BUILDING A STATISTICAL MODEL OF THE VOWEL SPACE FOR PHONETICIANS

Matthew Aylett

Human Communication Research Centre,
University of Edinburgh
email: matthewa@cogsci.ed.ac.uk

ABSTRACT

Vowel space data (A two dimensional F1/F2 plot) is of interest to phoneticians for the purpose of comparing different accents, languages, speaker styles and individual speakers. Current automatic methods used by speech technologists do not generally produce traditional vowel space models (See [6] for an overview); instead they tend to produce hyper dimensional code books covering the entire speakers speech stream. This makes it difficult to relate results generated by these methods to observations in laboratory phonetics. In order to address these problems a model was developed based on a mixture Gaussian density function fitted using expectation maximisation on F1/F2 data producing a probability distribution in F1/F2 space. Speech was pre-processed using voicing to automatically excerpt vowel data without any need for segmentation and a parametric fit algorithm [7] was applied to calculate likely vowel targets. The result was a clear visualisation of a speaker's vowel space requiring no segmented or labelled speech.

1. INTRODUCTION

The work reported in this paper was the result of the need to measure care of vowel articulation in a large corpus of connected speech in order to explore the relationships between redundancy, prosody and clarity. A speaker's vowel space was characterised using a statistical model. Vowels produced in spontaneous speech were then related to this model and results from this comparison were used to produce an objective measurement of care of vowel articulation.

Mapping a speakers vowel space by hand is resource intensive and in order to deal with the large amount of data needed to address the problem an automatic method was required. It was important that the method didn't require segmented speech and that the results from the modelling could be easily related to current phonetic literature on vowel production in order to answer such questions as whether a speaker's front vowels are more front in one phonetic context than in another or whether /i/ targets are more tightly grouped in a particular speaking style. In this way the model offers a means to compliment small vowel studies carried out in a carefully controlled laboratory setting with studies of vowel articulation in large connected speech corpora.

The model shares most of the advantages of a traditional code book characterisation of a speaker. It generalises, allowing noise reduction and it can be used to characterise vowels spaces produced by different speakers as well as categorising different vowels. It is also possible to use the model to compare different speech styles such as citation speech against running speech.

2. THE MODEL IN DETAIL

A formant is a concentration of acoustic energy that reflects the way air vibrates in the vocal tract. As the vocal tract produces sound, air vibrates at many frequencies at the same time. Peaks in the spectra reflect basic frequencies of the vibrations of air in the vocal tract. Areas within the spectrum with relatively high energy frequency components (i.e areas around these peaks) are termed formants. [For a more detailed definition see 8].

In vowels the frequency of formants, generally the first and second formant (F1, F2), can be used to categorise vowels. The higher the tongue in the mouth when producing the vowel the lower F1. The further forward the tongue in the mouth when producing the vowel the higher F2. So for example /i/ (in heed) which is a high front vowel (i.e. the tongue is high and to the front when producing this vowel) has a high F2 and a low F1 while /ɒ/ (in hod) which is a low back vowel (i.e. the tongue is low and to the

