

Auditory gist perception: an alternative to attentional selection of auditory streams?

Sue Harding¹, Martin Cooke¹, and Peter König²

¹ Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK

{s.harding,m.cooke}@dcs.shef.ac.uk

<http://www.dcs.shef.ac.uk/spandh>

² Neurobiopsychologie Labor, Institut für Kognitionswissenschaft, Universität Osnabrück, Albrechtstraße 28, 49069 Osnabrück, Germany

pkoenig@uos.de

<http://www.cogsci.uni-osnabrueck.de/~NBP>

Abstract. The idea that the gist of a visual scene is perceived before attention is focused on the details of a particular object is becoming increasingly popular. In the auditory system, on the other hand, it is typically assumed that the sensory signal is first broken down into streams and then attention is applied to select one of the streams. We consider evidence for an alternative: that, in close analogy with the visual system, the gist of an auditory scene is perceived and only afterwards attention is paid to relevant constituents. We find that much experimental evidence is consistent with such a proposal, and we suggest some possibilities for gist representations.

Key words: Attention, gist perception, auditory scene analysis

1 Introduction

Conventional theories of attention assume a hierarchy of processing, starting with low-level analysis of simple features which are then integrated to form more complex features and then mapped onto objects and categories. This *bottom-up* processing is assumed to be pre-attentive but also interacts with *top-down* processes which are mediated by attention. Details of all low-level features would thus be available for higher-level processing but, in order to reduce the processing load, it is assumed that at some stage an attentional selection process occurs. However, a long-standing debate exists over the stage at which attentional selection takes place (see [1] for a review).

Such bottom-up hierarchical theories of attention are common in vision (e.g. [2, 3]) and also in audition in models that use the principles of *auditory scene analysis* [4]. In auditory scene analysis, it is generally assumed that an auditory signal is broken down into streams, using differences in characteristics such as frequency, intensity and spatial location to segregate elements of the signal, while subsequent grouping of elements into streams occurs using principles such

as similarity, good continuity and common fate. Attention then acts to select one of the streams.

However, there are problems with such a hierarchical approach. In vision, a number of experiments have shown that an observer can rapidly become aware of the surroundings without being aware of the details (see [5] for a review). For example, observers are able to grasp the overall content of rapidly presented images [6] but they do not notice changes in detail, exemplified by the phenomenon of change blindness [7]. A further issue concerns the rapid ‘popout’ of an odd item from a set of items in the visual search paradigm: counter to the proposal that conjunctions of features should cause slower serial search [3], it has been shown that more complex items may pop out if they represent categories such as three-dimensional objects [5].

Similarly, in audition, listeners are able to perform tasks which are inconsistent with a bottom-up hierarchical view of auditory scene analysis and the subsequent attentional selection of an auditory stream. For example, listeners may perceive part of the signal in two ways, as in the phenomenon of duplex perception [8] in which a sine-wave chirp may be heard separately but may also be incorporated into a speech syllable. Speech perception occurs much more rapidly than would be expected if the signal were processed in a hierarchical fashion [9] (see also section 4.4). Other evidence suggests that top-down processing plays a major role [10]: for example, the same acoustic signal may be heard as two different words depending on the context.

In order to account for such results, an alternative has been proposed: that an observer rapidly processes the ‘gist’ (i.e. an imprecise overview) of a scene and then focuses attention on the detail of a limited region (e.g. [5, 11–14]). So, for example, an observer may be aware that there are people and items of furniture in a room, without noticing how many people or what type of furniture until attention is focused on a particular part of the room. Similarly, a listener may be aware that music is playing and people nearby are talking, without noticing what instruments are producing the music or what the people are saying. There are obvious advantages in becoming rapidly aware of the gist of a scene, as the observer or listener can obtain an overview of possible dangers and benefits in the environment and need only expend effort in scrutinising those objects or areas of particular interest.

Explorations of gist processing have so far mainly been concerned with the visual modality. In the rest of this paper, we review selected work exploring the idea that the gist of a visual scene is perceived before the detail and we consider evidence that, in audition too, we hear the gist of an auditory scene before we focus on the detail of a particular sound source.

2 The gist of gist processing

The idea of gist processing has been applied not only to explain visual perception [5, 6, 11, 12, 15–19] but also memory [20–23] and general theories of consciousness [24]. One approach [16] defines two types of gist:

- *perceptual* gist, which refers to the representations built during perception, and
- *conceptual* gist, which includes semantic information stored in memory.

