CYNTHIA CUMMINGS

Emily Myers is an associate professor in the department of speech, language, and hearing sciences and the department of psychological sciences at the University of Connecticut in Storrs.

From SOUND to MEANING

Emily B. Myers

Culture and experience contribute to the process that translates a complex acoustic stimulus into an intelligible message.

A man waits on a crowded train platform. His cell phone buzzes, and he recognizes his daughter's number. He answers, she reminds him to pick up milk on the way home, he says OK, and they chat briefly. The interaction might take only a minute and would be unremarkable for both father and daughter. Yet the transmission of even a simple message requires a multitude of physical and psychological processes that are phenomenally complex and as yet not fully understood. During the past decade or so, psychologists, neuroscientists, and acousticians have made tremendous strides in understanding the quasi-magical process of putting your thoughts into someone else's head.

The get-milk message began miles away from the train platform, when the daughter drew in a breath of air and began to speak. Speech production is an invisible ballet that requires precise and rapid coordination of the many muscle groups that control the lips, tongue, jaw, larynx, and respiration. The daughter's coordinated muscle movements, called speech gestures, result in an acoustic signal containing multiple acoustic cues that ultimately enable her father to decode the signal. The acoustic signal is transmitted via a cell phone, which compresses and filters it; the phone especially distorts the higher frequencies that allow listeners to distinguish the "s" sound in *sack* from the "sh" sound in *shack*. Back at the train station, the signal emerges from a cell phone, travels through the air, enters the father's ear, and impinges on his cochlea. He must sort the signal emanating from the cell phone from the screech of train brakes and the conversations of other commuters on the platform.

Next, the speech message is processed by the father's nervous system. Neural signals that represent the sound with high fidelity are transmitted along the auditory nerve, ascend through the brain stem, pass through the thalamus, and arrive in the cerebral cortex, where the language centers of the brain are located. There the message is further processed. The sounds in the speech stream are compared with the sounds that the father has learned over the course of his lifetime; the brain's task is to match the speech stream with words in the father's lexicon, or mental dictionary. Further transformations are required to turn strings of words into meaningful sequences such as "Don't forget to pick up milk."

What does the father actually perceive during the conversation? Most likely he is aware of only a few pieces of information—for example, that he is listening to his daughter's voice and that he's going to have to make a detour on the way home. Perception is built on such knowledge, and each step in the complex chain leading to that knowledge is worthy of

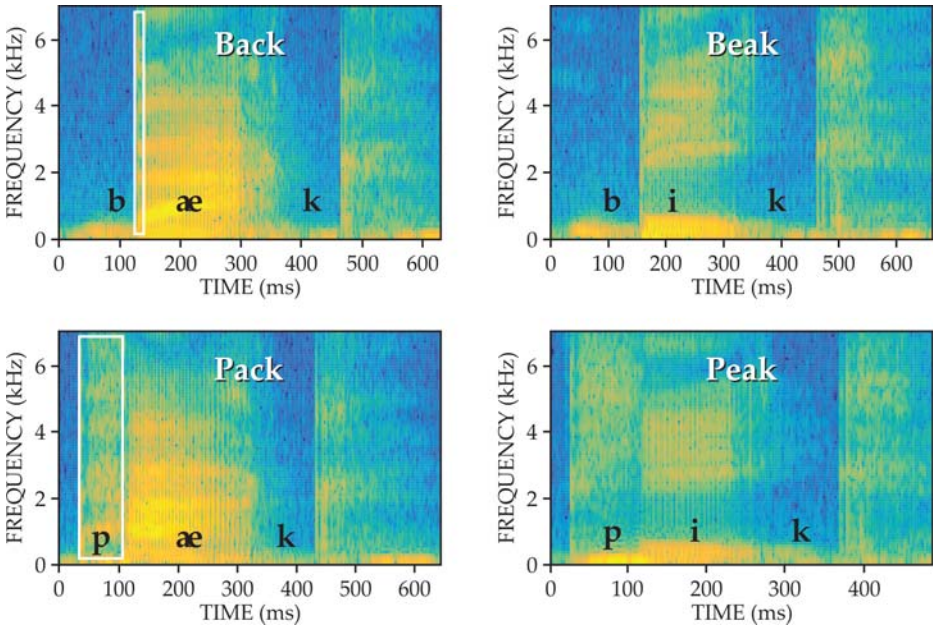

FIGURE 1. THE TIME SEPARATING the vowel sound from the burst of the initial consonant—the voice onset time (VOT)—is distinctly different for words like *back* and *pack* or *beak* and *peak*. The four spectrograms show the frequencies that are most intense at a given time (yellow is most intense; blue, least). The VOT is indicated by the width of the rectangles in the *back* and *pack* panels; it's about 20 ms for the "b" sound and near 80 ms for "p."

