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

Rebecca Scarborough[1], Patricia Keating[2], Sven L. Mattys[3], Taehong Cho[4], Abeer Alwan[2]

[1]*University of Colorado, U.S.A.*
[2]*UCLA, U.S.A.*
[3]*University of Bristol, U.K.*
[4]*Hanyang University, Korea*

Key words

audiovisual speech

lexical stress

phrasal stress

visual prosody

visual speech perception

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. The production of stressed and unstressed syllables from these utterances was analyzed along many measures of facial movement, which were generally larger and faster in the stressed condition. In a visual perception experiment, 16 perceivers identified the location of stress in forced-choice judgments of video clips of these utterances (without audio). Phrasal stress was better perceived than lexical stress. The relation of the visual intelligibility of the prosody of these utterances to the optical characteristics of their production was analyzed to determine which cues are associated with successful visual perception. While most optical measures were correlated with perception performance, chin measures, especially Chin Opening Displacement, contributed the most to correct perception independently of the other

Acknowledgements: This project was funded by the NSF grant IIS 9996088 to Lynne E. Bernstein, PI, at the House Ear Institute, Los Angeles, U.S.A. We thank her, Edward T. Auer, Jr., and the talkers and perceivers who participated in the experiments at House Ear Institute; Sumiko Takayanagi and Mutsumi Shamblin for help with statistical analyses; Ying Lin for the correlation analyses of eyebrow and F0 movements; Brian K. Chaney and John Jordan (House Ear Institute) for technical assistance with hardware, software, and analysis; Marco Baroni for contributing to early work on the project; Melissa Epstein for ToBI transcriptions; and Lynne Bernstein, Ed Auer, Marion Dohen, and an anonymous reviewer for their thoughtful comments on the manuscript. Earlier versions of this study were presented at the 2003 ICPhS in Barcelona and at Speech Prosody 2006 in Dresden.

Address for correspondence. Rebecca Scarborough, Linguistics Department, University of Colorado, 295 UCB, Boulder, CO 80309, U.S.A; e-mail: <rebecca.scarborough@colorado.edu>

Language and Speech 0023-8309; Vol 52(2/3): 135–175; 103165; DOI:10.1177/0023830909103165
http://las.sagepub.com

measures. Thus, our results indicate that the information for visual stress perception is mainly associated with mouth opening movements.

1 Introduction

When people talk, they produce optical signals that can be used by perceivers—hearing-impaired or normal-hearing—in understanding speech. Most work on speech reading/lipreading and audiovisual speech perception has focused on the identification of segmental aspects of speech: individual consonant and vowel sounds, isolated words, and words in sentences. Yet prosodic information also contributes to speech recognition. Intonation, which is the linguistic use of phrasal patterns of fundamental frequency (F0), is the most obvious aspect of prosody; but because the modulation of F0 is controlled by the vocal folds, invisible inside the larynx, intonation is only partially perceived from optical signals (for examples of studies of visual perception of intonational cues to sentential syntax, such as the distinction between questions and statements in English, see Auer, Bernstein, & Coulter, 1998; Bernstein, Eberhardt, & Demorest, 1989; Grant, Ardell, Kuhl, & Sparks, 1986; Lansing & McConkie, 1999; Srinivasan & Massaro, 2003; for visual perception of lexical pitch accent in Swedish, see Risberg & Lubker, 1978). However, prosody is more than intonation; it encompasses lexical (word) stress and phrasal stress, as well as phrasal groupings and boundaries. Differences in these types of prosody can also affect the meanings of utterances. This study focuses on how information about lexical and phrasal stress is conveyed on the face.

1.1 Stress

Stress can be defined as "the linguistic manifestation of rhythmic structure" (Hayes, 1995, p.1), a structure with different levels that can be marked by different levels of stress. In English, there is one strongest stressed syllable at the level of the content word and one strongest stressed word at the level of the intonational phrase (Hayes, 1995, p.24). Other syllables or words can also be stressed, but less strongly. Stress that is a property of words is *lexical stress*; stress that is a property of phrases is *phrasal stress*.

English marks phrasal stress by associating *pitch accents*—special patterns of F0—with stressed words (or one word within stressed constituents). A pitch accent indicates that a constituent contains new or important information: it is *focused*.[1] In the default or neutral case, a pitch accent falls on the last content word of the phrase, for example, *We drove to SCHOOL* (where capitalization indicates accent). This accent marks broad focus, that is, focus on a constituent larger than the word that bears the accent, such as the entire phrase; the sentence accented in this way (implicitly) answers general questions like "What happened?" However, accent on the same word may also indicate an implicit answer to a more specific question like "Where did we drive?" In fact, any word or constituent can be marked as focused by the presence of a pitch

[1] Accent is not the only way to signal focus; other prosodic means can include speech rate and pause, and non-prosodic means (in some languages at least) include lexical choice, special morphemes, and word order. See Hirst and Di Cristo, 1998, Ch. 1, section 2.3; Ladd, 1996, Ch. 5, for more on focus.

accent. A pitch accent used in this way, that is, a focal accent, marks narrow focus.[2] The present study examines visual perception of broad and narrow focus.

When a word bears a pitch accent, that accent is usually attached to the syllable bearing lexical stress (Bolinger, 1958; for a review, Ladd, 1996). As a result, a single syllable can be simultaneously lexically stressed and pitch accented, for example, the syllable *par* in *We drove to the ˈPARty*. For a word in isolation, then, lexical stress is necessarily confounded with phrasal pitch accent, and vice versa (see, e.g., Beckman, 1986, and Ladd, 1996, pp.45–51, for discussion of studies of these confounded stresses). The present study examines visual perception of utterances in which lexical and phrasal stress co-occur on pitch-accented syllables in isolated words and in sentences.

1.2
Visual perception of stress

It is known that visual perceivers can detect phrasal stress well above chance (Bernstein et al., 1989; Thompson, 1934, and references reviewed there). For example, Bernstein et al. (1989; see also Auer et al., 1998) found that the location of phrasal stress was perceived with 76% accuracy (vs. 33.3% chance). For Swedish, Risberg and Lubker (1978) showed that the stress pattern in isolated disyllables and emphasis on one of four keywords in a sentence could both be perceived visually well above chance, and in fact, better than vowel length, within-word pitch distinctions (found in Swedish but not in English), or word boundaries. However, these studies did not specify what cues the perceivers used to identify stress.

Several acoustic correlates of lexical and phrasal stress are known to serve as cues for auditory perception, for example, fundamental frequency (F0), duration, intensity, spectral balance, and full vowel quality. Some or all of these could have optical correlates.

F0 has been claimed to be an important auditory cue for pitch accents, which mark phrasal stress (and, consequently, instances of lexical stress as well) (e.g., Fry, 1958; Rietveld & Gussenhoven, 1985). Are there optical cues to F0? Larynx height varies with F0, but the movements are small, and the thyroid cartilage is often not prominent and can be covered by clothing or facial hair. Eyebrow movements are sometimes thought to be correlated with F0 change; however, evidence for this relation is currently weak at best (e.g., Cavé et al., 1996). Head movement might also be correlated with F0; Yehia, Kuratate, and Vatikiotis-Bateson (2002) showed a good relation between head movement and F0 within each of three sentences, but there was no relation when the three sentences were combined. Thus the visual cues for F0 *per se* are likely to be weak.

[2] Focal accents can also express emphasis, which either serves an expressive, intensifying purpose, or marks a contrast (Hirst & Di Cristo, 1998). Hirst and Di Cristo distinguish focus from emphasis, two very similar notions, with the idea that emphasis is paradigmatic (this thing compared to some other thing not in this phrase) while focus is syntagmatic (this thing compared to the other things in this phrase). When a focal accent marks an explicit contrast, as in *WE drove to school, not YOU*, or *We drove to SCHOOL, not HOME*, the prominence is often called contrastive stress or contrastive focus.

Other acoustic correlates include phonetic properties such as longer duration, greater intensity (or related measures), and full vowel quality (e.g., Beckman, 1986; Herment-Dujardin & Hirst, 2002; Kochanski, Grabe, Coleman, & Rosner, 2005; Lehiste, 1970; for French, Alexandre & Gérard, 2002; for Swedish, Fant, Kruckenberg, & Liljencrants, 2000; Heldner, 2003; for applications to automatic speech recognition (ASR), see Campbell, 1993; Choi, Hasegawa-Johnson, & Cole, 2005; van Kuijik & Boves, 1999). These cues are associated with articulations that are more likely than F0-related movements to be available to visual perceivers: the larger, faster, and longer jaw and lip movements that characterize the production of stressed syllables (e.g., Beckman & Edwards, 1994; Cho, 2005, 2006; de Jong, 1995; Erickson, 2002; Erickson, Fujimura, & Pardo, 1998; Harrington, Fletcher, & Beckman, 2000).[3]

Previous studies have shown that visual perceivers find information about stress in the lower half of the face, where these speech articulations can be easily tracked. Lansing and McConkie (1999) found that phrasal stress judgment was not affected when the face from the nose up was hidden from view, indicating that the mouth and chin area contains sufficient relevant information. Swerts and Krahmer (2008) found that either the top half or the bottom half of the face allowed for better-than-chance detection of phrasal stress. Even if perceivers are gazing at a talker's eyes, the lower face is readily seen because motion can be perceived by peripheral vision, and indeed locking onto the eyes would enhance a stable view of motion elsewhere on the face (for discussion, see Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998).

Finally, there are other potential optical cues not directly related to acoustic properties of speech. In particular, body movements could co-occur with pitch accents or with phrasal stress without being directly associated or correlated with F0 (Kendon, 1978). It is known that both eyebrow movements (e.g., Condon, 1976; Dohen, Loevenbruck, & Hill, 2005, 2006; Pentland & Darell, 1994) and rapid head movements (Hadar, Steiner, Grant, & Clifford Rose, 1983) can accompany pitch accents (see Lansing & McConkie, 1999, for references to earlier studies advancing this idea, e.g., Ekman, 1979). In fact, several studies have demonstrated the usefulness of eyebrow and head movements on synthetic talking heads as visual cues to audiovisual perception of phrasal stress, though acoustic cues generally dominate perception (Granström, House, & Lundeberg, 1999, and House, Beskow, & Granström, 2001, for Swedish; Krahmer, Ruttkay, Swerts, & Wesselink, 2002, and Krahmer & Swerts, 2006, for Dutch; Massaro & Beskow, 2002, and Srinivasan & Massaro, 2003, for English).[4] And with real speakers, the top half of the face alone has been shown to be nearly as effective as the complete face at conveying phrasal stress in Dutch (Swerts & Krahmer, 2008).

---

[3] In contrast, although recent studies have established that acoustic voice quality differences (measured as a balance of high vs. low frequencies in the speech spectrum) can be a more important auditory cue than overall intensity (Heldner, 2003; Kochanski et al., 2005; Sluijter, van Heuven, & Pacilly, 1997), these differences are unlikely to provide facial cues.

[4] In addition, Munhall, Jones, Callan, Kuratate, and Vatikiotis-Bateson (2004) showed that head and face movements in a Japanese synthetic talking head aided *word* intelligibility. The study does not establish how these movements helped, but perhaps as in these other languages, head movements help identify the locations of pitch accents, which in Japanese (as in Swedish but not in English) distinguish word meanings.