While it is clear that perceptual and conceptual gist are intimately linked, there is a significant body of work concentrating on the idea of conceptual gist or gist memory and, although this may shed light on possible representations, here we are mostly concerned with perceptual gist which concerns the rapid determination of the content of perceptual input (e.g. a visual or auditory signal).

A specific example of a theory concerned with perceptual gist in vision is Reverse Hierarchy Theory [5]; according to this theory, rapid bottom-up processing of the whole scene occurs (without attention) resulting in high-level awareness without detail (*vision-at-a-glance*); this is followed by top-down processing (with attention) to analyse the detail of that part of the scene within the focus of attention (*vision-with-scrutiny*), while the unattended parts of the signal do not require detailed processing. Thus, there are a number of facets of this theory:

- (a) only the gist of the scene or object is initially processed;
- (b) processing of the gist is rapid;
- (c) the focus of attention is deployed according to prior knowledge and the perception of the gist;
- (d) detailed analysis is performed on the part of the scene within the focus of attention;
- (e) unattended parts of the scene are undifferentiated.

In the following sections, we suggest some possible evidence for each of these for vision and, more extensively, for audition.

3 Gist perception in vision

The ‘gist of a scene’ describes a general overview of the global properties of the scene, as opposed to local properties of a specific object or area. Some early experiments showed that words [25] and large letters constructed from small letters [11] (figure 1) are perceived before their constituent letters. Navon [11] proposed the principle of *global precedence*, suggesting that a scene is processed from the top of the hierarchy to the bottom, from global to local, i.e. from the whole scene down to its constituent parts (discussed in [26]). More recently, it was shown that observers can categorise images that are degraded such that individual objects are not well defined [27], and can distinguish animals from man-made artifacts even when parts of the images have been jumbled [28], indicating that the gist of the scene can be discriminated in the presence of local distortions.

It should be noted that perceiving an image globally does not necessarily imply that the image or its parts have lower resolution; the global object may consist of either coarse or fine elements (figure 1; [27]). Similarly, local information need not necessarily consist of high-resolution elements, but simply describes a limited part of the scene. The issue of high and low resolution in gist representations is discussed further in section 5.1.

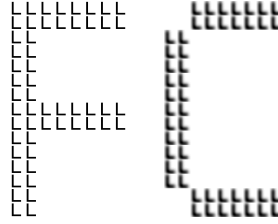


Fig. 1. Illustration of the difference between global/local and coarse/fine (after [27]). Left, global F, fine, local L; right, global C, coarse, local L.

There is increasing evidence that the gist of a scene is perceived quickly and pre-attentively [5, 15–17]. The rapid serial visual presentation (RSVP) paradigm, in which a series of images are presented in quick succession, has shown that basic categories such as animals or vehicles can be detected very rapidly. Recent studies found category detection was extremely rapid and needed only short presentation times [29]. The rapid ‘popout’ of complex items which represent categories provides further support for rapid processing of the gist, while the insensitivity of observers to the number of elements present when items popout suggests that the unattended regions of the image are not processed in detail [5].

The phenomenon of change blindness (reviewed in [7]), in which large changes in presented images go unobserved if there is a small disruption between the changes, may also be explained by the proposal that observers do not perceive all the details of the scene but only those within the focus of attention [5].

4 Gist perception in audition

4.1 A brief comparison of visual and auditory attention

The idea of gist perception has been discussed mainly in the context of the visual system but may also apply to audition. In audition, ‘objects’ (corresponding to auditory sources) are generally not static and instead there is the concept of the *auditory stream* which is the perceptual representation of an acoustic source determined during *auditory scene analysis* [4]. Such a stream, or the source from which it is assumed to emanate, may be the focus of attention.

Visual objects may overlap in space and may occlude one another, but are generally easily distinguishable from one another due to differences in spatial location, colour and luminance; although some objects may move, most remain static. In audition, the harmonics of complex tones and the inharmonic sounds making up auditory streams usually overlap in frequency, and very few sounds remain static over time. In addition, the spatial location of sound sources is determined by comparing the signals entering each ear, using interaural time and level differences between the binaural signals. The relationship between the interaural cues and the perceived location of the sound source is not simple (see

e.g. [30]), and interaural cues may be ignored: listeners are able to fuse distinct sounds entering each ear into a perceptual whole [31]. Therefore, attending to a particular sound source or stream requires complex analysis of the signal in order to separate one stream from another.

