Primer

The cocktail party problem

Josh H. McDermott

Natural auditory environments, be they cocktail parties or rain forests, contain many things that concurrently make sounds. The cocktail party problem is the task of hearing a sound of interest, often a speech signal, in this sort of complex auditory setting (Figure 1). The problem is intrinsically quite difficult, and there has been longstanding interest in how humans manage to solve it.

The problem is not unique to humans, however — many other species confront something similar. Nonhuman animals frequently must identify mates, offspring, or adversaries in crowded environments containing many animals vocalizing at once. Many species of frogs, for instance, must use conspecific vocalizations to locate mates at night, when there are few visual cues, amid hundreds of other frogs.

Cocktail party challenges

There are two conceptually distinct challenges for a listener in this situation (though much of the time they are closely related), and the term 'cocktail party problem' is often used in reference to both of them.

The first is the problem of sound segregation. The sounds in an auditory scene all sum together to generate the signal that enters the ear. But this mixture of sounds is itself of little use to an organism, which is typically interested in particular individual sound sources (a potential mate, for instance). The auditory system must derive the properties of individual sounds from the mixture entering the ears.

The second challenge is that of directing attention to the sound source of interest while ignoring the others, and of switching attention between sources, as when intermittently following two conversations. Most of our cognitive processes can operate only on one thing at a time, so we typically select a particular sound source on which to focus. This process is in practice intertwined with sound segregation, as segregation is biased by what we attend to.

The cocktail party problem was popularized in a classic paper by Cherry in 1953. Cherry focused on the attentional component of the problem, introducing the now-famous dichotic listening paradigm to study whether observers could select one speech signal over another, whether they retained anything about the non-selected signal, and how they could switch their attention between signals. Two decades later, Bregman began studying sound segregation, terming it auditory scene analysis. Most contemporary work on the cocktail party problem is rooted in this latter research program, and we know less about the mechanisms of auditory selective attention and their interaction with sound segregation. Reflecting the state of the field, this Primer will focus primarily on the problem of sound segregation.

Why are these hard problems?

Sound segregation is a classic example of an ill-posed perceptual problem. Many sets of sound signals are physically consistent with the mixture that enters the ear, only one of which is the actual set that occurred in the world. The brain has to infer the correct sounds, or at least correctly estimate the sound of primary interest. Because the problem is ill-posed, implicit knowledge of sound sources is needed to choose the sounds that are most likely given the mixture. Most of the time we do remarkably well.

The ear performs a frequency-decomposition on sound signals, and it can be useful to visualize signals in the same way. The left column of Figure 2 shows spectrograms (plots of frequency content over time) of four sound signals. The first is a single spoken sentence — this is the signal that would enter the ear if only one person were talking in a quiet room. Below it are the spectrograms of the same utterance combined with one, three, or seven additional utterances by different speakers. The task of the auditory system is to analyze mixtures like these (as might enter the ear at a party or restaurant) and derive a representation of the speech signal of interest that allows the listener to understand what is being said.

Viewing the auditory input in this way reveals two obstacles to recovering an individual sound signal from the mixture. The first is that it is not obvious which bits of sound in the mixture belong to the speech signal of interest (the 'target'), and which belong to the other utterances. In many places, the mixture has energy where the target has little (marked in red in the right column of Figure 2). Estimating the target necessitates identifying these parts of the mixture and segregating them from those belonging to the target. The second obstacle is that in some of the places where the isolated target speech has significant energy, one of the other signals has more (marked in green in the right column of Figure 2). That part of the target is thus physically masked by the other sounds, which could make it more difficult to recover from the mixture.

It is apparent that as the party gets larger, and the number of other speech signals increases, both of these problems become exacerbated (both the red and the green increase). When there are eight speakers in the mixture (Figure 2, bottom row), there is energy in most regions of the time-frequency plane, much of which is from signals other than the one of interest. Remarkably, the target speaker in this example remains at least partially intelligible to human listeners.