Some evidence about the cues used in visual perception of French focal accent comes from a recent series of studies by Dohen and colleagues (Dohen, Loevenbruck, Cathiard, & Schwartz, 2004a, 2004b, 2004c; Dohen & Loevenbruck, 2005; Dohen et al., 2005, 2006). The 2004 studies report on a one-talker corpus of 48 sentences that had either contrastive focus on one word or broad focus, and were produced either with real words or as reiterant speech. Studying the production of contrastive focus from video recordings of these utterances, Dohen et al. (2004a, 2004b) found that the talker used prefocal lengthening, focal lengthening, including initial consonant lengthening, increased lip area, larger jaw opening with greater opening velocity, and postfocal reduction. Dohen et al. (2004c) then perceptually tested the visual intelligibility of contrastive focus in these utterances, and found that it was perceived much better than chance and that it was best perceived when the duration, lip area, and jaw opening of the focused syllables were most different from neighboring syllables. The Dohen et al., 2005 and 2006 studies reported on a five-talker corpus, in which face point movements were recorded by an Optotrak, including movements of head, eyebrows, and cheeks, as well as lip opening, lip spreading, upper lip protrusion, and vertical chin opening. The largest correlation with focal accent was lip protrusion (note that French has more rounded vowels than English does). Some speakers moved their eyebrows and head, but not systematically.

1.3
This study

In sum, there are likely to be several optical correlates of lexical and phrasal stress in English, but it is not yet known which of these are used by perceivers. In this study, we compare the visual perception of prosody with its production by means of two experiments: a production experiment highlighting which of a large set of facial measures (including mouth, chin, head, and eyebrow movements) vary with stress, and a perception experiment highlighting which of these measures are associated with stresses that are more readily perceived. We examine lexical and phrasal stress marked by pitch accents both on isolated words and in sentences, in reiterant and real speech. Qualitatively, we examine how individual talkers varying in visual intelligibility for segments differ in their production and visual intelligibility of stress. Quantitatively, we examine correlations between production and perception measures to determine which optical correlates lead to successful stress perception.

2 Speech corpus

2.1
Selection of talkers

Talkers were recruited to the House Ear Institute from the University of Southern California by means of a newspaper ad and were paid for their participation. Fourteen talkers (9 females, 5 males) who (a) were native speakers of Southern California English and (b) had no facial hair around the mouth, tattoos, piercings, or braces participated in visual screening at the House Ear Institute. Each of these talkers was video-taped producing 20 sentences. Talkers were seated in front of a solid blue background, and the camera was focused such that the talkers looked directly into the camera and their faces filled most of the frame. Once recorded, the 280 sentences (14 talkers $\times$ 20 sentences) were

presented one at a time to five Deaf participants for lipreading transcription. Sentences were blocked by talker, and the order of presentation was the same for all participants.

Among the 14 talkers, percent words correct, averaged across the five lipreaders, ranged from 13.6% to 60.2%. Three talkers were selected for the present study based on their (segmental) visual intelligibility: T-LO, with a low mean intelligibility score (28.4%), T-MID, with a score in the middle of the range (45.5%), and T-HI, with a high score (55.1%).[5] Talkers were chosen across the range of visual intelligibility so that the relation between segmental and prosodic intelligibility could be examined. Additionally, if the talkers turned out to represent a corresponding range of prosodic visual intelligibility, then the factors that made one talker's prosody more perceptible than another's could be investigated. All three selected talkers were male, and their ages were 28, 27, and 42, respectively.

2.2
Speech materials

The data in our two studies were drawn from two corpora, one comprising words in isolation, the other comprising sentences. Although lexical and phrasal stress are necessarily confounded in production (as discussed above), for convenience, we will refer to stress in the isolated word corpus as "lexical stress" and stress in the sentence corpus as "phrasal stress."

2.2.1
Lexical stress

Four disyllabic minimal pairs for lexical stress were selected on the basis of the qualities of their vowels, which are relatively similar when stressed and unstressed: *'discharge—dis'charge*, *'discount—dis'count*, *'pervert—per'vert*, and *'subject—sub'ject*. When listed on a teleprompter for the talkers, the words had no overt indication of stress; instead, they were given with a disambiguating phrase, such as "*a discount store*." Eight additional disyllabic words with short lax vowels, four with initial stress and four with final stress but not comprising minimal pairs, were also selected: *business, instance, courage, debit*, and *submit, convince, gazelle, cassette*. These non-minimal pairs provided additional tokens from a wider variety of words that did not call attention to stress differences.

Reiterant nonsense syllables were used as well to facilitate direct comparisons of movements by minimizing segmental variation across stressed and unstressed syllables and across words. Two reiterant speech syllables were selected based on pilot data from the speech of the second author. The intent was to find syllables that contrast in their degree of mouth opening, and thus perhaps in their visual intelligibility. One syllable, [bʌ], was produced with a large mouth opening when stressed, but with a smaller mouth opening when unstressed ([bə]), while the other syllable, [fɝ], was produced with a similarly small mouth opening whether stressed or unstressed ([fɚ]).

Minimal pair words were produced in their real word form as well as with each of the reiterant syllables. Non-minimal pair words were produced only with reiterant

[5] A later study of sentence lipreading by hearing subjects—subjects like the perceivers in our perception study rather than Deaf lipreaders—gave a different result: T-LO and T-MID were similarly intelligible, while T-HI was still best.

Figure 1

Summary of lexical stress items and conditions

			Stress		
			1st syllable	2nd syllable	Talkers
Real speech		(minimal pair)	DIScharge DIScount PERvert SUBject	disCHARGE disCOUNT perVERT subJECT	T-LO T-MID T-HI
Reiterant speech	fer	minimal pair	DIScharge DIScount PERvert SUBject	disCHARGE disCOUNT perVERT subJECT	T-LO T-MID T-HI
		non-minimal pair	business instance courage debit	submit convince gazelle cassette	
	buh	minimal pair	DIScharge DIScount PERvert SUBject	disCHARGE disCOUNT perVERT subJECT	T-LO T-MID T-HI
		non-minimal pair	business instance courage debit	submit convince gazelle cassette	

syllables. In total, there were 40 disyllabic test words (including real and reiterant forms), each uttered by three talkers, yielding 120 tokens, or 240 test syllables. (Both stressed and unstressed syllables were analyzed for each token.) Of these, 48 were in real word minimal pairs, 96 were in reiterant minimal pairs, and 96 were in non-minimal pairs. Lexical stress items and conditions are summarized in Figure 1.

2.2.2
Phrasal stress

The sentence stimuli consisted of versions of "*So*, [name1] *gave*/*sang* [name2] *a song from*/*by* [name3]," in which one of the three names received a narrow focus accent, or the sentence received a neutral (broad focus) reading. The names comprised a set with initial labial consonants (*Mimi*, *Pammy*, *Bobby*) and a set with initial alveolar consonants (*Timmy*, *Debby*, *Tommy*). The labial names were preceded by either a velar-final (*sang*) or a vowel-final (*so*, *by*) word, while the alveolar names were preceded by either a labial-final (*gave*, *from*) or a vowel-final (*so*) word. (The words preceding the names were selected to provide articulatory contrast in terms of mouth opening between the final segment of the word and the initial segment of the following name.) 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*"—"*So Tommy gave Debby a song from TIMMY*"). Three different orders of each set of names were used such that each name appeared both accented and unaccented in each

Figure 2

Summary of phrasal stress items and conditions

	Stress				Talkers
	1st name	2nd name	3rd name	Netural	
Labial	Bobby Pammy Mimi	Bobby Pammy Mimi	Bobby Pammy Mimi	Bobby Pammy Mimi	T-LO T-MID T-HI
Alveolar	Tommy Debby Timmy	Tommy Debby Timmy	Tommy Debby Timmy	Tommy Debby Timmy	

position. The prosodic contours of the three focal stress conditions correspond roughly to the answers to the questions "Who gave/sang [name2] a song from/by [name3]?," "To whom did [name1] give/sing a song from/by [name3]?," and "From/by whom was the song [name1] gave/sang to [name2]?" For the neutral, broad focus condition, the prosodic contour corresponded to the answer to the question "What happened?"

With three orders of labial names, three orders of alveolar names, and four stress conditions, there was a total of 24 sentences. Each sentence was uttered by three talkers, giving 72 sentence tokens, each of which contained three test words, yielding a total of 216 test word tokens. (Only the stressed syllable, i.e., the first syllable, was analyzed for each word.) Phrasal stress items and conditions are summarized in Figure 2.

Note that in this corpus, both the lexical stress items and the phrasal stress items have a pitch accent. In the case of the phrasal stress sentences, the pitch accent is due either to narrow focus or to broad focus. In the case of the lexical stress items, the pitch accent is a result of the fact that the words were produced in isolation as individual intonational phrases; these pitch accents are likely due to broad rather than narrow focus, since talkers assigned the lexical stress pattern without any explicit marking of stress or instruction to produce stress contrasts.

2.3

Recording procedure

Video-recording took place in a sound-treated recording studio using a professional-quality video recorder (Sony Betacam SP) and camera (Sony DXC-30); recordings were high quality NTSC analog video. Talkers were seated in front of a dark blue background, and lighting was provided by two face-level 420-W video lights reflected off umbrellas at approximately 35 degrees to either side of the midline. The talker's nose level was slightly below the center of the video screen, and the face filled most of the screen area. Twenty retro-reflectors (small reflective dots) were attached to the talker's face for recording by a Qualisys facial motion analysis system. The 20 reflectors were attached at the locations shown in Figure 3, which were chosen on the basis of points of attachment of facial muscles, except for marker 2 on the bridge of the nose, chosen as a point where the skin moves least. This marker thus serves as an indicator of overall head position. Three Qualisys cameras tracked the locations of the reflectors in three dimensions using an infrared flash. The sampling frequency

Figure 3

Qualisys retro-reflectors. Labeled markers are those used for the measurements in this article: 1—eyebrow, 2—head, 3—upper and lower lips, 4—chin

was 120 Hz. Finally, a uni-directional Sennheiser microphone was used for acoustic recording onto a DAT machine at a sampling frequency of 44.1 kHz.

A teleprompter displaying the speech materials was positioned so that the talkers could look directly into the camera at all times. The teleprompter's scrolling speed was regulated manually by an experimenter located in a separate control room. Otherwise, there was no control of speaking rate. Each item was begun from a relaxed, closed-mouth position.

Single word tokens were recorded first. They were displayed in triplets on the teleprompter, with a real word (along with its disambiguating phrase for the minimal pair words) at the top of the screen, and the two reiterant syllables displayed under it. Accordingly, talkers were instructed to read the real word first and then mimic its stress pattern using *buh* and then *fer* (or *fer* and then *buh*). For example, in the case of *'discount*, the display showed *discount*, below which appeared *a discount store*, then below that *buh*, and at the bottom *fer*. In cases of non-minimal pairs, talkers read only the reiterant words. Reiterant speech was practiced prior to the actual recording. Triplet presentation was blocked by stress pattern (stress-initial followed by stress-final), and the entire list was read twice. If necessary, due to reading errors or technical considerations, individual items were repeated until two useable tokens had been recorded. Sentences were presented one at a time. Like the words, they were blocked by stress location (neutral, early, middle, and late), and the entire list was read twice.[6] The distinctions among the sentences—broad focus, and the different

[6] Blocking was done to lessen the likelihood of talkers producing contrastive focus; however, it might have led to repetition effects. Post hoc analysis showed significant positive correlations between order of presentation and both Chin Opening Displacement and Chin Opening Velocity, but r^2 was less than .04 in both cases. There were no other significant correlations for any measures in either the lexical or phrasal corpus. Thus, stress appears to have been produced consistently across the experimental session.

narrow focus locations—were explained to the talkers in terms of the questions that they answer, as outlined above.

