Overview

Speech perception and production

Elizabeth D. Casserly[1]* and David B. Pisoni[2]

Until recently, research in speech perception and speech production has largely focused on the search for psychological and phonetic evidence of discrete, abstract, context-free symbolic units corresponding to phonological segments or phonemes. Despite this common conceptual goal and intimately related objects of study, however, research in these two domains of speech communication has progressed more or less independently for more than 60 years. In this article, we present an overview of the foundational works and current trends in the two fields, specifically discussing the progress made in both lines of inquiry as well as the basic fundamental issues that neither has been able to resolve satisfactorily so far. We then discuss theoretical models and recent experimental evidence that point to the deep, pervasive connections between speech perception and production. We conclude that although research focusing on each domain individually has been vital in increasing our basic understanding of spoken language processing, the human capacity for speech communication is so complex that gaining a full understanding will not be possible until speech perception and production are conceptually reunited in a joint approach to problems shared by both modes. © 2010 John Wiley & Sons, Ltd. *WIREs Cogn Sci* 2010 1 629–647

Historically, language research focusing on the spoken (as opposed to written) word has been split into two distinct fields: speech perception and speech production. Psychologists and psycholinguists worked on problems of phoneme perception, whereas phoneticians examined and modeled articulation and speech acoustics. Despite their common goal of discovering the nature of the human capacity for spoken language communication, the two broad lines of inquiry have experienced limited mutual influence. The division has been partially practical, because methodologies and analysis are necessarily quite different when aimed at direct observation of overt behavior, as in speech production, or examination of hidden cognitive and neurological function, as in speech perception. Academic specialization has also played a part, since there is an overwhelming volume of knowledge available, but single researchers can only learn and use a small portion. In keeping with the goal of this series, however, we argue that the greatest prospects for progress in speech research over the next few years lie at the intersection of insights from research on speech perception and production, and in investigation of the inherent links between these two processes.

In this article, therefore, we will discuss the major theoretical and conceptual issues in research dedicated first to speech perception and then to speech production, as well as the successes and lingering problems in these domains. Then we will turn to several exciting new directions in experimental evidence and theoretical models which begin to close the gap between the two research areas by suggesting ways in which they may work together in everyday speech communication and by highlighting the inherent links between speaking and listening.

SPEECH PERCEPTION

Before the advent of modern signal processing technology, linguists and psychologists believed that speech perception was a fairly uncomplicated, straightforward process. Theoretical linguistics' description of spoken language relied on the use of sequential strings of abstract, context-invariant segments, or phonemes, which provided the mechanism of contrast between

*Correspondence to: casserly@indiana.edu

[1]Department of Linguistics, Speech Research Laboratory, Indiana University, Bloomington, IN 47405, USA

[2]Department of Psychological and Brain Sciences, Speech Research Laboratory, Cognitive Science Program, Indiana University, Bloomington, IN 47405, USA

DOI: 10.1002/wcs.63

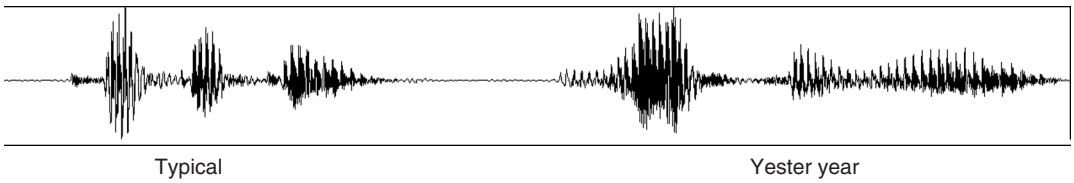

FIGURE 1 | Speech waveform of the words *typical* and *yesteryear* as produced by an adult male speaker, representing variations in amplitude over time. Vowels are generally the most resonant speech component, corresponding to the most extreme amplitude levels seen here. The identifying formant frequency information in the acoustics is not readily accessible from visual inspection of waveforms such as these.

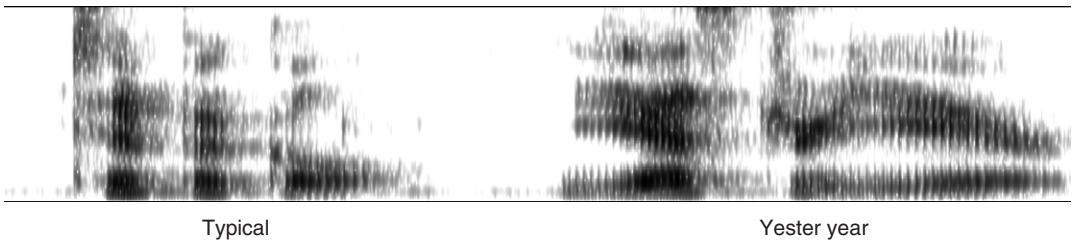

FIGURE 2 | A wide-band speech spectrogram of the same utterance as in Figure 1, showing the change in component frequencies over time. Frequency is represented along the y-axis and time on the x-axis. Darkness corresponds to greater signal strength at the corresponding frequency and time.

lexical items (e.g., distinguishing *pat* from *bat*).[1,2] The immense analytic success and relative ease of approaches using such symbolic structures led language researchers to believe that the physical implementation of speech would adhere to the segmental 'linearity condition,' so that the acoustics corresponding to consecutive phonemes would concatenate like an acoustic alphabet or a string of beads stretched out in time. If that were the case, perception of the linguistic message in spoken utterances would be a trivial matching process of acoustics to contrastive phonemes.[3]

Understanding the true nature of the physical speech signal, however, has turned out to be far from easy. Early signal processing technologies, prior to the 1940s, could detect and display time-varying acoustic amplitudes in speech, resulting in the familiar waveform seen in Figure 1. Phoneticians have long known that it is the component frequencies encoded within speech acoustics, and how they vary over time, that serve to distinguish one speech percept from another, but waveforms do not readily provide access to this key information. A major breakthrough came in 1946, when Ralph Potter and his colleagues at Bell Laboratories developed the speech spectrogram, a representation which uses the mathematical Fourier transform to uncover the strength of the speech signal hidden in the waveform amplitudes (as shown in Figure 1) at a wide range of possible component frequencies.[4] Each calculation finds the signal strength through the frequency spectrum of a small time window of the speech waveform; stringing the results of these time-window analyses together yields a speech spectrogram or voiceprint, representing the dynamic frequency characteristics of the spoken signal as it changes over time (Figure 2).

Phonemes—An Unexpected Lack of Evidence

As can be seen in Figure 2, the content of a speech spectrogram does not visually correspond to the discrete segmental units listeners perceive in a straightforward manner. Although vowels stand out due to their relatively high amplitudes (darkness) and clear internal frequency structure, reflecting harmonic resonances or 'formant frequencies' in the vocal tract, their exact beginning and ending points are not immediately obvious to the eye. Even the seemingly clear-cut amplitude rises after stop consonant closures, such as for the [p] in *typical*, do not directly correlate with the beginning of a discrete vowel segment, since these acoustics simultaneously provide critical information about both the identity of the consonant and the following vowel. Demarcating consonant/vowel separation is even more difficult in the case of highly sonorant (or resonant) consonants such as [w] or [r].

The simple 'acoustic alphabet' view of speech received another set-back in the 1950s, when Franklin Cooper of Haskins Laboratories reported his research group's conclusion that acoustic signals composed of

strictly serial, discrete units designed to corresponding phonemes or segments are actually impossible for listeners to process at speeds near those of normal speech perception.[5] No degree of signal simplicity, contrast between units, or user training with the context-free concatenation system could produce natural rates of speech perception for listeners. Therefore, the Haskins group concluded that speech must transmit information in parallel, through use of the contextual overlap observed in spectrograms of the physical signal. Speech does not look like a string of discrete, context-invariant acoustic segments, and in order for listeners to process its message as quickly as they do, it cannot be such a system. Instead, as Alvin Liberman proposed, speech is a 'code,' taking advantage of parallel transmission of phonetic content on massive scale through co-articulation[3] (see section 'Variation in Invariants,' below).

As these discoveries came to light, the 'speech perception problem' began to appear increasingly insurmountable. On the one hand, phonological evidence (covered in more depth in the 'Variation in Invariants' section) implies that phonemes are a genuine property of linguistic systems. On the other hand, it has been shown that the acoustic speech signal does not directly correspond to phonological segments. How could a listener use such seemingly unhelpful acoustics to recover a speaker's linguistic message? Hockett encapsulated early speech scientists' bewilderment when he famously likened the speech perception task to that of the inspector in the following scenario:

> Imagine a row of Easter eggs carried along a moving belt; the eggs are of various sizes, and variously colored, but not boiled. At a certain point, the belt carries the row of eggs between two rollers of a wringer, which quite effectively smash them and rub more or less into each other. The flow of eggs before the wringer represents the series of impulses from the phoneme source; the mess that emerges from the wringer represents the output of the speech transmitter. At a subsequent point, we have an inspector whose task it is to examine the passing mess and decide, on the basis of the broken and unbroken yolks, the variously spread out albumen, and the variously colored bits of shell, the nature of the flow of eggs which previously arrived at the wringer.
>
> (Ref 1, p. 210)

For many years, researchers in the field of speech perception focused their efforts on trying to solve this enigma, believing that the heart of the speech perception problem lay in the seemingly impossible task of phoneme recognition—putting the Easter eggs back together.

