Categorical vs dimensional perception of Italian emotional speech

Emilia Parada-Cabaleiro, Giovanni Costantini, Anton Batliner, Alice Baird, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Parada-Cabaleiro, Emilia, Giovanni Costantini, Anton Batliner, Alice Baird, and Björn Schuller. 2018. "Categorical vs dimensional perception of Italian emotional speech." In *Interspeech 2018, September 2-6, Hyderabad, India: Speech Research for Emerging Markets in Multilingual Societies*, edited by B. Yegnanarayana, C. Chandra Sekhar, Shrikanth Narayanan, S. Umesh, S. R. M. Prasanna, Hema A. Murthy, Preeti Rao, Paavo Alku, and Kumar Ghosh, 3638–42. Baixas: ISCA. https://doi.org/10.21437/interspeech.2018-47.

Interspeech 2018
2-6 September 2018, Hyderabad

Categorical vs Dimensional Perception of Italian Emotional Speech

Emilia Parada-Cabaleiro[1], *Giovanni Costantini*[2], *Anton Batliner*[1], *Alice Baird*[1], *Björn W. Schuller*[1,3]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Department of Electronic Engineering, University of Rome Tor Vergata, Italy
[3]GLAM – Group on Language, Audio & Music, Imperial College London, UK
emilia.parada-cabaleiro@informatik.uni-augsburg.de

Abstract

Culture and measurement strategies are influential factors when evaluating the perception of emotion in speech. However, multilingual databases suitable for such a study are missing, and there is no agreement on the most suitable emotional model. To address this gap, we present EmoFilm, a new multilingual emotional speech corpus, consisting of 1115 English, Spanish, and Italian emotional utterances extracted from 43 films and 207 speakers. We have performed a within-culture categorical vs dimensional perceptual evaluation, employing 225 native Italian listeners, who evaluated the Italian section of the database with the emotional states of anger, sadness, happiness, fear, and contempt. The aim of this study is to assess whether the emotional model (categorical or dimensional), taken as reference for measurement, influences a listener's perception of emotional speech, and—to what extent—both models are complementary or not. We show that the measurement strategy chosen does influence a listener's response, especially for some emotions, e. g., contempt. The confusion patterns typical of a categorical evaluation are not always mirrored by the dimensional assessment.

Index Terms: Listening test, multilingual corpus, film, emotion

1. Introduction

Some emotion categories, considered to be 'basic', have been thought to be universal [1], as they are expressed and perceived similarly by a diverse range of individuals across cultures. Nevertheless, the specific emotions that are known as basic [2], and the extent to which culture can impact expression and perception [1], are still a point of disagreement. Indeed, although a listener's level of accuracy in the perception of non-native emotional speech is higher than by chance [3], the identification accuracy of emotional states is substantially higher if made by native listeners [4]; this has been explained by the level of proximity, i. e., similarity, between the culture of the listener and that of the encoder [5]. Furthermore, the measurement strategy considered to evaluate listeners' perception of emotion could also influence their responses, since different subjects of the same culture may have diverse predispositions towards one model or another [6]. Although much research has been done on the suitability of the two main emotional models [7, 8, 9], i. e., the categorical [10] and the dimensional [11] model, no agreement has been reached so far whether the one or the other is more adequate for perceptual studies on emotional speech.

We present EmoFilm, a multilingual emotional speech corpus comprising 1115 utterances extracted from Italian, Spanish, and English films—original English, dubbed Italian and Spanish versions—thus an almost fully balanced distribution of speakers, utterances, and emotions. For a within-culture evaluation of the Italian part of the dataset, we contrasted categorical vs dimensional perception in three experiments with a total of 225 native Italian listeners. Our goal is to evaluate the extent to which categorical and dimensional assessments may be complementary methods to measure emotional speech perception. The rest of the manuscript is structured as follows: Section 2 presents related work; in Section 3, the dataset is described; Section 4 discusses the results of the perceptual study; and finally, in Section 5, conclusions and future work are given.