At the beginning of each recording, a custom circuit (Bernstein, Auer, Chaney, Alwan, & Keating, 2000), which analyzes signals from the Qualisys and video systems, invoked a 100-ms 1-kHz pure tone that was sent to the DAT line input for offline synchronization. The Qualisys system initiated the sync tone and sync pulse first; the audio recording could then be synchronized by finding the tone position.

The same talkers participated in a separate session in which they read a subset of this corpus while electromagnetic articulographic (EMA), as well as Qualisys, video-, and audio-recordings were made. Analyses of those recordings are reported elsewhere (Keating et al., 2000).

2.4
Post-processing of Qualisys data

The 3-D coordinates of the retro-reflectors were reconstructed from the 2-D output of each camera. Occasionally a reflector's position could not be recovered for a few frames, but generally these missing points did not occur near articulatory maxima or minima, so their omission had little effect on the measurements made in the present study. Although in some other studies head position is normalized around the position of the marker on the bridge of the nose, in the present study, the position of this marker was preserved to provide an index of whole-head movements.

2.5
Prosodic transcription

Aspects of the auditory prosody as actually produced by the talkers were recorded in a prosodic transcription by two transcribers trained in English ToBI transcription (Beckman & Elam, 1997; Silverman et al., 1992). There were no differences between transcribed and intended stress for the lexical stress corpus, but there were differences for the sentence stress corpus; in particular, talkers sometimes accented more than one of the three names. Both the scripted (prompted) prosody and the actually-produced prosody are taken into account in the analyses presented here.

3 Production analysis

3.1
Optical measures

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 tokens. In this article, only the tokens used as stimuli in the subsequent perception experiment (described below)—one of the two repetitions of each item (the first, unless it contained a hesitation, disfluency, or erroneous stress)—are examined unless otherwise specified. Table 1 shows the measures of facial position/movement made for all test syllables or test words, both accented or stressed, and unaccented or unstressed. In previous studies of focus, kinematic measures of jaw, lower lip, and interlip movements have been found to be useful and reliable; we substituted chin measures for jaw measures and added head and brow measures.

Table 1

Measures made from facial markers. The abbreviations for the measures will be used in all subsequent tables

Eyebrow displacement (in mm)	BROW DISPL
Head displacement (in mm)	HEAD DISPL
Interlip maximum distance (in mm)	LIP DIST
Interlip displacement in opening gesture (in mm)	LIP OPEN DISPL
Interlip displacement in closing gesture (in mm)	LIP CLOS DISPL
Chin displacement for opening gesture (in mm)	CHIN OPEN DISPL
Chin displacement for closing gesture (in mm)	CHIN CLOS DISPL
Lower lip opening peak velocity (in mm/sec)	LIP OPEN VEL
Lower lip closing peak velocity (in mm/sec)	LIP CLOS VEL
Chin opening peak velocity (in mm/sec)	CHIN OPEN VEL
Chin closing peak velocity (in mm/sec)	CHIN CLOS VEL

All measures are in the vertical (y) dimension only. Brow Displacement refers to the most extreme difference in position, in the y-dimension, of the marker on the left eyebrow (labeled point 1 in Figure 3) relative to the y-dimension of the reference point on the bridge of the nose (point 2 in Figure 3) during the test syllable (generally at the beginning of the syllable); Brow Displacement was typically in the upward direction. Head Displacement refers to the maximum difference in position within the test syllable, in the y-dimension, of the reference point on the bridge of the nose (point 2) (generally the beginning vs. the end of the test syllable); head movements in stressed syllables were consistently in the downward direction. Lip measures refer to the vertical difference between the midline markers on the upper and lower lips (both labeled 3). Chin measures refer to the vertical position of the chin retro-reflector (point 4) relative to the reference point on the bridge of the nose (point 2). Lower lip measures refer to the vertical movement of the lower lip retro-reflector (lower point 3) relative to the reference point on the bridge of the nose (point 2). For both chin and lower lip, the usual direction of opening movement was downward, with closing movements in the upward direction. Regardless of the typical direction of the facial movement, all displacements in the expected direction are reported as positive values. No other face markers seen in Figure 3 were measured for this study (see Jiang, Alwan, Keating, Auer, & Bernstein, 2002, for analyses of correlations among these markers and of these markers with articulatory (EMA) and acoustic recordings and Jiang, Auer, Alwan, Keating, & Bernstein, 2007, for an account of the relation between visual phonetic perception and physical stimulus attributes).

Opening and closing gestures were examined for chin and lips, but it was not possible to make consistent closing gesture measurements for all lexical stress items. Since both reiterant syllables (*buh* and *fer*) end in an open gesture (a vowel or a rhoticized vowel), no measurable closing gesture occurred where there were no following segments to require mouth closing, that is, in the second syllables of reiterant words. The same problem occurred for certain real word items. These unmeasurable data points resulted in small variations in degrees of freedom across analyses. Occasionally, even in non-closing gestures, local maxima and/or minima of the various movements could not be isolated; these data points were omitted from the analysis as well.

Table 2

Mean stressed and unstressed measurements (with standard deviations in parentheses), and differences between stressed and unstressed for all items in the lexical dataset. Dist and displ in mm; vel in mm/s. Significant effects for Stress are shown in bold; starred measures were analyzed with a 2-way rather than a 3-way ANOVA (omitting the Syllable factor). The different degrees of freedom reflect the different numbers of tokens that could be measured

	Unstressed	*Stressed*	*Diff.*	*Stress effect*	*p-value*
BROW DISPL	0	0	0	n/a	n/a
HEAD DISPL	.61 (.51)	.92 (.74)	**.32**	$F(1, 194) = 12.99$	$\mathbf{p < .001}$
LIP DIST	26.89 (4.69)	28.46 (4.89)	**1.57**	$F(1, 218) = 20.22$	$\mathbf{p < .001}$
LIP OPEN DISPL	9.84 (4.36)	11.27 (4.79)	**1.43**	$F(1, 214) = 8.82$	$\mathbf{p = .003}$
LIP CLOS DISPL*	8.79 (4.64)	9.70 (4.75)	.90	$F(1, 119) = 1.52$	$p = .220$
CHIN OPEN DISPL	6.86 (3.29)	7.83 (3.03)	**.97**	$F(1, 217) = 7.24$	$\mathbf{p = .008}$
CHIN CLOS DISPL*	6.12 (2.84)	6.97 (2.87)	.85	$F(1, 117) = 2.93$	$p = .089$
LIP OPEN VEL	.71 (.32)	.80 (.36)	**.09**	$F(1, 217) = 4.83$	$\mathbf{p = .029}$
LIP CLOS VEL*	.70 (.39)	.74 (.40)	.04	$F(1, 119) = 0.56$	$p = .456$
CHIN OPEN VEL	.52 (.27)	.58 (.24)	**.06**	$F(1, 218) = 4.79$	$\mathbf{p = .030}$
CHIN CLOS VEL*	.48 (.24)	.55 (.25)	.06	$F(1, 111) = 2.33$	$p = .130$

3.2
Results for lexical stress

As detailed above, lexical stress contrasts were produced for real-word minimal pairs (e.g., DIScharge vs. disCHARGE), reiterant versions of these pairs (e.g., BUHbuh vs. buhBUH), and reiterant versions of non-minimal pairs (e.g., FERfer vs. ferFER for business vs. submit). All reiterant versions were produced with two different reiterant syllables (*buh* and *fer*). Measurements were made for both syllables of all words (insofar as possible). Measurements were then compared (separately for each optical/production measure) using factorial ANOVAs. All comparisons included Stress (stressed or unstressed), Syllable Position (1st or 2nd), and Talker (T-LO, T-MID, or T-HI) as factors. Additional factors related to particular item types are described below when they were included. The alpha level for significance was set at $p < .05$; results with $p < .08$ were considered to be marginally significant.

3.2.1
All items

Data (pooled across all real and reiterant stimuli) were initially analyzed for possible main effects of Stress, Syllable Position, and Talker. However, because of the lack of closing gesture data for second syllables, the effect of Syllable had to be omitted from the analyses of all closing measures.

With respect to the main effect of Stress on our 11 measures, 6 of the 11 measures—mostly lip and chin opening measures—were found to reliably distinguish stress, as shown in bold in Table 2. These measures were larger or faster in stressed syllables than in unstressed ones. Another four measures showed the same pattern, but the effect

of Stress failed to reach statistical significance; these were the four closing gesture measures, though the effect for Chin Closing Displacement approached significance.[7] Finally, because the eyebrows showed no movement at all, Brow Displacement was ignored in the remainder of the lexical stress analyses.

The Syllable Position factor showed a significant effect for both Lip Distance and Chin Opening Velocity, with a greater peak Lip Distance and slower Chin Opening Velocity in second syllables. Talker and three interactions with Talker also reached significance: Talker for all measures (except Brow), Syllable Position by Talker for both Chin and Lip Opening Displacement, and Talker by Stress for Head Displacement. In each case, one talker showed a greater displacement difference for stressed vs. unstressed syllables than the others. The influence of the Talker factor and these interactions will be addressed further in the local discussion below. There were no significant three-way interactions.

3.2.2
Subsets of items

When the real words (all of which were minimal pairs) were considered alone (excluding reiterant items), the pattern of data looked similar to the overall pattern, but none of the differences was statistically reliable. As can be seen in Table 3, the means for stressed syllables suggested larger, faster movements than for unstressed syllables (with the exception of Lip Opening Velocity). This pattern held for each talker and for each syllable (with the exceptions of Lip Opening Displacement in first syllables and T-HI's closing gestures). Because of the small size of the dataset for the real word condition, an expanded dataset including both the originally analyzed data and the previously unused second repetitions of each word (which were not heard by perceivers) was analyzed as well. (Head measurements were not made for second repetitions and so are not included here.) In this analysis, the effect of Stress was significant for Lip Distance and Chin Opening Velocity measures and marginally significant for Chin Opening Displacement, as can be seen in Table 4. Other measures still showed a pattern in the expected direction, though it was non-significant.

In both the restricted and expanded analyses, Syllable was also a significant factor on all of the measures (except for Head Displacement and the closing gesture measures, for which it could not be included), with second syllables showing larger or faster movements (probably due to segmental differences between first and second syllables). In the expanded dataset, all non-closing measures also showed significant effects of Talker, which will be addressed in the local discussion below.

The patterns of results obtained with reiterant pairs were similar to the real word patterns, with all measures (except Brow Displacement) showing larger or faster

[7] These four closing gesture measures had to be analyzed with a 2-way rather than a 3-way ANOVA, since there was no Syllable factor, all of the measures coming from initial syllables. The question arises whether the six measures that showed significant effects of Stress would show different results if analyzed with the same 2-way analysis. Therefore, to align the statistical design on all measures, the other six measures were re-run in a 2-way analysis without a Syllable factor. Five of them were still significant while Chin Opening Velocity showed a marginally significant effect, $p = .053$.

