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

Matthew Aylett\textsuperscript{a)}
Rhetorical Systems Ltd. UK
Department of Linguistics and English Language, University of Edinburgh, United Kingdom

Alice Turk\textsuperscript{b)}
Department of Linguistics and English Language, University of Edinburgh, United Kingdom

(Received 24 November 2004; revised 19 February 2006; accepted 21 February 2006)

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 smooth signal redundancy hypothesis, which proposes that robust communication in a noisy environment can be achieved with an inverse relationship between language redundancy and the predictability given acoustic observations (acoustic redundancy). A general version of the hypothesis predicts similar relationships between the spectral characteristics of speech and language redundancy. However, investigating this claim 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 corpus 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 the first two formants, although these results vary considerably by vowel. In general, however, vowels show increased centralization with increased language redundancy. © 2006 Acoustical Society of America. [DOI: 10.1121/1.2188331]

PACS number(s): 43.70.–h, 43.71.–k, 43.71.Es, 43.70.Bk [RLD] Pages: 3048–3058

I. INTRODUCTION

Much of speech research has been devoted to the explanation and description of within-speaker variation in the production of the same phoneme in different contexts. One fairly consistent observation made by researchers going back to the 1960’s (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 be.

This predictability due to language structure can be termed language redundancy. Its correlation with duration is predicted by the hypothesis that the greater the language redundancy of a word, syllable or phoneme, the lower its acoustic redundancy is likely to be, where acoustic redundancy refers to its likelihood of recognition based on acoustic properties alone. On this view, duration reflects acoustic redundancy since longer segments are arguably more salient and recognizable, all other things being equal. Recent work has confirmed these earlier findings for duration (e.g., Aylett, 2000; Bybee, 2000; Bell \textit{et al.}, 2003; Wright, 2003; Munson and Soloman, 2004).

Spectral characteristics (e.g., formant frequencies in the central portion of a vowel, among other things) can also be considered as measures of acoustic redundancy, since these are known to relate to vowel distinctiveness and thus probability of recognition. However, although spectral differences have clearly been shown to occur due to differences in speaking style (Picheny \textit{et al.}, 1986; Lindblom, 1990; Moon and Lindblom, 1994; Bradlow \textit{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 measurements has been more elusive. Wright (2003), Munson and Soloman (2004), and Bybee (2000) all showed a relationship between word frequency and spectral measures, with vowel centralization more prevalent in more frequent words. Because these studies used a relatively small set of laboratory data, it is difficult to know whether these findings extend to a more general context. Relevant corpus studies include Aylett (2000, 2001), which show only a weak correlation between language redundancy and vocalic spectral change, and Jurafsky, Bell, and others (Jurafsky \textit{et al.}, 2001, 2003; Bell \textit{et al.}, 2003), which show a more robust result, but for vowel spectral reduction of function words only.

The relationship between language redundancy and acoustic redundancy has important implications for theories of speech production and perception because it offers a framework for predicting variation in articulation and, more crucially, a framework for explaining this variation in terms of speech perception. The idea of relating production to perception is by no means new. Lindblom (1990) in his H&H...
theory explicitly presents the idea that articulation varies in order to reduce effort while at the same time producing a signal which shows sufficient acoustic distinctiveness for the listener to correctly identify the linguistic content of the message. Language redundancy offers a means of quantitatively modeling differences in acoustic distinctiveness because the more predictable a section of speech, given its context, the less acoustic distinctiveness (or acoustic redundancy) is required for the listener to successfully identify the contents. Thus, the relationship between language redundancy and acoustic redundancy should implicitly mirror Lindblom’s findings. In addition, by offering a quantitative framework for measuring what is sufficiently acoustically distinct, it allows the investigation of these phenomena across large corpora and thus many more contexts, speakers, and speech styles than is possible in an experimental setting. Indeed, studies such as Jurafsky (2001), van Son and Pols (2003), Aylett (2000), and Aylett and Turk (2004) have all tested the relationship between language and acoustic redundancy for natural speech corpora. These authors share Lindblom’s view that the more predictable a section of speech the more articulatory reduction/centralization we expect to see (Moon and Lindblom, 1994; implicit in Lindblom, 1990). This view is referred to as the probabilistic reduction hypothesis (Jurafsky, 2001), speech efficiency (van Son and Pols, 2003), and is included as a fundamental element of Aylett and Turk’s (2004) smooth signal redundancy hypothesis.

