Phonetics of Emotion—A Window into the Meanings of Emotions

Yi Xu, University College London

Table of contents

Summary

The phonetics of emotion is about the acoustic-phonetic properties of the emotional facets of human vocalization. Conventionally, these properties are studied as correlates of a person's internal states arising from reactions to the environment, where the internal states are defined by influential psychological theories of emotion. A more recent perspective, however, views emotion as an evolved mechanism for motivating actions to *proactively* interact with other individuals, including, in particular, the production of emotional expressions. From this perspective, the acoustic properties of emotional vocalization are devised to actively influence the listeners in ways that may benefit the vocalizer. Interestingly, the meanings of these acoustic properties could be interpreted with knowledge of speech acoustics accumulated over the years. A key encoding mechanism is body-size projection, whereby vocal properties associated with emotions like anger make the vocalizer sound large to dominate the listener, while properties associated with emotions like joy make the vocalizer sound small to appease the listener. Body-size projection is encoded through three acoustic dimensions—pitch, voice quality and formant dispersion. Furthermore, body-size projection is likely accompanied by additional iconic encoding mechanisms also aimed at influencing the listener in specific ways. The acoustic properties associated with these mechanisms are not yet fully clear. Further exploration of the body-size projection principle and identification of additional mechanisms may drive much of the research activity in the coming decades.

Keywords
Morton-Ohala hypothesis, Pitch, Voice quality, Formant dispersion, Selection pressure, Body-size projection, Bio-informational dimensions, Discrete theories, Dimensional theories, Bodily/neural state theories

1. Introduction

As human beings, we have all experienced emotions of various kinds, and they feel as if they have arisen from within us in reaction to different situations. Such internal feelings then drive us, often without our conscious awareness (Ekman, 1992; Scherer, 2003), to act in various ways, including speaking with an emotional tone of voice. Theories of emotion have therefore been predominantly concerned with how to characterize and differentiate emotions according to these internal sensations, how to understand the processes that generate these sensations, and how to develop a taxonomy of emotion based on these sensations. Given the conventional theories of emotion, the task of studying the phonetics of emotional speech is therefore one of discovering the acoustic correlates of the feeling-based categories or dimensions of emotion. This general approach, however, has not led to the identification of distinctive phonetic cues for some of the most important emotional contrasts, such as happiness versus anger (Scherer 2003; Williams & Stevens 1972; Murray & Arnott, 1993; Mauss & Robinson, 2009). In the face of this difficulty, an alternative approach to the phonetics of emotion has been gradually gaining ground. The new approach is based on research originating from a hypothesis about animal calls (Morton, 1977), and its extension based on knowledge of speech acoustics (Ohala, 1984). This alternative approach has suggested functional connections between the acoustic properties of emotional vocalizations and the emotional categories and dimensions that conventional approaches have been unable to identify. More interestingly, the connection between acoustic properties and emotional categories or dimensions seems to also help unveil the nature of emotions from a

functional perspective that differs from conventional emotion theories. The following discussion will therefore start with an overview of the major theories of emotion.

2. Conventional Theories of Emotion

> *In 1884, William James, the American psychologist, famously posed the question: what is an emotion? After more than a century of scientific inquiry, however, emotions remain essentially contested concepts: scientists disagree on how they should be defined, on where to draw the boundaries for what counts as an emotion and what does not, on whether conscious experiences are central or epiphenomenal, and so on. Such disputes have sown great discord among scientists, leaving the field in perpetual upheaval, and without a unified framework for guiding scientific inquiry and accumulating knowledge.*
>
> *Adolphs, Mlodinow & Barrett (2019:R1)*

Decades of research has seen many theories of emotion, yet some of the most fundamental issues, including what is emotion, how different emotions should be defined and classified, whether conscious experiences of emotion are central or epiphenomenal, etc., remain unresolved, as recognized by Adolphs et al. (2019) quoted above as well as the most recent review by Scherer (2022). Interestingly, Adolphs' list does not include some even more basic questions, e.g., why do we have emotions in the first place? What general function do they serve? What is the specific function of each of the specific emotions? And why are emotions so often overtly expressed? The following discussion will start with a brief overview of some of the major theories of emotion by discussing, for each of them, how these key questions are addressed.

2.1 Discrete (basic) Emotions Theories

One of the most straightforward ways to refer to various emotions, even to this day, is to use common words of everyday language, such as happiness (joy), anger, fear, sadness, disgust, surprise, and so on. This is what Darwin did in his book over 150 years ago (Darwin, 1872), which established the study of emotion and emotional expressions as a scientific field. Even in the early 21 century, including in the present article, it is hard to avoid the use of these terms, because they provide convenient references to the emotions. Some of these terms, however, have been given augmented theoretical significance by being used to classify emotions. In particular, it is widely recognized that there are five or six basic emotions, known as the big 5 or big 6: happiness, anger, fear, sadness, disgust (and surprise). Models and frameworks that directly use these names to classify emotions are known as discrete or basic emotions theories. Ekman (1999), for example, has proposed the notion of basic emotions based on a variety of common terms. There are also continued efforts to add more discrete emotions. Ekman and Cordaro (2011) have proposed 10 additional pleasant emotions. Tracy (2004) has proposed pride as an independent emotion.

Given that everything about an organism is the product of evolution (Darwin, 1859), emotions, especially given its likely universality (Darwin, 1872; Ekman, 1973; Ekman et al., 1987), must have also been evolutionally adapted (Darwin, 1872), and both emotional feelings and their associated expressions have evolved under selection pressures. While the evolutionary origin of

emotion is widely assumed (Bryant, 2021; Ekman, 1992), it remains unestablished what exactly the selection pressures are that have driven the development of specific emotions as well as their expressions. This is not clearly specified by Darwin (1872), although a likely pressure is the need to motivate quick responses to situations arising in the environment. As for predicting auditory expression of emotions, Darwin (1872) had some extensive discussion of the associative service, with the most prototypical example of vomiting as the origin of emotions like disgust, distain, contempt, etc. Also Darwin's principle of antithesis states that opposite emotions, for example, anger and joy, would show opposing bodily gestures in every aspect. In general, however, discrete theories have offered little in the way of predicting phonetic properties of specific emotions (Ekman, 2009).

2.2 Dimensional (Affective State) Theories

To make up for the lack of explanatory power of discrete theories of emotion, dimensional theories have been proposed in an effort to identify common aspects shared by different emotions. The proposed commonalities are in terms of affective state or internal feelings associated with the emotional experience (Borod, 1993; Mauss & Robinson, 2009; Schlosberg, 1954; Zei Pollermann, 2002). The commonly assumed dimensions are all reflective of conscious human intuitions. The *valence* dimension refers to whether the emotion feels pleasant or unpleasant, agreeable or disagreeable. The activation or arousal dimension describes the level of activation of the emotional experience, or whether it is active or passive. And the power dimension describes power, control or attention/rejection of the emotional experience. The approach-withdrawal or approach-avoidance dimension, featured occasionally in some theoretical discussions (Borod, 1993; Zei Pollermann, 2002), is again about passive reactions driven by internal feelings. According to these theories, therefore, it is the dimensions that describe the internal feelings that define the different kinds of emotions. In this way, there is no explanation for the function of emotions and why and how emotions are overtly expressed. In particular, if we were to assume the that the dimensions were a kind of reward mechanism, it would be hard to explain why there is only one clearly positive emotion among the big five—happiness, but four negative ones—anger, sadness, fear, and disgust.

