Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei

Running Title: **Language redundancy and syllabic spectral characteristics**

Matthew Aylett[a]

Rhetorical Systems Ltd.UK

Department of Theoretical and Applied Linguistics, University of Edinburgh, UK

Alice Turk[b]

Department of Theoretical and Applied Linguistics, University of Edinburgh, UK

[a] Electronic mail: matthewa@cogsci.ed.ac.uk
[b] Electronic mail: turk@ling.ed.ac.uk

ABSTRACT

The language redundancy of a syllable, measured by its predictability given its context and inherent frequency, has been shown to have a strong inverse relationship with syllabic duration. This relationship is predicted by the hypothesis that an inverse relationship between language redundancy and the predictability given acoustic observations, the acoustic redundancy, makes speech more robust in a noisy environment (The smooth signal redundancy hypothesis). This hypothesis also predicts a similar relationship between the spectral characteristics of speech and language redundancy. However, the investigation of such a relationship is hampered by difficulties in measuring the spectral characteristics of speech within large conversational corpora, and difficulties in forming models of acoustic redundancy based on these spectral characteristics. This paper addresses these difficulties by testing the smooth signal redundancy hypothesis with a very high quality corpora collected for speech synthesis, and presents both durational and spectral data from vowel nuclei on a vowel by vowel basis. Results confirm the duration/ language redundancy results achieved in previous work, and show a significant relationship between language redundancy factors and F1/F2 formants. The results vary considerably by vowel. In general, vowels show increased centralization with increased language redundancy.

PACS numbers: 43.70.Fq, 43.71.Es, 43.70.Bk

I. INTRODUCTION

In spontaneous speech there is wide, within-speaker variation in the articulation of the same phoneme. Much of speech research has been devoted to the explanation and description of this variation. One fairly consistent observation made by many researchers going back to the 60s (Bolinger, 1963; Lieberman, 1963; Sharp, 1960) is, that the more predictable a section of speech, either because of context or inherent frequency, the shorter, or more reduced, it tends to become. This predictability due to language structure can be termed *language redundancy*, where the more predictable a word, syllable or phoneme, the greater its language redundancy. Duration change, because this leads to greater or less acoustic information can be termed *acoustic redundancy*, where the longer a word, syllable or phoneme, the greater its acoustic redundancy. These early observations support the hypothesis that the greater the language redundancy of a word, syllable or phoneme, the less its acoustic redundancy in terms of duration. Recent work has confirmed these findings (e.g. Aylett, 2000; Bybee, 2000; Bell et al, 2003; Wright, 2003; Munson and Soloman, 2004).

However, in addition to duration reduction, spectral effects have also been noted for the first two formants measured in the central portion of the vowel. Such change can be interpreted as varying how spectrally distinctive a vowel is. The more distinctive a vowel the more acoustic information can be said to be present. Thus spectral change can also be thought of in terms of acoustic redundancy. Although spectral change has clearly been shown to occur due to speaking style (Picheny et al., 1986; Lindblom, 1990; Moon and Lindblom, 1994; Bradlow et al 1996), prosodic variation (Summers, 1987; van Bergem, 1993) and also lexical neighborhood density

(Wright, 2003; Munson and Soloman, 2004), evidence showing a language redundancy relationship with spectral change has been more elusive. Wright (2003), Munson and Soloman(2004), Bybee (2000) all showed a relationship between word frequency and spectral change, with vowels becoming more centralized in more frequent words. However, these studies used a relatively small set of laboratory data which makes it difficult to extend these results to a more general context. On the other hand, corpus studies such as Aylett (2000; 2001), reported only a weak correlation between language redundancy and vocalic spectral change, whereas Jurafsky, Bell and others (Jurafsky et al 2001, 2003; Bell et al 2003) reported a more robust result, but for vowel spectral reduction of function words only.

The extent language redundancy relates to acoustic redundancy has significant implications for theories of language perception and production. Lindblom (1990) in his H&H theory argued that reduction and prominence, in terms of a hypo-articulation/hyper-articulation continuum, allows speakers to conserve effort while producing sufficiently distinctive speech. However, both language redundancy and acoustic redundancy affect what is acoustically sufficient for the listener, acoustic redundancy because the acoustics are more distinctive, language redundancy because of the increased probability of correctly recognizing the word due to context or inherent frequency. Because of this connection between acoustic redundancy, language redundancy and speech perception, both types of redundancy can be integrated into probabilistic explanations of speech production and perception. Examples of recent work in this area include The *Probabilistic Reduction Hypothesis* (Jurafsky, 2001), work by van Son and Pols on *Speech Efficiency* (van Son and Pols, 2003) and The *Smooth Signal Redundancy Hypothesis* (Aylett, 2000; Aylett and Turk, 2004).

The Probabilistic Reduction Hypothesis states that word forms are reduced when they have higher probability. This is more general than theories which have been concerned only with word frequency (e.g. Zipf, 1949) and predictability (Fowler and Housum, 1987) because, "The probability of a word is conditioned on many aspects of its context, including neighboring words, syntactic and lexical structure, semantic expectations, and discourse factors" (Jurafsky et al 2001). Van Son and Pols present the case that the underlying reason for this language redundancy/ acoustic redundancy relationship is to make speech an efficient communication channel. They present data on a segmental level across vowels and consonants from a corpus study to support this position (e.g. van Son and Pols, 2003). The Smooth Signal Redundancy Hypothesis also states that word forms are reduced when they have higher probability and agrees with van Son and Pols that this is in order to make the speech signal more efficient. However, it makes a wider claim, firstly that this efficiency is gained in order to make speech robust in a noisy environment, and secondly that such changes in reduction is linguistically implemented through prosodic structure. In contrast, the Probabilistic Reduction Hypothesis argues that "probabilistic relations between words are represented in the mind of the speaker" (Jurafsky et al 2001).

In this paper we carry out an analysis of a large speech corpora (approximately 500,000 syllables). We show, 1) the redundancy/duration effect is strongly present for this data, 2) the spectral effect is also significant, 3) that much of the predictive power of prosodic factors and redundancy factors is shared, and 4) the direction of the effect is of redundancy leading to more reduction in terms of vowel centralization. These results strongly support the Probabilistic

Reduction Hypothesis, the speech efficiency explanation, and go some way to supporting the Smooth Signal Redundancy Hypothesis within the spectral domain.

A. The smooth signal redundancy hypothesis

The Smooth Signal Redundancy Hypothesis argues that the acoustic consequences of differences in redundancy can be explained functionally within an information theoretical framework, by the drive for speakers to achieve robust information transfer in a potentially noisy environment while conserving effort. These pressures encourage speakers to produce utterances whose elements have similar probabilities of recognition, that is, utterances with a smooth signal redundancy profile. In Aylett and Turk (2004) we presented evidence showing that phrase-medial syllables with high language redundancy (i.e. highly predictable from lexical, syntactic, semantic, and pragmatic factors) were shorter than less predictable elements. This variability in duration, due to language redundancy, was strongly related to variability associated with prosodic prominence. We argued that a strong, but imperfect relationship between language redundancy and prosodic prominence is to be expected if prominence is the means speakers use to implement a smooth signal redundancy profile.

