Summary

The voice is a prime channel of communication in humans and other animals. Voices convey many kinds of information, including physical characteristics like body size and sex, as well as providing cues to the vocalizing individual’s identity and emotional state. Vocalizations are produced by dynamic modifications of the physiological vocal production system. The source-filter theory explains how vocalizations are produced in two stages: (a) the production of a sound source in the larynx, and (b) the filtering of that sound by the vocal tract. This two-stage process largely applies to all primate vocalizations. However, there are some differences between the vocal production apparatus of humans as compared to nonhuman primates, such as the lower position of the larynx and lack of air sacs in humans. Thanks to our flexible vocal apparatus, humans can produce a range of different types of vocalizations, including spoken language, nonverbal vocalizations, whispering, and singing.

A comprehensive understanding of vocal communication takes both production and perception of vocalizations into account. Internal processes are expressed in the form of specific acoustic patterns in the producer’s voice. In order to communicate information in vocalizations, those acoustic patterns must be acoustically registered by listeners via auditory perception mechanisms. Both production and perception of vocalizations are affected by psychobiological mechanisms as well as sociocultural factors. Furthermore, vocal production and perception can be impaired by a range of different disorders. Vocal production and hearing disorders, as well as mental disorders including autism spectrum disorder, depression, and schizophrenia, affect vocal communication.

Keywords: emotion, evolution, source-filter theory, vocal communication, vocal tract, voice, voice disorders, voice perception, voice production

Subjects: Cognitive Psychology/Neuroscience
Mechanisms of Vocal Production

The Physiological Basis of Vocalizing

There are three main physiological components of vocal production: the subglottal system, the larynx, and the supralaryngeal vocal tract (Lieberman & Blumstein, 1988). Figure 1 illustrates the three main physiological components of vocal production, showing the approximate locations of the vocal organs.

![Figure 1. Illustration of the physiological basis of vocal production. Left: The three main physiological components of voice production: the subglottal system, the larynx, and the supralaryngeal vocal tract. Right: The source and filter components. Adapted from Lieberman and Blumstein (1988).](image)

The subglottal system generates airflow that supplies the energy necessary for vocal production. The trachea (windpipe) is a single tube branching into two airways, each leading into a lung in the subglottal system. Lungs are mammals’ main respiratory organs, and their volume can be increased or decreased during vocal production by contracting the expiratory or inspiratory muscles.
The larynx converts the airflow into sound through the vibration of vocal folds. The pressure of the air coming from the lungs blows the vocal folds apart and sucks them together. Cyclic closure of the vocal folds relies on a negative pressure inside the glottis (i.e., the slit between the vocal folds) due to Bernoulli’s aerodynamic law and the elasticity of the tissue (Hirose, 2010). In the glottis, when velocity of airflow increases, pressure between the folds decreases and draws the folds together. When the vocal folds are sucked together, airflow through the glottis ceases, and the subglottal air pressure rises. The vocal folds are blown apart when subglottal air pressure becomes greater than the medial compression of the vocal folds, and air escapes into the subglottal space. Due to the loss of air, the subglottic air pressure falls, and the tissue elasticity causes the vocal folds to adduct again. The elasticity of the glottis allows the tissue to store energy and to revert to its initial length. The rapid and repetitive movements of the vocal folds modulate the airflow and result in the production of a vocalization (see Zhang, 2016a, on the physics of vocal fold vibration). Changes in the vocal fold physiology affect vocal fold vibration and the resulting voice acoustics (Zhang, 2016b). The physiology of the vocal folds can be modified by activation of the main muscles within the larynx, the cricothyroid and thyroarytenoid muscles. The activation of those muscles affects the perceptual characteristics of the voice. Medial surface thickness has a large impact on the vertical phase difference between the upper and lower margins of the medial surface and on higher order harmonics excitation. The medial surfaces between the two vocal folds form an angle controlling the resting glottal opening. Decreasing the resting glottal angle lowers the phonation threshold pressure (i.e., amount of lung pressure required to initiate vocal fold oscillation), reduces noise production at high frequencies, and increases fundamental frequency.

