This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

Chapter 6

Auditory Organization in Speech Perception

I now want to examine the role of the primitive scene-analysis processes in the perception of speech. As the discussion proceeds, however, we will see that the contribution of primitive processes is obscured by the contribution of speech schemas. I described some general properties of schemas in chapter 4. Among them was the property of being able to extract what is needed from mixtures. Since both primitive grouping and schema-based grouping operate at the same time, it will be difficult to know what is primitive and what is not. The strategy that I will follow is that when the organizational processes are the same in speech as in simpler signals, I will assume that they derive from primitive processes, whereas when particular capabilities for perceptual isolation are found in speech, I will assume that these are schema-based. However, I will not take any position on which schemas are innate and which are learned.

In 1953, Colin Cherry, a British researcher working at the Massachusetts Institute of Technology, reported research on what he called "the cocktail party problem."[665] How can we select the voice of a particular talker in an environment in which there are many others speaking at the same time? He did a number of experiments in which a person, listening over headphones, had to report what a recorded voice was saying when it was accompanied by a second recording made by the same talker. He found that two factors affected the ease with which this could be done. The first was whether the recordings were sent to the same headphone or to the opposite one. It was much easier to follow one voice when the other was sent to the opposite ear. The second was that when the two recordings were sent to the same ear, it was easier to follow one of them when the next part of what it was saying was predictable from the previous part. When this predictability decreased, the listener often switched attention to the other recording. The act of segregating and following one of the two recordings was called "filtering."

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

Cherry proposed other factors that would make it easier. He mentioned the assistance offered by cues such as differences in the quality of the two voices, differences in their mean speeds or mean pitches, differing accents, and even visual cues such as lip reading. Other researchers showed that when Cherry had made segregation easy by sending the messages to different ears, it was not the fact that each ear got only one signal that was important, but that the perceived spatial origins of the two sounds were different.[666] The same researchers showed that high-pass filtering one message above 1,600 Hz and low-pass filtering the other below that frequency also allowed an easy segregation of the messages.

Cherry mentioned several factors that he thought would help us to segregate voices. Some were raw physical qualities, but others had to do with the ability of the listener to predict the next moment of speech from the previous one. It is not clear whether this ability to predict is really used in segregating the message from others at the basic scene-analysis level or is only used in matching the sound to memories of words, phrases, and so on. However, there is no doubt that it is governed by the listener's knowledge of speech and language—a property that labels it as schema-based integration and therefore excludes it from the present discussion. I want to focus on an examination of the role of primitive scene-analysis processes in the separation of speech sounds.

A great number of theories were invoked to account for the selective attention demonstrated by Cherry. However, they were all similar in certain ways. They all mentioned that the physical properties of one of the voices could be used by the listener's attention to select that voice, and they all seemed to presuppose that factors such as location, pitch, timbre, and loudness were simple and easily available. Yet if we made a stereophonic tape recording at the position of the listener's ears at a cocktail party and then made a spectrogram from each channel of that recording, the resulting pictures would look very unlike the patterns that phoneticians have come to associate with speech sounds. Nor would the pitch and timbre of individual speakers' voices or their locations be obvious from an inspection of the pictures. Yet these were the "physical factors" that were assumed to be the strongest bases on which the process of attention could separate the mixture and track the desired target.

It is clear that we must introduce scene analysis as a preliminary process that groups the low-level properties that the auditory system extracts and builds separate mental descriptions of individual voices or nonvocal sounds, each with its own location, timbre, pitch, and so

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

on. Only then does it make sense to say that our attention can select a voice on the basis of one of these qualities.

What I am arguing is that factors such as pitch, timbre, and location are the *results* of segregating the mixture, not the *causes* of its segregation. "But," one might reply, "is it not true that a particular voice has its own spatial location, fundamental frequency, and so on, and that it is these properties that allow us to select it from the mixture?" The apparent contradiction, as in other cases that we have examined, is resolved by drawing a distinction between the physical and psychological realms. It is true, for example, that every signal at the cocktail party has a physical place of origin, but this does not guarantee that this origin is represented in the perceptual domain or that all the information that has been received from that origin will contribute to the appropriate mental description. For example, in certain illusions, the evidence received from one part of space is assigned to a perceptual sound that is heard as being at a different location.

A more exact account of causes and effects would be to say that the physical place of origin may be responsible for some physical properties that the received signal has, and if this evidence is used correctly, the perceptual image of the sound will have a mentally represented place that corresponds to the physical place and a set of perceived qualities that adequately represent the physical properties of the source. Until this is accomplished, however, there is no mentally integrated sound with its own location label that can be selected by higher mental processes.

If we look at the spectrogram of a mixture of sounds such as figure 1.4 of chapter 1, we see that there are two dimensions on which acoustic information must be grouped. The first is over time, to reconstruct the temporal pattern of the sound. The second is over the spectrum. Without the partitioning of evidence that is accomplished by such grouping, the evidence for particular speech sounds can be invisible. Christopher Darwin has given a nice example to show that the array of sound will be heard as having the phonetic patterns that we are familiar with only after it is partitioned into streams:

> Knowledge about the properties of phonetic categories [such as the phoneme "b"] must be represented by properties of the sound produced by a single . . . speaker. Yet properties that are apparent in the raw waveform are not specific to a single speaker or sound source; they are properties that are due to whatever sound sources are present at the time. For example, the silence necessary to cue an inter-vocalic stop consonant [such as the "b"

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

> in the word "about"] is silence of a single sound source; there may be no actual silence present in the waveform.[667]

We want to understand the segregation of speech sounds from one another and from other sounds for many practical as well as theoretical reasons. For example, current computer programs that recognize human speech are seriously disrupted if other speech or nonspeech sounds are mixed with the speech that must be recognized. Some attempts have been made to utilize an evidence-partitioning process that is modeled on the one that is used by the human auditory system. Although this approach is in its infancy and has not yet implemented all the heuristics that have been described in the earlier chapters of this book, it has met with some limited success. I will describe a number of these approaches in this chapter.

Sequential Organization of Speech Sounds

In this section we will look at the sequential integration of the speech signal. The rapid sequence of different types of sounds coming from a particular talker has to be held together into a single stream and, at the same time, must not connect up sequentially with the sounds coming from a different talker.

At the very microscopic level, even the identification of many of the speech sounds themselves depends on the relevant information being assigned to the same perceptual stream. For example, in the phrase "say chop", there is a brief silence before the "ch" noise burst that tells the listener that it is "ch" rather than "sh". The silence tells the listener that there has been a closing off of the air flow of the talker's voice. Yet the listener must interpret the silence as occurring between speech sounds made by the same voice. If one voice stops and another one starts, this does not signal a closure. An experiment done by Michael Dorman and his colleagues shows that the heuristics of scene analysis can contribute to the correct interpretation.[668] If the pitch of the voice changes suddenly from that of a male to that of a female between the two words, the perception of "chop" does not occur. Listeners will hear "shop". The dip in intensity signals "ch" only when it is interpreted as a within-stream dip and not the product of an accidental concatenation of two distinct sound sources.

Sequential integration also has to operate on a longer time scale. In the early observations by Cherry, in which a person was asked to shadow one verbal message while ignoring a second, whenever the acoustical basis for segregation was not good (for example, when the two messages were spoken by the same talker and not spatially segre-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

gated) the listener would frequently switch from tracking one message to the other. The problem did not seem to be one of segregating the simultaneous sounds; otherwise the listener would not have been able to track either message. It seemed, instead, to be a problem of the sequential grouping of the words from a single talker.

The need for sequential integration of speech sounds introduces a serious problem for primitive scene analysis. Speech is a succession of qualitatively different sounds. For example, an "s" is a type of noise burst whereas an "o" is a type of tone with a harmonic structure. We know that a noise burst will sequentially segregate from a tone; so why do the sounds of the speech stream hold together?

The basic building blocks of speech are usually described as phonemes, which are divided into vowels and consonants. For our rough purposes here, phonemes can be thought of as the simple sounds that are indicated by single letters or pairs of letters in writing. Examples are "s", "b", "ee", "sh", and "th". Phoneticians view these as the basic elements out of which the words of a particular language are built. Richard Warren gives the average rate of phonemes in normal speech as 10 or more per second.[669] We ought to recall that at these rates it is easy to get stream segregation of alternations of high and low pure tones. Yet Warren points out that speech is intelligible at much faster rates. Machines have been constructed that can speed up speech without altering its pitch. It has been reported that listeners can be trained to comprehend speech at a speed of about 30 phonemes per second without temporal order confusions.[670] Speeded-up speech remains partially intelligible even at four times the normal rate.[671] At these rates, if we listened to an alternation of high and low tones, substreams would form and would seem utterly unrelated. It is evident that speech is more cohesive than such tonal patterns. Still it is not utterly resistant to segregation. Van Noorden has reported that when speech is speeded up by a factor of 2.5 or more it tends to segregate into substreams.[672] Presumably the ability to perform at higher rates than this requires some training to allow the schema-based recognition process to overcome the effects of primitive segregation.

Warren has been struck by the inconsistency between peoples' performance on two tasks.[673] On the one hand they are able to understand a rapid sequence of speech sounds; on the other, they are unable to report the order of unrelated sounds (for example a hiss, a buzz, a whistle, and a vowel) in a short cycle. Yet the rate of sounds in the cycle of unrelated sounds may be much slower than the rates of phonemes in speech.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

It would seem that in order to understand speech, the listener would have to be able to determine the order of its sounds, because the same sounds in a different order would have a different meaning (for example, "serve" and "verse").[674] In an attempt to reconcile this contradiction between performance on speech and nonspeech sounds, Warren argues that listeners to a cycle of unrelated events have to decompose the signal into constituent parts, recognize each part, and then construct a mental representation of the sequence. Listeners to speech do not have to go through this process. Rather they can do some global analysis of the speech event and match it to a stored representation of the holistic pattern. After all, Warren continues, children can recognize a word and often have no idea of how to break it up into its constituent phonemes.

Even if this explanation were true, however, it would not solve the problem of stream segregation in speech. It seems that even direct recognition of holistic patterns should require that the parts of the sound that enter into the same analysis be part of the same stream. We have seen in chapter 2, for example, that listeners who are listening for a pattern that they have just heard have a great deal of difficulty if the tones that form the target pattern are absorbed into separate streams. The converse is also true; if irrelevant tones are absorbed into the same stream as the target pattern, they camouflage it.

Furthermore, only tones in the same stream enter into our recognition of another holistic property of a stream, its rhythm. It is likely that when we listen for a familiar melodic pattern or rhythm in a sequence, we can recognize it only if we can perform a global analysis on the pattern isolated from co-occurring ones. This was true, for example, for the familiar melodies studied by Dowling. As the reader may recall from chapter 5, if the notes of two familiar melodies were interleaved in time, there had to be a segregation of the tones of the two melodies into two separate streams, each containing the notes of one melody, before recognition was possible.. These facts about the recognition of tonal patterns for which recognition schemas are already available suggest that the recognizability of speech sounds as global patterns would not, in itself, exempt speech sounds from the effects of primitive grouping.

If speech holds together better than sequences of arbitrary sounds, it may not simply be due to the fact that schemas exist for the speech. Some years ago I became interested in how an infant who was in the process of learning language by being presented only with continuous streams of speech could discover what the component words were. Chapter 2 contains a description of the experiment I did. To briefly

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

recapitulate, I simulated the baby by using myself as a subject. I made a tape recording of a small number of words repeated many times in a random order without pauses. I played the tape backward to destroy the familiarity of the sounds (that is, their fit to preexisting schemas). After listening many times I was eventually able to discover (but not identify) a number of the backward words. However, I was not able to do the same thing with a similar tape that I made out of nonspeech sounds. I had created this by making artificial words that had 0.1-second mechanical sounds substituting for phonemes. In this sequence, the sounds did not cohere sequentially and I could not discover the repeating artificial words. Evidently there was something very different between a speech sequence and a sequence of artificial sounds.

People have tried to examine the extent to which speech sounds hold together by listening to rapid repetitions of short cycles. I have made some informal observations in my laboratory using synthesized vowels ("ee" and "ah") having the same fundamental frequency. I found that a rapid sequence in which the two occur in alternation split readily into two streams, one containing each vowel. This segregation may have been due to some simple factor such as brightness differences or to the fact that peaks in the two vowel spectra were at different frequencies.

Some of the research with cycles of speech sounds was stimulated by the finding of Richard Warren and his colleagues that listeners had a great deal of difficulty in reporting the order of events in a rapid cycle of unrelated short sounds, but could report the order more easily if the short events were spoken digits, each recorded separately and then spliced together to form a cycle.[675] Cycles that consisted of four different tape-recorded natural vowels, played at a rate of 200 msec per component, were also studied. It was very difficult to report their order if the individual 200-msec segments were chopped out of the steady-state portion of a longer vowel and the segments abutted against one another to form the cycle; it was easier if each individual segment was reduced to 150 msec and a 50-msec silence was introduced between segments; it was easiest of all when each segment was a full vowel with its own natural onset and decay. Further studies were able to detect some slight ability to report the order of steady-state vowel segments even when the segments were as short as 100 msec per segment provided there were brief silences between them.[676] This should be contrasted with two facts. First, it is much better than performance on a cycle of wholly unrelated sounds (such as hiss, buzz, and tone) at these rates. Second, it is much worse than the ability to integrate phonemes in rapid connected speech.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

The superiority in detecting the order of vowel sequences, as compared with unrelated sounds, may come, at least in part, from the ability of the listeners to encode a cycle of vowel sounds through the use of language skills. It is possible that they can hear the cycle as a four-syllable verbal utterance and use their verbal skills to remember it. In the case of cycles of fully meaningful words such as digits, verbal skills would be particularly effective. Verbal encoding would help account for why having natural beginnings and endings on the vowels makes the cycle easier and would explain why other researchers, using recycled sequences of vowels or vowel syllables, have found that the best performance occurred when the sounds were closest to those in normal speech.[677]

James Lackner and Louis Goldstein did an experiment to find out whether the existence of stop consonants (such as "b" and "p"), which introduce brief silences and rapid spectral transitions between the vowels, might assist in the detection of the order of sounds in a cycle.[678] They expected that it would be easier to distinguish the order when vowel (V) syllables such as "oo" were alternated with consonant-vowel (CV) syllables such as "bee" than when the cycle was formed entirely of vowels. The syllables were each 200 msec in duration. To their surprise, the hardest case was when two CV syllables, both of which started with the same consonant, alternated with two different V syllables (for example, "dee-oo-bee-ah-dee-oo-bee-ah-. . .").

The level of performance on such cycles was below chance. This occurred because the CV syllables formed one stream (such as "bee—dee—bee—dee—. . .") and the vowels formed another stream (such as "oo—ah—oo—ah—. . ."). They interpreted this as an example of stream segregation. In support of this, they reported that the listeners usually described the order of the cycle in a stream by stream order, first one of these two classes and then the other. Apparently the onsets of the consonants, rather than merely inserting convenient marker points into the sequence, had the effect of labeling some of the syllables as similar and causing them to group. This suggests that the onsets of sounds may play a role in grouping them over time, those with sudden onsets (such as syllables formed of stop consonants and vowels) segregating from those with smoother onsets. In nonspeech sounds, such a tendency would help to group a succession of events that had a common physical cause, such as the successive cries of the same bird or animal. Lackner and Goldstein also found that if recordings of the same syllable spoken by a male and a female were alternated, the two syllables formed separate streams. This segregation

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

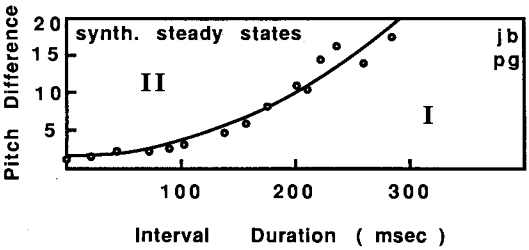

Figure 6.1
Stream segregation of vowels of alternating high and low fundamental frequencies. The horizontal axis shows the length of silence between the 100-msec tones, and the vertical axis shows the pitch difference in semitones. In region I, the listeners definitely hear one stream of vowels. In region II, they definitely hear two streams. (From Noteboom, Brokx, and de Rooij 1978.)

was probably based on differences in fundamental frequency, but since the syllables were recordings of natural speech, there may have been other differences as well.

Noteboom and his colleagues at the Institute for Perception Research in Eindhoven systematically studied the effects of differences in fundamental frequency on stream segregation of vowels.[679] They presented subjects with a sequence of nine artificially synthesized vowels. The vowels had steady pitches as though they had been sung. In the sequence of nine vowel sounds there was an alternation between a higher and a lower pitch. They varied the separations of the high and low fundamentals between 0 and 20 semitones. The duration of each vowel was 100 msec, but they also inserted silences ranging between 0 and 400 msec between successive vowels. When the jumps in pitch (fundamental frequency) were large and the silences short, the vowels segregated into two different streams as if sung by different singers.

