Reprinted from *Computational Auditory Scene Analysis*. Dave Rosenthal and Hiroshi Okuno (Eds.).Mahwah, NJ : Lawrence Erlbaum Associates, 1997.

3 A Critique of Pure Audition

Malcolm Slaney
Interval Research Corporation

All sound-separation systems based on perception assume a bottom-up or Marr-like view of the world. Sound is processed by a cochlear model, passed to an analysis system, grouped into objects, and then passed to higher-level processing systems. The information flow is strictly bottom up, with no information flowing down from higher-level expectations. Is this approach correct? In this chapter, I first summarize existing bottom-up perceptual models. Then, I examine evidence for top-down processing, describing many of the auditory and visual effects that indicate top-down information flow. I hope that this chapter generates discussion about what the role of top-down processing is, whether this information should be included in sound-separation models, and how we can build testable architectures.

3.1 THESIS

In this chapter,[1] I discuss the flow of information in a sound-analysis system. Historically, in perceptual models of audition, information has flowed from low-level filters up toward cortical or cognitive processes. The title for this chapter comes from a view that this approach, although it may offer a simple or pure way to model perception, faces increasing evidence suggesting that it is time for us to revisit this architectural model.

The question of bottom-up versus top-down processing is well known, especially in the artificial-intelligence (AI) community. In the bottom-up world, all information flows from the sensor. Bits of information are collected and combined, until finally an object is recognized. In the top-down view of the world, we know that there is a table out there somewhere; all we need to do is to collect the evidence that supports this hypothesis. Because decisions are based on sensor data, information in a top-down system flows both up and down. In real life no system lies at either extreme, but the categorization provides a useful framework to describe information flow qualitatively.

From an engineering point of view, there are many advantages to modeling the perceptual system with bottom-up information flow. As each process is studied and understood, the essential behavior is captured in a model. The results can then be passed to the next stage for further processing. Each stage of the model provides a solid footing that permits the work at the next stage to proceed.

The science of perception is bottom up. This assertion is true for both the visual system and the auditory system. Peripheral processes are studied and used as building blocks in the journey toward the cortex. It is relatively easy to understand what a neural spike near the retina or the cochlea does, but it is much harder to understand what a spike in the cortex signifies.

Churchland, Ramachandran, and Sejnowski, in their recent book chapter, "A Critique of Pure Vision" (Churchland et al., 1994), questioned the assumption that information flows exclusively bottom up. There is much evidence, both behavioral and neurophysiological, that suggests that the visual system uses significant information that flows top down. They define a *pure* system as one that is exclusively bottom up; the alternative model is a *top-down* or an *interactive* system.

We shall look at the arguments in "A Critique of Pure Vision" and shall discuss their applicability to the auditory world. Have those of us who build auditory-perception systems ignored the avalanche of information from higher cognitive levels? With gratitude to Churchland, Ramachandran, and Sejnowski, I hope that this chapter will promote discussion, and will provide a framework for describing computational auditory-scene-analysis systems.

Section 3.2 reviews the case for pure vision and pure audition systems. Section 3.3 surveys the evidence for an interactive approach to audition and vision. Section 3.4 concludes with observations about possible future research. This chapter does not describe the role of efferent projections in the auditory and visual pathways. I hope that the examples cited here will inspire more study of neural top-down connections.

3.2 MARR'S VISION

David Marr's book *Vision* was a conceptual breakthrough in the vision and AI worlds (Marr, 1982). Most important, for the present discussion, is the argument that the visual system can be described as a hierarchy of representations.[2] At the lowest level, an image represents intensity over an array of points in space. Simple processing converts these pixels into lines, curves, and simple blobs. This primal sketch can then be converted into a 2 1/2-D[3] sketch by finding orientations and noting the discontinuities in depth, all from the perspective of the camera. Later processing then converts this sketch into a world view of the objects in the scene.