The supralaryngeal vocal tract comprises the pharyngeal, oral, and nasal cavities. During speech production, the shape and length of the supralaryngeal vocal tract continually change, allowing the production of different sounds.

**Source-Filter Theory**

The source-filter theory explains how vocalizations are produced in terms of a two-stage process: (a) the production of a sound source and (b) the filtering of that sound (Fant, 1960; Titze, 1994; see Figure 1). The subglottal system generates and conducts airflow. The source transforms the airflow coming from the subglottal system into sounds through vibration of the vocal folds in the larynx. The sound spectrum is then shaped by the filter in the supralaryngeal vocal tract. The filter passes the sound at certain frequency bands and attenuates others. The spectral peaks of the sound spectrum are called formant frequencies (see Table 1). By manipulating the articulators such as tongue, lips, soft palate, and teeth, a wide range of articulatory possibilities is available. This is particularly important for speech production. For instance, different vowels can be produced by changing the position of the tongue within the oral cavity. Importantly, source and filter features can be controlled independently from each other in humans. The ability to independently control source and filter features also allows us to produce different types of vocalizations such as speaking, singing, or laughing. An animated magnetic resonance imaging (MRI) video of a speaking human adult is provided in Video 1 to illustrate the changes occurring in the vocal apparatus during speech production.
Watch Video 1 from https://emotionwaves.github.io/analysis/media/MyMovie.mp4

Video 1. An animated MRI video of a speaking human illustrating the changes occurring in the vocal apparatus during speech production.

Changes in the vocal apparatus during vocal production are reflected in the acoustic features of the vocalizations, which determine how the vocalizations sound to listeners. Table 1 shows definitions of common acoustic features and their perceptual correlates. The basic rate of vocal fold vibration is called fundamental frequency ($f_0$), which is the primary determinant of perceived pitch. Voice intensity is the energy through the unit area carried by a soundwave, which increases as the amplitude of the soundwave increases. Intensity and amplitude are perceived as loudness. Voice quality is the auditory coloring of the voice, which is determined by how the vocal folds vibrate (Laver, 1980). Three parameters of muscle tension (adductive tension, medial compression, and longitudinal tension) determine the configuration of the vocal folds and the interaction with aerodynamic factors related to subglottal pressure and glottal airflow (e.g., Gobl & Chasaide, 2010). These yield a variety of voice qualities (e.g., Esling et al., 2019). For instance, modal voice is the most frequent type of phonation during normal speech, and is produced when the vocal folds vibrate as a single unit. For modal voice quality, adductive tension, medial compression, and longitudinal tension are moderate. Creaky phonation is associated with intervals of glottal pulses, which are often characterized by some degree of aperiodicity. For creaky voice quality, high adductive tension and medial compression, but little longitudinal tension, are involved. When the vocal folds do not vibrate but are separated by a narrow gap, this results in a hissing sound, such as whispering. Whispering voice quality involves low adductive tension, moderate to high medial compression, and moderate longitudinal tension. Voice quality provides important information about the speaker, including their emotional state expressed in speech (e.g., Anikin, 2020; Grichkovtsova et al., 2012).

Table 1. Common Acoustic Features and Perceptual Correlates

<table>
<thead>
<tr>
<th>Acoustic parameter</th>
<th>Perceptual correlate</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_0$ (Fundamental Frequency)</td>
<td>Pitch</td>
<td>The basic rate of vocal fold vibration</td>
</tr>
<tr>
<td>$F_1, F_2, \ldots$ (Formant Frequencies)</td>
<td>Vowel sound</td>
<td>Frequency peaks in the spectrum with high degree of energy</td>
</tr>
<tr>
<td>Voice Intensity</td>
<td>Loudness</td>
<td>Energy through the unit area carried by a soundwave</td>
</tr>
</tbody>
</table>
Evolutionary Changes in the Vocal Anatomy

The mechanisms of vocal production are broadly similar in humans and other mammals, and the essential principles of source-filter theory apply to all vertebrate vocal production (Fant, 1960; Taylor & Reby, 2010). However, there are some morphological differences between the vocal production apparatus of humans as compared to other primates, including the lower positioning of the larynx and the lack of air sacs in humans.