As for predicting auditory expression of emotions, the activation dimension can lead to some obvious predictions. For example, at least for some emotions, the more a person is activated, the greater the intensity, loudness, pitch range, etc., may be (Mauss & Robinson, 2009; Goudbeek & Scherer, 2010). From the valence or approach/withdrawal dimension, however, no clear predictions can be derived, because there is no theoretical basis for establishing distinct links between these dimensions and specific acoustic properties of speech.

2.3 Bodily/neural State (James-Lange)

That consciously experienced subjective feelings are the essence of emotion has been questioned by the bodily state theories (James, 1884; Lange, 1885), also known as the peripheralist view (Paul et al., 2020). Those theories arise from the observation that subjective emotional sensations are often felt *after* the bodily actions have already taken place. James (1884), for example, has famously argued that we feel sad because we cry, and we feel afraid because we tremble. The bodily state view has been opposed to the centralist view (Cannon, 1929) that assumes that it is the central neural responses that lead to the mental states associated with specific emotions. But

as has also been recognized, both the peripheral and centralist views assume that emotions are about mental state or experience (Cabanac, 2002), or felt state or experience (Paul et al., 2020). In other words, they both assume that subjective feelings are what define emotions.

Like discrete theories and the dimensional theories, neither the centralist nor the peripheralist view has led to clear predictions about the acoustic correlates of emotional vocalizations. This is again due to the lack of theoretical basis for constructing a causal relation between subjective internal feelings and their vocal expressions, because the two opposing views differ only in terms of which comes first, bodily sensations or central nervous responses.

2.4 Other Theories

In addition to the three major frameworks mentioned above, there are also many other theories, including the appraisal theories (Arnold, 1960; Lazarus, 2006; Scherer, 2022), the componential theories (Scherer, 1984; Paul et al., 2020), the constructionist view (Barrett 2017), the survival theories (LeDoux, 2012), and the functional view (Keltner & Gross, 1999). The last one is worth particular mentioning, because it tries to understand the meanings of each of the specific emotions, which is also what the present article is trying to do. An apparent difficulty of this view for predicting the phonetics of emotion, however, is that the proposed functional definitions are in the language of the modern-day human world, at a very high-level. For example, the function of anger is to redress injustice (Haid, 2003) or demand respect (Parkinson, 1996); the function of fear is to avoid danger (Izard, 1993) or ask for help (Parkinson, 1996); and the function of sadness is to disengage from an unattainable goal (Van Dijk & Zeelenberg, 2002) or ask for comfort (Parkinson, 1996). From these definitions, it is hard to clearly link the proposed functions to the phonetic properties of the emotional expressions. As will be seen in Section 3, better predictions may be made based on underlying mechanisms shared with non-human animals, as Darwin (1872) originally suggested.

3. Conventional Analysis of Phonetic Cues of Emotional Speech

Modern instrumental studies of the acoustic properties of emotional speech can go back as early as Fairbanks and Pronovost (1939). The discovery of distinctive cues of emotional speech has been difficult, however, as can be seen in Table 1 which shows the main acoustic measurements of four major emotions reported by three of the most cited papers (Williams & Steven, 1972; Murray & Arnott, 1993; Scherer, 2003). From Table 1 there is a clear lack of consensus on the contrastive cues of even the most common emotions. Most worryingly, anger and happiness, two of the most frequently occurring emotions (Morrison, Wang & De Silva, 2007), both have raised pitch height according to two of the studies in the table (with no specification from Williams & Stevens, 1972), increased pitch range from two of the studies (again no specification from Williams & Stevens, 1972), conflicting changes for speech rate, though mostly for both anger and happiness. As for voice quality, both emotional voices are breathy according to Murray and Arnold (1993), irregular according to Williams and Stevens (1972). And there is no consensus on articulation precision. Similar lack of distinctive emotion-specific acoustic properties has been reported in other reviews (Ververidis & Kotropoulos, 2006; Murray & Arnott, 1993).

Table 1. Acoustic measurements reported in three highly cited papers. An up arrow or a down arrow indicates an increase or decrease from the neutral emotion, and two arrows indicate very much increased or decreased. Empty cells mean lack of report from that study.

Emotion / Measurement		Anger	Happiness	Fear	Sorrow
Pitch height	Williams & Stevens (1972)	↑↑		↑⇩	⇩
	Murray & Arnott (1993)	↑↑	↑↑	↑↑	⇩
	Scherer (2003)	↑	↑	↑	⇩
Pitch range		↑			⇩
		↑↑	↑↑	↑↑	⇩
		↑	↑	↑	⇩
Voice quality		Irregular		Irregular	Irregular
		Breathy	Breathy	Irregular	Resonant
Intensity		↑			
		↑	↑	Normal	⇩
		↑	↑	↑	⇩
Articulatory precision		↑		↑	
		Tense	Normal	↑	⇩
Speech rate		⇩		⇩	⇩⇩
		↑	↑⇩	↑↑	⇩
		↑	(↑)	↑↑	⇩

The lack of consensus is true of both properties for discrete emotions and those for affective dimensions. Mauss & Robinson (2009), citing multiple sources, conclude that no consistent cues have been found for the valence dimension (pleasant/unpleasant), while the activation/arousal dimension does show some consistent cues. Goudbeek & Scherer (2010) show that although activation/arousal can be well predicted by four acoustic measurements in logistic regression, valence and potency/control both had very low predictability with various acoustic measurements.

4. The Morton-Ohala Hypothesis—An Evolutional Perspective

As mentioned in Section 2.1, human emotions must have been evolutionally adapted under various selection pressures. What needs to be established is only what the specific selection pressures are. For this, it is helpful to start from some very basic considerations. Animals, by definition, make voluntary movements. Voluntary movements, however, need to be motivated. The most fundamental motivation would be the need for survival and procreation so as to guarantee the passing of genes to future generations. So a primary selection pressure would be to develop motivational mechanisms for actions that can serve these needs. There have therefore evolved urges like hunger and thirst to motivate the seeking of food and water, and sexual drives to motivate procreation behaviors. Likewise, survival and procreation both would involve interactions with other animal individuals, and the nature of the interaction would vary depending on the specificities of each situation. Thus there must have evolved drives that

motivate different types of inter-organism interactions. Would emotions, then, be the motivational mechanisms for such actions? There is indeed some evidence for it. For example, in anger, blood rushes to the arms and hands to prepare a person to strike, whereas in fear, blood rushes to the legs and feet to prepare for flight (Ekman, 1992; Ekman & Cordaro, 2011; Levenson, Ekman & Friesen, 1990). It is likely, therefore, that emotion is an evolutionally adapted mechanism for motivating actions to interact with other individuals (Ekman, 1992; Scarantino, 2014). For animals that do not rely on high-level cognitive controls, such motivational mechanism could be the primary drive for their interaction with other animals. For humans, emotional drives may reach consciousness in the form of affective sensations, and these sensations may enable high-level cognitive control of the emotion-triggered actions.