Cocktail parties and other comparably noisy environments thus present extreme examples of the challenges inherent to sound segregation, because the number of sound sources can be vast. Listeners are usually trying to follow a speech signal from one other person amid a cacophony of other sounds (other people talking and laughing, glasses clinking, music playing). In such situations, we are living on the edge — the speech signal of interest is often close to the threshold of intelligibility. If it decreases a bit in volume relative to the other sounds, as if the listener steps back a few feet, it often becomes impossible to understand.

Attentional selection in such environments is also at its most challenging. There are many sources competing for a listener's attention, so switching between them is difficult, as is the task of ignoring signals that are not of interest. The presence of multiple sources thus imposes a cognitive load even if the source of interest is recovered correctly from the mixture. Complex auditory environments typically demand the listener's full attention.

Cocktail party solutions

The cocktail party problem is partially solved with perceptual mechanisms that allow the auditory system to estimate individual sound sources from mixtures. 'Bottom-up' grouping cues derived from statistical regularities of sounds help tell us what goes with what (Figure 2, top). For instance, individual sounds tend to exhibit amplitude changes that are common across the different frequencies they contain. So if a mixture contains energy at multiple frequencies that start or stop at the same time, those frequencies are likely to belong to the same sound, and are interpreted as such by the brain. Mammalian vocalization and musical instrument sounds also tend to contain frequencies that are harmonically related — they occur at discrete integer multiples of a fundamental frequency (for speech, this is the frequency at which the vocal cords vibrate). In a spectrogram, these harmonics are evident as stacked bands of energy (visible in the top row of Figure 2, at the lower end of the spectrum). Frequencies in a mixture that have this harmonic relationship are thus likely to belong together, and tend to be heard as a single sound.

The listener's task is also aided by the fluctuations that occur in many natural sounds, as this lessens the extent to which they physically obscure each other. One sound may be high enough in amplitude to mask another at one moment in time, but not the next (evident in the intermittent regions of green in Figure 2). These fluctuations tend to provide 'glimpses' of each of the sounds in a mixture. If the various bits of sound that are glimpsed are grouped appropriately, the auditory system can often fill in the parts that are masked. The glimpses do not in themselves solve the grouping problem, but they help make a solution possible. If the competing sounds in Figure 2 did not fluctuate in level, they might completely obscure the target speech signal, in which case there would be nothing to group.

In many cases, however, it is likely that bottom-up segregation cues are not enough — listeners must also rely on their knowledge of specific sounds or sound classes. This is most evident for the perception of speech in a noisy background, which is substantially more accurate if the words form coherent sentences than if they are random sequences of words. Listeners are also better at comprehending speech in a cocktail party setting if the speaker has a familiar, rather than foreign, accent. In the examples of Figure 2, knowing what to listen for (by hearing the isolated target speaker first, for instance) makes it much easier to hear the target speech, especially in the eight-talker mixture. It is unclear whether these effects directly influence auditory grouping, or whether they simply reflect the benefit of linguistic and phonetic knowledge for interpreting an impoverished speech signal.