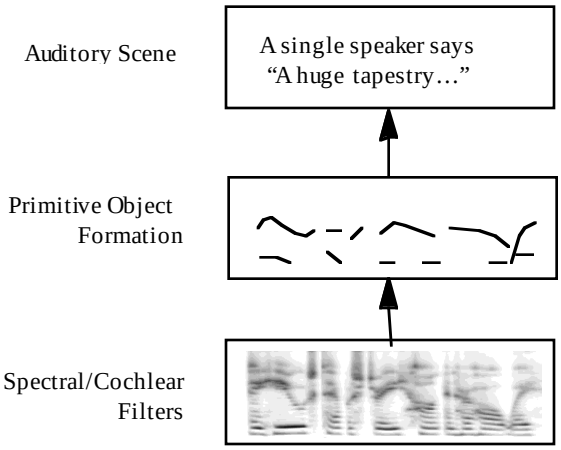

FIG. 3.1. A schematic of the pure-audition approach to auditory analysis.

Churchland and colleagues described a caricature of pure vision with the following attributes:

(1) We see the complete world. The retina records a complete image and we analyze it at our leisure.

(2) There is a hierarchy of information and representations.

(3) Information flows from bottom to top, with high-level representations depending on only the low-level processes, not vice versa.

This cartoon model of *pure vision* or *pure audition* serves as a reference point for one end of the pure–interactive scale. Whereas what Churchland called *interactive vision*, or what Blake (Blake & Yuille, 1992) called *active vision*, falls on the other end of the scale.

Many auditory systems have adapted the pure-vision philosophy. Figure 3.1 is an amalgam of system architectures as described by Mellinger (Mellinger, 1991), Cooke (Cooke, 1993), and Brown (Brown & Cooke, 1994) to do auditory sound separation. A filter stage feeds spectral information into an event analyzer and a detector. Later events are combined into objects by a process known as *scene analysis*. To my knowledge, all auditory-perception models, including my own, have assumed a bottom-up approach to the problem, often referring to Marr's notion as a guiding principle. Is this approach the best one?

Churchland, Ramachandran, and Sejnowski argued that the pure-vision view of the world is a dangerous caricature. Although computer vision has made much progress with this premise, the path could turn out to have a dead-end. They make these points:

> The idea of "pure vision" is a fiction, we suggest, that obscures some of the most important computational strategies used by the brain. Unlike some idealizations, such as "frictionless plane" or "perfect elasticity" that can be useful in achieving a core explanation, "pure vision" is a notion that impedes progress, rather like the notion of "absolute downess" or "indivisible atom" (Churchland et al., 1994, p. 24)

I worry that the same criticism applies to computational auditory scene analysis.

Churchland described—the opposite of pure vision—interactive vision or top-down processing, as follows:

(1) Perception evolved to satisfy distinct needs.

(2) We see only a portion of the visible world, although motion (or sudden sounds) can redirect our attention.

(3) Vision is interactive and predictive. It builds a model of the world and the visual system tries to predict what is interesting.

(4) Motion and vision are connected. We move to see more of the world.

(5) The neurophysiological architecture is not hierarchical; much information flows both ways.

(6) Memory and vision interact.

There is much evidence that the auditory system has many of the same properties. Perhaps our models should have them too.

A clear example of *interactive vision* is shown in Figure 3.2. Saccadic eye movements are plotted as a subject explores a visual scene. Clearly, the subject does not see the entire image at once. Instead she gradually explores pieces of it. Does a similar process occur in the auditory world?

3.3 EXAMPLES OF INTERACTIVE PROCESSES

There are many visual and auditory effects that are not what they seem. The following examples do not provide proof that the auditory and visual systems are interactive; instead, they serve to illustrate problems with a purely bottom-up view of processing flow. I describe global influences, motion, and categorization decisions that are influenced by the semantics, grouping, cross-modality influences, and the effect of learning. In all but the learning case, I give examples from the worlds of vision (from Churchland) and of audition.[4]

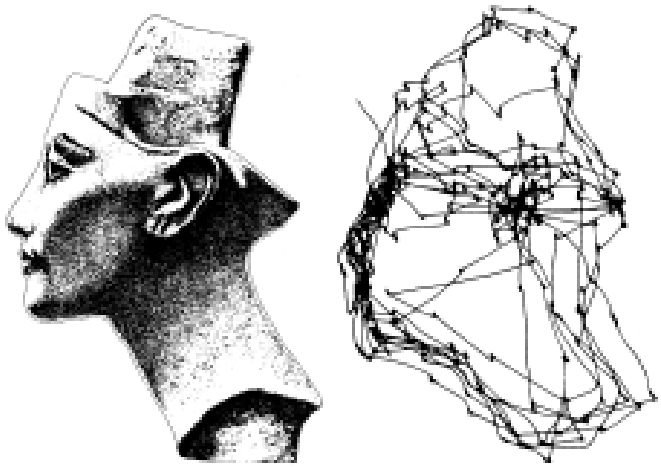

FIG. 3.2. The lines on the right are a plot the saccadic eye movements of a subject who is looking at the face on the left. (Source: Reprinted with permission from Yarbus, 1967.)

