UCLA Working Papers in Phonetics, No. 105, 118-124

Optical Cues to the Visual Perception of Lexical and Phrasal Stress in English*

Rebecca Scarborough†, Patricia Keating‡, Marco Baroni, Taehong Cho††, Sven Mattys‡‡, Abeer Alwan‡, Edward Auer Jr.**, & Lynne Bernstein°*

†Stanford University, USA
‡University of California, Los Angeles, USA
*University of Bologna, Italy
††Hanyang University, Korea
‡‡University of Bristol, England
**University of Kansas, USA
°House Ear Institute, USA
†rscar@stanford.edu, ‡keating@humnet.ucla.edu

Abstract

In a study of optical cues to the visual perception of stress, three American English talkers spoke words that differed in lexical stress and sentences that differed in phrasal stress, while video and movements of the face were recorded. In a production analysis, stressed vs. unstressed syllables from these utterances were compared along many measures of facial movement, which were generally larger and faster under stress. In a visual perception study, 16 perceivers identified the location of stress in forced-choice judgments of video clips of these utterances (without audio). Phrasal stress (54% correct vs. 25% chance) was better-perceived than lexical (62% correct vs. 50% chance). The relation of the visual intelligibility of the prosody of these utterances to the optical characteristics of production is discussed, with analysis of which cues are associated with successful visual perception.

1. Introduction

When people talk, they produce optical signals that can be used by perceivers – deaf, hearing-impaired, or normal-hearing – in understanding speech. Prosody, however, is often thought to be conveyed primarily by acoustic cues, since an important aspect of prosody, namely intonation, is associated with voice f0, which is not readily apparent on a talker's face. But another aspect of prosody, namely stress, is known to be perceivable from visual-only speech. For example, Bernstein et al. [2] showed that perceivers could distinguish the position of focal stress in sentences well above chance, even though the intonation of the same sentences was not well recovered visually. And Dohen et al. [5] showed that French perceivers were very successful at identifying the position of contrastive stress in reiterant French speech (86% correct vs. 25% chance). What optical phonetic characteristics allow visual perceivers to recognize stress?

From what is known about the articulation of prosody, a focal stress (a prominence due to a nuclear pitch accent) is likely to be associated with larger, longer, and faster articulations (e.g. [3,4]). However, talkers differ in the details of how they realize such prominence, making it likely that some talkers have higher visual intelligibility for stress. The current study is designed to compare the perception of prosody with talker-specific and utterance-specific differences in its production by means of a production analysis and a perception study. By examining the relation between the results of these two studies, we can hope to determine which aspects of production lead to successful perception, and which aspects are less important to perceivers.

* Appears in Speech Prosody 2006 (Proceedings of the 3rd International Conference on Speech Prosody), Dresden: TUDpress Verlag

2. Speech Corpus

2.1. Selection of talkers

Three male native speakers of Southern Californian English who had no obscuring facial hair, tattoos, piercings, or braces were selected from a larger group on the basis of their preliminary segmental visual intelligibility, as determined from five deaf adult's ability to transcribe 20 sentences from a videotape. The three talkers were selected because they had low (T06, age 28), medium (T02, age 27), and high (T14, age 41) levels of visual intelligibility for segments in these sentences. Further intelligibility testing showed T02 and T06 to be very similar in segmental intelligibility, while T14 was more intelligible.

2.2. Speech materials

Four disyllabic minimal pairs for lexical stress (*DIScharge-disCHARGE, DIScount-disCOUNT, PERvert-perVERT,* and *SUBject-subJECT*) were recorded in their real forms and in reiterant speech. Eight other disyllabic words with initial or final stress, but not forming minimal pairs (*business, instance, courage, debit, submit, convince, gazelle,* and *cassette*) were recorded only in reiterant speech. Two reiterant syllables, *buh* and *fer*, were used. *Buh* is produced with a large mouth opening when stressed ([bʌ]) but with a smaller mouth opening when unstressed ([bə]); while the other is produced with a similar, small mouth opening whether stressed or unstressed ([fɝ] or [fɚ]), and thus was expected to be visually less informative.