Synthetic Speech and the Haskins Pattern Playback

Soon after the speech spectrogram enabled researchers to visualize the spectral content of speech acoustics and its changes over time, that knowledge was put to use in the development of technology able to generate speech synthetically. One of the early research synthesizers was the Pattern Playback (Figure 3, top panel), developed by scientists and engineers, including Cooper and Liberman, at Haskins Laboratories.[6] This device could take simplified sound spectrograms like those shown in Figure 3 and use the component frequency information to produce highly intelligible corresponding speech acoustics. Hand-painted spectrographic patterns (Figure 3, lower panel) allowed researchers tight experimental control over the content of this synthetic, very simplified Pattern Playback speech. By varying its frequency content and transitional changes over time, investigators were able to determine many of the specific aspects in spoken language which are essential to particular speech percepts, and many which are not.[3,6]

Perceptual experiments with the Haskins Pattern Playback and other speech synthesizers revealed, for example the pattern of complex acoustics that signals the place of articulation of English stop consonants, such as [b], [t] and [k].[3] For voiced stops ([b], [d], [g]) the transitions of the formant frequencies from silence to the vowel following the consonant largely determine the resulting percept. For voiceless stops ([p], [t], [k]) however, the acoustic frequency of the burst of air following the release of the consonant plays the largest role in identification. The experimental control gained from the Pattern Playback allowed researchers to alter and eliminate many aspects of naturally produced speech signals, discovering the identities of many such sufficient or necessary acoustic cues for a given speech percept. This early work attempted to pair speech down to its bare essentials, hoping to reveal the mechanisms of speech perception. Although largely successful in identifying perceptually crucial aspects of speech acoustics and greatly increasing our fundamental understanding of speech perception, these pioneering research efforts did not yield invariant, context-independent acoustic features corresponding to segments or phonemes. If anything, this research program suggested alternative bases for the communication of linguistic content.[7,8]

Phoneme Perception—Positive Evidence

Some of the research conducted with the aim of understanding phoneme perception, however, did

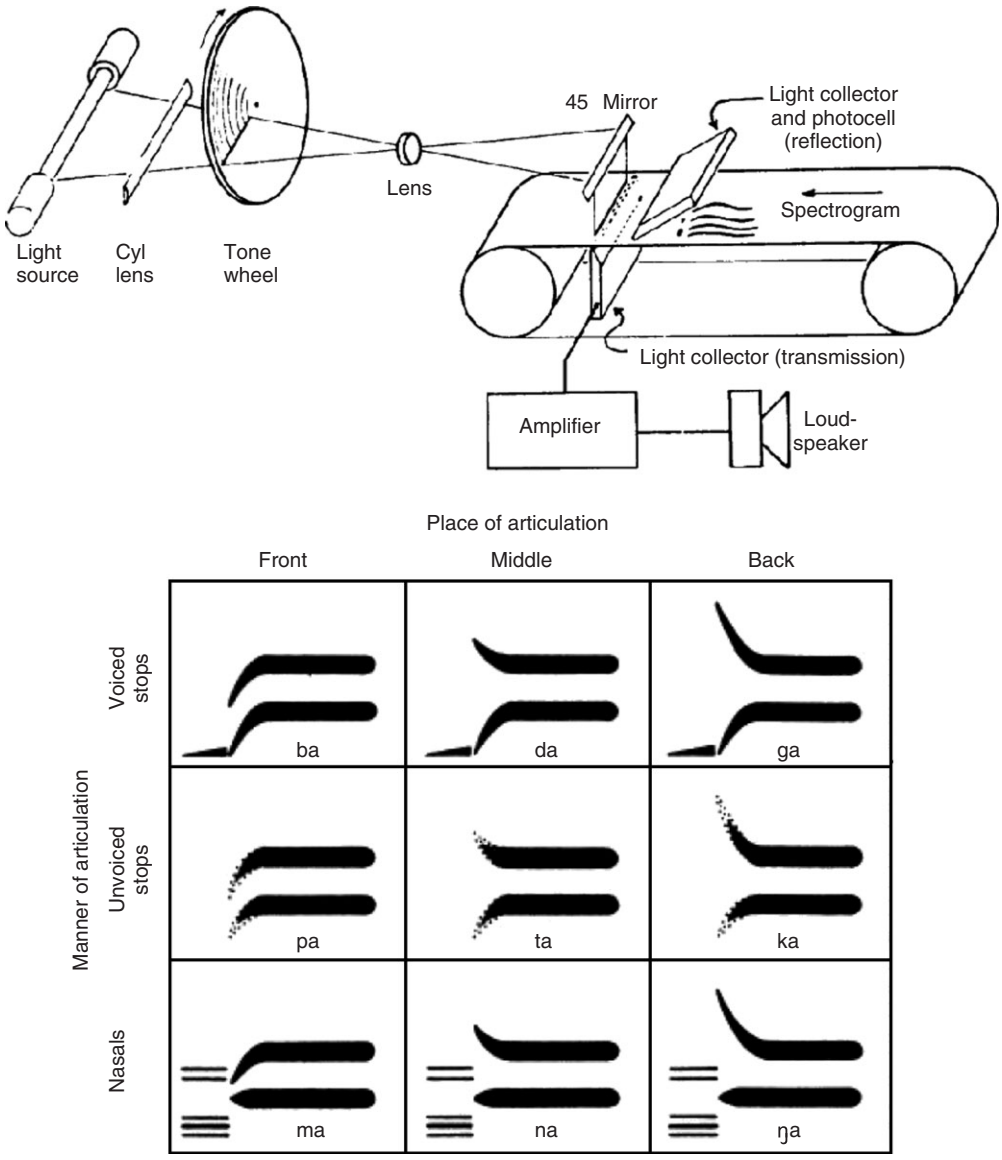

FIGURE 3 | Top panel: A diagram of the principles and components at work in the Haskins Pattern Playback speech synthesizer. (Reprinted with permission from Ref. 68 Copyright 1951 national Academies of Science.) Lower panel: A series of hand-painted schematic spectrographic patterns used as input to the Haskins Pattern Playback in early research on perceptual 'speech cues.' (Reprinted with permission from Ref. 69 Copyright 1957 American Institute of Physics.)

lead to results suggesting the reality of psychological particulate units such as phonemes. For instance, in some cases listeners show evidence of 'perceptual constancy,' or abstraction from signal variation to more generalized representations—possibly phonemes. Various types of such abstraction have been found in speech perception, but we will address two of the most influential here.

Categorical Perception Effects

Phoneme representations split potential acoustic continuums into discrete categories. The duration of aspiration occurring after the release of a stop consonant, for example, constitutes a potential continuum ranging from 0 ms, where vocalic resonance begins simultaneously with release of the stop, to an indefinitely long period between the stop release and the start of the following vowel. Yet stops in English falling along this continuum are split by native listeners into two functional groups—voiced [b], [d], [g] or voiceless [p], [t], [k]—based on the length of this 'voice onset time.' In general, this phenomenon is not so strange: perceptual categories often serve to break continuous variation into manageable chunks.

Speech categories appear to be unique in one aspect, however, listeners are unable to reliably discriminate between two members of the same category. That is, although we may assign two different colors both to the category 'red,' we can easily distinguish between the two shades in most cases. When speech scientists give listeners stimuli varying along an acoustic continuum, however, their discrimination between

different tokens of the same category (analogous to two shades of red) is very close to chance.[9] They are highly accurate at discriminating tokens spanning category boundaries, on the other hand. The combination of sharp category boundaries in listeners' labeling of stimuli and their within-category insensitivity in discrimination, as shown in Figure 4, appears to be unique to human speech perception, and constitutes some of the strongest evidence in favor of robust segmental categories underlying speech perception.

According to this evidence, listeners sacrifice sensitivity to acoustic detail in order to make speech category distinctions more automatic and perhaps also less subject to the influence of variability. This type of category robustness is observed more strongly in the perception of consonants than vowels. Not coincidentally, as discussed briefly above and in more detail in the 'Acoustic Phonetics' section, below, the stop consonants which listeners have the most difficulty discriminating also prove to be the greatest challenge to define in terms of invariant acoustic cues.[10]

Perceptual Constancy

Categorical perception effects are not the only case of such abstraction or perceptual constancy in speech perception; listeners also appear to 'translate' the speech they hear into more symbolic or idealized forms, encoding based on expectations of gender and accent. Niedzielski, for example, found that listeners identified recorded vowel stimuli differently when they were told that the original speaker was from their own versus another dialect group.[11] For these listeners, therefore, the mapping from physical speech characteristics to linguistic categories was not absolute, but mediated by some abstract conceptual unit. Johnson summarizes the results of a variety of studies showing similar behavior,[12] which corroborates the observation that, although indexical or 'extra-linguistic' information such as speaker gender, dialect, and speaking style are not inert in speech perception, more abstract linguistic units play a role in the process as well.

Far from being exotic, this type of 'perceptual equivalence' corresponds very well with language users' intuitions about speech. Although listeners are aware that individuals often sound drastically different, the feeling remains that something holds constant across talkers and speech tokens. After all, *cat* is still *cat* no matter who says it. Given the signal variability and complexity observed in speech acoustics, such consistency certainly seems to imply the influence of some abstract unit in speech perception, possibly contrastive phonemes or segments.