3.3.1 Global Influences

A basic feature of a pure system is that local features are all that the system needs to make decisions about the low-level properties of a stimulus. If a global property affects the local decision, then either the analysis of the two properties is different from that originally proposed, or a global or high-level information source is modifying the low-level percept.

Signal-Level Control. Both the auditory and the visual systems include control mechanisms to change the global properties of the received signal. The pupils of the visual system control the amount of light that falls on the retina. Likewise, at the lowest levels of the auditory system, efferent signals from the lower superior olivary complex affect the mechanical tuning of the cochlea, thus changing the size of the vibrations of the basilar membrane. Whereas both mechanisms are important, they do not change the information content of the signal and thus are not considered here.

Occlusion and Masking. A simple example of the type of information flow that we do want to consider is the way that we perceive occluded lines and tones. If the break is short, we see a continuous line. Likewise, if a rising chirp is partially replaced with a noise burst, we are convinced that we never heard the tone stop. The remainder of this section describes similar effects.

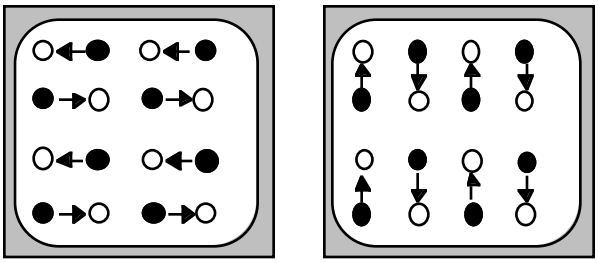

FIG. 3.3. Alternating white and black dots that create an illusion. Subjects see one uniform motion, either the motion indicated in the left or right image, and never see a combination of the two directions. (Source: Adapted with permission from Churchland et al., 1994).

3.3.2 Motion

Acoustic and visual motion provide evidence that perception is not a strictly hierarchical process. In some cases, local motion determines segmentation; in other cases, the segmentation and global properties determine the motion. A visual and an acoustic example show aspects of this hierarchy dilemma.

Vision: Bistable Quartets. Figure 3.3 illustrates a visual stimulus where the local motion is ambiguous. Motion can be perceived differently in different parts of the image; instead, however, when these two images are alternated, the subject sees all motion in the same direction. Similar examples are given in the Churchland chapter.

Audition: Deutsch Octave Illusion. Direct analogies to the bistable-quartet motion are difficult to find because acoustic-object formation is so strongly mediated by pitch and speech perception. A related auditory stimulus is presented by Diana Deutsch to show the effect of experience on perceived motion (Deutsch, 1990). Some people hear a two-tone pattern as ascending in pitch; when the pattern is changed to a different key, however, the same people hear it as descending. Deutsch reports that there is a correlation between the range of fundamental frequencies in the speaker's natural voice and the direction perceived.[5]

3.3.3 Categorization

In a purely bottom-up system, the semantic content of a stimulus does not affect the low-level perceptual qualities of a scene. Certainly, decisions such as object recognition and speech recognition are higher in the processing chain then are low-level perceptions like shape and sound characteristics.