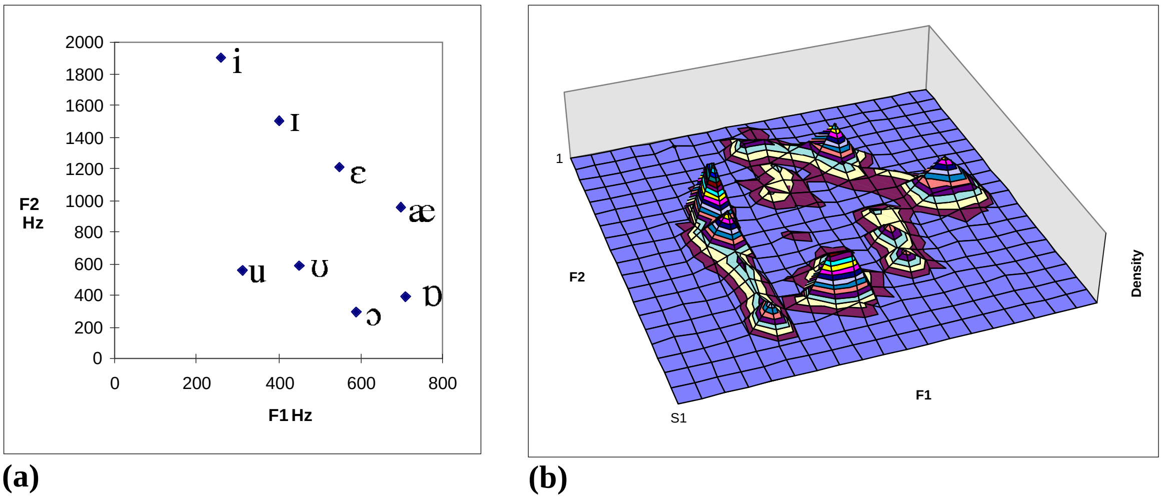

Figure 1: **(a)** The ‘vowel space’. A formant chart showing the frequencies of the first and second formant for eight American English vowels. heed /i/, hid /ɪ/, head /ɛ/, had /æ/, hod /ɒ/, hawed /ɔ/, hood /ʊ/ and who’d /u/. **(b)** Three dimensional view of citation speech. A scatter plot of F1/F2 values from vowels in citation speech show how actual values produced relate to the vowel space. If the density of the scatter is plotted as a third dimension a 3d plot of the vowel space is produced. No scale is marked due to pre-processing.

back when producing this vowel) has high F1 and a low F2. It is possible to plot the F1 value against the F2 value of different vowels (See Figure 1a).

This two dimensional space can be referred to as the vowel space. The triangular shape made by the three vowels /i, u, ɒ/ (heed, who’d, hod) is often referred to as the vowel triangle. The vowel space is of 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” [5, p78].

A scatter plot of F1/F2 values from vowels in citation speech show how actual values produced relate to the vowel space. If the density of the scatter is plotted as a third dimension a 3d plot of the vowel space is produced. From this (Figure 1b) the hills show locations of high density. The values in the hills would tend to correspond to an example of a particular vowel.

The vertical peaks to the left of the vowel triangle in Figure 1b are caused by nasals. 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, however nasals /mnŋ/ are very prominent. All models described here were built using unsegmented data and therefor have included nasals in the vowel space.

No scale is marked on these density plots because pre-processing includes:

Transformation from frequency in hertz to the Bark scale The transformation used to convert frequency into Barks is an approximation suggested by Zwicker and Terhardt [11]. The Bark scale represents the ability of the human ear to distinguish different tones at different frequencies [10, 11]. For example the human ear is more sensitive to tonal differences between 1000Hz and 2000Hz than between 4000Hz and 5000Hz. 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.

Use of a curve fitting algorithm to estimate steady state formant values within the vowel. 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 it was necessary to measure the F1/F2 values automatically. We used LPC (linear predictive coding) to calculate both the probability that voicing was taking place and the likely position of the formants. A parametric curve was then used to estimate the vowel formant targets by fitting the best parametric curve to a number of formant values over a time window [7]. The maximum or the minimum of the curve was regarded as the final spectral target that this formant was heading towards or away from (See Figure 2a). For more detail on this pre-processing stage see [1, 2].

Normalisation to give both dimensions a mean of 0 and a standard deviation of 1. This had the effect of stretching and squashing the F1/F2 dimensions so that nearly all the data fell within a square of size -2.5 sds to 2.5sds. This made it easier to compare different plots between different speakers.

The 3d plot (Figure 1b) can be related to Figure 1a showing the 'vowel triangle'. 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. A probability density function (pdf) constructed from a mixture of Gaussians does exactly this and the EM (expectation maximisation algorithm) is able to fit this pdf to a set of data.

2.1. 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 Gaussians can be added together to model a complex distribution. The expectation maximisation (EM) algorithm will, given a specified number of Gaussians, fit them to a distribution. I will not give a detailed account of the mathematical thinking behind the EM algorithm. This has been treated in some detail in other statistics and maths literature. For a clear and detailed account refer to [3, chapter 2] or [4].

The calculations that are required to run the algorithm are as follows.

Given a set of n points with vectors $\mathbf{x}$, $\mathbf{M}$ Gaussians, the initial probabilities of a jth Gaussian occurring $P(j)$, a covariance matrix Σ_j and a vector of means μ_j, recompute new $P(j)$, Σ_j and μ_j.

For the case where we allow no covariance between dimensions (in fact F1/F2 are fairly independent) the covariance matrix has only the variance for each dimension along the diagonal. To simplify the calculation this can be thought of as a vector of standard deviations σ_j.

The formulae to recompute the parameters are as follows:

To recompute the new means:

$$\mu_j^{new} = \frac{\sum_n P^{old}(j|\mathbf{x}^n)\mathbf{x}^n}{\sum_n P^{old}(j|\mathbf{x}^n)} \tag{1}$$

To recompute the new variances:

$$(\sigma_j^{new})^2 = \frac{\sum_n P^{old}(j|\mathbf{x}^n)(\mathbf{x}^n - \mu_j^{new})^2}{\sum_n P^{old}(j|\mathbf{x}^n)} \tag{2}$$

To recompute the new probabilities of a Gaussian occurring:

$$P(j)^{new} = \frac{1}{N}\sum_n P^{old}(j|\mathbf{x}^n) \tag{3}$$

Where:

$$P(x|j) = \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_j)^T\Sigma_j^{-1}(\mathbf{x}-\mu_j)\right\} \tag{4}$$

Taking $\mathbf{\Sigma}_j$ as the covariance matrix with σ_j^2 along the diagonals, this is the basic equation for a Gaussian.

