Speech Perception

ATHENA VOULOUMANOS

Abstract

Speech perception is the process by which listeners presented with a distribution of audible frequencies modulated in amplitude (loudness) and spectral (the frequency set) content across time turn this sound into a coherent unit of perception that is interpreted as language. Classic studies established that speech is not perceived by simply mapping sets of invariant acoustic properties onto different speech sounds. In fact, speech perception is robust even when the acoustic signal has been dramatically distorted. Current approaches focus on understanding how we perceive speech by investigating the neural basis of processing different physical aspects of the speech signal, the encoding of acoustic information in the speech signal at different time scales, the developmental of speech perception, and the multimodal representation of speech. Understanding how humans perceive speech will require the expertise of psychologists, neuroscientists, and engineers.

INTRODUCTION

Billions of humans use spoken language as a means of communication; speech perception is a critical, effortless daily activity. Speech is a sound consisting of a distribution of audible frequencies modulated in amplitude (akin to loudness) and spectral content (the set of frequencies) across time. Speech perception is the process by which listeners turn these modulated audible frequencies into a coherent unit of perception (or percept) that can be interpreted as language.

FOUNDATIONAL RESEARCH

Two fundamental questions guided early speech perception research. What do we perceive when we perceive speech? And how do we perceive it? The answer to these questions originally seemed obvious: speech sounds appeared to be characterized by specific acoustic properties (which needed to be discovered), and perceived by mapping these unique acoustic properties onto each discrete speech sound. But these assumptions were challenged by

Emerging Trends in the Social and Behavioral Sciences. Edited by Robert Scott and Stephen Kosslyn.
© 2015 John Wiley & Sons, Inc. ISBN 978-1-118-90077-2.

the early discovery of four remarkable properties of speech perception: the lack of acoustic invariance, categorical perception, the lack of segmentability, and the simultaneous perception of speech and nonspeech. Each of these points is considered in what follows.

Initial attempts to identify the set of invariant (unchanging) acoustic properties that allow us to perceive distinct speech sounds, for example, to identify the specific acoustic properties that characterize a "d" sound, revealed a major complication. No unique combination of acoustic properties could specify two different "d"s as being one and the same sound in different speech contexts. Specifically, the acoustic properties of the "transitions" (rapid frequency changes) that help specify the two "d" sounds in the syllables "da" and "di," even syllables spoken by the same speaker, were completely different. In fact, a given speech sound was found to be acoustically different depending on the specific speech sounds that come before and after it, so-called coarticulatory cues. Speech was discovered to lack acoustic invariance—coarticulation between neighboring speech sounds gave rise to too many acoustic differences and not enough similarities between the same speech sound in different speech contexts to identify a given speech sound based on unique (simple) acoustic properties. This surprising finding led some to argue that rather than perceiving speech directly through acoustic properties, speech perception is grounded in the (less variable) articulatory gestures that generated the speech signal. But invariant gestures for different speech sounds also proved elusive.

One of the ways people were proposed to adjust for the variability of speech sounds is through categorical perception. Categorical perception is the process that allows us to experience a continuously varying property as two or more discrete categories. For instance, although one of the properties that differentiates a "da" from a "ta"—voicing—varies continuously, listeners either perceive a "da" or a "ta" sound, never a sound in between the two. This process allows us to ignore differences within a category boundary and focus on differences between category boundaries. Perceiving discrete types of things rather than continuously variable things makes it computationally easier to think about and act on the world. Moreover, the ability to perceive speech sounds (especially consonants) categorically was thought to be uniquely human. But it was quickly shown that categorical perception was not unique to humans when, in rapid succession, chinchillas, budgerigars, and rhesus monkeys were shown to perceive human speech sounds categorically. Moreover, recent work has cast doubt on whether humans do, in fact, ignore within-category variation to the degree previously believed.

At the same time, even a simple consonant-vowel sound such as "da" cannot be divided into its component consonant and vowel segments. No matter how the syllable is split, it either sounds similar to the full consonant-vowel syllable or it sounds very like a nonspeech chirp or buzz, suggesting that syllables, rather than segments, might be more natural percepts. Counter to intuition, the signal does not have acoustic properties that mark boundaries between segments that we perceive as different (such as a vowel and a consonant), which is to say that speech lacks "segmentability"—it lacks clearly marked boundaries between what we hear as distinct speech sounds. (This problem persists at higher levels of perception. Words also lack clear silences to mark boundaries between words, in contrast to the blank spaces that mark word boundaries in written language.)

