aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–16

DOI: 10.1177/17456916231217722
www.psychologicalscience.org/PPS
Sage

The Sound of Emotional Prosody: Nearly 3 Decades of Research and Future Directions

Pauline Larrouy-Maestri[1,2,3], David Poeppel[3,4,5], and Marc D. Pell[2,6]

[1]Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany; [2]School of Communication Sciences and Disorders, McGill University; [3]Max Planck-NYU Center for Language, Music, and Emotion, New York, New York; [4]Department of Psychology and Center for Neural Science, New York University; [5]Ernst Strüngmann Institute for Neuroscience, Frankfurt, Germany; and [6]Centre for Research on Brain, Language, and Music, Montreal, Quebec, Canada

Abstract

Emotional voices attract considerable attention. A search on any browser using "emotional prosody" as a key phrase leads to more than a million entries. Such interest is evident in the scientific literature as well; readers are reminded in the introductory paragraphs of countless articles of the great importance of prosody and that listeners easily infer the emotional state of speakers through acoustic information. However, despite decades of research on this topic and important achievements, the mapping between acoustics and emotional states is still unclear. In this article, we chart the rich literature on emotional prosody for both newcomers to the field and researchers seeking updates. We also summarize problems revealed by a sample of the literature of the last decades and propose concrete research directions for addressing them, ultimately to satisfy the need for more mechanistic knowledge of emotional prosody.

Keywords

affective science, emotion, perception, acoustics, human communication

Prosody—the sound properties of vocal expressions—conveys linguistic as well as paralinguistic information, such as a speakers' intention (for the case of irony, see Larrouy-Maestri et al., 2023a) and a speakers' emotional state (Banse & Scherer, 1996). Prosody is thus a crucial tool for human communication.[1] When it comes to the communication of emotions, a minimal correspondence between the acoustics properties of the signal and the production/perception of a certain emotional or affective state is assumed.[2] For instance, an influential model of emotion expression and perception proposed by Bänziger et al. (2015), based on Brunswik's lens model and adapted from Scherer (2013a), distinguishes distal information (i.e., internal state of the speaker as estimated by acoustic analysis of their voice) and proximal information (i.e., listeners' perception). It addresses both the encoding and decoding processes involved in the vocal communication of emotions in terms of acoustic cues (for an introduction, see also Kamiloğlu & Sauter, 2021). However, the mapping between acoustic information and emotions, or what can be called the *sound of emotional prosody*, remains poorly defined. In the core of this article, we describe progress (and limits) in the search for the sound of emotional prosody and highlight ways to address current challenges.

As a prelude, and to convince skeptical readers about the relevance of emotional prosody to the social, natural, and computational sciences, we outline three of the (many) domains that will benefit from a deeper understanding of this topic. First, psychology (and its developmental, cognitive, and social aspects) would obviously profit from scientific advances because emotional prosody plays a central role in language and communication across the life span. On the perception side, it has been shown that we are sensitive to

Corresponding Author:
Pauline Larrouy-Maestri, Max Planck Institute for Empirical Aesthetics
Email: plm@ae.mpg.de

emotional prosody at an early age (e.g., event-related potential data in sleeping neonates; D. Zhang et al., 2014). The ability to correctly interpret affective states (i.e., positive, neutral, and negative) from expressive speech is already efficient around 5 years and improves with age, although with large individual differences (Sauter et al., 2013). On the production side, infants' vocalizations increase in complexity early on (Wermke et al., 2021), with intonation patterns found in the first months (Snow & Balog, 2002). Children quickly become proficient in using prosody to be understood (reviewed in Esteve-Gibert & Prieto, 2018). Over time, humans become experienced speakers and listeners, using prosody to form and maintain social positions relative to others (Cheng et al., 2016; Fischer & Manstead, 2008), which in turn influences the behavior of communication partners (Bandstra et al., 2011). Importantly, the effective use of emotional prosody is challenged by aging (Lima et al., 2014; Paulmann et al., 2008). Clarifying the life-span development curve (i.e., from emergence to decline) of emotional prosody, its relation to cognitive abilities, and its role in human interactions relies on a proper description of the sound of emotional prosody.

Second, research on the sound of emotional prosody has clinical implications and thus impacts the medical sciences. The use of emotional prosody in typical communicative contexts, although seemingly natural and effortless, reflects a complex array of perceptual, cognitive, and motor functions that can be selectively disrupted. Deficits in the perception and production of emotional prosody have been identified in children using cochlear implants (e.g., Geers et al., 2013), children with autism spectrum disorder (Rosenblau et al., 2017; Yoshimatsu et al., 2016), and children with attention-deficit/hyperactivity disorder (Chronaki et al., 2015). Difficulties can also appear in brain-damaged patients (e.g., Heilman et al., 2004; Pell & Baum, 1997; Van Lancker & Sidtis, 1992) and in adults with clinical conditions such as schizophrenia (Kantrowitz et al., 2015; Pinheiro et al., 2013), dementia of the Alzheimer's type (Horley et al., 2010), Parkinson's disease (Ariatti et al., 2008; Pell, 1996), and depression (e.g., Cummins et al., 2015; Kan et al., 2004; Schlipf et al., 2013). Difficulties using emotional prosody can understandably be debilitating and have considerable consequences for these individuals. It is thus necessary to develop precise diagnostic tools, rehabilitation programs, or coping strategies, all of which rely on a more comprehensive and mechanistic understanding of emotional prosody.

Finally, we live in a society in which the place and role of technology undeniably increase.[3] On the expression side, more and more devices incorporate artificial speech (Robinson & el Kaliouby, 2009) and aim at sounding as "human" as possible (Drahota et al., 2008) to facilitate human–computer interactions. On the recognition side, the objective is to build tools that can adequately capture the emotional state of a speaker.[4] Numerous applications of automatic emotion-tracking tools (e.g., Alonso et al., 2017; Wang et al., 2015) have already been proposed, for instance, to improve in-car safety systems (Eyben et al., 2010) and to detect stress or frustration or annoyance in speakers' voices (e.g., Ang et al., 2002; X. Zhang, Wang, et al., 2015; Zhou et al., 2001). Importantly, benefits of these tools are foreseen in pedagogical and medical contexts in which the communication through nonverbal behaviors between pupil/teacher or patient/physician is crucial (e.g., Alexander et al., 2015; Baruch et al., 2016; Dubey et al., 2016; Griol et al., 2014; Persky et al., 2016; Rochman & Amir, 2013). In addition to easing communication, such noninvasive tools appear promising for detecting disorders such as depression (e.g., Alghowinem et al., 2013; Pan et al., 2019) or autism (Asgari et al., 2021) and thus may be of benefit to public health.

State of the Art on the Sound of Emotional Prosody

Over the years, several attempts have been made to identify the relevant cues or features of emotional prosody (Murray & Arnott, 1993; Scherer, 1986). As summarized in Bänziger et al. (2015), emotional prosody has been examined from two different angles concurrently. Some studies have focused on acoustic aspects (i.e., encoding), whereas others have focused on the recognition of emotions by listeners (i.e., decoding). The number of encoding studies, in particular, has increased dramatically in tandem with technological advances (for a review of early studies, see Juslin & Laukka, 2003).

One major step toward identifying the acoustic characteristics of emotional prosody was attained by Banse and Scherer (1996). Their study represented a dramatic improvement in methods compared with previous work because they analyzed substantially more affective states (n = 14) and increased the number of acoustic features (n = 29 relative to pitch, spectral, and temporal dimensions). As reported in Table 1, listeners' recognition of specific emotions could be predicted by different constellations of features. Importantly, using a jackknifing procedure, the authors identified a subset of the 16 best performing features from the initial 29 parameters: four features concerning the fundamental frequency (f_0); one related to *speech rate*, an estimate of *loudness*; and the others related to *vocal quality/timbre*. The work of Banse and Scherer (1996) has inspired years of research on the sound of emotional

Table 1. Acoustic Predictors and General Description of Emotion Categories According to Banse and Scherer (1996)

Emotion category	Dimensions				General description
	Pitch	Temporal	Loudness	Timbre	
Hot anger	X			X	High and bright voice with limited pitch fluctuations
Panic fear	X				High-pitched voice with limited fluctuations
Anxiety	X		X		Quiet voice in the middle pitch range with limited pitch fluctuations
Desperation	X	X		X	High and bright voice with limited pitch fluctuations and a slow speech rate
Sadness			X	X	Quiet and thin voice
Elation	X			X	High-pitched voice with some fluctuations
Boredom	X	X	X		Low and quiet voice with slow speech rate
Shame			X		Quiet voice
Pride	X				Low-pitched voice
Contempt	X				Low-pitched voice with some pitch fluctuations

Note: The pitch dimension includes the mean and standard deviation of f_0. The temporal dimension refers to the duration of articulation periods (i.e., the duration of nonsilent periods). Loudness is estimated with the mean energy (mean of the log-transformed microphone voltage). Timbre includes the Hammerberg index (difference between the energy maximum in the 0–2000-Hz frequency band and in the 2000–5000-Hz band), the proportion of voiced energy up to 1000 Hz, and the slope of spectral energy above 1000 Hz. "X" denotes the significant contribution of acoustic dimensions in predicting the categorization for each emotion. Note that the fit of statistical models for happiness, cold anger, interest, and disgust were lower or the specific contribution of features was unclear. These emotional states are not reported here; for a full description, see Banse and Scherer (1996).

prosody and became a standard reference article cited by researchers from the computer sciences, social sciences, neurosciences, medical sciences, and the humanities.

