Categoricality in acceptability judgements for strong versus weak vowels

Anne Cutler
MRC Applied Psychology Unit, 15 Chaucer Rd., Cambridge CB2 2EF, UK.
and
Beverley Fear
Engineering Dept., Cambridge University, Cambridge CB2 1PZ, UK.

Abstract

A distinction between strong and weak vowels can be drawn on the basis of vowel quality, of stress, or of both factors. An experiment was conducted in which sets of contextually matched word-initial vowels ranging from clearly strong to clearly weak were cross-spliced, and the naturalness of the resulting words was rated by listeners. The ratings showed that in general cross-spliced words were only significantly less acceptable than unspliced words when schwa was not involved; this supports a categorical distinction based on vowel quality.

Introduction

The speech rhythm of English is principally determined by the opposition between strong and weak syllables. English is a stress language, and its rhythm is stress-based; whether a syllable is strong or weak could therefore be seen as wholly a function of whether or not it is stressed. According to this definition strong syllables are defined as stressed syllables, while weak syllables are defined as unstressed syllables (this is the definition used in verse metrics, for example; Halle & Keyser, 1971). An alternative definition, however, equates strong syllables with those containing full vowel quality, while weak syllables are defined as those with central, or reduced, vowels, usually schwa (see, e.g. Bolinger, 1981).

The two definitions are not equivalent, because some unstressed syllables have unreduced vowels; the first vowel of *automaton*, for instance, is non-central, but carries neither primary nor secondary stress. According to a stress-based definition it would be weak, but according to a vowel-based definition it would be strong. It could be argued, therefore, that the strong-weak distinction is more of a continuum than a categorical division; perhaps both stress and vowel quality play a role, and unstressed unreduced syllables occupy an intermediate position between stressed syllables and reduced syllables.

The question is important because although the distinction between strong and weak syllables is a phonological one, it plays a role in perception: studies of speech segmentation show that in English, listeners use the strong-weak distinction as a guide to locating word boundaries (Butterfield & Cutler, 1988; Cutler & Norris, 1988; Taft, 1984). Specifically, listeners treat strong syllables as if they are highly likely to be lexical word onsets (as indeed they are; Cutler & Carter, 1987). In most of these segmentation experiments a vowel-based definition of strong versus weak was assumed, but no actual tests of this definition were undertaken; it remains unclear whether listeners would treat unstressed unreduced syllables as more like reduced syllables (a stress-based distinction), as more like stressed syllables (a vowel-based distinction), or as a true intermediate category (a dual definition). The present experiment was designed to answer this question, by examining whether listeners perceive a set of contextually matched vowels ranging from indisputably strong to indisputably weak in a continuous or a categorical manner.

Method

Five sets of four words were constructed, each set having one word each with an initial vowel bearing (1) primary stress, (2) secondary stress, (3) no stress, but with unreduced vowel quality, and (4) no stress with reduced vowel quality. The phonetic context following the vowel was the same in each word in a set. The sets were: (i) *autumn, automation, automata, atomic;* (ii) *authorise, authorisation, authentic, authority;* (iii) *audiences, auditoria, audition, addition;* (iv) *idle, ideology, idolatry, adoption;* (v) *unity, unification, united, y'know.* The sets were compiled on the basis of the vowels and stress patterns given by Jones (1958).

A meaningful context was constructed for each word. In each context, the critical word occurred after the word *but*, in the beginning of a second clause. An example set of contexts is:

Summer is the time for berries, but autumn is the time for apples.
The factory once employed eighty, but automation reduced this by half.
The workers were treated as if they weren't humans, but automata to be programmed.
Armies used to be a country's main defence, but atomic weapons changed all that.

The 20 sentences were produced by a male speaker of Standard Southern British English, who did not know the purpose of the experiment, at normal and at fast speaking rates; each word was also recorded at normal rate in a neutral context ("Say the word — again"). The speaker's 60 productions were digitised, and, within each rate and context environment, three further versions of the words in each set were produced by cross-splicing each vowel with each decapitated word body. Thus *autumn* occurred in a version with its original vowel, as well as versions with the vowels of *automation, automata,* and *atomic; atomic* occurred with its own vowel and with the vowels of *autumn, automation* and *automata*; etc. These manipulations resulted in a total of 240 tokens: 5 sets x 4 words x 4 versions x 3 environments.