When the anatomy of the human vocal apparatus is compared with those of other mammals (and with human newborns), one key difference is the lowered position of the larynx in human adults (see Video 1). The lowering of the larynx in the human vocal tract occurs during early infancy and has important articulatory-acoustic implications. Thanks to the lowered larynx, the human tongue has more space to move in, which leads to more articulatory flexibility (Schön Ybarra, 1995). Also, humans have a particularly long vocal tract and can therefore create a wide variety of vocal tract shapes while vocalizing (Lieberman et al., 1969). An agile tongue and more flexible vocal tract allow for a uniquely large range of variation in formant frequencies, and thus an enhanced ability to create acoustically distinctive sounds through supralaryngeal vocal tract filtering. However, some other mammals have the capacity to dynamically lower the larynx in the vocal tract, and a permanently descended larynx is observed in various mammalian species including lions and koalas (Fitch, 2018). The descended larynx is thus not responsible for the inability of other primates to produce novel vocalizations.

Many nonhuman primates and other mammals have sac-like extensions of the larynx, yet humans do not (Fitch, 2000; Negus, 1949; Schön Ybarra, 1995). Modeling (de Boer, 2009) and empirical analyses (Harris et al., 2006) suggest that air sacs are used to lower formant frequencies, which is also the effect of a descended larynx. Air sacs thus enable vocalizers to sound bigger and stronger, which might intimidate rivals (de Boer, 2009). It is plausible that the lack of air sacs in humans is acoustically compensated for with a descended larynx (Finch, 2018).

The anatomical structures that humans use for speech and vocal production have thus been modified in two respects over the course of evolution: Humans lost laryngeal air sacs and gained a permanently descended larynx. In addition, there are several further anatomical differences among mammalian species, such as control over breathing, control of the tongue, vocal folds, and the shape of the chin (see de Boer, 2019, for a review).

Vocal Communication Between Producer and Receiver

Vocal communication involves an interaction between a producer and one or several receivers (Dusenbery, 1992). The producer sends a vocal signal that conveys information to receivers. The vocal signal contains a wide range of information, which can be conveyed involuntarily or voluntarily. Receivers consciously or unconsciously make use of the rich information conveyed by others’ voices, and accurate perception can help receivers orient themselves to the environment and adapt their social behavior (Belin, 2006). The wealth of socially relevant information present in vocalizations can influence success in a range of social contexts, from mate-choice decisions to job interviews and political decisions (e.g., Klofstad et al., 2015;
In humans as well as other species, accurate perception and responding to conspecific alarm calls in the presence of a predator can increase the chance of survival (Adolphs & Anderson, 2018).

An illustrative model for understanding the mechanisms involved in the production and perception of information during vocal communication is the modified Brunswikian functional lens model (Brunswik, 1952; Scherer et al., 2003, 2011). In this model, internal processes are expressed in the form of physical signals in the producer’s voice. This process is called encoding. Different types of information (e.g., the emotional state of the producer) are encoded in the producer’s voice via a series of distal cues. These cues consist of patterns of objectively measurable acoustic features. They are distal to the receiver, because in order to communicate the information to the receiver, the distal cues must first be accurately perceived. When the distal cues are captured by the auditory perceptual mechanisms of the receiver, they become proximal cues. Those cues are proximal in the sense of being close to the receiver. On the basis of the proximal cues, the receiver makes judgments via a process called decoding or cue utilization. In successful vocal communication, the receiver accurately perceives the proximal cues and infers the information expressed in the producer’s vocal signal.

