From Massaro, D.W., 2015. Speech Perception. In: James D. Wright (editor-in-chief), International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Vol 23. Oxford: Elsevier. pp. 235–242.
ISBN: 9780080970868

Speech Perception

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

Abstract

This psychological account of speech perception includes a description of the stimulus variables that support it and the internal processes that account for it. Research and theory indicate that speech perception is a form of pattern recognition that is influenced by multiple sources of bottom-up and top-down information.

Speech Perception warrants an entry in this encyclopedia along with Motion Perception and Music Perception. One possible implication of a separate entry for Speech Perception is that it differs significantly from other types of perceptions. In fact, much of the six or seven decades of research and theory centered on this topic has assumed that speech is special. If speech is special, then how it is perceived must also be special. In many ways, the specialty assumption is congruent with the more prominent belief that language and how it is acquired is special.

Speech Is Special

A highly influential proposal by Noam Chomsky 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 perception and learning theory. This mental organ, responsible for the human language faculty and our language competence, matures and develops with experience, but the matured system does not simply mirror this experience. The language user arrives on this planet with rule systems of a highly specific structure capable of acquiring languages that necessarily have a structure corresponding to these rules. This innate knowledge is necessary to become an expert in using language, even though the typical child has a paucity of the language input. The data of language experience are so limited that no learning 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 us to learn to use language appropriately without learning many of the formal intricacies of the language. This theory is highly controversial and there are many reasons to remain skeptical. As counterevidence, psychologists are demonstrating that infants' language is highly influenced by typical perception and learning processes, and linguists are documenting that the child's language input is not as sparse as the nativists claim (Scholz and Pullum, 2006).

Luckily for our purposes, we can evaluate the claim that speech is special without resolving the language is special controversy. There are several aspects of speech that appeared to demand specialty status. Perception researchers typically expect an orderly relationship between a stimulus and its resulting percept. The loudness of a tone increases as it is increased in intensity, for example, and two different intensities would give different percepts. However, the speech percept appeared to be characterized by a non-invariant relationship between it and the eliciting stimulus. An historical example gives the surprising perception of consonant–vowel (CV) syllables /di/ as in deed and /du/ as in do. We hear the phoneme (Smallest segment of spoken language that is functional in communicating meaning.) /d/ in both cases, but the speech signal differs significantly in the two cases. Of course, not every scientist was convinced that phonemes did not have invariant signal properties in different vowel contexts, and research in the six or so decades since the original claim has continued to attempt to uncover these properties (e.g., Kapoor and Allen, 2012). Even so, no one has yet provided convincing evidence of invariance.

Faced with this conundrum, researchers looked for a solution outside of the normal framework of perception in which the goal is to describe the stimulus characteristics responsible for perceptual outcomes. An obvious alternative was to believe that a motor theory could reinstate the regularity between signal and percept. A motor theory of speech perception would assume that we somehow perceive the vocal tract gestures creating the sounds rather than the sounds themselves. As recently observed by Hickok, motor theories have a long history in behavioral science. Over 100 years ago, Walter B. Pillsbury (1911: p. 84) said "The [motor] theory is so simple and so easy to present that everyone is glad to believe it. The only question that any one cares to raise is how much of it will the known facts permit one to accept." It is sobering how little has changed in the intervening century given that many scientists have not been swayed by Pillsbury's caveat. Speech provided a natural domain for a motor theory account because speech perception and speech production appeared to be so tightly linked. The perceiver simply perceives the intended phonetic gestures of the speaker rather than the variable auditory speech. For our /di/-/du/ example, the perceiver simply recovers the interlocutor's intention to articulate /d/ in both syllables. For most of us, it is difficult to intuit how the motor theory accounts for speech perception and how it resolves the lack of invariance problem. How do we access the talker's intended articulations? (Coarticulation: Pronunciation (articulation) of a speech segment is influenced by its neighboring segments.) If we could see the talker, we would obtain a few hints, and in fact observing the talker dramatically influences what we hear. The normal view is not sufficient, however,

International Encyclopedia of the Social & Behavioral Sciences, Second Edition, 2015, 235–242

because much of the articulation is hidden inside the mouth. More telling is the common observation of successful phone conversations when the talkers are not visible.

