A Study of Vowel Coarticulation in British English

James L. Hieronymus

Centre for Speech Technology Research, Edinburgh University
80 South Bridge, Edinburgh EH1 1HN, Scotland

Abstract

Coarticulation in continuous speech causes vowel formant frequencies to be affected by nearby phonemes. Generally continuous speech causes the vowel formant targets to be centralized relative to their isolated word counterparts. The present study concentrates on 660 phonetically hand labelled sentences from one male talker of the RP accent of British English. This allows the study of coarticulation without the confounding effects of accents, speech habits and differing individual formant ranges. The 12 monophthongal vowels of RP British English /i, I, ae, ɛ, ɑ, ʌ, ɔ, o, U, u,ə, ɚ/ have been studied using formant frequency and amplitude tracks and duration, and sentential stress (sentence stress as opposed to lexical stress). Generally the vowels are most affected by nearby semi–vowels /l, r, y, w/. No simple relationship between adjacent phoneme place of articulation and the vowel target change has been found when all the vowels are treated together. However, the data shows the presence of "robust vowels" which are not greatly effected by nearby semi–vowels. These vowels are not simply stressed vowels, but depend on duration and others factors being studied. The weak effect of duration is that the pre–pausal lengthened vowels are in the "robust" category, but shorter vowels can either be robust or ordinary. The categories of function word and content word do not account for robustness.

Introduction

Most coarticulation studies have considered isolated words. An early study by Shearme and Holmes [1] showed that vowels in continuous speech very seldom had steady states and often did not overlap the Peterson–Barney [2] 95 percentile contours in any part of their frequency trajectories in time. Generally the vowels are much more centralized in continuous speech and the vowel formant regions overlap considerably due to coarticulation.

Kuwabara [3] found a renormalization technique based on the theory of Lindblom and Studdert–Kennedy [4] which disambiguates Japanese vowels in continuous speech.

Hieronymus and Majurski [5] tried this technique on American English vowels and found that it did not work well. It has been speculated that the stress structure of English causes this method to fail. The presence of "robust" vowels as found in this study would cause this technique to fail, because the renormalization is applied uniformly to all vowels.

This is a report of an ongoing study of vowel properties and coarticulation in British English. The present approach is to study the speech of one talker at a time in detail to find the underlying mechanisms in coarticulation. Thus coarticulation can be studied without the confounding effects of regional accent, speaking styles, and formant ranges due to different talkers. Then speech data from other talkers will be studied and the pooling of the data explored to achieve speaker independent results later in this study.

It is postulated that some sort of hierarchical structure of linguistic factors modifies the effect of nearby phonemes such that the same vowel in the same phonetic context will have markedly different formant frequency trajectories in time. Some possibilities for factors which have been explored are sentential syllable stress, duration, and word identity. Originally it was thought that sentential stress would be the determining factor of vowel precision of production. Previous studies by us [5], [6] for American English have shown that sentential stress is not a determining factor, based on automatic stress labelling. The present study uses hand labelled stress and shows that, on the average, sententially stressed vowels are more precisely produced than their unstressed counterparts.

Method

The data was read by one male talker of a near RP dialect of British English in a sound isolation booth. The microphone was a Shure SM–10. The speech was digitized directly using a 16 bit a–to–d converter at 20 kHz sampling frequency with an anti–aliasing filter at 8 kHz. The talker was told to speak the sentence as if he was saying it in conversation and was prompted with the sentence on a computer screen. The speech was hand labelled by graduate phoneticians at a broad phonetic level with syllable stress marked using a PC based labelling workstation. The labellers were presented with a spectrogram and could play the segments. Subsequently the sentences were parsed by hand to provide loose bracketing of phrase boundaries, so that syntactic effects could be studied. Of the 660 sentences were designed for the CSTR/ATR database project to collect and label speech for speech technology studies. The other 460 sentences were Anglicized versions of the TIMIT compact sentences designed by the MIT Speech Group.

Each vowel formant is characterized by three values for each hand labelled vowel. The values are the first and second formant frequencies at points 10 %, 50 % and 90 % of the duration of the vowel. These values were chosen to minimize the effect of formant tracking errors. Formant tracks are obtained from a centroid based formant tracker developed by Crowe [8]. Except for low formant frequency values in the nasalized vowels the formant tracker seems to have a low error rate. These values are then fed into the APS system developed at CSTR by Watson [9] providing an interface to the S package to allow statistical studies of the data.

