VOWEL QUALITY IN SPONTANEOUS SPEECH: WHAT MAKES A GOOD VOWEL?

*Matthew Aylett** *Alice Turk***

Human Communication Research Centre*,
Department of Linguistics**,
University of Edinburgh

ABSTRACT

Clear speech is characterised by longer segmental durations and less target undershoot [9] which results in more extreme spectral features. This paper deals with the clarity of vowels produced in spontaneous speech in a large corpus of task-oriented dialogues. We present an automatic technique for measuring vowel clarity on the basis of a vowel's spectral characteristics. This technique was evaluated using a perceptual test. Subjects rated the 'goodness' of vowels with different spectral characteristics with controlled duration and amplitude and these results were compared with an automatic rating. Results indicated that although agreement between subjects and the automatic measurement was poor it was as poor as the agreement between subjects.

On the basis of these results we address the following questions:

1. Can subjects reliably judge the clarity of vowels excerpted from spontaneous speech without duration cues?
2. Can a statistical model [3] reliably predict the subjects' response to such vowels?

1. INTRODUCTION

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.

Work in articulatory phonetics has concentrated on the acoustic properties of 'clear speech' and the associated differences in articulation [9]. It has been shown that clear speech is easier to recognise and that it is more intelligible [10, 11]. This variation in spectral quality does not appear to be random but is closely related to prosodic structure [12], and to differences in redundancy [7, 6].

However spontaneous speech is often extensively reduced both in terms of duration and spectral clarity. It is possible that our sensitivity to such spectral change may differ significantly at different levels of reduction. Can a vowel only get so clear or so unclear so that any further change in spectral characteristics is perceptually insignificant?

By producing a model of vowel variation change and comparing the behaviour of the model to results from a perceptual experiment we hope to ascertain how sensitive human subjects are to spectral change and how successfully a model can predict the human response. If such a model reliably reflects human judgements it can be used as an objective measurement of vowel spectral clarity. Such an automatic measurement would offer an alternative to time consuming measurements by hand and allow the rapid spectral measurement of thousands of vowels in large spontaneous speech corpora.

2. MODELLING VOWEL CLARITY VARIATION

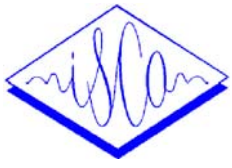
In order to model vowel clarity variation we first produce a statistical model which characterises a speaker's vowel space. Vowels produced in spontaneous speech are then related to this model and results from this comparison are used to produce an objective measurement of care of vowel articulation.

The model is based on 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. The pre-processing involves a transformation to the bark scale [14], use of a curve fitting algorithm to estimate steady state formant values within a vowel [2] and normalisation of both dimensions to give a mean of 0 and a standard deviation of 1. For a detailed description of the modelling and normalisation techniques see [3]. Figure 1a shows a model of a speaker's vowel space generated in this way. Compare this to a 3 dimensional plot of spontaneous speech from the same speaker (Figure 1b) where the centralisation and smearing of vowel groups is clearly visible.

