Reprinted from *Proceedings of the Computational Auditory Scene Analysis Workshop*, 1995 International Joint Conference on Artificial Intelligence, David Rosenthal and Hiroshi Okuno, Co-chairs, August 19-20, 1995, Montreal, Canada, pp. 13–18. A more refined version of this paper will be published in the book *Computational Auditory Scene Analysis*, edited by David Rosenthal and Hiroshi G. Okuno, Erlbaum, 1998.

A Critique of Pure Audition

Malcolm Slaney
Interval Research, Incorporated
1801 Page Mill Road, Building C
Palo Alto, CA 94304
malcolm@interval.com

Abstract

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 the right approach? This paper summarizes the existing bottom-up perceptual models, and the evidence for more top-down processing. This paper describes many of the auditory and visual effects that indicate top-down information flow. Hopefully this paper will generate discussion about the role of top-down processing, whether this information should be included in sound separation models, and how to build testable architectures.

1 Introduction[1]

This paper discusses the flow of information in a sound analysis system. Historically, in perceptual models of audition, information has flowed from low-level filters up towards cortical or cognitive processes. The title for this paper comes from a view that this is the pure way to approach the problem. But increasing evidence in the visual world suggests that it is time to step back and revisit this architectural issue.

The question of bottom-up versus top-down processing is well known, especially in the 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 collect the evidence that supports this hypothesis. Since decisions are based on sensor data, information in a top-down system flows both up and down the hierarchy. In real life no system is at either extreme, but this scale provides a useful framework to qualitatively describe information flow.

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 (hopefully) for the work at the next stage to proceed.

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

A recent book chapter, "A Critique of Pure Vision" [Churchland *et al.*, 1994], written by Churchland, Ramachandran and Sejnowski, questions the assumption that information flows exclusively bottom-up. There is much evidence, both behavioral and neurophysiological, that suggests that the visual system has a large amount of information that flows top-down. In this paper and the Churchland chapter, a pure system is exclusively bottom-up, while the opposite extreme is a top-down or an interactive approach.

This paper looks at the arguments in "A Critique of Pure Vision" and discusses their applicability to the auditory world. Have those of us building auditory perception systems been ignoring the avalanche of information that is descending from higher levels? With apologies and thanks to the authors of the "Pure Vision" chapter, I hope this paper will promote discussion and provide a framework for describing computational auditory scene analysis systems.

Section 2 reviews the case for Pure Vision and Pure Audition. Section 3 surveys some of the evidence for an interactive approach to audition and vision. Section 4 concludes with some observations about future directions. This paper does not describe the role of efferent projections in the auditory and visual pathways. I hope the examples cited here will inspire more study of the neural top-down connections.

2 Marr's Vision

David Marr's book Vision was a conceptual break-through 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 at 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 1/2D sketch by finding orientations and noting the discontinuities in depth, all from the perspective

1. This paper is also published as Interval Technical Report IRC1995-010.

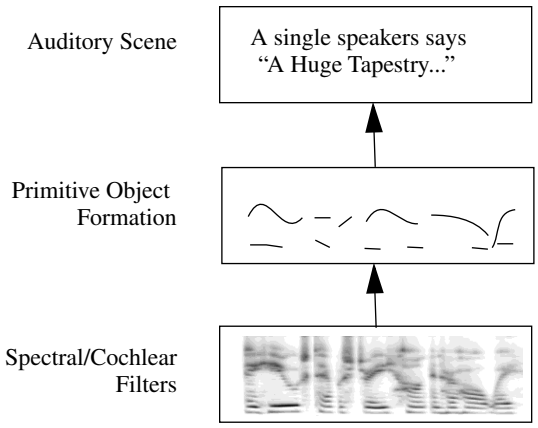

Figure 1. A schematic of the Pure Audition approach to auditory analysis.

of the camera. Later processing then converts this into a world-view of the objects in the scene.

Churchland *et al*, describe 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 only on the low-level processes, not visa-versa.

This is certainly a cartoon, but it serves as a reference point for one end of scale. Pure Vision or Pure Audition falls at this end of the scale while what Churchland calls Interactive Vision or Blake [Blake and Yuille, 1992] calls Active Vision falls on the other end of the scale.

