Further Reading

Borden GJ, Harris KS and Raphael LJ (1994) *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*, 3rd edn. Baltimore, MD: Williams & Wilkins.

Denes PD and Pinson EN (1993) *The Speech Chain: The Physics and Biology of Spoken Language*, 2nd edn. New York, NY: WH Freeman.

Fowler CA (1986) An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics* **14**: 3–28.

Fowler CA (1996) Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America* **99**: 1730–1741.

Johnson K (1997) *Acoustic and Auditory Phonetics*. Cambridge, MA: Blackwell.

Kent RD (1997) *The Speech Sciences*. San Diego, CA: Singular Publishing Group.

Ladefoged P (1993) *A Course in Phonetics*, 3rd edn. Fort Worth, TX: Harcourt Brace.

Lass NJ (1996) *Principles of Experimental Phonetics*. St Louis, MO: Mosby.

Liberman AM (1996) *Speech: A Special Code*. Cambridge, MA: MIT Press.

Pisoni DB (1997) Some thoughts on 'normalization' in speech perception. In: Johnson K and Mullennix JW (eds) *Talker Variability in Speech Processing*. San Diego, CA: Academic Press.

In Encyclopedia of Cognitive Science, Editor-in-chief, Lynn Nadel London; New York: Nature Publishing Group 2003, Vols. 1-4.

Speech Perception and Recognition, Theories and Models of

Intermediate article

Dominic W Massaro, University of California, Santa Cruz, California, USA

CONTENTS

Multiple sources of information are exploited to perceive and understand spoken language.

INTRODUCTION

It is not unusual 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, on the other hand, 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 perceptual and 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 become an active interdisciplinary endeavor, including the fields of psychophysics, neurophysiology, sensory perception, psycholinguistics, linguistics, artificial intelligence, and sociolinguistics.

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. The study of speech perception seems 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) resulting from the speech production. Somehow, this continuous input is transformed into a more or less meaningful sequence of discrete events.

One long-standing issue in research has been whether there is an invariance between the speech signal and its category membership. One of the most obvious perceptual categories for speech is the phoneme. Phonemes are the minimal units in speech that can change the meaning of a word. The word *ten* has three phonemes: we can change the /t/ to /d/ to make *den*, the /e/ to /æ/ to make *tan*, and the /n/ to /l/ to make *tell*. If phonemes were invariant perceptual categories, we would expect to find an orderly relationship between properties of the speech signal and phoneme categories. We would expect to find some constant characteristic in the speech signal for a given phoneme. However, this appears not to be the case. Figure 1 gives a visual representation of the sounds /di/ and /du/. Given that /d/ is the first phoneme of both sounds, we might expect to see the same signal at the beginning. However, we do not. The second

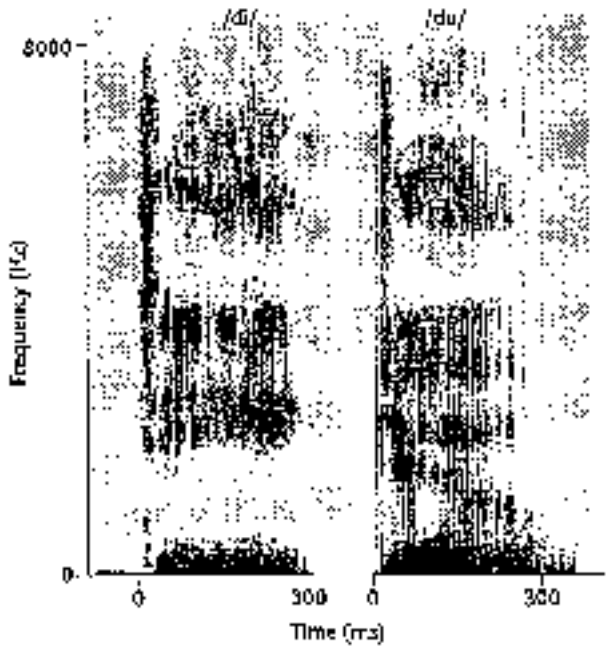

Figure 1. Spectrograms of the syllables /di/ and /du/, illustrating the lack of invariance between the acoustic signal and the phoneme. The second visible band of energy from the bottom (the second formant) rises in /di/ and falls in /du/, illustrating that the same phoneme /d/ has different acoustic characteristics in the different vowel contexts.

visible band of energy from the bottom (the second formant) rises in /di/ and falls in /du/. One of the original arguments for the special 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.

