MODELLING CLARITY CHANGE IN SPONTANEOUS SPEECH

Matthew Aylett
Human Communication Resource Centre, University of Edinburgh

December 8, 1997

1 Introduction

Spoken language can be regarded as the combination of two processes. The first is the process of encoding a message as an utterance. The second is the transmission process which ensures the encoded message is received and understood by the listener.

In this chapter I will argue that the clarity variation of individual syllables is a direct consequence of such a transmission process and that a statistical model of clarity change gives an insight into how such a process functions.

1.1 Clarity

We often don't say the same word the same way in different situations. If we read a list of words out loud we say them differently from when we produce them, spontaneously, in a conversation. Even within spontaneous speech there are wide differences in the articulation of the same word by the same speaker. If you remove these words from their context some instances are easier for a listener to recognise than others. The instances that are easier to recognise share a number of characteristics. They tend to be carefully articulated, the vowels are longer and more spectrally distinct and there is less co-articulation. These instances have been articulated more clearly than others. One extreme example of a clear instance of a word is when a speaker is asked to repeat a word because the listener doesn't understand it. For example:

A. Bread, Flour, Eggs, Margarine.

B. Sorry what was that last item?

A. MARGARINE.

The second instance of 'margarine' will be significantly different acoustically from the first instance. It will be much more clearly articulated.

Work in articulatory phonetics has concentrated on the acoustic properties of 'clear speech' and the associated differences in articulation (Moon and Lindblom, 1994). It has been shown that clear speech is easier to recognise and that it is more intelligible (Payton *et al.*, 1994; Picheny *et al.*, 1985). This variation in spectral quality does not appear to be random but is closely related to prosodic structure (van Bergem, 1988), and to differences in redundancy (Lieberman, 1963; Hunnicut, 1985).

1.2 Clarity and the Transmission Process

Some sections of speech are easy to predict. Lindblom, in his H&H theory (Lindblom, 1990), argues that to put the same amount of articulatory effort into saying a word that the listener should find trivial to recognise from context is not energy well spent. Rather it is better to concentrate one's energy on less redundant words or less redundant parts of words. You are more likely to get:

I'm going t- g- t- th- beach

than:

I- g- to go to the b-

We can regard redundancy as how easy it is to predict a word from context. I would like to extend this sense of context to include the acoustic observations of the word itself. The clearer a word is articulated the greater the probability of the word given the acoustic observations. The redundancy of a word is then the probability of the word given the context multiplied by the probability of the word given the acoustic observations.

In this way poor clarity can make a section of speech less redundant for the listener because it is harder to predict the word. Conversely good clarity can make a section of speech more redundant for the listener because it is easier to predict the word.

Much of the redundancy in language is produced by patterns within the lexicon and high level syntactic and semantic structure. Clarity variation can be regarded as a means of fine-tuning this redundancy in response to communication needs. This is important if the speech signal is degraded by a noisy environment in an unpredictable manner as "[redundancy] assists the transmission of a message over an error-prone communication channel." (Taylor, 1989, p.171).

More subtly we may wish to alter normal redundancy in language to convey meaning and protect what we personally regard as the core part of our message. The same sentence in different environments will require different levels of clarity. We may even wish to avoid making much articulatory effort at all if we are uninterested in the listeners needs.

1.3 The motivation for modelling clarity variation

There are two major problems with regards to showing a clear clarity/redundancy/recognition relationship. Firstly redundancy is by no means a simple measurement. There is a difference between a sound being likely (such as the more common /s/ as opposed to the /ʒ/ in fusion) and thus being more redundant and a word being inferable (such as "a stitch in ..."). Secondly the practical measurement of intelligibility and clarity make it very hard to gather sufficient quantities of data in order to explore traditional statistical measures of redundancy.

By building an effective statistical model of clarity variation the hope is as follows:

1. To be able to apply it to a large number spontaneous speech corpora in order to investigate, at a more statistical level, notions of redundancy and clarity in language.