Many auditory systems have adapted the pure vision philosophy. Figure 1 is an amalgam of system architectures as described by Mellinger [Mellinger, 1991], Cooke [Cooke 1993], and Brown [Brown and Cooke, 1994] to do auditory sound separation. A filter stage feeds spectral information into an event analyzer and 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 as a guiding principle. Is this the best approach?

Churchland, Ramachandran and Sejnowski argue that the Pure Vision view of the world is a dangerous caricature. While the computer vision field has made much progress with this premise, it could be a dead-end path. They point out:

> 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."

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

Churchland describes the opposite of Pure Vision, Interactive Vision or top-down processing, as having the following attributes:

1) Perception evolved to satisfy distinct needs.
2) We only see 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 tries to predict what is interesting.
4) Motion and vision are connected. We move to better see the world
5) The neurophysiology is not hierarchical. There is evidence that 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 too?

A clear indication of what is meant by interactive vision is shown in Figure 2. Saccadic eye movements are plotted as a subject explores a visual scene. Clearly the subject does not see the entire image at once, but instead gradually explores the image. Does the same thing happen in the auditory world?

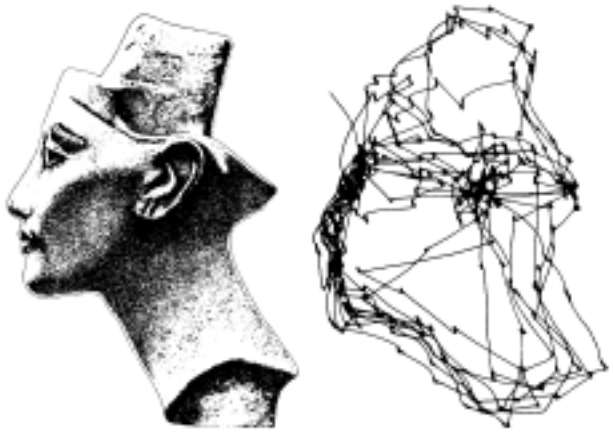

Figure 2. The lines on the right show the saccadic eye movements while a subject looks at the face on the left. (Reprinted with permission from [Yarbus, 1967])

3 Examples

There are many visual and auditory stimuli that are not what they seem. These examples are not proof that the auditory and visual systems are interactive, but instead serve to illustrate the problems with a purely bottom-up processing flow. The remainder of this section describes global influences, motion, categorization decisions that are influenced by the semantics, grouping, cross-modality influences, and the

2. A second important aspect of Marr's work deals with the representations of the "information processing" task. A *Computational Learning* theory specifies the goal of an algorithm 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.

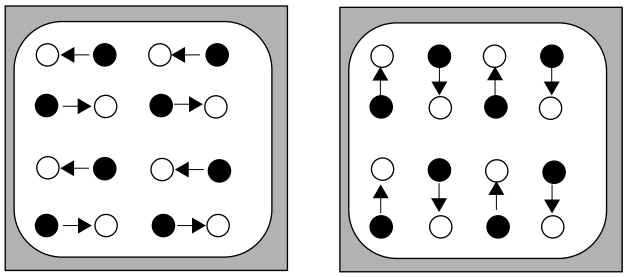

Figure 3. White and black dots alternate in the above illusion. Subjects see one uniform motion as indicated by the arrows, but never a combination of the two directions. (Adapted from [Churchland *et al.*, 1994].)

effect of learning. In all but the learning case, there are examples from both the worlds of vision (from Churchland) and audition.

3.1 Global Influences

A basic feature of a pure system is that local features are all that is needed 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 than original 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, efferents from the Lower Superior Olivary Complex affect the mechanical tuning of the cochlea, thus changing the size of the vibrations of the Basilar Membrane. While both of these mechanisms are important, they really don't change the information content of the signal and will not be considered here.

Occlusion and Masking

A simple example of the type of information flow we do want to consider is the way we perceive occluded lines and tones. If the break is not too long we see the line as being continuous. 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 other similar effects.