And where:

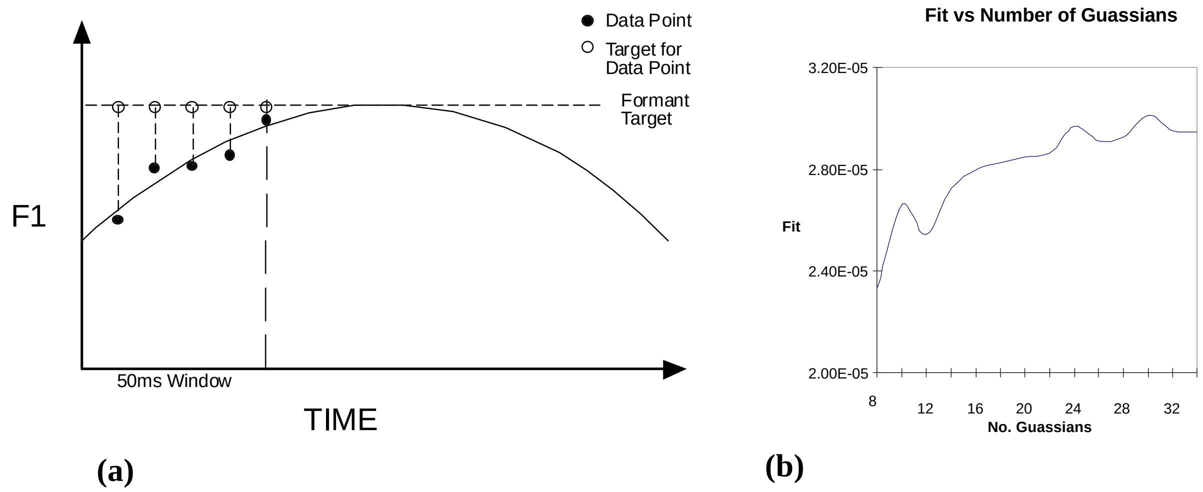

Figure 2: (a) Using a parametric curve to estimate formant targets for vowels. This data is then used to plot F1 v F2 for each 10ms frame within voiced speech. See [1] for more details. (b) Fit of models for different numbers of Gaussians. Fit is poor for too few Gaussians but becomes more unstable and risks over fitting with too many. 20 Gaussians were chosen for the modelling process.

$$P(x) = \sum_{j=1}^{M} P(\mathbf{x}|j)P(j) \tag{5}$$

And using Bayes theorem:

$$P(j|x) = \frac{P(x|j)P(j)}{P(x)} \tag{6}$$

The fit function being maximised is the average log likelihood of the data fitting the distribution:

$$Fit = \frac{1}{n}\sum \log(P(x)) \tag{7}$$

The EM algorithm is an iterative algorithm that will reach a maximum fit although the maximum fit it finds may only be a local maximum. This problem is general to all hill climbing algorithms such as the EM algorithm. 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 will 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, standard deviations and probability of occurring for that Gaussian. The fit is calculated by summing the probability of the pdf producing every point in the data set.

The algorithm is unsupervised. 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.