Since that landmark publication, extensive effort has been made to describe the mapping between acoustics and emotional prosody, in particular by extending the number of acoustic features examined. Figure 1 summarizes a chronological reading of English articles published between 1996 and 2021. By no means exhaustive, this list is grounded in a simple search procedure suited to this interdisciplinary topic: Google Scholar. Indeed, research on emotional prosody can be found in different types of publications that specific tools such as PubMed or Scopus do not necessarily cover. For instance, conference proceedings or patents are the main dissemination technique in engineering, whereas work in the humanities is reported in books and research in the social sciences is described in peer-reviewed journals. Concretely, we used Google Scholar without restriction regarding the format and looked at all entries citing the reference article (Banse & Scherer, 1996). Because of space, we limited the must-read empirical articles to a few references, but a large number of articles with experimental approaches, from both the social sciences and computer sciences, can be found throughout the selection of reviews in Figure 1.

Thanks to technological advances from individual teams and in response to scientific calls for innovations (e.g., INTERSPEECH 2009 Emotion Challenge; Schuller et al., 2009), the number of acoustic features found to be associated with emotional-prosody classification has dramatically grown. Much progress can be observed not only regarding the features quantifying quality/spectral features (see Fig. 1, blue dots) but also in the identification of other features or their interactions with information such as phonemic characteristics or semantic content (see Fig. 1, orange dots). As the number of acoustic features examined increased (Fig. 1), selection/reduction strategies became necessary to identify the most relevant ones (e.g., Dropuljić et al., 2013; Huang et al., 2009; McGilloway et al., 2000; Oudeyer, 2003; Schuller et al., 2004). In an attempt to standardize measurements, Eyben et al. (2016) proposed both a "minimalistic" parameter set (GeMAPS) containing 18 low-level descriptors (relative to frequency, energy/amplitude, and spectrum), some of their derivatives (leading to a total of 56 parameters), and six temporal features. The authors provided a publicly available implementation (with the openSMILE toolkit) to analyze a total of 62 parameters. This set of parameters can then be complemented by additional low-level features, cepstral parameters, as well as dynamic parameters (i.e., the "extended" version of 88 parameters, eGeMAPS), or by any potential additional features relevant for a specific research question or material.

Crucially, the idea of acoustic changes over time or dynamics, which was already found in Fairbanks and

Pronovost (1938), has been developed in the last decades (Fig. 1, red dots). Whereas most acoustic features reported in the literature are summary statistics over a unit (word or phrase or sentence), the role of dynamics or pitch changes over time has been repeatedly shown (e.g., Grichkovtsova et al., 2012; Pell & Kotz, 2011; van Rijn et al., 2023), and some attempts have been made to quantify them. For instance, Bänziger and Scherer (2005) and Rodero (2011) marked key points to describe contours through stylization, with tools such as the modeling melody algorithm MOMEL and the International Transcription System for Intonation (INTSINT) developed by Hirst (2005). Another method, proposed by Alonso et al. (2017), consists of modeling the pitch trajectory and interpreting the linear-regression coefficients to describe the pitch height and declination or trend of the pitch contour. More recently, van Rijn et al. (2023) quantified the pitch shape of sentences from existing emotional-prosody corpora in three different ways, including a morphometric method (for previous use in other domains, see M. A. Knoll & Costall, 2015; MacLeod, 1999). Although there is room for improvement of the measures, their study showed that such a method helps capturing the f_0 changes over time and improves the classification of emotions.

To describe the sound of emotional prosody, other approaches that make use of updated statistical methods have also emerged in recent years. Cowen et al. (2019) explored emotion recognition from prosody by analyzing the acoustic correlates of 2,519 speech samples and observed the acoustic features (of speech from

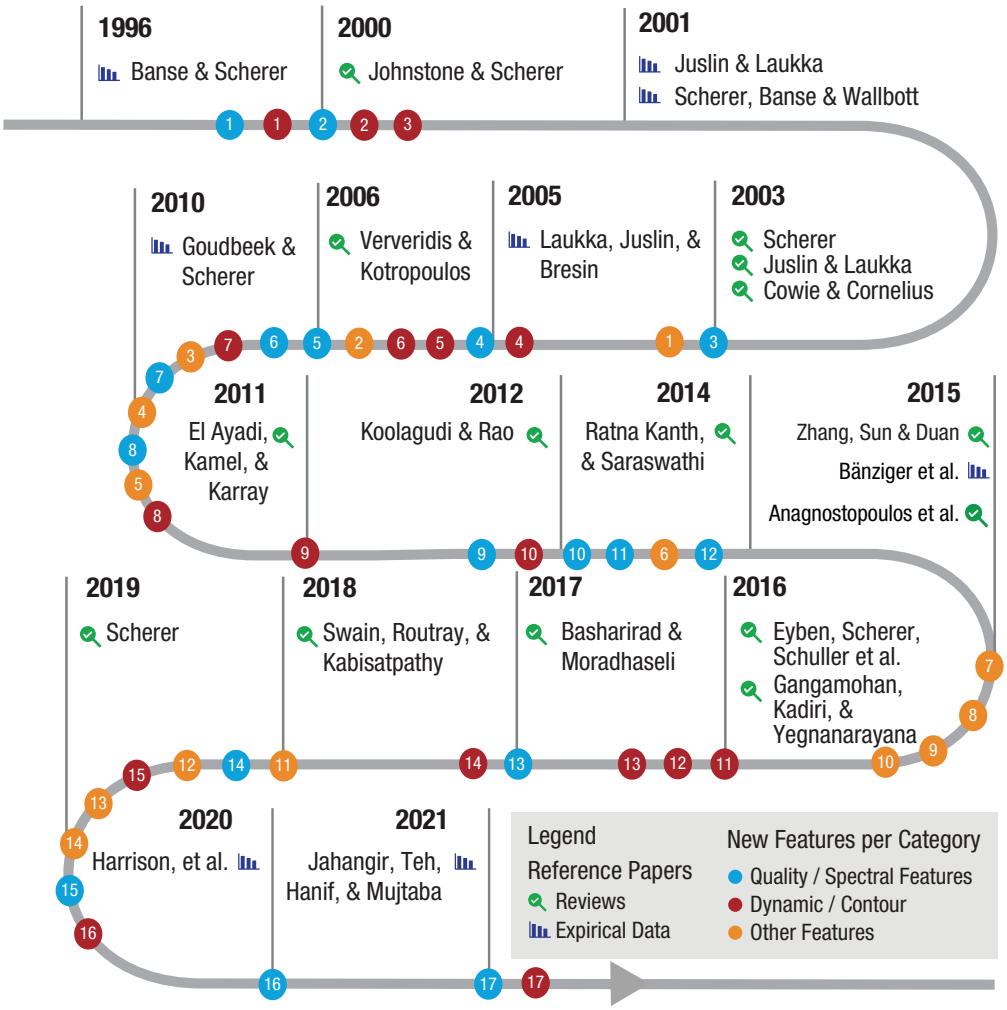

Fig. 1. *(continued on next page)*

1. Whiteside (1999): **Perturbation measures**
2. Burkhardt & Sendlmeier (2000): **Formants**
3. Gobl & Chasaide (2003): **Different types of voices (harsh, tense, modal, breathy, whispery, creaky, and lax–creaky)**
4. Borchert & Düsterhöft (2005): **Harmonic to noise ratio with different spectral bands**
5. Klabbers et al. (2007): **Four energy bands, overall energy, spectral tilt**
6. Tamarit et al. (2008): **Spectral slope measurements adapted to the speaker**
7. Bitouk et al. (2009): **Spectral features of specific phonemes**
8. Luengo et al. (2010): **Spectral envelope and voice quality measures**
9. Grichkovstsova et al. (2012): **Voice quality manipulation (exchange between utterances)**
10. Amarakeerthi et al. (2013): **Inter- and intra-subband energy variations**
11. Elbarougy & Akagi (2013): **Difference between the first and second harmonics**
12. Kabuta et al. (2013): **Spectral-tilt and higher-frequency bands**
13. Mouawad & Dubnov (2017): **Nonlinear dynamical features**
14. Zhu et al. (2018): **Modulation spectral features**
15. Agrima et al. (2019): **Six energy bands**
16. Kadiri et al. (2020): **Excitation features**
17. Singh et al. (2021): **Low frequencies**

1. Mozziconacci & Hermes (1999): **Intonational patterns**
2. McGilloway et al. (2000): **Number of inflections in F0 contours**
3. Fellenz et al. (2000): **Dynamical speech features**
4. Bänziger & Scherer (2005): **Stylization procedure**
5. Knoll et al. (2006): **Five contour categories (bell shaped, complex, falling, rising, level shaped)**
6. Ververidis & Kotropoulos (2006): **Downward and rising slopes**
7. Klabbers et al. (2007): **Decomposition of pitch curves**
8. Luengo et al. (2010): **Sentence-end features**
9. Rodero (2011): **Contrasted stylized contours**
10. Grichkovstsova et al. (2012): **Prosodic contours**
11. Eyben et al. (2016): **Mean and standard deviation of rising/falling slopes**
12. Madureira (2016): **Role of dynamic non-dichotomic models**
13. Gangamohan et al. (2016): **Irregular fluctuations**
14. Alonso et al. (2017): **Linear regression coefficients obtained from modelling F0 contours**
15. Rajković et al. (2018): **Short intonation contours**
16. Van Mersbergen & Lanza (2019): **Relative movements of F0**
17. van Rijn et al. (2023). **Slope description**

