Massaro, D. W. (1995). From Speech-Is-Special To Talking Heads: The Past To The Present. In Solso, R.L.,& Massaro, D.W. (Eds.), (1995). The Science of the Mind: 2001 and Beyond (pp. 203-220). New York: Oxford University Press.

FROM SPEECH-IS-SPECIAL TO TALKING HEADS: THE PAST TO THE PRESENT

Dominic W. Massaro

THE SETTING

Scene: Telecom Channel 46, January 7, 2101

Good audience, why partake in this antiquated pastime of scientific inquiry? Existence opened in mystery and will close in mystery. Our telecom channels have been designated to please, not puzzle. For those into delectation, the other telecom channels offer instantaneous desserts. Virtual Reality 3 presents Marilyn Monroe's rendezvous with Madonna III. If this tryst is too boring, there is the multistimulation of Bach's Brandenburg concerti guaranteed to bombard all sensory stations--sensory overload at its finest. Julia Child's gastronomic channel is serving up Stegosaurus, as reconstructed from simulations of the fossil record.

Figure 1: Talking Head of Communication Channel

If you haven't already dissolved my talking head *con* accompanying hand gestures, let me entice you not only to read my lips, but also to engage in the highest mode of thought of our ancestors. We do not have to dispute the consensus reached during the last century that *Homo* sapiens are not capable of knowing everything. Uncertainty in the outcome of our inquiry, however, should not preclude your participation. It's factual that thinking, problem solving, and inquiry are no longer taught in school or viewed as essential to success and happiness. Why do we need to penetrate the mysteries of the universe when we are nurtured and protected by intelligent machines of every kind? Who needs cogitation when we have every escapist philosophy imaginable, ranging from neurolinguistic programming to sleep learning to subliminal perception? Who can deny the exultation we find in walking over hot coals without pain, the ease of becoming an expert on some esoteric topic while asleep, and the ego boost we achieve in overcoming yet another frailty of our being by using the latest subliminal self-help disk. A recent discovery has revealed that the scientific puzzles deliberated during the twentieth century may still be worth pondering.

Supporting the rubble of the great quake of 1999, finally unearthed a century later, were books (bound pages of written language) of unprecedented importance. Although human readers of twentieth-century English were no longer available at the time of the discovery, the Smithsonian's computers and speech synthesizers were still functional and capable of translating this primitive written language reasonably well into the spoken language of that time. Although twentieth century speech sounds odd, it can be understood fairly easily. Why is our current spoken language no longer identical to this earlier form? There was an increased rate of sound change when we eliminated the written form of language. Remember that the major revision of written English early in the twenty-first century was aimed at establishing a regular correspondence between spelling and sound. Hypermediasts succeeded where Ben Franklin, Mark Twain, and George Bernard Shaw had failed. Now written language mirrored its spoken form. No longer could fish be spelled *ghoti* (gh as in rough, o as in women, and ti as in nation). Even with spelling-to-sound regularity, however, universal literacy was still beyond the reach of formal schooling. Notwithstanding the pledged intentions of youthful politicians, because of the huge budget deficit remaining from the twentieth century, the excessive cost and time required to teach literacy exceeded society's resources. Education and quality failed under Clintonomics in the same manner as under Reagonomics.

Given the dominance of the English language, all written language became extinct soon after the disappearance of written English. The scientists at that time consummated their research with the belief that reading was an unnatural act in contrast to the understanding of spoken language from which written language was derived. (As will be noted on the disk *Speech Perception and Cognitive Skills,* however, written language could also be acquired naturally without formal schooling.) Given simulated environments of talking beings that more closely engaged our natural processing, children no longer were required to struggle with written language. They could plug into (or be plugged into) any spoken lesson at any time and at any age. Contrary to the predictions of many of our intelligentsia of the time, the extinction of literacy was not accompanied by the fall of civilization. Written language was no more necessary for sagacious mentality than was color vision.

Within these ancient writings existed a review of speech perception research. Believe it or not, this topic evidently plagued the old sciences of the mind--dubbed cognitive science at the end of the twentieth century. With this discovery, we learn that some twentieth-century scientists anticipated the current view of perceptual, cognitive, and linguistic functioning. Today, we realize that evolution did not specifically give us the special skills required for our cognitive and technological world. Our sophisticated world is well beyond the comprehension of any one of us, but together we are an intelligent society. Multiple Pleistocene heads are better than one.

Speech perception

Understanding spoken language is only one of many domains of pattern recognition in which we impose meaning on an event by using multiple sources of information. We explore this domain because a) the availability of the ancient writings on the topic offer new insights, b) spoken language is typical of many of our worldly interactions, c) it is essential to appreciating the richness of communicating via talking heads, d) spoken language consumes many of our waking hours, and e) it is at least as important as fishing, even virtual fishing. We begin our inquiry by gaining an appreciation of the skill involved in speech perception.

