ROCZNIKI PSYCHOLOGICZNE /ANNALS OF PSYCHOLOGY
2019, XXII, 1, 15-34
DOI: http://dx.doi.org/10.18290/rpsych.2019.22.1-2

DANIELA HEKIERT[a]
MAGDALENA IGRAS-CYBULSKA[b]
[a]SWPS University of Social Sciences and Humanities, Warsaw, Poland
[b]AGH University of Science and Technology, Cracow, Poland

CAPTURING EMOTIONS IN VOICE: A COMPARATIVE ANALYSIS OF METHODOLOGIES IN PSYCHOLOGY AND DIGITAL SIGNAL PROCESSING

People use their voices to communicate not only verbally but also emotionally. This article presents theories and methodologies that concern emotional vocalizations at the intersection of psychology and digital signal processing. Specifically, it demonstrates the encoding (production) and decoding (recognition) of emotional sounds, including the review and comparison of strategies in database design, parameterization, and classification. Whereas psychology predominantly focuses on the subjective recognition of emotional vocalizations, digital signal processing relies on automated and thus more objective vocal affect measures. The article aims to compare these two approaches and suggest methods of combining them to achieve a more complex insight into the vocal communication of emotions.

Keywords: emotional vocalizations; emotional prosody; vocal bursts; process of encoding and decoding.

THE SUPERIORITY OF ONE APPROACH OVER THE OTHERS? IN SEARCH OF RELIABLE RESULTS

It is common knowledge that emotions are expressed and recognized via different channels. Psychological research on emotion recognition has long relied

Corresponding author: DANIELA HEKIERT – SWPS University of Social Sciences and Humanities, Faculty of Psychology, ul. Chodakowska 19/31, 03-815 Warszawa, Poland; e-mail: dhekiert@swps.edu.pl

exclusively on facial expressions. Nowadays, other modalities are gaining scientific attention as informative sources of knowledge for capturing various emotions. Nevertheless, there is still no agreement on which channel is superior over the others. Accordingly, at least four different research approaches can be identified: (1) one of the modalities is dominant in terms of recognizing all emotions, for instance voice-only communication is more informative than facial displays (e.g., Kraus, 2017) or facial expressions are a better source of information than prosody (e.g., Zhang et al., 2018); (2) differences in the accuracy of emotion recognition occur within one modality, for instance between affective prosody and vocal bursts (Hawk, Van Kleef, Fischer, & Van Der Schalk, 2009); (3) emotions are best recognized when dealing with multimodal channels as compared to the one-modal equivalent (e.g., Zaki, Bolger, & Ochsner, 2009); (4) each of the modalities is dominant in terms of recognizing selected, specific emotions; for instance, the social function of emotions can serve as an indication of the channel utilized for the communication of a particular emotion (e.g., App, McIntosh, Reed, & Hertenstein, 2011; Laukka et al., 2013).

In digital signal processing (DSP), whenever access to multimodal databases is provided, all modalities are usually used to train models, and such systems based on multimodal information tend to perform better than unimodal ones (e.g., Tzirakis, Trigeorgis, Nicolaou, Schuller, & Zafeiriou, 2017). The superiority of one channel over others is considered regarding its contribution to the performance of the automatic recognition system (e.g., the extent to which each modality contributed to the final decision of the classification system).

Apart from research concerning the superiority of one modality over the others, there have been studies that focused on testing the influence of different factors on the recognition of emotions – for instance, on the significant or insignificant role of culture (e.g., in relation to emotional vocalizations: Sauter, Eisner, Ekman, & Scott, 2010b; Cordaro, Keltner, Tshering, Wangchuk, & Flynn, 2016; Gendron, Roberson, van der Vyver, & Barrett, 2014). Additionally, there have been attempts to verify if individual differences in physiological reaction to emotional event change nonverbal expression (Pisanski, Nowak, & Sorokowski, 2016; Pisanski et al., 2018). Finally, some approaches focus on the representation of emotions in the brain (i.e., on the neural systems responsible for perceiving emotions from different channels; e.g., Schirmer & Adolphs, 2017). These include the criticized approach postulating the occurrence of neural fingerprints for specific categories of emotions (e.g., Saarimäki et al., 2015; for critique, see Clark-Polner, Johnson, & Barrett, 2017) or studies on brain structures respon-

sible for the processing of affective dimensions: arousal and valence (e.g., Bestelmeyer, Kotz, & Belin, 2017).

The aim of the review

The use of different theoretical frameworks and methodology by various researchers makes it difficult to identify the most reliable approach. For this reason, we have decided to perform a review, aiming (1) to present and compare different encoding and decoding strategies within and between psychological and digital signal processing (DSP) literature, and (2) to offer some examples of how the two scientific domains can complement each other. In particular, we have focused on research concerning emotional vocalizations, which has been limited to the recognition of emotions in the domain of psychology but addresses both production and recognition issues in digital signal processing (DSP). To the best of our knowledge, this review is the first to juxtapose the psychological and engineering methodologies applied in investigations devoted to the encoding (production) and decoding (recognition) of vocal manifestations of emotions.