A phenomenon known as *duplex perception* showed that a sound could be simultaneously perceived as speech and a nonspeech, by isolating and manipulating portions of the sound through a dichotic listening task in which different parts of the sound are presented to each ear. For example, if part of the transition (rapid frequency change) that differentiates a "ga" from a "da" is removed from the syllable and presented to one ear, it sounds similar to a chirp. If, at the same time, the rest of the syllable is presented to the other ear, the listener simultaneously hears the nonspeech chirp and the speech syllable. The ability to perceive a single sound simultaneously as both speech and nonspeech suggests there may be two distinct and differentiable ways of perceiving a sound: auditory—the way in which most sounds are perceived—and phonetic—a special way in which only speech sounds are perceived. However, later studies showed that duplex perception is also evident for nonspeech, with a clever study using the sound of slamming doors in which an analogous dichotic manipulation can induce listeners to simultaneously perceive a wooden door and a metal door.

Dichotic listening tasks also revealed that people were better at recognizing speech sounds when speech was presented to their right ear than their left, the so-called right-ear advantage. Because of the neural wiring of the auditory system, the right ear forms stronger connections to the left hemisphere of the brain, and this right-ear speech perception advantage was taken as evidence for specialized language processing in the left hemisphere. This behavioral work complemented the classic neurological work of Broca and Wernicke, which showed that focal lesions in the left hemisphere caused aphasias, disorders of speech and language.

These early discoveries highlight the fact that speech perception consists of more than perceiving and grouping sets of unique acoustic properties into discrete speech percepts. They fed into major theories of speech perception, which differ on several major questions that inform current debates. (i) What are the units of speech perception, acoustic or gestural, segmental or syllabic? (ii) Is speech perception a specialized mode of perception with unique processes and representations, or is speech perceived as just another environmental sound subject to general auditory processes and mapping onto general auditory representations? (iii) Does speech perception rely on specialized neural circuitry, or general-purpose neural substrates?

CUTTING-EDGE RESEARCH

Current approaches address major questions in speech perception by focusing on the nature of the speech signal as broadband frequencies subject to amplitude and spectral modulations. One current idea about how we turn time-varying frequencies into speech percepts is that modulations at different time scales reflect different aspects of the speech signal and are analyzed separately, using different areas of the brain. The rapid modulations that are important for perceiving segments such as consonants happen over short time scales (25–50 ms) that are processed in the left hemisphere. The slower modulations that contribute to the perception of syllables happen over longer time scales (200–300 ms) that are processed in the right hemisphere. This hemispheric division of labor of information encoded at different time scales might even be present in humans from birth. This may provide one answer to how we perceive the linguistic segments of speech: Information at different time scales in speech may map onto neural circuitry that can decode information at these different time scales.

One of the remarkable things about speech perception is how very good humans are at it. Adults can be induced to perceive a speech sound as speech even if we replace the smear of frequencies in the signal with three simple sine waves that track the main frequency changes in time, even if we rotate the sound around 2000 Hz so that all the energy that is usually at high frequencies is now at low frequencies and vice versa, even if we reverse the sound waves every 50 ms, and even if we band-pass filter the speech signal, removing much of its spectral information including formant (concentrations of energy around a particular frequency) transitions. Humans show a remarkable ability to perceive intelligible speech even in severely degraded signals. Human infants might also be able to perceive atypical speech as speech when other visual cues, such as the presence of a static human face, provide a supporting context. Examining how humans perceive a sound as intelligible speech, and the way intelligibility may be supported by neural responses reflecting the amplitude and spectral modulations in speech is a topic of current study.

Some important speech perception biases are present from birth. Even neonates discriminate between and prefer speech to many nonspeech sounds including backwards speech (which has scrambled temporal properties that make it unintelligible), and synthetic sounds that mimic some properties of speech by using sinusoidal waves to track the main spectral and temporal changes across time of natural speech sounds. This bias for speech is somewhat broadly tuned, as newborns appear to listen to other vocalizations such as rhesus monkey calls as much as speech. By 3 months, human infants prefer listening to speech, even speech in a foreign language they have not previously heard, to many other sounds including environmental sounds (running water, bells), rhesus monkey calls, and even human emotional vocalizations such as laughter. Important questions remain on what specific properties allow infants to recognize a sound as speech and prefer it to other sounds.