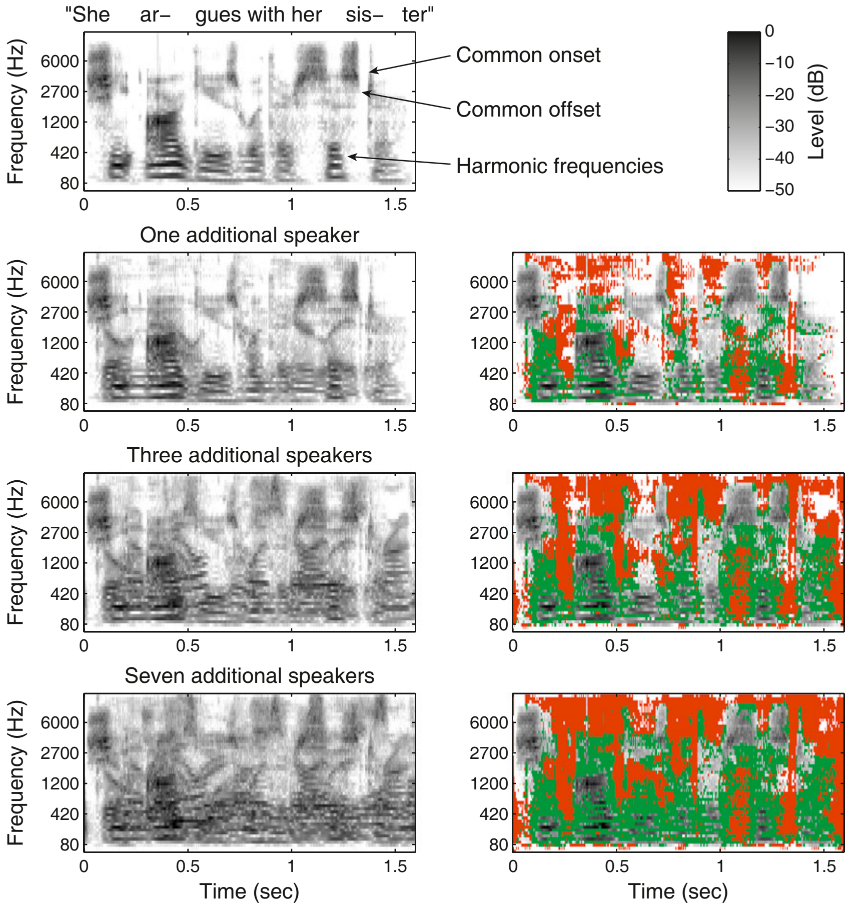
Figure 1. A typical Manhattan cocktail party.
The listener must follow the conversation of interest despite many concurrent sources of sound. (Image from Breakfast at Tiffany's: Paramount Pictures.)

Localization cues afforded by our two ears are another source of information — if a target sound has a different spatial location than distractor sounds, it tends to be easier to detect and understand. Visual cues to speech (as used in lip reading) also help improve intelligibility. Both location and visual cues may help in part by guiding attention to the relevant part of the auditory input, enabling listeners to suppress the parts of the mixture that do not belong to the target signal.

Biological versus machine algorithms for source separation

Recognizing sounds from mixtures is also a central challenge for machine algorithms that interpret sound. State-of-the-art speech recognition programs, for instance, are typically close to perfect if the speaker is alone in a quiet room, but perform much worse in real-world conditions with other concurrent sound sources. Because of its relevance to many audio applications, there has been much interest in the cocktail party problem from within the engineering community.

Although some machine hearing approaches are biologically inspired, many of the best-known engineering methods for source separation address a problem that is quite different from that solved by the brain in cocktail party settings. A class of algorithms known as independent component analysis (ICA) can be effective at separating sources from an auditory scene if the scene is recorded with multiple microphones positioned at different



Current Biology

Figure 2. Cocktail party acoustics.

Spectrograms of a single 'target' utterance (top row), and the same utterance mixed with one, three and seven additional speech signals from different speakers. The mixtures approximate the signal that would enter the ear if the additional speakers were talking as loud as the target speaker, but were standing twice as far away from the listener (as might occur in cocktail party conditions). Spectrograms were computed from a filter bank with bandwidths and frequency spacing similar to those in the ear. Each spectrogram pixel represents the rms amplitude of the signal within a frequency band and time window. The spectrogram thus omits local phase information, which listeners are insensitive to in most cases. The grayscale denotes attenuation (in dB) from the maximum amplitude of all the pixels in all of the spectrograms, such that gray levels can be compared across spectrograms. Acoustic cues believed to contribute to sound segregation are indicated in the spectrogram of the target speech (top row). Spectrograms in the right column are identical to those on the left except for the superimposed color masks. Pixels labeled green are those where the original target speech signal is more than –50 dB but the mixture level is at least 5 dB higher. Pixels labeled red are those where the target was less and the mixture was more than –50 dB in amplitude. The sound signals used to generate the spectrograms can be listened to at http://www.cns.nyu.edu/~jhm/cocktail_examples/.

locations in a room. As some sources are closer to a microphone than others, each microphone yields a different weighted combination of the sources. If there are as many microphones as there are sources, rudimentary assumptions about the statistical properties of sources (typically, that they are non-Gaussian according to some criterion) often suffice to derive the source signals from the mixtures.