For example, taking the utterance “I’m going to the beach”, the word “to” is more likely to occur than the word “beach”. In addition the amount of acoustic information both in terms of duration and in terms of spectral distinctiveness is lower in “to” than in “beach”. If we regard the overall probability of recognizing the lexical item as the probability given a language model multiplied by the probability given an acoustic model, we see how the overall signal probability is

smoothed by this inverse relationship. This smoothness adds robustness because the information content is spread more evenly across the signal.

The Smooth Signal Redundancy Hypothesis also argues that for articulation to adjust itself to reflect minor changes in redundancy is unrealistic. In this example a large pitch accent is placed on “beach”, whereas “to” undergoes lexical reduction. Prosodic structure appears to shadow redundancy effects. The small unique independent contribution redundancy factors made to our previous study (Aylett and Turk, 2004), led to the conclusion that prosodic structure is the linguistic structure used to implement smooth signal redundancy.

In this paper, we test, by examining spectral data, a more general version of the smooth signal redundancy hypothesis which claims any acoustic feature used to recognize the identity of a syllable could be used to smooth signal redundancy by co-varying inversely with language redundancy. The measures we used to test this version of the hypothesis are F1/F2 formant values at the temporal midpoint of vowel nuclei. We chose these measures because relationships between vowel centralization (as measured by F1/F2) and intelligibility are well documented (see section I.B), and like syllable duration, variation in F1/F2 values can affect acoustic redundancy. For example, the probability of identifying a particular vowel is higher if its formant values are unambiguously associated with a single vowel category, and not simultaneously associated with multiple vowel categories (see Figure 1).

B. The vowel space

The F1/F2 formant values for vowels can be measured by reference to the *vowel space*, a two dimensional space based on each vowel's F1/F2 formant values. The smaller the vowel space the more bunched and *centralized* the vowels, the larger the vowel space the more distinctive and peripheral the vowels. It has been shown that vowel space expansion (associated with hyper-speech) is correlated with speech intelligibility, and thus increased acoustic redundancy (Bradlow et al 1996; Bond and Moore 1994). In contrast casual or reduced speech (hypo-speech) is associated with centralization (Picheny et al., 1986; Lindblom, 1990; Moon and Lindblom, 1994; Bradlow et al 1996) which leads to a reduced vowel space and arguably lower acoustic redundancy. On this basis, a reasonable prediction would be, that as language redundancy increases, we would see average formant values to shift towards the neutral, centralized position, leading to lower acoustic redundancy and, overall, a smoother signal redundancy. Therefore, in the case of /i/ (as in 'lease'), as language redundancy increases, we would expect both F1 to rise and F2 to fall while for /ɑ/ (as in 'father') we might expect F1 to fall and F2 to rise.

This raises the question of how to compare such centralization results across vowels. Previous work has often looked at the size of the vowel space and argued that if the size increases the vowels are less reduced (e.g. Bradlow et al, 1996; Wright, 2003; Munson and Soloman, 2004). For a constrained set of laboratory stimuli this works quite well but has the disadvantage of masking the way results vary by-vowel and by-formant. Similar problems occur using a distance or "clarity" measurement. This approach involves relating all vowels to a model of the vowel space. A value is then calculated which measures how distinctive each vowel example is given

this model. The advantage of this approach is that a single comparable measurement can then be used for analysis. In previous work we applied a number of alternative vowel space models in an attempt to produce a consistent "clarity" score. In Aylett (2000) a citation vowel space model was used to measure degrees of hyper-articulation, and also a simple Euclidean distance to the centre of the vowel space in terms of the mean F1/F2 across all vowels. In Aylett (2001) we explored the use of Hidden Markov Models (HMMs), which are used extensively in speech recognition technology, as a model of vowel distinctiveness. In pre-published versions of Aylett and Turk (2004) we applied a model which was the Euclidean distance to the by-vowel mean F1/F2 calculated from a speaker's citation speech. Results for all techniques were disappointing. This was because the spectral effects can be quite small and can potentially vary across vowels. By using a vowel space model to calculate a "clarity" measure we are making perceptual and articulatory assumptions concerning the vowel space. For example, a model might assume that the vowel space was a linear or "flat" surface when transformed by a perceptual measure of frequency such as the Bark scale. Such assumptions can introduce noise into any metric and obscure the effects we are trying to measure.

This is not to discount the importance of modeling the vowel space. Much work has looked at means to improve such models such as using perceptual scales or by-vowel normalization schemes (See Rosner and Pickering, 1994, for a review). However, in this analysis the advantage of a single measure derived from a model were outweighed by the noise such a model might introduce. For this reason we carry out absolute F1/F2 measurements in Hertz. By doing so, we intentionally avoid making theoretical assumptions concerning human vowel perception and production. Instead we will show that raw F1/F2 values do vary significantly by a language

redundancy factor, that the direction of this change is one that can be interpreted as reduction, and the extent of the overlap between redundancy and prosodic factors in terms of predictive power.

II. METHODOLOGY

In order to test the smooth signal redundancy of spectral factors we used a very large corpus of citation speech (the Rhetorical Corpus detailed in section II.A). This corpus was chosen because the recording quality and the carefully articulated speech allowed reliable automatic formant tracking and segmentation.

Each syllable of the corpus was scored on the basis of an F1/F2 model of acoustic redundancy, a simple language redundancy model based on syllable n-gram probability, and a traditional prosodic model (detailed in section II.B). ANOVAs with post hoc tests together with regression analysis were applied to investigate the relationship between F1/F2 formant values, language redundancy and prosody (see section II.D for details on the statistical analysis).

A. Materials

1. The rhetorical corpus

The Rhetorical Corpus was collected by Rhetorical Systems Ltd. to create databases for speech synthesis. It contains data from eight General American speakers, 3 female and 5 male with

approximately 50,000 syllables recorded for each speaker. The speakers were all professional voice talents. The material was read in a recording studio environment and contained sentences of varying lengths taken from different genres (e.g. travel directions, financial news). If a sentence was mispronounced or disfluent the speaker was required to repeat the sentence.

The resulting corpus, compared to spontaneous speech corpora such as the Map Task (Anderson et al, 1991) or switchboard (Godfrey et al, 1992), is relatively easy to analyze spectrally. This is because of the high level of the recording environment, the rate of speech (lower than typical spontaneous speech), the quality of the speakers (chosen for their ability to produce clear natural citation speech), and the filtering out of disfluency.

2. Vowels

The results we present in section III. are for a limited set of the vowels within our corpus. We wished to concentrate on a set of vowels which would minimize noise caused by errors in formant analysis. Vowels with typically short durations such as /ə, ɪ, ʊ, ʌ/ are often too short to carry out spectral analysis automatically with much confidence. Diphthongs were also avoided due to their moving spectral targets which would require more than a single measurement. Finally the /ɒ/ vowel (as in 'lawn') was removed because of the extensive pronunciation variation of this vowel within the general American accent (Wells 1882). The resulting vowel set was: /ɑ/ (father), /æ/ (tap), /ɛ/ (less), /i/ (lease), /u/ (goose).

