Auditory grouping

Chris J. Darwin

Although our subjective experience of the world is one of discrete sound sources, the individual frequency components that make up these separate sources are spread across the frequency spectrum. Listeners use various simple cues, including common onset time and harmonicity, to help them achieve this perceptual separation. Our ability to use harmonicity to segregate two simultaneous sound sources is constrained by the frequency resolution of the auditory system, and is much more effective for low-numbered, resolved harmonics than for higher-numbered, unresolved ones. Our ability to use interaural time-differences (ITDs) in perceptual segregation poses a paradox. Although ITDs are the dominant cue for the localization of complex sounds, listeners cannot use ITDs alone to segregate the speech of a single talker from similar simultaneous sounds. Listeners are, however, very good at using ITD to track a particular sound source across time. This difference might reflect two different levels of auditory processing, indicating that listeners attend to grouped auditory objects rather than to those frequencies that share a common ITD.

Our conscious experience of sound usually consists of a number of different, separate sound-producing events each with their own location, pitch, loudness and quality. But these separate sources are not explicitly present in the sensory information. The sound waves from each source are mixed together when they enter the two ears to make a composite waveform. The problem of interpreting this sound mixture in terms of separate events is broadly similar to the visual problem of interpreting, in terms of three-dimensional objects, the pair of twodimensional images formed on the two retinas. An important difference is that, unlike vision, where adjacent retinal regions tend to be stimulated by light from the same object, frequency components from two separate sounds can be interleaved across the whole frequency spectrum. Each sound source contains many frequency components, and the components that make up a particular sound are not constrained to occupy a contiguous region of the auditory spectrum in the cochlea. How does the brain sort out which components belong to which sound source?

This problem has implications for a number of disciplines. For composers, it is important that the listener should hear separate sound sources for successful melodic lines and polyphony, but also be able to blend different instruments to give new tone colours. For engineers, the successful separation of a single voice from other, competing sounds and voices would increase recognition rates in automatic speech recognition and the effectiveness of aids to sensorineural hearing loss. For cognitive science, understanding of auditory attention integrated with auditory psychophysics, perception and object recognition is an important goal.

Although the early Gestalt psychologists could have taken advantage of music-box and pianola technology to carry out simple auditory analogues of their well-known work in vision, extensive experimental and computational work on the problem has occurred only over the past 20 years or so, with the advent of digital signal processing and simpler computer control over complex auditory stimuli. Modern experimental work on the segregation of complex sound mixtures into perceptual 'streams' has been led by Albert Bregman and reviewed in his book Auditory Scene Analysis. 1

Sequential grouping

The basic phenomenon of auditory stream segregation is compelling, and was particularly exploited by Baroque composers as 'implied polyphony' (one instrument sounding like two). In its simplest form, two rapidly alternating tones are heard as a trill when close in frequency, but as two separate monotonous 'streams' as they are pulled apart. Miller and Heise used 100 ms tones—a musically plausible trill rate—and found that a frequency difference of about 15% (a whole tone is 12%) was sufficient to force separation into two streams. More extensive work was done later on alternating tones by van Noorden, who distinguished between enforced separation at large frequency differences and voluntary separation at smaller frequency differences under attentional control.

A more elaborate demonstration of a similar effect was made in Bregman and Campbell's seminal study. They showed that a sequence of six rapid notes breaks up into two streams of three notes as the average frequency separation of the two groups of notes is increased. Once segregation has occurred, although listeners can readily distinguish the temporal order of notes within a stream, they find it very difficult to judge the temporal order of notes in different streams.
Fig. 1 The Wessel Illusion – streaming by timbre. (A) The rapid, ascending, three-tone motif in the left part of the figure streams into the two red and blue, slower descending motifs (in the right part of the figure) when alternate notes are played on sufficiently different timbres. In this example, the bottom few harmonics have been removed from the odd-numbered notes to give a brighter timbre, and the top few from the even-numbered notes to give a duller timbre. (B) The compressed time-scale makes explicit the percept of descending triplets in the streamed condition.

- a finding reminiscent of the difficulty that listeners have of accurately locating a click in a sentence.