4.2 Current proposals for auditory gist perception

Little research into gist perception in audition has taken place as yet, but a few authors have proposed that such a scheme may explain certain results involving attention and auditory perception. Nelken and Ahissar [13] suggest that Reverse Hierarchy Theory can be applied to auditory processing in order to resolve the conflict between global and local aspects of auditory processing. For example, during speech perception, a wide variety of acoustic signals can cause the perception of a particular phoneme or syllable; these variations occur not only across different talkers but even within the productions of a single talker, due to coarticulation between phonemes and other factors such as the accent and physical and emotional state of the talker [32, 33]. Similar problems exist in pitch determination: for example, the presence and intensity of individual harmonics may vary for the same pitch percept. Nelken and Ahissar suggest that crude high-level representations can be produced at an early stage of processing; for example, for pitch, a limited form of periodicity using narrow-band peripheral filters could be extracted rapidly; similarly a crude representation of a speech signal which preserves the most important global acoustic features of speech could be produced.

Cusack et al. [14] propose a hierarchical decomposition model as a parsimonious explanation for experimental results which show a build-up of streaming but only in the attended spatial location, frequency region or stream (where a stream in this case is a sequence of tones). Listeners were asked whether one or two streams were present when played a sequence of alternating tones; typically, when the listener is attending to the tone sequence, a single stream is heard first and after a few seconds two streams are heard. The listeners were also given a concurrent distracting task; several tasks were used, consisting of sequences of tones or noises at various frequencies, played in the same or a different ear. When the distracting task was included, the build-up of streaming did not occur until attention was applied to the tone sequence [14, 34]. Cusack et al. interpret this as evidence that attention is required for stream segregation. Also, since the build-up of streaming followed a similar pattern whether the sounds were in the same or different ear or frequency region, they suggest that there is some automatic segregation but then only the stream which is the current focus of attention is segregated further: unattended parts of the signal do not need to be differentiated, although a general idea of the whole signal is obtained. So, for example, if speech, music and traffic noise are all present in the auditory signal, it is not necessary for the auditory system to segregate individual singers or different car engines. Cusack et al. also suggest that some early attentional selection may take place, for example of peripheral frequency channels, which

would reduce demand on later processing stages involved in perceptual grouping and selective attention.

In order to gauge the feasibility of applying proposals for gist perception to audition, we consider the facets of the Reverse Hierarchy Theory listed above in section 2 in turn and discuss some possible evidence for each.

4.3 Perceiving the gist of an auditory scene

Intuitively it seems plausible that listeners are aware of the general content of an auditory signal before its detail: for example, music from an orchestra or pop/rock band is initially heard as a whole rather than hearing the individual instruments; a room full of talkers is heard as babble rather than individual conversations; and speech is heard as a whole rather than as individual formants and noise bursts. An unknown foreign language too is heard as a continuous babble, not as distinct acoustic features.

Speech can be understood in very difficult listening conditions, which suggests that only a limited part of the signal is necessary for accurate speech perception. Listeners are adept at understanding speech in noisy environments (unlike automatic speech recognisers) [35], and telephone transmission takes advantage of the fact that speech is still highly intelligible when presented with limited frequency bands. Even under extreme manipulations, speech perception is still robust: for example, Shannon et al. [36] modulated white noise using the temporal envelopes of speech extracted from a small number of broad frequency bands, which preserved amplitude and temporal cues while removing harmonics and formant structure within bands. Listeners were still able to identify consonants and vowels within a syllable context, even with as few as four frequency bands. Saberi and Perrott [37] time-reversed each brief segment of a sentence and found that listeners reported perfect intelligibility of the sentence when each segment was up to 50 ms in duration. Reverberation would also be expected to cause difficulties for speech perception, if the fine details of the signal were important, as the signal is smeared in the temporal domain, but speech remains highly intelligible under these conditions.

A further indication that only a general idea of the expected signal is required for speech perception is provided by the phenomenon of phonemic restoration [38]. Here, listeners perceive a complete speech sound when part of it is replaced with a non-speech sound: greater intelligibility occurs when the added non-speech sound is spectrally similar to the missing portion of speech (e.g. white noise replacing part of a fricative or a pure tone replacing part of a vowel).

Listeners can also interpret the auditory signal as speech when the formants are replaced by time-varying sinusoids that mimic the formant patterns of natural speech but do not contain harmonic structure [39]. In this case, the fine detail of the signal is rather unhelpful but the sinusoid patterns are sufficient to provide the gist of the speech for most listeners.