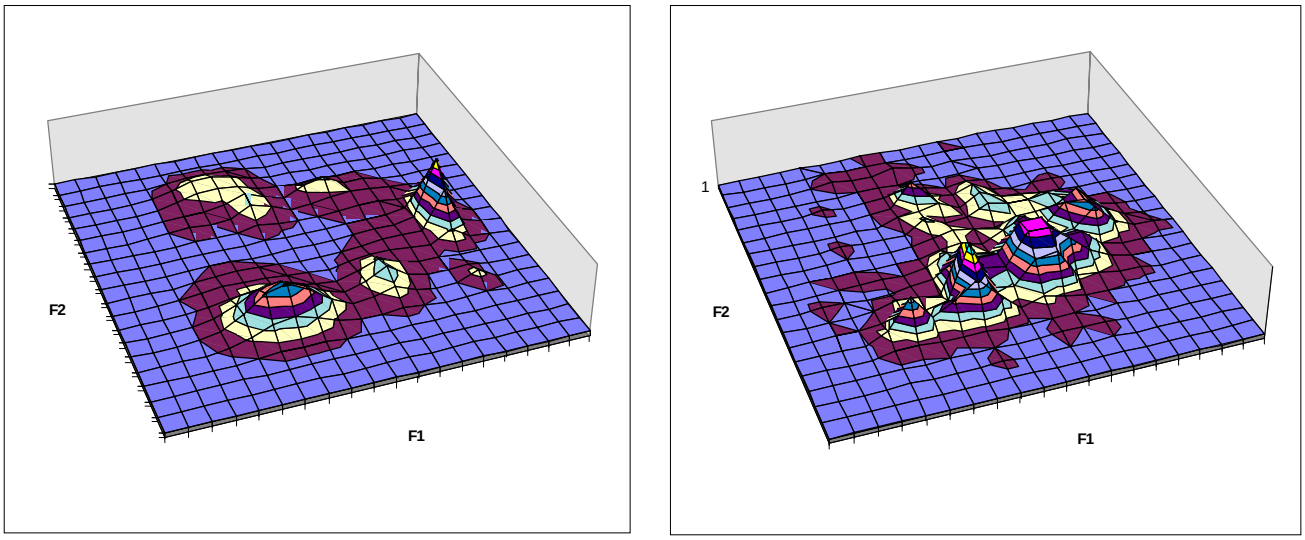
To use this model to score vowel clarity we take the vowel targets from the vowel in spontaneous speech (as computed using the pre-processing techniques) and calculate the probability of the targets appearing in the 'clear' citation speech. In other words how close to the 'hills' in the model are to the targets. These probabilities are then combined as an average log likelihood.

There are some clear drawbacks with this technique:

1. Although first and second formants are a good way to characterise a vowel many other acoustic factors are involved in a judgement of spectral characteristics.
2. Automatic formant trackers and automatic segmentation



5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998



(a) **(b)**

Figure 1: (a) Three dimensional view of a mixture of Gaussians statistical model of vowels produced in citation speech. Compare with a similar view of the same speaker's spontaneous speech. (b) Three dimensional view of spontaneous 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.

(both used to generate the model and select input data) are not completely reliable. These automatic approaches will introduce noise into both the model and the vowel clarity measurement.

3. The citation speech used to build the speaker models was not designed for this purpose. The clarity of this speech is sometimes questionable and no attempt was made to phonetically balance the sample.
4. The model is trained on all vowels. If an 'unclear' example of one vowel has similar spectral characteristics to another vowel the model cannot tell if it is a bad example of vowel 1 or a perhaps a better example of vowel 2.
5. The model is speaker dependent. A different model is produced for each speaker. Some speakers may be harder to model than others.
6. The model ignores the phonetic context. Human subjects could possibly compensate for perceived co-articulation when judging vowel 'goodness'. Although using parametric curve fitting to estimate vowel targets reduces co-articulatory effects to some extent the model itself does not explicitly take into account the identity of the surrounding phonemes.

However the technique allows the automatic measurement of fifteen hours of spontaneous speech produced by 64 different speakers. This produces about 75,000 data points covering all vowel types and many different phonetic contexts. In order to assess the model and perceptual salience of clarity variation materials were selected from this data and presented to human subjects. The perceptual experiment investigated the extent the model predicted human responses and the extent subjects could reliably perceive the spectral variation in different vowels.

3. PERCEPTUAL EXPERIMENT

3.1. Method

32 subjects (23 British English native speakers of which 12 had a Southern British accent, 7 were Northern British, 3 were Scottish and 1 Irish together with 4 North American English native speakers and 5 non-native speakers) were played 90 vowels excerpted from spontaneous speech together with 90 matched fillers taken from citation speech and asked to rate their 'goodness' using magnitude estimation. Magnitude estimation is a technique often used in psychophysics to validate and construct scales of physical sensations. The main advantage of magnitude estimation over more traditional rating scales or visual analogue scales is that the scale used to measure subjects response does not affect the response. In magnitude estimation a subject decides on their own scale based on the first stimulus and uses that first response as a yardstick to measure all others. In order to compare results between subjects the responses are log transformed. For a clear and concise introduction to magnitude estimation see [8].