The probabilistic reduction hypothesis (Jurafsky et al., 2001) states that word forms are reduced when they have higher probability, and is thus more general than theories 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, p. 229). van Son and Pols (1999) support the relationship between language redundancy and acoustic characteristics with phonemic data, and propose that the underlying reason for this language/ acoustic redundancy relationship is to make speech an efficient communication channel. Aylett and Turk’s (2004) smooth signal redundancy hypothesis also proposes that word forms are reduced when they are associated with a higher probability of recognition, and agrees with van Son and Pols’ claim that this relationship between language redundancy and acoustic characteristics is governed by the drive to make the speech signal more efficient. However, it makes a wider claim, first that this efficiency is gained in order to make speech robust in a noisy environment, and second that acoustic differences are linguistically implemented through prosodic prominence structure. In contrast to this view, the probabilistic reduction hypothesis argues that “probabilistic relations between words are represented in the mind of the speaker” (Jurafsky et al., 2001), and thus relate directly to their acoustic correlates without the intermediary of prosodic prominence structure.

In this paper an analysis of a large speech corpus (approximately 500,000 syllables) was carried out with the following goals: (1) to test for the expected relationship between language redundancy and syllabic duration; (2) to determine the relationship between language redundancy and measures of the first and second formants; and (3) to determine the degree to which prosodic prominence and language redundancy predict the same variation in each of these measures, as predicted by the smooth signal redundancy hypothesis.

A. The smooth signal redundancy hypothesis

The smooth signal redundancy hypothesis claims 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., they are highly predictable from lexical, syntactic, semantic, and pragmatic factors) were shorter than less predictable elements.

To take a concrete example, in the utterance “I’m going to the beach,” the word “to” is more likely to occur than the word “beach” since “to” is a relatively high-frequency closed class word, and often occurs immediately after the word “going,” whereas “beach” is a lower frequency word that is not predictable from its preceding context. Consequently, the amount of acoustic information both in terms of duration and in terms of spectral distinctiveness is lower in “to” than in “beach,” which contains a full vowel, and is likely to bear phrasal stress. If we regard the overall probability of recognizing the lexical item as the probability of recognition given a language model multiplied by the probability of recognition given an acoustic model, we can see that the overall signal probability is smoothed by this inverse relationship. This smoothness adds robustness to communication because information content is spread evenly across the signal.

Because prosodic prominence appears to largely shadow redundancy effects (e.g., the highly language redundant word to is unstressed in the above example while the less predictable beach bears a nuclear pitch accent (phrasal stress), and because it is arguably unrealistic for articulation to adjust itself to reflect minor changes in redundancy, the smooth signal redundancy hypothesis proposes that language redundancy is implemented through prosodic prominence structure, which itself has its own language-specific conventions and rules. Aylett and Turk’s (2004) present supporting evidence of a large shared influence of language redundancy and prosodic prominence structure on duration, with a small, expected, unique influence of prosodic prominence structure. In addition, a small unexpected unique influence of language redundancy factors was observed.

In this paper, we use spectral data to test a more general version of the smooth signal redundancy hypothesis that claims that any acoustic feature used to recognize the identity of a syllable could be used to smooth signal redundancy by covarying inversely with language redundancy. The measures used to test this version of the hypothesis were F1/F2
formant values at the temporal midpoint of vowel nuclei. These measures were chosen because relationships between vowel centralization (as measured by F1/F2) and intelligibility are well documented (see Sec. I B), and like syllable duration, variation in F1/F2 values relates straightforwardly to 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 Fig. 1).

B. The vowel space

The F1/F2 formant values for vowels can be measured by reference to the acoustic vowel space, a two-dimensional space defined by vocalic 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 expanded vowel spaces (associated with hyperspeech) are correlated with speech intelligibility, and thus increased acoustic redundancy (Bradlow et al., 1996; Bond and Moore, 1994). In contrast, casual or reduced speech (hypospeech) is characterized by centralization (Picheny et al., 1986; Lindblom, 1990; Moon and Lindblom, 1994; Bradlow et al., 1996), a reduced vowel space, and arguably lower acoustic redundancy. A reasonable prediction would therefore be, that as language redundancy increases, average formant values would shift towards the neutral, centralized position, leading to lower acoustic redundancy and, smoother signal redundancy overall. For example, in the case of /ɪ/ (as in “lease”), as language redundancy increases, F1 would be expected to rise and F2 to fall while for /ʌ/ (as in “father”) F1 would be expected to fall and F2 to rise.

