SPEECH PERCEPTION SEEN THROUGH THE EAR

C J Darwin

Experimental Psychology, University of Sussex

Abstract

Speech is normally heard against a background of other sounds. This paper reviews recent work on listeners' ability to separate speech from other sounds. Evidence is presented that both low-level grouping mechanisms and knowledge specific to speech are deployed in solving this difficult problem.

Preamble

I have called this talk "Speech perception seen through the ear" to contrast it with a title that might have been given to many other talks on speech perception namely "Speech perception seen through the synthesiser parameter control table". In making this contrast I do not wish to detract from the decades of painstaking work that has led to the present sophisticated understanding of the relation between acoustic cues and phonetic categories, rather I would like to draw attention to two aspects of speech perception which complement previous work. First how the sound that leaves the mouth of the speaker differs from the sound that typically enters the ear of the listener, and second how complex sounds (especially speech) are coded by the auditory system. These topics not only are important for a complete understanding of speech perception, but also (and especially at a conference such as this on Speech Communication and Technology) provide a particularly fertile ground for interchange between speech technologists and speech and hearing scientists.

Formant Extraction In Noise

The well-studied relationship between acoustic cue and phonetic category assumes that the acoustic cue can be extracted readily from the speech waveform. Although it is often difficult to, say, track formants and measure aspiration duration in carefully spoken speech in a sound-proof room, the problem becomes very much worse if the speech has been spoken in the presence of other sounds or when echoes are present. If the other sounds present are noise, then the problem is essentially one of extracting a signal from a noise background, one well-studied by engineers. But it is interesting that some of the more recent techniques used by engineers to solve this problem, extracting formants from noise, are very similar to mechanisms proposed from neurophysiology and psychoacoustics. In particular, Niederjohn's work on formant-tracking in noise (Niederjohn and Lahat, 1985) has used a combination of band-pass filtering and zero-crossing detection in regions of high spectral energy that mirrors both physiological work on the value of phase-locked timing information in the auditory nerve for estimating formant frequencies in noise, and also the parallel psychoacoustic claim that timing information is used to code the frequency of low-frequency tones (see review in Moore, 1982).

But different questions of particular interest to the psychologist emerge when we consider the perception of speech against competing sounds that are structured rather than random. Now both "signal" and "noise" have clearly apparent structure, but it is less obvious which of the various frequency components or complex features belong together to form the required signal. Two different types of knowledge can be used to try to separate them. Low-level knowledge about the properties of single sound sources can be used for a preliminary auditory parsing, identifying groups of frequency components that have an improved chance of originating from a common sound source. By contrast, the process of speech recognition itself might be able to identify relevant features in a pool that also contains those from other sound sources. Over the last few years there has been a rapid increase in experimental evidence relevant to these issues, and the overall consensus appears to be that both types of process occur in normal speech perception.

Grouping By FO Differences

Evidence for both types of process comes from experiments on the perception of "double vowels". Subjects hear two simultaneous vowel sounds and are asked to identify them. The original experiments using this paradigm by Scheffers (1983) showed that subjects could identify about 45\% of vowel pairs when both vowels (from a vocabulary of 10) were presented simultaneously on the same fundamental frequency. A small difference in pitch between the two vowels significantly increased the number of vowel pairs correctly identified to around 62\%. His results have since been replicated and extended by Zwicker (1984), Assmann and Summerfield (1989, forthcoming) and Culling (pers. comm). The improvement in vowel identification with fundamental
frequency difference has been modelled by Assmann and Summerfield (forthcoming) using an auditory model that incorporates a hair-cell trading non-linearity (after Meddis, 1988) and phase-locking. Their auditory model, which segregates sounds in a way broadly similar to the auto-correlation method used by Weintraub (1985, 1987) quite successfully predicts the increased identification performance produced by a small pitch difference between the vowels, and also the fact that this increase is not found for short-duration (50ms) vowels.

The auto-correlation method is an example of a general auditory mechanism that can separate any harmonic sound from a background that does not have the same or a harmonically close fundamental. It exploits the fact that an auto-correlation function computed on the output of a filter that is tuned to a region of energy of a complex periodic sound will show a peak at the fundamental frequency regardless of whether the harmonics of the complex sound are being resolved or not by that filter. Auditory grouping is achieved by partitioning the energy in each auditory filter according to the magnitude of the autocorrelation function at a particular fundamental. Filters that are predominantly excited by energy at a different fundamental will have most of their energy assigned to a different source. The auto-correlation peak at the fundamental thus provides a basis for grouping energy at different frequency regions that originated from a common sound source.