The vowels used all had durations between 90-110ms, had their amplitude normalised and were excerpted from the HCRC Map Corpus [1]. Segmentation was achieved by combining word segmentation done by hand with phonemic auto-segmentation carried out using the HTK toolkit [13] and hand corrected entries from the CELEX online dictionary [4]. The vowels represented 3 vowel types (one from each corner of the vowel triangle), 3 levels of clarity (high, medium, low) as calculated using the model described. Each cell of ten stimuli had a matching set of ten citation fillers with similar clarity scores, durations and speakers. The speakers who produced each of the ten stimuli in each cell

were different and split equally between male and female speakers. Where possible the same speakers were used in each cell.

Clarity groups were decided on the basis of the distribution of the clarity score of all 90-110ms vowels. The mean of the log likelihood clarity score of the vowels was -16.912. Any vowels with a clarity of less than -16.75 where regarded as low clarity items. Items above -16.5 were divided into two further groups, those with a clarity between -16.5 and -15.5 which were regarded as medium and those with a clarity of greater then -15.25 which were regarded as high clarity items. The standard deviation of the clarity score was 2.154.

Each subject was first given a practise exercise in Magnitude Estimation training them to use this technique to judge line lengths. They then listened to some randomly selected sections of spontaneous speech produced by Glaswegian Speakers and to some example vowels excerpted from this speech. They then carried out a short practise session judging the vowel quality of 10 vowels before taking part in the main experiment. In the main experiment they were played 60 randomised examples of each vowel (i as “ee” in “street”, o as “o” in “gold” and a as “a” in “cat”), they were given the word the vowel was taken from and asked to judge how good they thought the vowel sounded. The order of presentation of vowels was varied amongst subjects in case of an ordering effect.

Each vowel was presented twice with a 2 second gap between each presentation and a 4 second gap and a beep between each vowel. Vowels were blocked into groups of ten and data was captured using netscape and a web interface.

3.2. Results

The results were analysed as follows:

1. by-subjects ANOVA
2. by-materials ANOVA
3. Linear correlation between clarity as assigned by the statistical model and pooled subject responses
4. Cluster analysis of subjects responses

By-Subjects ANOVA. The by-subjects ANOVA used subject linguistic background (Native English, Native North American, Non-Native) as a grouping variable with vowel (i, o, a) and clarity as calculated by the model (high, medium, low) as crossed variables.

Surprisingly the linguistic background had no significant effect on the responses. Subjects from Germany and Poland rated vowels similarly to Native English speakers. As I will discuss later this probably has more to do with the basic difficulty of the task than some underlying similarity in vowel sensitivity.

Similarly vowel type alone had no significant effect on results although there was a vowel/clarity interaction $(F(4,96) = 4.15, p < 0.005)$. However clarity group $(F(2,48) = 20.75, p < 0.001)$ did have a significant effect on the subjects responses. The means of the responses for spontaneous speech within each clarity group were as follows:

By-Subjects Responses			
Clarity Group	High	Med	Low
Geometric Mean	0.883	0.799	0.777

This supported the hypothesis that the clarity model was modelling subjects response to some extent. Low, medium and high clarity groups as decided by the clarity model reflected low, medium and high responses from subjects.

By-Materials ANOVA. Following the non significant effect of subjects linguistic background these responses were pooled. In the by-materials ANOVA sex of speaker, vowel type and clarity group were used as grouping variables.

The clarity group result persisted in the by-materials analysis $(F(2,72) = 3.71, p < 0.05)$. Again the pattern of means supported the hypothesis:

By-Materials Responses			
Clarity Group	High	Med	Low
Geometric Mean	0.69	0.625	0.582

The difference in significance between by-subject and by-materials analyses suggests there is too much variance unaccounted for in the materials. This led to a re-examination of the clarity score. Noise is unquestionably in the system. This noise will produce spurious F1/F2. The likely effect of this is to produce very low clarity scores (i.e. nowhere near the distribution of the speakers vowels). Thus very low clarity scores (more than 2 standard deviations from the mean) should be treated with suspicion.

Linear correlation between clarity as assigned by the statistical model and pooled subject responses. Before carrying out a linear correlation between pooled subjects response and raw clarity score in terms of log likelihood it was decided to remove low valued outliers (that is with a value lower than 2 standard deviations from the mean.) Firstly because of suspicions concerning their validity and secondly because of the large effect outliers can have on linear correlation tests. This removed 7 data points from the 90 vowels taken from spontaneous speech. The result was a weak but significant correlation $(r = 0.313, p < 0.005)$.