Different specific predictions for formants of vowels with different height and backness specifications raise the issue of how to compare centralization results across vowels. Previous work has often measured 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). This method works quite well for a constrained set of laboratory stimuli, but has the disadvantage of masking the way results vary by vowel and by formant. Similar problems occur using distance or “clarity” measurements, whereby all vowels are related to a model of a citation “clear” version of the vowel space, using a value that reflects how distinctive or clear each vowel token is given this model. The advantage of this approach is that a single comparable measurement can then be used for analysis. In previous work, a number of alternative vowel space models were applied to a single dataset in an attempt to produce a clarity score that would show a systematic relationship with language redundancy. These included a citation vowel space model to measure degrees of hyperarticulation (Aylett, 2000), hidden Markov models (HMMs) (Aylett, 2001), as well as a simple Euclidean distance to the center of the vowel space in terms of mean F1/F2 (1) across all vowels Aylett (2000) and (2) from a speaker’s citation speech (prepublished versions of Aylett and Turk, 2004). Results for all techniques were disappointing, arguably because any modeling technique introduces articulatory and perceptual assumptions about degree of distinctiveness or clarity that obscure spectral effects that are subtle by nature and potentially vary across vowels.

Having said this, the importance of modeling the vowel space should not be discounted, for much progress has been made in developing perceptual scales and by-vowel normalization schemes (see Rosner and Pickering, 1994, for a review). However, for analyses relating language redundancy to spectral measures, the advantage of a single model-derived measure applied to all vowels is heavily outweighed by the potential noise such a model might introduce. For this reason, absolute F1/F2 measurements in hertz were used in the current study. By doing so, theoretical assumptions concerning the relationship between human vowel perception and production were intentionally avoided. By avoiding the pitfalls of these previous studies, the current study was able to show (1) that raw F1/F2 values do vary significantly with language redundancy factors; (2) that the direction of this change is one that can be interpreted as reduction, and (3) the extent of the overlap between redundancy and prosodic factors in terms of predictive power.

II. METHODOLOGY

A very large corpus of citation speech (the Rhetorical Corpus detailed in Sec. II A) was used to test the smooth signal redundancy hypothesis. This corpus was chosen because its recording quality and carefully articulated speech allowed reliable automatic formant tracking and segmentation. Each syllable of the corpus was scored on the basis of (1) a simple language redundancy model based on syllable n-gram probability; and (2) a traditional prosodic model (detailed in Sec. II B), and was represented in terms of its acoustic redundancy, i.e., in terms of (1) its F1/F2 values at the temporal midpoint of its vocalic nucleus; and (2) in terms of its duration. ANOVAs with posthoc tests, as well as regression analyses, were applied to investigate the relationship between language redundancy and prosodic prominence and these two types of acoustic redundancy measures. (See Sec. II D for details of the statistical analyses.)
A. Materials

1. The Rhetorical Corpus

The Rhetorical Corpus was collected by Rhetorical Systems Ltd. as a database for speech synthesis. It contains data from eight general American speakers, three female and five 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). Disfluencies and mispronunciations were avoided by requiring speakers to repeat errorful utterances. The resulting corpus, as compared to spontaneous speech corpora such as the Map Task (Anderson et al., 1991) or switchboard (Goffrey et al., 1992), lends itself to spectral analysis due to the high quality of the recording, the relatively slow rate of speech (lower than typical spontaneous speech), the quality of the speakers who were chosen for their ability to produce clear natural citation speech, and the filtering out of disfluency.

2. Vowels

Results presented in Sec. III are for a limited set of possible vowels within our corpus: /a/ (father), /æ/ (tap), /e/ (less), /i/ (lease), /o/ (goose). /ɔ/, /o/, /ʌ/ were avoided since these vowels were often too short for accurate automatic spectral analysis. In addition, diphthongs were avoided due to their moving spectral targets which would require more than a single measurement. And finally, the /n/ vowel (as in “lawn”) was avoided because of the extensive pronunciation variation of this vowel within the general American accent (Wells, 1982).

3. Segmentation and spectral analysis

As part of rhetorical systems’ synthesis process, the corpus was phonemically segmented using a mixture of propositional 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, log duration in milliseconds was calculated for each syllable in the corpus, and F1 and F2 values (Hz) at the temporal midpoint of each vowel were estimated using the ESPS formant tracker (Talkin, 1987).