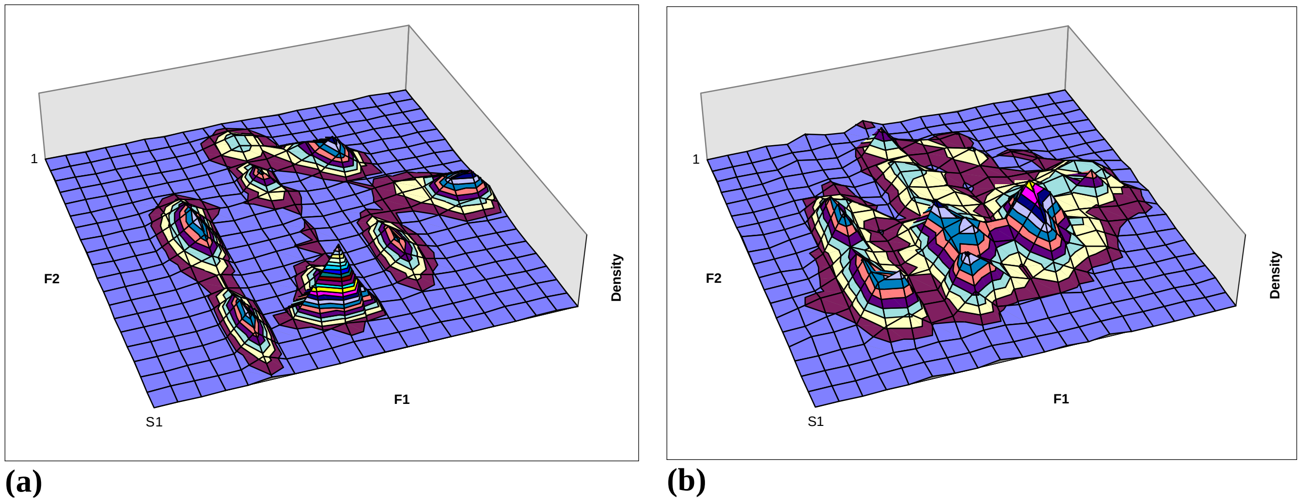

Figure 3: (a) Data from figure 1b modelled using the EM algorithm using 20 Gaussians. **(b)** F1/F2 density plot for spontaneous running speech.

There are, however, two disadvantages. Firstly it is necessary to choose the number of Gaussians in advance. On what basis do we choose this number? Secondly how can we ensure the algorithm does not get stuck in a local maxima? There is no theoretically bomb proof means of answering these questions. However a pragmatic approach to the problem can produce interesting results.

It is possible to look at final fit over different numbers of Gaussians. 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 1b) and because (as can been seen in Figure 2b) the improvement appears to both level off and become more unstable after about 20 Gaussians.

In order to avoid local maxima it is necessary to run the EM algorithm a number of times. The hope is that local maxima will generally be less stable than global maxima and thus it would be very unlucky, using random starting parameters to find the same local maxima on several occasions. Over 10 trials the results from the model appeared generally stable. The result of applying the 20 Gaussian mixture model to the data in Figure 1b is shown in Figure 3a. As can be seen the mixture function has successfully modelled the main peaks in the original distribution.

3. USING THE MODEL

3.1. Comparing citation speech to running speech

Figure 3b shows data taken from running speech. If Figure 1b is compared with Figure 3b a number of typical effects of running speech are clearly visible. There is a large amount of centralisation of vowels. The variance of the vowel types has increased merging the distinct hills and, finally, many vowels have been reduced to schwa (/ə/ the ‘a’ in ‘about’) 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 [9]; 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. 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.

3.2. 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 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.

3.3. 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.

4. CONCLUSION

Like many automatic methods this technique can suffer from noise. The formant tracker is not completely dependable. LPC trackers are particularly prone to error when confronted with nasalised vowels. The parametric curve fitting will only approximate the vowel target and in some circumstances might get it completely wrong especially when dealing with short vowels (less than 40ms). There is also no guarantee that the EM algorithm will produce the best model. However mapping a vowel space by hand is a lengthy business. A simple automatic alternative that produces a visual output that phoneticians can relate easily to previous work offers an important practical contribution to the study of vowel production and perception.

References

[1] Matthew Aylett. Using statistics to model the vowel space. In *Proceedings of the Edinburgh Linguistics Department Conference*, pages 7–17, 1996.

[2] Matthew Aylett. Modelling clarity change in spontaneous speech. In R. J. Baddeley, P. J. B. Hancock, and P. Foldiak, editors, *Information Theory and the Brain*. Cambridge University Press, New York, 1999.

[3] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[4] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[5] D. B. Fry. *The Physics of Speech*. Cambridge University Press, Cambridge, 1979.

[6] S. Furui. Speaker-dependent-feature extraction, recognition and processing techniques. In *ESCA Proceedings of the Tutorial and Research Workshop on Speaker Characterization in Speech Technology*, Edinburgh, 1990. CSTR.

[7] S. Isard. Formant targets. In *Phonetics and Phonology Workshop, Department of Linguistics, University of Edinburgh*, 1995. Unpublished.

[8] Peter Ladefoged. *Elements of Acoustic Phonetics*. University of Chicago Press, Chicago, 1962.

[9] C. F. Sotillo. *Phonological Reduction and Intelligibility in Task-Oriented Dialogue*. PhD thesis, University of Edinburgh, 1997.

[10] E. Zwicker. Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, 33:248–249, 1961.

[11] E. Zwicker and E. Terhardt. Analytical expressions for critical bandwidths as a function of frequency. *The Journal of the Acoustical Society of America*, 68:1523–1525, 1980.