For the phonetics of emotion, it is important to realize that the actions motivated by the emotional drives are not limited to fight and flight, but can also include the production of emotional expressions, both facial and vocal. The specific forms of the emotional expressions should also have been under various selection pressures. Many of these pressures would come from environmental factors, including, in particular, physical laws. A physics-based selection pressures that may have shaped animal calls was proposed by Morton (1977). Based on the calls of many avian and mammalian species, Morton summarized that "birds and mammals use harsh, relatively low-frequency sounds when hostile and higher frequency, more pure tone like sounds when frightened, appeasing, or approaching in a friendly manner" (Morton, 1977: 855). He further theorized that those sound qualities are used to give an impression of, or project, the animal's body size: a low fundamental frequency and harsh voice would project a large body size, while a high pitch and a pure-tone-like voice would project a small body size. Such projects are based on physical laws: the larger the animal, the more powerful it is likely to be, other things being equal. At the same time, the larger the animal, the longer and bulkier its vocal folds or syrinx are likely to be, hence its calls would likely to be low in fundamental frequency and non-modal (i.e., vibrating with subharmonics, Fitch, 2002; Sun, 2002). In fact, this natural correlation between body size and vocal properties is so strong that it would be evolutionarily disadvantageous not to develop adaptations to explore it. But there are two possible adaptation strategies. One is to grow larger, which would both enhance the bodily power and make the vocalization sound large. But growing larger is costly and takes a long evolutionary time. A more efficient strategy would be to mimic the visual and/or acoustic effects of a large body. Mimicry of natural environment is commonplace in nature, as seen in many cases of camouflage (Stevens & Merilaita, 2009). Visual mimicry of large body size can be seen in the case of erected hair or feathers during aggression (Morris, 1956), shrinkage of body size during submission (Reddon, Ruberto & Reader, 2021). Vocal mimicry of body size, therefore, would be just an application of the same strategy in the acoustic domain.

Morton's (1977) hypothesis was quickly extended to humans by Ohala (1984), and the extension went beyond just cross-species applications, in a number of ways. First, vocal resonances (formants) were added as a third acoustic dimension that can project body size, in addition to pitch and voice quality. This is based on the acoustic theory of speech production (Fant, 1960; Stevens, 1998), that is, other things being equal, the larger the animal, the longer its vocal tract is likely to be, and the lower all the resulting formants. Therefore, changing the length of the vocal tract could also be a means of projecting different body sizes. Second, based on the vocal tract length hypothesis, Ohala (1984) further theorized that the smile is for the sake of shortening the vocal tract so as to increase the frequency of formants in order to convey friendliness. Third,

sexual dimorphism in the vocal systems of humans and some other animals is also proposed to be driven by body size projection, such that male vocalizations show pitch and formants that are disproportionally lower than those of females relative to the actual differences in their body size, which serves to compete with other males for mating right. In this way, body size projection can be achieved through either permanent physiological changes or transient manipulations. Both strategies are seen in other cases of mimicry, e.g., permeant changes in body colour and pattern, and transient variations like changes performed by chameleon and octopus.

Neither Morton (1977) nor Ohala (1984), however, offered detailed discussion of how body size projection is applied in vocal emotional expressions. For non-human animals, Morton did not use the term emotion, presumably because the term has been too closely associated with conscious awareness of internal affective sensations, to which we have little access in the case of animals. In his discussion of "affective" use of pitch, Ohala (1984) mentioned politeness, submission, confidence, assertiveness, authority, aggression, but did not speculate on how those are linked to specific properties of body size projection. For empirical data, he presented a pilot experiment on the judgment of dominance, with no examination of its direct relevance for specific emotions.

4.1 Empirical Evidence

Since its initial proposal (Morton, 1977; Ohala, 1984), much research has been done on the Morton-Ohala hypothesis. Studies on animal calls have shown that caller's body size is correlated to the fundamental frequency (Davies & Halliday, 1978; Clutton-Brock & Albon, 1979; Pfefferle & Fischer, 2006) and vocal tract length (Fitch, 1997; Harris et al., 2006). Lower-pitched calls by male toads are more likely to prevent attacks during mating (Davies & Halliday, 1978). Female animals are attracted to male calls that have lower fundamental frequency (Ryan, 1980; Miyazaki & Waas, 2003) or narrower formant dispersion (average inter-formant distance) (Reby et al., 2010).

For humans, much has been done on static differences in the acoustic properties of speech on vocal preferences (Feinberg et al., 2005, 2006) and social dominance (Anikin, 2020; Anikin et al., 2021; Borkowska & Pawlowski, 2011; Puts, Gaulin & Verdolin, 2006; Wolff & Puts, 2010). Also, much is done on its implication for intonational structures than for emotional expressions (Bänziger & Scherer, 2005; Gussenhoven, 2002). In contrast, however, there has been little effort on testing the Morton-Ohala hypothesis on emotional expressions until recently. Scherer & Bänziger (2004) tested Ohala's hypothesis about the body size projection reflected by intonational contours, but found little evidence for emotion-specific intonation contours.

A major difficulty in studying the acoustic correlates of emotional expression is that, unlike phonological contrasts, emotional contrasts are not easy to elicit from speakers, even if they are trained actors (Anikin & Lima, 2018; Batliner et al., 2000; Campbell, 2000; Wilting, Krahmer & Swerts, 2006). The reason is that genuine emotional voice is produced only when the emotional conditions are really met, and the speaker is truly emotional. But inducing genuine emotions such as anger and sadness in the laboratory would be not only hard, but also ethically problematic or sometime even dangerous (in the case of anger, for example). Short of genuine emotions, researchers can only ask participants to act. But acting is hard, and consistency is especially difficult to guarantee. There have been many efforts to address this difficulty, and one of the most effective methods is to use perception of listeners to screen out the those acted

vocalizations that have low emotion recognition rates. This has been increasingly adopted as a necessary measure (Burkhardt et al., 2005; Hammami, 2018). Given the relevance of emotion perception, another method is to directly manipulate various acoustic parameters and test their effects on listeners. This approach is highly recommended by Scherer and Bänziger (2004) as a way to make a significant improvement over traditional exploratory method. But they also emphasize that such work should be based on hypotheses informed by earlier work rather than simply through trial and error.

