chensche Verlag. In: Cutler A, Fay D (eds.) John Benjamins, Amsterdam

Myers-Scotton C, Jake J 2001 Explaining aspects of code switching and their implications. In: Nicol J (ed.) *One Mind, Two Languages: Bilingual Language Processing*. Blackwell, Oxford, pp. 91–125

Motley M, Camden C, Baars B 1982 Covert formulation and editing of anomalies in speech production. *Journal of Verbal Learning and Verbal Behavior* **21**: 578–94

Shattuck-Hufnagel S 1987 The role of word onset consonants in speech production planning: New evidence from speech error patterns. In: Kellar E, Gopnik M (eds.) *Motor and Sensory Processing in Language*. Erlbaum Press, Hillsdale, NJ

Stemberger J 1985 An interactive activation model of language production. In: Ellis A (ed.) *Progress in the Psychology of Language*. Erlbaum, London, Vol. 1

Stemberger J, McWhinney B 1986 Form-oriented inflectional errors in language processing. *Cognitive Psychology* **18**: 329–54

Vousden J I, Brown G D A, Harley T A 2000 Serial control of phonology in speech production: a hierarchical model. *Cognitive Psychology* **41**: 101–75

M. Garrett

Speech Perception

It is a commonplace to have the impression that foreign languages are spoken much more rapidly than our own, and without silent periods between the words and sentences. Our own language, however, is perceived at a normal pace (or even too slowly at times) with clear periods of silence between the words and sentences. In fact, languages are spoken at approximately the same rate, and these experienced differences are solely due to the memory structures and psychological processes involved in speech perception. Thus we define speech perception as the process of imposing a meaningful perceptual experience on an otherwise meaningless speech input. The empirical and theoretical investigation of speech perception has blossomed into an active interdisciplinary endeavor, including the fields of psychophysics, neurophysiology, sensory perception, psycholinguistics, linguistics, artificial intelligence, and sociolinguistics.

1. Psychophysics of Speech Perception

In any domain of perception, one goal is to determine the stimulus properties responsible for perception and recognition of the objects in that domain (see *Psychophysics*). The study of speech perception promises to be even more challenging than other domains of perception because there appears to be a discrepancy between the stimulus and the perceiver's experience of it. For speech, we perceive mostly a discrete auditory message composed of words, phrases, and sentences. The stimulus input for this experience, however, is a continuous stream of sound (and facial and gestural movements in face-to-face communication) produced by the speech production process. Somehow, this continuous input is transformed into more or less a meaningful sequence of discrete events (see *Psychophysics*).

Although we have made great progress on this front, there is still active controversy over the ecological properties of the speech input that are actually functional in speech perception. One issue, revived by recent findings, is whether the functional properties in the signal are static or dynamic (changing with time). Traditionally, static cues (such as the location of formants (bands of energy in the acoustic signal related to vocal tract configuration), the distribution of spectral noise as in the onset of *saw* and *shawl*, and the mouth shape at the onset of a segment) have been shown to be effective in influencing speech perception. Dynamic cues such as the transition of energy between a consonant and the following vowel have also been shown to be important. For example, recent research has shown that the second formant (F2) transition defined as the change between the F2 value at the onset of a consonant–vowel (CV) transition and the F2 value in the middle of the following vowel is a reliable predictor of the place of articulation category (Sussman et al. 1998).

Controversy arises when research is carried out to argue for one type of cue rather than another. For example, investigators recently isolated short segments of the speech signal and reversed the order of the speech within each segment (Saberi and Perrott 1999). In this procedure, a sentence is divided into a sequence of successive segments of a fixed duration such as 50 ms. Each segment is time-reversed and these new segments are recombined in their original order, without smoothing the transition borders between the segments. In this fashion, the sentence could be described as between globally contiguous but locally time-reversed. The authors claimed that the speech was still intelligible when the reversed segments were relatively short (1/20th to 15th of a second). Their conclusion was that our perception of speech was demonstrated to be primarily dependent on higher-order dynamic properties rather than the short static cues normally assumed by most current theories. This type of study and logic follows a tradition of attempting to find a single explanation or influence of some psychological phenomenon. However, most successful research in psychology is better framed within the more general framework of *ceteris paribus* (all other aspects neutral). There is good evidence that perceivers exploit many different cues in speech perception, and attempting to isolate a single functionally sufficient cue is futile.

