Commentary

Dichotomies in the perception of speech

Perceptual systems in animals, including humans, seem to analyse different aspects of the sensory signals they receive in different ways. The most famous examples of this are the 'what' and 'where' pathways in vision: it appears that, in a simple way, the identity of a visually perceived object is perceived using a mechanism different from that used for estimating the spatial position of the object. A similar distinction has been proposed for the auditory system (e.g. Rauschecker and Tian 2000). But as far as audition is concerned, humans possess something that no other animal does, namely spoken language. A recent paper uncovers exciting dichotomies in our perception of speech which shed light on what/where issues in human auditory speech perception. Also, these data raise deeper issues regarding the nature of the human language system.

Smith *et al* (2002) use an elegant mathematical method to decompose the speech signal into a fast component (corresponding to the fine temporal details) and a slow component (the envelope). They do so with the help of the Hilbert transform, which provides a mathematically rigorous definition of the envelope and fine structure of the signal free of arbitrary parameters. For example, an exponentially damped sinusoidal oscillation can be viewed as the product of an undamped sinusoidal oscillation and an exponential. The exponential would be the slow component, the envelope, while the sinusoid would be the fast component, rapidly changing in time as compared to the slow component. To put it differently, the sound of a single tone fading can be factored into the tone itself and the shape of the decrease in amplitude.

The authors use two sound waveforms as input to an algorithm for generating what they call 'acoustic chimaeras'. These acoustic waveforms are either two sentences (for speech-speech chimaeras) or one sentence and a spectrally matched noise waveform (for speech-noise chimaeras). A filter bank extracts frequency components in different frequency bands. The filter bank they use is a set of filters that span different frequency ranges in such a manner that the width of each frequency band is approximately linearly spaced according to the receptivity of the human cochlea. The authors investigated the effect of having different numbers of frequency bands in the filter bank. The slow and fast components of the outputs of the different frequency bands are extracted, and the slow components from one waveform are multiplied with the fast components of the other waveform for each frequency band. Finally, the products are summed over all frequency bands to produce an auditory chimaera. Subjects who listened to the chimaeras had to report the words that they heard.

The most interesting result is with the speech-speech chimaeric sounds where, for one or two bands, speech reception is from the fast component (from the fine structure). On the other hand, for between 3 and 16 frequency bands, reception is dominated by the slow component. This reliance on the envelope for speech reception over the fine structure is also seen in speech-noise chimaeras. Here, with an increase in the number of frequency bands, the fast component (only) from speech causes poor to no reception, while the slow component (again, by itself) from speech causes near-perfect reception. These findings are in line with previous modulated-white-noise experiments described below.

The authors also manipulated the interaural time differences (ITDs) of the speech signal, the primary cues to localization in the horizontal plane. They created speech-speech chimaeras in which the slow components had ITDs pointing in one direction while the fast components had ITDs pointing the other way. Interestingly, now the pattern was reversed: while subjects still perceived the sentences that contributed the slow components, they were perceived to be at the *location* of the sentences that contributed the fast component.

If two melodies took the place of the two speech sounds, the result was reversed: the fast component dominated melody recognition for up to 32 frequency bands, and only for higher bands

was there evidence for the use of the envelope for melody recognition. The difference could be due to a difference in the nature of the stimuli or, more interestingly, due to different mechanisms for the analysis of speech-like versus other kinds of sounds.

These experiments suggest that the extraction of the content of the speech signal seems dissociable from the extraction of spatial location (using ITDs). Thus, different mechanisms, perhaps even anatomically distinct ones, might be involved in analysing different aspects of the speech signal.

Conventional theories of speech perception are "bottom-up" in the sense that the system is supposed to extract low level acoustic features from which it builds up phonemes (the basic sounds of a language), which in turn build up the syllables, and which finally build up the lexical items, the words. However, from the speech recognition community there has come strong support for models of speech perception that are driven largely by the lexicon, which can be thought of as a 'top-down' approach to hearing speech. Top-down in this context implies that speech input is approximated to the best possible sequence of words that the input might possibly represent. Notice that the strategy is not to decipher the precise details of the speech input (as a string of phonemes for example). Rather, one attempts to approximate the input to a given lexicon without strong dependence on individual low-level features. Interestingly, this approach seems to have provided the best route so far for implementing machine-based recognition of naturally spoken speech.

Several demonstrations support the top-down hypothesis. Shannon and coworkers (1995) used band-pass filtered white noise, which they then modulated with low-frequency components (at those frequencies) taken from actual spoken sentences. By doing so they were able to get rid of the rapid transitions in speech which are critical in defining the identity of individual phonemes. Under these conditions, listeners were still able to extract individual words. Saberi and Perrot (1999) demonstrated a more striking result. They chopped up waveforms of spoken sentences into small segments (about 50 millisecond-long segments) and digitally reversed each of them individually. Astonishingly, this seemingly drastic operation, in which the entire speech waveform was locally incoherent, still allowed for many or all of the words to be comprehended. Such experiments indicate that, at least in adults, precise detail in speech is not a pre-requisite for comprehension.

Put together, these results along with those from the work by Smith *et al* (2002) serve to illustrate that speech recognition in adults seems to be more dependent on broad characteristics of speech, and not so much on the precise identity of individual, smaller units. Although time differences between two phonemes in a language can be of the order of tens of milliseconds or less, such fine distinctions seem unnecessary for reception of a spoken sentence. This does not mean that these phonemic differences cannot be perceived. Rather, it means that the speech recognition mechanism inside our heads is able to handle large amounts of noise, presumably by taking a top-down approach to speech recognition. Clearly, the role of the context and pragmatics must play a large role in adding information to the language system in order to achieve best recognition. Finally, it is worth pondering about how is it that the initial state of the 'language organ' in the baby, which must by definition be in bottom-up mode, later operates in a top-down fashion.

References

Rauschecker J P and Tian B 2000 Mechanisms and streams for processing "what" and "where" in auditory cortex; *Proc. Natl. Acad. Sci. USA* **97** 11800–11806

Saberi K and Perrot D R 1999 Cognitive restoration of reversed speech; *Nature* (*London*) **398** 760

Shannon R V, Zeng F-G, Kamath V, Wygonski J and Ekelid M 1995 Speech recognition with temporal cues; *Science* **270** 303–304

Smith Z M, Delgutte B and Oxenham A J 2002 Chimaeric sounds reveal dichotomies in auditory perception; *Nature* (*London*) **416** 87–90

MOHINISH SHUKLA
Cognitive Neuroscience Sector,
International School for Advanced Studies (SISSA),
via Beirut 9,
34014 Trieste, Italy
(Email: Shukla@sissa.it)