What's most disappointing about motor theory is that its advocates do not offer specific testable hypotheses about how it accomplishes speech perception. Most of the explanations have taken the form of analysis by synthesis models. Given a speech event, the perceiver benefits from synthesizing possible alternatives during their processing of the speech input and somehow determining which of these possibilities is the best match. To accomplish this selection, however, the perceiver still requires access to the speech signal, and hence we have regressed back to the original challenge of understanding how the speech signal is processed.

In one of the few explicit accounts, Vihman (2002) describes how motor processes might work in speech acquisition. The infant practices canonical babbling and produces CV sequences at 6–8 months of age. This practice in production sensitizes the infant to similar input speech from their caregivers. The familiarity of CV sequences from babbling allows them to be easily recognized because they pop out of the acoustic stream. However, although these patterns would become familiar with practice and this increase in familiarity might facilitate perception, it does not mean that the motor processes involved in babbling were also functional during the infant's speech perception.

In fact, infant speech perception is much more sophisticated than what could be predicted from canonical babbling. We know that receptive language is acquired before productive language, so it is difficult to understand how motor behavior would contribute to speech perception. Motor theory does not solve the problem of speech perception. Infants understand speech well before they produce it intelligibly. Furthermore, infants are highly sensitive to the statistical properties of segmental speech input at 6–8 months (Saffran, 2003), which could not be anticipated by canonical babbling.

Roy (2011) documents 67 instances in which his son attempted to pronounce 'water' before he was able to pronounce it correctly. Many of these instances illustrate that he was able to perceive and understand the spoken word but was simply unable to produce it accurately. This scientific observation revives the anecdotal one in which a father is mimicking his son's mispronunciation of the word 'rabbit.' His son corrects his dad saying, "No dad, not rawwit, but rawwit." The son clearly could perceive the difference between 'rabbit' and 'rawwit' even though he was not able to produce the difference. Toddlers (13–15-month-olds) have few words in their productive vocabulary but can compute the relations in a spoken sentence (Hirsh-Pasek and Golinkoff, 1996). Hickok (2009) also reviews long-standing clinical findings that falsify motor theory. For example, persons with Broca's aphasia can perceive and understand speech but cannot produce it. Thus, research on language acquisition like the research and clinical findings with adults does not support motor theory.

The astute reader will notice that we have not solved the invariance problem that motivated a motor theory account. To solve the invariance problem between acoustic signal and phoneme, Massaro (1972) proposed the syllables V, CV, or VC as perceptual units in speech, where V is a vowel and C is a consonant or consonant cluster (unit of speech that is functional in speech perception). This might appear as too easy of a solution because one might argue that the bigger the unit the better. So why not assume a word, a phrase, or even a sentence as the perceptual unit? Luckily, there were other constraints on the size of the unit. The speech representing the perceptual unit must be maintained in a pre-perceptual auditory memory in order to be functional for recognition (Massaro, 1972). Backward masking experiments indicated that pre-perceptual auditory memory had a limit of roughly 250 ms, which can hold these syllables but not longer units (this assumption was built into the foundation of the model discussed in Section The Fuzzy Logical Model of Perception). Assuming that this larger segment is the perceptual unit that reinstates a significant amount of invariance between signal and percept, Massaro and Oden (1980: pp. 133–135) reviewed evidence that the major coarticulatory influences (that are responsible for the /di/-/du/ differences) on perception occur within these syllables, rather than between the syllables. Any remaining lack of invariance across these syllables could conceivably be disambiguated by additional sources of information in the speech stream.

Popular theories, such as the motor theory, will be reinvented and will evidently endure well beyond their many falsifications as we will discuss in the next section.

Mirror Neurons

The discovery of mirror neurons has apparently rejuvenated motor theories. A mirror neuron fires both when an animal performs an action and when the animal observes the same action performed by another animal (Rizzolatti and Craighero, 2004). Mirror neurons could serve as the basis for imitation and therefore for learning. Our understanding, however, is that mirror neurons cannot account for perception because they would overgeneralize. The macaque certainly experiences the difference between seeing a conspecific action and its own action, but the same mirror neurons are putatively activated by these very different events and experiences. If only the mirror neurons were responsible for speech perception, the interlocutor would not distinguish between her perception of the utterance from her production of the utterance. Thus, mirror neurons alone cannot account for perception (see Hickok, 2009; Lotto et al., 2009 for similar observations and additional critiques).

