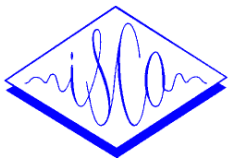

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

EFFECTS OF CONTEXT, STRESS, AND SPEECH STYLE ON AMERICAN VOWELS[1]

Caroline B. Huang

Department of Electrical Engineering and Computer Science and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
USA

ABSTRACT

In the present study, formant frequencies are measured in General American English vowels taken from a read story which is well-controlled with respect to consonant context, lexical stress, and speech style, including words in a carrier phrase, continuously read speech, and spontaneous speech. The vowels carry primary or secondary lexical stress. In total, the database consists of approximately 1500 vowel tokens from four speakers. Vowels from each speaker are analyzed separately. Results indicate that consonant context has a greater effect on formant frequencies of non-reduced vowels at the midpoint than lexical stress or speech style.

I. INTRODUCTION

Previous studies (Stevens and House, 1963; Lehiste, 1962) have investigated the effect of consonant context on the midpoint F1 and F2 frequencies of vowels. The effect is described as a shift of the midpoint on an F1-F2 plot relative to the position of the midpoint in some reference context. Similarly, midpoint shifts have been described for unstressed vowels relative to stressed vowels and vowels in running speech relative to isolated vowels (e.g., Delattre, 1969; Koopmans-van Beinum, 1980). The data of the present study were analyzed in a similar fashion. However, the studies of consonant context and lexical stress involved only isolated words or words in a carrier phrase, and the studies of continuous speech did not consider each consonant context separately. The corpus used for the present study is well-controlled with respect to all three factors and therefore allows comparison of these effects.

II. CORPUS

The corpus for the study was specially designed to be well-controlled with respect to the vowels of interest, their immediate consonant context, and their degree of lexical stress. The corpus consists of three kinds of text: a read story several paragraphs long, real words read in a carrier phrase, and nonsense words read in a carrier phrase. In addition, the speakers were asked to retell the story after they had read it. The spontaneous versions of words which appeared both in the read story and the retold story were collected to form the spontaneous speech corpus.

The vowels studied were /i/, /ɪ/, /e/, /ɛ/, and /ʌ/. The consonant contexts studied were /b/, /d/, /g/, /w/, /r/, /l/. The vowels were chosen to include minimum pairs in the features height (/i/-/e/, /ɪ/-/ɛ/), frontness (/ɛ/-/ʌ/), and tenseness (/i/-/ɪ/, (/e/-/ɛ/). The /w/, /r/, and /l/ contexts were chosen because they were found in a preliminary study to affect the vowel formant frequencies the most of the consonants. They are the consonants for which the tongue body is constrained in American English to be +BACK. The stop contexts were chosen as a contrast to the liquid and glide contexts, since the tongue body is less constrained in the articulation of the stops. The consonant context consists of one of the consonants mentioned above on one side of the vowel and an alveolar consonant on the opposite side. The alveolar context, which occurs frequently in English, was chosen to maximize the number of real English words which could represent the contexts. The manner of articulation of the alveolar consonant varies, with the restriction that nasals are not allowed, since nasals tend to obscure the F1 prominence of adjacent vowels.

For purposes of analysis, the vowel tokens must be categorized according to criteria independent of their acoustic and perceptual properties. Therefore, the vowels are categorized according to their dictionary pronunciation (*Webster's Ninth New Collegiate Dictionary*, 1985), which is the citation form pronunciation agreed upon by a large number of speakers of General American English.

The vowels to be studied carry primary or secondary lexical stress. The intention was to exclude schwas from the database, and it can be argued for almost all the vowels in the contexts to be studied that they are not reducible to a schwa. For example, in the word "disobedience," the first vowel, (/ɪ/), carries secondary stress by the principle of alternating stress in English.[2] That is, the first syllable is two syllables away from the primary-stressed syllable, so it must be stressed, and the vowel between those two syllables can be reduced. In other cases, the secondary-stressed syllable is in the less-stressed word in a compound word (e.g. /i/ in "bittersweet"). A few cases where the vowel might be reducible were unavoidable. Stress shift is another phenomenon which may cause discrepancies between the nominal level of lexical stress and the realized level of prominence (as judged by acoustic measurements or human perception) (Liberman and Prince, 1977). Again, the dictionary pronunciation, a criterion independent of acoustics and perception of individual tokens, is used to categorize the vowel by stress level.[3]

For each CVC sequence, two words were sought, one for each

[1]Supported by grant no. DC-00075 from NIDODS, grant no. NS07040 from NINCDS, and grant no. N00014-82-K-0727 from DARPA-ISTO.