Not only is there a lack of invariance between the phoneme and the speech signal, but it does not at first seem to be possible to isolate the phoneme in the speech signal. Consider the syllable /da/. It has two phonemes, /d/ and /a/. If we listen to this syllable in isolation, we hear /da/. Now if we repeatedly shorten the syllable by removing short segments from the end, we should eventually hear just /d/. But this is not what happens. 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/.

One way to instate some systematicity between the speech signal and perception is to postulate a somewhat larger unit of speech perception. Open syllables are defined as V, VC or CV items, where V is a vowel and C can be either a single consonant or a consonant cluster. There is evidence that these items might function as units of perception. Subjects can easily identify shortened versions of V, VC and CV syllables when most of the vowel portion is eliminated. Although there are clearly contextual effects on the signal properties of these syllables, the influences are much more minor than those on the phoneme. Some support for the relative invariance of the open syllable comes from concatenative speech synthesis systems that use diphones (pairs of adjacent phonemes) rather than phonemes as units that are concatenated. Based on psychological research, synthesis would be even better if units for synthesis were open syllables.

Regardless of the units that are used for perception, there is still 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). 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 *sew* and *shoe*, 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 difference between the F2 value at the onset of a consonant–vowel transition and the F2 value in the middle of the following vowel, is a reliable predictor of the category describing place of articulation (Sussman *et al.*, 1998).

Controversy arises over the type of cue on which speech perception primarily depends. 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 transitions between the segments. Thus, the sentence could be described as locally time-reversed. Saberi and Perrott claimed that the speech was still intelligible when the reversed segments were relatively short (about 50 to 65 ms). Their conclusion was that our perception of speech was primarily dependent on higher-order dynamic properties rather than the short static cues assumed by most current theories. However, most successful research in psychology is better framed within the framework of *ceteris paribus* (other things being equal). 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 support the perception, identification, and interpretation of spoken language. Auditory and visual cues from the speaker, as well as the situational, social, and linguistic context contribute to understanding. There is an ideal experimental paradigm that allows us to determine which subset 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 experiment involves the independent manipulation of two or more sources of information in a factorial or expanded factorial design. This 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). It addresses how different sources of information are evaluated and integrated, and can identify the sources of information that are actually used.

AMBIGUITY IN SPEECH PERCEPTION

When listening to speech, we have the impression of perceiving discrete categories. This fact has contributed to the development of the popular 'categorical perception' hypothesis, that listeners can discriminate syllables only to the extent that they can recognize them as different phoneme categories. This hypothesis was quantified in order to predict discrimination performance from the identification judgements. Many researchers concluded that discrimination performance was fairly well predicted by identification. However, discrimination performance is consistently better than predicted by identification, contrary to the predictions of the categorical perception hypothesis (Massaro, 1987).

In many areas of inquiry, a new experimental paradigm enlightens our understanding by helping to resolve theoretical controversies. 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 (Massaro and Cohen, 1983). Although communication forces us to partition the inputs into discrete categories for understanding, this does not imply that speech perception is categorical. To retrieve a toy upon request, a child might have to decide between *ball* and *doll*; however, the child can have information about the degree to which each toy was requested.

