Francesca Carbone
Università degli studi di Napoli "L'Orientale"

Identification of emotions by prosody and semantics in French spontaneous speech: A pilot study

ABSTRACT

When we talk about emotions, we refer to a complex and structured phenomenon that can influence and be influenced by external and internal elements involved in communication such as context, behavior and feelings of other individuals. During the expression of emotions, these internal and external factors can have effects on both verbal and non-verbal cues which can change according to the context and the intentions of the speakers. This complexity of the emotional phenomenon has induced researchers to focus the attention on individual cues (e.g., intonation, verbal content) in isolation from the others. Although it is possible that the transmission of emotional information is entrusted to a single type of cues, this is certainly the least frequent case in the spontaneous production of emotional speech, in which, usually, the emotional phonological information co-occurs with the other segmental features (e.g. syllables, words). Consequently, studies on individual emotional cues have not increased our understanding of how people integrate nonverbal and verbal cues in the expression of emotions. The aim of this study is to face this topic by assessing the contribution of prosodic parameters (i.e., fundamental frequency, intensity, speech rate) versus verbal content to the decoding of emotions by a perception test in which spontaneous speech is manipulated through the application of filters, validated in previous research. French listeners are asked to identify sentences that were annotated and selected from a French spontaneous corpus (EmoTV). Stimuli are presented in three versions: *integral* (non-manipulated), *verbal* (the prosodic information is hidden by applying filters) and *nonverbal* (the verbal content is filtered and masked). Results display a higher accuracy for integral stimuli suggesting that emotions are decoded by the interaction of the two different channels. This confirms the adequacy of a cognitive model in which verbal and nonverbal cues to emotions are theoretically integrated.

Keywords emotions, spontaneous speech, semantics, prosody

Received: 10.04.2019. Reviewed: 06.10.2019. Accepted: 24.11.2019. Published: 31.12.2019.

https://doi.org/10.32058/LAMICUS-2019-005

1. Introduction

During the spontaneous production of emotional speech, multiple channels (e.g., intonation, verbal content) are involved in conveying emotions and, although it is possible, it is quite rare that the transmission of emotional information is entrusted to a single type of cues in everyday life communication (Planalp 1999). To precisely capture the meaning, listeners must integrate information derived from segmental signals of language, like syllables, words, phrases (semantic cues), with suprasegmental information, such as voice tone and acoustic properties of the speech (e.g., prosodic cues, such as fundamental frequency, speech rate, intensity) (Planalp 1999; Nygaard & Queen 2008). Both of these aspects of spoken language, semantics and vocal cues, constitute essential components of successful interpersonal communication. Despite their joint importance, however, they have been studied separately by scientists and, consequently, research does not clarify how exactly they interact during communication and how they contribute to the recognition of emotions.

On the one hand, previous studies on emotional prosody have demonstrated that basic emotions can be quite accurately recognized through a single channel in a forced-choice response format (Pell et al. 2009, Paulmann & Kotz 2008; Juslin & Laukka 2003; Borod et al., 2000; Banse & Scherer 1996; Murray & Arnott 1993; Ekman 1992; Scherer et al. 1991). For instance, Banse and Scherer (1996) explored how accurately emotions can be conveyed only by prosody in sentences that did not contain any meaningful lexical-semantic information (pseudo-sentences). Their results showed that listeners are able to detect the emotional state of their communicative partner only by the acoustic modulation of the voice. On the other hand, studies on emotional semantics have pointed out the power of emotional words, suggesting that emotionally connotated words are processed faster and more easily than neutral ones, even when they are presented without any prosodic cue (Kousta et al. 2009; Schacht & Sommer 2009; Scott et al. 2009; Estes & Verges 2008; Kanske & Kotz 2007; Bradley & Lang 1999).

Studies dealing with the transmission of emotions from different channels have mostly presented multi-modal emotional stimuli (generally individual words) in paradigms that include mismatching or incongruent information between channels, frequently using a Stroop task methodology (Filippi et al. 2017; Kreifelts et al. 2007, De Gelder & Bertelson 2003; De Gelder & Vroomen 2000). When prosody and semantics are compared, outcomes are varied and do not reveal a unique trend. Electrophysiological data have shown that the semantics of an utterance is processed before the prosodic information during emotion processing (e.g., Kotz & Paulmann 2007). In contrast, behavioral studies have illustrated how prosody can give rise to

potential biases in the processing of emotional verbal content, even when conflicting with verbal content and faces simultaneously (Filippi et al. 2017). Furthermore, the selection of emotional meaning is significantly better when the voice tone is congruent with affective meaning, indicating that emotional prosody may be used to process the linguistic content of lexical items (Paulmann & Pell 2011; Nygaard & Lunders 2002). Moreover, the dominance of the two channels (i.e. semantics and prosody) could be related to cultural differences. For instance, Kitayama and Ishii (2002; Ishii et al. 2003), who applied an auditory Stroop task, have found that interference effects between word meaning and vocal emotion could be mediated by culture and language. In their task, comparing judgments of Japanese and American English listeners on the recognition of emotional words, it was found that for the first group (Japanese) vocal emotion interfered with judgments of word meaning to the greatest extent, while for the second (American English), word meaning interfered with judgments of vocal emotion to the greatest extent.

Although the interrelation of prosody and semantics in the transmission of the emotions is still not clear, however, there is some evidence that emotional expressions encoded by more than one information channel are recognized with greater accuracy, consistent with the idea of a multi-modal advantage (Frühholz et al. 2016; Milesi et al. 2014; Paulmann & Pell 2011). Paulmann & Pell (2011) ran an identification task using different kinds of stimuli (i.e., facial expressions, semantic information conveyed by texts, emotional prosody). Their results confirmed that emotion recognition is significantly better in response to multi-modal versus uni-modal stimuli. From a cognitive point of view, this advantage suggests an automatic integration of emotional information across channels, which enhances the accessibility of emotion-related knowledge in memory (cf. *emotion nodes* or *concepts*, Russell & Lemay 2000; Halberstadt et al. 1995).