1. Schuller et al. (2004): **Linguistic and acoustic information**
2. Airas & Alku (2006): **Glottal flow and inverse filtering parameters**
3. Hozjan & Kačič (2006): **Phonemes and pause durations**
4. Bitouk et al. (2009): **Phonetic classes**
5. Yang & Lugger (2010): **Features inspired from music theory (harmony)**
6. Chen et al. (2013): **Vocal folds excitation**
7. Gievska et al. (2015): **Bimodal (linguistic and prosodic) information**
8. Zhang et al. (2015): **Acoustic and linguistic features along with contexts**
9. Lee & Tashev (2015): **Long-range context and uncertainty of emotional label expressions**
10. Birkholz et al. (2015): **Phonation type (e.g., breathy, modal, and pressed)**
11. Shigeno (2018): **Semantic content of words**
12. Rilliard et al. (2018): **Paradigmatic variations in vowels**
13. Gudmalwar et al. (2018): **Feature vectors from several statistical parameters**
14. Dimitrova-Grekow & Konopko. (2019): **Laryngeal tone**

Fig. 1. Advances in the acoustic description of emotional prosody since 1996. A limited number of reference articles representing key reviews (magnifying-glass symbol) and highly cited empirical reports (histogram symbol) are especially marked. The numbers represent examples of references using additional features relative to the quality/spectral features (blue), dynamic/contour features (red), as well as other features (orange) over the years.

100 actors across five cultures) that tracked 12 dimensions or emotion categories. Their comparison between emotion judgments and acoustic properties across cultures highlighted the relevance of several features, namely duration, pause time, mean f_0, minimum/maximum f_0, first/second/third average formant frequencies, first/third quartiles of the frequency spectrum, spectral centroid, and pitch salience. Another example of the benefit of big data analysis can be found in van Rijn and Larrouy-Maestri (2023), who examined 3,000 min of recordings from various corpora across the globe. Whereas the mapping between acoustic features and emotions varied across corpora, seven acoustic factors named according to the type of features loading on each dimension explained a total variance of 57%: voice quality, loudness, pitch/formants, rhythm/tempo, shimmer, pitch variation, and mel-frequency cepstrum. The factor solutions were quite robust across the most common countries and languages in the data sets. With this elaborated approach, the work of Cowen

Table 2. Suggestions for Next Steps to Investigate the Acoustics–Emotion Mapping

Nonexhaustive sources of variability	Potential next steps
Speech material	
Length and language of the material	Examine the effect of length as well as the role of linguistic/phonological/semantic content of speech material on the acoustics–emotion mapping
Stereotypicality	Investigate the notion of stereotypicality (or caricature) in recorded material, potentially modulated by the type of speaker being recorded; increase variability in the material by recording professional singers who are used to being recorded but not trained in speech production (procedure used in Holz et al., 2022)
Acoustic characteristics	
Choice of the unit size	Identify the minimal size of relevant units, their distinct roles, and their integration in emotion communication through speech
Dynamic aspect	Quantify the dynamic aspect of emotional prosody and its role in emotion communication
Direction and magnitude	Describe the direction and magnitude of acoustic features responsible for the recognition of speakers' emotional states through prosody, for instance, by investigating listeners' perceptual thresholds (for methods proposed in the music domain, see Larrouy-Maestri et al., 2019)
Other factors	
Culture	Determine and quantify factors that affect the emotion-acoustic mapping
Emotion type	Extend research to the large range of affective expressions that reflect human emotional states and their nuances
Authenticity	Account for potential overacted material by including speaker or associated variables in statistical analyses (e.g., Zloteanu & Krumhuber, 2021); ensure the perceived authenticity of new data sets by using authenticity ratings instead of (or in addition to) emotion recognition as an inclusion threshold (as proposed by Holz et al., 2022); examine the "humanness" of emotional prosody by examining what makes synthetic voices nonauthentic

et al. (2019) and van Rijn and Larrouy-Maestri (2023) confirmed the relevance of key acoustic features in the communication of emotion through prosody but also highlighted the complexity (and opacity) of the mapping between acoustic features and emotion in speech.

Toward an Updated Definition of Emotional Prosody

Despite the tremendous progress that has occurred since Banse and Scherer (1996), our chronological reading of articles published since then does not lead to a comprehensive and definitive description of the sound of specific emotions. This conclusion was unexpected and probably disappointing (to us and to the reader). As a matter of fact, we observe a lack of consensus between studies, which makes a tentative description particularly speculative. In this section, we discuss possible sources of variability relative to the speech material and to the acoustic characteristics examined. We also reflect on the role of additional factors in the acoustics–emotion mapping. Without being exhaustive, we suggest directions for addressing each point raised in Table 2.

Factors relative to the speech material

The material found in existing data sets ranges from single vowels to full sentences. It has been shown repeatedly that emotions encoded in very short stimuli can be recognized (Paulmann & Kotz, 2008) and that emotion recognition improves as an utterance unfolds or accumulates (Pell & Kotz, 2011; Rigoulot et al., 2013). In addition to differences in terms of the amount of acoustic information available to the listener (Roche et al., 2015), the length of the speech material also affects the nature of acoustic information. For instance, the spectral characteristics of a single /a/ (Waaramaa et al., 2010) will be different from those of a sentence containing various consonants and a large variance among vowels. Note that the material is rarely phonologically balanced, that is, constituted of phonemes of equal frequency of occurrence in natural speech. Therefore, in addition to changing the amount of acoustic information available to the listener, the length of the speech material affects its acoustic characteristics.

In addition to the length of the speech samples, the specific language in which emotional prosody is embedded greatly differs between data sets. Although studying emotional prosody in the context of existing languages

enhances the "natural" aspect of the material, it has the disadvantage of allowing an interaction between the emotional prosody and the semantic content of the material (Pell et al., 2011). An alternative could be to use filtered speech (e.g., Bryant & Barrett, 2008) in which the information necessary to access lexical-semantic information is filtered out, thus rendering speech unintelligible (e.g., Flinker et al., 2019). However, removing spectral information might also affect emotional-prosody perception because voice timbre/quality plays a key role (see Fig. 1). Another alternative is to use pseudospeech (e.g., Banse & Scherer, 1996; Pell & Kotz, 2011). However, the creation of Jabberwocky sentences is not random but aims to preserve the rules of specific languages because listeners develop cultural expectations, and "foreign-sounding" material might influence emotional-prosody perception (Liu et al., 2015).

Another important decision for the creation of emotional speech material concerns the recordings and their selection. Ideal data sets should reflect real-life affective utterances produced in typical situations. However, examining the sound of emotional prosody usually requires a certain level of control with regard to the emotional content (i.e., what was specifically intended to be conveyed), the linguistic material (i.e., similar material across emotions), or the speaker (i.e., same performer for different emotions). Therefore, recordings are typically performed in laboratory settings by invited actors or nonactors. It has been shown that the acoustics of play-acted (or posed) recordings differ from those of spontaneous recordings (Jürgens et al., 2011; Juslin et al., 2018). One can assume that actors are able to express themselves in different (imagined) emotional states, thus providing different versions of specific sentences that can be directly compared. In addition, because actors are used to speaking in front of audiences and to being recorded, their stress level (documented as influencing vocal productions; Larrouy-Maestri & Morsomme, 2014; Paulmann et al., 2016) may be lower than that of nonactors in recording situations. Despite these advantages, several shortcomings are potentially associated with actors, such as the overuse of caricatures or stereotypes (Banse & Scherer, 1996; Drolet et al., 2012; Jürgens et al., 2013; Scherer, 2013b), and suggest that nonactors may be more suitable speakers. However, nonactors might be acting as well, without having adequate training to express emotions with plausible variability, and thus may also produce stereotypical stimuli.

In addition to potential factors linked to speakers' characteristics, the selection of the material itself can play a role in its stereotypicality and thus on the acoustics–emotion mapping. Banse and Scherer (1996) and subsequently several others included recognition tasks performed by small groups of judges or experimenters to discard stimuli that were poorly recognized. Such a procedure is often presented as a validation step. However, by reducing the initial set, this procedure reduces the acoustic variability (e.g., small standard deviations around the mean for each acoustic feature analyzed), which may likewise affect the quality of statistical models and thus bias the acoustic-emotional prosody association observed. In other words, the acoustic content of the material, and probably its stereotypicality as well, depends on the threshold applied to the recognition task for the selection of the speech material to examine.

Factors relative to the acoustic characteristics

Linguistic elements of different sizes, such as words, phrases, and sentences, are concurrently tracked and temporally integrated (e.g., Ding et al., 2016; Keitel et al., 2018). With regard to emotional prosody, it seems reasonable to hypothesize that units of different size exist and are integrated over time (Jiang et al., 2015; Pell & Kotz, 2011; Waaramaa et al., 2010). Previous research has focused on different units, such as sentences (Chen et al., 2012), segments (Schuller & Rigoll, 2006; Shami & Kamel, 2005), syllables (Agrima et al., 2019), phonemes (Bitouk et al., 2009; Hyun et al., 2010), or selected vowels (Goudbeek et al., 2009), thus supporting the role of acoustic information at these different levels. As a consequence, a realistic acoustics–emotion mapping would require a better understanding of how the acoustic features of speech units of different size potentially interact in longer segments.