It should be noted that, although much of the success of human listeners in understanding speech in difficult conditions is due to contextual and syntactic knowledge of the speech, several of the experiments described above required

listeners to identify nonsense syllables and to distinguish individual consonants and vowels. Memory does however play an important part in speech perception, but a detailed discussion of gist memory is outside the scope of this paper.

There is evidence to suggest that gist perception may apply to other aspects of an auditory signal besides speech intelligibility. For example, experiments have shown that timbre is identified before pitch [40, 41], suggesting that the overall spectral shape is perceived first, followed by the detailed harmonics responsible for the pitch.

More generally, experiments have indicated that fusion of the auditory signal is the default condition [4, 42] and that, unless attended to, the signal is processed holistically rather than in detail; for example, a sequence of alternating tones is segregated into two streams only after attention has been applied [14].

4.4 Rapid processing of the gist

The rate at which listeners can process speech has been given variously as 25 to 30 phonetic segments per second or 400 words per minute (the latter with some difficulty) [9], which is faster than would be expected if all the details of the signal were being processed, since the order of unrelated non-speech sounds (such as a hiss, tone or buzz) cannot be judged accurately when presented at only four sounds per second, and naive listeners require each sound to have a duration of at least 700 milliseconds (ms) in order to identify the order [43]. However, trained listeners can use qualitative differences to determine the order of very brief (around 4 ms duration) tones [44]. Other psychoacoustic studies indicate that very brief stimuli can be categorised rapidly using their timbre: different vowels or different musical instruments can be identified based on stimuli with duration of less than 10 ms [40, 41]. These results do not seem consistent with the idea of hierarchical processing of the detail of every sound. Instead, they suggest that in situations where the duration of the individual items is short, only the overall pattern is perceived; in speech, semantic memory enables the listener to make use of this pattern, while non-speech sounds do not have sufficient semantic content for their order to be noted without training.

On the other hand, evidence from studies of event-related brain potentials does not provide a consistent picture. Measurements of mismatch negativity (MMN), which indicates the detection of an irregularity in an auditory event and is believed to reflect pre-attentive processes, suggest vowel phonemes are extracted pre-attentively: deviant vowels elicited MMN while deviant complex tones did not [45]; but other evidence suggests that the timbre of harmonic complexes is perceived pre-attentively within 150 ms of stimulus onset [46]. It has also been found that different responses to man-made and natural sounds occur in the region of the auditory cortex thought to be associated with categorisation of sounds, within 70 ms of the stimulus onset [47].

Such results suggest a complex interaction of cues such as pitch and timbre within the first few hundred milliseconds after the auditory signal enters the ear.

4.5 Deployment of the focus of attention based on the gist

A study of event-related brain potentials in which pairs of concurrent vowels were presented at different fundamental frequencies indicates that low-level cues such as fundamental frequency differences are important at an early stage (after about 140 ms), perhaps for signalling to higher levels that more than one auditory object is present [48]. Further studies suggest that attention is required for sequential integration of sounds, but not for segregation of concurrent sounds [49], and that segregation occurs before integration [50]. Psychoacoustical experiments (reviewed in [51]) have shown that, on its own, interaural time difference cannot be used to segregate a single speaker from similar background sounds; however, it can be used to track a sound source over time. This suggests that listeners may use other cues, such as harmonicity or common onset, to segregate an auditory object and only then use attention to follow its spatial location. Taken together, these results are consistent with the proposal that pre-attentive concurrent segregation occurs before attentive sequential integration, and that listeners may use attention to track auditory objects that have been determined during the initial gist processing stage.

4.6 Detailed processing within the focus of attention

Picking out a single voice or other sound source within an environment containing several different voices is an everyday occurrence, and initially we are often not aware of the individual sources until we listen for them. Similarly, with training, listeners can hear a single instrument from a group playing together, or a single note from a chord.

Experiments have shown that under laboratory conditions listeners are able to hear details, when attending more closely to an auditory signal, that they do not hear initially. For example, when presented with a pair of vowels with the same pitch and spatial location and with similar intensity, listeners have the impression of one dominant vowel which is coloured by the other vowel, but they are able to identify both vowels at well over chance performance [52]. These results are consistent with the idea that scrutiny within the focus of attention enables details of the signal to be perceived that are not present in the initial gist representation. Training, requiring concentrated attention, may be necessary in order to hear the details: for example, listeners can be trained to discriminate formants within a vowel context [53] and to learn new phonemes [54].

