EFFECTS OF REPEATED INTONATION PATTERNS ON PERCEIVED WORD-LEVEL ORGANIZATION

Laura C. Dilley and Stefanie Shattuck-Hufnagel
Speech Communications Group, Massachusetts Institute of Technology

ABSTRACT

Musical and auditory perception theorists have suggested that listeners prefer to interpret parallel events, such as parallel sequences of pitches, as having parallel metrical structure. In this study, strings of full-vowel monosyllables such as *fort night club foot note book* were produced with an alternating high-low (HL) or low-high (LH) intonation pattern. These sequences could be bracketed in different ways (e.g. *fortnight clubfoot notebook* or *fort nightclub footnote book*). When sequences were produced as disyllabic words, and the initial and final syllables were removed from the utterance, listeners reorganized the syllable strings into a new sequence of disyllabic words, e.g. *nightclub footnote*. Moreover, some subjects reported different prominence status for identical syllables in original-stimulus and cut-stimulus versions at rates higher than chance. These results are consistent with a view that listener interpretations of syllable sequences reflect some of the processing constraints proposed for general auditory perception and music perception.

1. INTRODUCTION

When a perceiver confronts a stimulus, to what degree is the ensuing percept the result of physical characteristics of the stimulus, and to what degree is the percept the consequence of interpretive factors? This question has stimulated much debate in a broad range of fields, as well as substantial amounts of experimentation. Among the arguments often marshalled by those who emphasize the role of interpretation in perceptual processing is the observation that a given stimulus can be perceived differently under different circumstances. This fact been noted in a wide variety of perceptual domains, including auditory processing. Some of the factors that influence alternative interpretations of a given acoustic stimulus include prior context, repeated presentation, and listener knowledge and experience with particular forms [1]. Perception of repetition or parallelism in particular have been proposed to affect perceived *metrical structure*; parallel sequences of events tend to be interpreted with parallel metrical structures [2,3]. One factor which appears to contribute to the perception of parallel groups is pitch sequence [2]. Moreover, auditory listening experiments have shown that repetitions of simple tonal sequences can affect perceived grouping and meter. For sequences of isochronous tones which alternate between a higher pitch and a lower pitch, listeners tend to perceive higher-pitched elements as the initial accented elements of groups [4]. Listeners may perceive the lower-pitched tone in a sequence as the accented one in some cases, such as when it occurs less frequently in the sequence. These observations suggest that auditory sequences may be interpreted in more than one way with respect to grouping and metrical structure, and that interpretation may depend on many factors. Moreover, it has been noted that meter affects perception of the entire sequence, and that emerging information may cause listeners to reinterpret the organization of an entire sequence of events [3,1]. The fact that these observations hold for many aspects of audition raises the question of whether grouping and perceived metrical structure might operate similarly in language, so that the metrical structure and grouping of syllables might be influenced by pitch or other factors, as in music.

One apparently universal fact about the metrical structure of languages is that stresses tend to alternate strong and weak elements; music also evidences an alternation of strong and weak events. It has been suggested that the stress patterns of a listener's native language play a role in which syllables are perceived as strong or prominent; that is, the frequency of initial versus final main word stress may influence which syllable in a string is heard as prominent [5 and others]. The preference for alternating strong and weak elements appears to translate to a demonstrated *dispreference* in many languages for adjacent strong elements [6]. Moreover, it has been suggested that words with adjacent full-vowel syllables behave irregularly with respect to main stress placement and pitch accent [7]. Bolinger [8] has suggested that some of this confusion may arise because both syllables of such words are potential docking sites [9] for phrase-level intonational prominences or pitch accents. Alternatively, some instances of confusion may arise when adjacent FV syllables are produced with an F0 change from H to L or L to H, so that the location of prominence is ambiguous for the listener [10,11].

Dilley and Shattuck-Hufnagel [11] have reported pilot work on whether an alternating HLHL sequence could be heard with different prominence patterns depending on previous rhythmic context. For the phrase *they're all right now*, spoken with a HLHL intonation contour, more prominence reports were given on *they're* and *right* than on other words when the phrase was preceded by *maybe* (with trochaic rhythm), but more prominence reports were given on *all* and *now* when the phrase was preceded by *for sure* (with iambic rhythm). Work by Huss [12] also has suggested that preceding rhythmic context can influence listeners' reports of the location prominence on a following word.