In line with the idea of units and supported by empirical evidence (e.g., Grichkovtsova et al., 2012; van Rijn et al., 2023), the dynamics of speech, or how features change over time, greatly matters to listeners. As illustrated in Figure 1 (red dots), several attempts have been made to describe the dynamics of emotional prosody, with symbolic representations (e.g., tones and break indices: Silverman et al., 1992; INTSINT: Hirst, 2005), melodic contours (Cullen et al., 2008; see also Adams, 1976), linear and quadratic functions (Hoicka & Gattis, 2012), or using a morphometric approach (van Rijn et al., 2023). Although such tools and methods are promising, research on emotional prosody could also benefit from descriptors being proposed in adjacent research topics. For instance, pitch trajectories in single words have been quantified when studying trustworthiness perception (Belin et al., 2017), dominance (Ponsot et al., 2018), and certainty/honesty (Goupil et al., 2021). Note that acoustic changes are not limited to pitch but occur in the case of duration and loudness (Goupil

et al., 2021) or their combination, as shown in research on stress and prominence perception (for a discussion, see Cole & Shattuck-Hufnagel, 2016).

Finally, the identification of new acoustic features or of their changes over time does not necessarily inform us about the relevance of their direction and magnitude. For instance, low pitch is often associated with a "sad" emotional state relative to the same speaker performing a "happy" stimulus (Banse & Scherer, 1996), but that does not say "how much lower" the voice should be to sound sad. Laukka (2005) presented listeners vocal expressions that were created by morphing prototypical ones along continua (e.g., happiness–sadness or anger–fear). The results of the identification task supported the idea that changes of pitch, intensity, duration, and timbre shift the perception of the emotion. To the best of our knowledge, this promising finding has not been followed by explicit thresholding procedures as proposed in other domains. For instance, in the music domain, Larrouy-Maestri (2018) manipulated the magnitude of a relevant characteristic (i.e., enlarging or compressing pitch intervals within short tonal melodies) and identified thresholds above which performances were no longer perceived as in tune and were interpreted as out of tune. Of course, it is legitimate to wonder whether such approaches can be easily transferred across domains; however, one could argue that, even if there are differences in terms of the content (acoustic features and units of information) or functions between speech and music, there are similarities in terms of the processes underlying their perception such as their categorization (Larrouy-Maestri et al., 2023b). As a consequence, it seems realistic that the manipulation of single acoustic features, as applied in music, could be used to pursue the approach initiated by Laukka (2005) and determine boundaries between categories (i.e., specific emotions) in the case of emotional prosody.

Other factors affecting the acoustics–emotion mapping

A large number of studies have revealed an in-group advantage for the recognition of speakers' emotional states through emotional prosody (e.g., Jürgens et al., 2013; Koeda et al., 2013; Laukka et al., 2016; Paulmann & Uskul, 2014; Pell et al., 2009; Riviello & Esposito, 2012; Sauter, 2013; Sauter et al., 2010; Sauter & Scott, 2007; Scherer et al., 2001; Tisljár-Szabó & Pléh, 2014; Waaramaa & Leisiö, 2013). More recently, van Rijn and Larrouy-Maestri (2023) used large-scale Bayesian inference models to quantify the role of culture (country and language of the speaker) on the mapping between acoustic and intended emotions by analyzing a large set of collected speech corpora (more than 3,000 min of emotional speech). Unsurprisingly, culture substantially affected the correspondence between the intended emotional state of the speaker and acoustics of the vocal expressions, which confirms that growing up in a specific cultural and language environment may thus shape the acoustics–emotion association both in production and perception.

Another factor that has been overlooked refers to the granularity of emotions (Kamiloğlu et al., 2020). In the case of positive emotions, research has only recently focused on more than a very limited number of emotions (Sauter & Scott, 2007), and the comparison of the acoustic profiles of different positive emotions revealed differences in the acoustics. For instance, pitch was higher for joy and amusement but lower for lust and admiration, or speech rate was faster for joy and pride but slower for pleasure. Therefore, grouping all positive emotions under a single or limited number of terms (Scherer, 1986) is misleading. More generally, the emotional space in which we communicate is richer than previously studied (Cowen & Keltner, 2021; Keltner, 2019; Keltner et al., 2019), which supports the need for diversification from the six basic emotions studied by Ekman in the 1970s (Ekman & Friesen, 1971; for an extensive description, see Ekman, 1992, 1999). For the study of emotional prosody, such findings encourage researchers to further extend the usual number of emotional states or dimensions (e.g., 14 in Banse & Scherer, 1996; 12 in Cowen et al., 2019) to reach a more realistic view of the range of emotions communicated through prosody.

Last, the role of expression authenticity on the acoustics–emotion mapping is of great interest in a society in which humans are surrounded by synthetic speech. Text-to-speech tools and AI voice generators aim to create intelligible and realistic sounds but, whereas intelligibility is generally accomplished, the voices do not always sound natural and somehow lack "humanity." In the emotional-prosody literature, whoever is being recorded (actor, nonactor, singer) receives instructions ranging from the direct request of expressing a specific affective state to techniques to induce specific emotions in performers (Bänziger et al., 2012; see also Kamiloğlu et al., 2020), the latter encouraging spontaneity and thus increasing the genuineness of the expressions (Laukka et al., 2013; Lima et al., 2013). It has been suggested that the use of play-acted stimuli affects listeners' perceptions of the auditory signal (Anikin & Lima, 2017; Drolet et al., 2012; 2014; Lavan et al., 2016). Drolet et al. (2013) observed that the effect of the authenticity of speech (and its potential relation to acoustic features; see Dropuljić et al., 2017) on emotion categorization is reflected early in cortical processing. Whether authenticity is considered in terms of speakers'

ability to express convincing expressions or in terms of listeners' perception is currently under discussion (Zloteanu & Krumhuber, 2021), but in light of its relevance, it would certainly be an important factor to further investigate.

Conclusion

Although the existence of an acoustic signature for each possible emotional state is illusory, we (human speakers and listeners) use acoustic cues naturally and seemingly with ease to infer others' emotional states beyond words. Inspired by the work of Banse and Scherer (1996), the major advances of the last decades have set the stage for a much better understanding of these cues and how they are used in human communication. Nevertheless, the relentless enthusiasm of scientists in various fields has not been sufficient to fully define emotional prosody and clarify the nature of this crucial but complex phenomenon. We hope that by reflecting on potential issues that prevent a consensus about the acoustics–emotion mapping, future research will be in a better position to constructively move this field ahead. We invite the research community to address current challenges and establish a solid foundation for successfully characterizing the sound of emotional prosody, which is located at the nexus of the humanities, computational approaches, and the psychological and brain sciences.

Transparency

Action Editor: Laura A. King
Editor: Laura A. King
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
This work was supported by a Fonds de recherche du Québec Merit Scholarship for Foreign Students (to P. Larrouy-Maestri), as well as by a Social Sciences and Humanities Research Council of Canada Insight Grant 435-2017-0885 (to M. D. Pell), and by the Max Planck Society.

ORCID iDs

Pauline Larrouy-Maestri https://orcid.org/0000-0001-9245-0743
Marc D. Pell https://orcid.org/0000-0002-9947-1552

Acknowledgments

We are grateful to Pol van Rijn for help with the figure, Franziska Hannig and Lauren Fink for edits, and Nina Kazanina for valuable comments on a previous version of the manuscript.

Notes

1. Communication can be efficient using other modalities (reading, sign language, Braille, nonverbal communication), but here we focus specifically on speech and spoken language understanding.
2. The strength of the relation between the physical signal and the expression or perception of a specific emotion depends on the theoretical framework; for example, there is a strong acoustic-emotion relation in affect program theories but a more flexible relation in appraisal and constructivist ones. Interestingly, there is empirical evidence supporting both a straightforward mapping (e.g., association between roughness of screams and the expression of fear; see Arnal et al., 2015; for a review on neural response patterning, see Cowen & Keltner, 2021) and a more complex one (see Barrett, 2017; for an example in the visual domain, see Barrett et al., 2019; for a discussion on universality, see Gendron et al., 2018).
3. The humanization of machines, by improving the quality of the expression, and the improvement of recognition systems can also be viewed as threatening. On the recognition side, automatic systems to be used as assistance tools (e.g., to help the elderly with activities around the home) cannot fully substitute necessary social relations with other humans. On the expression side, the ability to mimic emotional states opens the door to manipulating human behavior because emotional prosody influences the behavior of communication partners (Bandstra et al., 2011; Kramer et al., 2014). This has implications in several contexts, such as negotiations (Sinaceur et al., 2015; Wubben et al., 2011), decision-making (Boidron et al., 2016), and politics (Banai et al., 2017; Dietrich, 2014; Marcus et al., 2000). Therefore, to be sensitive to positive versus negative implications, technological advances must be undertaken responsibly and in a well-considered ethical framework.
4. As discussed in Note 2, such a statement can be questioned in the case of constructivist theories of emotions because the relation between an acoustic and a specific emotional state is not straightforward.