There is now a large body of evidence indicating that multiple sources of information are available to

Massaro, D. W. (2001). Speech Perception. In N.M. Smelser & P.B. Baltes (Eds.) & W. Kintsch (Section Ed.), International Encyclopedia of Social and Behavioral Sciences (pp. 14870-14875). Amsterdam, The Netherlands: Elsevier.

support the perception, identification, and interpretation of spoken language. There is an ideal experimental paradigm that allows us to determine which of the many potentially functional cues are actually used by human observers, and how these cues are combined to achieve speech perception (Massaro 1998). The systematic variation of the properties of the speech signal and quantitative tests of models of speech perception allow the investigator to interpret the psychological validity of different cues. This paradigm has already proven to be effective in the study of audible, visible, and bimodal speech perception (Massaro 1998). Thus, this research strategy addresses how different sources of information are evaluated and integrated, and can identify the sources of information that are actually used.

2. *The Demise of Categorical Perception*

Too often, behavioral scientists allow their phenomenal impressions to spill over into their theoretical constructs. One experience in speech perception is that of categorical perception. Using synthetic speech, it is possible to make a continuum of different sounds varying in small steps between two alternatives. Listening to a synthetic speech continuum between /ba/ and /pa/ provides an impressive demonstration: students and colleagues usually agree that their percept changes qualitatively from one category to the other in a single step or two with very little fuzziness in-between. Consistent with this impression, there is no denying that we experience discrete categories in speech perception. This experience of discrete perception was developed as a description of the speech perception process and called categorical perception (CP), or the perceived equality of instances within a category. The CP of phonemes has been a central concept in the experimental and theoretical investigation of speech perception and has also spilled over into other domains such as face processing (Beale and Keil 1995, Etcoff and Mcgee 1992). CP was operationalized in terms of discrimination performance being limited by identification performance. Researchers at Haskins Laboratories (Liberman et al. 1957) used synthetic speech to generate a series of 14 consonant–vowel syllables going from /be/ to /de/ to /ge/ (/e/ as in gate). The onset frequency of the second formant transition of the initial consonant was changed in equal steps to produce the continuum. In the identification task, observers identified random presentations of the sounds as /b/, /d/, or /g/. The discrimination task used the ABX paradigm. Three stimuli were presented in the order ABX; A and B always differed and X was identical to either A or B. Observers were instructed to indicate whether X was equal to A or B. This judgment was supposedly based on auditory discrimination in that observers were instructed to use whatever auditory differences they could perceive.

The experiment was designed to test the hypothesis that listeners can discriminate the syllables only to the extent that they can recognize them as different phoneme categories. The CP hypothesis was quantified in order to predict discrimination performance from the identification judgments. The authors concluded that discrimination performance was fairly well predicted by identification. This rough correspondence between identification and discrimination has provided the major source of support for CP. There have been several notable limitations in these studies. Discrimination performance is consistently better than that predicted by identification, it is possible that participants are making their discrimination judgments on the basis of identification rather than on their auditory discrimination, and it is likely that any alternative theory would have described the results equally well (Massaro 1987).

In many areas of inquiry, a new experimental paradigm enlightens our understanding by helping to resolve theoretical controversies. For CP, rating experiments were used to determine if perceivers indeed have information about the degree of category membership. Rather than ask for categorical decisions, perceivers are asked to rate the stimulus along a continuum between two categories. A detailed quantitative analysis of the results indicated that perceivers have reliable information about the degree of category membership, contrary to the tenets of CP (Massaro and Cohen 1983). Although communication forces us to partition the inputs into discrete categories for understanding, this property in no way implies that speech perception is categorical. To retrieve a toy upon request, a child might have to decide between *ball* and *doll*; however, she can certainly have information about the degree to which each toy was requested.