Table 3

Mean stressed and unstressed measurements (with standard deviations in parentheses), and differences between stressed and unstressed for real word items in the lexical dataset. Dist and displ in mm; vel in mm/s. Starred measures were analyzed with a 2-way rather than a 3-way ANOVA (omitting the Syllable factor). None of the differences shown in this table represent statistically significant effects of Stress

	Unstressed	*Stressed*	*Diff.*	*Stress effect*	*p-value*
HEAD DISPL	.83 (.75)	1.23 (1.21)	.396	$F(1, 33) = 1.31$	$p = .261$
LIP DIST	28.47 (4.92)	29.56 (5.08)	1.090	$F(1, 35) = 3.17$	$p = .084$
LIP OPEN DISPL	9.19 (4.56)	9.61 (5.42)	.427	$F(1, 31) = 1.04$	$p = .315$
LIP CLOS DISPL*	6.43 (4.45)	7.21 (4.44)	.803	$F(1, 27) = 0.27$	$p = .606$
CHIN OPEN DISPL	5.86 (2.91)	7.00 (3.29)	1.138	$F(1, 35) = 2.18$	$p = .149$
CHIN CLOS DISPL*	5.21 (3.17)	5.99 (2.87)	.782	$F(1, 26) = 0.30$	$p = .591$
LIP OPEN VEL	.56 (.25)	.55 (.30)	–.002	$F(1, 33) = 0.02$	$p = .901$
LIP CLOS VEL*	.47 (.34)	.49 (.33)	.012	$F(1, 27) = 0.06$	$p = .806$
CHIN OPEN VEL	.39 (.16)	.48 (.22)	.090	$F(1, 36) = 2.82$	$p = .102$
CHIN CLOS VEL*	.35 (.23)	.43 (.26)	.077	$F(1, 20) = 0.65$	$p = .429$

Table 4

Mean stressed and unstressed measurements (with standard deviations in parentheses), and differences between stressed and unstressed for real word items in the lexical dataset plus the other produced repetition of each word. Dist and displ in mm; vel in mm/s. Significant effects for Stress are shown in bold; starred measures were analyzed with a 2-way rather than a 3-way ANOVA (omitting the Syllable factor)

	Unstressed	*Stressed*	*Diff.*	*Stress effect*	*p-value*
LIP DIST	27.88 (4.63)	29.28 (5.29)	**1.40**	$F(1, 78) = 7.75$	$\mathbf{p = .007}$
LIP OPEN DISPL	8.71 (4.68)	9.50 (5.44)	.80	$F(1, 71) = 2.57$	$p = .113$
LIP CLOS DISPL*	6.06 (4.32)	7.26 (4.66)	1.21	$F(1, 54) = 1.42$	$p = .239$
CHIN OPEN DISPL	5.71 (3.03)	6.86 (3.60)	**1.15**	$F(1, 79) = 3.75$	$\mathbf{p = .056}$
CHIN CLOS DISPL*	5.18 (2.94)	6.34 (3.48)	1.16	$F(1, 56) = 1.28$	$p = .263$
LIP OPEN VEL	.53 (.26)	.58 (.31)	.05	$F(1, 74) = 0.61$	$p = .439$
LIP CLOS VEL*	.48 (.33)	.52 (.35)	.03	$F(1, 53) = 0.33$	$p = .566$
CHIN OPEN VEL	.39 (.21)	.48 (.23)	**.09**	$F(1, 80) = 5.31$	$\mathbf{p = .024}$
CHIN CLOS VEL*	.39 (.23)	.45 (.28)	.06	$F(1, 45) = 0.84$	$p = .365$

mean movements in stressed syllables. The analyses of these data included factors of Reiterant Syllable (*buh* or *fer*) and Pair Type (minimal or non-minimal pair), in addition to Stress, Syllable, and Talker. In these reiterant words, where variation introduced by the segmental differences between stressed and unstressed syllables in real words was experimentally controlled, the effects of Stress were statistically robust. All ten of these measures were found to reliably distinguish stress, as shown in Table 5.

Table 5

Mean stressed and unstressed measurements (with standard deviations in parentheses), and differences between stressed and unstressed for all reiterant items in the lexical dataset. Dist and displ in mm; vel in mm/s. Significant effects for Stress are shown in bold; starred measures were analyzed with a 4-way rather than a 5-way ANOVA (omitting the Syllable factor)

	Unstressed	*Stressed*	*Diff.*	*Stress effect*	*p-value*
HEAD DISPL	.56 (.44)	.85 (.55)	**.29**	$F(1, 118) = 13.17$	$\mathbf{p < .001}$
LIP DIST	26.49 (4.57)	28.17 (4.83)	**1.68**	$F(1, 135) = 69.77$	$\mathbf{p < .001}$
LIP OPEN DISPL	9.99 (4.32)	11.65 (4.58)	**1.66**	$F(1, 135) = 38.82$	$\mathbf{p < .001}$
LIP CLOS DISPL*	9.53 (4.50)	10.70 (4.54)	**1.17**	$F(1, 68) = 13.72$	$\mathbf{p < .001}$
CHIN OPEN DISPL	7.11 (3.35)	8.05 (2.93)	**.94**	$F(1, 134) = 18.93$	$\mathbf{p < .001}$
CHIN CLOS DISPL*	6.41 (2.70)	7.36 (2.81)	**.95**	$F(1, 67) = 8.49$	$\mathbf{p = .005}$
LIP OPEN VEL	.75 (.33)	.86 (.35)	**.11**	$F(1, 136) = 36.07$	$\mathbf{p < .001}$
LIP CLOS VEL*	.77 (.38)	.84 (.38)	**.07**	$F(1, 68) = 7.94$	$\mathbf{p = .006}$
CHIN OPEN VEL	.56 (.27)	.61 (.24)	**.06**	$F(1, 134) = 12.40$	$\mathbf{p < .001}$
CHIN CLOS VEL*	.52 (.23)	.58 (.25)	**.06**	$F(1, 67) = 5.21$	$\mathbf{p = .026}$

The effect of Reiterant syllable was always significant as well. As might be expected, given the respective articulatory properties of [ʌ] and [ɝ], *buh* syllables showed bigger and faster movements than *fer* syllables, except with respect to head movements, which showed the reverse. However, there was no Reiterant Syllable by Stress interaction, indicating that *buh* measures, despite being bigger overall, did not show larger stress differences than *fer* measures. The effect of Pair Type was also significant for six of the measures (with larger or faster movements in non-minimal pairs), but there were no Pair Type by Stress interactions. Finally, there was a main effect of Talker for all measures except Head Displacement, which will be discussed in the local discussion.

3.2.3 Summary of lexical stress effects

To summarize, the eyebrows did not show any movement during word production. Six of the remaining 10 measures (all but the closing gestures) showed significant effects of Stress for the set of all lexical items considered together, and similar patterns were seen in separate analyses of both the real and reiterant word subsets of the corpus. However, while stress was differentiated by all of the measures examined in reiterant words, the same pattern of differentiation was only statistically reliable for Lip Distance and Chin Opening Velocity, and marginally reliable for Chin Opening Displacement, in real words.

In terms of effect size (expressed as partial eta-squared, which describes the proportion of total variability attributable to a factor), Lip Distance had the largest Stress effect in the set of all lexical items, with eta-squared of .085. Chin and Lip Displacement effects, with eta-squareds of .032 and .040, respectively, were smaller

than Lip Distance and also than Head Displacement, with an eta-squared of .063, but slightly larger than other measures, which had eta-squareds of .005–.024. Thus, displacement effects were greater than velocity effects, and Lip Distance and Head Displacement showed the largest effects overall. Effect sizes in the real word subset were comparable to those for the set of all lexical items, with eta-squareds of .093 and .045 for Lip Distance and Chin Displacement; however, effect sizes could only be considered for the three measures with significant Stress effects. Effect sizes in the reiterant subset were substantially higher, with eta-squareds of .341 for Lip Distance and .223 for Lip Displacement, and effect sizes greater than or equal to .1 for all other measures except Chin Velocity; eta-squared for Head Displacement was .1.

In terms of the numerical magnitude of the gestures differentiating stressed from unstressed syllables, movements of the lips were largest (up to almost 3 cm Lip Distance in stressed syllables, with about 1 cm of Lip Opening Displacement) and showed the greatest differentiation by stress condition, as expected (about 1.5 mm in the set of all lexical items). Chin movements were slightly smaller (about 5–8 mm) and showed a stress difference of about 1 mm. Head movements were quite small (about 1 mm in stressed syllables), with very small stress differences (less than .5 mm, despite their slightly larger statistical effect size). The movement and difference magnitudes were nearly the same for all of the lexical item subsets, except that Lip Opening Displacement (especially in stressed syllables) was reduced in the real words.

These effects in production can be used to make specific predictions of success in perception. For example, since there are more measures of facial movements distinguishing stressed from unstressed syllables for reiterant than real words, it can be predicted that stress in reiterant words will be better perceived than stress in real words. Additionally, it might be assumed that the measures that best distinguish stress in the real words as well as the reiterant ones, namely Lip Distance and the two Chin Opening measures, are relatively more important for perception—especially if real word stress perception is successful. Following similar logic for individual measures, it can be predicted that opening measures, which are more often available and more often significant, will be more helpful in perception. Because the head moves quite reliably with stress and because it is large and easy to see, its movements may be predicted to be perceptually helpful. On the other hand, because head movements are so small, they may not be salient to perceivers despite their consistency. And brows, because they did not move with lexical stress, are not expected to play a role in perception.

3.3
Results for phrasal stress

3.3.1
All measures

Recall that phrasal stress contrasts were produced as narrow focus on one of three names in a sentence (e.g., focus on *Tommy*, *Debby*, or *Timmy* in "*So Tommy gave Debby a song from Timmy*") or as broad focus, that is, "neutral." Measurements were made for the lexically stressed (initial) syllables of all test words (i.e., the names). Measurements were then analyzed separately for each production measure, again using

Table 6

Mean stressed and unstressed measurements (with standard deviations in parentheses), and differences between stressed and unstressed for all items in the phrasal dataset. Dist and displ in mm; vel in mm/s. Significant effects for Stress are shown in bold type

	Unstressed	*Stressed*	*Diff.*	*Stress effect*	*p-value*
BROW DISPL	0 (0)	.51 (.41)	**.51**	$F(2, 162) = 141.42$	$\mathbf{p < .001}$
HEAD DISPL	.96 (.48)	2.21 (1.56)	**1.25**	$F(2, 116) = 35.09$	$\mathbf{p < .001}$
LIP DIST	30.08 (5.69)	33.06 (5.77)	**2.98**	$F(2, 162) = 14.73$	$\mathbf{p < .001}$
LIP OPEN DISPL	12.06 (4.32)	15.70 (4.66)	**3.64**	$F(2, 161) = 18.12$	$\mathbf{p < .001}$
LIP CLOS DISPL	12.85 (3.71)	16.14 (3.55)	**3.29**	$F(2, 162) = 17.86$	$\mathbf{p < .001}$
CHIN OPEN DISPL	9.31 (3.10)	12.00 (3.23)	**2.69**	$F(2, 162) = 14.74$	$\mathbf{p < .001}$
CHIN CLOS DISPL	8.13 (2.83)	10.82 (3.10)	**2.69**	$F(2, 162) = 15.96$	$\mathbf{p < .001}$
LIP OPEN VEL	.85 (.31)	1.06 (.37)	**.21**	$F(2, 161) = 14.69$	$\mathbf{p < .001}$
LIP CLOS VEL	1.04 (.31)	1.32 (.30)	**.28**	$F(2, 162) = 17.60$	$\mathbf{p < .001}$
CHIN OPEN VEL	.67 (.23)	.85 (.25)	**.18**	$F(2, 162) = 13.43$	$\mathbf{p < .001}$
CHIN CLOS VEL	.66 (.25)	.88 (.26)	**.22**	$F(2, 162) = 14.67$	$\mathbf{p < .001}$

factorial ANOVAs. Analyses included factors of phrasal Stress (stressed, unstressed, or no stress), Position in Sentence (1st, 2nd, or 3rd name), and Talker (T-LO, T-MID, or T-HI). In addition, an Initial Consonant factor (alveolar or labial), referring to the place of articulation of the initial consonant in the test names, was included. The alpha level for significance was $p < .05$.