References

Adams, C. R. (1976). Melodic contour typology. *Ethnomusicology*, *20*(2), 179–215. https://doi.org/10.2307/851015
Agrima, A., Farchi, A., Elmazouzi, L., Mounir, I., & Mounir, B. (2019). Emotion recognition from Moroccan dialect speech and energy band distribution. In *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)* (pp. 471–476). Institute of Electrical and Electronics Engineers.
Airas, M., & Alku, P. (2006). Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*, *63*(1), 26–46. https://doi.org/10.1159/000091405
Alexander, S. C., Garner, D. K., Somoroff, M., Gramling, D. J., Norton, S. A., & Gramling, R. (2015). Using music[al] knowledge to represent expressions of emotions. *Patient Education and Counseling*, *98*(11), 1339–1345. https://doi.org/10.1016/j.pec.2015.04.019
Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Gedeon, T., Breakspear, M., & Parker, G. (2013). A comparative study

of different classifiers for detecting depression from spontaneous speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8022–8026). Institute of Electrical and Electronics Engineers.

Alonso, J. B., Cabrera, J., Travieso, C. M., López-de-Ipiña, K., & Sánchez-Medina, A. (2017). Continuous tracking of the emotion temperature. *Neurocomputing*, *255*, 17–25. https://doi.org/10.1016/j.neucom.2016.06.093

Amarakeerthi, S., Morikawa, C., Nwe, T. L., Silva, L. C. D., & Cohen, M. (2013). Cascaded subband energy-based emotion classification. *IEEJ Transactions on Electronics, Information and Systems*, *133*(1), 200–210. https://doi.org/10.1541/ieejeiss.133.200

Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review*, *43*(2), 155–177. https://doi.org/10.1007/s10462-012-9368-5

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In J. H. L. Hansen & B. Pellom (Eds.), *7th International Conference on Spoken Language Processing (ICSLP-2002)* (pp. 2037–2040). International Speech Communication Association.

Anikin, A., & Lima, C. F. (2017). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *The Quarterly Journal of Experimental Psychology*, *71*(3), 622–641. https://doi.org/10.1080/17470218.2016.1270976

Ariatti, A., Benuzzi, F., & Nichelli, P. (2008). Recognition of emotions from visual and prosodic cues in Parkinson's disease. *Neurological Sciences*, *29*(4), 219–227. https://doi.org/10.1007/s10072-008-0971-9

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051–2056. https://doi.org/10.1016/j.cub.2015.06.043

Asgari, M., Chen, L., & Fombonne, E. (2021). Quantifying voice characteristics for detecting autism. *Frontiers in Psychology*, *12*, Article 665096. https://doi.org/10.3389/fpsyg.2021.665096

Banai, I. P., Banai, B., & Bovan, K. (2017). Vocal characteristics of presidential candidates can predict the outcome of actual elections. *Evolution and Human Behavior*, *38*(3), 309–314. https://doi.org/10.1016/j.evolhumbehav.2016.10.012

Bandstra, N. F., Chambers, C. T., McGrath, P. J., & Moore, C. (2011). The behavioural expression of empathy to others' pain versus others' sadness in young children. *Pain*, *152*(5), 1074–1082. https://doi.org/10.1016/j.pain.2011.01.024

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path models of vocal emotion communication. *PLOS ONE*, *10*(9), Article e0136675. https://doi.org/10.1371/journal.pone.0136675

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, *12*, 1161–1179. https://doi.org/10.1037/a0025827

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, *46*(3–4), 252–267. https://doi.org/10.1016/j.specom.2005.02.016

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1–23.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Baruch, Y. K., Spektor-Levy, O., & Mashal, N. (2016). Preschoolers' verbal and behavioral responses as indicators of attitudes and scientific curiosity. *International Journal of Science and Mathematics Education*, *14*(1), 125–148. https://doi.org/10.1007/s10763-014-9573-6

Basharirad, B., & Moradhaseli, M. (2017). Speech emotion recognition methods: A literature review. *AIP Conference Proceedings*, *1891*, Article 020105. https://doi.org/10.1063/1.5005438

Belin, P., Boehme, B., & McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLOS ONE*, *12*(10), Article e0185651. https://doi.org/10.1371/journal.pone.0185651

Birkholz, P., Martin, L., Willmes, K., Kroger, B. J., & Neuschaefer-Rube, C. (2015). The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study. *Journal of the Acoustical Society of America*, *137*(3), 1503–1512. https://doi.org/10.1121/1.4906836

Bitouk, D., Nenkova, A., & Verma, R. (2009). Improving emotion recognition using class-level spectral features. In *10th Annual Conference of the International Speech Communication Association (INTERSPEECH-2009)* (pp. 2023–2026). International Speech Communication Association.

Boidron, L., Boudenia, K., Avena, C., Boucheix, J.-M., & Aucouturier, J.-J. (2016). Emergency medical triage decisions are swayed by computer-manipulated cues of physical dominance in caller's voice. *Scientific Reports*, *6*, Article 30219. https://doi.org/10.1038/srep30219

Borchert, M., & Düsterhöft, A. (2005, October 30–November 1). *Emotions in speech–Experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments* [Paper presentation]. 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Wuhan, China.

Bryant, G. A., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture*, *8*(1–2), 135–148. https://doi.org/10.1163/156770908X289242

Burkhardt, F., & Sendlmeier, W. (2000, September 5–7). *Verification of acoustical correlates of emotional speech using formant–Synthesis* [Paper presentation]. ISCA

Workshop on Speech and Emotion, Newcastle, United Kingdom.

Chen, L., Mao, X., Wei, P., & Compare, A. (2013). Speech emotional features extraction based on electroglottograph. *Neural Computation*, *25*(12), 3294–3317. https://doi.org/10.1162/NECO_a_00523

Chen, L., Mao, X., Wei, P., Xue, Y., & Ishizuka, M. (2012). Mandarin emotion recognition combining acoustic and emotional point information. *Applied Intelligence*, *37*(4), 602–612. https://doi.org/10.1007/s10489-012-0352-1

Cheng, J. T., Tracy, J. L., Ho, S., & Henrich, J. (2016). Listen, follow me: Dynamic vocal signals of dominance predict emergent social rank in humans. *Journal of Experimental Psychology: General*, *145*(5), 536–547. https://doi.org/10.1037/xge0000166

Chronaki, G., Benikos, N., Fairchild, G., & Sonuga-Barke, E. J. (2015). Atypical neural responses to vocal anger in attention-deficit/hyperactivity disorder. *Journal of Child Psychology and Psychiatry*, *56*(4), 477–487. https://doi.org/10.1111/jcpp.12312

Cole, J., & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology*, *7*(1), Article 8. https://doi.org/10.5334/labphon.29

Cowen, A. S., & Keltner, D. (2021). Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, *25*(2), 124–136. https://doi.org/10.1016/j.tics.2020.11.004

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, *3*(4), 369–382. https://doi.org/10.1038/s41562-019-0533-6

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*, 5–32. https://doi.org/10.1016/S0167-6393(02)00071-7

Cullen, C., Vaughan, B., & Kousidis, S. (2008, June 29–July 2). *LinguaTag: An emotional speech analysis application* [Paper presentation]. 12th World Multi-Conference on Systemics, Cybernetics and Informatics, Orlando, FL, United States. https://arrow.dit.ie/dmccon/17

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49. https://doi.org/10.1016/j.specom.2015.03.004

Dietrich, B. J. (2014). *It's not what you say, but how you say it. Anger, audio, and the U.S. House of Representatives* [Unpublished doctoral dissertation]. University of Illinois.

Dimitrova-Grekow, T., & Konopko, P. (2019, October 6–9). *New parameters for improving emotion recognition in human voice* [Paper presentation]. 2019 IEEE International Conference on Systems, Man and Cybernetics, Bari, Italy.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Drahota, A., Costall, A., & Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Communication*, *50*(4), 278–287. https://doi.org/10.1016/j.specom.2007.10.001

Drolet, M., Schubotz, R. I., & Fischer, J. (2012). Authenticity affects the recognition of emotions in speech: Behavioral and fMRI evidence. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(1), 140–150. https://doi.org/10.3758/s13415-011-0069-3

Drolet, M., Schubotz, R. I., & Fischer, J. (2013). Explicit authenticity and stimulus features interact to modulate BOLD response induced by emotional speech. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(2), 318–329. https://doi.org/10.3758/s13415-013-0151-0

Drolet, M., Schubotz, R. I., & Fischer, J. (2014). Recognizing the authenticity of emotional expressions: F0 contour matters when you need to know. *Frontiers in Human Neuroscience*, *8*, Article 144. https://doi.org/10.3389/fnhum.2014.00144

Dropuljić, B., Mršić, L., Kopal, R., Skansi, S., & Brkić, A. (2017, April 3–5). *Evaluation of speech perturbation features for measuring authenticity in stress expressions* [Paper presentation]. Asian Conference on Intelligent Information and Database Systems, Kanazawa-shi, Japan.

Dropuljić, B., Popović, S., Petrinović, D., & Ćosić, K. (2013). Estimation of emotional states enhanced by a priori knowledge. In *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 481–486). Institute of Electrical and Electronics Engineers.

Dubey, H., Mehl, M. R., & Mankodiya, K. (2016). BigEAR: Inferring the ambient and emotional correlates from smartphone-based acoustic big data. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (pp. 78–83). Institute of Electrical and Electronics Engineers.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*, 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. J. Power (Eds.), *Handbook of cognition and emotion* (pp. 45–60). John Wiley & Sons. https://doi.org/10.1002/0470013494.ch3

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*, 124–129.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

Elbarougy, R., & Akagi, M. (2013, October 29–November 1). *Cross-lingual speech emotion recognition system based on a three-layer model for human perception* [Paper presentation]. 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Kaohsiung, Taiwan.