Speech Perception: An Amazing Skill

Let us begin our study with the state of the art in speech science at the end of the twentieth century. Calling up the video archives from that time, we can experience the virtual reality of the following scenes. In a psychology laboratory, a six-week old infant in a baby chair has a pacifier in his mouth. As he sucks on the pacifier, the experimenter presents the sound /ba/ (as in *banana*) contingent on the infant's sucking. Given this feedback, the baby increases his sucking rate but soon becomes bored and sucks less. Now, however, the /ba/ sound is changed to /da/ and the infant increases his sucking rate again. The infant must have noticed the sound change from /ba/ to /da/. This initial research led to the development of the ingenious management devices available to today's caregivers. Bored infants are a thing of the past.

A petite three-year-old girl sits at a table of toy figures. She is told a short story and she must describe the story with the toy figures. To the child, she is playing a game, but to the psychologist and psycholinguist, she is displaying a remarkable ability to perceive and understand language. As an example, the child is told "The fence the horse kicks." The child takes the horse and has it kick the fence. This interpretation illustrates that the child has learned a constituent of the syntactic structure of English. The child's experience with subject-verb propositions is responsible for his understanding that the horse kicks the fence.

A small shipping company invests several thousand dollars to install an automatic speech recognition system. The operator reads the address on the package and simply calls out its destination for the machine to recognize. Contrary to the assurances of the manufacturer, the machine makes a variety of catastrophic errors. The system is most likely to fail when a talker speaks at a faster or slower rate than normal, when the talker forgets that he is talking to a machine and speaks with a lazy tongue, or when the talker has a cold. Why couldn't they design a machine to recognize speech as well as a three-year-old child?

The mysteries of understanding speech engaged speech scientists during the last four decades of the twentieth century. At the end of the twentieth century, scientists wondered how many more decades would be necessary to achieve enough understanding of spoken language understanding to build a machine to simulate this perhaps last specialization of *Homo Sapiens*. As expressed by George Miller (an ancestor of our hero in the Miller's Tale; see Chapter 9), "It enabled this big-brained, loudmouthed, featherless biped to overrun the earth..." (Miller, 1981, p. 1). Humans might not be able to claim language as uniquely theirs, but there can be no argument about speech. Chimpanzees and Apes can learn to sign but they aren't so constructed to speak. (It wasn't much later, however, that chimps at Yerkes laboratory and elsewhere were successfully learning how to understand spoken language, Savage-Rumbaugh et al., 1993.)

Perhaps because of the special nature of speech, the dominant belief at the end of the twentieth century was that *speech perception is special*. It had been reasoned, for example, that a speech "organ" (in the brain) had evolved to carry out this unique function. A speech organ is necessary because speech is a highly specialized domain that necessarily requires a specialized processing system. The minority alternative view was that understanding speech is just one domain of many that requires discrimination, categorization, and understanding. We also discriminate, categorize, and interact with everyday objects and events. Why should speech be any different? Although the controversy was not resolved, the specialization of speech perception became scientific dogma while the minority view was eventually forgotten. Spoken language became the dominant form of communication and we are now communicating via talking heads and accompanying hand gestures rather than by the written word. (Speed readers had to learn to search and skim spoken language with the same speed and prowess that they used on written language.) We will see that our current form of communication is consistent with this minority view. Your puzzle now is to undertake a retrospective examination of these two hypotheses of speech perception. Adopt the mental software of a twentieth century citizen confronted with this dilemma, and your disciplined inquiry will bear rewards that only a peek inside Father Nature's Trousers (or under Mother Nature's Skirt) can supply. To experience this twentieth century inquiry, we will enter its timeline to ponder how spoken language is understood. Put on your thinking caps because this inquiry from the twentieth century challenges inhabitants of the twenty-second.

SPEECH PERCEPTION SPECIALIZED?

A central issue in speech perception and psycholinguistics is the so-called modularity of speech and language. Noam Chomsky (1980) envisioned language ability as dependent on an independent language organ (or module), analogous to other organs such as our digestive system. This organ follows an independent course of development in the first years of life and allows the child to achieve a language competence that cannot be elucidated in terms of traditional learning theory. This mental organ, responsible for the human language faculty and our language competence, matures and develops with experience, but the mature system does not simply mirror this experience. The language user inherits rule systems of highly specific structure. This innate knowledge allows us to acquire the rules of the language, which cannot be induced from normal language experience because (advocates argue) of the paucity of the language input. The data of language experience are so limited that no process of induction, abstraction, generalization, analogy, or association could account for our observed language competence. Somehow, the universal grammar given by our biological endowment allows the child to learn to use language appropriately without learning many of the formal intricacies of the language. At the same time, however, other linguists are documenting that the child's language input is not as sparse as the nativists had argued (Sampson, 1989).