Studies using spontaneous emotional speech samples to compare multi-modal versus uni-modal versions are even more sporadic (Huang et al. 2017; Kim & Provost 2016; Jürgens et al. 2013), which could be attributed to a number of factors, including background noise and ethical reasons (Johnstone & Scherer 2000). In fact, basic emotions are still rare events in everyday conversations and it is difficult to observe them being expressed in uni-modal manner (Cowie & Cornelius 2003). In order to overcome these issues, a number of researchers have used stimuli produced by actors in laboratories. This procedure, however, has often led to more stereotyped voices, which do not always reflect the spontaneous expression of emotions (Drolet et al. 2012, Laukka et al. 2012; Bänziger & Scherer 2010; Scherer 2003). Since natural stimuli are considered more ecological (Scherer 2003), a possible solution to obtain spontaneous speech in uni-modal manner is to apply filters to hide the verbal content (e.g. *Low Pass filters* block the high

components of a sound and pass the low-frequency components hiding linguistic cues; Scherer & Oshinsky 1977; Van Bezooijen & Gooskens 1999 Knoll & Costall 2009) or to monotonize the prosodic contour (e.g. *Pitch Synchronous Overlap and Add* works by dividing the speech waveform in small overlapping segments; the pitch of a signal is changed by moving the segments further apart to decrease the pitch or closer together to increase the pitch, Moulines & Verhelst 1995; Van Bezooijen & Gooskens 1999).

This pilot study is a part of a broader research whose primary objective is to analyze the relationship of prosody and semantics in conveying emotions. More specifically, the goal of the present work is to test the contribution of semantics and prosody, separately and in combination, to the identification of emotions. In order to pursue this aim, in the present investigation natural speech samples were carefully selected from a French corpus of spontaneous speech, consisting of emotionally laden interviews. To separate the different channels and to create uni-modal stimuli, filters approved by previous literature (see above) were applied. I attempted to investigate how the recognition of spontaneous emotional sentences changes with the application of filters, resulting in the formation of uni-modal stimuli – i.e., conveyed by the prosodic or the semantic channel only – to express emotions. To reach this goal, listeners were presented with an identification task in a forced-choice response format, while reaction times were collected. Moreover, I aimed to test whether perceptual differences in intensity (i.e. how intense an emotion is felt) exist between multi-modal (not manipulated) and uni-modal emotional stimuli. In line with the literature promoting an enhanced recognition in a multi-modal condition (Kotz & Paulmann 2007), I expected to find a clear trend showing higher recognition rates for stimuli in which prosody and semantics are preserved. In light of the data which show that emotions are often explicitly categorized more easily from the semantic content of an utterance or from facial expressions (Borod et al., 2000; Johnstone & Scherer 2000; Pell 2002; Paulmann & Kotz 2008), I hypothesized to find lowest recognition rates when only prosodic cues were present. Concerning intensity, I expected that the integral stimuli (i.e. non-manipulated) would be evaluated as more intense.

2. Methods

This study consisted of two phases: a pre-experiment to annotate and validate the stimuli, followed by the main experiment, which tested the contribution of semantics and prosody to emotion identification in spontaneous speech.

2.1. Norming study

The annotation and the selection were based on the results of an identification test in which listeners had to indicate, in a forced choice task, the emotion expressed by positive and negative sentences, which were emotionally laden.

2.1.1. Material

The corpus consisted of 48 sentences with a duration between 2 and 5 seconds (5.2 syllables per second) taken from two different corpora: 30 sentences were extracted from EmoTV (Abrilian et al. 2005), while the other 18 were selected from the *Interactional Data Corpus* (CID, Bertrand et al. 2008). EmoTV is a French corpus of 100 videos taken from television interviews on emotionally intense topics; while the CID is composed of 8 hours of semi-spontaneous dialogues recorded in a soundproof room at the *Laboratoire Parole et Langage* (LPL-CNRS, Aix en Provence, France) in 8 dialogues in which conversations revolve around two pre-established central themes, namely the narration of unusual events and conflict situations.

Firstly, the 30 extracts of 2 to 5 seconds produced only by female speakers (18 different interviewees) were cropped from 20 videos of EmoTV with discrete acoustic quality and converted into audio files using Praat (Boersma 2001). Following the same criteria, the remaining 18 stimuli were selected from the CID (5 female speakers). Given that in the spontaneous production of the emotional discourse, the expression of emotional states is rarely conveyed through a single channel of communication (Planalp 1999), the verbal content of the selected sentences was variable and emotionally laden (e.g. EmoTV: *j'étais profondément en colère avec elle d'avoir menti* 'I was deeply angry with her for lying', CID: *il a pas du tout eu le temps pour moi* 'he had no time for me at all'). Totally, 48 short sentences conveying emotions both by prosody and semantics were collected.

2.1.2 Procedure

The perceptual validation of the stimuli was carried out in two steps. Firstly, two French native speakers (PhD students of linguistics, aged between 25 and 30) listened to the 48 stimuli and classified the sounds following the valence as criterion (namely the positivity or negativity of an emotion, Barrett 1998). They were instructed to classify the stimuli considering both their prosody and semantics. They found 24 utterances expressing positive emotions and 24 expressing negative ones. Secondly, eighteen French native speakers (12 F and 6 M), aged between 20 and 35, were recruited for a perception task. The experiment took place in a silent room at the Laboratoire Parole et Langage (LPL-CNRS, Aix en Provence). Listeners were instructed to sit comfortably and wore professional headphones. A

short practice session preceded the experiment consisting of two tasks. Firstly, they were asked to identify the emotion conveyed by the stimuli by pressing one of the six buttons on the keyboard, each corresponding to one of six basic emotions (i.e. *joy, anger, fear, sadness, neutral state, surprise*) respectively. The name of the emotions appeared on the computer screen. Once they identified the emotion, the listeners were asked to assess how well the chosen emotion was represented by the sounds on a 5-point Likert scale. The range of values was between 1 and 5 (with 1 corresponding to "not well represented" and 5 to "very well represented"). The stimuli were reproduced by a laptop using PERCEVAL (André & Ghio 2003). Each stimulus was heard twice and presented in 3 blocks; the presentation of the stimuli was randomized.

