Modelling Care of Articulation with HMMs is Dangerous

Matthew P. Aylett

Rhetorical Systems and
Department of Linguistics, University of Edinburgh
Edinburgh, Scotland
matthewa@cogsci.ed.ac.uk

Abstract

Changes in care of articulation (COA) affect both the spectral and durational characteristics of speech. This can have severe repercussions on both the success of speech recognition, and the quality of speech synthesis. Although auto-segmentation has proven useful for measuring the durational effects of COA, an automatic spectral measurement has proven more problematic [5]. In this paper, we will explore the use of the acoustic log likelihoods generated by HMM autosegmentation as a measure of these changes in comparison with two phonetically motivated modeling systems based on vocalic F1/F2 values. When duration variation is controlled, the HMM output does not correlate with the human perception of vowel goodness, whereas, the phonetically motivated models do.

1. Introduction

The amount of care, or effort, that a speaker makes when articulating speech sounds varies with particular circumstances. These circumstances may vary from the overall register, for example, speakers tend to hyper-articulate when speaking to small children [9] or non-native speakers [10], to particular instances of words, for example repeating a word because a listener did not understand [13]. There is considerable evidence that such variation can also be linked to prosodic structure and redundancy in language [5]. Such suprasegmental factors are important in the generation of high quality natural sounding synthesised speech as well as the automatic recognition of speech where modeling potential variation in segmental characteristics can aid recognition success.

Work on care of articulation (COA) in phonetics has concentrated on vowel production [11, 14, 8] and has shown that in poorly articulated speech (or hypo-articulated speech) the formants of a vowel tend to 'undershoot' the clear vowel target. In this work we present three models of COA based on the spectral characteristics of vowels. One is based on the acoustic log likelihood of a vowel from an HMM autosegmenter while the other two are based on F1/F2 formant models. The output of each model was compared to results from a perceptual test where subjects were asked to rate the 'goodness' of a vowel.

2. Materials

The models were trained and applied to spontaneous and citation speech taken from the HCRC Map Task Corpus [1]. The corpus contains approximately 15 hours of spontaneous dialogue spoken by 64 speakers. In addition each speaker was recording reading a list of map landmarks clearly and slowly.

3. The Models

3.1. Average Frame Log Likelihood (LL)

A monophonic, speaker dependent, single mixture HMM was trained for each speaker using a CELEX [7] based transcription and embedded training using HTK [15]. After five passes the model was further trained on the citation material for another five passes. This citation influenced model was then used to autosegment the spontaneous speech and output the log likelihood for each vowel averaged by duration (the number of 10ms frames).

3.2. F1/F2 Formant Preprocessing

The phonetically motivated models take normalised F1/F2 values calculated by fitting parametric curves on the output of an LPC formant tracker to establish achieved formant targets.

Vowels are traditionally described as having potentially both steady state and transition regions. Formants do not remain at a static value within a vowel but instead change value at the edge of the vowel and in the case of diphthongs within the vowel. The transitions at the edge of a vowel reflect the articulation of the surrounding phonemes.

A target model of vowel production assumes that the formant is moving towards and away from an ideal value that describes this vowel. Thus the ideal target value may not be reached depending on such factors as phonetic context, vowel duration and care of articulation. If the ideal value is not reached then the formant is said to undershoot the target [11].

The effects of care of articulation on vowel can be explained by undershoot. In the studies that examined F1/F2 values in carefully and less carefully articulated vowels it was found that the formants in the central region of the vowel tended to be less extreme in less carefully articulated speech and closer the centre of the vowel triangle [13]. This *centralisation* could be caused by the formant not reaching the extreme vowel target that it would in carefully articulated speech. This occurs because the speaker makes less effort to move the articulators to the extremes required to produce these ideal values.

In order to find these representative F1/F2 values of vowels a parametric curve is fitted to bark transformed formant values. The maxima or minima of the curve is then taken as the achieved vowel target (see Figure 1). The result of taking these values is to produce more granular data (see Figure 2).