Although speech does not have an advocate as charismatic and influential as Chomsky, a similar description is given for speech perception. In addition, advocates of the special nature of speech are encouraged by Fodor's influential proposal of the modularity of mind. Our magnificent capabilities result from a set of innate and independent systems, such as vision, hearing, and language (Fodor, 1983). Speech-is-special theorists now assume that a speech module is responsible for speech perception (Liberman & Mattingly, 1989; Mattingly & Studdert-Kennedy, 1991). Given the environmental information, the speech module analyzes this information in terms of possible articulatory sequences of speech segments. The perceiver of speech uses his or her own speech-motor system to achieve speech recognition.

The justification for a speech module is analogous to the one for language more generally. Performance is not easily accounted for in terms of the language input. In speech, it is asserted that the acoustic signal is deficient and that typical pattern recognition schemes could not work. Put another way, it is reasoned that speech exceeds our auditory information-processing capabilities. In terms of the modularity view, our speech perception system is linked with our speech production system--and our speech perception is somehow mediated by our speech production. For theorists in the speech-is-special camp, the objects of speech perception are articulatory events or gestures. These gestures are the primitives that the mechanisms of speech production translate into actual articulatory movements and are also the primitives that the specialized mechanisms of speech perception recover from the signal. Before evaluating experimental evidence and other relevant findings concerning the special nature of speech perception, we begin with a historical sketch of the psychological study of speech perception.

A Historical Glimpse of the Twentieth Century

Speech perception wasn't always considered specialized. The turn of the nineteenth century was a heady time for psychologists. Fechner, Donders, Wundt, and their converts had paved the way for an experimental study of mental life. With tools such as a tachistoscope to present visual displays for short measurable intervals, and named as such to tongue-tie undergraduates before computer monitors made them obsolete (the T-scopes, not the undergraduates), experimenters could gain control over stimuli and derive stimulus-response relationships. Some of the best known work involved reading written words (which also captivated many "cognitive" psychologists during much of the twentieth century). One of the main findings to surface from this research was the important influence of context on reading. As documented in Edmund B. Huey's (1908) seminal text, our knowledge about spelling, syntax, and meaning facilitates the recognition of the letters on a page of text.

In contrast to the plethora of studies carried out on the written word, apparently only one was done on the spoken word. William Chandler Bagley's dissertation under Edward Titchener showed influences in speech perception that were analogous to those found in written language. Members of Cornell's psychology department were asked to recognize mutilated words with missing segments. This manipulation is reminiscent of Pillsbury's (1897) studies of the recognition of written words with missing letters. As can be seen in the examples, readers easily recognized the words even though they were spelled without all of their letters.

Examples of the letters exposed and the word read by a subject in Pillsbury's (1897) study.

Letters Exposed	Word Read
Commonly	commonly

Fashxon	fashion
Foyever	forever
Disal	deal
Uvermore	evermore
Danxe	danger

In Bagley's (1900) experiment, the naturally spoken words were recorded and played back on Edison phonograph cylinders. The results demonstrated that the context of the sentence improved recognition (and even perception) of the mutilated words. Word recognition was improved if the word was placed in the middle of a sentence, for example. This intuitive result was published in the leading psychological journal of the time, but was quickly forgotten, and speech more or less fell outside the domain of experimental psychology. Bagley's seminal study was not cited in Woodworth's *Experimental Psychology* (1938) and a twentieth century survey of psychology in America omitted any reference to speech perception (Hilgard, 1987). It also remained somewhat foreign during the "cognitive revolution," at the end of twentieth century, and only the technical goal of speech recognition by machine delegated speech perception its fair share of attention from experimental psychologists and other explorers of the mind.

At the beginning of the twentieth century, the psychological study of speech perception came, not from within psychology, but from an applied problem: a reading machine for blinded veterans returning from World War II. The goal was to design a machine that would read typewritten English and convert the letters into distinct sounds. The nonsighted listener would learn to recognize these sounds and read by ear. The scientists quickly found that the words spoken by machine were very difficult to understand and were not easily learned. This led Alvin Liberman and his colleagues to question why humans recognize natural speech so easily. Their inspiration was that we perceive speech via the same mechanisms used to produce speech: Speech was special. The nonsense sounds emanating from the speaking machine had little to do with how speech was spoken and, therefore, were gibberish to the listener. The next three decades of research from Haskins Laboratory was centered on the theme of the specialized nature of speech perception.

Evolutionary History of Speech

If speech perception is a highly unique and modular function, we would expect it to have a relatively long evolutionary history. Our speech is critically dependent on the characteristics of our respiratory system and vocal tract. Thus, it is of interest to determine the evolutionary history of the biological system used for speech. That is, a unique process would be expected to have a unique evolutionary history. Speech as we know it, however, appears to be relatively recent in our evolutionary history. Before the artificial speech of the last few decades, speech could be produced only by biological entities.

