Overview

Psychology of auditory perception

Andrew Lotto[1]* and Lori Holt[2]

Audition is often treated as a 'secondary' sensory system behind vision in the study of cognitive science. In this review, we focus on three seemingly simple perceptual tasks to demonstrate the complexity of perceptual–cognitive processing involved in everyday audition. After providing a short overview of the characteristics of sound and their neural encoding, we present a description of the perceptual task of *segregating* multiple sound events that are mixed together in the signal reaching the ears. Then, we discuss the ability to localize the sound source in the environment. Finally, we provide some data and theory on how listeners *categorize* complex sounds, such as speech. In particular, we present research on how listeners weigh multiple acoustic cues in making a categorization decision. One conclusion of this review is that it is time for auditory cognitive science to be developed to match what has been done in vision in order for us to better understand how humans communicate with speech and music. © 2010 John Wiley & Sons, Ltd. *WIREs Cogn Sci* 2010

DOI: 10.1002/wcs.123

INTRODUCTION

Imagine sitting on a pier with a friend over a swimming beach where the swimmers are out of your direct view. Your friend offers you this challenge: from the movement of the water waves that you see below, can you tell how many swimmers there are? Where are these swimmers relative to each other? And what kind of stroke is each one doing? The offer of such a seemingly impossible challenge would probably lead you to question whether you needed a better class of friends. However, our auditory system performs similarly improbable feats every day.

Events in the environment can lead to perturbations of the air (changes in air pressure). When there are multiple events occurring at the same time, the disturbances of the air are summed, much like the mingling of waves from multiple swimmers. As a listener, we can use the sound waves that result from this summation to determine how many events occurred, whether the events that occurred are relative to each other, and exactly what are those events. If your mischievous friend challenged you to close your eyes and listen while a whistle blew, a person spoke and a dog barked and asked you to name the number, location, and identity of these sound sources—you would find the challenge much less intimidating (though you may still question your quality of friend). The ability of the auditory system to segregate, locate, and categorize events in the environment is a remarkable accomplishment given the complexity and transient nature of sound waves. A great deal of cognitive–perceptual processing must be involved in even the most basic auditory tasks in real-world environments. Although our understanding of auditory processing of complex sounds is well behind our understanding of visual processing of complex images, there has been substantial progress in auditory cognitive science. In this article, we provide some basics on sound and its neural encoding by the peripheral auditory system and then use the three tasks of segregation, localization, and categorization from the well-worn swimmer analogy to provide a brief overview of some classic problems in auditory perception along with some exciting new emerging research questions.

AUDITORY COGNITIVE SCIENCE

When compared to vision, audition seems like a rather unreliable perceptual system. At any point in time, the entire retina can be exposed to structured light and potentially result in activation of any of the approximately 100 million photoreceptors. However,

*Correspondence to: alotto@email.arizona.edu

[1] Department of Speech, Language, and Hearing Sciences, Tucson, AZ, USA

[2] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

DOI: 10.1002/wcs.123

the auditory system receives at most two inputs, the relative air pressure at each ear, at a single time point. And to complicate matters further, most sound events are relatively brief or transient. Although a visual object may be scanned across time from different angles, one usually only gets a single 'glimpse' of a sound event. In vision, if two objects are in the same 'line of sight', one will partially or totally occlude the other for the viewer. In audition, the perturbations of the air from two events in the same 'line of hearing' will mingle together in the sound wave reaching the listener. This makes the problems of segmentation and localization of events seemingly more complex than the task of segmenting and localizing objects in a visual scene. Even if segmentation is accomplished, the categorization of a sound event appears more difficult than visual object recognition. Sounds are rarely distinguishable by the mere presence or absence of discrete features. Instead, different sound events are typically distinguished by their values on multiple continuous acoustic dimensions. For example, the perceived difference between a clarinet and a flute playing the same note is not the result of the listener detecting a discrete feature but by the recognition of specific patterns across complex acoustic dimensions (such as the relative amplitude of harmonics and the duration/slope of intensity increase at note onset—these terms will have more meaning after we briefly discuss acoustics below). Similarly, one cannot distinguish the initial consonant sound in *bear* from *pear* by detecting a single invariably present 'b' feature. In fact, Lisker[1] cataloged 16 different acoustic dimensions that could provide information to a listener about whether a word started with 'b' or 'p'.

The preceding litany of complexities in audition would seem to render the sense unsuitable for gathering information about the world. However, despite these 'problems', we use sound as the basis for our dominant form of communication—speech—and for one of our major forms of artistic expression—music. It is clear that for audition to be so useful, a great deal of perceptual–cognitive processing must occur. Given the temporal nature of hearing and the fact that sounds are transient, memory must be important for audition. Given that sounds from different sources are intermixed in the signal and that we are often interested in a single source (such as a talker with whom we are conversing at a party), attention must be important for audition. Given the complexity of the acoustic variables that distinguish sound events and the need to classify them (especially for communication), categorization and pattern recognition must be important for audition. That is, the study of audition requires a cognitive science framework. Yet, auditory cognitive science is much less developed than visual cognitive science. There has been a great deal of excellent work on the encoding of simple sounds (such as tones and noise) in the peripheral auditory system and on the perception of pitch and loudness. However, there has traditionally been much less focus on the encoding and perception of complex sounds and on the roles of memory, attention, and categorization. For us to gain an understanding of how humans perceive speech, music, and other complex acoustic signals, auditory cognitive science will need to become a more vibrant and productive field.[2] Later, we outline some general issues in auditory cognitive science using the three specific tasks of segmentation, localization, and categorization. However, first, we set down some of the basics of acoustics and neural encoding of simple sounds.