The streaming of these vowels showed exactly the same pattern as the streaming of pure tones that we encountered in chapter 2. The results from two listeners are shown in figure 6.1. There was a tradeoff between pitch separation and temporal separation. It took less pitch separation to segregate the vowels if the sequence was more rapid.

Continuity of the Fundamental

When one takes either natural or synthesized vowel sounds and splices them together either with or without an intervening silence, there is more than one kind of discontinuity. The pitch may be dis-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

continuous but so are the peaks in the spectrum. There are also overall discontinuities in loudness when silences are introduced into the cycle. Therefore, when we observe that pitch differences can cause segregation, we wonder whether this would have happened if the other discontinuities had not also been present. It is possible to answer this question, but not with isolated vowels. We can use natural speech or an elaborate method that employs speech synthesis.

Although the vowels are not exactly complex tones, they are quasiperiodic, each repetition of the waveform being almost, but not quite, like the one before it. From here on, I will refer to the vowels as periodic sounds and let the reader supply the prefix "quasi" as desired. Because of the periodicity of vowel sounds, the human auditory system can hear them as having a pitch. This pitch will change over time, and the resulting *pitch trajectory* is experienced by the listener as a pattern of intonation. This pattern is governed by at least two constraints. The first is that the pitch of a voice changes relatively slowly. The second is that it follows the melodies that are inherent in the patterns of grammar and meaning that are part of any language. The first constraint is more general; the second is specific to a particular language. We know that the human listener takes advantage of both constraints to follow a voice over time.

Darwin used natural voices to study the effects of pitch continuity on sequential integration. His experiment was stimulated by an experiment reported in 1960 by Anne Treisman, who had asked listeners to shadow one passage of continuous prose while ignoring another simultaneous one.[680] In her experiment, the target message was sent to one ear over a headphone while the competing message was sent to the other. The listeners were asked to shadow the material that was sent to one ear (the "target ear"). At some point in the passage, the channels were suddenly switched so that the two passages continued on smoothly but in the ears opposite to those in which they had started. As a result, the listeners sometimes repeated the first few words after the switch from the nontarget ear. This showed that they could not prevent their attention from switching ears with the passage that they had been shadowing, even though they had been instructed to follow an *ear*, not a passage. Treisman used this result to argue that the tracking of voices in mixtures could be governed by the meaning content of the message.

Fifteen years later Darwin pointed out that when one of the passages was switched from the left to the right ear, not only did the material in the right ear continue the meaning of the passage that was formerly in the left, but it also continued its pitch contour.[681] He set about separating the effects of these two kinds of continuity. Like

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

Treisman, he sent different passages to the two ears of listeners and asked them to shadow the material sent to one of their ears. The material sent to the two ears was read by the same female speaker, but for convenience of explanation I will refer to the material in the two ears as being from two different voices. In one condition he introduced the semantic switch without any within-ear discontinuity of pitch. That is, at the switching point the voice in each ear took over the passage that had previously been spoken by the other voice but did so without a break in the pitch contour. In another condition, he introduced a pitch contour switch without any break in the semantic content. Another condition replicated Treisman's condition in which both semantic content and pitch contour changed together. Finally there was a control condition in which there was no switch at all.

Both kinds of switches induced the listeners to make errors, but the type of error seemed to depend on the type of switch. When the semantic content switched, the listeners were likely to miss some words from the ear that they were shadowing (presumably because the new words in that ear did not continue the previous ideas and therefore were unexpected) but they did not usually report words from the other ear. On the other hand, when the pitch contour suddenly was switched to the other ear, this often caused the listeners to report a few words from that ear. Apparently the continuity of pitch contour was controlling their attention to some degree.

In Darwin's experiment, the sequential integration was shown to be affected by both factors. It appeared that a primitive pitch-following process seemed to work in concert with a more sophisticated one that dealt with meaning. Since this book is about the primitive process, I will have little further to say about the more sophisticated one.

Very striking examples of the role of the continuity of the fundamental frequency in speech perception can be created by artificially synthesizing speech. Speech can be synthesized in two stages: The first stage simulates the creation of acoustic energy in the vocal tract and the second simulates the filtering of this sound by the shape of the vocal tract (the throat, mouth, and nose).

In the first stage, the initial sound takes one of two forms. The first is turbulent hissing noise created by forcing air through a constriction. This happens, for example, when we say "s". The second form is the periodic (pitch-like) sound that is created by our vocal cords, for example when we say "ee". The periodic sound is present whenever our vocal cords are active—that is, whenever there is voicing. For example, all vowels are predominantly periodic. Some consonants, such as "w" or "l", are also voiced. Consonants such as "s",

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

"f", "sh", and "t" are created entirely from turbulent noise without any presence of vocal cord sound and are therefore called "unvoiced." Some sounds, such as "z", involve both periodic sound and turbulent noise.

The pitch of speech is created at the sound-creation stage by the fundamental frequency of the voicing. The pitch contour is imposed at this stage by changing the fundamental frequency of the voicing over time.

At the second stage, the turbulent or vocal cord sound undergoes filtering that represents the effects of the changing positions of the throat, mouth, and nose on the signal. The filtering enhances certain bands of frequencies and imposes a pattern of peaks (formants) and valleys onto the spectrum. This pattern tells the listener which particular shapes have been assumed by the vocal tract of the speaker and hence which vowels and consonants have been spoken.

The vowels are distinguished from one another by their patterns of formants. A good replica of a vowel may be synthesized by starting with a periodic waveform that is rich in harmonics, with its higher harmonics less intense than its lower ones, and then passing this periodic sound through a series of resonators that simulate the filtering effect of the vocal tract. If the periodic sound that is fed into the resonators has a steady fundamental frequency, the resulting sound is experienced more as a tone than as a vowel. To make it more like a vowel, its fundamental frequency should be modified in two ways: (1) It should be jittered randomly by less than 1 percent to simulate the fact that the successive periods of the human voice are not all identical. (2) It should vary slowly in frequency to emulate the larger pitch variations of the human voice. Another variation that occurs in the spectrum of a vowel over time is in the position of its formants. Because the vowel is spoken in conjunction with other sounds, the movement of the speech apparatus from one sound to the next introduces movements in the formants.

The second stage, being purely a filtering operation, does not change the fundamental frequency of the periodic portions of the sound. Therefore it does not change the perceived pitch of the voiced sounds to any appreciable degree.

To summarize, the pitch of a vowel or voiced consonant is determined entirely by the fundamental frequency of the vocal cord vibration, while the identity (which phoneme it is) is determined by the filtering in stage 2. This means that in a two-stage synthesis of artificial speech, the pitch can be changed in stage 1 without changing which phonemes are heard.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

It has been noted that when individual words are prerecorded and then a sentence is made by splicing the individual words together, the resulting speech is often unintelligible. It sounds garbled. Listeners often report hearing the speech coming from different directions or different speakers and they often mistake the order of sounds.[682] It was thought that it was the discontinuity of fundamental frequency that caused the perceptual failure to integrate the sequence. Darwin and Bethell-Fox investigated this hypothesis. But rather than splicing different sounds together, a technique that introduces discontinuities in other properties as well as in pitch, they used the method of speech synthesis to demonstrate a dramatic effect of a sudden discontinuity in the fundamental frequency.[683]

Their stimulus pattern was a voiced spectrum that had a number of successive parts. In the middle section was the formant pattern that is characteristic of the vowel "a" as in "cat". It was held steady for 60 msec. Before and after the "a", and separated from it by 60-msec spectral transitions, were formant patterns that were characteristic of another vowel (for example, "u" as in "put"). That is, the first and third vowels were the same (and were not "a", of course). The spectrum of the two flanking vowels also stayed steady for a brief period. The parts that lay between the "a" and the flanking vowels were transitions of the spectrum that smoothly shifted the position of the formants from those of the first vowel to those of the "a" and from the "a" to those of the third. Each of these two transitions lasted for over a period of 60 msec. For example, the sequence might be represented as "u" . . . "a" . . . "u", with the dots representing the transitions.

Recall that these spectral patterns are due to the filtering stage of the synthesis and are independent of the fundamental frequency (pitch) of the voicing. When the pitch was held constant throughout the stimulus (at 130 Hz), the listeners heard not a sequence of three vowels (such as "u-a-u") but a syllable such as "wow" in which the "u" sound becomes a "w". Some consonants (for example, "w") simply involve a smooth change in the spectrum from a value that is sometimes used for a vowel ("u" in the case of "w") to the spectrum of the vowel that is adjacent to it. Such consonants are called sonorants and include "w", and "y", among others. The "wow" in our example was heard because the "u" and its transition were perceptually integrated with the vowel "a".

In another condition, the pitch had two different values, a high and a low one (101 and 178 Hz), and an abrupt shift was made from one to the other halfway through the transition on each side of the "a". For example, the pattern could be high-low-high, the pitch shift from

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

high to low occurring halfway through the "u-a" transition and the shift back up again occurring half way though the "a-u" transition This prevented the listeners from integrating the total formant pattern into a single syllable. Instead they heard (in our example) two low-pitched syllables and a high one. Apparently they began to hear a new voice at each place at which the pitch changed. When they were asked what consonant preceded the "a" they reported ones such as "b". Apparently they heard the short spectral transition (the part that adhered to the "a" when the longer transition was split in half by the pitch change) and interpreted it as the one that normally occurs at the onset of stop consonants such as "b", "g", and "d". The strong role that the smoothness of pitch change plays in holding natural syllables together is illustrated by the dramatic dissociation of the parts of the sequence when this continuity is violated.

It is easy to see why the voiced sounds of normal speech hold together during periods of spectral continuity. But we also have to account for why they hold together at moments when the spectrum undergoes a sudden discontinuity, as it does when a stop consonant such as "p" is interpolated between two vowels as in the word "repay". One obvious source of continuity is that the pitch contour of the sentence usually travels right through such interruptions, smoothly changing in pitch. But this holds only the two vowels together. What holds the intervocalic consonant in with them? Partly, this may be due to the fact that the air-stopping gesture of the vocal tract also introduces smooth transitions in the formants of the voiced sound just before and after the stop closure. These parts of the consonant may be integrated with the steady-state vowel formants as a result of smoothness of spectral transition. But what of the other parts, the silence that occurs during the brief closure of the "p" and the brief burst of noise as it is released? How does the pitch continuity affect our perception of these?

Suppose one synthesized a word in which the pitch that followed the silence was quite different from the pitch that preceded it, with no smooth transition to unite them. Extrapolating from the results of Darwin and Bethell-Fox, we would guess that the two pitches would be heard as separate voices.[684] It is possible that the silence would no longer be heard as a stop consonant embedded between two vowels. Indeed, it might not be registered as a silence at all, because its beginning would be interpreted as the end of voice A and its end as the onset of voice B. This example suggests that it is the continuity of pitch across a silence that allows it to be perceived as a closure of the vocal tract or even as a silence.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

An experiment on this issue was done by Michael Dorman and his colleagues.[685] It used a brief silence to cue the distinction between "sh" and "ch". In the sentence "Please say shop", if a silence of 50 msec or more was introduced before the "sh", the listeners tended to hear "Please say chop." However, if the words "Please say" were in a male voice and "shop" in a female, listeners heard "shop" despite the presence of the silent interval. The dip in intensity signaled "ch" only when primitive scene analysis decided that it was a within-stream dip and not the product of an accidental concatenation of two distinct sound sources.

Spectral Continuity

The acoustic nature of speech is commonly portrayed by means of a spectrogram. Figure 1.3 of chapter 1 showed a spectrogram of the word "shoe" spoken in isolation. The spectral peaks (formants) that continue over time are shown as dark streaks that run in a generally horizontal direction. Often a person viewing these formants thinks of them as if they were tones with their own pitches. However, formants are not pitches. In voiced sounds, where the formants are most noticeable (as in the vowel part of "shoe"), they are the groups of harmonics, all related to the same (much lower) fundamental, that happen to be emphasized by the vocal-tract resonances. In unvoiced sounds, which are turbulent noise (as in the "sh" part of "shoe"), the formants are enhanced bands of frequencies in the noise.

Since the formants are caused by the filtering that comes from the shape of the vocal tract, and this tract does not snap instantly from one setting to the next, the formants in successive sounds tend to be continuous with one another. This is most easily seen when the two successive sounds are both voiced, as in the syllable "you" shown in figure 6.2.

However, we do not always see continuity between the formants in noisy consonants such as "f" and the surrounding vowels. The formants in the noisy parts will be similar to those in the voiced ones only when the noise and the voiced sound have both been filtered in the same way. This will happen whenever both have traveled through the same cavities in the vocal tract on their way to the outside world. Since the voiced sounds are created in the vocal cords, and these are below the resonant cavities, voiced sounds travel through the whole system of cavities. However, different noisy sounds are created at different places. For example, the "h" in the word "aha" is created below all the cavities of the mouth and nose and therefore is filtered in the same way as a vowel. For this reason it will have formants that are continuous with those of adjacent vowels. On the

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

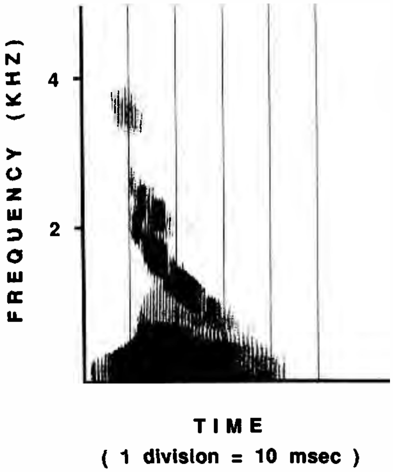

Figure 6.2
Spectrogram of the word "you" spoken in isolation.

other hand, the spectrum of "th" in "thaw", which is formed at the front of the mouth, will be affected only by the cavities in the very front of the mouth and will not have a formant structure that is strongly related to the "aw" that follows it. As the mouth moves between the positions that create the "th" and the "aw", that motion will be visible in the spectrum, but not always as the movements of clearly distinguishable formants. To summarize, the transitions between many speech sounds, but not all, will be visible in the form of continuous changes in the peaks visible in the spectrum.

Recall that I have been discussing formant trajectories and not trajectories in the pitch of the speaker's voice. Formant changes and pitch changes are separate acoustic phenomena. We have already seen how powerful the continuity in the fundamental of voiced speech sounds can be in holding them together. We can go on to consider the entirely separate question of whether continuity in the pattern of formants has a similar effect.

It was first reported by Peter Ladefoged that if a click was superimposed on a sentence, listeners could not tell where in the sentence it had occurred.[686] Our knowledge that the formation of separate streams can cause a loss of between-stream temporal relations leads us to believe that the click must have been heard in a separate stream from the speech. A similar effect occurs in the phonemic restoration phenomenon.[687] When a noise burst replaces a speech sound in a sentence, there are two effects on listeners. The first is that they will often

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

hear the partially obliterated word without any sense that part of it is missing. This is the actual phonemic restoration. The second is that the listeners will not be able to report, with any great accuracy, where in the sentence the noise burst has occurred. Again this is evidence for stream segregation. One is inclined to explain both these cases of segregation by appealing to the fact that there is a sudden change in the spectrum when the click (or noise) occurs, the spectrum of the click or noise being quite unlike that of neighboring speech sounds.

How, then, can speech sounds that resemble noise bursts or clicks remain coherent with neighboring speech sounds? For example, noise bursts such as "s" or "sh" occur in English, and clicks occur in African languages such as Xhosa. Since the meaning of words in these languages depends on the location of these clicks and noises in the words, the listener must not sort them into separate streams. The ability to hold the speech stream together may be partially due to learning; there are studies that suggest that stream segregation may be affected by practice. However, there may be subtle properties of the acoustic sequence that allow primitive organization to hold the stream together. I once tested myself by listening to a speaker of Xhosa (a click language that I do not understand at all) and experienced no difficulty in locating the position of the clicks relative to the other sounds in her sentences. This was done without visual cues and was verified as correct by the speaker.

It is worth noting that when a click is produced by speaking, it has a different acoustic relation to the surrounding speech sounds than it would if it were simply superimposed arbitrarily onto the speech stream. The spoken click occurs as a result of a consonantal gesture. The motion of the articulators that stops or releases the consonant will also have effects on the spectrum of the voicing that is going on. The voicing will therefore undergo major changes at the same time as the noise or click appears and disappears, and this synchronization of changes may tell the auditory system that the two effects are related to each other. A mechanically superimposed click, on the other hand, will often be placed so as to overlap a voiced sound (for example, in the middle of a vowel); so the normal covariation of changes will not occur.