Although CP has been discredited, it is often reinvented under new guises. Most recently, the perceptual-magnet effect (PME) has had a tremendous impact on the field, and has generated a great deal of research (Kuhl 1991). The critical idea is that the discriminability of a speech segment is inversely related to its category goodness. Ideal instances of a category are supposedly very difficult to distinguish from one another relative to poor instances of the category. If we understand that poor instances of one category will often tend to be at the boundary between two categories, then the PME is more or less a reformulation of prototypical CP. That is, discrimination is predicted to be more accurate between categories than within categories. In demonstrating its viability, the PME faces the same barriers that have been difficult to eliminate in CP research. In standard CP research, it is necessary to show how discrimination is directly predicted by identification performance. In the PME framework, it is also necessary to show how discrimination is directly predicted by a measure of category goodness. We can expect category goodness to be related to identification performance. Good category instances will tend to be identified equivalently, whereas poor instances will likely be identified as instances of different categories. Lotto et al. (1998) found that discriminability was *not* poorer for vowels with high category goodness, in contrast to the predictions of the PME. The authors also observed that category goodness ratings were highly context sensitive, which is a problem for the PME. If category goodness is functional in discrimination, it should be relatively stable across different contexts.

Notwithstanding the three decades of misinterpreting the relationship between identification and discrimination of auditory speech, we must conclude that it is perceived continuously and not categorically. Recent research reveals conclusively that both visible and bimodal speech are perceived continuously (Massaro 1987, 1998). This observation pulls the carpet from under current views of language acquisition that attribute to the infant and child discrete speech categories (Eimas 1985, Gleitman and Wanner 1982). Most importantly, the case for the specialization of speech is weakened considerably because of the central role that the assumption of CP has played (Liberman and Mattingly 1985). Finally, several neural network theories such as single-layer perceptrons, recurrent network models, and interactive activation have been developed to predict CP (Damper 1994): its nonexistence poses great problems for these models.

Given the existence of multiple sources of information in speech perception, each perceived continuously, a new type of theory is needed. The theory must describe how each of the many sources of information is evaluated, how the many sources are combined or integrated, and how decisions are made. The development of a promising theory has evolved from sophisticated experimental designs and quantitative model testing to understand speech perception and pattern recognition more generally. A wide variety of results have been accurately described within the Fuzzy Logical Model of Perception (FLMP).

3. The Fuzzy Logical Model of Perception (FLMP)

The three processes involved in perceptual recognition are illustrated in Fig. 1 and include evaluation, integration, and decision. These processes make use of prototypes stored in long-term memory. The evaluation process transforms these sources of information into psychological values, which are then integrated to give an overall degree of support for each speech alternative. The decision operation maps the outputs of integration into some response alternative. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. The assumptions central to the model are: (a) each source of information is evaluated to determine the continuous degree to which that source specifies various alternatives; (b) the sources of information are evaluated independently of one another; (c) the sources are integrated to provide an overall continuous degree of support for each alternative; and (d) perceptual identification and interpretation follows the relative degree of support among the alternatives. In the course of our research, we have found the FLMP to be a universal principle of perceptual cognitive performance that accurately models human pattern recognition. People are influenced by multiple sources of information in a diverse set of situations. In many cases, these sources of information are ambiguous and any particular source alone does not usually specify completely the appropriate interpretation.

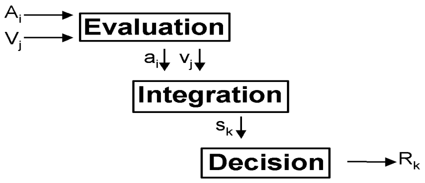

Figure 1
Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by upper-case letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j) These sources are then integrated to give an overall degree of support, s_k, for each speech alternative k. The decision operation maps the outputs of integration into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely

4. Multimodal Speech Perception

Speech perception has traditionally been viewed as a unimodal process, but in fact appears to be a prototypical case of multimodal perception. This is best seen in face-to-face communication. Experiments have revealed conclusively that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech (Massaro 1998). Consider a simple syllable identification task. Synthetic visible speech and natural audible speech were used to generate the consonant–vowel (CV) syllables /ba/, /va/, /a/, and /da/. Using an expanded factorial design, the four syllables

