DOMINIC W. MASSARO

MULTIMODAL SPEECH PERCEPTION: A PARADIGM FOR SPEECH SCIENCE

1. INTRODUCTION

Speech science evolved as the study of a unimodal phenomenon. Speech was viewed as a solely auditory event, as captured by the seminal speech-chain illustration of Denes & Pinson (1963) shown in Figure 1.

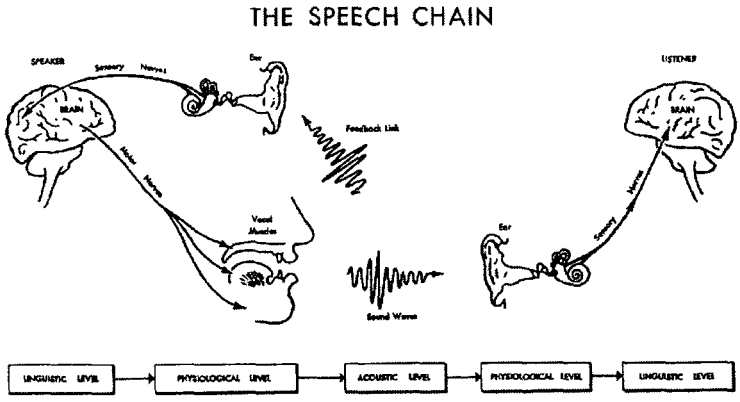

Figure 1. The classic speech-chain illustration of Denes & Pinson (1963).

This view is no longer viable as witnessed by this book as well as a burgeoning record of research findings. Although Denes & Pinson viewed speech as primarily an auditory phenomenon (rather than a multimodal one), they did acknowledge the important contribution of context to accurate recognition and understanding. In accepting the influence of both stimulus information and context on speech perception, the authors anticipated the approach taken in the present chapter. They stated,

> *"In speech communication, then, we do not actually rely on a precise knowledge of specific cues. Instead, we related a great variety of ambiguous cues against the background of the complex system we call our common language."* (Denes & Pinson, 1963, p. 8).

B. Granström et al. (eds.), Multimodality in Language and Speech Systems, 45–71.

Speech as a multimodal phenomenon is supported by experiments indicating 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, 1987, 1998). Many communication environments involve a noisy auditory channel, which degrades speech perception and recognition. Visible speech from the talker's face (or from a reasonably accurate synthetic talking head) improves intelligibility in these situations. Visible speech also is an important communication channel for individuals with hearing loss.

The number of words understood from a degraded auditory message can often be doubled by pairing the message with visible speech from the talker's face. The combination of auditory and visual speech has been called super-additive because their combination can lead to accuracy that is much greater than accuracy on either modality alone. Furthermore, the strong influence of visible speech is not limited to situations with degraded auditory input. 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, 1998).

There are several reasons why the use of auditory and visual information together is so successful. These include (a) robustness of visual speech, (b) complementarity of auditory and visual speech, and (c) optimal integration of these two sources of information. Speechreading, or the ability to obtain speech information from the face, is robust in that perceivers are fairly good at speech reading even when 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).

Complementarity of auditory and visual information simply means that one of the sources is strong when the other is weak. A distinction between two segments robustly conveyed in one modality is relatively ambiguous in the other modality. For example, the place difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the voicing difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. Two complementary sources of information make their combined use much more informative than would be the case if the two sources were non-complementary, or redundant (Massaro, 1998, pp. 424-427).

The final characteristic is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner. There are many possible ways to treat two sources of information: use only the most informative source, average the two sources together, or integrate them in such a fashion in which both sources are used but that the least ambiguous source has the most influence. Perceivers in fact integrate the information available from each modality to perform as efficiently as possible. A wide variety of empirical results has been accurately predicted by a model that describes an optimally efficient process of combination.

In this chapter, I will analyze the multimodality of spoken language understanding within an information-processing framework. After describing the framework, a specific theoretical model is described to help organize the descriptions of experiments and theories. Several alternative theories are then presented and evaluated. To test among the theories, we discuss how the theories account for the influence of multiple sources of stimulus information in speech perception. To structure our information-processing analysis of spoken language understanding, we use a specific theoretical framework that has received substantial support from a variety of experiments in speech perception.

2. THEORETICAL FRAMEWORK

The general theoretical framework provided by the information-processing approach is based on the assumption that there is a sequence of processing stages in spoken language understanding. Stages of information processing have guided, for example, much of the research in visual perception (Palmer, 1999). Visual perception is assumed to occur in three stages of processing: retinal transduction, sensory cues (features), and perceived attributes (DeYoe & Van Essen, 1988). Visual input is transduced by the visual system, a conglomeration of sensory cues is made available, and attributes of the visual world are experienced by the perceiver. In visual perception, there is both a one-to-many and a many-to-one relationship between sensory cues and perceived attributes. The sensory cue of motion provides information about both perceived shape of an object and its perceived movement. A case of the many-to-one relationship in vision is that information about the shape of an object is enriched not only by motion, but also by perspective cues, picture cues, binocular disparity, and shading (e.g., chicariscuro).

