[Frontiers in Bioscience 13, 148-169, January 1, 2008]

The auditory organization of complex sounds

Valter Ciocca

School of Audiology and Speech Sciences, Faculty of Medicine, The University of British Columbia, 5804 Fairview Avenue, Vancouver, B.C. V6T 1Z3, Canada

TABLE OF CONTENTS

1. ABSTRACT

This chapter reviews the existing evidence on the auditory processes that are responsible for the formation of auditory percepts in natural listening situations (‘the auditory scene’). The formation of the perceptual attributes of auditory events is explained as the result of the interaction of two types of auditory grouping processes, general-purpose and schema-based processes. A further distinction is made between attribute-specific and categorical schemas. After discussing the formation of perceptual attributes and of the timbre of familiar sounds, the chapter explores current knowledge on how the brain builds perceptual representations of simultaneous auditory events and of sequences of auditory events. The nature of auditory scene analysis processes and of their interactions is discussed, and a tentative interactive model is proposed as a framework for future research.

2. INTRODUCTION

The acoustic environment typically contains several sound sources that are active at any one moment. The auditory system processes the combined acoustic input from several sound sources in a similar fashion to conducting a spectral analysis that results in good frequency resolution at low frequencies, but poor frequency resolution at high frequencies (1). Following this analysis, the auditory system constructs representations of sound sources by organizing sets of components into distinct auditory events (2). In addition to assigning frequency components into groups that represent distinct sound sources, the auditory system also needs to integrate successive groups of frequency components into sequences of auditory events, since physical sound sources often produce sequences of sounds (think of the stream of speech, or the notes of a melody).

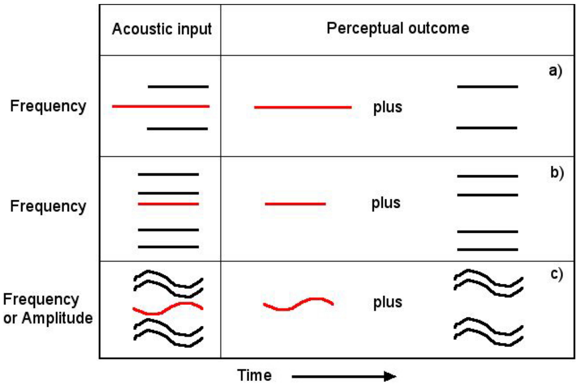

Figure 1. This figure illustrates the operation of three general-purpose principles of auditory organization. In Figure 1a, the asynchronous frequency component (in red color) is perceptually segregated from the other two components by virtue of its onset asynchrony. In Figure 1b, the component that does not fit in the regular frequency spacing of the remaining components is segregated from the rest of the complex tone. Finally, Figure 1c shows that the red component is segregated from the other components because the difference in its amplitude/frequency modulation pattern.

The process that organizes groups of frequency components into auditory percepts that represent individual sound sources has been called the 'auditory scene analysis' (2) or 'perceptual grouping' (3). While there are excellent reviews of the principles of auditory organization and their effect on the perception of both pure tones and complex sounds (see, e.g., 4-6), the present chapter will focus on the use of these principles for the perception of complex sounds, that is, sounds that contain two or more frequency components that overlap in time. Complex sounds constitute by far the most common types of sounds produced by physical sound sources since sounds that contain single frequency components (pure tones) can only be synthesized in the laboratory.

In this chapter, a description of the operation of auditory scene analysis processes will be followed by a discussion of the role played by these processes on the formation of perceptual attributes of complex sounds (timbre, phonetic percepts, pitch, loudness, duration and spatial location). Finally, current evidence about the auditory grouping of complex sounds will be reviewed. The terms 'auditory organization', 'auditory grouping' and 'auditory scene analysis' will be used interchangeably. Following (6), the term 'auditory event' will refer to the combination of perceptual attributes that corresponds to the complex sound produced by a single sound source, and that is delimited in time by a perceived onset and a termination (for example, a syllable, a vowel, a single musical note). This term can also apply to auditory events that change over time (although there is little empirical research on the grouping of this type of auditory events). More typically, auditory events will consist of relatively short complex sounds or noise bursts. The term 'auditory stream' or 'stream' will be used to refer to a sequence of auditory events (7).

3. THE NATURE OF AUDITORY GROUPING PROCESSES

The auditory grouping of complex sounds is governed by two basic principles: (i) portions of the auditory input that follow a regular pattern, over time and/or across frequency, are grouped together, and (ii) frequency components or auditory events that change gradually or minimally are also grouped together. All of the grouping processes that will be discussed in the following sections can be considered as instances of these two general grouping principles.

A basic question about auditory grouping processes concerns the units of grouping. One might ask 'what' is being grouped. Is it individual (or sets of) frequency components? Or is it auditory events? A related question concerns the nature of the grouping cues. Do auditory scene analysis processes operate on the basis of the characteristics of auditory cues that are associated with the physical attributes of the stimuli (such as frequency, intensity, inter-aural differences)? Or does grouping involve the perceptual organization of perceptual attributes (such as pitch, loudness, perceived location)? Although these questions could be asked about sequences of pure tones, it is easier to understand them in relation to sequences or mixtures of complex sound, whose perception involves a combination of frequency components that overlap in time. According to a simple serial model of auditory grouping, frequency components have to be grouped before perceptual attributes can be assigned to events, and before mixtures or sequences of complex sounds can be organized. Moreover, this model could state that auditory cues are the grouping units for frequency components, but that the auditory events corresponding to complex sounds are grouped on the basis of the properties of their perceptual attributes. As we will see in the later discussion in the Final Discussion section, such a simple model is not able to account for the complexity of the interaction among the different types of grouping processes.

Natural sounds are not only complex (that is, composed of co-occurring frequency components); they also change over time and/or can occur within sequences of sounds. Therefore, the auditory system needs to group frequency components both sequentially and simultaneously. *Simultaneous* grouping processes group two or more frequency components that overlap in time into a complex sound, in a chord-like manner (see, for example, the black frequency components in the 'Perceptual outcome' section of Figure 1). By contrast, *sequential* processes group successive frequency components into a sequence or into events that change over time. For example, the tones encircled by a red line in the 'Perceptual outcome' section of Figure 2 are being grouped into a sequence of tones. Although the categorization of grouping processes as sequential versus simultaneous is

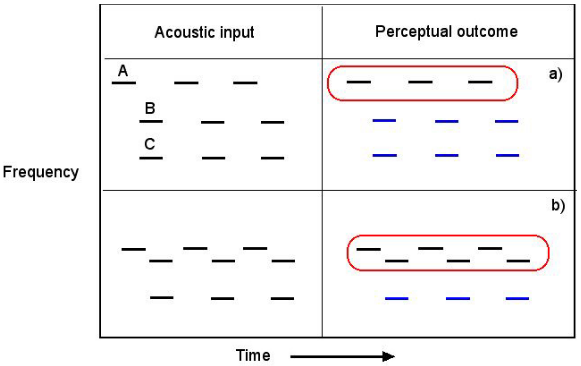

Figure 2. In part a) of this figure, the A and B sounds are far apart in frequency, therefore the B and C sounds are grouped together into a repeating complex tone, leaving the A sound to form a sequence by itself. In part b), the frequency proximity between A and B causes them to be sequentially grouped into a sequence, weakening the simultaneous grouping of B and C.

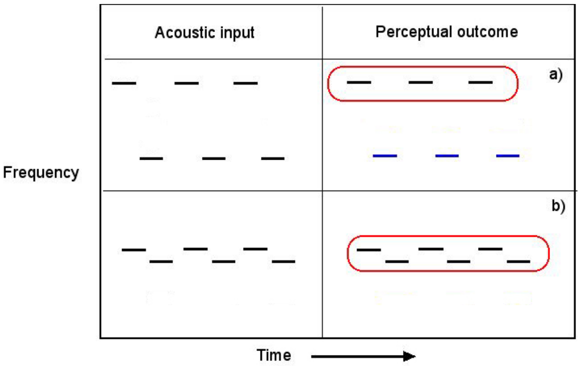

Figure 3. In part a) of this figure, the high-frequency sounds are grouped into a separate stream (circled in red) from the low-frequency sounds (in blue). In part b), the frequency proximity between the alternating tones causes them to be sequentially grouped into the same stream.

useful for understanding the operation of each process, it is important to realize that the two types of processes interact most of the time during the perception of natural sounds. Since the focus of this chapter is on the grouping of complex sounds, simultaneous grouping processes will always be involved in the grouping of frequency components. For this reason, studies on the sequential grouping of complex sounds involve by necessity an interaction with simultaneous grouping processes.

3.1. General-purpose and schema-based grouping processes

Gestalt psychologists proposed that the perception of events was mediated by the operation of two sets of grouping processes. According to them, one set of processes organizes the sensory input into patterns ("Gestalts") on the basis of principles (or laws) of perceptual organization. These principles were considered as innate processes that are used by the brain in an automatic (pre-attentive) fashion (8). Although Gestalt psychologists became famous mainly for stressing the role of these principles of perceptual organization, they also recognized that attention and other brain processes that make use of stored knowledge are also involved in the formation of perceptual objects and events. These processes rely on learning from prior experience for the purpose of object formation and recognition. For example, the processes that are used for recognizing the timbre of musical instruments make use of mental representations of musical timbres; these mental representations are acquired through exposure to the sounds of musical instruments.

Bregman made a similar distinction between "primitive" and "schema-based" processes in his review of auditory scene analysis (4). According to his proposal, primitive grouping processes operate in a 'bottom-up' fashion by using heuristic rules in order to group together parts of the auditory input; their operation is not affected by (nor does it require) the use of stored knowledge. An additional property of primitive processes is that they are 'general-purpose', rather than 'domain-specific'. That is, they group frequency components for purposes that are not specific to any perceptual attribute or any higher order organization; thus, they affect the perception of various perceptual aspects of complex sounds (pitch, timbre, etc.). A final characteristic of primitive processes is that they are 'pre-attentive', that is they do not require the use of attention.

The concept of a schema has been widely used by cognitive psychologists to refer to processes that make use of mental representations of regularities (or patterns) of perceptual objects or events that are stored in memory. These representations can also consist of patterns of motor skills. The schemas that are involved in the grouping of the auditory input are processes that select groups of frequency components in a 'top-down' fashion, either by (i) matching properties of the input with stored knowledge about categories of sounds (for example, representations of the various phonetic categories of the speech sounds of languages) or (ii) exploiting spectral or temporal regularities in the auditory input in order to generate perceptual attributes of sounds (such as loudness, pitch, and perceived location). A second feature of schema-based grouping processes is that they require the use of attention. Third, schema processes vary in terms of complexity. For example, there are schema-based processes that generate the pitch of a complex sound along high-low dimensions, and there are schemas for the linguistic categorization of pitch percepts as intonation patterns or lexical tones. An additional, and important, feature of schema-based grouping processes is that they are "domain-specific". That is, they exploit the types of regularities in the auditory input that are specific to each schema. For example, only the schema for perceived location is sensitive to differences in the level of sounds between the two ears; this schema is not responsive to auditory cues that are relevant to other schemas (such as the harmonic relations among frequency components). Finally, since schema processes depend on mental representations of stored knowledge, it would seem reasonable to deduct that they are processes that are learned. However, as pointed out by Bregman (4), the extent to which schema-based processes are learned or innate is still a matter of debate and may well differ for different schemas.

The distinction between primitive and schema-based processes relies on the validity of the above criteria. Some of these criteria, such as the " bottom-up versus top-down" and the "general-purpose versus domain-specific" criteria, might be accepted on the basis that they provide a conceptual distinction in terms of the defining features of the two processes. However, evidence for the pre-attentive nature of these processes is more controversial. This evidence came from demonstrations that, under certain conditions, listeners cannot prevent the perceptual segregation of a rapid sequence of alternating high- and low-frequency tones into two concurrent sequences: a sequence of high-pitch tones and a sequence of low-pitch tones (principle of "sequential grouping") (9). However, there is some evidence that the perceptual organization of such sequences of tones is affected by attention (9, 10, 11). In fact, it is difficult to test predictions about the role of attention on auditory organization with behavioral methods, as any kind of judgment by a listener requires the use of attention. Neurophysiological studies, in which listeners' responses can be measured independently of whether listeners pay attention to the stimuli, could in principle provide an answer to the question of whether attention is involved in stream formation. Sussman *et al.* have provided this kind of evidence by measuring the mismatch negativity (MMN) component of event-related potentials under conditions in which the sequential streaming of alternating high- and low-frequency tones was expected to occur (12). Since listeners were asked to read while they listened to tone sequences, Sussman *et al.* concluded that sequential streaming in their study occurred without the use of attention. However, as pointed out by Carlyon in his recent review of neurophysiological studies on auditory scene analysis (13), it is possible that listeners might have switched their attention from reading to listening during the recordings. A more recent study showed that MMN components measured during stream formation are likely to reflect the operation of attention as well as pre-attentive processes (14). It should also be pointed out that evidence concerning the pre-attentive nature of general-purpose processes was obtained from studies that investigated a single grouping process. Therefore, even though it was possible to convincingly demonstrate that sequential grouping is pre-attentive, it does not logically follow that all other general-purpose grouping processes are also pre-attentive. As discussed later in this chapter, the "grouping" (as opposed to "selecting") and the "innate" characteristics of primitive processes are also not strongly supported by evidence. If the bottom-up (versus top-down) and general-purpose (versus domain-specific) distinctions are the main criteria for differentiating the two processes, then the term "general-purpose" (rather than "primitive") might more accurately represent the different nature of these processes. For this reason, the term "general-purpose" will be used in the following parts of this chapter even though the term "primitive" has been commonly used in previous literature.