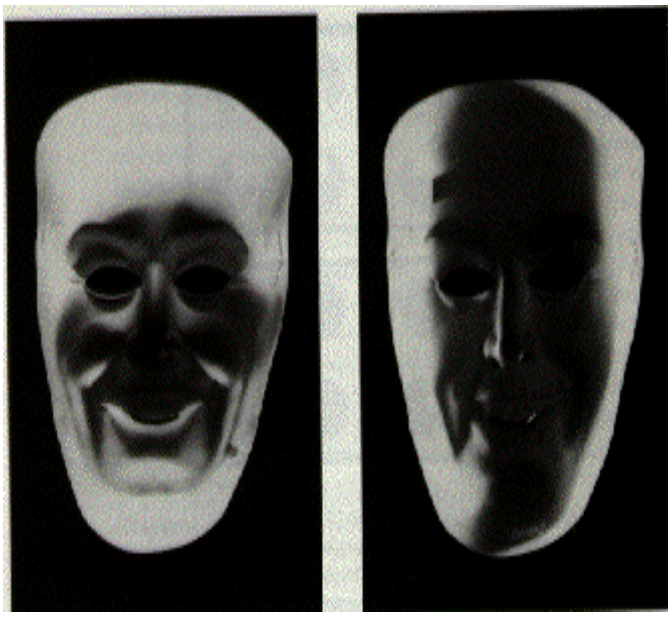

FIG. 3.4. Two concave masks photographed from their inside. The effect of faces on depth perception is illustrated. (Source: Reprinted with permission from Churchland et al., 1994.)

Vision: Faces and Shading. Figure 3.4 shows a simple example that illustrates visual ambiguity. Shading gives us important cues for determining what the shape of an object is. Most people see the masks in Figure 3.4 as having the nose projecting out of the page, even though the masks are in fact concave. Moving the lights from above—which is the direction from which we normally expect to see the light—to the sides does not change the perception that the nose is sticking out of the page as it would be from any normal face.

Audition: Ladefoged's Ambiguous Sentence. Context can dramatically affect the way speech is heard. Many people wonder whether speech is special and handled differently from other types of acoustic signals. I hope to illustrate how linguistic information and decisions can change our perception.

In Figure 3.5, the same introductory sentence is spoken by two different speakers. The final word, after each sentence, is the same—identical samples and waveforms. Yet most listeners hear the word at the end of the first sentence as "bit" and the word at the end of the second sentence as "bet." How can this difference occur if phonemes are recognized independent of their surroundings? Clearly, the words that we perceive, as shown by this example, are changed by our recent experience.

3.3.4 Grouping

Grouping together many components of a sound or scene is an efficient way for the perceptual system to handle large numbers of data. But can groups affect the low-level percepts? We would not expect a group to be formed unless all elements of the group have some property in common. Or is it possible that a

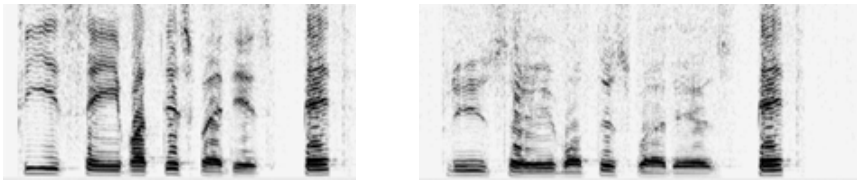

FIG. 3.5. Two spectrograms of the sentence “Please say what this word is: XX”. On the left, the last word is heard as “bit,” while on the right it is heard as “bet.” Identical waveforms are used in both cases. (Source: Audio courtesy of Peter Ladefoged.).

high-level decision is shorthand, so many low-level decisions are unneeded? The bistable quartets in Figure 3.3, and the dots in Figure 3.7 (described in Section 3.5) are also examples of visual grouping.

Audition: Sine-Wave Speech. Speech is often described as special because we hear spoken language as words, rather than as chirps, beeps, or random noise. A large orchestra produces sounds more complicated than those made by a single vocal tract, yet it is not hard for even untrained listeners to hear the piccolo part. Yet try as we might, we have a hard time describing more of the auditory experience of speech sounds than the pitch and the loudness. Language is certainly an important grouping process.

Sine-wave speech is an example of an acoustic signal that might or might not be heard as speech (Remez & Rubin, 1993). Figure 3.6 shows a spectrogram of a sine-wave speech signal. In sine-wave speech, the pitch of the acoustic signal is removed and the formants of the speech are modeled by a small number of sine waves (three in this case).

Most listeners first hear a sine-wave speech signal as a series of tones, chirps, and blips. There is no apparent linguistic meaning. Eventually, or after prompting, all listeners unmistakably hear the words and have a hard time hearing the individual tones and blips. Some of the tones remain, but it is as though the listener’s minds hear only the speech of normal speakers. With appropriate cueing, they hear the sounds as speech. The linguistic information in the signal has changed their perception of the signal.