B. Redundancy and prosodic models

1. The acoustic redundancy model

In these analyses, vowel nucleus first and second formant values were taken as measures of a syllable’s spectral quality, and centralization in terms of \(F1/F2\) was interpreted as reduced acoustic redundancy. Due to the possibility of centralization behavior differing markedly between vowels, \(F1/F2\) results are presented separately for each vowel. In addition, syllable duration was used as another measure of acoustic redundancy, and allowed for comparison with results in the literature, e.g., Aylett and Turk (2004).

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 phonemic identity (van Son and Pols, 1999). All of these measurements tended to support the hypothesis that there is a relationship between language redundancy and acoustic realization.

Syllabic probabilities were determined over unigram, bigram, and trigram contexts, i.e., the likelihood of a syllable without context, given the previous syllable, and given the previous two syllables. A log transformation was applied to these likelihoods in order to relate the values more closely to information content (Pierce, 1961) and to normalize the distributions. In order to calculate these likelihoods, the CMU language toolkit (Clarkson and Rosenfeld, 1997) was used to compute n-gram statistics based on the syllabic transcriptions of 187 million words found in news resources on the Internet. Transcriptions were automatically generated using the Rhetorical Systems speech synthesizer.

To make analyses of variance possible, and to help determine the direction of language redundancy effects, factor analysis was carried out on these likelihoods. Factor analysis has the effect of reducing the number of variables and, uses a linear transformation to create 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, and 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, s.d. 1.00; narrow factor mean 0.0, median 0.1130, s.d. 1.00. These distributions explained 96.8% of the variance in unigram/bigram/trigram likelihoods. Data points were then grouped into three redundancy groups on the basis of these distributions, for use in analyses of variance where language redundancy grouping was the independent variable of interest. The groups were constructed as follows:

1. High language redundancy group: Data points for which both narrow and wide redundancy factors were higher than the median value (+,+);

Aylett and Turk: Language redundancy and syllabic spectral characteristics

<table>
<thead>
<tr>
<th>TABLE I. Contribution from the original n-gram log likelihoods to wide and narrow factors. Values for the wide and narrow factors were produced by normalizing unigram, bigram, and trigram values using mean and standard deviation values and by multiplying them by the factor analysis components.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor analysis component</td>
</tr>
<tr>
<td>Unigram</td>
</tr>
<tr>
<td>Bigram</td>
</tr>
<tr>
<td>Trigram</td>
</tr>
</tbody>
</table>

\[ F_1 \]
TABLE II. Example syllables from each language redundancy group (in capitals). Example syllables are shown along with the two preceding syllables used to compute the target syllable’s language redundancy.

<table>
<thead>
<tr>
<th>Language Redundancy</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>to town</td>
<td>pause from Omaha</td>
<td>flag is NAILED</td>
<td></td>
</tr>
<tr>
<td>alternative</td>
<td>tanzaNA</td>
<td>upper TURBED</td>
<td></td>
</tr>
<tr>
<td>fathers such A</td>
<td>mediNA</td>
<td>radioactive gas</td>
<td>GUSHED</td>
</tr>
<tr>
<td>andy TO</td>
<td>pause in POentially</td>
<td>abeTROSS</td>
<td></td>
</tr>
<tr>
<td>kind of THE</td>
<td>flights from MALaysia</td>
<td>lab loCUSTS</td>
<td></td>
</tr>
</tbody>
</table>

(2) Medium language redundancy group: Data points for which either (1) narrow redundancy was higher than the median and wide redundancy was lower than the median or (2) narrow redundancy was lower than the median and wide redundancy was higher than the median. (+,−) or (−,+); and

(3) Low language redundancy group: Data points for which both narrow and wide redundancy factors were lower than the median value (−,−).

See Table II for an example of syllables that fell into each of 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 (see Aylett, 2000 and Aylett and Turk, 2004 for supporting durational data). The prosodic model used to test this hypothesis in the current study is a simple one that consists of three levels of prosodic prominence and three levels of boundary. It is identical to the one Aylett and Turk (2004) used for their automatically coded data, apart from the full/reduced vowel distinction that does not apply to the current, full-vowel only, dataset. It 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 >100 ms)

For practical reasons phrasal stress and phrasal boundaries were not hand-labeled, but were instead predicted on the basis of primary lexical stress on open-class words (for phrasal stress), and on the basis of following pauses (for phrasal boundaries). It should be noted that the occurrence of phrasal stress has probably been overpredicted, and the occurrence of phrasal boundaries has probably been underpredicted.