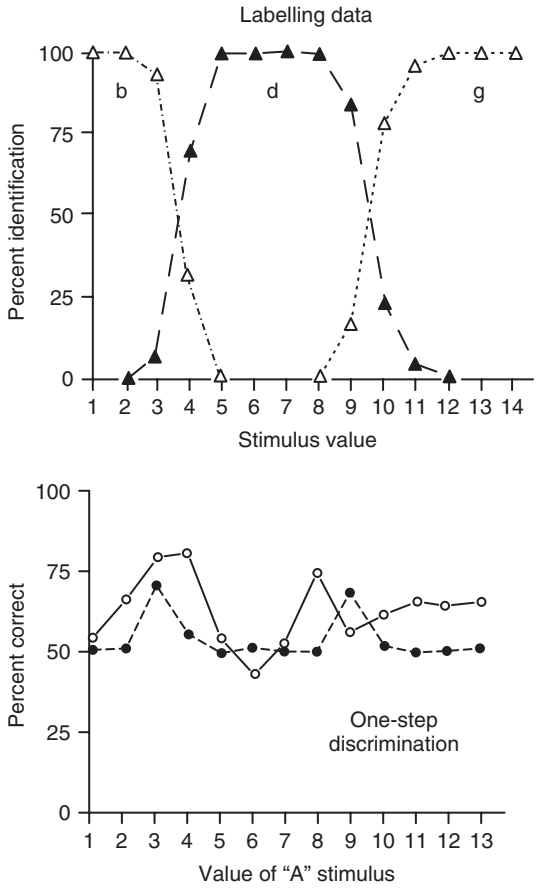

FIGURE 4 | Data for a single subject from a categorical perception experiment. The upper panel gives labeling or identification data for each step on a [b]/[g] place-of-articulation continuum. The lower graph gives this subject's ABX discrimination data (filled circles) for the same stimuli with one step difference between pairs, as well as the predicted discrimination performance (open circles). Discrimination accuracy is high at category boundaries and low within categories, as predicted. (Reprinted with permission from Ref. 9 Copyright 1957 American Psychological Association.)

Phoneme Perception—Shortcomings and Roadblocks

From the discussion above, it should be evident that speech perception research with the traditional focus on phoneme identification and discrimination has been unable either to confirm or deny the psychological reality of context-free symbolic units such as phonemes. Listeners' insensitivity to stimulus differences within a linguistic category and their reference to an abstract ideal in identification support the cognitive role of such units, whereas synthetic speech manipulation has simultaneously demonstrated that linguistic percepts simply do not depend on invariant, context-free acoustic cues corresponding to segments. This paradoxical relationship between signal

variance and perceptual invariance constitutes one of the fundamental issues in speech perception research.

Crucially, however, the research discussed until now focused exclusively on the phoneme as the locus of language users' perceptual invariance. This approach stemmed from the assumption that speech perception can essentially be reduced to phoneme identification, relating yet again back to theoretical linguistics' analysis of language as sequences of discrete, context-invariant units. Especially given the roadblocks and contradictions emerging in the field, however, speech scientists began to question the validity of those foundational assumptions. By attempting to control variability and isolate perceptual effects on the level of the phoneme, experimenters were asking listeners to perform tasks that bore little resemblance to typical speech communication. Interest in the field began to shift toward the influence of larger linguistic units such as words, phrases, and sentences and how speech perception processes are affected by them, if at all.

Beyond the Phoneme—Spoken Word Recognition Processes

Both new and revisited experimental evidence readily confirmed that the characteristics of word-level units do exert massive influence in speech perception. The lexical status (word vs non-word) of experimental stimuli, for example, biases listeners' phoneme identification such that they hear more tokens as [d] in a *dish*/*tish* continuum, where the [d] percept creates a real word, than a *da*/*ta* continuum where both perceptual options are non-words.[13] Moreover, research into listeners' perception of spoken words has shown that there are many factors that play a major role in word recognition but almost never influence phoneme perception.

Perhaps the most fundamental of these factors is word frequency: how often a lexical item tends to be used. The more frequently listeners encounter a word over the course of their daily lives, the more quickly and accurately they are able to recognize it, and the better they are at remembering it in a recall task (e.g., Refs 14,15). High-frequency words are more robust in noisy listening conditions, and whenever listeners are unsure of what they have heard through such interference, they are more likely to report hearing a high-frequency lexical item than a low-frequency one.[16] In fact, the effects of lexical status mentioned above are actually only extreme cases of frequency effects; phonotactically legal non-words (i.e., non-words which seem as though they could be real words) are treated psychologically like real words with a frequency of zero. Like cockroaches, these so-called 'frequency effects' pop up everywhere in speech research.

The nature of a word's 'lexical neighborhood' also plays a pervasive role in its recognition. If a word is highly similar to many other words, such as *cat* is in English, then listeners will be slower and less accurate to identify it, whereas a comparably high-frequency word with fewer 'neighbors' to compete with it will be recognized more easily. 'Lexical hermits' such as *Episcopalian* and *chrysanthemum*, therefore, are particularly easy to recognize despite their low frequencies (and long durations). As further evidence of frequency effects' ubiquitous presence, however, the frequencies of a word's neighbors also influence perception: a word with a dense neighborhood of high-frequency items is more difficult to recognize than a word with a dense neighborhood of relatively low-frequency items, which has weaker competition.[17,18]

Particularly troublesome for abstractionist phoneme-based views of speech perception, however, was the discovery that the indexical properties of speech (see 'Perceptual Constancy,' below) also influence word recognition. Goldinger, for example, has shown that listeners are more accurate at word recall when they hear stimuli repeated by the same versus different talkers.[19] If speech perception were mediated only by linguistic abstractions, such 'extra-linguistic' detail should not be able to exert this influence. In fact, this and an increasing number of similar results (e.g., Ref 20) have caused many speech scientists to abandon traditional theories of phoneme-based linguistic representation altogether, instead positing that lexical items are composed of maximally detailed 'episodic' memory traces.[19,21]

Conclusion—Speech Perception

Regardless of the success or failure of episodic representational theories, a host of new research questions remain open in speech perception. The variable signal/common percept paradox remains a fundamental issue: what accounts for the perceptual constancy across highly diverse contexts, speech styles and speakers? From a job interview in a quiet room to a reunion with an old friend at a cocktail party, from a southern belle to a Detroit body builder, what makes communication possible? Answers to these questions may lie in discovering the extent to which the speech perception processes tapped by experiments in word recognition and phoneme perception are related, and uncovering the nature of the neural substrates of language that allow adaptation to such diverse situations. Deeply connected to these issues, Goldinger, Johnson and others' results have prompted

us to wonder: what is the representational specificity of speech knowledge and how does it relate to perceptual constancy?

Although speech perception research over the last 60 years has made substantial progress in increasing our understanding of perceptual challenges and particularly the ways in which they are *not* solved by human listeners, it is clear that a great deal of work remains to be done before even this one aspect of speech communication is truly understood.

SPEECH PRODUCTION

Speech production research serves as the complement to the work on speech perception described above. Where investigations of speech perception are necessarily indirect, using listener response time latencies or recall accuracies to draw conclusions about underlying linguistic processing, research on speech production can be refreshingly direct. In typical production studies, speech scientists observe articulation or acoustics as they occur, then analyze this concrete evidence of the physical speech production process. Conversely, where speech perception studies give researchers exact experimental control over the nature of their stimuli and the inputs to a subject's perceptual system, research on speech production severely limits experimental control, making the investigators observe more or less passively, whereas speakers do as they will in response to their prompts.

Such fundamentally different experimental conditions, along with focus on the opposite side of the perceptual coin, allows speech production research to ask different questions and draw different conclusions about spoken language use and speech communication. As we discuss below, in some ways this 'divide and conquer' approach has been very successful in expanding our understanding of speech as a whole. In other ways, however, it has met with many of the same roadblocks as its perceptual complement and similarly leaves many critical questions unanswered in the end.

A Different Approach

When the advent of the speech spectrogram made it obvious that the speech signal does not straightforwardly mirror phonemic units, researchers responded in different ways. Some, as discussed above, chose to question the perceptual source of phoneme intuitions, trying to define the acoustics necessary and sufficient for an identifiable speech percept. Others, however, began separate lines of work aiming to observe the behavior of speakers more directly. They wanted to know what made the speech signal as fluid and seamless as it appeared, whether the observed overlap and contextual dependence followed regular patterns or rules, and what evidence speakers might show in support of the reality of the phonemic units. In short, these speech scientists wanted to demystify the puzzling acoustics seen on spectrograms by investigating them in the context of their source.

The Continuing Search

It may seem odd, perhaps, that psychologists, speech scientists, engineers, phoneticians, and linguists were not ready to abandon the idea of phonemes as soon as it became apparent that the physical speech signal did not straightforwardly support their psychological reality. Dating back to Panini's grammatical study of Sanskrit, however, the use of abstract units such as phonemes has provided enormous gains to linguistic and phonological analysis. Phonemic units appear to capture the domain of many phonological processes, for example, and their use enables linguists to make sense of the multitude of patterns and distributions of speech sounds across the world's languages. It has even been argued[22] that their discrete, particulate nature underlies humanity's immense potential for linguistic innovation, allowing us to make 'infinite use of finite means.'[23]

Beyond these theoretical gains, phonemes were argued to be empirically supported by research on speech errors or 'slips of the tongue,' which appeared to operate over phonemic units. That is, the kinds of errors observed during speech production, such as anticipations (''a leading list''), perseverations ('pulled a pantrum'), reversals ('heft lemisphere'), additions ('moptimal number'), and deletions ('chrysanthemum p_ants'), appear to involve errors in the ordering and selection of whole segmental units, and always result in legal phonological combinations, whose domain is typically described as the segment.[24] Without evidence to the contrary, these errors seemed to provide evidence for speakers' use of discrete phonological units.