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 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.

In a seminal paper, Miller and Nicely (1955) were interested in the acoustic properties of speech that were important for the recognition of English consonants. They asked participants to identify 16 initial consonants placed before the vowel /a/ as in father. Degradation of performance was measured under varying conditions of noise and filtering. The standard measure of performance is in terms of a confusion matrix which shows each of the responses to each of the test stimuli. One goal of this research is to determine which acoustic properties best explain the confusion matrix. The authors chose linguistic descriptors as rough indices of the acoustic properties. These included voicing (low frequency periodic energy), nasality (spectral changes caused by energy flow through the nasal tract), frication (high frequency energy), place of articulation (formant location and transitions), and duration. Although these properties were treated as binary (or trinary) linguistic dimensions, we now know that a more accurate description would be in terms of a continuous value (Massaro and Cohen, 1999). One disadvantage of these studies is that the speech signal is necessarily degraded in order to generate errors. The results may not generalize to typical situations in which the speech is not degraded.

Greenberg and his colleagues have pursued Miller and Nicely's original quest for understanding how the various acoustic properties representing linguistic features contribute to speech perception. They find that not all linguistic features are created equal. Based on systematic studies of confusion matrices in Danish, they find that the ease of resolving three features is ordered voicing, manner, and place, respectively. They also find that resolving place of articulation is highly correlated with correct consonant identification. They interpret this result to mean that resolving place of articulation is central to speech recognition. Allen (e.g., Kapoor and Allen, 2012) and his colleagues have provided some innovative techniques to assess which acoustic properties are important for speech recognition. They systematically manipulate the natural speech signal to emphasize or degrade an acoustic feature. They systematically study the burst of energy at the onset of stop consonants in the initial position. Based on their results, they conclude that the frequency range and relative intensity of the plosive burst and its temporal relationship to the sonorance of the following vowel are sufficient features for correct identification of the four stop consonants /da/, /ga/, /ta/, and /ka/. In an earlier study, they found that /ba/ is characterized by a broad band of energy in the burst. If this information is missing, the segment is confused for the consonant /v/.

The auditory speech signal is distributed across a frequency spectrum and it has been traditional to treat this spectrum as composed of independent frequency bands. These bands are putatively independent sources of information for intelligibility of the speech signal, which are important for the accepted measures, namely Articulation Index, Speech Intelligibility Index, and the Speech Transmission Index. However, this assumption is untenable as demonstrated nicely by the research of Christiansen and Greenberg (2012).

Although six decades later, 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, research has shown that the second formant (F2) transition defined as the change between the F2 value at the onset of a 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 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 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 1/15th 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 (e.g., Greenberg and Arai, 2004), 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. 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.

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 (Situation in which a speech stimulus is perceived in terms of its category membership rather than in terms of its surface properties.) meaning that we perceive an utterance as a specific syllable, word, or phrase. This phenomenal outcome is easily interpreted as a perceptual process that is categorical. This means that the perceptual system only has access to categorical information about the speech event rather than access to continuous information about the degree to which various events are possible. My experience, particularly now with an increasing hearing loss of high frequencies, is just the opposite. In many instances, I am stumped with a percept that fits several possible alternatives rather than one categorically. This experience is to be expected, given a large body of empirical research that reveals that speech perception is not categorical.

Using synthetic speech, it is possible to make a continuum of different speech sounds varying in small steps between two alternatives. Categorical perception 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). Categorical perception was operationalized in terms of discrimination performance being limited by identification performance. In a very influential study, Liberman et al. (1957) used synthetic speech to generate a series of 14 CV 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. Although this categorical perception hypothesis can be quantified in order to predict discrimination performance from the identification judgments (Massaro, 1987), only a qualitative analysis was carried out. A rough correspondence between identification and discrimination led the authors to incorrectly conclude that discrimination performance was fairly well-predicted by identification performance. Unfortunately, there are several notable limitations in these and many similar 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. Following developments in signal detection theory, we used rating experiments to determine if perceivers indeed have information about the degree of category membership (Massaro and Cohen, 1983). Rather than being asked for categorical decisions, perceivers were asked to rate the stimulus along a continuum between categories. A detailed quantitative analysis of the results indicated that perceivers have reliable information about the degree of category membership. 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, he/she can certainly have information about the degree to which each toy was requested.

