Synthese (2021) 199:661–687
https://doi.org/10.1007/s11229-020-02687-z

Bringing back the voice: on the auditory objects of speech perception

Anna Drożdżowicz[1,2]

Received: 1 October 2018 / Accepted: 2 May 2020 / Published online: 18 May 2020

Abstract
When you hear a person speaking in a familiar language you perceive the *speech sounds* uttered and the *voice* that produces them. How are speech sounds and voice related in a typical auditory experience of hearing speech in a particular voice? And how to conceive of the objects of such experiences? I propose a conception of auditory objects of speech perception as temporally structured mereologically complex individuals. A common experience is that speech sounds and the voice that produces them appear united. I argue that the metaphysical underpinnings of the experienced unity of speech sounds and voices can be explained in terms of the *mereological* view on sounds and their sources. I also propose a psychological explanation (the *Voice Shaping Speech* model) of how we form and individuate the auditory objects of experiences of listening to speech in a particular voice. Voice characteristics enable determining the identity of auditory objects of speech sound perception by making some features of the speech signal stable and predictable.

Keywords Speech perception · Voice perception · Auditory objects

1 Introduction

Speech perception is one of the key perceptual abilities in humans. It fulfills a unique role by enabling linguistic communication. When you hear a person speaking in a familiar language you hear what they say by means of, among others, your ability to perceive the *speech sounds* uttered. It is partly due to your perceiving of speech

✉ Anna Drożdżowicz
anna.drozdzowicz@gmail.com

1 Department of Philosophy, Classics, History of Arts and Ideas, University of Oslo, P.O. Box 1020, Blindern, 0315 Oslo, Norway

2 The Institute of Philosophy, School of Advanced Studies, London, UK

sounds that you can come to understand what a person wants to convey.[1] The auditory experience of speech is an experience of speech sounds retrieved from a complex acoustic speech signal. When listening to a person speaking in a familiar language you also hear the *voice* in which speech sounds are produced. An auditory experience of hearing speech sounds produced in a particular voice is markedly different from an auditory experience of the same speech sounds produced in a different voice. When hearing a person speaking in a familiar language you typically seem to hear the *speech sounds* uttered *in a particular voice*. In such typical auditory experiences of listening to speech, speech sounds and voice seem to go hand in hand—they *appear united*.

The nature of the commonly experienced relation between speech sounds and voices in auditory experiences of listening to speech, as well as its psychological underpinnings, have so far not been systematically investigated. Yet, a typical auditory experience of hearing speech in a particular voice is among the most common perceptual states in humans and provides the perceptual basis for linguistic communication. Moreover, we have compelling evidence that speech sound and voice perception rely on two psychological abilities that may give rise to experiences that are very different from the familiar case of hearing speech sounds in a particular voice. The fact that in normal cases speech sounds and voice are experienced together calls for an explanation. It is of great importance to understand (1) *how speech sounds and voice are related in typical cases of auditory experience of listening to speech*, and (2) *how the auditory objects* of such experiences are formed. The goal of this paper is to provide answers to these questions.

With regards to (1), I will discuss the nature of the commonly experienced relation between speech sounds and voices in the context of the recent philosophical debate concerning the phenomenological unity of sounds and their sources (Pasnau 1999; Kulvicki 2008; Nudds 2010). I will argue that when it comes to the metaphysics of a typical auditory experience of listening to speech in a particular voice, Casey O'Callaghan's (2017) *mereological* view makes good sense of the experienced unity between speech sounds and voices. Drawing on the mereological view, I will argue that the audible relation between speech sounds and their sources can be explained as a part-whole relation between speech sounds and voicing events. Normally, the voice of the speaker is a participant in such events. In listening to voicing events, hearers can also become aware of the speaker to whom the voice belongs. I will also briefly indicate how other views on the relation between sounds and sources could contribute to our understanding of the relation between speech sounds and voices. Regarding (2), I will develop a model compatible with the mereological view (the *Voice Shaping Speech model*) that explains how we form and individuate the auditory objects when listening to speech in a particular voice. Drawing on recent empirical and theoretical research, I will argue that voice characteristics make some features of the speech signal stable and pre-

[1] Another common observation is that when listening to an utterance spoken in a familiar language, one becomes aware of the meaning conveyed with that utterance. This observation has led some to adopt the view on which meanings can also be perceived (e.g. Bayne 2009; Siegel 2010; cf. O'Callaghan 2011a; Brogaard 2018). Here I leave it open whether such awareness of meanings is perceptual [I discuss some problematic aspects of this view in Drożdżowicz (2019)]. In any case, such awareness would be compatible with the claim that speech sounds are perceived when listening to a familiar language and seems to be largely independent from issues of speech and voice perception discussed here. Because of that, this paper does not engage in the debate about 'hearing meanings'.

dictable and in this way enable listeners to determine the identity of auditory objects of speech sound perception. The proposal provides a psychological explanation of the experienced unity between voice and speech.

Until recently, the topic of voice perception has received relatively less attention than that of speech perception (Belin et al. 2004; Smith 2009; Schweinberger et al. 2014). This paper helps to bring back the voice into the overall picture of speech perception. In order to properly characterize and explain the experienced relation between speech sounds and voices, it is important that we get a good understanding of how speech and voice are perceived. To this end, I start by sketching a recent conception of auditory objects as mereologically complex individuals that exhibit a *temporal structure* (as in O'Callaghan (2008, 2017) (Sect. 2) and apply it to the case of speech sound and voice perception (Sect. 3). In Sect. 4, I present evidence suggesting that speech sound and voice perception rely on two psychological abilities that may give rise to a variety of experiences of speech and voice. Material presented in Sects. 2–4 will be instrumental for developing a psychological model that explains how we form the auditory objects when hearing speech sounds in a particular voice (Sect. 6). The metaphysical underpinnings of the experienced unity between speech sounds and voices are discussed in Sect. 5.

2 Auditory perception and its objects

Vision may be king, but *auditory perception* is without doubt vital for providing us with information about the world. It gives us, for example, the advantage of being able to hear sounds before we can see their sources, as in the case of a siren of an ambulance rushing through the streets or the footsteps of a villain creeping up behind. Auditory stimuli are patterns in the displacement of air molecules resulting from vibrating objects (Plack 2014). A *sound* can be defined as a vibration that propagates as a typically audible mechanical *wave* of pressure and displacement, through a medium such as air or water or as a sensation, i.e. a stimulation of the hearing mechanism that results in the perception of sound.[2] In humans, the sense of hearing is stimulated by acoustical energy, i.e. sound waves that enter the outer ear. Sound waves set into vibration the eardrum and the attached bones, which transfer the mechanical energy to the inner ear, the cochlea. The mechanical energy is then transduced into neural impulses within the cochlea due to the stimulation of the sensory hair cells that connect with the eighth cranial, or auditory nerve and reach the brain (Plack 2014).

So described, the human auditory system produces experiences in which hearers become aware of and can attend to distinct units—so called *auditory objects* (Griffiths and Warren 2004; Bizley and Cohen 2013; O'Callaghan 2016). An auditory object is a result of auditory psychological processes of parsing a perceptual scene into distinct units that can be attended to, distinguished from the background and from each other

[2] A sound wave is characterized by its *frequency*, i.e. the number of cycles that a wave goes through per unit time, and *amplitude*, i.e.—is the maximum pressure change from the wave's mean value. Frequency determines most pitch (De Cheveigne 2005), amplitude normally corresponds to the psychological experience of loudness (Plack and Carlyon 1995). Other sound wave parameters include duration, timbre, sonic texture and spatial location (Burton 2015).

(O'Callaghan 2016, p. 2.2). This notion of an auditory object is different from audition's intentional objects and from the notion of ordinary material objects, such as bells or trumpets, that hearers can also become aware of in auditory experience. According to O'Callaghan (2008), auditory objects are mereologically complex individuals that are individuated and tracked by the human auditory system due to their *temporal* and pitch characteristics. This conception of auditory objects is based on our knowledge about psychological regularities that govern the human auditory system in parsing the auditory scene into distinct units. In much of the empirical research on perception the notion of a perceptual object is tightly connected with the *object analysis* in a sensory domain (Griffiths and Warren 2004). Following this line, auditory objects are characterized as "computational results of the auditory system's capacity to detect, extract, segregate and group spectrotemporal regularities in the acoustic environment" (Bizley and Cohen 2013, p. 693). I now briefly explain how such auditory objects are formed. In Sect. 3, I apply this notion of auditory objects to the cases of speech sound and voice perception.

A helpful way to understand how auditory objects are formed is by means of the notion of the *auditory stream* (Bregman 1990; Bendixen 2014; O'Callaghan 2008, 2017, ch. 2; Winkler et al. 2012). The human auditory system analyses the auditory scene into distinct auditory streams on the basis of similarities in frequency, amplitude, etc. For example, each step of a walking person is a unique acoustic event, but the auditory system tends to group separate stimuli together into a temporal sequence of footsteps. As a result, such streams or sequences are typically perceived as individual auditory objects. Auditory streams are characterized by features such as pitch, timbre, and loudness at a time. Several auditory streams might occur simultaneously. Such streams can be understood as perceptual individuals that have auditory qualities. So conceived, auditory streams can be individuated in hearers' experience as auditory objects, which persist through time and survive changes to their audible qualities. For example, an auditory stream may start as low-pitched and quiet but gradually shift into being high-pitched and loud and still be individuated as one continuing stream. In discussing the nature of auditory streams O'Callaghan observes (2008, p. 820):

> (…) since the identities of many recognizable sounds, such as the sounds of spoken words, police sirens, bird calls, and tunes, are tied to distinctive patterns of change in audible qualities through time, the constraints imposed upon auditory perception by the needs of recognition and concept application require that audible individuals exhibit *temporal structure* (emphasis added).