Although there have been dissenters[25] and shifts in the conception of the units thought to underlie speech, abstract features, or phoneme-like units of some type have remained prevalent in the literature. In light of the particulate nature of linguistic systems, the enhanced understanding gained with the assumption of segmental analysis, and the empirical evidence observed in speech planning errors, researchers were and are reluctant to give up the search for the basis of phonemic intuitions in physically observable speech production.

Acoustic Phonetics

One of the most fruitful lines of research into speech production focused on the acoustics of speech. This body of work, part of 'Acoustic Phonetics,' examines the speech signals speakers produce in great detail, searching for regularities, invariant properties, and simply a better understanding of the human speech capacity. Although the speech spectrograph did not immediately show the invariants researchers anticipated, they reasoned that such technology would also allow them to investigate the speech signal at an unprecedented level of scientific detail. Because speech acoustics are so complex, invariant cues corresponding to phonemes may be present, but difficult to pinpoint.[10,26]

While psychologists and phoneticians in speech perception were generating and manipulating synthesized speech in an effort to discover the acoustic 'speech cues,' therefore, researchers in speech production refined signal processing techniques enabling them to analyze the content of naturally produced speech acoustics. Many phoneticians and engineers took on this problem, but perhaps none has been as tenacious and successful as Kenneth Stevens of MIT.

An electrical engineer by training, Stevens took the problem of phoneme-level invariant classification and downsized it, capitalizing on the phonological theories of Jackobson et al.[27] and Chomsky and Halle's *Sound Patterns of English*[28] which postulated linguistic units below the level of the phoneme called distinctive features. Binary values of universal features such as [sonorant], [continuant], and [high], these linguists argued, constituted the basis of phonemes. Stevens and his colleagues thought that invariant acoustic signals may correspond to distinctive features rather than phonemes.[10,26] Since phonemes often share features (e.g., /s/ and /z/ share specification for all distinctive features except [voice]), it would make sense that their acoustics are not as unique as might be expected from their contrastive linguistic function alone.

Stevens, therefore, began a thorough search for invariant feature correlates that continued until his retirement in 2007. He enjoyed several notable successes: many phonological features, it turns out, can be reliably specified by one or two signal characteristics or 'acoustic landmarks.' Phonological descriptors of vowel quality, such as [high] and [front], were found to correspond closely to the relative spacings between the first and second harmonic resonances of the vocal tract (or 'formants') during the production of sonorant vowel segments.[10]

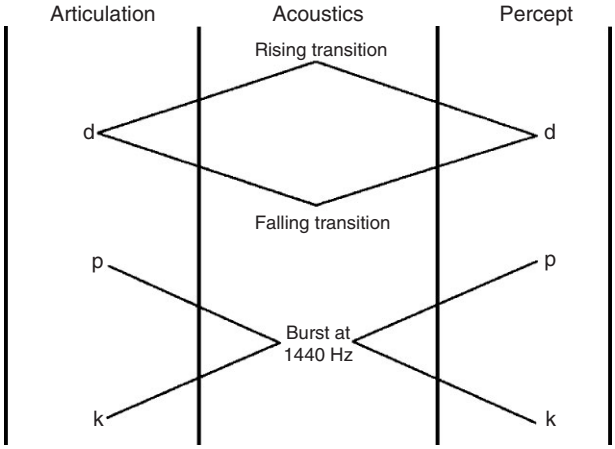

FIGURE 5 | Observations from early perceptual speech cue studies. In the first case, two different acoustic signals (consonant/vowel formant frequency transitions) result in the same percept. In the latter case, identical acoustics (release burst at 1440 Hz) result in two different percepts, depending on the vocalic context. In both cases, however, perception reflects articulatory, rather than acoustic, contrast. (Adapted and reprinted with permission from Ref. 29 Copyright 1996 American Institute of Physics.)

Some features, however, remained more difficult to define acoustically. Specifically, the acoustics corresponding to consonant place of articulation seemed to depend heavily on context—the exact same burst of noise transitioning from a voiceless stop to a steady vowel might result from the lip closure of a [p] or the tongue-dorsum occlusion of a [k], depending on the vowel following the consonant. Equally problematic, the acoustics signaling the alveolar ridge closure location of coronal stop [t] are completely different before different vowels.[29] The articulation/acoustic mismatch, and the tendency for linguistic percepts to mirror articulation rather than acoustics, is represented in Figure 5.

Variation in Invariants

Why do listeners' speech percepts show this dissociation from raw acoustic patterns? Perhaps the answer becomes more intuitive when we consider that even the most reliable acoustic invariants described by Stevens and his colleagues tend to be somewhat broad, dealing in relative distances between formant frequencies in vowels and relative abruptness of shifts in amplitude and so on. This dependence on relative measures comes from two major sources: individual differences among talkers and contextual variation due to co-articulation. Individual speakers' vocal tracts are shaped and sized differently, and therefore they resonant differently (just as different resonating sounds are produced by blowing over the necks of differently

sized and shaped bottles), making the absolute formant frequencies corresponding to different vowels, for instance, impossible to generalize across individuals.

Perhaps more obviously problematic, though, is the second source: speech acoustics' sensitivity to phonetic context. Not only do the acoustics cues for [p], [t], or [k] depend on the vowel following the stop closure, for example, but because the consonant and vowel are produced nearly simultaneously, the identity of the consonant reciprocally affects the acoustics of the vowel. Such co-articulatory effects are extremely robust, even operating across syllable and word boundaries. This extensive interdependence makes the possibility of identifying reliable invariance in the acoustic speech signal highly remote.

Although some researchers, such as Stevens, attempted to factor out or "normalize" these co-articulatory effects, others believed that they are central to the functionality of speech communication. Liberman et al. at Haskins Laboratories pointed out that co-articulation of consonants and vowels allows the speech signal to transfer information in parallel, transmitting messages more quickly than it could if spoken language consisted of concatenated strings of context-free discrete units.[3] Co-articulation therefore enhances the efficiency of the system, rather than being a destructive or communication-hampering force. Partially as a result of this view, some speech scientists focused on articulation as a potential key to understanding the reliability of phonemic intuitions, rather than on its acoustic consequences. They developed the research program called 'articulatory phonetics,' aimed at the study of the visible and hidden movements of the speech organs.

Articulatory Phonetics

Techniques

In many ways articulatory phonetics constitutes as much of an engineering challenge as a linguistic one. Because the majority of the vocal tract 'machinery' lies hidden from view (see Figure 6), direct observation of the mechanics of speech production requires technology, creativity, or both. And any potential solution to the problem of observation cannot disrupt natural articulation too extensively if its results are to be useful in understanding natural production of speech. The challenge, therefore, is to investigate aspects of speech articulation accurately and to a high level of detail, while keeping interference with the speaker's normal production as minor as possible.

Various techniques have been developed that manage to satisfy these requirements, spanning from the broadly applicable to the highly specialized. Electromyography (EMG), for instance, allows researchers to measure directly the activity of muscles within the vocal tract during articulation via surface or inserted pin electrodes.[30] These recordings have broad applications in articulatory phonetics, from determining the relative timing of tongue movements during syllable

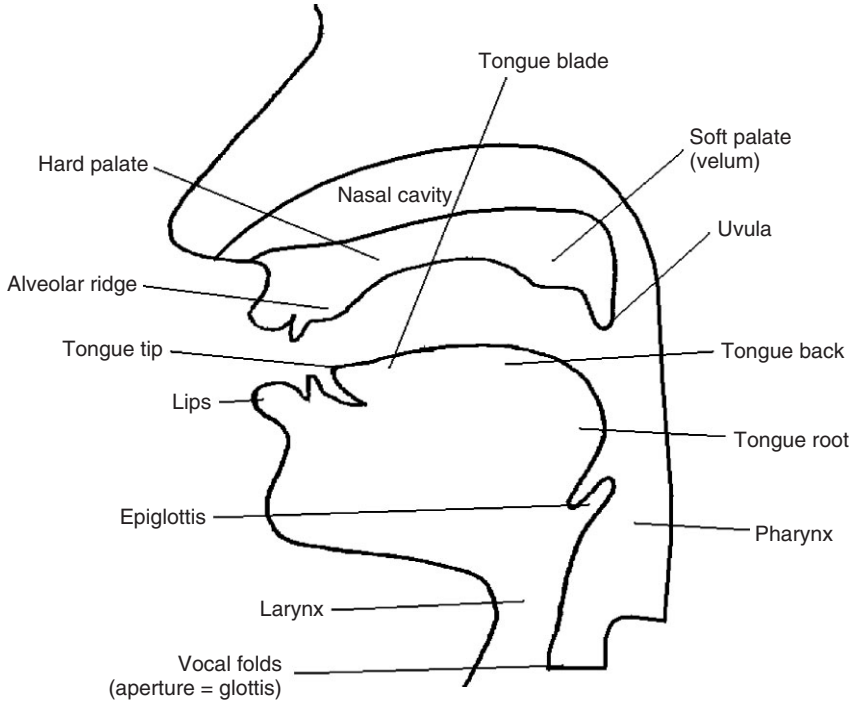

FIGURE 6 | A sagittal view of the human vocal tract showing the main speech articulators as labeled. (Reprinted with permission from Ref. 70 Copyright 2001 Blackwell Publishers Inc.)

production to measures of pulmonary function from activity in speakers' diaphragms to examining tone production strategies via raising and lowering of speakers' larynxes. EMG electrode placement can significantly impact articulation, however, which does impose limits on its use. More specialized techniques are typically still more disruptive of typical speech production, but interfere minimally with their particular investigational target. In transillumination of the glottis, for example, a bundle of fiberoptic lights is fed through a speaker's nose until the light source is positioned just above their larynx.[31] A light-sensitive photocell is then placed on the neck just below the glottis to detect the amount of light passing through the vocal folds at any given moment, which correlates directly with the size of glottal opening over time. Although transillumination is clearly not an ideal method to study the majority of speech articulation, it nevertheless provides a highly accurate measure of various glottal states during speech production.