Biological organisms, however, have but two microphones (our ears), and routinely must segregate sounds of interest from scenes with more than two sources. Moreover, although sound segregation is aided by binaural localization cues, we are not dependent on them — humans generally can separate sources from a monaural mixture. Indeed, much popular music of the 20th century was recorded in mono, and listeners can nonetheless hear different instruments and vocalists in such recordings without trouble. The target speech of Figure 2 is also readily understood from monaural mixtures of multiple talkers. It is thus clear that biological auditory systems are doing something rather different from standard ICA algorithms, though it remains possible that some of the principles of ICA contribute to biological hearing. At present we lack effective engineering solutions to the 'single microphone' source separation problem that biological auditory systems typically solve with success.

Auditory versus visual segmentation

Segmentation is also a fundamental problem for the visual system. Visual scenes, like auditory scenes, usually contain multiple objects, often scattered over complex backgrounds, and the visual system must recognize individual objects amid this clutter (Figure 1). However, several salient differences between visual and auditory scenes make the auditory segmentation problem particularly difficult. The first is that visual objects tend to occupy local regions on the retina. Sound sources are, by comparison, spread out across the frequency map of the cochlea (as is apparent in Figure 2). As a result, their sensory representation often overlaps considerably more than do those of visual objects. A second difference compounds this problem — sound sources add linearly to create the signal entering the ears, whereas visual objects occlude each other. If objects are opaque, as they usually are, the object nearest the viewer determines the image content at its location. The sound energy at a particular time and frequency, by contrast, is a sum across every sound source in the auditory scene. The upshot is that the more people there are at a party, the harder it will be to hear the person standing next to you, as the sounds made by each additional person will at various points in time mask the speaker of interest. The speaker's face, in contrast, will remain perfectly visible unless someone steps in front of them (Figure 1).

The situation with sound is perhaps most analogous to what the visual world would be like if objects were transparent. Recognition tasks in such a world would no doubt be more difficult, as the image would be determined not by the closest object, but rather by a mixture of many objects, each of which could potentially mask the others. This difference is perhaps one reason why listening in noisy environments often feels effortful, a feeling we rarely get when performing visual recognition tasks. Another difference is that much of the information we need from the visual world is relatively stable, at least over short time scales. Sounds, in contrast, are here and then gone, and so must be carefully monitored to avoid missing something of interest.

Sound segregation in music

Music provides spectacular examples of sound segregation in action — recordings often have more instruments or vocalists than can be counted, and in the best such examples, we feel like we can hear every one of them. Why does following a particular sound source in a piece of music often feel effortless? Unlike a naturally occurring auditory scene, music is often specifically engineered to facilitate sound segregation. Recording engineers apply an extensive bag of tricks to make instruments audible in their mix, filtering them, for instance, so that they overlap less in frequency than they normally would, thus minimizing masking. The levels of different instruments are also carefully calibrated so that each does not overwhelm the others. Real-life cocktail parties unfortunately do not come with a sound engineer.

Sound segregation in music also no doubt benefits from our familiarity with instrument sounds and musical structure — we often have well-defined expectations, and this knowledge of what is likely to be present surely helps us distinguish instruments and voices.

Music also provides interesting examples where sound segregation is intentionally made difficult for aesthetic effect. For instance, sometimes a producer may want to cause two instruments to perceptually fuse to create a new sort of sound. By carefully coordinating the onset and offset of two instrument sounds, our auditory system can be tricked into thinking the two sounds are part of the same thing.