The claim that typical objects of auditory perception exhibit a temporal structure is supported by several observations about how auditory streams are typically organised in perceptual experience (O'Callaghan 2008, pp. 820–824). First, the organisation of auditory objects *in time* is determined by auditory processes operating according to principles similar to those that determine the organization of visual objects in space. Audition provides temporal boundaries for the individuation of auditory streams. According to one such principle, the principle of exclusive allocation (Bregman 1990, pp. 14–15), a sequence of tones *p p p q r r r* may appear as either *p-p-p-q* and *r-r-r* or as *p-p-p* and *q-r-r-r*, depending on the relative pitch distance between the tones. In auditory experience, where two streams are normally perceived, tone *q*

must belong to one of them. Second, selective attention in listening may affect experienced auditory streams resulting in figure-ground and shift effects analogous to those commonly observed in vision, but dependent on time rather than on spatial organisation (Bregman 1990, p. 18). Third, auditory streams individuated in experience persist through noise and masking and are subject to *amodal* completion analogous to the one observed in vision. For example, auditory streams with removed segments later replaced with white noise are experienced as continuing despite signal degradation.[3]

According to O'Callaghan, the above regularities provide evidence for the claim that the individuation at a time and identification over time of auditory streams is critically determined by the features of a signal's temporal profile (2008, p. 822). *Time*, according to O'Callaghan, determines the internal structure and composition of auditory objects understood as auditory streams, in a similar manner as space determines internal organisation of visual objects. *Pitch*, i.e. the sound's quality determined by its frequency, is also important in the process of individuation of auditory streams (Bizley and Cohen 2013).[4] For example, we can hear noises coming from separate speakers but having the same pitch as one audible individual. Pitch is important for spatial location of auditory streams determining the structural relations between auditory objects at a time (O'Callaghan 2008, p. 823).

Online studies on auditory processing seem to support the claim that auditory objects have a temporal structure. According to Bizley and Cohen (2013), the process of determining the component objects within the auditory scene crucially depends on their temporal structure (p. 693). In natural listening conditions, *onset time* is one of the most important cues for the grouping and identification of auditory streams. Sound components which have the same onset time are likely to be perceived as originating from the same source and strongly indicate the presence of a single auditory individual (Bizley and Cohen 2013, p. 695). Moreover, natural sounds, such as speech, are often *harmonic*, i.e. they have energy at integer multiples of the fundamental frequency. Individual harmonics tend to change coherently over time and are thus grouped together in auditory experience (Bizley and Cohen 2013, p. 695). Studies investigating neural activity during listening suggest that *temporal coherence* is an important factor in the forming of auditory streams: when two tones are presented simultaneously they are typically perceived as one auditory stream (Elhilali et al. 2009).

Time is a defining parameter for the formation of auditory objects. The identity and organisation of auditory objects depend on varying time and pitch characteristics of the acoustic signal. In this sense, auditory objects can be seen as having a temporal structure. So conceived, auditory objects do not seem to *wholly* exist at a given moment, rather, they appear to occur, unfold, or take place, and thus occupy time (O'Callaghan 2008; for discussion see also Skrzypulec 2018). In what follows I provide evidence suggesting that auditory objects of speech sound and voice perception are no different in this respect, in that they also crucially depend, among others,

[3] Auditory amodal completion processes complete the temporal contours of a stream through temporal occlusion. Distinct audible individuals at a time are sequentially integrated into distinct auditory streams (O'Callaghan 2008).

[4] Sounds with pitch correspond to regular or periodic patterns of vibration that differ in fundamental frequency and complexity (O'Callaghan 2016).

on the temporal characteristics of the signal. The material presented in Sect. 3 will be instrumental for developing a psychological model explaining how the abilities to perceive speech sounds and voices are jointly used at the level of processing, giving rise to the experienced unity of the two (Sect. 6).

3 Auditory objects in speech perception and voice perception

Among various auditory stimuli, human speech is an extraordinarily important and complex one (Nygaard and Pisoni 1995). Speech sound perception concerns the earliest stages of speech processing, during which the listener maps the time-varying acoustic signal of speech into a set of discrete linguistic representations. These representations are typically construed in terms of sequences of phonetic segments, i.e. consonants and vowels. Phonetic segments form the words of a language (Pisoni and Remez 2005). For example, the word *keep* is composed of three phonetic segments: an initial consonant (in phonetic notation, symbolized as /k/), a medial vowel (/i/), and a final consonant (/p/). Phones and phonemes are the building blocks of speech. Phonemes are the smallest contrastive linguistic units with a capacity to bring about a change of meaning. Phones, on the other hand, are audible features which may occur in distinct utterances and which may differ in otherwise indistinguishable utterances (as [t] and [d] in [mat] and [mad]). Different languages deploy different sets of these building blocks (O'Callaghan 2015).

Human speech perception is an extraordinarily complicated process due to a considerable and highly systematic complexity in the mapping between acoustic signal and phonetic structure. Researchers agree that there is not a simple one-to-one mapping between the phonetically relevant acoustic properties of speech and the phonetic structure of an utterance (O'Callaghan 2015).[5] Given the complex nature of the articulatory process, in most cases phonetic contrasts are specified by multiple acoustic properties. As a result, it is necessary that hearers rely on some specialized mechanisms in order to analyse the complex, context-dependent speech signal (Holt and Lotto 2008).[6] Theories of speech perception provide competing explanations for how humans are capable of succeeding in this task. Such theories provide two broad accounts of the objects of speech sound perception (O'Callaghan 2015). On the *motor* and *direct realist* theories, perceiving speech involves recovering information about articulatory gestures from the acoustic signal. According to the motor theory, speech perception has as its objects articulatory gestures, i.e. motor commands of producing phonemes (Liberman et al. 1967; Galantucci et al. 2006), whereas for the direct realist theorists such objects are actual vocal tract movements (Fowler et al. 2003). *General* and *learning* theories of speech perception (Stevens and Klatt 1974), on the other hand, propose that the objects of speech perception are best construed as complex acoustic structures corresponding to perceived phonemes.

[5] Many factors contribute to this complexity in mapping, see Kent (1977), Fowler (1984) and Nearey (1989).

[6] The objects of speech perception are multi-modally accessible, as has been convincingly shown by examples of visual influences on speech sound perception in the McGurk effect (McGurk and Macdonald 1976) and tactile influences (Gick and Derrick 2009; Yeung and Werker 2013).

Theories of speech perception disagree about the nature and the locus of the objects of speech perception. The general and learning theories' conception that construes such objects as complex acoustic structures corresponding to perceived phonemes seems to fit well the notion of auditory objects as mereologically complex individuals adopted in this paper. All of the above theories would, however, largely agree on some basic facts about how such objects are typically formed. The identity and qualities of objects in speech perception are determined by the *temporal* parameters of the acoustic signal. Such parameters determine how the articulated speech signal is organised into phonetic segments by the hearer's auditory system. First, one of the key parameters responsible for discriminating between speech sounds is the *voice onset time.* Different speech sounds differ in their *voice onset time,* i.e. the time from the beginning of an utterance until voicing begins. In the word 'pad' the voicing starts at the end of /a/ since /p/ is a voiceless consonant. In the word 'bad' the voicing starts from the very beginning with the production of voiced /b/ phoneme. In this way voice onset time partly determines which speech sounds are perceived.

Next, the auditory analysis of speech sounds typically requires information about the signal that comes before and after the articulation of a specific sound. For example, whether a certain auditory stream produced by a speaker is perceived by a hearer as /k/ may depend on the preceding and following contextual acoustic information. Moreover, speech, as many other natural sounds, has a *harmonic* structure (Bizley and Cohen 2013). Individual speech harmonics change coherently over time and are thus grouped together as auditory streams of speech coming from the same speaker. Finally, neuropsychological evidence strongly suggests that the representation of temporal characteristics of the speech signal is critical for the ability to perceive speech.[7] While hearers can tolerate spectral, i.e. pitch-related, impoverishment of speech to a remarkable degree (Shannon et al. 1995; Lorenzi et al. 2006), temporal manipulations cause observable failures of perception (Adank and Janse 2009). This is also supported by studies showing that deficits such as *cochlear hearing loss* reduce the ability to use temporal fine structure cues and affect hearers' ability to recognize speech sounds, as well as other types of complex harmonic acoustic signal (Moore 2008).

The above evidence supports the claim that the identity and internal organization of the auditory objects of speech perception are determined by the time-dependent parameters of the acoustic signal. In this sense, auditory objects of speech perception exhibit a temporal structure. While the auditory experience of speech is studied primarily as an experience of speech sounds mapped onto a complex acoustic speech signal, it is normally also an auditory experience of a *voice* in which speech sounds are typically presented to the hearer. Information about voice and speech is contained in and recoverable from a single source, the speech signal. Speech formants are quite special: they carry both linguistic information about speech sounds and voice information, including paralinguistic cues that are directly related to the size of the vocal tract and can therefore provide estimates of body size and other speaker features (Belin et al. 2004).

