‘typical’—would present less variability in training, and therefore less available information when performing the test. Yet, we speculate that when considering all emotions as a target, i. e., in the 7-class classification problem, the robustness offered by the variability of non-selected samples would not be in any case sufficient to discriminate between different classes, due to their high inter-class similarity impairing recognition. This question is addressed in subsection 7.2.

7.2 Seven-class Classification

In order to evaluate the extent to which ML systems might be affected by the typicality and sample size of the corpus, a seven-class classification problem to discriminate between the seven emotional categories was performed (cf. Classification target in Table 10). Experiments were carried out by taking into

Table 10 Overview of the 7-class classification of seven emotions for within- and inter-group approaches—the considered groups are selected (sel), non-selected (n-sel), and all (i. e., the previous together). Actual size (act-s) and down-sampling (down-s) of the partitioning are also indicated.

Classification target	**7-class problem** ANG / SAD / HAP / FEA / DIS / GUI / SUR
Experimental set-up	**Within-group** 6 permutations of Train, Dev, & Test **Inter-group** 2 permutations of Train & Dev (group 1) / Test (group 2)
Sample size	**Within-group (Speaker independent)** Balanced: sel (act-s), n-sel (act-s), n-sel (down-s), all (act-s) **Inter-group (Speaker independent)** Balanced: Train & Dev with sel (act-s & down-s) / Test with n-sel (down-s) Balanced: Train & Dev with n-sel (down-s & down-s) / Test with sel (act-s) Unbalanced: Train & Dev: sel (act-s) / Test: n-sel (act-s) Unbalanced: Train & Dev: n-sel (act-s) / Test: sel (act-s)
# experiments	**Within-group: 8 individual experiments (2 x 4) x 10 partitioning** Gender (F, M) x Group (sel, n-sel, n-sel down-s, all) **Inter-group (balanced): 4 individual experiments (2 x 2) x 10 partitioning** Gender (F, M) x Group (n-sel, sel) **Inter-group (unbalanced): 4 individual experiments (2 x 2) x 10 partitioning** Gender (F, M) x Group (n-sel, sel)

account two different approaches: within-group and inter-group classification. For the **within-group classification**, the following groups were considered: selected, non-selected, all (selected + non-selected), and the down-sampled version of the non-selected group (we employed the downsampling procedure described in Section 7.1); for the **inter-group classification**, the groups were only selected and non-selected. In the within-group approach, the samples of one group were considered for training, development, and test; in the inter-group approach, samples of one group were considered for training and development, whereas samples of the other group were considered for test. For both within-group and inter-group approaches, the experiments were speaker independent, i. e., different speakers were considered for training, development, and test (cf. Experimental set-up in Table 10), and the experiments were performed ten times according to the different partitionings (cf. Table 8). For the inter-group classification, unbalanced and balanced sample sizes were taken into account: The unbalanced experiments were performed considering the actual size per fold for each group according to the partitionings. The balanced experiments were performed by down-sampling the two bigger folds, in order to match the size of the smallest one (cf. Sample size in Table 10). Permutations between training/development and test set were not performed as they belong to different groups (selected and non-selected). The experiments, performed ten times according to the 10 different partitionings, were carried out considering gender, group, and sampling separately (cf. # experiments in Table 10). Results are given in Table 11 and Table 12 for within- and inter-group approaches respectively.

7.2.1 Results

For the **within-group classification**, experiments were carried out by considering selected and non-selected samples both separately and together. Results

Table 11 Within-group classification for females (F) and males (M). Mean recall per class and UAR (in %) for the 7-class speaker independent problem (the seven emotions as a target) achieved in the test phase after the six permutations. For train, dev, and test samples of the same group, i. e., selected (sel), non-selected (n-sel), or both together (all), are considered. Results for the groups: sel, n-sel, n-sel down-sampled (down-s), and all; mean number of cases per fold (#) are given. Values above 50% are highlighted in bold.

F	ANG	SAD	HAP	FEA	DIS	GUI	SUR	UAR	# train	# dev	# test
sel (508)	**67.3**	**54.5**	34.1	35.2	40.8	29.0	45.0	43.7	199	151	158
n-sel (2,824)	**52.7**	**66.8**	**57.6**	48.2	**61.9**	48.7	**60.5**	**56.6**	926	971	927
n-sel down-s (508)	38.1	**60.6**	39.9	36.1	20.8	36.8	35.8	38.3	199	151	158
all (3332)	**58.5**	**67.7**	**58.7**	47.9	**61.8**	47.7	**61.7**	**57.7**	1,125	1,122	1,085
M	ANG	SAD	HAP	FEA	DIS	GUI	SUR	UAR	# train	# dev	# test
sel (1056)	**82.8**	**70.8**	37.6	**65.2**	23.2	32.4	**69.1**	**54.4**	357	329	370
n-sel (4977)	**66.4**	**72.8**	**59.6**	**58.9**	**69.0**	**57.7**	**65.4**	**64.2**	1,568	1,753	1,656
n-sel down-s (1056)	**54.2**	**65.4**	29.2	42.7	25.0	**51.8**	**58.9**	46.8	357	329	370
all (6033)	**68.0**	**71.2**	**69.7**	**68.9**	**75.4**	**68.3**	**70.5**	**70.3**	1,925	2,082	2,026

in Table 11 show that the classification considering the actual sample size yielded lower UAR for selected than for non-selected speech; yet, when performing a fair comparison by down-sampling the non-selected group, this strategy yields the lowest performance (38.3% for females and 46.8% for males). Thus, the hypothesis that the good accuracy revealed in the binary classification (i. e., selected vs non-selected) would depend of the experimental task is confirmed. As expected, the sample size plays an important role in ML approaches, as shown by the improvement in UAR when both selected and non-selected samples were considered together, which yielded highest performance with UAR = 57.7% for female and UAR = 70.3% for male (cf. rows ‘all’ for F and M in Table 11). This is supported by the better performance achieved for male voices in all the evaluated groups—note that the number of samples is much higher for males—and it is also shown when considering the recall for disgust, guilt, and happiness, which decreases considerably for a small sample size, i. e., for selected and non-selected down-sampled speech. Indeed, for these emotions the elicitation was not really successful; due to this, a small sample size would especially affect the performance of the model for these emotions.

For the **inter-group classification**, experiments were carried out by considering selected and non-selected samples for training/development and test alternatively. Results in Table 12 show that a higher sample size (thus higher variability) in training and development yielded the highest UAR for both females (53.5%) and males (55.0%), i. e., non-selected samples for training/development and selected for test. A balanced sample size for the three partitions yielded a higher accuracy when considering non-selected samples (down-sampled) for training (39.5% for females, 42.2% for males), unlike when considering the non-selected down-sampled group for test, which, as expected, yielded similar results as the unbalanced task (30.9% for females, 32.2% for males).[7] The presented results indicate that training a model with many samples, espe-

[7] The down-sampling in test (cf. row 2 for F and row 6 for M in Table 12) was made to allow a comparison with the down-sampling in Train and Dev with really everything kept equal (cf. row 4 for F and row 8 for M in Table 12), by processing in both cases fully

Table 12 Inter-group classification results for females (F) and males (M). Mean recall per class and UAR (in %) for the 7-class problem (the seven emotions as target) achieved in the test phase. Samples for train/dev and for test belong to different groups, i. e., selected (sel), non-selected (n-sel), and n-sel down sampled (down-s); mean number (#) of cases per fold are given. Values above 50% are highlighted in bold.