BASIC ACOUSTICS AND PERCEPTION

To begin, it is necessary to review some of the foundations of acoustics/audition (for a more detailed description, we recommend Ref 3). When an object moves or vibrates, it disturbs air (or other media) causing the density of molecules to fluctuate. Such fluctuations in air pressure are the basis of sound waves. Just like the ripples in our hypothetical ocean carry information about the nature of the splashing and swimming, the structure of sound waves carries information about the event that disturbed the air. These principles are demonstrated by the simplest sound in auditory science, the sine-wave tone. Figure 1 illustrates that three characteristics define a sine-wave sound. *Frequency* relates to the rate at which the object vibrates (and, thus, the rate at which it affects air pressure) and is measured in Hertz (Hz), the number of times the pressure changes from high to low in one second. The range of human hearing spans approximately 20–20,000 Hz (although the upper limit decreases markedly with age), defining the human region of sound in the same way we define visible light in the electromagnetic spectrum. Changes in frequency of a sine wave are perceived as changes in *pitch*. That is, pitch is the perceptual representation of the rate of vibration of a sound-producing event. Variations in the magnitude of air pressure disturbances are characterized by the sounds *amplitude* (or *intensity*, if one is measuring power). Amplitude is represented in logarithmic units called decibels (dB). The decibel is a measure of the ratio between two sounds—the sound you are measuring and a standard level, which is typically the amplitude at which a normal hearing individual can detect a 1000-Hz tone. Note that because decibels are calculated as a ratio, 0 dB does

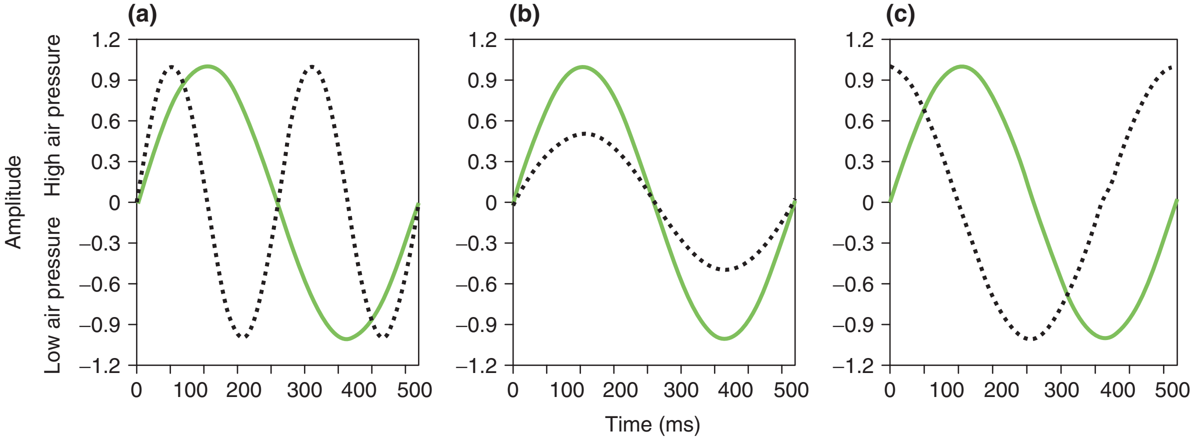

FIGURE 1 | The three characteristics that define a sine-wave sound. Each of the three graphs shows two sounds' air pressure changes as a function of time. (a) The two sounds differ in *frequency*, with the sound illustrated by the solid line cycling between periods of higher and lower air pressure at a lower rate, or frequency, than the sound shown by the dotted line. The two sounds in (b) have the same frequency, but differ in *amplitude;* the sound illustrated by the solid line elicits greater changes in air pressure and has greater amplitude. The two sounds in (c) have the same frequency and amplitude, but differ in *phase*.

not mean no sound—it means the measured amplitude is equivalent to the standard. The perceptual correlate of amplitude is *loudness*. That is, loudness is the perceptual representation of the extent of the movement of the original sound event. The dynamic range of humans (from the lowest detectable sound to pain threshold) is approximately 140 dB (or an incredible 14 orders of magnitude in linear amplitude measures). The final acoustic attribute that defines the sine-wave tone is its starting *phase*—the point in the sine cycle at which the wave starts. Phase is measured in radians or degrees, with 360° or 2π radians for one full cycle of the sine wave. For individual sounds, changes in starting phase do not result in noticeable perceptual changes. This does not mean that phase is not encoded by the auditory system, as phase *differences* between the sounds reaching each ear result in differences in perceived location of the sound source.

The sine wave is a simple sound example that nicely demonstrates sound acoustics. However, alone it is not very ecologically valid as it exists only in the acoustician's laboratory or as a consequence of electronic synthesis. Its power and prevalence in auditory sciences comes when sine waves are combined. Sine waves can be thought of like building blocks of the auditory world. Just as the simple plastic building block can be organized and configured to build fantastically detailed replicas of cities, landscapes, or machines, sine waves can be added (with the proper characteristics of frequency, phase, and amplitude) to create more complex sounds. In fact, *Fourier's theorem* assures that with a sufficient number of sine waves, *any* sound can be created. The flip side of this is that more complex sounds such as speech and music can be broken down into their constituent sine-wave tones using the mathematical techniques of Fourier analysis.