This has been done in a series of studied by using synthetic manipulations of speech utterances to test the Morton-Ohala hypothesis (Chuenwattanapranithi et al., 2008; Hsu & Xu, 2014; Noble & Xu, 1011; Xu, Kelly & Smillie, 2013; Liu et al., 2021). These studies started from testing the hypothesis that the expression of ager is to project a large body size so as to intimidate the observer, and it is done by lowering pitch and reducing formant dispersion, and happiness is to express appeasement, and would therefore show raised pitch and increased formant dispersion. This was found to be indeed the case whether the synthetic manipulation was done with an articulatory synthesizer (Chuenwattanapranithi et al., 2008; Xu et al., 2013), or resynthesized speech with acoustic modifications (Hsu & Xu, 2014; Noble & Xu, 2011; Xu, Kelly & Smillie, 2013). The findings of these studies have provided the first set of direct evidence in support of the Morton-Ohala hypothesis.

4.2 Need for Additional Dimensions

Body-size projection, however, cannot explain all the acoustic cues of vocal emotional expressions. In particular, it is critical to also account for the major inconsistencies in the reported acoustic patterns associated with various emotions, as shown in Table 1. For example, pitch is frequently reported to be high in both angry and happy speech, yet perception experiments have consistently found that lower pitched utterances are more likely to be heard as angry (Chuenwattanapranithi et al., 2008; Noble & Xu, 2011; Xu et al., 2013). A likely reason is that other factors also contribute to the expressive aspects of speech that are independent of body-size projection. This has led to the proposal of the bio-information dimensions theory.

5. A Bio-informational Dimensions Theory

In the bio-informational dimensions (BID) theory (Xu, Kelly & Smillie, 2013), body-size projection is only one of the dimensions, albeit one of the most important. Similar to the conventional dimensional theories (Borod, 1993; Mauss & Robinson, 2009; Schlosberg, 1954), dimensions in the BID theory are more primitives than basic emotions such as the big six. These dimensions, however, differ from those of the feeling-defined dimensions in that they are based on the assumption that emotional expressions have evolved to *proactively* elicit behaviours that may benefit the vocalizer. The current version of the BID theory posits four dimensions: *size projection*, *dynamicity*, *audibility,* and *association*. The following are brief definitions of the four dimensions from Xu, Kelly and Simillie (2013):

> The *size projection* dimension is to project either a large body size to achieve an effect of repelling or dominating the receiver, so as to express threat or assertiveness, or a small body size to achieve an effect of attracting or appeasing the receiver, so as to express friendliness, subordination or request for sympathy. At least three parameters are likely

involved in this dimension — vocal tract length, as reflected by formant dispersion, pitch and voice quality.

The *dynamicity* dimension controls how vigorous the vocalization sounds, depending on whether it is beneficial for the vocalizer to appear strong or weak. A vigorous vocalization has a large movement range with high velocity, in terms of both pitch and formant movements, whereas a less vigorous vocalization has a narrow movement range with low velocity.

The *audibility* dimension controls how far a vocalization can be transmitted, depending on whether and how much it is beneficial for the vocalizer to be heard over long distance. The control of audibility is mainly through glottal effort, which will affect sound intensity. But there may be significant interactions with voice quality.

The *association* dimension controls associative use of sounds typically accompanying a non-emotional biological function in circumstances beyond the original ones. For example, the disgust vocalization seems to mirror the sounds made when a person orally rejects unpleasant food (Darwin, 1872). Articulating this kind of sounds involves tightening the pharynx, which would result in raised F1 (Stevens, 1998) as well as devoicing.

The following discussion illustrates how BID theory can be applied to interpret and predict the phonetic cues of some of the major emotions. Note that the application of dynamicity, audibility and association is more tentative than that of body-size projection, because they have been less empirically tested, and their independent contributions are far from clear.

5.1 Anger/happiness

These two emotions are discussed together because they are the most frequently occurring among the emotions (Morrison, Wang & De Silva, 2007). However, most of the phonetic cues reported for the two emotions are not highly distinctive from each other, as shown in Table 1. From the BID perspective, what may best separate them is the size projection dimension. That is, angry expressions would resemble aggressive calls by animals (Morton, 1977) that project a large body size. Happy expressions, in contrast, may resemble the submission calls that project a small body size. This has been supported by perceptual studies in which acoustic parameters are manipulated to project an enlarged or reduced body size through pitch (higher for joy) and formant dispersion (narrower for anger) (Chuenwattanapranithi et al., 2008; Hsu & Xu, 2014; Noble & Xu, 2011; Xu et al., 2013; Xu, Kelly & Smillie, 2013; Xu et al., 2013). Nobel and Xu (2011), Xu et al. (2013) and Hsu and Xu (2014) further showed that breathy voice was consistently heard as happy, while pressed voice was consistently heard as angry.

There is also initial evidence for the other dimensions. Xu, Kelly and Smillie (2013) found that an expanded pitch range, hence increased dynamicity, was a major perceptual cue for happy speech. Noble and Xu (2011), interestingly, found that perceived happiness and friendliness both involved higher median pitch and wider pitch range, but the amount of increase needed was greater for joy than for friendliness. The difference seems point to two rather different functions that are likely involved in the joy expression. The first is related to the kind that is for the sake of appeasement or submission (Fernandez-Dols & Ruiz-Belda 1995; Kraut & Johnston 1979). For this kind of joy expression, neither dynamicity nor audibility needs to be high. The second function is the *play instinct* (Pellegrini et al., 2007) which is often associated with laughter. This instinct is shared by many animals (Panksepp, 2005) for its evolutionary benefit of motivating

the practice of critical survival skills through play. Thus play-related joy is likely to show greater dynamicity and greater audibility than smile-related joy, as both would indicate that the vocalizer is enthusiastic and energetic.

The dynamicity and audibility dimensions may also be critical for anger expressions. For hot anger, in particular, both dynamicity and audibility are likely high, because it makes sense for the vocalizer to sound energetic and be easily heard. Audibility, however, may have a tricky interaction with the body-size projection. That is, to make a vocalization louder, pitch is also made higher due to the Lombard effect (Brumm & Zollinger, 2011). This means that the often-reported high pitch in anger (Table 1 and also Hammerschmidt & Jürgens, 2007) is likely for the sake of sounding louder rather for projecting a smaller body size. The Lombard effect therefore introduces a major confound, making it hard to distinguish hot anger from joy when pitch and intensity are treated as the only acoustic cues.

5.2 Sadness

Like joy expression, there are at least two rather different types of sad vocal expressions (Burkhardt & Sendlmeier, 2000; Scherer, 1979), one is relatively quiet and passive, and the other is an active grief often seen in mourning. The first kind may be characterized with very low energy, which would entail low loudness and low pitch. The grieving type of sadness would often be associated with a sobbing voice. It is interesting to note, however, that the sad vocalization elicited under experimental conditions is often the passive type, as can be seen from the reported acoustic measurements (Banse & Scherer, 1996; Scherer, 2003). But this type of sadness, curiously, does not fit the typical sadness logos and emojis, which almost invariably portray a face with dropped lip corners that is apparently associated with weeping or sobbing (Ekman, 1998).