[7] Neuroanatomical models of speech perception explain how human brain represents the temporal structure of the speech signal. According to Giraud and Poeppel (2012), delta, theta and gamma oscillations in the brain track the multi-timescale, quasi-rhythmic properties of speech.

For humans, voice is typically the most prevalent and most important sound of the auditory environment. It can convey rich information not only about a speaker's arousal and emotional state (Banse and Scherer 1996), but also provide extra-linguistic cues that reflect more stable speaker characteristics, including identity (Baumann and Belin 2010), biological sex and social gender (Owren et al. 2007), age (Mulac and Giles 1996), and sometimes even the socioeconomic or regional background of a speaker (e.g. Rakić et al. 2011). Our ability to analyze and understand information that is contained in voices is, therefore, crucial for a variety of social interactions (Belin et al. 2004; Campanella and Belin 2007). However, until recently, the topic of voice perception, i.e. the perception of paralinguistic information retrievable from voices, has been far less investigated than that of speech perception (Belin et al. 2004; Schweinberger et al. 2014).[8] I now briefly explain how human voices are typically produced and perceived.

Vocal sounds are generated by the interaction of a source (the vocal folds in the larynx) and a filter, i.e. the vocal tract above the larynx (Ghazanfar and Rendall 2008). Voiced sounds correspond to a periodic oscillation of the vocal folds with a well-defined *fundamental frequency* (f0). Although, for an individual, the range of f0 values can vary quite a lot during normal phonation or singing, the average f0 of a particular speaker is to a large extent a function of the size of the vocal folds (Latinus and Belin 2011). Human voices tend to vary extensively. Small differences in the dimensions and histology of the individual body parts that speakers use in phonation result in great individual variability among speakers in the individual acoustic patterns they can produce (Schweinberger et al. 2014).

The perception of fundamental frequency of a speaker's voice is a key parameter in recognizing the voice (Baumann and Belin 2010). Voice perception results in auditory objects whose identity and internal organisation are determined in a way analogous to that in which the identity and organisation of other auditory objects are determined. The key parameters utilized in identifying and tracking vocal sounds at a time, and individuating them across time, such as the voice's fundamental frequency, pitch, shimmer and tempo, are tied to distinctive patterns of change in *temporal* characteristics of the acoustic signal. This conception of auditory objects in speech sound and voice perception provides an empirically sound picture on how humans form and individuate auditory objects when listening to speech and voice. This picture will be crucial for developing a psychological model in Sect. 6 explaining how the abilities to perceive speech sounds and voices are related at the level of processing, giving rise to the experienced audible relation between speech and voice.

[8] Two notable exceptions in the recent philosophical literature are Smith's (2009) paper on voice and speech (I discuss part of it in Sect. 5) and Di Bona's (2017) paper on gender, a feature that is typically revealed in voice perception, in which she argues that gender can figure in the contents of hearers' perceptual experience.

4 Voice perception and speech perception: two abilities, various experiences

Human voices can be used for a variety of purposes, such as singing, yodeling and other forms of non-linguistic signaling. In this section I discuss evidence which suggests that speech sound and voice perception rely on two psychological abilities that may give rise to a variety of atypical experiences of combining speech sounds and voices. However, in a typical auditory experience of hearing speech, we hear speech sounds in a particular voice. Given that speech sound and voice perception rely on two psychological abilities and that there are various possibilities of combining voice and speech, the *audible relation between speech sounds and voice* experienced in such typical cases calls for an explanation. In Sect. 5 I explain its metaphysical basis. In Sect. 6 I explain how the two psychological abilities are combined at the level of processing giving rise to the commonly experienced relation between the two.

Several observations suggest that speech sound perception and voice perception involve and rely upon two abilities. First, there is an *evolutionary argument* for the distinct origin of the abilities to perceive voice and to perceive speech sounds. The ability to recognize calls and vocalizations of one's own species, which is likely to underlie our ability to perceive voices, has been prominent in the auditory environment of vertebrates for millions of years before speech emerged and can be found in other species (Belin et al. 2004). The ability to perceive speech sounds, on the other hand, evolved relatively recently as an abstract and complex use of voice by the human species (Fitch 2000). Second, *studies on neural correlates* of speech sounds and voice perception suggest a dissociation between middle superior temporal sulcus (STS) regions that are more responsive to the presence of speech but not to the meaning, i.e. they respond to backwards speech but not to understandable modulated noise, and more anterior regions of the left STS/superior temporal plane that are claimed to be more involved in comprehension of the speech signal, including signal with a degraded acoustic structure (Hickok and Poeppel 2000; Scott and Johnsrude 2003). Third, *neuropsychological* studies show that although deficits in discriminating unfamiliar speakers and deficits in the recognition of familiar speakers ('phonagnosia') can be dissociated, they tend to occur more often after lesions in the right hemisphere (e.g. Neuner and Schweinberger 2000). Furthermore, cases of preserved speech perception but impaired speaker recognition, as well as cases of aphasia with normal voice recognition (Assal et al. 1981; Belin et al. 2004) provide evidence for a *double dissociation* between speech perception and voice perception. These three types of observations are often taken to support a model where retrieving information about speech sounds and retrieving information about speaker's voice and identity rely on two psychological abilities and are processed in partially dissociated cortical regions (Belin et al. 2004).

The two abilities to perceive voice and speech can give rise to various experiences in which voice and speech sounds can be combined in various interesting ways or come apart. *Attention* may be one important factor here: attending to sounds is not strictly necessary for the formation of auditory objects (Micheyl et al. 2003), but it can influence auditory perception (Shinn-Cunningham 2008). A quick look at auditory phenomenology reveals that we can attend separately to some of the characteristics of

speech and voice in our experience (Bizley and Cohen 2013, p. 694). When listening to a person speaking in a familiar language we can, by focusing our attention on the sounds, distinguish different qualities of both the speech sounds that are spoken (e.g. as rolled, trilled, blended) and the voice that produces them (e.g. as harsh, soft, melodic).

Phenomenological observations about some special cases reveal some interesting ways in which hearing speech sounds and voices can be combined.[9] For example, when listening to a choir singing in a familiar language, we can hear the speech sounds and understand the lyrics. But in some such cases the common experience of hearing speech in a particular voice is lost when individual choir voices singing the lyrics are blended into one, unified voice. As a result, the normal one to one correspondence between a string of speech sounds and a voice is absent. Another interesting example is that of listening to speech in switching or rotating voices, so that speech sounds belonging to a particular word are always heard in a different voice. Peter Hall's production of Oresteia (Spencer 1999) provides a good example of a Chorus that never speaks in unison. Instead, the script is divided into words, each of them spoken out by a different voice. As a result, one experiences individual rotating voices throughout the Chorus lines.[10] In this case the voice in which speech is delivered appear interestingly disunified. Finally, early speech synthesizers provide a good example of how we can perceive speech sounds provided in the medium that is very different from human voice. Most voiced phonemes can be recreated as synthesized sounds with energy at formant frequencies. Synthetic speech sounds, such as *sinewave speech* composed of pure tones at the first three formant frequencies, lack typical vocal qualities, but can be perceived (Remez et al. 1981).

Such atypical cases show that there are various ways in which speech and voice may be combined. What typically comes together in a particular format, need not necessarily come together in that format. Hearing speech in a particular voice points to a familiar but not essential connection. The above evidence suggests that speech sound perception and voice perception rely on two psychological abilities that can give rise to experiences of speech and voice in which the two can be combined in different ways.

However, a familiar auditory experience is that of hearing speech sounds in a particular voice. In such typical experiences, speech sounds and voice *appear audibly related*. Try as you might, when actually listening to speech sounds spoken in a voice, it is impossible to hear them in an abstract, voice-less way. Listening to speech sounds in one voice results in an auditory experience markedly different from the experience of listening to the same speech sounds spoken in another voice. The audible impression of the speaker's vocal characteristics seems normally inextricable from the impression of speech sounds. There is a sense in which speech sounds appear *bound with* the voice that produces them resulting in a particularly salient audible impression of

[9] I thank Barry Smith for helpful discussions about how voice and speech can be combined and for drawing my attention to some of these cases.

[10] You can experience the effect by listening to the recording available on: https://www.youtube.com/watch?v=n7WKMovLLho.

their *sourcehood*.[11] A typical auditory experience of speech sounds spoken in a voice doesn't present speech sounds and voice as two independent phenomena. It is in special cases—when a speaker has a cold affecting her voice or when she pronounces a speech sound in a peculiar manner—that our attention is drawn to either one or the other. But even in such cases, speech sounds are heard as closely tied to a particular voice. These observations suggest that in typical cases speech sounds and voice are not merely *co-present*, as the previous atypical cases of combining speech and voice might seem to suggest.[12] Rather, there is a sense in which speech sound and the voice *appear united*.