Using fossil records, Lieberman (1991) argued that speech as we know it was not possible just over 100,000 years ago. As can be seen in Figure 2, Neanderthal had a larynx positioned high, close to the entrance to the nasal cavity. The tongue was also positioned almost entirely in the mouth as opposed to being half in the pharynx, as it is in our mouths. Computer modeling showed that the Neanderthal vocal tract could not make many of our everyday speech sounds and would speak in a highly nasalized fashion. These characteristics would make speech a less than optimal communication system, primarily because the primitive segments of speech would be highly similar to one another. If Lieberman is correct, it wasn't until Homo sapiens evolved around 100,000 years ago that speech could have taken the form we know today. Although Lieberman's analysis is still being debated (Bradshaw & Rogers, 1993), it seems certain that speech is relatively novel by evolutionary standards.

Figure 2: Picture of Neanderthal Fossil

Because speech (as we know it) is so recent in our evolutionary history, it seems unlikely that a unique skill evolved to perceive speech and understand language. It appears that the astonishing brain growth of our ancestors occurred sometime before the development of speech and language as we know them. Given that the fundamental stuff of thought and language were probably already present, it is unlikely that specific brain structures had to evolve to empower speech production and speech perception. Our gift of language, thought, and culture must be due to exploiting the plasticity of the brain for communication. Although spoken language eventually emerged as the higher-level programing language of human computer systems, there doesn't appear to be anything in our evolutionary history that forces the conclusion that speech is special.

The Mystery of the Missing Phoneme

Linguists had invented the phoneme as the building block of speech. Phonemes are the minimal units in speech that can change the meaning of a word. The word *ten* has 3 phonemes: We can change the /t/ to /d/ to make *den*, the /e/ to /ae/ to make *tan*, and the /n/ to /l/ to make *tell*. Psychologists believed that recognizing speech must, therefore, necessarily involve recognizing phonemes. However, it did not seem to be possible to find the phoneme in the speech signal. Consider the syllable /da/: It has two phonemes /d/ and /a/. If we play this syllable in isolation, we hear /da/. Now if we repeatedly shorten this syllable by removing short segments from the end, we should eventually hear just /d/. Not true. Our percept changes from /da/ to nonsense, not from /da/ to /d/. Therefore, some magic must be involved in hearing both /d/ and /a/ given the syllable /da/.

The magic didn't stop here. We would expect to find some constant characteristic in the speech signal for a given phoneme. However, this was not the case. Figure 3 gives a visual representation of the sounds /di/ and /du/. Given that /d/ is first phoneme of both sounds, we should see the same signal at the beginning. We don't: The higher band of energy increases in /di/ and falls in /du/. One of the original arguments for the specialized nature of speech perception implicated this uncertain relationship between properties of the speech signal and a given phonemic category. It was emphasized that, in contrast to other domains of pattern recognition, one could not delineate a set of acoustic properties that uniquely defined a phoneme.

Figure 3: Picture of Spectrograms of /di/ and /du/

This argument holds very little force under close scrutiny. First, the psychological reality of phonemes can be questioned. Preliterate children and illiterates have trouble accessing the phonemes in spoken language. We modern illiterates, for example, have difficultly perceiving eight different speech segments in the word strategy. Most of us would say that it has just three segments. Similarly, a subjective experience of ma can occur without individual percepts of /m/ and /a/. It follows that phonemes might not be perceived at all, and much of the mystery can be overcome if the perceptual units of speech are larger than phonemes. In addition, some variability between the actual signal and the perceived pattern is not unusual in human pattern recognition. Therefore, the relationship between signal and percept in speech does not require us to accept that speech perception is specialized.

If phonemes were functional in speech perception, we would expect them to be ordered sequentially one after the other. However, they appear to be squashed together. This smudging of phonemes and their contextual variation is due to coarticulation--the articulation of one segment being influenced by the articulation of preceding and following segments. As visualized by Hockett (1955), phonemes are like a conveyer belt of eggs run through a wringer so that it is difficult to discern at what point one egg ends and the next begins. This overlapping of phoneme segments in

speech has also been enlisted in service of the argument that speech is special. However, the absence of a strict sequence of phonemic units does not necessarily require a specialized speech perception process. Perhaps, the most comparable situation is handwriting in which the visible characteristics of a letter are influenced by its adjacent neighbors.
(*Narrator's Overlay: This example is lost on us because we do not read, but a similar situation holds for movements in dance.)*

Figure 4: Picture of Eggs in Wringer

Rate of Speech Processing
One traditional argument for a special processor for speech is that the transmission rate of the speech signal appears to exceed our perceptual capacity. Phonetic segments--the minimum linguistic units of speech that are approximated by the letters of the alphabet--occur at a rate of between 10 and 20 per second. Supposedly, humans cannot identify nonspeech signals at even half this rate. There are several counterarguments to the rate argument, however. First, speech has a fast rate only when phonetic segments are taken as the psychologically real unit of analysis. Although many linguists promote the linguistic reality of these phonetic segments, there is no evidence that these segments are psychologically functional in speech perception. If larger units (such as syllables) are assumed to be the functional perceptual units in speech perception, then the rate of presentation of these signals is well within the range of our information-processing capability.