Esteve-Gibert, N., & Prieto, P. (2018). Early development of prosody-meaning interface. In N. Esteve-Gibert & P. Prieto (Eds.), *The development of prosody in first language acquisition* (pp. 228–246). John Benjamins.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & 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. https://doi.org/10.1109/taffc.2015.2457417

Eyben, F., Wöllmer, M., Poitschke, T., Schuller, B., Blaschke, C., Färber, B., & Nguyen-Thien, N. (2010). Emotion on the road—Necessity, acceptance, and feasibility of affective computing in the car. *Advances in Human-Computer Interaction*, *2010*, Article 263593. https://doi.org/10.1155/2010/263593

Fairbanks, G., & Pronovost, W. (1938). Vocal pitch during simulated emotion. *Science*, *88*(2286), 382–383. https://doi.org/10.1126/science.88.2286.382

Fellenz, W. A., Taylor, J. G., Cowie, R., Douglas-Cowie, E., Piat, F., Kollias, S., Orovas, C., & Apolloni, B. (2000, July 24–27). *On emotion recognition of faces and of speech using neural networks, fuzzy logic and the ASSESS system* [Paper presentation]. International Joint Conference on Neural Networks, *Como, Italy*.

Fischer, A. H., & Manstead, A. S. R. (2008). Social functions of emotions. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (pp. 456–468). Guilford Press.

Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nature Human Behaviour*, *3*(4), 393–405. https://doi.org/10.1038/s41562-019-0548-z

Gangamohan, P., Kadiri, S. R., & Yegnanarayana, B. (2016). Analysis of emotional speech—A review. In A. Esposito & L. C. Jain (Eds.), *Toward robotic socially believable behaving systems* (Vol. 1, pp. 205–238). Springer.

Geers, A. E., Davidson, L. S., Uchanski, R. M., & Nicholas, J. G. (2013). Interdependence of linguistic and indexical speech perception skills in school-age children with early cochlear implantation. *Ear Hear*, *34*(5), 562–574. https://doi.org/10.1097/AUD.0b013e31828d2bd6

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science*, *27*(4), 211–219. https://doi.org/10.1177/0963721417746794

Gievska, S., Koroveshovski, K., & Tagasovska, N. (2015, September 21–24). *Bimodal feature-based fusion for real-time emotion recognition in a mobile context* [Paper presentation]. 2015 International Conference on Affective Computing and Intelligent Interaction, Xi'An, China.

Gobl, C., & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*, 189–212. https://doi.org/10.1016/S0167-6393(02)00082-1

Goudbeek, M., Goldman, J. P., & Scherer, K. R. (2009). Emotion dimensions and formant position. In *10th Annual Conference of the International Speech Communication Association (INTERSPEECH-2009)* (pp. 1575–1578). International Speech Communication Association.

Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, *128*(3), 1322–1336. https://doi.org/10.1121/1.3466853

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., & Aucouturier, J.-J. (2021). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature Communication*, *12*, Article 861. https://doi.org/10.1038/s41467-020-20649-4

Grichkovtsova, I., Morel, M., & Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, *54*(3), 414–429. https://doi.org/10.1016/j.specom.2011.10.005

Griol, D., Molina, J. M., & Callejas, Z. (2014). Modeling the user state for context-aware spoken interaction in ambient assisted living. *Applied Intelligence*, *40*(4), 749–771. https://doi.org/10.1007/s10489-013-0503-z

Gudmalwar, A. P., Rama Rao, C. V., & Dutta, A. (2018). Improving the performance of the speaker emotion recognition based on low dimension prosody features vector. *International Journal of Speech Technology*, *22*(3), 521–531. https://doi.org/10.1007/s10772-018-09576-4

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/7880d7226e872b776d8b9f23975e2a3d-Paper.pdf

Heilman, K. M., Leon, S. A., & Rosenbek, J. C. (2004). Affective aprosodia from a medial frontal stroke. *Brain and Language*, *89*(3), 411–416. https://doi.org/10.1016/j.bandl.2004.01.006

Hirst, D. J. (2005). Form and function in the representation of speech prosody. *Speech Communication*, *46*(3–4), 334–347. https://doi.org/10.1016/j.specom.2005.02.020

Hoicka, E., & Gattis, M. (2012). Acoustic differences between humorous and sincere communicative intentions. *British Journal of Developmental Psychology*, *30*(4), 531–349. https://doi.org/10.1111/j.2044-835X.2011.02062.x

Holz, N., Larrouy-Maestri, P., & Poeppel, D. (2022). The Variably Intense Vocalizations of Affect and Emotion (VIVAE) corpus prompts new perspective on nonspeech perception. *Emotion*, *22*(1), 213–225. https://doi.org/10.1037/emo0001048

Horley, K., Reid, A., & Burnham, D. (2010). Emotional prosody perception and production in dementia of the Alzheimer's type. *Journal of Speech, Language, and Hearing Research*, *53*, 1132–1146. https://doi.org/10.1044/1092-4388(2010/09-0030)

Hozjan, V., & Kačič, Z. (2006). A rule-based emotion-dependent feature extraction method for emotion analysis from speech. *Journal of the Acoustical Society of America*, *119*(5), 3109–3120. https://doi.org/10.1121/1.2188647

Huang, C., Jin, Y., Zhao, Y., Yu, Y., & Zhao, L. (2009). Recognition of practical emotion from elicited speech. In *2009 First International Conference on Information Science and Engineering* (pp. 639–642). Institute of Electrical and Electronics Engineers.

Hyun, K. H., Kim, E. H., & Kwak, Y. K. (2010). Emotional feature extraction method based on the concentration of phoneme influence for human–robot interaction.

Advanced Robotics, *24*(1–2), 47–67. https://doi.org/10.1163/016918609x12585530487822

Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimedia Tools and Applications*, *80*(16), 23745–23812. https://doi.org/10.1007/s11042-020-09874-7

Jiang, X., Paulmann, S., Robin, J., & Pell, M. D. (2015). More than accuracy: Nonverbal dialects modulate the time course of vocal emotion recognition across cultures. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 597–612. https://doi.org/10.1037/xhp0000043

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. In M. Lewis & J. Haviland (Eds.), *The handbook of emotion*. Guilford Press.

Jürgens, R., Drolet, M., Pirow, R., Scheiner, E., & Fischer, J. (2013). Encoding conditions affect recognition of vocally expressed emotions across cultures. *Frontiers in Psychology*, *4*, Article 111. https://doi.org/10.3389/fpsyg.2013.00111

Jürgens, R., Hammerschmidt, K., & Fischer, J. (2011). Authentic and play-acted vocal emotion expressions reveal acoustic differences. *Frontiers in Psychology*, *2*, Article 180. https://doi.org/10.3389/fpsyg.2011.00180

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, *1*(4), 381–412. https://doi.org/10.1037/1528-3542.1.4.381

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. https://doi.org/10.1037/0033-2909.129.5.770

Juslin, P. N., Laukka, P., & Bänziger, T. (2018). The mirror of our soul? Comparisons of spontaneous and posed vocal expression of emotion. *Journal of Nonverbal Behavior*, *42*(1), 1–40. https://doi.org/10.1007/s10919-017-0268-x

Kabuta, Y., Taniguchi, T., Hatoko, M., Matsui, H., Fukumoto, T., & Takeda, S. (2013). Comparison of spectral-tilt features of emotional speech depending on the degree of emotions. *International Journal of Affective Engineering*, *12*(2), 161–167. https://doi.org/10.5057/ijae.12.161

Kadiri, S. R., Gangamohan, P., Gangashetty, S. V., Alku, P., & Yegnanarayana, B. (2020). Excitation features of speech for emotion recognition using neutral speech as reference. *Circuits, Systems, and Signal Processing*, *39*(9), 4459–4481. https://doi.org/10.1007/s00034-020-01377-y

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review*, *27*(2), 237–265. https://doi.org/10.3758/s13423-019-01701-x

Kamiloğlu, R. G., & Sauter, D. A. (2021). Voice production and perception. In O. Braddick (Ed.), *Oxford research encyclopedia of psychology*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190236557.013.766

Kan, Y., Mimura, M., Kamijima, K., & Kawamura, M. (2004). Recognition of emotion from moving facial and prosodic stimuli in depressed patients. *Journal of Neurology, Neurosurgery & Psychiatry*, *75*(12), 1667–1671. https://doi.org/10.1136/jnnp.2004.036079

Kantrowitz, J. T., Hoptman, M. J., Leitman, D. I., Moreno-Ortega, M., Lehrfeld, J. M., Dias, E., Sehatpour, P., & Javitt, D. C. (2015). Neural substrates of auditory emotion recognition deficits in Schizophrenia. *Journal of Neuroscience*, *35*(44), 14909–14921. https://doi.org/10.1523/JNEUROSCI.4603-14.2015

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLOS Biology*, *16*(3), Article e2004473. https://doi.org/10.1371/journal.pbio.2004473

Keltner, D. (2019). Toward a consensual taxonomy of emotions. *Cognition and Emotion*, *33*(1), 14–19. https://doi.org/10.1080/02699931.2019.1574397

Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of Nonverbal Behavior*, *43*(2), 133–160. https://doi.org/10.1007/s10919-019-00293-3

Klabbers, E., Mishra, T., & van Santen, J. (2007, August 22–24). *Analysis of affective speech recordings using the superpositional intonation model* [Paper presentation]. 6th ISCA Workshop on Speech Synthesis, *Bonn, Germany*.