Although it is helpful to represent complex sounds in terms of combinations of simpler sine waves, the mapping between acoustics and perception is not as easily derived from the perceptual consequences of the component sine waves. Figure 2 shows line spectra for a flute and clarinet playing the same note. The lines represent the amplitude and frequency of sine waves that would create each sound (if the sound was stable in time). Each sound is composed of a *fundamental frequency* ($f0$) and *harmonics* at multiples of the fundamental frequency. When these harmonics (including $f0$) are added, listeners do not hear a number of simultaneous pitches corresponding to each sine wave. Instead, the listener hears one sound with a pitch that matches that of the fundamental frequency. Hence, the two sounds presented in the figure will have the same pitch and, thus, be perceived as the same musical note. If this is true, how do we distinguish between instruments playing the same note? The answer is that *timbre* or quality of the sound is a function of the amplitudes of the harmonics. (The term *timbre* is often used to speak broadly about differences in sound not covered by pitch, loudness, or location and may, therefore, be related to other acoustic variables such as 'attack' and 'decay'—how rapidly the amplitude increases or decreases at the beginning and end of a sound, respectively). The patterns of relative amplitudes differ for the clarinet and flute because of the resonant characteristics of the instruments. Thus, the melody of a song is carried by changes in the fundamental frequency, whereas

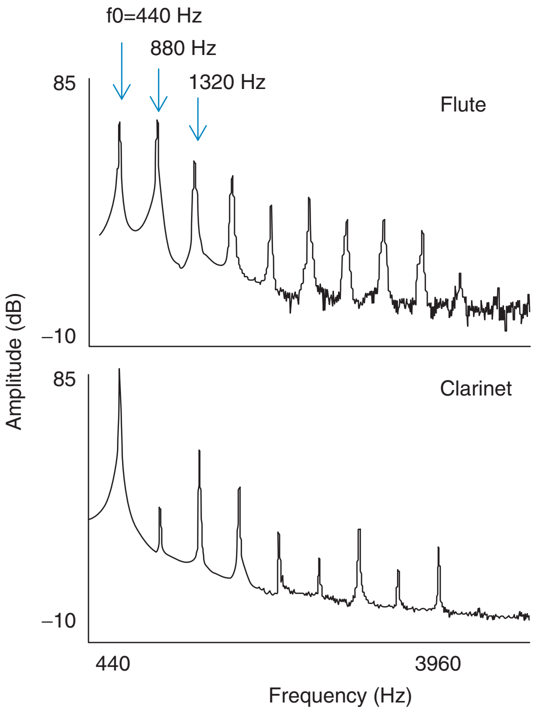

FIGURE 2 | Spectra for a flute and clarinet playing the same note. The lines represent the amplitude and frequency of sine waves that would create each sound (if the sound were stable in time). Each sound is composed of a *fundamental frequency* (*f*0) and *harmonics* at multiples of the fundamental frequency, highlighted by the arrows in the graph. Note that there are differences in the relative amplitude of different harmonics for flute *versus* clarinet. This contributes to the *timbre*, or quality, of the sound and differentiates the sounds of a flute and a clarinet playing the same note.

our identification of the instruments playing the melody is a result of the amplitude patterns across the harmonics. One may wonder whether one could devise a system like music in which variation in timbre was more important than variation in pitch. We do have such a system—it is speech. Vowel sounds are differentiated in large part to differences in timbre—they vary in the relative amplitude pattern across the harmonics due to the different resonances that result from changes in the shape of our vocal tract. Thus, speech production is much like a musician playing the same note but changing the instruments on which it is played. (Voice pitch does play a role in speech communication as well, especially for tone languages such as Mandarin Chinese, but changes in timbre carry most of the linguistic information).

The description of the mapping of acoustics and perception above is a good starting point for understanding audition, but it is still a gross oversimplification when one examines more complex sounds. Even the basis for the perception of the pitch of complex sounds has not been satisfactorily resolved.[4]

NEURAL ENCODING OF SOUND

For humans, perception begins with sound pressure waves entering the outer ear, or pinna, and setting the delicate tympanic membrane (eardrum) into vibration in the middle ear, the movement of which is transferred to three tiny bones that are attached (the ossicles), which push up against the end of the fluid-filled inner ear (the cochlea), setting up a wave that displaces a flexible structure called the basilar membrane. Like a swimming flipper, the basilar membrane is wide and floppy at the apex and narrower and stiffer at the base. These physical properties influence how it is displaced by sound; higher frequencies vibrate the stiffer base to a greater extent than do lower frequencies, creating a place code along the basilar membrane such that different locations are maximally displaced by different sound frequencies. The result of this place code is that the cochlea acts much like a Fourier analyzer, performing a spectral analysis by separating out the frequency components of complex sounds. Auditory receptors called inner hair cells reside on the basilar membrane, each with hairlike stereocilia. As the basilar membrane is displaced by the wave, the stereocilia are bent, setting in place a cascade of chemical events that translates the mechanical energy into a neural code along the auditory nerve. Outer hair cells, another type of auditory receptor, are also present along the basilar membrane. These cells appear to act as miniature motors that amplify the movement of the basilar membrane for low-intensity sounds by actively changing their shape, which alters the movement of the basilar membrane and causes the hairs on the inner hair cell to bend more, thereby improving transduction to the auditory nerve. This active *cochlear amplifier* increases sensitivity to low-amplitude sounds but does not affect large-amplitude sounds, thereby increasing the effective dynamic range for the auditory system. This sensitivity boost is frequency specific as the outer hair cell amplifier occurs only at the place along the basilar membrane tuned to the incoming frequency.

Because different inner hair cells communicate with different auditory nerve fibers, the frequency-place code of the basilar membrane is maintained in the eighth nerve and, in fact, remains pervasive in auditory processing through cortical levels. Stimulus frequency is also encoded in the temporal pattern of neural firing with neurons tending to fire at the same place in a sine wave's cycle, a phenomenon known as *phase locking*. Stimulus intensity is encoded by the number of active fibers and the firing rates of those fibers. The neural code at the auditory nerve is actually much more complicated than these simple principles suggest, especially when one looks at complex signals such as speech and music. Auditory nerve fibers exhibit adaptation and their temporal patterns can be dominated by component frequencies that are well removed from the frequency to which that neuron normally is tuned (*synchrony capture*[5]). Thus, for complex signals, there is quite a bit of neural processing and stimulus component interaction occurring even at the level of the auditory nerve. Substantial processing and feature extraction occurs all along the auditory pathway, which is more complex than the visual pathway with significant subcortical processing and interactions between the signal from both ears.