3.2 Motion

Motion is often a confusing part of the hierarchy. In some cases motion determines segmentation, while in other cases the segmentation and global properties determine the motion. The next two stimuli show aspects of this hierarchy dilemma

Vision: Bistable Quartets

Figure 3 illustrates a visual stimulus where the local motion is ambiguous, motion can be perceived differently in different parts of the image. But instead when these two images are alternated, all motion is seen in the same direction. Other similar examples are shown in the Churchland chapter.

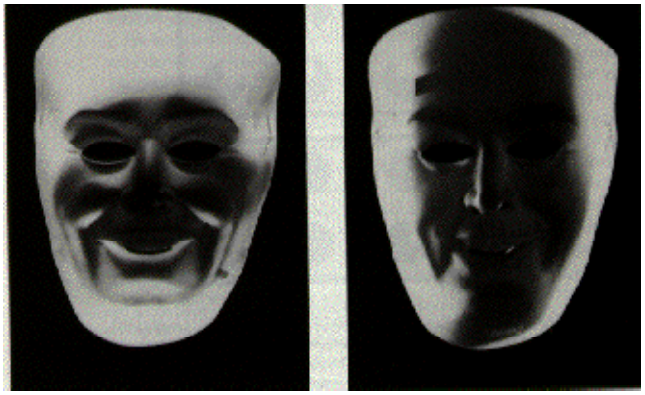

Figure 4. Two masks, photographed from their inside, illustrate the effect of faces on depth perception. (Reprinted with permission from [Churchland *et al.*, 1994].)

Audition: Deutsch Octave Illusion

Direct analogies to the bistable quartet motion is difficult because acoustic object formation is so strongly mediated by pitch and speech perception. But a related auditory stimuli is presented by Diana Deutsch showing the effect of experience on perceived motion [Deutsch, 1990]. A two-tone pattern is heard as ascending in pitch by some people, but when changed to a different key is heard by the same people to descend. Deutsch reports that there is a correlation between the range of fundamental frequencies in the speaker's natural voice and the perceived direction.[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 concepts such as object recognition and speech recognition are higher in the processing chain then low-level decisions like shape and sound characteristics.

Vision: Faces and Shading

A simple example from Churchland is shown in Figure 4 and illustrates the visual case. Shading is an important part of the way that we decide on the shape of an object. The masks in Figure 4 are seen normally with the nose projecting out of the page, even though they were photographed from the inside (concave side) of the mask. Moving the lights from above, as we normally expect to see the light, to the sides does not change the perception that the noise is sticking out of the page like any normal face.

Audition: Ladefoged's Ambiguous Sentence

Speech is certainly not always what it seems. Many people wonder if speech is somehow special and handled differently then other types of acoustic signals. For the purposes of this discussion the speech-is-special distinction is not important. Instead I hope to illustrate how linguistic information and decisions can change our perception.

Peter Ladefoged has prepared the two sentences shown in Figure 5 [Ladefoged, 1989]. The same introductory sentences are spoken by two different speakers. The last word in

3. Furthermore, there is a strong difference in perception between subjects that grew up in California versus those that grew up in the South of England.

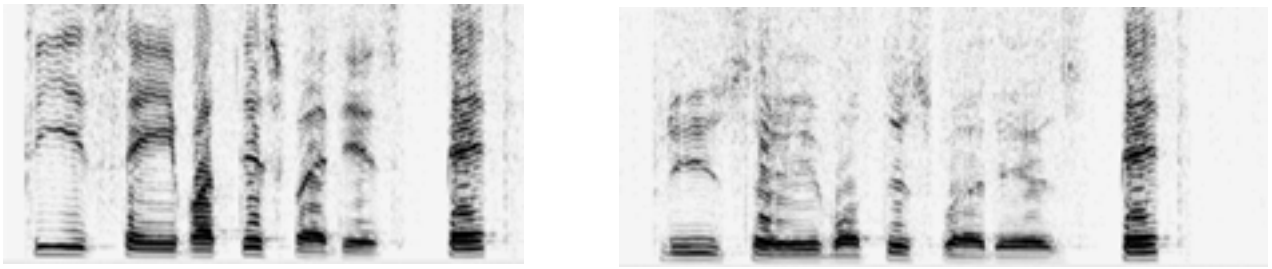

Figure 5. Two spectrograms of the sentence "Please say what this word is: XX". On the left, the last word is heard as "bit," while on the right it is heard as "bet." In both cases the last word are identical waveforms. (Example courtesy of Peter Ladefoged.)