Knoll, M., Uther, M., MacLeod, N., O'Neill, M., & Walsh, S. (2006, May 2–5). *Emotional, linguistic or just cute? The function of pitch contours in infant- and foreigner-directed speech* [Paper presentation]. 3rd International Conference on Speech Prosody, *Dresden, Germany*.

Knoll, M. A., & Costall, A. (2015). Characterising $F_{(0)}$ contour shape in infant- and foreigner-directed speech. *Speech Communication*, *66*, 231–243. https://doi.org/10.1016/j.specom.2014.10.007

Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., & Okubo, Y. (2013). Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners. *Frontiers in Psychology*, *4*, Article 105. https://doi.org/10.3389/fpsyg.2013.00105

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*(2), 99–117. https://doi.org/10.1007/s10772-011-9125-1

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, USA*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1412469111

Larrouy-Maestri, P. (2018). "I know it when I hear it:" On listeners' perception of mistuning. *Music and Science*, *1*. https://doi.org/10.1177/2059204318784582

Larrouy-Maestri, P., Harrison, P. M. C., & Müllensiefen, D. (2019). The mistuning perception test: A new measurement instrument. *Behavior Research Methods*, *51*(2), 663–675. https://doi.org/10.3758/s13428-019-01225-1

Larrouy-Maestri, P., Kegel, V., Schlotz, W., van Rijn, P., Menninghaus, W., & Poeppel, D. (2023a). Ironic twists of sentence meaning can be signaled by forward move of prosodic stress. *Journal of Experimental Psychology: General*, *152*(9), 2438–2462. https://doi.org/10.1037/xge0001377

Larrouy-Maestri, P., & Morsomme, D. (2014). The effects of stress on singing voice accuracy. *Journal of Voice*, *28*(1), 52–58. https://doi.org/10.1016/j.jvoice.2013.07.008

Larrouy-Maestri, P., Poeppel, D., & Pfordresher, P. Q. (2023b). Pitch units in music and speech prosody. In R. Wiese & M. Scharinger (Eds.), *How language speaks to music: Prosody from a cross-domain perspective* (pp. 17–42). DeGruyter. https://doi.org/10.1515/9783110770186-002

Laukka, P. (2005). Categorical perception of vocal emotion expressions. *Emotion*, *5*(3), 277–295. https://doi.org/10.1037/1528-3542.5.3.277

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*, Article 353. https://doi.org/10.3389/fpsyg.2013.00353

Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, *111*(5), 686–705. https://doi.org/10.1037/pspi0000066

Laukka, P., Juslin, P. N., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, *19*(5), 633–653. https://doi.org/10.1080/02699930441000445

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Laugh like you mean it: Authenticity modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behavior*, *40*(2), 133–149. https://doi.org/10.1007/s10919-015-0222-8

Lee, J., & Tashev, I. (2015, September 6–10). *High-level feature representation using recurrent neural network for speech emotion recognition* [Paper presentation]. *INTERSPEECH* 2015, Dresden, Germany.

Lima, C. F., Alves, T., Scott, S. K., & Castro, S. L. (2014). In the ear of the beholder: How age shapes emotion processing in nonverbal vocalizations. *Emotion*, *14*(1), 145–160. https://doi.org/10.1037/a0034287

Lima, C. F., Castro, S. L., & Scott, S. K. (2013). When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing. *Behavior Research Methods*, *45*(4), 1234–1245. https://doi.org/10.3758/s13428-013-0324-3

Liu, P., Rigoulot, S., & Pell, M. D. (2015). Cultural differences in on-line sensitivity to emotional voices: Comparing East and West. *Frontiers in Human Neuroscience*, *9*, Article 311. https://doi.org/10.3389/fnhum.2015.00311

Luengo, I., Navas, E., & Hernaez, I. (2010). Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, *12*(6), 490–501. https://doi.org/10.1109/tmm.2010.2051872

MacLeod, N. (1999). Generalizing and extending the eigenshape method of shape space visualization and analysis. *Paleobiology*, *25*(1), 107–138.

Madureira, S. (2016). Intonation and variation: The multiciplicity of forms and senses. *Dialectologia*, *6*, 57–74. https://doi.org/10.1344/Dialectologia2016.2016.4

Marcus, G. E., Neuman, W. R., & MacKuen, M. (2000). *Affective intelligence and political judgement*. University of Chicago Press.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., & Stroeve, S. (2000). Approaching automatic recognition of emotion from voice: A rough benchmark. In *Speech and Emotion, ISCA Tutorial and Research Workshop* (pp. 207–212). International Speech Communication Association.

Mouawad, P., & Dubnov, S. (2017). On modeling affect in audio with non-linear symbolic dynamics. *Advances in Science, Technology and Engineering Systems Journal*, *2*(3), 1727–1740. https://doi.org/10.25046/aj0203212

Mozziconacci, S. J. L., & Hermes, D. J. (1999, August 1–7). *Role of intonation patterns in conveying emotion in speech* [Paper presentation]. 14th International Congress of Phonetic Sciences, San Francisco, CA, United States.

Murray, I. R., & Arnott, J. L. (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. https://doi.org/10.1121/1.405558

Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: Features and algorithms. *International Journal of Human-Computer Studies*, *59*(1–2), 157–183. https://doi.org/10.1016/s1071-5819(02)00141-6

Pan, W., Flint, J., Shenhav, L., Liu, T., Liu, M., Hu, B., & Zhu, T. (2019). Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders. *PLOS ONE*, *14*(6), Article e0218172. https://doi.org/10.1371/journal.pone.0218172

Paulmann, S., Furnes, D., Bokenes, A. M., & Cozzolino, P. J. (2016). How psychological stress affects emotional prosody. *PLOS ONE*, *11*(11), Article e0165022. https://doi.org/10.1371/journal.pone.0165022

Paulmann, S., & Kotz, S. A. (2008). An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context. *Brain and Language*, *105*(1), 59–69. https://doi.org/10.1016/j.bandl.2007.11.005

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). How aging affects the recognition of emotional speech. *Brain and Language*, *104*(3), 262–269. https://doi.org/10.1016/j.bandl.2007.03.002

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition and Emotion*, *28*(2), 230–244. https://doi.org/10.1080/02699931.2013.812033

Pell, M. D. (1996). On the receptive prosodic loss in Parkinson's disease. *Cortex*, *32*(4), 693–704. https://doi.org/10.1016/s0010-9452(96)80039-6

Pell, M. D., & Baum, S. R. (1997). The ability to perceive and comprehend intonation in linguistic and affective contexts by brain-damaged adults. *Brain and Language*, *57*(1), 80–99. https://doi.org/10.1006/brln.1997.1638

Pell, M. D., Jaywant, A., Monetta, L., & Kotz, S. A. (2011). Emotional speech processing: Disentangling the effects of prosody and semantic cues. *Cognition and Emotion*, *25*(5), 834–853. https://doi.org/10.1080/02699931.2010.516915

Pell, M. D., & Kotz, S. A. (2011). On the time course of vocal emotion recognition. *PLOS ONE*, *6*(11), Article e27256. https://doi.org/10.1371/journal.pone.0027256

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*(2), 107–120. https://doi.org/10.1007/s10919-008-0065-7

Persky, S., Ferrer, R. A., & Klein, W. M. P. (2016). Nonverbal and paraverbal behavior in (simulated) medical visits related to genomics and weight: A role for emotion and race. *Journal of Behavioral Medicine*, *39*(5), 804–814. https://doi.org/10.1007/s10865-016-9747-5

Pinheiro, A. P., Del Re, E., Mezin, J., Nestor, P. G., Rauber, A., McCarley, R. W., Gonçalves, Ó. F., & Niznikiewicz, M. A. (2013). Sensory-based and higher-order operations contribute to abnormal emotional prosody processing in schizophrenia: An electrophysiological investigation. *Psychological Medicine*, *43*(3), 603–618. https://doi.org/10.1017/S003329171200133X

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences, USA*, *115*(15), 3972–3977. https://doi.org/10.1073/pnas.1716090115

Rajković, M., Jovičić, S., Grozdić, D., Zdravković, S., & Subotić, M. (2018). A note on acoustic features in pitch contours for discrimination of happiness and anger. *Acta Acustica United With Acustica*, *104*, 369–372. https://doi.org/10.3813/AAA.919179

Ratna Kanth, N., & Saraswathi, S. (2014). A survey on speech emotion recognition. *Advances in Computer Science and Information Technology*, *1*(3), 135–139.

Rigoulot, S., Wassiliwizky, E., & Pell, M. D. (2013). Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition. *Frontiers in Psychology*, *4*, Article 367. https://doi.org/10.3389/fpsyg.2013.00367

Rilliard, A., d'Alessandro, C., & Evrard, M. (2018). Paradigmatic variation of vowels in expressive speech: Acoustic description and dimensional analysis. *Journal of the Acoustical Society of America*, *143*(1), 109–122. https://doi.org/10.1121/1.5018433

Riviello, M. T., & Esposito, A. (2012). A cross-cultural study on the effectiveness of visual and vocal channels in transmitting dynamic emotional information. *Acta Polytechnica Hungarica*, *9*(1), 157–170.