These observations, taken together, suggest the hypothesis that strings of full-vowel syllables may in some cases be ambiguous with respect to their metrical and/or grouping structure. Moreover, we hypothesize that if confronted with a sequence of full-vowel syllables which may group into lexical items in more than one way, listeners will prefer an interpretation which is consonant with their linguistic and world knowledge as well as emerging information about the signal and innate and learned constraints. In particular, we were interested in sequences

of full-vowel syllables produced with a repeated pattern of H and L tones. For example, a syllable string such as *fort night club foot ball room* could potentially be heard as 6 monosyllabic words, or as three two-syllable words (*fortnight clubfoot notebook*), or as a combination (*fort nightclub footnote book*). If the words are produced with e.g. an alternating HL tone pattern, will the binary repetition in tone influence the listener to hear a sequence of disyllabic words? Furthermore, if the initial and final syllables *fort* and *book* are removed from the original utterance, resulting in a new LH tonal sequence, will the resulting string be heard as a different string of two-syllable words–*nightclub football*–which results in more parallel groups of syllables, intonationally and metrically? Or will it be heard with the original bracketing, *-night clubfoot ball-*? If the former, it will suggest that the listener is willing to interpret the same L tone syllables as compatible with main lexical stress in the LH condition, and with absence of main lexical stress in the HL condition, offering some support for the interpretive view of the perception of prominence patterns in strings of spoken syllables. Moreover, it would point out a further example of symmetry between language and music, as well as other modes of audition.

Thus, the experiments described in this paper address the following questions. (i) If a sequence of full-vowel syllables is produced with an intonation contour consisting of repeated HL or LH alternations, will listeners group the syllables according to the repeated parallel tone pattern, thus hearing two-syllable words? (ii) If so, will they reanalyze such strings into new two-syllable words when the initial and final syllables are removed? (iii) Will listeners hear a different set of syllables as prominent for original and cut versions?

2. METHOD

2.1. Stimuli

The stimulus utterances consisted of three types of word strings, each made up of strings of monosyllabic full-vowel words: original, cut and controls. Pairs of original and cut utterances were derived from the same string, where the string was such that each successive pair of monosyllabic words could form a new disyllabic word, as in *fortnight clubfoot notebook* or (*fort*) *nightclub footnote* (*book*). Possible compound words always carried lexical stress on the first syllable. Strings were recorded in the frame sentence *It takes [____] THEN*, where upper case indicates emphasis. (The word *then* was included at the end of the frame sentence to attract prosodic phenomena associated with the end of the phrase.) The target string was produced on a sequence of alternating H and L tones (or L and H tones), one tone per syllable, resulting in a HLHLHL (or LHLHLH) contour for this portion of the utterance. The speaker intended the target string to contain solely disyllabic words with normal prominence on the syllable with lexical main stress, i.e. on the initial syllable of each word.

Each recorded utterance was digitized at 10 kHz using Klattools running on a VAX machine; utterances were then transferred to a UNIX enviroment for modification using Xwaves software from Entropics Inc. From it, two types of experimental utterances were constructed: the original, and a cut-and-concatenated version. The original was produced by removing the final *then*. The cut-concatenated version was then constructed by splicing out the first and last syllables of the sequence of full-vowel target words, so that e.g. *fortnight clubfoot notebook* (originally HLHLHL for instance) became *-night clubfoot note-* (now LHLH). This sequence was then concatenated with the original *It takes* from the utterance. In some cases a variety of splicing locations (e.g. in the final /s/ of *takes*) were tried, to arrive at a natural-sounding cut version.

Sixteen such utterance pairs were devised, each recorded originally with both a repeated HL intonation and also with repeated LH intonation. Ten contained 8 syllables in the original string and 6 in the cut string, and six contained 6 and 4 syllables in the original and cut versions, respectively. Sixteen additional control utterances were recorded in the frame sentence using both repeated HL and LH intonation. (The word *then* was later truncated.) Control strings of monosyllabic full-vowel words could also form compound disyllabic words, but possible compound words could only be formed by one bracketing, e.g. *back space hair cut mail bag* could be *backspac*e *haircut mailbag* but not (*back*) **spacehair *cutmail* (*bag*). Controls were intended to discourage listeners from adopting listening strategies for detecting bracketing ambiguities.

