Proceedings of The Institute of Acoustics 4A22

SPECTRAL AND TEMPORAL-DOMAIN QUESTIONS FOR AN AUDITORY MODEL OF VOWEL PERCEPTION

R.A.W. BLADON
UNIVERSITY COLLEGE OF NORTH WALES, BANGOR

BJÖRN LINDBLOM
STOCKHOLM UNIVERSITY

Abstract

Drawing on the work of Zwicker (e.g. Zwicker and Feldtkeller 1967), recently formalized by Schroeder, Atal and Hall (1979), we have elaborated a specific version of a theory of peripheral auditory representation of steady-state vowels (Bladon and Lindblom 1979). This model, together with a distance metric which follows Plomp (1970), has been tested by hypothesizing that listeners in vowel-matching tasks of a natural or experimental nature make their judgements of vowel distance in accordance with the model. Very largely, it seems they do, and this is an encouraging result.

Possibly more interesting, though, are the two residual cases from our experiments where the model does not predict the auditory distance correctly. It is in search of an explanation for these irregularities in the data that we pose some questions here which would permit some fine-tuning of the model, either in respect of its amplitude-spectrum characteristics, or in respect of the effect of temporal processing on perceived vowel quality. These questions are raised without as yet fully knowing the answers, but with the hope that a preliminary airing of them may narrow down the choice of priorities for the next stage of our research.

Outline of an auditory model of vowel perception

Before turning to the questions referred to, and since a published report of the model is not yet widely available (but see Bladon and Lindblom 1979; a full version is to be published elsewhere soon), a brief account of the model itself is in order.

Figure 1 is a block diagram of the model. The input to the computation is a harmonic power spectrum, designated (1), of a steady-state vowel of $F_0 = 120$ Hz (in this case), whose frequencies are then converted by formula to a Bark scale of perceived pitch - see panel (2). This procedure follows an established view of auditory perception in psychoacoustics, which holds that the frequency-to-place transformation along the basilar membrane of the inner ear is in terms of critical bands whose bandwidth is one Bark (approximately one-quarter octave above 250 Hz). The next stage is the application of a masking device in the form of a frequency-smearing function (3), due to Schroeder, an "auditory filter" intended to simulate the spreading of energy distributions along the basilar membrane, such that the resulting smeared spectrum (4) may be imagined as corresponding to the mean-square amplitude of

SPECTRAL AND TEMPORAL-DOMAIN QUESTIONS FOR AN AUDITORY MODEL OF VOWEL PERCEPTION

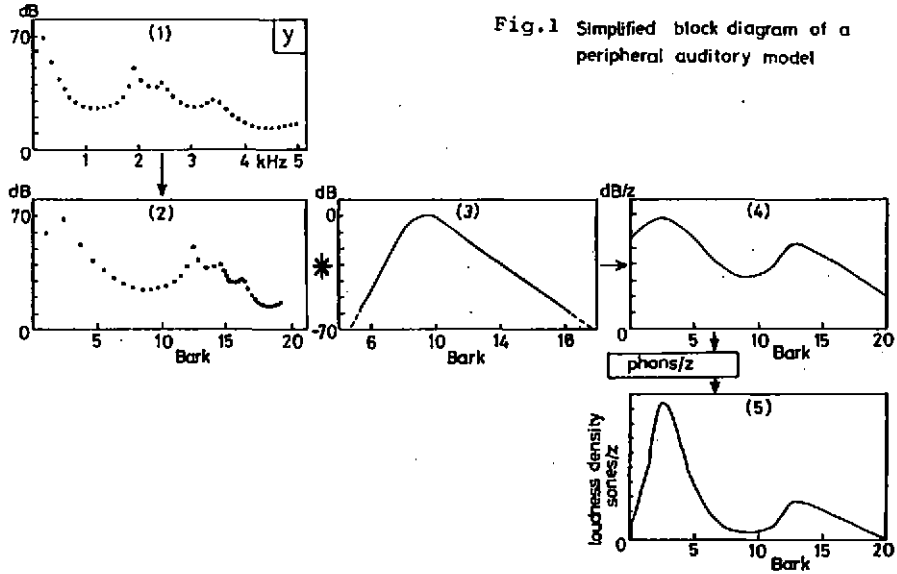

Fig.1 Simplified block diagram of a peripheral auditory model

the basilar membrane motion. For mid-range sound-pressure levels, of perhaps 40 to 70 dB above the threshold of hearing, it is entirely appropriate to postulate the filtering function in this invariant form, with positive and negative skirts of + 25 dB/Bark and -10 dB/Bark respectively.