3.3. Vowel Clarity (CL)

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

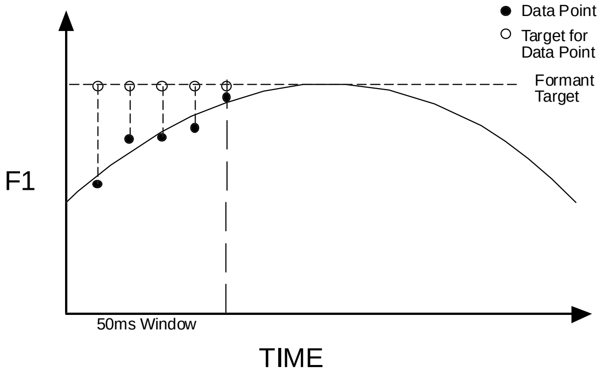

Figure 1: *Using a parametric curve to calculate the achieved spectral target of a formant.*

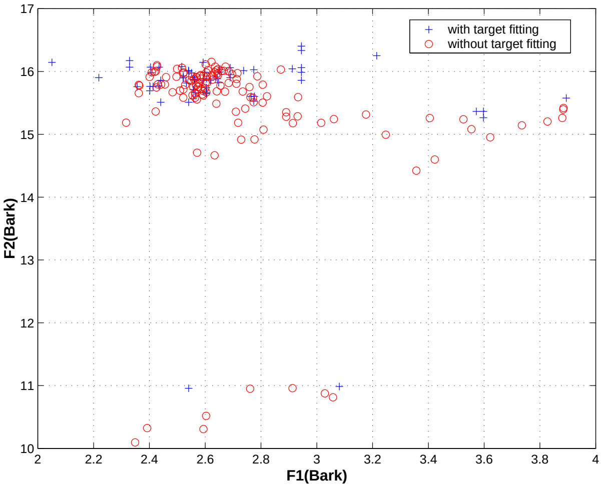

Figure 2: *Comparison of F1/F2 values for the vowel /i/ with (n=125) and without (n=128) target fitting.*

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 [16], 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 modeling and normalisation techniques see [3]. Figure 3 shows a model of a speaker's vowel space generated in this way.

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 achieved targets are to the 'hills' in the model. These probabilities are then combined as an average log likelihood.

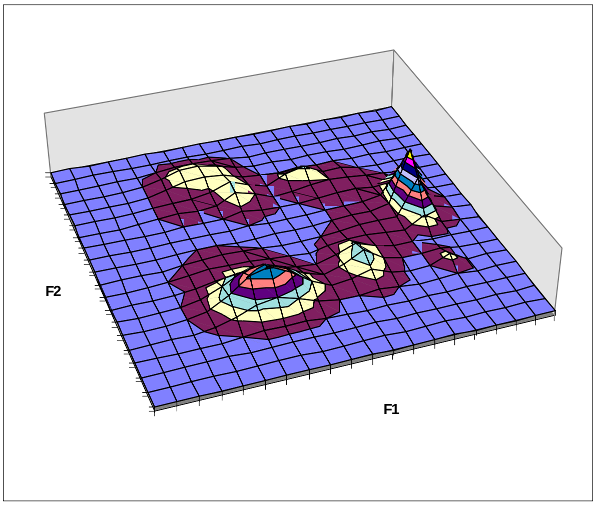

Figure 3: *A 20 Gaussian model generated using expectation maximisation. The model was built from achieved targets generated by applying parametric curve pre-processing to citation speech.*

3.4. Vowel Centralisation (TR)

A simpler model of vowel articulation is to regard unclear vowels as just more centralised (closer to the centre of the vowel space). This model calculates the average Euclidean distance of a vowel's pre-processed F1/F2 values from the central point (F1/F2 mean) of the vowel space.

4. PERCEPTUAL EXPERIMENT

4.1. 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 [12].