The phrasal stress stimuli consisted of 24 versions of "*So,* [name1] *gave/sang* [name2] *a song from/by* [name3]*"*. The names began either with labials (*Mimi, Pammy, Bobby*) or alveolars (*Timmy, Debby, Tommy*). One of the 3 names in each sentence received a narrow-focus accent, or the sentence received a neutral (broad focus) reading. Three different orders of each set of names were used, and the location of the focal stress was varied over the first, second, and last name (e.g., *"So TOMMY have Debby a song from Timmy", "So Tommy gave DEBBY a song from Timmy", etc.*)

The words in the lexical stress corpus were produced in isolation as individual utterances. Thus, the words had both lexical stress and phrasal stress in the same location, and both the lexical stress items and the phrasal stress items involve a nuclear pitch accent.

2.3. Recording procedure

Videorecording took place in a sound-attenuated recording studio using professional-quality equipment and lighting. Twenty retroreflectors (small reflective dots) were attached to the talker's face, as in Figure 1. Three Qualisys^{TM} facial motion analysis system cameras tracked the locations of the retroreflectors using an infrared flash at a sampling frequency of 120 Hz. The 3-D coordinates of the retroreflectors were later reconstructed from the 2-D output of each camera. The data collection system, including data channels not reported here, is described in [1].

A teleprompter displaying the speech materials was positioned just below the camera so that talkers looked into the camera at all times. Each item was begun from a relaxed, closed-mouth position. Words were recorded first, displayed in triplets on the teleprompter with a real word at the top of the screen and its two reiterant versions displayed under it. Takers were instructed to read the real word first and then mimic it using *buh* or *fer*, or, for non-minimal words, to read only the reiterant versions. Reiterant speech was practiced prior to the recording. Sentences were presented one at a time. Items in both lists were blocked by stress location; each list was read twice.

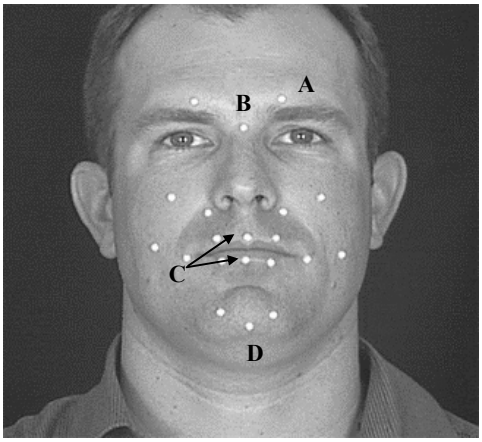

Figure 1: *Talker's face showing Qualisys retroreflectors. Labeled markers are those used for measurements in this paper: A-eyebrow, B-head, C-upper & lower lips, D-chin.*

3. Production Analysis

The articulatory correlates of word- and sentence-level stress were examined by comparing several articulatory measures for their ability to distinguish stressed and unstressed syllables. Table 1 shows the measures of facial position / movement. All differences cited are based on comparisons by ANOVA in which $p < .01$, unless otherwise specified.

Table 1: *Measurements made from facial markers. Displacement in mm; velocity in mm/s.*

1.	eyebrow displacement
2.	head displacement
3.	interlip maximum distance
4.	interlip displacement for opening gesture
5.	interlip displacement for closing gesture
6.	chin displacement for opening gesture
7.	chin displacement for closing gesture
8.	lower lip opening peak velocity
9.	lower lip closing peak velocity
10.	chin opening peak velocity
11.	chin closing peak velocity

3.1. Lexical stress

Measurements for stressed and unstressed syllables were compared. Overall, stress was well distinguished in reiterant items. All of the chin, lip, and head measures examined (#2-11) showed larger or faster mean movements in stressed syllables, as shown in Figure 2. Eyebrow movement (#1) showed no differences with stress, however, as the brow did not in fact move during lexical stress productions. As predicted, *buh* syllables showed larger and faster movements and bigger stress differences than *fer* syllables for all lip and chin measures (#3-11), suggesting more salient cues to stress in *buh* words. For head displacement (#2), however, the pattern is reversed: stressed *fer* syllables showed greater head movement than stressed *buh* syllables, possibly compensating for the reduced lip and chin markings of stress on *fer*.