2. By doing this to build a more complex theory of language structure from a statistical perspective and relate this to suprasegmental structure in language.

I will first give a detailed description of the modelling technique used and then discuss to what extent these objectives have been addressed.

2 Modelling Clarity Variation

A potential approach to modelling clarity variation is to model an individual speaker's clear speech and then compare normal spontaneous speech with this model. The degree with which the spontaneous speech is predicted by the clear speech model gives us our clarity measurement. Given a set of acoustic observations and a clear speech model M this measurement can be expressed as the average log likelihood of the observations given the model.

$$Clarity = \frac{1}{n}\sum_{i=1}^{n}\log(p(x_i|M)) \tag{1}$$

The method I have chosen to model clear speech and compare this model to running speech is summarised as follows:

1. The model is a probability density function in two dimensions described by a mixture of gaussians. The dimensions relate to 1st and 2nd formant frequencies of voiced speech. The model is built by applying the expectation maximisation (EM) algorithm to pre-processed, normalised citation speech.

2. It is based on a large corpus of spontaneous speech; The HCRC Map Task Corpus (Anderson *et al.*, 1991). In order to investigate clarity it

is important to study natural speech as it is 'sloppy', casual speech that exhibits major lack of clarity. A large quantity of data was also required in order to sensibly build and test a statistical model. The HCRC Corpus offers over fifteen hours of spontaneous speech as well as more than two hours of citation speech produced by the same speakers.

3. The approach is one based on 'self organisation' in that the statistical model is formed from underlying structure within the data.

4. The method used to judge clarity is based on vowel quality only. This choice was made in order to simplify the process and because evidence suggests that the stressed vowels within a word make the greatest overall contribution to the intelligibility of the word (Sotillo, 1997). In order to establish the independent contribution of the spectral information within the vowel, duration information is withheld. It is hoped to amalgamate both spectral and duration information in future models.

3 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 Ladefoged, 1962).

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 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 1).

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" (Fry, 1979, 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 2) the hills show locations of high density. The values in the hills would tend to correspond to an example of a particular vowel.

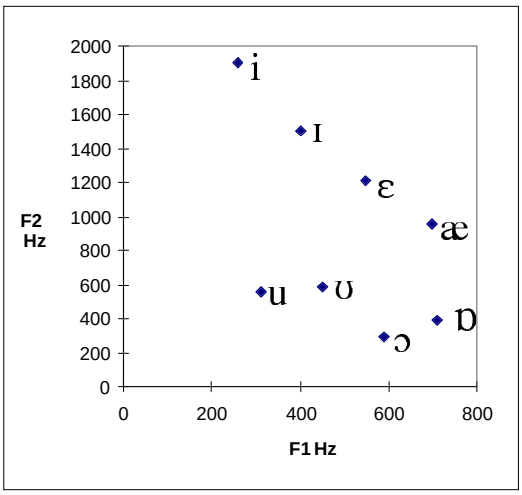

Figure 1: The 'vowel space'. A formant chart showing the frequencies of the first and second formant for eight American English vowels. heed /i/, hid /ɪ/, head /ɛ/, had /æ/, hod /ɒ/, hawed /ɔ/, hood /ʊ/ and who'd /u/.

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

1. 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 (1980). It is a mixture of two arctan curves as follows:

   $$z = 13 \arctan\left(\frac{f}{1000} 0.76\right) + 3.5 \arctan\left(\frac{f}{7500}\right) \tag{2}$$

   Where z is Barks and f is the frequency in Hz.

   The Bark scale represents the ability of the human ear to distinguish different tones at different frequencies (Zwicker, 1961; Zwicker and Terhardt, 1980). 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.