Following a review of each of the general-purpose grouping processes, this section will discuss the characteristics of attribute-specific schema that group simultaneous frequency components -although there is some evidence that pitch schema can integrate into a single pitch frequency components that do not overlap over time (15, 16). These schema-based processes include schemas for timbre, loudness, perceived location, perceived duration, and pitch. Categorical schema for speech and nonspeech sounds can integrate frequency components both simultaneously and sequentially for the purpose of sound recognition (think, for example, about music played by several instruments or fluent speech); they will be introduced in the last part of this section.

3.2. General-purpose grouping processes

Among general-purpose processes, sequential grouping occurs on the basis of 'frequency similarity' and 'temporal proximity', while simultaneous grouping is carried out by "common onset', 'spectral regularity', 'spatial cues', and 'coherent modulation'. Each of these grouping processes will be explained in the following sub-sections.

3.2.1. Sequential grouping by frequency similarity and temporal proximity

This principle states that successive frequency components that have similar frequency are grouped into sequences of frequency components (streams) on the basis of both frequency and temporal proximity. This principle is based on evidence from early studies of auditory organization, in which it was found that rapid sequences of alternating high- and low-frequency pure tones are perceived such that tones of similar frequency are grouped into separate sequences. Figure 3a shows that when the frequency difference between the high- and low-frequency tones is large, the sequence is perceptually split into a high-pitch sequence (encircled in red) and a low-pitch sequence (blue tones). These sequences are perceived as overlapping in time. This phenomenon was called the "trill threshold" (17), "doppio trillo" ("double trill"; (18, 19)) and "stream segregation" (7). Stream segregation is more likely to occur when the onset-to-onset duration between successive tones is short. For example, when the frequency separation between alternating high- and low-frequency tones is about five semitones listeners typically report hearing two co-occurring sequences with onset-to-onset intervals of about 100 msec or shorter, even though listeners are instructed to try as hard as they can to hear a single sequence of alternating high- and low-frequency tones; under the same task, listeners report hearing a single sequence if the onset-to-onset duration is increased (9). There is an inverse relationship between frequency and temporal proximity, such that stream segregation occurs at slower rates if the frequency separation is large; as frequency separation gets smaller, faster presentation rates are required for segregation to occur. While stream segregation can be considered as an "illusory" phenomenon, its occurrence demonstrates that the auditory system generates sequences of auditory events by grouping parts of the auditory input that have similar frequencies and that are close together in time.

3.2.2. Common onset

When frequency components are partially overlapping in time, one of the most powerful grouping principles is to from a group of frequency components that start and stop at about the same time, and to segregate into separate groups those components that start and/or stop at different times (Figure 1a). Since it has been shown that a common offset has a relatively small effect on the grouping of frequency components (20, 21), the following discussions about the role of this principle will refer to common onset (or the lack of common onset; 'onset asynchrony'), unless otherwise specified.

3.2.3. Spectral regularity and harmonicity

This principle operates by grouping frequency components that conform to a regular pattern of frequency spacing (Figure 1b). Spectral regularity is a general-purpose grouping process that is responsible for the fusion of frequency components into a single percept associated with a complex sound. When frequency components are thus fused, it is difficult to hear them as separate pure tones. This regularity most obviously occurs in harmonic complex sounds, but is also present in frequency-shifted harmonic sounds (that is, when the same frequency increment is added to each harmonic) as well as compressed/stretched harmonic series (in which each harmonic receives an increment/decrement specified as a percentage of the harmonic frequency) (22). Roberts and colleagues provided compelling evidence that spectral regularity can affect the grouping of frequency components into a complex sound (23, 24). They showed that an even-ordered harmonic can be 'heard out' more easily than neighboring odd-numbered harmonics within a complex tone containing odd-numbered harmonics of the same f0. Roberts and colleagues interpreted these results as evidence for the existence of a grouping principle that operates on the basis of regularities in the spectral pattern of frequency components. Roberts and Bailey demonstrated that this principle is distinct from the principle of harmonic grouping ('harmonicity') (23). On the basis of the principle of harmonicity, harmonics of the same fundamental frequency (f0) are grouped into a single percept. Roberts and Bailey found that spectrally regular, but inharmonic, complex sounds are also fused into a single percept, and that frequency components that break the pattern of regular frequency spacing are also heard out of these inharmonic sounds. Roberts and Bailey suggested that some of the grouping effects attributed to harmonicity could in fact be the result of grouping by spectral regularity. In fact, the principle of harmonicity could be considered as a special case of spectral regularity. Grouping by spectral regularity (or harmonicity) is distinct by the grouping of harmonics into the pitch of a complex sound which is carried out by pitch schema. This distinction can account for the finding that a mistuned harmonic can contribute to the pitch of a harmonic series even though it can also be heard out of the harmonic series as a separate pure tone. For example, Moore and his colleagues showed that a resolved harmonic can be heard out ('segregated') from the other harmonics of a harmonic series when it is mistuned by about 2% or more for f0s up to about 300 Hz (25); in spite of this, the mistuned harmonic still contributes to the pitch of the harmonic series for mistunings of about 8% (26).

3.2.4. Spatial cues

Cherry (27) first proposed the idea that grouping sounds through the use of spatial cues is one of the principles that people use to segregate attended sounds from extraneous sounds ('the cocktail-party effect'). On the basis of this principle, sounds that have the same location are grouped together. There is evidence that perceptual grouping of a sequence of pure tones of different frequency is affected by spatial location. O'Connor and Sutter (28) found that differences in the spatial location between a background and a target sequence of tones improved the perceptual segregation of the two sequences. Moreover, several studies have shown that ITD and ILD cues are likely to play a minor role in auditory scene analysis, since their effect can be overridden by other grouping cues. For example, it has been shown that a frequency component and a harmonically related complex sound can perceived at the same location even though the complex sound and the frequency component have contrasting IPD cues (and hence should be perceived at different locations) (29, 30). It is unclear whether grouping by spatial cues occurs by virtue of interaural level differences (ILD) and/or interaural time or phase differences (ITD or IPD), or whether it takes place on the basis of differences in perceived location. It is possible that either type of spatial cue could be used depending on whether a listener makes use of attention.

3.2.5. Coherent modulation

The principle that common changes in frequency components should result in their grouping (fusion) into the same percept is consistent with the principle of 'common fate' proposed by Gestalt psychologists (8) (Figure 1c). The dynamic changes in the amplitude waveform that are present in natural periodic sounds, such as the modulations caused by the opening and closing of the vocal chords, are imposed on all of the frequency components in a similar fashion. Therefore, it is reasonable to think that the auditory system has evolved to exploit coherent amplitude modulation (AM) for grouping frequency components. The "comodulation masking release" phenomenon could be considered as a demonstration of the importance of coherent AM for grouping frequency components (31). In this phenomenon the masking of a pure tone by an amplitude-modulated, narrow-band masking noise can be reduced if additional bands of noise are added, as long as the additional bands of noise are modulated at the same rate as the masking noise. Although this phenomenon has shown that coherent AM can affect performance in detection tasks, there is no clear evidence that AM coherence is effective at higher S/N ratios in tasks that require listeners to recognize auditory events such as speech sounds (32).

In addition to AM, natural periodic sounds have dynamic f0 changes that can be relatively small (like the jitter in normal phonation) or large (such as the f0 changes that occur in lexical tones and intonation patterns). Whenever there is a change in f0, the harmonics of periodic sounds undergo frequency changes in the same direction and at the same time. Therefore, grouping frequency components that have common pattern of frequency modulation (FM) should be a powerful principle of perceptual organization. However, when f0 changes in a natural periodic sound, its harmonics do not only share a common pattern of FM, but also remain harmonically related as f0 moves up or down. Therefore, it can be difficult to attribute the effects of common FM to the coherence of FM *per se*, rather than to the operation of grouping by spectral regularity or harmonicity.

3.3. Schema-based grouping processes

While typical examples of auditory schema include the processes that are used to perceive speech and music, for the purpose of this chapter the perceptual attributes of complex sounds (such as pitch, timbre, loudness, perceived location and perceived duration) will also be considered as the perceptual outcomes of schema-based grouping processes. Therefore, the meaning of 'stored knowledge' in the definition of schema will be expanded in the present chapter to include both i) those higher-level processes that involve the learning (through past experience) of complex systems of categories (such as speech and music), and ii) those processes that are used for the formation of perceptual attributes (such as the perceived location, loudness and timbre of auditory events) and which may be innate or learned. The distinction between schemas associated with perceptual attributes and higher-level schemas is an example of the different degree of complexity of schema-based processes. The existence of schemas of different levels of complexity can be justified by observing that the auditory system assigns perceptual attributes to musical and speech sounds, but these sounds are also matched with stored knowledge in the form of perceptual categories -linguistic and musical schemas of even higher level of abstraction, such as those used for the processing of syntax in speech, are outside the scope of the present discussion. In order to understand the domain-specific feature of schemas, it is useful to refer to attribute-related schemas. For example, perceived location relies on the use of ITD and ILD cues; these cues *per se* are irrelevant to the perception of other attributes such as pitch or loudness. Two types of schemas will be discussed in this chapter: schemas that generate the perceptual attributes of complex sounds and higher-level schemas. These schema-based processes will be referred to hereafter as 'attribute-specific' and 'categorical' schema, respectively. Examples of categorical schema processes used for the recognition of speech sounds, musical timbre, and other nonspeech sounds (e.g., the sounds produced by animals and familiar objects).

3.3.1. Attribute-specific schemas

Timbre, loudness, perceived location and pitch are perceptual attributes of sounds that are generated for each auditory event, whether the event can be recognized as familiar sounds or not. The auditory processes that generate these perceptual attributes are perceptual schemas that are specific to each attribute. The following sections will briefly introduce schema-based processes for the formation of perceptual attributes, followed by a discussion of higher –level schema for the perception of speech.

The standard definition of **timbre** is that it is the attribute that distinguishes two auditory events that are equal in pitch, perceived duration, and loudness (33). This definition is rather vague, and it does not take into account the interaction between timbre and pitch percepts that will be discussed later on, but it is sufficient for the purpose of the present chapter. Schema processes for timbre perception make use of static and dynamic spectral characteristics, as well as amplitude envelope properties of complex sounds (see, for example, 34), to generate timbre percepts for auditory events. Three perceptual dimensions have been identified for timbre percepts: 'brightness', 'spectral fluctuation', and 'attack quality' (35). Brightness is associated with the spectral distribution of energy; sounds with relatively more energy at the high frequencies are perceived as bright. Spectral fluctuation is related to the degree of change in the amplitude pattern of frequency components over time. Attack quality is related to the presence of high-frequency, noisy energy in the attack portion of a sound.

Loudness is the perceptual attribute that is associated with the perceived level of an auditory event. On the basis of this attribute, auditory events can be ranked from soft to loud. While it is not clear exactly how the brain estimates the loudness of complex sounds, it is likely that schema processes for loudness integrate auditory information across frequency bands in order to generate this perceptual attribute (36). Plack and Carlyon stated that the perception of the loudness of complex sounds can be modeled by transforming the estimates of the amount of excitation of the basilar membrane into loudness levels for each frequency channel (37). An overall measure of loudness is estimated as a combination of the estimates for each frequency channel.

The **perceived location** of a complex sound is an attribute that specifies the placement of an auditory event in space, along horizontal (left-right, front-back) and vertical (high-low) coordinates. The term 'lateralization' is used to describe the perceived location between the two ears for sounds that are presented over headphones; in this presentation mode, sound are typically localized within the head of the listener. Schema processes for perceived location make use of both binaural and monaural cues to assign a perceived location to a complex sound (38). Binaural cues consist of interaural differences in the level (interaural level difference, or ILD) or in the time of arrival (interaural time difference, or ITD) of the same acoustic input between the two ears. Monaural cues result from the filtering properties of the pinnae for sounds that are generated from different locations in physical space.