discussion. In this article, though, I concentrate on what happens once the signal arrives in the brain—that is, on how a listener takes a processed auditory signal and maps it to a meaningful message.

When light sounds right

Speech is made up of units that linguists call phonemes, abstract units of perception and production that, when swapped, produce a change in the word. Linguists use forward slash marks to denote phonetic symbols: For example, /l/ denotes the beginning sound in the word *light*. In the English language, /l/ and /r/ are different phonemes because replacing the /l/ sound in *light* with an /r/ sound results in a new word, *right*. Japanese, by contrast, has no instances in which changing an /l/ phoneme to an /r/ will result in a new word. For that reason, many Japanese listeners find it difficult to hear the difference between words like *lock* and *rock*.

Some languages have distinct phonemes that are not distinguished in English. Hindi, for example, has two /d/-like sounds: one, the dental /d̪/, made with the tongue placed behind the teeth and the other, the retroflex /ɖ/, made with the tongue curled back along the hard palate. In Hindi, those two sounds, when swapped, can change the meaning of a word. People who use English as their native language simply perceive them as two slightly different varieties of /d/. Therefore, the dental and retroflex sounds make up a phonemic contrast in Hindi but not in English, whereas the /l/ and /r/ sounds are contrasting in English but not in Japanese.

Speech sounds are differentiated with the help of multiple acoustic cues. For instance, the stop sounds that begin words like *pack* and *back* are distinguished primarily in terms of the timing of two articulatory movements. Both sounds are made with an initial closure of the lips. What differs is the time—on the order of tens of milliseconds—between the pop open of the lips and the beginning of the vowel sound. Figure 1 illustrates that lag, which is called voice onset time (VOT). The same type of time lag enables listeners to distinguish between words like *duck* and *tuck* or *goal* and *coal*.[1]

Different acoustic properties distinguish other speech sounds. For instance, vowels are primarily determined by the patterns of energy maxima, or formants, in the frequency spectrum. Fricative speech sounds such as the /s/ in *sack* and the /ʃ/ ("sh") in *shack* are determined by a combination of factors that include their duration, their amplitude, and the concentration of energy across the frequency spectrum (see figure 2). In fact, most of the time people like the father on the train platform have to assemble multiple pieces of information to determine the identity of the sounds they are hearing, a process known as cue weighting.[2]

The lack of invariance problem

Even after the father has extracted the acoustic cues in his daughter's request that he pick up milk, his task is hardly over. His next major challenge is that no two utterances of a particular phoneme—for example, the /p/ in *pick*—are identical. His daughter might sometimes produce her /p/ sound with a VOT of 70 ms, sometimes with a VOT of 90 ms. His wife might produce /p/ with relatively shorter VOTs, even as her sister tends to pronounce the sound with longer VOTs.[3] Add to that the fact that the sounds abutting the /p/ will bleed into the consonant; the /p/ sound in *pick*, for example, is acoustically different from the /p/ sound in *poke*. An infinite number of acoustic patterns can map into a single speech sound. Ostensibly, that "lack of invariance" problem presents an enormous challenge to the father. It is not enough for him simply to note the acoustic cues of speech. He also must figure out how to categorize the sound he is hearing on the basis of what he knows about the talker (she's his daughter), speech rate (she's speaking quickly), coarticulatory context (the /p/ in *pick* is next to an /I/ sound), and other information (*pick* makes sense in context, whereas *bick* does not).

To convince yourself of how difficult it can be to translate acoustic cues into words, try any commercially available speech-recognition interface such as Apple's Siri or Amazon's Alexa. Say a single, monosyllabic word such as *pack* clearly and slowly, and the system is reasonably likely to identify it correctly. However, if you repeat the word *pack* quickly, you may get a multitude of responses; in different tries, Siri thought I was saying *back*, *beck*, *talk*, and *part*.