2. 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

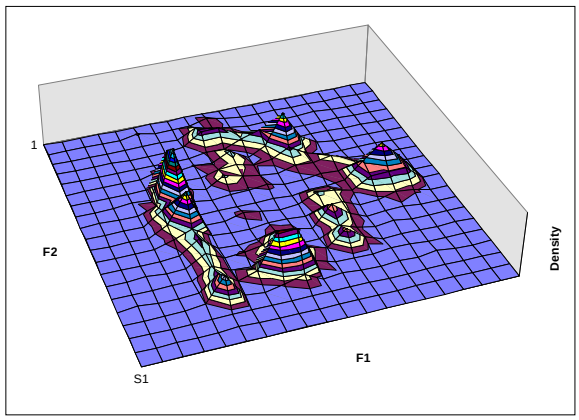

Figure 2: Three dimensional view of citation speech. A scatter plot of F1/F2 values from vowels in citation speech show how actual values produced relate to the vowel space. If the density of the scatter is plotted as a third dimension a 3d plot of the vowel space is produced. No scale is marked due to pre-processing. See details in section 3.

automatically. 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. A parametric curve can then be used to estimate the vowel formant targets by fitting the best parametric curve to a number of formant values over a time window.[1] 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 (Figure 3). For more detail on this pre-processing stage see Aylett (1996).

3. Normalisation to give both dimensions a mean of 0 and a standard deviation of 1.

   This has the effect of stretching and squashing the F1/F2 dimensions so that nearly all the data falls within a square of size -2.5 sds to 2.5sds. This makes it easier to compare different plots between different speakers.

The 3d plot (Figure 2) can be related to Figure 1 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.

[1] My implementation of this technique is based on a talk given by Steve Isard to the Phonetics and Phonology group at Edinburgh University in 1996

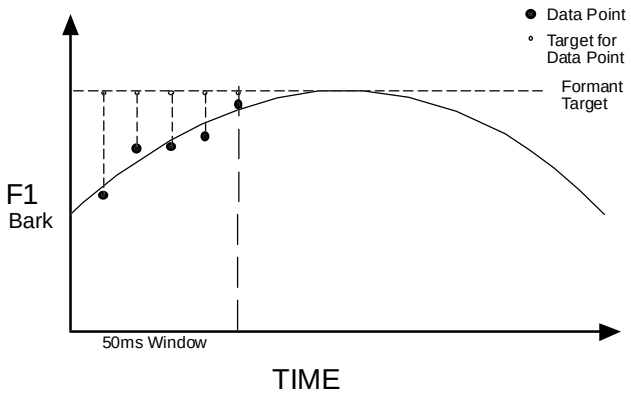

Figure 3: Using a parametric curve to estimate formant targets for vowels. This data is then used to plot F1 v F2 for each 10ms frame within voiced speech. See (Aylett, 1996) for more details

3.1 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 Bishop (1995, chapter 2) or Duda and Hart (1973).

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:

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

To recompute the new variances:

$$(\sigma_j^{new})^2 = \frac{\sum_n P^{old}(j|\mathbf{x}^n)(\mathbf{x}^n - \mu_j^{new})^2}{\sum_n P^{old}(j|\mathbf{x}^n)} \tag{4}$$

To recompute the new probabilities of a Gaussian occurring:

$$P(j)^{new} = \frac{1}{N}\sum_n P^{old}(j|\mathbf{x}^n) \tag{5}$$

Where:

$$P(x|j) = \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_j)^T\Sigma_j^{-1}(\mathbf{x}-\mu_j)\right\} \tag{6}$$

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:

$$P(x) = \sum_{j=1}^{M} P(\mathbf{x}|j)P(j) \tag{7}$$

And using Bayes theorem:

$$P(j|x) = \frac{P(x|j)P(j)}{P(x)} \tag{8}$$

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

$$Fit = \frac{1}{n}\sum \log(P(x)) \tag{9}$$

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.

The problem of local maxima 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.

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 (see Figure 4). 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 4) 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 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.

