# THE AUDITORY MODELLING DILEMMA, AND A PHONETIC RESPONSE

ANTHONY BLADON

Phonetics Laboratory, University of Oxford
41 Wellington Square
Oxford, OX1 2JF, U.K.

## A. A DILEMMA IN CURRENT AUDITORY MODELLING

In recent years, results in psychoacoustics and auditory physiology have become routinely available to speech researchers. Numerous computational models of peripheral auditory processing have been published, some being only partial models, but some, including those by the following authors, being rather more complete (see the Blomberg et al. review [5], papers by Cohen, Divenyi, Lyon, Seneff in [14], Dolmazon and Boulogne [9], Cooke [6].)

However, at the time of writing there are many uncertainties about what should go into an auditory model for speech processing. Different models will result, depending on how the investigator views such matters as the following:

(a) which of the many reported psychophysical effects the model incorporates (a partial list could include a tonality scale, frequency masking and resolution, temporal masking and resolution, saturation, equal loudness curves, total loudness, lateral suppression, combination tones, retention of phase information);

(b) which of the physiological findings it seeks to replicate (such as phase-locking, adaptation to a steady-state signal, recovery, probabilistic neural firing, onset/offset asymmetry, efferent intervention, interactions at various stages of the auditory process);

(c) whether it is safe to extrapolate to speech signals, from data of the above types obtained mostly with simpler stimuli; and if not, what modifications to make;

(d) parameters of these models which are intended to be variable (e.g. time windows, bandwidths);

(e) parameters which are empirically variable because we do not yet know what values they should have;

(f) design considerations (more functional versus less so, more data reduction versus less).

In all of these general ways, including the extent to which they have taken specific account of speech, published auditory models reveal considerable differences.

As if this indeterminacy were not itself enough of a nuisance, there is also a multiplicity of answers to the question of methods of evaluation of such models. One method is to use an auditory model to preprocess the signal at the front end of an automatic speech recogniser, and to consider the model to be improved when the recognition rate improves. This brings with it the enormous variable of the recogniser characteristics themselves, for which there is no foreseeable standard. An alternative possibility is to calibrate our auditory models against human perceptual data, such as confusion matrices, perceptual distance judgements, recognition against noise, etc. The main problem here is that there is an acute shortage of such data; but problems of language bias and task differences also add to the difficulty of interpretation. Finally, since almost all auditory models presuppose a calculation of distance between a stored reference pattern and an incoming candidate signal, there is the open question of a distance metric.

Putting together all these uncertainties, the researcher is confronted with a dilemma. It is that, at the current stage of knowledge, we are faced with more variants of auditory models than we can ever possibly test experimentally; and yet, if we do not test the models, there is no way to identify a better model and know when progress has been made. The essence of a modelling exercise is to advance by successive testing and refinement.

Inevitably therefore we need to identify some factors to help limit the search among candidate auditory models. Some expedients which may assist in this task include:

* cost, computability
* best guessing
* functional overlap
* limiting the objective

- * improved data on the auditory processes
- * improved knowledge in speech perception.

Further elaboration of these possibilities, and of individuals' answers to them, could form a worthwhile discussion issue at this Congress. Our own personal decisions are implicit in the Appendix, in which we briefly sketch the current implementation of an auditory model at Oxford.

## B. ONE RESPONSE: SPEECH PERCEPTION

Meanwhile in this paper we concentrate on two strands of research, illustrated mainly from our own work, which can contribute to constraining the search among auditory models. The first, and more familiar, exercise involves trying to refine existing knowledge about speech perception. There is of course nothing new in that, as a research programme. However, the strategy we wish to advocate adds to that position, by suggesting that advances in speech perception research can, when cautiously intepreted, help us to infer (or, more realistically, to state speech-based preferences about) properties of auditory analysis which might underlie the findings. We reason back from these findings, as it were, so as to shape our expectations about an auditory model for speech.