F	ANG	SAD	HAP	FEA	DIS	GUI	SUR	UAR	# train	# dev	# test
Train / Dev sel – Test n-sel	43.6	41.3	21.9	21.9	24.8	23.5	39.6	30.9	199	151	927
Train / Dev sel – Test n-sel down-s	38.8	38.6	22.3	20.1	34.3	21.5	40.9	30.9	151	151	151
Train / Dev n-sel – Test sel	**66.7**	**60.8**	44.0	49.5	**51.3**	35.4	**66.7**	**53.5**	926	971	158
Train / Dev n-sel down-s – Test sel	39.5	42.4	29.3	40.4	33.2	37.7	**54.1**	39.5	158	158	158
M	ANG	SAD	HAP	FEA	DIS	GUI	SUR	UAR	# train	# dev	# test
Train / Dev sel – Test n-sel	47.7	**57.1**	23.3	24.3	5.1	28.8	47.1	33.3	357	329	1,656
Train / Dev sel – Test n-sel down-s	42.4	**55.7**	25.1	25.9	7.5	26.6	42.1	32.2	329	329	329
Train / Dev n-sel – Test sel	**76.8**	**50.5**	43.4	**55.1**	**73.3**	31.4	**54.6**	**55.0**	1,568	1,753	370
Train / Dev n-sel down-s – Test sel	**61.1**	**60.9**	12.2	40.3	43.5	24.8	**52.8**	42.2	370	370	370

cially if these are not emotionally characteristic (i. e., non-selected samples), particularly increases its robustness due to the great intra-class diversity of the samples. This variability between samples of the same emotional class is encouraged on one side due to their scarce typicality, on the other side due to their sample size. This is confirmed by the low accuracy (regardless of the sample size in the test set) achieved when training the classifier with selected samples and performing the test with the non-selected ones. Indeed, selected speech, due to its high intra-class similarity (shown by the acoustic analysis, cf. Section 8), would not offer enough information for classifying samples with high intra-class diversity, i. e., non-selected samples (cf. the acoustic analysis in Section 8), which is particularly evident when looking at the low recall for disgust in male voices.

8 Acoustic evaluation: Selected vs non-selected

For assessing the acoustic differences between selected and non-selected samples of each emotion on an exemplary basis, we decided in favour of two robust features—F0 range and range of energy—that have shown to differ prominently between emotions (Williams and Stevens, 1972). In Figure 4, results for the F0_range (represented in the ComParE feature set as *F0final_sma_pctlrange0-1*), i. e., the range of the smoothed fundamental frequency (F0) contour, are given for both selected and non-selected samples, produced by females and males. Non-selected samples display more similarity between different emotions, which is shown by a comparable F0 range for all emotions, i. e., there is a small range of variances across emotions: For females, F0 ranges are from minimal values of 37 Hz for sadness to 67 Hz for fear, to maximum values of 390 Hz for happiness to 510 Hz for fear; cf. first and fourth quartiles of non-selected (female) in Figure 4. For males, F0 ranges show a minimum value from 36 Hz for sadness to 50 Hz for surprise, and a maximum value from 421 Hz for guilt to 460 Hz for disgust; cf. first and fourth quartiles of non-selected (male)

balanced groups; as expected, the classification results for the down-sampled test did not differ noticeably from those obtained for the unbalanced group.

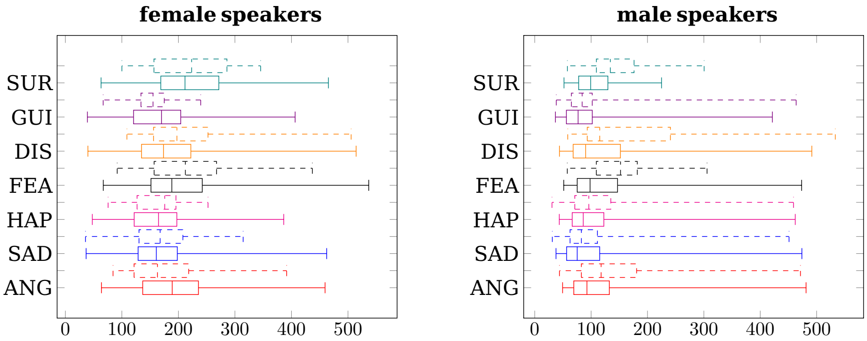


Fig. 4 Representation of the F0_range (x-axis) considering seven emotions (y-axis), non-selected (solid line) and selected (dashed line) samples, female and male speakers. The median and the four quartiles are also indicated: first quartile, second quartile, median, third quartile, and fourth quartile (from left to right).

in Figure 4; all emotions except surprise are included in the given ranges.[8] Differently, selected speech has a more characteristic F0 range for each emotion, i. e., there is a greater variability across emotions: For females, the F0 ranges are from a minimum value of 35 Hz for sadness to 108 Hz for disgust, and from a maximum value of 239 Hz for guilt to 506 Hz for disgust; cf. first and fourth quartiles of selected (female) in Figure 4. For males, F0 ranges are from a minimum value of 30 Hz for sadness to 58 Hz for fear, and from a maximum value of 300 Hz for surprise to 533 Hz for disgust; cf. first and fourth quartiles of selected (male) in Figure 4. The difference on the F0 ranges between non-selected and selected speech are most prominent for the maximum values: 390 Hz − 510 Hz vs 239 Hz − 506 Hz for female (Pearson chi squared yielded $p < .001$); 421 Hz − 460 Hz vs 300 Hz − 533 Hz for male ($p < .001$). Differently, the differences on the minimum values of the F0 range were small: 37 Hz − 67 Hz vs 35 Hz − 108 Hz for female ($p = .058$); 36 Hz − 50 Hz vs 30 Hz − 58 Hz for male ($p = .290$).

Results for the Energy_range (*pcm_RMSenergy_sma_pctlrange0-1* in the `ComParE` feature set), i. e., the range of the smoothed Root Mean Square (RMS) energy contour, display a similar tendency to the one described for the F0_range. Again, we see a homogeneous representation across different emotions for non-selected samples, and more unique representations for each emotion for the selected samples. This is more evident for females, as shown by the highly similar box plots of non-selected samples, whose median and second and third quartiles coincide almost perfectly across emotional categories: second quartile from .099 for guilt to .103 for surprise; third quartile from .132 for fear to .137 for guilt; median from .115 for fear to .118 for anger; cf. non-selected for female speakers (solid line) in Figure 5. This difference is evident when comparing emotions commonly related to one arousal level, such as sadness (usually identified as low aroused), with those commonly related to more

[8] Note that we do not compare F0 values across but only within gender; thus, we do not have to take into account the different mean pitch ranges of males and females.

