Proceedings of the Edinburgh Linguistics Department Conference '96 pp. 7–17

Using statistics to model the vowel space

Matthew Aylett
matthewa@ling.ed.ac.uk

Abstract

In much experimental phonetics the concept of a vowel space is discussed. The density of points within this vowel space can be modelled using a mixture gaussian density function. There are a number of advantages in modelling the vowel space in this way. The model can generalise, producing a smooth representation of sometimes noisy data. The model can be used to normalise vowel spaces produced by different speakers. It can be used to categorise different vowels. It can be used to categorise and compare different types of speech, such as citation speech against running speech. To further improve the model a parametric curve fitting technique is applied. This paper describes the technique in detail, reports some tentative results and finally discusses the use of the model as a normalisation technique and as a method for measuring care of articulation.

1 Introduction

Formants are areas of high energy in the frequency spectrum of voiced speech. In vowels the frequency of formants, generally the first and second formant (F1 & F2), can be used to categorise vowels. The higher the vowel, the lower F1. The more front the vowel, the higher F2. So for example /i/, which is a high front vowel, has a high F2 and a low F1; while /ɑ/, which is a low back vowel, has high F1 and a low F2, almost meeting at around 1000Hz. It is possible to plot the F1 and F2 values of vowels. See Figure 1 for an example adapted from Ladefoged.

This is of particular interest because it has been argued that F1/F2 differences play a major role in vowel perception:

> For vowel sounds generally, and this is true of the English system, a significant part of the information listeners use in distinguishing the sounds is carried by the disposition of F1 and F2. (Fry 1979:78)

A scatter plot of F1/F2 values can be used to show how actual vowel values correspond with each other. If the density of the scatter is plotted as a third dimension, a 3D plot of the vowel space is produced. From this (Figure 2) the hills show locations of high density. The values in the hills would tend to correspond to an example of a particular vowel.

The 3D plot is interesting in itself and shows the 'vowel triangle'. However it is not easy to manipulate a plot of this type mathematically. The plot also shows minor variation