Acoustic features of the grieving type of sadness have indeed been reported. A number of studies with experimental methods that explicitly elicited grieving sadness have found rather different phonetic cues: raised rather than lowered pitch (Costanzo, Markel & Costanzo, 1969; Erickson et al., 2006), or even falsetto voice (Burkhardt & Sendlmeier, 2000). Xu, Kelly and Smillie (2013) found in a perception experiment that, in addition to raised pitch, perceived sad utterances also had reduced formant dispersion, indicating a lengthened vocal tract. This suggests that the function of a weeping voice, which is best typified by children's cries, is to *demand* rather than *plead for* attention, care and sympathy. This function could also be additionally helped by a harsh voice to further enhance the urgency of the demand, although this is yet to be empirically attested. Similar to depressed sadness, the grieving sadness also shows a reduced pitch range and reduced speech rate (Xu, Kelly & Smillie, 2013), but probably for different reasons. For depressed sadness, the reduced pitch range and speech rate are likely due to the vocalizer's lack of will to exert much speaking effort. For grieving sadness, pitch range is reduced but pitch level is raised. This, when combined with a slow speech rate, creates a sustained high-pitched wailing that can drown out most of the competing sounds, thus serving to maximally capture the attention of the listeners. Again, however, these predictions need to be empirically tested.

5.3 Fear

From the BID perspective, fear presents yet another case of complexity in regard to the exact function its expression serves. In terms of facial expressions, fear is featured with wide-open

eyes, mouth and even nostrils, which help to gather as much sensory information as possible in a very short period of time (Susskind et al., 2008). But this facial expression cannot directly lead to predictions about fear vocalization. Morton (1977) places fear calls in clear opposition to hostile calls, positing that its function is to signal the caller's submission or appeasement. If this is indeed the case, and if the facial expression of fear were consistent with vocal expression, it would be similar to the smile, which is apparently not the case. When Xu, Kelly and Smillie (2013) used listener perception to obtain prototypical emotional cues, two attributes were found to be the most relevant for fear: high pitch median and small formant dispersion. The high pitch median is consistent with many early reports (Burkhardt & Sendlmeier, 2000; Cowie et al., 2001; Protopapas & Lieberman, 1997; Mozziconacci, 2001; Murray & Arnott, 1993; Ververidis & Kotropoulos, 2006). The lengthened vocal tract, however, goes against Morton's (1977) submission hypothesis for fear calls. To interpret the contradiction, it may be helpful to note that there can be three very different scenarios of facing a powerful adversary. In one the opponent is a conspecific or even a member of the same herd, whose goal is to achieve dominance in the social ladder. In that case, submission would spare the vocalizer a fight that may result in a grave injury. In another scenario, the adversary is a predator, in which case submission would only mean one thing: to be eaten. To avoid this fate, the vocalizer is likely to put on a fight and the accompanying vocalization would be more likely to signal the willingness to fight by incorporating features similar to hostile calls. In yet a third scenario, the fear-evoked call could be a warning signal for fellow members of the same family or herd. In this case, it also makes sense for the vocalization to project a large body size to sound authoritative rather than tentative. Interestingly, there have been reports that fear vocalization is often perceived as reproach, suppressed anger or indignation (Fónagy, 1978; Mozziconcci, 2001). Further research is needed to identify the likely function behind the vocal expression of fear.

5.4 Surprise, Disgust

These two emotions are not related. They are discussed in the same section because of the lack of extensive research on their phonetic cues.

Though often treated as a separate emotion, surprise has sometimes been questioned for its independence (Ekman et al., 1987). Indeed, surprise shares similarities with both fear and joy. Facially, it is similar to fear as both involve wide eyes and open mouth (Chamberland et al., 2017; Kim et al., 2004; Zhao et al., 2017). It is also similar to fears vocally as both involve high pitch and wide pitch range (Belin et al., 2008; Liu et al., 2021; Murray & Arnott, 1993; Xu, Kelly & Smillie, 2013). Interestingly, in perception both visually (Boucher & Carlson, 1980; Ekman, 1973) and auditorily (Van Bezooijen, Otto & Heenan, 1983), fear is confusable with surprise but not vice versa. This then seems to be related to the fact that surprise can be either positive or negative (Vrticka, 2014), and it is the negative surprise that is similar to fear (Neta et al., 2017), while positive surprise is actually similar to joy (Van Bezooijen, Otto & Heenan, 1983). Note that if the features shared with joy and fear are removed from the surprise facial expression, only a startle reaction (Susskind et al., 2008) is left, which is brief and likely mainly shown only in the facial expression. The subsequent vocalization, if any, is to express the resulting fear or joy, which may leave no need for surprise to have its own unique phonetic cues.

For disgust, its facial expression shows features that are claimed to be the exact opposite of fear (Susskind et al., 2008), with the eyes, mouth and even nostrils all narrowed, so as to reduce the

amount of sensory exposure upon detecting something unpleasant. Like surprise, however, these facial features cannot lead to clear predictions of vocal features, suggesting that the facial and vocal expressions of disgust may serve different purposes. The former may be mainly for self-protection, while the latter mainly for warning others. The warning function likely involves the association dimension in the BID theory. That is, disgust may imitate properties of vomiting. To produce a vomit-like sound, the pharyngeal cavity would be tightened, which would raise the first formant (Stevens, 1998) and generating devoicing due to reduced transglottal pressure. Initial support for this association hypothesis can be seen in the finding of *negative intensification* (Niebuhr, 2010), which involves devoicing of both consonants and vowels in words that refer to the unpleasant things in an utterance.

5.5 Separation of Emotional Feelings and Emotional Expressions

An important implication of the BID theory is that, because emotional expressions are proactive and communicational rather than passive and reactional, they may not be fully aligned with consciously felt emotions. For example, happiness as a felt emotion is mainly about the pleasure a person experiences. But happy expressions such as smiling are not well correlated with the level of felt pleasure. For instance, Olympic gold medallists are found to smile frequently only when interacting with other people, although their feeling of happiness was judged to be intense throughout a medal awarding ceremony (Fernandez-Dols & Ruiz-Belda 1995). Also, customers in a supermarket are more likely to smile when being apologetic because they have mistakenly asked for assistance from someone who does not work there than when they express thanks to a person who has actually provided the requested assistance (Kraut & Johnston 1979). Likewise, bowlers are more likely to smile when facing fellow playmates than when they have just scored a spare or a strike (Kraut & Johnston 1979).

Such separation of emotional feelings and expressions is closely related to another important issue that has been barely scratched in the study of emotional expressions, namely, display rules. Proposed by Ekman & Friesen (1969), display rules refer to socially learned conventions for controlling the manifestation of emotions according to social settings and social roles. Under these rules, the overt expression of certain emotions can be subdued or even supressed, depending on the cultural norms. The acquisition of display rules is quite early in life (Ekman & Friesen, 1969; Saarni, 1979), and their application is likely automatic and subliminal. These rules therefore introduce further complications to the study of the phonetic cues of emotional expressions. For example, the inclination for human actors to express sadness with the quiet rather than the sobbing type of vocalization in laboratory recording (Section 5.2) could be related to some kind of social norms in the cultures where the research is conducted.