Discussion of the Data

Figure 1 shows a scatter plot of the first and second formant values measured at the temporal center for the long British English vowels extracted from 358 sentences with ellipses representing 66 percent of the data (/ae/ is a long vowel durationally in this data even though it is phonologically lax). Figure 2 shows data for short vowels. The normal range for a male talker is 200–1000 Hz for the first formant and 800–2300 Hz for the second formant. The minimum perceptible differences (DL) in formants were measured by Flanagan [ ] and found to be +/– 50 Hz for F1 and +/– 75 Hz for F2. Thus a measure of precision of production is how large the standard deviation of the data is relative to the DL. The cross hatched area in each vowel region is the ellipse for the sententially stressed vowels.

The formant regions for most vowels are as expected except that this talker has a very fronted /u/. The vowel /ɔ / is the highest back vowel for this talker with a median second formant of approximately 800 Hz. The long schwa is more precisely produced than the reduced vowel schwa (not plotted because of its large standard deviation) with the long form having a significantly lower second formant.

While the stressed vowels are more compact in the 66 percentile ellipses, there are a a considerable number of wide ranging outliers. Secondly there is a concentration of data points towards the outer edge of these ellipses. These are the "robust" vowels as will be shown.

The short vowels have more scatter and thus seem to be produced with less precision. Once again the stressed vowels are statistically more compact than the unstressed vowels. A superposition of these plots shows a considerable overlap between vowels in the tense–lax pairs. Duration plots show that the durations of the tense vowels pairs are statiscally longer than the vowels in their lax counterpart, but that there is considerable overlap in the distributions, especially for /i/ and /I/ and /I/ and fronted /u/ for this talker.

The presence of "robust" vowels is shown in Figure 3 which shows the stylized formant trajectories for /i/ in the environment of preceding semi–vowel /w/. The smaller font characters are the preceding context and the large character represent the following context.

The question to be answered is: why do some examples of the vowel /i/ have second formant "targets" above 2100 Hz. even in this environment? The primary stress vowels in this set are shown by a round circle and the secondary stressed vowels are highlighted by a square. As we