Audition: The Wedding Song. Parts of speech can also be heard as music. Mariam Makeba recorded a musical piece called the *Wedding Song.* In the introduction, she names the song in Xhosa, an African click language. When she says the title, an American listener hears the click as part of the word. Yet when the listener hears the same type of click in the song, she hears it as separate from the speech, as part of the instrumental track. To my American-English ears, a click is not normally part of a language; when the click is placed into an ambiguous context, such as in a song, it is not heard as a part of the speech signal.

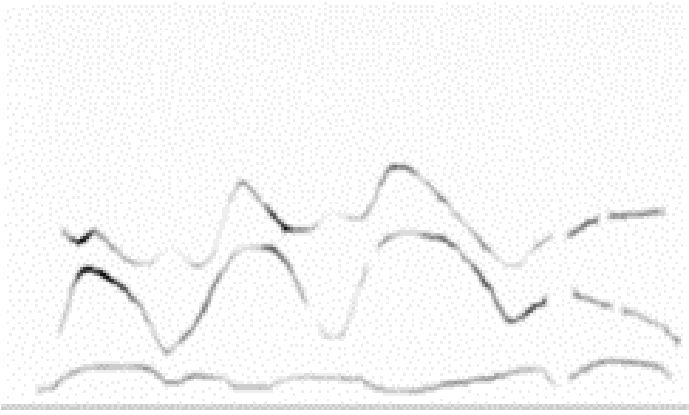

FIG. 3.6. A spectrogram of sine-wave speech. The sentence is “Where were you a year ago.” (Source: Audio courtesy of Richard Remez.)

3.3.5 Cross-Modality

Thus far, we have considered the auditory and visual processing systems only independently. Surely, in a pure system, an auditory signal would not affect what we see, and visual stimuli would not affect our auditory perception. However, we do get cross-perception effects. We all perceive the voices of television actresses as coming from their mouths, rather than from the television’s speakers, even if they are placed away from the screen. Two other such examples are described next.

Audition Affecting Vision: Behind the Occluder. Churchland and colleagues described a stimulus that illustrates illusory motion; it is shown in Figure 3.7. In each of three experiments, the dots in Column A are turned on and off in opposition to those in Column B (the square is always present). In the first experiment, the subject sees all three dots as moving back and forth, with the middle dot occluded by the square. In the second experiment, she sees the same dot as just blinking on and off. (These two experiments also provide an example of global changes affecting local perception.) Finally, in the third experiment, a tone is played in her left ear when the dot in Column A is shown. The dot and the left tone alternate with a tone played in her right ear. Apparent motion returns. Here an auditory event changed perception of the visual scene. How did this happen? The auditory stimuli added information to disambiguate the visual experience.

Vision Affecting Audition: The McGurk Effect. Vision can change the acoustic perception. The *McGurk effect*, an example of this cross-modality influence, is illustrated in Figure 3.8 (Cohen & Massaro, 1990). With our eyes closed, we hear a synthesized voice saying “ba.” When we open our eyes, and watch the

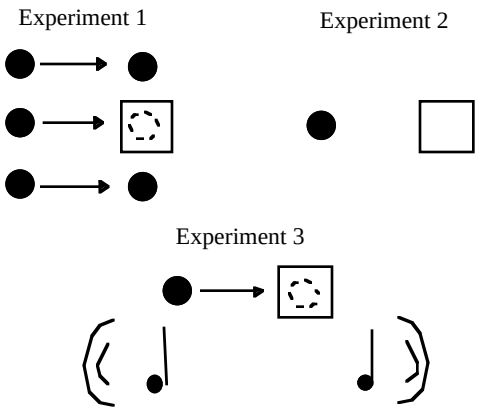

FIG. 3.7. Three experiments demonstrating illusory motion. The stimuli on the left alternate with those on the right. The arrows indicated perceived motion; the dashed circle indicates the object is perceived under the square. Global change cause the perception of motion in experiment 1; the tones played in left and right ears lead to motion in experiment 3. (Source: Adapted with permission from Churchland et al., 1994.)

artificial face, we hear "va." The acoustic signal is clearly "ba," yet the lips are making the motions for "va." Thus, our brains put together these conflicting information sources and, for this sound, trust the information from the eyes.

3.3.6 Learning

At the highest level, learning and training affect our perception over long periods. Most of the effects that we have discussed are immediate. Our perception is instantaneous and does not change much over time.