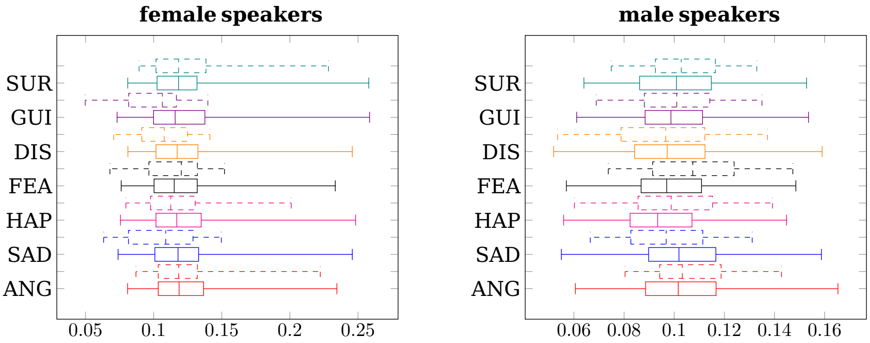


Fig. 5 Representation of the Energy_range (x-axis) considering: seven emotions (y-axis), non-selected (solid line) and selected (dashed line) samples, female and male speakers. Median and the four quartiles are also indicated: first quartile, second quartile, median, third quartile, and fourth quartile (from left to right).

than one arousal level such as anger (usually produced as both cold and hot anger): Selected samples of sadness present a smaller energy range (from .063 to .149) in comparison to the non-selected (from .074 to .246), while selected samples of anger show a large range (from .087 to .222) which is comparable to that for non-selected (from .081 to .234). The difference on the maximum values between sadness and anger for selected and non-selected speech yielded $p = .001$, cf. first and fourth quartiles for SAD and ANG (female speakers) in Figure 5. However, this tendency is not observed in male voices; this may relate to the cultural stereotype that emotional expressions related to 'weakness' would not be appropriate for males (Fischer, 1993): Males might often replace sadness with other culturally accepted emotions as, e. g., anger. Thus, selected speech for both sadness and anger show similar energy variability across samples: from .066 to .131 for sadness, and from .080 to .142 for anger. This is comparable to that displayed for non-selected speech: from .050 to .158 for sadness, and from .060 to .165 for anger ($p = .820$ and $p = .968$ for maximum and minimum values, respectively), cf. first and fourth quartiles for SAD and ANG (male speakers) in Figure 5. This variability would explain the high recall achieved for the recognition of sadness in males also when the classifier was trained with non-selected samples (cf. Table 12), by that supporting the idea that not only sample size but also greater variability in the training phase increases the robustness of speech emotion recognition systems.

In order to assess the relevance of these two features in an ML task, we performed binary classification of selected vs non-selected samples for each emotion individually (cf. Section 7.1), but considering this time only the two previously evaluated features. Our results are especially interesting for the classification of guilt in female voices—an emotion which showed great dissimilarity between selected and non selected samples for both F0_range and energy_range (cf. guilt for female speakers in Figures 4 and 5). Indeed, when considering these two features, the selected samples of guilt produced by females were classified much more accurately than the non selected ones, showing a recall of 65.4% vs 48.3% (selected vs non selected) in up-sampled experiments (UAR = 56.9%); and 69.8% vs 51.7% in down-sampled experiments (UAR = 60.8%). The mean UAR value across all emotions is 51.8% for females and 54.7% for males in up-sampled experiments, and 51.7% for females and 55.1% for males in down-sampled experiments.

From our acoustic evaluation, we can conclude that non-selected samples present high intra-class diversity and low inter-class homogeneity, i. e., the non-selected samples of each emotion are acoustically dissimilar to each other, thus similar across different emotional classes. This is shown by a great variability in the F0 and energy ranges for non-selected samples within each emotion—for all emotions, the first quartile is mostly low and the fourth quartile is high. Differently, selected samples are characterised by high intra-class similarity and high inter-class diversity, i. e., the selected samples of each emotion are acoustically similar to each other, thus dissimilar across different emotions. This is shown by a low variability in the F0 and energy ranges for selected samples within each emotional class—mostly, the four quartiles and median do not coincide across emotions.

9 Perceptual study

Aiming at a better understanding of the typicality of the different emotions, a subjective perceptual evaluation of the selected samples of the corpus (1,564 samples produced in seven emotional states) was conducted. For that, a listening test based on the categorical (Ekman, 1984) and bi-dimensional (Russell, 1980) models of emotion was performed by 86 Italian natives (42 females, 44 males). All listeners were students of computer science and received academic credits for their voluntary participation (mean age 21.4 years, std 1.1 years); none of them had taken part in the induction sessions. The listening test was performed through a computer-based interface developed in the visual programming tool *Max MSP*[9], and the samples were presented randomly to each listener. For the categorical assessment, a forced-choice task to decide between the 7 emotion classes of the corpus, i. e., anger, sadness, happiness, fear, disgust, guilt, and surprise, was considered. For the dimensional assessment, the two dimensions *arousal* (emotional intensity, from low to high) and *valence* (emotional hedonistic value, from negative to positive) were employed with a five-level rating scale (0 standing for low arousal and less positive valence; 4 standing for high arousal and more positive valence). For each sample, the listeners chose first a single category, then a single level for each dimension. To avoid fatigue, the corpus was divided into six sub-tasks (each lasted around 90 minutes); the samples were randomly assigned. The participants were instructed to perform the test with headphones.

[9] `https://cycling74.com/products/max/`

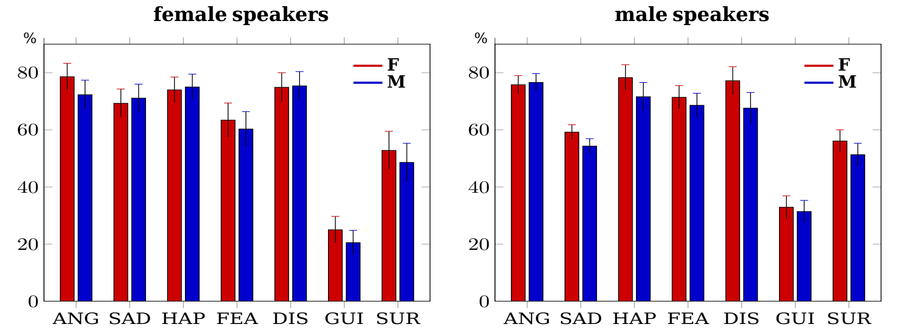


Fig. 6 Recall per class (in %) achieved by female (F) and male (M) listeners in the perception of the seven evaluated emotions. Results are given for the selected speech produced by female and male speakers; 95% coefficient intervals are also indicated on the top of each bar.