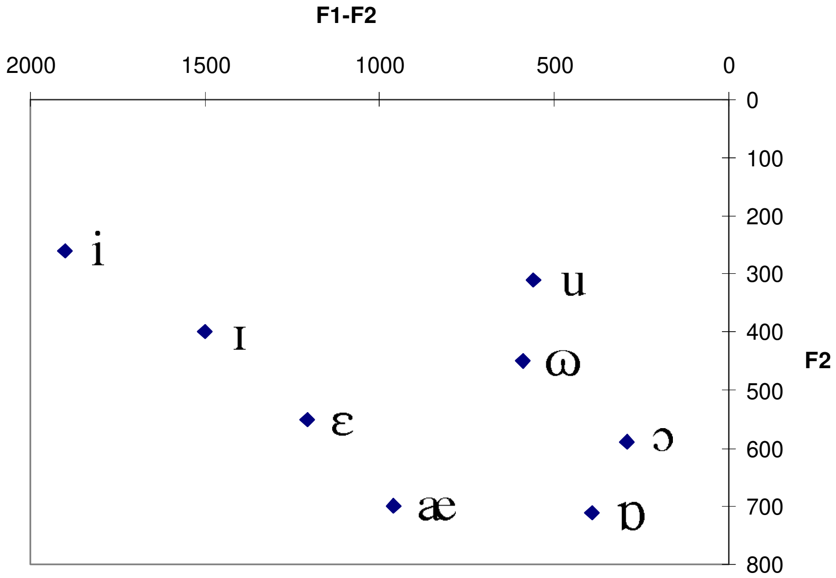

Figure 1: *The 'vowel space' adapted from Ladefoged 1975, chapter 8.*

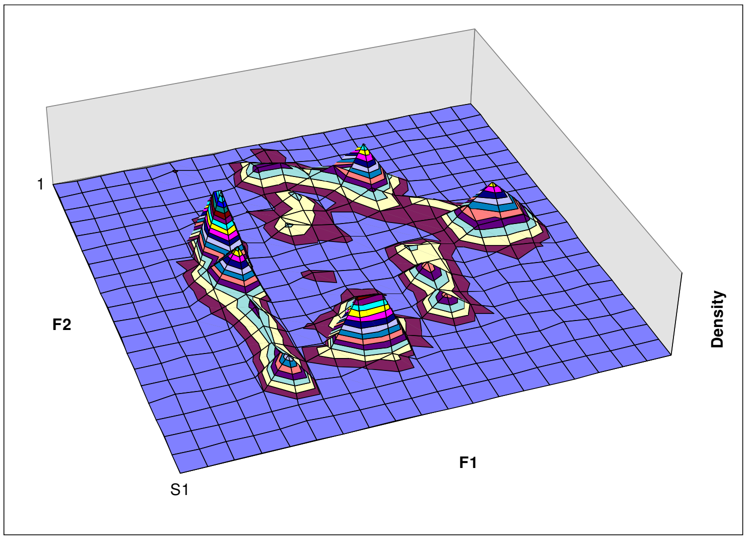

Figure 2: *A 3D scatter plot of the vowel space.*

which may be caused by noise. It does not generalise well because some of the small peaks are caused by minor variation that may not be significant to vowel production. In contrast, if you were to replace the scatter plot with a number of specified hills this could potentially characterise the shape of the plot very well. It would also allow you to regard the hills as categorical groupings, produce a probability that any point in the F1/F2 space was part of the distribution and allow comparison and mapping between hills.

The EM (expectation maximisation algorithm) applied to a mixture of gaussians is able to model the data in precisely this manner.

2 The EM algorithm

A two dimensional gaussian curve resembles a hill. The height of the hill is the probability of the gaussian occurring, the north/south width of the hill is the variance of the gaussian in one dimension and the east/west width is the variance in the second dimension. The location of the peak of the hill is the mean of the gaussian.

A number of these types of gaussians can be added together to model a complex distribution. There is also an algorithm that will, given a number of gaussians, fit them to a distribution. This algorithm is the EM algorithm. It is an iterative algorithm that will reach a maximum fit, although the maximum fit it finds may only be a maximum locally.

The problem of local maxima is general to all hill climbing algorithms such as the EM algorithm. By hill climbing we mean that the algorithm always improves the model with each iteration, and the algorithm stops when no change in parameters improves the model. This is analogous to getting to a top of a mountain by taking a step in a direction which always takes you up hill. Unfortunately it is always possible that you may end up at the top of a small nearby hill rather than the mountain you intend to climb. This small hill is the same as a local maximum. The number of local maxima depends on many complex interactions in what is a multi-dimensional search space. The more local maxima, the more sensitive the algorithm becomes to starting criteria and the more likely it is to find not the best solution, but a secondary solution. The EM algorithm will find a fit for a set of n gaussians, but in order to feel secure that this fit is a good fit it may be necessary to run the algorithm a number of times from different random starting positions.

The algorithm works as follows:

1. Pick a number of gaussians.

2. Randomly place them on the distribution with random standard deviations, random probabilities of occurring and random means.

3. While the fit continues to improve, take the points that 'belong' to each gaussian and use them to recompute the means, sds and probability of occurrence for that gaussian.

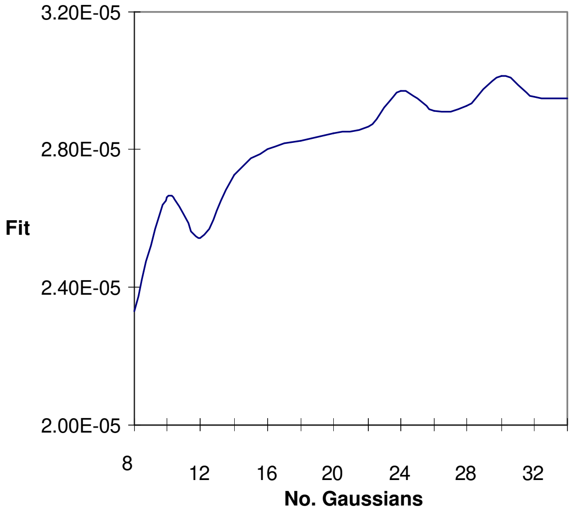

Figure 3: *The relationship between number of gaussians and how well the model fits the data.*

One of the major advantages of the algorithm is that it is unsupervised. That is, it is only necessary to specify the number of gaussians used in the model. It is not necessary to specify what the data points in the distribution represent.

There are, however, some serious disadvantages. It is necessary to choose the number of gaussians in advance. On what basis do we choose this number? How can we ensure the algorithm does not get stuck in a local minimum?

There is no theoretically bomb proof means of answering these questions. This is one reason why fitting with gaussians (and other cluster analysis techniques) are often avoided by statisticians. However a pragmatic approach to the problem can produce interesting results. The pragmatic answer to the questions posed above is to use trial and error.