2.1.3. Results

Statistical analyses were performed on 864 observations (18 listeners * 48 stimuli) on the R software (version 3.5.1). To test the agreement between listeners in the categorization of stimuli, Kappa of Fleiss (1971) was applied (k = 0.0293, z = 3.21, p <0.0001), showing a limited agreement among the listeners. Also, the internal agreement was equally poor for both corpora (CID: k = 0.11, z = 8.84, p < 0.01; EmoTV: k = 0.0516, z = 4.27, p = 0.0000199). Concerning representativity, a repeated measure ANOVA with *emotions* (5) as within-subjects factor showed that utterances expressing joy and sadness were evaluated as more representative (F (5.4) = 5.32, p < 0.0001). Moreover, a Student t-test was performed to verify if one of the two corpora (EmoTV/CID) was judged to be more significant. No significant differences emerged (t = 2.55, df = 862, p = 0.798). However, EmoTV collected a tiny advantage in the mean value of representativity (see Figure 1).

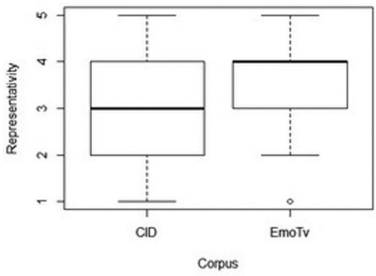

Figure 1. Boxplot of representativity value (y-axis) for the factor corpus (i.e. EmoTV, CID; *x-axis)*

Since for both corpora the agreement was scarce and there were no significant differences between them, I decided to select the corpus presenting the more "spontaneous" sentences. Utterances were taken from EmoTV because they were spontaneously spoken in a natural setting, while sentences from the CID were task-orientated and recorded in a laboratory. This choice is also

supported by the evidence of the small advantage in representativity scores for EmoTV. Indeed, I selected from EmoTV stimuli judged in the same way by 80% to 100% of the listeners and obtained high scores in terms of representativity (see Table 1).

Table 1. Verbal content, agreement percentage and representativity score (mean) of the sentences chosen for the main experiment

Verbal content	Emotion	Agreement	Representativity score
Forte odeur /âcre/ âcre/ je pus plus respirer dans ma maison / c'était infernal 'Strong smell / acrid / acrid / I could not breathe in my house / it was hellish'	*Colère* 'Anger'	89%	3.9
J'étais profondément en colère avec elle d'avoir menti 'I was deeply angry with her for lying'	*Colère* 'Anger'	89%	3.7
Ça fait quand même plus de deux mois /c'est très longue / très longue /les gens en ont peut-être marre 'It's still over two months / it's very long / very long / people may be fed up'	*Colère* 'Anger'	89%	3.9
Plus les enfants grandissent/ s'ils sont pas suivi /c'est/ c'est très idiot 'Children grow up more and more/ if they are not followed / it's / it's very silly'	*Tristesse* 'Sadness'	94%	3.8
Il ne dit pas du tout la vérité 'He does not tell the truth at all'	*Tristesse* 'Sadness'	94%	3.8
Pour une maman/ c'est / il n'y a pas pire de voir ses enfants souffrir de cette façon-là 'For a mom / there's/ nothing is worse than to see her children suffer in this way'	*Tristesse* 'Sadness'	89%	4.1
On essaie d'avoir une ville jolie /agréable / accueillante 'We try to have a nice city / pleasant / welcoming'	*Neutre* 'Neuter'	89%	3.9
Pour le tourisme d'Andorre, Pas de Calais, et voyez ce qu'on peut avoir 'For tourism in Andorra, Pas de Calais, and see what we can have'	*Neutre* Neuter	83%	3.8
Bon / ça se passe bien/ ça fait un petit salut/ c'est intéressant 'Well/ it's okay /it was a short greeting/ it's interesting'	*Joie* 'Joy'	94%	4.5
Là aujourd'hui / tout me plait 'There, today / I like everything'	*Joie* 'Joy'	100%	4.5
Et le matin/ quand on se lève/ on se dit/ la vie est belle And in the morning/ when we get up/ we tell ourselves/ life is beautiful	*Joie* 'Joy'	94%	3.7
Anglement /c'était une famille très apathique / forcément ça choque Anglement /it was a very apathetic family /this is necessarily surprising	*Surprise* 'Surprise'	94%	4.1

To sum up, for the main experiment, 12 stimuli (pronounced by 10 female speakers) were selected from EmoTV. They consisted in short sentences with a duration from 2 to 5 seconds. They always contained a congruent set of cues to express one of the following emotions: 3 stimuli expressed *joy*, 3 *sadness*, 3 *anger*, 1 *surprise*, 2 *neutral state*. Utterances conveying surprise were inserted as fillers to balance the number of positive and negative emotions. Sentences judged as neutral were used as control states.

2.2. Main experiment

2.2.1. Corpus

The 12 speech fragments selected from the norming study were presented to the listeners in three different versions:
1. *Nonverbal.* By lowpass filtering the signal at 350 Hz (standard value) the speech was rendered unintelligible. In this condition, the only cue provided to the listeners is prosodic information.
2. *Verbal.* By means of electronic monotonization (at a fixed value of 220 Hz, i.e., the mean pitch over all listeners) the intonation was removed from the signal. Monotonization was affected through *Pitch Synchronous Overlap and Add* (PSOLA) analysis and resynthesis (Moulines & Verhelst 1995). Verbal information is maintained, and so are temporal and loudness variation. The fragments are completely intelligible, but perfectly monotonous.
3. *Integral.* In this version all prosodic and verbal information is present.
As a result, the corpus included 34 stimuli, 24 of which were uni-modal (i.e. 12 preserving only verbal content, 12 presenting only prosody), expressing 4 emotions (i.e. joy, anger, sadness, surprise) and neutral state.

2.2.2. Procedure

Eighteen French native speakers (10 F and 8 M), aged 20 to 35, were selected for the test. The experiment took place in a silent room at the University of Naples "L'Orientale". The listeners were volunteers, recruited through social networking sites. They were asked to sit in a comfortable chair and wore professional headphones. Concerning the task, they were firstly asked to identify the emotion conveyed by the stimuli by pressing one of the six buttons on the keyboard corresponding to the list of five basic emotions (i.e. *joy, anger, sadness, neutral, surprise*) appearing on the computer screen. Secondly, after identifying the emotions, listeners were asked to rate the intensity of the emotion expressed by the sounds on a 5-point Likert scale. The range of the values was from 1 to 5 (with 1 meaning "not intense emotion" and 5 "very intense emotion"). Stimuli were played on a laptop by PERCEVAL software

(André & Ghio 2003). The fragments were presented in separate blocks of *integral*, *verbal*, and *nonverbal* fragments; in two sequences and in two random counterbalanced orders. Prior to starting the experiment, instructions were displayed on the screen and, in order to get familiarized with the experimental procedure, listeners ran three practice trials.