More accurate perception of stress in *buh* than in *fer* words would suggest that lip and/or chin movements were more important to perception, whereas more accurate perception of *fer* word stress would suggest that head movements were important. On the other hand, if the reiterant syllable makes no difference, then a combination of measures, or other measures altogether, must drive perception.

The real word (non-reiterant) data show similar patterns of larger or faster movements in stressed syllables, but due to the noise introduced by the segmental differences between stressed and unstressed syllables in real words as well as a smaller data set (just one-fourth the size of the reiterant word set), there were fewer significant results. To counter this statistical disadvantage, both repetitions of real word productions (not just those used later in the perception experiment) were included in the analysis. Interlip distance (#3) was a reliable marker of stress, as were the chin opening measures (#6,10) at $p < .05$. There were no effects of stress on either head or brow movements. Thus, because it has more (and larger) cues, we would expect perception of lexical stress to be easier in reiterant than non-reiterant words. If speech type makes no difference, then the cues marking stress in real words (#3,6,10) can be taken to be primary.

3.2. Phrasal stress

Measurements made for the lexically stressed (initial) syllables of the test words were compared across phrasal stress conditions (focused vs. not). As sentences were not produced in reiterant versions, all comparisons are of stressed and unstressed instances of real words. Every measure examined clearly distinguished phrasally stressed and unstressed words (#1-11). Even the eyebrow measure (#1) marked stress; although unstressed words showed very little or no brow movement, talkers raised an eyebrow on almost all stressed words.

Thus, the measures that varied with lexical stress also varied with phrasal stress, but the differences between stressed and unstressed syllables identifying phrasal stress are larger, as can be seen in Figure 2. And especially with respect to real words, there are also more differentiating measures for phrasal than for lexical stress (#1-11 vs. #3,6,10). Thus phrasal stress may be expected to be easier to perceive visually than lexical stress.

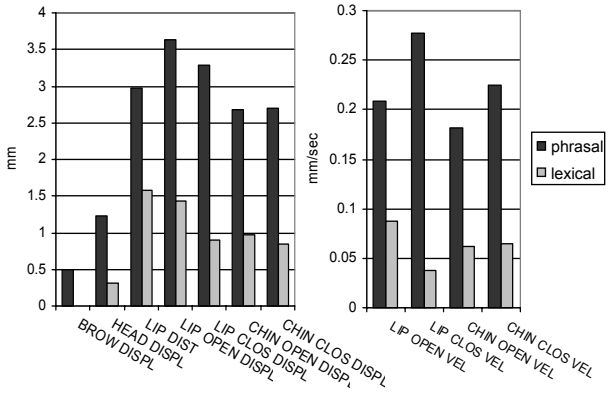

Figure 2: *Differences between stressed and unstressed measurements for reiterant lexical (in grey) and phrasal (in black) datasets. Left graph shows displacement measures; right graph shows velocity.*

3.3. Talker differences

The results reported above are overall results, across talkers. But there were also talker effects on some measures and, critically, talker interactions with stress. It is of especial interest for an intelligibility study whether any talker stands out as making particularly large or small differences along some measure(s).

Among lexical stress items, only head displacement (#2) showed such an interaction, indicating that T14 differentiated stress conditions with larger head movements than the other talkers. There were more talker differences for phrasal stress marking: again, T14 produced larger differences in head movement than the other two talkers, as well as larger lip displacement differences (#4,5) than T02; T06 also marked lip displacement (#4) more extremely than T02, and made other larger lip and chin distinctions (#7,8) than the other talkers.

Note that these talker effects do not correspond perfectly to the talker differences in visual segmental intelligibility. Based on their production characteristics, if movements in the mouth area are most important for marking prosodic differences, the most- to least-intelligible talkers should be T06>T14>T02, where one of the segmentally least-intelligible talker (T06) would be the prosodically most-intelligible. On the other hand, if head movements are most important, talkers should rank T14>T02,T06 in intelligibility, matching their segmental ranking.