The physiological background of the production and recognition of emotional vocalizations

Although the voice apparatus is invisible from the outside and thus may seem inconspicuous, it can in fact reveal rich information, for instance about the individual's age, gender, identity, and emotions (Johar, 2016). Voice is a reflection of primary stimuli, as in the case of the physiological fight-or-flight response. This is due to the fact that emotions evoke autonomic nervous system reactions.

Production of sounds

As noted by Fitch (2000; cited in Waaramaa-Mäki-Kulmala, 2009) the production and recognition of vocal sounds is enabled by the vocal tract and facilitated by vocal control and vocal learning processes. The vocal sound is produced when air from the lungs triggers the oscillation of the vocal folds (commonly known as vocal "cords"), located in the larynx (Fitch, 2000). The rate of vocal fold oscillation – fundamental frequency, F0 – usually determines the perceived pitch of the sound. The signal is generated and subsequently filtered while passing through the vocal tract (the pharyngeal, oral, and nasal cavities). It finally

reaches the environment through the nostrils and lips (Fitch, 2000). In other words, voiced speech is generated at the vocal cords and modulated by the vocal tract (Johar, 2016). During the unvoiced parts of speech, the source of the signal originates from turbulence noise, which occurs when air flows rapidly through a narrow constriction of the vocal tract in the oral cavity. The biocybernetic model of speech production (the source–filter model) accommodates the process described above.

Recognition of emotional sounds

The sound can be identified along its path from the outer, middle, and inner ear, via the auditory nerve, to the auditory fields. Emotional vocal information travels to the primary auditory cortex (lemniscal pathway), to the secondary auditory cortex (non-lemniscal pathway) (Schirmer & Adolphs, 2017), and to the amygdala (e.g., Fecteau, Belin, Joanette, & Armony, 2007). Mirror neurons also play a role in processing emotional vocalizations; for example, in one study skin conductance response (SCR) was enhanced for listening to the vocal emotions of fear and anger, indicating the increased activation of amygdala; SCR elicited by thinking of what it would be like to sound a certain emotion was higher than SCR for listening (Ramachandra, Depalma, & Lisiewski, 2009). Schirmer and Adolphs (2017) pointed out that each modality (voice, face, and touch) activated a specific sensory system, although there was an early integration of perceptual representations from the systems. Kuhn, Wydell, Lavan, McGettigan, and Garrido (2018) reflected on the shared representations of emotions via voices and faces, claiming that similar coding processes for emotions might exist across modalities, despite input differences. As noted by Baart and Vroomen (2018), discrepant imaged information recalibrates the recognition of vocal emotions in such a way that a signal originating in one channel can influence and modify signals in the other channel.

VOCAL CUES OF EMOTIONS

Indicators of emotions in the acoustic parameters of speech

In order for the features of voice to be described numerically, acoustic parameters of the signal are extracted using speech processing algorithms. Three groups of acoustic speech parameters can be distinguished: source parameters, vocal tract parameters, and prosody. The first and the second groups are usually

analyzed in the frequency domain in short, 20-to-30-milisecond frames of speech. Therefore, they are referred to as Low-Level Descriptors (LLD). Signal processing algorithms such as filtering or linear prediction enable the extraction of the acoustic characteristics of the source and the filter separately. The most popular source parameters include the Linear Prediction residuals, jitter or shimmer, which describe the stability of F0 production, and voice trembling. Examples of typical vocal tract features are formants or Mel-frequency cepstral coefficients (MFCC), which show the amount of energy in the frequency bands. On the other hand, prosodic features are observed in larger frames, such as syllables, phrases, and sentences. For this reason, prosody parameters are also called suprasegmental or high-level features. Prosody describes intonation (modulation of F0 within the utterance), intensity (loudness, energy), and rhythm of speech (pauses, duration of speech segments, speech tempo). To sum up, prosody refers to some aspects of speech that could be applicable to music. Recently, there have been some advances in research pointing to the similarities and differences between the acoustic production of emotions in the speaking and singing voices (Scherer, Sundberg, Tamarit, & Salomão, 2015; Scherer, Sundberg, Fantini, Trznadel, & Eyben, 2017). Prosodic properties of speech can be interpreted as carriers of emotional information (emotional prosody) or linguistic information (Rymarczyk, 1999) such as lexical accent or ascending intonation at the end of a question (Marczewska & Osiejuk, 1994, as cited in Rymarczyk, 1999). Prosody is considered to be the most emotion-sensitive among the three groups of acoustic speech parameters, although all groups have been found to be affected by emotions (Koolagudi & Rao, 2012). A review of 104 studies of vocal expression identified acoustic markers for different emotions, such as sadness, anger, fear, tenderness, and happiness (Juslin & Laukka, 2003). For instance, the increase in the mean frequency of vocal cord oscillation (F0), its intensity and variability (Johnstone & Scherer, 2000), have been attributed to the acoustic parameters of joy. As suggested by Ekman (2003), although there is only one facial expression that signals happiness, positive emotions can be expressed by distinct vocal signals. To this effect, research on speech prosody has placed emphasis on the distinction between different positive emotions, for example between happiness and elation (Banse & Scherer, 1996) or between contentment, sensual pleasure, amusement, and triumph (Sauter, 2006).