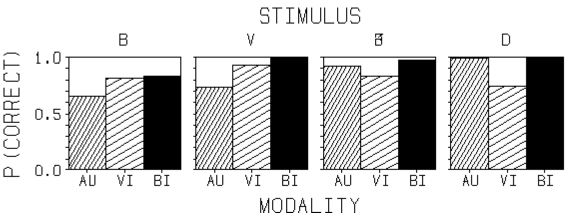

Figure 2
Probability correct identification of the unimodal (AU = Auditory, VI = Visual) and bimodal (BI) consistent trials for the four test syllables

were presented auditorily, visually, and bimodally. Each syllable was presented alone in each modality for $4\times2 = 8$ unimodal trials. For the bimodal presentation, each audible syllable was presented with each visible syllable for a total of 4×4 or 16 unique trials. Thus, there were 24 types of trials. Twelve of the bimodal syllables had inconsistent auditory and visual information. The 20 participants in the experiment were instructed to watch and listen to the talking head and to indicate the syllable that was spoken.

Accuracy is given in Fig. 2 for unimodal and bimodal trials when the two syllables are consistent with one another. Performance was more accurate given two consistent sources of information than given either one presented alone. Consistent auditory information improved visual performance about as much as consistent visual information improved auditory performance. Given inconsistent information from the two sources, performance was poorer than observed in the unimodal conditions. These results show a large influence of both modalities on performance with a larger influence from the auditory than the visual source of information.

Although the results demonstrate that perceivers use both auditory and visible speech in perception, they do not indicate how the two sources are used together. There are many possible ways the two sources might be used. We first consider the predictions of the FLMP.

In a two-alternative task with /ba/ and /da/ alternatives, the degree of auditory support for /da/ can be represented by a_i, and the support for /ba/ by ($1–a_i$). Similarly, the degree of visual support for /da/ can be represented by v_j, and the support for /ba/ by ($1–v_j$). The probability of a response to the unimodal stimulus is simply equal to the feature value. For bimodal trials, the predicted probability of a response, $P(/\mathrm{da}/)$ is equal to

$$P(/\mathrm{da}/) = \frac{a_i v_j}{a_i v_j + (1-a_i)(1-v_j)} \qquad (1)$$

In previous work, the FLMP has been contrasted against several alternative models such as a weighted averaging model (WTAV), which is an inefficient algorithm for combining the auditory and visual sources. For bimodal trials, the predicted probability of a response, $P(/\mathrm{da}/)$ is equal to

$$P(/\mathrm{da}/) = \frac{w_1 a_1 + w_2 v_j}{w_1 + w_2} = wa_i + (1-w)v_j \qquad (2)$$

The WTAV predicts that two sources can never be more informative than one. In direct contrasts, the FLMP has consistently and significantly outperformed the WTAV (Massaro 1998).

More generally, research has shown that the results are well-described by the FLMP, an optimal integration of the two sources of information (Massaro and Stork 1998). A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. For example, if the ambiguous auditory sentence, *My bab pop me poo brive*, is paired with the visible sentence, *My gag kok me koo grive*, the perceiver is likely to hear, *My dad taught me to drive*. Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro and Stork 1998).

Recent findings show that speech reading, or the ability to obtain speech information from the face, is not compromised by oblique views, partial obstruction or visual distance. Humans are fairly good at speech reading even if they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Massaro 1998).

5. *Higher-order or Top-down Influences*

Consistent with the framework being developed, there is now a substantial body of research illustrating that speech perception is influenced by a variety of contextual sources of information. Bottom-up sources correspond to those sources that have a direct mapping between the sensory input and the representational unit in question. Top-down sources or contextual information come from constraints that are not directly mapped onto the unit in question. As an example, a bottom-up source would be the stimulus presentation of a test word after the presentation of a top-down source, a sentence context. A critical question for both integration and autonomous (modularity, models is how bottom-up and top-down sources of information work together to achieve word recognition. For example, an important question is how early can contextual information be integrated with acoustic-phonetic information. A large body of research shows that several bottom-up sources are