These simple demonstrations, using pure tones, illustrate a grouping principle of frequency similarity, where the auditory system reduces the rate of change of frequency over time within a stream by increasing the number of streams. With complex tones, though, the situation becomes more complex. For sequences of complex tones, consisting of a series of harmonics (of frequency nFo) of a fundamental (Fo), streaming can be determined by changes in timbre and by changes in pitch. In addition, streaming can be produced by changes in spatial location for both pure and complex tones.

The Wessel illusion⁴ is a particularly neat demonstration of streaming by timbre, and is illustrated in Fig. 1. A rapid, ascending, three-tone motif streams into two slower, descending motifs when alternate notes are played on sufficiently different timbres.⁵ Removing the bottom few harmonics from the odd-numbered notes gives a brighter timbre, whereas removing the top few harmonics from the even-numbered notes gives a duller timbre. Tones that have different gross average spectral properties (as in Fig. 1) are more likely to stream, but dynamic attributes, including attack duration and the change of the spectrum over time, also have an influence, so it is unlikely that such streaming can be accounted for on the basis of simple frequency channels.

An example of streaming produced by repeated changes in pitch comes from an early experiment by Darwin and Bethell-Fox⁶ (Fig. 2). The frequencies of three formant resonances are varied over time. When these resonances are excited by a monotonous pulse train (giving a constant Fo) listeners hear the repeating syllables ‘yang…’. But if the exciting pulse train is made to alternate between two different frequencies then, after a couple of cycles or so, the speech breaks up into two different voices, each speaking on one of the two pitches. Moreover, because each voice is heard to be silent while the other is speaking, one of the voices produces a formant pattern and silence that is heard as ‘gagag…’ – the other ‘voice’ produces a curious sound because the rising first formant before the implied closure is phonetically implausible.

Another grouping principle for the sequential organization of sound is that of spatial location. The Wessel illusion can be easily obtained if alternate notes are played to the left and right ears rather than on different timbres. The story, though, becomes more complex if more than one note is present at a time. Deutsch⁷ has shown that a tone that has become unrecognizable when its notes are alternated across the ears, does become recognizable again if each note is accompanied in the opposite ear by a simultaneous drone note on a constant pitch. Deutsch’s result, and others like it⁸, indicate that grouping by fundamental frequency can dominate over grouping by location, particularly when the location cues are weakened by another simultaneous note.

Our own observations with the Wessel stimulus qualify Deutsch’s conclusion. If we alternate the sounds of the Wessel stimulus between the ears and put a drone tone in the opposite ear, the percept is initially as Deutsch’s observations predict: we hear the original single, rapid, ascending pattern of three notes, indicating that organization by frequency is dominating over organization by location. However, after a second or two, the percept changes to one corresponding to a perceptual organization by location: a slow descending pattern of three notes is heard in each ear. The conclusion here is that the weakening of grouping by location produced by the contralateral drone tone changes over time. Grouping by spatial location becomes stronger as a stimulus continues and might become dominant at longer durations.
In summary, continuity of timbre, pitch and spatial location are important for maintaining the perceptual integrity of a sound source over time.

Simultaneous grouping
Another type of grouping is the separation of a mixture of sounds into separate, broadly simultaneous sources. An example is our ability to hear the separate instruments playing a chord. This ability breaks down when players produce notes that are simultaneous and whose frequencies are unisons (equal) or simple ratios such as an octave (2:1) or a twelfth (3:1, as in the ‘octave quint’ compound stop of the organ). It is easier to hear the individual contributions of less accomplished players who are slightly out of time or tune, or of those playing less consonant intervals. For the practically significant problem of separating simultaneous voices, these two dimensions of pitch and asynchrony are of prime importance.

Pitch
If the speech of two talkers is artificially modified to a monotonous pitch and then mixed together, it is considerably harder to hear one target voice if the voices are speaking on the same pitch than if each is on a different pitch. A difference of four semitones increases the number of content words correctly reported from about 40% to about 60%. In normal, natural speech there are many stops and starts of the voice — during stop-consonant closures and pauses — so that some of the job of perceptually separating the two speech streams could be achieved using common onset time as a grouping principle, and by listening to one voice during the gaps in the other. If the speech is constrained so that it has no stop-consonants or pauses, it becomes more difficult to hear the target voice when the two voices are speaking on the same fundamental, and the improvement with a difference in Fo is correspondingly greater.