Indicators of emotions in nonverbal vocalizations

The second strategy, vocal bursts, focuses on the analysis of various sounds such as groans, shrieks, and others (e.g., "whaaaa"; Scherer, 1994). The so-called *nonlinguistic affect vocalizations* are defined as "short, emotional non-speech expressions, comprising both clear non-speech sounds (e.g., laughter) and interjections with a phonemic structure (e.g., 'Wow!')" (Schröder, 2003, p. 103, as cited in Hawk et al., 2009, p. 103). These include pancultural sounds with high individual variance in patterning and in culturally moderated "emblems" with a constant phonemic structure (Scherer, 1994, as cited in Hawk et al., 2009). In their study, Sauter and Scott (2007) found that accuracy levels for vocal bursts in two cultural groups reached 70.1%.

Table 1. *Vocal Bursts and Relative Changes of Example Acoustic Parameters of Corresponding Emotions in Accordance With the Universalist Perspective*

Emotions	Vocal bursts	Relative change in selected acoustic parameters in relation to the reference level		
		F0 mean	F0 range	Tempo
Joy	Laughter	+50%	+100%	+30%
Sadness	Crying	-1	-5	-10
Fear	Scream	+150%	+20%	+30%

Source: Cordaro et al., 2016; Hawk et al., 2009; Schröder, 2001.

Table 1 presents a summary illustrating examples of acoustic parameters and vocal bursts for different emotions in accordance with the universalist approach. Universalists claim that each vocalization can be linked to a specific emotional category, whereas culturalists do not postulate the existence of any discrete acoustic emotion indicators, which is why their typology will not be represented in a table format. The discrete and dimensional approaches to emotional vocalizations will be discussed in the next subsection.

Discrete vs. dimensional approaches to emotional vocalizations

The universality of basic emotions has been extensively debated in the academic writings of twentieth-century psychology (e.g., Ekman, Friesen, & Ellsworth, 1972). Following the evolutionary ideas of Darwin (1872/1998) and Tomkins (1955), Ekman demonstrated that members of illiterate and literate cultures recognize basic emotions with similar accuracy. Although Ekman and Friesen (1969) coined the term "display rules" to stress the social and cultural overlap in the primal display of emotions, Ekman did not abandon his attempts to prove the universal nature of emotional expression. Nevertheless, some researchers have argued that there is no convincing evidence for the universality of emotional recognition and the expressed emotions are culture-dependent (e.g., Birdwhistell, 1970). This point of view was further developed by Russell (1980). According to Russell's circumplex model of affect, emotions arise in two neurophysiological systems and can thus be explained using two dimensions: arousal and valence. Without context people are unable to recognize distinct emotions, as initially proposed by Ekman. It is the linear combination of the valence and arousal dimensions that frames context and provides information about a person's emotional state (Sauter, 2006). Henceforth, two contrasting approaches emerged: one emphasizing the universal properties of basic emotions and the other stressing the cultural character of expressing and perceiving emotions. The former is associated with the evolutionary development of basic emotions and the latter highlights the independent process of constructing one's own representation of emotions (Sauter, 2006).

Universalist perspective on emotional vocalizations

Initially, vocalizations were thought to merely signal arousal rather than distinctive emotions (Scherer, 1986). Subsequently, researchers began to test the discrete concept of emotions. For instance, Bryant and Barrett (2008) established that members of a tribe living in Amazonian Ecuador were able to identify the vocalizations of sadness, anger, fear, or happiness produced by an English speaker. Scherer, Banse, and Wallbott (2001) tested the recognition accuracy of the vocal display of five emotions in several locations including Indonesia, the U.S., and nine European countries. They found that the overall recognition rate amounted to 66%, and participants from the same country made similar errors. Albas, McCluskey, and Albas (1976) observed that Caucasian and Cree males scored higher on the level of emotion recognition displayed by members of their

own cultural groups. This phenomenon is referred to as in-group advantage and has been reported in a large number of publications (Chronaki, Wigelsworth, Pell, & Kotz, 2018). For instance, research suggests with high certainty that sadness, anger, and disgust are universally recognized (through vocal bursts), whilst moderately high certainty applies to surprise, awe, and triumph (Cordaro et al., 2016). The universality argument has been considerably weakened by more recent studies, which have found that pride, guilt, and shame are not recognized in the same manner across cultures (Cordaro et al., 2016).

Non-universalist (culturalist) perspective on emotional vocalizations

Recent neuroimaging studies do not support the occurrence of neural fingerprints for specific categories of emotions (e.g., Clark-Polner et al., 2017), which means that the discrete model is scientifically limited. As noted by Gendron et al. (2014), there is a stronger (i.e., the same set of emotions is recognized all over the world) and a weaker (i.e., cultural dialects overlap universally recognized emotions) version of the universality hypothesis. Culturalists have tried to prove that universalists are wrong by pointing out that the correspondence between vocalizations and emotions, as between smiling and happiness, is not equally accurate across cultures (Gendron et al., 2014). In this regard, Gendron et al. (2014) claimed that it was the very perception of emotional valence that was pancultural, not the recognition of discrete emotions itself. Russell (1994) claimed that context was important in the recognition of emotions, on the grounds that it enabled the recognition of specific emotions (Sauter, 2006). Given that culture creates context, the former can influence the perception and interpretation of emotional prosody (Chronaki et al., 2018).