All natural complex sounds have physical onsets and offsets. Schema processes for perceived duration have to assign onsets and offsets to auditory events for generating the perception of duration. If this assignment cannot take place, auditory events are perceived as continuous. As demonstrated by Nakajima and his colleagues (39), the decision of when a sound starts and when it stops is analogous to the assignment of edges to visual objects. They presented a long frequency glide with a 100 ms silent gap around its middle portion. This glide was crossed in the region of the gap by a shorter continuous glide varying in the opposite direction (for example, the long interrupted glide was rising while the short continuous glide is falling in frequency over time). When listening to this crossing-glide pattern, listeners consistently perceived a long continuous glide together with a *short interrupted* glide. In other words, the physical gap in the long glide was perceptually assigned to the short glide. Nakajima and colleagues proposed that this illusion, which they named the “gap-transfer illusion”, occurs because the onsets and terminations of frequency components are grouped together according to the principle of temporal proximity. Following this proposal, the onset of the short glide is grouped with the earliest “available” offset, which happens to be that of the long glide at the point that defines the beginning of the gap. This grouping causes the perception of a short interrupted glide. Likewise, the onset of the long glide at the end of the gap is grouped with the offset of the short glide, to generate the perception of another short glide. Since the two onsets and the two offsets around the middle portion of the long glide have already been assigned to two auditory events (short glides), the long glide is perceived as uninterrupted. Although this phenomenon has been demonstrated by using pure tones, there is no reason to believe that the same principle of assigning onset and offsets to auditory events would not apply to the perception of complex sounds as well. This illusion shows that the assignment of onsets and offsets affects the perceived duration of auditory events.

Pitch is the attribute that allows us to order sound percepts from low to high, as it is done when listening to an ascending musical scale (33). There is considerable evidence to show that the pitch of a complex sound (also called “virtual” or “residue” pitch) is mainly determined by the frequency of low-numbered, resolved harmonics of a complex sound (40-43). Although various models of pitch perception have been formulated to contrast the use of temporal versus spatial cues for pitch perception (42), models based on a pattern matching mechanism have enjoyed considerable success in explaining the experimental findings about the pitch of complex tones (44-46). According to such models, the pitch schema generates a pitch percept that corresponds to the f0 of the harmonic series that provides the best fit to the frequency components of a complex sound. He grouping of frequency components into a single pitch has been likened to the operation of a ‘harmonic sieve’, since energy at frequencies that are equal or close to the harmonic frequencies of the same f0 will be ‘grouped’ into the same pitch, while other frequency components are excluded (47). Pitch is an attribute that is not assigned to the same extent to all complex sounds. For example, some auditory events such as noise and other inharmonic sounds may not have a clear, single pitch. By contrast, perceived location, loudness and timbre are attributes that apply to any auditory event, no matter what its spectral, temporal and amplitude characteristics; these perceptual attributes can therefore be considered as ‘obligatory’ while pitch is an ‘optional’ attribute of complex sounds. For this reason, it has been argued that the pitch schema has a different status from schemas for the formation of other perceptual attributes (see, for example, 48).

3.3.2 Categorical schema

There is an extensive literature on the on the acoustic characteristics that allow listeners to generate **speech** (or phonetic) percepts (see, for example, 49). Very few researchers would argue against the idea that speech perception is achieved through processes that are speech-specific, and that speech has a special status as the main modality of human communication. However, there are different views about the role of speech schema in the auditory organization of the auditory input. According to Bregman’s proposal, schemas can only select among the potential groupings of frequency components that have been generated by general-purpose grouping processes (4). The opposite view is that speech processes have the ability to extract frequency components from the input independently of the operation of general-purpose grouping processes (50-52). Even though the studies reviewed in the following sections demonstrate that phonetic perception is affected by auditory organization principles, it is still necessary to explain why speech percepts are sometimes formed in apparent violation of the principles of general-purpose grouping processes. One such case is the phenomenon of across-ear integration of formants into the same phonetic percept (53, 54). The other case is the perception of sine-wave replicas of speech (55). The existence of these phenomena could be explained by postulating that speech schemas are also involved in the organization, and not just in the selection, of frequency components.

The recognition of familiar complex sounds requires a complex system of stored knowledge that in many ways parallels that of speech and language. For example, the recognition of the timbre of musical instruments is analogous to the recognition of speech sounds. Moreover, the rules of composition of melody and harmony within a musical system bear strong similarities with the role of syntax in languages. For the purpose of the present chapter, the discussion of **nonspeech schemas** will mainly focus on those aspects that have received most attention in studies of auditory organization, that is, musical timbre and melody recognition. In one of the first demonstration of the effect of the knowledge of melodies on auditory scene analysis, Dowling (10) showed that prior knowledge of a target melody facilitated the ability to hear this melody as separate from a distracter melody, when the notes of both melodies are alternated in rapid succession (interleaved melodies paradigm). The melodies consisted of square-wave tones, with an onset-to-onset interval of 125 ms between successive notes. In one condition, the target melody was a familiar melody, and listeners were told that such a melody was presented together with the distracter melody. In another condition the target melody was unfamiliar, but listeners heard the target melody in isolation prior to judging whether this melody was present in the subsequently presented interleaved melodies. In both cases, prior knowledge of the target melody improved the listeners’ ability to segregate it from the distracter melody. Bey and McAdams (56) replicated and expanded these results by showing that hearing the target melody after the two interleaved melodies resulted in worse performance than hearing the target melody before the interleaved melodies.

4. THE AUDITORY GROUPING OF FREQUENCY COMPONENTS INTO THE PERCEPTUAL ATTRIBUTES OF COMPLEX SOUNDS

The interaction of general-purpose and schema-based processes for the perception of perceptual attributes of complex sounds will be reviewed in this section. Auditory events corresponding to complex sounds can be considered as the results of a combination of perceptual attributes. Perceived duration, perceived location, loudness and timbre are assigned to all auditory events, while pitch may be attributed only to auditory events that are periodic or quasi-periodic (that is, whose frequency components are harmonically related). This section will review the interaction between general-purpose processes and attribute-specific schemas that results in the formation of perceptual attributes. Because of the lack of studies on perceived location and duration, the following subsections will focus on pitch, loudness, and timbre. The formation of the timbre of unfamiliar sounds, typically complex sounds synthesized in the laboratory as a sum of frequency components, will be discussed in this section, while the formation of the timbre of speech and other familiar sounds will be reviewed in the next section as part of the discussion of categorical schemas.

4.1. Pitch schema

Since many natural sounds are quasi-periodic, it is hardly surprising that the human auditory system has evolved a strategy for the perceptual integration of frequency components that are integer multiples of the same fundamental frequency (f0). Moore, Glasberg and Peters (26) carried out a study to determine to what extent harmonics that were mistuned could still contribute to the pitch of a harmonic series (the 'mistuned component' paradigm). They found that when a low-numbered harmonic of a 12-harmonic complex tone was mistuned, the pitch of the harmonic series shifted in the same direction as the mistuning. Pitch shifts increased as a function of the amount of mistuning up to 3%; for larger amounts of mistuning (above 3%) pitch shifts gradually decreased, until they became zero with a mistuning of 8%. These findings have been interpreted as demonstrations of the operation of the grouping harmonics for pitch perception (5, 6).

While the existence of the principle of spectral regularity has been demonstrated in tasks requiring subject to hear out individual frequency components (22-24), one may ask whether this grouping principle has an effect on the pitch of complex sounds. Ciocca (57) compared pitch shifts produced by the 3rd, 4th and 5th harmonics within two harmonic series: one composed of all harmonics up to the 12th (complete-H tone) the other composed of odd-numbered harmonics up to the 11th (odd-H tone). The idea behind the experiment was that if spectral regularity affects the grouping of frequency components into a single pitch, then a mistuning of the 3rd and 5th harmonics should produce larger pitch shifts in the odd-H than in complete-H tone. Although the results of the first experiment seemed to support this prediction, a second experiment demonstrated that larger pitch shifts produced by mistuning the 3rd and 5th harmonics could be accounted for by the smaller amount of masking by adjacent harmonics. Therefore, these results indicate that the principle of spectral regularity *per se* does not affect pitch perception, and that the pitch schema groups components on the basis of their proximity to harmonic frequencies of the same f0.

Darwin, Hukin and Al-Khatib (58) investigated the effects of sequential grouping by using the 'mistuned component' paradigm (75). Darwin and colleagues measured the pitch shifts produced by a mistuned component that was preceded by four identical components. They found that the preceding components captured the mistuned component away from the other harmonics, thereby greatly reducing the size of pitch shifts in the complex sound. In order to prove that the effect of sequential grouping was not due to adaptation, Darwin and colleagues varied the level of a mistuned component that was presented either ipsilaterally (at the same ear as a harmonic series) or contralaterally (at the opposite ear). They found that a decrease in level produced a reduction in pitch shifts for the ipsilateral but not for the contralateral component. Since adaptation is functionally equivalent to a decrease in level, the absence of level effects for contralateral components suggest that adaptation is not a likely explanation for the sequential grouping effects observed with contralateral mistuned components.

Darwin and Ciocca studied the effect of spatial cues on pitch perception (59). They measured the contribution of a mistuned component, presented at the right ear, to the pitch of a contralateral harmonic series presented at the right ear. They found that pitch shifts caused by a contralateral component were large, though slightly smaller than shifts produced by an ipsilateral component. This effect of across-ear grouping of frequency components into a single pitch was replicated by Darwin, Hukin and Al-Khatib, who also showed that the sequential grouping effect occurred when both mistuned component and capturing tones were presented to the opposite ear as the complex tone (58). Therefore, it seems that the decision about the pitch of complex sounds is affected by sequential grouping by frequency similarity. By contrast, the across-ear integration of frequency components into a single pitch shows that the effect of the spatial cues is very small.

Darwin and Ciocca employed the mistuned component paradigm to investigate the effect of common onset (onset asynchrony) on pitch perception (59). They measured pitch shifts in a complex sound composed of the first 12 harmonics of a 155-Hz f0 as a function of the mistuning of the 4th harmonic. The mistuned component could start either at the same time or up to 640 ms before the other harmonics. The results showed that an onset asynchrony of 160 ms produced a reduction in the contribution of the mistuned component to the pitch of the complex sound. With an asynchrony of 320 ms or longer this contribution was eliminated, showing that a frequency component is not integrated into the pitch of a complex sound if it starts long before the other harmonics of the complex sound. Ciocca and Darwin (60) showed that the effects of asynchrony could be reduced by the presence of a complex tone that was i) synchronous with the leading portion of the asynchronous component, and ii) harmonically related to the mistuned component. In a second experiment, Ciocca and Darwin found that the expected amount of adaptation produced by the leading portion of the asynchronous frequency component (sine precursor) was increased by embedding the precursor into a complex tone composed of unresolved frequency components (complex precursor). The complex precursor started and stopped at the same time as the sine precursor. They found that the complex precursor produced a decrease in pitch shifts, relative to the shifts produced a sine precursor, only at an 80-ms asynchrony. This outcome was expected if adaptation is responsible for the effect of onset asynchrony. However, a complex precursor that lasted 160 ms or longer produced *larger* pitch shifts than the corresponding sine precursor. These findings indicate that adaptation is likely to be responsible for decrease in pitch shifts at short asynchronies, but that perceptual grouping is the main factor at long asynchronies.

The pitch schema has been found to be very sensitive to harmonic relations among frequency components, but to be tolerant of relatively large amounts of onset asynchrony. For example, Hukin and Darwin (60) showed that the amount of onset asynchrony required to remove a harmonic from a vowel's pitch is much longer (160 ms or longer) than the asynchrony that prevented the same component from affecting the same vowel's identity (80 ms). The tolerance of pitch schema for onset time differences is consistent with the findings from two studies that employed different experimental paradigms. Ciocca and Darwin (14) used the mistuned component paradigm (26), but employed a mistuned component that either stopped before the complex sound started (pre-target condition), or started after the complex tone had stopped (post-target condition). In both conditions, a silent gap of various durations was introduced between the complex sound and the nonsimultaneous components. The results showed that pre-target components had relatively little effect on the pitch of the complex tone, but post-target components made a contribution even when a silent gap of 80 ms. Since both the complex tone and the nonsimultaneous mistuned component had a duration of 90 ms, they estimated that the pitch integration period has a duration of 170-250 ms. This estimate is in agreement with measurements of the virtual pitch generated by four 40-ms, nonsimultaneous frequency components presented in a background of broadband noise (16). With such sequences, reliable pitch percepts could be obtained with intervals as long as 45 ms, corresponding to an integration period of about 200 ms. These studies show that pitch schema can accumulate evidence about the pitch of an auditory event over a period of about 200 ms ("pitch integration period) (15).

Darwin, Ciocca and Sandell found that coherent FM affects the grouping of frequency components into the same pitch (62). They used the mistuned component paradigm described earlier, and showed that coherent FM applied to all frequency components produced larger pitch shifts and over a larger range of mistuning than unmodulated stimuli. It is unclear whether the effect was due to the coherence rather than to the mere presence of FM. Darwin, Ciocca and Sandell also investigated the effects of common AM on pitch perception. They showed that the contribution of a mistuned component to the pitch of a complex sound was not affected by the presence of (6 or 17 Hz) sinusoidal AM (62). When harmonic series are used to study coherent FM, it is difficult to separate effects of coherent modulations from the effects of spectral regularity/harmonicity because FM-modulated frequency components also maintain their harmonic relationships.