The result of applying the 20 Gaussian mixture model to the data in Figure 2 is shown in Figure 5. As can be seen the mixture function has successfully modelled the main peaks in the original distribution.

4 Using the Model to Calculate Clarity

4.1 Comparing Citation Speech to Running Speech

Figure 6 shows data taken from running speech. If Figure 2 is compared with Figure 6 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 (Sotillo, 1997); 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

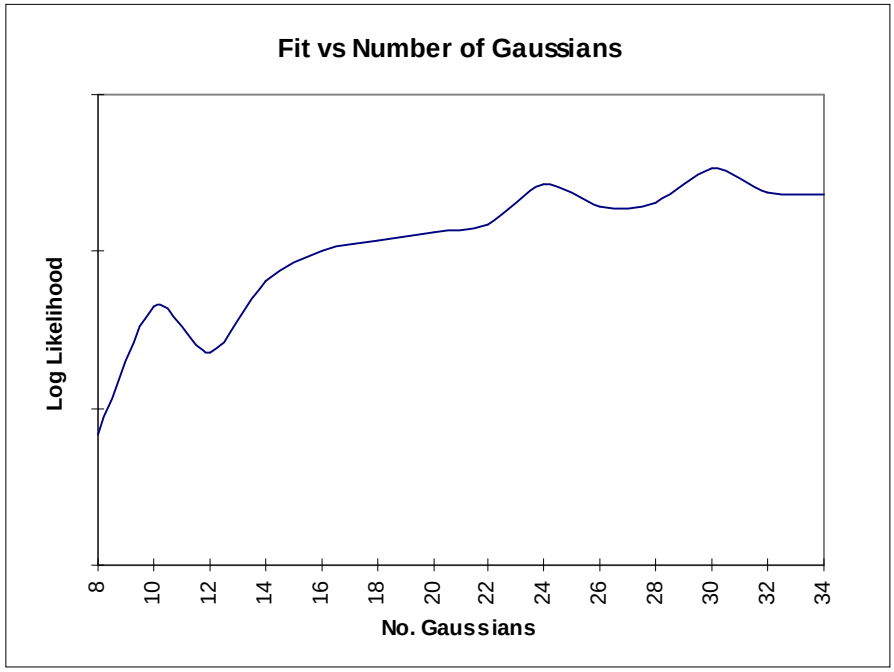

Figure 4: Fit of models for different numbers of Gaussians. Fit is poor for too few Gaussians but becomes more unstable and risks over fitting with too many. 20 Gaussians were chosen for the modelling process.

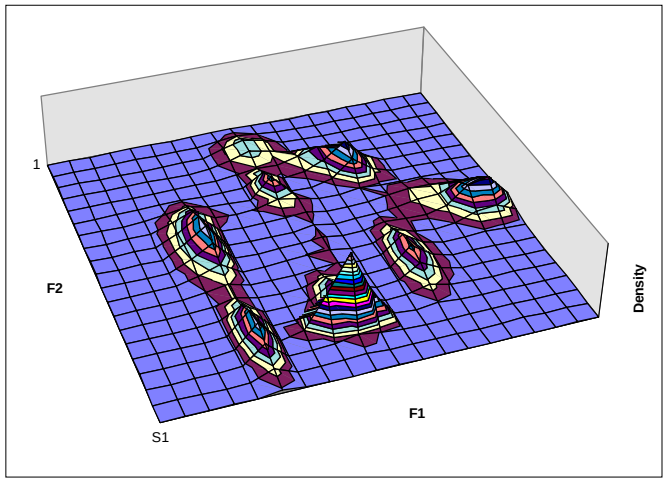

Figure 5: Data from figure 2 modelled using the EM algorithm using 20 Gaussians.