Perhaps the most currently celebrated articulatory phonetics methods are also the least disruptive to speakers' natural articulation. Simply filming speech production in real-time via X-ray provided an excellent, complete view of unobstructed articulation, but for health and safety reasons can no longer be used to collect new data.[32] Methods such as X-ray microbeam and Electromagnetic Mid-Sagittal Articulometer (EMMA) tracking attempt to approximate that 'X-ray vision' by recording the movements of speech articulators in real-time through other means. The former uses a tiny stream of X-ray energy aimed at radio-opaque pellets attached to a speaker's lips, teeth, and tongue to monitor the movements of the shadows created by the pellets as the speaker talks. The latter, EMMA, generates similar positional data for the speech organs by focusing alternating magnetic fields on a speaker and monitoring the voltage induced in small electromagnetic coils attached to speaker's articulators as they move through the fields during speech. Both methods track the movements of speech articulators despite their inaccessibility to visible light, providing reliable position-over-time data that minimally disrupts natural production.[33,34] However, comparison across subjects can be difficult due to inconsistent placement of tracking points from one subject to another and simple anatomical differences between subjects.

Ultrasound provides another, even less disruptive, articulatory phonetics technique that has been gaining popularity in recent years (e.g., Refs 35,36). Using portable machinery that does nothing more invasive than send sound waves through a speaker's tissue and monitor their reflections, speech scientists can track movements of the tongue body, tongue root, and pharynx that even X-ray microbeam and EMMA cannot capture, as these articulators are all but completely manually inaccessible. By placing an ultrasound wand at the juncture of the head and neck below the jaw, however, images of the tongue from its root in the larynx to its tip can be viewed in real-time during speech production, with virtually no interference to the speech act itself. The tracking cannot extend beyond cavities of open air, making this method inappropriate for studies of precise place of articulation against the hard palate or of velum movements, for example, but these are areas in which X-ray microbeam and EMMA excel. The data recently captured using these techniques are beginning to give speech scientists a more complete picture of speech articulation than ever before.

Impact on the Search for Phonemes

Unfortunately for phoneme-based theories of speech production and planning, the results of recent articulatory studies of speech errors do not seem to paint a compatible picture. As discussed above, the categorical nature of speech errors has served as important support for the use of phonemic units in speech production. Goldstein, Pouplier, and their colleagues, however, used EMMA to track speakers' production of errors in a repetition task similar to a tongue twister. Confirming earlier suspicious (e.g., Ref 25), they found that while speakers' articulation sometimes followed a categorically 'correct' or 'errorful' gestural pattern, it was more frequently somewhere between two opposing articulations. In these cases, small 'incorrect' movements of the articulators would intrude upon the target speech gesture, both gestures would be executed simultaneously, or the errorful gesture would completely overshadow the target articulation. Only the latter reliably resulted in the acoustic percept of a speech error.[37] As Goldstein and Pouplier point out, such non-categorical, gradient speech errors cannot constitute support for discrete phonemic units in speech planning.

Importantly, this finding was not isolated in the articulatory phonetics literature: speakers frequently appear to execute articulatory movements that do not result in any acoustic consequences. Specifically, X-ray microbeam tracking of speakers' tongue tip, tongue dorsum, and lip closures during casual pronunciation of phrases such as *perfect memory* reveals that speakers raise their tongue tips for [t]-closure, despite the fact that the preceding [k] and following [m] typically obscure the acoustic realization of the [t] completely.[38] Although they could minimize their

articulatory effort by not articulating the [t] where it will not be heard, speakers faithfully proceed with their complete articulation, even in casual speech.

Beyond the Phoneme

So far we have seen that, while technological, methodological and theoretical advances have enabled speech scientists to understand the speech signal and its physical production better than ever before, the underlying source of spoken language's systematic nature remains largely mysterious. New research questions continue to be formulated, however, using results that were problematic under old hypotheses to motivate new theories and new approaches to the study of speech production.

The theory of 'Articulatory Phonology' stands as a prominent example; its proponents took the combination of gradient speech error data, speakers' faithfulness to articulation despite varying means of acoustic transmission, and the lack of invariant acoustic speech cues as converging evidence that speech is composed of articulatory, rather than acoustic, fundamental units that contain explicit and detailed temporal structure.[8,38] Under this theory, linguistic invariants are underlyingly motor-based articulatory gestures which specify the degree and location of constrictions in the vocal tract and delineate a certain amount of time relative to other gestures for their execution. Constellations of these gestures, related in time, constitute syllables and words without reference to strictly sequential segmental or phonemic units. Speech perception, then, consists of determining the speech gestures and timing responsible for creating a received acoustic signal, possibly through extension of experiential mapping between the perceiver's own gestures and their acoustic consequences, as in Liberman and Mattingly's Motor Theory of Speech Perception[7] or Fowler's Direct Realist approach.[39] Recent evidence from neuroscience may provide a biological mechanism for this process[40] (see 'Neurobiological Evidence—Mirror Neurons,' below).

And although researchers like Stevens continued to focus on speech acoustics as opposed to articulation, the separate lines of inquiry actually appear to be converging on the same fundamental solution to the invariance problem. The most recent instantiation of Stevens' theory posits that some distinctive phonological features are represented by sets of redundant invariant acoustic cues, only a subset of which are necessary for recognition in any single token. As Stevens recently wrote, however, the distinction between this most recent feature-based account and theories of speech based on gestures may no longer be clear:

> The acoustic cues that are used to identify the underlying distinctive features are cues that provide evidence for the gestures that produced the acoustic pattern. This view that a listener focuses on acoustic cues that provide evidence for articulatory gestures suggests a close link between the perceptually relevant aspects of the acoustic pattern for a distinctive feature in speech and the articulatory gestures that give rise to this pattern.
>
> (Ref 10, p. 142)

Just as in speech perception research, however, some speech production scientists are beginning to wonder if the invariance question was even the right question to be asked in the first place. In the spirit of Lindblom's hyper-articulation and hypo-articulation theory[41] (see 'Perception-Driven Adaptation in Speech Production,' below), these researchers have begun investigating control and variability in production as a means of pinning down the nature of the underlying system. Speakers are asked to produce the same sentence in various contextual scenarios such that a target elicited word occurs as the main element of focus, as a carrier of stress, as a largely unstressed element, and as though a particular component of the word was misheard (e.g., in an exchange such as 'Boy?' 'No, *toy*'), while their articulation and speech acoustics are recorded. Then the data are examined for regularities. If the lengths of onset consonants and following vowels remain constant across emphasized, focused, stressed, and unstressed conditions, for example, that relationship may be specified in the representation of syllables, whereas the absolute closure and vocalic durations vary freely and therefore must not be subject to linguistic constraint. Research of this type seeks to determine the articulatory variables under active, regular control and which (if any) are mere derivatives or side effects of deliberate actions.[42–44]

Conclusion—Speech Production

Despite targeting a directly observable, overt linguistic behavior, speech production research has had no more success than its complement in speech perception at discovering decisive answers to the foundational questions of linguistic representational structure or the processes governing spoken language use. Due to the joint endeavors of acoustic and articulatory phonetics, our understanding of the nature of the acoustic speech signal and how it is produced has increased

tremendously, and each new discovery points to new questions. If the basic units of speech are gesture-based, what methods and strategies do listeners use in order to perceive them from acoustics? Are there testable differences between acoustic and articulatory theories of representation? What aspects of speech production are under demonstrable active control, and how do the many components of the linguistic and biological systems work together across speakers and social and linguistic contexts? Although new lines of inquiry are promising, speech production research seems to have only begun to scratch the surface of the complexities of speech communication.

SPEECH PERCEPTION AND PRODUCTION LINKS

As Roger Moore recently pointed out, the nature of the standard scientific method is such that 'it leads inevitably to greater and greater knowledge about smaller and smaller aspects of a problem' (Ref. 45, p. 419). Speech scientists followed good scientific practice when they effectively split the speech communication problem, one of the most complex behaviors of a highly complex species, into more manageable chunks. And the perceptual and productive aspects of speech each provided enough of a challenge, as we have seen, that researchers had plenty to work on without adding anything. Yet we have also seen that neither discipline on its own has been able to answer fundamental questions regarding linguistic knowledge, representation, and processing.