9.1 Results and discussion

Supporting previous research (Parada-Cabaleiro et al., 2017), the high similarity between females' and males' responses for all the evaluated emotions shows that gender seems not to play a role in the perception of emotional speech (cf. Figure 6). Therefore, the responses for all listeners will be evaluated together, irrespective of gender. Emotional expression produced by females and males seems to be comparable, too, since perceived by the listeners as similar (cf. Figure 6), thus showing analogous confusion matrices (cf. chromatic patterns in Table 13.A). Guilt is the emotion perceived with lowest accuracy (22.4 % for females and 32.3 % for males), followed by surprise (50.5 % for females and 53.8 % for males). This relates to the ambiguity of these emotions, guilt being a secondary emotion (Plutchik, 1991), surprise having undefined valence (Ortony and Turner, 1990). Guilt was mainly misidentified as sadness, by that contradicting the idea that guilt is a secondary emotion at the intersection between happiness and fear (Plutchik, 1991). However, the confusion of guilt with sadness supports previous research in the identification of guilty facial expressions (Keltner, 1996), who has shown that guilt is an emotion not reliably identified by observers, since mainly confused with sadness and other secondary emotions. Surprise, elicited as 'positive' surprise, was mainly misidentified as happiness, given the common level of valence in both emotions. These confusion patterns might also relate to the utilisation of the same induction procedure, e. g., the elicitation method A (cf. Figure 1) would yield similar results for guilt and sadness (both emotions with negative valence), and for surprise and happiness (both emotions with positive valence). Finally, although the induction of sadness showed to be very effective for males, its perception was less accurate than for females, which may again be explained by the cultural idea that 'weakness' is not an appropriate expression for males (Fischer, 1993). In this regard, sadness—since masked by males with emotions more 'aggressive' and culturally appropriated, as e. g., anger—would present a higher acoustic variability (cf. Section 8) and thus be easily misidentified by the listeners. Guilt and—up to a certain extent—surprise, being ambigu-

Table 13 Confusion matrices of the perception and automatic recognition of the selected samples of the corpus (results are expressed in %). Each row gives the ‘reference’ (emotion labels in capitals); each column ‘identified as’ (emotion labels in small letters). The darker the shadowing, the higher the accuracy. Number of cases per emotion (#), for female (F) and male (M) speakers are given.

Table 13.A Perception results of the listening test. UAR (in %) for F = 61.4, for M= 62.6.

F	*ang*	*sad*	*hap*	*fea*	*dis*	*gui*	*sur*	#
ANG	75.0	2.4	0.9	8.5	3.7	2.8	6.6	73
SAD	4.8	70.3	4.2	7.3	4.4	5.2	3.7	82
HAP	2.1	2.1	74.6	1.7	2.0	0.8	16.7	88
FEA	10.3	6.8	3.1	61.7	2.5	3.8	11.8	61
DIS	3.2	8.0	4.2	3.5	75.1	1.2	4.7	71
GUI	6.2	36.8	15.4	5.9	6.0	22.4	7.2	80
SUR	5.0	3.8	32.6	2.8	4.3	1.1	50.5	53
M	*ang*	*sad*	*hap*	*fea*	*dis*	*gui*	*sur*	#
ANG	76.2	2.3	0.9	5.6	6.0	2.0	7.0	173
SAD	6.1	56.9	4.0	8.8	9.1	6.2	8.9	340
HAP	1.3	8.8	75.0	1.6	1.4	0.6	11.4	79
FEA	6.0	4.5	2.4	70.1	1.1	2.6	13.3	116
DIS	2.8	8.3	5.0	1.5	73.8	1.4	7.1	69
GUI	6.1	33.8	6.3	4.0	7.9	32.3	9.6	129
SUR	5.5	1.7	31.7	2.7	3.9	0.8	53.8	150

Table 13.B Test results for the automatic classification. Selected samples were considered for test, non-selected samples for train/dev (cf. the rows Train/Dev n-sel - Test sel in Table 12). The rows diff$_F$ and diff$_M$ give the differences between perception and classification for female and male respectively. UAR (in %) for F = 53.5, for M= 55.0.

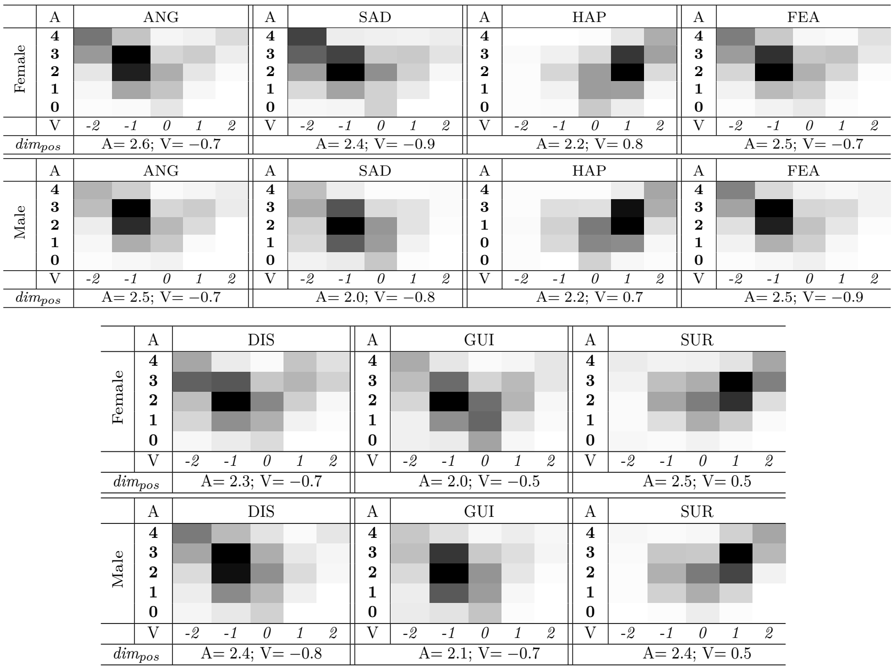
F	*ang*	*sad*	*hap*	*fea*	*dis*	*gui*	*sur*	#
ANG	66.7	11.2	3.2	6.0	4.2	5.3	3.3	21
SAD	2.2	60.8	8.0	15.6	2.2	11.1	0.0	26
HAP	8.6	7.3	44.0	4.9	10.8	6.5	17.9	23
FEA	11.2	5.2	11.7	49.5	6.5	6.0	9.9	21
DIS	9.3	12.5	2.4	7.6	51.3	10.6	6.3	25
GUI	1.0	21.6	15.2	16.7	2.7	35.4	7.5	28
SUR	2.1	3.4	14.7	4.6	7.6	0.8	66.7	17
diff$_F$	8.3	9.5	30.6	12.2	23.8	−13.0	−16.2	−350
M	*ang*	*sad*	*hap*	*fea*	*dis*	*gui*	*sur*	#
ANG	76.8	1.6	0.7	3.4	4.4	2.0	11.1	59
SAD	10.3	50.5	11.3	6.2	11.2	7.2	3.2	105
HAP	8.3	0.6	43.4	19.9	18.7	2.2	6.8	26
FEA	12.7	2.9	3.6	55.1	10.9	1.6	13.2	41
DIS	0.0	7.1	6.7	8.6	73.3	1.6	2.6	42
GUI	11.1	8.3	14.4	18.7	12.0	31.4	4.0	48
SUR	10.9	0.2	5.5	15.6	9.6	3.5	54.6	45
diff$_M$	−0.6	6.4	31.6	15.0	0.5	0.9	−0.8	−686

ous emotions and thus characterised by low typicality—were identified less accurately.