### 1. Diphthongs

Consider diphthong sounds, for example, as a test case for dynamic auditory modelling of speech. We know that confusions between steady-state vowels and diphthongs are rare; the spectral change in diphthongs is somehow auditorily salient. And it is quite well established that there are auditory mechanisms (e.g. neurons in the cochlear nucleus and inferior colliculus) which respond specifically to a change in a stimulus. However, it is possible to imagine more than one way in which the auditory system might assign importance to this particular kind of changing signal. We can formulate the issue as a speech perception experiment: are diphthongs perceived in terms of their endpoints, or, irrespective of the targets achieved, in terms of a constant rate-of-change? Several authors have addressed this issue experimentally, but in our opinion (see [1]), inconclusively.

In our presentations of diphthong stimuli, which had been artificially cut back in a variety of ways, and when offered a good range of possible transcriptions. of the diphthong quality, our trained listeners consistently responded in terms of the endpoints actually achieved (and not the rate-of-frequency-change). Moreover, when listening to diphthongs whose transitional interval was excised completely, 100% identification was maintained, and the fact that there was an instantaneous spectral jump in these edited diphthongs was hardly noticed at all. At the same time, listening to stimuli consisting of the transition alone (in running speech, but without the early and late steadier-states of the diphthong segment) led to many confusions.

From these studies our first conclusion had to be that "the one thing the ear is not doing, during the transitional part of a diphthong, is estimating the spectral shape change over time" [1, p.152]. Instead, the data were interpreted as suggesting the following role for spectral change in the auditory processing of diphthongs (and perhaps other speech sounds as well). Recall that, whether the spectral change in a diphthong lasts 100 ms or (artificially) 0 ms, it suffices to tell the listener that the sound is diphthongal in quality. The auditory role of spectral change in a diphthong may therefore be, first, as a weighting flag, alerting the system to assign extra distinctiveness to the current stimulus (because it contains spectral change); and second, as a temporal pointer, designating temporal regions of the signal (here, the adjacent endpoints) which the system should inspect more closely for their spectral content.

What do these interpretations mean for a physiologically-based auditory model? Some evidently quite appealing parallels can be drawn - for example, with peaks in neural discharge rate, with adaptation and with recovery in the auditory nerve. Such data show that, when spectral change intervenes, adapted fibres may recover leading to an enhancement of contrast in the adjacent segments (cf. [8]). On the other hand, other prospective model components would fare less well: lateral suppression, for example. We might be justified in inferring that modelling this behaviour would not be productive in the case of a diphthong transition. This is because lateral suppression would predict a frequency-sharpening effect, whereas our findings seem to confirm the other view (cf. [10, 17]) that when listening to rapidly changing signals, the frequency analysis of the ear is much coarser than otherwise.

### 2. Laterals

Lateral consonants have been another focus of our recent interest. It turns out that, in a limited way which is however reinforced from other speech data, laterals shed light on the question of auditory integration (versus resolution) of frequency. We (Bladon and Burleigh) recently manipulated lateral consonants, both in isolation and in a CV context, in respect of several variables including the width (from 0 to 5 Bark) of an antiformant-like notch in the spectrum. A 1-Bark notch, see Figure 1, which is typical of what regularly occurs in production samples, was essentially undifferentiable from no notch at all. Consistent with other experiments on fricatives, this finding suggests that the auditory filter for speech "smoothes over" a typical lateral consonant's spectral notch. Our experiments were not very sensitive, but a JND for notch width in the region of 2-4 Bark was indicated. The straightforward notion of psychophysical critical band (corresponding to a resolution of 1 Bark) is not, it seems, an appropriate model. A wider-scale integration is operative.