3. Results

3.1. Correct identification

Statistical analyses were performed on 648 observations (18 listeners * 36 sentences). A logistic regression (*logit model*) was used to describe the relationship between the correct identification and the different kind of emotions and cues. This model is particularly appropriate for categorization tasks because it is specifically designed for analyzing binary (e.g. yes/no) and categorical response variables (Hosmer et al. 1989). In this model, the dummy variable was *correct/incorrect*, while the factors *emotions* (5 levels: *neutral state*, *joy*, *anger*, *sadness*, *surprise*), *type* (3 levels: *integral*, *verbal*, *nonverbal*) and *intensity* (5 points Likert scale) were the predictors, allowing for interaction effects. Final model according to AIC stepwise variable selection is reported in Table 2, describing the coefficients associated to each predictor variable (first column).

Table 2. Coefficients of LOGIT model estimated for the factor emotions (anger, joy, surprise, sadness), type (nonverbal, verbal), intensity

	B	Odds Ratio	Std. Error	z value	Pr(>|z|)
(Intercept)	0.4243	1.5285	0.3843	1.104	0.269649
Colère 'Anger'	-0.5816	0.5590	0.2694	-2.159	0.030882
Joie 'Joy'	-0.9623	0.3820	0.2979	-3.231	0.001236
Surprise 'Surprise'	0.7069	2.0278	0.3624	1.951	0.051109
Tristesse 'Sadness'	-0.9996	0.3680	0.2720	-3.676	0.000237
Nonverbal	1.6170	5.0380	0.2368	6.829	0.0001
Verbal	0.3480	1.4163	0.2280	1.526	0.126892
Intensity	-0.2798	0.7559	0.0824	-3.396	0.000684

In line with the hypotheses and the previous literature, as shown by the bar plot (Figure 1), the emotions were better identified in the *integral* version (*M*: 75.2 %) followed by *verbal* type (*M*: 66.2 %). The lowest recognition rates are found when only prosodic cues were present in utterances (*M*: 36.8 %). Significant effects are found for *nonverbal* stimuli expressing sadness ($p<0.0001$) neutral state ($p<0.0001$), anger ($p<0.0001$), surprise ($p<0.0001$). For these stim-

uli the recognition rates are lower (sadness: 49 %; anger: 38%; neutral: 17%, surprise: 14%).

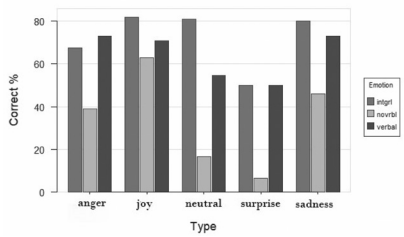

Figure 2. Listener's correct identification percentage (y-axis) for the five emotions (x-axis), split by type (i.e. integral, nonverbal, verbal)

3.2. *Intensity*

Concerning intensity, a repeated measures ANOVA revealed a significant main effect for the *emotion* ($F(5.37) = 38.06$, $p<0.0001$) and *type* ($F(9.74) = 37.61$, $p<0.0001$) factors. As shown in Figures 3-4, the more intensely judged stimuli were globally better recognized (t-test: $t=6.490$, $df=645$, $p<0.0001$). More importantly, the *integral* stimuli were considered the most intense ($M=3.725581$, $SD=1.043022$), followed by the *verbal* ($M=3.50463$, $SD=1.153348$), while the *nonverbal* stimuli ($M=2.916667$, $SD=1.138134$) were judged the least intense ($p<0.0001$) confirming the predictions made about intensity.

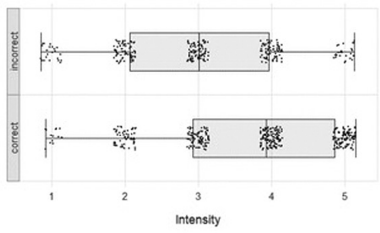

Figure 3. Boxplot of intensity value (x-axis) for the factor correct/incorrect *(y-axis)*

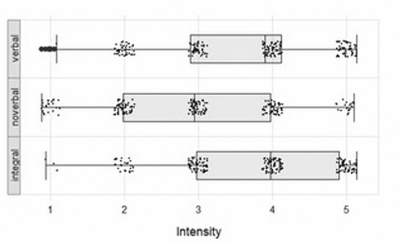

Figure 4. Boxplot of intensity value (x-axis) for the factor type *(i.e.* integral, nonverbal, verbal; *y-axis)*

3.3. *Reaction time*

Before setting up the statistical analyses, because of the positive asymmetry of data (skewness=3.899778), a log transformation was applied in order to normalize them (skewness of logarithm of reaction time =0.04336222). A repeated measures ANOVA with *type* as the within-factor (three different levels: i.e. integral, verbal, nonverbal) was performed on reaction time (log trans-

formed). Analyses revealed significant main effect (F (7.4) =38.69, p<0.0001) with important differences between *integral-nonverbal* stimuli (p<0.0001) and *integral-verbal* (p<0.0001); while a not-so-significant effect is found by comparing *verbal-nonverbal* (p=0.327). Boxplot (Figure 5) shows how listeners identified *integral* stimuli faster than the other ones. The reaction time was slowest for stimuli presenting only prosodic cues.