[2]Morris Halle, personal communication.

[3]A vowel token which was "reduced," as judged by acoustic measurements or human perception, was not omitted from the database unless it was im-

level of lexical stress. Five repetitions of these words were then embedded in a story to be read by the speakers. In an attempt to lessen the effects of factors other than consonant context and lexical stress on the vowel of interest, restrictions were applied to the words. To lessen the duration variation of the vowels, only polysyllabic words were used. (Port, 1981, showed that vowel durations in a word tend to vary inversely with the number of syllables in the word, but that the duration difference was greatest between mono- and bisyllabic words.) Also, in the read text, a syllable containing a vowel to be studied was never placed in a prepausal position. It was not always possible to find a suitable word, and therefore, coverage of the contexts was not quite complete. In some cases, a word containing a voiceless stop instead of the voiced stop consonant of the same place of articulation was accepted.

Other factors which may affect the vowels were not controlled in the read story. Segmental factors include the manner and voicing of the alveolar consonant of the CVC and transconsonantal context of the vowel. An example of a prosodic factor is phrasal stress, in particular, the presence or absence of pitch accent in the word, as defined by Pierrehumbert (1980). Whether syllable boundaries occur in or adjacent to the CVC is another factor. Syntactic factors include word class (whether the word is a noun, verb, adjective, etc.) and morpheme class (whether the vowel occurs in an affix, word root, or compound word). Semantic factors include the effect of given versus new information and the predictability of the word. Finally, the production of a vowel in a word may be affected if there are other words in the lexicon which differ only in the vowel.

A total of 854 tokens was analyzed for one male speaker (JS). Approximately 200 tokens were analyzed for each of the remaining three speakers.

III. COMPARISON OF EFFECTS

The effect of the factors on the midpoint formant frequencies will be examined. Figures 1 through 4 show vowel distributions from the speaker (JS) from whom 854 tokens were collected. Table 1 lists the number of tokens represented in each plot. For the different plots, vowel tokens were grouped according to consonant context, lexical stress, or speech style, and modelled as two-dimensional Gaussian distributions. The ellipses are equal probability contours one standard deviation away from the mean. A different subset of JS's full vowel set was taken for each plot in order to balance the factors. (For example, there is the same proportion of primary-stressed to secondary-stressed vowels in each of the /i/ distributions in the consonant context plot.) In addition, the nonsense words were excluded from the lexical stress plot because the speakers seemed inconsistent in stressing nonsense syllables meant to receive secondary stress. The spontaneous tokens were excluded from all plots except for the separate plot showing only read and spontaneous speech, because the spontaneous set was too small to balance the contexts and stress levels. For the separate plot, read tokens were only included when there was also a spontaneous token from the same word, so tokens appeared in matched pairs. From the plots, it can be seen that different consonant contexts shift the vowel midpoint distributions in F1-F2 space more than lexical stress (primary and secondary) or speech style (from nonsense words to spontaneous speech). Plots of the data from the three other speakers, not shown, are similar.

possible to measure formants. This only occurred once. Another token of the word was substituted.

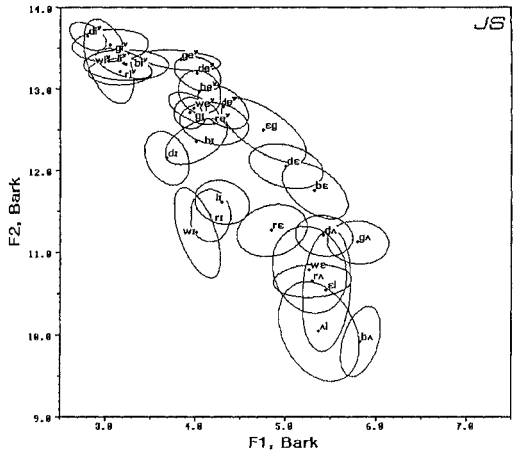

Figure 1: F1-F2 plots of midpoints for JS vowel tokens grouped according to consonantal context. Distributions modelled as two-dimensional Gaussians. Ellipses are equal-probabilty contours one standard deviation from the mean.