There is still no agreement in academic circles as regards the extent to which nature vs. nurture contributes to the vocal communication of emotions. Subscribing to the universalist or culturalist hypothesis is linked with the choice of research design and, consequently, to the strategies selected for the encoding and decoding of vocal emotions. For instance, if researchers assume the pancultural nature of the recognition of emotions, they usually ask participants to recognize emotions from a close-ended list, whereas if they believe that the recognition of emotions is culture-dependent they tend to ask about the level of arousal and valence. As a result, both universalists and culturalists find support for their initial assumptions because the techniques they use usually leave no other option. Accordingly, using different theoretical assumptions and methodologies, one

may end up with discrepant findings: emotions and affective states are culturally dependent or independent. In search of the most reliable approach, it seems necessary to assess not only the methodology of emotion recognition, but also the production of sounds. In this paper we distinguish encoding and decoding processes, providing some useful information about how stimulus is produced (encoded) and how it is used to recognize emotions (decoding).

Strategies of encoding: Production of vocalizations

As noted by Scherer et al. (2017), strong emphasis on perception and recognition is responsible for the dearth of empirical work on encoding (i.e., the production of expression). This is especially visible in the domain of psychology.

Encoding strategies in psychology

The cross-cultural study conducted by Sauter et al. (2010b) utilized sound recordings from a validated data base (see Sauter, Eisner, Calder, & Scott, 2010a), which stores nonverbal emotional vocalizations produced in artificial settings by native English speakers with no formal training in acting. Among the Himba,[1] the production of stimuli was similar to the English stimulus production (Sauter et al., 2010b). Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, and Abramson (2009) collected material from non-trained participants who received a list of emotions and scenarios (theoretical descriptions of the affective states) and were asked to produce one to five sounds for each emotion. A similar methodology was adopted by Cordaro et al. (2016). In contrast, Hawk et al. (2009) decided to ask acting students as opposed to non-trained individuals to participate in stimulus production. They were instructed on how to think about specific emotions and asked, for example, to vocally (without verbal content and then using speech-related expressions) act out a situation in which they felt a given emotion. Two evaluators made assessments using various criteria; as a result, a specific pool was selected for the decoding phase (Hawk et al., 2009). In another research project (Laukka et al., 2013), authors made use of an already available corpus (VENEC) in which actors (trained professionals) from different countries were asked to convey emotions: prosodic expressions and nonlinguistic vocalizations as convincingly as they could. Instructions were provided to participants in which situations a given emotion could be felt (Laukka et al., 2010). Laukka et al. (2013) only used stimuli from VENEC that contained vocal bursts.

[1] An illiterate tribe from Namibia.

Oleszkiewicz Pisanski, Lachowicz-Tabaczek, and Sorokowska (2017) collected recordings from four males and four females who produced monophthong vowels (e.g., /i/, /ɛ/, /o/). Then, the pitch of each voice was artificially changed: raised or lowered (Oleszkiewicz et al., 2017).

In summary, psychological research relies primarily on encoding strategies as applied by professional actors (e.g., Hawk et al., 2009) or by untrained individuals (e.g., Simon-Thomas et al., 2009), who are asked to express specific emotions using their own emotional memory, (one-sentence) stories, or single words/syllables. Researchers employ various methods of introducing stimuli, such as the standard contents paradigm (i.e., each emotion is expressed with the same sentence/word) or pseudo-speech (i.e., composing of phonemes from various languages into a sentence) (Sauter, 2006). Alternatively, researchers make use of readily available databases (e.g., Laukka et al., 2013). In general, descriptions of the encoders and the process of encoding are not as precise as information about decoding. For instance, in Sauter (2010b) the English vocalizers are called "actors," whereas in Sauter (2010a) it is explained that they were not trained professionals.

Encoding strategies in DSP

Many monographic works devoted to the problem of collecting vocal recordings (e.g., Douglas-Cowie, Campbell, Cowie, & Roach, 2003) stress the importance of choosing a proper database. A corpus of recordings is often supplemented by a structured database with detailed metadata. In speech technology it is required to specify technical parameters with high precision (e.g., sampling rate, bit depth, file format, or signal to noise ratio).

Sidorova (2007) distinguished the following types of emotional speech corpora:

– Acted – emotional speech simulated by actors; when gathering data for acted corpora, professional actors are asked to produce a set of utterances expressing a given emotion;

– Authentic – real life situations (e.g., call center conversations, emergency phone calls, or live TV coverage) as the most natural resources of emotional speech in terms of spontaneity and authenticity;

– Elicited – emotions triggered by affective stimuli such as emotional movies, stories, pictures, or games.

In 2003, Ververidis and Kotropoulos found that in more than 20 of a total of 32 emotional speech databases reviewed the emotional speech recorded was pro-

duced by actors. One of the most frequently used databases of acted emotions is the *German Emotional Speech Database* (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005). Ten actors (five women and five men) simulated seven emotions: neutral, anger, fear, joy, sadness, disgust, and boredom, producing ten German utterances (five short and five longer sentences) which could be used in everyday communication. The complete database was evaluated in a perception test regarding the recognizability of emotions and their naturalness.

An example of a widely used elicited emotions corpus is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database in English (Busso, Bulut, Lee, Kazemzadeh, Mower, Kim, ... & Narayanan, 2008). The actors performed selected emotional scripts and improvised hypothetical scenarios designed to elicit specific types of emotions (happiness, anger, sadness, frustration, and neutral state).