3. Segmentation and spectral analysis

As part of rhetorical systems' synthesis process, the corpus was segmented using a mixture of proprietal automatic segmentation and hand correction. Results are similar, although superior, to automatic segmentation using HTK (Young et al 1996) which was used as a basis for word medial segmentation in Aylett and Turk (2004). Using this segmentation, for each syllable in the corpus we calculated its log duration in milliseconds and, using the ESPS formant tracker (Talkin 1987), we extracted the F1 and F2 value in Hertz at the temporal midpoint of each vowel.

B. Redundancy and prosodic models

1. The acoustic redundancy model

In this analysis we use the F1/F2 formant values of the vowel nucleus as a measure of a syllable's spectral quality. We interpret centralization, in terms of F1/F2, as reduced acoustic redundancy. Due to the possibility of centralization behavior differing markedly between vowels, results are presented separately for each vowel (showing actual F1/F2 differences).

2. The language redundancy model

A wide variety of possible measurements can be used to represent language redundancy. Examples in previous work have included word frequency, syllabic trigram models, accessibility

(Aylett 2000), joint probability and conditional probability (Bell et al 2003), and Bayesian probability given phoneme contents (van Son and Pols 1999). All these measurements have gone some way to supporting the hypothesis that there is a relationship between language redundancy and acoustic realization. However, in this work, because of the potentially small spectral changes we expect to see and because of data sparsity caused by analyzing the vowels separately, we considered syllabic probabilities only and used factor analysis so that the direction of the effects was easy to analyze.

The language model used in this work is as follows:

1. We used syllabic likelihood over a unigram, bigram and trigram context, i.e. how likely a syllable is without context, how likely given the previous syllable, and how likely given the previous two syllables. A log transformation was applied to these likelihoods because it relates the values more closely to information content (Pierce 1961) and helped normalize the distributions. In order to calculate these likelihoods the CMU language toolkit (Clarkson and Rosenfield 1997) was used to compute n-gram statistics based on the syllabic transcriptions of 187 million words found in news resources on the internet. The transcriptions were automatically generated using the Rhetorical Systems speech synthesizer.
2. To make an ANOVA analysis possible, and to help determine the direction of a language redundancy effect, we carried out a factor analysis. Factor analysis has the effect of reducing the number of variables and, using a linear transformation, creating new factors which have no correlation with each other. The analysis produced two factors which we called *wide context* and *narrow context* measures of redundancy where:

-wide context was roughly an average of all three likelihoods.

-narrow context was the unigram likelihood with bigram and trigram likelihoods subtracted.

(see Table I).

The resulting factors had the following normalized distributions: wide factor mean 0.0, median 0.132, sd 1.00, narrow factor mean 0.0, median 0.1130, sd 1.00. These distributions explained 96.8% of the variance of unigram/bigram/trigram likelihoods. The language redundancy grouping variable used in ANOVA analysis was constructed as follows:

We grouped data points into three redundancy groups:

1. High language redundancy: both narrow and wide redundancy factors higher than median value (+, +)
2. Medium language redundancy: narrow redundancy higher than median and wide redundancy lower than median or narrow redundancy lower than median and wide redundancy higher than median. (+, -) or (-, +)
3. Low language redundancy: both narrow and wide redundancy factors lower than median value (-,-)

See Table II for an example of some syllables that fell into these three groups.

3. The prosodic model

The smooth redundancy hypothesis argues that an important function of prosodic prominence is as a linguistic means for increasing smooth signal redundancy. Results in Aylett (2000) and Aylett and Turk (2004) showed that much of the predictive power of a prosody model was shared with that of a language redundancy model.

The prosodic model we used in this work is identical to the automatic coding reported in Aylett and Turk (2004) differing only by the removal of the full/reduced vowel distinction which was used there as a level of prosodic prominence. This change was required in order to explore the F1/F2 relationships with language redundancy and prosodic prominence for full vowels only.

It is a simple model consisting of three levels of prosodic prominence and boundary. For practical reasons we did not hand label phrasal stress or phrasal boundaries, but instead predicted them on the basis of primary lexical stress on open class words (for phrasal stress), and using following pauses (for phrasal boundaries). It should be noted that we have probably over predicted the occurrence of phrasal stress and under predicted the occurrence of phrasal boundaries. The model can be summarized as follows:

-Prominence:

- -None
- -Primary lexical stress
- -High probability of having a phrasal stress (primary lexical stress + open class)

-Boundaries:

-No following prosodic boundary

-Followed by a word boundary

-High probability of a following phrase boundary (followed by a pause > 100ms)

We did not take into account changes in f0, accent types or distinguish between intermediate and full phrase boundaries. We also did not distinguish between syllables which were unstressed and those which could have secondary stress (for example /i/ in "city" and /i/ in "reply"). However, this simple model made it possible to automatically code the half million syllables present in the Rhetorical Corpus.

Further work should address the weaknesses in this system, but a simple prosodic model was a strong, solid, starting position for this analysis.

C. Summary of values present in our data for each syllable

To summarize, we chose a simple language redundancy model, a simple prosodic model and used both to code each syllable in our corpus. We used the segmentation already present in our corpus to determine syllabic duration and used the ESPS formant tracker (Talkin 1987) to extract F1/F2 formant values at the temporal midpoint of each vowel to obtain a spectral measure for each syllable nucleus.

D. Data Analyses

Our results and their statistical analysis will be carried out using the following framework for duration and F1/F2 dependent variables.

The weak form of the smooth signal redundancy hypothesis accepts that the expected language redundancy relationship with acoustic redundancy may be confounded by prosodic boundary markers. In order to address this concern, in the data presented here, boundaries are controlled by discarding any syllables, 1) before a phrase boundary, and 2) with no word boundary following them, i.e all syllables are discarded not followed by a break index of one. This approach differs from both Aylett (2000) and Aylett and Turk (2004) where only syllables from monosyllabic words were used. This change was necessary because of differences between the two corpora. The Map Task Corpus (Anderson et al, 1991) used in the earlier work had a high preponderance of monosyllabic words (around 75%), whereas the Rhetorical Corpus had a much higher incidence of polysyllabic words (only 40% were monosyllabic). Whereas the monosyllabic words in the Map Task Corpus arguably provided a fair representation of the overall behavior of data within the corpus, this is not the case within the Rhetorical Corpus. In addition, given the requirement of by-vowel analysis where data sparsity becomes a significant problem, it was decided to relax the boundary control requirements.

As well as controlling for prosodic boundaries, spectral measure results for male and female speakers were analyzed separately. An alternative approach would have been to manipulate the underlying values using a speaker normalization scheme (e.g. Nearey 1992). However, there are

problems with these techniques (see Adank 2003 for a review) which would be further compounded by artifacts caused by formant tracking as high female f0 can lead to significant systematic formant tracking errors. By keeping the spectral analysis separate by sex we have avoided any normalization errors, and made it clear which results may be compromised by formant tracking errors.