2. Related work

2.1. Emotional model: Categorical vs Dimensional

When evaluating the perception of emotional speech, the two models mainly considered are categorical (identifies emotions as discrete classes) [10], and dimensional (places them in a continuous hyper-space characterised by different 'dimensions'—mostly arousal and valence) [11]. The measurement strategy commonly used for the categorical model is the forced-choice test, as this minimises the spread of the responses; yet, it has been discussed that its accuracy may relate more to a 'discrimination' than to a 'recognition' task [12]. To deal with this risk, additional emotional labels not represented in the evaluated utterances such as 'neutral' or other emotional states (so called *distractor labels*), may be considered [13]. Even though it has been suggested that listeners' perception of emotional speech is categorical [7], this model has been criticised due to the one-to-one correspondence between categories and emotions, making it difficult to identify ambiguous emotional states, i. e., made up of different emotions manifested simultaneously [9].

The limitation of the categorical model for describing such states (also known as mixed or non-prototypical emotions) would mostly depend on an evaluation based on the use of a minimal set of categories. Yet, a reduced set of dimensions faces similar problems [14]. Indeed, the typical dichotomous representation of the dimensional model, i. e., arousal and valence, has also shown to be insufficient for discrimination between basic emotional states, such as anger and fear [15], leading to great overlap in the perception of mixed emotions [16]. The use of alternative dimensions, e. g., potency [8] or interaction [17], has previously been proposed; it has been suggested that at least four dimensions might be necessary, yet still not sufficient to discriminate between related non-prototypical emotions, e. g., shame, guilt, and embarrassment [18].

2.2. Emotional speech: Multilingual corpora

The emotional speech corpora (available for research) is biased towards some languages such as English [19]. A vast amount of languages are not considered, and commonly large speaker populations, e. g., Russian [20], are under-represented. Furthermore, multilingual datasets are quite uncommon [19]; therefore,

10.21437/Interspeech.2018-47

cross-cultural studies in machine learning are often forced to consider different monolingual corpora for training and evaluation [21]. Although this might be valuable for robust development and real-world applications [22, 23], an unbalanced distribution across speakers, utterances, and emotions may limit the performance of experimental techniques and bias results. Researchers have made efforts towards multilingual datasets, based on several types of speech, including acted emotional speech [4], natural [24], or speech gathered in so-called Wizard-of-Oz experiments [25]; yet so far, only a few cross-lingual and cross-cultural databases exist. Multilingual film replicas present the same emotional speech in a variety of languages and are well suited for the development of balanced multilingual corpora; however, these are still rare [26] and focus on single films and by that, on a limited number of speakers and emotions.

3. EmoFilm: A multilingual corpus

3.1. Emotional speech from films

Acted emotional speech creates acoustic profiles more intense than natural emotional speech [27]; yet, it has been critisised for not being natural enough [28]. Following Shakespeare's theory of acting, it can be seen as an intense representation of true reality [29]. On the other hand, humans are social beings that develop and interact in specific environments, following *display rules* [1], i. e., culturally prescribed rules that influence their emotional expressions in order to adequately handle social situations [30]—which can result in hiding or exaggerating emotions. Acting techniques such as the *Stanivslavsky method* [31] claim to guarantee the validity of acted emotions as they are based on self-induction techniques. All in all, the extent to which acted speech and natural speech can be considered to be realistic expressions of emotions is still unclear, especially from a cross-cultural perspective: Culture influences individuals in different ways. We can say that acting and displaying emotions in—original and dubbed—films represents one out of many types (styles) and follow their own specific 'display rules'. They are thus valid objects of investigation without representing everyday emotions in a one-to-one relationship.

Monolingual emotional speech databases from films have been collected and considered to be suitable for research purposes. Yet, such corpora are predominantly in English [32, 33, 34, 35, 21]; only rarely, other languages are taken into account, such as Turkish [36]. Multilingual databases from films have scarcely been collected [26], and languages such as Italian have, to the best of the authors' knowledge, not yet been considered.