To summarize, pitch schema select components on the basis of their proximity to harmonic frequencies of the same f0, and can integrate frequency components into a single pitch over a period of up to about 200 ms. Pitch schema are insensitive to spatial cues, but are affected by sequential grouping and common onset, although they are tolerant of relatively long onset asynchrony. Finally, the presence of FM but not AM has been found to facilitate the integration of frequency components into the same pitch. However, it is not clear whether the effect of coherent FM is effective *per se*, or whether frequency components have been grouped because they are harmonically related.

4.2. Loudness schema

It had been shown that when a soft, low-pass filtered noise (S sound) and a loud wide-band noise (L sound) are alternated, listeners typically hear the S sound continuing through the L sound. Moreover, the continuation of the S sound causes the perceptual grouping of the low-frequency portion of the L sound into a continuous soft sound, which causes the perception of a softer, pulsating L sound. This phenomenon, known under the name of 'auditory continuity' (4) or auditory 'induction' (63), shows that the perceptual segregation of part of the L sound causes its loudness to decrease. McAdams, Botte and Drake conducted a systematic investigation of this phenomenon (64). They studied the perception of loudness in sequences of either pure tones or narrowband noise bursts, which alternated in level (High-Low-High-Low-…) and whose duration and levels were selected so as to give rise to auditory continuity. McAdams and colleagues reasoned that if the loudness of the stimuli is affected by the auditory organization of the bands of noise, then the perception of continuity should cause the subtraction of the level of the (continuous) low-level sound from the (intermittent) high level sound. Therefore, they predicted that the perception of auditory continuity would result in the high level, intermittent sound having a softer loudness than when it is heard in isolation. They results were in agreement with this prediction. Moreover, the matched levels of the intermittent sound supported the idea that loudness is computed after the perceptual segregation of the high level sound into two percepts, one intermittent and one that is grouped with the softer sounds to produce a continuous percept.

4.3. Timbre schema

Bregman and Pinker (65) provided a first demonstration of the effect of sequential grouping on the timbre of complex sounds. They alternated a pure tone (A sound) with a complex sound that contained two harmonically-unrelated pure tones (B and C) of equal duration and amplitude. They found that when sounds A and B were far apart in frequency, listeners heard a high frequency tone (A) alternating with a complex tone (B plus C; Figure 2a). By contrast, A and B were grouped into an ABAB... sequence when the frequency of A was close to the frequency of B. As a consequence of this grouping, sound B was perceptually segregated from sound C (Figure 2b); this segregation caused a change in the timbre of the complex sound, which was perceived as less 'rich'. Using a similar paradigm, Steiger and Bregman (66) investigated the grouping of complex sounds constituted by simultaneous gliding frequency components. They found that sequential grouping was strongest when the glides corresponding to the A and B sounds had similar frequency range and direction of change. The alignment of the A and B glides along a log frequency trajectory did not improve grouping, relative to the grouping effect of the similarity of frequency range.

Bregman & Pinker (65) also investigated the competition between two general-purpose grouping processes, sequential grouping by frequency similarity and the simultaneous grouping by common onset. To do this, they manipulated the onset asynchrony between the B and C sounds. They found that if sound C started 29 or 58 ms prior or after the onset of sound B, then it was easier for the listeners to hear the ABAB… sequence. As expected, the asynchrony between sounds B and C caused the complex sound to have a 'pure' timbre. Moreover, the segregation between the B and C sounds was stronger as the asynchrony between them increased. These findings indicate that a frequency component makes a smaller contribution to the timbre of a complex sound when its onset time differs from that of other components. Using a similar paradigm, Dannenbring and Bregman (20) demonstrated that the effect of asynchrony is stronger if the asynchronous component started before (onset asynchrony) or ended after (offset asynchrony) the other frequency components (rather than starting after, or ending before the other component). Moreover, they found that the segregation effect of onset asynchrony was stronger than that of offset asynchrony.

While there is a lack of studies on the effects of spatial cues on timbre by humans (although some findings about categorical schema will be reviewed in a later section), there is evidence that other animals such as frogs display a similar disregard for spatial cues in timbre perception as humans do for pitch perception. Farris, Rand and Ryan (67) found that the responses of female frogs to spatially separate frequency components of male calls were affected by the degree of spatial separation. They showed that frequency components from different locations were not as strongly fused as those presented from the same location. However, the across-location integration occurred even for relatively large separation angles (up to 135 degrees) showing that these animals are insensitive to spatial cues.

There is some evidence that coherent AM affects the grouping of frequency components. Bregman, Abramson, Doehring and Darwin (68) studied the perceptual fusion of a complex sound composed of two simultaneous pure tones. They showed that the perceptual fusion of the two tones was stronger when both had the same modulation frequency than when they had a different modulation frequency, even when the tones were not harmonically related. Bregman, Levitan and Liao (69) replicated these findings by using a larger range of carrier and modulation frequencies. Since both of these studies measured the capacity to hear out one of the modulated frequency components, rather than directly measuring the timbre of the complex sound, it is unclear whether the timbre of the complex sound was affected.

To summarize, the contribution of a frequency component to the timbre of a complex sound can be reduced if the component is perceptually segregated from the other components. A frequency component can be segregated by either embedding it into a sequence of pure tones of similar frequency, or by introducing an asynchrony between the component and other components of a complex sound. Differences in spatial cues do not strongly affect the integration of frequency components into the same timbre. There is indirect evidence that common AM can affect the timbre of complex sounds by increasing the perceptual integration of its frequency components.

The discussion about the factors that affect the grouping of frequency components into auditory events - that is, percepts that have a perceptual onset and offsets, and that have perceived duration, timbre, pitch loudness and perceived location- has focused on the interaction between general-purpose grouping and attribute-specific schemas. The next section will present evidence for a similar interaction that occurs in the formation of percepts that can be recognized as belonging to perceptual categories that are stored in memory and that have been learnt through past experience.

5. THE EFFECTS OF AUDITORY GROUPING ON THE RECOGNITION OF COMPLEX SOUNDS

Although phonetic identity could be considered as a more specific type of timbre perception, a much larger number of studies have been conducted on speech perception because of the special status of speech perception for normal human communication. For the same reason, timbre and speech have historically been investigated as distinct research areas. The discussion about the interaction between general-purpose and categorical schemas for the formation of speech or familiar timbre percepts will therefore be presented in separate sections for speech and nonspeech schemas.

5.1. Speech schema

Research on the effects of general-purpose grouping processes on the perception of a single speech percept can be classified into two lines of research. The first line of research was motivated by the intention to challenge the 'motor theory' view of speech perception proposed by Alvin Liberman and his colleagues at Haskins Laboratories (70, 71). This view proposes that speech is a separate module from other auditory processes, and that speech perception is not affected by the operation of (general-purpose) grouping processes or other nonspeech schema (‘independence’ idea). This view relies on evidence from the perception of single (usually consonant-vowel, or CV) syllables, such as the phenomenon discovered by Rand (53). He presented the F2 and F3 transitions of synthetic /da-ga/ syllables to one ear, and the rest of the syllable (base) to the opposite ear. Rand found that this stimulus was perceived as two auditory events: a syllable whose identity was determined by the fusion of the F2-F3 transitions with the base, and a nonspeech (chirp-like) sound. This phenomenon was later called “duplex perception” to highlight the fact that the “isolated” F2-F3 transitions were perceived at the same time as part of two percepts (72). The simultaneous use of the information provided by the isolated formant transition for building a speech and a nonspeech percept was interpreted as evidence of the distinct operation of processes for the perception of speech and nonspeech sounds (71). The second line of research, mainly adopted by Chris Darwin and his colleagues at the University of Sussex, explored the role of auditory grouping processes on the perception of vowels (specifically the /i/-/e/ contrast) and of the /ru/-/li/ contrast.

5.1.1. Sequential grouping

Ciocca and Bregman used a modified version of the duplex perception paradigm to challenge the independence view (73). They embedded the isolated F3 transition of /da-ga/ duplex stimuli into a sequence of transitions of identical or higher center frequency. As expected on the basis of sequential grouping, the identical transitions -but not the higher-frequency transitions- “captured” the isolated transition away from the base. As a result, the integration of the isolated transition with the base was significantly reduced, although not completely eliminated, causing the phonetic percept to sound like the base stimulus alone. Ciocca and Bregman also demonstrated that the effect of sequential grouping could not be entirely due to adaptation because a similar reduction in the integration of the isolated transition with the base occurred when the isolated transition was preceded and followed by formant transitions that were aligned with it on a linear frequency trajectory. Another demonstration of the effect of sequential grouping by frequency similarity was given by Darwin and Hukin (74) who employed the /I/-/e/ paradigm. In this paradigm, the energy of a 500-Hz harmonic in the first formant (F1) region of a synthetic vowel was manipulated. This manipulation causes the perceived F1 frequency to shift. For example, a lowering in the energy level of the 500-Hz harmonic results in the perception of a higher F1 frequency, which in turn makes the vowel identity to change from /I/ to /e/. Darwin and Hukin showed that the perceptual capturing of the 500-Hz harmonic into a sequence of 500-Hz pure tones was equivalent to physically removing the harmonic from the vowel. This effect demonstrates that the grouping of frequency components into speech percepts is subject to the same grouping effects as those that occur for the perception of other perceptual attributes of complex sounds.

5.1.2. Common onset

In fluent speech, different harmonics of the same voiced sound can start at different times, as pointed out by Darwin (21). For example, the formant frequency changes that occur during a formant transition can cause adjacent harmonics to start at different times. For the formant transitions of glides and liquids, which typically have durations of about 100-200 ms, at an f0 of about 100 Hz one can observe onset asynchronies of 10-20 ms among low-numbered harmonics. Since in these speech sounds the different harmonics of a formant transition are grouped into a single phonetic percept, one may wonder whether onset asynchrony has an effect on the perception of speech.

A demonstration that a lack of common onset affects speech perception was given by Cutting (54). He presented the two formants of a synthetic, two-formant CV syllable to separate ears, a condition he labeled ‘spectral fusion’. He also included a condition in which only the second formant transition was presented to one ear and the rest of the syllable to the opposite ear (the earlier described ‘duplex perception’ stimulus). He found that the percentage of ‘fused’ percepts decreased with increasing asynchrony with both stimuli, although listeners were still able to integrate the formants into a single speech percept with asynchronies up to 80 ms. Repp and Bentin replicated these results with duplex stimuli and observed an asynchrony of 100 ms did not completely eliminate the integration of the isolated transition with the base (75). Additional evidence for the effect of onset asynchrony on the perceptual grouping of formants comes from a study by Darwin, who used an ingenious paradigm consisting of a four-formant synthetic syllable (3). This syllable is identified as /ru/ when all formants are synthesized at the same f0 but, when the second formant (F2) is physically removed from this stimulus, listeners perceive the remaining three formants as the syllable /li/. Darwin showed that with an onset asynchrony of 300 ms between F2 and the other formants, there was an increase of about 15% in the number of /li/ responses.

In attributing the effects of onset asynchrony to perceptual grouping, it is necessary to rule out alternative explanations based on peripheral auditory mechanisms. One possibility is that the decreased integration of an asynchronous component is due to the fact that the leading portion of the asynchronous component produces peripheral adaptation (76). This explanation can be rejected on the basis of evidence from two findings. First, Darwin and Sutherland demonstrated that the leading portion of an asynchronous component (“precursor”) could be perceptually removed from the portion of the asynchronous component that continued through a vowel. To this effect, they employed a captor tone that i) started at the same time as the precursor, ii) stopped as the remaining components of the vowel started, and iii) was harmonically related to the asynchronous component (77). Since adaptation at the frequency of the precursor should also have occurred in the presence of the captor tone, Darwin and Sutherland concluded that adaptation could not account for the reduced integration of an asynchronous harmonic with the vowel. However, Roberts and Holmes recently provided some evidence that the captor effect may be due the inhibition properties of cells within the cochlear nucleus instead of perceptual grouping (78, 79). Second, it has been demonstrated that the contribution of an asynchronous harmonic to vowel identity is also reduced when the harmonic ends after the other harmonics of the vowel (offset asynchrony), although the effect of offset asynchrony is smaller than that of onset asynchrony (77, 80). Therefore, the current evidence supports the view that grouping by common onset affects the integration of frequency components for the purpose of processing vowel identity.

5.1.3. Spatial cues

A demonstration of the ability of the speech schema to integrate information across distinct spatial locations comes from the earlier described phenomenon of duplex perception. Although the across-ear fusion of stimuli into a phonetic percept indicates that spatial cues do not strongly affect phonetic perception, it would be a mistake to conclude that spatial cues have no effect at all. Ciocca, Bregman, and Capreol (81) reported an effect of spatial cues on speech perception. They used a variation of the duplex perception paradigm in which the periodic isolated transition was replaced by a sine-wave glide, whose frequency followed the center frequency of the original formant (50). Ciocca and colleagues reported that the effect of the sine-wave transition on the identity of the base decreased significantly if the transition was presented either binaurally (to both ears), or to the opposite ear as the base (contralateral presentation), relative to an ipsilateral presentation (same ear as the base). The fact that the binaural transition produced a decrease in phonetic integration is particularly significant if one considers that a sine-wave transition was also presented to the same ear as the base under this condition, as was the case in the ipsilateral condition. If speech schema were able to group components by ignoring the perceptual grouping on the basis of spatial cues, one would have expected the same amount of integration between the base and the sine-wave transition to occur in both conditions.