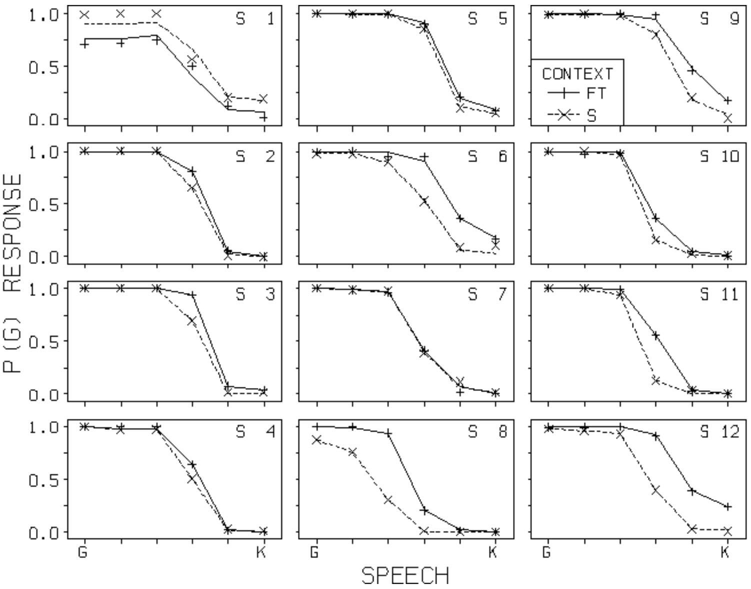

Figure 3
Observed (points) and FLMP's predictions (lines) of /g/ identifications for FT and S contexts as a function of the speech information of the initial consonant. Results from Pitt's (1995) Experiment 3a

evaluated in parallel and integrated to achieve recognition (Massaro 1987, 1994). An important question is whether top-down and bottom-up sources are processed in the same *manner*. A critical characteristic of autonomous models might be described as the language user's *inability* to integrate bottom-up and top-down information. An autonomous model must necessarily predict no perceptual integration of top-down with bottom-up information.

Pitt (1995) studied the joint influence of phonological information and lexical context in an experimental paradigm developed by Ganong (1980). A speech continuum is made between two alternatives, and the contextual information supports one alternative or the other. The initial consonant of the CVC syllable was varied in six steps between /g/ and /k/. The following context was either /Ift/ or /Is/. The context /Ift/ favors or supports initial /g/ because *gift* is a word whereas *kift* is not. Similarly, the context /Is/ favors or supports initial /k/ because *kiss* is a word whereas *giss* is not. Pitt improved on earlier studies by collecting enough observations to allow us to perform a subject-by-subject evaluation of the ability of specific models of language processing to account for the results. Previous tests of models using this task have been primarily dependent on group averages which may not be representative of the individuals that make the averages up.

The points in Fig. 3 give the observed results for each of the 12 subjects in the task. For most of the subjects, the individual results tend to resemble the average results reported by Pitt and earlier investigators. Ten of the 12 subjects were influenced by lexical context in the appropriate direction. Subject 1 gave an inverse context effect and Subject 7 was not influenced by context.

This FLMP was applied to the identification results of the 12 individual subjects in Pitt's Experiment 3a for which the greatest number of observations (104) were obtained for each data point for each subject. According to the FLMP, both the bottom-up information from the initial speech segment and the top-down context are evaluated and integrated. If s_i is the degree of support for the voiced alternative given by the initial segment and c_j is the support given by the following context, the total support for the voiced alternative i is given by

$$S(\text{voiced} \mid S_iC_j) = s_i \times c_j \qquad (3)$$

The support for the voiceless alternative would be

$$S(\text{voiceless} \mid S_iC_j) = (1 - s_i) \times (1 - c_j) \qquad (4)$$

The predicted probability of a voiced response is simply

$$P(\text{voiced} \mid S_iC_j) = \frac{S(\text{voiced} \mid S_iC_j)}{S(\text{voiced} \mid S_iC_j) + S(\text{voiceless} \mid S_iC_j)} \qquad (5)$$

In producing predictions for the FLMP, it is necessary to estimate parameter values for each level of each experimental factor. The initial consonant was varied along six steps between /g/ and /k/ and the following context was either /Ift/ or /Is/. Thus, there were six levels of bottom-up phonological information s_i and two contexts c_j. A free parameter is necessary for each level of bottom-up information but it is reasonable to assume that the contextual support given by /Is/ is one minus the lexical support given by /Ift/ so that only one value of c_j needs to be estimated. Thus, seven free parameters are used to predict the 12 independent points.