As noted earlier, the human speech system does not deliver the entire auditory content to the point of conscious awareness. Rather, we usually can perceive only acoustic differences that matter for meaning. Consider, for example, a series of sounds that lie along a continuum between two speech categories in English—say, /d/ and /t/. People whose native language is English will easily perceive the difference between sounds that fall into one class or the other. Those same listeners, however, will struggle to hear a difference between two examples of the same sound—for example, two types of /d/ sound—that have the same degree of acoustic distinctiveness as /d/ and /t/ sounds that are readily distinguished. The tendency of listeners to perceptually collapse sounds in the same category leads to difficulties in distinguishing the Hindi /ḍ/ and /ɖ/, two phonemes that fall into the /d/ category in English. (Native English listeners can confirm this for themselves with the sound files available online.) That phenomenon, known as categorical perception, may be in place to help our brain's limited resources focus on only the most important aspects of the speech signal—what is the message, and who is doing the talking.[4]

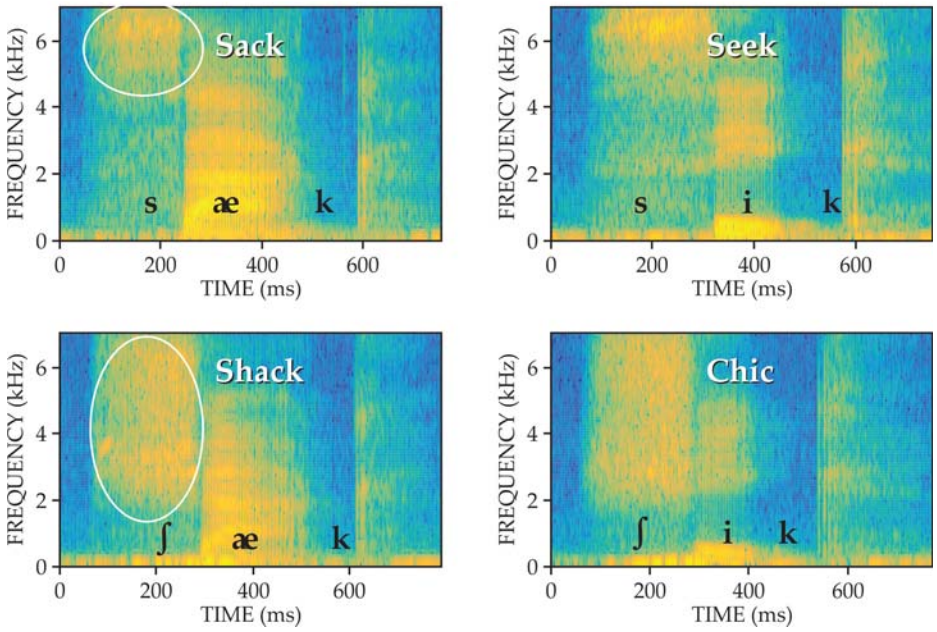

FIGURE 2. FREQUENCY RANGE influences phoneme perception. The four spectrograms show the frequencies that are most intense at a given time (yellow is most intense; blue, least). As the ovals indicate, for the "s" sound in *sack*, the average frequency is higher and the distribution is narrower than for the "sh" sound in *shack*. Those features persist when the vowel sound is changed, but as the right panels comparing the words *seek* and *chic* attest, altering the vowel sound significantly changes the details of the acoustic cues.

So what we hear is not what we perceive. The pressure waves that impinged on the commuting father's cochlea were full of details that he couldn't tell you if you asked him—how long was the VOT for that stop sound? Were the formants for the vowel close together or spaced far apart? Studies suggest that the brain encodes both the fine-grained acoustic details of the speech signal and the information about the identity of the speech sound itself (Is it a /p/ or is it a /b/?).