We can distinguish two ways in which a difference in pitch might help the listener to separate the simultaneous sounds of two talkers: helping the listener either to identify formant peaks, or to group together formants from the same vowel. Both these processes are impaired when two vowels are spoken on the same pitch. To illustrate the first point, the two lower curves of Fig. 3 show the spectra of two vowels, /i/ as in ‘heed’ and /u/ as in ‘hood’, sung on a steady monotone of 150 Hz so that they both have harmonics at the same frequencies. The frequency of the first broad peak in the spectral envelope of each vowel (the first formant) is important for vowel identification. The upper curve shows the spectrum of the sum of the two vowels (raised by 20 dB for clarity). The first formant peak of /u/ has disappeared. If the vowels were on different fundamentals, their harmonics would form different harmonic series and could, in principle, be at least partly separated.

The pitch of a complex sound is determined predominantly by the frequencies of its resolved harmonics, and it is likely that the detailed time pattern of auditory nerve firing conveys information about these frequencies to the brain. In the absence of any resolved harmonics, the brain can perceive pitch solely on the basis of the beating of the unresolved harmonics at the fundamental frequency. Information about this beat rate is conveyed in the auditory nerve firing pattern. Human pitch discrimination is, however, an order of magnitude better when based on resolved harmonics than when based only on unresolved harmonics. In male speech the resolved harmonics occur in the frequency range of the first formant (less than about 1 kHz), with the higher formants having largely unresolved harmonics. Female speech
is generally at a higher fundamental frequency (with a relatively smaller increase in formant frequencies), so resolved harmonics are more evident in female speech in the higher formants.

The separation of two voices on the basis of fundamental frequency is also sensitive to the difference between resolved and unresolved harmonics. For small differences in Fo (less than two semitones) between two speakers, there is considerable improvement compared with unison, but the improvement is due to better formant frequency estimation among the resolved harmonics, rather than to the grouping together of different formants that share the same Fo. Only at larger Fo separations does a difference in Fo help listeners to group together those formants that have a common Fo (Refs 12,17).

**Frequency modulation**

Common movement is a very powerful cue for segregating visual scenes, but the auditory system is surprisingly insensitive to its obvious auditory analogue—common frequency modulation. Although listeners are very good at detecting slight misstunings of one or more components of an otherwise harmonic sound, it is surprisingly difficult to tell whether two inharmonic sounds are being frequency-modulated coherently or incoherently18. When two concurrent voices are heard, with independent changes in their fundamental frequencies, listeners do not use the patterns of common movement within the two subsets of harmonics to segregate the voices19.

**Location**

Classically, the two main cues for localizing a sound in azimuth (along the horizon) are the differences in intensity and in phase of a sound at the two ears. Intensity cannot be used as a cue at low frequencies for sounds heard naturally because the head casts little of an acoustic shadow for frequencies that are comparable in wavelength to its size. For complex sounds, such as speech, the dominant cue is the timing difference in the low frequency region20, brought about by the time it takes sound to travel round the head. Surprisingly, such interaural timing differences (ITDs) are remarkably ineffective at producing simultaneous grouping. We cannot segregate simultaneous speech sounds purely on the basis of differences in ITD. For example, if listeners hear four formant-like noise bands that can make different pairs of vowels depending on how they are combined, the vowels that they hear are not influenced by which pairs of noise bands have the same ITD (Ref. 21). Similarly, we have shown that although a single, resolved harmonic in a vowel can be segregated from the vowel by altering its onset time or its tuning, it cannot be segregated solely by altering its ITD (Ref. 22).

