Speech Perception*

David Lane

This work is produced by OpenStax-CNX and licensed under the Creative Commons Attribution License 1.0

For most of us, listening to speech is an effortless task. Generally speaking, speech perception proceeds through a series of stages in which acoustic cues are extracted and stored in sensory memory and then mapped onto linguistic information. When air from the lungs is pushed into the larynx across the vocal cords and into the mouth nose, different types of sounds are produced. The different qualities of the sounds are represented in formants, which can be pictured on a graph that has time on the x-axis and the pressure under which the air is pushed, on the y-axis. Perception of the sound will vary as the frequency with which the air vibrates across time varies. Because vocal tracts vary somewhat between people (just as shoe size or height do), one person’s vocal cords may be shorter than another’s, or the roof of someone’s mouth may be higher than another’s, and the end result is that there are individual differences in how various sounds are produced. You probably know someone whose voice is slightly lower in pitch than yours or higher in pitch. Pitch is the psychological correlate of the physical acoustic cue of frequency. The more frequently the vibrations of air occur for a particular sound, the higher in pitch it will be perceived. Less frequent vibrations are perceived as being lower in pitch. When language is the sound being processed, the formants are mapped onto phonemes, which are the smallest unit of sound in a language. For example, in English the phonemes in the word "glad" are /g/, /l/, /æ/, and /d/.

The nature of speech, however, has provided researchers of language with a number of puzzles, some of which have been researched for more than forty years.

NOTE: To demonstrate one of these problems, click here. The waveform you see shows speech as a function of amplitude, which is measured in decibels (dB), and frequency of the sound waves, measured in hertz (Hz). As the cursor passes over the waveform, you may notice various sections that correspond to the words and individual sounds you hear; for example, you can detect where the word "show" begins and where the word "money" ends. After a bit of experimentation, however, you notice that it is difficult to pinpoint precisely where one phoneme ends and another begins. Try to find the "th" sound in the word "the", for example; and where can the "uh" sound in "the" be located? Often the acoustic feature of one sound will spread themselves across those of another sound, leading to the problem of linearity; that is, for each speech sound phoneme, if phonemes were produced one at a time, or linearly, there should be a single corresponding section in the waveform. As "the" shows, however, speech is not linear.

Another problem that investigators have studied is the problem of invariance. Invariance refers to a particular phoneme having one and only one waveform representation; that is, the phoneme /i/ (the "ee" sound in "me") should have the identical amplitude and frequency as the same phoneme in "money". As you can see again, that is not the case; the two differ. The plosives, or stop consonants, /b/, /d/, /g/, /k/, provide particular problems for the invariance assumption.

*Version 2.9: Jun 18, 2008 5:22 pm -0500

1http://creativecommons.org/licenses/by/1.0

2http://psych.rice.edu/mmhtml/language/sPerception/Showmethemoney.html

http://cnx.org/content/m11175/2.9/
The problems of linearity and invariance are brought about by co-articulation, the influence of the articulation (pronunciation) of one phoneme on that of another phoneme. Because phonemes cannot always be isolated in a spectrogram and can vary from one context to another depending on neighboring phonemes, speakers' rate of speech, and loudness, perceptually identifying one phoneme among a stream of others, the process of segmentation, also seems like a daunting task. Theories and models of speech perception have to be able to account for how segmentation occurs in order to provide an adequate account of speech perception. We will discuss some accounts of speech perception below.

Some clues as to how identifying phonemes occurs arise from investigation into the ability to perceive voiced consonants, or consonants in which the vocal cords vibrate. To understand the concept of voicing, say the phoneme, /p/, followed by the phoneme, /b/, while touching your throat. You will feel the vibration of your vocal cords during /b/ but not during /p/. Both of these phonemes are bilabial; that is, they are produced by pressing the lips together, and are released with a puff of air. Since the discriminating difference between these two phonemes relevant to English is in their voicing, the ability to adequately perceive voicing is crucial for an adept listener; for example, as the rate of speech increases, listeners are able to shift their criterion of what constitutes a voiceless phoneme. The criterion shift allows them to accept phonemes that are pronounced with shorter VOT durations. Although shifting criteria during the perception of phonemes may be one process that allows accurate identification of phonemes despite changing conditions, what supports the criterion shifts is still a matter of investigation. These skills effortlessly become highly automatic and are probably acquired and fine-tuned during early childhood, a topic we talk about in infant speech perception (p. 2).

1 Is speech special?

In visual perception, people discriminate among colors based on the frequency of the wave length of light. Low frequencies are perceived as red and high frequencies are perceived as violet.