The smooth signal redundancy hypothesis predicts a strong inverse relationship between language redundancy and acoustic redundancy and a positive relationship between prosodic prominence and acoustic redundancy. In addition, it expects the predictive power of the language redundancy model and the prosodic prominence model to be substantially shared.

The significance and direction of these relationships are addressed by way of a univariate ANOVA analysis followed by post hoc t-tests with Bonferroni correction. The size of the relationships and the shared proportion of the two models predictive power were analyzed by way of multiple linear regression together with a likelihood ratio test (Neter et al 1990).

1. ANOVA and post hoc t-tests

The prosodic variables in our study were categorical and therefore ideal for an ANOVA model with either syllabic duration or F1/F2 formant value as a dependent variable. For language redundancy it is necessary to use the redundancy grouping variable described in section II.B.2.

All post hoc t-tests were carried out on cells which were significant in the ANOVA. There is some debate concerning when a Bonferroni correction should be made in these circumstances. We opted for a conservative approach to the data and used the correction for all tests carried out within the same analysis. For example, comparing three redundancy groups across five vowels resulted in a correction factor of 15. Using this correction factor, if a t-test returned a significance of 0.01 the result would be recomputed as 0.15 and be regarded as non-significant.

2. Linear regression and likelihood ratio test

The main role of the linear regressions was to show how much the prosodic and redundancy models could predict the dependent variable and to what extent the two models overlapped.

Multiple linear regressions with fixed factor ordering were carried out with syllabic duration and F1/F2 formant values as dependent variables and with language redundancy and prosodic factors as independent variables. The result of the regression revealed the size of the relationship between dependent and independent variables in terms of predicted variance.

In addition, to help interpret by-vowel results, a table is included which shows how these factors affected each vowel (see section III.A.3). In some cases a dependency exists between the vowel identity and the variance of these factors. For example, for the /ɛ/ vowel (as in 'set'), in our data 90% of all tokens were lexically stressed and 84% were open class. Such homogeneity should be taken into account when comparing the contribution of prosodic and redundancy models to predicting the dependent variables by-vowel.

We then carried out a likelihood ratio test to show to what extent the proportion of variance predicted by the independent factors was unique. This was accomplished by carrying out the linear regression 1) with all factors, 2) without language redundancy factors, and 3) without prosodic factors. The extent the r-squared value decreased when factors were removed was used to represent the unique contribution of these factors. We calculated the shared contribution of the independent factors by subtracting all unique contributions from the predicted variance of the entire model.

III. RESULTS

A. Syllabic duration

1. Do the results for syllabic duration found for spontaneous speech apply to the citation speech in this data base?

In Aylett and Turk (2004), the evidence supporting the smooth redundancy hypothesis was taken from a spontaneous spoken dialogue. It was by no means clear that the citation speech in the Rhetorical Corpus would show the same relationship. It could be argued that the natural patterns of redundancy would persist, despite the unnatural nature of the recording environment. Alternatively, we might expect significant differences between the duration results found in the Rhetorical Corpus compared to a spontaneous speech corpus such as the Map Task Corpus. To investigate this question the analysis we carried out on the Rhetorical Corpus was matched as

closely as possible to the analysis carried out using the Map Task Corpus in Aylett and Turk (2004).

Univariate ANOVAs carried out across language redundancy group, and prominence group both showed a strong significant effect, (Language redundancy group – high, medium, low, $F(2, 313941) = 67726$, $p < 0.001$, Prominence Group – no stress, lexical stress, phrasal stress, $F(2,313941) = 72118$, $p < 0.001$). Figure 2 shows the mean values of each group, all were significantly different from each other (post hoc t-test with Bonferroni correction, $p < 0.001$). As language redundancy increases syllables become shorter, and conversely, as the level of prosodic prominence increases syllables become longer.

Linear regression analysis, with boundaries controlled, found that regression models using these factors predicted a significant proportion of the Rhetorical Corpus data ($r = 0.7219$, $r^2 = 0.5211$ - 52% of the variance). The independent contribution of the redundancy and prosodic model and the shared predictive power of both models is shown in Figure 3. Close to 20% of the predictive power was shared.

These results were similar to our Map Task Corpus results (Aylett and Turk, 2004) and support the view that the smooth redundancy hypothesis holds for this citation speech corpora. On that basis, we first analyzed syllabic duration with regards to our vowel set and then looked at the spectral results in terms of F1/F2 in the vowel nuclei.

2. How does the relation between language redundancy and log syllable duration vary by vowel?

Figure 4 shows the duration results (log ms) by-vowel. The relationship between duration and language redundancy by-vowel, for all the strong vowels we analyzed, showed the same pattern, i.e. longer = less language redundancy. However, the relationship between the medium language redundancy group and duration did vary for different vowels. For example for /i/ we see a smooth decrease but for some vowels like /u/ the result is not as easy to interpret. All differences were significant at $p < 0.001$ (post hoc t-test with Bonferroni correction).

Figure 5 shows how much of the variance in syllabic duration (log ms) was predicted by full regression models, containing both prosodic prominence and language redundancy factors. The combined regression/prosody model significantly predicted ($p < 0.001$) between 18-50% of the syllabic duration variation across the vowels we analyzed. Much of the predictive power was shared between language redundancy and prosodic prominence although the extent of this shared contribution varies for different vowels.

3. Duration results: Conclusion

The language redundancy effect on duration holds for this data. However the by-vowel interpretation of the results is more complex. Some vowels, such as /i/, behave as might be predicted, while others, for example /ɛ/, seem to exhibit a very different relationship between the prosodic model and the redundancy factors. One complication we must bear in mind is that

neither the prosodic factors nor the redundancy factors have similar variances across different vowels (see Table III). For example, /i/ is a common vowel which is often used in unstressed contexts (such as the last syllable of the word 'spongy'), as well as in stressed syllables. Compare this to /ɛ/ where a lower variance in the wide redundancy and a low mean suggest this vowel is never in a very predictable syllable in any context. In addition, with boundaries controlled, nearly all the /ɛ/ tokens are stressed and in open class words. This lack of variance available to both prominence and redundancy models explain the low overall performance of the models in the /ɛ/ context.

B. Spectral results

1. Spectral results: t-tests

Given a simple centralization model of acoustic redundancy, we predicted the F1/F2 variation shown in Table IV as language redundancy increased and the vowels became more centralized.

F1/F2 results were analyzed across the five chosen strong vowels with prosodic boundaries controlled. Language redundancy was significantly related to the F1/F2 values of vowels but the significance and the direction of the effect varied by vowel, see Table V. Results are presented separately for male and female speakers.

Results which confound a simple centralization prediction are shown in bold face. Results which were predicted to remain constant need to be judged separately. Arguably, providing the results

do not change dramatically, a small change could be accommodated by the centralization model. Overall, the results were reasonably consistent with a simple centralization hypothesis except for /u/ over all speakers where F1 drops and /ɑ/ for female speakers where the F2 rises.