4.7 Unattended streams are undifferentiated

In early work on the ‘cocktail party effect’, Cherry [55] found that, when listeners attend to one of two talkers, they may not be aware of the meaning or even the language of the unattended speech, although they are aware of basic physical properties such as pitch range and the end of the message.

Two studies specifically investigating the role of attention have also found that unattended parts of the signal appear to be undifferentiated. Brochard et

al. [56] presented listeners with complexes consisting of one to four concurrent subsequences of tones at various frequencies, each with a different repetition rate. Listeners were asked to detect a temporal irregularity in one of the subsequences; performance was not affected by the number of subsequences, indicating that the unattended subsequences were not differentiated. A further example of a lack of awareness of unattended characteristics occurs in the phenomenon of ‘change deafness’, analogous to change blindness: listeners were asked to repeat words spoken by a male talker, and more than 40% of participants failed to notice that halfway through the task a different male talker presented the words [57].

Some experiments have indicated that unattended parts of the signal are processed to some degree: a word in an unattended stream can cause semantic priming of words in the attended ear [58, 59]. This may however simply be an indication that the gist of the unattended stream is sufficient to activate semantic processes.

5 Representations of visual and auditory scenes

5.1 The nature of gist representations

A representation of the gist of a scene can take more than one form: at one extreme, it may cover the whole of the scene at a very low resolution, while at the other extreme it may only sparsely cover the scene but use fine detail (figure 2). In between, a representation may consist of medium-resolution elements and/or these elements may not cover the scene completely. It is feasible that representations are created at a range of scales, in a ‘hierarchy of gist’ as suggested for semantic gist [23]. Some examples of possible representations are described below.

There is evidence that coarse information is normally processed before fine detail: Schyns and Oliva [18] presented observers with hybrid images formed by superimposing two images, belonging to two easily distinguishable categories, that had been filtered at high or low spatial frequencies producing an outline or blurred image respectively. When the hybrid images were presented for 30 ms, the coarse or blurred image was identified more frequently, but when presented for 150 ms the fine outline was seen. However, if observers are presented with images at the fine scale before the hybrid images, they can be primed to see the fine scale of the hybrid image preferentially [27]. Oliva and Schyns [16, 27] propose that attention is driven to the scale that is diagnostic of the task; so, if fine detail is required, the observer will use that scale first. For example, detail may be required to distinguish an image of one city from another city, but not to distinguish an indoor from an outdoor scene. Thus, although the perception of the gist may be pre-attentive, the scale of the representation used may be affected by attention.

Oliva and Torralba [16, 60] suggest that visual scenes can be represented and categorised using a ‘spatial envelope’ consisting of representations at different scales. Each representation is based on statistical distributions of image properties, such as orientation or contrast, believed to be important in the early visual



Fig. 2. Illustration of two possible representations: left, broad and coarse; right, fine and sparse (50% coverage).

system, defining a number of perceptual dimensions such as naturalness, openness and roughness. Scenes with similar perceptual dimensions share the same semantic category. Representations at different scales may correspond to receptive fields, from local to global, according to the image property represented. A computational model exploring these ideas is under development [60].

In addition, there is evidence that accurate speech perception can be based on glimpsing a surprisingly small number of spectro-temporal regions of the auditory signal when the speech is masked by other talkers [61]. These regions of high energy tend to be sparsely distributed and support the idea of a sparse representation of gist. In fact, using a coarse and broad spectro-temporal representation might cause problems in audition, since information such as frequency harmonics and temporal onset differences, required to distinguish auditory objects, would be smeared out. However, such differences could perhaps be processed independently at an early stage, as suggested in section 4.5.

A model with some similarities to that of Oliva and Torralba has been proposed for sounds [62]. Although designed to classify audio signals, this model is inspired by mechanisms in the human auditory system. Auditory features based on spectral and temporal modulations at multiple scales determined after initial spectro-temporal analysis are used to distinguish types of sounds. Using this system, speech can be discriminated from non-speech (animal vocalisations, music and environmental sounds) based on differential responses to fast or slow changes, and narrow- or broadband spectra.

5.2 Saliency maps

A somewhat different type of representation is the *saliency map*. It was first proposed by Koch and Ullman (described in [63]) as a biologically plausible representation of bottom-up visual attention. Different implementations share a basic conceptual structure: features of an image corresponding to low-level properties such as orientation, intensity and colour are extracted in parallel to produce topographical maps which are then combined into a single saliency map indicating the salience or perceptual influence of each part of the image; attention is drawn towards the most salient event. In its pure form, it is complementary

to a high level gist representation: the saliency map gives no indication of the stimulus category, but instead indicates relevant locations. Indeed, recent work merges these two concepts and uses low level statistics for a categorization of the scene, thereby modulating a prior of the probability distribution of relevant locations [64].