---

As we move from low to high frequencies, we perceive a continuum of colors from red to violet. Notice that as we move from red to orange, we pass through a middle ground that we call "red orange." Speech

---

2http://www.blackdiamondsound.com/download.htm
3http://psych.rice.edu/nmbtn/language/sPerception/infantIntro_h.html
4http://psych.rice.edu/nmbtn/language/sPerception/infantsucking_h.html
5http://psych.rice.edu/nmbtn/language/sPerception/infantHeaddturn_h.html
6http://psych.rice.edu/nmbtn/language/sPerception/infantlooking_h.html
sounds lie on a physical continuum as well. For example, an important dimension in speech perception is **voice onset time**. This refers to the time between the beginning of the pronunciation of the word and the onset of the vibration of the vocal chords. For example, when you say "ba" your vocal chords vibrate right from the start. When you say "pa" your vocal chords do not vibrate until after a short delay. To see this for yourself, put one of your fingers on your vocal chords and say "ba" and then "pa."

The only difference between the sound "ba" and the sound "pa" is that the voice onset time for "ba" is shorter than the voice onset time for "pa". An important difference between speech perception and visual perception is that we do not hear speech sounds as falling halfway between a "ba" and a "pa." We hear a sound one way or the other. This means that a range of voice onset times are perceived as "ba" and a different range of voice onset times are perceived as "pa". This phenomenon is called **categorical perception** and is very helpful for understanding speech.

The sounds "ba" and "pa" differ on the continuous dimension of voice onset time. The sounds "ga" and "da" also differ on a continuous dimension. However, the continuous dimension for these stimuli is more complex than the dimension of voice onset time (it is called the second formant but that is a little beyond the scope of this text). What is important here is that there is a continuum of sounds from "da" to "ga." The following demonstration uses computer generated speech sounds. Ten sounds were generated in equal steps from "da" to "ga." The experiment uses sounds numbered 1, 4, 7, and 10. Sounds 1 and 4 are both heard as "da" whereas sounds 7 and 10 are heard as "ga." In the task, subjects are presented with a randomly-ordered series of sound pairs and asked, for each pair, to judge whether the sounds are the same or different. Since sounds 1 and 4 are both heard as "da" it should be very hard to tell them apart. Therefore, subjects usually judge these sounds as identical. By contrast, Sound 4 is heard as "da" while Sound 7 is heard as "ga." Since Sound 4 and Sound 7 are on opposite sides of the "categorical boundary" it is easier to hear the difference between these sounds than the difference between Sounds 1 and 4. **This occurs even though the physical difference between Sounds 1 and 4 is the same as the difference between Sounds 4 and 7.** By similar logic, the difference between Sounds 7 and 10 should be hard to hear.

The results from one subject in this demonstration experiment are shown below and can be interpreted as follows: When the comparison was between Sounds 1 and 4, the subject judged them to be different once and the same 4 times. When the comparison was between Sounds 4 and 7 (which cross the border), the subject correctly judged them to be different 5/5 times. Finally, in comparing Sounds 7 and 10, the subject always judged the sounds to be the same. Thus, the only time this subject heard a difference between sounds that were three steps apart was for Sounds 4 and 7.

<table>
<thead>
<tr>
<th>Sound Pair</th>
<th>Judged different</th>
<th>Judged same</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 vs. 4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4 vs. 7</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>7 vs. 10</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1

Not all results are as clear cut as those shown above. Many people need more time to become familiar with the task than is possible in this demonstration. In any case, you should get a sense of how this kind of experiment works.

**NOTE:** Try this categorical discrimination task yourself.²

²http://psych.rice.edu/mmmbn/language/sPerception/dagaSound/catDis.html
variations of /p/ carries no meaning. (There are languages in which two different /p/ sounds are used, and, in such cases, perception would be categorical).

The first experiment to demonstrate categorical perception was conducted by Liberman, Harris, Hoffman and Griffith (1957) [8], and in it they presented consonant-vowel syllables along a continuum. The consonants were stop consonants, or plosives, /b/, /d/, and /g/, followed by /a/; for example, /ba/. When asked to say whether two syllables were the same or different, the participants reported various forms of /pa/ to be the same, whereas /pa/ and /ba/ were easily discriminated.

Another categorical perception task presents two syllables followed by a probe syllable, and participants have to say which of the first two syllables the probe matches. If the first two sounds are from two different categories - for example, /da/ and /ga/ - participants accurately match the probe syllable. If the first two syllables are taken from the same category, however, participants cannot differentiate them well enough to do the matching task, and their performance is at chance.

Does the categorical perception of speech mean that speech is perceived via a specialized speech processor? Kewley-Port and Luce (1984) [5] did not find categorical perception in some non-speech stimuli, indicating that there may be something special about speech.

For there to be a specialized speech processor, categorical perception should occur during the perception of all phonemes. However, Fry, Abramson, Eimas, and Liberman (1962) [3], failed to find categorical perception with a vowel continuum. So, there are vowels and consonants that do not behave the same in that respect. Additionally, chinchillas have been shown to categorically perceive speech, despite their obvious lack of speech-processing mechanism (Kuhl, 1987 [6]).

2 How is speech perceived?

One theory of how speech is perceived is the Motor Theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967 [7]). The motor theory postulates that speech is perceived by reference to how it is produced; that is, when perceiving speech, listeners access their own knowledge of how phonemes are articulated. Articulatory gestures such as rounding or pressing the lips together are units of perception that directly provide the listener with phonetic information. The motor theory can account for the invariance problem; that is, the ways that phonemes are produced and perceived have more in common than the ways they are acoustically represented and perceived.