Although the categorical perception hypothesis has been refuted, it is often reinvented under new guises. Recently, the 'perceptual magnet effect' hypothesis has generated a great deal of research (Kuhl, 1991; Iverson and Kuhl, 2000). The 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 perceptual magnet effect hypothesis is more or less a reformulation of the categorical perception hypothesis. That is, discrimination is predicted to be more accurate between categories than within categories. In the perceptual magnet effect 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 tend to 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 perceptual magnet effect hypothesis. They also observed that category goodness ratings were highly context-sensitive, because they changed systematically with changes in the task context. This reliable context-sensitivity is a problem for the perceptual magnet effect hypothesis. If category goodness is functional in discrimination, it should be reasonably stable across different contexts.

It was once commonly believed that speech is perceived categorically. Current research suggests, however, that speech is perceived continuously and not categorically (Massaro, 1987, 1998). These results challenge 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 special nature of speech perception is weakened considerably, because of its dependence on the assumption of categorical perception (Liberman and Mattingly, 1985). Several neural network theories, such as single-layer perceptrons, recurrent network models, and interactive activation, have been developed to predict categorical perception (Damper, 1994; Damper and Harnad, 2000), and its probable 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. 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).

THE FUZZY-LOGICAL MODEL OF PERCEPTION

The three processes involved in perceptual recognition are illustrated in Figure 2. They are 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,

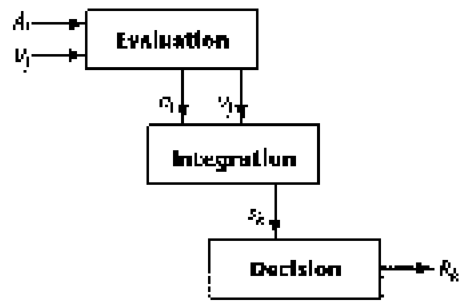

Figure 2. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown proceeding from left to right (in time) to illustrate their successive but overlapping operation. These processes make use of prototypes stored in long-term memory. Sources of information 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 lower-case 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.

which are then integrated to give an overall degree of support for each speech alternative. The decision operation maps the outputs of integration to 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: (1) each source of information is evaluated to determine the continuous degree to which that source specifies various alternatives; (2) the sources of information are evaluated independently of one another; (3) the sources are integrated to provide an overall continuous degree of support for each alternative; and (4) perceptual identification and interpretation follows the relative degree of support among the alternatives. The FLMP appears 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 no one source specifies completely the appropriate interpretation.

MULTIMODAL SPEECH PERCEPTION

Speech perception has traditionally been viewed as a unimodal process, but in fact appears to be multimodal. This is best seen in face-to-face

communication. Experiments have shown conclusively that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as by 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 syllables /ba/, /va/, /da/, and /da/. Using an expanded factorial design, the four syllables were presented audibly, visibly, and bimodally. Each syllable was presented alone in each modality for $4 \times 2 = 8$ unimodal trials. For the bimodal presentation, each audible syllable was presented with each visible syllable for a total of $4 \times 4 = 16$ unique trials. Thus, there were 24 types of trial. Of the bimodal syllables, 12 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.

Figure 3 shows the accuracy for unimodal and bimodal trials when the two syllables were 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 strong influence of both modalities on performance, with a stronger influence from the auditory than from the visual source of information.

Although the results demonstrate that perceivers use both auditory and visible cues in speech 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 $P(/da/)$ is

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

In previous work, the FLMP has been contrasted against several alternative models, including 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 $P(/da/)$ is

$$P(/da/) = \frac{w_1 a_i + w_2 v_j}{w_1 + w_2} = w a_i + (1 - w) v_j \tag{2}$$

where W_1 and W_2 are the weights and $W = \frac{W_1}{W_1 + W_2}$.

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

Furthermore, 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 bib pop me poo brize* is paired with the visible sentence *my gag kak me koo grive* the perceiver is likely to hear *my did taught tee to drive*. Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro and Stork, 1998).

