Proceedings of The Institute of Acoustics

1D1

SPEECH FOR HEARING OR HEARING FOR SPEECH? IMPLICATIONS FOR THEORY AND PERCEPTUAL TESTING.

M.P. HAGGARD

MRC Institute of Hearing Research University of Nottingham
University Park Nottingham NG7 2RD

Perceiving speech is now the most general justification for humans being able to hear. This has led to the interpretation of certain correspondences between hearing and speech as favouring evolutionary speculation that the peripheral auditory and speech mechanisms may have coevolved or that one has determined properties of the other. The central neural systems have not escaped attention either; process models have been postulated in which the central neural circuitry is partially common so that speech may be produced with reference to its auditory effects or vice versa, and one theory constrains the articulatory repertoire in languages to movements producing particular values of acoustical dimensions that suit central auditory property detectors. This paper affords a general escape route from such conceptual incarceration. Man's articulatory advantage over the apes lies in his control of his acoustical output in the mid audio frequencies (500-2500 Hz), arising from enhancement of the tongues effective mobility, by its central location in a rectangularised vocal tract; this was probably accompanied by brain evolution giving (lateralised) control over the midline muscular structures. However the essential properties of the mammalian peripheral auditory system seem not to have changed, so that all that can be said is that auditory frequency resolution must already have been adequate in mid-frequencies to justify the change in vocal tract anatomy.

PROPORTIONAL FREQUENCY RESOLUTION

There is an apparent isomorphism between functions relating measures of the ability to generate differences in frequency in speech production to frequency and corresponding fluctuations in hearing or speech production on the other However these may be seen as the result of common acoustical constraints in the acoustical parts of the two systems. In production, bandwidth of formants is roughly proportional to frequency, therefore greater spectral amplitude differences are created by those articulatory differences which change lower formants. In addition, the size of articulatory organs that also chew and lick is too large to change high frequency formants reliably, with the exception of the resonance of the front cavity between tongue tip and lips. As frequency is roughly proportional to the reciprocal of cavity length, articulatory variability has disproportionately large effects for short front cavities, ie for high resonant cavities. The next limit is neural. It is impossible to exert precise enough articulatory control over the front cavity's length, ie over place of articulation, to permit differences of the order of 200 Hz to be reliably produced above a mean resonance frequency of about 2.5 kHz. Further back in the vocal tract such neural control would in theory be possible, but the necessary articulatory control in the relatively massive tongue body is lacking. Differences below 500 Hz however are chiefly the outcome of differences in the degree of opening of the vocal tract over which relatively fine control can be exerted via a variety of articulatory parameters,

by the use of tactile/kinesthetic feedback or by shapes of the tongue in which degree of opening on tongue contact is pre-programmed, as in fricatives, semivowels and liquids. Control over F_1 to the nearest 100 Hz is not difficult to achieve. Lengthening the vocal tract likewise scales all formants in a roughly proportional way, giving precise control over the lower ones. Finally it is hard to envisage how a biological larynx could produce appreciable energy above 3 kHz without raising its fundamental frequency and loosing the ability to act as a carrier for the information in low frequencies already discussed. Thus speech contains its own coherent constraints upon the ways in which frequency can carry information.

In sound reception, anatomical constraints upon genetically programmable variations in mass and elasticity have lead to gradations in the ratio of these quantities along the basilar membrane. These properties apparently operated millions of years ago to entail that the low pass and band pass properties of the basilar membrane travelling wave also lead it to behave more like a constant-Q than a constant-bandwidth filter. But there is probably a neural contribution to this proportionality as well, spotlighted by the renewed popularity of periodicity theory. Apparently the first, mechanical, filter and the second, physiological filter both serve to limit the range of periodicities a group of nerve fibres is required to convey. Because of neural refractoriness the 'cost' in numbers of parallel fibres for conveying a temporal representation of high frequencies increases exorbitantly, hence resolution decreases with frequency. It has recently been shown that in representing a complex sound such as a vowel, the nominally high frequency fibres receive a significant low frequency input; due to the extra demands upon the neural population for statistically transmitting a high periodicity, any lower periodicity of equal SPL will dominate. In vowels this happens at high levels where F_1 'captures' the high-frequency units, causing a form of masking; the root of the phenomena is in mechanical asymmetry and nonlinearity, but its final realisation is neural.

IMPLICATIONS FOR LANGUAGE

How do these constraints of frequency resolution in production and perception influence the phonetic structures employed in languages, in particular their functional load? We may take statistical counts of various single feature minimal pairings as indices of their functional loads. Doing this it is apparent that the place-of-articulation feature is not given as much weight in English consonants as the manner and voicing features are. This is congruent with the fact that the larynx spectrum and hence the long term spectrum of speech is weighted towards the low frequencies. The intelligibility centroid of the spectrum is about 1.8 kHz for phonemically balanced (PB) monosyllables and somewhat lower for continuous speech whereas the energy centroid is much lower - around 700 to 800 Hz. The spectrum from 1800 Hz to 6 kHz conveys as much intelligibility as the lower portion, on less energy and slightly less resolving power as measured by the critical band function. Why then does the place-of-articulation

feature have a lower functional load than other consonant features? The answer lies in the fact that the language does contain many phonemes of which the high frequency components are used in full communication under favourable conditions with normal hearing. The articulatory repertoire for (a limited number of) front vowels and consonant places of articulation leads to the importance of frequencies above 1.8 kHz and gives frontal place of articulation a role in the language. But other factors diminish its auditory importance in continuous speech. These factors are possibly extra pitch and timing information and probably the coarticulation and imprecise articulation that decreases the place information available. These factors can otherwise be offset by visibility of a speakers face. Put differently, the phonemes with high frequency content may be no less 'important' in conversation but may be supplemented non-auditorily. A corollary of the frontal places-of-articulation that produce the high frequency resonances is that the articulatory position is fairly readily seen in lipreading.

In the past it has been suggested that the importance function for frequency bands in speech intelligibility is determined by auditory frequency resolution. However in the extreme there must be slight differences in the function between languages;the interesting questionis whether languages tend to distribute the effort between articulatory control, acoustical consequences of articulation, auditory resolution and visibility of articulations in any systematic way. The recent attempts to explain the location of vowels in the entire vowel space for vowel systems of various sizes by auditory modelling are a great advance on purely phonological diachronic speculations. However the postulate of maximal auditory contrast within the vowel space as a whole does not work, at least if psychophysiological models of the internal spectrum are used; the large distances are reduced too much. Given that gross regions of the vowel space are hardly ever confused auditorily or visually, effort should now be concentrated on determining whether the 'sufficient' contrast between confusible neighbours within regions tends in the parameters of an auditory model to be constant across various vowel systems.

IMPLICATIONS FOR TESTING

It is appropriate when assessing telephony to include phonemes having high frequency content by use of PB monosyllables, although there can be no such thing as an absolutely correct and representative statistical sampling of materials at the phonemic level alone. For assessing the finer points of an ear, a transducer or a network the more demanding higher frequencies should be even more highly weighted. For differentiating among substantial degrees of auditory disability high frequencies can be given low weighting as the disability is mostly determined by the effectiveness of the interplay between intense low frequencies and the visual signal. In may applications some other requirement may take precedence over phonemic balance, when reconciling the sensitivity of a test with its range.