What is the nature of the audible relation between speech sounds and voice that are commonly experienced as a unity? And can we account for this phenomenological observation by pointing to specific processing regularities in speech sound and voice perception? The latter question requires explaining how the two abilities to perceive speech sounds and voices are connected. In the next section I continue discussing the audible relation between speech sounds and sources and explain the metaphysical underpinnings of this phenomenological datum, a task that requires taking a closer look at recent philosophical debates on sounds and their sources. In Sect. 6 I explain the psychological underpinnings of the phenomenological datum, a task that requires a conception of how the abilities to perceive speech sounds and voices are connected at the processing level to enable the formation of auditory objects of experiences of speech sounds in a particular voice.

5 Voice, speech sounds and their sources

In typical cases of listening to speech sounds of a familiar language, a hearer experiences speech sounds as coming in a particular voice. There is a sense in which in such common experiences, voice and speech appear *united* to the hearer. This phenomenological datum calls for an explanation. How are speech and voice related in a common experience of listening to speech in a particular voice? In this section I explain the metaphysical underpinnings of this phenomenological unity. I start by addressing some ambiguities concerning the notion of voice. On the one hand, voice can be understood as a physical source of vocal sounds grounded in the physiology of the speaker's vocal folds and vocal tract and resulting in a relatively stable set of audible characteristics. Stable vocal tract characteristics give rise to a particular kind of audible profile, i.e. they determine the audible qualities of one's voice. So understood, a voice typically belongs to a particular *speaker*. Voice is what hearers can discriminate and recognise when listening to vocalisations from different people. In many ordinary cases, hearing a voice will also allow the hearer to recognies a person to whom the voice belongs, a speaker (Smith 2009). On the other hand, 'voice' is often used to speak of particular events of producing voice signal, i.e. vocal sounds or vocalisations. Vocalisations are produced on a particular occasion by a specific physical source, i.e. the vibrating vocal

[11] Leddington (2014) provides an interesting discussion of the phenomenology of audible relation between environmental sounds and their sources and argues that sounds are normally experienced as bound or fused to their sources.

[12] I thank an anonymous reviewer for raising this issue.

folds that result in specific audible characteristics. I will call such events of producing vocalisations *voicing events* or simply *voicings*. The two senses of voice are closely connected. In vocal production, a physical source or object, i.e. speaker's vocal folds and tract, gives rise to voicing events where vocal sounds are produced, and such vocal sounds have a particular profile, i.e. exhibit specific audible characteristics. The above sketched distinction, between *voice*, i.e. a physical source—speaker's vocal folds and vocal tract resulting in a particular profile of vocal features, a *voicing,* i.e. an event of producing voice sounds on a particular occasion, and a *speaker* to whom a voice normally belongs, will be employed in the discussion presented in this section.

In typical cases of linguistic communication, speech sounds are heard as coming in a particular voice, which typically belongs to a particular speaker. In such cases, the voice and the speech appear united. However, as O'Callaghan observes (2017, p. 109), there is a variety of ways in which two things may be experienced as united: elements of a scene may appear unified within the same spatial field, two properties may be experienced simultaneously, a property may be experienced as belonging to its bearer, a perceptible part may seem unified with the whole, two things may appear united through an experience of causation. Not every kind of phenomenological unity would adequately capture the unity experienced in the case of voice and speech sounds. What is the nature of this phenomenological unity? In order to investigate this question, I will follow a recent philosophical debate on the relation between *sounds* and their *sources*. In recent years, the audibly apparent close relation between sounds and their sources has been a subject of an in-depth discussion (O'Callaghan 2017, ch. 6, see also Kulvicki 2008; Nudds 2010; Leddington 2014).[13] Given limited space, I will briefly discuss three accounts and apply them to the case of voice and speech sounds. I will suggest that when it comes to the metaphysical underpinnings of a typical auditory experience of listening to speech in a particular voice, O'Callaghan's *mereological* view makes good sense of the unity between voice and speech sounds. I will also discuss whether and how other views can contribute to our understanding of the experienced unity of speech sounds and voices.

Consider first the *property* view on the relation between sounds and their sources, according to which sounds are audible attributes of physical objects in a way analogous to visible colours or shapes, i.e. properties of physical objects. According to Pasnau (1999), sources of sounds, i.e. vibrating objects, possess or bear sounds. For Kulvicki (2008), sounds are best explained as the standing dispositions of their sources, i.e. material objects, to vibrate in response to being 'thwacked'. On these views, the relation between sounds and their sources is a relation of sources audibly *instantiating* sounds. In the case of speech sounds heard in a particular voice, the property view would imply that speech sounds are best seen as properties of voices that produce them. A voice would possess or bear speech sounds, as in Pasnau (1999) or, following Kulvicki, one could try to construe speech sounds as stable dispositions to vibrate that voices have when set into motion. When applied to the case of hearing speech sounds and voice, both versions of the property view face some problems. The main problem is this: rather than being properties of anything, speech sounds seem to be event-like

[13] This section follows in its structure the discussion presented in Chapter 6 from O'Callaghan (2017) and applies some of the observations that O'Callaghan makes to the case of auditory experience of speech and voice.

individuals or units on their own that exhibit audible properties themselves (e.g. can be heard as rolled or blended) and can persist through time—call it the *event* view. A similar complaint is made by O'Callaghan (2017, pp. 98–99), when assessing the property view for environmental sounds and their sources. Like environmental sounds, speech sounds and their strings, do not seem to fit the metaphysical bill of properties.

The proponents of the property view could try to object to the event-construal of sounds. Kulvicki (2008, 2014) defends a view on which sounds are stable dispositional properties by undermining one of the main intuitive observations presented in favour of the event view: that sounds have durations. The *appearance* of sound duration, according to Kulvicki (2014), can be explained by the fact that mechanical stimuli have stable vibratory dispositions that are focused and brief. On this account, constancies involved in the mechanical stimulation of a particular object source account for our experience of sounds as extending in time. The proposal could be an interesting alternative explanation for the experienced duration of many environmental and mechanical sounds. However, it seems to me that, unlike the property view, the *event* view can nicely capture the fact that speech sounds in particular, although typically experienced in a particular voice, not only appear to have duration, but also to be interestingly self-standing individuals. The latter observation follows from speech sounds' functional significance, i.e. the linguistic information they carry. Abstract information which determines phonetic segments, i.e. building blocks of linguistic utterances, goes well beyond information about voice's stable characteristics. For this reason, I think that the voicing of a speech sound is better construed as an individual event, rather than a stable dispositional property of the material object (i.e. speaker's vocal folds). The property view might be difficult to square with our knowledge about how the auditory objects of speech perception are individuated, as explained in Sect. 3. To sum up, it is still unclear, I think, whether a version of the property view would have resources to account for the above observations.[14]

The second view in the debate is that the audible relation between sounds and their sources is *causal*. On that view, material objects produce or make sounds, whereas sounds are audible *effects* of their audible sources. The *causal* view gets several things right. It succeeds in capturing an intuitive idea that sounds and their sources are distinct and that in hearing sounds we typically become aware of what produced them, we seem to *mediately* perceive their sources (O'Callaghan 2017, p. 101). A recent, promising version of this line of explaining the relation between sounds and their sources comes from Matthew Nudds (2010). According to Nudds, in normal circumstances we typically hear sounds *as apparently having been produced* by their sources (2010, p. 118). The view seems applicable to the case of hearing speech sounds in a particular voice: a typical auditory experience of speech sounds is an experience of speech sounds as apparently having been produced by the voice. Voice can be

[14] An alternative interesting construal of the property view can be found in Leddington (2019). Leddington defends a view according to which sounds are audible properties of their *event sources*. Interestingly, he argues that the causal relation between sounds can be maintained: sounds are argued to be properties caused by events that bear them. This type of property view could perhaps account for some of the observations concerning the nature of speech sounds and voicing events. It might also be compatible with the idea that we hear articulatory events (as postulated by the motor and direct realist theories of speech perception). I leave this possibility open for further discussion. I thank an anonymous reviewer for drawing my attention to this material. For critical discussion of the event-property view see also O'Callaghan (2016).

understood here as a physical source, i.e. the speaker's vocal folds and tract, that when put into use, results in a stable pattern of audible characteristics that, in normal circumstances, can be attributed to a particular speaker.

This seems like a *prima facie* promising way of capturing the relation between speech sounds and their ordinary source, a voice. A possible obstacle for this proposal in the case of audible relation between environmental sounds and their sources is raised by O'Callaghan (2017, pp. 104–105). The view suggests that sources are presented in auditory experience in a descriptive way (*as having been produced by*), which arguably does not allow for demonstrative thought about sources of sounds. According to O'Callaghan, this makes the view implausible given that in cases of auditory experience we can typically demonstratively refer to what appears to be an ordinary sound source as, e.g. 'these footsteps', 'that bell'. Similarly, in the case of speech sounds and voices, auditory experience of speech in a particular voice typically furnishes information that allows for referring to voice demonstratively ('that voice') and without descriptive contents (*as having been produced by*). Thus, if we want our conception of how speech sounds and voices are related to allow for demonstrative thought about voices, Nudds' version of the causal view may not suffice to capture the relation.[15]

An alternative version of the causal view would be to say that we perceive not only sounds and their sources, but also the *causal relation* between them (inspired by e.g. Siegel 2010). In the case of typical auditory experience of speech sounds coming in a particular voice, we could try to explain the apparent phenomenological unity of the two in terms of perceiving the causal relation between the two. The problem for this conception is that it seems to run against the typical phenomenology of sounds and their sources. O'Callaghan observes (2017, pp. 109–110) that on this view, the proposed perceived causal relation implies that events, i.e. causal relata, are experienced as *wholly distinct*. This, according to him, runs against a typical auditory experience of sounds and their sources, where the phenomenological unity appears to involve an *audible sourcehood of sounds*, but does not present sounds and sources at distinct discrete locations and occurring at different times. The unity involved in such cases does not seem to require that sounds and sources differ in their audible appearance.[16] On that basis, O'Callaghan argues that auditory perceptual experience of a causal relation is not necessary to account for the phenomenological unity between sounds and their sources, i.e. it cannot explain the apparent sourcehood of sounds and the relation in which audible sounds stand to their audible sources (2017, p. 110).

