Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English

Patricia Keating†, Marco Baroni‡, Sven Mattys*, Rebecca Scarborough†, Abeer Alwan†, Edward T. Auer, and Lynne E. Bernstein****

† University of California, Los Angeles, U.S.A.

‡ University of Bologna, Italy

* University of Bristol, England

** House Ear Institute, U.S.A.

ABSTRACT

Three male 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 study, stressed vs. unstressed syllables from these utterances were compared along many measures of facial movement, which were generally larger and faster. 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 visual intelligibility to the optical characteristics of production is discussed.

1. INTRODUCTION

Prosody 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. However, another aspect of prosody, namely stress, is known to be perceivable from visual-only speech (e.g. [2,3,4,7]). For example, Bernstein et al. [2] showed that perceivers could distinguish the position of focal stress in sentences well above chance (76% correct vs. 33% chance), while the intonation of the same sentences was not well recovered visually. 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 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. Thus we can hope to determine which aspects of production lead to successful perception, and which aspects are less important to perceivers.

2. SPEECH CORPUS

2.1 SELECTION OF TALKERS

Three male native speakers of So. Californian English who had no 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 adults' ability to transcribe 20 sentences from a videotape. The three talkers were selected because they had low (T02, age 27), medium (T06, age 28), and high (T14, age 42) levels of visual intelligibility for segments in these sentences; however, a subsequent visual perception experiment with these talkers has shown that T02 and T06 do not differ reliably.

2.2 SPEECH MATERIALS

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

The phrasal stress stimuli consisted 24 sentences which were all versions of "*So*, [name1] *gave/sang* [name2] *a song from/by* [name3]". The names began 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 was used, and the location of the focal stress was varied over the first, second, and last name (e.g. "*So TOMMY gave Debby a song from Timmy" – "So Tommy gave DEBBY a song from Timmy"*).

ISBN 1-876346-48-5

The words in the lexical stress corpus were utterances by themselves, and thus the lexical stress was also the location of the phrasal stress of the utterance. When the words formed minimal pairs, this phrasal stress was likely to be a narrow, focal stress. Thus in an important sense the prosody of the words and sentences was likely to be similar.

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 for use with the Qualisys™ facial motion analysis system. Three Qualisys™ cameras tracked the locations of the retroreflectors in three dimensions by 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].

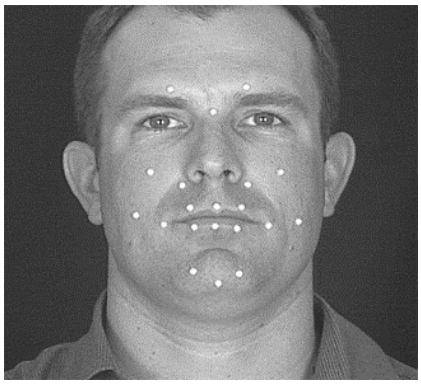

Figure 1: Talker's face showing arrangement of the 20 retroreflectors used to track facial movements.

A teleprompter displaying the speech materials was 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. Talkers 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 extensively 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.

3. PRODUCTION ANALYSIS

The articulatory correlates of word-level and sentence-level stress were examined by comparing several articulatory measures for their ability to distinguish the stressed and unstressed syllables. Table 1 shows the measures of facial position/movement. No duration measurements were included, as preliminary work indicated that peak velocity measures gave the same information. In general, measures of closing movements (#6,7,10,11) showed few differences and are not reported here. All differences cited are based on comparisons by ANOVA in which $p < .01$.

1.	left eyebrow displacement (re nose bridge)
2.	head (bridge of nose) displacement
3.	maximum midline interlip distance
4.	lip displacement for opening gesture
5.	lower lip opening peak velocity
6.	lip displacement for closing gesture
7.	lower lip closing peak velocity
8.	chin displacement for opening gesture
9.	chin opening peak velocity
10.	chin displacement for closing gesture
11.	chin closing peak velocity

Table 1: Measurements made from facial markers.

3.1 LEXICAL STRESS

Measurements were made for all the words produced in the lexical stress corpus (not just those used later in the perception experiment, as that set has only one token of each item). We compared the first to the second syllable within words, a syntagmatic comparison that corresponds to what perceivers might do in deciding which syllable is stressed. For reiterant pairs, two of the lip measures (#4,5) distinguished stressed from unstressed reiterant syllables (larger and faster opening when stressed), but the differences were larger for *buh* than for *fer*. Figure 2 shows this effect for measure #4, lip displacement for opening gesture (results for #5 are almost identical). Another measure (#3) distinguished stressed from unstressed *buh* but not *fer*, while measure #9 (and a trend for #8) distinguished stressed from unstressed *fer* but not *buh*.