The vowels used, all had durations between 90-110ms, had their amplitude normalised and were excerpted from the HCRC Map Corpus [1]. 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 models 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 assigned as follows:

Model LL - mean -135.8 sd 20.5
Low: $x < -140$, Medium $-140 > x < -127$, High $x > -127$
Model CL - mean -16.912 sd 2.154
Low: $x < -16.75$, Medium $-16.75 > x < -15.5$, High $x > -15.5$
Model TR - mean 1.32 sd 0.35
Low: $x < 1.2$, Medium $1.2 > x < 1.5$, High $x > 1.5$

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.

4.2. Results

4.2.1. Model LL.

A 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. This probably has more to do with the basic difficulty of the task than some underlying similarity in vowel sensitivity as agreement between subjects was generally not higher than 10% [4].

Effect of Vowel Type - Non Significant Effect of Clarity Group - $F(2, 52) = 70.76, p < 0.001$ Interaction of Vowel/Clarity - $F(4, 105) = 84.45, p < 0.001$			
Geometric Mean By-Subjects Responses			
Clarity Group	High	Med	Low
Vowel i	0.44	0.78	1.31
Vowel a	1.85	0.36	0.28
Vowel o	0.90	1.00	0.71

As we can see, although there is a significant relationship between clarity group and response there is no linear relationship between log likelihood and subject response. The very strong clarity/vowel type interaction shows the vowel type has a strong effect on the results. For the LL model, not only does the magnitude of a clarity effect vary between vowels, but more importantly, the direction of this effect varies.

4.2.2. Model TR.

As above linguistic background had no significant effect on subjects responses. In contrast to the LL model vowel type was significant.

Effect of Vowel Type - $F(2, 52) = 3.25, p < 0.5$ Effect of Clarity Group - $F(2, 52) = 47.71, p < 0.001$ Interaction of Vowel/Clarity - $F(4, 104) = 59.4, p < 0.001$			
Geometric Mean By-Subjects Responses			
Clarity Group	High	Med	Low
Vowel i	0.99	0.81	0.66
Vowel a	0.66	1.14	0.72
Vowel o	0.65	0.90	0.81

Again the strong interaction between vowel type and clarity group meant that no clear linear relation was shown between subjects responses and clarity scoring using the TR model.

4.2.3. Model CL.

As above, linguistic background had no significant effect on subjects responses. In contrast to the both previous models vowel/clarity group interactions were not as strong.

Effect of Vowel Type - $F(2, 52) = 3.26, p < 0.5$ Effect of Clarity Group - $F(2, 52) = 37.12, p < 0.001$ Interaction of Vowel/Clarity - $F(4, 104) = 4.87, p < 0.005$			
Geometric Mean By-Subjects Responses			
Clarity Group	High	Med	Low
Vowel i	0.91	0.81	0.75
Vowel a	0.87	0.84	0.80
Vowel o	0.85	0.75	0.76

As we can see, the results from the vowel space model followed human perception more closely, grouping vowels similarly to human judgments of vowel quality. This suggests the CL model could be used as a basis for an automatic measure of the spectral effects of COA.

4.2.4. Correlations.

A linear correlation was then carried out to investigate the correspondence between each models’ clarity score for each vowel and the average geometric mean of the pooled subjects. Because of the undue influence of outliers on linear correlations, points outside 2 standard deviations of the mean clarity score for each model were removed (between 4-5 points for each model). Results were as follows:

Model LL -	no significant correlation		
Model TR -	$r = 0.31$	$r^2 = 0.10$	$p < 0.005$
Model CL -	$r = 0.34$	$r^2 = 0.11$	$p < 0.001$

Although neither the CL or TR models could be regarded as robustly predicting human responses to vowel quality they do agree with subjects approximately as well as subjects do with each other [4]. More interesting and, considering the results from the by-subjects ANOVA, expected was that log likelihood showed no linear correlation with subjects responses.

Some caution is required when interpreting these results. Results from the perceptual test do not equate directly with the motor movement (and effort) made by a speaker. However numerous studies have shown that human perception of clarity is strongly related to care of articulation e.g. [13]. Any serious automatic measurement of COA would at least show some correlation with human responses. The fact the the LL model shows no such effect, as well as the differences, not only in magnitude but direction on a by vowel basis, in the by-subjects analysis suggest that the LL model is a very poor measure of COA indeed.