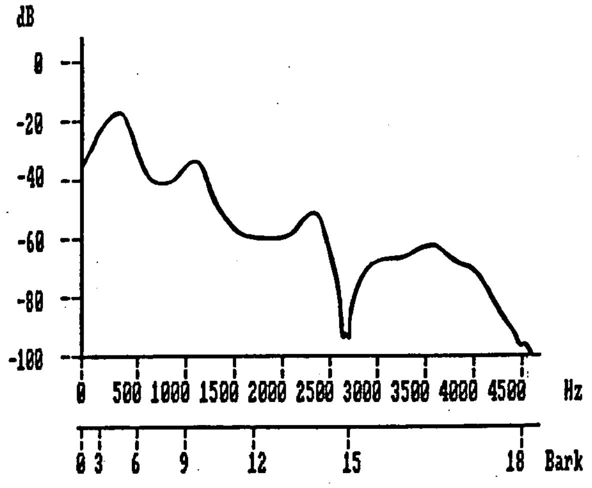

Figure 1. Spectrum of a synthetic lateral consonant used in Experiment 1. The bandwidth of the anti-formant notch was varied from 1 Bark (shown here) to 5 Bark.

Could the lateral's antiformant be detected instead, at least when adjacent to a vowel, by temporal auditory mechanisms, such as enhanced onset/offset of discharge in certain auditory channels? Our results revealed not: notches remained barely detectable, and one can only surmise that the notch is not salient enough (in a word such as "law") to survive temporal masking. Our experiments went on to suggest that the auditory signature of lateralness relies instead on grosser characteristics such as transition duration and overall amplitude envelope.

The concept of a wider-scale (>>1 Bark) auditory integration, for speech sounds, will in due course merit some further attention. We shall return to it at a later stage of the next section, in which we address a second methodological response to the modelling dilemma.

## C. A RESPONSE FROM LINGUISTIC PHONETICS

A second way of limiting our testing of auditory models is by virtue of the objective we set. One well established objective, which underlies much of the philosophy of our model given in the Appendix, is to use it for pre-processing the signal supplied to an automatic speech recogniser. But that is not the objective we wish to pursue here. Suppose instead as an interesting objective, that auditory modelling should equip us better to understand the auditory constraints upon language systems and language use. After all, speech is designed not only to be spoken but also to be heard. It turns out that this fact can be inferred to lie at the basis of a whole gamut of properties of sound-systems, their long-term structural trends, distinctive features, and aspects of sound change.

These inferences, and the explanatory value they have for linguistic phonetics, have been fleshed out elsewhere, [2]. Generalising from them to the theme of this paper, it can be said that sound-system properties show evidence of long-term influence from two main kinds of auditory behaviour: one, the asymmetry in auditory representation of energy onsets (which are disproportionately more salient) versus offsets; and two, the wide-scale spectral integration mentioned earlier.

The inclusion of onset/offset asymmetry in an auditory model for speech processing seems well justified by numerous linguistic examples. Summarising [2], there are various instances of unaccounted directionality in phonological behaviour which could be due to the stronger representation of auditory onsets. For instance, phonological nasalisation of vowels spreads very commonly onto a preceding vowel (as it did in the history of French), but only rarely onto a following one. Lateral consonants can vocalise (as in Cockney "field") but commonly do so after, and rarely before, a vowel. The rarity of aspiration after (but not before) a vowel, as in word-final /h/ or in preaspiration, is another often-noted directional asymmetry. In all these cases, a general tendency to spectral energy offset is what characterises the rare occurrence; whereas the common member contains more of an onset. All the cases (as well as others, such as patterns of syllable consonant formation) could well have this auditory foundation.

Now to pick up the earlier reference to the bandwidth of auditory integration. The limited evidence of the lateral consonant notch can be supplemented very considerably, so as to show that much of speech behaviour, especially the long-term organisational properties of sound systems, is consistent with an auditory resolution as wide as some 3.5 Bark. The psychophysical evidence for this idea, it must be said, is still not large; the linguistic evidence, however, is mounting.

If we suppose, then, that two vowel formants are integrated into a single auditory percept when they are less than 3.5 Bark apart, a number of interesting observations follow. Syrdal and Gopal [20] showed how the vowels of American English partition into categories, defined by formant integration versus resolution, which align impressively with the distinctive-feature classification of these vowels into grave/acute, diffuse/compact. The same kind of partition applies if we reanalyse, in Bark-scale integration terms, the Lehiste data [13] for /r,l/ of American English, see Figure 2; and likewise, though not shown here, if we reanalyse in the same way the fricative spectra of Polish reported in [12]. In a nutshell, the identification of certain sounds in language (probably those with a strong spectral pattern), seems to be favoured, on a long-term basis, if they maintain boundaries some 3.5 Bark apart.