Thus the set of available utterances comprised 96 stimuli: 16 originals (with *then* eliminated) in an HLHLHL version, the same 16 in an LHLHLH version, 32 corresponding cut-concatenated stimuli, and 32 controls. A stimulus tape was prepared which consisted a randomized sequence of original and cut versions, as well as controls. Each utterance was recorded three times in succession, with a 5 second pause between each such group of three.

2.2. Subjects and task

Seven subjects participated in the study, three males and four females. All were native speakers of English with normal hearing between the ages of 18 and 35. Subjects were asked to write the words that they heard, to write them in a vertical column, and to circle the syllables that they heard as prominent. No further instructions were provided. The experimenter stopped the tape after each set of three utterances if the subject was still writing; this occurred mainly for the longer utterances with 8 monosyllabic target words.

Results were transferred to a score sheet and analysed for number of disyllabic and monosyllabic responses, as well as the nature of the perceived bracketings of syllables into words (to determine whether cut stimuli were reanalyzed into new disyllabic words as predicted). The correspondence of reported prominence with the position of main lexical stress in the reported words was also assessed.

3. RESULTS

3.1. Monosyllabic versus disyllabic report

Subjects overwhelmingly reported hearing disyllabic words. Four subjects gave disyllabic responses 100% of the time for original, cut and control stimuli, with other subjects higher than 94%. Results are shown in Table 1. All subjects reported control stimuli as disyllabic words at a rate of 100%.

3.2. Word-level reassociation of syllables

There were several possibilities for how subjects would interpret the bracketing of cut stimulus sentences. One possibility was that subjects would perceive the original intended bracketing of

syllables into words, so that *It takes -night clubfoot note-* would be heard as *It takes night clubfoot note*. Another possibility was that subjects would reorganize the sequence to form new compound words.

Subject	% of syllables reported as disyllabic words
DH	100
EP	100
HC	100
KR	100
SB	99.7
EH	97.0
JS	94.7

Table 1. Rate of report of disyllabic words (n ≥ 594).

Results show that subjects overwhelmingly heard a reorganization of syllables into new compound words. These results are given in Figure 1.

3.3. Reanalysis of prominence

One hypothesis was that if subjects heard cut versions as rebracketed with respect to the originals, they would perceive as prominent a set of syllables corresponding to the lexically stressed syllables of newly indicated lexical items. One way of determining this is to ask what proportion of syllables were heard with one prominence status (+prominent or –prominent) in the original version and with the opposite prominence status in the cut version. First, however, it was necessary to determine that subjects could reliably locate prominence on the lexically main stressed syllables intended by the speaker in unmodified utterances, and more importantly, that they refrained from indicating prominence on syllables without lexical main stress.

Subjects' ability to locate prominence on syllables with lexical main stress was gauged by assessing how reliabily they indicated prominence on main-lexical-stress syllables of control words compared with the total number of prominences indicated. (All subjects had indicated control stimuli as disyllabic words at a rate of 100%.) Only three of seven subjects located prominence on main-lexical-stress syllables a high proportion of the time for control stimuli: subjects KR, EH, and SB indicated prominence on lexically stressed syllables 95%, 89%, and 76% of the time. The remaining subjects, JS, HC, EP, and DH, indicated prominence on syllables with lexical main stress 62%, 54%, 52%, and 42% of the time, respectively.

Responses of subjects KR, EH, and SB were selected for analysis of prominence reassessment in cut stimuli compared with original stimuli, since their responses on control stimuli indicated that they were able to correctly locate intended prominence on syllables with lexical main stress. These subjects' rate of prominence reassessment was calculated as follows. For those syllables which were heard in original versions with the intended prominence pattern (+prominent on lexically stressed syllables and –prominent on lexically unstressed syllables) and with the intended bracketing (i.e. all disyllabic words), the prominence indication on the corresponding syllable in the cut version was noted. If subjects indicated opposing prominence status for original versus cut versions, this was taken as an instance of "prominence reassesment".

Rates of prominence reassessment for subjects EH, SB, and KR are given in Figure 2. All subjects indicated opposite prominence status for syllables in cut versions compared with original versions at rates higher than chance of 50% ($p < 0.00001$ for all three subjects).