We apply this same framework to speech perception and spoken language understanding. Speech perception via the auditory modality is characterized by a transduction of the acoustic signal along the basilar membrane, sensory cues, and perceived attributes. A single sensory cue can influence several perceived attributes. The duration of a vowel provides information about vowel identity (bit vs. beet), information such as lexical stress (the noun and verb pronunciations of the word *permit*), and syntactic boundaries in sentences. Another example is that the pitch of a speaker's voice is informative about both the identity of the speaker and intonation. The best-known example of multiple cues to a single perceived attribute in speech is the case of the many cues for the voicing of a medial stop consonant (Cohen, 1979; Lisker, 1978). These include the duration of the preceding vowel, the onset frequency of the fundamental, the voice onset time, and the silent closure interval. A multimodal example is the impressive demonstration that both the speech sound and the visible mouth movements of the speaker influence perception of place of articulation of a stop consonant (Massaro & Cohen, 1983; McGurk & MacDonald, 1976).

Our research and that of many others has demonstrated a powerful influence of visible speech in face-to-face communication. The influence of several sources of information from several modalities provides a new challenge for theoretical

accounts of speech perception. For theories that were developed to account for the perception of unimodal auditory speech (Diehl & Kluender, 1987, 1989), it is not obvious how they would account for the positive contribution of visible speech. Some extant theories view speech perception as a specialized process and not solely as an instance of pattern recognition (Liberman & Mattingly, 1985; Mattingly & Studdert-Kennedy, 1991). We take a different approach by envisioning speech perception as an instance of a more general process of pattern recognition (Massaro, 1998). In language processing, recognition is achieved via a variety of bottom-up and top-down sources of information. Top-down sources include contextual, semantic, syntactic, and phonological constraints; bottom-up sources include audible and visible features of the spoken word. A top-down source might be the overall frequency of a speech segment in the perceiver's language. A bottom-up source might be the degree of jaw rotation while talking.

3. THEORETICAL/EMPIRICAL INQUIRY

Our general framework documents the value of a combined experimental/theoretical approach. The research has contributed to our understanding of the characteristics used in speech perception, how speech is perceived and recognized, and the fundamental psychological processes that occur in speech perception and in pattern recognition in a variety of other domains.

We evaluate the contribution of visible information in face-to-face communication and how it is combined with auditory information in the ecologically valid condition of bimodal speech perception (face-to-face communication). Psychophysical and pattern-recognition tasks are carried out to analyze which audible and visible features are used by human observers in auditory, visual, and auditory-visual (bimodal) speech perception. Quantitative models of feature evaluation and integration are tested against identification judgments, ratings, and confusion matrices from perceptual tests. The results are used to determine which features influence performance.

The results are also to test formal models of speech perception. The models are formalized to make quantitative predictions of the judgments of the test items. Multiple models are tested to preclude a confirmation bias and to adhere to a falsification strategy of inquiry (Massaro, 1989, chapter 5). Each model is tested against the results of single subjects in order to avoid the pitfalls of averaging results across subjects. We also test a variety of participants to explore a broad variety of dimensions of individual variability. These include (1) life-span variability, (2) language variability, (3) sensory impairment, (4) brain trauma, (5) personality, (6) sex differences, and (7) experience and learning. In addition, a large variety of experimental procedures and test situations are used in our investigations (Massaro, 1998, Chapter 6). Generally, we need to know to what extent the processes uncovered in our research generalize across (1) sensory modalities, (2) environmental domains, (3) test items, (4) behavioural measures, (5) instructions, (6) and tasks.

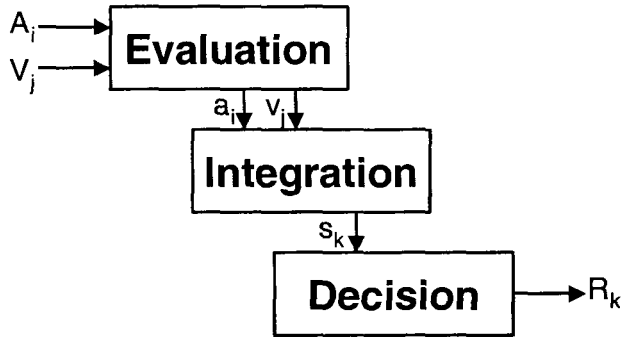

Figure 2. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by uppercase 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.

We believe that our empirical work would be inadequate and perhaps invalid without the corresponding theoretical framework. Thus, the research addresses both empirical and theoretical issues. At the empirical level, experiments are carried out to determine how visible speech is combined with auditory speech for a broad range of individuals and across a wide variation of situational domains. At the theoretical level, the assumptions and predictions of several models are formalized, analyzed, contrasted, and tested. Various types of model fitting strategies have been employed, with similar outcomes. These model tests have been highly informative with respect to improving our understanding of how spoken language is perceived and understood.

4. FUZZY LOGICAL MODEL OF PERCEPTION

We have learned that a variety of empirical results can be successfully described within a framework of a fuzzy logical model of perception (FLMP). The FLMP assumes necessarily successive but overlapping stages of processing, as shown in Figure 2. The perceiver of speech is viewed as having multiple sources of information supporting the identification and interpretation of the language input. The model assumes that (1) each source of information is evaluated to give the continuous degree to which that source supports various alternatives, (2) the sources of information are evaluated independently of one another, (3) the sources are