**Prepared Mechanisms of Emotional Vocalizations**

The production of distal cues and the perception of proximal cues might be determined by hard-wired mechanisms that do not require learning, that is, innate forms of vocal production and perception shaped by evolution (Mortillaro et al., 2013). One method for probing evolutionarily prepared features of vocal communication is to look for homologues of human vocal expressions in other species. The clearest evidence of homologues comes from the domain of nonverbal vocalizations. For instance, humans and other great apes produce laughter when tickled or during play (e.g., Davila-Ross et al., 2009), and many species produce screams when being threatened (e.g., Linnankoski et al., 1994). Moreover, when listening to vocalizations of other species, humans are able to accurately infer affective information about the particular behavioral context in which the vocalizations were produced (e.g., Kamiloğlu et al., 2020). These findings suggest phylogenetic continuity in vocal expression of emotions and inherent capacities to perceive acoustic regularities conserved across species.

Another approach for understanding prepared features of vocal behaviors is to examine to what extent vocalizations are affected by the absence of auditory learning. Examining deaf individuals who have had reduced opportunities for auditory learning can consequently provide important insights into prepared mechanisms. One study examined the acoustics of laughter produced by hearing and deaf adults and found that their laughs were acoustically similar (Makagon et al., 2008). Subsequent work has shown that hearing individuals can reliably infer many emotional states from nonverbal vocalizations produced by bilaterally congenitally deaf adults, demonstrating that some emotional information is preserved in the absence of auditory learning opportunities (Sauter et al., 2020). Auditory learning is thus not necessary for the development of nonverbal vocalizations in deaf individuals, suggesting that humans might have prepared mechanisms for vocal emotional expressions.
The human mind and behavior are dependent on processes that are both innate and learned (Sasaki & Kim, 2017). It is thus important to acknowledge that biological preparedness and learning interact in human vocal communication production and perception. One approach that can illuminate the interplay between evolutionarily prepared and culture-specific learning processes is the use of cross-cultural comparisons.

**Cultural Comparisons of Emotions in the Voice**

Cross-cultural studies have demonstrated cultural similarities as well as differences in vocal production and perception. An illustrative example is human laughter. The association between laughter and mirth is consistently observed across cultures (e.g., Bryant et al., 2018; Sauter et al., 2015). However, laughter signals are found to be different across cultural groups and understood more easily by members within a cultural group (Sauter, 2013; Sauter & Scott, 2007). Differences in production and perception of laughter across cultural groups suggests systematic differences reflecting culture-specific learning.

More broadly, vocal expressions of a wide range of positive and negative emotions are characterized by acoustic patterns that are shared by people across cultural backgrounds (e.g., Laukka et al., 2014). However, vocalizations produced by individuals from the same cultural group are acoustically more similar, demonstrating consistent cultural differences in emotional vocalizations. While vocal expressions of some emotions are well recognized across cultures, other vocal expressions differ across groups (e.g., Cordaro et al., 2016; Cowen, Laukka, et al., 2019; Sauter et al., 2010). These findings highlight the role of cultural learning in vocal production and perception, and suggest that shared evolutionary and cultural processes work together to shape vocal communication.

**Types of Vocalizations**

The encoding and decoding of information during vocal communication can occur via different kinds of vocalizations, including speech and nonverbal vocalizations, which differ in terms of source and filter characteristics (Frühholz & Grandjean, 2013).

**Speech**

In order to communicate specific meaning to others, we often use speech. Speech is the vocal form of language, a human adaptation that allows us to transmit symbolic information in an efficient manner. Speech production is a highly complex motor act that involves the subglottal system, the larynx, and the supralaryngeal vocal tract articulators, which work together in a highly coordinated manner. This complex motor act readily integrates auditory, somatosensory, and motor information, which are represented in the temporal, parietal, and frontal cortex, respectively (Kearney & Guenther, 2019).

Precursors of speech have been found in different mammalian groups (e.g., Hauser et al., 2002; Seyfarth et al., 1980). Control over fundamental frequency and formants for expressing, exaggerating, and faking physical traits in other animal species might be precursors of the
more complex formant modulation involved in speech production. For instance, red deer and
Diana monkeys can effectively modulate formants, thereby communicating body size-related
information (Reby et al., 2005; Riede et al., 2005).