3.2. Corpus description

A total of 43 films (from 1993–2009, English originals), were selected from genres including comedy, drama, horror, and thriller. Sequences with emotional content were chosen (collected under creative-commons) and segmented, extracting the audio in wave mono format (48 kHz sample rate and 16-bit). For selection, we started with the so-called 'big six', i. e., the basic emotions anger, sadness, happiness, fear, surprise, and disgust [10]. Surprise and disgust could be found rarely—something common in films [32]—contempt more often, which has been identified as a basic emotion too [37]. Due to this, we excluded surprise and disgust, and considered contempt as well as the remaining emotions from the 'big six', i. e., anger, sadness, happiness, and fear. The same 828 utterances (447 produced by females) have been extracted for each of the three languages: English (EN) produced by English actors in the original version, and over-dubbed versions by Italian (IT) and Spanish (SP) speaking actors, i. e., a total of 2,848 utterances.

A first selection was made manually by two affective computing researchers, rejecting clips affected by background noise and/or music. Subsequently, a pre-test was conducted employing 10 Italian listeners who evaluated the whole database; we rejected all clips with a rater-agreement lower than 6. This two-layered evaluation constitutes our reference ('gold standard'). Note that the same sentence may be considered as emotional in one language but not in another; thus, the number of utterances is not the same across the three languages. Furthermore, as the same dubbing actor can dub more than one original actor in different films, the number of actors in IT and SP is mostly lower than in EN. All in all, there are 207 speakers: 94 females (35 EN, 35 IT, and 24 SP) and 113 males (44 EN, 36 IT, and 33 SP). The final version of EmoFilm consists of 1115 clips with a mean length of 3.5 sec. (std 1.2 sec.), resulting in 360 EN clips (182 produced by females), with an average of 34.3 utterances per emotion (std 6); 413 IT audio clips (190 by females), with an average of 41.3 utterances per emotion (std 6.8); and 342 SP clips (165 by females), with an average of 35.9 utterances per emotion (std 9). The higher number of IT clips might be due to Italian being a more 'emotionally expressive' language; this could also relate to the pre-test made by Italian listeners, who may be better at perceiving emotions in their own language.

4. The perceptual study

4.1. Listening test design

Emotional Italian speech is under-researched [38, 39], thus, we start with a within-culture evaluation of the EmoFilm dataset by evaluating only the Italian instances. We conducted three listening tests in order to evaluate the extent to which measurement strategies might influence the perception of Italian emotional speech by natives: a dimensional test (T. 1) and two forced-choice categorical test (T. 2 and T. 3):

T. 1. Five-level rating-scale dimensional test (151 listeners). We evaluated arousal and valence in two scales from 0 to 4 (from less to more intense, and from negative to positive).

T. 2. Forced-choice categorical test with 'real' labels (151 listeners). We evaluated anger, sadness, happiness, fear, and contempt; thus, there is a one-to-one correspondence between the emotions of the corpus and the reference labels of the test.

T. 3. Forced-choice categorical test with 'distractor' labels (74 listeners). In addition to the 'real' emotions of the corpus, we considered two 'distractor' labels—surprise and neutral.

The three perception tests were performed using a computer-based interface (developed in the visual programming tool *Max MSP*[1]), with 225 Italian native listeners, who evaluated the Italian emotional speech of the database, i. e., 413 emotional utterances. T. 1 and T. 2 were performed together in the same perceptual session by 151 listeners (48 females), age between 19 and 42 years (mean 21.2, std 2.7). To avoid categories becoming anchors, for each utterance the listeners first did the dimensional and then the categorical annotation. T. 3 was done by 74 listeners, different from those who participated in T. 1 and T. 2 (40 females); age was between 20 and 26 years (mean 21.3, std 1.3). The subjects were all students of engineering and obtained credits for their voluntary participation in the test. In order to prevent listeners' fatigue, each test was divided into four sessions, of around 100 utterances each (T. 1 and T. 2 lasting together around 80–90 min; T 3 around 45–50 min), with

[1] https://cycling74.com/products/max/

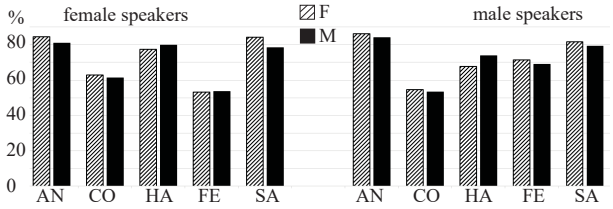

Figure 1: *Results of T. 2; recognised correctly in %: AN (anger), CO (contempt), HA (happiness), FE (fear), and SA (sadness); perceived by female (F) and male (M) listeners.*