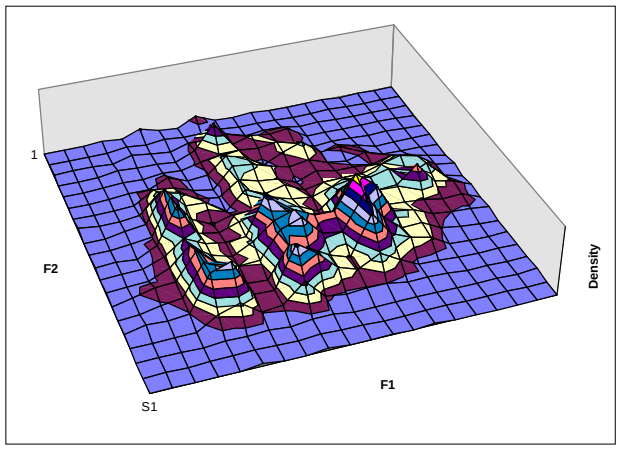

Figure 6: F1/F2 density plot for spontaneous running speech..

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.

4.2 Calculating a clarity score for a section of speech in the HCRC map task corpus

The process for calculating the clarity of a section of speech is as follows:

1. Build a pdf mixture Gaussian model of a speaker's citation speech using the EM algorithm.

2. Take the target speech and preprocess in the same way as the citation speech.

3. Calculate the log likelihood that the citation speech pdf would produce the data points.

Figure 7 shows the log likelihood of a section a speech with regards to a citation speech model. The sentence is as follows - "right, you got a map with an extinct volcano?". The /ɑi/ in right, the /æ/ in map and the /l/ in /v l k ei n əu/ appear clearly articulated.