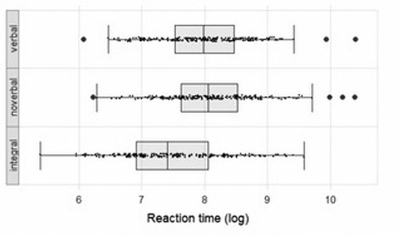

Figure 5. Boxplot of log reaction time data (x-axis) for the factor type *(i.e.* integral, nonverbal, verbal; *y-axis)*

4. Discussion

The aim of the study was to investigate the contribution of semantic and prosodic cues separately and when they are both present, in the identification of emotions in spontaneous speech. In particular, the goal was to test whether emotion recognition is facilitated by multi-modal versus uni-modal stimuli. This was attempted by filtering utterances from spontaneous speech. In addition, I explored whether the perception of an emotion's intensity changed depending on the channel of transmission. In contrast to most of the methodological approaches adopted by previous studies, the issue was investigated by using spontaneous congruent stimuli, which always conveyed the same emotions as represented by both prosody and semantics. In line with Paulmann and Pell (2011), my results showed that, in spontaneous speech, the recognition of emotions increases in accuracy when multiple channels cooperate to convey them in a congruent manner, i.e. *integral* stimuli were recognized significantly better than uni-modal stimuli. This aligns with previous studies which show that emotional judgments tend to improve when more than one source of congruent information about the intended emotion is available (Collignon et al. 2008; Pell 2005; De Gelder & Vroomen 2000; Massaro & Egan 1996). This assumption is also confirmed by the reaction times, showing that the access to emotions is faster when they are conveyed by both prosodic and semantic cues. From a cognitive point of view, since information in each emotional channel activates shared conceptual knowledge about an emotion (*emotion concepts*, cf. Bower 1981; Russell & Lemay 2000; Halberstadt et al. 1995), it is possible that, during emotion identification tasks, knowledge about emotions is more readily accessible when more congruent cues cooperate to convey the emotional states (Paulmann & Pell 2011). In line with this hypothesis, Borod et al. (2000) has suggested two steps

in the processing of emotions: first, emotional channels are independently processed by separate sensory modality systems and, after that, they pass through a *general affective processor.* Following this, it could be assumed that prosodic and verbal cues are incorporated during emotional recognition processes, leading to systematically higher accuracy rates, as observed here. Interestingly, confirming results of the study of Paulmann and Pell (2011), the described process seems to also involve the recognition of spontaneous neutral states that were better identified when presented in the *integral* version with both prosodic and semantic cues. This result supports those received by Cornew et al. (2009). In two gating experiments, they showed that participants identified neutral prosody more rapidly and accurately than happy or angry prosody, suggesting an actual prominent advantage for processing neutral over emotional content.

Concerning intensity, in line with the literature (Wells et al. 2016), the results illustrated that the integral stimuli were perceived as more intense, meaning that the interaction between the prosodic and semantic channel could co-occur to strengthen the perception of emotional intensity. Not surprisingly, the stimuli judged as more intense were identified better, signifying that perceived emotional intensity might be a strong signal for discrimination when usable (Audibert et al. 2008).

Moreover, this study showed that spontaneous emotions conveyed by semantics were recognized more accurately and faster than through prosodic information. A possible reason may be that there are some differences in how prosody and semantics trigger *emotion nodes* during the expression of emotion (Bower 1981; Paulmann & Pell 2011; Halberstadt et al. 1995; Russell & Lemay 2000). Previous studies have argued that emotions are conveyed through prosody in a categorical manner over a protracted time period (Juslin & Laukka 2003; Rigoulot et al. 2013). In contrast, the semantic content of stimuli reflected, in a prototypical manner, the selected emotions, therefore activating the underlying emotional conceptual knowledge more strongly and faster (Paulmann & Pell 2011). Furthermore, the low recognition rates collected for stimuli presenting only prosodic parameters (*M*: 36%) could be explained in light of recent research on the prosody of authentic emotions. In fact, the acoustic analyses on spontaneous speech presented weaker acoustic differences between the emotions for the authentic expressions compared to studies on portrayal (Laukka et al. 2012), compromising the capacity of listeners to detect emotions in absence of other speech signals and the context. Future studies are necessary to verify whether this semantic dominance is culture/language-dependent, as suggested by studies on Japanese and American English listeners (Ishii et al. 2003; Kitayama & Ishii 2002). Another important issue is the gender variation in the perception of emotional prosody and semantics. Schirmer and Kotz (2003) in an ERP study have shown a

better integration of emotional cues for women than men. This study, as a pilot, did not address this matter but it will be investigated in the wider study by balancing the number of male and female participants and by presenting to the listeners stimuli pronounced both by men and women.

In conclusion, this study showed that the concept of multimodal advantage could be extended also for the recognition of spontaneous emotions. Although the identification of emotions could be possible by separate channels of transmission, congruent and simultaneous information conveyed by prosodic and semantic patterns at the same time increase the accuracy and speed in perceiving emotions. As underlined, this could confirm the adequacy of applying a cognitive model of emotional processing in which verbal and nonverbal cues to emotions are integrated.

Acknowledgments

This study was set up with a PhD grant of the author from University of Naples "L'Orientale" (Italy) in a context of a Doctoral Thesis whose Director is Prof. Alberto Manco. The Perceptual test (*norming*) was conducted at the *Laboratoire Parole et Langage* (LPL) in Aix en Provence (France). Thanks to Caterina Petrone for helping with the stimuli selection and Francesco Santelli for some comments on statistical analyses.

References

Abrilian, Sarkis, Jean-Claude Martin, Laurence Devillers 2005: A corpus-based approach for the modeling of multimodal emotional behaviors for the specification of embodied agents. *Proceeding from HCI International Las Vegas*, USA.

André, Carine, Alain Ghio 2003: PERCEVAL: une station automatisée de tests de PERCeption et d'EVALuation auditive et visuelle. *Travaux Interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence 2003* (TIPA) 22, 115-133.

Audibert, Nicolas, Véronique Aubergé, Albert Rilliard 2008: Acted vs. spontaneous expressive speech: perception with inter-individual variability. *Programme of the Workshop on Corpora for Research on Emotion and Affect*, 23-26.

Banse Rainer, Scherer Klaus 1996: Acoustic profiles in vocal emotion expression, *Journal of Personality and Social Psychology* 70(3), 614-636.

Bänziger, Tanja, Klaus R. Scherer 2010: Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for Affective Computing: A Sourcebook*, 271-294.

Barrett, Feldman Lisa 1998: Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion* 12, 579-599.

Bertrand, Roxane, Philippe Blache, Robert Espesser, Ferré, Christine Meunier, Béatrice Priego-Valverde, Stéphane Rauzy 2008: Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues* 49(3), 105-134.