Measures without talker effects are also important, as they indicate movements that are produced consistently across talkers and thus cues that would be consistently available to a perceiver. These measures include interlip distance (#3) and the chin measures (#6,10,11), as well as lip closing velocity (#9).

4. Perception experiment

4.1. Methods

Stimuli for the perception experiment consisted of one token of each item in the production study, video recording only. Sixteen paid native English speaking volunteers, aged 18-40 with normal hearing and vision and no self-reported learning disabilities, were tested individually in a sound-proof booth, about .5 m away from two 14-inch color monitors. A trial began with the presentation on the left monitor of the response choices, each of which was clickable, followed on the right monitor by a test video-clip. For both real and reiterant word stress items, the response choices were the real words, with the stressed syllable written in all capital letters. For phrasal stress items, response choices were the three names, shown in capital letters in the sentence, and also a 4th choice "No Stress". The order of conditions was always real words, *buh* reiterant words, *fer* reiterant words, and finally sentences; each item was presented twice.

4.2. Results

Overall, lexical stress was perceived significantly above chance (62% correct, vs. 50% chance). No variation in perceiver accuracy was found across conditions: real and reiterant words, *buh* and *fer* words, and minimal pair and non-minimal pair words were similarly well perceived. Likewise, all three talkers' lexical stress was perceived above chance overall, but T06 was perceived reliably better than the others on reiterant words, as shown in Figure 3.

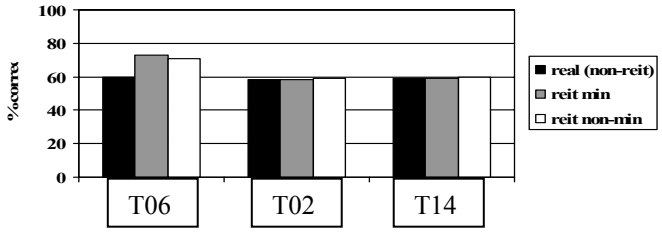

Figure 3: *Percent correct on lexical stress items, by talker.*

Note that perceivers' success with lexical stress perception is more consistent than predicted by the production patterns. Despite greater production differences for reiterant relative to real words and *buh* relative to *fer* words, both real words and *buh* words are as well perceived as others. However, T06, who was predicted to be more intelligible on the basis of his greater *phrasal* stress differentiations, is more intelligible on *lexical* stress for reiterant items.

Overall, phrasal stress was also perceived well above chance (54% correct, vs. 25% chance), and the stress of all three talkers was perceived better than chance as well. There was, however, an effect of stress location on perception, in that the no-stress condition was less-well perceived (33%) than stress in any of the three name positions (61%). Additionally, some talkers are easier to perceive correctly for stress in some positions within the sentence: T14 is best perceived in first position, T02 is best perceived in second position, and T06 is best perceived in final position.

Again, perceivers' success is more consistent than predicted by the production patterns. Despite the talker differences found in production (in particular, the greater stress distinctions made for certain measures by T06 and T14), the phrasal stress of all three talkers is perceived similarly better than chance. However, the predicted intelligibility difference between lexical and phrasal stress is borne out: there were more differentiating measures and the differences were larger for phrasal than for lexical stress, and perceivers' percent correct scores are statistically further above chance for phrasal stress than for lexical stress.

Comparing the accuracy of perception of these talkers' stress productions with their segmental intelligibility, it can be noted that there is little relation between these two types of intelligibility. Despite their range of segmental

intelligibility, talkers showed few differences with respect to their prosodic intelligibility. Only T06, who was least well perceived in the segmental task, stood out in any way in prosodic intelligibility, being best perceived on lexical stress in reiterant words.

5. Production-perception relation

Based on talker effects seen in the production measures, T06 was expected to be the most intelligible talker if movements in the mouth area mattered most, or T14, if head movement mattered most. T06 was more intelligible, suggesting that lip and chin movements are more important than head movements, but only for lexical stress in reiterant words; otherwise there was no one most intelligible talker.

Real words and *fer* words were also expected to be less-well perceived, as the production analysis showed reduced differences between their stressed and unstressed syllables. Nonetheless, real words and *fer* words were perceived no worse overall than *buh* words, indicating either that perceivers can indeed use relatively subtle information, or that the production measures reported here do not include all visually available information.