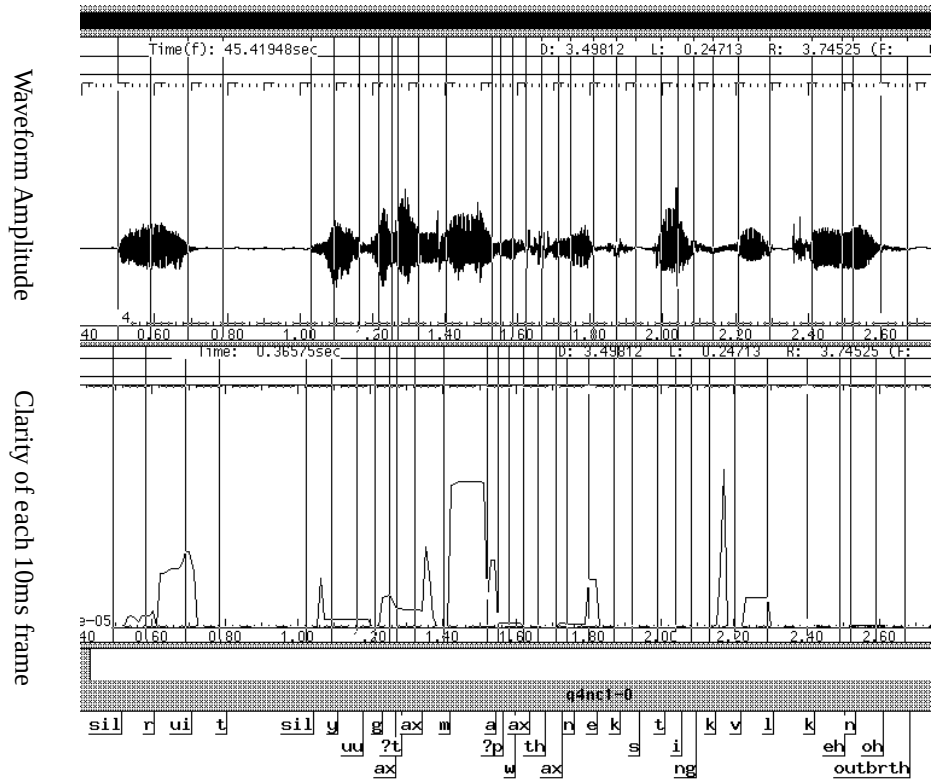

Figure 7: Clarity scores expressed as the probability of an F1/F2 point within voiced speech ocurring given the citation model for the same speaker. The transcription shown at the bottom is in non standard machine readable phonetic alphabet. It translates as follows: 'sil' - silence, 'r ui t' - right /rɑit/, 'y uu' - you /yu/ , 'g ax ?t ax' - gotta /ɡ ə t ə/, 'm a ?p' - map /m æ p/, 'w ax th' - with /w ə θ/, 'ax n' an /ə n/, 'e k s t i ng t' extinct /e k s t ɪ ŋ t/, 'v l k eh n oh' volcano /v ɒ l k eɪ n əu/.

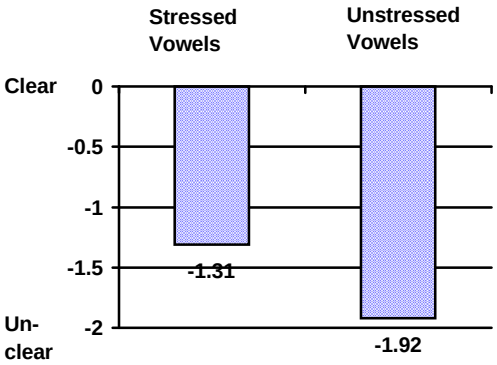

Figure 8: Clarity variation between vowel type and lexical stress. A model was built for 4 of the 64 speakers in the HCRC Map task corpus. Each speakers models was applied to their first dialogue. The average log liklehood of each vowel over its duration within each syllable was calculated according to the citation model. A crossed design ANOVA shows a significant effect ($F(1,5581)=118.63; p<0.001$) between stressed and unstressed syllables. Syllables with no vowel and syllables with secondary stress were ignored. This left 5583 syllables from 7187.

5 Evaluating the model

5.1 Does the clarity score agree with the expectations of vowel clarity in different phonetic contexts?

The clarity measurement depends on relating the vowel targets of a speaker (In this case automatically determined through a speaker's citation speech) to the vowel targets attained in running speech. Lexical stress affects vowel targets. Vowels in stressed syllables have less variation in their spectral target than vowels in unstressed syllables. In running speech closed class words (such as the, an, a, of etc.) tend to be realised lexically unstressed. We would therefore expect vowels in monosyllabic closed class words in running speech and in unstressed syllables in open class words to have a lower clarity score than in the stressed syllables.

A significant effect ($F(1,5581)=118.63; p<0.001$) was shown to exist between stressed and unstressed syllables. See Figure 8. In this case clarity scores reflect our expectations. Stressed vowels are clearer than unstressed vowels.

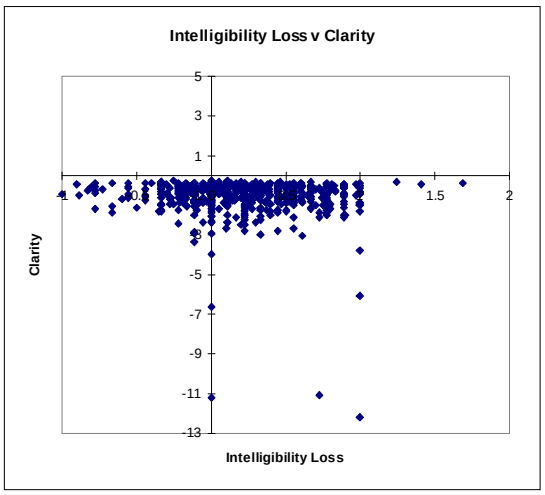

Figure 9: A weak negative correlation exists between the clarity score of a running speech token and the token's loss of intelligibility between citation form and running speech form (n=806, r =-0.119, p<0.001)