The initial trend of using artificially created emotions has been recently replaced by searching for opportunities to investigate natural and spontaneous emotional speech, for example the data collected from an emergency call center in Poznań (Demenko & Jastrzębska, 2012) or a corpus of real-life conversations retrieved from an emergency call center, tagged with emotions described by Gałka and colleagues (2015). Kamińska and Sapiński (2017) collected a corpus of emotional speech containing samples extracted from discussions in TV programs.

The most desired properties of emotional speech corpora are authenticity, diversity, intensity and explicitness of emotions, good quality of recordings, a large number of speakers, and uniform speech content (e.g., Sidorova, 2007). However, their mutual and simultaneous coexistence is a rather remote possibility.

STRATEGIES OF DECODING: RECOGNITION OF VOCALIZATIONS

Decoding strategies in psychology

In the previously mentioned study conducted by Sauter et al. (2010b), English and Himba participants were first asked to listen to a story that reflected on a particular emotion – for instance, about a person who felt very scared due to a sudden encounter with a dangerous animal. Next, they heard recordings of two vocalization sounds and were required to choose the most suitable one. In this way the researchers avoided problems with direct translations. Similarly, Cor-

daro et al. (2016) asked participants from ten globalized and one remote culture to pair vocal bursts with 16 one-sentence emotional stories that they were required to read (for remote Bhutanese villagers the stories were read by a translator). They had to match one of three vocal burst sounds with each of the 16 stories (Cordaro et al., 2016). In another study (Gendron et al., 2014), which tested the reliability of the universal approach as compared to the culturalist perspective, participants were asked to use a single word or a phrase to name the emotion (non-word vocalization) that they had heard. If they spontaneously reported that a given sound reminded them of a specific situation or behavior, they were subsequently asked to think about one word that best described that emotion. The respondents' answers were assessed by two coders in accordance with Russell's criteria (Russel, 1990, as cited in Gendron et al., 2014); for instance, coders decided whether the participants' responses were consistent with the stimulus according to discrete emotions or the affective dimensions (valence and arousal). Hawk et al. (2009) invited students from one of the Dutch universities to rate 80 stimuli (10 emotion categories x 8 encoders) presented on a computer according to three channels (affect vocalizations, speech, and face). One person rated 80 stimuli from only one channel, which means that he or she had to choose one of 10 labels for each emotion (Hawk et al., 2009). Laukka et al. (2013) asked Swedish female students to rate recordings, giving each vocalization produced by actors from different countries one emotional label. Participants received a dictionary and the scenarios for each emotion, which had previously been presented to the actors. The recognition of nine positive and nine negative emotions was examined separately, but practically the same procedure was used for both. In the research project conducted by Simon-Thomas et al. (2009), students (only women) matched a list of emotions – nine negative and 13 positive emotions – with vocal burst sounds played on the computer in a so-called forced-choice procedure. There was also an option of selecting "none of the above." Kraus (2017) designed a series of five experiments. In one of them participants from the U.S. were required to evaluate an interaction between individuals. The interaction was tested according to three modalities, namely: voice-only, visual-only, or dual voice and visual communication using video and/or audio recordings. Raters were assigned to one of three conditions and estimated 23 discrete emotion words that matched the stimuli using 9-point Likert scales.

In summary, most studies methodologically rely on findings that are often based on forced-choice replies to questions about emotions and their categorization, without recognizing the possible impact of culture, represented by the target audience or present in the culture-laden meanings attached to questions. As

a result, researchers have frequently arrived at the conclusion that a given sound is linked to a given emotion, which could be interpreted in favor of affect universality.

It is worth mentioning that in psychology researchers usually confirm or disconfirm their hypotheses (e.g., the hypothesis on the universality of emotional vocalizations), whereas in DSP they tend first to define a goal and then to report on the achievement of results (i.e., success or failure). For instance, a possible DSP goal can be defined as the improvement of the state-of-the-art accuracy level of emotion recognition in a given corpus.

Decoding strategies in DSP

One of the aims of speech technology, next to automatic speech recognition and speaker verification, is the automatization of the emotion recognition process using artificial intelligence tools. Speech engineers use algorithms for speech processing and decision making in order to build systems that classify a recording into a particular affect category. In the typical machine learning (ML) approach, there are two phases: training and classification. In the training phase, models of emotions are created in three major steps: pre-processing, parameterization, and stochastic modelling. In the pre-processing phase, voice activity detection and signal normalization are performed. Emotion-dependent acoustic features are extracted in the parameterization step. Sometimes, the number of extracted acoustic features can reach even several thousands, together with their statistical descriptors (Eyben et al., 2016). In the modelling step, common patterns describing each category of emotion are extracted on the basis of corpora of emotional speech.

The identification process is divided into pre-processing and parameterization (the same as in the training), multiple verification, and calculation of the final scores. In the last step, based on the calculated similarity scores, the system decides which emotion from the database is the most likely to generate an acquired voice sample. As a result, each fragment of the recording is labelled with the emotional states showing the highest confidence rate (Figure 1).