Such weak simultaneous grouping by ITD contrasts with the substantial improvement in intelligibility that occurs when a target sentence comes from a different direction from distractor speech. Some of the intelligibility gain in the natural condition comes from head shadowing improving the signal-to-noise ratio at one ear, and some from sounds being more detectable through binaural interaction22,24. We have recently shown that ITD can also help the listener to attend across time to one sound source in the presence of another25. The experiment used two different carrier sentences, presented on every trial. The listeners’ task was to attend to one of the sentences and to say which of two possible target words occurred in the sentence. The other target word was in the other sentence. The two sentences were spoken by the same speaker and could differ in their (monotonous) fundamental frequency and/or their ITD. When the sentences differed in ITD by only ±45 μs, subjects were very good at telling which target word came from which sentence. An ITD difference of ±45 μs corresponds to an angular separation between the two sources of about 10 degrees.

In striking contrast to their inability to use ITD for simultaneous grouping, subjects continued to use continuity of ITD to define a particular sound source when it was opposed by a difference in fundamental frequency. If the target word that shared the same ITD as the attended sentence had the same fundamental frequency as the other sentence, subjects continued to report the target word with the same ITD. This result is surprising because, as we saw above, in even the presence of different vowels, subjects do use continuity of fundamental frequency to define a particular sound source across time6.

The evident power of ITD to define a sound source across time contrasts with the weakness of ITD to specify simultaneous perceptual groups. This contrast might reflect a distinction between the use of ITD at two different levels.
of auditory processing: one where the auditory system calculates the ITDs of individual frequency components, and a second, later stage where auditory perceptual objects are assigned subjective locations.

**ITD at different processing stages**

The first stage, determining the interaural time difference in individual frequency channels, is very well established both psychophysically and physiologically. Coincidence detectors in the brainstem that receive input from both ears via a delay line (Fig. 5) will respond best when the delay line compensates for the interaural delay introduced by sound travelling different distances to the two ears - a form of cross-correlation. However, there are two reasons why animals might not want to use these raw time differences at their face value. The first involves the inherent ambiguity of the coincidence information; the second involves the stability of the percept of a complex sound.

In any one frequency channel, a given time-delay between the ears will excite more than one coincidence detector because the signal in that frequency channel will be somewhat periodic. If the frequency of the channel is \( f \) and the actual delay between the ears is \( T \) then coincidence detectors with delays of \( -T \sin f \) will be excited (provided that the auditory nerve is able to indicate the signal's phase by means of phase locking). If the frequency is low, the ambiguities fall outside the range of possible ITDs and are not a problem. For higher frequencies, the problem can be solved by comparing information from the same sound source across frequency channels. Because \( f \) varies across these channels, so will the delay associated with coincidences at \( -T \sin f \) when \( n \neq 0 \); the only coincidences that will be common across all these frequencies will be those for which \( n = 0 \). Across-frequency integration of coincidence information can thus remove phase ambiguity. Cells in both the owl and the cat midbrain perform such across-frequency integration (though in somewhat different ways) enabling them to recover the true ITD for broadband signals\(^{26,27}\).

Human listeners can also localize sound by performing a similar across-frequency comparison of binaural coincidence information. Jeffress\(^{28}\) made the striking observation that the subjective location of a bandpass noise that is delayed to one ear can shift from the leading side to the lagging side simply by changing the pass band from narrow to broad.

For example, if a narrow band of noise around 500 Hz leads in the right ear by 1.5 ms, it will actually be heard on the left, lagging side. The reason (Fig. 6) is that the narrowband noise has enough periodicity at 500 Hz to give substantial coincidence at a delay of 0.0015 - \( 1/500 = -0.0003 \) s = -0.5 ms. Coincidence at this delay is apparently preferred by the system to the actual delay of +1.5 ms because it is closer to zero. However, as the bandwidth of the signal is increased, the consistency of the +1.5 ms coincidences across frequency becomes stronger and the sound moves across the head to the leading side\(^{29}\). In human listeners, for whom phase ambiguity is less problematic than it is for owls, the advantage of such an across-frequency
comparison of ITD might be to produce a reliable and stable subjective location for complex sounds in an environment where echoes can produce large changes to ITDs.