the two sentences are 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 the this be if phonemes are recognized independent of their surroundings? Clearly the words we perceive, as shown by this example, are changed by recent experience.

3.4 Grouping

Grouping many components of a sound or scene together is an efficient way for the perceptual system to handle large amounts of data. But can groups affect the low-level percepts? One wouldn't expect a group to be formed unless all elements of the group have the same property. Or is it possible that a high-level group is a shorthand so that many low-level decisions are unneeded? The bistable quartets in Figure 3 and the dots in Figure 7 (to be described in Section 3.5) are also visual grouping examples so they won't be discussed here.

Audition: Sine Wave Speech

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

Sine-wave speech is an example of an acoustic signal that is close to the boundary. Figure 6 shows a spectrogram of a signal produced by Richard Remez of Barnard College. In sine-wave speech the pitch of the acoustic signal is removed and the formants of the speech are modeled by a small number sine waves (three in this case) [Remez and Rubin, 1993].

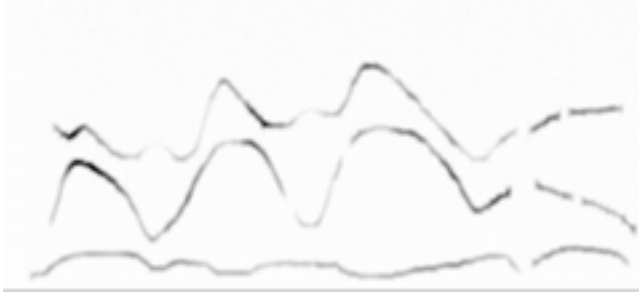

Figure 6. A spectrogram of sine-wave speech. The sentence is "Where where you a year ago." (Courtesy of Richard Remez.)

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 separating the tones and blips. Some of the tones remain, but it is as if our minds only hear the speech of normal speakers. With appropriate cuing we hear the utterances as speech. The linguistic information in the signal has changed our perception of the signal.

Audition: The Wedding Song

The speech versus tone/chirp/blip distinction also works the other way. Mariam Makeba has recorded a piece called the *Wedding Song*. In the introduction she names the song in its original tongue, an African click language. When she says the title the click is definitely heard as part of the word. Yet when the same type of click is heard in the song it separates from the speech and becomes part of the instrumental track. To my American-English ears, a click is not normally part of a language and when placed into an ambiguous context, such as a song, the click is no longer heard as a part of the speech signal.

3.5 Cross-Modality

So far we have only considered the auditory and visual processing systems independently. Surely in a pure system an auditory signal would not affect what we see, and visual stimuli would not affect our auditory perception. Unfortunately, this is not true. We all perceive the voices of TV actresses as coming from their mouths, not from the TV's speakers at some distance from the screen. Two other such examples are described next.

Audition affecting Vision: Behind the Occluder

Churchland *et al.* describe a stimulus that illustrates illusory motion. This example is shown in Figure 7. In each of the 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 all three dots are seen to move back and forth, with the middle dot occluded by the square. Yet in the second experiment the very same dot just blinks on and off. (These two experiments are also an example of global changes affecting local perception.) Finally in the third experiment a tone is played to the left ear when the dot in Column A is shown. The dot and the left tone alternate with a tone in the right ear. The apparent motion has returned. Here an auditory event has changed our perception of the visual scene. How can this be?