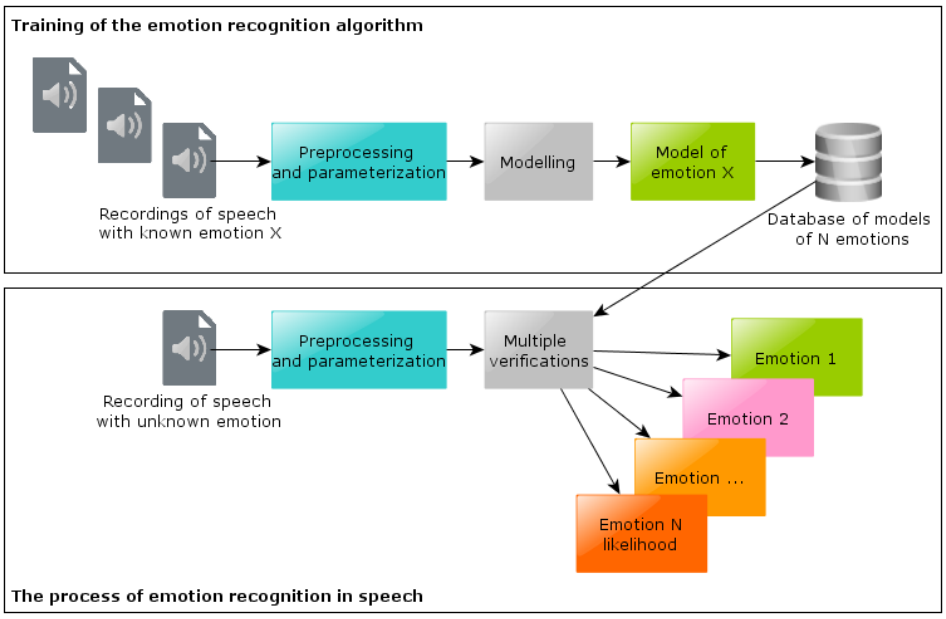

Figure 1. The scheme of training and the identification of emotions in speech. Source: own elaboration based on Witkowski et al. (2016).

Examples of pattern recognition algorithms used for speech emotion classification include the Gaussian Mixture Models, decision trees, the Hidden Markov Models, or neural networks. The use of classifiers mainly depends on the nature of the data. If known in advance, the task of deciding on the type of the classifier is made easier, but in most real-life situations information on the nature of data is rarely available in advance (Koolagudi & Rao, 2012).

All these methods have achieved a recognition efficiency of approximately 60-85% (e.g., according to reviews by Koolagudi & Rao, 2012; Basu, Chakraborty, Bag, & Aftabuddin, 2017). Recently, Deep Learning (DL) has gained much popularity due to its superiority in terms of accuracy, conditional on being trained on big data. As of late, deep learning architectures have been applied to speech emotion recognition and authors have reported outperformance compared to previous results (over 90% accuracy; e.g., Harár, Burget, & Dutta, 2017; Fayek, Lech, & Cavedon, 2017). Traditional ML algorithms are better suited for small data. Another drawback of DL is that the system is a kind of "black box," which makes it difficult to discern why it has opted for a particular solution or decision. This method is therefore perceived as having a limited contribution to the broadening of our understanding of speech signals.

Shortcomings of research methodologies

One of the main sources of problems faced by both branches of science is the complex nature of the processes underlying the expression and perception of emotions. The most popular categories of basic emotions are constituent parts of a model, while real life usually offers complex and intercorrelated emotions. Moreover, emotions are multimodal phenomena. Some are more clearly expressed by facial mimic or gestures, and others – by physiological reactions. In this process, the voice conveys only partial affect information. Furthermore, the differences between individuals (e.g., their personality) predispose the deployment of strategies for acoustic manifestation of emotions. Certain aspects of emotional expression are language- or culture-specific.

A listener's individual ability to recognize emotions in speech can influence the results of the human perception test during the labeling of recordings. The subjectivity of a recipient who perceives and assesses the emotional content of speech is dependent on individual sensitivity and empathic ability. Moreover, while planning experiments with intense emotional elicitation, ethical rules need to be followed. At the same time, emotions which are too weak may not necessarily achieve an adequate level of vocal manifestation. Perceptual tests are time-consuming and expensive. As the result, emotional speech corpora are usually small or have not been sufficiently evaluated.

Conclusions and future implications

Studies on emotional vocalizations are already popular in DSP but are also likely to become an important component of psychological research. For instance, non-linguistic vocalizations may garner recognition as a highly convenient cross-cultural research method (Laukka et al., 2013), which allows to avoid equivalence testing altogether. Moreover, vocal emotion recognition can be instrumental as a lie detection tool (Kraus, 2017), given that emotional vocalizations are not easy to control. Nevertheless, certain challenges exist in connection with research design and sampling, as stated in the previous sections. In this regard, the amplification of research effort using various cross-cultural and multimodal paradigms as well as increased collaboration between psychologists, engineers, and computer science specialists is a step in the right direction (Johar, 2016). While psychology focuses primarily on the subjective recognition of emotional vocalizations, DSP relies on automation, which is believed to generate more objective vocal affect measures. Below we would like to propose some examples of how the two scientific domains can complement each other.