The mere fact that one sound goes off at the same moment as another goes on, the situation that occurs at the boundary between speech and noise in the phonetic restoration stimulus, is not enough to tell the auditory system that the two sounds are related, nor should it be. The replacement of one sound by another of a different type often occurs when a loud sound masks a softer one coming from a different physical source. We have already discussed the cues for the

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

masking of one sound by another in the section in chapter 3 on the continuity illusion. Such cues cause segregation, not unification. The auditory system seems to be prepared to integrate two sounds only when the cues for masking are not present. If the transition can be interpreted as the beginning or end of masking, the replacement of one sound by another does not count as evidence that the two were affected by the same cause. However, when there are changes in the two sounds near the temporal boundary between them, such as sudden movements in the spectrum, these may be used as evidence that the disappearance of one sound and the beginning of another may have arisen from a common cause.[688]

There are a number of reasons for our interest in formants. One is that they are affected in fairly direct ways by the resonances of the vocal tract. Another is that, being compact frequency bands, they are readily discernible to our eyes as we look at speech spectrograms. We therefore expect them to be important for our ears as we listen to speech. Furthermore, their frequencies or frequency trajectories seem to be sufficient to allow us to discriminate many phonemes from one another. Therefore the formant seems to be a plausible unit for the auditory system to be using for scene analysis. Accordingly, there have been experimental investigations of the role of formant trajectories in holding speech sounds together.

Dorman, Cutting, and Raphael studied the effects of spectral continuity in synthesized speech sounds.[689] They wanted to find out how formant transitions contributed to the sequential integration of syllables. They excluded those speech sounds that lack significant formants (such as the noise bursts that occur in consonants such as "s"). A number of different types of cycles were created, each containing four synthesized vowels ("ee", "a", "uh", and "oo"), all with the same steady fundamental frequency and each having three formants.

The different types of cycles are illustrated in figure 6.3. The first type (long vowels) contained 120-msec vowels with instantaneous transitions from one to the next. This type was intended to simulate the stimuli of the experiments, described earlier in this chapter, in which steady-state portions of natural vowels were spliced into cycles. The second type consisted of consonant-vowel-consonant (CVC) syllables in which the consonant was "b" (for example, "boob"). The consonant "b" was used because it can be convincingly synthesized using only formant transitions. A listener will hear an initial "b" when there are rising formant transitions leading into the beginning of the vowel's formants and a final "b" when there are falling transitions at the end. In these CVC stimuli, the durations of the steady-state vowel portions were 30 msec and those of the transi-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

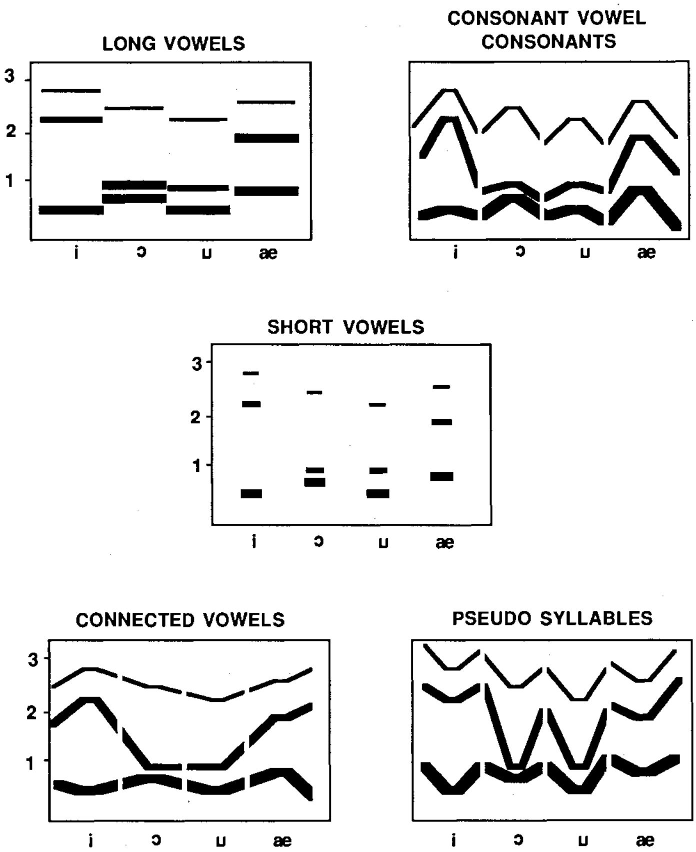

Figure 6.3
Schematic spectrograms of the five types of stimulus sequences used by Dorman, Cutting, and Raphael (1975).

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

tions were 45 msec. In a third type (short vowels), only the 30-msec steady-state vowels from the CVC syllables were used, the transitions being replaced by silence. In the fourth type (connected vowels), the formants of successive vowels were connected directly to one another by formant transitions. The final type was made out of sounds that the experimenters called pseudosyllables. These were created by extending the formants of the steady-state vowel formants upward in frequency for 45 msec on both sides of the 30-msec steady-state portion rather than downward as in the case of the "b-vowel-b" syllables. No real consonant can be synthesized by such upward transitions, and, according to these researchers, there is no movement of the vocal tract that will produce such transitions.

The task of the listeners was to write down the vowels in their order of occurrence in the repeating cycle. We have to collect the results across a set of experiments in a sort of intuitive way, because no single experiment compared all the types of cycles and the results were not fully consistent with one another. Nonetheless, it is clear that there were two conditions that were relatively easy: the one with the vowels connected directly to one another by formant transitions and the one with the CVC syllables. Those consisting of the vowels that were not connected by smooth transitions were substantially worse, and those consisting of pseudosyllables, in which the transitions were unnatural, were the worst. The research also showed that the smooth transitions of the lowest formant were most effective in holding the sequence together and those of the highest formant were the least effective.

The subjective impressions of the sequences, as reported by the experimenters, suggested that these results were due to differences in the degree of perceptual integration. In the unconnected vowel sequences, the listeners often heard two streams of vowels, with "ee" and "oo" in one stream and the other two in a second.

The perceptual integrity of the sequences in which vowels were directly joined by transitions is explainable as a continuity effect analogous to the one found by Bregman and Dannenbring with alternations of high and low tones joined by frequency glides.[690] But the advantage of having "b" consonants between the successive vowels cannot be explained in this way. Rather than joining the formants of successive consonants by a direct frequency transition, the consonant introduces a frequency change (a dip and then a rise) in all the formants. There is some continuity; if the listeners tracked the frequency transitions they would never be faced with a sudden discontinuity as in the case of the transitionless vowels. However it is not apparent why the CVC condition should be as good as the condition that in-

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

troduced direct and smooth transitions from one vowel to the next. Perhaps the listener's knowledge of language played a role in the success of the CVC condition. This is suggested by the failure of the listeners to derive any benefit from the unnatural transitions in the pseudosyllables. However, it is also conceivable that the acoustic organizing principle involved here is a primitive one rather than one that depends on the structure of language. We will know only when we have done experiments on nonspeech cycles (perhaps sequences of tones) in which successive tones are connected by brief dips in frequency, as the CVC formants were, or by brief rises in frequency, as the pseudosyllable formants were.

Earlier we saw that an experiment that shifted the fundamental frequency of a voice before and after a silence prevented the silence from being interpreted as a stop consonant closure. However, the absence of continuity in the fundamental can be compensated for by a continuity in the spectral peaks. Silence retains its effectiveness as a cue for closure despite a shift from a male to a female voice across the silence, as long as the general articulatory pattern is continuous across the two speakers.[691] This confirms the findings, described in chapter 2, that fundamental frequency and spectral peaks are independent contributors to sequential integration.

The research of Cole and Scott, which has already been described in chapter 4, did employ, in addition to vowel sounds, unvoiced sounds such as "s" that are not normally thought of as having formants.[692] To briefly summarize their procedure, they took consonant-vowel syllables such as "sa", spliced out the formant transitions leading from the consonant into the vowel, and created rapid cycles of transitionless syllables. Their listeners had more difficulty in judging the order in cycles formed of these syllables than in cycles formed out of normal syllables, which presumably were heard as coherent units. The researchers interpreted the perceptual coherence of the normal syllables as due to the fact that formant transitions in the vowel pointed to important spectral regions in the noise burst of the consonant, and related this to the findings of Bregman and Dannenbring on spectral continuity in cycles of tones.[693] In chapter 4, I offered three other explanations of the findings of Cole and Scott. Regardless of which of these is correct, the experiment shows that formant transitions can affect the grouping of successive sounds. This is, however, somewhat mysterious because there was no marked formant structure in some of the consonants that they employed.

It is difficult to explain why noise bursts (such as the "s" in the word "essay") will group with adjacent voiced sounds. Even if the voiced sounds group with one another because of pitch continuity,

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

why is the interrupting frication noise not rejected into a separate stream and the vowel extrapolated right through it as in the case of the continuity illusion? Part of the evidence that prevents the perception of the two vowels as a single sound is the change that the formants undergo just before and after the "s" occurs. This change takes the form of formant transitions that are caused by the fact that the mouth has to move to a new position to produce the "s". In an earlier chapter on the continuity illusion I described the "No discontinuity in A" rule (where A is the sound that occurs before and after the interruption). Violating this rule by introducing a change in sound A just before and after an interruption by sound B can reduce the perceived continuity of A.[694]

Applying this rule to speech, the auditory system may have a choice between integrating the noise or rejecting it as an interruption. The discontinuity of the voiced sound before and after a consonantal noise may inhibit its rejection into a separate stream. This explanation implies that clicks or bursts of noise that are spliced into a speech stream at arbitrary points (so that they do not fall between two discontinuities in the voicing) will probably be rejected as extraneous sounds. The research on inserting clicks or noise bursts into sentences generally shows that the speech is heard as continuing behind the added sound, which is merely perceived as an interruption. As far as I know there has been no examination of whether there are alterations in the placement of the click or noise that will cause it to be integrated into the speech stream rather than heard as an interruption.

We might suspect that the loudness of the noise burst would play a role. If the noise is really masking a continuous voiced sound, it should be loud enough to do so. To obtain the continuity illusion the interruption has to be a good deal louder than the ongoing sound. People who have tried to synthesize speech sounds have noticed that if the consonant noise bursts are too loud they do not become part of the speech. Perhaps this is explainable by the preceding argument.

Another fact that may be related to the sequential integration of different types of sounds is that the onsets of many instrumental sounds, such as those of the struck or plucked instruments (guitar, piano, and the like), are very noisy. That is, the sound is not periodic at the onset but gradually evolves into a periodic sound. Yet this noisiness is integrated with the sound that follows it and is heard as a quality of the onset of the sound (as, indeed, it is). We do not understand how the auditory system registers this transition, but I know from my own experience in synthesizing sounds that it is not sufficient to splice a bit of loud noise onto the beginning of a periodic sound for the noise to be heard as a noisy onset of the second sound.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

Typically it is rejected as a separate sound. Again, we do not know why this happens, but it bears a suspicious resemblance to the case of integrating consonant noise bursts with voiced sound that precedes or follows them.

Obviously we need fundamental research on the sequential grouping of voiced and unvoiced sounds. For example, it would be useful to know whether imposing a formant pattern onto an unvoiced sound would help it to group with a voiced sound that followed it. We would not necessarily want to use real consonants in studying this problem, but could pass a wideband noise through filters that imposed formants on it that matched those of a similarly filtered complex tone. If such a filtered noise were alternated with the tone in a cycle, we could compare the perceptual integration in this case to what would be obtained if the noise had no formants or if the formants were less sharp or did not match those of the tone. To the best of my knowledge, nobody has carried out such research.

I have also been told by researchers who are trying to get computers to produce intelligible speech that a continuity problem arises when they try to generate a continuous speech stream by concatenating stored segments. Apart from the discontinuity of the fundamental frequency and of the formants at the segment boundaries, there is a problem with formant bandwidth.[695] If one of the segments has formants that have a different bandwidth than its predecessor, even though the center frequencies of the formants may match up correctly, the spectral discontinuity tends to be heard as the intrusion of a new speaker. A difference in bandwidth means that at the moment of change there will be an abrupt drop or rise in the energy in certain frequency regions. Therefore the change can be viewed as one of loudness instead of one of bandwidth, and it makes us suspect that a sudden change in the intensity of any spectral region may disrupt the perception of a single stream.

Spatial Continuity We have seen how continuities in the fundamental frequency and in the formants are important in keeping speech sequences integrated. Both these forms of continuity derive from the fact that they are produced by gradual changes in the speaker's articulation; the human vocal tract cannot snap instantaneously from one setting to another. There is another continuity that derives from a property that speakers share with inanimate sound sources: the fact that they tend to stay in the same place or to move relatively slowly through space. As we have seen in chapter 2, sounds that are alternated rapidly between the ears often group themselves into different streams, each localized at a different ear, and this makes it hard to

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

integrate the entire sequence. This difficulty of integration has been shown to apply to speech as well as nonspeech sounds. When successive segments of continuous speech were presented to alternate ears, the capability of recognition broke down.[696] Later it was shown that this happens because the residual portion at each ear is treated as if the material sent to the other ear were missing; that is, the switch of the material to the other ear leaves a perceptual gap that acts as a silence, creating a false segmentation of the signal.[697] We can conclude that if discontinuities in spatial location act to disrupt the integration of speech then the spatial continuity that is found in normal circumstances must be holding it together.

The breakdown of identification when material is switched from one ear to another is also due to the loss of information about individual phonemes. When, in a nasal-plus-vowel syllable such as "me", the "m" is sent to one ear and the "e" to the other, listeners find it very hard to identify the "m". Apparently the sudden increase of the bandwidth of the spectrum is what signals the release of a nasal consonant, and this increase can be assessed only by comparing the successive spectra. When primitive scene analysis favors the interpretation that the two spectra are in separate streams, this comparison of spectra is not carried out.

Are These Acoustic Continuities Enough? We have seen that a number of acoustic factors influence the integration or segregation of speech sounds. Most of these are concerned with similarities or continuities of one sort or another. Do we have enough in these acoustic factors to explain why normal speech holds together sequentially? After all, the segregation of unrelated sounds, such as noise and tone, presented in cycles in the laboratory is quite strong. There are different points that can be made in addressing this question.

One is that the use of cycles in the laboratory creates a much stronger segregation of sounds than is normally obtained. This is true for two reasons: The repetitive nature of the alternation allows the evidence for separate streams to build up over time, and the competition that occurs in cycles can promote the segregation of sounds. Let me explain. In a short cycle, there will be only a short delay between successive occurrences of the same sound. Therefore, if the grouping process must choose between integrating a successive pair of sounds of different types (A and B) or waiting only a bit longer and integrating two sounds of the same type (A and A), it will skip over the different tone, B, and group the two A's, causing the B tone to be rejected from the stream. Happily, the B tone will have repetitions of itself to group with. If the cycle is slowed down, the choice of wait-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

ing a long time to group two tones of the same type or grouping each tone with the next is resolved in favor of grouping unrelated tones. The competition between grouping by temporal proximity and grouping by similarity is always part of the stream-forming process.

In speech, in contrast with artificial cycles of sound, even though two sounds of different types, A and B, are next to one another, it is unlikely that there will be another occurrence of an A-type sound for the A to group with (and a B-type sound for the B to group with) in the short time interval in which the qualitative similarity can dominate the grouping process. Also, even if this does happen once, it is unlikely to be repeated immediately again and again so as to permit the formation of separate streams containing the two types of sounds. It is not that normal speech sounds are immune from segregation. The experiment by Cole and Scott on the repetitive looping of single syllables showed that they are not.[698] Even when the normal formant transitions were left in the syllables, syllables such as "sa" would eventually split into two streams, one formed of unvoiced sounds deriving from the "s" and the other of voiced sounds deriving from the "a".

Another factor that may affect the tendency of speech sounds of different types to be integrated sequentially may be their grouping into larger units such as syllables. While it is possible that this is accomplished by schemas (for example, schemas for familiar syllables or schemas that represent the way that syllables are produced), it is also possible that this grouping could also be derived from primitive scene-analysis principles. The second syllable of the word "repay" may hold together as a unit, with the brief noise burst caused by the release of the "p" being heard as an onset property of the syllable "pay" rather than a separate acoustic event. The acoustic factors that cause this might resemble those that unite the noisiness at the onset of the plucked string of a guitar with the later, more periodic part of the note. One such factor might be a sudden rise in intensity followed by a slower decay. It may be that a brief noisiness that accompanies a sudden rise in intensity tends to be grouped with the loud (but decaying) sound that follows it. This might be a primitive, unlearned grouping tendency. Its role in nature would be to hear as single units the events created by impacts, where the onset is determined by the brief period of energy transfer in which the sound is never periodic, followed by the vibration of the struck body, which may or may not be periodic depending on the nature of the struck body. The perceptual formation of syllables in speech may be accomplished by the same primitive process.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

Of course it is hard to distinguish whether this form of integration is based on specifically speech-related schemas, more general *learned* schemas (such as one for an impact), or primitive unlearned grouping tendencies. We might think it possible to find out through the study of newborn infants, but studying the perceptual grouping of sounds in such young babies is difficult, if not impossible. Perhaps we could achieve the more limited goal of finding out whether the grouping was determined by knowledge of one's own specific language if we studied the grouping of unfamiliar sounds in unfamiliar languages, such as the phonemic clicks of Xhosa.

