Gotland workshop: Session ”Vowel perception

Ainsworth, W.

journal: STL-QPSR
volume: 21
number: 1
year: 1980
pages: 008-012

http://www.speech.kth.se/qpsr
Bill Ainsworth gave a short introduction to the stages of speech signal processing in the auditory system (Ainsworth, 1976). The mechanical tuning characteristics of the basilar membrane were compared with the tuning properties of the fibres of the cochlear nerve, and it was suggested that the hair cells might be responsible for this additional frequency selectivity. This point was taken up by René Carré who described a model involving an interaction between the inner and outer hair cells (Dolmazon and Bastet, 1979).

The higher levels of the auditory system were mentioned briefly, and it was pointed out that there were some units in the auditory cortex which responded only to onsets or offsets of signals and others which responded only to frequency changes. The responses of these units are much more labile and easier to fatigue than the units at the lower levels of the auditory system.

Intensity coding was discussed next. Although the dynamic range of hearing is greater than 100 dB, the dynamic range of fibres in the auditory nerve is only 20-50 dB (Sachs and Abbas, 1974). At the level of the cochlear nucleus, however, units have been found which signal differences in intensity over a range of more than 100 dB (Evans and Palmer, 1975). As these cells receive their input from the cochlear nerve, it is believed that intensity is coded in part by the time structure of the discharge pattern of primary cochlear nerve fibres.

The results of Sachs and Young on the way in which the auditory system responds to vowel sounds were then introduced (1979). If the normalized discharge rates of a large number of cochlear nerve fibres are plotted against their characteristic frequencies, a spectrum containing formant peaks is obtained at low intensity levels (40 dB), but the peaks disappear when the intensity is increased to 80 dB. If, instead, the average amplitude of the corresponding frequency component of the Fourier transform of the instantaneous rate of response of fibres with characteristic frequencies within half an octave of that frequency is plotted, the formant pattern remains throughout the dynamic range.
In the following discussion the model was criticised because it gave rise to spurious points on the spectrum. These were thought to be harmonics and intermodulation products of the first two formants. Adrian Pourcin suggested that the high frequency resolution would be improved if the pinna-oval window transformation was included in the model. Peter Ladefoged wondered how a temporal model would fit in with this frequency model.

Louis Pols then introduced some psychophysical approaches to vowel perception. In particular he mentioned several 'psychoacoustical facts' which are often neglected in speech perception models.

Most models employ a spectral data representation which is based on either a detailed spectral analysis (FFT), or on a spectral envelope (cepstrum, LPC), or on a bandfilter analysis (many overlapping or a few contiguous (critical band) filters). Subsequently some form of data reduction is applied resulting in formants, F1-F2', or a principal component representation. The ear itself, however, is neither a formant nor a principal component extractor. We know that the ear has properties like critical bandwidth, that there is spectral dominance for pitch detection, and that 7 or 8 harmonics can be heard separately. We also know that forward and backward masking in time exist. Danaher et al (1978) has shown that F1 can mask F2 for hard of hearing people. The role, if any, of nonlinearities, such as combination tones and lateral suppression, in everyday speech perception is not understood. In addition, we do not know what the pulsation threshold paradigm can teach us about the internal representation of vowel spectra, or what scaling procedures can tell us about distance measures. In general the questions can be asked about how to apply psychoacoustic knowledge to speech perception, and how to study the perceptual relevance of such measured 'facts' as formants, their transitions and loci, and VOT.

The pulsation threshold paradigm (Houtgast, 1974) was next explained, and some conflicting results of Eva Holmberg-Lundström et al (1977) and Richard Tyler on the one hand, and of Louis Pols (Houtgast, 1974, Pols, 1979) on the other, were discussed. Sven Öhman wondered what the underlying mechanism of this continuity phenomenon
could be, and how phase sensitive it was. Wolfgang Hess reported some masking experiments of Terhardt which seemed relevant. Björn Lindblom then played a tape, prepared by Plomp (1979), to demonstrate several continuity effects. It showed that noise during the silent gaps allows even dynamically varying sounds (such as a sweep tone or a singing voice) to be heard continuously. John Holmes recalled that Cherry and Wiley (1967) had performed similar experiments by presenting only the voiced portions of speech. This became intelligible if the gaps were filled with noise.

Louis Pols then posed a question about the perceptually relevant details in the spectrum. Were these the energy maxima, formants, or F1-F2'? It is only when we know more about this that we can develop realistic distance measures which specify the similarity between vowel sounds (Bladon and Lindblom, 1979, Dolmazon et al, 1977). Rolf Carlson described some experiments by Carlson, Granström, and Klatt (1979) in which a synthetic vowel /æ/ was slightly modified in various ways with respect to its spectral and waveform characteristics. Subjects were asked to estimate the psychophysical distance between each of these variants and a reference vowel. The results were correlated with distance measured derived from various models (Plomp, 1970, third-octave bands, Schroeder et al, 1979, sone/bark, Carlson et al, 1970, zero-crossings in frequency bands).

Although in general all models predicted the perceptual results equally well, there were several exceptions. The Schroeder model was bad for formants (because the masking is too wide), the zero-crossing model was insensitive to amplitude (although this can be modified), and all models were poor at frequencies below F1. For more detailed information, see Carlson and Granström (1979).

Louis Pols stressed the fact that analysis procedures as well as most models neglect low-frequency information, whereas perceptually this information seems to be very important. He also suggested that the experiment should be repeated with noise in order to evaluate the importance of the information in the spectral valleys. John Holmes (1979) pointed out that high quality speech synthesis requires a very thorough spectral match below F1. J.M. Pickett wondered why there was as yet no model which included suppression (Chistovich, 1979).
Peter Ladefoged and Gunnar Fant discussed whether these types of similarity judgements of artificial stimuli could be considered to be speech mode perception. Hiraya Fujisaki, however, mentioned some Japanese work which showed that it is hard to avoid using linguistic knowledge when listening to stimuli of this type. Rolf Carlson said that phonetic labelling had been avoided by choosing minor variations around the particular vowel /æ/. Sven Öhman expressed a great interest in these experiments but questioned their relevance for speech perception. Björn Lindblom suggested that the level in the auditory system where the distance measures operated was at the point where the 'top down' and 'bottom up' processes interact.

References


Sachs, M.B. and Young, E. (1979): "Effects of nonlinearities on speech encoding in the auditory nerve", J.Acoust.Soc.Am. 65, Suppl. 1, S102A.