Second, a word could be recognized without necessarily recognizing the phonemes that make it up. If a sequence of arbitrarily selected sounds is presented, listeners have trouble identifying the order of the elements that make up the sequence unless each sound is presented for a quarter of a second or so. On the other hand, these same listeners can *discriminate* one of the sequences from another when the sounds are much shorter--in the range of 5 to 100 milliseconds (Warren, 1982). A sequence of short speech segments produces a unique percept that is necessarily informative for a communication system. Two different sequences of identical components are discriminated from each other because one arrangement is heard as different from the other. One might sound "bubbly" and the other like a "shrill," and people can even learn to label and identify these sequences.

A final problem with the argument that the rate of speech processing is greater than other forms of auditory information processing is the positive contribution of context. Our ability to process speech at a fast rate holds only for familiar speech. Even linguists have great difficulty transcribing a language that they do not know. Knowing a language allows us to perceive and understand speech given a deficient signal or very little processing time. For example, we can hear the first /s/ in the word *legislatures*, even when the relevant segment has been replaced by a tone (Warren, 1970). Similarly, we can perceive the speech of a language we know when it is speeded up at 2 or 3 times its normal rate (Foulke & Sticht, 1969). Finally, when spoken language is represented in written form, literates can read as quickly as they can listen. The impressively fast rate of processing spoken language does not require a specialized processor.

Categorical Perception
Categorical perception serves as the cornerstone for the view that speech is special. We can usually discriminate among more instances than there are categories of the instances. For example, we can discriminate among thousands of different colors, but have only a few dozen or so labels for them. Speech is believed to be different. The dogma is that perceivers are limited in their ability to discriminate differences among different speech sounds belonging to the same phoneme category. According to this view, the speech sounds within a category are identified only

absolutely, and discrimination is possible for only those sounds that can be identified as belonging to different categories.

Psychology and the speech sciences seem imprisoned by the notion of categorical perception perhaps, in part, because of phenomenal experience. One's phenomenal experience in speech perception is usually that of perceiving categories. If perception simply refers to our reported linguistic experience, then we cannot deny categorical perception because we naturally attend to the different categories of language. We cannot be swayed by linguistic experience because we have learned that it does not necessarily mirror the underlying processing. If perception refers to the psychological processing, however, then it is clear that the processing system is not limited to categorical information. Many empirical investigations have now demonstrated that perceivers are capable of perceiving differences within a speech category. For example, the ambiguity of tokens of a given syllable can be made synthetically, and presented as test items. People can reliably indicate the degree to which these different tokens represent the speech category. In addition, the ambiguous tokens require more time for categorization than do clear tokens. These results indicate that people can discriminate differences within a speech category and are not limited to just categorical information. The richness of the representation of a speech token is not obscured during speech perception, but retains its graded composite of information. Most likely because of the discrete structure of human communication via spoken language, however, decision processes simply map the rich continuous information into one of the discrete categories used in our language. The toddler must choose between perhaps a ball and a doll when his caregiver asks him to put away his doll, but may mutter something that is roughly a good match for either of these two words. Given that speech is not perceived categorically, the case for the modularity or specialization of speech is weakened considerably.

Development of Speech Perception

Modularity of speech necessarily has a large innate component. It is still common to attribute categorical perception to infants as well as adults (Eimas, 1985; Gleitman & Wanner, 1982). Although early studies appeared to find that infants noticed differences only between sounds from different speech categories and not between sounds from within the same speech category, followup studies quickly demonstrated that infants discriminate differences within, as well as between, categories. More generally, research with infants reveals that they discriminate the multiple dimensions of the auditory speech signal. However, the meaning of these differences in the language must be learned and infants are not prewired to categorize the signals into innate phonetic categories. The infant is analogous to the adult learning to label and identify sequences of meaningless sounds. It is as false to attribute categorical perception to the infant and child as it is to claim that fully developed adults are categorical perceivers (Massaro, 1987).

It was also experimentally demonstrated that infants and young children do not discriminate and categorize speech signals as well as adults. Their caregivers seem to be aware of this limitation because there is also a substantial amount of "motherese" during the first years of life. "Motherese" is spoken when the caregiver speaks clearly and slowly to the child. As with most skills, the child manifests a slow acquisition of the fundamental distinctions of our spoken language. Children have difficulty discriminating speech categories and their ability to discriminate increases gradually throughout childhood.