Because some sentences were produced with a stress pattern that differed from the pattern that was scripted, separate analyses were initially performed using the designation of stress as it was scripted and as it was actually produced (as auditorily transcribed by two linguist listeners). Because the differences between scripted and produced stress were slight and turned out to make no difference in the statistical results, only the results using the scripted stress designations are reported.

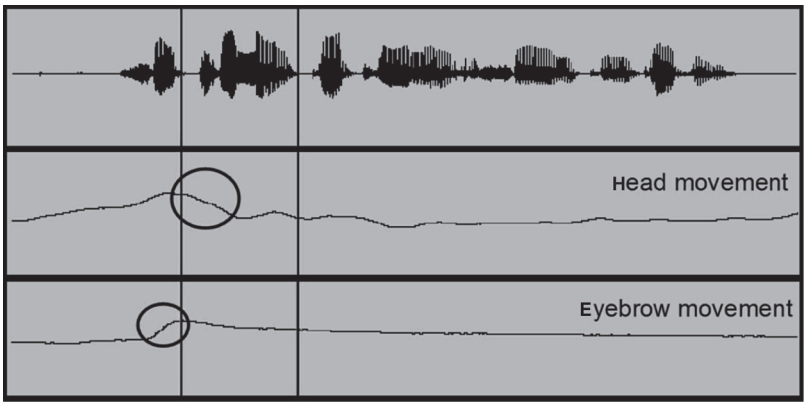
In these analyses, all 11 measures showed significant main effects of Stress, with larger or faster movements in the stressed condition, as shown in Table 6. On all measures other than Brow and Head Displacement, Talker, and Initial Consonant were also significant, with larger, faster movements for the labial names than for the alveolar ones. On two of the measures (Lip Distance and Chin Opening Velocity), Position-in-Sentence was significant, with larger or faster movements in first position than in final position. All significant interactions involved Talker; that is, talkers differed in exactly where they made significant stress differences. Talker effects and interactions will be discussed below.

3.3.2 Head and eyebrows

Although head and eyebrow movements are only two among 11 measures that reliably marked phrasal stress, they differ from the others in that they are not directly associated

Figure 4

Head and eyebrow movements under phrasal stress in *So TIMMY gave Tommy a song from Debby*. The focused word *TIMMY* is shown between the two vertical lines



with segmental articulation. Therefore, their movements could be purely prosodic, free to cue stress without being constrained by segmental context. As noted above, the typical movement of the head associated with stress was in a downward direction, while brow movements were typically upward. From our qualitative observation of the tokens, these movements were coordinated such that the eyebrows (or at least one of them) rose just before the stressed word and fell slowly during the word, and the head fell, or nodded, beginning during the closure of the initial stop in stressed words, as illustrated in Figure 4. Note that although head and brow movements spread over a domain slightly larger than the stressed syllable, their extrema tended to fall within the stressed syllable, as was measured. Still, the fact that these gestures are temporally extended might contribute to their perceptibility.

As mentioned in the introduction, however, it has been suggested that the relation between eyebrow movement and prosodic marking might be even stronger than simply indicating the position of stress, in that eyebrow movement could directly reflect F0 (Cavé et al., 1996). Given that no eyebrow movements were found in lexical stress items (which were pitch accented), this hypothesis seems unlikely, but it is worth explicit investigation. To this end, 40 sentences were randomly chosen from the phrasal stress corpus. Brow Displacement (both left, as reported elsewhere in this article, and right brows) and F0 (extracted from the audio signal) were sampled from these utterances at 12 ms intervals. Significant correlations were found between Brow Displacement and F0, but these accounted for only 1–4% of variance. Thus it seems unlikely that much if any direct information about F0 could be available from brow movements.

3.3.3 Summary of phrasal stress effects

In summary, all facial movement measures reliably distinguished stressed from unstressed words. Note that these measures even included Brow Displacement, which

was not found to contribute to lexical stress production. Furthermore, these measures were significant despite the fact that the sentence items involved real words (albeit a relatively limited group of six names) rather than reiterant syllables.

In terms of effect size (expressed as partial eta-squared), Brow and Head Displacement showed the largest stress effects, with eta-squared values of .636 and .377, respectively. The other measures were not distinguished much in terms of their effect sizes, with eta-squareds between .142 and .184 (except Chin Closing Displacement, at .021), though the two Lip Displacement measures were at the top of that range, with eta-squareds over .180. All of these phrasal effects (except Chin Closing Displacement) were much larger than the effects seen in the lexical stress corpus as a whole, but they are comparable to (or even slightly smaller than) the effect sizes seen for the reiterant word subset of the corpus.

As was the case with lexical stress, the magnitudes of the movements of the lips with phrasal stress were also largest (over 3 cm for Lip Distance) and showed the greatest stress-conditioned differences (about 3 mm). The chin showed slightly smaller movements than the lips, but stress differences of nearly the same magnitude (2.7 mm). Note that these stress-conditioned differences were about two times or more the magnitude of those seen in the lexical corpus. Both head and eyebrow movements were much smaller than the lip and chin movements (.5 and 2.2 mm, respectively, in stressed syllables), but the differences across stress conditions were somewhat larger relative to the displacements and were large relative to the lexical stress differences for these measures: the difference for the head was 1.3 mm (or four times the lexical difference), and the difference for the eyebrow was .5mm (as opposed to no movement at all in the isolated words).

Because all of the measures examined showed highly significant effects of Stress, it may be predicted that phrasal stress will be well perceived visually, even if only some of the measures are ultimately important for perception. Because there are more differentiating measures for phrasal than for lexical stress, phrasal stress may be easier to perceive than lexical stress.

3.4 Discussion

3.4.1 Talker differences

As mentioned in the sections above, Talker was a significant factor in a majority of the lexical and phrasal analyses performed. When talkers differed (except for with respect to Head Displacement), T-MID showed larger, faster movements than one or both of the other talkers. In other words, T-MID's articulations were either the biggest and fastest or were tied as the biggest and fastest. T-HI consistently showed the smallest, slowest movements (either by himself or tied) for all of the lexical analyses (again, except for head) and for six of the sentence analyses. A main effect of Talker, however, simply reflects that one talker moved more than the others overall, perhaps simply reflecting that one talker (T-MID) was bigger than the others.

It is Talker by Stress interactions that indicate differences among talkers with respect to their ability to mark differences between stressed and unstressed

Table 7

Significant Talker by Stress interactions for all lexical and phrasal stress items

	Lexical	*Phrasal*
BROW DISPL	–	–
HEAD DISPL	T-HI > T-MID, T-LO	T-HI > T-MID > T-LO
LIP DIST	–	–
LIP OPEN DISPL	–	T-LO, T-HI > T-MID
LIP CLOS DISPL	–	T-LO, T-HI > T-MID
CHIN OPEN DISPL	–	–
CHIN CLOS DISPL	–	T-LO > T-MID, T-HI
LIP OPEN VEL	–	T-LO > T-MID, T-HI
LIP CLOS VEL	–	–
CHIN OPEN VEL	–	–
CHIN CLOS VEL	–	–

conditions. Among lexical measures, Head Displacement uniquely produced a significant interaction. Pairwise post-hoc comparisons indicated that T-HI differentiated stress conditions with more extreme head movements than the other talkers. There were more talker differences for phrasal stress marking: again, T-HI produced the largest differences in head movement, as well as larger Lip Distance differences (opening and closing) than T-MID; T-LO also marked Lip Distance more extremely than T-MID and made larger Chin Closing Displacement and Lip Opening Velocity differences than the other talkers. These interactions are summarized in Table 7.

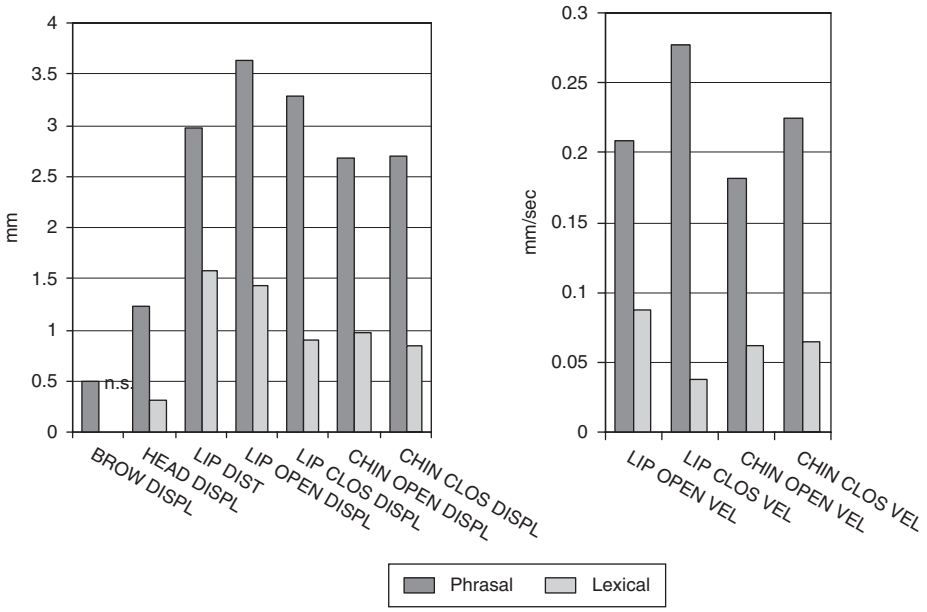
Summarizing, the largest phrasal stress differences were due to T-LO, followed by T-HI who made the largest Head Displacement differences, and was the only talker to stand out with respect to lexical stress for any measure. Individual talker differences were greater for phrasal than for lexical stress. Based on these results, individual talker effects would be predicted for perception of phrasal stress but not for lexical stress, unless head movement is perceptually salient. T-LO is predicted to be most intelligible and T-MID least intelligible, for phrasal stress.

3.4.2 Best overall measures

In the results above, we saw that both lexical and phrasal stress were well-marked by the obtained measures. Phrasal stress was reliably differentiated by all of the measures. Lexical stress expressed using reiterant speech resulted in significant differences on 10 of the 11 measures (i.e., all except for Brow Displacement, as brows showed no movement in lexical stress items). These stress effects are summarized in Figure 5 below. In general, the production of lexical and phrasal stress involved the same measures, but larger differences marked stressed versus unstressed syllables that identified phrasal stress as opposed to lexical. When just real words were considered, only three of the measures (Lip Distance, Chin Opening Displacement, and Velocity) showed significant or nearly significant effects for lexical items (although phrasal stress, which also involves only real words, was differentiated by all the measures.

Figure 5

Difference between all stressed and all unstressed measurements for lexical (all items) and phrasal stress datasets. Only the Brow Displacement difference for lexical stress was *not* statistically significant



(Although the difference between lexical and phrasal stress was large, it may be noted that this difference might be due in part to the fact that the script for phrasal stress marked the stressed word explicitly with capital letters, while the script for lexical stress was not explicitly contrastive for stress, eliciting the correct stress pattern with a disambiguating phrase.)