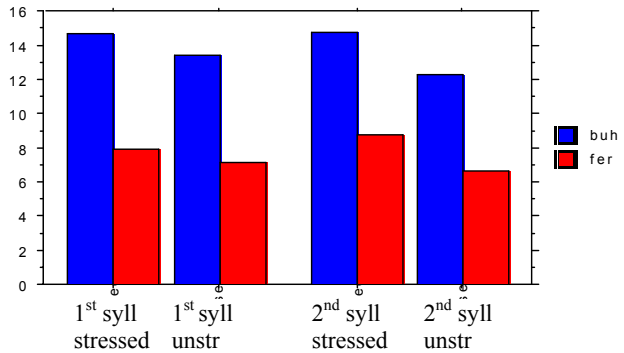

Figure 2: Lip displacement for opening gesture for reiterant words, comparing stressed and unstressed first and second syllables for reiterant syllables *buh* and *fer*. Compare 1st and 4th pairs of columns to see words with initial stress, and 2nd and 3rd pairs to see words with final stress.

However, additional information about stress on reiterant words comes from head movements (#2): for all three talkers, the head was more likely to move, or moved farther, on stressed syllables, regardless of their position in the word. These movements were the same for *buh* and *fer* words, so could compensate for the reduced lip marking of stress for *fer*. Thus if stress is more difficult to perceive on *buh* than on *fer*, or the reverse, certain measures can be taken to be more important to perception, whereas if the reiterant syllable makes no difference, then other measures must be more important. Eyebrow movement (#1), however, showed no differences with stress.

ISBN 1-876346-48-5

The non-minimal words were quite different. In these, the second syllable had larger values for facial measures #3,5,8,9 (trend for #4). For measures #3,8 stress also had a significant effect, but that was swamped by the effect of position. At most, perceivers could detect stress not through a syntagmatic comparison (which syllable is bigger?) but relative to some understood baseline value for unstressed syllables, and relative to the vowel qualities in each syllable. Furthermore, there was no effect of stress (or of any factor) on head movements in non-minimal words. Thus we would expect perception of lexical stress to be easier in reiterant than non-reiterant words.

3.2 PHRASAL STRESS

Measurements were made for the vowels of the lexically stressed (initial) syllables of the test words when phrasally stressed (focused) and not. The sentences were not produced in reiterant versions, and so all comparisons are of stressed and unstressed instances of each word. Almost every measure distinguished the stressed vs. unstressed words. These include the head and eyebrow measures (#1,2). Talkers raised an eyebrow on almost all stressed words, and moved their heads more on stressed words. Thus, the measures which varied with lexical stress also varied with phrasal stress, plus additional measures varied with phrasal stress. Thus we would expect phrasal stress to be easier to perceive visually than lexical stress.

3.3 TALKER DIFFERENCES

The results given above are overall results, across talkers. There were also talker effects on some measures. It is of especial interest for an intelligibility study whether any talker stands out as making distinctions along more measures, or making particularly large (or small) differences along some measure(s), or showing rather different results for lexical vs. phrasal stress.

For most lip or chin measures where there was a speaker difference, T06 had the largest effects and T14 the smallest and fewest. For example, measure #5 reflected lexical stress for the other talkers but not T14, while measure #8 reflected lexical stress for T06 but not the other talkers. Note that these speaker effects do not correspond to the speaker differences in visual segmental intelligibility, where T14 had higher intelligibility. On the other hand, the head movement results pattern differently, at least for phrasal stress: T14 had by far the largest head movements on stressed words, while T06 had virtually no head movement. Thus if movements in the mouth area are most important, the most- to least-intelligible talkers should be T06>T02>T14, while if head movements are most important, they should be T14>T02>T06.

At a more detailed level for lexical stress, T02 shows a split between first and second syllables. Unlike the other talkers, he has significantly more chin displacement for opening gesture (#8) and peak velocity (#9) in initial syllables, especially when stressed, but even when unstressed. The only cue to final stress by these measures is the smaller advantage of the initial syllable when unstressed. As a result, we would expect T02 to be more intelligible than the others for first syllable stress, but less intelligible for second syllable stress.

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 videoclip. For both real and reiterant word stress items, the response choices were the real words. For phrasal stress items, response choices were the three names, shown in capital letters in the sentence, and also a 4th choice "NoStress". The order of conditions was always real words, *buh* reiterant words, *fer* reiterant words, and finally sentences; there were 2 repetitions of each item.

4.2 RESULTS

Overall, lexical stress was perceived above chance (62% correct, vs. 50% chance). Accuracy for each talker's productions in two of the conditions is shown in Figure 3. All three talkers' lexical stress was perceived above chance overall, but for 2 of the talkers (T02, T14), some of the speech conditions were perceived at chance, and T06 was perceived better than the others on reiterant words. Overall, phrasal stress was also perceived well above chance (54% correct, vs. 25% chance). All three talkers' phrasal stress was perceived above chance overall, but the neutral condition sentences (with no focal stress) were perceived at chance for two of the three talkers (again T02, T14), and overall, T02 was perceived less accurately than the others.