By evaluating the ML approach which achieved best results in classifying the selected speech, i. e., considering non-selected samples for training/development and the selected ones for test, i. e., UAR for F = 53.5%, for M= 55.0% (cf. the rows Train/Dev n-sel - Test sel in Table 12), the classification results show a high similarity with those obtained by the perceptual assessment, especially for male speakers (cf. Table 13.B for ML results, Table 13.A for perception). Anger is the emotion classified and perceived best, with 76.8% and 76.2%, respectively (cf. -0.6% of difference in the diff$_M$ row in Table 13.B). Guilt is classified and perceived worst, with 31.4% and 32.3%, respectively (0.9% of difference). This relates to anger having a prototypical representation in the selected corpus, whereas guilt has none. This was shown in the acoustic evaluation (cf. Figures 4 and 5), where anger showed less variability of F0 and energy range for the selected samples than for the non-selected, unlike guilt, an emotion with similar variability for both selected and non-selected samples. As previously discussed, this relates to anger being a basic emotion and therefore having a prototypical representation which can be modelled by selection procedures, whereas guilt, being a secondary emotion, does not present such a typical representation in any of the groups, be these selected or non-selected. The recognition of happiness and fear (and also disgust for females) displays a generally lower performance in the ML approach (the biggest differences are 44.0% vs 74.6% and 43.4% vs 75.0% for classification vs perception of happiness in the female and male voices respectively), whereas performance for surprise is higher, especially for females (66.7% vs 50.5% for classification vs perception). For happiness and fear, this could relate to the lower efficiency of the induction procedures; yet, the higher accuracy of surprise and the lower of disgust for females do not seem to relate to the efficiency of the emotional induction. The most evident confusion is between guilt and sadness for females: It is mostly marked for perception with 36.8% of the cases of guilt misidentified vs 22.4% correctly identified, less for ML with 21.6% misclassified vs 35.4% correctly classified. Moreover, confusion is asymmetric: Guilt is more confused with sadness than vice versa. This supports the idea that secondary emotions are made up of the combination of primary ones, thus, they do not display a unique acoustic representation.

Supporting the categorical findings, the dimensional perception of samples produced by females and males turned out to be very similar (cf. the analogous chromatic patterns in Table 14). Similarities between different emotions are also displayed by comparable dimensional constellations in the arousal-valence space; this mirrors the confusion patterns displayed in the categorical evaluation, i. e., the confusion between surprise and happiness, and between guilt and sadness. However, this tendency was not consistent, since some emotions are highly similar in the dimensional space, as, e. g., fear and anger, while not yielding relevant confusion patterns in the categorical domain (cf. Table 13.A). Furthermore, the categorical confusion patterns present themselves mostly in one direction, i. e., one of the emotions of the confusion pattern is more affected by the misidentification than the other, which cannot be displayed in a bi-dimensional space. This is clear when evaluating the confusion pattern of happiness vs surprise, where around 30% of the samples of surprise are misidentified as happiness, whereas only around 15% of the samples of happiness are misidentified as surprise (cf. Table 13.A). From the categorical point of view, the direction of this confusion pattern is explained by the fact that

Table 14 Dimensional evaluation of each considered emotion, for both female and male speakers. Arousal (A) in the y-axis and valence (V) in the x-axis (from 0–lower to 4–higher level). Per cell, sum of listeners' scores, normalised to 0–100 is given; grey shadowing represents frequencies (the darker, the higher); dimensional position dim$_{pos}$ (overall mean score) given for arousal (A) an valence (V).



happiness is an emotion less ambiguous than surprise (thus easily identified). Such a confusion cannot be displayed in the dimensional model. Indeed, in the dimensional assessment, both emotions (happiness and surprise) show a similar arousal-valence representation, mainly clustered in the upper-right section of the bi-dimensional space, with a mean valence around 1 and arousal around 2, e. g., in males A = 2.2 and V = 0.7 for happiness, A = 2.4 and V = 0.5 for surprise (cf. dim$_{pos}$ for happiness and surprise in Table 14).

Thus, sometimes emotion A is confused more with emotion B than it is the other way round. This is evident in the categorical confusion between sadness and guilt, with many samples of guilt wrongly identified as sadness (36.8% for females, cf. Table 13.A) and only a few samples of sadness wrongly identified as guilt (5.2% for females, cf. Table 13.A). This pattern is mirrored by the ML approach as well: 21.2% of the samples of guilt were misclassified as sadness, while only 6.1% of the samples of sadness were misclassified as guilt (cf. females in Table 13.B). From the categorical point of view, this unidirectional confusion pattern would be explained with sadness having been successfully induced (especially for females), thus presenting a high typicality, i. e., low dissimilarity across selected samples. Still, this might also relate to the fact that sadness—since characterised by low pitch, tone, and energy—may be mostly perceived as an emotion less 'prominent', i. e., more similar to a 'neutral' or an 'undefined' category. Indeed, this has been shown by evaluating the perception of emotional speech in background noise, a condition in which sadness was the emotion best recognised due to an increment in the confusion, i. e., in background noise the percentage of samples wrongly identified as sadness was higher than in clean condition (Parada-Cabaleiro et al., 2017). Guilt, however, was not only unsuccessfully induced, but since it is a secondary emotion, it presents itself with a low typicality (i. e., high dissimilarity across selected samples). Again, this tendency cannot be mirrored in the bi-dimensional space, which simply displays a high similarity of the arousal-valence representation for sadness and guilt (cf. Table 14). This confirms that the bi-dimensional model is not sufficient to discriminate between emotions with comparable levels of arousal and valence (Devillers et al., 2005b; Parada-Cabaleiro et al., 2018), suggesting that more dimensions would be needed, especially to discriminate between emotions without a prototypical representation, such as guilt or other secondary emotions (Fontaine et al., 2007).