Narrator's Overlay: These discoveries of the slow and gradual development of speech are relevant to reading written language. When both were present in society, it was easy to conclude that reading was an unnatural act relative to speech. Speech seemed to be acquired naturally whereas reading required some formal instruction. With hindsight, however, we now understand that the advantage speech may have enjoyed was primarily its persistent presence from the womb *onward. On the other hand, the infant and toddler did not interact intensively with the written word and most children are shielded from it until some formal schooling. You are probably unaware of the scandalous infant-read experiment that was undertaken at about time that literacy was becoming extinct. Infants, from the time of birth, were equipped with specialized goggles that presented the written transcription of all spoken language in her environment. (The experiment originally used females because of their assumed superior language skills, but follow-up studies with males showed the same result.) The state of the art in speech recognition by machine had improved sufficiently to translate the spoken language of the infant's caregivers into a written form as it was being spoken. These infants procured literacy with no formal schooling and hand-in-hand with its spoken form. These results were suppressed by government agencies that had just renovated all the libraries (buildings containing books) so they could be sold as high-income housing. The shredded books made excellent insulation and this innovative housing easily satisfied the current stringent energy requirements.*

Retrospectively, we can comprehend that the research attempting to prove that speech is special was verificationist in approach, rather than adhering to the sacred scientific tenets of falsification and strong inference. Speech researchers weren't alone: The most striking example from the end of the twentieth century involved the putative language of bees (Wenner & Wells, 1990). Experiments had convinced most scientists that bees communicate the direction hive. These experiments and the language hypothesis even earned a Nobel Prize. It was only many years later that a few investigators seriously considered alternative hypotheses. It was then possible to design experiments without a confirmation bias. When these experiments were carried out, the bee language hypothesis failed. Similarly, advances in the understanding of speech perception
End of Narrator's Overlay

THE NONSPECIALIZED NATURE OF SPEECH PERCEPTION

Not only is the documentation for the special nature of speech perception weak, there is also corroboration to support the idea that speech perception is simply one of many domains of pattern recognition. No specialization is required for speech any more than for recognizing objects, melodies, and written language. We review a few of these sources of evidence here.

Contextual Effects in Speech Perception

A strong source of evidence against the modularity of speech perception involves the strong contribution of linguistic and situational context to speech perception. We perceive language more easily when we have some expectation of what the talker is going to say. Many of our conversations involve situations in which we find ourselves predicting exactly what the talker will say next. One hundred years after Bagley's first demonstration, experiments are still demonstrating that sentential context can facilitate word recognition. Situational context can also improve word recognition (Pollack & Pickett, 1963).

Speech Perception by Nonhumans

There is another source of evidence against the hypothesis that speech perception is carried out by a specialized module unique to humans. If speech perception is special and mediated in any way by speech production, then discrimination and recognition of fundamental speech categories should be impossible for nonhumans. However, some nonhuman animals can discriminate fundamental speech segments. Chinchillas (a small rodent with auditory capabilities close to humans) can discriminate fine distinctions in our spoken language. Even quail can learn to discriminate a set of syllables beginning with the stop consonant /d/ from a set of syllables beginning with the stops /b/ and /g/ (Kluender, Diehl, & Killeen, 1987). If there is information in the auditory speech signal that can be processed using normal perceptual mechanisms, we would expect that speech perception would not be limited to humans. More recently, chimps at Yerkes appear to be learning how to understand spoken language when regularly paired with other meaningful symbols (Savage-Rumbaugh et al., 1993)

Reading Speech Spectrograms

As shown in Figure 5, speech spectrograms are visible representations of the acoustic characteristics of speech. When a device that translated speech into spectrograms was invented after World War II, there was the obvious hope that people with impaired hearing could learn to read spectrographic displays. This would give these individuals direct access to spoken language and, therefore, would allow them into the dominant linguistic community. The initial training studies and followup experiments and applications were very promising, but somehow these positive results were camouflaged by the Zeitgeist that speech is special. It wasn't until an expert spectrogram reader was reported (Cole et al., 1980) and some additional positive training studies were completed (Greene et al., 1984) that researchers had to acknowledge that the signal for speech might well be within the purview of a general pattern recognition system. It is not unreasonable that people might be capable of utilizing both acoustic and spectrographic information in speech recognition. The value of reading spectrographic patterns has been substantiated in telecommunications because speech spectrograms are one of the additional sources of information currently available on most channels.

Figure 5: Speech Spectrogram of Speech That You May See

Speech Perception and Cognitive Skills

If speech perception is governed by a specialized noninteractive module, we would expect no relationship between speech and other skills. However, there is a positive correlation between motor skills and language, and also one between cognitive functioning and vocabulary size. For example, there is a positive correlation between cognitive development and the learning of new words (Gopnik & Meltzoff, 1984). It seems that speech perception can be considered as one of several perceptual or cognitive functions that can be understood in terms of basic skills.

In conclusion, if you have persisted in our deliberations, I trust that you agree that the research we have reviewed weakens the claim that speech perception requires a specialized module. It is now time to consider speech perception as one of many different forms of pattern recognition.

Speech Perception as Pattern Recognition

Speech perception might be best understood in terms of general perceptual, cognitive, and learning processes. The guiding assumption for this framework is that humans use multiple sources of information in the perceptual recognition and understanding of spoken language. In this regard, speech perception resembles other forms of pattern recognition and categorization because integrating multiple sources of information appears to be a natural function of human endeavor. Integration appears to occur to some extent regardless of the goals and motivations of the perceiver. A convincing demonstration for this fact is the Stroop color-word test.