As I hinted above, another way to account for the integration of speech is to move to a wholly different approach that argues that it is our learned knowledge of speech sounds that allows us to integrate them. This knowledge might take a more specific form, such as memories of specific words, or a more general form, such as stored records of the inventory of the different sounds of our language and the probabilities with which they follow one another. Encountering a familiar word pattern or a highly probable sound transition, we would integrate the sequence of sounds.

There is even a more profound type of knowledge that a listener might have: a mental model (but not a conscious one) of the vocal tract of the human speaker, together with some methods of deriving the possible sounds and sound sequences that it could make. Then any incoming sequence would be passed through the model, and if any subset of the acoustic components could have been produced by a single speaker it would be grouped and labeled as such. Once again the theories may take two forms, divided along the nature-nurture dimension. One version would hold that the model is formed largely through learning. The more provocative view is that the perceiver's model of the human vocal tract is innate. This latter is the approach that has been taken by Alvin Liberman and his associates at Haskins laboratories, and we shall examine its properties later in more detail.[699]

Integration based on any of these types of knowledge would fall into the category of schema-driven integration discussed in chapter 4. A plausible candidate for this form of integration is the fricative-vowel sequence (as in the word "say"), because of the lack of spectral similarity or continuity between the fricative and vowel portions. Indeed, persons who try to synthesize speech with machines have found that unless the balance of spectral components and the loudness of fricatives are tuned just right, the fricative is experienced as an added sound and not as part of the speech. It is possible that the requirement

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

for an exact replication of the natural spectrum and loudness of fricatives stems not from requirements imposed by primitive grouping processes but from our knowledge of the acoustical patterns of speech (as stored in our speech schemas).

The process of sequential integration of speech is involved in the perceptual restoration of speech when some of the sounds in a speech sample are deleted and replaced by noise bursts. Particular cases of this, called phonemic restoration or the "picket fence" effect, have been discussed already in chapter 3. I will not repeat that discussion here except to point out that the integration of speech sounds before and after an interruption should depend on the same principles of sequential grouping that hold speech together even when parts of it are not obliterated. For example, suppose we were to do an experiment on phonemic restoration in which the part of the sentence that preceded the loud noise burst was heard as coming from a female speaker to the left of the listener and the part that followed the noise was heard as coming from a male speaker to the right of the listener. Even though the two parts went together quite well to form a meaningful sentence, I would not expect the two parts to be integrated or the missing phoneme to be perceptually restored.

I would like to draw a comparison between the laboratory examples of sequential stream formation and the examples that occur in speech. In the basic laboratory demonstration of streaming the tones are discrete and often separated by silences. In speech, on the other hand, the sounds are continuous; so how could a tendency to group sounds that were separated in time be useful? One way is to group speech sounds that have been interrupted by a masking sound. The similarity of the sounds before and after the masking sound would be grouped by their similarity and this, in turn, would activate an analysis of the spectrum during the interrupting sound for evidence of the continuation of the interrupted sound. I have already described this in chapter 3.

Often, however, an added sound does not really mask the underlying sound, but simply creates a jumble of properties. Think, for example, of what happens when the voices of two speakers are mixed. Due to the natural stopping and starting of speech and its variations in intensity, there are moments when the spectrum mainly reflects the properties of one of the voices. There are other moments when the spectrum is mainly shaped by the second speaker. Clear acoustic properties may be ascertainable in both cases. There is a third class of moments in which the sounds can be easily segregated by the use of primitive principles such as grouping by differences in fun-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

damental frequency or in spatial origin, or by asynchrony of onset. We can think of these three types of moments as yielding "high-quality" data. However, there are times when the mixed spectrum is very complex and the primitive scene-analysis processes cannot decompose it. We can think of these as regions of "low-quality" data. It would be appropriate to treat the low-quality moments in the same way as we treat an interrupting noise burst. We would simply group the events on both sides of it according to their similarities and then look for continuations of the high-quality sounds in the low-quality regions.

I can think of two ways of doing this. The first would be to have a test for the quality of data in a temporal region so that when the region was judged as high-quality the horizontal grouping process would use its properties to establish sequential links and favor the formation of one or more streams. This process would make use of the brain's proven capacity to link discontinuous but related sounds, as in the streaming phenomenon. Then the low-quality areas would be analyzed for the presence of properties that matched those within one or another of the primitive groupings of evidence (perhaps we could call them "protostreams") that had been established through the organization of the high-quality data.

Another approach would be for the auditory system to always look for continuations of the previously detected regions of spectral activity. The system would look for matches even for the accidental spectral patterns arising from the mixture of two voices but would not find any sequential material to link it up with, so the memory for these patterns would disappear. However, patterns that were parts of individual voices would find matches in subsequent time periods, and the streams of which they were a part would be strengthened. This method would not need to make a decision about the quality of regions of data, but would require more memory to carry along the often meaningless patterns that were derived from mixed data. Since we do not know much about the memory capacity of the auditory system, we do not know how plausible this alternative might be.

Simultaneous Organization

So far we have looked only at the grouping and segregation of speech sounds that arrive at different times. However, it is obvious that there must be principles that group and segregate the acoustic components of speech that arrive together. These principles must involve both segregation and grouping. They must segregate the acoustic compo-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

nents that belong to different voices so that we do not hear sounds that are the accidental composite of a number of voices. They must also integrate the components that come from the same voice so that we do not hear, as separate sounds, the formants or the harmonics of the individual voice. Part of the integration of harmonics can be accomplished by the limited resolution of the auditory system, especially with respect to the higher harmonics, but its resolution is perfectly adequate to register the lower harmonics separately.

If we assume that the first stage of the auditory system creates something that resembles a neural spectrogram, then to detect and recognize the sounds made by a particular voice, when they are part of a mixture, the auditory system must allocate each spectral component either to that voice or to some other sound. We have seen, in our examination of nonspeech sounds, that simultaneous components can be segregated by a number of factors, including spectral region, pitch, spatial location, and independence of changes. We should examine the role that each of these factors plays in the segregation of speech sounds. Part of the motivation for doing this is to find out the extent to which such primitive acoustic differences play a role in the parsing of the neural spectrogram so that we will know the extent to which they have to be supplemented by schema-based processes of grouping.

In trying to decide this issue, it is necessary to have at least a rough idea of how the primitive processes and the schema-driven ones are different in their modes of activity. I am going to try to propose such a difference.

I would like to propose, as I did in chapter 4, that they differ in the way in which they partition a mixture. Primitive segregation acts in a symmetrical way to partition a signal. If the signal is partitioned into two parts on the basis of a difference in location, for example, both of the parts are equally accessible to more sophisticated analysis. Therefore we can say that the primitive process *sorts* the signal. Schema-driven processes select and integrate material from the signal rather than sort it. They look for certain patterns in the data, and when they find them they extract them. The residual that they leave behind is not organized or coherent in any way. Schema-driven processes create a figure and a ground (in Gestalt terminology). Primitive processes do not; all the partitioned streams have equal status.

Segregation and integration have been treated as if they were opposite sides of the same coin. Perhaps they actually are for the primitive processes, which perform a sorting function: what they do not segregate they integrate. But the schema-driven functions do not have this

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

property of sorting, or of creating a symmetry between segregation and integration.

Second, I think that the primitive processes will be found to be sensitive to variables to which the schema-governed ones are not sensitive. For example, the fundamental frequency of a set of harmonics will turn out to be very significant to the primitive processes and less so to the schema-governed ones.

I am not going to say very much about the schemas that are involved in the interpretation of speech. Doing so would require not just a whole book, but a library. I am hoping that despite the large number of schemas of different types involved in the understanding of speech, there are some ways in which their modes of operation are similar enough that we can determine when they are involved in a perceptual process.

The research on the grouping of simultaneous components has mainly been concerned with the grouping of formants. The apparent reason for choosing the formant as the unit of analysis is that the pattern of formants is what defines a vowel sound. The pattern of change in the formants also provides a strong cue for the identity of stop consonants.

There are other reasons for the choice of formants as materials. The theory of how they are produced is fairly well worked out. Also, they are fairly simple acoustic phenomena and are therefore easy to synthesize well enough to create plausible-sounding vowels or voiced stops.

A final reason is that they are sustained events and therefore can be heard as speech even without the participation of other types of spectral events. By way of contrast, the noise burst that defines a fricative, if produced in isolation, will simply sound like a burst of noise.

We do not know whether formants are meaningful perceptual entities in themselves. Most scientists who work in speech research believe that they are, and that the auditory system assesses the frequencies of these spectral peaks and tracks their changes over time as a partial step on the way to recognizing speech. We may, however, just be fooled by the prominence of the formants in the spectrogram. The eye of the observer certainly treats the formants as separable features of the spectrogram. Whether the ear does so too or whether the cue value that we attribute to the formants themselves is really the result of a different form of analysis will have to be decided by future research. (There are researchers who doubt whether formants play a central role. They think that the formants are merely the acoustic basis for the perceptual analyses of global qualities such as the compactness and tilt of the spectrum, or the rapidity of spectral change.[700])

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

Role of Harmonic Relations and F0

Everyday observation tells us that we can hear two separate pitches when tones of different fundamental frequencies are played. The segregative process that makes this possible also helps us to sort out a mixture of voices. Effects of pitch have been found in the studies of selective attention that I mentioned earlier, in which listeners are asked to shadow one of a pair of voices. It is easier for them to do so when voices of two different pitch ranges, such as a male and a female voice, are used than when the two voices are in the same pitch range.[701] It also helps to have the spectra of the two signals restricted to different ranges by filtering.[702] These studies of selective attention used natural voices reading connected materials, and so the acoustic situation in these cases was very complex. The studies that I shall describe in the following pages were more analytical from an acoustic point of view.

Two-Voice Research

The auditory system faces a serious problem in using the different fundamental frequencies of two voices to segregate them from one another. At each moment of time, the auditory system must not only detect that there are two fundamentals and use this to derive two different pitches but must somehow form a separate view of the two different spectra. Each view must consist of a description of not only the harmonics that belong to each voice, but (in the ideal case) the intensities of each of them. This is necessary because the identity of the voiced sounds depends on the relative intensities of the different harmonics and how they change over time. To decide which vowel is present in each of two voices, the complex pattern of intensities that is registered in the neural spectrogram must be decomposed into two separate formant patterns.

First, of course, it is necessary to show that the auditory system can use information about the existence of different fundamentals to segregate voices. Relevant experiments have been done by Brokx and Noteboom at the Institute for Perception Research at Eindhoven in the Netherlands.[703] In one experiment listeners were asked to repeat aloud a series of nonsense sentences that were heard mixed with a second voice that was reading a story.

In an ingenious experiment, a male speaker was trained to produce high-pitched imitations of the same sentences spoken by a female speaker using the female's pattern of intonation (pitch variation). His imitations yielded sentences spoken by a male voice but with a high average pitch (about 160 Hz). To obtain a contrasting set of low-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

pitched sentences, he simply spoke them in his normal voice. Each of the resulting two sets of sentences was mixed with the same interfering speech, a story spoken in his natural pitch range. This yielded two test conditions, one in which both voices in the mixture (test sentences and interfering story) were low pitched and one in which the sentences were high but the story was low. As might be expected, the sentences were harder to perceive when they were in the same pitch range as the interfering story.

In another set of conditions, the pitch of the test sentences was controlled more exactly. They were recorded by the same indefatigable male speaker, who was now trained to speak in a monotone at the same pitch as a tone that was provided to him through earphones. Two versions were made of his sentences, one at a fundamental frequency of 110 Hz and one at 220 Hz. The story that was used to mask the sentences was spoken in his natural voice, which, though it varied normally in pitch, had an average fundamental of about 110 Hz. The listeners made more mistakes on the low-pitched sentences, which were nearer to the pitch of the masking speech from the story, than they did on the high-pitched sentences.

To be even more precise about the fundamental frequencies that were used, another experiment was done in which a digital signal processing technique was used to produce speech with an absolutely constant pitch.[704] The interfering story was produced with a fundamental frequency of 100 Hz. The test sentences were generated on different fundamentals in different experimental conditions. As the fundamentals were separated in frequency, the number of errors decreased. They dropped from about 60 percent when the fundamentals were the same to about 40 percent when they were three semitones apart. When the mixed voices were exactly an octave apart, the number of errors went up again, presumably because there are so many overlapping harmonics in spectra in which the fundamental of one is exactly double that of the other. (The reader is referred to the discussion of dissonance and perceptual fusion in chapter 5.) The amount of separation in fundamentals that is needed before listeners hear two separate voices is not large. The experimenters reported their own impressions as follows: "Whereas at zero semitones [separation] one definitely hears only a single auditory stream of garbled but speech-like sound, at one half semitones [separation] one hears very clearly two voices, and it is possible to switch one's attention from one to the other." In normal speech, the pitch keeps changing all the time and it is therefore unlikely that the fundamentals of two simultaneous voices in a natural situation will remain within half a semitone of one another for more than a brief moment at a time.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

To examine the effects of fundamental frequency even more precisely, Michaël Scheffers, working at the same laboratory in the Netherlands, carried out a series of experiments in which listeners were asked to identify pairs of synthetic vowels presented simultaneously.[705] Such stimuli are ideal for studying the effects of differences in fundamental frequency. A first reason is that the vowel can be recognized even when its spectrum is unchanging. A second is that it can be created from a wholly harmonic spectrum. Finally, the use of mixtures of single vowels, rather than whole sentences, excludes the sequential role of pitch continuity in uniting a sequence of syllables as part of the same message.

In one study Scheffers used a computer to synthesize eight Dutch vowels, each 200 msec in duration.[706] Six different separations in fundamental were used: 0, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, and 4 semitones. The two fundamentals in a mixture were always placed symmetrically above and below 150 Hz. The subjects were asked to identify both vowels in each mixture. Performance was better when the fundamentals were different, but virtually all the improvement, from about 68 percent to about 79 percent correct, was caused by changing separation of the fundamentals from 0 to 1 semitone. In fact, an earlier study showed that the percentage correct dropped when the vowels were separated by 8 or 12 semitones. At these separations, the performance was considerably worse than when the vowels had the same fundamental.

Scheffers interpreted his results as consistent with the "harmonic sieve" model of pitch perception, developed in the same laboratory.[707] It is an example of the class of pattern-recognition models of pitch perception discussed in chapter 3. The auditory process of finding the fundamental of a set of harmonics is thought of as analogous to a sieve. The sieve for any given fundamental has "holes" in it that are spaced apart such that only the harmonics of that fundamental can "fall through." Others are blocked. If the harmonics of a second vowel are mixed with those of a first, the sieve that is responsible for detecting the first fundamental will block the harmonics of the second (except for the occasional one whose frequency happens to be near one of the harmonics of the first). If, at any given moment, there are two subsets of harmonics (belonging to different harmonic series), they will be detected by two different sieves. We can suppose that the harmonics that fall through the same sieve will tend to be grouped together for purposes of vowel recognition. Scheffers' finding suggest that the sieve mechanism can block unwanted partials that are as little as one semitone away from the harmonics that do fall through the sieve. Other research has given similar results.[708]

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

In research carried out in our laboratory, Magda Halikia (Chalikia) included some conditions that were like those of Scheffers but did not find any anomalies at large differences in fundamental frequency.[709] Her stimuli, however, were different. They were four different English vowels rather than the eight Dutch ones of Scheffers, and her synthesis of each vowel used only three formants rather than the five that Scheffers had used. It also appears from the spectra that were published by her and Scheffers that the bandwidths of her formants were larger. Whatever the reason, her results were much simpler. Averaging over three of her experiments, we can derive the following percentages of correct responses for separations of 0, 0.5, 3, 6, and 12 semitones (one octave): 58, 83, 84, 86, and 73 percent. (Note that the chance probability in her experiment would be 50 percent, on the average, if the subject made two guesses on every trial.) In one of her experiments, in which a 9-semitone separation was included, the percentage correct for this separation was the same as for the 6-semitone separation. These averaged results were also quite representative of those for the individual experiments. Each one showed a striking improvement between the 0-semitone separation and 0.5-semitone. The only drop at higher separations was at the octave, and even there performance remained quite high. Apparently the exact coincidence of harmonics in the two vowel spectra in the case of the 0-semitone separation makes them much harder to segregate and the coincidence of every second harmonic at the octave introduces a lesser difficulty.