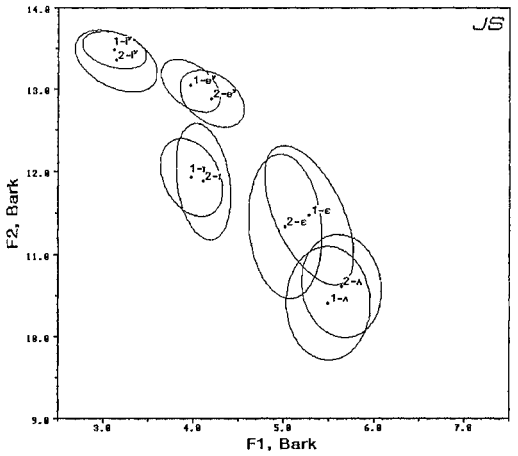

Figure 2: F1-F2 plots of midpoints for JS vowel tokens grouped according to level of lexical stress. Distributions modelled as two-dimensional Gaussians. Ellipses are equal-probabilty contours one standard deviation from the mean.

To quantify the shifting of the distributions from the effect of the factors, the Fisher Criterion was calculated. The Fisher Criterion is the ratio of a measure of the distance between the means of the Gaussian models to a measure of the scatter of the model (Duda and Hart, 1973). Figure 5 is a bar graph showing the maximum distance between any two distributions within a vowel class for each F1-F2 plot for JS. The maximum sum of distances from nonsense to read and read to spontaneous is shown for speech style. The distance is greatest among the vowel distributions when the vowels are grouped according to their consonant contexts. The great distance between vowels in different contexts is even more apparent if F3 is taken into account as well as F1 and F2. In Table 2, the Fisher Criterion distances for the other speak-

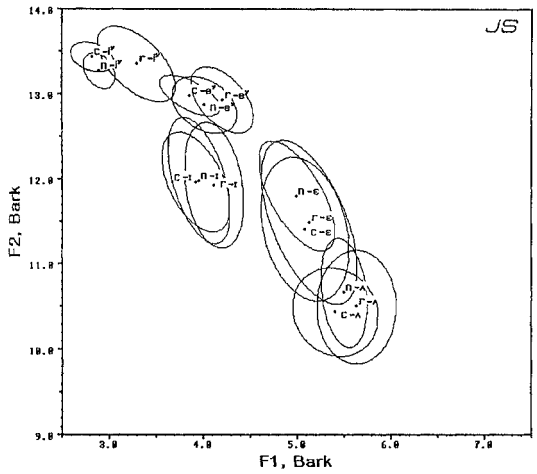

Figure 3: F1-F2 plots of midpoints for JS vowel tokens grouped according to speech style. Distributions modelled as two-dimensional Gaussians. Ellipses are equal-probabilty contours one standard deviation from the mean. Key for labels of means: n = nonsense words, c = real words in carrier phrase, r = read.

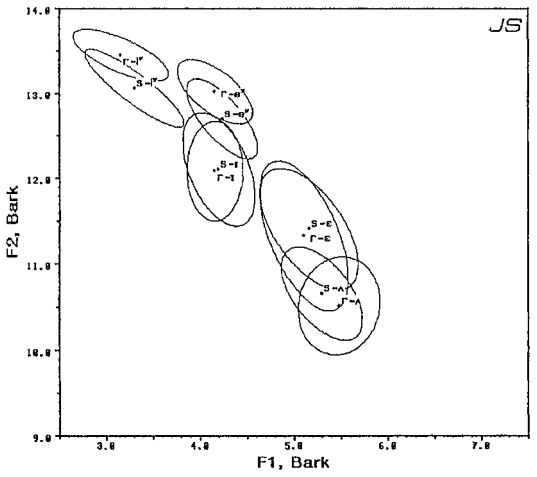

Figure 4: F1-F2 plots of midpoints for JS vowel tokens grouped according to read or spontaneous style. Each read token has a corresponding spontaneous token. Distributions modelled as two-dimensional Gaussians. Ellipses are equal-probabilty contours one standard deviation from the mean.

ers (RU, EE, and MP) are shown. (Some of the values are very high because the small number of tokens for some distributions leads to very small variances.) For each speaker and each vowel class, the distances between distributions within a vowel class are greatest for the distributions divided by consonant context.

The maximum distance between any two out of a number of distributions is likely to increase with an increasing number of distributions. A greater number of consonant contexts was examined than levels of stress or speech style. Therefore, it can be argued that the greater maximum distance between consonant context distributions arises by chance. To quantify the separation among distributions while accounting for the difference in number of distributions for each factor, a multivariate analysis of variance

Table 1: Number of tokens in distributions in each JS F1-F2 plot.