Meaning in speech can be conveyed via semantics and/or speech prosody. Semantics refers to
linguistic phrases and constructions in speech, and is used to communicate lexical meaning
10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-297?
rskey=AZvvRb&result=1>). Speech prosody refers to modifications of suprasegmental cues
such as pitch, amplitude, and speech rate within a verbal utterance. By varying these
elements, producers are able to convey a wide range of information, including intentions and
emotions (Grandjean et al., 2006). For instance, a rise in pitch at the end of a sentence can
indicate whether the sentence is interrogative (a question) as opposed to declarative (a
statement). Changes in pitch also play a key role in communicating emotions (Scherer, 2018).

Nonverbal Vocalizations

Nonverbal vocalizations (sometimes referred to as affect bursts, see Scherer, 1994) are
nonlinguistic vocalizations such as sighs, laughs, and grunts. Nonverbal vocalizations are
considered a relatively “pure” type of vocal expressions, since they are not constrained by
linguistic structures (Scott et al., 2009). Since nonverbal vocalizations do not require precise
articulations like speech does, they are less reliant on detailed supralaryngeal movements.
Laughter, for instance, is considered to be a form of modified breathing rather than something
akin to speaking (Kohler, 2008). In fact, in evolutionary terms nonverbal vocalizations are
thought to predate speech, since nonhuman primates produce calls related to food,
dominance status, sex, affiliation, and aggression that are similar to those produced by
humans (Snowdon, 2003).

Nonverbal vocalizations are potent carriers of emotional information. They develop early in
ontogeny: Young children can communicate their emotions via nonverbal vocalizations like
laughter, which occurs in response to physical play (e.g., Panksepp, 2005), recognize emotions
from nonverbal vocalizations (Sauter et al., 2013), and reason about probable causes of
emotional vocalizations (Wu et al., 2017). In adults, nonverbal vocalizations can communicate
a variety of positive and negative emotions (e.g., Cowen, Elfenbein, et al., 2019; Hawk et al.,
2009; Sauter et al., 2010; Schröder, 2003), which are rapidly detected (Sauter & Eimer, 2010)
and largely consistent across cultural groups (e.g., Laukka et al., 2013; Sauter et al., 2010).

Singing

Singing differs from other types of vocal expressions in terms of the changes in voice
characteristics (Sundberg, 2019). For instance, narrow ranges of lung volume are typically
used in speech, while use of wider ranges of lung volume is required for singing. Moreover,
while subglottal pressure is used for varying loudness in speech, loudness also varies with
pitch during singing. Unlike speech production, when singing, subglottal pressure needs to be
constantly adjusted, taking both loudness and pitch into account.
During singing, vibrato is produced by periodic modulations when the pressure is exerted by the lungs. This is an important characteristic of learning singing technique. One idea about the origin of vibrato is that it is created by a pulsating contraction of pitch-raising muscles (Sundberg, 2019). Vibrato rate is generally stable, decreasing typically about 0.5 Hz per decade with advancing age (Sundberg et al., 1998).

Singing voices are often classified on the basis of pitch range, from basses (80–330 Hz) to sopranos (200–1,200 Hz). Different singing styles, such as Western operatic singing, pop, and rock, differ in glottal adduction and formant frequencies, as well as the use of voice quality (Hakanpää et al., 2019; Sundberg, 2019). One factor determining the development of vocal characteristics in different styles of singing is the acoustic environment. For example, opera needs to be loud to be heard over the orchestra without sound amplification, and thus opera singers learn to adjust their vocal apparatus to take maximum advantage of vocal tract resonances.

**Information Conveyed by the Voice**

Vocal signals convey enduring features like speaker identity and gender, as well as a wide range of transitory states, such as emotion and power (Kreiman & Sidtis, 2011).

**Physical Characteristics**

When listening to another person’s voice, receivers are typically able to very accurately infer whether the voice is from a man or a woman, even when listening to very brief segments (e.g., Andrews & Schmidt, 1997; Bachorowski & Owren, 1999) or degraded speech like whispers (Tartter, 1989). Acoustic features related to both the source (e.g., pitch) and filter (second formant frequency) are important for listeners to be able to infer sex from vocalizations (Bachorowski & Owren, 1999; Pernet & Belin, 2012).