Split-Formant Research

So far, in our discussion of the role of harmonic relations in segregating voices, we have proceeded ever more analytically, from mixtures of natural voices to mixtures of synthesized vowels. The next step is to see what happens when parts of the spectrum that are smaller than an individual vowel are given the same or different fundamentals. Being in different parts of the spectrum, the parts cannot actually contain the fundamental itself but can contain harmonics whose frequencies are multiples of the same fundamental. Most of the experiments that I shall discuss synthesize formants separately, using harmonics related to a chosen fundamental, and then mix these formants for presentation to a listener.

In 1955, Donald Broadbent published a brief report of some research in which the two ears of a listener received signals that were filtered differently.[710] The signal sent to one ear was high-pass filtered at 2,000 Hz and thus contained the higher frequencies. The other ear got the lower ones, resulting from low-pass filtering at 450 Hz. Because the filters were not very sharp, there was some overlap of fre-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

quencies at the two ears.[711] When normal speech was used, a large majority of the listeners reported that they heard only a single voice despite the division of the two frequency ranges. Broadbent attributed the fusion to two factors: first to the synchrony of onsets of energy in the two channels that occurred as consonants were released, and second to a common amplitude modulation in the two divided spectral regions. He was referring to the amplitude pulsation at the period of the fundamental that is seen as vertical striations in spectrograms such as the one shown in figure 6.2.

Two years later Donald Broadbent and Peter Ladefoged did a related experiment, but this time the signals were synthesized.[712] They were the first to raise the issue of the correct grouping of formants when the ear of the listener was presented with a mixture of formants. How, they asked, could the auditory system know which combination of formants to group to form a vowel? They proposed that it could be done on the basis of similarities within subsets of formants, and one similarity that they proposed was the factor that Broadbent had noted earlier—the overall repetition rate of the waveform. This rate would be determined by the fundamental of the voice, and would be visible in each formant as a pulsation (the vertical striation referred to earlier). Formants in which there were identical repetition rates would be grouped as parts of the same speech sound.

They produced a simple sentence—"What did you say before that?"—by speech synthesis. They created the voiced sounds as follows. They fed the pulse train from an electronic sound generator that was meant to simulate the vocal cord pulsation—and therefore to provide the fundamental frequency (F0)—into a set of resonant circuits that created formants. Sometimes the same F0 generator was used to create formants 1 and 2 and sometimes a different generator was used for each. In the latter case, the fundamental frequencies would be different. Either the resulting formants were both sent to the same ear of the listener, or else formant 1 was sent to one ear and formant 2 to the other. Their listeners were asked whether there were one or two voices and whether they were in the same or different spatial positions.

When the same F0 generator was used for both formants, either varying the pitch in a natural way or creating a monotone, the listeners usually heard only a single voice, even if the formants were sent to different ears. (Later research showed that the voice was heard at whichever side of the head received the lower formant.[713]) Having the same fundamental seemed to bind the formants together. However, when a different fundamental was used for each formant, the listeners usually heard two voices. For example, if one of the

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

generators (say the one feeding the first formant resonator) was programmed to give a natural pitch contour to the sentence, but the other was programmed to stay 10 Hz higher than the first one at every instant, the listeners usually reported two voices, even when the two formants were mixed and presented to both ears. A particularly interesting condition was one in which there were two generators, one for each formant, both programmed to follow the same pattern of F0 over time. Due to slight inaccuracies in the equipment, the fundamentals tended to go in and out of phase randomly as one got slightly ahead of the other. In this case the binding of the two formants was weaker and they could no longer be fused if they were sent to separate ears. This suggests that a great precision in timing is needed to induce the maximum fusion of different spectral regions and that just having harmonics that are related to the same fundamental is not as powerful.

In a second experiment by these researchers, simpler sounds were used—just two formants, sustained without change for 15 seconds. Lacking the normal variations of a voice, these were heard as buzzers rather than as vocal sounds. As before, when the two fundamental were the same, the listeners tended to hear one sound, and when they were different to hear two. I think that the most important result, and one which agrees with data from other studies, was that even if the two resonances (formants) were centered at exactly the same frequency in the spectrum, and hence were totally overlapped spectrally, if they had different fundamentals they were heard as two voices. This illustrates the role of the pitch-detection mechanism in segregating two overlapping periodic sounds.

These experiments are important in that they gave exactly the same results for speech sounds in sentences and those that were not even heard as speech. But we must be careful to remember that in these experiments the listeners were only asked about the number and location of the sounds. They were not required to recognize what was being said.

James Cutting studied the effects of fundamental frequency in binding tögether formants that were presented to different ears.[714] The formants sometimes carried complementary and sometimes contradictory information about the identity of a syllable containing a stop consonant and a vowel. Such syllables can be convincingly synthesized using only two formants. The first result was concerned with the number of sounds heard. Even when the same pair of formants (the syllable "da") was sent to both ears, if the left and right ear signals were synthesized with different fundamentals, they were heard as two separate sounds. There was a striking difference in the percentage

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

of times that the listeners judged that there was only one sound present when they both had the same fundamental (100 percent) to when their fundamentals were separated by only 2 Hz (nearly 0 percent).

When one ear got one formant of a syllable such as "da" and the other ear got the second, again the listeners were much more likely to report two different sounds when the formants had different fundamentals, but this segregation did not prevent the listeners from being able to recognize that it was "da" that they had heard and not "ba" or "ga". This result is remarkable. The listeners could combine the left and right ear information to derive the identity of the speech sound. At the same time, they appeared to segregate them to derive independent percepts that included their locations and pitches. This sort of result has occurred in many experimental contexts. It has been used as the basis of a claim by Alvin Liberman that ordinary principles of auditory analysis do not apply to speech perception. Because it is so important, I will discuss it in more detail later.

Christopher Darwin, together with a number of collaborators at the University of Sussex, has tried to show in a number of ways that the perceptual fusion and segregation of speech sounds are governed by the principles of auditory scene analysis and that this fusion affects the perceived identity of the sounds.

Darwin, in 1981, reported a series of experiments which, like those of Broadbent and Ladefoged and of Cutting, synthesized formants with different properties from one another and looked at the extent to which the listener recombined them perceptually into a single sound.[715] While he manipulated several properties of the formants, I shall mention only their fundamental frequencies here. In every case in which a speech sound was created by combining formants that had different fundamentals, the listeners tended to report hearing more than one sound. At the same time, there was usually no tendency for a difference in the fundamental to inhibit the perception of the phoneme that resulted from the combination of those formants. In short, the listeners did segregate the formants, in the sense of hearing more than one of them, but at the same time combined them to derive the correct speech sound.

There was one exception to this general pattern. One of his experiments used the pattern of formants shown in figure 6.4. One combination of these, the formants F1, F2, and F3, if heard in isolation, will give the syllable "roo"; another combination, F1,F3,F4, will be heard as "lee". It is important to notice that the first and third formants appear in both syllables and therefore there is a competition for the belongingness of these formants. Unless formants are used twice, the listener cannot hear both syllables. With this setup, the listeners

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

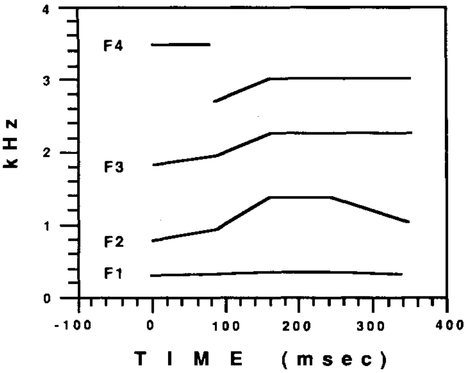

Figure 6.4
Frequency tracks of the formants used to synthesize the syllables "lee" and "roo". (From Darwin 1981.)

tended to hear the syllable whose component formants were labeled as going together by all having the same fundamental frequency. Nonetheless, in the alternative condition, where the formants were broken up by not being labeled with the same fundamental, the syllable was still heard on a significant number of occasions in a series of tests. This suggests that while a common fundamental *can* contribute to the perceptual decision of how to group formants to get a speech sound, its effect is seen only in situations in which there is a competition between speech-recognition schemas.

The conclusion that we can draw from the studies of both Cutting and Darwin is that the phoneme-recognition system, a schema-based process, has some way of selecting what it needs out of a mixture of sounds without the assistance of primitive scene-analysis processes that operate on the acoustic similarities. The results tend to show, however, that the schema-based ones can improve their performance in difficult situations by using the information provided by primitive grouping.

Harmonics in Nonoverlapping Frequency Ranges We have now reviewed research that has tried to manipulate the perceptual segregation either of whole voices or of individual formants. We should note one important difference between the two cases. With whole voices the material to be segregated is in overlapping frequency ranges, and we find that the segregating effect of different fundamental frequencies in the two voices (and presumably the integrating effect of all the spectral components from the same voice having the

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

same fundamental) plays a very important role in segregating the voices. On the other hand, in the case of manipulation of individual formants, the main energy of each formant is usually in a fairly distinct frequency region. In this case the phonetic integration is not usually very sensitive to the relation between the fundamentals of the formants.

This points to the possibility that in normal speech the main use of the fundamentals in scene analysis is to disambiguate cases in which a number of alternative phoneme recognitions are supported equally well by the mixture (as in whole mixed voices). It appears that each phoneme-recognizing schema can look for the formant patterns that it needs, and if crucial evidence is not simultaneously being claimed by another phoneme recognizer, the fundamental frequencies in different parts of the spectrum will have no effect on the identity of the derived phoneme. This latter description probably tells us what happens in natural listening to a single voice and in the formant-manipulating experiments in which there is no competition among phoneme schemas.

Scene Analysis in the Defining of Formants

So far we have looked at the process of segregating voices and grouping of formants. But formants are not the only sounds in an environment. Other sounds can interfere with the extraction of the peak frequencies of the formants, a value that listeners may need on the way to arriving at a description of the speech sounds. Most linguists believe that the description of a formant that is most relevant to phoneme recognition is this peak frequency. It is not a direct property of the formant itself but something more abstract, a peak in the spectral envelope. The latter is an imaginary smooth line that represents the intensity that each harmonic *ought to have* if it occurs in a spectrum that has been filtered in a certain way.

The issues can be clarified with the help of figure 6.5, which shows the spectrum of a spoken vowel with an extra pure tone added near one of the peaks. The vertical lines represent the intensities of different partials that are present in the spectrum. The dotted line represents the relative intensities that the harmonics *might* have had if they had been at different places along the spectrum. The peaks in this envelope are important because they come from resonances in the vocal tract of the speaker; therefore they tell the listener about how the speaker's tongue, jaw, and so on are positioned. The line labeled T represents an additional nonspeech tone that has been mixed with the speech sound. If we focus on the first formant, labeled F1 in the

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

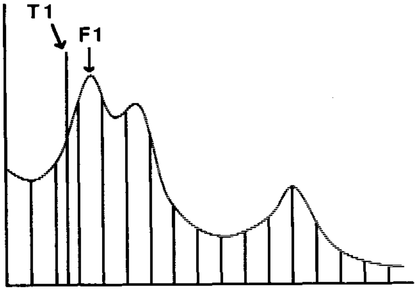

Figure 6.5
Spectrum of a vowel mixed with a pure tone in the region of the first formant. T1: pure tone. F1: first formant peak.

figure, we see that the point that defines it is not the frequency of the most intense partial in the region, because this partial is the added tone. It is not even that of the second most intense, which is part of the voice but is not at the resonance peak. Depending on the fundamental frequency, it will often be the case that there is no harmonic exactly at the resonance peak. Somehow two results have to be achieved. The added tone must be rejected and the remaining partials used to estimate the frequency of the peak. The first of these two involves scene analysis, and Darwin and his colleagues have examined how principles of auditory grouping can help the listener to achieve it.

Darwin and his associates have investigated a number of heuristics whereby the auditory system could determine that the tone did not belong to the vowel. Their general method has been to use a vowel spectrum that yields a sound that is intermediate in quality between two standard vowels, say the "i" in "bit" and the "a" in "bate". Then they introduce an extra tone in the region of the first formant. If the tone is accepted as part of the spectrum of the vowel, it shifts the identity of the vowel toward one of the two standard vowels. To the extent that it is rejected, the shift is smaller or nonexistent.

One experiment by Darwin and Gardner examined the factor of harmonicity.[716] It was based on the premise that the tone can be rejected if it does not fit into the harmonic series defined by the spectrum of the vowel. This expectation was derived from the harmonic sieve model of pitch perception that I described in relation to the research of Scheffers and also in chapter 3. The experiment mixed a vowel spectrum with a pure tone whose frequency was mistuned by varying amounts from one of the harmonics of the vowel. With

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

about 8 percent mistuning, the tone was effectively segregated from the vowel so that the vowel's identity was affected very little by its presence. The effect has nothing to do with the fact that one of the sounds was a vowel. In Brian Moore's laboratory the same amount of mistuning was found to be sufficient to cause a mistuned harmonic in a complex tone to stand out as a separate sound.[717]

Earlier we saw that when only one synthesized sound was present, the mistuning of different formants relative to one another (in the sense that their harmonics belonged to different fundamentals) did not prevent their integration into a single speech sound. We attributed this to the fact that a phoneme-recognition schema was capable of simply looking for energy in the spectral regions in which it expected it, and could overcome the segregating effects of scene analysis. Yet we saw in Darwin and Gardner's research that the phoneme schema does not seem to be able to overcome scene analysis when a tone is added to a formant. Why the difference?

One possibility is that it arises from the sorts of evidence that the phoneme-recognition schemas can use. These schemas, while they map patterns of formants onto particular speech sounds, and are therefore capable of taking what they need from mixtures, may contain no definition of what it means to be a formant. So they may be fully dependent upon more primitive processes for this decision.

A second possibility is that primitive scene analysis may have a more dominant role in spectral grouping when the sounds that are to be separated occupy the same spectral region. Recall that in the experiments on the spectral grouping of synthesized formants, where scene analysis did not affect the phonetic interpretation, the mistuned formants mainly occupied different spectral regions.

The effects of the similarity in fundamentals of co-occurring formants must depend upon a neural mechanism that can calculate whether or not one or more partials are multiples of a particular fundamental (F0). One proposed method is the harmonic sieve. Since this method involves measuring the frequency of each harmonic to determine whether it is close enough to qualify as a multiple of some particular F0, it will succeed only if the frequencies of the harmonics are separately resolvable by the auditory system.

An alternative view is that the F0 can be assessed without having to resolve the harmonics on which it is based. This can be done by responding to the pattern of timing in the neural impulses activated by the sound. We have examined this theory in chapter 3.

Often, in psychology, when you ask whether the nervous system does it this way or that, the answer is both. The methods of the nervous system are highly redundant so that it can continue to do a

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

reasonably good job when particular types of evidence are missing. Basilar-membrane place analysis may be carried out to determine the fundamental or fundamentals in spectral regions in which the harmonics are resolved, and periodicity may be used where they are crowded too close together.

Computer Models for Segregation of Two Voices

Attempts to segregate simultaneous voices by their fundamentals have used one or another of these methods. Thomas Parsons chose the approach of analyzing an incoming spectrum to find harmonics from two different fundamentals; Mitchel Weintraub chose to use the timing information that is located within narrow spectral regions.[718]

Parsons' program used Fourier analysis to derive a spectrum for the mixture in successive 51-msec segments. It then examined the peaks in the spectrum and used some tricks to decompose them into separate partials. Next it looked for the fundamental (F0) that would best account for the set of detected components. It then put aside all the harmonics that were accounted for by this F0 and tried to find another one for the partials that remained. If it succeeded, it treated the harmonics associated with this second F0 as a second set.

Next it had to decide how to group the discovered sets of harmonics over time. It did so by assuming that the pitch of each voice would not change radically from one analyzed segment to the next. For each voice, it predicted the F0 in the current segment from the pattern of F0's found in previous segments and accepted the one that best matched this prediction. It therefore used sequential grouping based on F0 similarities and trajectories. This method allowed certain F0 estimates to be rejected as erroneous.

I have argued earlier that the purpose of auditory segregation is to make subsequent recognition easier. However, rather than passing these separated sets of harmonics on to a speech recognizer, Parsons' program used them to resynthesize each voice so that a human listener could judge how well the two had been segregated by the procedure.