		/i/	/ɪ/	/e/	/ɛ/	/ʌ/
stress	prim.	42	42	42	42	35
	sec.	42	42	42	42	35
context	b-init.	18	18	18	18	18
	d-init.	18	18	18	18	18
	g-init.	18	18	18	0	18
	g-fin.	0	0	0	18	0
	w-init.	18	18	18	18	0
	r-init.	18	18	18	18	18
	l-init.	18	18	18	0	0
	l-fin.	0	0	0	18	18
style	nons.	30	28	28	28	24
	car.ph.	30	28	28	28	24
	read	75	70	70	70	60
style (read-sp.)	read	15	11	29	21	19
	spont.	15	11	29	21	19

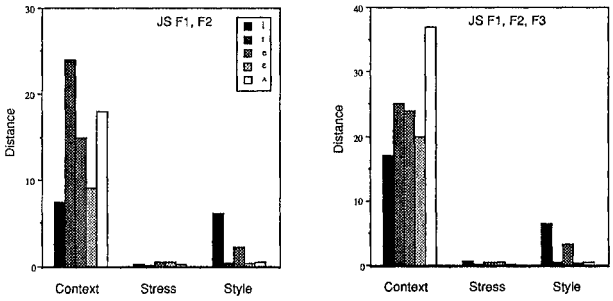

Figure 5: Maximum Fisher Criterion distance for JS vowel midpoint distributions (F1, F2 or F1, F2, F3). For example, the leftmost bar in the top graph is the distance between the distributions of F1-F2 midpoints of /bi/ and /wi/, which are furthest apart among the /i/ distributions when /i/ tokens are grouped according to context.

(MANOVA) was calculated. The MANOVA statistic (Bartlett's statistic) is similar to the Fisher Criterion but includes a factor which depends on the number of distributions within each vowel class. Effectively, when there are more distributions, the distance must be greater for the MANOVA statistic to reach the same level of significance as for a set of fewer distributions. The MANOVA was calculated on the intersection of the vowel sets used for the F1-F2 plots. Table 3 lists the values of $(1 - p)$ for each speaker, each vowel class, and each factor dividing the data, where p is the smallest level of significance for which the null hypothesis could be rejected. The quantity $(1-p)$ is the probability that the centroids within the vowel class arise from different distributions, that is, that shifts in the centroids do not arise by chance alone. The table shows that the distributions are significantly different to a level of .001 for most of the vowel classes for all speakers when divided by consonant context, though significant differences do occur elsewhere.

Since a large number of independent MANOVAs are computed, it cannot be claimed that all the statistical differences are significant at the level of the specified p value. Rather, the probability of all shifts in centroids arising from different distributions is the product of the $(1 - p)$ values obtained for each MANOVA (the "multiple comparison problem"). However, though the rep-

Table 2: Maximum Fisher Criterion distance between any two distributions within each set of vowels having the same phonemic label. Vowels having the same phonemic label are divided according to their consonant context, lexical stress, or speech style, and means and standard deviations of the smaller distributions are computed. Style data are from carrier phrase/read/spont. data for speakers RU, EE, and MP and from sum of maxima for nonsense/carrier phrase/read and read/spont. for speaker JS.

		Context		Stress		Style	
		F1,F2	F1,F2,F3	F1,F2	F1,F2,F3	F1,F2	F1,F2,F3
RU	/i/	27	27	1.1	2.1	.40	.62
	/ɪ/	20	53	.57	.61	.38	.41
	/e/	83	230	.053	.16	.40	1.1
	/ɛ/	38	82	.34	.67	1.0	1.0
	/ʌ/	28	74	.67	.86	.65	.66
EE	/i/	30	30	.56	.84	6.3	6.3
	/ɪ/	74	340	.065	.076	.040	.20
	/e/	7.7	12.8	.79	.86	.59	.59
	/ɛ/	11	11	.66	.66	.44	.44
	/ʌ/	13	17	.59	.63	1.1	1.1
MP	/i/	30	43	.97	.97	2.5	.36
	/ɪ/	170	170	.12	.21	6.7	13
	/e/	55	97	.14	.18	2.0	2.1
	/ɛ/	19	96	.56	.59	.33	2.6
	/ʌ/	62	100	.20	1.0	.17	1.8
JS	/i/	7.5	17	.19	.58	6.1	6.5
	/ɪ/	24	25	.16	.22	.36	.52
	/e/	15	24	.52	.55	2.2	3.2
	/ɛ/	9.0	20	.51	.52	.32	.37
	/ʌ/	18	37	.22	.24	.46	.52