DSP could benefit from deeper insight into the discrete and dimensional models of emotions which are used in automatic speech recognition, especially their nature and applicability to the usage context of the speech technology. Usually, a particular model is used without theoretical justification and different models are rarely compared in the same investigation. In some applications, new models or new selections of categories of emotions could bring better results than standard models. Psychology could also have an impact in improving the quality of emotional speech corpora with more complex perceptual evaluation of recordings as well as an assessment of other dimensions of speaker characteristics affecting emotional expression, e.g., personality or expressivity. The understanding of such connections might result in better signal descriptors being designed by speech engineers.

On the other hand, psychological empirical studies on emotions often lack in detailed descriptions of the production of emotional speech corpora, including signal quality, the recruitment and description of speakers, and the design of recording setup, while in DSP literature such specifications are essential and provide better comparability of research. Moreover, in DSP it is common to create and share emotional databases, and some of them became standard corpora that are used as a universal benchmark for algorithm evaluation. We suggest that this could be adopted as a good practice in psychological research. Finally, both fields of science may draw on psychoacoustic methods in which the measurement of sound perception meets high technical and methodological standards, as this field bridges biology, psychology, and engineering.

The future promising area for the mutual cooperation of engineers and psychologists is the use of crowdsourcing in the process of collecting and evaluating databases of emotional recordings. In recent years, the dynamic increase in the popularity, quality, and performance of microphone-equipped computational devices (e.g., smartphones, Internet of Things), as well as their miniaturization, has led to the development of novel voice interface solutions. This context creates good conditions for the increase of the popularity of voice user interface in society, which provides opportunities to collect emotional speech recordings from users or ask them to help in the tagging of recordings while using voice services. The collection of large amounts of emotional speech data is crucial for applying deep learning methods in emotion recognition, which is considered nowadays as one of the main trends in computing.

REFERENCES

Albas, D. C., McCluskey, K. W., & Albas, C. A. (1976). Perception of the emotional content of speech: A comparison of two Canadian groups. *Journal of Cross-Cultural Psychology, 7*(4), 481-490.

App, B., McIntosh, D. N., Reed, C. L., & Hertenstein, M. J. (2011). Nonverbal channel use in communication of emotion: How may depend on why. *Emotion, 11*(3), 603-617.

Baart, M., Vroomen, J. (2018). Recalibration of vocal affect by a dynamic face. *Experimental Brain Research, 236*(7), 1911-1918.

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

Basu, S., Chakraborty, J., Bag, A., & Aftabuddin, M. (2017, March). A review on emotion recognition using speech. In *International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 109-114). IEEE.

Bestelmeyer, P. E., Kotz, S. A., & Belin, P. (2017). Effects of emotional valence and arousal on the voice perception network. *Social Cognitive and Affective Neuroscience, 12*(8), 1351-1358.

Birdwhistell, R. L. (1970). *Kinesics and context: Essays on body motion communication.* Philadelphia, PA, US: University of Pennsylvania Press.

Bryant, G. A., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture, 8*(1), 135-148.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In *EUROSPEECH-2005: Ninth European Conference on Speech Communication and Technology* (pp. 1517-1520). Lisbon, Portugal.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., . . . Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42*(4), 335-359.

Chronaki, G., Wigelsworth, M., Pell, M. D., & Kotz, S. A. (2018). The development of cross--cultural recognition of vocal emotion during childhood and adolescence. *Scientific Reports, 8*(1), 8659.

Clark-Polner, E., Johnson, T. D., & Barrett, L. F. (2017). Multivoxel pattern analysis does not provide evidence to support the existence of basic emotions. *Cerebral Cortex, 27*(3), 1944-1948.

Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion, 16*(1), 117-128.

Darwin, C. (1872/1998). *The expression of emotion in man and animals.* New York, NY, US: Oxford University Press.

Demenko, G., & Jastrzębska, M. (2012). Analysis of voice stress in call centers conversations. In *Proceedings Speech Prosody. 6th International Conference* (pp. 183-186). Shanghai, China.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication, 40*(1-2), 33-60.

Ekman, P. (2003). *Emotions revealed.* New York, NY, US: Times Books.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*(1), 49-98.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and a review of findings.* New York, NY, US: Pergamon Press Inc.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., . . . Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*(2), 190-202.

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks, 92*, 60-68.

Fecteau, S., Belin, P., Joanette, Y., & Armony, J. (2007). Amygdala responses to nonlinguistic emotional vocalizations. *NeuroImage, 36*, 480-487.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Sciences, 4*(7), 258-267.

Gałka, J., Grzybowska, J., Igras, M., Jaciów, P., Wajda, K., Witkowski, M., & Ziółko, M. (2015). System supporting speaker identification in emergency call center. *Sixteenth Annual Conference of the International Speech Communication Association – INTERSPEECH* (724-725). Dresden, Germany.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science, 25*(4), 911-920.

Harár, P., Burget, R., & Dutta, M. K. (2017). Speech emotion recognition with deep learning. In *4th International Conference on Signal Processing and Integrated Networks (SPIN)* (pp. 137-140). New Delhi, India.

Hawk, S. T., Van Kleef, G. A., Fischer, A. H., & Van Der Schalk, J. (2009). "Worth a thousand words": Absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion, 9*(3), 293-305.

Johar, S. (2016). Psychology of voice. In S. Johar (Ed.), *Emotion, affect and personality in speech* (pp. 9-15). Berlin: Springer.