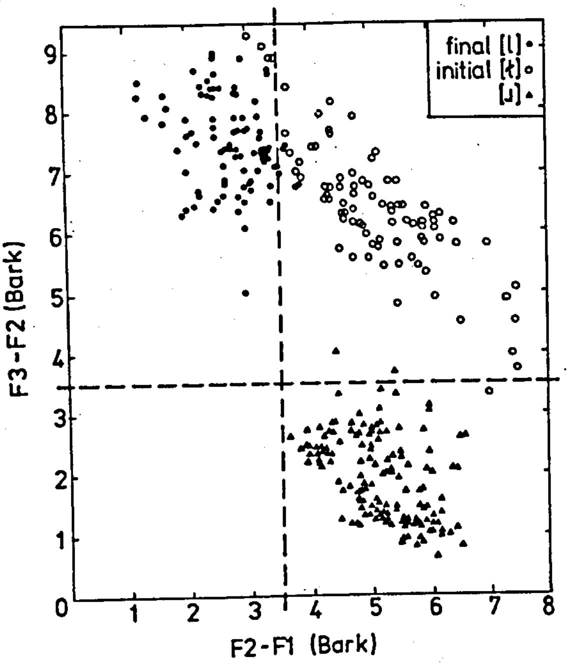

Figure 2. Liquid consonants of American English, their F3-F2 distance plotted against their F2-F1 distance (both in Bark). Each data point is one male (of 6) in one context (of 15).

Space limitations here do not permit the other examples of this kind to be elaborated in detail. In brief, though, the assumption is of a 3.5 Bark band of spectral integration, within which formants will be clearly integrated, outside which they will be clearly resolved, but if falling near the boundary formants will be auditorily less distinct, hence perhaps disfavoured in language and unstable. In these terms, it becomes possible to understand that there could be an auditory motivation for several properties of vowel systems. One such is the under-population, whether in actual languages or in computational simulations of vowel systems, of the "close" region of vowel space. Another is the dimension of "brightness", often noted to be a consistent reality for naive listeners; and a third is the auditory dimension of "rhotacised". We can also understand the strong disfavouring, in languages, of "interior" vowels. Finally, if we imagine vowel space to be a juxtaposition, in (perhaps) three dimensions, of zones of auditory integration/resolution, then we can understand further general properties such as that the number of height distinctions in back vowels is rarely more than the number in front vowels. All of these observations follow from the same basic assumption mentioned at the start of the paragraph; all can be appreciated, with a little patience, from the (rather conjectural) diagramming of cardinal vowels, Figure 3.

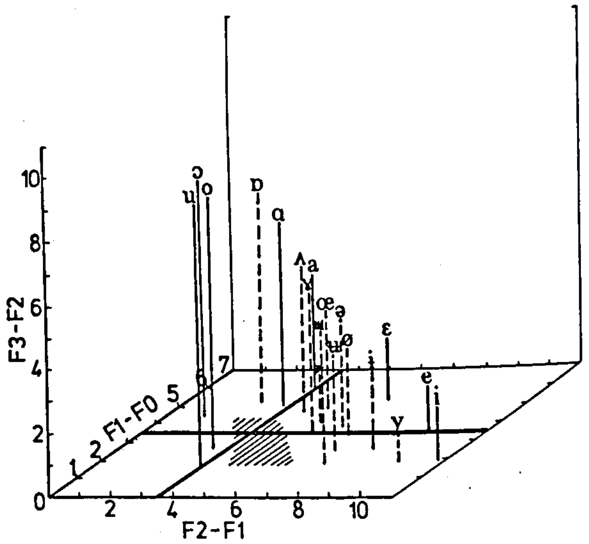

Figure 3. Cardinal vowels visualized in a three-dimensional Bark space defined by whether there is integration or resolution of their various spectral peaks, within a 3.5 Bark band. Primary cardinal vowels as solid lines, secondary ones as dashed lines.