These observations can be applied to the case of auditory experience of speech coming in a particular voice. The audible unity between speech sounds and the voice does not seem to present audible speech sounds *as being audibly caused* by the audible voice. The experienced connection between the voice and speech sounds is more intimate than the causal perception view would imply, given that a voice and speech

[15] Another worry for the causal view is that causal relation might not be coarse grained enough to allow hearers of sounds to discriminate their sources from their surroundings, which does appear to be one of the key abilities enabled by audition. Awareness as of an effect does not itself typically furnish epistemically unmediated awareness of its cause. For discussion see O'Callaghan (2017, pp. 106–108).

[16] See also Leddington's (2014) observations in support of the view that sounds and their sources appear fused or bound in auditory experience.

sounds do not seem to be individuated in our experience as two distinct audible events standing in a causal relation. One might reply to this point by observing that in the case of visual experience of causation the two relata are often experienced as separate because they typically (but not always) take place one after another, whereas in the case of auditory experience of sounds and sources a similar separation may be rare. If that were the case, couldn't we say that we hear the cause (source) as simultaneous with the effect (sounds) and in this way support the causal view? Couldn't we say that we hear the speech sounds as caused by voices just as we see, for example, a doorstop *holding a door open*?[17]

This kind of move might make the dialectic situation somewhat problematic. If the cause and effect were experienced simultaneously, then it is not clear whether the causal relation between the two would be *experienced* as well. We can of course stipulate or come to know via inference that the causal relation obtains and, in that sense, become aware of its existence, as in the case of the door and doorstop. But it is less clear whether in the simultaneous case we could perceptually experience the causal relation between the two relata. Of course, this is not a knockdown argument against the causal view and other versions of that view could be better equipped to deal with these observations (see footnote 18). Nevertheless, it is worth noting that the *simultaneity* interpretation of the causal view seems to leave something interesting out in the case of speech sounds experienced as produced in a particular voice. Neither the simultaneous co-occurrence of voicing event and speech sounds, nor our knowledge that these are causally connected seem able to account for our experience that presents speech sounds and voice as fused. I think the view that I will present now is well equipped to account for these observations.

How can we capture the unity of voice and speech sounds, as experienced, in typical cases of hearing speech sounds in a voice? A third proposal in the debate on the relation between audible sounds and their audible sources is O'Callaghan's *mereological* view (2017; see also 2011b). On that view, an audible relation between sounds and sources is mereological, i.e. it is a relation between part and a whole (O'Callaghan 2011b). Sounds are constituent parts of everyday audible events: collisions and vibrations. Sounds, which on this account are event-like individuals, are parts of bigger, encompassing events. For example, the event of the ringing of a bell involving a pattern of vibrations and collisions in environment includes as a part the ringing sound. On this view, sound sources are such encompassing events that include sounds as their parts. But the epistemic direction goes the other way round, given that auditory perceptual awareness as of the whole on this view occurs thanks to experiencing the part, i.e. the sound. On this view, sounds and sources overlap in space and time. It is in this way, according to O'Callaghan, that we can make sense of the idea that in typical cases of auditory experience we experience sounds and their sources, i.e. encompassing events, as phenomenologically united. Sound sources, i.e. events, typically involve material bodies and ordinary objects as *participants*. For example, a bell is a participant in the event of ringing that has the ringing sound as its part. The event of the ringing of a bell and the bell as a participant of that event are revealed to a listener by the audible part of the event, a sound that the bell makes.

[17] I thank an anonymous reviewer for raising this point and for interesting examples.

We can now apply the mereological view to the case of the experienced audible unity of speech sounds and voices. As defined at the beginning of this section, a *voicing* is an event of using a voice on a particular occasion, whereas a voice is a physical source determined by the speakers' vocal folds and tract resulting in a stable acoustic pattern of vocal characteristics. A voicing event involves a vibration of a speaker's vocal folds against the filter and in the case of auditory experience of speech it produces speech sounds. Following the mereological view, the event of voicing is a source of speech sounds. On this account, speech sounds are heard as united with their vocal sources because they are a part of the voicing event, as described above. The voice, i.e. a physical source determined by the speaker's vocal folds and vocal tract resulting in a set of stable vocal characteristics, is a *participant* in the event of voicing.

So described, the mereological model makes good sense of the relation between speech sounds and voices. The apparent unity is the *unity of speech sounds and the events of voicing*. The unity is captured here as a mereological relation of part and whole. The view explains why the speech sounds and their sources, the voicing events, appear to overlap in space and time resulting in the audible impression of unity. The view can also explain why in a typical auditory experience of speech a hearer can become aware of a material source of the signal—a voice. The voice is a participant in the event of voicing. Stable characteristics of a particular voice are audible in such an event. This allows the hearer to attribute a particular voicing event to a particular voice, and frequently even to a particular speaker to whom the voice belongs. The view can also accommodate the observation that speech sounds are event-like individuals caused by the happenings in the environment, a vibration of the vocal folds against a filter, a voicing event.

According to the above picture, on which the voicing event is the source of speech sounds, and given how I defined voice (i.e. as a physical object - the speaker's vocal folds and tract), neither voice, nor a speaker to whom the voice belongs are a *source* of speech sounds. One might therefore worry that on the mereological view the speaker, in particular, seems to be left out of the picture. The worry is motivated by the following intuitive observations. Typically, a voice, i.e. a physical source of stable vocal characteristics, belongs to a particular person. Such observations speak in favour of the intuitive idea that the voice of a speaker, and mediately, the speaker themselves, are normally *somehow experienced* by hearers. In line with these observations, Smith (2009) has argued for a stronger theoretical claim that when listening to speech sounds, we also listen to what and who is producing them: "A voice belongs to a person, an embodied subject who intentionally produces the sounds we hear" (p. 204). Since the heard properties of the voice often enable us to recognize who is speaking, according to Smith, recognition of the voice in speech sounds normally involves a recognition of a person, as well as the recognition of some characteristics of that person (e.g. age, gender). For Smith, hearing a voice is normally a case of hearing a person.

I believe that our pre-theoretical intuitive observations should not be dismissed: they seem to capture the fact that when listening to speech, hearers typically become aware of and gain information about speakers. However, I am less sure whether we need to capture these observations in terms of *person/speaker perception*. Instead, I will now suggest how such intuitive observations can be accommodated within the mereological view just sketched. When a voice belongs to a speaker, as it typically

happens, there is an intuitive sense in which the speaker *appears* to be a source of speech sounds. According to the mereological view, in such cases a speaker's voice, i.e. the physical source for the voicing event, is a participant in the voicing event. In ordinary situations, by hearing speech sounds, a hearer also often gains information about and becomes aware of the speaker to whom the voice belongs. This awareness is ecologically important from the hearer's perspective, since in many cases it leads to the recognition of the speaker who, by means of the voicing event, has intentionally produced the speech sounds to communicate a thought.

The above observations seem to suggest that we could treat both the physical object, i.e. the speaker's vocal tract, and the speaker themselves, a person, as *participants* in the event of voicing speech sounds. But one might wonder whether the speaker as a person is a participant in such an event in a similar manner that the physical object (their vocal folds and vocal tract)? There are clear differences here: vocal folds and vocal tract participate mechanically, as a ringing bell does, but the person does not seem to participate in that way. The answer to the question also depends on what the conditions for being a participant in an auditory event are. I have suggested that becoming an object of hearers' awareness might function as such a condition. But a reader might worry that this way of putting things makes the relation between hearers and speakers somewhat unspecified, because it leaves it open whether when becoming aware of the person speaking we *hear them* or merely *infer* their presence.[18]

For the purpose of this paper, I would like to leave it as an open possibility that in cases of hearing speech in a particular voice we might be *inferring* the presence (and characteristics) of the speaker, often a particular one. In order to explain the experienced unity between speech sounds and voices we need not postulate that we perceive the person per se. The unity on the mereological model is explained by the part-whole relation between speech sounds and voicing events. When a voice belongs to a speaker, as it typically does, the presence of the speaker can be revealed to a hearer in such voicing episodes. This is explained on the mereological account by the fact that vocal folds that participate in the voicing event normally belong to a particular body that belongs to particular person. But the revealing of the person in such cases need not go via perception. It could go via a quick *inferential* step a speaker has to make when confronted with their perceptual experience of hearing speech in a voice. Such an inference may be spontaneous, automatic and effortless. It may thus result in a particularly intimate experiential link between the perception of speech sounds in a voice and the awareness of the speaker. It is an open possibility that should not, I think, be excluded.[19]

[18] One interesting argument in favour of idea that we might also *hear* the person speaking could perhaps come from observations that humans are perceptually sensitive to intentionally directed action (e.g. Teufel et al. 2010). If hearing speech sounds in a particular voice is normally identified as an instance of intentional action, then one might perhaps argue that in the voicing event produced by the speaker's vocal tract we also hear the agent intentionally producing the sounds. However, the exact interpretation of these observations might still require further work, given that we would need to determine whether the tracking and understanding of intentional action in such cases is best explained in terms of social cognition, social perception, an inference or a mix of these. I thank an anonymous reviewer for pressing this problem and for mentioning this piece of evidence.