Recent findings show that speechreading, or the ability to obtain speech information from the face,

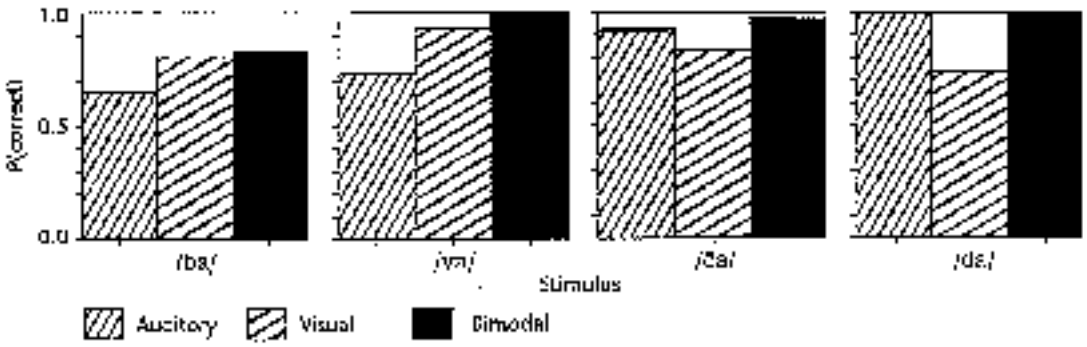

Figure 3. Probabilities of correct identification of syllables in unimodal (auditory and visual) and bimodal consistent trials for the four test syllables /ba/, /va/, /da/ and /da/.

is not compromised by oblique views, partial obstruction or visual distance. Humans are fairly good at speechreading 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).

CONTEXTUAL, HIGHER-ORDER, OR TOP-DOWN INFLUENCES

There is now a substantial body of research showing that speech perception is influenced by a variety of contextual sources of information. Bottom-up sources 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 to the unit in question. An example of 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 contextual information can be integrated with acoustic and phonetic information. A large body of research shows that several bottom-up sources are evaluated in parallel and integrated to achieve recognition (Massaro, 1987, 1994). Another 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 a CVC syllable was varied, taking six values between /g/ and /k/ inclusive. The following context was either /ift/ or /is/. The context /ift/ supports initial /g/ because *gift* is a word whereas *kift* is not. Similarly, the context /is/ supports initial /k/ because *kiss* is a word whereas *giss* is not.

The lines in Figure 4 give the predictions of the FLMP. The model generally provides a good description of the results of this study. 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.

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

CONNECTIONIST MODELS

In connectionist models, information processing occurs through excitatory and inhibitory interactions among a large number of simple processing units. These units are meant to represent the functional properties of neurons or neural networks. Three levels or sizes of units are used in the TRACE model of speech perception (McClelland and Elman, 1986; McClelland, 1991): feature, phoneme, and word. Features activate phonemes which activate words, and activation of some units at a given level inhibits other units at the same level. In addition, an important assumption of interactive activation models is that activation of higher-order units activates their lower-order units; for example, activation of the /b/ phoneme would activate the features that are consistent with that phoneme.

In the TRACE model, word recognition is mediated by feature and phoneme recognition. The input is processed online in TRACE, all words are activated by the input in parallel, and their activation is context-dependent. In principle, TRACE is continuous, but its assumption about interactive activation leads to categorical-like behavior at the sensory (featural) level. In the TRACE model, a stimulus pattern is presented and activation of the corresponding features sends more excitation to some phoneme units than others. Given the assumption of feedback from the phoneme to the feature level, the activation of a particular phoneme feeds back and activates the features corresponding to that phoneme (McClelland and Elman, 1986, p. 47). This effect enhances sensitivity around category boundaries, exactly as predicted by the categorical perception hypothesis. Evidence against phonemes as perceptual units and against the categorical perception hypothesis is, therefore, evidence against the TRACE model.