The application of the auditory filter (3) to derive the smeared spectrum (4) is accomplished by a process of convolution, symbolized in Figure 1 by a star. It is then further postulated, in accordance with known characteristics of human auditory behaviour, that vowel timbre perception is in terms of patterns of loudness density per critical band. A loudness density pattern, such as panel (5) in Figure 1, can be derived from the dB values by two successive processes: first, via curves of equal loudness level, and second, via a conversion to sones per Bark by formula.

Experiments have been performed to calibrate the model auditorily, and have been reported more fully elsewhere (Bladon and Lindblom 1979). To summarize, the procedure was to obtain a large body of listener judgements of perceived vowel distance by presenting synthetic 4-formant and 2-formant vowels in pairs and eliciting judgements of how similar in quality each pair of stimuli was. In several cases, the judgements called for were of a rather fine degree, since certain vowel pairs consisted of a 4-formant vowel and

its best 2-formant match (as reported by Carlson, Fant and Granström 1975). The results shown in Figure 2 give the median of the listener-judged auditory distances plotted as a function of the model-calculated distances D_{AX}. It is reasonable to say that, if we leave aside for the moment the data points for [α] and [o] , the correlation between auditory and model-calculated distance is high.

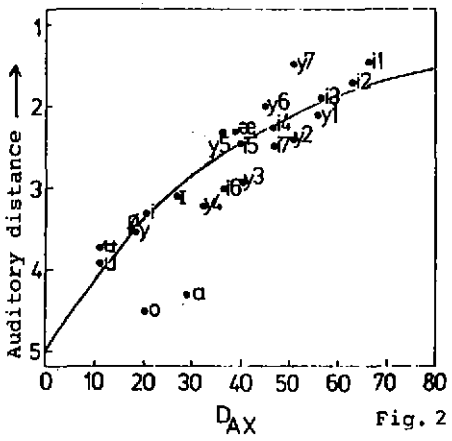

Fig. 2

But the validity of the model, as revealed in Figure 2 by its ability to predict auditory judgements of vowel distance, is restricted to the extent that its prediction of the [α,o] results was unsuccessful. In what directions might we look for an explanation of this behaviour? At this stage we do no more than touch on some possibilities in framing the questions which follow.

Q.1 Does loudness disparity contribute much to vowel distance judgements?

By definition, a disparity in total loudness is reflected as a disparity in area under the loudness density curves. An examination of the loudness density patterns for 4- and 2-formant [α]in Figure 3 suggests that in its high calculated D_{AX} value, our model over-estimates the importance of loudness differences, especially in the F_1 region. To test the extent of this effect, all the output loudness density curves were normalized for total loudness. After application of the distance metric, the model yielded an improvement in the [a] result itself, but at the cost of a poorer correlation among many of the other vowels. Noticeably also, the result for [o] did not improve with loudness normalization; indeed upon inspection of the middle panel of Figure 3 this is not surprising, since although the two stimuli show an imbalance in F1 and F2 levels (and hence an undesirably high D_{AX}), they do not show an appreciable total loudness difference. Provisionally, then, desirable as the inclusion of loudness normalization is for vowel quality judgement tasks, it seems to add little to our present results.

Q.2 Is a pattern of loudness density per Bark currently the best representation of the auditory spectrum?

An alternative approximation to the auditory spectrum might correspond to stage (4) of our model, namely a basilar membrane excitation pattern specified not in loudness density terms but as dB per Bark. This hypothesis is worth considering since, broadly

SPECTRAL AND TEMPORAL-DOMAIN QUESTIONS FOR AN AUDITORY MODEL OF VOWEL PERCEPTION

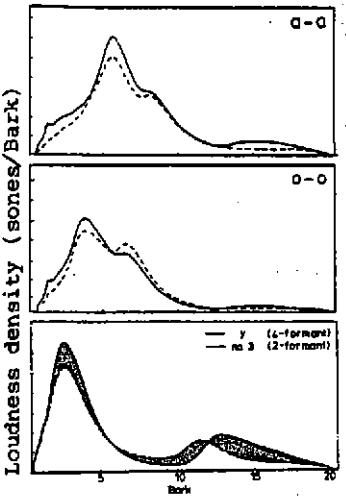

Fig. 3. Loudness density patterns for 3 vowel-pairs

speaking, it has been implemented in hardware in a number of third-octave spectrum analyzers, and such an analysis has in fact been applied to the same vowels of our work by Fant (1978:77). It is not clear to us what the justification would be for such a model in speech perception theory. To terminate the modelling at the dB/Bark stage implies a rejection of the well-founded work on loudness summation by Plomp and others, and a correspondingly less close approximation to our current knowledge.