SEGMENTATION

As difficult as it has been to characterize the encoding and processing of single complex sounds, determining how the auditory system manages to accommodate multiple concurrent sound events is much more challenging. The sound wave reaching the ear includes the superimposed effects of multiple sound events. That sound wave will be deconstructed into frequency components by the cochlea, which leaves the listener with two daunting tasks—to segregate the components coming from the independent sound events and to integrate the components that belong to the same event. Normally, we perform these tasks with ease. Cherry[6] wondered how we could so readily listen to someone with whom we were engaged in conversation even when surrounded by many other conversations, which he labeled the *cocktail party problem* (do people go to cocktail parties anymore?). What is remarkable about this ability is not just that we can segregate the parts of the signal specific to our interlocutor, but that we can shift our attention to another talker if our current conversation becomes uninteresting (as when our friend is discussing determining swimmers from water waves).

Bregman has referred to the abilities to segregate and integrate sounds relative to their sources as *auditory scene analysis*.[7] The visual analogy implied in that title was carried further by Bregman, who wondered if the basic perceptual principles of sound event segmentation may resemble those for visual organization developed by the Gestalt psychologists. There will be regularities in the structure of sounds rising from a single source that can be used as heuristics or principles for segregation and integration (sometimes called 'grouping'). Acoustic components arising from the same source tend to be similar across time, to be harmonically related (frequencies being integer multiples of each other), to begin and end together, and to continue without abrupt discontinuities. A great deal of empirical evidence has demonstrated that listeners tend to segregate complex sounds using these principles of similarity, harmonicity, contemporaneity, and good continuation when there is no other evidence available with which to segregate sounds.[7]

One of the clearest and best-known demonstrations of auditory scene analysis is a phenomenon called *auditory stream segregation*. A sequence of tones with alternating high and low frequencies (e.g., HLHLHL...) is heard as 'galloping' between the two frequencies at slow presentation rates. At faster presentation rates, however, the perceived organization of the tones changes; they are heard as two pulsating simultaneous sound events or 'streams', one high-frequency and the other low-frequency grouping by frequency similarity.[8,9] These perceptual streams no longer interact with each other perceptually, so that listeners can no longer tell the relative ordering of the tones in the two streams or hear rhythmic or melodic patterns that include tones from both streams. Although listeners usually report being able to 'hold' streams together at some rates of presentation, at faster presentation rates it is no longer possible for listeners to integrate the streams. It is also the case that stream segregation seems to 'build up' over time. Initially, one may hear a single stream that separates into two streams with greater exposure.[10,11] Auditory stream segregation has been demonstrated in nonhuman animals including macaques,[12] birds,[13] and even goldfish,[14] suggesting that the perceptual principles involved are quite general. In fact, one can see the 'build-up' effects of stream segregation in the responses of single neurons in auditory cortex (A1) of macaques[15] and even in neural responses in the brain stem of Guinea pigs (at the cochlear nucleus[16]).

Although the preceding depicts stream segregation as an obligatory, stimulus-driven, and 'lower level' perceptual phenomenon, auditory scene analysis, in general, is affected by attention, context, and knowledge. Bregman[7] points out that auditory organization may also be 'schema based' with expected or previously resolved patterns or schema guiding perception. For example, prior exposure to an otherwise unfamiliar target melody assists listeners later in segregating it from a more complex auditory scene.[17] Likewise, the perceptual interpretation of a preceding auditory stream may influence the way that subsequent streams are perceived.[18] Modulating effects of attention have also been demonstrated. For example, listeners are less likely to lose the galloping rhythm described above when they are distracted by a challenging auditory discrimination task presented to the opposite ear.[19] The nature of the interaction of 'schema-based' processing with the Gestalt-like basic principles of grouping remains an important empirical and theoretical question in auditory cognitive neuroscience. Our ability to segregate a single talker's speech from a crowd—the cocktail party problem—is likely a result of both our experience with patterns of speech across time (schema based) and of the fact that the speech signal from one talker satisfies many of the principles described above, such as good continuation, harmonicity, and similarity.[20]

Other questions that remain unanswered are how many streams or auditory events a person can represent at a time and how well represented are auditory events that are not currently the focus of attention. Gregg and Samuel[21] have developed an interesting methodology for answering these types of questions using an auditory analog of the well-established 'change blindness' for visual displays.[22] In this 'change deafness' paradigm, listeners are presented an 'auditory scene' consisting of four to six simultaneously presented sound events (e.g., dog barking, bell ringing). After an intervening 350 milliseconds of (white) noise, the 1-second auditory scene is repeated either identically or with one of the auditory events replaced with a novel sound. Listeners find this task remarkably difficult even with only four events in the scene; they fail to hear the change in approximately half of the change trials. It does not appear that this is due to a lack of encoding of the separate events because if the same presentation design is followed by an forced recognition task (which of these two sounds was presented in the previous scene?) listeners have little trouble identifying the presented sounds whether they were in the initial auditory scene, the second scene, or both. Thus, it appears that multiple sound events can be encoded by a listener, but that there is some difficulty in comparing sets of events across time (similar to some models of change blindness, e.g., Ref 23).