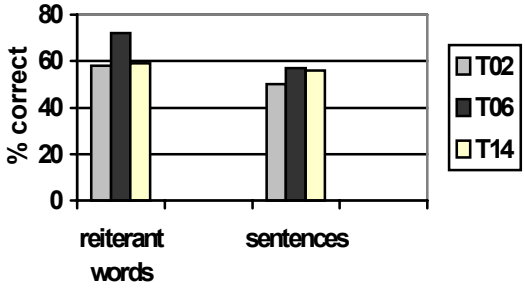

Figure 3: Mean accuracy of 16 perceivers' perception of intended lexical and phrasal stresses as produced by three talkers. Chance for words is 50%, for sentences 25%.

Comparing accuracy of perception of these talkers' stress productions with their segmental intelligibility, we can see a similarity and a difference. It is known that T14 has higher visual segmental intelligibility than T06 and T02. However, in no condition did T14 have the highest pro-

ISBN 1-876346-48-5

sodic intelligibility. For the perception of phrasal stress, T02 was less intelligible, but T06 and T14 did not differ. T02 seemed to have some trouble with the stress production task, which could explain his lower intellgibility for stress, but it would not explain his lower segmental intelligibility. T06 was the most intelligible for the perception of word stress in reiterant speech, but his segmental intelligibility was not the highest.

There were no effects of speech condition on perception of lexical stress; perceivers were equally good at perceiving intended stress in all the word stress conditions: not only in real vs. reiterant words, but for *buh* vs. *fer* reiterant syllables. However, some individual items were perceived below chance, for example, T02's reiterant words with final stress, meaning that T02's intended final stress is sometimes consistently mis-perceived as initial stress in reiterant speech.

5. PRODUCTION-PERCEPTION RELATION

Based on talker effects seen in the production measures, we expected the most intelligible talker to be T06 if the mouth area matters most, or T14 if head movement matters most. T06 was more intelligible, but only for lexical stress in reiterant words. Otherwise there was no one most-intelligible talker. We also expected T14 to be the least intelligible if mouth area matters the most, T06 if head movement matters most. Instead, T02 was less intelligible for phrasal stress. This result thus suggests that the mouth area is most important, but head movement can help make up for lack of information in the mouth area. We also expected that T02 could be more intelligible than other talkers for first syllable stress, but less intelligible for second syllable stress. This was in part the case; his intended final stress was often misperceived as initial stress. This result indicates that the parameters that he varied, chin opening displacement and velocity, are important in cueing stress (whether correctly or incorrectly).

Next, we expected some difficulty in perceiving stress on non-reiterant words and perhaps on *fer* words, as the production analysis showed reduced differences between their stressed and unstressed syllables. Non-reiterant words did not even have head movements to mark stress. Nonetheless, they were perceived overall no worse 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 shows that the movements associated with those non-reiterant stresses (interlip distance, chin opening displacement) provide some visual information.

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. When many more aspects of the articulation distinguish stressed from unstressed (as they did for phrasal as opposed to lexical stress) perception is better. This result has an implication for intellgibility of optical speech synthesis: it is better to vary several articulatory parameters than just one. We also found that talkers can differ in their visual intelligibility for stress, in a way that does not necessarily correspond to segmental intelligibility. Thus, in basing optical synthesis on a model speaker, care should be taken to model the synthetic talking face's prosody on a talker who in fact produces clear prosodic distinctions.

ACKNOWLEDGMENTS

Many thanks to Brian Chaney, Taehong Cho, Christina Esposito, and Kuniko Yasu Nielsen for help with data collection and analysis. This work was supported by NSF grant IIS-9996088 to Lynne E. Bernstein.

REFERENCES

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

[2] L. E. Bernstein, S. P. Eberhardt and M. E. Demorest, "Single-channel vibrotactile supplements to visual perception of intonation and stress," *Journal of the Acoustical Society of America,* vol. 85, pp.397-405, 1989.

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

[4] K. de Jong, "The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation," *Journal of the Acoustical Society of America,* vol. 97, pp.491-504, 1995.

[5] K.W. Grant, L.A. Ardell, P.K. Kuhl and D.W. Sparks, "The transmission of prosodic information via an electrotactile speech reading aid," *Ear Hearing* vol. 7, pp. 328-335, 1986.

[6] A. Risberg and J. Lubker Prosody and Speechreading, Speech Transmission Laboratory-Quarterly Progress Report, Status Report vol. 4, pp.1-16, 1978.

[7] D.M. Thompson, "On the detection of emphasis in spoken sentences by means of visual, tactual, and visual-tactual cues," *Journal of General Psychology*, vol.11, pp.160-172, 1934.