Although categorical perception has been discredited, it is often reinvented under new guises. The hypothesis of a 'perceptual magnet' is that prototypical instances of a speech category function like magnets for surrounding sounds that are less prototypical members of that category, creating sharper boundaries between categories (Kuhl, 1991, 2004). The perceptual magnet putatively creates representations which produce good discriminability of speech segments near category boundaries and poor discriminability near the category prototype. (Ideal or central member of a category or concept: Robins are prototypical birds, whereas eagles and penguins are not). Thus, the perceptual magnet effect a reformulation of categorical perception. Evidence for the perceptual magnet requires an experiment to show how discrimination is directly predicted by a measure of category goodness. Lotto et al. (1998) found, however, that discriminability was *not* poorer for vowels with high category goodness, in contrast to the predictions of the perceptual magnet theory. The same authors also observed additional negative evidence that category goodness ratings were highly context-sensitive. If category goodness is functional in discrimination, it should be relatively stable across different contexts. (Prototype Description: A summary description of a category in terms of a representation for the ideal member of the category.)

Some proponents of categorical perception are also skeptical of the view that speech is special. As evidence, they cite several experimental studies that demonstrate that nonhuman animals also produce categorical perception. The belief is that these studies reveal that spoken language capitalized on natural auditory discontinuities for defining speech categories. Perhaps the most telling negative evidence is that the 6000 or so languages of the world have carved up the speech stimulus space in very different ways, precluding this possibility (Lindau and Ladefoged, 1983). My belief is that there are too few possible discontinuities for just a single language, let alone a plethora of languages. Most importantly, the 'categorical perception' results in nonhuman animals, like the corresponding results with humans, can be equally or better explained in terms of continuous perception.

It was common to attribute categorical perception to infants as well as adults (Eimas, 1985; Gleitman and 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, follow-up studies quickly demonstrated that infants discriminate differences within, as well as between, categories (Massaro, 1987: p. 239; McMurray and Aslin, 2005). Other infant studies revealed how rapidly infants develop an affinity for their language. Infants have been shown to behave differently to two different speech segments at 4 months of age but not at 14 months, if the two segments are not informative

in their language (Werker and Tees, 1983). This ability to quickly learn the meaningful distinctions in their language is not unique to speech. Research in sign language has revealed exactly parallel results. Infants raised with American Sign Language show a sensitivity to different handshapes at 4 months of age but do not behave differently to them at 14 months, if they are not functional (Baker et al., 2006).

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. It is just 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).

Although categorical perception is about as old as speech research itself, we must conclude that speech is perceived continuously. Speech decisions are necessarily categorical. Research reveals conclusively that audible and visible sources of bottom-up information and top-down sources of contextual information generate continuous representations. As discussed in Section The Fuzzy Logical Model of Perception, these representations are integrated to provide a usually robust interpretation of the language input (Massaro, 1987, 1998). Most importantly, the case for the specialization of speech is weakened considerably because of the central role that the assumption of categorical perception 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 categorical perception (Damper, 1994): its nonexistence poses great problems for these models.

Speech Perception as Pattern Recognition

The study of speech perception has matured into an interdisciplinary endeavor, which involves a varied set of experimental and theoretical approaches. It includes the fundamental psychophysical question of which 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 pursue not only the psychophysical question but also the issue of how the multiple cues are used together for perception and understanding. We have learned that spoken language understanding is influenced by multiple sources of information from several modalities and that perceivers derive continuous information from these many sources. In addition to these bottom-up sources, higher-order context is also used in speech understanding. Several productive theoretical approaches address the complex question of how we so easily understand another's utterances.

It is 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 rates that can differ by about a factor of two (Pellegrino et al., 2011). The reason for these different rates appears to be due to the information density of the speech: some languages are denser than others, and it appears that all languages are spoken to have about the same density. If your language is less dense, you speak more quickly so your listener stays attentive. Even with these differences, foreign languages give the impression that they are being spoken more quickly. Our troubling experience with a foreign language is simply that we do not know it. These experienced differences with known and unknown languages are solely due to the memory structures and psychological processes involved in speech perception.

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.