LOCALIZATION

Once a listener has segregated out a sound event, they may wish to determine the location from which that sound originated. Auditory localization is possible, of course, but performance pales in comparison to vision. The poorer performance for auditory localization may be one reason that it has often been considered a 'secondary sense' to vision. When visual and auditory spatial cues are put into conflict, the spatial location determined by the visual system tends to win out. This is the basis of the ventriloquist effect in which we perceive the speech coming from the visually moving dummy's mouth instead of from its true source, the ventriloquist's mouth.[24,25]. This 'visual dominance' appears to be due to the fact that visual spatial acuity is superior to auditory acuity in most situations. Alais and Burr[26] demonstrated that the auditory-determined location can dominate over the visual location when the visual spatial cues are degraded (by blurring the visual object). That is, perceivers actually integrate information across the senses but the more reliable information is more heavily 'weighted' in the final perception.[27] In most cases of localization, that information comes from the visual system, which is best at recognizing *objects* in *space*. However, before we relegate the auditory system to a lower position on the sensory esteem scale, it should be noted that the auditory system is better at recognizing *events* in *time* (which is why we have used the term auditory/acoustic 'event' instead of auditory 'object'). The temporal resolution of the auditory system far exceeds that of the visual system. And similar to the dominance of vision in spatial tasks, the rate of a flickering light tends to be perceived as synchronized to the rate of a repeating sound—a phenomenon called *auditory driving*.[28,29]

Despite the poorer spatial resolution, auditory localization does have its advantages: it works at night or when one's eyes are closed; it works for events that are occluded behind other objects; and it can determine locations of events occurring behind one's back. This flexibility may be one major reason that sound is used as our major conduit for communication. Unlike vision, in which the spatial representation is part of the encoding from the retina on, there is no spatial map in the auditory periphery. To localize a sound source, the listener needs to compute the direction from which a sound arrives at the ears from rather indirect cues. The two main cues that are used to localize an event in the horizontal (azimuth) plane are the differences in timing and intensity at the two ears. A sound to one's right is going to arrive at the right ear first. The difference in arrival times is referred to as an interaural time difference (ITD) or as an interaural phase difference (IPD) because the temporal difference for a sine wave will result in a shift in the relative phase at each ear. Note that differences in relative phase are perceived as changes in location of the sound source. It is also the case that a sound coming from one's right will be more intense at the right ear. This is not only due to the loss of power with distance (the inverse square law) but also due to the fact that the head produces a 'sound shadow'. The resulting difference in intensity between the two ears is referred to as an interaural level difference (ILD). Both of these cues are used by listeners to localize sounds, but ITDs tend to be more reliable for frequencies under 1500 Hz and ILDs for sounds above 4000 Hz. Between 1500 and 4000 Hz, localization ability is poorer.[30] These results support Lord Rayleigh's[31] original *duplex theory* of localization.

Note that the cues to location described above cannot resolve all spatial differences between sound sources. There will be no change in ITD or ILD for sound events differing only in the vertical plane (elevation) or directly in front or back of the listener. Although these types of localization tasks do lead to more errors, listeners can perform above chance, suggesting that there must be other localization cues. It is presumed that this ability is due to *head-related transfer functions* (HRTFs)—the shifts in amplitudes of different frequencies due primarily to the particular sound shadow cast by one's outer ear (the pinna) and ear canal. The exact perturbation of the incoming signal is a function of both the location of the sound source and the individual structure of the listener's ear. Thus, the auditory system can use this information to localize sounds in the vertical plane, but the mapping of the HRTF to spatial location will be specific to the individual listener. Presumably, this mapping has to be learned and altered across development. Hofman et al.[32] provided a demonstration of this perceptual plasticity by fitting adult humans with ear molds that significantly changed the HRTFs. Although localization in the vertical plane was abysmal on the first day with the molds, performance reached near normal levels after a month. Interestingly, immediately after removal of the molds, the listeners still demonstrated normal localization, suggesting that their initial HRTF mapping was preserved in parallel with the newly learned mapping.

While our understanding of localization of single sound sources (especially for simple sounds such as tones) is fairly well developed, there is still much to be learned about how spatial location is determined for multiple simultaneous sounds. One chicken-or-egg question is whether we segregate sound events prior to localization or whether we use spatial location cues to segregate sound events. It may be surprising to most readers that spatial cues do not appear to play a strong role in segmenting events especially when there are multiple events coming from multiple locations.[33] For example, listeners easily integrate harmonically related tones when presented to opposite ears,[34] even though this is clear evidence that they do not arise from the same spatial source. However, streaming based on spatial location can be observed as long as there are not other cues, suggesting a different organization.[35] Darwin[36] proposes that the relative weakness of spatial cues for segmentation may be due to the lack of reliability of these cues when multiple sound sources are allowed to interact or when reverberations and echoes are part of the listening environment. The complexity of the relationship of segmentation and localization in audition may be due to intervening effects of attention and the nature of the task.[37] In fact, attention, memory, and expectations all are likely to be important for creating the spatial representation for multiple sound sources, given the dearth of spatial information in the initial encoding of sound.

CATEGORIZATION

Once a listener has segregated and localized an event, they will probably wish to identify it. When presented sounds of struck bars, listeners can classify the material,[38,39] size,[38,40]] and shape[41] of the bar. However, listeners' performance is typically well below what would be considered optimal if one examines the relevant acoustic information in the signal.[42] Again, the auditory system does not appear to be particularly good at 'object' recognition. On the other hand, listeners can make robust classifications of some sound events based on subtle changes in the acoustics. In particular, listeners have little problem categorizing the dynamic speech signal into phonemes.