Boersma, Paul 2001: Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341-345.

Borod, Joan C., Lawrence H. Pick, Susan Hall, Martin Sliwinski, Nancy Madigan, Loraine K. Obler, Joan Welkowitz, Elisabeth Canino, Hulya M. Erhan, Mira Goral, Chris Morrison, Matthias Tabert 2000: Relationships among facial, prosodic, and lexical channels of emotional perceptual processing. *Cognition & Emotion* 14, 193-211.

Bower, Gordon H. 1981: Mood and memory. *American Psychologist* 36, 2, 129-148.

Bradley, Margaret M., Peter J. Lang, 1999: Affective norms for English words (ANEW): Instruction manual and affective ratings. *Technical Report C-1, the Center for Research in Psychophysiology*, 30(1), 25-36.

Collignon, Oliver, Frédéric Gosselin, Sanjoy Roy, Dave Saint-Amour, Maryse Lassonde, Franco Lepore 2008: Audio-visual integration of emotion expression. *Brain Research* 1242, 126-135.

Cornew, Lauren, Leslie Carver, Tracy Love 2009: There's more to emotion than meets the eye: A processing bias for neutral content in the domain of emotional prosody. *Cognition & Emotion* 24(7), 1133-1152.

Cowie, Roddy, Randolph Cornelius 2003: Describing the emotional states that are expressed in speech. *Speech Communication* 40(1/2), 5-32.

De Gelder, Beatrice, Jean Vroomen 2000: The perception of emotions by ear and by eye. *Cognition & Emotion* 14(3), 289-311.

De Gelder, Beatrice, Paul Bertelson 2003: Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences* 7(10), 460-467.

Drolet, Matthis, Ricarda I. Schubotz, Julia Fischer 2012: Authenticity affects the recognition of emotions in speech: behavioral and fMRI evidence. *Cognitive, Affective, & Behavioral Neuroscience* 12(1), 140-150.

Ekman, Paul 1992: An argument for basic emotions. *Cognition & Emotion* 6(3/4), 169-200.

Estes, Zachary, Michelle Verges 2008. Freeze or flee? Negative stimuli elicit selective responding. *Cognition* 108(2), 557-565.

Filippi, Piera, Sebastian Ocklenburg, Daniel L Bowling, Long Heege, Onur Güntürkün, Newen Albert, Bart de Boer 2017: More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing. *Cognition & Emotion* 31(5), 879-891.

Fleiss, Joseph L. 1971: Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5), 378-382.

Frühholz, Sascha, Wiebke Trost, Sonja A Kotz. 2016: The sound of emotions—Towards a unifying neural network perspective of affective sound processing. *Neuroscience & Biobehavioral Reviews* 68, 96-110.

Halberstadt, Jamin B., Paula M. Niedenthal, Julia Kushner 1995: Resolution of lexical ambiguity by emotional state. *Psychological Science* 6(5), 278-282.

Hosmer, David W., Stanley Lemeshow, Rodney X. Sturdivant 1989: The multiple logistic regression model. *Applied Logistic Regression* 1, 25-37.

Huang, Gao, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, July 2017: Densely connected convolutional networks. *Proceedings from the IEEE conference on computer vision and pattern recognition*, 4700-4708.

Ishii, Keiko, Alberto Jose Reyes, Shinobu Kitayama 2003: Spontaneous attention to word content versus emotional tone: Differences among three cultures. *Psychological Science* 14(1), 39-46.

Johnstone, Tom, Klaus R. Scherer 2000: Vocal communication of emotion. *Handbook of Emotions* 2, 220-235.

Jürgens, Rebecca, Matthis Drolet, Ralph Pirow, Elisabeth Scheiner, Julia Fischer 2013: Encoding conditions affect recognition of vocally expressed emotions across cultures. *Frontiers in Psychology* 4, 111.

Juslin, Patrik N., Petri Laukka, 2003: Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin* 129(5), 770.

Kanske, Philipp, Sonja A. Kotz 2007: Concreteness in emotional words: ERP evidence from a hemifield study. *Brain Research* 1148, 138-148.

Kim, Yelin, Emily Mower Provost 2016: Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions. *Proceedings from the 18th ACM International Conference on Multimodal Interaction, October 2016*, 92-99.

Kitayama, Shinobu, Keiko Ishii 2002: Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition & Emotion* 16, 1, 29-59.

Knoll, Monja A., Maria Uther, Alan Costall 2009: Effects of low-pass filtering on the judgment of vocal affect in speech directed to infants, adults and foreigners. *Speech Communication* 51(3), 210-216.

Kotz, Sonja A., Silke Paulmann 2007: When emotional prosody and semantics dance cheek to cheek: ERP evidence, *Brain Research* 1151, 107-118.

Kousta, Stavroula-Thaleia, David P. Vinson, Gabriella Vigliocco 2009: Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition* 112(3), 473-481.

Kreifelts, Benjamin, Thomas Ethofer, Wolfgang Grodd, Matthias Erb, Wildgruber Dirk 2007: Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. *Neuroimage* 37(4), 1445-1456.

Laukka, Petri, Nicolas Audibert, Véronique Aubergé 2012: Exploring the determinants of the graded structure of vocal emotion expressions. *Cognition & Emotion* 26(4), 710-719.

Massaro, Dominic W., Peter B. Egan 1996: Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review* 3(2), 215-221.

Milesi, Valérie, Sezen Cekic, Julie Péron, Sascha Frühholz, Chiara Cristinzio, Margitta Seeck, Didier Grandjean 2014: Multimodal emotion perception after anterior temporal lobectomy (ATL). *Frontiers in Human Neuroscience* 8, 275.

Moulines, Eric, Werner Verhelst 1995: Time-domain and frequency-domain techniques for prosodic modification of speech. In: Kleijn, Bastiaan, W., Kuldip K. Paliwal, (eds.) *Speech Coding and Synthesis*, Amsterdam: Elsevier, 519-555.

Murray, Iain R., John L. Arnott 1993: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America* 93(2), 1097-1108.

Nygaard, Lynne C., Erin R. Lunders 2002: Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition* 30(4), 583-593.

Nygaard, Lynne C., Jennifer S. Queen 2008: Communicating emotion: linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception & Performance* 34(4), 1017.