Phonetic Integration

Although such low-level grouping mechanisms can improve performance in speech perception, a recent experiment by John Culling in this laboratory has illustrated that speech perception can still integrate information from different sound groups. Culling repeated the double-vowel experiment but with a slight variation: instead of synthesising the whole of one vowel on one fundamental and the whole of the other vowel on a different fundamental, he split each vowel into two frequency regions (mid-way between its first and second formants) and then assigned one fundamental to the first formant of one vowel and the higher formants of the other vowel, and vice versa for the other fundamental. Low-level grouping mechanisms should then provide the wrong answer - the first formant of one vowel should be grouped with the higher formants of the other vowel and performance should be worse than when both vowels are on the same fundamental. His results are shown in Figure 1.

Performance on the split vowel pairs is only slightly worse than on the normal vowel pairs. So split fundamental frequency between the vowels of a pair gives almost as much improvement with increasing pitch difference as do normal vowels. Why does performance improve at all in the split vowel condition? Presumably because a difference in fundamental frequency between the two sounds within a formant region helps to identify formant frequencies by reducing overlap of harmonics. The small though significant decrease in performance for split over normal vowels reflects the contribution of grouping across frequency regions. Why is this contribution so small when such across-formant grouping by F0 has been demonstrated clearly in other paradigms (Darwin, 1981; Gardner, Gaskill and Darwin, 1989)? The inappropriate low-level grouping between first and higher formants regions induced by the split F0 can evidently be substantially overcome by our speech-specific knowledge of the structure of (particular) speech sounds. In other words, we can still hear a phonetic category that corresponds to formants on different fundamentals. Low-level grouping is not a pre-requisite for phonetic categorisation.

A similar conclusion can be drawn from a recent study of Remez, Bressel, Klapwald and Rubin (pers comm). They played subjects synthetic sentences in which formant frequency contours were realise as frequency-modulated sine-waves: each sine wave tracked a formant frequency. It has been known for some time that subjects can identify sentences synthesized in this way, and the stimuli are interesting from the point of view of auditory grouping since there is no common fundamental frequency to help bind the individual formants together. Previous experiments had presented all the tone contours to the same ear. In Remez...
et al.'s experiment some of the subjects heard the second formant-tone in one ear and the remaining three formant-tones in the other ear. Although this manipulation reduced performance (from 69% to 45% correct syllables), it was still substantially above the score for either the second formant-tone presented by itself (2%) or the remaining three presented by themselves (18%). Subjects can clearly integrate into a phonetic percept information that appears to come from different directions, and which has no common fundamental frequency.

Pre-empting By Speech?

Both Culling's experiment and Remez et al.'s emphasise the ability of the speech processing system to gather together sounds that lower-level mechanisms do not group together. Perhaps we do not need such low-level mechanisms at all, and should embrace the principle espoused by Liberman and Mattingly (in press) that speech processing mechanisms "pre-empt" auditory mechanisms by choosing those features present in a sound that together give a phonetically plausible result. I have argued elsewhere (Darwin, 1989) that such a claim is unwarranted. There are clearly occasions where the choices of auditory features that the speech mechanism makes are constrained by auditory grouping principles. These include:

(i) the beneficial effect of local (in frequency) differences in fundamental frequency on double-vowel identification (see discussion above);

(ii) the effect of differences of onset- and offset-time in grouping out energy at a harmonic frequency of a vowel (Darwin, 1984);

(iii) the effect of a surrounding repeated sequence of tones in grouping out energy at a harmonic frequency (Darwin, Patterson and Gardner, 1989);

(iv) the effect of a repeated "chirp" in one ear on preventing integration of a chirp into a syllable on the opposite ear (Ciocca and Bregman, 1989).

None of these effects can be explained away by peripheral processes, yet they all constrain phonetic perception. So it is clear that auditory grouping mechanisms can influence the features that phonetic perception interprets; but equally it is clear that phonetic mechanisms can in some circumstances group together sounds that are otherwise treated as separate groups. Further work is needed to illuminate what those circumstances are.

Fundamental Frequency Modulation

We discussed above the use that the perceptual system makes of fundamental frequency differences in grouping together simultaneous vowels. But one of the most obvious characteristics of the voice is that the fundamental changes over time. Does perception use a common pattern of frequency modulation to group together the harmonics of a particular voice? Our laboratory has recently been looking at this question both from the point of view of speech perception and of psycho-phonetics.