It is possible to look at final fit over different numbers of gaussians (see Figure 3). Again the improvement appears to level off and become more unstable (probably due to more local minima with models containing more gaussians). This levelling off together with an inspection of the actual density distribution we wish to model can be used to estimate a good number of gaussians. Models with a similar number of gaussians behave in similar fashions so it is not necessary to be absolutely correct. The number I chose for my model was 20, partly because that seemed a sufficient number to model the data by inspection (Figure 2), and because (as can been seen in Figure 3) the improvement appears to both level off and become more unstable after about 20 gaussians.

In order to avoid local minima it is necessary to run the model a number of times. The hope is that local minima will generally be less stable than global minima and thus it would be very unlucky, using random starting parameters, to fall in the same local minima. Over 10 trials the results from the model appeared generally stable.

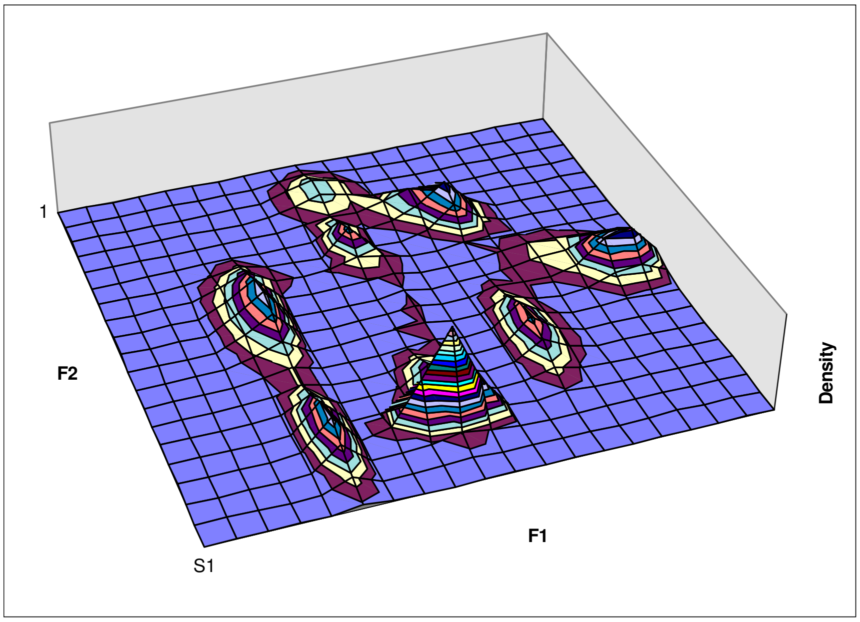

Figure 4: *The data from Figure 2 modelled with 20 gaussians.*

In brief, using a purely pragmatic approach I chose a 20 gaussian model and ran each model 10 times choosing the best model in order to avoid local minima. The result of applying the 20 gaussian mixture model to the data in Figure 2 is shown in Figure 4. As can be seen the mixture function has successfully modelled the main peaks in the original distribution.

Before discussing how this model can be used I will first talk about the automatic measurement of the vowel F1/F2 values from a large corpus of speech.

3 Automatic measurement of F1/F2 values

Traditionally the F1/F2 values of a vowel are measured by hand using a spectrogram. The phonetician listens to the speech and decides on the location of the beginning and the end point of the vowel. Looking at the spectrogram of the vowel the phonetician chooses the most steady state location — where the dark formant bands in the spectrogram are flattest (for diphthongs this would be two locations, one for each part of the diphthong). Either the phonetician would take the centre of these bands to represent the frequency of F1 or F2, or, if the computing power is available, the phonetician might compute a fourier transform of this steady state area and pick out the frequency of the formant peaks from this fourier plot. For the example in Ladefoged 1975, chapter 8, (Figure 1), a number of measurements for the same vowels are made for the same speaker and the average F1/F2 values are used to plot the location of the vowel in the vowel space. This is an intensely time consuming process and it is not possible to apply this technique to large amounts of speech data.

In order to apply statistical modelling techniques to data such as the EM algorithm it is necessary to have a large number of data points, certainly in the thousands. Therefore in order to apply the modelling techniques it is necessary to automatically measure the F1/F2 values.

The simplest method for doing this is to use LPC (linear predictive coding) to calculate both the probability that voicing is taking place and the likely position of the formants. The Waves software supplied by Entropics has functions that do exactly this.

Therefore the first stage in the automatic measurement of the F1/F2 values is to calculate the probability of voicing and the formant values over each 10 ms section of speech and output F1/F2 values when the probability of voicing is high enough.