Yet training has been shown to change a owl monkey's ability to perform a discrimination task (Recazone et al., 1993). Over time, with much training, the owl monkey improved its ability to make frequency discriminations. Most important, the neurons in the AI section of monkey's cortex had reorganized themselves such that more neural machinery than before learning was dedicated to the task. A similar effect was seen with visual discrimination

3.4 FUTURE WORK

I know of no study that quantifies the information flow down the processing chain. Clearly, the centrifugal or descending auditory pathways are important. At the lowest levels, efferent signals from the superior olivary complex affect the mechanical tuning of the cochlea.

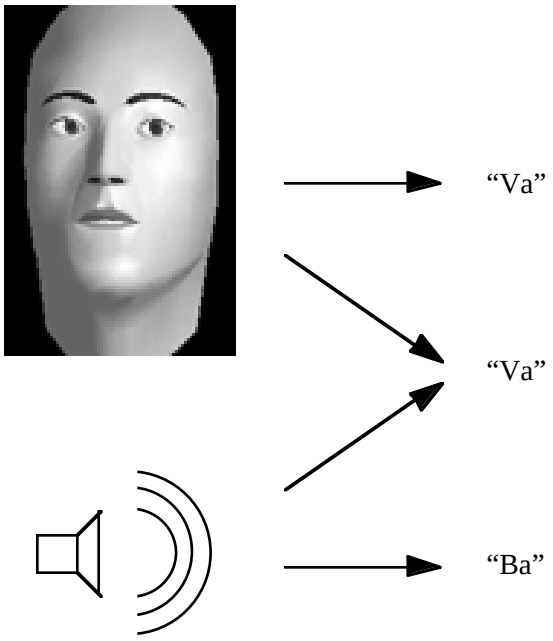

FIG. 3.8. The McGurk effect illustrates how visual stimuli can overrule the auditory perception. (Source: Image courtesy of Michael Cohen, University of California, Santa Cruz.)

Many of the examples in this and Churchland's chapter can be explained easily if low-level detectors generate all possible hypothesis. Higher-level processes then evaluate all the ideas, and suppress the inconsistent results. It is impossible for psychophysical experiments to rule out one or the other of these alternatives. To answer this question, we must perform experiments on efferent projections.

There are auditory systems that use top-down information. Most speech-recognition systems today use linguistic information and knowledge about the domain to guide the word-hypothesis search (Lee, 1989). A system proposed by Varga and Moore (1990) uses two hidden Markov model (HMM) recognizers to separate speech and noise. Works by Carver and Lessor (1992), Nawab (1992) and Ellis (1993) discuss blackboard systems that allow expectations to control the perception. Recent work by the Sheffield group (Cooke et al., 1995) used Kohonen nets and HMMs to recognize speech with missing information. These systems, however, are not tied to physiology or psychoacoustics. Is there common ground between speech recognizers and systems that perform auditory-scene analysis?

I do not mean to imply that pure audition is inherently bad. Interactive and top-down systems are hard to design and test. The world of perception offers little guidance in the design of these systems.

Instead, I hope to find a middle ground. I hope that those of us who design top-down systems will learn what has made the perceptual system successful. We wish to discover which attributes of the perceptual representation are important and should be incorporated into the top-down systems. Clearly the mel-frequency cepstral-coefficient (MFCC) representation in the speech-recognition world (Hunt et al. 1980) is one such win for perception science.

Likewise, those of us who design pure-audition systems need to acknowledge all the top-down information that we are ignoring in the pursuit of our sound-understanding systems. Much information is processed without regard to high-level representations. We clearly perceive the voice of somebody speaking a language we have never heard as being one sound source, rather than as isolated chirps and tones. Yet many problems—such as understanding how we separate speech from background at a noisy cocktail party—might be easier to solve if we pay attention to our understanding of the linguistic content. I, unfortunately, do not know how to do so yet.

ACKNOWLEDGMENTS

Earlier versions of this work were presented to the Perception Group at Interval Research and to the Stanford CCRMA Hearing Seminar. I am grateful for the feedback, criticism, and suggestions I received from all these people. Specifically, Subutai Ahmad, Michele Covell, and Lyn Dupré had many suggestions that greatly improved my ideas and their presentation.

REFERENCES

Blake, A. & Yuille, A. (Eds.). (1992). *Active Vision*. Cambridge, MA: MIT Press.

Brown, J. G. & Cooke, M. P. (1994). Computational auditory scene analysis. *Computer Speech and Language*, 8 (4), 297–336.