Human listeners can also extract body size information from the voice (e.g., Patterson et al., 2008). In terms of acoustic features, higher pitch and formant frequencies are associated with smaller body size in most mammals (e.g., Pisanski et al., 2014; Taylor et al., 2010). In humans, mean pitch and formants are on average lower among men than women due to men having larger vocal folds (Rendall et al., 2005).

**Speaker Identity**

The voice is a dominant way of expressing speaker identity and also allows for conveying cues to speaker characteristics such as regional accents and individual differences in pronunciation (Lavan et al., 2019). From vocal signals, perceivers can discriminate between individuals and identify a specific person (Kreiman & Sidtis, 2011). Many factors can affect accurate perception of speaker identity in the voice. For example, speaker discrimination and recognition are more accurate for voiced speech as compared to whispered speech (Bartle & Dellwo, 2015; Yarmey et al., 2001), and discriminating speaker identity is easier from volitional as compared to spontaneous vocalizations (Lavan et al., 2016, 2018). For accurate voice identity perception, listeners do not only need to discriminate voices of different speakers but also to generalize percepts of identity across variable voice signals of the same speaker (Lavan et al., 2019). For unfamiliar voices, listeners have difficulty in generalizing
identity information of a single speaker and tend to perform worse in speaker discrimination for judgments of linguistically less similar words (Narayan et al., 2017), different types of vocalizations (Lavan et al., 2016), and across sung and spoken speech (Peynircioğlu et al., 2017). The effects of within-speaker variability are reduced when listeners are familiar with the voice: Listeners can accurately perceive variable recordings of the same speaker as a single identity even when there is substantial within-speaker variability. For instance, listeners can rapidly recognize familiar others just from hearing their voice over the phone, even after only having met them once. These differences in perception of speaker identity from familiar and unfamiliar voices are ascribed to access to robust representations of familiar voices which enable the listener to link voices to a single identity (e.g., Burton et al., 2016; Lavan et al., 2019).

**Emotion**

Emotions are expressed in vocalizations of many species including humans (Darwin, 1872). Mammalian species such as bats (Bastian & Schmidt, 2008), silver foxes (Gogoleva et al., 2010), and three shrews (Schehka & Zimmerman, 2009) change their voices when vocalising in agonistic or affiliate contexts. In humans, researchers have shown that emotional information can be reliably conveyed via the voice. For instance, nonverbal vocalizations can communicate negative emotions including anger, fear, and sadness and positive emotions like awe, compassion, and interest (e.g., Laukka et al., 2013; Simon-Thomas et al., 2009). Audio 1 presents examples of negative and positive vocalizations. Understanding how emotions affect the human voice requires considering the effects of emotions on the physiology of the vocal production system. Emotions affect our somatic and autonomic nervous systems (SNS and ANS). In turn, changes in the SNS affect the tension and action of the musculature used for vocal production in the larynx, vocal tract, and articulatory organs while ANS is associated with changes in the respiratory system in the lungs and trachea (see Scherer, 1986). Modifications in the vocal apparatus result in specific patterns of voice parameters. For instance, a change in the respiratory system can alter the loudness, duration, and pitch of the voice. Different emotional states that are associated with different physiological changes can thus be mapped onto distinct acoustic patterns in the voice (e.g., Scherer, 2003).
Audio 1a. Nonverbal vocalizations of emotions: (a) achievement

Audio 1b. Nonverbal vocalizations of emotions: (b) amusement.

Audio 1c. Nonverbal vocalizations of emotions: (c) anger.

Audio 1d. Nonverbal vocalizations of emotions: (d) disgust.

Audio 1e. Nonverbal vocalizations of emotions: (e) fear.

Audio 1f. Nonverbal vocalizations of emotions: (f) relief.