5.6 Non-emotional Use of Bio-informational Dimensions

The bio-informational dimensions are relevant not only for vocal emotional expressions, but also for many non-emotional functions of speech, including vocal attractiveness, charisma, social attitude, etc.

As mentioned in Section 4, male animals often use their calls not only to dominate other males during the mating season, but also to attract females (Reby et al., 2010). Ohala (1984) proposes that body-size project applies to humans as well, and that it is for the sake of projecting a large body size that human males have evolved longer vocal folds and more descended larynx than

females, which can generate lower pitch and narrower formant dispersion. In other words, such sexual dimorphism is not merely for the distinction between the sexes, but actually for the sake of attracting the opposite sex. This has been confirmed by the finding that female listeners judge men with lower pitch and narrower formant dispersion as more attractive (Collins, 2000; Xu et al., 2013). But body-size projection is used not only by human males, but also by female speakers. Male listeners find female voices with higher pitch and higher formant frequencies as more attractive (Puts et al., 2011).

Most surprisingly, voice quality has been found to be one of the most important properties for vocal attractiveness. In Xu et al. (2013), breathy voice contributed the most to an attractive female voice, much more than pitch and formant dispersion. The interpretation was that breathy voice, due to its increased spectral slope, pushes the speaker's voice toward a pure-tone-like sound (at the fundamental frequency), which would be the most appeasing voice in animal calls (Morton, 1977). Given that human speech needs higher-frequency energy to carry consonants and vowels, making the voice breathy seems to be the second-best strategy for the speaker to sound appeasing. Even more interestingly, breathy voice has also been found to make a male voice attractive to female listeners (Xu et al., 2013). This seems to have resolved the otherwise baffling question how a male voice could sound attractive when its pitch, formant dispersion and voice quality all signal a man in anger. It also further suggests that voice quality, the as yet not well studied dimension despite Morton's (1977) original proposal, may play a much bigger role in expressive speech for conveying many emotional and social meanings.

Finally, bio-information dimensions may also be highly relevant for many speaking style issues. One of them is about the nature of child-directed speech or motherese. Early on, motherese is predominantly considered to facilitate baby's vocal learning by enhancing clarity (Golinkoff et al., 2015). But there is growing evidence that the observed vocal exaggerations are to make the caregiver sound happy to attract the baby's attention (Benders, 2013; Golinkoff et al., 2015; Kalashnikova, Carignan & Burnham, 2017; Singh, Morgan & Best, 2002). But voice quality has not yet been seriously considered in this debate. Another stylistic issue is about charisma of public speakers (Rosenberg & Hirschberg, 2009). A number of acoustic properties have been identified for charismatic speakers (Niebuhr, Voße & Brem, 2016), among which is an exceptionally high pitch. Intriguingly, this high pitch could also be related to the Lombard effect discussed in Section 5.1 for hot anger. That is, the high pitch could again be for the sake of enhancing loudness rather than for projecting a small body size, as charisma should be associated with increased rather than reduced authority.

6. Final Remarks

The overview in this article has demonstrated that the phonetics of emotion may not be merely passive reflections of the internal feelings behind the associated emotions, but are more likely to be communicative signals used to proactively (albeit subconsciously) influence the listener in ways that may benefit the speaker. It is shown that vocal emotional expressions are a communication system that have evolved under various selection pressures, which include not only the need to influence the listener, but also physical laws of acoustic vibration and resonance, and articulation mechanisms. The phonetics of emotion therefore offers a unique window into the meaning of emotions, making the emotional expressions interpretable, and the related hypotheses testable.

It has also been demonstrated that the phonetics of emotions is highly multi-dimensional, especially in light of the bio-information dimensions. The size projection dimension alone, for example, involves three acoustic dimensions, pitch, formant dispersion and voice quality, which work either together or independently to create impression of body size to influence the listener. Also, dynamicity, audibility and association are additional dimensions that help to further influence the listener. The high multidimensionality makes it possible for emotional meanings to be conveyed alongside the rich linguistic meanings carried by phonetic structures: consonant, vowel, stress, tone and intonation (Xu, 2019). This kind of parallel encoding (Xu, 2005) means that the emotional and attitudinal use of the acoustic cues is constantly blended with the linguistic phonetics of speech. The intimate blending could be a source of the frequent (yet by no means regular) cases of sound symbolism (Hinton, Nichols & Ohala, 1995; Svantesson, 2017), i.e., the iconic (as opposed to arbitrary) representation of meanings in speech. Compared to sound symbolism, which requires phonologization or lexicalization of the sound-meaning association, however, the noncategorical phonetics of emotion is richer, more nuanced, more obligatory, and more frequently occurring.

The evolutionary-functional approach reviewed in the present article is relatively new in the research of the phonetics of emotion, but its predictive power can already be seen. The next step in research along this line may gain insight on 1) the critical interaction of body-size projection with dynamicity and audibility for the understanding of anger vocalization, charisma in public speaking and vocal confidence. 2) the importance of voice quality, and 3) the role of display rules.

Finally, research on the phonetics of emotion can be facilitated by a number of tools developed over the years. For perception-oriented studies, the synthetic manipulation of pitch level, pitch range and formant dispersion can be performed with Praat (Boersma, 2001; Xu et al., 2013a). The synthetic manipulation of voice quality (from breathy to tense voice) can be done with VocalTractLab (Birkholz et al., 2015; Xu et al., 2013), although the range of voice qualities covered is still far from enough. For analysis-oriented research, a full set of BID measurements, including many for voice quality, can be obtained with ProsodyPro (Xu, 2013b).

7. References List

Adolphs, R., Mlodinow, L. and Barrett, L. F. (2019). What is an emotion? *Current Biology* **29**(20): R1060-R1064.

Anikin, A. (2020). The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations. *Bioacoustics* **29**(2): 226-247.

Anikin, A. and Lima, C. F. (2018). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Quarterly Journal of Experimental Psychology* **71**(3): 622-641.

Anikin, A., Pisanski, K., Massenet, M. and Reby, D. (2021). Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size. *Proceedings of the Royal Society B: Biological Sciences*.

Bänziger, T. and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication* **46**: 252-267.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* **70**: 614-636. [Google: 152]

Batliner, A., Fischer, K., Huber, R., Spilker, J. and Nöth, E. (2000). Desperately seeking emotions or: Actors, wizards, and human beings. In *Proceedings of ISCA tutorial and research workshop (ITRW) on speech and emotion*

Belin, P., Fillion-Bilodeau, S., and Gosselin, F. (**2008**). "The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing," *Behavior research methods* **40**, 531-539.

Benders, T. (**2013**). "Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent," *Infant Behavior and Development* **36**, 847-862.

Birkholz, P., Martin, L., Willmes, K., Kröger, B. J. and Neuschaefer-Rube, C. (2015). The contribution of phonation type to the perception of vocal emotions in German: an articulatory synthesis study. *The Journal of the Acoustical Society of America* **137**(3): 1503-1512.

Borkowska, B. and Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour* **82**(1): 55-59.