Table 3: Values of $(1-p)$ from MANOVA, i.e., probability that the true means of formant midpoints (F1, F2, F3) of vowels from different context, stress, and style conditions are different. For example, the .969 in the RU, /i/, Context cell means that the probability is .969 that /i/'s in at least one of the contexts (e.g., /bi/) have mean formant midpoint frequencies which are different from /i/'s in other contexts. MANOVA performed on same set of vowels for context, stress, and style. Tokens in "style (read/spont.)" set are matched pairs of read and spontaneous versions. Number of tokens (n and n_s) noted for each set.

		n	Context	Stress	Style	n_s	Style(read/spont.)
RU	/i/	24	.969	.965	.415	33	.618
	/ɪ/	24	1.000	.591	.001	36	.532
	/e/	24	1.000	.148	.111	51	.990
	/ɛ/	24	1.000	.637	.750	39	.910
	/ʌ/	20	1.000	.641	.176	30	.667
EE	/i/	24	.774	.730	.978	33	1.000
	/ɪ/	24	1.000	.577	.004	42	.320
	/e/	24	.990	.736	.879	54	.924
	/ɛ/	24	1.000	.628	.126	42	.731
	/ʌ/	20	1.000	.510	.024	27	.675
MP	/i/	24	.829	.785	.783	15	.839
	/ɪ/	24	1.000	.206	.982	30	1.000
	/e/	24	1.000	.170	.979	45	.991
	/ɛ/	24	1.000	.583	.156	27	.835
	/ʌ/	20	.999	.717	.001	24	.799
JS	/i/	84	1.000	.987	1.000	30	.987
	/ɪ/	84	1.000	.773	.391	22	.143
	/e/	84	1.000	.984	1.000	58	1.000
	/ɛ/	84	1.000	.980	.014	42	.160
	/ʌ/	70	1.000	.732	.257	38	.396

resentation of the data as MANOVA statistics has some problems, this representation, as well as the plots of the distributions and the Fisher Criterion measure all suggest that consonant context shifts a vowel's midpoint more than primary and secondary lexical stress or the different speech styles studied.

IV. DISCUSSION

Evidence from this study suggests that, in continuously spoken American English, consonant context affects the vowel formant midpoints more than level of lexical stress (primary or secondary) or speech style. The effects of stress and style were small compared to the variance of the vowel distributions. This result is especially relevant now for speech synthesis and speech recognition systems, which are moving from handling isolated words and read speech to handling more spontaneous styles. However, the result seems to conflict with results of the previous studies by Delattre (1969) and Koopmans-van Beinum (1980), which found larger effects of stress and style on vowel midpoints, respectively. Delattre compared stressed vowels to vowels which would be considered schwas by the criteria of the present study. The present study compared primary- and secondary-stressed vowels. Combining the results of both studies, it may be proposed that primary- and secondary-stressed vowels are more similar to each other than to schwa vowels. An alternative explanation would be that the vowels in the database which carry secondary stress in the lexicon often received phrasal prominence in a sentence context due to the phenomenon of stress shift. Delattre's study would not have been affected by stress shift phenomena for two reasons. First, schwas do not normally receive stress in a stress shift. Second, Delattre's study involved isolated words only, and at least a phrasal context is required for stress shift to occur.

Koopmans-van Beinum's study found a larger difference between read and spontaneous vowels in a Dutch database than found in the present study in an American English database. The size of the effect of speech style may be language-dependent. Alternatively, Koopmans-van Beinum's spontaneous speech, which was collected in free conversation, may have been more casual than the spontaneous speech in the present study, which was collected by asking subjects to retell what they had just read for the experimenter. The more casual spontaneous speech may differ from read speech more than less casual speech.

REFERENCES

Delattre, P. (1969): "The general phonetic characteristics of languages. An acoustic and articulatory study of vowel reduction of four languages," Final Report, Univ. of Calif., Santa Barbara, CA, USA.

Duda, R., and Hart, P. (1973): *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, pp. 44-49, 114-118.

Koopmans-van Beinum, F. (1980): "Vowel contrast reduction, An acoustic and perceptual study of Dutch vowels in various speech conditions" Ph.D. thesis, University of Amsterdam, Amsterdam, The Netherlands.

Liberman, M., and Prince, A. (1977): "On stress and linguistic rhythm," *Linguistic Inquiry* 8, pp. 249-336.

Port, R. (1981): "Linguistic timing factors in combination," *JASA* 69(1), pp. 262-274.

Stevens, K., and House, A. (1963): "Perturbation of vowel articulations by consonantal context: An acoustical study," *Journal of Speech and Hearing Research* 6, pp. 111-128.

Webster's Ninth New Collegiate Dictionary (1985), F. Mish, Editor in Chief, Merriam-Webster Inc., Springfield, MA, USA.