Table 1: *Results of T. 2; 'recognised as' in %, cf. caption of Fig. 1, by all listeners. In each row, the 'gold standard' is given; in each column, 'recognised as' is given. For the number of evaluated instances, cf. caption of Table 2.*

T. 2	female speakers					male speakers				
	an	*co*	*ha*	*fe*	*sa*	*an*	*co*	*ha*	*fe*	*sa*
AN	82	11	03	03	01	85	10	02	02	01
CO	16	62	13	03	06	19	54	21	02	04
HA	05	10	79	02	04	03	19	72	03	03
FE	10	04	01	54	31	12	02	03	70	13
SA	02	03	03	12	80	03	04	02	11	80

different randomisation of the utterances for each listener.

4.2. Results and discussion

As our dimensional evaluation relates to the categorical results obtained in T. 2, we will start with analysing T. 2 and T. 3, followed by T. 1. From now on, when we refer to the gold standard, i. e., the emotion assigned to an utterance by at least 6 out of 10 Italian evaluators, we use upper-case, e. g., ANGER (AN); when we refer to the emotion category perceived by our listeners in e. g., T. 2, we will use lower-case, e. g., *anger* (*an*).

T. 2. Forced-choice categorical test with real labels

Results for T. 2 support prior research [21], showing that anger and sadness are more accurately perceived than fear (female speakers) and contempt (male speakers); cf. Figure 1. No significant differences have been found for gender: By computing Pearson's chi-square tests, the largest differences were shown in the perception of sadness (female speakers) and happiness (male speakers), both yielding p values of .27, i. e., substantially above the conventional threshold of $p < .05$. Thus, we do not analyse female and male listeners separately. Fear and contempt yield the worst recognition: Fear was mainly misclassified as sadness; contempt as happiness and anger (cf. Table 1). By computing the mean value of the upper and lower diagonal matrix, e. g., $12 + 31/2$ for the confusions of SADNESS with *fear* (SA-*fe*) and FEAR with *sadness* (FE-*sa*, female speakers), we conclude that the main confusion patterns are fear vs sadness (female speakers), contempt vs anger, and contempt vs happiness (both male speakers). Notice that although the number of cases can be low for some cells, e. g., SA-fe 12%, i. e., 4.9 cases, this relates to 151 listeners, i. e., 744 judgements.

Fear was consistently confused with sadness (female speakers), which might be 'culturally' explained, mirroring the idea that in western cultures females, unlike males, are socially allowed to express weakness but not aggression [40], and thus usually express fear as a passive and low aroused emotion, acoustically similar to the typical expression of sadness [41]. The confusion of contempt with anger may relate to the similarities between the former and the low aroused expression of the latter ('cold anger'). The confusion of contempt with happiness may relate to this emotion being often expressed via sarcasm—commonly indicated by false smiles and laughs, traits typical of positive emotions such as happiness, which present the inverse confusion pattern especially for male voices. Fleiss' kappa, as a measure of inter-rater reliability, displays a moderate agreement for both: female ($k = 0.44$) and male ($k = 0.45$) speakers.

Table 2: *Results of T. 3; 'recognised as' in %, by all listeners, of: 'true' labels (cf. caption of Fig. 1) and distractors: su (surprise) and ne (neutral); female and male speakers separated; number of evaluated instances (#) for each emotion.*

T. 3		Real labels					Distractor		#
		an	*co*	*ha*	*fe*	*sa*	*ne*	*su*	
female	**AN**	59	10	01	07	06	12	06	35
	CO	22	31	02	02	06	28	09	30
	HA	01	02	47	03	04	20	23	40
	FE	11	03	01	44	27	07	06	41
	SA	02	07	04	10	63	09	05	44
male	**AN**	63	17	01	04	02	08	05	44
	CO	26	22	05	01	02	33	11	38
	HA	02	07	37	03	04	26	21	54
	FE	05	02	01	64	15	05	08	38
	SA	03	03	02	09	68	10	04	49