It is obvious that Parsons' program is not suitable for unvoiced sounds (which have no F0) and therefore cannot be a general method for voice separation. However, when it was tested on sentences consisting mainly of voiced sounds ("Where were you a year ago? We were away in Walla Walla."), each of the resynthesized versions of each voice was relatively free of intrusions from the second original voice. The quality tended to be poorer when the unwanted voice was louder than the target voice.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

Weintraub's program also used the fundamental frequencies of the speakers' voices to segregate speech. The speech was from two talkers, a male and a female, speaking sequences composed of digits. Unlike Parsons' program, it assessed the two fundamentals by analyzing the pulsation of the signal within narrow spectral bands. As its first stage of analysis, it employed the cochlear model of Richard Lyon that we discussed in chapter 3.[719] A sample of one of Lyon's cochleagrams was shown in figure 3.10 of that chapter. A computer implementation of Lyon's model transformed the incoming signal into a temporal pattern of pulsations in each of 85 frequency channels. Then Weintraub's program looked for regularities in the pulsations in each channel. It combined these across channels to find two different periodicities in the timing information. These were treated as the two fundamentals. Like Parsons' program, but using different methods, it linked successive estimates of pitch into two F0 tracks, each representing the successive F0's in one of the voices over time. Then it allocated the spectral energy in each frequency region, giving all or part to each voice. Unlike Parsons' program, it did not segregate individual harmonics, and therefore it had more difficulty with the problem of partitioning the spectral energy in each frequency band to the two voices (Parsons just gave the total energy from each harmonic to the voice spectrum to which that harmonic was assigned.) Like Parsons' program, Weintraub's assessed the degree to which it had been successful by reconstructing the two separated voices.

The programs were successful enough that they established the plausibility of estimating the fundamentals in an incoming signal and the utility of using them to segregate different voices. Yet each had only a limited degree of success and purchased that by restricting the signal to a mixture of only two voices speaking primarily voiced sounds. Obviously the use of F0 to segregate voices applies only to those speech sounds that have harmonically related partials, not to any others. To make it worse, the possibility of finding different fundamentals gets harder as the number of voices goes up. How does the human deal with complex mixtures of voices? Not, I think, by improving on the computer's ability to use F0 for segregation, but by increasing the number of scene-analysis factors that it uses.

Energy that "sticks out" inappropriately from a spectrum may also tend to be segregated. Darwin drew this conclusion from an experiment in which he used his method of mixing a tone with a synthesized vowel.[720] The tone was at the exact frequency of one of the harmonics in the first formant and started at the same time as the vowel. The only thing that was altered was the intensity of the added tone. At lower levels of intensity, the tone was incorporated into the

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

vowel, affecting its identity, but at higher intensities some of the energy from the added tone was rejected from the perceptual computation of the vowel's identity. Darwin attributed this to some ability of the auditory system to reject tones whose intensity could not fit into a speech spectrum (or perhaps, more generally, one that was shaped by a system of resonances). It was not apparent whether this implied that the auditory system had a special method for processing speech sounds or whether it was using a more general method for rejecting spectral energy that sticks out too far from the spectrum. If it is truly the result of a specialized analyzer for speech, then that analyzer must have a property that I have so far attributed only to primitive scene-analysis processes and not to schema-governed ones. This is the property of being able to partition a spectrum. To do this, a process must be able not only to find, in a dense spectrum, components that it is looking for, but also to remove this set of components and leave behind a coherent residual from which the components have been removed. The evidence that I cited in chapter 4 suggested that schema-based segregation did not make a coherent residual available to other pattern-recognition processes.

Common-Fate Cues

Summary of Findings with Nonspeech Sounds

The next class of cues for segregating speech sounds involves correlated changes in different parts of the spectrum. In chapter 3, these were called common-fate cues. They included changes in both frequency and amplitude.

As far as frequency was concerned, we saw that parallel changes of partials (in log frequency) in different parts of the spectrum promoted the grouping of subsets of harmonics. These changes included slow glides in frequency as well as micromodulation. There was some uncertainty, however, as to whether the effects could not be attributed to an explanation that had nothing to do with a common-fate heuristic (such as the fact that a subset of partials, changing in parallel, will maintain their harmonic relations).

The major use of amplitude differences was to segregate sounds that started at different times. Associated with this factor was the old-plus-new heuristic that tried to subtract the effects of an earlier-starting tone out of a mixture. We saw also that a sudden change in the amplitude or phase of a certain spectral component in a mixture tended to segregate that component from other components that did not change. Sudden sharp rises in the intensity of a number of frequency components, all at the same time, tended to cause them to

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

fuse. Another form of amplitude change that was shown to be important was the rapid changes in amplitude that occur throughout the spectrum when adjacent harmonics beat with one another at the fundamental frequency of the harmonic series.

Correlated Frequency Changes

Frequency Modulation of Harmonics Let us begin by looking for evidence that parallel frequency changes are involved in the segregation of speech sounds from one another. One would imagine that they should be. The human voice changes in pitch over time. This causes the harmonics that are parts of the same voice to move in parallel on a log-frequency scale. As the fundamental glides up in frequency by 25 percent, for example, each harmonic also goes up by 25 percent. When two people are talking at the same time, it is unlikely that their fundamentals (perceived as pitches) will be changing in the same way at the same time. A betting man would wager that any harmonics that were moving up or down in parallel and maintaining their harmonic relations to one another all belonged to the same sound.

In chapter 3, we found some evidence that our ears bet in the same way. Different frequency modulations of two sets of partials cause them to be segregated and grouped into two sounds. But it was not clear whether this grouping merely told the auditory system that there were two sounds or also partitioned the two spectra well enough to permit separate descriptions of their two timbres. Speech perception provides a good vehicle in which to study this question because you can ask the listeners not only whether they hear more than one sound but also what the identities of the speech sounds are. If recognition is possible, the segregation must have been sufficient for the building of a description of the individual spectra.

However, there is a phenomenon called "the tracing out of the spectral envelope" that makes research on this subject hard. Earlier in this chapter we saw that formants are defined by peaks in the spectral envelope and that most theories of speech recognition require the auditory system to estimate where the peaks are. This estimation may be very hard when the harmonics are widely spaced. This occurs when the fundamental is high, since successive harmonics are separated from one another by the fundamental frequency. An example is the spectrum of a female voice. Part 1 of figure 6.6 illustrates a typical vowel spectrum and shows the spectral envelope (E1) that results from a momentary setting of a speaker's vocal tract. Part 2 focuses on the region of the second formant. The curve labeled E1 is the true spectral envelope created by the vocal resonances of the speaker.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

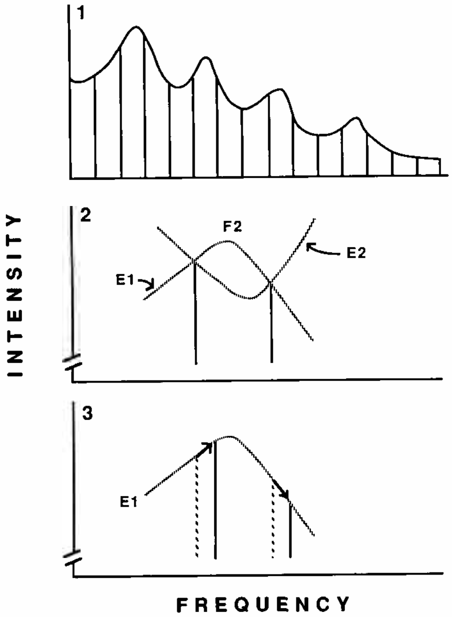

Figure 6.6
Illustration of "the tracing out of the spectral envelope" by a moving partial. 1: spectrum. 2: two possible spectral envelopes that fit the same harmonic intensities. 3: evidence for the local slope of the envelope as the harmonics go up in frequency.

However, as E2 shows, there are other envelope shapes that are equally consistent with the intensities of the two nearest harmonics. Part 3 shows what happens when there is a rising pitch and the frequencies of all the harmonics move upward. The dotted lines show the original harmonics and their intensities, while the solid lines show their new frequencies and intensities. Stephen McAdams and Xavier Rodet have argued that if the evidence provided by these movements and changes in intensity is taken into account, alternative envelopes such as E2 can be ruled out. Rodet has demonstrated that this information can be used by computer algorithms to achieve better estimates of formant peak frequencies, and McAdams and Rodet have shown that human listeners seem to employ this type of information in deciding on the identity of a vowel.[721]

If people used this information it would make it harder to determine the contribution of the parallel movement of harmonics in segregating them from a mixture. If we use, as our criterion, the identification of a vowel in a mixture, any improvement that is

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

found to result from moving all the harmonics in parallel may be the consequence of tracing out the spectral envelope rather than segregating the right subset of harmonics.

The reader might be inclined to ask at this point how the listeners could benefit from a better estimate of the spectral envelope if they could not even segregate the correct subset of harmonics. The answer to this question depends upon how the evidence furnished by spectral tracing is used by the human brain. One possibility is that it is used by schemas for speech recognition. We must recall that there is evidence to suggest that the speech sound schemas can extract what they need from a mixture even without the assistance of primitive scene analysis. The tracing out of the spectral envelope could help these schemas to confirm that certain formants were present.

The second possibility is that the envelope tracing is used by processes more primitive than those of speech recognition. The auditory system may be built to make use of the fact that it is often dealing with acoustic sources whose resonances are either stable over time or are changing fairly slowly. If changes in the intensities of certain harmonics accompanied changes in the fundamental in a way that was consistent with their all being passed through the same resonance system, this might contribute to their being grouped together. For example, if there were two speakers or two musical instruments present in a mixture, the auditory system might be able to build up a picture of the two sets of resonances and then, at later moments, assign to each resonance the harmonics that were consistent with it. How this could be done is not clear at the moment, but we cannot say that it is impossible.

To establish the role of common fate in frequency in binding harmonics together, it is necessary to test the effects of *covariation* in the movement of harmonics and not just the effects of movement itself. The simplest way to do this is to have two sets of harmonics. The changes in each set must be parallel but the changes in the two sets must be independent. We must compare this with the case in which the movements of all the harmonics are parallel.

Even when speakers think that the pitch of their voices is perfectly steady, it jitters slightly from moment to moment. I described this in chapter 3 as micromodulation. John Chowning has shown in acoustic demonstrations that adding this jitter to a synthesis of the human singing voice makes it sound much more natural and causes the harmonics to fuse into a speech sound.[722] Stephen McAdams carried out experiments with a mixture of synthesized voices.[723] He wanted to find out whether imposing independent micromodulation patterns onto the fundamentals of the different voices would cause the listener

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

to fuse the subset of harmonics that defined each voice and to segregate the subsets that defined different voices. The micromodulation that he used was a mixture of regular vibrato and random jitter. He synthesized the vowels "ah", "ee", and "oh" in a male singing voice and presented them in mixtures, each vowel at a different pitch (five-semitone separations between adjacent pitches). The listeners knew which three vowels could be present and were asked to judge the salience of each of these or their certainty that it was present. Sometimes none of the vowels had any modulation, sometimes only one did, and sometimes they all did. When they all did, either they all had exactly the same pattern or else two of them had the same pattern and the third had a different one.

Generally a vowel was heard better if it was modulated than if it was not. However, it did not matter whether the other vowels were modulated or not. Nor did it matter, when the vowels were modulated, whether the modulation was parallel or independent in different vowels. It appears, then, that the segregation of the vowels was not assisted by the independence of the modulation patterns. This result was subsequently confirmed in an experiment by Cecile Marin.[724] Nonetheless, McAdams did notice one effect of the modulation that was not just due to tracing out the spectral envelope. This is that often, when the harmonics were steady, many of the listeners heard more pitches than the three they were supposed to hear. Apparently, the accidental harmonic relations in the mixed spectrum fooled the pitch-detection system into computing extra pitches. However, when the harmonics were being modulated, there was a clear perception of three pitches.

Marin extended the research of McAdams by asking whether it was the listeners' sensitivity to tracing out of the spectral envelope that strengthened the perception of modulated vowels as compared with unmodulated ones. She designed two forms of modulation. In the first, the listener could use the modulation to get a "fix" on the spectral envelope. In this form of modulation, the spectral envelope was fixed and the amplitude changes in the moving harmonics traced it out. In the second type, the spectral envelope that defined a given vowel moved up and down in frequency with the harmonics. This meant that the harmonics remained constant in intensity and there was no fixed spectral envelope. The two forms of modulation strengthened the perception of the modulated vowels equally well. This means that it is not the tracing of formants that makes modulated vowels clearer. Perhaps the modulation simply evokes a stronger response in neural systems that are computing the vowel quality.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

We know from a large number of demonstrations that concurrent sets of harmonics that have different modulation patterns will be given separate pitch analyses. Therefore there is a contrast between the effectiveness of micromodulation in giving separate pitch identities to subsets of harmonics and its ineffectiveness in affecting the recognition of speech sounds. This is not the first time that we have encountered this lack of correspondence between the effects of primitive grouping and the identification of speech sounds. As we shall see, the absence of effects of scene analysis on the recognition of speech sounds is one of several pieces of evidence that has encouraged Liberman and his colleagues at Haskins Laboratories to argue that the brain's system for speech identification *never* makes use of primitive scene-analysis processes.

There is, in addition to micromodulation, another form of pitch change that takes place in the human voice. As we talk, our intonation rises and falls over the course of words, phrases, and sentences.

Magda Halikia (Chalikia), in our laboratory, studied the segregation of simultaneous pairs of synthetic vowels whose pitch was moving in this slower way.[725] She used various mixtures involving the vowels "ee", "ah", "u" (as in "put"), and "e" (as in "bed"). The listeners were asked to identify the two vowels in each mixture. Halikia studied two aspects of the role of fundamental frequency in the segregation of the vowels: the frequency separation of their fundamentals and how the fundamentals changed over time. The conditions are illustrated in figure 6.7. The frequencies of the fundamentals were either steady or else moved in either parallel or crossed trajectories over a 1-second period. Figure 6.8 shows the results.

I have already mentioned Halikia's findings on frequency separation so I will say no more here. The effect of the movement of the fundamentals was interesting. Both crossing and parallel gliding helped the listeners to identify the vowels. The help from the parallel gliding could not have been due to the segregation of partials based on different patterns of movement. Halikia attributed it to the tracing of the spectral envelope as the spectra moved. There was a small additional benefit to having the fundamentals move in different directions, but the difference was generally not significant. The one strong effect was at the octave separation of the fundamentals, where the different direction of modulation in the crossing condition prevented the continuation over time of a fixed octave relation between the two fundamentals. This continued relation tended to fuse the vowels when they were steady or moving in parallel.

Although the assistance from the crossing of the fundamentals was not, on the whole, better than from their parallel movement, we

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

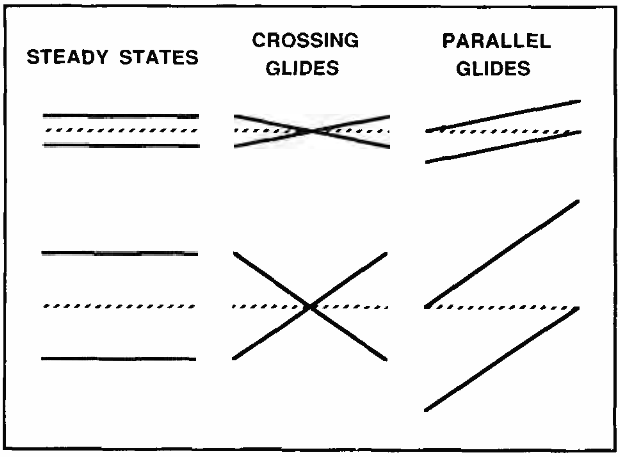

Figure 6.7
The pitch transitions in a pair of mixed vowels: steady states, crossing glides, and parallel glides. Top: small frequency separation. Bottom: large frequency separation.

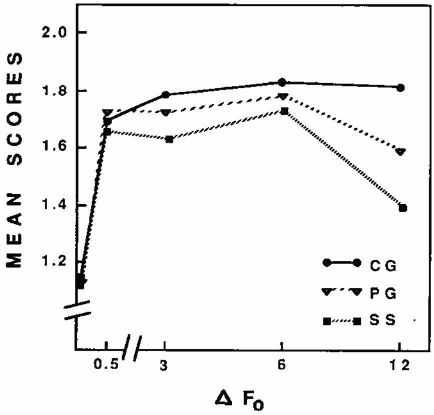

Figure 6.8
Mean identification scores for pairs of vowels with steady-state pitches (SS), parallel glides (PG), or crossing glides (CG). (From Halikia 1985.)