Figures 6 and 7 show these changes across the vowel space for male and female speakers. Groups 1 (high language redundancy) and 3 (low language redundancy) show the centralization pattern. However, for many of the vowels this occurs because a general drop in F1. Group 2, formed from a mixture of high/low and low/high items (See section II.B.2) is more volatile, suggesting its members are less homogeneous.

2. Spectral results: linear regression

The t-tests showed that redundancy groups significantly affect the F1/F2 values of the mid point of the six chosen vowels. The extent the pattern of change reflects a centralization model of acoustic redundancy varies by vowel with /i/ fitting well, and /ɑ/, /æ/, /ɛ/, /u/ fitting to a certain extent. However, the t-tests do not show the extent the redundancy factors can predict changes in F1/F2, nor how these predictions relate to a prosodic prominence model.

By carrying out linear regression on the spectral results we were able to show:

1. The amount of variance predicted by both prosodic and redundancy models.
2. How much predictive power is shared across between prosodic and language redundancy models by vowel.

The regression analysis showed that the redundancy and prosodic models could predict between 0.5%-7.9% of variation of F1 ($p < 0.001$), and 0.2%-5.9% of the variation of F2 ($p < 0.001$) see Table VI.

The differences between male and female speakers might have been caused by poor performance of the formant tracker for female speakers. Given this concern we only compared the relationships between prosodic and redundancy models for the male speakers (Figure 8).

Figure 8 shows how much the relationship between the prosodic and redundancy models varied by vowel. Some vowels (e.g. /u/**-f1**) have a strong redundancy component which is orthogonal to the prosodic model, others (e.g. /i/) have a very low unique contribution from the redundancy model. In all cases (except condition /ɛ/**-f2** and /u/**-f2**) the redundancy model regression was significant ($p < 0.001$). Unlike syllabic duration, for some vowels the shared contribution was quite small or non-existent.

IV. CONCLUSIONS

The aim of this paper was to test whether, the Probabilistic Reduction Hypothesis, the idea of Speech Efficiency, and the Smooth Signal Redundancy Hypothesis could be generalized to explain spectral results. We have shown that significant effects were present for F1/F2 formant values. The more language redundant and less prominent the syllable, the more centralized and less acoustically redundant the vocalic nucleus is.

We have also shown that the effect of redundancy and prosodic prominence considerably overlap. This supports the concept that prosodic structure could be a functional method of achieving smooth signal redundancy and thus making the speech signal more robust in a noisy environment. Overall, we believe this finding substantially supports the Smooth Signal Redundancy Hypothesis.

However, it is interesting to note that our results vary considerably by vowel. Some of this variation could be ascribed to inherent differences in the range of prosodic and language redundancy variation for these vowels in the lexicon. This was touched upon in section III.A.3 where differences between /ɛ/ and /i/ were highlighted. However, limitations of our prosodic, redundancy and acoustic models could also contribute to this variation between vowels.

There are many possible improvements that could be made to the prosodic model. For example, our prosodic model does not include secondary lexical stress, degrees of variation in phrasal stress, or accent type differentiation based on f0 contours. Refining the model in this way could lead to a greater overlap between the prosodic and the redundancy models. In addition, interaction between prominence and prosodic boundaries could also be examined as a source of predictive power.

With regards to the redundancy model we could add many additional measurements such as: word frequency, likelihood of the phone sequence, a neighborhood density metric, the probability of the syntactic category given a probabilistic syntax, and, at a discourse level,

factors such as whether a referent has been mentioned previously. Using factor analysis as described in section II.B.2 this wide variety of factors could be distilled to a smaller orthogonal set and added to the redundancy model.

It may, however, be unwise to complicate either the redundancy or the prosodic model until limitations in our acoustic modeling are addressed. Dynamic cues are not taken into account in our model and these alone could account for significant by-vowel differences. It is not, after all, just the vowel which conditions the acoustic redundancy of a syllable, consonants are equally important. In fact co-articulation with consonants, which may cause centralization in the vowel, may also add information concerning the surrounding segments. For example, nasalization in a vowel means it is easier to predict the subsequent segment is nasal thus increasing the acoustic redundancy. Such co-articulation, even while centralizing the middle portion of the vowel, can also add to acoustic redundancy by adding dynamic perceptual cues to the vowels' identity. Such cues at the onset and offset of a vowel have been shown to have a fundamental effect on speech perception (e.g. Strange 1989).

In addition, the characterization of the vowel space we have used could also be improved. F1/F2 values could be modified by different types of vowel space normalization. There is a strong argument for the use of a perceptual scaling metric, such as Bark or JNDs, which we avoided in this paper to aid transparency. Finally, a more complex model than centralization may do much better at modeling this change in acoustic redundancy, for example, one taking into account vowel distinctiveness at different points in the vowel space.

This issue raises the question of what underlying perceptual model is most appropriate for building a model of acoustic redundancy. Unlike duration alone, where longer segments are generally agreed to be more intelligible, it is far from clear how factors such as F1/F2, amplitude, f0 change and spectral tilt may combine both with dynamic and static cues to increase or decrease intelligibility.

Investigating these factors and their relationship to perception is an ongoing task but we believe that a language redundancy framework can offer a solid quantitative approach for their examination.

REFERENCES

Adank, P. (**2003**) *Vowel Normalization: A perceptual study of Dutch Vowels*. Ph.D. thesis, University of Nijmegan.

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, pp. 351-366.

Aylett, M.P. (**2000**) *Stochastic Suprasegmentals: Relationships between Redundancy, Prosodic Structure and Care of Articulation in Spontaneous Speech (http://www.cogsci.ed.ac.uk/matthewa/thesis_sum.html)*. Ph.D. thesis, University of Edinburgh.

Aylett, M. (**2001**) "Modeling care of articulation with hmms is dangerous," in *Proceedings Eurospeech 01*. pp 1491-94.

Aylett, M. and Turk, A. (**2004**) "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech.," *Language and Speech*, **47**, pp. 31-56.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (**2003**) "Effects of disfluencies, predictability, and utterance position on word form variation in english conversation.," *The Journal of the Acoustical Society of America*, **113**, pp. 1001-1024.

van Bergem, D. R. (**1993**) "Acoustic vowel reduction as a function of sentence accent, word stress, and word class," *Speech Communication*, **12**, pp. 1-23.

Bolinger, D. (**1963**) "Length, vowel, juncture," *Linguistics*, **1**, pp. 5-29.

Bond, Z. and Moore, T. (**1994**) "A note on the acoustic-phonetic characteristics of inadvertently clear speech," *Speech Communication*, **14**, pp. 325-37.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**) "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Communication*, **20**, pp. 255-72.