Although the scientific method serves to separate aspects of a phenomenon, the ultimate goal of any scientific enterprise is to unify individual discoveries, uncovering connections and regularities that were previously hidden.[46] One of the great scientific breakthroughs of the 19th century, for example, brought together the physics of electricity and magnetism, previously separate fields, and revealed them to be variations of the same basic underlying principles. Similarly, where research isolated to either speech perception or production has failed to find success, progress may lie in the unification of the disciplines. And unlike electricity and magnetism the *a priori* connection between speech perception and speech production is clear: they are two sides of the same process, two links in Denes and Pinson's famous 'speech chain'.[47] Moreover, information theory demands that whatever signals generated in speech production match those received in perception, a criteria known as 'signal parity' which must be met for successful communication to take place; therefore, the two processes must at some point even deal in the same linguistic currency.[48]

In this final section, we will discuss theories and experimental evidence that highlight the deep, inherent links between speech perception and production. Perhaps by bringing together the insights achieved within each separate line of inquiry, the recent evidence pointing to the nature of the connection between them, and several theories of how they may work together in speech communication, we can point to where the most exciting new research questions lie in the future.

Early Evidence—Audiovisual Speech Perception

Lurking behind the idea that speech perception and production may be viewed as parts of a unified speech communication process is the assumption that speech consists of more than just acoustic patterns and motor plans that happen to coincide. Rather, the currency of speech must somehow combine the domains of perception and production, satisfying the criterion of signal parity discussed above. Researchers such as Kluender, Diehl, and colleagues take the former, more 'separatist' stance, believing that speech is processed by listeners like any other acoustic signal, without input from or reference to complementary production systems.[49,50] Much of the research described in this section runs counter to such a 'general auditory' view, but none so directly challenges its fundamental assumptions as the phenomenon of audiovisual speech perception.

The typical view of speech, fully embraced thus far here, puts audition and acoustics at the fore. However, visual and other sensory cues also play important roles in the perception of a speech signal, augmenting or occasionally even overriding a listener's auditory input. Very early in speech perception research, Sumby and Pollack showed that simply seeing a speaker's face during communication in background noise can provide listeners with massive gains in speech intelligibility, with no change in the acoustic signal.[51] Similarly, it has been well documented that access to a speaker's facial dynamics improves the accuracy and ease of speech perception for listeners with mild to more severe types of hearing loss[52] and even deaf listeners with cochlear implants.[53,54] Perhaps no phenomenon demonstrates this multimodal integration as clearly or has attracted more attention in the field than the effect reported by McGurk and MacDonald in 1976.[55] When listeners receive simultaneous, mismatching visual and auditory speech input, such as a face articulating the syllable

ba paired with the acoustics for *ga*, they typically experience a unified percept *da* that appears to combine features of both signals while matching neither. In cases of a closer match—between visual *va* and auditory *ba*, for example—listeners tend to perceive *va*, adhering to the visual rather than auditory signal. The effect is robust even when listeners are aware of the mismatch, and has been observed with conflicting tactile rather than visual input[56] and with pre-lingual infants.[57] As these last cases show, the effect cannot be due to extensive experience linking visual and auditory speech information. Instead, the McGurk effect and the intelligibility benefits of audiovisual speech perception provide strong evidence for the inherently multimodal nature of speech processing, contrary to a 'general auditory' view. As a whole, the audiovisual speech perception evidence supports the assumptions which make possible the discussion of evidence for links between speech perception and production below.

Phonetic Convergence

Recent work by Pardo builds on the literature of linguistic 'alignment' to find further evidence of an active link between speech perception and production in 'real-time,' typical communicative tasks. She had pairs of speakers play a communication game called the 'map task,' where they must cooperate to copy a path marked on one speaker's map to the other's blank map without seeing one another. The speakers refer repeatedly to certain landmarks on the map, and Pardo examined their productions of these target words over time. She asked naive listeners to compare a word from one speaker at both the beginning and end of the game with a single recording of the same word said by the other speaker. Consistently across pairs, she found that the recordings from the end of the task were judged to be more similar than those from the beginning. Previous studies have shown that speakers may align in their patterns of intonation,[58] for example, but Pardo's are the first results demonstrating such alignment at the phonetic level in an ecologically valid speech setting.

This 'phonetic convergence' phenomenon defies explanation unless the processes of speech perception and subsequent production are somehow linked within an individual. Otherwise, what a speaker hears his or her partner say could not affect subsequent productions. Further implications of the convergence phenomenon become apparent in light of the categorical perception literature described in 'Categorical Perception Effects' above. In these robust speech perception experiments, listeners appear to be unable to reliably detect differences in acoustic realization of particular segments.[9] Yet the convergence observed in Pardo's work seems to operate at the sub-phonemic level, affecting subtle changes within linguistic categories (i.e., convergence results do not depend on whole-segment substitutions, but much more fine-grained effects).

As Pardo's results show, the examination of links between speech perception and production has already pointed toward new answers to some old questions. Perhaps we do not understand categorical perception effects as well as we thought—if the speech listeners hear can have these gradient within-category effects on their own speech production, then why is it that they cannot access these details in the discrimination tasks of classic categorical perception experiments? And what are the impacts of the answer for various representational theories of speech?

Perception-Driven Adaptation in Speech Production

Despite the typical separation between speech perception and production, the idea that the two processes interact or are coupled within individual speakers is not new. In 1990, Björn Lindblom introduced his 'hyper-articulation and hypo-articulation' (H&H) theory, which postulated that speakers' production of speech is subject to two conflicting forces: economy of effort and communicative contrast.[41] The first pressures speech to be 'hypo-articulated,' with maximally reduced articulatory movements and maximal overlap between movements. In keeping with the theory's roots in speech production research, this force stems from a speaker's motor system. The contrasting pressure for communicative distinctiveness pushes speakers toward 'hyper-articulated' speech, executed so as to be maximally clear and intelligible, with minimal co-articulatory overlap. Crucially, this force stems from listener-oriented motivation. Circumstances that make listeners less likely to correctly perceive a speaker's intended message—ranging from physical factors like presence of background noise, to psychological factors such as the lexical neighborhood density of a target word, to social factors such as a lack of shared background between the speaker and listener—cause speakers to hyper-articulate, expending greater articulatory effort to ensure transmission of their linguistic message.

For nearly a hundred years, speech scientists have known that physical conditions such as background noise affect speakers' production. As Lane and Tranel neatly summarized, a series of experiments stemming from the work of Etienne Lombard in

1911 unequivocally showed that the presence of background noise causes speakers not only to raise the level of their speech relative to the amplitude of the noise, but also to alter their articulation style in ways similar to those predicted by H&H theory.[59] No matter the eventual status of H&H theory in all its facets, this 'Lombard Speech' effect empirically demonstrates a real and immediate link between what speakers are hearing and the speech they produce. As even this very early work demonstrates, speech production does not operate in a vacuum, free from the influences of its perceptual counterpart; the two processes are coupled and closely linked.

Much more recent experimental work has demonstrated that speakers' perception of their own speech can be subject to direct manipulation, as opposed to the more passive introduction of noise used in inducing Lombard speech, and that the resulting changes in production are immediate and extremely powerful. In one experiment conducted by Houde and Jordan, for example, speakers repeatedly produced a target vowel [ɛ], as in *bed*, while hearing their speech only through headphones. The researchers ran the speech through a signal processing program which calculated the formant frequencies of the vowel and shifted them incrementally toward the frequencies characteristic of [æ], raising the first formant and lowering the second. Speakers were completely unaware of the real-time alteration of the acoustics corresponding to their speech production, but they incrementally shifted their articulation of [ɛ] to compensate for the researchers' manipulation: they began producing lower first formants and higher second formants. This compensation was so dramatic that speakers who began by producing [ɛ] ended the experiment by saying vowels much closer to [i] (when heard outside the formant-shifting influence of the manipulation).[60]

Houde, Jordan, and other researchers working in this paradigm point out that such 'sensorimotor adaptation' phenomena demonstrate an extremely powerful and constantly active feedback system in operation during speech production.[61,62] Apparently, a speaker's perception of his or her own speech plays a significant role in the planning and execution of future speech production.

The Role of Feedback—Modeling Spoken Language Use

In his influential theory of speech production planning and execution, Levelt makes explicit use of such perceptual feedback systems in production.[63] In contrast to Lindblom's H&H theory, Levelt's model (WEAVER++) was designed primarily to provide an account of how lexical items are selected from memory and translated into articulation, along with how failures in the system might result in typical speech errors. In Levelt's model, speakers' perception of their own speech allows them to monitor for errors and execute repairs. The model goes a step further, however, to posit another feedback loop entirely internal to the speaker, based on their experience with mappings between articulation and acoustics.

According to Levelt's model, then, for any given utterance a speaker has several levels of verification and feedback. If, for example, a speaker decides to say the word *day* the underlying representation of the lexical item is selected and prepared for articulation, presumably following the various steps of the model not directly relevant here. Once the articulation has been planned, the same 'orders' are sent to both the real speech organs and a mental emulator or 'synthesizer' of the speaker's vocal tract. This emulator generates the acoustics that would be expected from the articulatory instructions it received, based on the speaker's past experience with the mapping. The expected acoustics feed back to the underlying representation of *day* to check for a match with remembered instances of the word. Simultaneous to this process, the articulators are actually executing their movements and generating acoustics. That signal enters the speaker's auditory pathway, where the resulting speech percept feeds back to the same underlying representation, once again checking for a match.