Narrator's Overlay. This test, well-known to most children and adults, became extinct with the loss of literacy. Literates, asked to name the color of the print of words that are color names printed in a color other than the word (for example, the word red printed in blue type), became tongue-tied and had difficulty naming the colors. Evidently, literates cannot stop themselves from reading the color word, and this interferes with naming the color of the print. The analogous demonstration for illiterates is to identify the pitch of a speaker's voice as high or low. It is much harder to do so when the speaker says the word "low" in a high pitch, requiring the perceiver to say high while perceiving *the word "low" (McClain, 1983). This is where we leave our twentieth-century inquiry and close with a few observations linking research at that time with today's communication media.*

SPEECH PERCEPTION BY EYE AND EAR

Good participants, have you wondered why we are communicating face-to-face rather than via just sound. A century ago, virtual-reality videophones (VRVPs) had not yet replaced telephones that provided only the speaker's voice. Experiments had revealed conclusively that our perception and understanding are influenced by the visible movements in the speaker's face and the accompanying gestural actions. These experiments have shown that the speaker's face is particularly helpful when the auditory speech is degraded as a result of noise, bandwidth filtering, or hearing-impairment (Massaro, 1987; Summerfield, 1991). Although the influence of visible speech is substantial when auditory speech is degraded, visible speech also contributes to performance even when paired with intelligible speech sounds. The importance of visible speech is most directly observed when conflicting visible speech is presented with intelligible auditory speech. One famous example resulted from the dubbing of the auditory syllable /ba/ onto a videotape of a talker saying /ga/ (McGurk & MacDonald, 1976). A strong effect of the visible speech is observed because a person will often report perceiving (or even hearing) the syllable /da/, /va/, or /tha/, but seldom /ba/ corresponding to the actual auditory stimulus.

One attractive aspect of providing or using audible and visible speech jointly is the complementarity of audible and visible speech. Visible speech is usually most informative for just those distinctions that are most ambiguous auditorily. For example, place of articulation (such as the difference between /b/ and /d/) are difficult via sound but easy via sight. Voicing (such as the difference between /b/ and /p/, on the other hand, is difficult to see visually but is easy to resolve via sound. Thus, audible and visible speech not only provide two independent sources of information, these two sources are often productively complementary. One is strong when the other is weak.

Speech researchers quickly saw the value of studying bimodal speech perception. Given the importance of visible speech and the perceiver's natural ability to integrate multiple sources of information, a new experimental paradigm was possible. To control the visible speech, it was necessary to develop an animation system for visible speech synthesis. A critical assumption of this effort concerned the experimental, theoretical, and applied value of synthetic speech. Auditory synthetic speech had proven to be valuable in all three of these domains. Much of what we know about speech perception has come from experimental studies using synthetic speech. Synthetic speech gives the experimenter control over the stimulus in a way that is not always possible using natural speech. Synthetic speech also permits the implementation and test of theoretical hypotheses, such as which cues are critical for various speech distinctions. The applied value of auditory synthetic speech was apparent even then, before the extinction of written language, in the multiple everyday uses for text-to-speech systems that appealed to both normal and visually-impaired individuals.

It was believed that visible synthetic speech would prove to have the same value as audible synthetic speech. Synthetic visible speech could provide a more fine-grained assessment of psychophysical and psychological questions not possible with natural speech. For example, testing people with synthesized syllables intermediate between several alternatives gives a more powerful measure of integration relative to the case of unambiguous natural stimuli. It was also obvious that synthetic visible speech had a valuable role to play in alleviating some of the communication disadvantages of the deaf and hearing-impaired. Analogous to the valuable contribution of using auditory speech synthesis in speech perception research, visible speech synthesis permitted the type of experimentation necessary to determine (1) what properties of visible speech are used, (2) how they are processed, and (3) how this information is integrated with auditory information and other contextual sources of information in speech perception.

The development of a realistic, high-quality, facial display provided a powerful tool for investigation of a number of questions in auditory-visual speech perception. The analysis of the articulation of real speakers guided the development of visible speech synthesis. In addition, perception experiments indicated how well the synthesis simulated real speakers. The results of this research were used to implement automatic lipreading to enhance speech recognition by machine. Just as human perceivers achieved robust recognition of speech by using multiple sources of information, the same was true for machine recognition.

One applied value of visible speech was its potential to supplement other (degraded) sources of information. Visible speech is particularly beneficial in poor listening environments with substantial amounts of background noise. Its use is also important for hearing-impaired individuals because it allows effective spoken communication--the universal language of the community. Just as auditory speech synthesis has proved a boon to our visually-impaired citizens in human-machine interaction, visual speech synthesis should prove to be valuable for the hearing-impaired. Finally, synthetic visible speech had an important part in building synthetic "actors" (Thalmann & Thalmann, 1991) and played a valuable role in the then exciting new sphere of virtual reality.