Physically, the largest differences between stressed and unstressed items were found in the Lip Distance and Lip Opening and Closing Displacement for both lexical and phrasal stress. Chin Opening and Closing Displacement differences for phrasal stress were nearly as large. In terms of the statistical effect size for stress (i.e., the proportion of variability attributable to stress), however, recall that there were relatively few differences among measures within each corpus (though Lip Distance did stand out somewhat in the word corpus, and Head and Brow Displacement stood out in the sentence corpus, with Lip Displacement effects slightly greater than those for the other measures). These statistics also indicated that phrasal effects were indeed much larger than lexical effects (at least for real word speech).

The results with reiterant speech and phrasal stress showed that all our measures other than Brow Displacement mark stress. The only measures that remain significant or marginally significant as markers in real word lexical stress, however, are Lip Distance and Chin Opening Displacement and Velocity. These three measures, along with Lip and Chin Closing Velocities, are also most consistent across talkers, as these

measures showed no interaction of Stress by Talker. These measures, then, may be most important in allowing perceivers to perceive stress visually across words and talkers. The next experiment tests visual stress perception.

4 Perception experiment

4.1 Methods

4.1.1 Stimuli

Stimuli consisted of the single token of each test item that was analyzed in the main part of the production study. Thus, there were 120 word tokens and 72 sentence tokens (192 stimuli).

4.1.2 Participants

Sixteen paid volunteers aged 18–40 years (mean age 26.2 years) with no self-reported learning disabilities, vision of 20/30 or better in each eye as determined with a standard Snellen chart, and self-reported use of English as the native language participated in the experiment. To assess the segmental lipreading ability of these participants, they were administered a lipreading screening test. Scores ranged from 2–49% words correct. No participant was excluded on the basis of the screening score.

4.1.3 Procedure

Participants were tested individually in a sound-treated booth, seated about .5 m from a 14-inch color monitor and a 14-inch Sony Trinitron video display. Video was presented off a Betacam UVW-18000 industrial video recorder/player that was under computer control by purpose-written software. Video was in its native recording format, played at a frame rate of 29.9 7Hz.

The word stimuli were presented first, followed by the sentence stimuli. Among the words, real word stimuli were presented first, followed by [bʌ] reiterant stimuli, and then [fɜ˞] reiterant stimuli, with minimal and non-minimal pairs randomized within the reiterant blocks. Within each block (real-speech words, [bʌ] reiterant words, [fɜ˞] reiterant words, and sentences), each stimulus was presented twice. Stimuli were pseudo-randomized such that (a) a given stimulus (i.e., a given word or sentence with a specific stress pattern, regardless of the talker who produced it) never appeared sequentially and (b) no more than three consecutive stimuli were produced by the same talker. Two such pseudo-randomizations were generated, and participants were assigned to them randomly.

A word trial began with the presentation of two stress-contrasted alternatives (e.g., PERmit and perMIT), presented side by side on the left computer monitor. The stressed syllable was written in capital letters. The stress-initial alternative was always to the left of the stress-final alternative. After 2 seconds, a video-taped test word was presented on the right display. The image of the face took up the full height of the display (8.4 inches); thus, the image the participants saw was approximately 90% of life-size.

At the end of the clip, participants clicked on one of the two printed alternatives. Their response triggered the next trial. In the reiterant speech conditions, the alternative responses were the real words that the talkers had mimicked during the recording, and participants were asked to choose the word that the talker had mimicked. The sentence stimuli were presented using a similar format. First, a full sentence, in orthography, appeared on the left monitor, with the three names in red capital letters. Each word was clickable. One centimeter to the right of the transcribed sentence was a fourth clickable alternative, also in red letters, reading "No Stress." After 2 seconds, a video-taped test sentence matching the printed transcription was presented on the right display. At the end of the clip, participants indicated which word they thought the talker had stressed or emphasized by clicking on the corresponding name. If they thought that none of the three names had received stress, they were to click the "No Stress" choice.

4.2
Results

The lexical dataset is not a balanced design in that reiterant items comprise both minimal and non-minimal pairs (as well as *buh* and *fer* syllable types), while real word items comprise only minimal pairs (with no reiterant syllable type). Therefore, the lexical type (real or reiterant) and the pair type (minimal or non-minimal) cannot both be considered in the same analysis. Because the production results predict possible differences in the intelligibility of real and reiterant items, the lexical type condition is included in the omnibus analysis, and the pair type condition is put aside. Pair type (along with reiterant syllable type) is investigated in a separate analysis of just the reiterant items.

Responses were scored as correct or incorrect based on the intended, scripted stress, even though talkers did not always produce their utterances exactly as scripted, inasmuch as the net effect of these deviant stresses on perception (as on the overall production patterns) was very small.[8]

4.2.1
Lexical stress effects

Perceivers correctly identified the stress in 62.2% of words (for which the chance level was 50%). A one-sample test for above chance identification indicated that the visual perception of lexical stress significantly exceeded chance, $t(15) = 7.56$, $p < .001$.

A repeated measures ANOVA was performed on percent correct responses for lexical items. The within-subjects factors were Lexical type (real or reiterant), Syllable (1st or 2nd), and Talker (T-LO, T-MID, or T-HI). (Minimal vs. non-minimal and *buh*

[8] Of the 72 items used in the perception experiment, 20 were produced in one of four less-than-ideal ways. First, 5 items that were scripted as NoStress were produced with a weak accent on one name (the final name in 4/5 cases). The perception of these items shows no pattern; like NoStress items generally, they elicited a variety of responses, which were not related to the location of the unintended weak accent. Second, one item was produced with only a weak accent on the correct name. Third, 9 items were produced with the correct accent, but with another, weak, accent on another name. These last two categories of errors accounted for half of the problem items, but all maintained the strongest stress on the scripted name, and they were generally perceived as intended. Finally, 5 items had two equal stresses, one on the scripted name and one on another name. Two of these 5 items elicited responses that could be attributed to the extra stress.

Figure 6

Percent correct on lexical stress items, by talker

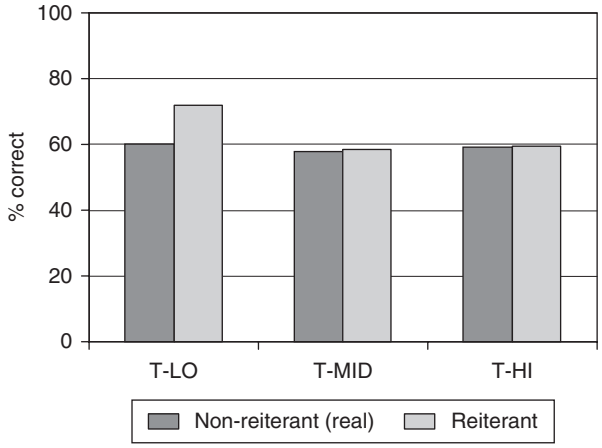
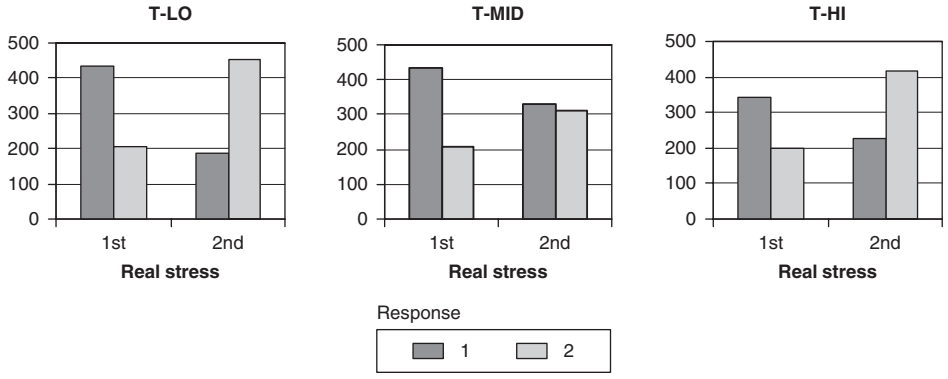


Figure 7

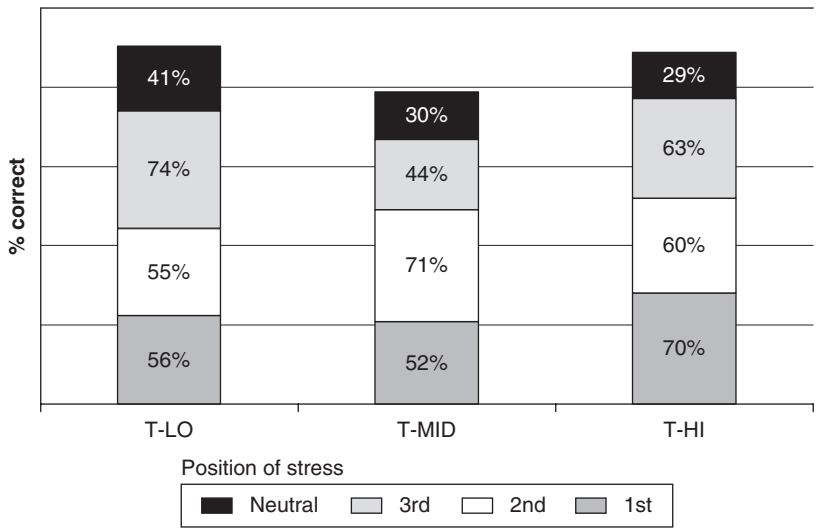
Distribution of listeners' responses (syllable 1 or syllable 2), correct or not, by real position of lexical stress (1st or 2nd syllable) and by talker



vs. *fer* distinctions were collapsed in the analysis.) The analysis revealed a main effect of Talker, $F(2, 15) = 8.66$, $p = .001$. Pairwise post-hoc comparisons indicated that items produced by T-LO were better perceived (65% correct) than those by the other talkers (55% and 58% correct for T-MID and T-HI, respectively). However, Talker interacted with Lexical type, $F(2, 15) = 4.61$, $p = .018$. Pairwise post-hoc comparisons indicated that this interaction was due to better performance on T-LO's productions of reiterant items. There were no significant differences in performance between the other two talkers' items or between reiterant and real items produced by the other talkers. These effects are summarized in Figure 6. None of the other effects or interactions were significant.

Figure 8

Correct perception by talker and position of phrasal stress (1st, 2nd, or 3rd word, or neutral NoStress). Talker differences (T-MID<T-LO, T-HI) and Position differences (Neutral < 1st, 2nd, 3rd) are significant



To specifically examine the results within the reiterant condition, an additional repeated measures ANOVA was performed on percent correct responses to reiterant lexical items with within-subjects factors of Reiterant-type (*buh* or *fer*) and Pair type (minimal or non-minimal), as well as Syllable (1st or 2nd), and Talker (T-LO, T-MID, or T-HI). As expected, there was a main effect of Talker, $F(2, 15) = 35.73$, $p < .001$. Post-hoc analyses showed T-LO's (reiterant) items to be better perceived, as in the previous analysis (75% correct for T-LO versus 59% and 60% correct for T-MID and T-HI, respectively). However, there were no significant effects of either Reiterant type or Pair type and no interactions, suggesting that *buh* and *fer* pairs and minimal and non-minimal pairs were similarly well perceived.

Independent of correctness, responses in the lexical stress task (without respect to whether they were right or wrong), were evenly divided between the two syllables (50.7% first syllable, 49.3% second syllable), χ^2 (1, $N = 3840$) = .76, $p = .384$. Responses broken down by stressed syllable and talker are summarized in Figure 7.