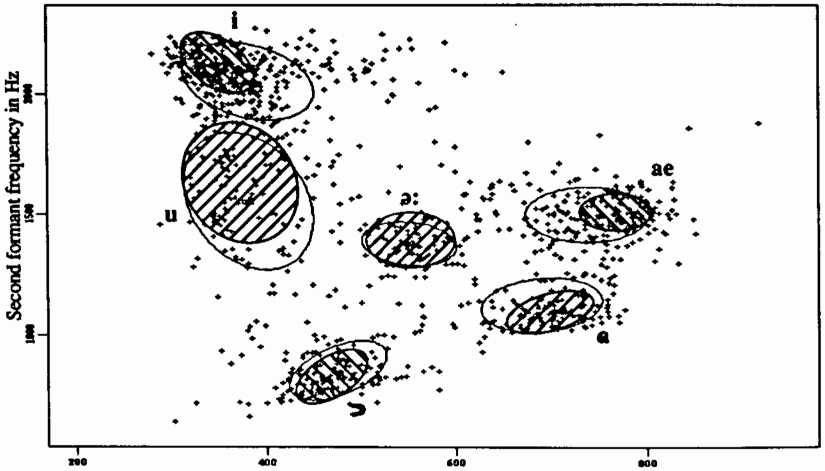

Figure 1 : First and Second Formants for British English Long Vowels

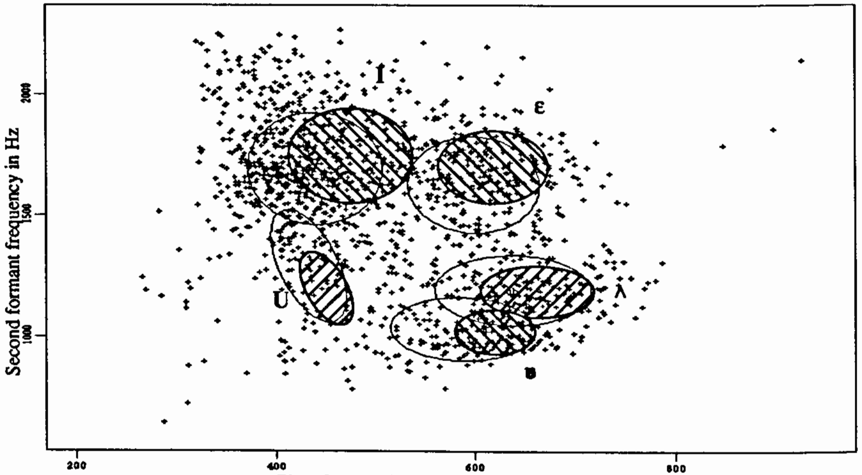

Figure 2 : First and Second Formants for British English Short Vowels

see, all of the stressed vowels and one unstressed vowel are the ones with high second formant. Consider the context /w i k/ in the figure. Two of the tracks are for the stressed vowel and appear in the robust region. One of these vowels in the same context (shown by the arrow) is not robust, and is unstressed. The /w i n/ context shows two unstressed examples, one of which is robust and one not. Thus stress does not seem to be a reliable correlate of "robustness" for vowel production. Similar plots of prevocalic /r/ show an even greater tendency for unstressed vowels to be robust but are more difficult to see at this scale.

The vowel /o/ was also examined for the presence of robust exemplars in the presence of /j/ the palatal glide, and they were found. There was a weak correlation between stress and robustness.

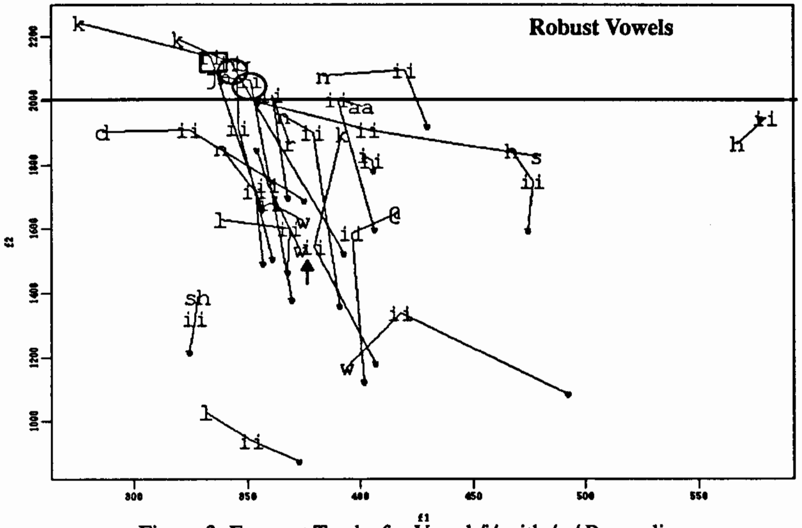

Figure 3: Formant Tracks for Vowel /i/ with /w/ Preceeding

Two possible factors, the type of word and duration, were examined to determine whether or not they determined robustness of the vowel. The longest, pre–pausal lengthened vowels are all robust. However for shorter vowels, duration is not a good correlate of robustness. Both function and content words were found to contain robust vowels. The word "between" was found to have a robust vowel on one occasion and a coarticulated vowel in another.

Acknowledgement

The programs written by Alan Crowe and Gordon Watson were used in this study. The speech labelling by Marie Alexander, Cathy Bennett, Delyth Davies, Imogen Cohen, Keith Edwards, Hazel Sydeserff, Ivan Uemlianin and Christina Barr were instrumental in this study. The collection and labelling project was supported by ATR, Japan. This work was supported by the Alvey Integrated Speech Technology Demonstrator project in collaboration with Marconi Speech and Information Systems and HUSAT, Loughborough University. The demonstrator is supported by the UK Department of Trade and Industry and the British Science and Engineering Research Council under grants D/29611, D/29628, D/29604 and F/10309.

References

1. Shearme, J.N. and Holmes, J.N. , Fourth Congress of Phonetic Sciences, Helsinki 1961 (Mouton and Co.), 233–240 (1962).
2. Peterson, G.E. and Barney, H.L., *JASA*, **24**, 175–184 (1952).
3. Kuwabara, H., *JASA*, **77**, 686 – 694 (1985).
4. Lindblom, B. E. F. and M. Studdert–Kennedy, *JASA*, **42**, 830 – 843 (1967).
5. Hieronymus, J. L. and W. J. Majurski, *Proc. ICASSP86*, Tokyo, Japan , 2787–2790 (1986)
6. Hieronymus, J. L., *Proc. ICASSP89*, Glasgow, Scotland , 608 – 611 (1989).
7. Cutler, A. and D.M. Carter, *Computer Speech and Language*, **2** (1988).
8. Crowe, A. S. and Jack, M. A., *Electonics Letters*, **23**, 1019–1020 (1988).
9. Watson, G. S., *Proc. Eurospeech* **89**, 300–303.