In addition, due to practical limitations, several known distinctions were not taken into account. These include accent type, boundary tone type, phrase type (e.g., full vs intermediate), and distinctions between unstressed syllables and those that could bear secondary lexical stress (e.g., /i/ in “city” and /i/ in “reply”). Despite these obvious limitations, this prosodic model represents a reasonable number of factors known to affect duration and segmental phonetic characteristics, and as such, is a rough approximation of current prosodic phonological theory. For our purposes, it made it possible to automatically code the half million syllables in the Rhetorical Corpus in terms of prosodic factors.

C. Data summary

To summarize, each syllable in our corpus was coded for language redundancy, prosodic prominence, and position relative to prosodic boundaries, according to the two models presented above. Syllable durations were determined on the basis of segmentation already present in the Rhetorical Corpus, and spectral measures ($F1/F2$) at the temporal midpoint of each syllable were determined using the ESPS formant tracker (Talkin, 1987).

D. Data analyses

1. Control for prosodic boundary location and speaker sex

The weak form of the smooth signal redundancy hypothesis accepts that the expected language/acoustic redundancy relationship may be confounded by prosodic boundary markers, which, among other things, act to increase the duration of phrase-final syllables. 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 1. 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 was not the case for 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 formant tracking errors. Separate by-sex spectral analyses have pre-
vented any normalization errors, and have made it possible to evaluate results which may have been compromised by formant tracking errors.

2. Statistical analyses

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 predicts that the predictive power of the language redundancy model and the prosodic prominence model will be substantially shared, while acknowledging that the conventions and rules of English prosodic prominence may yield a unique prediction of variation due to prosodic prominence over and above the shared effects. The significance and direction of these relationships for our two types of dependent variables, syllable duration and vowel $F_1/F_2$ frequency, were addressed by way of a univariate ANOVA followed by posthoc 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).

a. ANOVA and posthoc t-tests. The prosodic variables in our study were categorical and therefore ideal for an ANOVA model with either syllabic duration or $F_1/F_2$ formant value as a dependent variable. Continuous language redundancy values were coded in terms of the categorical redundancy grouping variable described in Sec. II B 2 to make ANOVA possible. All posthoc t-tests were carried out for effects that were significant at $p < 0.05$ in the ANOVA. There is some debate concerning when a Bonferroni correction should be made in these circumstances. A conservative approach to the data was taken here, whereby the correction was used for all tests carried out within the same analysis. To give a concrete example, a comparison of three redundancy groups across five vowels would result in a correction factor of 15, which would then be multiplied by the $p$ value generated by each t-test. Thus, with a correction factor of 15, if a posthoc t-test returned a significance of 0.01, the $p$ value would be recomputed as a nonsignificant 0.15.

b. Linear regression and likelihood ratio tests. The main goal 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. To this end, multiple regression analyses with fixed factor ordering were carried out with syllabic duration and $F_1/F_2$ formant values as dependent variables, and with language redundancy and prosodic factors as independent variables. The result of the regression analysis revealed the size of the relationship between dependent and independent variables in terms of predicted variance. Likelihood ratio tests were then performed to reveal the unique contributions of each type of independent variable to the regression model. 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 $r^2$-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.

Because $F_1/F_2$ analyses were conducted on a by-vowel basis, a table is included which shows the prosodic prominence and redundancy characteristics for each vowel type (see Sec. III A 2). In some cases a dependency exists between vowel identity and the variance of these factors. For example, for the /s/ vowel (as in “set”), in our data 90% of all tokens were lexically stressed and 84% were open class. Such homogeneity must be taken into account when interpreting the relative contribution of prosodic and redundancy models to predicting the dependent variables by vowel.

III. RESULTS

A. Syllabic duration

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

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, significant differences might be expected between the duration results in the Rhetorical Corpus and those for a spontaneous speech corpus such as the Map Task Corpus. To investigate this question the analysis carried out on the Rhetorical Corpus was matched as closely as possible to the Map Task Corpus analysis reported in Aylett and Turk (2004).