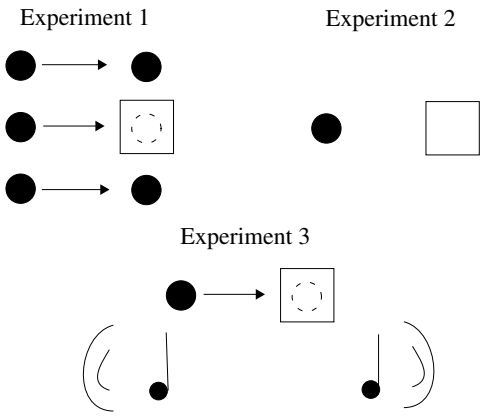

Figure 7. Three experiments demonstrate illusory motion, the arrows indicate perceived motion. Global change cause the perception of motion in Experiment 1, while the tones played in left and right ears lead to motion in Experiment 3. (Adapted from [Churchland *et al.*, 1994].)

Vision affecting Audition: The McGurk Effect

The opposite situation, vision changing the acoustic perception, is also possible. The McGurk effect is an example of this cross-modality influence and is illustrated in Figure 8 [Cohen and Massaro, 1990]. With our eyes closed we hear a synthesized voice saying "ba." When we open our eyes and watch the artificial face we hear "va." The acoustic signal is clearly "ba" yet the lips are making the motions for a "va." Thus our brains put together these conflicting information sources and for this sound trusts the information from the eyes.

3.6 Learning

At the highest level, learning and training affect our perception over long periods of time. Most of the effects we've discussed so far are immediate. Our perception is instantaneous and doesn't change much with time.

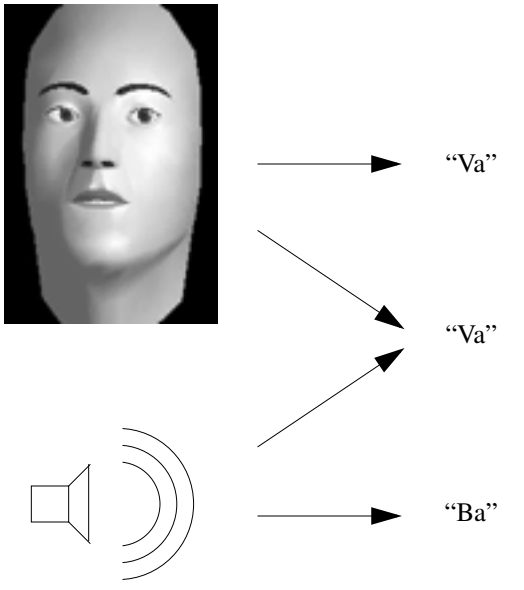

Figure 8. The McGurk effect illustrates how visual stimuli can overrule the auditory perception. (Example courtesy of Michael Cohen, Univ. of California, Santa Cruz.)

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, an Owl Monkey was found to have much better frequency discrimination ability. Most importantly, it was found that neurons in the AI section of cortex had reorganized themselves so that more neural machinery was dedicated to the task. A similar effect has been seen with visual discrimination

4 Where Do We Go From Here?

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 very lowest levels, efferents from the Superior Olivary Complex affect the mechanical tuning of the cochlea.

Many of the examples in this paper and Churchland's chapter can be explained 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 of the other of these alternatives. Experiments on efferent projections will be needed to answer this question.

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 search [Lee, 1989]. A system proposed by Varga [Varga and Yuille, 1990] uses two HMM recognizers to separate speech and noise. Work described by Carver [Carver and Lessor, 1992], Nawab [Nawab, 1992] and Ellis [Ellis, 1993] discuss blackboard systems that allow expectations to control the perception. Recent work by the Sheffield group [Cooke *et al.*, 1995] has looked at using Kohonen nets and HMMs to recognize speech with missing information. But alas, these systems are not tied to physiology or psychoacoustics. Is there common ground?

I do not mean to imply that Pure Audition is inherently bad. Interactive and top-down systems are hard to design and test. There is not much guidance from the world of perception to drive the design of these systems.

Instead I hope for some middle ground. Hopefully those of us that design top-down systems will learn what has made the perceptual system successful. Which attributes of the perceptual representation are important and should be incorporated into the top-down systems? Clearly the MFCC representation in the Speech Recognition world [Hunt et al. 1980] is one such win for perception science.