Audio 1g. Nonverbal vocalizations of emotions: (g) sadness.
Audio 1h. Nonverbal vocalizations of emotions: (h) sensory pleasure.

Listen Audio1h from https://emotionwaves.github.io/analysis/media/h.mp3

Audio 1i. Nonverbal vocalizations of emotions: (i) surprise.

Listen Audio1i from https://emotionwaves.github.io/analysis/media/i.mp3

Another approach to defining emotional states is core affect dimensions (i.e., arousal and valence). Arousal refers to the degree of physiological alertness or attentiveness, and valence is the degree of pleasure or displeasure (Russell, 1980). Arousal is reflected in the voice via pitch and loudness, as well as temporal features such as duration and speech rate. High arousal emotions such as fear and joy are associated with loud and high-pitched vocalizations, while vocalizations of low-arousal emotions such as boredom tend to be low in pitch and characterized by slow speech rate (e.g., Bachorowski & Owren, 1995; Banse & Scherer, 1996; Zei Pollermann & Archinard, 2002). Emotional valence, on the other hand, is associated with differences in intonation patterns and voice quality (i.e., energy distribution patterns in the spectrum; Scherer, 1986). For instance, energy is lower in frequency in voices expressing positive as compared to negative emotions (Goudbeek & Scherer, 2010; Zei Pollermann & Archinard, 2002).

Effects of Disorders on Voice Production and Perception

Vocal Production and Hearing Disorders

Voice disorders are characterized by abnormal production of pitch, loudness, and/or voice quality, resulting from disturbances in one or more subsystems of vocal production (Raming & Verdolini, 1998). Voice disorders range from mild symptoms to complete voice loss, and their prevalence is affected by sex, age, and occupation (Cohen et al., 2012; Van Houtte et al., 2009). Causes of vocal production disturbances can be broadly defined as organic or functional (Titze, 1994). Organic causes include vocal fold abnormalities and problems in laryngeal structures, due to, for example, trauma or cancer. Other organic causes of voice pathology include neurological disorders of the central nervous system (e.g., Parkinson’s disease) and of the peripheral nervous system (e.g., recurrent laryngeal nerve paralysis). Functional causes include behavioral disorders involving vocal misuse and hyperfunction,
such as phonothroma (e.g., screaming and yelling) and muscle tension imbalance. Organic and functional causes of vocal production disorders are not mutually exclusive. For instance, a functional cause like misuse of the voice can lead to organic disorders such as laryngeal trauma.

Hearing disorders are characterized by impaired auditory sensitivity of the physiological auditory system (American Speech-Language Hearing Association, 1993). Impairments range from mild hardness of hearing to profound deafness. Hearing is considered essential for normal vocal development (Brainard & Doupe, 2000), and an absence of auditory input affects a range of vocal production features. In recent years, many individuals with hearing impairments receive cochlear implants (CI). CIs provide little pitch information but do provide other cues, such as duration and intensity (Chatterjee et al., 2019). CI users typically experience difficulties in perceiving emotional prosody (e.g., Chatterjee et al., 2015; Hopyan-Misakyan et al., 2009; Paquette et al., 2018), and identifying producers’ sex or identity (e.g., Fu et al., 2005; Fuller et al., 2014; Mühler et al., 2009). However, there are factors that affect these abilities. For instance, early implantation of cochlear implants improves higher-order auditory functions, including enabling better voice discrimination (Zaltz et al., 2018).

Mental Disorders

Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impairments in social communication or interpersonal interaction and by limited and repetitive behaviours, interests, or activities (DSM-5, American Psychiatric Association, 2013). Voice processing deficits are observed in individuals with ASD and are thought to contribute to difficulties in social interactions and communication.