Another source of information is tactile, which appears to be naturally integrated with auditory or visual speech in the same way that auditory and visual speech are integrated. For example, deaf individuals benefit from both tactile and visual speech in the same way that hearing-impaired individuals benefit from both auditory and visual information. The value of tactile speech is also illustrated by deaf nonsighted individuals who can perceive speech by holding their hands on the speaker's face. This Tadoma method has proved to be a successful channel of communication, and it has even been demonstrated that hearing individuals can exploit tactile information in speech perception. Speech can be translated into a tactile form and transmitted via a virtual reality glove to the perceiver. The tactile output of our communication devices has proved to be a valuable source of information.

Good audience, we have now traveled through time to learn why our current dialogue is embellished with multiple sources of information. The consistent downsizing and downpricing of communication technology permitted the multifaceted interaction we take for granted. However, it was speech science and psychological theory that laid the foundation for our virtual world of communication. I hope to meet you in person someday (or have I already

Acknowledgment
The writing of this paper was supported, in part, by grants from the Public Health Service (PHS R01 NS 20314) and the National Science Foundation (BNS 8812728).

REFERENCES

Bagley, W. C. (1900). The apperception of the spoken sentence: A study in the psychology of language. American Journal of Psychology, 12,80-130.

Bradshaw, J., & Rogers, L. (1993). The evolution of lateral asymmetries, language, tool use, and intellect. San Diego, Calif.: Academic press.

Chomsky, N. (1980). Rules and representations. Oxford: Blackwell.

Cole, R. A., Rudnicky, A. I., Zue, V. W., & Reddy, D. R. (1980). Speech as patterns on paper. In R. A. Cole (Ed.) Perception and production of fluent speech. Hillsdale, N. J.: Erlbaum.

Eimas, P. D. (1985, January). The perception of speech in early infancy. Scientific American, 252(1), 46-52.

Fodor, J. A. (1983). Modularity of Mind. Cambridge, Mass.: Bradford books.

Foulke, E., & Sticht, T. G. (1969). A review of research on time compressed speech. Psychological Bulletin, 72, 50-62.

Gleitman, L. R., & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner, & L. R. Gleitman (Eds.), Language acquisition: The state of the art. (pp. 3-48). Cambridge: Cambridge University Press.

Gopnik, A., & Meltzoff, A. N. (1984). Semantic and cognitive development in 15- to 21-month-old children. Journal of Child Language, 11, 495-513.

Greene, B. G., Pisoni, D. B., & Carrell, T. D. (1984). Recognition of speech spectrograms. Journal of the Acoustical Society of America, 76, 32-43.

Hilgard, E. R. (1987). Psychology in America. .R San Diego: Harcourt Brace Jovanovich.

Hockett, C. F. (1955). A manual of phonology, Memoir No. 11 of International Journal of American Linguistics. Baltimore: Waverly Press.

Huey, E. B. (1968). .I The psychology and pedagogy of reading. .R Cambridge, Mass: MIT Press. (Original work published 1908)

Kluender, K. R., Diehl, R. L. , & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. Science, 237, 1195-1197.

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. Science, 243,

Lieberman, P. (1991). Uniquely human. Cambridge, Mass: Harvard University Press.

McClain, L. (1983). Stimulus-response compatibility affects auditory Stroop interference. Perception & Psychophysics, 33, 266-270.

Massaro, D. W. (1987). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, N. J.: Erlbaum.

Mattingly, I. G., & Studdert-Kennedy, M. (1991) Modularity and the motor theory of speech perception. Hillsdale, N. J.: Erlbaum.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Miller, G. A. (1981). Language and speech. San Francisco: Freeman.

Pillsbury, W. B. (1897). A study in apperception. .I American Journal of Psychology, 8, .R 315-393.

Pollack, I., & Pickett, J. M. (1963). The intelligibility of excerpts from conversation. .I Language and Speech, 6, .R 165-171.

Sampson, G. R. (1989). Language acquisition: Growth or learning? Philosophical Papers, 18, 203-240.

Savage-Rumbaugh, E. S., Jurphy, J., Sevcik, R. A., Brakke, K. E., Williams, S. L., & Rumbaugh, D. M. (1993). Language comprehension in ape and child. Monographs of the Society for Research in Child Development, Vol. 58, Nos. 3-4.

Summerfield, Q. (1991). Visual perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy, M. (Eds.) Modularity and the motor theory of speech perception (pp. 117-137). Hillsdale, N. J.: Erlbaum.

Thalmann, N. M., & Thalmann, D. (1991). Computer Animation '91. Heidelberg: Springer-Verlag.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. Science, 167, 392-393.

Warren, R. M. (1982). Auditory perception: A new synthesis. New York: Pergamon.

Wenner, A. M., & Wells, P. H. (1990). Anatomy of a Controversy: The Question of a "Language among Bees New York: Columbia University Press, 1990.

Woodworth, R. S. (1938). Experimental psychology. .R New York: Holt.

List of Figures