Finally, phrasal stress was perceived more accurately than word stress for all three talkers. This was expected, as almost every production measure distinguished focal stress in the sentences, while few measures consistently distinguished lexical stress in the words. However, the fact that lexical stress was perceived above chance, even for real words, shows that the movements associated with those real word stresses (interlip distance (#3) and chin opening movements (#6,10)) provide some visual information.

5.1. Correlations

The relative contribution of various production dimensions to visual perception can be further explored using correlation analysis. Because phrasal stress was reliably differentiated on more production dimensions and was better perceived than lexical stress, phrasal stress items are the focus of this analysis.

Ten of the eleven measures made in stressed syllables (all but #3), considered independently, correlated significantly with successful phrasal stress perception, with 7-40% of the variance accounted for by the relation. Chin measures accounted for the most variance, from 24% for opening velocity (#10) to 40% for opening displacement (#6). Since information about the location of phrasal stress might be contained in unstressed parts of a sentence as well, and since stress and the absence of stress might be indicated by different cues, unstressed syllables were also analyzed. In unstressed syllables, eight of the measures correlated significantly with perception (#1-6,9,11), with lip closing measures (#9,5) accounting for the most variance: 12% and 11%, respectively.

5.2. Partial correlations

Although ten of the production measures are correlated with the perception results, many of these measures are also correlated with one another, so it is not yet clear how many *independent* production variables affect the perception of phrasal stress. Partial correlations were examined to assess the relative importance of individual production dimensions, controlling for the contributions of others. Since stressed syllable measures were found to correlate better with perception than unstressed syllable measures, only those measures are considered in the following analysis.

Because of the close articulatory relation between the lips and the chin, which move largely together in mouth opening, partial correlations of the chin measures (#6,7,10,11) with perception, controlling for the corresponding contributions of the lips (#4,5,9,10), were examined. In all cases, the partial correlations reveal a significant unmatched contribution of the chin over the lips. Chin opening displacement accounted independently for the greatest amount of variance: 25%. Similarly, corresponding displacement and velocity measures are highly correlated, and partial correlations indicate that velocity never shows an independent contribution to perception over the contribution of displacement. On the other hand, displacement accounts independently for up to 26% of the variance in perception (in the case of chin opening displacement, #6) even when velocity is controlled for.

6. Conclusion

This study has shown that larger and faster mouth opening movements, more open mouth positions, and head movements can allow visual perceivers to recover information about lexical and phrasal stress. Across conditions (distinguishing stress in sentences, reiterant, and real words) and across talkers, though, interlip distance (#3) and chin opening gestures (#6,10) were most consistently produced. And chin opening, with an average displacement

of 2.7 mm in stressed syllables, seems to be one of the most important production variables for visual perception, as it has the greatest unmatched predictive power for successful perception. This is not to say, however, that visual perceivers read lexical and phrasal stress either directly or exclusively from the chin. Perceptual robustness depends generally on the presence of a variety of correlated cues. This study has revealed a number of such cues and has shown as well that when more aspects of the articulation distinguish stressed from unstressed (as they did for phrasal as opposed to lexical stress), perception is more accurate.

7. References

[1] Bernstein, L.E., Auer, E.T., Chaney, B., Alwan, A., Keating, P.A. 2000. Development of a facility for simultaneous recordings of acoustic, optical (3-D motion and video), and physiological speech data. *JASA* 107, 2887.

[2] Bernstein, L.E., Eberhardt, S.P., Demorest, M.E. 1989. Single-channel vibrotactile supplements to visual perception of intonation and stress. *JASA* 85, 397-405.

[3] Cho, T. 2002. *The Effects of Prosody on Articulation in English*. New York: Routledge.

[4] de Jong, K. 1995. The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation. *JASA* 97, 491-504.

[5] Dohen, M., Loevenbruck, H., Cathiard, M.-A., Schwartz, J.-L. 2004. Visual perception of contrastive focus in reiterant French speech. *Speech Communication* 44, 155-172.