Regions in the superior temporal gyrus, a part of the brain that specializes in auditory processing, respond to the complex acoustic landscape of speech sound. They also show sensitivity to tiny acoustic differences that the father may not be able to consciously perceive and that may not even be important for understanding the message.[5,6] As the neural processing advances away from his superior temporal gyrus to other areas in the temporal lobe and toward left frontal brain areas, the representation of sounds appears to lose some of the fine-grained acoustic detail. Instead, it seems to represent something closer to what he actually is aware of hearing.[7] Figure 3 shows the locations in the brain of both the temporal-lobe system that can access all the acoustic complexity in the signal and the frontal system that discards that detail in favor of preserving the things that are important for meaning. Having both those systems may allow us to ignore the minute acoustic variation in the signal while still processing that information in case it is relevant for other purposes.

Entrenchment and flexibility

Newborn infants, as psychologist Peter Eimas and others showed, can detect differences between most, if not all, of the sound contrasts in the world's languages.[8,9] Yet over the first year of life, babies begin to ignore sound contrasts that are not represented in their native language and to preserve those that are found in their language. Patricia Kuhl and other scientists have called that process "perceptual narrowing." By the time they reach adulthood, people in an English-speaking environment will struggle to hear the difference between sounds like the dental and retroflex speech contrast that is used in Hindi, a contrast that poses no problem for adults who were raised in India by Hindi-speaking parents. For reasons that are not fully understood, by adulthood we have become perceptually entrenched—that is, we do not appear to show the flexibility in learning new speech sounds that we had as children.

Perceptual entrenchment is easy to observe. Many of us, for example, are acquainted with excellent speakers of English who learned the language late in life and retain strong traces of their native language. It is rare to speak any language like a native speaker if you have learned that language after some critical juncture sometime around puberty. In addition to perceptual entrenchment, a second obstacle to speaking a new language like a native is so-called motor entrenchment: It may be difficult for an adult to learn the movements of the lips, tongue, and larynx that are necessary for new speech sounds. The same phenomenon observed in accented speech production is present in speech perception as well; we also "listen with an accent," meaning that we often cannot distinguish sounds that