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

should take into account how the three kinds of patterns, steady, parallel, and crossing, were equated on the frequency separation of the two fundamentals. It was done by equating their maximum separations. This penalized the crossing pattern. For the other two, the maximum separation continued for the full 1 second. For the crossing pattern, it defined only the separation at the beginning and end. Between these two points, the separation was much less. We can see in figure 6.8 that frequency separation was the largest single effect, so the penalty paid was large; yet the crossing fundamental was still always easiest. This suggests that the crossing patterns are easier than parallel ones when the effects of frequency separation are removed. This additional advantage, if real, cannot come from envelope tracing, since both kinds of movement would have done this. It would have been due to common fate, the independent movement of the different subsets of harmonics in the mixture.

The idea that the partials of a voice can be segregated from partials that are unrelated to that voice if the two sets of partials have different patterns of motion was supported by another of Halikia's observations. She mixed a steady-state vowel with a gliding nonvowel tone. The latter sound consisted of a large number of partials of equal amplitude, covering the frequency range of the partials of the vowel. This tone was either swept in fundamental frequency or steady. If it was swept, it started at a fundamental frequency that was three semitones in frequency above or below that of the vowel and ended at the same frequency as the vowel. If steady, it remained three semitones above or below the vowel.

We already know that a separation in fundamental frequency favors the segregation of two sounds. On this basis, the gliding tone should have hurt the identification of the vowel more than the steady tone, since it moved nearer to it in fundamental frequency. However, the opposite results were found. The vowel paired with the gliding tone was more intelligible. Since the vowel remained steady, the results cannot be attributed to the tracing of the spectral envelope. A genuine segregation based on differences in the motion of the two fundamentals must be involved. Of course it is possible that steady-state motion is not really a type of motion at all, and that rather than dealing with nonparallel frequency motion we are dealing with a difference between two entirely different sorts of beasts—moving harmonics and steady-state ones—with a tendency to segregate from one another.

The results of an experiment by Gardner and Darwin suggests that common fate in FM may not be effective in segregating speech sounds from other ones.[726] These researchers followed a technique

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

pioneered in Darwin's laboratory: They altered the perceived identity of a vowel by inducing some of the energy in the spectrum to be perceptually isolated from the vowel spectrum. They first added a tone to the spectrum of a vowel at one of its harmonic frequencies. This caused it to sound more like a different vowel. Then they tried to induce the auditory system to isolate the added tone so that the vowel would sound more like it was before the tone had been added. The acoustic cue that they used in this experiment was correlated FM. The added tone was modulated either in parallel with the harmonics of the vowel or at a different rate or phase. The frequencies were varied by 2 percent and at rates of 6 or 10 times per second. All to no avail. Perceptual segregation was induced, but it did not affect phonetic identity. That is, the listeners heard an extra sound source, but the modulated partial still contributed to the vowel's quality.

The report of an extra sound source in this speech sound resembles a finding of Stephen McAdams with nonspeech sounds.[727] McAdams studied the perception of complex tones (built out of 16 equal-intensity partials) in which all the partials were subjected to micromodulation. In some tones the modulation was identical for all the partials (coherent modulation), and in others one partial was modulated in a different temporal pattern from the others (incoherent modulation). On each trial the listener heard one signal of each type and was required to judge which one seemed to contain more sound sources. McAdams found that when the tones were harmonic the choice of the incoherent tone as having more sources increased with the depth of the modulation (the variation in frequency caused by the modulation) and was greater when the independently moving partial was a higher harmonic. Since higher harmonics are *less* resolvable from their neighbors by the auditory system, this latter result suggested to McAdams that the multiplicity of the sound, in the case of the higher harmonics, was being judged not by being able to really hear out the partial but by the fact that incoherent modulation created the sort of phase rolling or chorus effect that is heard when slightly different sounds are played at the same time. On the other hand, when one of the first five harmonics was uncorrelated with the others it sounded like a separate sinusoidal tone. Since the harmonic that Gardner and Darwin had modulated in a speech sound was one of the first five, their discovery that it was heard as a separate sound agrees with McAdams' result.

In all these results on differential modulation of subsets of partials, there was a kind of segregation that clearly affected the perceived number of tones. However, it is doubtful whether the segregation

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

improved the assignment of a separate phonetic identity to one of these subsets.

More recently, however, unpublished research by Chalikia and Bregman on mixtures of two vowels with gliding fundamentals has demonstrated a clear superiority of the crossing glides over the parallel glides even when their maximum separations were equated. But this superiority was found only when the spectrum of each vowel was *inharmonic*. The inharmonicity was designed by taking a harmonic spectrum and deriving a new spectrum from it by displacing each partial upward or downward by a random amount, keeping it within one-half a harmonic spacing of its original frequency. When this spectrum glided, it did so in a way that maintained the ratios among its partials over time. To create an inharmonic vowel, the resulting inharmonic spectrum was passed through filters that created the vowel-defining formants. The superiority of the crossing movement over the parallel movement in separating the vowels was greater when they had these inharmonic spectra than when they had normal harmonic spectra. Apparently it becomes more helpful to unify a spectrum by giving it a distinct pattern of movement when it is not already being unified by the property of harmonicity.

Segregation by Independent Movement of Formant Center Frequency? There is another kind of change that can occur in voices and that we might expect to be used in segregation. This is the movement of formants. It is caused by the changing resonances in the speaker's vocal tract as the sizes and shapes of the cavities change over the course of an utterance. If we used this cue, however, to segregate formants that were not moving in parallel, we would just as often be segregating the formants of a single voice from one another than the formants of different voices.

When we look at a spectrogram of a single person talking, we see that the formants of the voiced segments act somewhat independently of one another, often moving in different directions. For example, in the word "yaw", a schematic spectrogram of which is shown in figure 6.9, the first formant moves up as the second one moves strongly downward. Why does this not cause them to segregate from one another so that they are heard as separate whistles? It is probably their harmonicity that holds them together. We must remember that a formant is merely a group of adjacent partials in the spectrum of a complex tone that has been enhanced through resonance. When we speak of a formant as moving down in frequency we mean that the region of enhancement has moved down, not that the partials themselves have moved down. Indeed, if the pitch of the voice was rising

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

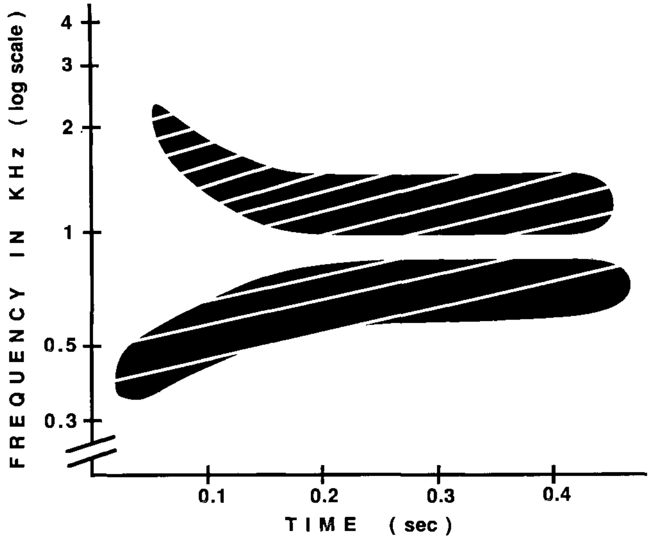

Figure 6.9
Schematic spectrogram of the word "yaw" spoken with rising pitch. The thin rising lines represent harmonics of a rising fundamental and the darkened regions represent formants.

during the word "yaw", all the individual partials would rise in concert. Therefore the second formant could be dropping as all the partials were rising. As the formant dropped it would enhance different (rising) partials. This effect is shown in figure 6.9, where the thin rising parallel lines represent partials and the larger outlined regions represent formants. If it were subsets of partials that were moving in two different trajectories, as would occur if they were coming from two different voices, their independent motions would guarantee that within each subset the harmonic relations would be preserved but that there would be no sustained harmonic relations that held across the two subsets. In such a case we would expect segregation to occur. However, we can see in figure 6.9 that when formants move in different directions, the change maintains the parallelness of the partials, and therefore their harmonicity.

If we think back to the experiment done by Dorman, Cutting, and Raphael on cycles of synthesized vowels, we can recall a condition in which they joined adjacent vowels by connecting the corresponding formants by formant transitions.[728] Their interest was in the fact that the vowels were now more sequentially integrated. But they did not comment on the fact that the high level of identification of the order

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

of the vowels by the listeners suggested that the vowels were well integrated *spectrally* as well, despite the fact that the formant transitions were not moving in parallel.

We should think a bit about the ecology of such matters. Since formants are resonances imposed on a single spectrum, their independent motion does not represent the existence of different sound sources, but tells us about changes in the dimensions of the chambers in which the sound is resonating. We would want to read this as a single sound source with changing properties or moving through changing enclosures.

On the other hand, it is not entirely impossible to segregate the formants that have been imposed on a single harmonic spectrum. For example, if we take a complex tone with only a few lower harmonics (tone A) and repeatedly alternate it with a complex tone (B) that has all the harmonics of A as well as some additional higher ones, in such a way that there is no break between the tones, tone A will be heard as continuous, and a second sound (the extra harmonics that are present in B) will be heard beeping in the background. B will get decomposed into two parts. A sound pattern like this would correspond in the natural world to two unrelated sounds that happened to have the same fundamental frequency but had different harmonics. This would be a rare occurrence but it can happen, at least for brief moments.

Notice, however, that the heuristic being used in this last example is not the independent changes of the *frequency* of the resonance peaks but a large asynchrony in their onset. The latter is much stronger evidence for the existence of two voices. Apparently the independent movement of formants does not constitute such persuasive evidence for the auditory system.

Nonetheless Stephen McAdams has reported that Jean-Baptiste Barrière, in a computer music piece, manipulated the way the individual formants changed over time to make the spectral forms coalesce into vowels or disintegrate into the several formants as individual images. McAdams has also reported synthesizing an example of Tibetan chant, a musical form in which vocal resonances are caused to move over time in a nonspeechlike manner. In his example, the voices are slowly made to disintegrate by decorrelating the formant movements.[729] Why does this not happen to all voices, since, as I pointed out earlier, their formants will often move in different directions at the same time? This requires more research. Possibly the synthesized formants moved over wider ranges than do normal speech formants or perhaps other factors that normally favor the integration of the whole spectrum were weaker. For example, in Tibe-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

tan chant, the perceptual segregation of the formants, as they move in different patterns, is assisted by the fact that the fundamental frequency is held constant. This would eliminate the coherence that is contributed by the parallel movements of all the harmonics. Also in Tibetan chant, the movement of the formants is much slower than in speech. Clearly, then, the existence of the two contrasting cases, speech and Tibetan chant, opens up opportunities for research.

Correlated Amplitude Changes

There are two other facts to mention about the duration and timing of sounds. The first is that unrelated sounds tend to go on and off at different times. This property of the world is exploited by a tendency of the auditory system to segregate sets of partials or parts of the spectrum that start at different times. The second is that often a sound will continue or be repeated as a second sound joins it. The auditory system exploits this fact by extracting from mixtures any components that are close in frequency to sounds that were heard in isolation just prior to the mixture. In chapter 3, we called this the old-plus-new heuristic. That chapter gave a number of examples of the perceptual results of these two facts.[730] It is possible that a part of the tendency to segregate mixed components that start and stop asynchronously may be an extension of the old-plus-new heuristic, because the auditory system gets a good look at the components that start earlier and therefore knows what to look for in the mixture. However, it cannot explain why components that end asynchronously will be segregated.

Sequential Capturing (Onset and Offset Asynchrony) Asynchronous changes in the amplitudes of formants can occur in natural speech. For example, sometimes in a long nasal sound, such as the "n-m" sequence in the phrase "in Maine", the higher formants are quite attenuated during the nasals and start up again quite strongly when the following vowel begins. One would think that this asynchrony would cause the spectral bands to segregate so that the components that were heard in the nasals would be experienced as continuing past the onset of the vowel, with the added formants coming from the vowel's onset sounding like the beginning of an accompanying sound. Actually I think that there is some tendency for this organization to occur. If you say "mamamama . . .", continuously to yourself, you will be able to hear a continuous tone with a repeating tone accompanying it. However, the segregation is not strong because the sounds are held together by the continuity of the fundamental frequency. We need experiments to measure the extent to which such asynchronies affect the integrity of speech signals.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

The factor of asynchrony has been used to explain why, in musical ensembles, there is so little masking of one note by another. Asynchrony has similar effects in speech. Scheffers made an interesting observation when he was in the process of designing his experiments on the segregation of two synthesized vowels by differences in fundamental frequency. (We discussed these experiments earlier.) He noticed that if the target vowel started some tenths of a second later than the masking vowel, the recognition of the target was virtually as good as it would have been in the absence of a second vowel.[731] Furthermore, the ability to recognize it no longer depended upon the differences in fundamental frequency. It is interesting to compare these results with conclusions drawn by Rasch on the asynchrony of musical tones.[732] He attributed the value of asynchrony to the brief opportunity for the auditory system to hear the target unaccompanied by the masker. This would not account for any advantage of turning the target on second; yet Scheffers found such an advantage with vowels. His result resembles that of Kubovy, who found that changing the intensity of a component tone inside a mixture caused that tone to be perceptually segregated from the mixture.[733] In natural speech such sudden changes of energy would occur at the beginning of syllables such as "ba" or "ga", making such syllables resistant to masking.

Earlier in this chapter, we discussed the work of Darwin and his colleagues on the segregation of a pure tone from a vowel spectrum with which it was mixed. We saw how the auditory system could use the harmonic relations between the tone and the spectrum to decide whether they should be segregated. The same research showed that the system could also use the asynchrony of the onsets of the tone and vowel.[734] The added pure tone was made to coincide exactly with one of the harmonics of one of the formants in the vowel spectrum so that the only factor that allowed the auditory system to segregate the tone was its asynchrony from the rest of the harmonics. The results showed that when the tone was added synchronously with the rest of the harmonics, its energy was incorporated into the perceptual estimation of the frequency of the formant. If it started earlier it was increasingly rejected as its asynchrony became greater and greater, until at an asynchrony of about a quarter of a second, its energy had no effect at all on the perception of the vowel.

The effect of asynchrony can also be viewed as sequential capturing of the part of the tone that is mixed with the vowel spectrum by the part that precedes the spectrum. This perspective is more obvious when a silent gap is left between an earlier portion of the tone and the tone-vowel mixture. Then we can think of the earlier part of the tone

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

as a captor analogous to the captor tone in the experiment by Bregman and Pinker discussed extensively in chapter 3. In the research of Darwin and his co-workers, when a 1-second captor tone preceded the tone-vowel mixture, its effect depended on how close it was in time to the mixture. When it was continuous, the extra sound was rejected completely and, at the other extreme, when it was separated from the mixture by a 300-msec silence, it had no ability to capture the added tone out of the mixture. An intermediate separation gave intermediate results.

If a sequence of captor tones preceded a short tone-plus-vowel mixture, the capturing became greater as the number of capturing tones was increased and was also improved if the tone sequence continued after the vowel. Capturing was also found when the preceding and following tones could be grouped with the tone in the mixture to form a sequence that descended smoothly in frequency. But it was found that it was not the trajectory itself that was responsible for the capturing, but merely the fact that the two tones that immediately preceded the mixture were near in frequency to the target tone.[735]

A first attempt to explain these results might be to argue that the preceding tone had habituated the auditory system and made it less sensitive to that frequency. The experimenters pointed out that this theory could not account for a second observed fact. When the tone started at the same time but ended later, it tended again to be rejected from the spectrum of the vowel. Although this offset effect was not as strong as when the asynchrony was at the onset, the fact that it occurred at all is not explainable by a habituation theory. Incidentally, this asymmetry between onset and offset asynchrony is not unique to tone-vowel mixtures. It was observed earlier by Dannenbring and Bregman in the perceptual isolation of a harmonic from a complex tone.[736] Evidently although retrospective grouping effects do exist, they are weaker than forward capturing. Perhaps a better refutation of the idea that the captor tone works by desensitizing the auditory system to a particular frequency is that the tone that is part of the mixture is not heard *less well* than it would have been if it had not been preceded by a captor; it is actually heard *better*, but it is heard as a separate sound, not as an undifferentiated part of the mixture. The effects on perceived loudness are more like one would obtain by adding some energy to the target tone rather than by subtracting some.