Such a system may seem redundant, but each component has important properties. As Moore points out for his own model (see below), internal feedback loops of the type described in Levelt's work allow speakers to repair errors much more quickly than reliance on external feedback would permit, which translates to significant evolutionary advantages.[45] Without external loops backing up the internal systems, however, speakers might miss changes to their speech imposed by physical conditions (e.g., noise). Certainly the adaptation observed in Houde and Jordan's work demonstrates active external feedback control over speech production: only an external loop could catch the disparity between the acoustics a speaker actually perceives and his or her underlying representation. And indeed, similar feedback-reliant models have been proposed as the underpinnings of non-speech movements such as reaching.[64]

As suggested above, Moore has recently proposed a model of speech communication that also incorporates multiple feedback loops, both internal and external.[45] His Predictive Sensorimotor Control and Emulation (PRESENCE) model goes far beyond the specifications of Levelt's production

model, however, to incorporate additional feedback loops that allow the speaker to emulate the *listener's emulation of the speaker*, and active roles for traditionally 'extra-linguistic' systems such as the speaker's affective or emotional state. In designing his model, Moore attempts to take the first step in what he argues is the necessary unification of not just research on speech perception and production, but the work related to speech in many other fields as well, such as neuroscience, automated speech recognition, text-to-speech synthesis, and biology, to name just a few.[45]

Perhaps most fundamental to our discussion here, however, is the role of productive emulation or feedback during speech perception in the model. Where Levelt's model deals primarily with speech production, Moore's PRESENCE incorporates both speech perception and production, deliberately emphasizing their interdependence and mutually constraining relationship. According to his model, speech perception takes place with the aid of listener-internal emulation of the acoustic-to-articulatory mapping potentially responsible for the received signal. As Moore puts it, speech perception in his model is essentially a revisiting of the idea of 'recognition-by-synthesis' (e.g., Ref 65), whereas speech production is (as in Levelt) 'synthesis by recognition.'

Neurobiological Evidence—Mirror Neurons

The experimental evidence we considered above suggests pervasive links between what listeners hear and the speech they produce. Conversational partners converge in their production of within-category phonetic detail, speakers alter their speech styles in adverse listening conditions, and manipulation of speakers' acoustic feedback from their own speech can dramatically change the speech they produce in response. As we also considered, various theoretical models of spoken language use have been proposed to account for these phenomena and the observed perceptual and productive links. Until recently, however, very little neurobiological evidence supported these proposals. The idea of a speaker-internal vocal tract emulator, for instance, seemed highly implausible to many speech scientists; how would the brain possibly implement such a structure?

Cortical populations of newly discovered 'mirror neurons,' however, seem to provide a plausible neural substrate for proposals of direct, automatic, and pervasive links between speech perception and production. These neurons 'mirror' in the sense that they fire both when a person performs an action themselves and when they perceive someone else performing the same action, either visually or through some other (auditory, tactile) perceptual mode. Human mirror neuron populations appear to be clustered in several cortical areas, including the pre-frontal cortex, which is often implicated in behavioral inhibition and other executive function, and areas typically recognized as centers of speech processing, such as Broca's area (for in-depth review of the literature and implications, see Ref 66).

Neurons which physically equate (or at least directly link) an actor's production and perception of a specific action have definite implications for theories linking speech perception and production: they provide a potential biological mechanism. The internal feedback emulators hypothesized most recently by Levelt and Moore could potentially be realized in mirror neuron populations, which would emulate articulatory-to-acoustic mappings (and vice versa) via their mutual sensitivity to both processes and their connectivity to both sensory and motor areas. Regardless of their specific applicability to Levelt and Moore's models, however, these neurons do appear to be active during speech perception, as one study using Transcranial Magnetic Stimulation (TMS) demonstrates elegantly. TMS allows researchers to temporarily either attenuate or elevate the background activation of a specific brain area, respectively inducing a state similar to the brain damage caused by a stroke or lesion or making it so that any slight increase in the activity of the area causes overt behavior when its consequences would not normally be observable. The later excitation technique was used by Fadiga and colleagues, who raised the background activity of specific motor areas controlling the tongue tip. When the 'excited' subjects then listened to speech containing consonants which curled the tongue upward, their tongues twitched correspondingly.[67] Perceiving the speech caused activation of the motor plans that would be used in producing the same speech—direct evidence of the link between speech perception and production.

Perception/Production Links—Conclusion

Clearly, the links between speech perception and production are inherent in our use of spoken language. They are active during typical speech perception (TMS mirror neuron study), are extremely powerful, automatic and rapid (sensorimotor adaptation), and influence even highly ecologically valid communication tasks (phonetic convergence). Spoken language processing, therefore, seems to represent a linking of sensory and motor control systems, as the pervasive effects of visual input on speech perception suggest. Indeed, speech perception cannot be

just sensory interpretation and speech production cannot be just motor execution. Rather, both processes draw on common resources, using them in tandem to accomplish remarkable tasks such as generalization from talker to talker and acquiring new lexical items. As new information regarding these links comes to light, models such as Lindblom's H&H, Levelt's WEAVER++, and Moore's PRESENCE will both develop greater reflection of the actual capabilities of language users (simultaneous speakers and listeners) and be subject to greater constraint in their hypotheses and mechanisms. And hopefully, theory and experimental evidence will converge to discover how speech perception and production interact in the highly complex act of vocal communication.

CONCLUSIONS

Despite the strong intuitions and theoretical traditions of linguists, psychologists, and speech scientists, spoken language does not appear to straightforwardly consist of linear sequences of discrete, idealized abstract, context-free symbols such as phonemes or segments. This discovery begs the question, however; how does speech convey equivalent information across talkers, dialects, and contexts? And how do language users mentally represent both the variability and constancy in the speech they hear?

New directions in research on speech perception include theories of exemplar-based representation of speech and experiments designed to discover the specificity, generalized application, and flexibility of listeners' perceptual representations. Efforts to focus on more ecologically valid tasks such as spoken word recognition also promise fruitful progress in coming years, particularly those which provide tests of theoretical and computational models. In speech production, meanwhile, the apparent convergence of acoustic and articulatory theories of representation points to the emerging potential for exciting new lines of research combining their individual successes. At the same time, more and more speech scientists are turning their research efforts toward variability in speech, and what patterns of variation can reveal about speakers' language-motivated control and linguistic knowledge.

Perhaps the greatest potential for progress and discovery, however, lies in continuing to explore the behavioral and neurobiological links between speech perception and production. Although made separate by practical and conventional scientific considerations, these two processes are inherently and intimately coupled, and it seems that we will never truly be able to understand the human capacity for spoken communication until they have been conceptually reunited.

REFERENCES

1. Hockett CF. *A Manual of Phonology*. Baltimore: Waverly Press; 1955.
2. Chomsky N, Miller GA. Introduction to the formal analysis of natural languages. In: Luce RD, Bush RR, Galanter E, eds. *Handbook of Mathematical Psychology*. New York: John Wiley & Sons; 1963, 269–321. DOI:10.1016/S0010-0277(98)00034-1.
3. Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. *Psychol Rev* 1967, 74:431–461. DOI:10.1037/h0020279.
4. Potter RK. Visible patterns of sound. *Science* 1945, 9:463–470.
5. Cooper FS. Research on reading machines for the blind. In: Zahl PA, ed. *Blindness: Modern Approaches to the Unseen Environment*. Princeton: Princeton University Press; 1950, 512–543.
6. Cooper FS, Borst JM, Liberman AM. Analysis and synthesis of speech-like sounds. *J Acoust Soc Am* 1949, 21:461–461.
7. Liberman AM, Mattingly IG. The motor theory of speech perception revised. *Cognition* 1985, 21:1–36.
8. Goldstein, L, Fowler, CA. Articulatory phonology: a phonology for public language use. In: Schiller NO, Meyer AS, eds. *Phonetics and Phonology in Language Comprehension and Production*. Berlin: Mouton de Gruyter; 2003, 159–207.
9. Liberman AM, Harris KS, Hoffman HS, Griffith BC. The discrimination of speech sounds within and across phoneme boundaries. *J Exp Psychol* 1957, 54:358–368.
10. Stevens KN. Features in speech perception and lexical access. In: Pisoni DB, Remez RE, eds. *Handbook of Speech Perception*. Malden: Blackwell Science; 2005, 125–155.
11. Niedzielski N. The effect of social information on the perception of sociolinguistic variables. *J Lang Social Psychol* 1999, 18:62–85.
12. Johnson KA. Speaker normalization in speech perception. In: Pisoni DB, Remez RE, eds. *Handbook of Speech Perception*. Malden: Blackwell Science; 2005.