Saliency maps are not limited to the visual domain. Kayser et al. [65] proposed auditory saliency maps, analogous to visual saliency maps, which use features such as intensity, frequency contrast and temporal contrast. The individual maps can be produced at various scales and combined into a single map using competition between the scales and features. This model was able to replicate findings in [66] indicating the relative salience of short, long and temporally modulated tones in a noisy background. Although there is neurophysiological evidence for activity consistent with the use of a mechanism similar to the saliency map, the most salient features have also been shown to be task-dependent, indicating that top-down mechanisms are important at an early stage of processing [67], providing further support for an attentional model such as the Reverse Hierarchy Theory.

6 Conclusions

In general, there is evidence that auditory processing is consistent with the ideas in Reverse Hierarchy Theory and other similar proposals [5, 11–17] suggesting that rapid processing of the gist of a scene occurs followed by detailed scrutiny of that part of the signal within the focus of attention. Rather than attention being applied after stream segregation has taken place, as is typical in models of auditory scene analysis, under this proposal an initial bottom-up gist processing stage (*‘audition-at-a-glance’*, comparable to hearing) provides an overview of the whole auditory scene, stimulating large receptive fields and/or category detectors. The default assumption would be that features of the signal are part of a single source unless there is clear evidence for segregation, such as large differences in pitch. Early processing would indicate the likely number of sources and category of each source, but top-down processes would focus on the attended source and analyse its detail (*‘audition-with-scrutiny’*, comparable to listening), determining the features of the attended stream via small receptive fields.

Further experimental evidence is required to determine which aspects of the signal would be available during initial gist processing and which aspects would be processed in detail, as well as whether detailed analysis would occur subsequent to, or in parallel with, gist processing (e.g. [68]).

Acknowledgments. This work was funded by the European Union Cognitive Systems STREP project POP (Perception On Purpose), FP6-IST-2004-027268.

References

1. Driver, J.: A selective review of selective attention research from the past century. *Brit. J. Psychol.* **92** (2001) 53–78

2. Biederman, I.: Visual object recognition. In Kosslyn, M., Osherson, D.N., eds.: *An Invitation to Cognitive Science: Visual Cognition*. Volume 2. (1995) 121–165
3. Treisman, A.M., Gelade, G.: Feature-integration theory of attention. *Cognitive Psychol.* **12**(1) (1980) 97–136
4. Bregman, A.S.: *Auditory Scene Analysis: The perceptual organization of sound*. MIT Press, Cambridge, MA (1990)
5. Hochstein, S., Ahissar, M.: View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* **36**(5) (2002) 791–804
6. Potter, M.C.: Short-term conceptual memory for pictures. *J. Exp. Psychol. Hum. L.* **2**(5) (1976) 509–522
7. Simons, D.J.: Current approaches to change blindness. *Vis. Cogn.* **7**(1-3) (2000) 1–15
8. Liberman, A.M., Isenberg, D., Rakerd, B.: Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Percept. Psychophys.* **30** (1981) 133–143
9. Jusczyk, P.W., Luce, P.A.: Speech perception and spoken word recognition: Past and present. *Ear and Hearing* **23**(1) (2002) 2–40
10. Slaney, M.: A critique of pure audition. In Rosenthal, D., Okuno, H., eds.: *Proc. 1st Workshop CASA, IJCAI, Montreal, Canada* (1995) 13–18
11. Navon, D.: Forest before trees: The precedence of global features in visual perception. *Cognitive Psychol.* **9** (1977) 353–383
12. Di Lollo, V., Enns, J.T., Rensink, R.A.: Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *J. Exp. Psychol. Gen.* **129**(4) (2000) 481–507
13. Nelken, I., Ahissar, M.: High-level and low-level processing in the auditory system: The role of primary auditory cortex. In Divenyi et al., P., ed.: *Dynamics of speech production and perception*. IOS Press, Amsterdam (in press 2006)
14. Cusack, R., Deeks, J., Aikman, G., Carlyon, R.P.: Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J. Exp. Psychol. Hum. P.* **30**(4) (2004) 643–656
15. Li, F.F., VanRullen, R., Koch, C., Perona, P.: Rapid natural scene categorization in the near absence of attention. *P. Natl. Acad. Sci. USA* **99**(14) (2002) 9596–9601
16. Oliva, A.: Gist of the scene. In Itti, L., Rees, G., Tsotsos, J., eds.: *Neurobiology of Attention*. Academic Press, Elsevier (2005) 251–256
17. Evans, K.K., Treisman, A.: Perception of objects in natural scenes: Is it really attention free? *J. Exp. Psychol. Hum. P.* **31**(6) (2005) 1476–1492
18. Schyns, P.G., Oliva, A.: From blobs to boundary edges: evidence for time-scale-dependent and spatial-scale-dependent scene recognition. *Psychol. Sci.* **5**(4) (1994) 195–200
19. Rousselet, G.A., Joubert, O.R., Fabre-Thorpe, M.: How long to get to the “gist” of real-world natural scenes? *Vis. Cogn.* **12**(6) (2005) 852–877
20. Bransford, J., Franks, J.J.: The abstraction of linguistic ideas: A review. *Acta Acust. Acust.* **1**(2-3) (1972) 211–249
21. Roediger, H.L., McDermott, K.B.: Creating false memories: remembering words not presented in lists. *J. Exp. Psychol. Learn.* **21**(4) (1995) 803–814
22. Koutstaal, W., Schacter, D.L.: Gist-based false recognition of pictures in older and younger adults. *J. Mem. Lang.* **37**(4) (1997) 555–583
23. Reyna, V.F., Brainerd, C.J.: Fuzzy-trace theory: An interim synthesis. *Learn. Individ. Differ.* **7**(1) (1995) 1–75
24. Crick, F., Koch, C.: A framework for consciousness. *Nat. Neurosci.* **6**(2) (2003) 119–126