Bybee, J. (**2000**) "Lexicalization of sound change and alternating environments," in *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, edited by M. Broe and J. Pierrehumbert. (Cambridge University Press, Cambridge)

Clarkson, P. and Rosenfeld, R. (**1997**) "Statistical language modeling using the CMU-Cambridge toolkit," in *Proceedings of Eurospeech 97*, pp. 2707-10.

Fowler, C. A. and Housum, J. (**1987**) "Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction," *Journal of Memory and Language*, **26**, pp. 489-504.

Godfrey, J., Holliman, E., and McDaniel, J. (**1992**) "SWITCHBOARD: Telephone speech corpus for research and development," In Proceedings of *ICASSP-92 San Francisco*, pp. 517-520.

Jurafsky, D., Bell, A., and Girand, C. (**2003**) "The role of the lemma in form variation," in *Laboratory Phonology VII*, edited by N. Warner and C. Gussenhoven. (Mouton de Gruyter, Berlin)

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. (**2001**) "Probabilistic relations between words: Evidence from reduction in lexical production," in *Frequency and the*

Emergence of Linguistic Structure, edited by J. Bybee and P. Hopper. (John Benjamins, Amsterdam)

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

Lindblom, B. (**1990**) "Explaining phonetic variation: a sketch of the H & H theory," in *Speech Production and Speech Modeling*, edited by W.J. Hardcastle and A. Marchal, pp. 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**, pp. 40-55.

Munson, B. and Soloman, N. (**2004**) "The influence of phonological neighborhood density on vowel articulation," *Journal of Speech, Language and Hearing Research*, **47**, pp. 1048-58.

Nearey, T. (**1992**) "Applications of generalized linear modeling to vowel data," in *Proceedings ICSLP 92*. pp 583-587

Neter, J., Wasserman, W., and Kutner, M. H. (**1990**) *Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Design, 3rd edition.* (Irwin, Boston)

Picheny, M., Durlach, N., and Braida, L. (**1986**) "Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech," *Journal of Speech and Hearing Research*, **29**, pp. 434-445.

Pierce, J.R. (**1961**) *Symbols, Signals and Noise: The Nature and Process of Communication.* (Harper, New York)

Rosner, B. and Pickering, J. (**1994**) *Vowel Perception and Production.* (Oxford Science Publications, Oxford).

Son van, R. J. J. H. and Pols, L. C. W. (**1999**) "An acoustic profile of consonant reduction," *Speech Communication*, **28**, pp. 125-140.

Son van, R. J. J. H. and Pols, L. C. W. (**2003**) "How efficient is speech?," *Institute of Phonetic Sciences, University of Amsterdam, Proceedings*, **25**, pp. 171-184.

Sharp, A. E. (**1960**) "The analysis of stress and juncture in English," *Transactions of the Philological Society*, pp. 104--135.

Strange, W. (**1989**) "Dynamic specification of coarticulated vowels spoken in sentence context," *The Journal of the Acoustical Society of America*, **85**, pp. 2135-53.

Summers, W. V. (**1987**) Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *The Journal of the Acoustical Society of America*, **82**, 847{863.

Talkin, D. (**1987**) "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *The Journal of the Acoustical Society of America*, **82, Suppl.1,S55.**

Wells, J.C. (**1982**) *Accents of English*, vol.3. (Cambridge University Press, Cambridge)

Wright, R. (**2003**) "Factors of lexical competition in vowel articulation," in *Papers in Laboratory Phonology VI: Phonetic Interpretation*, edited by J. Local, R. Ogden, and R. Temple,. (Cambridge University Press, Cambridge)

Young, S., J. Jansen, J. Odell, D. Ollason, and P. Woodland (**1996**) *The HTK Book*. Entropic. Version 2.00.

Zipf, G. K. (**1949**) *Human Behavior and the Principle of Least Effort*. (Addison-Wesley, Reading, Mass.)

TABLES

Table I: Contribution from the original n-gram log likelihoods to the wide and narrow factors. We normalized the unigram, bigram and trigram values using the mean and sd values and multiplied them by the factor analysis components. This transformation produced the values for the wide and narrow factors.

	Factor Analysis Component		Mean	Standard Deviation
	Wide	Narrow		
UNIGRAM	0.315	1.090	-6.924	1.956
BIGRAM	0.421	-0.314	-5.480	2.931
TRIGRAM	0.407	-0.520	-5.063	3.411

Table II: The syllables in the group are shown in capitals together with the word context they are from. The actual language redundancy value of each syllable will change depending on the context of the previous two syllables.

Language Redundancy		
High	***Medium***	***Low***
to town **OF**	*pause* from **O**maha	flag is **NAILED**
alternative **IN**	tanza**NI**a	unper**TURBED**
fathers such **A**	medi**NA**	radioactive gas **GUSHED**
andy **TO**	*pause* in **PO**tentially	alba**TROSS**
kind of **THE**	flights from **MA**laysia	lab lo**CUSTS**

Table III. Variation in prosodic prominence and language redundancy by vowel (Prosodic boundaries controlled).

	Prominence		**Wide redundancy**		**Narrow redundancy**	
	lexical stress	high prob phrasal stress	mean	sd	mean	sd
ɑ	91%	63%	-0.58	0.98	-0.19	1.01
æ	95%	54%	-0.85	0.78	-0.24	0.87
ɛ	92%	76%	-0.53	0.89	-0.33	0.92
i	52%	37%	0.4	0.94	0.02	0.89
u	89%	54%	-0.03	0.83	0.46	0.98

Table IV: F1/F2 predictions for each vowel as acoustic redundancy drops due to centralization.

F1/F2 Prediction with centralization model of acoustic redundancy		
Vowel	***F1 change***	***F2 change***
ɑ	minus	plus
æ	minus	same
ɛ	minus	same
i	plus	minus
u	plus	plus

Table V. Changes in F1/F2 mean values from low redundancy to high redundancy syllables.

F1/F2 change from low to high redundancy prediction/result ***low redundancy group F1/F2 - high redundancy group F1/F2 (Hz)***		
	Male Speakers (n=3)	
Vowel	**F1 change**	**F2 change**
ɑ	minus/ -42.0	plus/ +22.2
æ	minus/ -21.0	same/ +15.1
ɛ	minus/ -72.2	same/ -36.3
i	plus/ 8.4	minus/ -48.8
u	**plus/ -18.43**	plus/ 31.72
	Female Speakers (n=5)	
Vowel	**F1 change**	**F2 change**
ɑ	minus/ -188.6	**plus/ -13.2**
æ	minus/ -44.1	same/ 37.65
ɛ	minus/ -119.06	same/ 12.93
i	plus/ 8.76	minus/ -37.0
u	**plus/ -5.0**	plus/ 82.2

Table VI. Linear regression results for the combined redundancy and prosodic prominence models for F1 and F2, split by vowel, and by the sex of the speaker.