T. 3. Forced-choice categorical test with distractor labels

As expected, *distractor labels* reduced the accuracy for all perceived emotions. This is shown by a high level of confusion with the distractors (see grey shadowing in Table 2), resulting in strong degradation of accuracy for happiness and contempt ($> 30\%$ difference between T. 1 and T. 2), medium degradation for anger and sadness ($< 25\%$ and $< 20\%$ difference respectively), and a smaller degradation for fear ($< 10\%$). This trend is displayed for both female and male speakers. Happiness was mostly misclassified as neutral and surprise, contempt commonly as neutral. In both cases, this confusion may be due to the listeners' interpreting neutral as an 'undecided' option [36]. Happiness misclassified as surprise relates to the similarity between the former and positive surprise. Contempt and surprise are indeed ambiguous emotions [2], difficult to identify without context and visual information. There is a fair inter-rater agreement for both: female ($k = 0.20$) and male ($k = 0.25$).

T. 1. Five-level rating-scale dimensional test

Now, we compare the responses of T. 2 with those of T. 1 for the main confusion patterns of T. 2: fear vs sadness (female speakers), contempt vs anger, and contempt vs happiness (both male speakers). In Table 3, above, we show, for the utterances correctly classified in T. 2, e. g., for FE-*fe* (54%), a 2-dim plot with Arousal (A) on the y-axis and Valence (V) on the x-axis with the sum of scores per cell; the darker the shadowing, the higher the frequencies; 'dimensional reference position' dim_{pos} (overall mean scores) is given as well. Table 3, below, displays the same information for the misclassified utterances, e. g., for FE-*sa* (31 %). The correctly identified, categorically different emotions of fear and sadness (female speakers) display similar 2-dim patterns and similar dim_{pos}; this is different, however, for contempt vs anger (male speakers). Thus, an unequivocal categorical difference is sometimes mirrored in dim_{pos} (markedly different for contempt and anger, male speakers), sometimes not (strikingly similar for fear and sadness, female speakers).

To quantify and evaluate these differences, we computed a one-way ANOVA employing the distances for Arousal (A_{dist})

Table 3: *Results of T. 1 for the utterances correctly identified (above, e. g., FEAR identified as fear—FE-fe), and misclassified (below, e. g., FEAR identified as sadness—FE-sa) in T. 2; % of utterances encoded in each matrix is given. 2-dim plot of Arousal–A (y-axis) and Valence–V (x-axis), from 0–lower to 4–higher level considering the confusion patterns: fear vs sadness (female speakers); contempt vs anger, and contempt vs happiness (both male speakers). Per cell, sum of listeners' scores, normalised to 0–100; grey shadowing represents frequencies (the darker, the higher); dimensional position dim_{pos} (overall mean score) given for arousal (A) an valence (V).*

female speakers — **fear vs sadness**			male speakers — **contempt vs anger**			male speakers — **contempt vs happiness**		
A	FE-*fe* (54%)	SA-*sa* (80%)	*A*	CO-*co* (54%)	AN-*an* (85%)	*A*	CO-*co* (54%)	HA-*ha* (72%)
4			**4**			**4**		
3			**3**			**3**		
2			**2**			**2**		
1			**1**			**1**		
0			**0**			**0**		
V	***0 1 2 3 4***	***0 1 2 3 4***	***V***	***0 1 2 3 4***	***0 1 2 3 4***	***V***	***0 1 2 3 4***	***0 1 2 3 4***
dim_{pos}	A:2.5, V:1.3	A:2.6, V:1.0	dim_{pos}	A:1.9, V:1.7	A:3.3, V:1.0	dim_{pos}	A:1.9, V:1.7	A:2.5, V:2.9
A	FE-*sa* (31%)	SA-*fe* (12%)	*A*	CO-*an* (19%)	AN-*co* (10%)	*A*	CO-*ha* (21%)	HA-*co* (19%)
4			**4**			**4**		
3			**3**			**3**		
2			**2**			**2**		
1			**1**			**1**		
0			**0**			**0**		
V	***0 1 2 3 4***	***0 1 2 3 4***	***V***	***0 1 2 3 4***	***0 1 2 3 4***	***V***	***0 1 2 3 4***	***0 1 2 3 4***
dim_{pos}	A:2.3, V:1.3	A:2.4, V:1.2	dim_{pos}	A:2.0, V:1.6	A:2.9, V:1.3	dim_{pos}	A:1.8, V:2.4	A:2.2, V:1.9