13. Ganong WF. Phonetic categorization in auditory word perception. *J Exp Psychol* 1980, 6:110–125. DOI:10.1037/0096-1523.6.1.110.
14. Howes D. On the relation between the intelligibility and frequency of occurrence of English words. *J Acoust Soc Am* 1957, 29:296–305.
15. Oldfield RC. Things, words and the brain. *Q J Exp Psychol* 1966, 18:340–353.
16. Goldiamond I, Hawkins WF. Vexierversuch: the log relationship between word-frequency and recognition obtained in the absence of stimulus words. *Journal of Experimental Psychology*. 1958, 56:457–463.
17. Luce PA, Pisoni DB. Recognizing spoken words: the neighborhood activation model. *Ear Hearing* 1998, 19:1–36.
18. Luce PA, Goldinger SD, Auer ET, Vitevitch MS. Phonetic priming, neighborhood activation, and PARSYN. *Perception Psychophys* 2000, 62:615–625.
19. Goldinger SD. Echoes of echoes? An episodic theory of lexical access. *Psychol Rev* 1998, 105:251–279.
20. Nygaard LC, Sommers MS, Pisoni DB. Speech perception as a talker-contingent process. *Psychol Sci* 1994, 5:42–46.
21. Port R. How are words stored in memory? Beyond phones and phonemes. *New Ideas Psychol* 2007, 25: 143–170.
22. Abler WL. On the particulate principle of self-diversifying systems. *J Social Biol Struct* 1989, 12:1–13.
23. Humboldt Wv. *Linguistic Variability and Intellectual Development*. Baltimore: University of Miami Press; 1836.
24. Fromkin VA. Slips of tongue. *Sci Am* 1973, 229: 110–117.
25. Mowrey RA, MacKay IRA. Phonological primitives: electromyographic speech error evidence. *J Acoust Soc Am* 1990, 88:1299–1312.
26. Stevens, KN. 1986., *Models of phonetic recognition II: a feature-based model of speech recognition*. Montreal Satellite Symposium on Speech Recognition (Twelfth International Congress of Acoustics). 67–68.
27. Jackobson R, Fant G, Halle M. *Preliminaries to Speech Analysis: The Distinctive Features*. Cambridge: MIT; 1952.
28. Chomsky N, Halle M. *The Sound Pattern of English*. New York: Harper and Row; 1968.
29. Fowler CA. Listeners do hear sounds, not tongues. *J Acoust Soci Am* 1996, 99:1730–1741. DOI:10. 1121/1.415237.
30. Borden GJ, Harris KS, Raphael LJ. *Speech Science Primer: Physiology, Acoustics and Perception of Speech*. 4th ed. Baltimore: Lippincott Williams & Wilkins; 2003.
31. Lisker L, Abramson AS, Cooper FS, Schvery MH. Transillumination of the larynx in running speech. *J Acoust Soc Am* 1969, 45:1544–1546. DOI:10.1121/1.1911636.
32. Perkell JS. *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Cambridge: MIT Press; 1969.
33. Nadler, R, Abbs, JH, Fujimora, O. 1987., *Speech movement research using the new x-ray microbeam system*. 11th International Congress of Phonetic Sciences; Tallinn. 221–224.
34. Perkell J, Cohen M, Svirsky M, Matthies M, Garabieta I, et al. Electromagnetic mid-sagittal articulometer (EMMA) systems for transducing speech articulatory movements. *J Acoust Soc Am* 1992, 92: 3078–3096.
35. Stone M, Lundberg A. Three-dimensional tongue surface shapes of English consonants and vowels. *J Acoust Soc Am* 1996, 99:3728–3737.
36. Stone M. A guide to analysing tongue motion from ultrasound images. *Clin Linguist Phon* 2005, 19: 455–501.
37. Goldstein L, Pouplier M, Chen L, Saltzman E, Byrd D. Dynamic action units in speech production errors. *Cognition* 2007, 103:386–412. DOI:10.1016/j.cognition.2006.05.010.
38. Browman C, Goldstein L. Tiers in articulatory phonology, with some implications for casual speech. In: Kingston J, Beckman M, eds. *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. Cambridge: Cambridge University Press; 1990, 341–376.
39. Fowler CA. An event approach to the study of speech-perception from a direct realist perspective. *J Phon* 1986, 14:3–28.
40. Rizzolatti G, Fadiga L, Gallese V, Fogassi L. Premotor cortex and the recognition of motor actions. *Cogn Brain Res* 1996, 3:131–141.
41. Lindblom, BEF. Explaning phonetic variation: a sketch of H&H theory. In: Hardcastle HJ, Marchal A, eds. *Speech Production and Speech Modeling*,, *NATO ASI Series D: Behavioural and Social Sciences*. Dordrecht: Kluwer Academic Press; 1990, 403–439.
42. Kent RD, Netsell R. Effects of stress contrasts on certain articulatory parameters. *Phonetica* 1971, 24: 23–44.
43. de Jong K. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J Acoust Soc Am* 1995, 97:491–504.
44. de Jong K. Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *J Phon* 2004, 32:493–516. DOI:10.1016/j.wocn.2004.05.002.
45. Moore RK. Spoken language processing: piecing together the puzzle. *Speech Commun* 2007, 49: 418–435.

46. Wilson EO. *Consilience: The Unity of Knowledge*. New York: Knopf; 1998.
47. Denes P, Pinson E. *The Speech Chain*. Garden City: Anchor Press/Doubleday; 1963.
48. Mattingly, IG, Liberman, AM. Specialized perceiving systems for speech and other biologically significant sounds. In: Edelman GM, Gall E, Cowan WM, eds. *Auditory Function: Neurological Bases of Hearing*: New York, NY: John Wiley & Sons; 1988, 775–793.
49. Kluender, KR. Speech perception as a tractable problem in cognitive science. In: Gernsbacher M, ed. *Handbook of Psycholinguistics*. San Diego: Academic Press; 1994, 172–217.
50. Diehl RL, Lotto AJ, Holt LL. Speech perception. *Annu Rev Psychol* 2004, 55:149–179.
51. Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am* 1954, 26: 212–215.
52. Geers A, Brenner C. Speech perception results: audition and lipreading enhancement. *Volta Rev* 1994, 96:97–108.
53. Lachs L, Pisoni DB, Kirk KI. Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: a first report. *Ear Hearing* 2001, 22:236–251.
54. Bergeson, TR, Pisoni, DB. Audiovisual speech perception in deaf adults and children following cochlear implantation. In: Calvert GA, Spence C, Stein BE, eds. *The Handbook of Multisensory Processes*. Cambridge: MIT Press; 2004, 749–772.
55. McGurk H, McDonald J. Hearing lips and seeing voices. *Nature* 1976, 264:746–748.
56. Fowler CA, Dekle DJ. Listening with eye and hand: crossmodal contributions to speech perception. *J Exp Psychol* 1996, 17:816–828.
57. Rosenblum LD, Schmuckler MA, Johnson JA. The McGurk effect in infants. *Percept Psychophys* 1997, 59:347–357.
58. Gregory SW, Webster S. A nonverbal signal in voices of interview partners effectively predicts communication accomodation and social status perceptions. *J Pers Soc Psychol* 1996, 70:1231–1240.
59. Lane H, Tranel B. The Lombard sign and the role of hearing in speech. *J Speech Hearing Res* 1971, 4:677–709.
60. Houde JF, Jordan MI. Sensorimotor adaptation in speech production. *Science* 1998, 279:1213–1216.
61. Houde JF, Jordan MI. Sensoriomotor adaptation of speech I: compensation and adaptation. *J Speech Hearing Res* 2002, 45:295–310.
62. Villacorta VM, Perkell J, Guenther FH. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J Acoust Soc Am* 2007, 122:2306–2319. DOI:10.1121/1.2773966.
63. Levelt WJ, Roelofs A, Meyer AS. A theory of lexical access in speech production. *Behav Brain Sci* 1999, 22:1–75.
64. Kawato M. Adaptation and learning in control of voluntary movement by the central nervous system. *Adv Robot* 1989, 3:229–249.
65. Halle M, Stevens KN. Speech recognition: a model and a program for research. *IRE Trans Information Theory* 1962, 8:155–159. DOI:10.1109/TIT.1962.1057686.
66. Rizzolatti G, Sinigaglia C. *Mirrors in the Brain: How our Minds Share Actions and Emotions*. Oxford: Oxford University Press, 2008.
67. Fadiga L, Craighero L, Buccino G, Rizzolatti G. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur J Neurosci* 2002, 15:399–402. DOI:10.1046/j.0953-816x.2001.01874.x.
68. Cooper FS, Liberman AM, Borst JM. The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proc Natl Acad Sci* 1951, 37:318–325.
69. Liberman AM. Some results of research on speech perception. *J Acoust Soc Am* 1957, 29(1):117–123.
70. Cleary M, Pisoni DB. Speech perception and spoken word recognition: Research and theory. Pp 499–534 in Goldstein EB, ed. Handbook of Perception. 2001. Blackwell Publishers Inc., Malden, MA.

FURTHER READING

Speech Perception

Johnson K, Mullinex JW, eds. *Talker Variability in Speech Processing*. San Diego: Academic Press; 1997.
Pisoni DB, Remez RE, eds. The Handbook of Speech Perception. Malden: Blackwell Sciece; 2005.

Acoustic & Articulatory Phonetics

Lagefoged P. *A Course in Phonetics*. 5th ed. Boston: Thomson/Wadsworth; 2005.
Stevens KN. *Acoustic Phonetics*. Cambridge: MIT Press; 1998.

Perception/Production Links

Bradlow AR, Pisoni DB, Akahane-Yamada R, Tohkura Y. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *J Acoust SocAm* 1997 101: 2299–2310.
Iacoboni, M. *Mirroring People: The New Science of How we Connect with Others*. New York: Farrar, Straus and Giroux; 2008.