10 Limitations

In our work, we have considered categorical emotions, which allows for the comparative evaluation of our findings with those previously presented in emotional speech research (mostly based on the categorical model of emotions). Yet, emotion categories are not fully comparable to the emotional states typical of real life. In this regard, the outcomes of our work, even though presenting a compromise between the categorical model and emotional speech authenticity, might not be fully generalisable to the 'natural' emotional states typical of everyday interactions. Therefore, we consider our study to be an intermediate step towards the collection and evaluation of more 'natural' emotional speech. Such a process might also consider the utilisation of emotional induction procedures, but having in mind a larger array of emotions to be elicited (which would not be limited to the 'big six' emotion categories).

Concerning the variety of Mood Induction Procedures (MIPs) employed, these were not equally efficient for the elicitation of all the emotional states—being particularly unsuccessful in the induction of fear. Furthermore, the use of emotionally connoted texts and utterances, i. e., mood induction procedures that require suitable reading aloud skills, reduced also the naturalness of the speech productions for some participants (as shown by their lack in speaking fluency). This resulted in the rejection of many samples during the selection process, which massively reduced the sample size of the selected corpus. In this regard, other mood induction procedures should be investigated, especially for the elicitation of fear, and future research should, if ever possible, prioritise spontaneous speech production over reading aloud procedures.

As for the selection procedures, even though the self-assessment turned out not to be fully reliable, it is not clear whether this depends on the instrument of measurement considered, i. e., the proposed categorical/dimensional diagram, or on the self-assessment itself (which has shown not to be fully reliable). Also, as expected, the perceptual evaluation through the bi-dimensional model proved to be insufficient to mirror all the categorical confusion patterns. In this regard, the evaluation of other measurement instruments in self-assessment procedures, as well as the consideration of additional emotional dimensions in perceptual assessment, are research questions still open to further investigation.

Finally, regarding the ML techniques, these were affected by the unbalanced sample size of the corpus: a predominance of samples produced by males and, after selection, a very small amount of prototypical instances produced by females. This made the ML results for within-group classification (in selected samples) unstable and less generalisable for females than those achieved for males. Furthermore, given the unequal distribution of samples per gender, also the comparative evaluation of ML and perceptual findings between males and females might be subjected to a certain degree of bias.

11 Conclusions

We presented `DEMoS`, the first database of induced emotional speech in Italian, an almost unrepresented language in emotional speech research. `DEMoS` encompasses 9,365 samples in anger, sadness, happiness, fear, disgust, and surprise (the emotions most prominently considered in speech emotion research) as well as guilt (a secondary emotion scarcely considered in emotional elicitation studies). In addition to these, the corpus also presents 332 samples in neutral mood produced by the 68 speakers who participated in the study. To evoke the emotional states in the participants, an induction program to better suit the optimal induction of each emotion was developed by taking into account the combination of at least three Mood Induction Procedures (MIPs). Combining music, autobiographical memories, and emotionally connoted texts and sentences showed to be particularly effective for the elicitation of sadness, but not for happiness, guilt, or surprise; film sequences and emotional texts and sentences were effective to elicit anger but not fear; pictures, texts, and sentences succeed in eliciting disgust. After discussing the difficulties of creating a corpus of elicited emotional speech, we also assessed the extent to which the collected samples are typical of each emotion. The listeners' evaluation displays that the selected cases (prototypes) of fear and happiness, even though the elicitation of these emotions was not fully successful, are well identified by the listeners, showing that these emotions can be successfully induced; this encourages the investigation of more adequate elicitation methods. Differently, when assessing the perception of surprise and especially of guilt, the listeners' accuracy is considerably lower, which suggests that not only the induction method might not be adequate, but also that these two emotions, given their ambiguity (guilt as secondary emotion, surprise as void of a specific valence) could be particularly challenging to be elicited and perceived. This is partially mirrored by the ML approach, which confirms that the emotional speech collected for guilt is hardly classified correctly, given its lack of typicality. On the contrary, the classification of anger—an emotion for which the collected samples seem to be prototypical—yielded higher accuracy for both the perceptual and ML approaches. Yet, emotional typicality, in contrast to sample size, seems not to be essential for successfully training ML models.

By employing acoustic analysis, perception study, and self-assessment, our research showed the influence of cultural rules in emotion expression, which can be observed when evaluating the selected samples of sadness produced by male participants. Sadness, even though successfully elicited, was expressed by males as masked with more aggressive emotions as, e. g., anger, which can be explained by the cultural idea that the expression of weakness is not adequate for males. This is shown by the acoustic variability of the selected samples for sadness produced by males, unlike those produced by females (characterised by a high acoustic homogeneity across selected samples). Furthermore, this is also supported by perceptual and ML evaluation, which demonstrates that the selected samples of sadness produced by males are perceived and classified worse than those produced by females. Given that the selection process comes with a massive reduction of the corpus, in order to assess the extent to which sample size and typicality influence the performance of ML systems for speech emotion recognition, the role of the selected and non-selected sections of the corpus in an ML framework were comparatively evaluated. As expected, sample size plays an important role in training, unlike emotional typicality. Prototypical samples, since presenting a high intra-class acoustic homogeneity, i. e., samples of one emotion are similar to each other and different to those of another emotional class, would reduce the robustness of a system when considered for training; yet, samples typical of an emotion are more adequate to reliably evaluate the performance of the model, i. e., these would be suitable for testing. Indeed, listeners' perception and the ML approach showed comparable results when training the system with non-selected samples and testing with selected ones. Finally, and confirming previous research, the bi-dimensional model proved to be insufficient to completely mirror all the aspects displayed by the categorical model, as, e. g., the hierarchy of the confusion patterns. This became evident in the confusion with guilt, which is hardly induced, perceived, and classified, something that relates to the scarce acoustic homogeneity typical of secondary—thus ambiguous—emotions, not presenting a prototypical expression.

In this work, we evaluated the challenges of performing mood induction procedures when collecting emotional speech data, demonstrating evidence for successful techniques and pinpointing research questions that are still open to further investigation. By presenting DEMoS, we also aim to encourage the study of Italian in affective computing research, as well as the consideration of other emotions (such as guilt) currently underrepresented in most of the available emotional speech corpora. Finally, through our perceptual study, we also attempt to motivate emotional speech research from an integrative perspective which would comparatively evaluate the emotion model most prominently considered in speech emotion corpora (i. e., the categorical model) with the other main emotion model in psychology (i. e., the dimensional model).

12 Acknowledgements

This work was supported by the European Union's 7th Framework programme under grant agreement No. 338164 (ERC StG iHEARu). We would also like to thank Arianna Mencattini for her help with the texts used in the induction procedure as well as Barbara Basile for her advices in the psychological measurement.

References

Amir, N., Ron, S., Laor, N. (2000). Analysis of an emotional speech corpus in Hebrew based on objective criteria. In: Proc. of ITRW, ISCA, Newcastle, UK, pp. 29–33.