The model appears to predict only about 10% of the subjects responses.

Cluster analysis of subjects responses In order to investigate agreement between subjects a cluster analysis was carried out on subjects responses. The clustering was carried out using correlation as a distance measurement and maximum similarity (minimum distance), single linkage to combine clusters [5]. No grouping effect was apparent. Agreement between subjects varied considerably. The average correlation between any two subjects is quite low $(r = 0.33)$ but the significance of the agreement between subjects is generally high (79% with a $p <= 0.05$) between all pairwise comparisons. Bearing in mind the difficulty faced by subjects when carrying out the task of rating vowel goodness the statistical model performs comparatively well $(r = 0.313, p < 0.005)$.

4. DISCUSSION

Can subjects reliably judge the clarity of vowels excerpted from spontaneous speech without duration cues? The answer is yes but it's hard. They reliably agree with each other about 10% of the time. Can a statistical model [3] reliably predict the subjects' response to such vowels? Again the answer appears to be yes but, again, it's quite hard only predicting about 10% of the subjects responses. Basically the model is roughly as good – or bad – a predictor of any one listener's judgement as any other listener's judgement is.

Vowel quality in spontaneous speech does contribute to subjects' perception of vowel 'goodness'. However the failure of subjects to agree on individual vowels suggests that this contribution is not a strong one. Duration is a primary factor. Of the 170,000 vowels segmented in the HCRC Map task nearly 100,000 are either too short to measure the spectral target reliably (less 40ms) or were unvoiced. The materials we used in our perceptual experiment did not reflect these short vowels or devoiced vowels. In contrast to materials generated in 'clear speech' experiments, where the scale of vowel articulation varies from clear to very clear, in spontaneous speech the spectral quality of vowels often varies from poor to very poor. Perhaps in these conditions the difficulty in relying on spectral cues alone to perceive vowel quality leads to more reliance on segmental duration. However, in order to establish this, further experiments varying the duration of the segments used would be required.

Finally a clear problem with the approach taken in the modelling strategy is the fact that phonetic context is not taken into account. Rather than the model assigning a clarity score based solely on the F1/F2 targets of the vowel it might be more productive to assign this score on these values given the pre and/or post segmental context. However modelling these factors effectively using the statistical approach described here would require substantial quantities of controlled citation data from each speaker. It is also important to bear in mind that other acoustic factors such as spectral tilt, f0 and amplitude might also make an important contribution to any judgement of a vowel's 'goodness' in spontaneous speech. Although the model could be altered to take such factors into account it is not entirely clear how such factors should be automatically measured and incorporated.

To conclude I would argue that a corpus approach to the analysis such phonetic factors offers a useful contrast to hand measured laboratory studies. Large corpora of spontaneous speech are now available making such an approach tractable as well as producing interesting scientific results to support or confound previous findings in laboratory phonetics.

REFERENCES

[1] Anne H. Anderson, Miles Bader, Ellen G. Bard, Elizabeth Boyle, Gwyneth M. Doherty-Sneddon, Simon Garrod, Stephen Isard, Jacqueline C. Kowtko, Jan M. McAllister, Jim E. Miller, Catherine F. Sotillo, Henry S. Thompson, and Regina Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.

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

[3] 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.

[4] R. H. Baayen, R. Piepenbrock, and L. Gulikers. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995. Version 2.5.

[5] J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

[6] Sheri Hunnicut. Intelligibility versus redundancy – conditions of dependency. *Language and Speech*, 28:45–56, 1985.

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

[8] Milton Lodge. *Magnitude Scaling: Qunatitative Measurement of Opinions*. Sage Publications, Beverly Hills, California, 1981.

[9] Seung-Jae Moon and Björn Lindblom. Interaction between duration, context and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America*, 96:40–55, 1994.

[10] K. L. Payton, R. M. Uchanski, and L. D. Braida. 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, 1994.

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

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

[13] Steve Young, Joop Jansen, Julian Odell, Dave Ollason, and Phil Woodland. *The HTK Book*. Entropic, 1996. Version 2.00.

[14] 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.