Hearing impairment and sound segregation

As people get older, they often complain of being unable to follow conversations in noisy restaurants and parties. Their difficulties can be traced at least in part to the hearing impairment that is common in the elderly. The problem is not audibility — even when sounds are presented to the ear at high levels (for instance, courtesy of a hearing aid), hearing impaired listeners tend to struggle in noisy environments. The reasons for this are an area of active research, but seem to involve an inability to benefit from the 'glimpses' of sources provided by sound fluctuations. When recognizing speech embedded in noise, normal hearing listeners perform better when the noise fluctuates in time and frequency than when it does not; this advantage tends to be reduced in hearing impaired listeners. A big part of this may derive from the altered frequency response of the ear, which becomes more coarsely tuned in people with hearing impairment. Broader peripheral filters are less able to resolve fine spectral detail, and in general produce more masking than would occur for a normal listener. The factors that make sound segregation difficult to begin with are thus likely to be aggravated in the hearing impaired.

What don't we understand about the cocktail party problem?

Despite having a crude outline of how humans succeed at segregating sounds, we cannot produce machine algorithms that achieve anything close to human competence. In this sense there is still much we do not know. Even situations that are fairly trivial for a normal human listener, such as following one speaker in the presence of another, are a challenge for automated speech recognition programs.

Intriguing lines of research await at all levels of the problem. The bottom-up component of sound segregation has been studied primarily with simple synthetic stimuli; it remains to be seen how our understanding will change as we consider how natural sounds are grouped and segmented from mixtures. There may be additional bottom-up grouping cues, for instance, that we do not currently appreciate but that will become apparent with closer examination of natural sounds. We also have much to learn about the role of 'top-down' influences, be it of linguistics, speech acoustics, or specific sound identities. It is clear that listeners benefit from familiarity with the sounds to be segregated, but how stored representations of sound interact with bottom-up segregation processes is poorly understood. The mechanisms of auditory attention and the way they interact with sound segregation have only been studied sporadically, and will surely be a rich area of future research. Finally, the representation of segregated sounds in neural circuitry is an area of great recent interest.

Acknowledgments

The author thanks Mark Bee, Trenton Jerde, Umesh Rajashekar, and Brett Vintch for comments on an earlier draft of this Primer.

Further reading

Bee, M., and Micheyl, C. (2008). The "cocktail-party problem": What is it? How can it be solved? And why should animal behaviorists study it? J. Comp. Psychol. *122*, 235–251.
Bregman, A.S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound (Cambridge, MA: MIT Press).
Bronkhorst, A.W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acustica *86*, 117–128.
Brungart, D.S., Chang, P.S., Simpson, B.D., and Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. *120*, 4007–4018.
Carlyon, R.P. (2004). How the brain separates sounds. Trends Cogn. Sci. *8*, 465–471.
Cherry, E.C. (1953). Some experiments on the recognition of speech, with one and two ears. J. Acoust. Soc. Am. *25*, 975–979.
Cooke, M. (2006). A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. *119*, 1562–1573.
Darwin, C.J. (1997). Auditory grouping. Trends Cogn. Sci. *1*, 327–333.
Elhilali, M., and Shamma, S. (2008). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. J. Acoust. Soc. Am. *124*, 3751–3771.
McDermott, J.H., and Oxenham, A.J. (2008). Spectral completion of partially masked sounds. Proc. Natl. Acad. Sci. USA *105*, 5939–5944.
Narayan, R., Best, V., Ozmeral, E.J., McClaine, E., Dent, M., Shinn-Cunningham, B.G., and Sen, K. (2007). Cortical interference effects in the cocktail party problem. Nat. Neurosci. *10*, 1601–1607.
Pressnitzer, D., Sayles, M., Micheyl, C., and Winter, I.M. (2008). Perceptual organization of sound begins in the periphery. Curr. Biol. *18*, 1124–1128.
Shinn-Cunningham, B.G. (2009). I want to party, but my hearing aids won't let me! Hearing J. *62*, 10–13.
Wang, D., and Brown, G.J. (2006). Computational Auditory Scene Analysis: Principles, Algorithms, and Applications (Hoboken, NJ: Wiley-IEEE Press).

Center for Neural Science, New York University, New York, NY 10003, USA.
E-mail: jhm@cns.nyu.edu