and for Valence (V_{dist}) between the dim_{pos} displayed in Table 3. We interpret Cohen's *d* as a measure of effect size. CO-*co* vs AN-*an* (male speakers) show a large $A_{dist} = 1.33$; $d = 1.51$ and a smaller $V_{dist} = 0.68$; $d = 0.77$. CO-*co* vs HA-*ha* (male speakers) are less similar for valence ($V_{dist} = 1.26$; $d = 1.70$) and more similar for arousal ($A_{dist} = 0.59$; $d = 0.65$). FE-*fe* vs SA-*sa* (female speakers) show a similarity both for valence ($V_{dist} = 0.29$; $d = 0.33$) and arousal ($A_{dist} = 0.08$; $d = 0.09$). Although even these distances are significant ($p < .05$ in Tukey's post hoc test), due to the relatively high number of cases, the small effect size for valence and especially for arousal suggest that these dimensions are not sufficient to discriminate between fear and sadness in female speakers; this confusion can be traced back to cultural factors (cf. Section 4. 2, T. 2).

Now we analyse whether the categorical confusion patterns of T. 2 are mirrored in the dimensional assessment. In Table 3, below, we display the 2-dim plot and the dim_{pos} for the confused categories, i. e., fear vs sadness (female speakers, FE-*sa*: 31%, SA-*fe*: 12%), contempt vs anger (male speakers, CO-*an*: 19%, AN-co: 10%), and contempt vs happiness (male speakers, CO-*ha*: 21%; HA-*co*: 19%). We want to know whether the categorical 'similarity' in T. 2—assuming that two categories are perceived as 'similar' when misclassified reciprocally—can be seen in the dimensional space as well. Yet, the confusion pattern contempt vs anger seems not to be mirrored in the dimensional space, since the utterances misclassified (CO-*an* and AN-*co*) show 2-dim representations similar to those of the utterances correctly classified (CO-*co* and AN-*an*, respectively), cf. Table 3, above, and dissimilar from each other. In T. 2, the prominent confusion was CO-*an*, i. e., more instances of CONTEMPT were misclassified as *anger* (19%) than vice versa (AN-*co* 10%). This is not mirrored in Table 3, below, since the dim_{dist} between CO-*an* vs CO-*co* is very small ($A_{dist} = 0.08$, $d = 0.09$; $V_{dist} = 0.09$, $d = 0.13$).

The confusion pattern contempt vs happiness is reciprocal in T. 2 (CO-*ha*: 21%; HA-*co*: 19%). This is mirrored by the 2-dim representation (cf. Table 3), where the utterances categorically misclassified (CO-*ha*; HA-*co*) seem to be perceived as dimensionally different from those correctly classified (CO-*co*; HA-*ha*, respectively), by showing an inverse tendency in the valence dimension (CO-*ha* perceived as more positive and similar to HA-*ha*; HA-*co* as more negative and similar to CO-*co*), cf. Table 3, above. A_{dist} is small: CO-*ha* vs CO-*co* ($A_{dist} = 0.14$; $d = 0.14$), and HA-*co* vs HA-*ha* ($A_{dist} = 0.35$; $d = 0.39$); V_{dist}, however, is large: CO-*ha* vs CO-*co* ($V_{dist} = 0.73$; $d = 1.02$), and HA-*co* vs HA-*ha* ($V_{dist} = 1.07$; $d = 1.33$). The confusion pattern fear vs sadness (prominently displayed in T. 2 in the direction FE-*sa*: 31%), is mirrored in the dimensional evaluation: The distance between FE-*sa* and FE-*fe* is very small for arousal ($A_{dist} = 0.16$; $d = 0.16$), and non-existing for valence ($V_{dist} = 0.00$; $d = 0.00$). This goes together with the similarity between these two categories when correctly identified, cf. Table 3, above. Thus, such similarities should not be interpreted as a sign of confusion but as a confirmation (supporting previous work [42]) of arousal and valence as insufficient dimensions in discriminating between some emotions.