Readers who have had any exposure to the speech perception literature are likely to have heard about categorical perception—the hypothesis that speech sounds are not perceived as sounds, *per se*, but only in terms of their phoneme categories.[43] The support for this hypothesis is that it is very difficult to discriminate acoustic changes for sounds within a single phoneme category but very easy to discriminate sounds from different categories. It has been demonstrated that this discrimination pattern is not specific to speech sounds[44] and can be predicted readily from models of general auditory categorization.[45,46] Although many researchers (perhaps most) believe that speech perception—mapping acoustics to phonemes—is an example of perceptual categorization,[47] there has been a real lack of empirical work on the processes and constraints involved in categorizing sounds, generally.

Although there have been several studies that have examined how listeners can learn to categorize nonspeech sounds varying on a single acoustic dimension,[48–50] most auditory categories including speech categories are defined across multiple dimensions. One of the important considerations when one examines multidimensional categories is that listeners can selectively 'weight' or attend to one dimension or 'cue' more than the other. In speech sound categorization, this relative *cue weighting* changes over development[51,52] and is specific to one's native language. The difficulties producing and perceiving speech sounds in a nonnative language (such as Japanese speakers have difficulties with English 'l' and 'r') appear to be due, in part, to mismatches of cue weighting strategies between the first and second language.[53,54] Clearly, these cue weights are learned, but what determines which cue gets greater weight, and can we modify these weights when they are inefficient? Although there has been some research devoted to these questions using speech stimuli,[55] it is difficult to assert the necessary control over listeners' experience with speech sounds in order to evaluate the importance of any variable in determining the cue weights.

One solution is to use novel nonspeech sound categories that allow the researcher to know the stimuli that the listener has experienced perfectly. Holt and Lotto[56] created a set of arbitrary categories using a set of sine-wave tones that repetitively increased and decreased in frequency at a certain rate—modulation frequency (MF), which is heard as a difference in 'roughness'—around a particular center frequency (CF), which is perceived as a the basic pitch of the tone. Distributions of these sounds were created by varying both dimensions MF and CF. Figure 3 shows the distributions with each dot representing one exemplar of each category, which were simply named A and B. The shape of these distributions in the two-dimensional shape resembles distributions used previously for studies of vowel categorization with human infants[57] and birds.[58] Listeners were trained to categorize each sound as it was played in isolation with feedback. Although both the MF and CF cues are equally informative for the task, listeners weighted CF more heavily than MF—which could be seen in both their patterns of errors and by correlating their responses with values on each dimension. That is, listeners appear to have a bias to use pitch as a cue for auditory categorization. Similar biases in cue weighting have also been witnessed in categorization of struck bars.[59] This may be one reason that listeners' accuracy in judging the size, shape, and materials of these bars is lower than would be expected if they used all of the information available in the acoustics.

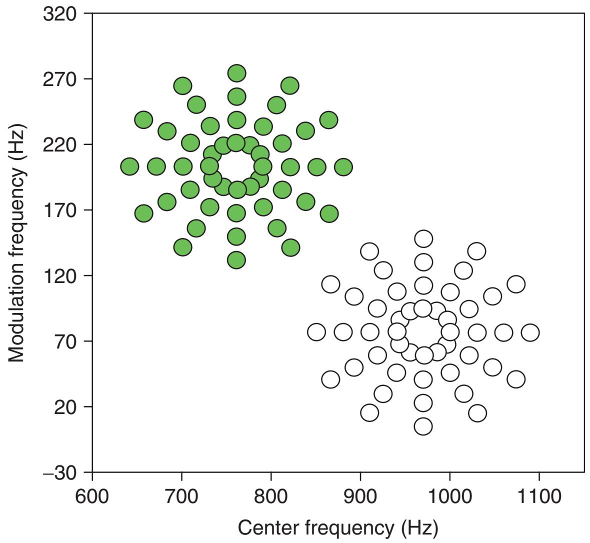

FIGURE 3 | Distributions of two arbitrary, novel sound categories. Each dot represents one sound; filled dots comprise one category and open dots define the other. Listeners learned to categorize these novel sounds as 'A' or 'B'. Although either of the dimensions could be used to differentiate the categories, listeners relied primarily upon the center frequency (CF) dimension.[56]

The biases in the cue weighting from this experiment provided an opportunity to see how one might modify the weights by modifying the training set. Holt and Lotto[56] moved the distributions closer together on the CF dimension in order to decrease the reliability of this cue (one would make more errors relying on it). This had no effect on the cue weights; listeners continued to rely more on CF despite the inefficiency of this strategy. In the third experiment, CF was made more variable within each distribution. In this case, the cue weights flipped so that MF became dominant. The key to modifying weights appears to be related to variance of a variable within categories. One may infer then that the best way to teach a learner of a second language to stop relying on an inefficient phonetic cue is to present that cue with substantial variability. In fact, it has been demonstrated that learning is enhanced by hearing multiple speakers produce target phonemes in a training set.[60] Presumably multiple speakers provide more variance in acoustic cues that are unrelated to the phonemes of concern, which helps to modify the learner's cue weights.

The previous work on cue weighting in auditory categorization demonstrates that learning and attention play a strong role in functional audition. It is still unclear how cue weighting or selective attention affects the representation of sounds and we can only speculate as to why listeners show the biases in weighting that have been demonstrated. There are sure to be more surprises as cognitive models of complex audition are developed. One possibility is that attention can modulate the encoding of sound all the way down to the movement of the basilar membrane. A feedback-efferent pathway exists from the brain stem (superior olivary complex) back to the outer hair cells in the cochlea. The activation of these efferents affect the gain of the cochlear amplifier benefit discussed in the section on neural encoding (see Ref 61 for an in-depth overview). There are also descending pathways from cortex to the brainstem-cochlear feedback system. This suggests the provocative possibility that cognitive processes could modulate the gain at the earliest stages of neural encoding. There is some evidence (though not definitive) that listeners who are selectively attending to a particular frequency region change the mechanics of transduction specifically at the point tuned for those frequencies.[62] If selective attention could affect the mechanical interface between the world and the central nervous system, it would provide a very different view of the relationship between perception and cognition than has been traditional.