[19] It is an interesting question how to describe the experience of the speaker and their characteristics (e.g. age, gender, place of origin) and the underlying process. The results might point to an interesting difference

6 Hearing speech sounds and voice together: the Voice Shaping Speech model

I have argued that the mereological view, as applied in the previous section to the case of auditory experience of speech heard in a particular voice, provides a plausible account of the metaphysical underpinnings for the apparent phenomenological unity between speech sounds and voices. It construes the relation between speech sounds and their source, the voicing event, in mereological terms, as the relation between a part and the whole. The proposal provides an independent metaphysical explanation. However, it is not the whole story to be told about the experienced unity of voice and speech sounds. We also need to make sense of the experienced unity between speech and voice *at the level of processing*, i.e. in terms of psychological regularities that determine how the *auditory objects* of the experience of listening to speech in a particular voice are typically formed and individuated. It is important to see whether our metaphysical picture is compatible with our best knowledge of how auditory objects of such experiences are formed. This task is pressing also because, as presented in Sect. 4, speech sound and voice perception are commonly taken to rely on two psychological abilities (Belin et al. 2004). In order to account for the experienced unity, the connection between the two abilities should be explained in light of current best empirical knowledge. Drawing on the material presented in the first part of the paper, I will now sketch a proposal that aims to address this issue. The mereological view and the psychological model presented in this section are meant to work in tandem in explaining the nature and source of the experienced unity between speech sounds and voices.

In Sects. 2 and 3 I argued that auditory objects of speech perception and voice perception, as other auditory objects, have a *temporal structure*. Time is one of the main organising principles that determines how we individuate objects in the auditory experience of listening to speech. Are the processing facts of how such objects are typically formed compatible with the above sketched mereological view of the relation between speech sounds and their sources, i.e. voicings? In general, the mereological model is compatible with the claim that auditory objects are individuated according to psychological principles based on time and thus exhibit temporal structure. According to O'Callaghan (2008, 2016), both visual and auditory perceptual objects are mereologically complex individuals. Visual objects possess a *spatial* mereology—in human visual experience such objects are tracked and individuated thanks to their spatial features. Auditory objects have a *temporal* mereology—in human auditory experience such objects are tracked and individuated in terms of their temporal characteristics and pitch. Time determines the internal structure of auditory objects, whereas pitch normally helps to locate them in space and distinguish them from other audible individuals. According to O'Callaghan, this way of carving the mereologically complex audible individuals, i.e. sounds and their sources, is related to the functional significance of auditory perception, namely informing us about our extra-acoustic environment.

Footnote 19 continued
between ordinary auditory perception and speech perception. I thank an anonymous reviewer for comments and suggestions on this topic. Given space limitations, I leave detailed discussion of this interesting issue for another occasion.

Consider now the mereological view on speech sounds and their sources and the conception of how the auditory object of speech and voice perception are formed. As I have argued in Sect. 3, the identity of typical objects in auditory experience of speech is determined by the temporal characteristics of speech signal (e.g. the voice onset time). In that sense, the objects of speech perception, as other auditory objects, exhibit a temporal structure. Similarly, the key parameters utilised in voice perception, i.e. fundamental frequency, suggest that the objects of voice perception are individuated according to psychological principles governed by temporal characteristics of the signal. Now, according to the mereological view sketched in Sect. 5, speech sounds are parts of their sources—voicing events—and in this way they audibly reveal the voicing events that produced them. Auditory objects of hearing speech in a voice are complex mereological individuals that are individuated in our experience according to psychological principles governed by time—it is in this sense that such objects have a temporal mereology.

How is the experienced unity of speech sounds and voicing events, as explained in Sect. 5, achieved at the processing level? I will now present a model that addresses this question by pointing to specific processing regularities that are crucially involved in speech sounds and voice perception and provide a psychological source of the experienced unity. According to what I will call the *Voice Shaping Speech* model, information about speech sounds and information about voice are processed together: in particular, information about the voice signal is utilized in determining the identity and audible qualities of the auditory objects of speech sound perception. The characteristics of the voice signal, determined by the speaker's vocal folds and vocal tract, provide *stability* and *predictability* that enable listeners' perceptual system to determine the audible qualities and identity of the objects of speech perception, i.e. specific phonetic segments. The proposal is meant to explain how the two abilities to perceive speech sounds and voices are connected at the processing level in typical cases of hearing speech in a particular voice and result in the experienced unity of the two. In what follows I provide theoretical and empirical material in support of the model.[20]

Consider again the case of rotating voices in the Oresteia production, where the verses of the play are split word by word between individual voices of the Chorus. In this case, we can perceive speech sounds and understand verses they form. At the same time when hearing individual voices pronouncing specific words, we experience a particular kind of *discontinuity* in the sources of speech sounds—the events of voicing. Consider now a parallel case where instead of splitting the verses into specific words pronounced by individual voices, it is specific words that are being split into meaningful phonetic units, i.e. phonetic segments corresponding to the speech signal. As a result, each phonetic segment is heard as pronounced in a different voice. Arguably, such a manipulation would greatly increase the experienced discontinuity and would obstruct the perception of speech sounds. It is also likely that the experienced discontinuity and difficulty in perceiving speech sounds would be even greater in the case where *two* different streams of verses were actually spoken out in that peculiar manner *in parallel* and where in hearing both of them the hearer's goal was to make sense of only one of

[20] I thank two anonymous reviewers of this journal for particularly helpful comments and suggestions for how to develop this positive proposal.

them. These imaginary cases draw our attention to a particular, unappreciated *stability* and inherent structure that voices provide in a typical experience of listening to speech. Hearing speech sounds in a particular voice ensures a certain level of *predictability*. These and the following observations provide a theoretical foundation for the Voice Shaping Speech model. I now explain how the conception of temporally structured auditory objects can be put into use to explain the experienced unity of speech and voice.

When investigating how auditory objects are formed by the auditory system in cases of hearing environmental sounds, Nudds (2010, pp. 114–115) discusses the role of sound sources more generally. He observes that, according to our current best empirical knowledge, there are relationships between frequency components produced by the same source that are unlikely to exist between components coming from different sources. For example, components produced by the same source tend to be harmonically related. Moreover, soundwaves produced by a single event will have frequency components with the same *temporal* characteristics: all components will begin at the same time, they are likely to be in phase with one another and change over time in similar ways with respect to their amplitude and frequency (2010, p. 114).[21] This is very unlike components produced by distinct sources. The auditory system exploits this fact: when it detects the above mentioned relationships between components it groups them together as belonging to one auditory object and as produced by the same source.

According to the Voice Shaping Speech model, this general feature of the auditory system applies to the case of experience of hearing speech in a particular voice. The model explains how it is realised in this case at the processing level. The auditory system exploits *harmonic* and *temporal* relationships between frequency components characteristic for soundwaves coming from a single source, a particular event of voicing, in grouping these components as auditory objects of speech sounds perception. At the level of processing, information about the voice signal is utilised in determining the identity of the auditory objects of speech sound perception (phonetic segments). Temporal characteristics of voice signal are deployed by the hearer's auditory system in the process of identifying speech sounds uttered.

In order to explain which specific audible features of a voice enable the perception of speech sounds and how they do so, I now present recent empirical evidence that supports The Voice Shaping Speech model. The presented material explains how the two psychological abilities to perceive speech sounds and voices, as described in Sect. 4, underlie *joint* processing regularities that give rise to experiences of phenomenologically united speech sounds and voice. In the empirical work on speech perception the conventional view has been that since speech sound recognition and speaker identification based on voice perception rely on two separate functions, they are implemented by two independently operating neural mechanisms. Speech sounds processing takes place predominantly in the left hemisphere, whereas processing of voice-specific information is located in the right hemisphere (von Kriegstein et al. 2010).

[21] These observations concern basic processes of auditory object formation and as such they do not fall prey to O'Callaghan's (2017) criticism of the version of the causal view defended by Nudds in the same paper, which I have discussed in Sect. 5.

This conventional view is now under pressure from a series of recent studies on speech sound and voice perception. Some earlier findings from behavioural (Cutler et al. 2010; Magnuson and Nusbaum 2007; Nygaard 2005) and neuroimaging studies (e.g. Kaganovich et al. 2006) suggested that there may be important interdependencies between speech sounds perception and voice perception (Laing et al. 2012). But it is only recently that the idea has been further pursued in the context of speech perception in a *multiple-speaker environment*. Notably, humans are particularly expert in understanding speech coming from different speakers, whereas artificial speech recognition systems still struggle with this task (e.g. Scharenbrog 2007). Recent neuroimaging studies on speech perception in the simulated multi-speaker environment suggest that this unique ability is possible due to the role that different voice characteristics play in the process of incremental construction of the representations of speech sounds in the hearers (von Kriegstein et al. 2010).