An experiment containing all 240 stimuli, and hence lasting about an hour and a half, would be very fatiguing for listeners; therefore three separate tapes were made, each tape containing one-third of the experimental stimuli. Every tape contained all 16 original and cross-spliced words, in context, in one environment for each set. All five sets occurred on each tape, and all three environments were represented at least once on each tape. (One tape contained, for example, the *autumn* and *unity* sets in the normal-rate meaningful context, the *audiences* and *idle* sets in the fast-rate meaningful context, and the *authorise* set in the neutral context.) A practice set of words (*upper, upset, appeal*) was recorded in similar meaningful and neutral contexts, at normal and at fast rate, and a few cross-spliced versions were made, to provide a small set of practice examples which was recorded at the beginning of each tape.

24 listeners, students and staff of Magdalene College, Cambridge, took part in the experiment, which lasted approximately 25 minutes. All were native speakers of British English. They were tested in groups of four, two such groups hearing each tape. They were given written instructions to rate the naturalness of each critical word on a scale from 1 to 5, with a rating of 1 signifying that the pronunciation of the word on the tape could not be recognised as the intended word, and a rating of 5 signifying that the pronunciation on the tape was exactly as would be expected for that word.

The first question to be asked is whether listeners treat the sets of vowels as lying along a continuum or as falling into two categories. If listeners perceive each set as continuous, then the acceptability rating received by any cross-spliced version should be a simple function of distance between the two original versions: any cross-splicing between original versions three steps apart in the set should be less acceptable than cross-splicings two steps apart which in turn should be less acceptable than cross-splicings one step apart. Cross-splicings the same

number of steps apart should not differ in acceptability. On the other hand, if there is a category boundary within the set, then any cross-splicing which crosses the category boundary should be less acceptable than any cross-splicing which does not. If there is indeed a category boundary, then a second question can be posed; a stress-based definition of the strong-weak difference predicts a category boundary between the two stressed vowels and the two unstressed vowels, whereas a vowel-based definition predicts a category boundary between the three unreduced vowels and the one reduced vowel.

Results

Mean acceptability ratings were computed for each version of each word in each environment. The mean ratings for the 16 word types (4 words with 4 vowels), averaged across the five sets and the three environments, are shown in the left half of Table 1. They are ranked in order of rated acceptability (recall that 5 signifies most acceptable, 1 least acceptable).

How should these results best be evaluated? If we average the acceptability ratings for all unspliced words (i.e. a mean of the 1-1, 2-2, 3-3 and 4-4 rows) versus spliced versions combining words one (1-2, 2-1, 2-3, 3-2, 3-4, 4-3), two (1-3, 3-1, 2-4, 4-2) and three (1-4, 4-1) steps apart, it might seem that the results support a continuity claim: the four unspliced word types receive the highest mean (4.51), and splicings between words one, two and thee steps apart are ranked after them in order (3.86, 3.31, 2.51 respectively). However, Table 1 shows that the different combinations do *not* rank strictly in the order predicted by the continuity hypothesis. Specifically, combinations of primary stress and unreduced zero stress (1-3 and 3-1, two steps apart) are ranked higher than would be expected, while combinations of reduced and unreduced zero stress (3-4 and 4-3, one step apart) are ranked lower than expected.

Moreover, there appears to be, just as the categorically hypothesis predicts, a break in the rankings, signified by a much larger gap between the ratings for the combinations 4-3 and 4-2 than between other adjacent versions.

All Environments			Neutral Context Only		
Vowel	Body	Rating	Vowel	Body	Rating
1	1	4.57	1	1	4.88
4	4	4.55	3	3	4.78
3	3	4.54	4	4	4.70
2	2	4.39	2	3	4.68
2	3	4.34	2	2	4.65
1	2	4.21	3	2	4.53
2	1	4.20	1	2	4.50
3	2	3.99	2	1	4.28
1	3	3.90	3	1	4.28
3	1	3.62	1	3	4.20
4	3	3.53	4	3	3.35
4	2	2.97	4	2	3.00
3	4	2.87	3	4	2.88
2	4	2.75	2	4	2.78
4	1	2.53	4	1	2.70
1	4	2.50	1	4	2.70

Table 1. Ranked mean acceptability ratings as a function of vowel and word body, on the left for all environments, and on the right for the neutral context only (1 = primary stress, 2 = secondary stress, 3 = zero stress unreduced, 4 = zero stress reduced).

The neutral context is of particular interest because it offers an index of the simple acceptability of given combinations of vowel and word body, irrespective of whether a particular meaning is intended. The ratings for the neutral context only are shown in the right half of Table 1. The order is similar to the ordering for all contexts, with 1-3 and 3-1 being ranked higher than 3-4 and 4-3, and again there is a clear break in the rankings, with the gap between 1-3 and 4-3 being much larger than all other gaps between adjacent versions. Again, this supports the categoricality hypothesis.