25. Johnston, J., McLelland, J.L.: Perception of letters in words: Seek not and ye shall find. *Science* **184** (1974) 1192–1194
26. Kimchi, R.: Primacy of wholistic processing and global/local paradigm: A critical review. *Psychol. Bull.* **112**(1) (1992) 24–38
27. Oliva, A., Schyns, P.G.: Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychol.* **34**(1) (1997) 72–107
28. Levin, D.T., Takarae, Y., Miner, A.G., Keil, F.: Efficient visual search by category: Specifying the features that mark the difference between artifacts and animals in preattentive vision. *Percept. Psychophys.* **63**(4) (2001) 676–697
29. Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmidt, A.M., Dale, A.M., Hamalainen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., Halgren, E.: Top-down facilitation of visual recognition. *P. Natl. Acad. Sci. USA* **103**(2) (2006) 449–454
30. Moore, B.C.J.: An introduction to the psychology of hearing. 5th edn. Academic Press, London (2003)
31. Broadbent, D.E.: A note on binaural fusion. *Q. J. Exp. Psychol.* **7** (1955) 46–47
32. Lindblom, B., Brownlee, S., Davis, B., Moon, S.J.: Speech transforms. *Speech Commun.* **11**(4-5) (1992) 357–368
33. Green, K.P., Tomiak, G.R., Kuhl, P.K.: The encoding of rate and talker information during phonetic perception. *Percept. Psychophys.* **59**(5) (1997) 675–692
34. Carlyon, R.P., Cusack, R., Foxtton, J.M., Robertson, I.H.: Effects of attention and unilateral neglect on auditory stream segregation. *J. Exp. Psychol. Hum. P.* **27**(1) (2001) 115–127
35. Lippmann, R.P.: Speech recognition by machines and humans. *Speech Commun.* **22**(1) (1997) 1–15
36. Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M.: Speech recognition with primarily temporal cues. *Science* **270** (1995) 303–304
37. Saberi, K., Perrott, D.R.: Cognitive restoration of reversed speech. *Nature* **398**(6730) (1999) 760
38. Bashford, J.A., Riener, K.R., Warren, R.M.: Increasing the intelligibility of speech through multiple phonemic restorations. *Percept. Psychophys.* **51**(3) (1992) 211–217
39. Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D.: Speech perception without traditional speech cues. *Science* **212** (1981) 947–950
40. Robinson, K., Patterson, R.D.: The stimulus-duration required to identify vowels, their octave, and their pitch chroma. *J. Acoust. Soc. Am.* **98**(4) (1995) 1858–1865
41. Robinson, K., Patterson, R.D.: The duration required to identify the instrument, the octave, or the pitch chroma of a musical note. *Music Perception* **13**(1) (1995) 1–15
42. Moore, B., Gockel, H.: Factors influencing sequential stream segregation. *Acta Acust. Acust.* **88**(3) (2002) 320–333
43. Warren, R.M., Obusek, C.J., Farmer, R.M., P., W.R.: Auditory sequence: Confusion of patterns other than speech or music. *Science* **164**(3879) (1969) 586
44. Green, D.M.: Temporal acuity as a function of frequency. *J. Acoust. Soc. Am.* **54** (1973) 373–379
45. Jacobsen, T., Schroger, E., Alter, K.: Pre-attentive perception of vowel phonemes from variable speech stimuli. *Psychophysiology* **41**(4) (2004) 654–659
46. Tervaniemi, M., Winkler, I., Naatanen, R.: Pre-attentive categorization of sounds by timbre as revealed by event-related potentials. *Neuroreport* **8**(11) (1997) 2571–2574