It has been suggested however that the loss in prediction accuracy induced by basing a distance metric on dB/Bark patterns would in cases such as ours be only slight (Plomp 1976:95). In order to test this hypothesis our data were again re-processed, but this time using as a basis for the distance calculation the output of stage (4) of our model. The results clearly show a marked worsening of prediction accuracy; and hence do not appear to challenge the view that loudness density patterns offer a more suitable basis for vowel distance judgement.

Q.3 Should auditory-nerve rate-saturation be directly modelled?

It is noticeable that a property shared by the spectra of the two problematic vowel pairs [ɑ-ɑ] and [o-o] is the relatively small frequency separation of F1 and F2. It is possible to suggest that such vowels might be especially susceptible to rate saturation of auditory nerve fibres, given the relatively small dynamic range of average discharge rates observed. The high-level F1-F2 peaks in our [ɑ, o] spectra might therefore be processed as a single, broad response pattern: or, at the least, our model appears to overestimate the importance of level differences in the F1 and F2 regions. An explanation involving saturation effects would be an attractive option, since if both the F1 and F2 loudness differences were neutralized by saturation in [ɑ] and [o] , those vowels would achieve a much improved correlation with auditory distance.

However, the rate-saturation explanation should not be adopted uncritically. For instance, it has been shown that the presence of bandstop noise in the stimulus can lower the threshold for saturation. It is conceivable that the presence in vowel sounds of harmonic components may induce a response more akin to the noisy condition than to the one without noise. Then energy at the formant frequencies might not, after all, result in saturation in the corresponding fibres. A different argument might be derived from the psychoacoustic technique of pulsation-threshold measurement (Houtgast, 1974), which makes strong claims to reveal the characteristics of the auditory

spectrum, and in so doing further emphasizes the auditory relevance of formant peaks.

Q.4 What allowance should be made for the coding of vowel stimuli in the temporal domain?

Frequency information relevant to the distinction of vowel qualities is coded in the auditory nerve not merely as spectrally-related average discharge rates, but also temporally, in the phase-locking of the auditory nerve. To summarize recent findings in this field, the general effect of information from the temporal response pattern of auditory nerve fibres seems to be one of restitution of the formant-dominated pattern, At high stimulus levels, the phase-locking of responses to harmonics outside the formants is with very few exceptions suppressed by locking to formant harmonics. Thus, even at high levels, temporal synchronization to the formant frequencies ensures that they remain perceptually dominant.

Thus it seems at least possible that temporal processing of vowel quality may add to or even sometimes override the loudness-density analysis which we have modelled. Interestingly, this possibility then opens up a route to explain the residually problematic result obtained for [ɑ, o] vowel-pairs: formant-frequency information would be available not only indirectly as a part of the excitation pattern, but at the same time coded temporally, thus promoting the importance of that information in the perceptual system (and in its distance metric), to the detriment of place-coded excitation level discrepancies. The high similarity judgements of the [ɑ] and [o] pairs become explained in formant-frequency terms.

References

R.A.W. BLADON and B. LINDBLOM 1979 in Speech Communication Papers (eds. J.J. Wolf and D.H. Klatt), 1-4. Auditory modeling of vowels.

R. CARLSON, G. FANT and B. GRANSTRÖM 1975 in Auditory Analysis and Perception of Speech (eds. G. Fant and M.A.A. Tatham), 55-82. Two-formant models, pitch and vowel perception.

G. FANT 1978 Rivista Italiana di Acustica 2, 69-87. Vowel perception and specification.

T. HOUTGAST 1974 Acustica 31, 320-324. Auditory analysis of vowel-like sounds.

R. PLOMP 1970 in Frequency Analysis and Periodicity Detection in Hearing (eds. R. Plomp and G. Smoorenburg), 397-414. Timbre as a multidimensional attribute of complex tones.

R. PLOMP 1976 Aspects of Tone Sensation.

M.R. SCHROEDER, B.S. ATAL and J.L. HALL 1979 in Frontiers of Speech Communication Research (eds. B. Lindblom and S. Öhman). Objective measure of certain speech signal degradations based on masking...

E. ZWICKER and R. FELDTKELLER 1967 Das Ohr als Nachrichtenempfänger.

The support of SRC grant no. GR/A/03894 is acknowledged.