What would be the evidence that listeners use articulatory features when perceiving speech? Here, an accidental discovery made by two film technicians led to one of the most robust and widely discussed findings in language processing. A researcher, Harry McGurk, was interested in whether auditory or visual modalities are differentially dominant during infants’ perceptual development. To find out, he asked his technician to create a film to test which modality captured infants’ attention. In this film, an actor pronounced the syllable "ga" while an auditory "ba" was dubbed over the tape. Would babies pay attention to the "ga" or the "ba"? The process of making the film, however, led to a surprising finding about adults. The technician (and others) did not perceive either a "ga" or a "ba". Rather, the technician perceived a "da".

In an experiment that formally tested this observation, McGurk and McDonalki (1976) [10] showed research participants a video of a person saying a syllable that began with a consonant formed in the back of the mouth at the velum-that is, a velar consonant, "ga"-while playing an auditory tape of a consonant which is formed in the front of the mouth at the two lips; that is, a bilabial, "ba". When viewers were asked what they heard, like the film technician, they replied "da". Perceiving a "da" was the result of combining articulatory information from both visually and auditorily presented stimuli.

NOTE: You can experience McGurk effect by clicking here.

(To return to the question Harry McGurk originally asked about infants, neither modality seems to have dominance; infants as young as 5-months old take in the visual and auditory information about words in the same way as adults: both influence perception).


http://cnx.org/content/m11175/2.9/
In addition to being interpreted as evidence that listeners perceive phonetic gestures, an account that suggests an explanation based on memory has been raised. Because perceivers have ample experience with both hearing and seeing people speak, they may have built memories of these events that have subsequently become associated with the phoneme’s mental representation, so that when the phoneme is perceived, memories based on the visual information are recalled (Massaro, 1987).

To test this possibility, Fowler and Dekle (1991) introduced research participants to one of two experimental conditions. In one, the participants were presented with either a printed ba or printed ga syllable, while listening to a syllable from the auditory /ba/-/ga/ continuum. In the other, the printed syllables were replaced with their haptic presentations; that is, participants were able to feel how the syllables were being produced. Since there are no previously made associations to how syllables feel when a speaker produces them, by the memory account there should be no McGurk effect. The experimenters found no effect of the printed syllables on the auditory ones, as expected, and they found that the feel of how a syllable is produced affected the perception of the auditory syllables, indicating that articulatory gestures are indeed perceived by listeners.

The TRACE model of speech perception, TRACE 1, developed by Jay McClelland and Jeff Elman (1986; Elman & McClelland, 1988), depicts speech as a process in which speech units are arranged into levels and interact with each other. There are three levels: features, phonemes, and words. The levels are comprised of processing units, or nodes; for example, within the feature level, there are individual nodes that detect voicing.

Nodes that are consistent with each other share excitatory activation; for example, to perceive a /k/ in "cake", the /k/ phoneme and corresponding featural units share excitatory connections. Nodes that are inconsistent with each other share inhibitory links. Such nodes are nodes within a level. In this example, /k/ would have an inhibitory connection with the vowel sound in "cake", /eI/.

To perceive speech, the featural nodes are activated initially, followed in time by the phoneme and then word nodes. Thus, activation is bottom-up. Activation can also spread top-down, however, and TRACE can model top-down effects such as the fact that context can influence the perception of individual phonemes.

Perception of speech can be influenced by contextual information, indicating that perception is not strictly bottom-up but can receive feedback from semantic levels of knowledge. In 1970, Warren and Warren took simple sentences, such as "It was found that the wheel was on the axle", removed the /w/ sound from "wheel", and replaced it with a cough. They found that listeners were unable to detect that the phoneme was missing. They found the same effect with the following sentences as well:

\[
\text{It was found that the *eel was on the shoe.}
\]

\[
\text{It was found that the *eel was on the orange.}
\]

\[
\text{It was found that the *eel was on the table.}
\]

Listeners perceived heel, peel, and meal, respectively. Because the perception of the word with the missing phoneme depends on the last word of the sentence, their finding indicates that perception is highly interactive.
time to identify a word in context is about a fifth of a second, whereas it takes a third of a second in isolation. It is thought that the grammar and meaning of the preceding part of the sentence limit the range of possibilities for the gated word, such that it can be identified sooner in a sentence than on its own. The point at which there is only one possible candidate is called the \textit{uniqueness point}. The uniqueness point and the isolation point need not correspond: on the one hand, the word may be recognized before there is one remaining candidate, if the context is helpful (i.e., strongly biasing); on the other hand, there may be a delay in isolating the word. There is a third point, called the \textit{recognition point}. This is the point at which the person is confident in his/her identification of the gated word.

The guesses people make on this task indicate that the perceptual identity of the word is also important to spoken word recognition, even before the context has its effect. In other words, people's early guesses resemble the perceptual aspects of the word and not the contextually signaled candidate.

\textbf{References}