5. Discussion

It is perhaps not surprising that vowel log likelihoods’ generated by an HMM do not have any linear relationship with care of articulation. After all, HMM recognition systems were not designed with this role in mind. However there is a persistent tendency to regard the log likelihood values as meaningful in terms of how canonical a segment is. Given this, if a model is trained on 'clear speech' you may expect such a model to tell you how close to a clear speech token a segment is. Perhaps the crucial aspect of this is what is meant by 'close' and what is meant by canonical in this context. Given that an HMM operates on many parameters and deals with a sequence of values over time it is not surprising that such closeness is very difficult to define phonetically.

However, by contrasting the CL and LL models, we can gain insight into what is being represented by acoustic log likelihood. If you gave a single state multiple mixture HMM the same pre-processed F1/F2 information you would have a more or less an identical model to model CL. This leads to some conclusions concerning why model CL appears to model COA to some extent, while the raw output from model LL (MFCC HMM) does not.

1. The preprocessing deals (to some extent) with co-articulation effects when co-articulation does not affect the achieved formant target. Without this pre-processing it is quite possible for a segment to be regarded as canonical simply because it has a common context, not because the central state has reached some ideal target.
2. F1/F2 are directly related to the movements of the tongue while generating a vowel. Thus a spectral model of COA based on these is probably more likely to relate actual production effort than MFCCs which do not have such a clear relationship.
3. Log likelihood from a traditional HMM is much more closely correlated to duration than the output from either of the other two models (see correlations below).

   Model LL - $r = 0.26, r^2 = 0.07, p < 0.001$
   Model TR - $r = 0.0052, r^2 = 0.00, p - NS$
   Model CL - $r = 0.001, r^2 = 0.0001, p < 0.005$

   This may be because, when normalised for duration, the output represents a long sequence of pretty predictable values. The shorter a vowel the greater the transitions tend to be and despite the fact that violent transitions towards idealised formant targets are a sign of careful articulation the log likelihood will tend to be lower. In the experiment we carried out, duration was controlled for. If it is not, then log likelihood might well relate to COA but only as a noisy duration measurement.

The CL model is far from perfect [6] but it shows that HMMS could be used to model COA if the above problems are addressed. If not then the acoustic log likelihood should not be used to model COA. In addition log likelihood values should only used to represent how 'close' a segment is to a particular HMM model with great care.

6. 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, 2000.

[4] Matthew Aylett and Alice Turk. Vowel quality in spontaneous speech: What makes a good vowel?. In *ESCA Workshop: Sound Patterns of Spontaneous Speech*, 1998.

[5] Matthew P. Aylett. *Stochastic Suprasegmentals.* PhD thesis, University of Edinburgh, 2000.

[6] Matthew P. Aylett. *Stochastic Suprasegmentals: Relationships between Redundancy, Prosodic Structure and Care of Articulation in Spontaneous Speech (http://www.cogsci.ed.ac.uk/m̃atthewa/thesis_sum.html).* PhD thesis, University of Edinburgh, 2000.

[7] 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.

[8] D.J. Broad and F. Clermont. A methodology for modelling vowel formant contours in CVC context. *The Journal of the Acoustical Society of America*, 81:1572–1582, 1987.

[9] C.A. Ferguson. Baby talk as a simplified register. In C.E. Snow and C.A. Ferguson, editors, *Talking to Children.* Cambridge University Press, Cambridge, 1977.

[10] B.F. Freed. *Foreign talk: A study of Speech Adjustments made by Native Speakers of English when in Conversation with Non-native Speakers.* PhD thesis, University of Pennsylvania, 1978.

[11] B. Lindblom. Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, 35:1773–81, 1963.

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

[13] 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.

[14] R. van Son. *Spectro-Temporal Features of Vowel Segments.* PhD thesis, University of Amsterdam, 1993.

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

[16] 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.