Univariate ANOVAs testing effects of (1) language redundancy group; and (2) prominence group both showed highly significant effects (language redundancy group—high, medium, low, $F(2,313,941)=67,726, p<0.001$; prominence Group—no stress, lexical stress, phrasal stress, $F(2,313,941)=72,118, p<0.001$). Figure 2 shows the mean values of each group; all were significantly different from each other (posthoc t-tests with Bonferroni correction, all with $p<0.001$). Syllables were shorter for higher values of language redundancy and, conversely, syllables were longer for higher levels of prosodic prominence.

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 Fig. 3. Close to 20% of the predictive power was shared. These results were similar to the Map Task Corpus results reported in Aylett and Turk, 2004, where a comparable proportion of predictive power was shared. The results support the view that the smooth signal redundancy hypothesis holds for this citation speech corpus, and raise the possibility that other smooth signal redundancy effects might be found for this corpus, e.g., in the spectral domain. Because spectral effects were to be analyzed on a by-vowel basis, we analyzed the
syrllabic duration results on a by-vowel basis as well, in order
to see if smooth signal redundancy effects were more reliable
for some vowels as compared to others.

2. By-vowel durational evidence for smooth signal
redundancy

Durational results analyzed separately for each vowel
showed general support for the smooth signal redundancy
hypothesis, both in terms of the direction of the language
redundancy effects, and in terms of the shared predictive
power of language redundancy and prosodic prominence.
Figure 4 shows that vowels in the low language redundancy
group were generally longer than vowels in the high redu-
dancy group. However, there were differences across vowels
in the relationship between the durations associated with the
medium language redundancy group vs those associated with
the high redundancy group. For example, for /i/ there was a
smooth decrease across redundancy groups, but for /u/ the
result was not as easy to interpret given that durations asso-
ciated with the medium redundancy group were actually
shorter than durations in the high redundancy group. All dif-
fferences were significant at $p < 0.001$ (posthoc t-test with
Bonferroni correction).

Figure 5 shows how much of the variance in syllabic
duration (log ms) for each vowel was predicted by full re-
gression models containing both prosodic prominence and
language redundancy factors. The combined regression/prosody model significantly predicted $p < 0.001$ between
18%–57% of the syllabic duration variation across vowels.
Much of the predictive power was shared between language
redundancy and prosodic prominence, although the extent of
this shared contribution varied for different vowels. Some
vowels, such as /i/, behave similarly to the pooled results,
while others, for example /e/, seem to exhibit a very different
relationship between the prosodic model and the redundancy
factors. One complication is that neither the prosodic factors
nor the redundancy factors have similar variances across dif-
frent 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. In contrast, /e/ has lower variance and a lower mean
in the wide redundancy category which suggests this vowel
rarely if ever occurs in a very predictable syllable in any
context. In addition, with boundaries controlled, nearly all
the /e/ tokens are stressed and in open-class words. This lack
of variance available to both prominence and redundancy

FIG. 2. Syllabic duration means (log ms), presented by redundancy group
and by prosodic prominence, with prosodic boundaries controlled as speci-
fied in Sec. II D. All differences in means are significant. As a reference:
2.00 log ms=100 ms, 2.20 log ms=158.5 ms, 2.40 log ms=251.2 ms, 2.60
log ms=398.1 ms.

FIG. 3. Unique contributions of prosodic prominence and redundancy fac-
tors to the linear regression model.

FIG. 4. Syllabic duration means (log ms), by redundancy group and by
vowel. All differences in means are significant. As a reference: 2.00
log ms=100 ms, 2.20 log ms=158.5 ms, 2.40 log ms=251.2 ms, 2.60
log ms=398.1 ms.

FIG. 5. Unique contributions of prosodic prominence and redundancy fac-
tors to the linear regression model.
models can thus explain the low overall performance of the models in the /æ/ context (less than 20% of total variance explained).

B. Spectral results

1. Spectral results: t-tests

Our hypothesis as far as $F_1/F_2$ effects are concerned was that vowels should show increased centralization with increased language redundancy, and that the direction of $F_1/F_2$ differences would be different for each vowel type (See Table IV for predicted effects for each analyzed vowel). Separate data for males vs females were analyzed for each vowel, with prosodic boundaries controlled (see Sec. II D 1).

Results show that language redundancy did indeed have a significant effect on $F_1/F_2$, but the significance and the direction of the effect varied by vowel. The extent to which our predictions were met can be seen in Table V, in which effects that clearly run counter to our predictions are shown in bold face. Overall, the results were reasonably consistent with a simple centralization hypothesis except for /u/, where $F_1$ drops and /æ/ for female speakers, where $F_2$ rises. Note that cases where no difference in formant value was predicted were not evaluated in this way, on the assumption that small differences could easily be accommodated by the centralization model.