5.2 Does the clarity score relate to psycholinguistic measurements of intelligibility?

5.2.1 Intelligibility

Intelligibility data produced at The HCRC, University of Edinburgh, from the HCRC Map corpus (Bard *et al.*, 1995) was compared to clarity scores. 806 words were excised from spontaneous speech together with a citation form which is used as a control. These words were played to subjects who tried to recognise each word (See Bard *et al.*, 1995, for more details). The citation control is used to minimise intelligibility effects caused by word frequency, word structure, context and speaker. The assumption is, that when these factors are controlled for, the intelligibility loss from the citation form to the token represents a difference in the acoustic properties between these two forms. The clarity score is attempting to measure the difference in the acoustic properties of a vowel in running speech with a vowel in citation form. It was therefore hoped that the clarity of a token would be negatively correlated with the intelligibility loss of a token. It was also hoped that speakers that produced more intelligible speech would also produce clearer speech in terms of higher clarity scores.

A weak negative correlation exists between the clarity score of a token and the tokens loss of intelligibility between citation form and running speech form (n=806, r =-0.119, p<0.001). See Figure 9.

The low value of r is possibly due to:

1. The clarity machine is too noisy to effectively predict the clarity of indi-

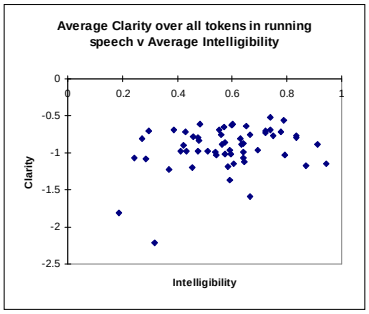

Figure 10: A positive correlation exists between the average clarity of a speaker's running speech tokens and the average intelligibility of the tokens (n=55, r=0.326, p<0.05)

vidual tokens.

2. Noise in the Intelligibility measurement.
3. Other factors obscuring the relationship such as duration and amplitude.
4. The relationships are non-linear. There is no actual reason why the relationship between any of these factors should be linear.

Some speakers appear to be easier to understand than others. The average intelligibility of all their running speech tokens is higher. A positive correlation exists between the average clarity of a speaker's tokens and the average intelligibility of the running speech tokens (n=55, r=0.326, p<0.05). See Figure 10. To a certain extent the model appears to reflect some inter-speaker differences in intelligibility.

The mixture of Gaussian models achieved different levels of fit for different speakers. The final log likelihood of the model representing the citation data correlates with the intelligibility of words spoken in citation form by the same speaker. (n=60, r=0.296, p<0.05) See Figure 11. This result suggests that the acoustic features that make citation speech difficult to understand are the same features that make it hard to model.

This is not a consequence of higher entropy. It is not that unclear speakers have more unpredictability in their citation speech than clear speakers. There is no obvious relationship between the entropy of a speaker's citation model and the average citation intelligibility of a speaker. It is the type of structure in the citation speech not the existence of structure that seems to relate to speaking unclearly. For example clear speakers appear to have a broader range in their F2 values.

Overall the evaluation is promising. Given the imperfect nature of the model together with difficulties in measuring intelligibility and the acoustic features of running speech the model appears to make sensible predictions about vowel quality and word intelligibility.

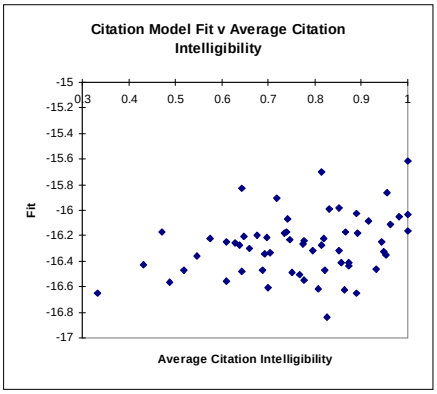

Figure 11: The final log likelihood of the model representing the citation data correlates with the intelligibility of words spoken in citation form by the same speaker. (n=60, r=0.296, $p<0.05$)

6 Summary of Results

6.1 The Relationship between the Modelling Process and Psycholinguistic Measurements

1. There is a weak relationship between intelligibility loss, which is a psycholinguistic measurement, and clarity, which is an acoustic measurement based on a statistical model of citation speech. Figure 9.