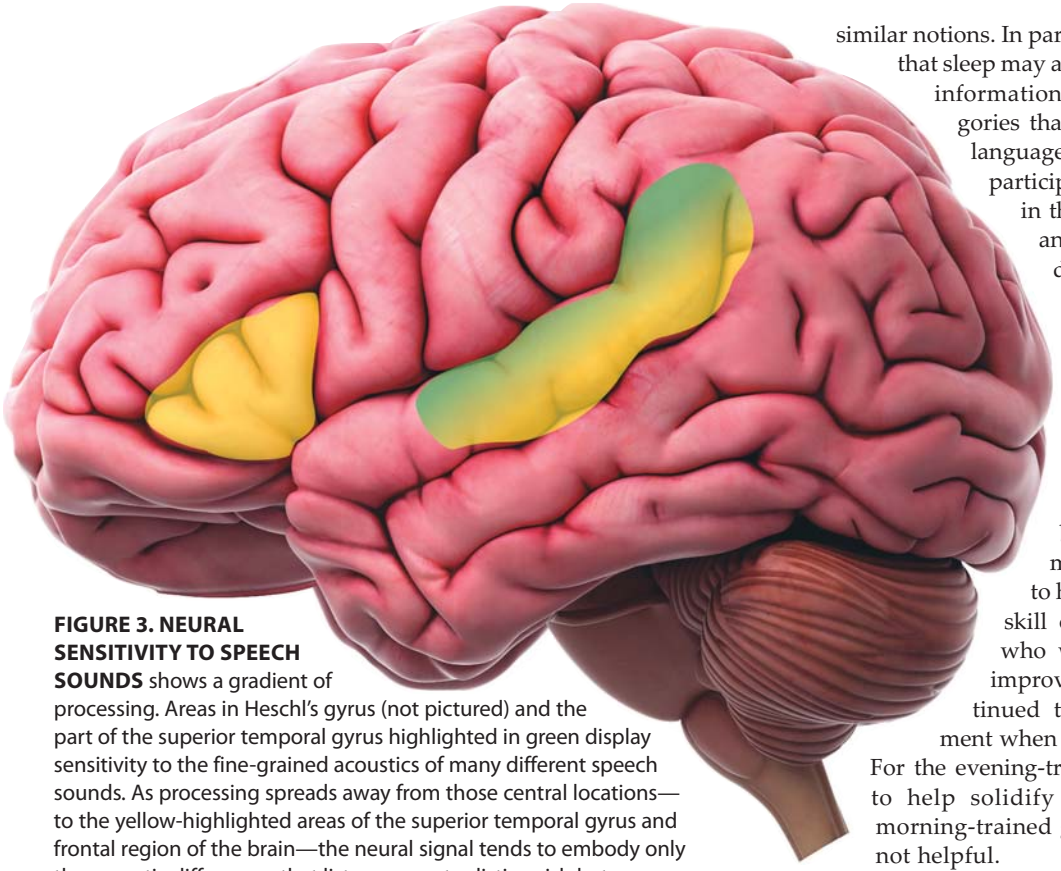

FIGURE 3. NEURAL SENSITIVITY TO SPEECH SOUNDS shows a gradient of processing. Areas in Heschl's gyrus (not pictured) and the part of the superior temporal gyrus highlighted in green display sensitivity to the fine-grained acoustics of many different speech sounds. As processing spreads away from those central locations—to the yellow-highlighted areas of the superior temporal gyrus and frontal region of the brain—the neural signal tends to embody only the acoustic differences that listeners use to distinguish between words. (Adapted from an image by Sebastian Kavlitski.)

aren't part of our native language's repertoire. It may be that our binning of sounds into distinct speech categories is itself an obstacle to learning new sounds.

Adults can learn the sounds of a new language, but with highly varying degrees of success. Attempts to train motivated learners on difficult sound contrasts—for example, efforts to teach native speakers of Japanese to hear differences between /l/ and /r/ sounds—usually show modest gains after many hours of training.[10] Some people learning a new language reach native-like performance, but most fall short of that goal, even after a lifetime's immersion.[11]

Many factors make it hard to learn the sounds of a new language. For instance, as adults, we have many demands on our personal time that make language learning a lower priority than it is for an infant. Differences in motivation, in auditory acuity, in ability to remember speech sounds that we've heard before, and in neural plasticity likely affect how well we will do in picking up a new language. New data from my lab, the Language and Brain Lab at the University of Connecticut, hint at two less-explored factors that may explain some of the variability in adults' speech-sound learning—sleep, and interference from sounds from the native language.

Sleep has various effects related to memory. One is that it facilitates the transfer of learned information from the hippocampus to cortical regions, a move that allows learners to generalize from concrete experiences to abstract categories and also protects learned information from being confused with similar notions. In particular, recent studies suggest that sleep may also help learners "lock down" information about speech-sound categories that are not part of their native language.[12,13] In those investigations, participants came to the lab either in the morning or in the evening and were trained in the Hindi dental–retroflex speech-sound contrast that English listeners usually hear as two varieties of the /d/ sound (see figure 4).

Participants who came in the morning appeared to retain what they had learned over the course of the day, but as tests the following morning revealed, they seemed to have lost their newly acquired skill overnight. In contrast, those who were trained in the evening improved overnight, and they continued to show increased improvement when tested the following evening. For the evening-trained group, sleep appeared to help solidify training, whereas for the morning-trained group, sleep apparently was not helpful.

My colleagues and I speculated that the difference in the two groups' performance lay in participants' exposure to sounds that are similar to the trained sounds. For our specific experiment, we reasoned that participants who were trained in the morning likely were exposed to many examples of the English /d/ sound before they went to sleep, whereas participants who were trained in the evening heard many fewer English /d/ sounds before bed.

We tested our hypothesis by training two groups of participants in the evening. One group was exposed to many examples of the /d/ sound after training; the other group heard many examples of the /b/ sound. As predicted, listeners who heard many /d/ sounds showed less benefit from sleep than those who heard the /b/ sounds. In fact, the evening-trained /d/ group did a poor job of discriminating the new Hindi sounds, much as the morning group had in our previous study. A good deal remains to be learned about the process that led to our results, but our findings hint that learners of a new language pay a real perceptual price when they switch between languages. English-speaking adults who take an Italian class to prepare for a vacation may lose ground when they leave the class and listen to English for the rest of the day—particularly if they hear their native language before the protective effects of sleep.