Figures 6 and 7 show these changes across the vowel space presented separately for male and female speakers. Groups 1 (high language redundancy) and 3 (low language redundancy) differed in a direction consistent with the hypothesis that language redundancy correlates with centralization, with /i/ showing a particularly good match to prediction. That is, average values for group 1 tended to be more central than average values for group 3. However, for many of the vowels, the difference between groups is reflected primarily in differences in $F_1$, with little contribution from $F_2$. Group 2, formed from a less homogeneous mixture of high/low and low/high items (see Sec. II B 2) did not show the same relationship with groups 1 and 3 across vowels.

2. Spectral results: Linear regression

Linear regression analysis showed that the redundancy and prosodic models could together predict between 0.5%–7.9% of variation in $F_1$ ($p<0.001$), and 0.2%–5.9% of the variation in $F_2$ ($p<0.001$), as shown in Table VI. The smaller percentage of explained variance for female speakers may be due in large part to poor performance of the formant tracker for these speakers. Given this concern about formant tracker reliability for the female speakers, relationships between prosodic and redundancy models were compared for male speakers only. Figure 8 shows that for many of the vowel formants, redundancy factors and prosodic prominence factors predicted a large proportion of shared variance. But, in contrast to results for syllable duration, this pattern did not hold for all cases: the shared contribution was quite small or nonexistent for e.g., /u/−$f_1$ and /æ/−$f_2$. In addition, the unique contributions of redundancy and prosodic prominence factors seemed to vary considerably across vow-

---

**TABLE III. Variation in prosodic prominence and language redundancy by vowel (prosodic boundaries controlled).**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Lexical stress</th>
<th>High prob phrasal stress</th>
<th>$F_1$ change</th>
<th>$F_2$ change</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>91%</td>
<td>63%</td>
<td>−0.58</td>
<td>0.98</td>
</tr>
<tr>
<td>æ</td>
<td>95%</td>
<td>54%</td>
<td>−0.85</td>
<td>0.78</td>
</tr>
<tr>
<td>e</td>
<td>92%</td>
<td>76%</td>
<td>−0.53</td>
<td>0.89</td>
</tr>
<tr>
<td>i</td>
<td>52%</td>
<td>37%</td>
<td>0.4</td>
<td>0.94</td>
</tr>
<tr>
<td>u</td>
<td>89%</td>
<td>54%</td>
<td>−0.03</td>
<td>0.83</td>
</tr>
</tbody>
</table>

**TABLE IV. $F_1/F_2$ predictions for each vowel as acoustic redundancy drops due to centralization.**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Prediction from a centralization model of acoustic redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>minus plus</td>
</tr>
<tr>
<td>æ</td>
<td>minus same</td>
</tr>
<tr>
<td>e</td>
<td>minus same</td>
</tr>
<tr>
<td>i</td>
<td>plus minus</td>
</tr>
<tr>
<td>u</td>
<td>plus plus</td>
</tr>
</tbody>
</table>
els: /u/−f1 had a strong redundancy component orthogonal to the prosodic components of the model, whereas /i/ had a very low unique contribution from the redundancy model. However, in all cases except /e/−f2 and /u/−f2, both redundancy and prosodic prominence factors showed significant unique contributions to the regression model (p < 0.001).

IV. CONCLUSIONS

The primary aim of this paper was to test whether the predictions of the probabilistic reduction hypothesis, speech efficiency, and the smooth signal redundancy hypothesis could be generalized to spectral dimensions. General support for these views was found in the form of significant effects of language redundancy on F1/F2 formant values: vowels with higher language redundancy and less prominence showed more centralization. Evidence for the smooth signal redundancy hypothesis in particular was found in terms of considerable overlap in effects of redundancy and prosodic prominence, supporting the view that prosodic prominence is a functional structure for achieving smooth signal redundancy and thus making the speech signal more robust in a noisy environment.