Aubergé, V., Audibert, N., Rilliard, A. (2003). Why and how to control the authentic emotional speech corpora. In: Proc. of Interspeech, ISCA, Geneva, Switzerland, pp. 185–188.

Baiocco, R., Giannini, A. M., Laghi, F. (2005). SAR - Scala Alessitimica Romana. Valutazione delle capacità di riconoscere, esprimere e verbalizzare le emozioni. Erickson, Trento, Italy.

Bänziger, T., Pirker, H., Scherer, K. (2006). GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. In: Proc. of LREC, ELRA, Genova, Italy, pp. 15–19.

Barkhuysen, P., Krahmer, E., Swerts, M. (2010). Crossmodal and incremental perception of audiovisual cues to emotional speech. Language and Speech 53(1), 3–30.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E. (2000). Desperately seeking emotions or: Actors, wizards, and human beings. In: Proc. of ITRW, ISCA, pp. 195–200.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M. J., Wong, M. (2004). 'You stupid tin box' – children interacting with the AIBO Robot: A cross-linguistic emotional speech corpus. In: Proc. of LREC, ELRA, Lisbon, Portugal, pp. 171–174.

Batliner, A., Steidl, S., Hacker, C., Nöth, E., Niemann, H. (2005). Tales of tuning – prototyping for automatic classification of emotional user states. In: Proc. of Interspeech, ISCA, Lisbon, Portugal, pp. 489–492.

Bennett, M. J. (1979). Overcoming the golden rule: Sympathy and empathy. Annals of the International Communication Association 3(1), 407–422.

Bonny, H. L. (2002). Music & consciousness: The evolution of guided imagery and music. Barcelona Publishers, Gilsum, NH.

Bradley, M. M., Lang, P. J. (2000). Affective reactions to acoustic stimuli. Psychophysiology 37(2), 204–215.

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

Cavanagh, S. R., Urry, H. L., Shin, L. M. (2011). Mood-induced shifts in attentional bias to emotional information predict ill-and well-being. Emotion 11(2), 241–248.

Chiţu, A. G., van Vulpen, M., Takapoui, P., Rothkrantz, L. J. M. (2008). Building a Dutch multimodal corpus for emotion recognition. In: Workshop on Corpora for Research on Emotion and Affect, LREC, Marrakesh, Morocco, pp. 53–56.

Ciceri, M. R., Anolli, L. M. (2000). La voce delle emozioni: Verso una semiosi della comunicazione vocale non-verbale delle emozioni. Franco Angeli, Milan, Italy.

Costantini, G., Iaderola, I., Paoloni, A., Todisco, M. (2014). EMOVO Corpus: An Italian emotional speech database. In: Proc. of LREC, ELRA, Reykjavik, Iceland, pp. 3501–3504.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M. (2000). Feeltrace: An instrument for recording perceived emotion in real time. In: Proc. of ITRW, ISCA, Newcastle, UK, pp. 19–24.

Cullen, C., Vaughan, B., Kousidis, S., Wang, Y., McDonnell, C., Campbell, D. (2006). Generation of high quality audio natural emotional speech corpus using task based mood induction. In: Proc. of InSciT, Dublin Institute of Technology, Mérida, Spain.

Dan-Glauser, E. S., Scherer, K. R. (2011). The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. Behavior Research Methods 43(2), 468–477.

Devillers, L., Abrilian, S., Martin, J.-C. (2005a). Representing real-life emotions in audiovisual data with non basic emotional patterns and context features. ACII pp. 519–526.

Devillers, L., Vidrascu, L., Lamel, L. (2005b). Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.

Van der Does, W. (2002). Different types of experimentally induced sad mood? Behavior Therapy 33(4), 551–561.

Douglas-Cowie, E., Cowie, R., Schröder, M. (2000). A new emotion database: Considerations, sources and scope. In: Proc. of ITRW, ISCA, Newcastle, UK, pp. 39–44.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P. (2003). Emotional speech: Towards a new generation of databases. Speech Communication 40(1), 33–60.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., Karpouzis, K. (2007). The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In: Proc. of ACII, AAAC, Lisbon, Portugal, pp. 488–500.

Douglas-Cowie, E., Cox, C., Martin, J.-C., Devillers, L., Cowie, R., Sneddon, I., Mcorie, M., Pelachaud, C., Peters, C., Lowry, O., Batliner, A., Hönig, F. (2011). Data and databases. In: Petta, P., Pelachaud, C., Cowie, R. (eds) Emotion-oriented systems: The HUMAINE handbook, Springer, Berlin, Germany, pp. 163–284.

Ekman, P. (1984). Expression and the nature of emotion. Approaches to emotion 3, 19–344.

El Ayadi, M., Kamel, M. S., Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44(3), 572–587.

Eyben, F., Wöllmer, M., Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proc. of ACM Multimedia, ACM, Florence, Italy, pp. 1459–1462.

Eyben, F., Salomão, G. L., Sundberg, J., Scherer, K. R., Schuller, B. W. (2015). Emotion in the singing voice – a deeper look at acoustic features in the light of automatic classification. EURASIP Journal on Audio, Speech, and Music Processing 1, 1–9.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J. (2008). Liblinear: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874.

Fernandez, R., Picard, R. W. (2003). Modeling drivers' speech under stress. Speech Communication 40, 145–159.

Fischer, A. H. (1993). Sex differences in emotionality: Fact or stereotype? Feminism & Psychology 3, 303–318.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. Psychological Science 18(12), 1050–1057.

Gerrards-Hesse, A., Spies, K., Hesse, F. W. (1994). Experimental inductions of emotional states and their effectiveness: A review. British journal of psychology 85(1), 55–78.

Grichkovtsova, I., Morel, M., Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. Speech Communication 54(3), 414–429.

Gross, J., Levenson, R. (1995). Emotion elicitation using films. Cognition & Emotion 9, 87–108.

Husain, G., Thompson, W. F., Schellenberg, E. G. (2002). Effects of musical tempo and mode on arousal, mood, and spatial abilities. Music Perception: An Interdisciplinary Journal 20(2), 151–171.

Iida, A., Campbell, N., Higuchi, F., Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. Speech Communication 40(1–2), 161–187.

Johnstone, T., Scherer, K. R. (1999). The effects of emotions on voice quality. In: Proc. of ICPhS, UCLA, San Francisco, CA, pp. 2029–2032.

Johnstone, T., van Reekum, C. M., Hird, K., Kirsner, K., Scherer, K. R. (2005). Affective speech elicited with a computer game. Emotion 5(4), 513.

Keltner, D. (1996). Evidence for the distinctness of embarrassment, shame, and guilt: A study of recalled antecedents and facial expressions of emotion. Cognition & Emotion 10, 155–172.

Klasmeyer, G., Johnstone, T., Bänziger, T., Sappok, C., Scherer, K. R. (2000). Emotional voice variability in speaker verification. In: Proc. of ITRW, ISCA, Newcastle, UK, pp. 213–218.