Male F1/F2				
	F1		F2	
	r	r^2	r	r^2
ɑ	0.261	0.068	0.113	0.013
æ	0.215	0.046	0.190	0.036
ɛ	0.281	0.079	0.081	0.007
i	0.214	0.046	0.243	0.059
u	0.228	0.052	0.205	0.042
Female F1/F2				
	F1		F2	
	r	r^2	r	r^2
ɑ	0.228	0.052	0.071	0.005
æ	0.071	0.005	0.189	0.036
ɛ	0.266	0.071	0.040	0.002
i	0.115	0.013	0.223	0.050
u	0.075	0.006	0.096	0.009

FIGURE CAPTIONS

Figure 1. Illustration of how F1/F2 formant values could affect acoustic redundancy.

Figure 2. The means of syllabic duration (log ms), by redundancy group and by prosodic prominence. (Prosodic boundaries are controlled for all means see section II.D). All means are significant. As a reference: 2.00 log ms = 100ms, 2.20 log ms = 158.5ms, 2.40 log ms = 251.2 ms, 2.60 log ms = 398.1 ms.

Figure 3. The unique contribution to the linear correlation from the prosodic prominence and redundancy models.

Figure 4. The means of syllabic duration (log ms), by redundancy group and by vowel. All means are significant. As a reference: 2.00 log ms = 100ms, 2.20 log ms = 158.5ms, 2.40 log ms = 251.2 ms, 2.60 log ms = 398.1 ms.

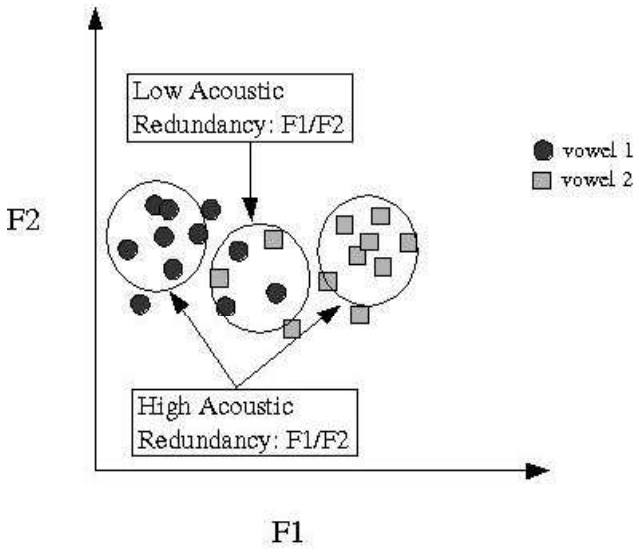
Figure 5. The unique contribution to the linear correlation from the prosodic prominence and redundancy models shown for each vowel.

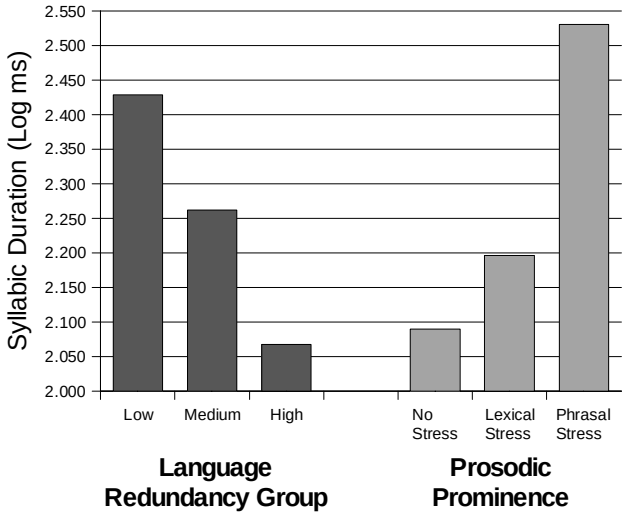
Figure 6. Male spectral results by vowel (n=3).1 = High Language Redundancy, 2 = Medium Language Redundancy, 3 = Low Language Redundancy. All differences between 1&2, 2&3, 1&3 are significant ($p < 0.005$) for both F1 and F2 (unless noted) on the basis of a post hoc t test

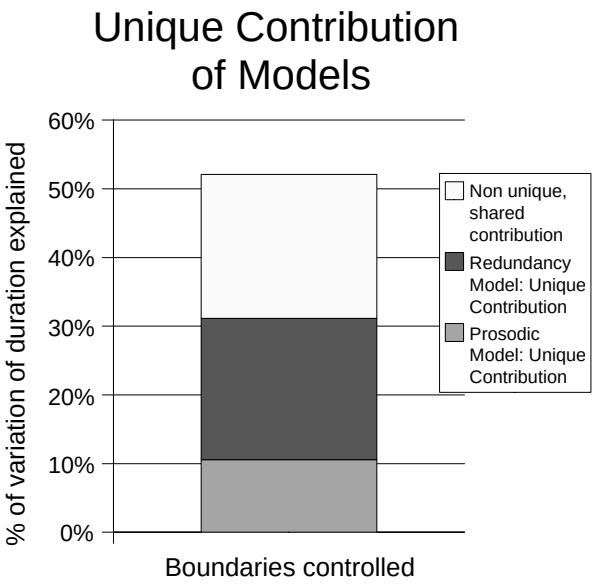
with Bonferroni correction. Non significant group differences: /ɑ/ F1 1&2, /ɑ/ F2 1&2&3, /ɛ/ F1 2&3, /ɛ/ F2 1&2.

Figure 7. Female spectral results by vowel (n=5).1 = High Language Redundancy, 2 = Medium Language Redundancy, 3 = Low Language Redundancy. All differences between 1&ˌˇˌˇ are significantˇp ⁿ˜ for both ɣ and ɣˇunless notedð on the basis of a post hoc t test with βonferroni correction. ŋon significant group differences[1] /æ/ F1 1&2, /æ/ F2 1&2, /i/ F1 2&3, /u/ F2 1&2.

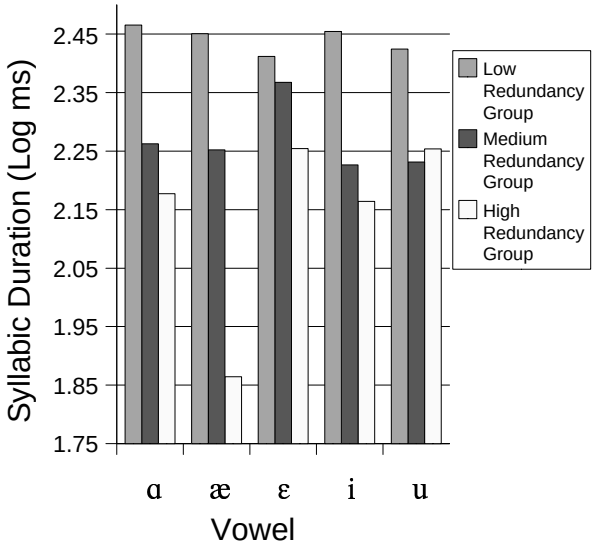
Figure 8: Relationship between the predictive power of redundancy and prosodic prominence linear regression models, male speakers only.