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 these 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). (A logic based on the premise that propositions can be graded in truth value, rather than simply being true or false. Degree of truth is represented by a truth value between 0 (perfectly false) and 1 (perfectly true). Fuzzy logic also gives algorithms for computing the negation, conjunction, and disjunction of continuous truth values.)

The Fuzzy Logical Model of Perception

The three processes involved in perceptual recognition are illustrated in Figure 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: (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.

Given multiple sources of information, it is useful to have a common metric representing the degree of match of each feature. Two features that define a prototype can be related to

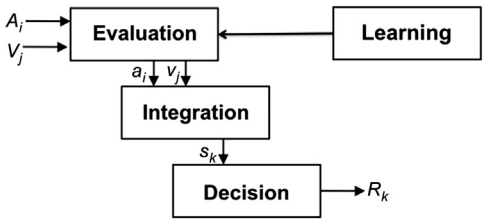

Figure 1 Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The different 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.

one another more easily if they share a common currency. To serve this purpose, fuzzy-truth values (Goguen, 1969; Zadeh, 1965) are used because they provide a natural representation of the degree of match. Fuzzy-truth values lie between 0 and 1, corresponding to a proposition being completely false and completely true. The value 0.5 corresponds to a completely ambiguous situation, whereas 0.7 would be more true than false and so on. Fuzzy-truth values, therefore, not only can represent continuous rather than just categorical information, but also can represent different kinds of information. The truth values in speech perception would correspond to the likelihood of a speech category given the sensory information. The variability of the sensory information signaling a specific speech category would also influence the truth value (Bejjanki et al., 2011). (Mental Representation: The mental content of perceptions, ideas, images, beliefs, thoughts, memories, and hypotheses. These are symbols because they stand for something else. There is no reason that the representation or symbol has to be discrete).

Figure 1 also illustrates how learning is conceptualized within the model by specifying exactly how the feature values used at evaluation change with experience. Learning in the FLMP can be described by the following algorithm (Friedman et al., 1995; Kitzis et al., 1999). The initial feature value representing the support for an alternative is initially set to 0.5 (since 0.5 is neutral in fuzzy logic). A learning trial consists of a feature (such as closed lips at onset) occurring in a test item followed by informative feedback (such as the syllable /ba/). After each trial, the feature values would be updated according to the feedback, as illustrated in Figure 1. Thus, the perceiver uses the feedback to modify the prototype representations and these in turn will become better tuned to the informative characteristics of the patterns being identified. This algorithm is highly similar to many contemporary views of language acquisition (Best and McRoberts, 2003; Best et al., 2001; Werker and Logan, 1985)

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.

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). 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, $P(/\text{da}/)$, is equal to.

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

In the 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, $P(/\text{da}/)$ is equal to.

$$P(/\text{da}/) = \frac{w_1 a_1 + w_2 v_j}{w_1 + w_2} = w a_i + (1 - w) v_j \qquad [2]$$

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).

Higher-Order or Top-Down Influences

William Chandler Bagley published perhaps the first experimental study of speech perception in his dissertation in 1890. In Bagley's experiment, naturally spoken words were recorded and played back to eight members of the Cornell University Psychology Department on Edison phonograph cylinders. The words were pronounced by deleting one of the consonant sounds in the word. Bagley evaluated the effects of context on word perception: The word was presented alone, with one or two related words, and at the beginning, middle, and end of complete sentences. This experiment may have been the first systematic distortion of the spoken message, although people doubtless have been playing with the words they utter since they began uttering them (witness Pig Latin, for example). The results demonstrated that subjects were often able to correctly recognize the distorted words and, in addition, the word's context word recognition. Correct word recognition was better if the word was placed in the middle of a sentence, for example, relative to being presented alone.

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 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.

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, which assume that perceptual outcomes are impenetrable by context effects. Evidence of multiple sources of information and perceptual influences of context effects validates our critique of categorical perception. If the auditory input constrained the perception to a single speech category, then context could have no productive influence. When the context agreed with the speech category, there would be no benefit. When context disagreed with the categorization, the perceiver would be at a loss as to what to do. Having continuous information from both bottom-up and top-down sources allows for their productive integration.