Vocal tract length is one important cue in voice recognition (Lavner et al. 2000), and regions responding to this information are commonly taken to be involved in recognising other humans by voice (Belin et al. 2004; von Kriegstein and Giraud 2004). In their functional magnetic resonance (fMRI) study, von Kriegstein et al. (2010) used resynthesized speech stimuli to evoke speaker changes by variations in *vocal tract length parameters*. The study provided neuroimaging evidence for a dynamic speech-processing network that seems to undermine the conventional view that speech sound perception and voice are processed by two largely independent brain networks. In particular, von Kriegstein et al. found that speech recognition regions in *left* posterior superior temporal gyrus/superior temporal sulcus (STG/STS) are also involved in the encoding of the speaker-related vocal tract parameters. It was also shown that the *right* posterior STG/STS regions were activated specifically more to a speaker-related vocal tract parameter change during a speech recognition task than in a voice recognition task. The results suggest that the left and right posterior STG/STS are functionally connected and jointly responsible for processing of speech and voice specific parameters of the signal. According to the authors the results support a *network account of speech recognition* in which information about the speech formants and the perceived parameters of the speaker's vocal tract are jointly utilized to solve the difficult task of understanding speech from different speakers.

The above results led to a proposal for a potential mechanism responsible for the processing of speaker-related vocal tract variability in speech recognition. Von Kriegstein et al. argue that the speech perceptual system operates in a *predictive* manner (e.g. Kiebel et al. 2008; von Kriegstein and Giraud 2006; Overath et al. 2007). They support this claim with observations that information about the relatively constant vocal tract length of a speaker helps the hearer, among other constraints, to identify possible formant positions that determine the phonemes produced by that speaker. On their view, predictions of speech trajectories involve information about perceived uncertainty about voice and speech specific parameters. The prediction mechanism based on information about voice characteristics is particularly helpful in a multi-speaker environment, but it is also useful in everyday conversational situations (von Kriegstein et al. 2010). This view can also accommodate findings that speech produced in the same voice and normally belonging to the same speaker is usually more intelligible to hearers than speech from changing speakers who have different voices (e.g. Pisoni

and Levi 2007). By showing that a specific feature of a voice, i.e. vocal tract length, takes part in identifying speech sounds, the above results point to relevant specific processing regularities in speech and voice perception and lend support to the Voice Shaping Speech model.

According to this recent strand of research, the vocal length tract is just one among many voice-related parameters that are used to adjust internal models responsible for predicting and computing the speech signal.[22] A following study by Kreitewolf et al. (2014) showed an interaction similar to the one established above between perceived vocal tract parameters and speech sounds, but for *glottal fold* parameters. Glottal fold parameters are the other main acoustic feature crucially involved in voice recognition and in the perception of speech prosody. Pitch, in particular, is determined by the glottal pulse rate. In their 2014 study Kreitewolf's et al. modified pitch in order to induce changes in the perception of speaker's voice and in the prosodic features of speech signal. The study provided neuroimaging evidence for, among other, increased functional connectivity between right and left Heschl's gyri during recognition of linguistic prosody when speakers differed in glottal fold parameters. The authors argued that these findings cannot be easily reconciled with the conventional view that speech-specific and voice-specific parameters of the signal are processed independently. Instead, they provide further support for the view according to which the processing of speech and voice parameters interact at the neural and behavioural level (von Kriegstein et al. 2010; Nygaard 2005). These results show that the processing of glottal fold helps the task of recognizing speech sounds from a particular speaker.[23]

The empirical observations presented above strongly suggest that the processing of voice-specific characteristics reflected in speech signal, such as the vocal tract length and glottal folds parameters, is utilised in identifying the auditory objects of speech perception. The mechanism involved in that process is argued to operate in a predictive manner. Although the above studies focus primarily on speech sound perception in a multi-speaker auditory environment, the results are taken to reveal mechanisms operating also in a one speaker situation (von Kriegstein et al. 2010). Hearers exploit information about voice-specific characteristics of the signal in order to make predictions and identify speech sounds produced in a particular voice. By pointing to systematic processing regularities in speech sound and voice perception, these results provide empirical support for the Voice Shaping Speech model.

The model is compatible with recent studies on the auditory scene analysis that investigate psychological mechanisms responsible for producing different auditory streams and look into how they are implemented in the brain (Griffiths and Warren 2004; Ding and Simon 2012; Bizley and Cohen 2013). It has been observed that at the processing level speech signal is represented in the brain *as coming from a particular source* and *as locked to time characteristics* of the voice that produces it. Mesgarani and Chang (2012) provide evidence that when listening to more than one speaker

[22] Other parameters could be: speaking rate, visual information from the face of a speaker, and various types of social information about the speaker (e.g. accent).

[23] A recent study by Zhang et al. (2016) provides complementary evidence for the role of pitch in identifying speech sounds in the multi-speaker auditory environment in *tone* languages, thereby expanding evidence to many non-European languages, where differences in tone change the meaning of a word despite the fact that the word preserves the same pronunciation.

the spectral and temporal features are represented as if hearers were listening to that speaker alone.

Drawing on the above presented empirical and theoretical evidence, the Voice Shaping Speech model explains how information about the voice and speech sounds is jointly utilised in identifying the auditory objects of such experience and results in the phenomenological unity of the two. On this view, the auditory objects of such experiences are produced by an interacting dynamic network where stable voice characteristics enable and shape the perception of speech sounds. The identity and audible qualities of auditory objects of speech perception depend on stable voice characteristics exercised in a particular event of voicing.[24] The latter ones are retrieved on the basis of *time*-dependent characteristics of the acoustic signal, such as voice onset time, pitch, fundamental frequency, timbre or shimmer. So conceived, the objects of speech perception are no different from other auditory objects in that they have a temporal mereology.[25]

7 Concluding remarks

Although speakers can use their voices to produce a whole variety of sounds, we commonly hear voices when they are used to produce speech. In such typical experiences of listening to speech in a voice, speech sounds uttered and the voice appear united. In this paper I have proposed a metaphysical and psychological explanation for this apparent unity. First, I have argued that we can make sense of the phenomenological unity between speech sounds and their sources in terms of the *mereological* view (O'Callaghan 2017). On that proposal, the intimate audible relation between speech sounds and their sources is a part-whole relation between speech sounds and voicing events that overlap in time and space. The voice, i.e. a physical object, is a participant in the voicing event. When a voice belongs to a speaker, as it typically does, hearers can also become aware of his or her presence when listening to a voicing event. Second, I have sketched the Voice Shaping Speech model to explain the psychological source of the experienced unity between voice and speech. The explanation draws on our current best empirical knowledge of how the auditory objects of such experiences are formed on the basis of the temporal characteristics of the signal. The mereological view and the Voice Shaping Speech model work in tandem in explaining the sense in which auditory objects of speech perception exhibit a temporal mereology.

[24] A reviewer for this journal suggests that the Voice Shaping Speech Model might be compatible with observations made in Young (2018), where it is argued that to hear *source events* of sounds is to hear a physical object as extending through time and as having a certain temporal shape. In the present case, the idea would be, I think, that to hear a voicing event is to hear vocal folds as extending through time and having a certain temporal shape. Given space limitations, I leave discussion of the details of this proposal for another occasion.

[25] The Voice Shaping Speech model is compatible with the general psychological accounts of auditory perception that explain how auditory objects are individuated (e.g. Bregman 1990; Nudds 2010). It goes beyond such accounts, because: (1) it concerns a particularly complicated case of speech perception, and (2) draws on recent empirical research that points to specific processing regularities in speech sound and voice perception.

An important feature of the proposal developed in this paper is that it helps to bring back the voice into overall conception of auditory experiences of speech perception. The speech signal provides relatively abstract information about the phonetic structure of an utterance that is retrieved by the hearer's speech perception mechanisms.[26] Explaining how this information can be retrieved from speech signal is a daunting empirical task, which might explain why speech sound perception has so far received more attention than voice perception. But we should not forget about the voice. Information about vocal characteristics, as I argued here, takes part in structuring the objects of auditory speech experience.[27] Voice is also important for social interactions based on linguistic communication. Empirical research suggests that paralinguistic cues retrieved from speakers' voices about, for example, accent and pronunciation or perceived gender, often bias hearers and affect how much credence they give to speakers (e.g. Lev-Ari and Keysar 2010; Hosoda and Stone-Romero 2010; Rakić et al. 2011). Various possibly good and bad ways in which we are sensitive to speakers' voices call for further empirical and theoretical investigation (e.g. Lev-Ari 2015).

Acknowledgements Open Access funding provided by University of Oslo. I would like to thank Jen Hornsby, Barry Smith, audiences at the University of Oslo and Department of Philosophy at Birkbeck College (UoL), as well as the anonymous reviewers for this journal for extremely helpful comments on previous versions of the paper and/or discussions on the topic.

Funding This work was supported by and developed as part of the Mobility Grant Fellowship Programme (FRICON) funded by The Research Council of Norway and the Marie Skłodowska-Curie Actions (Project Number: 275251).