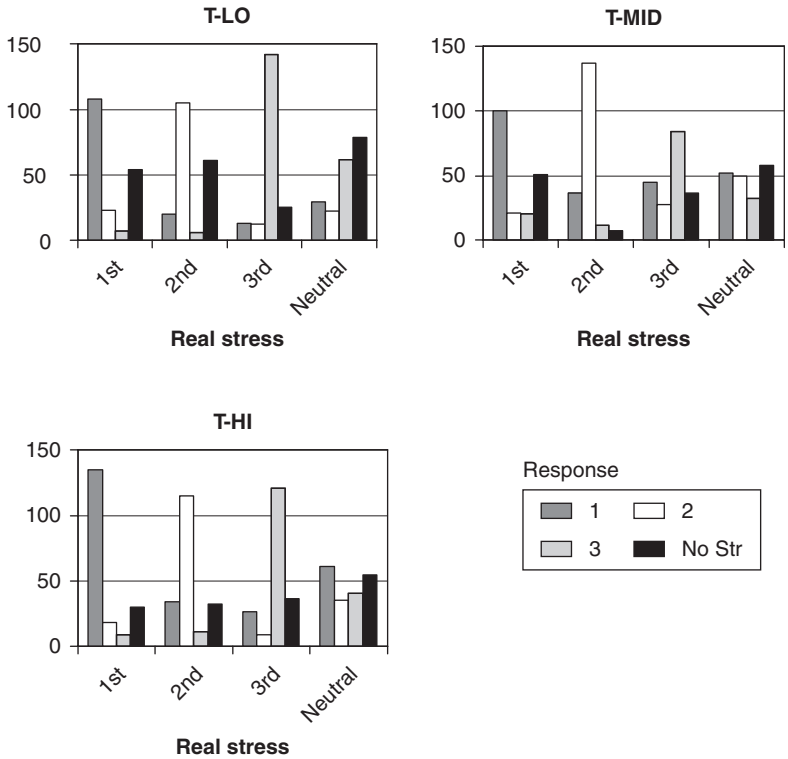
4.2.2 Phrasal stress effects

Perceivers correctly identified the (scripted) stress in 53.8% of sentences (for which chance was 25%). A one-sample test for above chance identification indicated that the visual perception of phrasal stress significantly exceeded chance, $t(15) = 12.20$, $p < .001$.

A repeated measures ANOVA with within-subject factors of Position of Stress (1st, 2nd, 3rd, or NoStress) and Talker (T-LO, T-MID, or T-HI) was performed on percent correct responses for phrasal items. The analysis revealed a main effect of Position of Stress, $F(3,15) = 8.96$, $p = .002$. Post-hoc comparisons showed that the NoStress

Figure 9

Distribution of listeners' responses (syllable 1, 2, 3, or "No Stress"), correct or not, by real position of phrasal stress (1st, 2nd, or 3rd word, or neutral/NoStress)



(neutral or broad focus) stress condition was less well perceived (33.3% correct overall) than any of the other three stress conditions (59.5–60.2% correct), but there were no differences in perceiver success among the narrow focus stress conditions (1st, 2nd, or 3rd position stress) (cf., Dohen, et al., 2004a, 2004c, who found first position stress to be perceived best).[9]

The analysis also revealed a significant effect of Talker, $F(2,15) = 5.04$, $p = .013$. Although the mean differences in perception between the talkers were small (ranging from 49–56% correct), pairwise post-hoc comparisons indicated that T-MID's phrasal stress was slightly less-well perceived than that of the other two talkers. The positional and talker effects are summarized in Figure 8.

[9] Note that when any narrow focus stress was incorrectly judged, the response was most often NoStress (57.9%, 45.7%, and 42.4% for scripted first, second, and third position stress, respectively). When NoStress was inaccurately judged, the response was either initial (37%) or final (35.2%) stress. Thus, not only was NoStress the least-frequent right answer, it was also the most frequent wrong answer. Perceivers were apparently unsure about when to use that response. Furthermore, the pattern of these errors revealed no particular tendency by perceivers to hear either earlier or later stress than intended.

Responses in the phrasal stress task (independent of correctness) revealed a perceiver preference for selecting the first word, but were otherwise fairly evenly divided (28.6% first word vs. 24.9%, 23.7%, and 22.7% for second, third, and NoStress, respectively), χ^2 (1, $N = 2304$) = 18.51, $p < .001$. (While Granström et al. 1999 and Jensen 2003 found that listeners tend to perceive last words as well as first words as prominent, it can be seen that last (third) words had no priority in our data.) Responses broken down by position of stress and talker are summarized in Figure 9.

4.2.3
Analysis of items

Perceptual success can also be analyzed for individual items. Exact binomial tests (based on 32 independent trials per item) were used to determine the numbers of correct responses for the lexical and phrasal stress items that indicated significant deviation from chance. Items were then divided into groups of those perceived better than chance, at chance, or worse than chance. Of the 120 lexical items, 45 were perceived better than chance, and 7 were perceived worse than chance. Of the 72 phrasal items, 47 were perceived above chance, and none were perceive below chance.

Most of the lexical items perceived above chance were reiterant as opposed to real words. In fact all but 7 of the successfully perceived words were reiterant, and 25 of the 38 successful reiterant items were *buh* as opposed to *fer* items. Recall that in the production experiment, stress in reiterant items was better differentiated than in real items. Thus, the relative success of reiterant items in perception is consistent with the production results. Also, while *buh* items showed bigger movements than *fer* items, the bigger movements did not lead to significantly greater differences across stress conditions. However, the greater perceptual accuracy for *buh* items could be attributed to their overall larger movements, relative to *fer*.

By talker, T-LO produced 26 lexical items that were perceived reliably above chance and 2 items perceived below chance. T-MID produced 10 items that were perceived above chance and 3 items perceived below chance. T-HI produced 9 items that were perceived above chance and 2 items perceived below chance. Thus T-MID and T-HI were similarly effective overall (although on different items), while T-LO was better at conveying lexical stress by this measure as well.

For phrasal stress, no items were perceived reliably below chance. 18 of T-LO's 24 sentences were perceived better than chance, as were 15 of T-HI's and 14 of T-MID's. Again, T-LO was the easiest talker to read, though talkers were perceived similarly-well overall. Except that NoStress sentences were rarely perceived above chance, T-MID's and T-LO's above-chance items were fairly evenly divided across positions. Almost all errors, however, were on the names containing high vowels (*Mimi* and *Timmy*). T-HI's stress was better perceived in the first and second positions than in the third, but errors were still on the high vowel names.

4.3
Discussion

Stress was well-perceived overall. As noted above, both lexical and phrasal stress were perceived above chance, though in fact, phrasal stress was even better perceived than lexical stress. When perceivers' scores were subtracted from the above-chance

levels (determined by exact binomial tests), a paired *t*-test comparing these differences in the lexical vs. phrasal conditions showed a significant difference in performance, $t(15) = 9.75$, $p < .001$. This result is consistent with the greater number of differentiating measures for phrasal than for lexical stress in the production study. Note, however, that perceivers' success with lexical stress perception was better than predicted by the production patterns. Despite a greater number of differentiating measures for stress in the production of reiterant than real words, reiterant words were no better perceived. This suggests that it is the measures that are marked in real as well as reiterant words, namely Lip Distance and Chin Opening measures, that may be most important for the perception of lexical stress. So while it is possible that the better perceptibility of phrasal stress over lexical stress is attributable to more and larger production differences, similar perceptual success among the subgroups of lexical items indicates that more and larger differences alone do not ensure better perception.[10]

Talker differences in perception were relatively minor. T-LO, though, stood out for lexical stress, and T-LO and T-HI were both more successful at conveying phrasal stress information than T-MID. These findings are consistent with the production results showing that T-LO (along with T-HI) produced better stress differences on several measures (see Table 7). However, these stronger differences were seen only in sentences, whereas a perceptual advantage was found for T-LO in (reiterant) words as well.

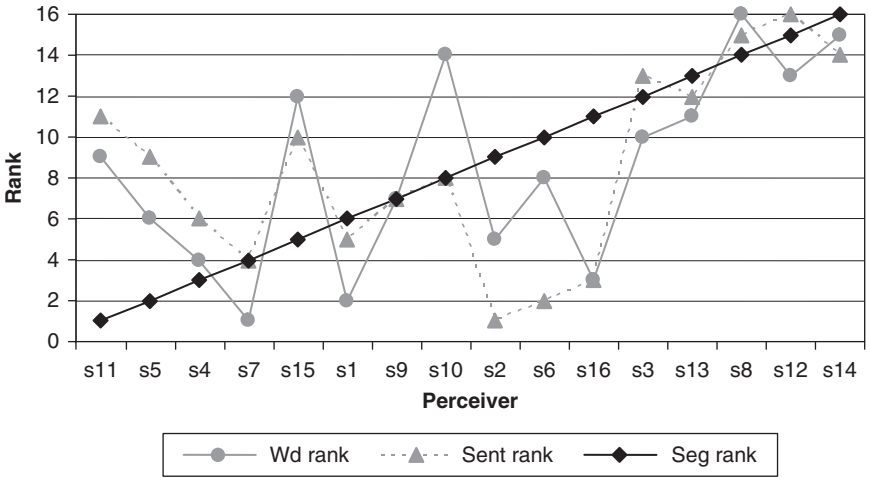
Recall that talkers were chosen across a range of segmental intelligibility, reflected in their names, LO, MID, HI, in order to be able to relate segmental and prosodic intelligibility. While there was some range of prosodic intelligibility across talkers, the segmental intelligibility levels did not seem highly related to the prosodic perceptual results.

A correlation between lexical and phrasal stress scores for individual perceivers indicated that those perceivers who were more successful with one type of stress were also more successful with the other: adjusted $R^2 = .45$; $F(1, 15) = 13.13$, $p = .002$. However, when lexical and phrasal stress success rates were compared with perceivers' success in the segmental lipreading task, there was no such general relation, as can been seen in Figure 10. The five most accurate segmental perceivers, nonetheless, were also the five most accurate phrasal stress perceivers and five of the top seven perceivers of lexical stress. This result makes sense if the ability to perceive prosody aids in the perception of a sentence as a whole, since here "segmental" perception in fact refers to words correct in a coherent grammatical sentence. Prosodic perception ability is not sufficient, though, for good segmental perception, as the two worst segmental perceivers have fairly strong stress perception ability. Conversely, strong segmental perception is insufficient for good prosodic perception as well, as the perceivers ranking just below the top five have relatively good segmental perception

[10] Another possible source for the difference in perception between lexical and phrasal stress is the difference in the domain over which the cues are produced. For example, Dohen et al. 2006 have shown that the entire focused word (both stressed and unstressed syllables) contains phrasal stress cues, whereas lexical stress cues are necessarily constrained to the stressed syllable. Although we only analyzed the stressed syllable (which was the pitch accented syllable) for both lexical and phrasal stress items, both syllables may have contributed to perceivers' success with phrasal stress.

Figure 10

A comparison of individual perceivers' rankings relative to one another in a preliminary lip reading screening (of segmental, not prosodic materials), and in the lexical and phrasal stress perception tasks. Subjects are ranked from 1 to 16, worst to best



but poor prosodic perception. High visual perception ability in both domains, then, may be driven by some other general language skill.

5 Relation between production and perception

The main result of the production experiment was that most of the measured dimensions varied according to stress, both lexical and phrasal, while the main result of the perception study was that stress was reliably perceived above chance. We now ask whether some production dimensions contributed more than others to perception, using correlation analysis and focusing on the phrasal stress items, since phrasal stress was reliably differentiated on more production dimensions and was better perceived than lexical stress.