I can now explain what the vertical peak to the left of the vowel triangle in Figure 2 represents. Voiced speech also includes nasals and voiced approximants and fricatives. It is not unreasonable to include /wylr/ in the vowel triangle, voiced fricatives /vðzʒ/ will not have a strong effect and can be discounted as noise, while nasals /mnŋ/ are very prominent. The vertical peak beside the vowel triangle in Figure 2 is caused by /mnŋ/. It may be possible to use other techniques to filter out voiced speech not belonging to vowel segments. However, if the main objective is the comparison between vowel spaces then this inclusion of all voiced speech will not necessarily be a major disadvantage.

The use of LPC analysis also introduces another major problem that as yet has not been resolved. The LPC model is a poles only model. When a vowel is nasalised a zero is introduced into the filter function (see Holmes 1988 for a review). This zero causes formant tracking software to perform poorly when encountering nasalised vowels. One major problem in this automatic technique is that it deals badly with nasalised vowels.

Accepting the limitations above still leaves us with a major problem. When the F1/F2 measurements are made by hand it is necessary to find the central steady state part of the vowel (or both steady state locations in the case of diphthongs). In contrast, so far all the voiced values are being used. In many cases coarticulation affects the formants. At onset and offset of the vowel the F1/F2 values will often be quite different from the F1/F2 values in the central part of the vowel.

In citation speech the steady state part of a vowel can often be found by setting a criterion on how long something must stay the same to count as steady state, and how close measurements must be to count as being the same. However, in running speech these same criteria may produce virtually no data points at all as so many vowels are reduced and the time for a formant to remain constant may be negligible.

In order to deal with this problem a technique using parametric curve fitting to calculate the spectral target of the central part of the vowel was applied.

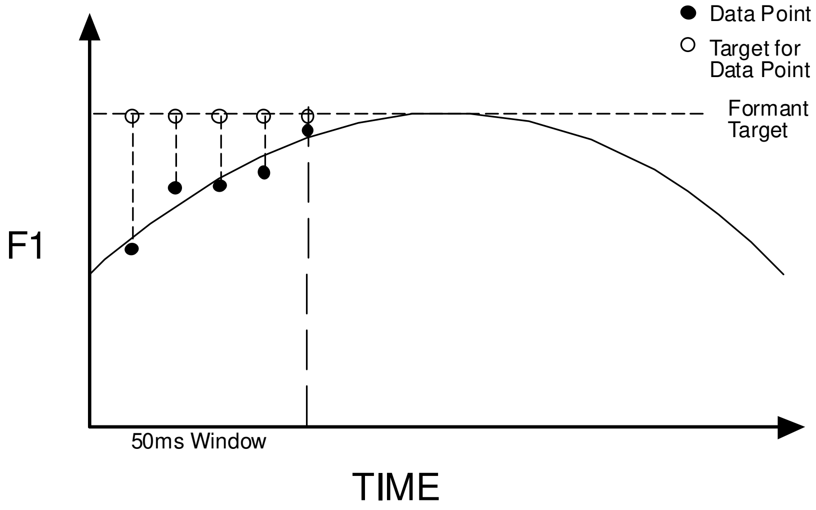

Figure 5: *Using a parametric curve to calculate the spectral target of a formant.*

4 Using parametric curve fitting to model coarticulation

A different variety of mathematical functions can be used to model formant transitions.[1] The most well known approach is that used by Lindblom (1963). Lindblom, in proposing a target model for vowel production modelled the formant transitions using exponential functions based on the mathematics of a damped spring. Van Son (1993) gives an excellent review of the target/undershoot model and its variations since 1963. He also discusses the use of Legendre polynomials to model formant tracks (1993, chapter 4)

There are two advantages of using simple parametric curves of the form $y = ax^2 + bx + c$. The first is that it is possible to calculate the maximum or minimum of the curve and use this value as the vowel target or approximation to the steady state central formant value in the vowel. The second advantage is that it is possible to calculate which parametric curve of the above form will best fit a set of points using mean squared error.

The method can be used to estimate the mid vowel formant targets by fitting the best parametric curve to a number of formant values over a time window. The maximum or the minimum of the curve can be regarded as the final spectral target that this formant is heading towards or away from (see Figure 5).

A major problem in applying this technique to unsegmented speech is to decide on how many points to use (or the window size) and whether such a window should overlap. If windows do overlap or a number of window sizes are used, it is necessary to choose between different values predicted by different curves for the same point in time. See Figure 6 for

[1]My implementation of this technique is based on a talk given by Steve Isard to the Phonetics and Phonology group at Edinburgh university.