To further demonstrate that the effect of the preceding tone is one of grouping, not habituation, Darwin and Sutherland did an experiment in which the preceding tone was itself captured into a mixture so that it was less free to group with its counterpart in the tone-vowel

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

mixture.[737] The sound pattern that they studied was constructed as follows: There was a mixture of a pure tone (tone A) and a vowel as before, and again the tone started before the mixture. However, another pure tone (tone B), an octave higher than tone A, and therefore harmonically related to it, started at the same time as A, but stopped just as the vowel came on. It was expected that tone B, starting at the same time as A and harmonically related to it, would group with the leading part of tone A and prevent it from grouping with its own tail end that was accompanying the vowel. The tail end should then be free to be interpreted as part of the vowel. By being an octave above A, tone B would be too far away to mask or suppress it by any mechanism that is known to exist in the peripheral part of the auditory system, and therefore its effects would have to be attributed to grouping. The results showed that while the presence of tone B did not completely nullify the effect of starting tone A ahead of the mixture, it did reduce it by about a half. That is, the leading part of tone A was less able to capture its own later part out of the tone-vowel mixture. Consequently the vowel identity was more altered by the part of A that was within the vowel.

Asynchronous Onsets Can Be Integrated The experiments described so far in this section have shown that asynchronous onsets or antecedent capturing tones can promote the perceptual segregation of parts of the spectrum. This has been found in two contexts, Darwin's studies of the isolation of a partial from a formant and Scheffers' studies of the segregation of two mixed vowels. However, the findings are more complex when the experiment is concerned with the fusion or segregation of two separately synthesized formants. Cutting found that when the first and second formants of a two-formant synthesized stop consonant were played to different ears, listeners could still identify the consonant with over 90 percent accuracy, but when an onset asynchrony was introduced between the two formants, the accuracy got worse as the asynchrony increased, dropping to about 50 percent at an asynchrony of 160 msec (the chance value was 33 percent).[738] Actually this level of accuracy is not bad when you consider that the consonant is defined by a short 50-msec spectral transition at the onset of each of the formants. However, if the recognition process has to pick up the *pattern* formed between the transitions in the two formants, then, when the formants are asynchronous it must not actually fuse them, in the sense of remixing them. Instead it has to *coordinate* the evidence derived from the two. Therefore it is not certain whether the result should actually be described as fusion, in the sense of losing the identities of the indi-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

vidual components in favor of the computation of an emergent property of their combination. Perhaps the word coordination would be better.

The ability to coordinate the evidence from the two formants was disrupted much less by differences in fundamental frequency than by asynchrony. This should not be surprising since the temporal course of the evidence for the consonant's identity is not distorted. Indeed, we should think of the effect of asynchrony in the case of two formants as not doing its harm by damaging the coherence of the consonant but by distorting the evidence for it.

Darwin also looked at the power of asynchrony to prevent the integration of formants but did it a little differently.[739] He used vowels synthesized from three steady-state formants. The correct guessing of the identities of these vowels was only minimally affected (between 5 and 10 percent drop) by asynchronies of as much as 200 msec between the onsets and offsets of the first and third formants. Yet at the same time, these asynchronies led the listeners to report that they heard more than one sound. We saw a similar pattern of results earlier when I described Darwin's manipulation of the difference in fundamental of the different formants. The listeners heard more than one sound, but still could identify the phoneme.

In a different experiment Darwin did find an effect of asynchrony of onset on phonetic identity.[740] This occurred under conditions when four formants were mixed. The perceptual result, either the syllables "lee" or "roo", would depend upon which three formants were grouped by the auditory system. One can think of the situation as involving two syllable descriptions that compete for the formants. If Darwin lengthened the second formant so that it started 300 msec ahead of the 330-msec mixture, the auditory system tended to remove this formant from the computation of the identity of the syllable so that the description that used the other three would win. However, the removal was not complete.

Taken overall, the effects of asynchrony seem to be the same as those of differences in fundamental frequency in speech sounds. Whole speech sounds are strongly affected, with asychronous ones much easier to segregate. Asynchronous tones are separated from speech sounds more easily. Finally, asynchronies do segregate formants from one another but the segregation does not prevent the recognition of the phonemes, except where alternative descriptions are competing for the possession of a formant.

We have talked about the asynchronous onset of signals. Another acoustic phenomenon that produces synchrony of amplitude changes in the voice is the beating among its partials. If you look at a wide-

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

band spectrogram such as was shown in figure 6.2, you see small vertical striations that occur in all voiced regions. These all occur with the same period, the period that the fundamental of the voice has at that moment. The reason for their occurrence is that in a wide-band spectrogram the filters are not narrowly tuned. Therefore, instead of responding to (resolving) separate harmonics, they respond to the interaction between the nonresolved ones. This factor was thoroughly discussed in chapter 3, so we will not repeat that discussion except to note that any ability of the auditory system to integrate spectral regions that were showing a common periodicity would contribute to the separation of mixed voices.

It should be pointed out that this striation is a good source of information about the fundamental frequency of the voice and that this can supplement the information contained in the spacing of the harmonics. The latter is useful only in lower frequency regions where (due to the structure of the human cochlea) the harmonics are more resolvable and it is easier to estimate their individual frequencies. The amplitude fluctuations that the auditory system receives, on the other hand, occur most strongly in higher regions of the spectrum where the ear's frequency resolution is much lower. We should not be misled by the fact that ordinary spectrograms show the striation equally well throughout the spectrum. This happens because the computations that create a spectrogram do not have the human ear's decreasing ability to resolve higher frequencies. Lyon's cochleagram (figure 3.10 of chapter 3) shows this change in sensitivity with frequency much better. Weintraub's computer program to segregate mixed voices uses the amplitude fluctuation to estimate two fundamentals at any given moment, one for each voice.[741] However, it does not use the actual *synchrony* of the fluctuations as a way of deciding which frequency regions should go together.

It may seem as though this amplitude fluctuation is helpful only in the voiced components of speech which are constructed out of harmonics that can beat together. However, a matching amplitude modulation can also help the listener to integrate certain kinds of noisy spectral components with harmonic ones. For example, the "z" sound in the word "zoo" is formed when the periodic variation in pressure coming from the vocal chords is used to drive air through a constriction at the front of the mouth. Therefore the "z" has both periodic properties and noisy ones. In a simple *mixture*, noisy and periodic sound tend to be perceptually segregated, but the "z" is not a mixture. The periodicity is present both in the "tone" (harmonic sound) coming from the vocal cords and the noisy sound that is created in the front of the mouth. If it were possible to subtract the

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

harmonic sound out of the "z", a periodic amplitude modulation of the noise would still be seen, since the harmonic and noisy sounds are not simply additive but also multiplicative. In the "z", the noise has the periodicity of the vocal chords imprinted on it. This could assist in the perceptual fusion of the harmonic and noisy components of the sound. It would be important to verify the role of AM in the fusion of harmonic and nonharmonic sounds using synthetic stimuli in which the AM of the noisy sound could be controlled separately from the frequency of the harmonic sound.

To summarize, we see that correlated amplitude changes in different parts of the spectrum can contribute to the assignment of the right spectral components to a perceived source. Such correlations occur at the stopping and starting of environmental sounds and, at a more microscopic level, when groups of nonresolved harmonics beat with one another at the period of their fundamental. In speech, the period of the fundamental also affects noisy sounds that are being formed at the same time as harmonically structured sounds.

Spatial Location Cues

In chapter 3 we saw that a major factor in the segregation of simultaneous sounds is their separation in space and in chapter 5 we found that "vertical" relations in music are suppressed if the participating notes come from quite different spatial positions. The contribution of spatial separation to auditory scene analysis applies equally well to speech sounds. For example, people who lose their hearing in one ear report having a difficult time in picking out a single conversation from a mixture of conversations. The fact that differences in spatial origin can, in principle, serve as powerful information in segregating voices is illustrated in the reasonable degree of success that engineers have had in programming computers to segregate co-occurring speech signals by their locations (see chapter 3).

The research in the 1950s in which listeners were asked to shadow one speech stream in a mixture of two showed that it was enormously easier to do this when the voices came from very different positions in space. In fact, it was so easy to concentrate on one voice when they were spatially separated that often the listener did not even notice what language the other voice was speaking.[742]

Nonetheless, some later research of a more analytical type found only modest effects of spatial separation. Schubert and Schultz played pairs of voices to listeners over headphones. Each ear got both voices but the interaural phase relations for each voice were adjusted so as to suggest different origins in space. Such phase-altered mixtures were

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

compared to simple mixtures of two voices presented to both ears. These researchers found that the ratio of intelligibility scores of phase-altered to unaltered signals ranged from about 1.00 to 2.00 and the benefit from altering the phase was stronger when the distracting sounds were broad-band noise than when they were voices.[743] There are obvious differences between the techniques of the earlier studies (one voice to each ear) that found strong effects, and the later ones (phase manipulations) that found weaker ones. Notice that the phase-altering technique does not allow the individual voice to have a different intensity in the two ears, while the presentation of a separate voice on each headphone maximizes the intensity difference. It is likely that the cocktail party effect depends in some important way on this intensity difference. One possible explanation is that the intensity difference merely improves the estimation of separate spatial locations for the various frequency components of each sound. In this case, the segregation is based on improved grouping of spectral components by spatial origin. Another possibility is that the ear nearer to the target signal (or exclusively getting it, in the case of dichotic presentation) simply benefits from the enhanced signal-to-noise ratio and the reduced masking within that ear. It is also possible that the problem with mixed voices in the Schubert-Schultz study was not primarily due to an ambiguity of spectral grouping but to a raw masking of one voice by energy in the other. Scene analysis may not be able to contribute as much in such cases.

Because so many factors could be operating when the task is to segregate signals coming from different locations, this is not, perhaps, the best test for whether or not the auditory system uses spatial location as a basis for partitioning of the spectrum. Perhaps the best test would be a situation in which differences in spatial origin *prevented* the auditory system from putting the information from different spatial locations together. Unfortunately, we rarely obtain this compelling segregation effect with speech because of the existence of schemas for the recognition of speech sounds. These schemas appear to be very powerful and not to care about the spatial origins of acoustic components.

Here is an example of the fact that spatial separation does not compel segregation of speech components. Donald Broadbent presented the high-frequency components of speech to one ear and the low-frequency components to the other.[744] The listeners fused the two signals and reported hearing only a single sound. Yet this fusion was not due to the destruction of the information that was needed to segregate the two signals. Two things point in this direction. First, there is evidence that spatial location can be assessed independently in

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

different frequency bands (see chapter 3). Second, the signals sent to the two ears in Broadbent's experiment were in different frequency regions, making an independent assessment of their location quite possible.[745] The fusion was probably a scene-analysis decision. If you give the auditory system sufficient information that the signals at two ears derive from a common source, the brain will fuse them. In the Broadbent example, there would have been a number of commonalities: fundamental frequency, synchrony of starts and stops, and so on. Broadbent himself attributed the fusion to these. I conclude that the existence of separate spatial estimates for different frequency bands is only one factor among many that lead to the grouping and segregation of components of signals.

Segregation of Formants

Other studies have looked at the tendency of the brain to fuse parts of speech signals whose formants are presented to different ears.

James Cutting carried out a series of experiments on the fusion of phonetic information from a two-formant synthetic syllable with the formants presented to different ears.[746] It is interesting to note that this dichotic presentation still allowed the syllable to be identified between 75 and 100 percent of the time as long as the two signals were properly synchronized. This happened even when the two formants had different fundamentals.

The results were different when listeners were asked how many different sounds they heard. When the two formants had the same fundamental, they were still heard as a single sound 60 percent of the time, even though presented to different ears. This shows that segregation by location is not obligatory. But when the fundamentals were only 2 Hz apart (100 and 102 Hz), the formants were almost always heard as two sounds. Since this difference in fundamentals was only one-third of a semitone, it is unlikely that the listeners experienced much of a difference in pitch. This pitch difference alone would probably not have been sufficient to segregate the two sounds if they had not been in separate ears.

It is possible that differences in location and in fundamental frequency act in combination to produce a segregation that is stronger than the sum of the individual effects. To quote Christopher Darwin:

> . . . in speech . . . it is common to find the higher formants predominantly noise-excited and the lower ones predominantly excited by a periodic source. . . . An heuristic that grouped only harmonically related tones together would fail to include components from the same speaker for a substantial proportion of what

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

> is conventionally regarded as voiced speech as well as for all whisper, friction and aspiration. It is perhaps not surprising that the auditory system is prepared to tolerate differences in harmonic structure provided that they originate from a common location; and since even prolonged listening to dichotic formants on the *same* pitch failed to abolish the fused percept, the auditory system is also apparently prepared to tolerate a difference in location if sounds share a common periodicity.[747]

We must note, as we have many times before, the discordance, in Cutting's findings, between the increased number of sounds that the listeners experienced (when dichotic presentation was combined with pitch differences) and the apparent lack of effect on the identification of speech sounds. Similar effects were found by Darwin with dichotic presentations.[748] Yet even the phonetic fusion can be broken down if dichotically presented syllable patterns are repeated over and over, with the end frequency of each formant in one repetition being the same as its starting frequency in the next repetition, and with the two formants on different fundamentals in the two ears. The sequential grouping of the separate formants causes two separate streams to form, one for each ear, and the phonetic identity of the syllable pattern changes, suggesting that phonetic integration has been prevented. This happens when the two sounds are differentiated both by location and by fundamental and does not occur when only one of these factors distinguishes them.

We are led to this conclusion: if you increasingly pile up cues that suggest that the two signals that are being presented to different ears are actually from different environmental sources, the first effect is that the perceived number of signals increases, and then with more evidence, the phonetic integration is disrupted. The phonetic integration probably requires more evidence to disrupt it because of the participation of speech-sound schemas. It seems plausible that such schemas work by detecting energy in certain parts of the spectrum with certain temporal relations. If the schema that is looking for a particular speech sound is able to find the pattern of evidence that it is looking for, this may vote heavily in favor of that speech sound. The relative insensitivity of the phonetic schemas to factors such as perceived location and fundamental frequency is probably due to the fact that neither of these factors is probably used by the recognition schema since neither is involved in the definition of a speech sound. Since the time-varying spectral pattern does form part of the definition, we would expect any factor that distorted it to retard recognition. As for the links between the bits of evidence that are supplied by scene

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a portion of the eBook at doi:10.7551/mitpress/1486.001.0001

analysis, these may be decisive in governing the recognition only in competitive cases in which more than one speech-sound schema can fit the undifferentiated spectral pattern.

A striking example of the power of speech-sound schemas to integrate spectral information without the assistance of scene analysis occurred in an experiment by Scheffers.[749] Pairs of vowels, synthesized with a fixed fundamental, were presented to listeners to determine whether the difference in the fundamentals of the two vowels helped in their segregation. It did. I proposed that the difference in fundamental frequency allowed the auditory system to group the set of harmonics for each vowel and thereby "see" its spectral shape. Otherwise, I assumed, the formants from the two vowels would be grouped inappropriately to create illusory vowels. However, there was a striking result in Scheffers' experiment that opposes this idea. Even when the two vowels had exactly the same fundamental frequency and spatial location, and had synchronized onsets and offsets, they were guessed correctly 68 percent of the time.

The only explanation for this astounding success is that there must be schemas that, if activated, are capable of taking what they need from a dense mixture of sound. This should not surprise us. A similar effect was illustrated in chapter 3 in the case of phonemic restoration. A sentence, interrupted by a loud masking sound, can be heard as continuing right through the interruption even if the part of the sentence that coincides with the masking sound has actually been physically removed. I pointed out that the neural stimulation that corresponded to the missing sounds was actually present, having been activated by the masking sound. Therefore the speech-sound recognizers must have been extracting the evidence that they needed from the complex neural activity stimulated by the masking sound.

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

This is a section of doi:10.7551/mitpress/1486.001.0001

Auditory Scene Analysis

The Perceptual Organization of Sound

By: Albert S. Bregman

Citation:

Auditory Scene Analysis: The Perceptual Organization of Sound
By: Albert S. Bregman
DOI: 10.7551/mitpress/1486.001.0001
ISBN (electronic): 9780262269209
Publisher: The MIT Press
Published: 1994

The MIT Press

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025

MIT Press Direct

First MIT Press paperback edition, 1994

© 1990 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in Bembo
by Asco Trade Typesetting Ltd. in Hong Kong
from computer disks provided by the author,
and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Bregman, Albert S.
Auditory scene analysis: the perceptual organization of sound /
Albert S. Bregman.

p. cm.
"A Bradford book."
Includes bibliographical references.
ISBN-13: 978-0-262-02297-2 (hc. : alk. paper)—978-0-262-52195-6 (pbk. : alk. paper)
ISBN-10: 0-262-02297-4 (hc. : alk. paper)—0-262-52195-4 (pbk. : alk. paper)
1. Auditory perception. I. Title.
[DNLM: 1. Auditory Perception. WV 272 B833a]
QP465.B74 1990
152.1'5—dc20
DNLM/DLC
for Library of Congress 89-14595
CIP

10 9 8 7

Downloaded from http://direct.mit.edu/books/monograph/chapter-pdf/2444585/9780262269209_caf.pdf by UNIV OF MELBOURNE user on 08 December 2025