47. Murray, M.M., Camen, C., Andino, S.L.G., Bovet, P., Clarke, S.: Rapid brain discrimination of sounds of objects. *J. Neurosci.* **26**(4) (2006) 1293–1302
48. Alain, C., Reinke, K., He, Y., Wang, C.H., Lobaugh, N.: Hearing two things at once: Neurophysiological indices of speech segregation and identification. *J. Cognitive Neurosci.* **17**(5) (2005) 811–818
49. Alain, C., Izenberg, A.: Effects of attentional load on auditory scene analysis. *J. Cognitive Neurosci.* **15**(7) (2003) 1063–1073
50. Sussman, E.S.: Integration and segregation in auditory scene analysis. *J. Acoust. Soc. Am.* **117**(3) (2005) 1285–1298
51. Darwin, C.J.: Auditory grouping. *Trends Cogn. Sci.* **1**(9) (1997) 327–333
52. McKeown, J.D., Patterson, R.D.: The time-course of auditory segregation: Concurrent vowels that vary in duration. *J. Acoust. Soc. Am.* **98**(4) (1995) 1866–1877
53. Kewley-Port, D.: Vowel formant discrimination II: Effects of stimulus uncertainty, consonantal context, and training. *J. Acoust. Soc. Am.* **110**(4) (2001) 2141–2155
54. Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., Yamada, T.: Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *J. Acoust. Soc. Am.* **96**(4) (1994) 2076–2087
55. Cherry, E.C.: Some experiments on the recognition of speech with one and with two ears. *J. Acoust. Soc. Am.* **25** (1953) 975–979
56. Brochard, R., Drake, C., Botte, M.C., McAdams, S.: Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation. *J. Exp. Psychol. Hum. P.* **25**(6) (1999) 1742–1759
57. Vitevitch, M.S.: Change deafness: The inability to detect changes between two voices. *J. Exp. Psychol. Hum. P.* **29**(2) (2003) 333–342
58. Mackay, D.: Aspects of the theory of comprehension, memory and attention. *Q. J. Exp. Psychol.* **25** (1973) 22–40
59. Banks, W.P., Roberts, D., Ciranni, M.: Negative priming in auditory attention. *J. Exp. Psychol. Hum. P.* **21**(6) (1995) 1354–1361
60. Oliva, A., Torralba, A.: Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res.* (in press 2006)
61. Cooke, M.: A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.* **119**(3) (2006) 1562–1573
62. Mesgarani, N., Slaney, M., Shamma, S.A.: Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE T. Audi. Speech. Lang. P.* **14**(3) (2006) 920–930
63. Itti, L., Koch, C.: Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**(3) (2001) 194–203
64. Torralba, A.: Modeling global scene factors in attention. *J. Opt. Soc. Am. A* **20**(7) (2003) 1407–1418
65. Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K.: Mechanisms for allocating auditory attention: An auditory saliency map. *Curr. Biol.* **15**(21) (2005) 1943–1947
66. Cusack, R., Carlyon, R.P.: Perceptual asymmetries in audition. *J. Exp. Psychol. Hum. P.* **29**(3) (2003) 713–725
67. Fecteau, J.H., Munoz, D.P.: Saliency, relevance, and firing: a priority map for target selection. *Trends Cogn. Sci.* **10**(8) (2006) 382–390
68. Cooke, M.: Auditory organisation and speech perception: Arguments for an integrated computational theory. In Ainsworth, W., Greenberg, S., eds.: *Proc. ESCA Workshop Aud. Basis Speech Perc.*, Keele, Worth Printing Ltd (1996) 186–193