Likewise those of us that 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've never heard as being one sound source, and not as isolated chirps and tones. Yet many problems, such as understanding how we separate speech in a noisy cocktail party might be easier if some understanding of the linguistic content is included. I, unfortunately, don't know how to do this yet.

Acknowledgments

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

References

[Blake and Yuille, 1992] Andrew Blake and Alan Yuille, editors, *Active vision*, MIT Press, Cambridge, Mass. 1992.

[Brown and Cooke, 1994] G.J. Brown and M.P. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, 8, 297-336. 1994.

[Carver and Lessor, 1992] Norman Carver and Victor Lessor, "Blackboard systems for knowledge-based signal understanding," in *Symbolic and Knowledge-based Signal Processing*, Alan V. Oppenheim and S. Hamid Nawab, eds., Prentice Hall, 1992.

[Churchland *et al.*, 1994] Patricia Churchland, V. S. Ramachandran, Terrance Sejnowski, "A Critique of Pure Vision," in *Large-Scale Neuronal Theories of the Brain*, edited by Christof Koch and Joel Davis, MIT Press, 1994.

[Cohen and Massaro, 1990] M.M. Cohen and D. Massaro, "Synthesis of visible speech," *Behaviour Research Methods, Instruments and Computers*, vol.22, no.2, pp. 260-263, April 1990

[Cooke 1993], Martin Cooke, *Modelling auditory processing and organisation,* Cambridge University Press, Cambridge UK, 1993.

[Cooke *et al.*, 1995] Martin Cooke, Malcolm Crawford, and Phil Green, "Learning to recognise speech in noisy environments," ATR TR-H-121, *Proceeding of the ATR workshop on A Biological Framework for Speech Perception and Production*, 1995.

[Deutsch, 1990] Diana Deutsch, "A link between music perception and speech production, *Abstracts for the 120th Meeting of the Acoustical Society of America, ournal of the Acoustical Society of America, Supplement 1*, Vol. 88, Fall 1990.

[Ellis, 1993] Daniel P. W. Ellis, "Hierarchic models of hearing for sound separation and reconstruction," *1993 IEEE-Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, New York, NY, USA, p. 157-160, 1993.

[Hunt et al. 1980] M. J. Hunt, M. Lennig, and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," *Proceedings of the 1980 ICASSP*, Denver, CO, pp. 880-883, 1980.

[Ladefoged, 1989] Peter Ladefoged, "A note on 'Information conveyed by vowels'," *Journal of the Acoustical Society of America*, 85, pp. 2223-2234, 1989.

[Lee, 1989] Kai-Fu Lee, *Automatic speech recognition: the development of the SPHINX system*, Kluwer Academic Publishers, Boston, 1989.

[Marr, 1982] David Marr, *Vision*, W. H. Freeman and Company, 1982.

[Mellinger, 1991], David K. Mellinger, "Event formation and separation in musical sound," Unpublished Ph.D. Thesis, CCRMA, Dept. of Music, Stanford University, 1991.

[Nawab, 1992] S. Hamid Nawab, "Integrated processing and understanding of signals," in *Symbolic and Knowledge-based Signal Processing*, Alan V. Oppenheim and S. Hamid Nawab, eds., Prentice Hall, 1992.

[Recazone *et al.*, 1993] G. H. Recanzone, C. E. Schreiner, M. M. Merzenich, "Plasticity in the Frequency Representation of Primary Auditory Cortex following Discrimination Training in Adult Owl Monkeys," *The Journal of Neuroscience*, 13(1), pp. 87-103, January 1993.

[Remez and Rubin, 1993] R. E. Remez and P. E. Rubin, "On the intonation of sinusoidal sentences: contour and pitch height," *Journal of the Acoustical Society of America*, 94 (4), pp. 1983-8, Oct. 1993.

[Varga and Moore, 1990] A. P. Varga, and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in the *Proceedings of ICASSP-90*, Volume 2, pp. 845-848, Albuquerque, NM, 1990.

[Yarbus, 1967] Alfred L. Yargus, *Eye Movements and Vision*, Plenum Press, New York, NY, 1967.