3.4. Subject variability in prominence report and word-level rebracketing

Subjects appeared to have very different strategies for reporting prominence. Three subjects (EH, SB, and KR) tended to indicate prominence on syllables corresponding to the lexically stressed syllables of the words they indicated, and tended to refrain from marking prominence on syllables which were not lexically stressed. The other subjects evidenced a higher tolerance for marking prominence on lexically unstressed syllables.

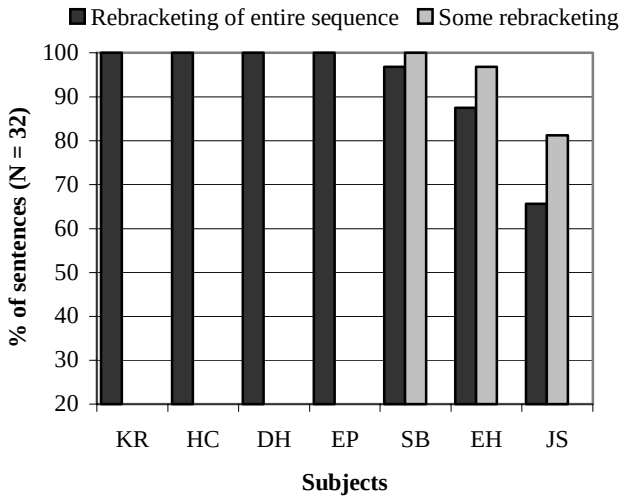

Figure 1. Word-level rebracketing of cut stimulus utterances.

4. DISCUSSION

4.1. Summary of findings

Results of these experiments suggest that listeners prefer to interpret a string of full-vowel syllables, produced on an intonation contour of alternating H and L tones, with binary word-level groupings which resulted in parallel metrical structure for perceived words. This is compatible with proposals in the music and auditory perception literatures that listeners prefer to interpret parallel groups, such as parallel sequences of pitches, with parallel metrical structure. For example, subjects showed a strong tendency to interpret the stimulus sequences as strings of disyllabic compound words. When the first and last syllables of original target utterance strings were removed, listeners easily reanalyzed the remainder into a new string of words, so that e.g. the *-stand bypass word-* portion of *grandstand bypass wordplay* was reorganized to *standby password* by all subjects in cut stimulus versions. This suggests that the intonational pattern produced for e.g. *-stand* in original versions of stimuli was heard as appropriate for either a word-*initial* syllable *with* lexical stress (in the cut stimulus), or for a word-*final* syllable *without* main

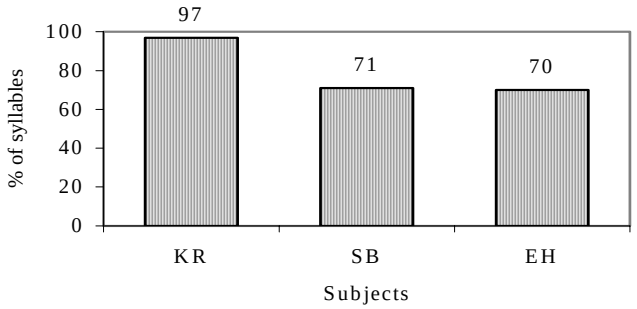

Figure 2. Percentage of syllables heard with different prominence status in cut versus original stimulus versions.

main lexical stress (in the original stimulus, and as intended by the speaker).

Additionally, some subjects reported different prominence status for identical syllables in original-stimulus and cut-stimulus versions at rates significantly higher than chance (see Figure 2). These results are consistent with the interpretation that interpretation of F0 cues to prominence depends on a number of factors, including context.

4.2. Variability in subject report

Although all seven listeners reported predominantly two-syllable words, not all listeners consistently reported prominence on the main-stressed syllables of the words they heard. Some of the variability in prominence report can be accounted for in terms of listener-specific strategies. Two of the subjects, DH and HC, who frequently marked prominence on syllables without lexical main stress, also showed a strong tendency to report prominence on *alternating* syllables, beginning with either the first or second syllable in the target string. Subject DH reported prominence on only even- or only odd-numbered syllables for 75 of 96 stimuli (78%), and subject HC did so fo 70 of 96 utterances (73%). Furthermore, this listener preferentially heard prominence *on H tones*; of the 70 sentences for which he indicated only the even- or odd-numbered syllables as prominent, 64 consisted of all H tones. Moreover, 93% of his reports of prominence on non-main-stressed syllables of the indicated words can be explained by a preference for marking H tones as prominent. This may reflect a general listener proclivity for hearing relatively higher elements in a repeating sequence as the prominent ones [3,4].