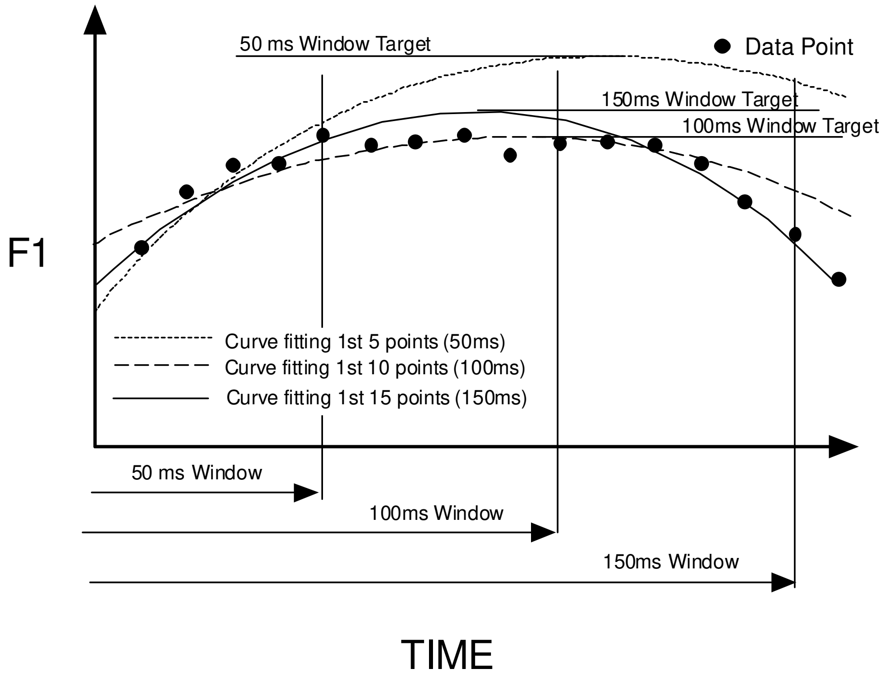

Figure 6: *The effects of differing time windows on fitting parametric curves.*

an example of this effect using different sized windows.

The method used to choose one target predicted over another is to use the target associated with the best fitting curve. One problem that this can cause is erratic values when the formant transition is very sudden. It was therefore necessary to filter out the erratic values by setting a limit on possible F1/F2 targets. In practice a uniform limit of 4kHz removed most erratic values.

The final effect of fitting the parametric curves is to produce a set of more discrete values that take into account the onset and offset of the vowel as well as the steady state period. This has the effect of making the data more granular and of increasing the definitions of the hills in the distributions.

Finally, to accept output the target had to remain within 1 bark for a minimum of 4 frames (40ms). This helped reduce noise and again make the vowel space more distinct. It also had the effect of reducing the number of lax vowels in the data. This was an advantage for my purposes as the lax vowels, especially the schwa, have higher variance and can make the vowel space less distinct. This is the reason why there are no hills in the centre of the vowel space as most of this central data tends to be schwa.

5 Normalisation

The scale on Figures 2 and 4 is unmarked. This is because the output from the parametric curve fitting and thus the parameters of the mixture model have been normalised. The normalisation is as follows:

All data is first converted into barks. The aim of the model is to produce as separately distinct hills as possible. The bark scale represents the ability of the human ear to distinguish different tones at different frequencies (Zwicker 1961; Zwicker & Terhardt 1980). For example, the human ear is more sensitive to tonal differences between 1000kHz and 2000kHz than between 4000kHz and 5000kHz. The use of the bark scale has the effect of stretching the vowel space where the human ear is most sensitive and contracting the space where tonal differences are difficult for the ear to perceive. The bark scale produces a more distinct mapping than either raw frequency or the mel scale (an alternative psycho-acoustic scale).

The F1 and F2 are normalised to have a mean of 0 and a standard deviation of 100. This has the effect of stretching and squashing these dimensions so that nearly all the data falls within a square of size minus 200 to plus 200. This makes it easier to compare different plots between different speakers.

6 Using the model

6.1 Comparing citation speech to running speech

Figure 7 shows data taken from running speech. It is interesting to compare Figure 2 with Figure 7. A number of typical effects of running speech are clearly visible. There is a large amount of centralisation of vowels. The variance between vowel types has increased merging the distinct hills and finally many vowels have been reduced to schwa filling up the central area of the vowel space.