Borod, J. C. (1993). Emotion and the brain -- Anatomy and theory: An introduction to the special section. *Neuropsychology* **7**: 427-432.

Boucher, J. D. and Carlson, G. E. (1980). Recognition of facial expression in three cultures. *Journal of cross-cultural psychology* **11**(3): 263-280.

Brumm, H. and Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* **148**(11-13): 1173-1198.

Burkhardt, F. and Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proceedings of ISCA Workshop on Speech and Emotion: A conceptual framework for research*, Belfast

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. and Weiss, B. (2005). A database of German emotional speech. In *Proceedings of Ninth European Conference on Speech Communication and Technology*

Cabanac, M. (2002). What is emotion? *Behavioural Processes* **60**(2): 69-83.

Campbell, N. (2000). Databases of emotional speech. In *Proceedings of ISCA Workshop on Speech and Emotion: A conceptual framework for research*, Belfast**:** 34-38.

Cannon, W. B. (1929). *Bodily Changes in Pain, Hunger, Fear and Rage*. New York: Appleton.

Chamberland, J., Roy-Charland, A., Perron, M. and Dickinson, J. (2017). Distinction between fear and surprise: an interpretation-independent test of the perceptual-attentional limitation hypothesis. *Social neuroscience* **12**(6): 751-768.

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. and Maneewongvatana, S. (2008). Encoding emotions in speech with the size code -- A perceptual investigation. *Phonetica* 65: 210-230.

Clutton-Brock, T. H. and Albon, S. D. (1979). The Roaring of Red Deer and the Evolution of Honest Advertisement. *Behaviour* **69**: 145-170. [Emphasizing honesty]

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour* **60**: 773–780.

Costanzo, F. S., Markel, N. N. and Costanzo, P. R. (1969). Voice quality profile and perceived emotion. *Journal of Counseling Psychology* **16**(3): 267-270.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE* **18**(1): 32-80.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. London: John Murray.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London, England: John Murray.

Davies, N. B. and Halliday, T. R. (1978). Deep croaks and fighting assessment in toads Bufo bufo. *Nature* **274**(5672): 683-685. [Correlation with true body size]

Ekman, P. (1973). Cross-cultural studies of facial expression. In *Darwin and facial expression: A century of research in review*. P. Ekman (ed.). New York: Academic. pp. 169-222.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion* **6**: 169-200.

Ekman, P. (1999). Basic Emotions. In *The Handbook of Cognition and Emotion*. T. Dalgleish and T. Power. Sussex, U.K: John Wiley & Sons, Ltd. pp. 45-60.

Ekman, P. (2009). Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**(1535): 3449-3451.

Ekman, P. and Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion review* **3**(4): 364-370.

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

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M. and Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology* **53**(4): 712-717.

Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T. and Shibuya, Y. (2006). Exploratory Study of Some Acoustic and Articulatory Characteristics of Sad Speech. *Phonetica* **63**: 1-25.

Fairbanks, G. and Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs* **6**: 87.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Feinberg, D. R., Jones, B. C., Law-Smith, M. J., Moore, F. R., DeBruine, L. M., Cornwell, R. E., Hillier, S. G. and Perrett, D. I. (2006). Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. *Hormones and Behavior* **49**: 215-222.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M. and Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behavior* **69**: 561-568.bork

Fernandez-Dols, J.-M. and Ruiz-Belda, M.-A. (1995). Are Smiles a Sign of Happiness?: Gold Medal Winners at the Olympic Games. *Journal of Personality & Social Psychology* **69**(6): 1113-1119.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of the Acoustical Society of America* **102**: 1213-1222.

Fónagy, I. (1978). A new method of investigating the perception of prosodic features. *Language and Speech* **21**: 34-49.

Golinkoff, R. M., Can, D. D., Soderstrom, M. and Hirsh-Pasek, K. (2015). (Baby) talk to me: the social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science* **24**(5): 339-344.

Goudbeek, M. and 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.

Gussenhoven, C. (2002). Intonation and interpretation: Phonetics and Phonology. In *Proceedings of The 1st International Conference on Speech Prosody*, Aix-en-Provence, France**:** 47-57.

Hammami, A. (2018). *Towards developing a speech emotion database for Tunisian Arabic*, Itä-Suomen yliopisto.

Hammerschmidt, K. and Jürgens, U. (2007). Acoustical Correlates of Affective Prosody. *Journal of Voice* **21**(5): 531-540.

Harris, T. R., Fitch, W. T., Goldstein, L. M. and Fashing, P. J. (2006). Black and White Colobus Monkey (Colobus guereza) Roars as a Source of Both Honest and Exaggerated Information About Body Mass. *Ethology* **112**(9): 911-920.

Hinton, L., Nichols, J. and Ohala, J. J. (1995). *Sound symbolism*. Cambridge: Cambridge University Press.

Hsu, C. and Xu, Y. (2014). Can adolescents with autism perceive emotional prosody? *Interspeech 2014*, Singapore.

Izard, C. E. (1993). Organizational and motivational functions of discrete emotions. In *Handbook of emotions*. M. Lewis and J. M. Haviland: The Guilford Press pp. 631–641.

James, W. (1884). What is emotion? *Mind; A Quarterly Review of Psychology and Philosophy* **9**: 188-205. https://doi.org/10.1093/mind/ os-IX.34.188

Lange, C. G. (1885). The mechanisms of the emotions. In *The Classical Psychologists, Houghton Mifflin*. B. Rand. Boston**.** pp. 672-684. (Classics Editor's note: This translation of a passage from Lange's Om Sindsbevaegelser (1885) from Lange's Ueber Gemüthsbewegungen. Eine psycho-physiologische Studie (1887)).

Kalashnikova, M., Carignan, C. and Burnham, D. (2017). The origins of babytalk: Smiling, teaching or social convergence? *Royal Society Open Science* **4**: 170306.

Keltner, D. and Gross, J. J. (1999). Functional Accounts of Emotions. *Cognition & Emotion* **13**(5): 467-480.

Kim, H., Somerville, L. H., Johnstone, T., Polis, S., Alexander, A. L., Shin, L. M. and Whalen, P. J. (2004). Contextual modulation of amygdala responsivity to surprised faces. *Journal of cognitive neuroscience* **16**(10): 1730-1745.

Kraut, R. E. and Johnston, R. E. (1979). Social and emotional messages of smiling: An ethological approach. *Journal of Personality & Social Psychology* **37**: 1539-1553.

Lazarus, R. S. (2006). Stress and emotion: A new synthesis. Springer Publishing Company.

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron* **73**(4): 653-676.

Levenson, R. W., Ekman, P. and Friesen, W. V. (1990). Voluntary Facial Action Generates Emotion-Specific Autonomic Nervous System Activity. *Psychophysiology* **27**(4): 363-384.

Liu, X., Xu, Y., Zhang, W. and Tian, X. (2021). Multiple prosodic meanings are conveyed through separate pitch ranges: Evidence from perception of focus and surprise in Mandarin Chinese. *Cognitive, Affective, & Behavioral Neuroscience.*

Mauss, I. B. and Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion* **23**(2): 209-237.