Robinson, P., & el Kaliouby, R. (2009). Computation of emotions in man and machines. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3441–3447. https://doi.org/10.1098/rstb.2009.0198

Roche, J. M., Peters, B., & Dale, R. (2015). "Your tone says it all": The processing and interpretation of affective language. *Speech Communication*, *66*, 47–64. https://doi.org/10.1016/j.specom.2014.07.004

Rochman, D., & Amir, O. (2013). Examining in-session expressions of emotions with speech/vocal acoustic measures: An introductory guide. *Psychotherapy Research*, *23*(4), 381–393. https://doi.org/10.1080/10503307.2013.784421

Rodero, E. (2011). Intonation and emotion: Influence of pitch levels and contour type on creating emotions. *Journal of Voice*, *25*(1), e25–e34. https://doi.org/10.1016/j.jvoice.2010.02.002

Rosenblau, G., Kliemann, D., Dziobek, I., & Heekeren, H. R. (2017). Emotional prosody processing in autism spectrum disorder. *Social Cognitive and Affective Neuroscience*, *12*(2), 224–239. https://doi.org/10.1093/scan/nsw118

Sauter, D. A. (2013). The role of motivation and cultural dialects in the in-group advantage for emotional vocalizations. *Frontiers in Psychology*, *4*, Article 814. https://doi.org/10.3389/fpsyg.2013.00814

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology*, *63*(11), 2251–2272. https://doi.org/10.1080/17470211003721642

Sauter, D. A., Panattoni, C., & Happe, F. (2013). Children's recognition of emotions from vocal cues. *British Journal of Developmental Psychology*, *31*(1), 97–113. https://doi.org/10.1111/j.2044-835X.2012.02081.x

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. https://doi.org/10.1007/s11031-007-9065-x

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*(2), 143–165. https://doi.org/10.1037/0033-2909.99.2.143

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*, 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Scherer, K. R. (2013a). Emotion in action, interaction, music, and speech. In A. Arbib (Ed.), *Language, music, and the brain: A mysterious relationship* (pp. 107–139). MIT Press. https://doi.org/10.7551/mitpress/9780262018104.003.0005

Scherer, K. R. (2013b). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech and Language*, *27*(1), 40–58. https://doi.org/10.1016/j.csl.2011.11.003

Scherer, K. R. (2019). Acoustic patterning of emotion vocalizations. In S. Frühholz & P. Belin (Eds.), *The Oxford handbook of voice perception* (pp. 61–92). Oxford University Press.

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. https://doi.org/10.1177/0022022101032001009

Schlipf, S., Batra, A., Walter, G., Zeep, C., Wildgruber, D., Fallgatter, A. J., & Ethofer, T. (2013). Judgment of emotional information expressed by prosody and semantics in patients with unipolar depression. *Frontiers in Psychology*, *4*, Article 461. https://doi.org/10.3389/fpsyg.2013.00461

Schuller, B., & Rigoll, G. (2006, September 17–21). *Timing levels in segment-based speech emotion recognition* [Paper presentation]. Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, United States.

Schuller, B., Rigoll, G., & Lang, M. (2004, May 17–21). *Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine–belief network architecture* [Paper presentation]. ICASSP, Montreal, QC, Canada.

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *10th Annual Conference of the International Speech Communication Association (INTERSPEECH-2009)* (pp. 312–315). International Speech Communication Association.

Shami, M. T., & Kamel, M. S. (2005, July 6–8). *Segment-based approach to the recognition of emotions in speech* [Paper presentation]. 2005 IEEE International Conference on Multimedia and Expo, *Amsterdam, the Netherlands*.

Shigeno, S. (2018). The effects of the literal meaning of emotional phrases on the identification of vocal emotions.

Journal of Psycholinguistic Research, *47*(1), 195–213. https://doi.org/10.1007/s10936-017-9526-7

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. B., & Hirschberg, J. (1992). ToBI: A standard for labelling English prosody. In *ICSLP 92 Proceedings: International Conference on Spoken Language Processing* (Vol. 92, No. 2, pp. 867–870). University of Alberta Press.

Sinaceur, M., Kopelman, S., Vasiljevic, D., & Haag, C. (2015). Weep and get more: When and why sadness expression is effective in negotiations. *Journal of Applied Psychology*, *100*(6), 1847–1871. https://doi.org/10.1037/a0038783

Singh, P., Saha, G., & Sahidullah, M. (2021, January 27–29). *Non-linear frequency warping using constant-Q transformation for speech emotion recognition* [Paper presentation]. 2021 International Conference on Computer Communication and Informatics, Coimbatore, India.

Snow, D., & Balog, H. L. (2002). Do children produce the melody before the words? A review of developmental intonation research. *Lingua*, *112*(12), 1025–1058.

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology*, *21*(1), 93–120. https://doi.org/10.1007/s10772-018-9491-z

Tamarit, L., Goudbeek, M., & Scherer, K. (2008, June 4–6). *Spectral slope measurements in emotionally expressive speech* [Paper presentation]. ISCA Tutorials and Research Workshops, Speech Analysis and Processing for Knowledge Discovery, Aalborg, Denmark.

Tisljár-Szabó, E., & Pléh, C. (2014). Ascribing emotions depending on pause length in native and foreign language speech. *Speech Communication*, *56*, 35–48. https://doi.org/10.1016/j.specom.2013.07.009

Van Lancker, D., & Sidtis, J. J. (1992). The identification of affective-prosodic stimuli by left- and right-hemisphere-damaged subjects: All errors are not created equal. *Journal of Speech & Hearing Research*, *35*(5), 963—970. https://doi.org/10.1044/jshr.3505.963

van Mersbergen, M., & Lanza, E. (2019). Modulation of relative fundamental frequency during transient emotional states. *Journal of Voice*, *33*(6), 894–899. https://doi.org/10.1016/j.jvoice.2018.07.020

van Rijn, P., & Larrouy-Maestri, P. (2023). Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour*, *7*, 386–396. https://doi.org/10.1038/s41562-022-01505-5

van Rijn, P., Poeppel, D., & Larrouy-Maestri, P. (2023). *Contribution of pitch measures over time to emotion classification accuracy*. PsyArXiv. https://doi.org/10.31234/osf.io/pnysd

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181. https://doi.org/10.1016/j.specom.2006.04.003

Waaramaa, T., Laukkanen, A. M., Airas, M., & Alku, P. (2010). Perception of emotional valences and activity levels from vowel segments of continuous speech. *Journal of Voice*, *24*(1), 30–38. https://doi.org/10.1016/j.jvoice.2008.04.004

Waaramaa, T., & Leisiö, T. (2013). Perception of emotionally loaded vocal expressions and its connection to responses to music. A cross-cultural investigation: Estonia, Finland, Sweden, Russia, and the USA. *Frontiers in Psychology*, *4*, Article 344. https://doi.org/10.3389/fpsyg.2013.00344

Wang, W.-C., Chien, C. S., & Moutinho, L. (2015). Do you really feel happy? Some implications of voice emotion response in Mandarin Chinese. *Marketing Letters*, *26*(3), 391–409. https://doi.org/10.1007/s11002-015-9357-y

Wermke, K., Robb, M. P., & Schluter, P. J. (2021). Melody complexity of infants' cry and non-cry vocalisations increases across the first six months. *Scientific Reports*, *11*(1), Article 4137. https://doi.org/10.1038/s41598-021-83564-8

Whiteside, S. P. (1999). Acoustic characteristics of vocal emotions simulated by actors. *Perceptual and Motor Skills*, *89*(3), 1195–1208. https://doi.org/10.2466/pms.1999.89.3f.1195

Wubben, M. J. J., de Cremer, D., & van Dijk, E. (2011). The communication of anger and disappointment helps to establish cooperation through indirect reciprocity. *Journal of Economic Psychology*, *32*(3), 489–501. https://doi.org/10.1016/j.joep.2011.03.016

Yang, B., & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal Processing*, *90*(5), 1415–1423. https://doi.org/10.1016/j.sigpro.2009.09.009

Yoshimatsu, Y., Umino, A., & Dammeyer, J. (2016). Characteristics of the understanding and expression of emotional prosody among children with autism spectrum disorder. *Autism-Open Access*, *6*(4), Article 185. https://doi.org/10.4172/2165-7890.1000185

Zhang, D., Liu, Y., Hou, X., Sun, G., Cheng, Y., & Luo, Y. (2014). Discrimination of fearful and angry emotional voices in sleeping human neonates: A study of the mismatch brain responses. *Frontiers in Behavioral Neuroscience*, *8*, Article 422. https://doi.org/10.3389/fnbeh.2014.00422

Zhang, X., Sun, Y., & Duan, S. (2015). Progress in speech emotion recognition. In *TENCON 2015-2015 IEEE Region 10 Conference* (pp. 1–6). Institute of Electrical and Electronics Engineers.

Zhang, X., Wang, H., Li, L., Zhao, M., & Li, Q. (2015). Negative emotion recognition in spoken dialogs. In Maosong Sun, Z. Liu, M. Zhang, & Y. Liu (Eds.), *Chinese computational linguistics and natural language processing based on naturally annotated big data. Lecture notes in computer science* (Vol. 9427, pp. 103–115). Springer.

Zhou, G., Hansen, J. H. L., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, *9*(3), 201–216. https://doi.org/10.1109/89.905995

Zhu, Z., Miyauchi, R., Araki, Y., & Unoki, M. (2018). Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech. *Acoustical Science and Technology*, *39*(6), 379–386. https://doi.org/10.1250/ast.39.379

Zloteanu, M., & Krumhuber, E. G. (2021). Expression authenticity: The role of genuine and deliberate displays in emotion perception. *Frontiers in Psychology*, *11*, Article 611248. https://doi.org/10.3389/fpsyg.2020.611248