5.1.4. Common modulation

While the effects of AM on the grouping of frequency components on the perception of single speech sounds are unknown, Gardner and Darwin (82) found that neither rate or phase differences in FM between a harmonic of a vowel and the other harmonics had an effect on vowel quality. An investigation on the effect of FM incoherence between F2 and the other formants of a /ru-li/ syllable also reported no effects of this cue on the grouping of formants (83). These results argue against a role of coherent FM as an important grouping factor, and are consistent with findings that listeners cannot detect incoherent FM, once within-channel mistuning cues are removed (84).

5.1.5. Harmonicity

Since many speech sounds are voiced, it is reasonable to expect that harmonics of the same f0 are grouped into the same speech sound, and are segregated from harmonics of a different f0. Different processes may be at play when harmonics of the same f0 are integrated into a single vowel sound. First, it is possible that harmonics of the same f0 are grouped together because of the principle of spectral regularity since harmonic relationships are a special case of spectral regularity. This grouping principle may be responsible for the fact that harmonics of the same f0 are fused into a single percept, rather than being heard as separate pure tones. Second, the pitch schema is also likely to be active since voiced sounds typically have a clear pitch. Therefore, the pitch schema is likely to group harmonically-related frequency components into a single pitch. Third, speech processes may also group harmonics of the same f0 into a single voiced sound on the basis of a harmonicity-like rule for the purpose of phonetic perception. Darwin and Gardner (85) employed the /I-e/ paradigm to determine whether the mistuning of a harmonic of a vowel can cause a reduction in its contribution to vowel identity, in the same way in which it reduces the contribution of a harmonic to the pitch of an auditory event (21). They found that both downward and upward mistunings of the 500-Hz harmonic of the vowel of produced results expected on the basis of grouping by harmonicity (i.e., a perceptual exclusion of the harmonic from vowel quality). However, even relatively large mistunings (8% of the harmonic frequency) failed to eliminate the mistuned harmonic's contribution to vowel identity. These findings replicated by Gardner, Gaskill and Darwin (83). They introduced f0 differences between F2 and the other formants of a /ru-li/ stimulus, and found that relatively large f0 differences (32 Hz or larger) were needed to change the phonetic percept from /ru/ (perceived with no f0 differences) to /li/ on about half the trials. These studies demonstrate that the speech schema is tolerant of violations of the harmonicity principle.

5.2. Nonspeech schemas

While it is clear that the operation of principles of organization can affect the timbre of unfamiliar auditory events such as the complex sounds used in the experiments discussed in the previous section, there is also evidence that some degree schema of onset asynchrony among frequency components is tolerated in the recognition of familiar sounds. For example, Risset and Matthews observed that some harmonics of trumpet tones can reach their maximum amplitude about 40 ms earlier than other harmonics (86). The slow rise in level of some harmonics and the rapid rise time of other harmonics approximate the onset asynchrony conditions that have been studies in laboratory experiments. In spite of this onset asynchrony, the harmonics of a trumpet tone are fused into a single timbre.

There is very limited experimental evidence about effects of grouping on the recognition of familiar nonspeech sounds. Fowler and Rosenblum carried out a study on the integration of spatially separate signals into a single familiar timbre (87). They obtained two spectrally distinct stimuli by filtering a slamming metal door sound around a 3 kHz cut-off frequency. The low-pass portion of the spectrum (by itself) sounded like a 'wooden door' sound, while the high-pass portion sounded like a 'shaking' sound. When the sounds were played simultaneously to both ears at their original levels, the wooden sound and the shaking sound fused into a single percept of a slamming metal door. However, when the level of the shaking sound was increased listeners heard both a metal door sound and a shaking sound. In other words, the high-pass portion of the original sound contributed at the same time to two percepts: the metal door and the shaking sounds. This study provides a striking demonstration of 'duplex perception', a phenomenon that was discussed in detail in the subsection on the speech schema. But the analogy with the speech phenomenon does not stop here, because when the two sounds were presented simultaneously to opposite ears, the most common response was the perception of the original metal door sound together with the shaking sound. These findings further demonstrate that the simultaneous use of the same piece of auditory input by different schemas is not confined to the perception of speech. Moreover, the across-ear grouping of two (spectrally different) sounds shows that spatial cues have little effect on the organization of frequency components into a single familiar sound.

5.3. The interaction between auditory organization processes and categorical schemas

The research reviewed in this section has shown that the speech schema is not affected by differences in spatial cues. Harmonic relations among frequency components affect their grouping into a single speech sound, although a mistuned harmonic of a vowel can contribute to vowel quality for much larger mistunings than those required to remove a harmonic from the pitch of a complex sound. Darwin and Gardner remarked that the need for larger amounts of mistuning to exclude a harmonic from vowel quality suggest that speech schema do not give as much weight to this grouping principle as does the pitch schema (85). Although the grouping principle of asynchrony operates in the same basic way across different types of stimuli, it is possible that the there are differences in the way frequency components and formants are grouped into speech percepts. Speech schema may be more tolerant of onset asynchrony among formants because natural speech sounds often contain formants that start and stop at different times (such is the case for the formants of aspirated stops and liquids). By contrast, speech schema may be less tolerant of asynchronous frequency components, since vowel sounds with a single asynchronous frequency component (like the stimuli used in the studies reviewed in this section) are not normally found in natural speech. Common FM or AM cues do not appear to have a role in the grouping of frequency components into speech percepts. The latter conclusion is somewhat surprising given the presence of coherent frequency and amplitude modulations contained in natural speech sounds. Finally, sequential grouping is a very effective cue in the segregation of frequency components or formants from speech.

Although the conclusions about the relative strength of the various general-purpose grouping processes on the recognition of familiar speech and nonspeech sounds was derived by examining the results of the effects of individual grouping processes, similar conclusions can be drawn by considering the findings of studies that measured the effects of combinations of general-purpose grouping processes. For example, Darwin and Hukin used the /I-e/ paradigm to measure the effects of ITD differences on the contribution of a 500-Hz harmonic to vowel quality (88). They found that the contribution was minimally affected by ITD differences *per se*, but that the segregation of the harmonic from the vowel was strong when the harmonic was embedded in a sequence of one or more 500-Hz, "capturing" tones that had the same ITD as the harmonic. Moreover, the combined effect of ITD and sequential grouping was found even though the target harmonic and the vowel were lateralized at one ear, while the capturing tones were lateralized at the opposite ear. In a follow-up study, Darwin and Hukin reported similar cumulative effects of the combination of ITD cues with onset asynchrony and harmonicity cues (89). These studies further demonstrate that i) sequential grouping and onset asynchrony have strong effects on the perception of speech sounds, while spatial cues have relatively little effect, and ii) the combination of two or more grouping cues provides stronger grouping effects than individual cues alone.

6. THE AUDITORY ORGANIZATION OF AUDITORY STREAMS AND SIMULTANEOUS AUDITORY EVENTS

While the research findings presented in the previous sections discussed the interaction between general-purpose and schema processes for the formation of the perceptual attributes of individual auditory events and for the recognition of familiar sounds, most of the sounds we encounter in natural listening environments consist of sequences of complex sounds ('auditory streams'). Moreover, in natural listening situations auditory streams are not heard in isolation but are perceived in the presence of simultaneous complex sounds or streams. For example, one might listen to one talker in the presence of extraneous noises. This section will first discuss the sequential grouping of complex sounds into auditory streams. The perceptual organization of co-occurring auditory streams will then be reviewed.

6.1. The sequential organization of auditory streams

The sequential grouping principles that apply to auditory streams are generally similar to the principles of proximity that are effective for the sequential grouping of pure tones, although the more general principle of spectral or f0 proximity will replace the frequency proximity principle of sequential grouping. Most of the studies on sequential grouping of complex sounds use paradigms in which two-tone sequences are presented in an ABABAB… or ABA-ABA… patterns. Listeners perceive these patterns as composed of one or two sequences, depending on the spectral or f0 differences, and on the time interval between the onsets of the A and B sounds. For example, in order to clearly hear two sequences that are overlapping in time - one sequence of A sounds (A-A-A-…) and one of B sounds (-B-B-B… or –B---B-…)- the frequency separation should be about 5 semitones or larger for intervals shorter that 100 ms, and about 12 semitones or larger for intervals of 150 ms (9).

6.1.1. Nonspeech sounds

The strong segregation observed with sequences of alternating pure tones and narrowband noise bursts indicates that large differences in timbre can provide strong evidence for timbre-based grouping (90, 91). Although these findings were obtained with non-natural stimuli, there is strong experimental support about the role of timbre in the perception of sequences of complex sounds. Bey and McAdams (92) investigated grouping by timbre similarity by using an interleaved melody paradigm in which the notes of a target melody were alternated with those of a distracter melody. They used interleaved melodies that differed in timbre as well as in mean pitch. Listeners had to judge whether they heard a probe melody as one of the melodies in the interleaved sequence. The probe melody was presented after, instead of prior to, the interleaved sequence to minimize the use of prior knowledge of the probe melody on the ability to hear it within the interleaved sequence. They found that the amount of segregation was positively correlated with the amount of timbre difference between the target and the distracter melodies.

While timbre is perceived on the basis of the properties of i) the spectral distribution of energy and ii) the amplitude envelope (34), there is evidence that the similarity in spectral characteristics might be more important than amplitude envelope cues for the sequential grouping of auditory events. On the basis of their findings, Hartmann and Johnson concluded that i) the sequential organization of auditory streams is the outcome of the grouping of peripheral frequency channels that are stimulated by spectrally similar signals ('peripheral channeling'), and that ii) amplitude envelope cues do not affect stream segregation (93). Both of these conclusions have been challenged. Iverson employed sequences composed of notes of the same pitch (middle C, 272 Hz), but with alternating instrumental timbres (T1T2T1T2…) He showed that the sequences could be perceptually organized on the basis of timbre similarity -specified by manipulating spectral and onset-time cues, and that both cues contributed to the observed ratings of timbre-based segregation (94). Singh and Bregman extended Iverson's findings (95). They conducted a study to determine the relative importance of spectral and amplitude envelope cues in the grouping of ABA-ABA… sequences. Timbre differences between A and B tones were obtained by combining one of two spectral compositions (first two vs. first four harmonics) and two amplitude envelopes (fast vs. slow attack time). Although spectral differences produced the strongest segregation, Singh and Bregman found a significant effect of amplitude envelope, supporting Iverson's finding that envelope cues do play a role in the perception of sequences of complex sounds. Regarding the second conclusion, several studies have demonstrated that peripheral channeling is not a necessary condition for sequential grouping (11, 96, 97). In a recent study by Roberts, Glasberg, and Moore varied the phase relationships among the frequency components of complex sounds composed of unresolved harmonics in order to obtain differences in pitch and timbre, but not in spectral distribution of energy, between successive A and B sounds of ABA-ABA… sequences (98). Their results provided clear evidence that differences in pitch and timbre, and not in spectral distribution, were responsible for the perceptual organization of their sequences of complex sounds.

Although there is strong evidence that peripheral channeling is not the only grouping mechanism for stream segregation, spectral cues have been found to generate stronger perceptual grouping effects than f0-based cues (11, 99). These findings might indicate that grouping by spectral similarity is an obligatory (pre-attentive) general-purpose grouping process, while grouping by pitch similarity (on the basis of f0 cues) is a schema-based process requiring attention. Alternatively it is possible that spectral differences among sounds are processed as timbre cues, and that grouping by timbre similarity is a stronger factor than grouping by pitch similarity. Other studies demonstrated that the interaction between the effects of grouping by timbre and pitch similarity (100, 101). For example, Singh employed sequences of four complex tones composed of four consecutive harmonics (101). These sequences were constructed so that if listeners grouped the tones by pitch similarity then they would hear a low-pitch pair followed by a high-pitch pair in rapid succession. By contrast, if listeners grouped the tones by timbre similarity then they heard two ascending intervals at a slower tempo. Singh found that grouping by timbre prevailed when successive notes had large differences in the spectral region while pitch differences between successive pairs of notes were small, but the opposite pattern of responses was observed with large pitch differences but similar timbre. The most interesting aspect of these results is that the same timbre difference could give rise to a pitch- or timbre-based organization depending on the size of the pitch differences between tone pairs. The finding that grouping by timbre can either enhance or override grouping by pitch was later replicated using interleaved sequences of musical notes (92, 94, 102).