Miyazaki, M. and Waas, J. R. (2003). Acoustic properties of male advertisement and their impact on female responsiveness in little penguins Eudyptula minor. *Journal of Avian Biology* **34**(3): 229-232.

Morris, D. (1956). The feather postures of birds and the problem of the origin of social signals. *Behaviour* **9**(1): 75-111.

Morton, E. W. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist* **111**: 855-869.

Morrison, D., Wang, R. and De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication* **49**: 98-112.

Mozziconacci, S. J. L. (2001). Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Modeling and User-Adapted Interaction* **11**: 297-326.

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

Neta, M., Tong, T. T., Rosen, M. L., Enersen, A., Kim, M. J. and Dodd, M. D. (2017). All in the first glance: first fixation predicts individual differences in valence bias. *Cognition and Emotion* **31**(4): 772-780.

Niebuhr, O. (2010). On the phonetics of intensifying emphasis in German. *Phonetica* **67**(3): 170-198.

Niebuhr, O., Voße, J. and Brem, A. (2016). What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice. *Computers in Human Behavior* **64**: 366-382.

Noble, L. and Xu, Y. (2011). Friendly Speech and Happy Speech – Are they the same? *The 17th International Congress of Phonetic Sciences*, Hong Kong: 1502-1505.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* **41**: 1-16.

Panksepp, J. (2005). Beyond a Joke: From Animal Laughter to Human Joy? *Science* **308**(5718): 62-63.

Parkinson, B. (1996). Emotions are social. *British Journal of Psychology* **87**: 663-683.

Paul, E. S., Sher, S., Tamietto, M., Winkielman, P. and Mendl, M. T. (2020). Towards a comparative science of emotion: affect and consciousness in humans and animals. *Neuroscience & Biobehavioral Reviews* **108**: 749-770.

Pellegrini, A. D., Dupuis, D. and Smith, P. K. (2007). Play in evolution and development. *Developmental review* **27**(2): 261-276.

Pfefferle, D. and Fischer, J. (2006). Sounds and size: identification of acoustic variables that reflect body size in hamadryas baboons, Papio hamadryas. *Animal Behaviour* **72**(1): 43-51.

Protopapas, A. and Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America* **101**: 2267-2277.

Puts, D. A., Barndt, J. L., Welling, L. L., Dawood, K. and Burriss, R. P. (2011). Intrasexual competition among women: Vocal femininity affects perceptions of attractiveness and flirtatiousness. *Personality and Individual Differences* **50**(1): 111-115.

Puts, D. A., Gaulin, S. J. C. and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior* **27**(4): 283-296.

Reby, D., Charlton, B. D., Locatelli, Y. and McComb, K. (2010). Oestrous red deer hinds prefer male roars with higher fundamental frequencies. *Proceedings of the Royal Society B: Biological Sciences* **277**(1695): 2747-2753.

Reddon, A. R., Ruberto, T. and Reader, S. M. (2021). Submission signals in animal groups. *Behaviour* **159**(1): 1-20.

Rosenberg, A. and Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Communication* **51**(7): 640-655.

Ryan, M. J. (1980). Female mate choice in a neotropical frog. *Science* **209**(4455): 523-525.√

Saarni, C. (1979). Children's understanding of display rules for expressive behavior. *Developmental psychology* **15**(4): 424.

Scarantino, A. (2014). The motivational theory of emotions. In *Moral psychology and human agency: Philosophical essays on the science of ethics.* New York, NY, US: Oxford University Press**.** pp. 156-185.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. *Approaches to emotion* **2293**(317): 31.

Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In *Emotions in personality and psychopathology*. C. E. Izard. New York: Plenum Press. pp. 493-529.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication* **40**: 227-256.

Scherer, K. R. (2022). Theory convergence in emotion science is timely and realistic. *Cognition and Emotion* **36**(2): 154-170.

Scherer, K. R. and Bänziger, T. (2004). Emotional expression in prosody: a review and an agenda for future research. In *Proceedings of Speech Prosody 2004***:** 359-366.

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

Singh, L., Morgan, J. L. and Best, C. T. (2002). Infants'Listening Preferences: Baby Talk or Happy Talk? . *INFANCY* **3**: 365-394.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: The MIT Press.

Stevens, M. and Merilaita, S. (2009). Animal camouflage: current issues and new perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**(1516): 423-427.

Sun, X. (2002). *The determination, analysis, and synthesis of fundamental frequency*. Ph.D. dissertation, Northwestern University, 2002.

Susskind, J. M., Lee, D. H., Cusi, A., Feiman, R., Grabski, W. and Anderson, A. K. (2008). Expressing fear enhances sensory acquisition. *Nat Neurosci* **11**(7): 843-850.

Tracy, J. L. and Robins, R. W. (2004). Show Your Pride: Evidence for a Discrete Emotion Expression. *Psychological Science* **15**(3): 194-197.

Van Bezooijen, R., Otto, S. A. and Heenan, T. A. (1983). Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology* **14**(4): 387-406.

Van Dijk, W. W. and Zeelenberg, M. (2002). Investigating the appraisal patterns of regret and disappointment. *Motivation and Emotion* **26**(4): 321-331.

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

Vrticka, P., Lordier, L., Bediou, B. and Sander, D. (2014). Human amygdala response to dynamic facial expressions of positive and negative surprise. *Emotion* **14**(1): 161.

Williams, C. E. and Stevens, K. N. (1972). Emotion and speech: Some acoustical correlates. *Journal of the Acoustical Society of America* **52**: 1238-1250.

Wilting, J., Krahmer, E. and Swerts, M. (2006). Real vs. acted emotional speech. In *Proceedings of Interspeech***:** 9th.

Wolff, S. E. and Puts, D. A. (2010). Vocal masculinity is a robust dominance signal in men. *Behavioral Ecology and Sociobiology* **64**: 1673-1683.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication* **46**: 220-251.

Xu, Y. (2013). ProsodyPro — A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France**:** 7-10.

Xu, Y. (2019). Prosody, tone and intonation. In *The Routledge Handbook of Phonetics*. W. F. Katz and P. F. Assmann: Routledge. pp. 314-356.

Xu, Y., Kelly, A. and Smillie, C. (2013). Emotional expressions as communicative signals. In S. Hancil and D. Hirst (eds.) *Prosody and Iconicity*, John Benjamins Publishing Co, pp. 33-60.

Xu, Y., Lee, A., Wu, W.-L., Liu, X. and Birkholz, P. (2013). Human vocal attractiveness as signaled by body size projection. *PLoS ONE* **8**(4): e62397.

Zei Pollermann, B. (2002). A Place for Prosody in a Unified Model of Cognition and Emotion. In *Proceedings of The 1st International Conference on Speech Prosody*, Aix-en-Provence, France**:** 17-22.

Zhao, K., Zhao, J., Zhang, M., Cui, Q. and Fu, X. (2017). Neural Responses to Rapid Facial Expressions of Fear and Surprise. *Frontiers in Psychology* **8**.