Nevertheless, these general effects were complicated by considerable by-vowel variation, some of which may 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 Sec. III A 2, where differences in the redundancy profiles between /e/ and /i/ in the Rhetorical Corpus were highlighted. However, limitations of our prosodic, redundancy, and acoustic models could also contribute to this variation between vowels. Indeed, many possible improvements could be made to the prosodic model. For example, the prosodic model used here does not include secondary lexical stress, degrees of variation in phrasal stress, accent type differentiation based on f0 contours, nor any interaction between prominence and prosodic boundaries. Refining the model along these lines might well lead to

---

**TABLE V.** Changes in F1/F2 mean values from low to high redundancy.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Low Redundancy change</th>
<th>High Redundancy change</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>minus/42.0</td>
<td>plus/ 8.76</td>
</tr>
<tr>
<td>e</td>
<td>minus/−21.0</td>
<td>same/ 15.1</td>
</tr>
<tr>
<td>e</td>
<td>minus/−72.2</td>
<td>same/−36.3</td>
</tr>
<tr>
<td>i</td>
<td>plus/ 8.4</td>
<td>minus/−48.8</td>
</tr>
<tr>
<td>u</td>
<td>plus/−18.43</td>
<td>plus/ 31.72</td>
</tr>
</tbody>
</table>

**Female speakers (n=5)**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Low Redundancy change</th>
<th>High Redundancy change</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>minus/188.6</td>
<td>plus/−13.2</td>
</tr>
<tr>
<td>e</td>
<td>minus/−44.1</td>
<td>same/ 37.65</td>
</tr>
<tr>
<td>e</td>
<td>minus/−119.06</td>
<td>same/ 12.93</td>
</tr>
<tr>
<td>i</td>
<td>plus/ 8.76</td>
<td>minus/−37.0</td>
</tr>
<tr>
<td>u</td>
<td>plus/−5.0</td>
<td>plus/ 82.2</td>
</tr>
</tbody>
</table>

---

**TABLE VI.** F1/F2 linear regression results for combined redundancy and prosodic prominence models, split by vowel and by speaker sex.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Male F1/F2</th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.261</td>
<td>0.068</td>
<td>0.113</td>
</tr>
<tr>
<td>e</td>
<td>0.215</td>
<td>0.046</td>
<td>0.190</td>
</tr>
<tr>
<td>e</td>
<td>0.281</td>
<td>0.079</td>
<td>0.081</td>
</tr>
<tr>
<td>i</td>
<td>0.214</td>
<td>0.046</td>
<td>0.243</td>
</tr>
<tr>
<td>u</td>
<td>0.228</td>
<td>0.052</td>
<td>0.205</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Female F1/F2</th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.228</td>
<td>0.052</td>
<td>0.071</td>
</tr>
<tr>
<td>e</td>
<td>0.071</td>
<td>0.005</td>
<td>0.189</td>
</tr>
<tr>
<td>e</td>
<td>0.266</td>
<td>0.071</td>
<td>0.040</td>
</tr>
<tr>
<td>i</td>
<td>0.115</td>
<td>0.013</td>
<td>0.223</td>
</tr>
<tr>
<td>u</td>
<td>0.075</td>
<td>0.006</td>
<td>0.096</td>
</tr>
</tbody>
</table>
A greater overlap between the prosodic and the redundancy models. And, similarly, the redundancy model could be improved with additional specifications such as word frequency phone sequence likelihood, a neighborhood density metric, the probability of the syntactic category given a probabilistic phone sequence likelihood, a neighborhood density metric, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether a word is subject to intonation, and, at a discourse level, factors such as whether A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech, “The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech,” Lang Speech 47, 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.” J. Acoust. Soc. Am. 113, 1001–1024. Bolinger, D. (1963). “Length, vowel, juncture,” Linguistics 1, 5–29. Bond, Z., and Moore, T. (1994). “A note on the acoustic-phonetic characteristics of inadvertently clear speech.” Speech Commun. 14, 325–337. 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 Commun. 20, 255–272. 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. Pirenne (Cambridge University Press, Cambridge). Clarkson, P., and Rosenfeld, R. (1997). “Statistical language modeling using the cmuCambridge toolkit,” in Proceedings of Eurospeech 97, pp. 2707–2710. Fowler, C. A., and Housum, J. (1987). “Talkers’ signaling of ‘new’ and ‘old’ words in speech and listeners’ perception and use of the distinction.” J. Mem. Lang. 26, 489–504. Godfrey, J., Hoffman, 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 Girard, 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 (Benjamin, Amsterdam, pp. 229–254. Lieberman, P. (1963). “Some effects of semantic and grammatical context on the production and perception of speech.” Lang Speech 6, 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. Hard...