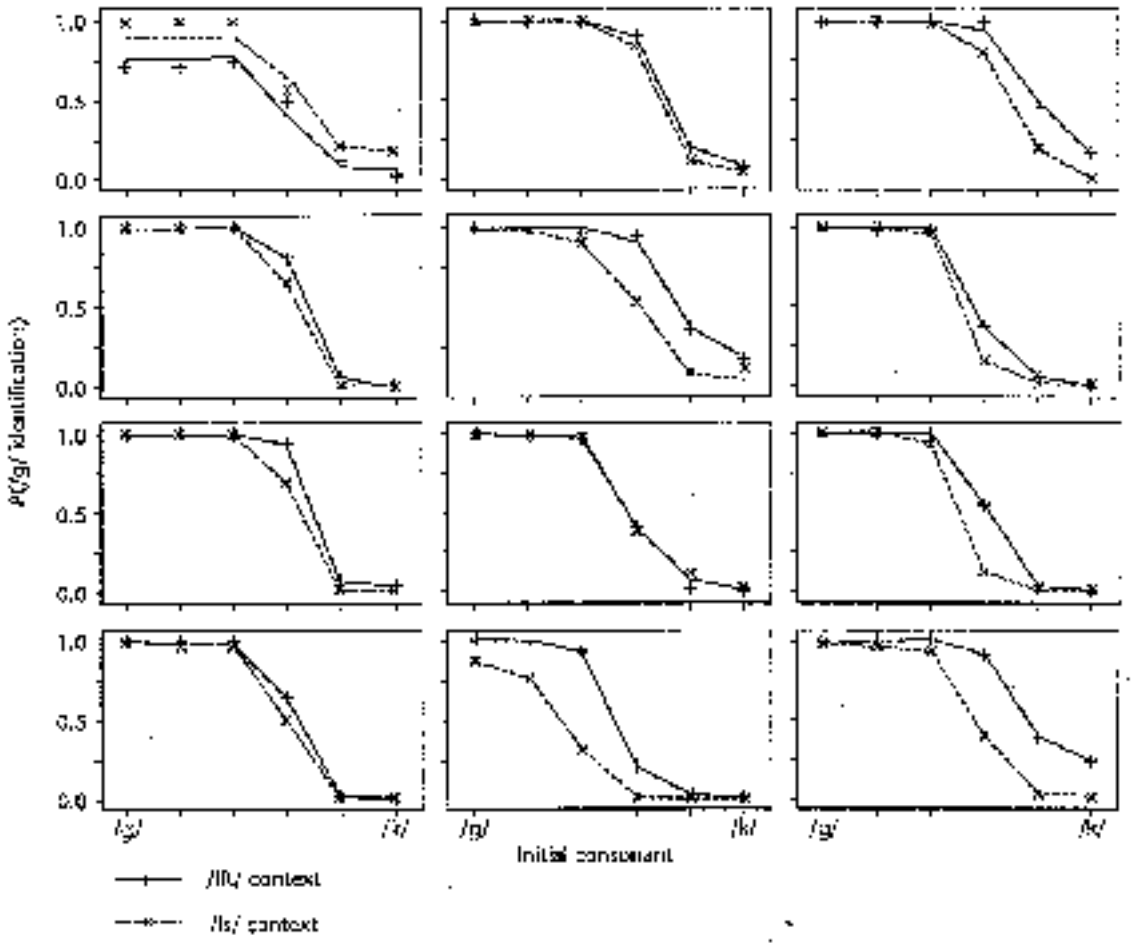

Figure 4. Observed probabilities (points), and FLMP's predictions (lines), of /g/ identifications for /l/- and /s/- contexts as a function of the speech information of the initial consonant which ranges from /g/ to /k/. Results for 12 subjects from Pitt (1995).

PHONEMIC RESTORATION

In the original type of phonemic restoration study (Warren, 1970), a phoneme in a word is removed and replaced with some other stimulus, such as a tone or white noise. Subjects perceive the word as relatively intact and have difficulty indicating what phoneme is missing. This illusion has been taken to support interactive activation: the lexical information supposedly modifies the sublexical auditory representation. This interpretation contrasts with that given by the FLMP, in which the lexical context simply provides an additional source of information. There is no reason to assume that the auditory representation was modified. In terms of signal detection analysis, the lexical influence is one of bias and not sensitivity. Several experimenters have addressed this issue. Although Samuel (1981) concluded that his results favored interactive activation, the changes in top-down context were confused with different stimulus conditions (Massaro, 1989).