John Culling has investigated whether adding different frequency-modulation functions to the fundamental frequency of two simultaneous vowels helps their perceptual separation. Following McAdams' (1984) work on prominence ratings with three simultaneous vowels, and also with work by Gardner, Gaskell and Darwin (1989) using a composite /ru/ - /l/ syllable. None of these studies had found evidence for grouping by differential frequency modulation.

Culling found that the different FM phases only improved vowel separation when there was no baseline fundamental frequency difference between the vowels. In this condition, the different phase of the FM function introduces a fundamental frequency difference between the two vowels, which is probably responsible for the effect. This result is consonant with McAdams' (1984) work on prominence ratings with three simultaneous vowels, and also with work by Gardner, Gaskell and Darwin (1989) using a composite /ru/ - /l/ syllable. None of these studies had found evidence for grouping by differential frequency modulation.

Culling's result is also consonant with some recent work from Robert Carlyon (1989) in this laboratory found that listeners were unable to detect changes in the phase of frequency modulation of one of three frequency components (using FM functions similar to Culling's) unless the three components
were harmonically related. (Harmonically related components have a clarity and pitch which is disturbed when they suffer incoherent frequency modulation). Since differences in the phase of a vibrato-like FM are undetectable, it is unlikely that they could be used to selectively group different sound sources. (By contrast, Carlyon and Stubbs (in press) have found that our ability to detect the presence of frequency modulation in a complex sound - against a noise background - was improved if the components were harmonically related.) In all the experiments described so far there has been no evidence that listeners were sensitive to differential frequency modulation of components except insofar as it led to a breakdown of harmonic relations.

A recent study from Magda Chalikia, working in Albert Bregman's laboratory (pers comm) has cast doubt on the generality of this conclusion. She first synthesised pairs of vowels on two different fundamentals (as described earlier) and made the two vowels of a pair either have the same fundamental, different but unchanging fundamentals, or fundamentals that changed in different ways. The fundamentals could be either have in parallel (maintaining a constant ratio) or in opposite directions, so that they crossed over. Her results from these conditions were quite different in the original conclusions from our own report. However, she then went on to produce a new type of stimulus in which the vowels were excited by an inharmonic pattern of frequency components. The frequency of each harmonic is shifted by a random amount (upto half the harmonic spacing) to produce inharmonic vowels. The various FM functions are then put on the underlying fundamentals, so that the (now harmonically unrelated) frequency components move either in the same or different directions. These inharmonic sounds gave a substantial effect - when the fundamentals contours crossed, subjects identified the vowel pairs better than when the contours were parallel.

Chalikia's experiment demonstrates that a common direction of movement can be used to group together frequency components with the absence of harmonic cues. But her results were much clearer using a large (6 semitone) frequency excursion than a small (1 semitone) that was similar to the vibrato functions used by McAdams, by Culling and by Carlyon. In the light of Carlyon's results it is unlikely that differential vibrato phase would give improved recognition of inharmonic vowel pairs.

In summary, frequency components can be grouped perceptually on the basis of both harmonic relations and common direction of movement. For large frequency excursions, a common direction of movement appears to contribute little to perceptual grouping, even when the frequencies form a harmonic series. However, even large differences in fundamental frequency across formants can be overridden to give an appropriate vowel percept, indicating that speech-specific knowledge can integrate information from different auditory groups.

Auditory Modelling

A simple linear bandpass filter model of the auditory system (see Moore, 1988) is immensely useful to the psychologist as a first approximation to the response of the auditory system. Various groups have found it a useful front-end for speech recognition probably because it emphasises low frequencies at the expense of high. The compressive non-linearity found in cochlear mechanics has also been found useful in modelling the perception of simultaneous vowels (Assmann and Summerfield, in preparation). More detailed modelling of the auditory system, however, may be less useful for speech technology, since the auditory system has had to evolve sensory coding strategies that ensure that a system composed of elements with a limited dynamic range can show overall a large dynamic range.

Modelling the physiological responses of the wide variety of cell types found in the dorsal cochlear nucleus to complex sounds is exercising a number of different research groups. Simple speech stimuli such as vowels provide effective stimuli for investigating their properties, although there is a clear lack of psychoacoustic phenomena that correlate with the observed properties of such cells. Research in vision can point to many parallels between perceptual phenomena and second- and higher-order cell types. Perhaps the great parallelism of the auditory pathways after the auditory nerve is responsible for this lack of evidence from hearing.

Acknowledgments

I am grateful to John Culling for permission to show his unpublished data and to the SERC for grants.

References


Assmann,P. and Summerfield,A.Q. (forthcoming) Modelling the perception of concurrent vowels: vowels with different fundamentals frequencies.