You hear what you expect to hear

If the above story about the difficulties in perceiving the sounds of a new language painted a pessimistic picture, here's a sunny antidote: In the context of our native language, we have a remarkable ability to use knowledge about what words and sounds are likely to appear to solve difficult perceptual prob-

FIGURE 4. BEFORE TRAINING there is troubleshooting. Sahil Luthra and Pamela Fuhrmeister, University of Connecticut graduate students and members of the Language and Brain Lab, make sure that the equipment for a speech perception experiment functions properly. In the study, people are trained to associate Hindi speech sounds with novel objects on a computer screen. The participants then return to the lab several times to measure how well they remember what they have learned.

lems. Consider the father on the train station platform, chatting with his daughter. Cell-phone service being what it is, his connection may have dropped a few times and cut out bits and pieces of the conversation. Further, noises from the approaching train and the conversations of passengers around him probably obscured certain sounds. The speech perception system is built to fill in those gaps. Even if a whole speech sound such as the /s/ in *Tennessee* is replaced with a cough, listeners will report hearing the full word along with a superimposed cough; interestingly, they don't report the cough as overlapping with the restored sound.[14] The phenomenon, discovered in 1970, is called phoneme restoration.

Listeners attempting to understand ambiguous speech sounds lean heavily on their expectations about which sounds and words are likely to appear. For instance, when they hear an ambiguous sound between a /g/ and a /k/ at the beginning of the syllable *ift*, they will assume that the sound is a /g/, which corresponds to the real word *gift*. In contrast, if that same sound is inserted into the syllable *iss* they will come to the conclusion that they are hearing a /k/ sound, completing the real word *kiss*.[15] Other work shows that people can learn to adapt to speech sounds that are out of the norm. For example, researchers Ann Bradlow and Tessa Bent have demonstrated that with the right kind of experience, listeners can improve their ability to understand accented speech, and they can even generalize what they have learned about an accent to new talkers with the same accent.[16]

The process of mapping speech to meaning is laden with contradictions. On the one hand, understanding speech seems effortless to the listener. On the other, a lot is going on under the hood before the message is delivered. The speech system is plastic, able to adapt to various listening conditions. Yet it is also rigid in the sense that adults struggle to learn the sounds of a new language. Speech scientists have made progress in mapping the brain architecture that allows people to take sound vibrations and turn them into meaning, but we still have much to learn.

REFERENCES

1. L. Lisker, A. S. Abramson, *Word* **20**, 384 (1964).
2. L. L. Holt, A. J. Lotto, *J. Acoust. Soc. Am.* **119**, 3059 (2006).
3. R. M. Theodore, J. L. Miller, D. DeSteno, *J. Acoust. Soc. Am.* **125**, 3974 (2009).
4. A. M. Liberman, K. S. Harris, H. S. Hoffman, B. C. Griffith, *J. Exp. Psychol. Gen.* **54**, 358 (1957).
5. E. B. Myers, S. E. Blumstein, E. Walsh, J. Eliassen, *Psychol. Sci.* **20**, 895 (2009).
6. E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, R. T. Knight, *Nat. Neurosci.* **13**, 1428 (2010).
7. Y.-S. Lee, P. Turkeltaub, R. Granger, R. D. S. Raizada, *J. Neurosci.* **32**, 3942 (2012).
8. P. D. Eimas, E. R. Siqueland, P. Jusczyk, J. Vigorito, *Science* **171**, 303 (1971).
9. J. F. Werker, R. C. Tees, *Annu. Rev. Psychol.* **50**, 509 (1999).
10. A. R. Bradlow, R. Akahane-Yamada, D. B. Pisoni, Y. Tohkura, *Percept. Psychophys.* **61**, 977 (1999).
11. J. E. Flege, I. R. A. MacKay, D. Meador, *J. Acoust. Soc. Am.* **106**, 2973 (1999).
12. F. S. Earle, E. B. Myers, *J. Acoust. Soc. Am.* **137**, EL91 (2015).
13. F. S. Earle, E. B. Myers, *J. Exp. Psychol. Hum. Percept. Perform.* **41**, 1680 (2015).
14. R. M. Warren, *Science* **167**, 392 (1970).
15. W. F. Ganong, *J. Exp. Psychol. Hum. Percept. Perform.* **6**, 110 (1980).
16. A. R. Bradlow, T. Bent, *Cognition* **106**, 707 (2008). PT