In sum, this demonstration brings home quite forcibly how an auditory model of the identification of vowels in actual languages may well need to incorporate specifics which are not typical of most psycho-physical models on offer today.

In addition, some components of existing models may need to be emphasised at the expense of others. To further refine those models, as phoneticians, we may have to resort to the circularity of picking out persuasive trends in the very data we want an auditory model to explain, as a way of focussing the forbiddingly large search space.

## APPENDIX: CURRENT OXFORD AUDITORY MODEL

Our current version of auditory modelling routines is built on the foundations of the vowel model used by [4], extending it to include important aspects of the auditory processing of dynamically changing events.

The model is a modular piece of analysis and display software, written in C to run under Unix on a Masscomp 5500 computer with array processor and high-resolution colour graphics system.

The modules which are currently available are outlined below, with skeleton comments: for a fuller description, see [3]. For the honing and encoding of the algorithms in question we are indebted to C. deSilva.

### 1. Middle ear transfer function

Two alternatives are embodied, following [7], with some simplification. One alternative relates the pressure at the eardrum to the displacement of the stapes, and has essentially the form of a low-pass filter; the other, a stapes velocity function, resembles a pass-band filter with broad skirts.

(a) Displacement function (normalized to unity at 0 Hz):

$$( 25.0 / ( 1.0 + t^{2} (25.0 - 6.0 t^{4} ) ) )^{0.5}$$

(b) Velocity function (normalized to unity at its maximum):

$$\sqrt{\frac{20.8925937\, t^{2}}{1.0 + t^{2} (25.0 - 6.0\, t^{4})}}$$

where t = frequency in Hertz/1500.

### 2. Frequency conversion to the Bark scale

The conversion from the physical Hertz scale of frequency to the Bark scale of tonality or perceived pitch is accomplished by the following formula from Traunmüller (unpublished):

$$\text{Bark} = \left[ \frac{26.81\, h}{1960.0 + h} \right] - 0.53$$

where h is the frequency in Hertz.

### 3. Freq. conversion to the ERB-rate scale

This is intended as an alternative to the Bark scale, or more strictly, a compromise among several different suggested scales of tonality. The conversion from Hertz to ERB-rate is accomplished by the formula from [15]:

$$\text{ERB-rate} = 11.17 \log \left[ \frac{k + 0.312}{k + 14.575} \right] + 43.0$$

where k is the frequency in kiloHertz.

### 4. Frequency smearing

Spectral masking effects are modelled by convolving the spectral values in linear units with a function derived from [18]:

$$10.0 \log \left[ 15.81 + 7.5\, x - 17.5 \sqrt{1 + x^{2}} \right]$$

where x = 0.474 k (in Bark), and k is a scaling factor which enables a selection of different -3dB bandwidths for this function, according to the relationship:

k=1.429046/required -3dB bandwidth (Bark).

### 5. Decibel spectra

Spectral values, s, scaled in linear units of pressure are converted to decibels by the familiar relationship:

$$\text{dB} = 20.0 \log(s)$$

### 6. Equal-loudness curves (phons)

The equal-loudness curves of [16], Table 8, Appendix 4, are used to convert decibel values to phons. These curves are used principally because of the wide range covered, 0 to 15000 Hertz. The phon values are determined by substituting the spectral values in decibels into quadratic functions whose coefficients are functions of the frequency.

### 7. Total loudness (sones)

Loudness levels in phons are converted to total loudness values in sones by the use of the loudness indices given in [19] Table I, interpolating when necessary. Below 18 phons, loudness is approximated by:

$$\text{sones} = (\text{phons}/40)^{2.884}$$

## 8. Enhancement of spectral change

Two alternative models (8 and 9 below) are being explored, each of which incorporates some enhancement of the input signal at instants of rapid spectral change - [11], Divenyi in [14]. In the simpler implementation, enhancement of spectral change is carried out by adding, at each frequency point, a multiple of the rate of spectral change at that point. The formula used is:

$$o(h,t) = i(h,t) + a\left[i(h,t+1) - i(h,t-1)\right]$$

where:

$i(h,t)$ is the input spectral value at freq h, time t.