The lines in Fig. 3 also give the predictions of the FLMP. As can be seen in the figure, the model generally provides a good description of the results of this study. The root mean squared deviation (RMSD) between predicted and obtained is 0.017 on the average across all 12 independent fits. For the 10 subjects showing appropriate context effects, the RMSD ranges from 0.003 to 0.045 with a median of 0.007. Thus, for each of these individuals, the model captures the observed interaction between phonological information and lexical context: the effect of context was greater to the extent that the phonological information was ambiguous. This yields a pattern of curves in the shape of an American football, which is a trademark of the FLMP.

The model tests have established that perceivers integrate top-down and bottom-up information in language processing, as described by the FLMP. This result means that sensory information and context are integrated in the same manner as several sources of bottom-up information. These results pose problems for autonomous models of language processing.

See also: Audition and Hearing; Speech Production, Neural Basis of; Speech Production, Psychology of; Speech Recognition and Production by Machines

Bibliography

Beale J M, Keil F C 1995 Categorical effects in the perception of faces. *Cognition* **57**(3): 217–39

Damper R I 1994 Connectionist models of categorical perception of speech. *Proceedings of the IEEE International Symposium on Speech, Image Processing and Neural Networks*, Vol. 1, pp. 101–4

Etcoff N L, McGee J J 1992 Categorical perception of facial expressions. *Cognition* **44**: 227–40

Eimas P D 1985 The perception of speech in early infancy. *Scientific American* **252**: 46–52

Fitch H L, Halwes T, Erickson D M, Liberman A M 1980 Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics* **27**: 343–50

Ganong III W F 1980 Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance* **6**: 110–25

Gleitman L R, Wanner E 1982 Language acquisition: The state of the state of the art. In: Wanner E, Gleitman L R (eds.) *Language Acquisition: The State of the Art*. Cambridge University Press, Cambridge, UK

Kuhl P K 1991 Human adults and human infants show a ‘perceptual magnet effect’ for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics* **50**: 93–107

Liberman A M, Harris K S, Hoffman H S, Griffith B C 1957 The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* **54**: 358–68 , 753–71

Liberman A M, Mattingly I G 1985 The motor theory of speech perception revised. *Cognition* **21**: 1–36

Lotto A J, Kluender K R, Holt L L 1998 Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America* **103**: 3648–55

Massaro D W 1987 *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Erlbaum Associates, Hillsdale, NJ

Massaro D W 1994 Psychological aspects of speech perception: Implications for research and theory. In: Gemsbacher M (eds.) *Handbook of Psycholingristics*. Acadamie Press, New York pp. 219–63

Massaro D W 1998 *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press, Cambridge, MA

Massaro D W, Cohen M M 1983 Categorical or continuous speech perception: A new test. *Speech Communication* **2**: 15–35

Massaro D W, Stork D G 1998 Sensory integration and speechreading by humans and machines. *American Scientist* **86**: 236–44

Pitt M A 1995 The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition* **21**: 1037–52

Saberi K, Perrott D R 1999 Cognitive restoration of reversed speech. *Nature* **398**: 760

Sussman H M, Fruchter D, Hilbert J, Sirosh J 1998 Linear correlates in the speech signal: The orderly output constraint. *Behavioral & Brain Sciences* **21**(2): 241–99

D. W. Massaro

Speech Production, Neural Basis of

1. *Brain Regions Involved in Speech Production*

Over a century ago, the French neurologist Paul Broca demonstrated that speech mechanisms could be localized in the human brain. He did this by interviewing a patient with a severe speech production disorder with output limited to the recurring utterance, ‘*tan*.’ Upon the patient’s death, Broca examined the brain and concluded that the patient’s inability to speak was due to a lesion in the inferior part of the frontal lobe (Broca 1861b). A second patient also had severely reduced speech and was subsequently found to have a similar cortical lesion (Broca 1861a). Since the time of Broca, scientists have found that lesions to Broca’s area alone are not enough to produce lasting speech deficits (e.g., Alexander et al., 1989; Dronkers et al. 2000; Mohr 1976). Many have attempted to diagram

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7