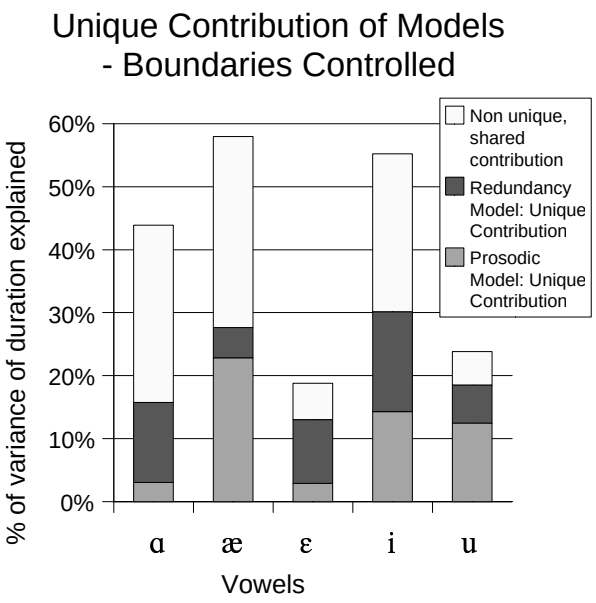

Low Acoustic
Redundancy: F1/F2
vowel 1
vowel 2
F2
High Acoustic
Redundancy: F1/F2
F1

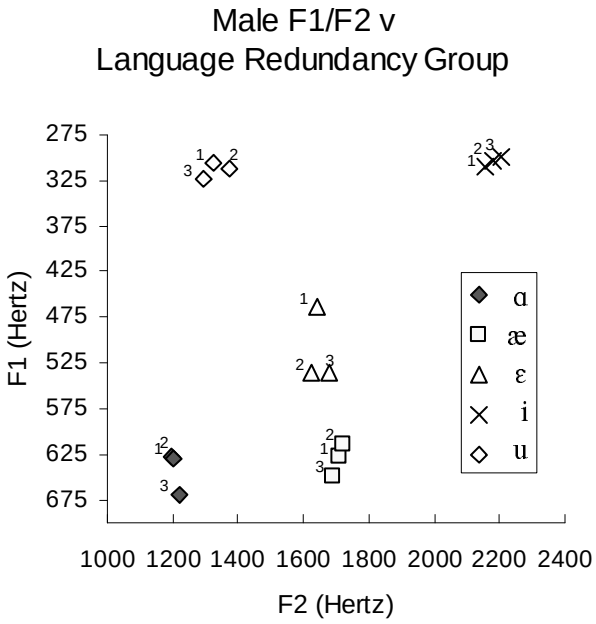

Syllabic Duration (Log ms)
2.550
2.500
2.450
2.400
2.350
2.300
2.250
2.200
2.150
2.100
2.050
2.000
Low
Medium
High
No Stress
Lexical Stress
Phrasal Stress
Language Redundancy Group
Prosodic Prominence

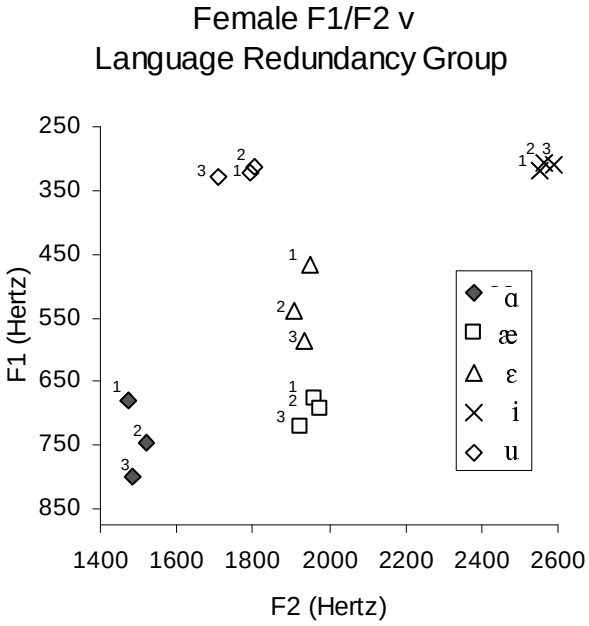

Unique Contribution of Models
60%
50%
40%
30%
20%
10%
0%
% of variation of duration explained
Boundaries controlled
Non unique, shared contribution
Redundancy Model: Unique Contribution
Prosodic Model: Unique Contribution

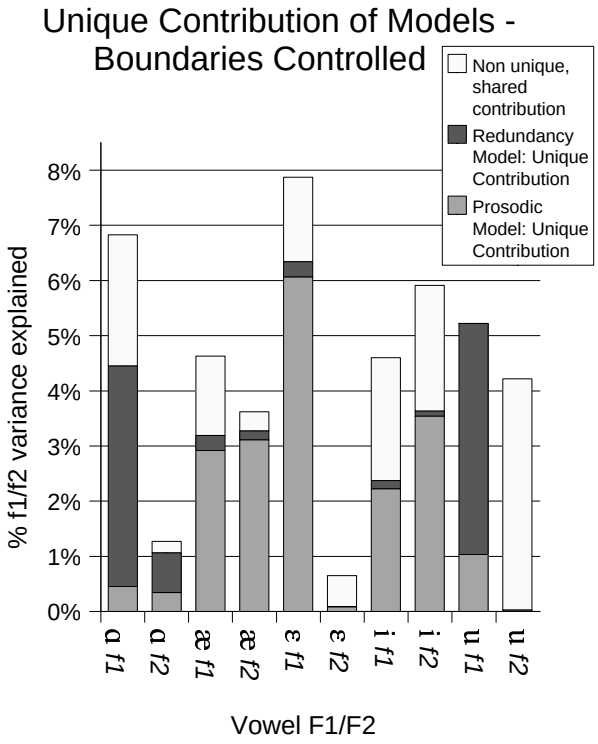
Duration v Redundancy (by vowel)




Unique Contribution of Models
- Boundaries Controlled
% of variance of duration explained
60%
50%
40%
30%
20%
10%
0%
ɑ
æ
ɛ
i
u
Vowels
Non unique, shared contribution
Redundancy Model: Unique Contribution
Prosodic Model: Unique Contribution


Male F1/F2 v
Language Redundancy Group
F1 (Hertz)
275
325
375
425
475
525
575
625
675
1000 1200 1400 1600 1800 2000 2200 2400
F2 (Hertz)
ɑ
æ
ɛ
i
u


Female F1/F2 v
Language Redundancy Group
F1 (Hertz)
250
350
450
550
650
750
850
1400
1600
1800
2000
2200
2400
2600
F2 (Hertz)
ɑ
æ
ɛ
i
u


Unique Contribution of Models - Boundaries Controlled
Non unique, shared contribution
Redundancy Model: Unique Contribution
Prosodic Model: Unique Contribution
% f1/f2 variance explained
8%
7%
6%
5%
4%
3%
2%
1%
0%
ɑ f1
ɑ f2
æ f1
æ f2
ɛ f1
ɛ f2
i f1
i f2
u f1
u f2
Vowel F1/F2