CONCLUSION

Although auditory perception played an important role in the early development of cognitive science,[63] auditory cognitive science has subsequently lagged behind its visual counterpart. Using the functional tasks of segmentation, localization, and categorization of sound events as examples, we have tried to demonstrate that significant perceptual/cognitive processing is involved in everyday audition. We have tried hard not to use the banal phrase 'much more research must be done', but much more research needs to be done. We have a fairly good understanding of the neural encoding and perception of simple sounds (tones and noises) presented in isolation, but there are few coherent models of how complex sounds, especially presented simultaneously, are perceived. For example, the perception of complex sounds, such as speech, is often affected by the acoustic makeup of preceding and following sounds.[64] To determine the mechanisms underlying these effects, researchers must move away from studying sounds in isolation. The development of more complex auditory models is essential for us to make continuing progress in understanding how humans communicate through speech and music.

REFERENCES

1. Lisker L. "Voicing" in English: a catalogue of acoustic features signaling /b/ versus /p/ in Trochees. *Lang Speech* 1986, 29:3–11.
2. Holt LL, Lotto AJ. Speech perception within an auditory cognitive science framework. *Curr Dir Psychol Sci* 2008, 17:42–46.
3. Yost WA. *Fundamentals of Hearing: An Introduction.* Massachusetts: Academic Press; 2007.
4. de Cheveigné A. Pitch perception models. In: Plack CJ, Oxenham AJ, Fay RR, Popper AN, eds. *Pitch: Neural Coding and Perception.* New York: Springer Science and Business Media; 2005, 169–233.
5. Javel E, McGee J, Walsh EJ, Farley GR, Gorga MP. Suppression of auditory nerve responses. II. Suppression threshold and growth, iso-suppression contours. *J Acoust Soc Am* 1983, 74:801–813.
6. Cherry EC. Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am* 1953, 25:975–979.
7. Bregman AS. *Auditory Scene Analysis.* Massachusetts: MIT Press; 1990.
8. Miller GA, Heise GA. The trill threshold. *J Acoust Soc Am* 1950, 22:637–638.
9. Bregman AS, Campbell J. Primary auditory stream segregation and perception of order in rapid sequences of tones. *J Exp Psychol* 1971, 89:244–249.
10. Bregman AS. Auditory streaming is cumulative. *J Exp Psychol* 1978, 4:380–387.
11. Anstis S, Saida S. Adaptation to auditory streaming of frequency-modulated tones. *J Exp Psychol Hum Percept Perform* 1985, 11:257–271.
12. Izumi A. Auditory stream segregation in Japanese monkeys. *Cognition* 2002, 82:113–122.
13. MacDougall-Shackleton SA, Hulse SH, Gentner TQ, White W. Auditory scene analysis by European starlings (Sturnus vulgaris): perceptual segregation of tone sequences. *J Acoust Soc Am* 1998, 103:3581–3587.
14. Fay RR. Spectral contrasts underlying auditory stream segregation in goldfish (Carassius auratus). *J Assoc Res Otolaryngol* 2000, 1:120–128.
15. Micheyl C, Carlyon RP, Gutschalk A, Melcher JR, Oxenham AJ, Rauschecker JP, Tian B, Courtenay WE. The role of auditory cortex in the formation of auditory streams. *Hear Res* 2007, 229:116–131.
16. Pressnitzer D, Sayles M, Micheyl C, Winter I. Perceptual organization of sound begins in the auditory periphery. *Curr Biol* 2008, 18:1124–1128.
17. Bey C, McAdams S. Schema-based processing in auditory scene analysis. *Percept Psychophys* 2002, 64:844–854.
18. Snyder JS, Carter OL, Lee SK, Hannon EE, Alain C. Effects of context on auditory stream segregation. *J Exp Psychol* 2008, 34:1007–1016.
19. Carlyon RP, Cusack R, Foxton JM, Robertson IH. Effects of attention and unilateral neglect on auditory stream segregation. *J Exp Psychol Hum Percept Perform* 2001, 27:115–127.
20. Dorman MF, Cutting JE, Raphael LJ. Perception of temporal order in vowel sequences with and without

formant transitions. *J Exp Psychol Hum Percept Perform* 1975, 1:121–129.