Repp (1992) carried out a systematic set of experiments to clarify the locus of the phonemic restoration phenomenon. Essentially, he asked the question whether the restoration was localized at the auditory level or at a higher linguistic level. Within the framework of the FLMP, the effect is attributed to the independent contribution of lexical constraints, with no top-down influence on the auditory level. Subjects were asked to judge the perceived pitch (timbre, brightness) of the extraneous noise that replaced the speech sound. If restoration involved auditory processing, we would expect these pitch judgments to be influenced by lexical context. The overall results of five experiments were negative: auditory effects were not observed, even though phonemic restoration did occur. Repp concludes that perceivers' reports of what they are hearing cannot be taken as an index

of auditory process or representation (the McGurk effect). A positive influence of a top-down constraint on perceptual report does not necessarily mean that the representation of the bottom-up influences has been modified. Thus there is no evidence for top-down activation of a lexical representation on a lower-level auditory representation.

HIDDEN MARKOV MODELS

Hidden Markov Models (HMMs) are used in speech recognition when we conceptualize the signal in speech as a sequence of states (Rabiner, 1989; Jurafsky and Martin, 2000). The goal is to determine which model best accounts for the observed sequence of states. For example, the states might correspond to a sequence of phonemes or to a sequence of smaller segments, and the goal is to determine which word best accounts for the observed sequence. A pedagogical example involves an observation of a series of coin tosses, such as HHTHHTTTH. We know there are many possible models that could generate these observed results. The most obvious (and parsimonious) model would be to assume that only a single (possibly biased) coin was being tossed. This is an observable model, and the only parameter needed to completely specify the model is the bias value. In another model, one could assume that two different coins are being tossed. In this case, it is necessary to specify both the biases of the two coins and how the system moves from state to state (coin to coin). A third model assumes three biased coins, and choosing among the three on the basis of some probabilistic event. The three models vary in the number of free parameters: the first model requires only a single free parameter; the second model requires four free parameters; and the third model requires nine free parameters. Given that the goal is simply speech recognition accuracy rather than psychological validity, it is not a problem that larger HMMs will necessarily provide a better description of the observed sequence of events. The only limitation on the size of the HMMs is computational complexity.

The three basic problems to solve given HMMs are: (1) computing the probability of an observed sequence for a given model; (2) choosing a state sequence that is optimal given the observation sequence and the model; and (3) determining how the model's parameters are adjusted to maximize the probability of an observed sequence for a given model. What are the fundamental assumptions in using HMMs? It is assumed that the observations are independent, so that the sequence of observations can be written as a product of individual observations. Furthermore, the probability of being in state i depends only on the state at time $t-1$. Finally, the distributions of individual observation parameters can be represented by a mixture of Gaussian or autoregressive densities.

HMMs have found limited application in psychology perhaps because they are not easily applied to empirical studies of speech perception, they are not grounded in psychological processes, and they may not be falsifiable.

CONCLUSION

The study of speech perception is an interdisciplinary endeavor, which involves a varied set of experimental and theoretical approaches. It includes the fundamental psychophysical question of what properties of spoken language are perceptually meaningful and how these properties signal the message. Independent variation of several properties, along with a quantitative theoretical analysis, is a productive paradigm to address not only the psychophysical question but also the issue of how multiple cues are used together for perception and understanding. Spoken language consists of multiple sources of information from several modalities, and people have continuous information from these many sources. In addition to these bottom-up sources, higher-order context is exploited in speech processing. There are several productive theoretical approaches to the complex question of how we so easily communicate with one another.