By using a statistical model of the citation speech of the same speaker it is possible to make a measurement of care of articulation. The premise is as follows: Citation speech is carefully articulated; if a segment of running speech could just as easily be citation speech then this segment has been carefully articulated. If a segment is unlikely to have been produced using citation speech then this segment is not carefully articulated. Of course there is a certain amount of noise in the system as well as occasions when segments in citation speech are not carefully articulated. However the above process can be carried out automatically over a large amount of speech. In any study looking at care of articulation a large enough sample would hopefully overcome the effects caused by noise.

6.2 Speaker normalisation

A major problem in comparing data from different speakers is the variation between the vowel space generated from one speaker to the next. Using the modelling technique it is

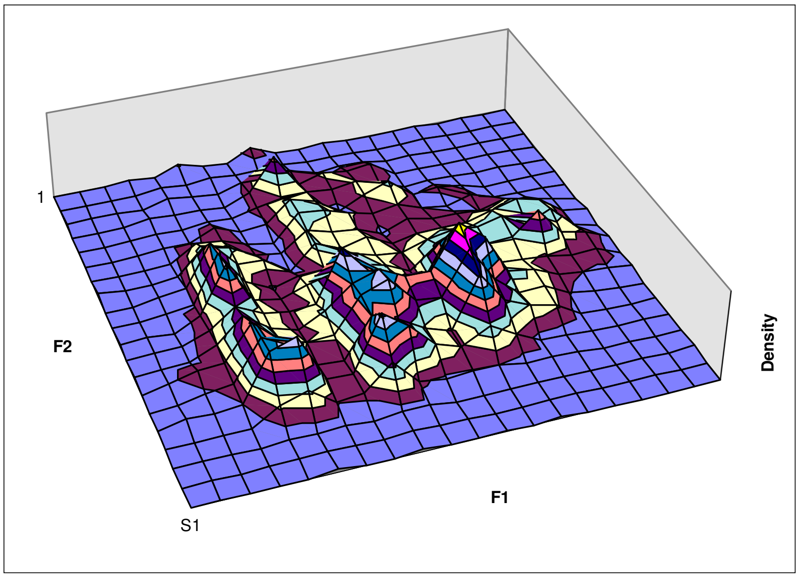

Figure 7: *A 3D scatter plot of running speech.*

possible to generate two 20 gaussian models of two speakers' citation speech. It is then possible to map the closest matching gaussians onto each other. It would then be possible to use this mapping as a non-linear normalisation function. Potentially this could be used to normalise the speakers' running speech. This offers a potential method for adding the vowel spaces of different speakers together as well as a mathematical measurement of vowel space differences embodied in the mapping.

6.3 Categorisation

It is possible to regard each separate gaussian as a different vowel sound. The accuracy of using the model to categorise vowels in this way is not particularly dependable, although it could be used to generate formant values of the most typical vowels in the speakers' vowel space. This could be useful when investigating different accents or language vowel spaces.

7 Conclusion

By using automatic modelling techniques as described above it is possible to characterise the vowel spaces of different speakers. It allows the development of non-linear normalisation techniques which could potentially allow accurate comparison between speakers. It allows characterisation of different types of speech such as citation speech and normal running speech. But perhaps the most exciting feature of this work is that these methods expose structure and regularity in a highly variant environment without even the pre-conception of the phoneme. The modelling techniques find their own segmentation and their own categorisation without any more information than formant frequencies and the probability of voicing. Unlike a phonetician the model has no concept of meaning, either phonetic or semantic. The structure we see captured by these techniques simply emerges from the data. It is a true bottom up approach. As such it offers only a part of the story, but what it does offer is a solid data driven basis for examining vowel production.

References

Fry, D.B. 1979. *The Physics of Speech.* Cambridge: Cambridge University Press.

Holmes, J.N. 1988. *Speech Synthesis and Recognition.* Wokingham: Van Nostrand Reinhold.

Ladefoged, P. 1975. *A Course in Phonetics.* New York: Harcourt Brace College.

Lindblom, B. 1963. Spectrographic study of vowel reduction. *Journal of the Acoustic Society of America* 35: 1773–1781.

van Son, R. 1993. *Spectro-Temporal Features of Vowel Segments.* Amsterdam: Uitgave IFOTT.

Zwicker, E. 1961. Subdivision of the audible frequency range into critical bands. *Journal of the Acoustic Society of America* 33: 248–249.

—— & Terhardt, E. 1980. Analytical expressions for critical bandwidths as a function of frequency. *Journal of the Acoustic Society of America* 68: 1523–1525.