Paulmann, Silke, Sonja A. Kotz 2008: Early emotional prosody perception based on different speaker voices. *Neuroreport* 19(2), 209-213.

Paulmann Silke, Mark David Pell 2011. Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation & Emotion* 35, 192-201.

Pell, Marc David 2002: Evaluation of nonverbal emotion in face and voice: Some preliminary findings on a new battery of tests. *Brain & Cognition* 48(2/3), 499-504.

Pell, Marc David 2005: Prosody–face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior* 29(4), 193-215.

Pell, Marc David, Laura Monetta, Sally Paulmann, Sonja A. Kotz 2009: Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior* 33(2), 107-120.

Planalp, Sally 1999: *Communicating Emotion: Social, Moral, and Cultural Processes*. Cambridge: Cambridge University Press.

Rigoulot, Simon, Eugen Wassiliwizky, Marc David Pell 2013: Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition. *Frontiers in Psychology* 4, 367.

Russell, James A., Ghyslaine Lemay 2000: Emotion concepts. In: M. Lewis and J.M. Haviland-Jones (Eds.), *Handbook of Emotions* (2nd ed.). New York: Guilford Press.

Schacht, Annekathrin, Werner Sommer 2009: Emotions in word and face processing: early and late cortical responses. *Brain & Cognition* 69(3), 538-550.
Scherer, Klaus R. 2003: Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40(1-2), 227-256.
Scherer, Klaus R., James S. Oshinsky 1977: Cue utilization in emotion attribution from auditory stimuli. *Motivation & Emotion* 1(4), 331-346.
Scherer, Klaus R., Rainer Banse, Harald G. Wallbott, Thomas Goldbeck 1991: Vocal cues in emotion encoding and decoding. *Motivation & Emotion* 15(2), 123-148.
Schirmer, Annett, Sonja A. Kotz 2003: ERP Evidence for a sex-specific Stroop effect in emotional speech. *Journal of Cognitive Neuroscience* 15(8), 1135-1148.
Scott, Graham G., Patrick J. O'Donnell, Hartmut Leuthold, Sara C. Sereno 2009: Early emotion word processing: evidence from event-related potentials. *Biological Psychology 80*, 95-104.
Van Bezooijen, Renée, Charlotte Gooskens 1999: Identification of language varieties: The contribution of different linguistic levels. *Journal of Language and Social Psychology* 18(1), 31-48.
Wells, Laura Jean, Steven Mark Gillespie, Pia Rotshtein 2016: Identification of emotional facial expressions: effects of expression, intensity, and sex on eye gaze. *PloS One* 11, 12, e0168307.

Streszczenie

Rozpoznawanie emocji na podstawie prozodii i semantyki w języku francuskim w mowie spontanicznej: Badanie pilotażowe

Celem tego artykułu jest zbadanie wpływu sygnałów prozodycznych i semantycznych – osobno oraz współwystępujących ze sobą – na rozpoznawania emocji w mowie spontanicznej. Istotą eksperymentu było zbadanie, czy rozpoznawanie emocji jest łatwiejsze w przypadku bodźców monomodalnych czy multimodalnych. Aby to osiągnąć, zastosowano filtrowanie wypowiedzi w mowie spontanicznej. Ponadto zbadano również, czy rozpoznawanie intensywności emocji zmieniało się w zależności od kanału przekazu. W przeciwieństwie do metodologii używanych w większości wcześniejszych badań, zastosowano bodźce zgodne, które zawsze przekazywały tę samą emocję w obu modalnościach: prozodii i semantyce. Tak jak w badaniu Paulmann i Pella (2011), wyniki obecnego badania wskazują, że dokładność w rozpoznawaniu emocji w mowie spontanicznej wzrasta w reakcji na bo-

dziec zgodny. To znaczy, że bodźce zintegrowane były rozpoznawane znacząco lepiej niż bodziec jednomodalny. Wyniki te potwierdzają wcześniejsze badania, które wskazują, że rozpoznanie emocji wzrasta, jeśli dostępne jest więcej niż jedno źródło informacji zgodnych na temat danej emocji (Collignon et al. 2008; Pell 2005; De Gelder & Vroomen 2000; Massaro & Egan 1996). Czasy reakcji również wskazują, że dostęp do emocji jest szybszy, kiedy przekazywane są one przy pomocy obu modalności: semantycznej i prozodycznej. Z kognitywnego punktu widzenia, w związku z tym, że informacja w każdym z kanałów komunikacji emocji aktywuje wspólną wiedzę o danej emocji (por. Bower 1981; Russell & Lemay 2000; Halberstadt et al. 1995), uzasadnione wydaje się twierdzenie, że w czasie zadań polegających na rozpoznawaniu emocji wiedza na temat danej emocji jest łatwiej dostępna, jeśli więcej bodźców zgodnych współdziała w przekazywaniu stanów emocjonalnych (Paulmann & Pell 2011). W zgodzie z tą hipotezą, Borod et al. (2000) zasugerowali dwa etapy w przetwarzaniu emocji. Najpierw kanały emocjonalne są przetwarzane niezależnie, poprzez oddzielne systemy sensoryczne, a następnie przechodzą przez ogólny procesor afektywny (*general affective procesor)*. Z tego wynika, że bodźce prozodyczne i werbalne są integrowane w procesie rozpoznawania emocji, co prowadzi do systematycznie coraz lepszej dokładności, jak wykazano w obecnym badaniu. Co ciekawe, zgodnie z wynikami Paulmann i Pella (2011), opisany proces dotyczy również rozpoznawania spontanicznych stanów neutralnych, które lepiej odczytywano w obecności bodźców zintegrowanych: prozodycznych i semantycznych. Potwierdzają to badania Cornew et al. (2009). W dwóch eksperymentach wykorzystujących zadania stopniowego odkrywania/udostępniania wyrazów (*gating experiment*)pokazali oni, że uczestnicy eksperymentu rozpoznawali prozodię neutralną szybciej i dokładniej niż prozodię wyrażającą złość lub radość, co wskazuje na znaczącą przewagę przetwarzania bodźców neutralnych nad emocjonalnymi.