6.1.2. Speech sounds

Although most experimental studies on speech perception have used CV or CVC syllables as the target stimuli, speech communication in natural listening environments typically involves listening to sequences of words, each containing one or more syllables. The sequential integration of speech sounds into fluent speech sentences is problematic because speech sounds include phonemes with very different timbres; for example, the noisy timbre of fricatives, such as the /s/ and /sh/ sounds in English), is very different from the timbre of periodic vowel sounds. And yet, when listening to fluent speech listeners do not perceptually segregate into sequences of voiced and noisy sounds as it happens with sequences of nonspeech sounds. It would be incorrect to conclude that general-purpose auditory grouping processes do not play a role on the sequential grouping of speech sounds since there is clear evidence that sequences of vowels and consonants can be perceptually segregated on the basis of timbre (103) or pitch differences (104, 105) between successive speech sounds. In an early study, Broadbent and Ladefoged investigated the interaction between general-purpose and schema-based grouping processes in the perception of sentences (106). They studied the combined effects of harmonicity and spatial cues by presenting the first and second formants in the synthetic sentence "What did you say before that" to opposite ears. They found that most listeners heard a single voice at one location as long as the formants were generated at the same f0. However, differences in f0 between the formants caused the perception of two sounds at two locations. In spite of these demonstrations of the effects of auditory grouping processes on speech perception, it is likely that the speech schema play a major role in preventing the perceptual segregation of the sequences of vowels and consonants in the sentences produced by a single talker.

Perhaps the most striking demonstration of the capacity of speech schema to fuse together frequency components into sequences of words comes from the earlier described phenomenon of sine-wave speech (55). It is important to point out that the sinusoidal components that follow the center frequency of the formants are not harmonically related to each other, since the frequency changes of formants are not related to f0 changes in voiced speech. Moreover, the sinusoidal components for a given word may have asynchronous onsets, and there can be sudden discontinuities in the frequencies of successive portions of the same sine-wave formant tracks. There is little doubt that speech processes can derive intelligible speech from such unusual stimuli because correct phoneme identification rates of about 70 % can be achieved (51). The crucial question is how can this happen in spite of the violation of some of the most effective principles of auditory organization? In fact, the perception of sine-wave sentences has been cited as evidence for the view that auditory grouping does not affect speech perception (51). A first response to this view is that there is in fact evidence that the perception of these stimuli as meaningful speech probably involves a considerable amount of "top-down" knowledge on the part of the listeners. Most listeners do not perceive sine-wave sentences as speech stimuli at first, and need to be told explicitly to hear them (one could say "interpret them") as speech (55). Therefore, the ability to identify phonemes in sine-wave speech is likely to require linguistic knowledge beyond the level of phonetic categories, and does not necessarily support the idea that acoustic signals are automatically converted into articulatory patterns by a phonetic module (52). Second, it has been found that general-purpose grouping processes affect the perception of sine-wave speech. Remez, Rubin, Berns, Pardo and Lang presented four-formant sine-wave replicas of sentences either binaurally, or with the F2-equivalent sinusoid presented to one ear and the remaining sinusoids to the opposite ear (dichotic presentation) (51). The intelligibility of the sentences decreased in the dichotic condition, relative to the binaural condition, showing that spatial location affected the integration of the F2 sine-wave with the other sine-waves. Interestingly, speech recognition in the dichotic condition was better than in a condition in which the F2 sine-wave was physically removed, providing yet another example of the 'duplex' use of the auditory input. Carrell and Opie showed that the perception of sine-wave sentences is affected by another general-purpose grouping process, namely common modulation (107). They reported that the intelligibility of sine-wave sentences improved when the sine-wave components were amplitude-modulated at a rate of 50 or 100 Hz. They also showed that intelligibility was higher when the sine-wave formant analogues had the same frequency of AM (coherent modulation) than when they were modulated at different frequencies (incoherent modulation). To the best of my knowledge, this is the only evidence that common AM is effective in grouping frequency components for the recognition of familiar sounds.

6.2. Auditory grouping of simultaneous auditory events

Although the situation in which several sound sources are active at the same time is the most common in natural environments in which several nonspeech sound sources are active at the same time, research on this topic has mainly focused on the perception of simultaneous speech sounds. This section will first focus on the few available studies on the perceptual organization of simultaneous nonspeech sounds, and then present a larger body of evidence on the simultaneous grouping of speech sounds.

6.2.1. Nonspeech sounds

After the previous review of studies on the effects of spatial cues on the grouping of auditory events, it will come as no surprise that spatial factors can be overridden by other factors in the sequential organization of complex sounds. A clear demonstration of this conclusion is provided by the 'scale illusion' (108). In this illusion, tones belonging to an ascending and a descending musical scale are alternated between the ears, such that at any one time the left ear receives a note from the ascending scale while the right ear receives a note from the descending scale (or vice versa). Most listeners perceive this dichotic stimulus as a sequence of high tones at one ear and a sequence of low tones at the other ear. This percept corresponds to a grouping of tones by pitch similarity, rather than by ear of presentation, since each ear is physically presented with sequences of alternating high and low tones. Although in the original version of the illusion the notes of the musical scales were pure tones, a subsequent study replicated these findings with complex tones showing that, although listeners mainly grouped notes on the basis of pitch similarity, other attributes such as timbre, loudness and ear of presentation also had an effect (109).

6.2.2. Speech sounds

While the harmonicity principle can be employed to decide which frequency components belong to a single pitch or vowel, the same principle could also be used to segregate two simultaneous speech sounds on the basis of f0 differences. In a series of experiments, Brokx and Nooteboom (110) found that the intelligibility of nonsense sentences presented at the same time as other speech sounds increased as the f0 separation between the sentences and the distracter sentences increased from 0 to 3 semitones. Similar results have been observed in several studies that investigated the perception of two simultaneous vowels ("double-vowel experiments). Scheffers (111) replicated Brokx and Nooteboom's findings by presenting two simultaneous vowels, and by asking listeners to identify both vowels. He found that small f0 differences (up to about 1 semitone, relative to a 150-Hz f0) accounted for most of the improvement in recognition. These basic results were subsequently replicated and extended, although substantial improvements in performance were typically reported only for small f0 separations (up to 0.5 semitones) (112-114). Culling and Darwin (112) found that the improvement at small f0 separations was largely due to differences in the frequency of harmonic in the F1 region, while larger f0 differences in higher formants are needed to improve performance. Culling and Darwin (115) subsequently demonstrated that the beating among harmonics of different vowels in the F1 region allows one vowel to be more clearly heard than the other during each beating cycle. Since vowel duration of about 200 ms is needed to exploit this cue (116), it is likely that improvements in the identification of simultaneous vowels at small f0 separations could not be observed with the sentences used by Brokx and Nooteboom as vowels produced within fluent speech are typically shorter than 200 ms.

McAdams studied the effect of common modulation on the perception of simultaneous vowels (117). He presented three simultaneous vowels to his listeners, a target vowel and two background vowels, and found that introducing FM in the target vowel made it sound more prominent when the background vowels were unmodulated. However, he did not find any difference between introducing coherent or incoherent modulation between target and background vowels. Summerfield and Culling (118) reached a similar conclusion of the effects of coherent FM by measuring thresholds for the identification of a vowel in the presence of another vowel that acted as a masker. Summerfield and Culling also investigated the effects of the coherence of the phase of AM by modulating two vowels, a target and a masker vowel, at the same rate of modulation but either in or out of phase. They found no evidence that coherent AM *per se* had any effect on vowel identification.

Spatial grouping based on ITD was found to be a relatively weak cue for the grouping of simultaneous frequency components into vowels or for segregating frequency components from vowel sounds (section 5.1.3). By contrast, Darwin and Hukin (119) proposed that ITD might be a stronger cue when it comes to grouping auditory streams. They tested this hypothesis by presenting two simultaneous carrier sentences each containing a target word. The sentences were processed so that they had a flat f0 contour, and differed in f0 by 0, 1, 2, or 4 semitones. Moreover, interaural time differences were varied so that the sentences could be heard either at the same or at different lateralizations over headphones. The results showed that listeners most often selected the target word that had the same ITD as the attended sentence, independent of whether the target word had the same or different f0 than the attended sentence. Darwin and Hukin further demonstrated that i) listeners group target words and carrier sentences on the basis of their common perceived location, and ii) this location is determined on the basis of a combination of ITDs of those individual frequency components that are assigned to each auditory stream. These findings might be interpreted as evidence that perceived location is more important than f0 cues for the grouping of auditory streams. However, Darwin and Hukin (120) later reported that, when the carrier sentences had natural prosody (instead of sentences with a flat f0 contour) and were spoken by different talkers (instead of the same talker), the advantage of grouping by perceived location disappeared. In conclusion, these studies demonstrate that continuity in pitch and vocal tract size are the most important factors for assigning words to a stream of speech, although perceived location may also play a role.

In typical sine-wave speech studies such as those reviewed earlier, listeners are asked to identify words from a single “sine-wave speaker”. This way of presenting the stimuli does not allow the testing of the hypothesis that speech schema can independently separate speech from other sounds independently of auditory grouping processes. To properly test this hypothesis, one should for example measure speech perception performance when two simultaneous sine-wave sentences are presented. If speech schema can group frequency components into phonetic percepts independently of other nonspeech grouping processes then the intelligibility of simultaneous sine-wave speech sentences should be as high as when only one such sentence is heard at a time. Barker and Cooke directly tested this prediction (121). They first replicated Carrell and Opie’s (107) findings by employing a larger set of sentences spoken at a normal speech rate, and containing both voiced and voiceless sounds. Barker and Cooke then performed a second experiment in which they compared the identification of two simultaneous sentences in sine-wave speech and in filtered speech (that is, normal speech that was low-pass filtered with a cut-off frequency of 3300 Hz). In order to maximize the intelligibility of simultaneous sine-wave sentences, Barker and Cooke selected those sentences that were highly intelligible in isolation, and recruited only the best performing listeners from the first experiment. Nevertheless, while filtered speech sentences were identified with a phoneme accuracy of about 80 to 95%, all but one listener scored below 20% correct for the most intelligible sentence in each pair (the best listener scored just below 50% on the most intelligible sentences). These results show that the speech schema is not able to group sinusoidal components that mimic the characteristics of speech if there is ambiguity about the grouping of sinusoidal components. By contrast, the high intelligibility of single sine-wave speech sentences could be due to the fact that as perceptual fusion might be the default option for grouping sinusoidal components that are asynchronous and are not harmonically related (121). This outcome further supports the view that phonetic perception is affected by the operation of auditory grouping processes, and that the speech schema interacts with general-purpose processes for the purpose of speech perception.

7. FINAL DISCUSSION

The evidence presented in this chapter supports the idea that general-purpose processes, attribute-specific schemas and categorical schemas interact for the organization of the auditory input into simultaneous auditory events and streams. A frequency component may or may not contribute to the formation of a perceptual attribute depending on the specific processing properties of each schema, all other general-purpose auditory grouping cues being equal. For example, pitch, timbre and speech schema require different amounts of onset asynchrony in order to exclude a frequency component from the computation of the respective perceptual features of an auditory event. Although general-purpose grouping processes operate in the same fashion across attribute-specific and categorical schemas, the strength of their effects on the perceptual outcomes of auditory organization also depends on the specific function of different schemas. For example, grouping by common onset is relatively more important for the perception of timbre, and for speech and nonspeech recognition, than for pitch perception. By contrast, sequential grouping by frequency similarity and spectral regularity play a role in the processing of all perceptual attributes and for sound recognition. Finally, spatial and common modulation cues have a weak effect on the auditory organization of complex sounds across all schemas. The failure to find consistent effects of spatial cues is particularly surprising, since it is reasonable to believe that being able to locate a sound source is helpful in carrying out auditory scene analysis (27). The poor use of spatial cues for auditory grouping by humans and other animals suggests that these cues may not be particularly advantageous from an evolutionary perspective, possibly because the presence of reverberation in natural listening environments renders these cues unreliable. Being able to detect a meaningful event using auditory cues may be more important than knowing its precise location. For example, an animal would normally orient itself towards the *general* direction of a danger or of a potential mate, and then use visual cues in conjunction with hearing for precise localization. It is possible that animals that typically rely on hearing, instead of vision, for precise localization (for example, nocturnal animals such as owls) might be more selective in the use of spatial cues for the grouping of frequency components.

A second conclusion from the review of existing findings is that the effect of general-purpose processes is not of an "all-or-none" type (5). That is, different amounts of a given cue (say, frequency separation or onset asynchrony) produce varying amounts of segregation or fusion. For example, the presence of small or intermediate amounts of onset asynchrony does not automatically result in the complete segregation of one component from other components, although a large asynchrony may exclude a component from a given grouping, perceptual attribute or familiar timbre. Consistent with this conclusion is the idea that general-purpose grouping processes produce groupings of frequency components, and that such groupings are of variable strength (4). According to this hypothesis, the ability of speech schema to select frequency components *across* groupings depends on the strength of the evidence for such groupings. For example, we have seen in sections 4 and 5 that the amount of asynchrony and the specific properties of each schema affect the extent to which an asynchronous component will be included in the formation of pitch and speech percepts.