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. *Handbook of Emotions, 2*, 220-235.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*(5), 770-814.

Kamińska, D., & Sapiński, T. (2017). Polish emotional speech recognition based on the committee of classifiers. *Przegląd Elektrotechniczny, 93*, 101-105.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology, 15*(2), 99-117.

Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. *American Psychologist, 72*(7), 644-654.

Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garrido, L. (2017). Similar representations of emotions across faces and voices. *Emotion, 17*(6), 912-937.

Laukka, P., Elfenbein, H. A., Chui, W., Thingujam, N. S., Iraki, F. K., Rockstuhl, T., & Althoff, J. (2010). Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. In *Proceedings of the LREC 2010 Workshop on Corpora for Research on Emotion and Affect* (pp. 53-57). Paris, France: European Language Resources Association.

Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Iraki, F. K. E., . . . Thingujam, N. S. (2013). Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in Psychology, 4*, 353. DOI: 10.3389/fpsyg.2013.00353

Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., & Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic Bulletin & Review, 24*(3), 856-862.

Pisanski, K., Kobylarek, A., Jakubowska, L., Nowak, J., Walter, A., Błaszczyński, K., . . . Sorokowski, B. (2018). Multimodal stress detection: Testing for covariation in vocal, hormonal and physiological responses to Trier Social Stress Test. *Hormones and Behavior, 106*, 52-61.

Pisanski, K., Nowak, J., & Sorokowski, P. (2016). Individual differences in cortisol stress response predict increases in voice pitch during exam stress. *Physiology & Behavior, 163*, 234-238.

Ramachandra, V., Depalma, N., & Lisiewski, S. (2009). The role of mirror neurons in processing vocal emotions: Evidence from psychophysiological data. *International Journal of Neuroscience, 119*(5), 681-691.

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

Russell, J. A. (1994). Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. *Psychological Bulletin, 115*, 102-141.

Rymarczyk, K. (1999). Zaburzenia prozodii emocjonalnej i lingwistycznej u pacjentów z uszkodzeniami mózgu [Disorders of emotional and linguisticprosody in patients with brain damage]. *Przegląd Psychologiczny, 42*, 135-150.

Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I. P., Lampinen, J., Vuilleumier, P., Hari, R., . . . Nummenmaa, L. (2015). Discrete neural signatures of basic emotions. *Cerebral Cortex, 26*(6), 2563-2573.

Sauter, D. (2006). *An investigation into vocal expressions of emotions: The roles of valence, culture, and acoustic factors* (Doctoral dissertation). University College London.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010a). Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology, 63*(11), 2251-2272.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010b). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences, 107*(6), 2408-2412.

Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion, 31*(3), 192-199.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143-165.

Scherer, K. R. (1994). Affect bursts. In S. H. M. van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 161-193). Hillsdale, NJ, US: Erlbaum.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76-92.

Scherer, K. R., Sundberg, J., Fantini, B., Trznadel, S., & Eyben, F. (2017). The expression of emotion in the singing voice: Acoustic patterns in vocal performance. *The Journal of the Acoustical Society of America, 142*(4), 1805-1815.

Scherer, K. R., Sundberg, J., Tamarit, L., & Salomão, G. L. (2015). Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language, 29*(1), 218-235.

Schirmer, A., & Adolphs, R. (2017). Emotion perception from face, voice, and touch: Comparisons and convergence. *Trends in Cognitive Sciences, 21*(3), 216-228.

Schröder, M. (2001). Emotional speech synthesis: A review. In *EUROSPEECH-2001: Seventh European Conference on Speech Communication and Technology* (pp. 561-564). Aalborg, Denmark.

Sidorova, J. (2007). Speech emotion recognition. *DEA report, doctoral program Ciencia Cognitiva i Llenguatge*. Universitat Pompeu Fabra, Barcelona.

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion, 9*(6), 838-846.

Tomkins, S. S. (1955). Consciousness and the unconscious in a model of the human being. In *Proceedings of the 14th International Congress of Psychology* (pp. 160-161). Amsterdam: North-Holland Publishing Co.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing, 11*(8), 1301-1309.

Vallee, M. (2017). The science of listening in bioacoustics research: Sensing the animals'sounds. *Theory, Culture & Society, 35*(2), 47-65.

Ververidis, D., & Kotropoulos, C. (2003, October). A state of the art review on emotional speech databases. In *Proceedings of 1st Richmedia Conference* (pp. 109-119). Laussane, Switzerland.

Waaramaa-Mäki-Kulmala, T. (2009). *Emotions in voice. Acoustic and perceptual analysis of voice quality in the vocal expression of emotions* (Doctoral dissertation). University of Tampere.

Witkowski, M., Gałka, J., Grzybowska, J., Igras, M., Jaciów, P., & Ziółko, M. (2016). Online caller profiling solution for a call centre. *Odyssey 2016: The Speaker and Language Recognition Workshop*. Bilbao, Spain.

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion, 9*, 478-487.

Zhang, H., Chen, X., Chen, S., Li, Y., Chen, C., Long, Q., & Yuan, J. (2018). Facial expression enhances emotion perception compared to vocal prosody: Behavioral and fMRI studies. *Neuroscience Bulletin, 34*(5), 801-815.