5. Conclusions

Our study shows that expression of emotions seems to relate to gender, leading to confusion patterns that can be explained by cultural influences, e. g., fear is prominently misclassified as sadness for female speakers; contempt as anger and happiness for male. This categorical finding is, however, dimensionally supported only for valence in the confusion pattern contempt vs happiness, which shows that the measurement strategy influences listeners' evaluation. For the categorical model, a decline in accuracy was shown when considering distractor labels, suggesting that the accuracy of this model relates to the restricted amount of choices given to the listeners. Thus, each of the two models has advantages and disadvantages. This is mostly due to some 'systematic lacunas' in the models, which demonstrates the principle WYALFIWYG [43]—What You Are Looking for Is What You Get: When you leave out relevant dimensions and categories, you don't get them; and when you include irrelevant ones, you will get these as well. We plan a further evaluation of the models, by considering the other languages of `EmoFilm`.

6. Acknowledgements

This work was supported by the European Union's Seventh Framework under grant agreements No. 338164 (ERC StG iHEARu) and by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation Bavaria (ZD.B).

7. References

[1] P. Ekman, "Biological and cultural contributions to body and facial movement," in *Anthropology of the body*, J. Blacking, Ed. New York NY, USA: Academic Press, 1977, pp. 39–84.

[2] A. Ortony and T. J. Turner, "What's basic about basic emotions?" *Psychological Review*, vol. 97, no. 3, pp. 315–331, 1990.

[3] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural Psychology*, vol. 32, pp. 76–92, 2001.

[4] P. Laukka, "Vocal expression of emotion: Discrete emotions and dimensional accounts," Ph.D. diss., University of Uppsala, 2004.

[5] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychological Bulletin*, vol. 128, pp. 203–235, 2002.

[6] L. F. Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cognition & Emotion*, vol. 12, no. 4, pp. 579–599, 1998.

[7] P. Laukka, "Categorical perception of vocal emotion expressions," *Emotion*, vol. 5, no. 3, pp. 277–295, 2005.

[8] P. Laukka, P. Juslin, and R. Bresin, "A dimensional approach to vocal expression of emotion," *Cognition & Emotion*, vol. 19, no. 5, pp. 633–653, 2005.

[9] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1, pp. 5–32, 2003.

[10] K. R. Scherer and P. Ekman, "Expression and the nature of emotion," in *Approaches to emotion*, P. Ekman, Ed. Taylor & Francis Group, 1984, pp. 329–343.

[11] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[12] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1–2, pp. 227–256, 2003.

[13] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Communication*, vol. 16, no. 4, pp. 369–390, 1995.

[14] L. Devillers, S. Abrilian, and J.-C. Martin, "Representing real-life emotions in audiovisual data with non basic emotional patterns and context features," *ACII*, pp. 519–526, 2005.

[15] R. J. Larsen and E. Diener, "Promises and problems with the circumplex model of emotion," *Review of Personality and Social Psychology*, pp. 25–59, 1992.

[16] E. Mower, A. Metallinou, C. Lee *et al.*, "Interpreting ambiguous emotional expressions," in *Proc. of ACII*. Amsterdam, Netherlands: IEEE, 2009, pp. 1–8.

[17] A. Batliner, S. Steidl, C. Hacker *et al.*, "Private emotions versus social interaction: A data-driven approach towards analysing emotion in speech," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1, pp. 175–206, 2008.

[18] J. R. Fontaine, K. R. Scherer, E. B. Roesch *et al.*, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[19] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[20] V. Makarova and V. A. Petrushin, "RUSLANA: A database of russian emotional utterances," in *Proc. of Interspeech*. Denver, CO, USA: ISCA, 2002, pp. 2041–2044.

[21] N. Kamaruddin, A. Wahab, and C. Quek, "Cultural dependency analysis for understanding speech emotion," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5115–5133, 2012.

[22] B. Schuller, B. Vlasenko, F. Eyben *et al.*, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[23] F. Eyben, A. Batliner, B. Schuller *et al.*, "Cross-corpus classification of realistic emotions-some pilot experiments," in *Proc. of LREC*. Valettea, Malta: ELRA, 2010, pp. 77–82.