References

Adank, P., & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *The Journal of the Acoustical Society of America, 126*(5), 2649–2659.

Assal, G., Buttet, J., & Jolivet, R. (1981). Dissociations in aphasia: A case report. *Brain and Language, 13*(2), 223–240.

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

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research PRPF, 74*(1), 110.

[26] Although this observation may be perhaps seen as setting apart the case of speech perception from the case of auditory perception of non-linguistic signal, this is far from saying that speech is special among other objects of auditory perception. For a critical discussion of such a claim see O'Callaghan (2015).

[27] An interesting comparison worth further investigation is that between *voice* as a source of stable vocal characteristics and of speech and *face* as pattern or feature of a body giving rise to emotional expressions. Given limited space, I leave a detailed discussion of the analogy between voice and face for another occasion.

Bayne, T. (2009). Perception and the reach of phenomenal content. *Philosophical Quarterly, 59*(236), 385–404.
Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129–135.
Bendixen, A. (2014). Predictability effects in auditory scene analysis: A review. *Frontiers in Neuroscience, 8,* 60.
Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience, 14*(10), 693.
Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge: MIT Press.
Brogaard, B. (2018). Defense of hearing meanings. *Synthese, 195*(7), 2967–2983.
Burton, R. L. (2015). The elements of music: what are they, and who cares? In J. Rosevear & S. Harding (Eds.), *ASME XXth national conference proceedings*.
Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*(12), 535–543.
Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory Phonology, 10,* 91–111.
De Cheveigne, A. (2005). Pitch perception models. In C. J. Plack, A. J. Oxenham, & R. R. Fay (Eds.), *Pitch: Neural coding and perception* (pp. 169–233). New York: Springer.
Di Bona, E. (2017). Towards a rich view of auditory experience. *Philosophical Studies, 174*(11), 2629–2643.
Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, 109*(29), 11854–11859.
Drożdżowicz, A. (2019). Do we hear meanings? Between perception and cognition. *Inquiry*. https://doi.org/10.1080/0020174X.2019.1612774.
Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron, 61*(2), 317–329.
Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Sciences, 4*(7), 258–267.
Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics, 36*(4), 359–368.
Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49*(3), 396–413.
Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*(3), 361–377.
Ghazanfar, A. A., & Rendall, D. (2008). Evolution of human vocal production. *Current Biology, 18*(11), R457–R460.
Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature, 462*(7272), 502.
Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*(4), 511.
Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience, 5*(11), 887.
Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences, 4*(4), 131–138.
Holt, L. L., & Lotto, A. J. (2008). Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science, 17*(1), 42–46.
Hosoda, M., & Stone-Romero, E. (2010). The effects of foreign accents on employment-related decisions. *Journal of Managerial Psychology, 25*(2), 113–132.
Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research, 1114*(1), 161–172.
Kent, R. D. (1977). Coarticulation in recent speech production. *Journal of Phonetics, 5*(1), 15–133.
Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4*(11), e1000209.
Kreitewolf, J., Gaudrain, E., & von Kriegstein, K. (2014). A neural mechanism for recognizing speech spoken by different speakers. *Neuroimage, 91,* 375–385.
von Kriegstein, K., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage, 22*(2), 948–955.

Kulvicki, J. (2008). The nature of noise. *Philosophers' Imprint, 8*(11), 1–16.
Kulvicki, J. (2014). Sound stimulants. In Dustin Stokes, Stephen Biggs, & Mohan Matthen (Eds.), *Perception and its modalities* (pp. 205–221). New York: Oxford University Press.
Laing, E. J., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in psychology, 3,* 203.
Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology, 21*(4), R143–R145.
Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication, 30*(1), 9–26.
Leddington, J. (2014). What we hear. In R. Brown (Ed.), *Consciousness inside and out: Phenomenology, neuroscience, and the nature of experience* (pp. 321–334). Dordrecht: Springer.
Leddington, J. P. (2019). Sounds fully simplified. *Analysis, 79*(4), 621–629.
Lev-Ari, S. (2015). Comprehending non-native speakers: Theory and evidence for adjustment in manner of processing. *Frontiers in psychology, 5,* 1546.
Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology, 46*(6), 1093–1096.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431.
Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences, 103*(49), 18866–18869.
Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human perception and performance*, *33*(2), 391.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748.
Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature, 485*(7397), 233.
Micheyl, C., Carlyon, R. P., Shtyrov, Y., Hauk, O., Dodson, T., & Pullvermüller, F. (2003). The neurophysiological basis of the auditory continuity illusion: A mismatch negativity study. *Journal of Cognitive Neuroscience, 15*(5), 747–758.
Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology, 9*(4), 399–406.
Mulac, A., & Giles, H. (1996). 'You're Only As Old As You Sound': Perceived vocal age and social meanings. *Health Communication, 8*(3), 199–215.
Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America, 85*(5), 2088–2113.
Neuner, F., & Schweinberger, S. R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and Cognition, 44*(3), 342–366.
Nudds, M. (2010). What are auditory objects? *Review of Philosophy and Psychology, 1*(1), 105–122.
Nygaard, L. C. (2005). *Linguistic and paralinguistic factors in speech perception. Handbook of speech perception*. Oxford: Blackwell Publishers.
Nygaard, L. C., & Pisoni, D. B. (1995). Speech perception: New directions in research and theory. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition: Speech, language, and communication*. San Diego: Academic Press.
O'Callaghan, C. (2008). Object perception: Vision and audition. *Philosophy Compass, 3*(4), 803–829.
O'Callaghan, C. (2011a). Against hearing meanings. *The Philosophical Quarterly, 61*(245), 783–807.
O'Callaghan, C. (2011b). XIII—Hearing properties, effects or parts? In *Proceedings of the Aristotelian Society (Hardback)* (Vol. 111, No. 3pt3, pp. 375–405). Oxford: Blackwell Publishing Ltd.
O'Callaghan, C. (2015). Speech perception. In M. Matthen (Ed.), *Handbook of the philosophy of perception* (pp. 475–494). Oxford: Oxford University Press.
O'Callaghan, C. (2016). Auditory perception. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 Edition). https://plato.stanford.edu/archives/win2016/entries/perception-auditory/.
O'Callaghan, C. (2017). *Beyond vision: Philosophical essays*. New York: Oxford University Press.
Overath, T., Cusack, R., Kumar, S., Von Kriegstein, K., Warren, J. D., Grube, M., et al. (2007). An information theoretic characterisation of auditory encoding. *PLoS Biology, 5*(11), e288.
Owren, M. J., Berkowitz, M., & Bachorowski, J. A. (2007). Listeners judge talker sex more efficiently from male than from female vowels. *Perception and Psychophysics, 69*(6), 930–941.

Pasnau, R. (1999). What is sound? *The Philosophical Quarterly, 49*(196), 309–324.
Pisoni, D. B., & Levi, S. V. (2007). Some observations on representations and representational specificity in speech perception and spoken word recognition. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 3–18). Oxford University Press.
Pisoni, D. B., & Remez, R. E. (Eds.). (2005). *The handbook of speech perception*. Oxford: Blackwell.
Plack, C. J. (2014). *The sense of hearing*. New York: Psychology Press Ltd.
Plack, C. J., & Carlyon, R. P. (1995). Loudness perception and intensity coding. In B. C. J. Moore (Ed.), *Handbook of perception and cognition* (2nd ed., pp. 123–160). Hearing.
Rakić, T., Steffens, M. C., & Mummendey, A. (2011). Blinded by the accent! The minor role of looks in ethnic categorization. *Journal of Personality and Social Psychology, 100*(1), 16.
Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*(4497), 947–949.
Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication, 49*(5), 336–347.
Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(1), 15–25.
Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences, 26*(2), 100–107.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303–304.
Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182–186.
Siegel, S. (2010). *The contents of perception*. New York: OUP.
Skrzypulec, B. (2018). Visual endurance and auditory perdurance. *Erkenntnis*, *85*, 467–488.
Smith, B. C. (2009). Speech sounds and the direct meeting of minds. In M. Nudds & C. O'Callaghan (Eds.), *Sounds: New essays in perception*. London: Oxford University Press.
Spencer, C. (1999). An inexhaustible masterpiece is transformed into a glib anti-war morality play. *The Telegraph*. https://www.telegraph.co.uk/culture/4719184/An-inexhaustible-masterpiece-is-transformed-into-a-glib-anti-war-morality-play.html.
Stevens, K. N., & Klatt, D. H. (1974). Current models of sound sources for speech. In *Ventilatory and phonatory control systems: and international symposium*. New York: Oxford University Press.
Teufel, C., Fletcher, P. C., & Davis, G. (2010). Seeing other minds: attributed mental states influence perception. *Trends in Cognitive Sciences, 14*(8), 376–382.
von Kriegstein, K., Smith, D. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience, 30*(2), 629–638.
Winkler, I., Denham, S., Mill, R., Bőhm, T. M., & Bendixen, A. (2012). Multistability in auditory stream segregation: A predictive coding view. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1591), 1001–1012.
Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science, 24*(5), 603–612.
Young, N. (2018). Hearing objects and events. *Philosophical Studies, 175*(11), 2931–2950.
Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., et al. (2016). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *NeuroImage, 124,* 536–549.