Carver, N. & Lessor, V. (1992). Blackboard systems for knowledge-based signal understanding. In *Symbolic and Knowledge-based Signal Processing,* Alan V. Oppenheim and S. Hamid Nawab, (Eds.). Englewood Cliffs, NJ: Prentice-Hall.

Churchland, P., Ramachandran, V. S., & Sejnowski, P. (1994). A critique of pure vision. In *Large-Scale Neuronal Theories of the Brain,* Christof Koch and Joel Davis, (Eds.). Cambridge, MA: MIT Press.

Cohen, M. M. & Massaro, D. (1990) Synthesis of visible speech. *Behavior Research Methods, Instruments and Computers*, 22(2), 260–263.

Cooke, M. (1993). *Modelling Auditory Processing and Organisation.* Cambridge, UK: Cambridge University Press.

Cooke, M., Crawford, M., & Green, P. (1995). Learning to recognise speech in noisy environments. ATR TR-H-121, *Proceedings of the ATR workshop on A Biological Framework for Speech Perception and Production*, Kyoto Japan, 13–17.

Deutsch, D. (1990). A link between music perception and speech production. *Abstracts for the One Hundred Twentieth Meeting of the Acoustical Society of America, Journal of the Acoustical Society of America*, Vol. 88 (Suppl. 1).

Ellis, D. P. W. (1993). Hierarchic models of hearing for sound separation and reconstruction. *1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, New York, pp. 157–160.

Hunt, M. J., Lennig, M., & Mermelstein, P. (1980). Experiments in syllable-based recognition of continuous speech. *Proceedings of the 1980 ICASSP*, Denver, CO, pp. 880–883.

Ladefoged, P. (1989). A note on 'information conveyed by vowels.' *Journal of the Acoustical Society of America*, 85 (5), 2223–2234.

Lee, K. (1989). *Automatic Speech Recognition: The Development of the SPHINX System*. Boston, MA: Kluwer Academic Publishers.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman and Company.

Mellinger, D. K. (1991). Event formation and separation in musical sound. Doctoral Thesis, CCRMA, Dept. of Music, Stanford University, Stanford, CA.

Nawab, S. H. (1992). Integrated processing and understanding of signals. In *Symbolic and Knowledge-Based Signal Processing*, Alan V. Oppenheim and S. Hamid Nawab, (Eds.), Englewood Cliffs, NJ: Prentice-Hall.

Recazone, G. H., Schreiner, C. E., Merzenich, M. M., (1993). Representation of primary auditory cortex following discrimination training in adult owl monkeys. *The Journal of Neuroscience*, 13(1), 87–103.

Remez, R. E. & Rubin, P. E. (1993). On the intonation of sinusoidal sentences: Contour and pitch height. *Journal of the Acoustical Society of America*, 94 (4), 1983–1988.

Varga, A. P. & Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. In *Proceedings of ICASSP-90*, Albuquerque, NM, Vol. 2, 845–848.

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York: Plenum Press.

NOTES

[1]An earlier version of this chapter was published as Interval Technical Report IRC1995-010.

[2]A second important aspect of Marr's work deals with the representations of the "information processing" task. A *computational learning theory* specifies what the goal of an algorithm is and why it is important. The *representation* and *algorithm* specify how the computational theory should be implemented. Finally, the *hardware representation* describes how the algorithm is realized. These distinctions are important in both the audition and vision worlds, and should be kept clearly in mind.

[3]Marr's book (Marr, 1982, pp. 128–129) describes a 2 1/2-D sketch as follows. "According to our emerging theory of intermediate visual information processing, however, a key goal of early visual processing is the construction of something like an orientation-and-depth map of the visible surfaces around a viewer. In this map, information is combined from a number of different and probably independent processes that interpret disparity, motion, shading, texture, and contour information. These ideas are called the 2 1/2-D sketch. … The full 2 1/2-dimensional sketch would include rough distances to the surfaces as well as their orientations; contours where surface orientation change sharply, which are shown dotted; and contours where depth is discontinuous (subjective contours), which are shown with full lines."

[4] Many of the examples in this chapter can be found online at
http://www.interval.com/papers/1997-056

[5] Furthermore, there is a strong difference in perception between subjects who grew up in California and those who grew up in the South of England.