Although the rebracketing of syllable pairs into new words was more consistent across subjects than the prominence reports, there was nevertheless some degree of variability in rebracketing as well. For example, subject JS showed a strong tendency to hear disyllabic words, but she did not reanalyze the cut versions as frequently as other listeners in this study. The pattern of subject JS's responses suggests that she adopted a different listening strategy from other subjects. However, JS still heard the majority of cut versions (66%) as a sequence of new compound words, and she showed the same overwhelming tendency as other subjects to report two syllable words in all stimulus conditions.

Finally, subjects varied in how frequently they reported single syllable words. Subject JS had the most single-syllable reports: 32 of 600 syllables. Subjects EH and SB reported 18 and 2 single-syllable lexical items, respectively. All remaining subjects reported no single-syllable items.

4.3. Compatibility with an emerging framework for rhythm and intonation

Parallel groups, such as parallel sequences of pitches, are reported to elicit a sense of parallel metrical structure in listeners [2,3]. In this experiment, subjects showed a strong preference for interpreting sequences of full-vowel syllables produced with a repeating, binary HL or LH intonation pattern as grouped into disyllabic words, and listeners' preferred word-level groupings suggested a preference for parallel metrical structure for sequences with repeating tone patterns. Our findings are compatible an emerging framework for the perception of rhythm and intonation [13], and with the notion that perceived organization is determined at least in part in a top-down manner. Moreover, these results suggest that perceived prominence not entirely determined by the acoustic signal, but is rather the result of interpretation, since the same piece of acoustic signal may be heard as compatible with more than one prominence pattern and word-level organization, in different contexts. The results suggest further symmetries between language and music, as well as general auditory perception.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the NIH (grants no. RO1-DC02125 and RO1-DC02978), as well as the NSF Graduate Fellowship Program and the MIT Undergraduate Research Opportunities Program (UROP).

REFERENCES

[1] Bregman, Albert S. 1990. Auditory Scene Analysis: The Perceptual Organization of Sound}, MIT Press, Cambridge, MA.

[2] Lerdahl, F. and Jackendoff, R. 1983. *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA.

[3] Handel, Stephen. 1993. *Listening: An Introduction to the Perception of Auditory Events*. MIT Press, Cambridge, MA.

[4] Thomassen, J. M. 1982. Melodic accent: Experiments and a tentative model. *JASA* 71(6), 1596, 1605.

[5] Jakobson, Roman, Gunnar Fant, and Morris Halle. 1952. *Preliminaries to speech analysis.* Cambridge, MA: MIT Press.

[6] Nespor, Marina and Irene Vogel. 1989. On clashes and lapses. *Phonology* 6, 69-116.

[7] Shattuck-Hufnagel, S. 1995. Pitch accent patterns in adjacent-stress vs. alternating-stress words in American English. *Proceedings of the XIII International Congress of the Phonetic Sciences*, Stockholm, Vol. 3, 656-659.

[8] Bolinger, D. 1965. Pitch accent and sentence rhythm. In *Forms of English: Accent, Morpheme, Order*, Cambridge, Mass.: Harvard University Press, 163.

[9] Beckman, M. and Edwards, J. 1994. Articulatory evidence for differentiating stress categories. In *Phonological structure and phonetic form: Papers in laboratory phonology III*, ed. P.A. Keating, Cambridge, England: Cambridge University Press, 7-33.

[10] Bolinger, D. 1961. Ambiguities in Pitch Accent. *Word* 17, 309-317. Reprinted 1965 in *Forms of English: Accent, Morpheme, Order*, 119-127.

[11] Dilley, L. and Shattuck-Hufnagel, S.. 1998. Ambiguity in prominence perception in spoken utterances of American English, in *Proceedings of the 16th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America*, Vol. II, 1237-1238.

[12] Huss, Volker. 1978. English word stress in the post-nuclear position. *Phonetica* 35, 86-105.

[13] Dilley, L. A theory of regular rhythm and parallel intonation patterns in speech and language. Manuscript in preparation.