Third, research on the perceptual grouping of complex sounds suggests that general-purpose and schema-based processes are likely to be active at the same time. For example, the recognition of speech sounds in a noisy environment requires the perceptual segregation of speech from noise, as well as the assignment of a perceived duration, location, pitch (for voiced sounds), and timbre. The simple serial model that was introduced in the introduction of section 3 to describe the nature of auditory grouping processes is obviously inadequate to explain the complexity of the interaction among the perceptual processes involved in auditory scene analysis. Therefore, it is more likely that the stages of the formation of auditory events and of combinations of events into sequential and simultaneous streams occur in parallel rather than in series, and that general-purpose grouping and schema processes interact to generate perceptual representations. At the auditory event formation stage, general-purpose grouping processes group frequency components into potential groupings from which schemas select relevant auditory representations of the acoustic properties of the stimuli (frequencies, amplitudes, etc.). At the stream formation stage, general-purpose grouping processes generate potential streams by grouping auditory events on the basis of the similarity of their perceptual attributes (pitch, timbre, etc). Hence, the units of auditory grouping might differ depending on the processing stage at which auditory grouping is taking place. This interactive model might provide a solution to the "chicken-and-egg" question of whether schema-based percepts (e.g. pitch or timbre) are effects or causes of auditory organization. For example, several studies have shown that the grouping of frequency components determines the perception of the pitch or vowel quality of auditory events (see sections 4 and 5). On the other hand, it has also been shown that pitch and timbre are used for organizing auditory events into streams (see section 6). According to this model, schema-based percepts are affected by general-purpose grouping of frequency components at the auditory event formation stage, but the perceptual outcomes of schema-based processes become the primary cues for the grouping of auditory events into (simultaneous) streams. The parallel functioning of general-purpose and schema-based processes is also consistent with the numerous instances of 'duplex perception' with both speech and nonspeech complex sounds. For example, the same mistuned harmonic can be heard out of a vowel (probably because of a violation of spectral regularity), but at the same time it can also contribute to that vowel's pitch and identity as a result of the operation of pitch and speech schemas. This violation of the exclusive allocation of information in the auditory modality is widespread, and it shows that different schema-based processes can simultaneously share the same piece of auditory evidence (4).

An interactive view of auditory scene analysis processes can provide an insight on the relationship between the general-purpose grouping process of harmonicity and the pitch schema. The distinction between general-purpose grouping by spectral regularity/harmonicity and schema-based grouping by the pitch schema can account for the well-known phenomenon that a mistuned harmonic can be heard out of a complex sound but at the same time still contribute to that sound's pitch (25, 26). The perceptual segregation of the mistuned component may be the result of the operation of the general-purpose grouping by spectral regularity/harmonicity, while the pitch schema is responsible for the integration of the mistuned component into the pitch of the complex sound. Grossberg, Govindarajan, Wyse and Cohen recently proposed a computational model of auditory scene analysis (the ARTSTREAM model), which involves the interaction between (bottom-up) spectral representations of frequency components and (top-down) pitch perception templates (122). In this model, frequency components activate multiple spectral representations in a "spectral stream" layer. Each spectral representation activates filters that are sensitive to harmonic relations among components; these filters in turn activate pitch categories. The pitch category that receives the largest activation then sends top-down feedback to the corresponding spectral stream layer, such that frequencies that are not related to the pitch category are suppressed. Suppressed frequencies are "free' to be grouped into alternative pitch percepts. While Grossberg and his colleagues did not make an explicit distinction between general-purpose and schema-based processes in relation to their model, the top-down operation of harmonic sieve templates at the pitch stream layer, which according to them is attention-driven, is equivalent to operation of the pitch schema that was discussed in this chapter. This model provides an insight on the possible interaction among auditory grouping processes for the generation of auditory streams.

Although the number of studies on auditory scene analysis during the last twenty years or so has been growing steadily, there are still many unanswered questions in this research field. It is still not clear to what extent attention processes are involved in the operation of general-purpose or schema-based processes. For example, section 3.1 reviewed some of the evidence in the use (or lack of use) of attention in sequential grouping by frequency similarity, but there are no studies that I am aware of that investigated the role of attention in other general-purpose grouping processes. In fact, it is also debatable whether schema-based processes such as pitch or speech schemas rely on attention processes to generate the corresponding percepts. A second set of unanswered questions concerns the relationship between general-purpose and schema-based processes for the perception of loudness, perceived location and perceived duration as there are very few studies on the auditory organization of frequency components into these perceptual attributes. Finally, the issue about the units of grouping has not been directly addressed. For example, we do not know whether streams or simultaneous auditory events are grouped on the basis of the properties of the perceptual attributes of each event, or whether grouping is based on auditory cues in the same way that it happens for the formation of individual auditory events.

8. ACKNOWLEDGMENT

I would like to thank Brian Gygi, Gary Kidd and an anonymous reviewer for providing helpful and incisive comments on previous drafts of this manuscript.

9. REFERENCES

1. Moore, B. C. J.: An Introduction to the Psychology of Hearing. Academic Press, London, UK (1997)

2. Bregman, A. S.: Auditory Scene Analysis. *Proceedings of the Seventh International Conference on Pattern Recognition*, IEEE Computer Press (1984).

3. Darwin, C. J.: Perceptual Grouping Of Speech Components Differing In Fundamental Frequency And Onset-Time. *Q J Exp Psychol Sect A: Hum Exp Psychol*, 33, 185-207 (1981)

4. Bregman, A. S.: Auditory Scene Analysis: The Perceptual Organization of Sound. Bradford Books, MIT Press, Cambridge, Mass. (1990)

5. Darwin, C. J. & R. P. Carlyon: Auditory grouping. In: Handbook of perception and cognition, Volume 6, Hearing. Ed: B. C. J. Moore. Academic Press, London, UK 387-424 (1995)

6. McAdams, S. & C. Drake: Auditory Perception and Cognition. In: Stevens' Handbook of Experimental Psychology. Eds: H. Pashler & S. Yantis. Wiley, New York, NY 397-452 (2002)

7. Bregman, A. S. & J. Campbell: Primary auditory stream segregation and perception of order in rapid sequences of tones. *J Exp Psychol*, 89, 244-249 (1971)

8. Wertheimer, M.: Untersuchungen zur Lehre der Gestalten. Psychologische Forschung, 4, 310-350 (1923)

9. van Noorden, L. P. A. S.: Temporal coherence in the perception of tone sequences. Technische Hoogschool, Eindhoven (The Netherlands) (1975).

10. Dowling, W. J.: The perception of interleaved melodies. *Cognitive Psych*, 5, 322-337 (1973)

11. Vliegen, J., B. C. J. Moore & A. J. Oxenham: The role of spectral and periodicity cues in auditory steam segregation, measured using a temporal discrimination task. *J Acoust Soc Am*, 106, 938-945 (1999)

12. Sussman, E., W. Ritter & J. H. G. Vaughan: An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology*, 36, 22-34 (1999)

13. Carlyon, R. P.: How the brain separates sounds. *Trends Cog Sci*, 8, 465-471 (2004)

14. Winkler, I., R. Takegata & E. Sussman: Event-related brain potentials reveal multiple stages in the perceptual organization of sounds. *Cog Brain Res*, 25, 291-299 (2005)

15. Ciocca, V. & C. J. Darwin: The integration of nonsimultaneous frequency components into a single virtual pitch. *J Acoust Soc Am*, 105, 2421-2430 (1999)

16. Grose, J. H., J. W. Hall, 3rd & E. Buss: Modulation gap detection: effects of modulation rate, carrier separation, and mode of presentation. *J Acoust Soc Am*, 106, 946-53 (1999)

17. Miller, G. A. & G. A. Heise: The trill threshold. *J Acoust Soc Am*, 22, 637-638 (1950)

18. Bozzi, P. & G. Vicario: Due fattori di unificazione fra note musicali: la vicinanza temporale e la vicinanza tonale. *Riv Psicol*, 54, 235-258 (1960)

19. Vicario, G. B.: Vicinanza spaziale e vicinanza temporale nella segregazione degli eventi. *Riv Psicol*, 59, 843-863 (1965)

20. Dannenbring, G. L. & A. S. Bregman: Streaming vs. fusion of sinusoidal components of complex tones. *Percept Psychophys*, 24, 369-376 (1978)

21. Darwin, C. J.: Perceiving vowels in the presence of another sound: constraints on formant perception. *J Acoust Soc Am*, 76, 1636-47 (1984)

22. Roberts, B. & P. J. Bailey: Spectral regularity as a Factor Distinct From Harmonic Relations in Auditory Grouping. *J Exp Psychol Human*, 22, 604-614 (1996)

23. Roberts, B. & A. S. Bregman: Effects of pattern of spectral spacing on the perceptual fusion of harmonics. *J Acoust Soc Am*, 90, 3050-3060 (1991)

24. Roberts, B. & P. J. Bailey: Regularity of spectral pattern and its effects on the perceptual fusion of harmonics. *Percept Psychophys*, 58, 289-99 (1996)

25. Moore, B. C., B. R. Glasberg & R. W. Peters: Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *J Acoust Soc Am*, 80, 479-83 (1986)

26. Moore, B. C. J., B. R. Glasberg & R. W. Peters: Relative dominance of individual partials in determining the pitch of complex tones. *J Acoust Soc Am*, 77, 1853-1860 (1985)

27. Cherry, C.: Some experiments on the recognition of speech with one and with two ears. *J Acoust Soc Am*, 26, 975-979 (1953)

28. O'Connor, K. N. & M. L. Sutter: Global spectral and Location Effects in Auditory Perceptual Grouping. *J Cog Neurosci*, 12, 342-354 (2000)

29. Buell, T. N. & E. R. Hafter: Combination of binaural information across frequency bands. *J Acoust Soc Am*, 90, 1894-900 (1991)

30. Hill, N. I. & C. J. Darwin: Effects of onset asynchrony and of mistuning on the lateralization of a pure tone embedded in a harmonic complex. *J Acoust Soc Am*, 93, 2307-2308 (1993)

31. Hall, J. W., III., M. P. Haggard & M. A. Fernandes: Detection in noise by spectro-temporal analysis. *J Acoust Soc Am*, 76, 50-56 (1984)

32. Grose, J. H. & J. W. Hall, 3rd: Comodulation masking release for speech stimuli. *J Acoust Soc Am*, 91, 1042-50 (1992).

33. A. N. S. I.: H960 USA Standard Acoustical Terminology (Including Mechanical Shock and Vibration) Sl.1. American National Standards Institute. (1960 (R1976)).

34. Risset, J.-C. & D. L. Wessel: Exploration of timbre by analysis and synthesis. In: The Psychology of Music. Ed: D. Deutsch. Academic Press, San Diego 113-168 (1999)

35. Grey, J. M.: Multidimensional perceptual scaling of musical timbres. *J Acoust Soc Am*, 61, 1270-1277. (1977)

36. Zwicker, E., G. Flottorp & S. S. Stevens: Critical bandwidth in loudness summation. *J Acoust Soc Am*, 29, 548-557. (1957)

37. Plack, C. J. & R. P. Carlyon: Loudness perception and intensity coding. In: Handbook of perception and cognition, Volume 6, Hearing. Ed: B. C. J. Moore. Academic Press, San Diego, CA 123-160 (1995)

38. Blauert, J.: Spatial hearing, 2nd ed. MIT Press, Cambridge, Mass. (1997)

39. Nakajima, Y., T. Sasaki, K. Kanafuka, A. Miyamoto, G. Remijn & G. t. Hoopen: Illusory recoupling of onsets and terminations of glide tone components. *Percept Psychophys*, 62, 1413-1425 (2000)

40. Plomp, R.: Pitch of complex tones. *J Acoust Soc Am*, 41, 1526-33 (1967)

41. Ritsma, R. J.: Frequencies dominant in the perception of the pitch of complex sounds. *J Acoust Soc Am*, 42, 191-8 (1967)

42. Houtsma, A. J. M.: Pitch perception. In: Handbook of perception and cognition. Ed: B. C. J. Moore. Academic Press, New York, NY (1995)

43. Hartmann, W. M.: Pitch, periodicity, and auditory organization. *J Acoust Soc Am*, 100, 3491-3502 (1996)

44. Terhardt, E., G. Stoll & M. Seewann: Algorithm for extraction of pitch and pitch salience from complex tonal signals. *J Acoust Soc Am*, 71, 679-688 (1982)

45. Terhardt, E., G. Stoll & M. Seewann: Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions. *J Acoust Soc Am*, 71, 671-678 (1982)

46. Goldstein, J. L.: An optimum processor theory for the central formation of the pitch of complex tones. *J Acoust Soc Am*, 54, 1496-1516. (1973)

47. Duifhuis, H., L. F. Willems & R. J. Sluyter: Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. *J Acoust Soc Am*, 71, 1568-1580 (1982)