Jeśli chodzi o intensywność emocji, jak pokazują wyniki Wells et al. (2016), bodźce zintegrowane odczytywano jako bardziej intensywne, co wskazuje, że interakcja pomiędzy kanałem prozodycznym i semantycznym wzmacnia rozpoznawanie intensywności emocjonalnej. Nie dziwi fakt, że bodźce oceniane jako intensywniejsze były lepiej rozpoznawane, co oznacza, że postrzegana intensywność emocji może stanowić silny sygnał do rozróżniania emocji, o ile jest dostępna (Audibert et al. 2008).

Co więcej, obecne badanie pokazało, że emocje spontaniczne przekazywane w modalności semantycznej rozpoznawano dokładniej i szybciej niż te, które przekazywano w modalności prozodycznej. Zjawisko to można wyjaśnić na gruncie różnic tym, jak bodźce prozodyczne i semantyczne pobudzają węzły emocjonalne (*emotion nodes*) w trakcie wyrażania emocji (Bower 1981; Paulmann & Pell 2011; Halberstadt et al. 1995; Russell & Lemay 2000). Wcze-

śniejsze badania pokazują, że emocje wyrażone przez prozodię mają charakter kategoryczny w przedłużonym okresie. Natomiast emocje wyrażone semantycznie mają charakter prototypowy, co sprawia, że wybrane emocje aktywują wiedzę na temat pojęcia danej emocji mocniej i szybciej (Paulmann & Pell 2011). Niskie wskaźniki rozpoznania w reakcji na bodźce prozodyczne (M: 36%) można wyjaśnić w świetle najnowszych badań dotyczących prozodii emocji wyrażanych spontanicznie. Analizy akustyczne wskazały na mniejsze różnice akustyczne pomiędzy emocjami w mowie spontanicznej niż w przypadku emocji wyrażanych aktorsko (Laukka et al. 2012), co utrudniło słuchaczom rozpoznanie emocji pod nieobecność innych elementów mowy i kontekstu. Konieczne są dalsze badania, które sprawdziłyby, czy ta dominacja modalności semantycznej jest charakterystyczna dla danego języka/kultury, na co wskazują badania na Japończykach i Amerykanach (Ishii et al. 2003; Kitayama & Ishii 2002). Inną ważną zmienną w percepcji prozodii i semantyki emocji jest płeć. Schirmer i Kotz (2003) w badaniu ERP pokazali lepszą integrację bodźców emocjonalnych u kobiet niż u mężczyzn. Obecne badanie pilotażowe pominęło ten aspekt, ale w badaniu głównym liczba kobiet i mężczyzn zostanie wyrównana zarówno w grupie eksperymentalnej, jak i w nagraniach bodźców.

Podsumowując, obecne badanie pokazuje, że przewaga multimodalna jest również zauważalna w rozpoznawaniu spontanicznych emocji. Chociaż rozpoznawanie emocji na podstawie tylko jednej modalności jest możliwe, bodźce zgodne i równoczesne przekazane zarówno w warstwie prozodycznej, jak i semantycznej prowadzą do wzrostu dokładności i tempa rozpoznawania emocji. Te wyniki można traktować jako potwierdzenie kognitywnego modelu przetwarzania emocji, w którym bodźce werbalne i niewerbalne są integrowane.

Tłum. Małgorzata Fabiszak

Bibliografia

Audibert, Nicolas, Véronique Aubergé, Albert Rilliard 2008: Acted vs. spontaneous expressive speech: perception with inter-individual variability. *Programme of the Workshop on Corpora for Research on Emotion & Affect*, 23-26.

Borod, Joan C., Lawrence H. Pick, Susan Hall, Martin Sliwinski, Nancy Madigan, Loraine K. Obler, Joan Welkowitz, Elisabeth Canino, Hulya M. Erhan, Mira Goral, Chris Morrison, Matthias Tabert 2000: Relationships among facial, prosodic, and lexical channels of emotional perceptual processing. *Cognition and Emotion* 14, 193–211.

Bower, Gordon H. 1981: Mood and memory. *American Psychologist* 36(2), 129-148.

Collignon, Oliver, Frédéric Gosselin, Sanjoy Roy, Dave Saint-Amour, Maryse Lassonde, Franco Lepore 2008: Audio-visual integration of emotion expression. *Brain Research* 1242, 126-135.

Cornew, Lauren, Leslie Carver, Tracy Love 2009: There's more to emotion than meets the eye: A processing bias for neutral content in the domain of emotional prosody. *Cognition & Emotion* 24(7), 1133–1152.

De Gelder, Beatrice, Jean Vroomen 2000: The perception of emotions by ear and by eye. *Cognition & Emotion* 14(3), 289-311.

Halberstadt, Jamin B., Paula M. Niedenthal, Julia Kushner 1995: Resolution of lexical ambiguity by emotional state. *Psychological Science* 6(5), 278-282.

Ishii, Keiko, Alberto Jose Reyes, Shinobu Kitayama 2003: Spontaneous attention to word content versus emotional tone: Differences among three cultures. *Psychological Science* 14(1), 39-46.

Kitayama, Shinobu, Keiko Ishii 2002: Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition & Emotion* 16(1), 29-59.

Laukka, Petri, Nicolas Audibert, Véronique Aubergé 2012: Exploring the determinants of the graded structure of vocal emotion expressions. *Cognition & Emotion* 26(4), 710-719.

Massaro, Dominic W., Peter B. Egan 1996: Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review* 3(2), 215-221.

Paulmann Silke, Mark David Pell 2011. Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation & Emotion* 35, 192-201.

Pell, Marc David 2005: Prosody–face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior* 29(4), 193-215.

Russell, James A., Ghyslaine Lemay 2000: Emotion concepts. W: M. Lewis and J.M. Haviland-Jones (Red.), *Handbook of emotions* (Wyd. 2). New York: Guilford Press.

Schirmer, Annett, Sonja A. Kotz 2003: ERP Evidence for a sex-specific Stroop effect in emotional speech. *Journal of Cognitive Neuroscience* 15(8), 1135–1148.

Wells, Laura Jean, Steven Mark Gillespie, Pia Rotshtein 2016: Identification of emotional facial expressions: effects of expression, intensity, and sex on eye gaze. *PloS One* 11, 12, e0168307.