It is important to mention that the FLMP have been proven to be mathematically equivalent to Bayes theorem, which is an optimal algorithm for combining two information sources (Massaro, 1987; Massaro and Friedman, 1990). However, we prefer the common metric to be fuzzy-truth values because they more naturally represent continuous information than do the probabilities of Bayes theorem. If language processing evolved across many millennia of language use, it should not be surprising that speech communication settled on an optimal situation. The many experimental and theoretical sources of evidence for the FLMP thus in turn supports a Bayesian analysis of speech perception, which invites it within an acceptable general framework of perception and learning.

Neurological Underpinnings of Speech Perception

This article is dedicated to a psychological level of description of speech perception. Of course, there is a neurological level that can be equally informative (Massaro, 1989). Many researchers are actively involved in discovering what brain mechanisms are involved in speech perception and how they work together to allow speech understanding and to support speech production. Speech perception or recognition is distinguished from speech production in a recently proposed dual-stream model (Hickok and Poeppel, 2007). A ventral stream supports perception and a dorsal stream interfaces with articulatory networks to allow speech production. A few assumptions of the model that find support at the neurological level are relevant to our review of speech perception. First, there are brain processes that allow the influence of multiple sources of information on speech perception. Second, the model rejects outright that motor behavior intervenes in speech perception (Hickok et al., 2009). Both these assumptions at the neurological level are consistent with our conclusions at the psychological level.

Bibliography

Baker, S.A., Michnick-Golinkoff, R., Petitto, L.A., 2006. New insights into old puzzles from infants' categorical discrimination of soundless phonetic units. Language Learning and Development 2, 147–162.

Beale, J.M., Keil, F.C., 1995. Categorical effects in the perception of faces. Cognition 57 (3), 217–239.

Bejjanki, V.R., Clayards, M., Knill, D.C., Aslin, R.N., 2011. Cue integration in categorical tasks: insights from audio-visual speech perception. PLoS One 6, e19812. http://dx.doi.org/10.1371/journal.pone.0019812.

Best, C., McRoberts, G., Goodell, E., 2001. American listeners' perception of nonnative consonant contrasts varying in perceptual assimilation to English phonology. Journal of the Acoustical Society of America 1097, 775–794.