For our statistical evaluation of the results we carried out multiple comparisons between all possible pairs of mean ratings and computed the studentised range statistic *q* for each comparison. A conventional way of presenting the results of multiple comparisons is to list the values in ranked order and draw an association line under any set of values which do not show any statistically significant differences between one another. In such a representation, then, any pair of values linked by a common association line are not significantly different; any pair of values which do not share a common association line are significantly different.

The predictions from the continuity and categoricality hypotheses can also be depicted in this form. The continuity hypothesis predicts the order described above, and four statistically defined groupings among the values:

1-1 2-2 3-3 4-4 1-2 2-1 2-3 3-2 3-4 4-3 1-3 3-1 4-2 2-4 4-1 1-4

A prediction from the categoricality hypothesis must incorporate an assumption about *where* the category boundary should fall. Assuming it to fall between reduced vowels and all others, the categoricality hypothesis predicts:

1-1 2-2 3-3 4-4 1-2 2-1 2-3 3-2 1-3 3-1 3-4 4-3 4-2 2-4 4-1 1-4

The multiple (15!, or 136) comparisons between the means averaged over all continua showed a pattern of statistical significance which can be represented as follows:

1-1 4-4 3-3 2-2 2-3 1-2 2-1 3-2 1-3 3-1 4-3 4-2 3-4 2-4 4-1 1-4

The apparent break in the distribution, between the values for 4-3 and 4-2, can be seen to be reflected in a break in the statistical association lines. The comparisons of the means from the neutral context alone showed an even more categorical pattern:

1-1 3-3 4-4 2-3 2-2 3-2 1-2 2-1 3-1 1-3 4-3 4-2 3-4 2-4 4-1 1-4

The category boundary appears to be that predicted by the vowel-based definition of the strong-weak distinction. All the cross-spliced words involving either substitution of schwa for another vowel or substitution of another vowel for schwa were rated as less acceptable than all others; moreover, the acceptability ratings for the cross-spliced words *not* involving schwa did not differ significantly either from each other or from the ratings for unaltered words. This result

held even within a neutral context, where the pattern of significance was exactly that predicted by the vowel-based categoricality hypothesis. Across all contexts, the pattern of significance was very similar, but version 4-3 (reduced vowel on the word body originally produced with an unstressed unreduced vowel) produced an anomalous result in that its rating did not differ significantly from the rating for *any* other version.

The apparent support for the continuity hypothesis provided by the average ratings of combinations one, two and three steps apart can be seen to be an artefact of the category boundary placement. Only two out of six combinations one step apart (33%) are combinations which cross the category boundary; but two out of four combinations two steps apart (50%) and all combinations four steps apart (100%) cross it. Since the categoricality hypothesis predicts that a greater proportion of cross-category combinations will produce lower mean acceptability, it would also predict this result in this case.

Conclusion

The results support the categoricality hypothesis over the continuity hypothesis, and vowel-based categories over stress-based. We conclude that listeners treat vowels as falling into two categories, with the boundary defined by vowel quality. One category contains all vowels with full, or unreduced, quality, irrespective of stress level. The other contains central, or reduced, vowels. A purely vowel-based distinction between the categories strong and weak therefore appears to be perceptually justified.

Acknowledgements

This research forms part of a dissertation research project conducted by the second author, under the supervision of the first author, for the MPhil degree in Computer Speech and Language Processing at the University of Cambridge. Financial support was provided by a grant from the Science and Engineering Research Council. We are grateful to Sally Butterfield for much assistance and to Ian Nimmo-Smith for statistical advice.

References

Bolinger, D.L. (1981) *Two Kinds of Vowels, Two Kinds of Rhythm.* Bloomington: Indiana University Linguistics Club.

Butterfield, S. and Cutler, A. (1988) Segmentation errors by human listeners: Evidence for a prosodic segmentation strategy. *Proceedings of SPEECH '88, Seventh Symposium of the Federation of Acoustic Societies of Europe*, Edinburgh; Vol. 3, pp. 827-833.

Cutler, A. and Carter, D.M. (1987) The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language, 2*, 133-142.

Cutler, A. and Norris, D.G. (1988) The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception & Performance, 14*, 113-121.

Halle, M. & Keyser, S.J. (1971) *English Stress: Its Form, its Growth, and its Role in Verse.* New York: Harper & Row.

Jones, D. (1958) *Everyman's English Pronouncing Dictionary.* London: Dent; 11th edition.

Taft, L. (1984) *Prosodic Constraints and Lexical Parsing Strategies.* PhD Dissertation, University of Massachusetts.