[24] R. Brueckner, M. Schmitt, M. Pantic *et al.*, "Spotting social signals in conversational speech over IP: A deep learning perspective," in *Proc. of Interspeech*. Stockholm, Sweden: ISCA, 2017, pp. 2371–2375.

[25] A. Batliner, C. Hacker, S. Steidl *et al.*, "'You stupid tin box' – children interacting with the AIBO Robot: A cross-linguistic emotional speech corpus," in *Proc. of LREC*. Lisbon, Portugal: ELRA, 2004, pp. 171–174.

[26] A. Braun and M. Katerbow, "Emotions in dubbed speech: An intercultural approach with respect to F0," in *Proc. of Interspeech*. Lisbon, Portugal: ISCA, 2005, pp. 521–524.

[27] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *JASA*, vol. 52, pp. 1238–1250, 1972.

[28] E. Douglas-Cowie, N. Campbell, R. Cowie *et al.*, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1–2, pp. 33–60, 2003.

[29] E. Douglas-Cowie, C. Cox, J.-C. Martin *et al.*, "Data and databases," in *Emotion-oriented systems: The HUMAINE handbook*, P. Petta, C. Pelachaud, and R. Cowie, Eds. Berlin, Germany: Springer, 2011, pp. 163–284.

[30] M. Lewis, J. M. Haviland-Jones, and L. Feldman Barrett, "Social functions of emotion," in *Handbook of emotions*, 3rd ed., A. H. Fischer and A. S. Manstead, Eds. New York, NY, USA: Guilford Press, 2008, vol. 3, pp. 456–468.

[31] J. Benedetti, *Stanislavski and the actor: The method of physical action*. New York, NY, USA: Routledge, 2013.

[32] T. S. Polzin and A. Waibel, "Emotion-sensitive human-computer interfaces," in *Proc. of ITRW*. Newcastle, UK: ISCA, 2000, pp. 201–206.

[33] I. Vasilescu, L. Devillers, C. Clavel *et al.*, "Fiction database for emotion detection in abnormal situations," in *Proc. of Interspeech*. Jeju, Jeju Island, South Korea: ISCA, 2004, pp. 2277–2280.

[34] C. Clavel, I. Vasilescu, L. Devillers *et al.*, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008.

[35] S. G. Karadoğan and J. Larsen, "Combining semantic and acoustic features for valence and arousal recognition in speech," in *Proc. of CIP*. Baiona, Spain: IEEE, 2012, pp. 1–6.

[36] C. Oflazoglu and S. Yildirim, "Recognizing emotion from Turkish speech using acoustic features," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–11, 2013.

[37] P. Ekman, "Universality of emotional expression? a personal history of the dispute," in *The expression of the emotions in man and animals*, 3rd ed., C. Darwin, Ed. New York, NY, USA: Oxford University Press, 1998, pp. 363–393.

[38] E. M. Caldognetto, P. Cosi, C. Drioli *et al.*, "Modifications of phonetic labial targets in emotive speech: Effects of the co-production of speech and emotions," *Speech Communication*, vol. 44, no. 1–4, pp. 173–185, 2004.

[39] G. Costantini, I. Iaderola, A. Paoloni *et al.*, "EMOVO corpus: An italian emotional speech database," in *Proc. of LREC*. Reykjavik, Iceland: ELRA, 2014, pp. 3501–3504.

[40] A. H. Fischer, "Sex differences in emotionality: Fact or stereotype?" *Feminism & Psychology*, vol. 3, no. 3, pp. 303–318, 1993.

[41] E. Parada-Cabaleiro, A. Baird, A. Batliner *et al.*, "The perception of emotions in noisified non-sense speech," in *Proc. of Interspeech*. Stockholm, Sweden: ISCA, 2017, pp. 3246–3250.

[42] P. N. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *JNMR*, vol. 33, pp. 217–238, 2004.

[43] A. Batliner, "Eine Frage ist eine Frage ist keine Frage. Perzeptionsexperimente zum Fragemodus im Deutschen," in *Zur Intonation von Modus und Fokus im Deutschen*, H. Altmann, A. Batliner, and W. Oppenrieder, Eds. Tübingen, Germany: Niemeyer, 1989, pp. 87–109.