Best, C.T., McRoberts, G.W., 2003. Infant perception of non-native consonant contrasts that adults assimilate in different ways. Language and Speech 46, 183–216.
Christiansen, T.U.C., Greenberg, S., 2012. Perceptual confusions among consonants, revisited – cross-spectral integration of phonetic-feature information and consonant recognition. IEEE Transactions on Audio, Speech, and Language Processing 20, 147–161.
Damper, R.I., 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.
Etcoff, N.L., McGee, J.J., 1992. Categorical perception of facial expressions. Cognition 44, 227–240.
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, 343–350.
Friedman, D., Massaro, D.W., Kitzis, S.N., Cohen, M.M., 1995. A comparison of learning models. Journal of Mathematical Psychology 39, 164–178.
Ganong III, W.F., 1980. Phonetic categorization in auditory word recognition. Journal of Experimental Psychology: Human Perception and Performance 6, 110–125.
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.
Goguen, J.A., 1969. The logic of inexact concepts. Synthese 19, 325–373.
Greenberg, S., Arai, T., 2004. What are the essential cues for understanding spoken language? IEICE Transactions on Information and Systems E87, 1059–1070.
Hickok, G., 2009. Eight problems for the mirror neuron theory of action understanding in monkeys and humans. Journal of Cognitive Neuroscience 21, 1229–1243.
Hickok, G., Holt, L.L., Lotto, A.J., 2009. Response to Wilson: what does motor cortex contribute to speech perception? Trends in Cognitive Sciences 13, 330–331.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393–402.
Hirsh-Pasek, K., Golinkoff, R.M., 1996. The Origins of Grammar. Evidence from early language comprehension. MIT Press, Cambridge, MA.
Kapoor, A., Allen, J.B., 2012. Perceptual effects of plosive featur modification. Journal of the Acoustical Society of America 131 (1), 478–491.
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.
Kuhl, P.K., 2004. Early language acquisition: Cracking the speech code. Nature Reviews Neuroscience 5, 831–843.
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, 358–368, 753–771.
Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21, 1–36.
Lindau, M., Ladefoged, P., 1983. Variability of feature specifications. In: Perkell, J., Klatt, D. (Eds.), Invariance and Variability in Speech Processes. Lawrence Erlbaum Associates, Inc., New Jersey, pp. 464–479.
Lotto, A.L., Hickok, G., Holt, L.L., 2009. Reflections on mirror neurons and speech perception. Trends in Cognitive Sciences 13, 110–114.
Lotto, A.J., Kluender, K.R., Holt, L.L., 1998. Depolarizing the perceptual magnet effect. Journal of the Acoustical Society of America 103, 3648–3655.
McMurray, B., Aslin, R.N., 2005. Infants are sensitive to within-category variation in speech perception. Cognition 95, B15–B26.
Massaro, D.W., 1972. Preperceptual images, processing time, and perceptual units in auditory perception. Psychological Review 79, 124–145.
Massaro, D.W., 1987. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Erlbaum Associates, Hillsdale, NJ.
Massaro, D.W., Friedman, D., 1990. Models of integration given multiple sources of information. Psychological Review 97, 225–252.
Massaro, D.W., 1994. Psychological aspects of speech perception: implications for research and theory. In: Gemsbacher, M. (Ed.), Handbook of Psycholingristics. Acadamic Press, New York, pp. 219–263.
Massaro, D.W., 1998. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press, Cambridge, MA.
Massaro, D.W., Cohen, M.M., 1999. Speech perception in hearing-impaired perceivers: Synergy of multiple modalities. Journal of Speech, Language, and Hearing Research 42, 21–41.
Massaro, D.W., 2012. Speech perception and reading: two parallel modes of understanding language and implications for acquiring literacy naturally. American Journal of Psychology 125 (3), 307–320.
Massaro, D.W., Chen, T.H., 2008. The motor theory of speech perception revisited. Psychonomic Bulletin & Review 15 (2), 453–457.
Massaro, D.W., Cohen, M.M., 1983. Categorical or continuous speech perception: a new test. Speech Communication 2, 15–35.
Massaro, D.W., Oden, G.C., 1980. Speech perception: a framework for research and theory. In: Lass, N.J. (Ed.), Speech and Language: Advances in Basic Research and Practice, vol. 3. Academic Press, New York, pp. 129–165.
Massaro, D.W., Stork, D.G., 1998. Sensory integration and speechreading by humans and machines. American Scientist 86, 236–244.
McMurray, B., Aslin, R.N., 2005. Infants are sensitive to within-category variation in speech perception. Cognition 95, B15–B26.
Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America 27, 338–352.
Ouni, S., Cohen, M.M., Ishak, H., Massaro, D.W., 2007. Visual contribution to speech perception: measuring the intelligibility of animated talking heads. EURASIP Journal on Audio, Speech, and Music Processing 2007. http://dx.doi.org/10.1155/2007/47891.
Pellegrino, F., Coupé, C., Marsico, E., 2011. A cross-language perspective on speech information rate. Language 87 (3), 539–558.
Pillsbury, W.B., 1911. The Essentials of Psychology. Macmillan, New York.
Pitt, M.A., 1995. The locus of the lexical shift in phoneme identification. Journal of Experimental Psychology: Learning, Memory, & Cognition 21, 1037–1052.
Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annual Review of Neurosciemce 27, 169–192.
Roy, D., 2011. http://www.ted.com/talks/deb_roy_the_birth_of_a_word.html.
Saberi, K., Perrott, D.R., 1999. Cognitive restoration of reversed speech. Nature 398, 760.
Saffran, J.R., 2003. Statistical language learning: Mechanisms and constraints. Current Directions in Psychological Science 12, 110–114.
Scholz, B.C., Pullum, G.K., 2006. Irrational nativist exuberance. In: Stainton, R. (Ed.), Contemporary Debates in Cognitive Science. Basil Blackwell, Oxford, pp. 59–80.
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), 241–299.
Vihman, M.M., 2002. The role of mirror neurons in the ontogeny of speech. In: Stamenov, M., Gallese, V. (Eds.), Mirror Neurons and the Evolution of Brain and Language. Benjamins, Amsterdam, pp. 305–314.
Werker, J.F., Tees, R.C., 1983. Developmental changes across childhood in the perception of non-native speech sounds. Canadian Journal of Psychology 37, 278–286.
Werker, J.F., Logan, J.S., 1985. Cross-language evidence for three factors in speech perception. Perception & Psychophysics 37, 35–44.
Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 338–353.