Konečni, V. J., Brown, A., Wanic, R. A. (2008). Comparative effects of music and recalled life-events on emotional state. Psychology of Music 36(3), 289–308.

Labov, W. (1972). Sociolinguistic patterns. University of Pennsylvania Press, Philadelphia, PA.

Martin, M. (1990). On the induction of mood. Clinical Psychology Review 10(6), 669–697.

Mayer, J. D., Allen, J. P., Beauregard, K. (1995). Mood inductions for four specific moods: A procedure employing guided imagery vignettes with music. Journal of Mental Imagery 19(1–2), 151–159.

McCraty, R., Barrios-Choplin, B., Atkinson, M., Tomasino, D. (1998). The effects of different types of music on mood, tension, and mental clarity. Alternative therapies in health and medicine 4(1), 75–84.

Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M., Di Natale, C. (2014). Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowledge-Based Systems 63, 68–81.

Mikula, G., Scherer, K. R., Athenstaedt, U. (1998). The role of injustice in the elicitation of differential emotional reactions. Personality and social psychology bulletin 24(7), 769–783.

Mower, E., Metallinou, A., Lee, C., Kazemzadeh, A., Busso, C., Lee, S., Narayanan, S. (2009). Interpreting ambiguous emotional expressions. In: Proc. of ACII, IEEE, Amsterdam, Netherlands.

Murray, I. R., Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Communication 16, 369–390.

Ortony, A., Turner, T. J. (1990). What's basic about basic emotions? Psychological Review 97(3), 315–331.

Parada-Cabaleiro, E., Baird, A., Batliner, A., Cummins, N., Hantke, S., Schuller, B. (2017). The perception of emotions in noisified non-sense speech. In: Proc. of Interspeech, ISCA, Stockholm, Sweden, pp. 3246–3250.

Parada-Cabaleiro, E., Costantini, G., Batliner, A., Baird, A., Schuller, B. (2018). Categorical vs Dimensional perception of Italian emotional speech. In: Proc. of Interspeech, ISCA, Hyderabad, India, pp. 3638–3642.

Philippot, P. (1993). Inducing and assessing differentiated emotion-feeling states in the laboratory. Cognition & emotion 7(2), 171–193.

Plutchik, R. (1991). The emotions. University Press of America, Lanham, MD.

Roedema, T. M., Simons, R. F. (1999). Emotion-processing deficit in alexithymia. Psychophysiology 36(3), 379–387.

Rosch, E. H. (1973). Natural categories. Cognitive psychology 4(3), 328–350.

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology 39(6), 1161–1178.

Russell, J. A. (1991). In defense of a prototype approach to emotion concepts. Journal of Personality and Social Psychology 60, 37–47.

Scherer, K. R. (2005). What are emotions? And how can they be measured? Social Science Information 44(4), 695–729.

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. Computer Speech & Language 27(1), 40–58.

Scherer, K. R., Ceschi, G. (1997). Lost luggage: A field study of emotion–antecedent appraisal. Motivation and emotion 21(3), 211–235.

Scherer, K. R., Shuman, V., Fontaine, J. R., Soriano, C. (2013). The grid meets the wheel: Assessing emotional feeling via self-report. In: Fontaine, J. R., Scherer, K. R., Soriano, C. (eds) Components of emotional meaning: A sourcebook, Oxford University Press, Oxford, UK, pp. 281–298.

Schienle, A., Schäfer, A., Stark, R., Walter, B., Vaitl, D. (2005). Relationship between disgust sensitivity, trait anxiety and brain activity during disgust induction.

Neuropsychobiology 51, 86–92.

Schlosberg, H. (1954). Three dimensions of emotion. Psychological review 61(2), 81.

Schröder, M. (2004). Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. PhD thesis, Saarland University.

Schuller, B., Batliner, A., Steidl, S., Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53(9-10), 1062–1087.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S. (2013). The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In: Proc. of Interspeech, ISCA, Lyon, France, pp. 148–152.

Schuller, B., Steidl, S., Batliner, A., Marschik, P. B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F., Rathner, E.-M., Bartl-Pokorny, K. D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., Tzirakis, P., Zafeiriou, S. (2018). The Interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. In: Proc. of Interspeech, ISCA, Hyderabad, India, pp. 122–126.

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. Personality and Individual Differences 25(2), 167–177.

Singhi, A., Brown, D. G. (2014). On cultural, textual and experiential aspects of music mood. In: Proc. of ISMIR, ISMIR, Taipei, Taiwan, pp. 3–8.

Sobin, C., Alpert, M. (1999). Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. Journal of Psycholinguistic Research 28(4), 347–365.

Tato, R., Santos, R., Kompe, R., Pardo, J. M. (2002). Emotional space improves emotion recognition. In: Proc. of ICSLP, ISCA, Denver, CO, pp. 2029–2032.

Tolkmitt, F. J., Scherer, K. R. (1986). Effect of experimentally induced stress on vocal parameters. Journal of Experimental Psychology: Human Perception and Performance 12(3), 302–313.

Truong, K. P., Van Leeuwen, D. A., de Jong, F. M. G. (2012). Speech-based recognition of self-reported and observed emotion in a dimensional space. Speech Communication 54(9), 1049–1063.

Türk, U. (2001). The technical processing in smartkom data collection: a case study. In: Proc. of Eurospeech, ISCA, Aalborg, Denmark, pp. 1541–1544.

Utay, J., Miller, M. (2006). Guided imagery as an effective therapeutic technique: A brief review of its history and efficacy research. Journal of Instructional Psychology 33, 40–44.

Västfjäll, D. (2001). Emotion induction through music: A review of the musical mood induction procedure. Musicae Scientiae 5(1), 173–211.

Vaughan, B. (2011). Naturalistic emotional speech corpora with large scale emotional dimension ratings. PhD thesis, Dublin Institute of Technology.

Velten, E. (1968). A laboratory task for induction of mood states. Behaviour Research and Therapy 6(4), 473–482.

Ververidis, D., Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech Communication 48(9), 1162–1181.

Wasserstein, R. L., Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. The American Statistician 70, 129–133.

Westermann, R., Stahl, G., Spies, K., Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. European Journal of Social Psychology 26, 557–580.

Williams, C. E., Stevens, K. N. (1972). Emotions and speech: Some acoustic correlates. The Journal of the Acoustical Society of America 52(4B), 1238–1250.

Zhang, T., Hasegawa-Johnson, M., Levinson, S. (2004). Children's emotion recognition in an intelligent tutoring scenario. In: Proc. of Interspeech, ISCA, Jeju Island, Korea, pp. 1441–1444.

Zou, C., Huang, C., Han, D., Zhao, L. (2011). Detecting practical speech emotion in a cognitive task. In: Proc. of ICCCN, IEEE, Maui, HI, pp. 1–5.