Acknowledgement

This work was supported in part by NSF CHALLENGE grant CDA-9726363, Public Health Service grant PHS R01 DC00236, National Science Foundation grant 23518, Intel Corporation, and the University of California Digital Media Innovation Program.

References

Damper RI (1994) Connectionist models of categorical perception of speech. In: *Proceedings of the IEEE International Symposium on Speech, Image Processing and Neural Networks*, vol. 1, pp. 101–104. Hong Kong.

Damper RI and Harnad SR (2000) Neural network models of categorical perception. *Perception and Psychophysics* 62(4): 843–867.

Eimas PD (1985) The perception of speech in early infancy. *Scientific American* 252: 46–52.

Ganong WF (1980) Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance* 6: 110–125.

Gleitman LR and Wanner E (1982) Language acquisition: the state of the state of the art. In: Wanner E and Gleitman LR (eds) *Language Acquisition: The State of the Art*, pp. 3–48. Cambridge, UK: Cambridge University Press.

Iverson P and Kuhl PK (2000) Perceptual magnet and phoneme boundary effects in speech perception: do they arise from a common mechanism? *Perception and Psychophysics* 62(4): 874–886.

Jurafsky D and Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kuhl PK (1991) Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50: 93–107.

Liberman AM and Mattingly IG (1985) The motor theory of speech perception revised. *Cognition* 21: 1–36.

Lotto AJ, Kluender KR and Holt LL (1998) Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America* 103: 3648–3655.

Massaro DW (1987) *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Erlbaum.

Massaro DW (1989) Testing between the TRACE model and the Fuzzy Logical Model of speech perception. *Cognitive Psychology* 21: 398–421.

Massaro DW (1994) Psychological aspects of speech perception: implications for research and theory. In: Gernsbacher M (ed.) *Handbook of Psycholinguistics*, pp. 219–263. New York, NY: Academic Press.

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

Massaro DW and Cohen MM (1983) Categorical or continuous speech perception: a new test. *Speech Communication* 2: 15–35.

Massaro DW and Stork DG (1998) Sensory integration and speechreading by humans and machines. *American Scientist* 86: 236–244.

McClelland JL (1991) Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology* 23: 1–44.

McClelland JL and Elman JL (1986) The TRACE model of speech perception. *Cognitive Psychology* 18: 1–86.

Pitt MA (1995) The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory and Cognition* 21: 1037–1052.

Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77: 257–286.

Repp BH (1992) Perceptual restoration of a 'missing' speech sound: auditory induction or illusion? *Perception and Psychophysics* 51: 14–32.

Seberi K and Perrett DR (1999) Cognitive restoration of reversed speech. *Nature* 398: 760.

Samuel AG (1981) Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General* 110: 474–494.

Sussman HM, Fruchter D, Hilbert J and Sirosh J (1998) Linear correlates in the speech signal: the orderly output constraint. *Behavioral and Brain Sciences* 21: 241–299.

Warren RM (1970) Perceptual restoration of missing speech sounds. *Science* 167(3917): 392–393.

Further Reading

Allport DA, MacKay DG, Prinz W and Scheerer E (eds) (1987) *Language Perception and Production: Shared Mechanisms in Listening, Speaking, Reading and Writing*. London: Academic Press.

Altmann GTM (ed.) (1990) *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press.

Campbell R, Dodd B and Burnham D (eds) (1998) *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*. Hove, UK: Psychology Press.

Jurafsky D and Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

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

Schouten MEH (ed) (1992) *The Auditory Processing of Speech*. Berlin: Mouton de Gruyter.

Speech Processing: Perception, Analysis, and Synthesis. [http://www.mp.uchicago.edu/overview/language/speech/]

Speech Research. [http://mambo.ucsc.edu/psl/speech.html]

Tohkura Y, Vatikiotis-Bateson E and Sagisaka Y (eds) (1992) *Speech Perception, Production and Linguistic Structure*. Tokyo: Ohmsha.

UCLA Speech Processing and Auditory Perception Laboratory. [http://www.icsl.ucla.edu/~spapl/]