Speech prosody production differs between individuals with and without ASD. Atypical prosody features in individuals with ASD include increased or decreased intonation variability, atypical stress patterns and volume modulation, and inappropriate phrasing (e.g., Green & Tobin, 2009; McCann & Peppé, 2003; Nadj & Shaw, 2012; Paul et al., 2005). However, results are inconsistent in terms of differences between those with and without ASD in recognizing emotions expressed in the voice (see Matsumoto et al., 2016). There are multiple studies that have found that individuals with ASD have difficulty in understanding others’ emotions from the voice (e.g., Mazefsky & Oswald, 2007; Philip et al., 2010), while other studies have not found any such differences (e.g., Grossman et al., 2010; Jones et al., 2011). There are indications that individuals with ASD have difficulties in recognizing speaker identity from the voice (Boucher et al., 1998; Schelinski et al., 2014) and that this difficulty is more pronounced for learning novel voices and in the recognition of unfamiliar speaker voices (Schelinski et al., 2017).

Depression

Disturbances in speech production and perception are considered to contribute to the onset and maintenance of mood disorders such as depression (Leppänen, 2006). Vocal expressions of individuals with depression have recognizable characteristics, including low intensity,
increased monotonicity (i.e., less pitch variation and flat speech), reduced articulation rate, and varied pause duration (e.g., France et al., 2000; Moore et al., 2003; Zlochower & Cohn, 1996). Given that voice analysis is a noninvasive and easy-to-acquire method, researchers have attempted to use it to detect depression. Voice features including spectral, prosodic, and glottal features have been used to help diagnose depression (e.g., Alghowinem et al., 2013; Moore et al., 2008; Scherer et al., 2013). However, the validity of these studies is challenged by potentially confounding factors, such as age, gender, and emotion (Pan et al., 2019).

In the perception of emotional prosody, individuals with depression show a negativity bias for ambiguous stimuli (Kan et al., 2004) and reduced perception of positive prosody (Schlipf et al., 2013). More broadly, individuals with depression perform worse in recognition of discrete emotions from prosody compared to healthy controls and display a general bias toward negative emotions, which might reflect a deficit in executive functioning (Péron et al., 2011; Uekermann et al., 2008).

**Schizophrenia**

Voice atypicalities are a characteristic feature of schizophrenia and are considered important for diagnosis, treatment evaluation, and prognostic considerations (see DSM-5; American Psychiatric Association, 2013). Atypical voice patterns in schizophrenia include increased pauses, decreased variability in intensity, and tonal abnormalities in vocalizations (see Parola et al., 2020 for review and meta-analysis). These differences are considered a key component of impaired social communication in individuals with schizophrenia (e.g., Parola et al., 2018), which can in turn result in social withdrawal and further social cognitive impairments (e.g., Bambini et al., 2016; Tahir et al., 2019). However, the results on the atypicalities in acoustic patterns in schizophrenia are not conclusive due to large heterogeneity between studies.

Individuals with schizophrenia have difficulties in recognizing others’ emotions from the voice (see Hoekert et al., 2007 for a review; Leitman & Haigh, 2018). Errors include misperception—that is, the inability to detect changes in acoustic patterns expressing emotions in vocalizations, as well as misattribution errors, such as interpreting happy prosody as sounding angry. Such impairments are thought to be important contributors to psychosocial functioning difficulties (e.g., Hoekert et al., 2007), auditory hallucinations (Leitman et al., 2005), and delusions (Rossell & Boundy, 2005).

**Conclusion**

What is a voice? The voice is a primary way in which we communicate with others, which highlights the two main processes involved in vocal communication: production/encoding and perception/decoding. Both vocal production and the perception of basic biological meanings have been conserved to some extent across mammalian species, although the physiological apparatus of vocal production has undergone major changes in human evolution, which allow us to produce spoken language. The voice carries a wealth of information, including physical features of the producer, as well as their emotional state. This information can be conveyed via different types of vocal expression, such as nonverbal vocalizations, singing, and speech. Vocalizations allow listeners to perceive, interpret, and make use of a wealth of information.
However, the extent to which information and meaning are conveyed depends on a combination of psychobiological mechanisms and sociocultural factors. Furthermore, both vocal production and perception can be impaired by a range of different disorders.

**Further Reading**


**References**


### Related Articles

- Speech Comprehension and Cognition in Adult Aging
- Social Markers in Language and Speech
- Hearing and Cognitive Aging