$o(h,t)$ is the output spectral value at freq h, time t.

$a$ is a user-controllable sharpening (= change enhancement) factor.

## 9. Neural adaptation/recovery effects

As an alternative to the preceding, it is possible to combine some change-related enhancement with other properties of auditory nerve behaviour, specifically neural adaptation/recovery effects. They are modelled by combining a filter which models exponential decay to an equilibrium level with one whose output is related to the derivative of the input. The model has four parameters: (i) equilibrium output level, that is, the steady-state output of the filter when the input is identically zero; (ii) adaptation time constant, which represents the time taken for the output to decay to 1/e of the difference between its initial value and the equilibrium value; (iii) recovery time constant; (iv) input response factor, which determines the amount to which changes in the input are reflected in the output. The formula has:

$$o(h,t) = c.o(h,t-1)+(1-c).y0+r\left[i(h,t) - i(h,t-1)\right]$$

where:

$i(h,t)$ is the input spectral value at frequency h, time t.

$o(h,t)$ is the output spectral value at frequency h, time t.

$y0$ is the equilibrium output level.

$r$ is the input response factor.

$c$ is related to the time constants as follows:

$$c = \exp\left[-\left[\frac{\text{interval between spectra}}{\text{ad/rec time constant}}\right]\right]$$

## REFERENCES

[1] A. Bladon, Sp.Comm. 4, 145-154, 1985.

[2] A. Bladon, In G. McGregor, "Language for Hearers", Pergamon, 1-24, 1986.

[3] R.A.W. Bladon, C.J. Clark, C. deSilva, P.F.D. Seitz, Prog.Rep.Oxf.Un.Phonet. Lab. 2, 28-42, 1987.

[4] R.A.W. Bladon, B. Lindblom, J.Acoust. Soc.Am. 69, 1414-1422, 1981.

[5] M. Blomberg, R.Carlson, K. Elenius, B. Granström, In R. Carlson and B. Granström, "The Representation of Speech in the Peripheral Auditory System", Elsevier, 197-201, 1982.

[6] M.P. Cooke, Sp.Comm. 5, 261-281, 1986.

[7] P. Dallos, M.C. Billone, J.D. Durrant, C.-Y. Wang, S. Raynor, Science, 177, 356-359, 1972.

[8] B. Delgutte, N.Y.S. Kiang, J.Acoust. Soc.Am. 75, 897-907, 1984.

[9] J.M. Dolmazon, M. Boulogne, Sp.Comm. 1, 55-73, 1982.

[10] P. Escudier, J.L. Schwartz, Sp.Comm. 4, 189-198, 1985.

[11] S. Furui, M. Akagi, Proc.12th.Int. Cong.Acoust., A2-6, 1986.

[12] W. Jassem, In B. Lindblom, S. Ohman, "Frontiers of Speech Communication Research", Academic, 77-91, 1979.

[13] I. Lehiste, "Acoustical Characteristics of Selected English Consonants", Indiana Univ, 1964.

[14] P. Mermelstein (ed.), "Proc. Montreal Symposium on Speech Recognition", Canadian Acoust. Assn, 1986.

[15] B.C.J. Moore, B.R. Glasberg, J.Acoust. Soc.Am., 74, 750-753, 1983..

[16] D.W. Robinson, R.S. Dadson, Brit.J. App.Phys., 7, 166-181, 1956.

[17] M.R. Schroeder, IEEE Comms.Magazine, 23, 54-61, 1985.

[18] M.R. Schroeder, B.S. Atal, J.L. Hall, In B. Lindblom, S. Ohman, "Frontiers of Speech Communication Research", Academic, 217-229, 1979.

[19] S.S. Stevens, J.Acoust.Soc.Am., 33, 1577-1585, 1961.

[20] A.K. Syrdal, H.S. Gopal, J.Acoust.Soc. Am., 79, 1086-1100, 1986.