48. Neuhoff, J. G.: Pitch variation is unnecessary (and sometimes insufficient) for the formation of auditory objects. *Cognit*, 87, 219-224 (2003)

49. Nygaard, L. C. & D. B. Pisoni: Speech perception: New directions in research and theory. In: Speech, language, and communication. Eds: J. L. Miller & P. D. Eimas. Academic Press, San Diego, CA (1995)

50. Whalen, D. H. & A. M. Liberman: Speech perception takes precedence over nonspeech perception. *Science*, 237, 169-171 (1987)

51. Remez, R. E., P. E. Rubin, S. E. Berns, J. S. Pardo & J. M. Lang: On the perceptual organization of speech. *Psychol Rev*, 101, 129-156 (1994)

52. Mattingly, I. & A. M. Liberman: Speech and other auditory modules. In: Signal and sense: Local and global order in perceptual maps. Eds: G. M. Edelman & W. E. Gall. New York, NY W.M. Cowan, Wiley (1990)

53. Rand, T. C.: Dichotic release from masking for speech. *J Acoust Soc Am*, 55, 678-680 (1974)

54. Cutting, J. E.: Auditory and lingustic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychol Rev*, 83, 114-140 (1976)

55. Remez, R. E., Rubin, P. E., Pisoni, D. B. & T. D. Carrell: Speech perception without traditional speech cues. *Science*, 212, 947-950 (1981)

56. Bey, C. & S. McAdams: Schema-based processing in auditory scene analysis. *Percept Psychophys*, 64, 844-854 (2002)

57. Ciocca, V.: Evidence against an effect of grouping by spectral regularity on the perception of virtual pitch. *J Acoust Soc Am*, 106, 2746-51 (1999)

58. Darwin, C. J., R. W. Hukin & B. Y. Al-Khatib: Grouping in pitch perception: evidence for sequential constraints. *J Acoust Soc Am*, 98, 880-885 (1995)

59. Darwin, C. J. & V. Ciocca: Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *J Acoust Soc Am*, 91, 3381-3390 (1992)

60. Ciocca, V. & C. J. Darwin: Effects of onset asynchrony on pitch perception: Adapation or grouping? *J Acoust Soc Am*, 93, 2870-2878 (1993)

61. Hukin, R. W. & C. J. Darwin: Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. *Percept Psychophys*, 57, 191-6 (1995)

62. Darwin, C. J., V. Ciocca & G. J. Sandell: Effects of frequency and amplitude modulation on the pitch of a complex tone with a mistuned harmonic. *J Acoust Soc Am*, 95, 2631-2636 (1994)

63. Warren, R. M., C. J. Obusek & J. M. Ackroff: Auditory induction: perceptual synthesis of absent sounds. *Science*, 176, 1149-51 (1972)

64. McAdams, S., M. C. Botte & C. Drake: Auditory continuity and loudness computation. *J Acoust Soc Am*, 103, 1580-1591 (1998)

65. Bregman, A. S. & S. Pinker: Auditory streaming and the building of timbre. *Can J Psychology*, 32, 19-31 (1978)

66. Steiger, H. & A. S. Bregman: Capturing frequency components of glided tones: Frequency separation, orientation, and alignment. *Percept Psychophys*, 30, 425-435 (1981)

67. Farris, H. F., A. S. Rand & M. J. Ryan: The Effects of Spatially separated call Components on Phonotaxis in Tungara Frogs: Evidence for Auditory Grouping. *Brain Behav Evol*, 60, 181-188 (2002)

68. Bregman, A. S., J. Abramson, P. Doehring & C. J. Darwin: Spectural integration based on common amplitude modulation. *Percept Psychophys*, 37, 483-493 (1985)

69. Bregman, A. S., R. Levitan & C. Liao: Fusion of auditory components: Effects of the frequency of amplitude modulation. *Percept Psychophys*, 47, 68-73 (1990)

70. Liberman, A. M. & I. G. Mattingly: The motor theory of speech perception revised. *Cognition*, 21, (1985)

71. Liberman, A. M.: On finding that speech is special. *Am Psychol*, 37, 148-167 (1982)

72. Liberman, A. M., D. Isenberg & B. Rakerd: Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Percept Psychophys*, 30, 133-143 (1981)

73. Ciocca, V. & A. S. Bregman: The effects of auditory streaming on duplex perception. *Percept Psychophys*, 46, 39-48 (1989).

74. Darwin, C. J. & R. W. Hukin: Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *J Acoust Soc Am*, 102, 2316-2324 (1997)

75. Repp, B. H. & S. Bentin: Parameters of spectral/temporal fusion in speech perception. *Percept Psychophys*, 36, 523-530 (1984)

76. Smith, R. L.: Adaptation, saturation, and physiological masking in single auditory-nerve fibers. *J Acoust Soc Am*, 65, 166-178 (1979)

77. Darwin, C. J. & N. S. Sutherland: Grouping frequency components of vowels: When is a harmonic not a harmonic? *Q J Exp Psychol*, 36, 193-208 (1984)

79. Roberts, B. & S. D. Holmes: Asynchrony and the grouping of vowel components: Captor effect revisited. *J Acoust Soc Am*, 119, 2905-2918 (2006)

79. Holmes, S. D. & Roberts, B.: Inhibitory influences on asynchrony as a cue for auditory segregation. *J Exp Psychol Human*, 32, 1231-1242 (2006)

80. Roberts, B. & B. C. Moore: The influence of extraneous sounds on the perceptual estimation of first-formant frequency in vowels under conditions of asynchrony. *J Acoust Soc Am*, 89, 2922-2932 (1991)

81. Ciocca, V., A. S. Bregman & K. L. Capreol: The Phonetic Integration of speech and Non-speech Sounds: Effects of Perceived Location. *Q J Exp Psychol Sect A: Hum Exp Psychol*, 44, 577-593 (1992)

82. Gardner, R. B. & C. J. Darwin: Grouping of vowel harmonics by frequency modulation: Absence of effects on phonemic categorization. *Percept Psychophys*, 40, 183-187 (1986)

83. Gardner, R. B., S. A. Gaskill & C. J. Darwin: Perceptual grouping of formants with static and dynamic differences in fundamental frequency. *J Acoust Soc Am*, 85, 1329-1337 (1989)

84. Carlyon, R. P.: The psychophysics of concurrent sound segregation. *Philos T Roy Soc B*, 336, 327-255 (1992)

85. Darwin, C. J. & R. B. Gardner: Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. *J Acoust Soc Am*, 79, 838-845 (1986)

86. Risset, J.-C. & M. V. Matthews: Analysis of musical instrument tones. *Physics Today*, 22, 23-30 (1969)

87. Fowler, C. A. & L. D. Rosenblum: Duplex Perception: A Comparison of Monosyllables and Slamming Doors. *J Exp Psychol Human*742-754 (1990)

88. Darwin C. J. & R. W. Hukin: Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *J Acoust Soc Am*, 102, 2316-2324 (1997)

89. Darwin, C. J. & R. W. Hukin: Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony. *J Acoust Soc Am*, 103, 1080-1084 (1998)

90. Dannenbring, G. L. & A. S. Bregman: Stream segregation and the illusion of overlap. *J Exp Psychol Human*, 2, 544-555 (1976)

91. Cusack, R. & B. Roberts: Effects of differences in timbre on sequential grouping. *Percept Psychophys*, 62, 1112-1120 (2000)

92. Bey, C. & S. McAdams: Postrecognition of Interleaved Melodies as an Indirect Measure of Auditory Stream Formation. *J Exp Psychol Human*, 29, 267-279 (2003)

93. Hartmann, W. M. & D. Johnson: Stream segregation and Peripheral Channeling. *Music Percept*, 9, 155-184 (1991)

94. Iverson, P.: Auditory Stream Segregation by Musical Timbre: Effects of Static and Dynamic Acoustic Attributes. *J Exp Psychol Human*, 21, 751-763 (1995)

95. Singh, P. G. & A. S. Bregman: The influence of different timbre attributes on the perceptual segregation of complex-tone sequences. *J Acoust Soc Am*, 102, 1943-1952 (1997)

96. Vliegen, J. & A. J. Oxenham: Sequential stream segregation in the absence of spectral cues. *J Acoust Soc Am*, 105, 339-346 (1999)

97. Rogers, W. L. & A. Bregman: An experimental evaluation of three theories of auditory stream segregation. *Percept Psychophys*, 53, 179-189 (1993)

98. Roberts, B., B. R. Glasberg & B. C. Moore: General-purpose stream segregation of tone sequences without differences in fundamental frequency or passband. *J Acoust Soc Am*, 112, 2074-85 (2002).

99. Grimault, N., C. Micheyl, R. P. Carlyon, P. Arthaud & L. Collet: Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency. *J Acoust Soc Am*, 108, 263-71 (2000)

100. Wessel, D. L.: Timbre space as a musical control structure. *Comput Music J*, 3, 45-52 (1979)

101. Singh, P. G.: Perceptual organization of complex-tone sequences: A tradeoff between pitch and timbre. *J Acoust Soc Am*, 82, 886-899 (1987)

102. Gregory, A. H.: Timbre and Auditory streaming. *Music Percept*, 12, 161-174 (1994)

103. Lackner, J. R. & L. M. Goldstein: Primary auditory stream segregation of repeated consonant-vowel sequences. *J Acoust Soc Am*, 56, 1651-1652 (1974)

104. Noteboom, S. G., J. P. L. Brokx, & J. J. De Roji: Contributions of prosody to speech perception. In: Studies in the perception of language. Eds: Levelt W. J. M. & G. B. Flores d'Arcais. Wiley, Chichester (1976)

105 Darwin, C. J. & C. E. Bethell-Fox: Pitch continuity and speech source attribution. *J Exp Psychol Human*, 3, 665-672 (1977)

106. Broadbent, D. E. & P. Ladefoged: On the fusion of sounds reaching different sense organs. *J Acoust Soc Am*, 29, 708-710 (1957)

107. Carrell, T. D. & J. M. Opie: The effect of amplitude comodulation on auditory object formation in sentence perception. *Percept Psychophys*, 52, 437-445 (1992)

108. Deutsch, D.: Musical illusions. *Sci Am*, 233, 92-8,103-4 (1975)

109. Smith, J., S. Hausfeld, R. P. Power & A. Gorta: Ambiguous musical figures and auditory streaming. *Percept Psychophys*, 32, 454-64 (1982)

110. Brockx, J. P. L. & S. G. Nooteboom: Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10, 23-36 (1982)

111. Scheffers, M. T. M.: Sifting vowels: auditory pitch analysis and sound segregation. Rijksuniversiteit Groningen, The Netherlands(1983).

112. Culling, J. E. & J. C. Darwin: Perceptual separation of simultaneous vowels: Within and across-formant grouping by f0. *J Acoust Soc Am*, 93, 3454-3467 (1993)

113. Chalikia, M. H. & A. S. Bregman: The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel segregation. *Percept Psychophys*, 46, 487-496 (1989)

114. Assmann, P. F. & A. Q. Summerfield: Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *J Acoust Soc Am*, 88, 680-697 (1990)

115. Culling, J. E. & J. C. Darwin: Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating. *J Acoust Soc Am*, 95, 1559-1569 (1994)

116. Assmann, P. F. & A. Q. Summerfield: The contributions of waveform interactions to the perception of concurrent vowels. *J Acoust Soc Am*, 95, 471-484 (1994)

117. McAdams, S.: Segregation of concurrent sounds. I: Effects of frequency modulation coherence. *J Acoust Soc Am*, 86, 2148-2159 (1989)

118. Summerfield, A. Q. & J. F. Culling: Auditory segregation of competing voces: Absence of FM or AM coherence. *Philosophical Transactions of the Royal Society of London*, 336, 357-366 (1992)

119. Darwin, C. J. & R. W. Hukin: Auditory Objects of Attention: The Role of Interaural Time Differences. *J Exp Psychol Human,* 25, 617-629 (1999)

120. Darwin, C. J. & R. W. Hukin: Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J Acoust Soc Am*, 107, 970-977 (2000)

121. Barker, J. & M. P. Cooke: Is the sine-wave speech cocktail party worth attending? *Speech Commun*, 27, 159-174 (1999)

122. Grossberg, S., K. K. Govindarajan, L. L. Wyse & M. A. Cohen: ARTSTREAM: a neural network model of auditory scene analysis and source segregation. *Neural Networks*, 17, 511-536 (2004)

Key Words: Auditory Scene Analysis, Auditory Grouping, Complex Sounds, Schema, Pitch, Timbre, Loudness, Localization, Speech Perception, Perceived Duration, Review

Send correspondence to: Valter Ciocca, PhD, Professor and Director, School of Audiology and Speech Sciences, Faculty of Medicine, The University of British Columbia, 5804 Fairview Ave., Vancouver, B.C. V6T 1Z3, Canada, Tel: 604-822-2266, Fax: 604-822-6569, E-mail: vciocca@audiospeech.ubc.ca

http://www.bioscience.org/current/vol13.htm