2. There is a stronger relationship between speakers' average intelligibility in running speech and the average clarity of their running speech. Figure 10.

3. The more unintelligible a speaker's citation speech the poorer the fit of the final statistical model based on the EM algorithm. This is not caused by variation in entropy. Figure 11.

6.2 Redundancy/Articulation/Recognition

1. Clear tokens are more intelligible and thus easier to recognise. Figure 9

2. Stressed syllables are clearer than unstressed syllables. Figure 8.

3. 90% of tokens start with a stressed syllable (Cutler and Norris, 1988). The beginning of a word is a hot spot of low redundancy.

Thus the performance of the model supports the premise that high redundancy items in language are articulated poorly and are more difficult to recognise out of context. In turn this supports the assertion that we are controlling levels of redundancy in order to improve the robustness of the transmission process.

6.3 Conclusion

None of these results, given work in psycholinguistics and experimental phonetics are very surprising. Results from both fields suggest this redundancy/articulation/recognition relationship exists. However I believe that using a statistical method and a corpus of spontaneous speech to investigate speech at a phonetic level is a powerful approach particularly as we have a large body of data from laboratory phonetics to guide both the modelling process and application of such a model.

7 Discussion

What does this work tell us about brain function? Overall a simplistic model such as presented here cannot say very much. The results from this work do suggest a special significance for citation speech but they don't establish this fact. Perhaps citation speech has a special status in language because it is clear speech and clear speech is characterised, at a statistical level, as something that unsupervised cluster analysis algorithms can model easily.

I have presented a simple model of the clarity change within spontaneous speech and tried to relate results from it to the structure of language. The hope is that this chapter can be viewed in the context of what has been presented elsewhere in the book. I am neither a neuroscientist nor an expert on information theory. I do however know something about speech and I do believe that ideas emerging from these disciplines are of fundamental importance in the understanding and structure of spoken language.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty-Sneddon, G. M., Garrod, S., Isard, S., Kowtko, J. C., McAllister, J. M., Miller, J. E., Sotillo, C. F., Thompson, H. S., and Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, **34**(4), 351–366.

Aylett, M. (1996). Using statistics to model the vowel space. In *Proceedings of the Edinburgh Linguistics Department Conference*, pages 7–17.

Bard, E. G., Sotillo, C. F., Anderson, A. H., Doherty-Sneddon, G. M., and Newlands, A. (1995). The control of intelligibility in running speech. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 4, pages 188–191.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Cutler, A. and Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**(1), 113–121.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.

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

Hunnicut, S. (1985). Intelligibility versus redundancy – conditions of dependency. *Language and Speech*, **28**, 45–56.

Ladefoged, P. (1962). *Elements of Acoustic Phonetics*. University of Chicago Press, Chicago.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, **6**, 172–187.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H & H theory. In W. J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Moon, S.-J. and Lindblom, B. (1994). Interaction between duration, context and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America*, **96**, 40–55.

Payton, K. L., Uchanski, R. M., and Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, **95**, 1581–1592.

Picheny, M., Durlach, N., and Braida, L. (1985). Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, **28**, 96–103.

Sotillo, C. F. (1997). *Phonological Reduction and Intelligibility in Task-Oriented Dialogue*. Ph.D. thesis, University of Edinburgh.

Taylor, M. M. (1989). Response timing in layered protocols: a cybernetic view of natural dialogue. In M. M. Taylor, F. Neel, and D. G. Bouwhuis, editors, *The Structure of Multimodal Dialogue*, pages 403–439. Elsevier Science Publishers (North Holland), Amsterdam, The Netherlands.

van Bergem, D. R. (1988). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, **12**, 1–23.

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

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