21. Gregg MK, Samuel AG. Change deafness and the organizational properties of sounds. *J Exp Psychol Hum Percept Perform* 2008, 34:974–990.
22. Luck SJ, Vogel EK. The capacity of visual working memory for features and conjunctions. *Nature* 1997, 390:279–281.
23. Mitroff SR, Simons DJ, Levin DT. Nothing compares 2 views: change blindness can occur despite preserved access to the changed information. *Percept Psychophys* 2004, 66:1268–1281.
24. Pick HL, Warren DH, Hay JC. Sensory conflict in judgments of spatial direction. *Percept Psychophys* 1969, 6:203–205.
25. Warren DH, Welch RB, McCarthy TJ. The role of visual-auditory "compellingness" in the ventriloquism effect: implications for transitivity among the spatial senses. *Percept Psychophys* 1981, 30:557–64.
26. Alais D, Burr D. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol* 2004, 14:257–262.
27. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 2002, 415:429–433.
28. Welch RB, DuttonHurt LD, Warren DH. Contributions of audition and vision to temporal rate perception. *Percept Psychophys* 1986, 39:294–300.
29. Recanzone GH. Auditory influences on visual temporal rate perception. *J Neurophys* 2003, 89:1078–1093.
30. Stevens SS, Newman EB. The localization of actual sources of sound. *Am J Psychol* 1936, 48:297–306.
31. Rayleigh L. On our perception of sound direction. *Philos Mag* 1907, 13:214–232.
32. Hofman PM, Van Riswick JG, Van Opstal AJ. Relearning sound localization with new ears. *Nat Neurosci* 1998, 1:417–421.
33. Darwin CJ. Auditory grouping. *Trends Cogn Sci* 1997, 1:327–333.
34. Beerends JG, Houtsma AJM. Pitch identification of simultaneous dichotic two-tone complexes. *J Acoust Soc Am* 1986, 80:1048–1056.
35. Hartmann WM, Johnson D. Stream segregation and peripheral channeling. *Music Percept* 1991, 9:155–184.
36. Darwin CJ. Spatial hearing and perceiving sources. In: Yost WA, Popper AN, Fay RR, eds. *Auditory Perception of Sound Sources*. New York: Springer Science and Business Media; 2008, 215–232.
37. Ihlefeld A, Shinn-Cunningham B. Disentangling the effects of spatial cues on selection and formation of auditory objects. *J Acoust Soc Am* 2008, 124:2224–2235.
38. Gaver WW. *Everyday Listening and Auditory Icons*. California: University of California; 1988.
39. Lutfi RA, Oh EL. Auditory discrimination of material changes in a struck-clamped bar. *J Acoust Soc Am* 1997, 102:3647–3656.
40. Tucker S, Brown GJ. Modeling the auditory perception of size, shape and material: applications to the classification of transient sonar sounds. *114th Auditory Engineering Society Convention*. Amsterdam: The Netherlands; 2003, 22–25.
41. Kunkler-Peck AJ, Turvey MT. Hearing shape. *J Exp Psychol Hum Percept Perform* 2000, 26:279–294.
42. Lutfi RA. Human sound source identification. In: Yost WA, Popper AN, Fay RR, eds. *Auditory Perception of Sound Sources*. New York: Springer Science and Business Media; 2008, 13–42.
43. Liberman AM, Harris KS, Hoffman HS, Griffith BC. The discrimination of speech sounds within and across phoneme boundaries. *J Exp Psychol* 1957, 54:358–368.
44. Cutting JE. Plucks and bows are categorically perceived, sometimes. *Percept Psychophys* 1982, 31:462–476.
45. Massaro DW. Categorical partition: a fuzzy logical model of categorization behavior. In: Harnad S, ed. *Categorical Perception: The Groundwork of Cognition*. Massachusetts: Cambridge University Press; 1987, 254–283.
46. Schouten B, Gerrits E, Van Hessen A. The end of categorical perception as we know it. *Speech Commun* 2003, 41:71–80.
47. Lotto AJ. Language acquisition as complex category formation. *Phonetica* 2000, 57:189–196.
48. Guenther FH, Husain FT, Cohen MA, Shinn-Cunningham BG. Effects of categorization and discrimination training on auditory perceptual space. *J Acoust Soc Am* 1999, 106:2900–2912.
49. Sullivan SC, Lotto AJ, Newlin ET, Diehl RL. Sensitivity to changing stimulus distribution characteristics in auditory categorization. *J Acoust Soc Am* 2005, 118:1896.
50. Smits R, Sereno J, Jongman A. Categorization of sounds. *J Exp Psychol* 2006, 32:733–754.
51. Hazan V, Barrett S. The development of phonemic categorization in children aged 6–12. *J Phon* 2000, 28:377–396.
52. Nittrouer S. The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *J Acoust Soc Am* 2004, 115:1777–1790.
53. Iverson P, Kuhl PK, Akahane-Yamada R, Diesch E, Tohkura Y, Kettermann A, Siebert C. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 2003, 87:47–57.
54. Lotto AJ, Sato M, Diehl RL. Mapping the task for the second language learner: the case of Japanese acquisition of /r/ and /l/. In: Slifka J, Manuel S, Matthies M, eds. *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*. Massachusetts: MIT Press; 2004, C181–C186.

55. Francis AL, Kaganovich N, Driscoll-Huber C. Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *J Acoust Soc Am* 2008, 124:1234–1251.

56. Holt LL, Lotto AJ. Cue weighting in auditory categorization: implications for first and second language acquisition. *J Acoust Soc Am* 2006, 119:3059–3071.

57. Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 1992, 255:606–608.

58. Kluender KR, Lotto AJ, Holt LL, Bloedel SL. Role of experience for language-specific functional mappings of vowel sounds. *J Acoust Soc Am* 1998, 104:3568–3582.

59. Lutfi RA, Liu CJ. Individual differences in source identification from synthesized impact sounds. *J Acoust Soc Am* 2007, 122:1017–1028.

60. Lively SE, Logan JS, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *J Acoust Soc Am* 1993, 94:1242–1255.

61. Guinan JJ Jr. Olivocochlear efferents: anatomy, physiology, function, and the measurement of efferent effects in humans. *Ear Hear* 2006, 27:589–607.

62. Maison S, Micheyl C, Collet L. Influence of focused auditory attention on cochlear activity in humans. *Psychophysiology* 2001, 38:35–40.

63. Broadbent DE. *Perception and Communication*. London: Pergamon Press; 1958.

64. Lotto AJ, Holt LL. Putting phonetic context effects into context: a commentary on Fowler (2006). *Percept Psychophys* 2006, 68:178–183.

FURTHER READING

McAdams S, Bigand E, eds. *Thinking in Sound: The Cognitive Psychology of Human Audition*. New York: Oxford University Press; 1993.

Yost WA, Popper AN, and Fay RR, eds. *Auditory Perception of Sound Sources*. New York: Springer Science and Business Media; 2008.