Although we think of speech as primarily an auditory medium, speech perception is inherently multimodal. There is a significant visual component to speech perception. Not only can adults and infants match vowels, gender, and affect in voices and faces but classic studies also showed that when adults and infants hear a voice producing "ba" and see a silent face producing "ga," they integrate the information to perceive a sound that is never actually specified by either stimulus, a "da." Visual information alone is sufficient to allow infants and adults to detect when a bilingual speaker switches to a different language. But vision is not the only integrated sense. By applying a puff of air to irrelevant parts of the body such as the ankle, adults perceive a "pa" sound where previously they had perceived a "ba" because the puff of air provides tactile information consistent with aspiration (an expulsion or air) produced during a "pa" sound. Even infants' perception of speech is affected by their own lip movements. By sucking on differently shaped objects, infants produced lip movements consistent with "ee"-like lip spreading or "oo"-like lip rounding, and these lip movements affected their perception of those vowels. Speech processing is inherently multimodal from early in infancy. The multimodal representation of speech including visual, tactile, and somatosensory information is a key area of current study.

Many of these approaches tackle how we perceive speech by building up from acoustic properties of signals, but another approach examines the influences of higher level information from words, syntax, semantics, and cognition on how we perceive individual speech sounds. When a portion of a speech sound is replaced by a nonspeech sound such as a cough, listeners are very poor at detecting which speech sound has been replaced, suggesting that they were able to automatically "restore" the missing speech portion. At the same time, when a particular speech segment is made ambiguous, say, by being halfway between a "b" and a "p," listeners perceive it unambiguously as one or the other depending on whether it forms a real word ("pork" rather than "bork"), or whether the sentence makes sense with one rather than the other ("I ate a pear for dessert"). These observations suggest that speech perception is not based entirely on the acoustic information in the signal but is instead modulated by the higher order linguistic context. Speech perception is a dynamic process that allows us to perceive, segment, and recognize words from different speakers and in different contexts. Current research examines whether the mechanisms that support speech perception in simple perceptual tasks contribute to our understanding of dynamic spoken language.

KEY ISSUES FOR FUTURE RESEARCH

Speech perception has been studied intensively since the 1930s. But key aspects of how we perceive speech are incompletely understood: What do we perceive when we perceive speech? Do we perceive speech differently than we perceive other sounds? How is speech perception instantiated in neural circuitry and do physical aspects of the auditory signal map neatly onto distinct neural processes or regions? Is the perception of speech different in humans than in other animals? Advances in understanding human speech perception will likely require the collaboration of psychologists specializing in perception, neuroscientists specializing in functional neuroanatomy, and engineers specializing in digital signal processing. The way in which humans perceive speech is still very much a matter of debate.

FURTHER READING

Blumstein, S. E., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, *10*(1–3), 25–32.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–7.

Liberman, A., & Whalen, D. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5), 187–196.

McGettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: What’s wrong, what’s right and what’s left? *Trends in Cognitive Sciences*, *16*(5), 269–276.

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology*, *62*, 49–72.

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science*, *10*(2), 159–164.

ATHENA VOULOUMANOS SHORT BIOGRAPHY

Athena Vouloumanos is a Professor of Psychology at New York University where she directs the NYU Infant Cognition and Communication Lab. Vouloumanos’s research addresses fundamental questions on speech perception, language acquisition, and the development of communication. With funding from the Natural Sciences and Engineering Research Council of Canada, the Fonds québécois de recherche sur la société et la culture, and the National Institutes of Health, Vouloumanos has been exploring the linguistic and cognitive abilities of adults and young infants, including newborns and infants at high risk for autism spectrum disorder. Her findings have been published in journals such as *Science, Cognition, Cognitive Science, Child Development*, *Developmental Science*, and the *Proceedings of the National Academy of Sciences.*

Personal webpage: http://www.psych.nyu.edu/vouloumanos/
Laboratory webpage: http://www.psych.nyu.edu/niccl/

RELATED ESSAYS

Mental Models *(Psychology)*, Ruth M. J. Byrne
Misinformation and How to Correct It *(Psychology)*, John Cook *et al*.
Construal Level Theory and Regulatory Scope *(Psychology)*, Alison Ledgerwood *et al.*
Resource Limitations in Visual Cognition *(Psychology)*, Brandon M. Liverence and Steven L. Franconeri
Neural and Cognitive Plasticity *(Psychology)*, Eduardo Mercado III
Attention and Perception *(Psychology)*, Ronald A. Rensink