Across-frequency comparison of ITD will only work in practice if the comparison is confined to frequencies that have come from a single sound source. There is evidence from human listeners that this is indeed the case\(^4\). The Jeffer effect can also be obtained by varying the number of harmonics in a complex tone rather than by varying the bandwidth of a noise. As described above, a 500-Hz tone that leads by 1.5 ms in one ear will, through phase-ambiguity, be heard in the lagging ear. If this tone is now flanked by six harmonics of 100 Hz steps (from 200 to 800 Hz) which are also delayed by the same amount, listeners hear a single complex tone on the side of the leading ear. However, if the 500 Hz component is perceptually segregated from the other harmonics, either by slightly mistuning it, or by delaying its onset, then it is heard as a separate sound source localized back in the lagging ear. The perceptual segregation of the 500 Hz component from the remaining harmonics has prevented it from sharing with them across-frequency comparison of ITD.

We can now return to the problem of why ITD is a weak cue for defining simultaneous sound sources, but a strong one for defining a speech source across time. One possible resolution of the difference is to recognize that the tasks on which ITD provides a strong sequential grouping cue are ones where ITD is giving a location to complex sounds (such as natural speech) that have already been grouped by a number of other cues, such as harmonicity, common onset time and so on. These cues define the frequencies across which a stable estimate of the location of the sound source can be made by combining ITDs. Once this estimate has been made and other considerations, such as interaural intensity differences and, perhaps, visual cues\(^3\), have been taken into account, the listener can attend to a particular spatial location. Conversely, on simultaneous grouping tasks, where ITD is the only cue that the listener has for defining different sound sources, it provides only weak grouping.

Schema-based perceptual segregation

Bregman\(^1\) distinguishes between a bottom-up process of sound segregation based on the automatic operation of primitive grouping mechanisms, and a top-down process that uses more specific schema-based knowledge. An example of schema-based knowledge being applied to integrate sounds, which would otherwise be heard as separate sources, is in sine-wave speech\(^1,3\). Sine-wave speech is analogous to a line-drawing of an object in that an abstract property of the stimulus is made explicit. The time-varying formant frequencies (the resonant frequencies of the vocal tract that serve to enhance some harmonics and attenuate others) are represented explicitly in sine-wave speech as frequency-modulated sine waves. A spectrogram of a sine-wave speech rendition of the sentence ‘My dog Bingo ran around the wall’ is shown in Fig. 7.

Although auditory grouping cues are minimal in sounds of this type (with the possible exception of common onset time), listeners can nevertheless hear the underlying speech. Does this observation render the concept of auditory grouping irrelevant for speech perception? Two points can be made. First, sine-wave speech caricatures a very abstract aspect of speech: the formant frequency trajectories. It is a difficult problem to extract these tracks even from the clear speech of a single talker. Heuristics for auditory grouping, such as those described in this paper, are demonstrably useful in performing the sort of signal processing that precedes such an abstract representation as the formant trajectory. Second, when additional information relevant to auditory grouping is added to the sine-wave speech, its intelligibility improves substantially\(^5\). On the other hand, it is likely that a sound source as specialized as the human vocal tract does impose unique constraints on the production of sound, which the human listener has probably learned to exploit\(^5\).
Conclusion

Work on auditory grouping has succeeded in identifying a variety of simple cues that help listeners to segregate frequency components that originate from different sound sources. These cues are useful both for the segregation of simple tonal patterns and for the more complex sounds of speech. The questions raised by the need to segregate sound into different putative sources link basic psychophysical work on complex sounds with the much less developed field of auditory attention. Increased understanding of basic psychophysical constraints on the processing of highly constrained complex, artificial sounds on the one hand, together with experiments using more natural stimuli on the other, are helping to identify how the information available at the auditory periphery is used by more central mechanisms.

References

12 Bird, J. and Darwin, C.I. Effects of a difference in fundamental frequency in separating two sentences, in Psychophysical and Physiological Advances in Hearing (Palmer, A.R. et al., eds), Whurr (in press)
33 Remez, R.E. et al. (1981) Speech perception without traditional speech cues Science 212, 947–950
39 Denbigh, P.N. and Zhao, J. (1992) Pitch extraction and separation of overlapping speech Speech Comm. 11, 119–126

Students

Subscribe to TICS at a 50% discount using the bound-in card