Correlations were computed between the production measures and the percent correct scores for both stressed and unstressed syllables. First, simple correlations were performed with individual production measures. Then partial correlations were performed to look for measures with strong unique contributions to perception. Finally, multiple regression was used to examine combined contributions of various measures. The response for a stressed or unstressed syllable was considered to be correct if the response for the whole sentence in which it was embedded corresponded to the scripted stress. By examining stressed and unstressed syllables separately, we considered that information about the location of phrasal stress might be contained in both the stressed and the unstressed parts of a sentence and allowed for the possibility that stress and the lack of stress might be indicated by different optical cues.

Table 8

Variance accounted for (r^2) for the correlations between production and percent correct perception of stressed syllables in the phrasal dataset. Significant correlations are shown in bold type

	r^2	*p-value*
BROW DISPL	**.127**	***p* = .008**
HEAD DISPL	**.169**	***p* = .003**
LIP DIST	.051	$p = .102$
LIP OPEN DISPL	**.213**	***p* = .001**
LIP CLOS DISPL	**.235**	***p* < .001**
CHIN OPEN DISPL	**.397**	***p* < .001**
CHIN CLOS DISPL	**.315**	***p* < .001**
LIP OPEN VEL	**.074**	***p* = .048**
LIP CLOS VEL	**.248**	***p* < .001**
CHIN OPEN VEL	**.243**	***p* < .001**
CHIN CLOS VEL	**.298**	***p* < .001**

5.1
Correlation results

5.1.1
Stressed syllables

Ten of the eleven measures made on stressed syllables correlated significantly with the percent correct perception of phrasal stress, as shown in Table 8. Only Lip Distance failed to show a significant correlation, which is somewhat of a surprise given that this measure looked promising based on its consistent stress differentiation across word types, stress types, and talkers. That is, talkers were making stress distinctions in Lip Distance, but perceivers were apparently not attending to them. This was so even though the lower lip, because it is carried on the jaw, effectively conveys information about stress from both the lips and the chin. The variance in perception success accounted for by each of the visible production measures ranged from 7% to 40%, with chin measures accounting for the most variance (at least 24%).[11] Additionally, displacement measures seemed to be more explanatory than velocity measures, at least for opening gestures.

5.1.2
Unstressed syllables

Eight of the eleven measures correlated significantly with the percentage correct non-perception of stress on unstressed syllables, as shown in Table 9. The variance accounted for by each of the measures in the unstressed syllables was less than in stressed syllables, ranging from 5–12%.

[11] In considering the relatively low correlations with Brow and Head Displacements, it should be kept in mind that both of these measures had many zero values, that is, tokens with no movement at all. For Brow Displacement, 154 of the 216 tokens had no movement, while for Head Displacement it was 44 of the 216 tokens. Clearly, with so many cases of no movement, there is less information to be had from those variables.

Table 9

Variance accounted for (r^2) for the correlations between production and percent correct perception of unstressed syllables in the phrasal dataset. Significant correlations are shown in bold type

	r^2	*p-value*
BROW DISPL	**.048**	$\boldsymbol{p = .005}$
HEAD DISPL	**.053**	$\boldsymbol{p = .011}$
LIP DIST	**.065**	$\boldsymbol{p = .001}$
LIP OPEN DISPL	**.102**	$\boldsymbol{p < .001}$
LIP CLOS DISPL	**.112**	$\boldsymbol{p < .001}$
CHIN OPEN DISPL	**.112**	$\boldsymbol{p < .001}$
CHIN CLOS DISPL	.002	$p = .549$
LIP OPEN VEL	.001	$p = .780$
LIP CLOS VEL	**.123**	$\boldsymbol{p < .001}$
CHIN OPEN VEL	.011	$p = .194$
CHIN CLOS VEL	**.105**	$\boldsymbol{p < .001}$

Note that in unstressed syllables, lip measures accounted for at least as much variance as chin measures, and were more reliable as correlates of perceptual success, with four significant correlations from five possible lip measures as against two significant correlations from four possible chin measures. Thus, while the presence of stress seemed to be most visible on the chin, the absence of stress seemed to be most apparent from the lips. Also, in unstressed as in stressed syllables, eyebrow and head movements were significantly correlated with correct perception, but the variance accounted for was lower than for the other measures (and lower in unstressed than in stressed syllables). In what follows, only correlations with stress will be considered.

5.2
Partial correlations

In the correlations reported in the previous section, all but one of the production variables were correlated with the perception results. However, because many of the production measures are also highly correlated with one another, it is not yet clear whether there are indeed 10 independent production variables that affect the perception of phrasal stress, or whether certain of these variables subsume the effect of certain others. Partial correlations show the correlation between an independent and dependent variable when the linear effects of other independent variables have been controlled for. Such partial correlations allow us to assess the relative importance of individual independent variables in providing predictive information above and beyond the information provided by other independent variables.

5.2.1
Related articulators: chin and lips

It is unsurprising that at least some of the production measures are highly correlated with others, given the close physical relation between certain articulators: in particular, the lips and the chin, which move largely together in mouth opening. Therefore, we

tested whether the lips and chin contributed independently to the success of perceivers or whether one of the two articulators drives the effect of the other. We could expect that since the jaw moves the lower lip rather than the reverse, the lips' contribution would be largely dependent on the chin's.

Partial correlations for each of the chin measures, controlling for the contributions of the corresponding lip measures, showed significant independent contributions of the chin to perception in all cases. Chin Opening Displacement, for instance, which showed the highest simple correlation with perception, $R^2 = .40$, still accounted independently for 25% of the variance when Lip Opening Displacement was controlled for. Furthermore, Lip Displacement measures showed no significant correlation with perception when Chin Displacement was controlled for. In fact, lip measures never showed an effect independent of chin measures, even in unmatched chin–lip pairs. All of the other chin measures showed similar amounts of information added over corresponding lip measures.

5.2.2
Displacement and velocity

Similarly, corresponding displacement and velocity measures were highly correlated. Considering all displacement–velocity pairings (corresponding or not), velocity never showed an independent contribution to perception over the contribution of displacement. Chin Opening Displacement, on the other hand, contributed to perception even when Chin Opening and Closing Velocity were controlled for, accounting independently for 26% and 14% respectively of the variance, and Chin Closing Displacement accounted for 10% of the variance over and above Chin Opening Velocity. Displacement, then, provided more information for perception than velocity. Opening and closing measures were correlated as well; however, the partial correlations did not so straightforwardly disentangle their relation. For displacement measures, closing never showed a correlation with perception that was independent of the effect of opening. But with respect to velocity, closing measures (both chin and lip) showed a consistent, unmatched independent contribution to percent correct.

5.3
Combined contributions

The partial correlations show which measures are most explanatory, above and beyond the explanatory contributions of other measures, but they still only assess the individual contributions of the different measures. However, the combined effect of multiple measures may benefit perception more than any individual measures. Multiple regression analyzes the simultaneous effects of multiple variables, but it cannot be reliably used for our data because so many of our measures are correlated with one another. However, we eliminate the collinearity problem by using the partial correlations to make predictions about which of the intercorrelated measures are most important and removing the less explanatory measures. Thus, lip measures were omitted due to the contribution of chin measures, and velocity measures were omitted in favor of displacement, leaving Chin Opening and Closing Displacement, Head Displacement, and Brow Displacement as possible predictors. Stepwise regression yielded a significant model which accounted for 58% of the variance in perception, as indicated by adjusted R^2, $F(2, 46) = 34.8$, $p < .0001$. Two variables were significant in this model: Chin Opening

Table 10

Variance accounted for (R^2) for the independent correlations between Chin Opening Displacement and percent correct perception of phrasal stress when contributions of each of five other variables are partialed out

CHIN OPEN DISPL indep. corr. over	R^2	*p-value*
CHIN OPEN VEL	.26	$p < .0001$
LIP OPEN VEL	.38	$p < .0001$
LIP OPEN DISPL	.25	$p < .0001$
CHIN CLOS DISPL	.13	$p = .009$
CHIN CLOS VEL	.14	$p = .005$

Displacement (beta = –.700, $p < .0001$) and Eyebrow Displacement (beta = .400, $p < .0001$). In this model, Brow Displacement contributed 14.6% to the explanation of variance, beyond the contribution of Chin Opening. (Head displacement was not a significant predictor in this model.)

5.4 Discussion

To summarize, chin gestures were informative for perception beyond what lip gestures could show. And displacement was generally informative independent of velocity. In fact, Chin Opening Displacement, with an average displacement difference between stress conditions of 2.7 mm, seems to be one of the most important production variables for the perception of phrasal stress. It showed an unmatched independent correlation with correct perception over the largest number of variables; these independent correlations are summarized in Table 10.

Chin Opening Displacement was among the predicted best measures based on its reliability across stimulus conditions (lexical and phrasal, real and reiterant) and talkers in the production experiment. These criteria identified Lip Distance and Chin Opening Velocity as good measures as well. Closing measures and brow movements were less consistently available, particularly in the lexical items where second syllable closing gestures were lost, and the eyebrows did not move and so were predicted to be *less* useful in visual perception. The observations from the production study were largely shown to be associated with perception, although a multiple regression analysis showed that Eyebrow Displacement did have additional predictive value when combined with Chin Opening Displacement for phrasal stress perception.

In comparing optical measures of movements with visual perception, we do not know how physical displacements of different articulators are scaled psychologically. It is entirely possible that the just noticeable difference for chin movements is smaller than for, say, lip movements, such that physically smaller chin movements can be visually more salient than larger lip movements. Our data can be taken to suggest such a hypothesis.

6 General discussion and conclusions

In this study, we investigated associations between optical phonetic characteristics and perceivers' recognition of lexical and phrasal stress. First, we looked at selected optical aspects of talkers' faces that reliably were correlated with stress. Then we investigated which types of utterances and stresses were most reliably/accurately perceived. Lastly, we assessed the associations between optical measures and the perception of phrasal stress.

The production experiment showed that both lexical stress (that is, phrasal stress on isolated words) and phrasal stress (in sentences) were well marked by a number of potentially visible articulatory correlates. In general, the effects of phrasal stress were larger than those of lexical stress. All eleven of the measures examined reliably distinguished phrasal stress, ten of the eleven distinguished lexical stress in reiterant words, and three distinguished stress in real words. These three measures, Lip Distance, Chin Opening Displacement and Chin Opening Velocity, were thus the most consistent measures across conditions (distinguishing stress in sentences, reiterant, and real words). They were also most consistent across talkers.

The large literature on the production of stress has shown that stressed syllables have greater acoustic intensity (presumably related to mouth opening) and larger, longer, and faster movements of the jaw, lips, and other articulators. Accordingly, our measures included Lip Distance (roughly, vertical mouth opening), and displacements and velocities of the opening and closing movements of the chin and lips, all of which varied significantly with stress in the reiterant and sentence corpora. The Lip Distance and chin movement measures, which are directly visible, extend the measures in the literature, which include lip and jaw measures. At the same time, our result that two chin measures were among the most consistently distinctive for stress supports the focus of the speech production literature on jaw movements rather than lip movements in the study of speech prosody, given that the chin is a close proxy for the jaw. We found a mean Chin Opening Displacement for phrasally stressed syllables of 12 mm, which was 2.7 mm greater than the displacement for unstressed syllables. These numbers are slightly larger than those reported by Erickson (2002) for the jaw, even for her largest talker, but generally similar.

Previous production studies have also shown that head and eyebrow movements can accompany pitch accents. In our study, with respect to eyebrow movement, this was true only when the talkers produced entire sentences. However, the mean difference for Brow Displacement between stressed and stressless syllables in those sentences was the smallest of all the measures: .5 mm (.5 mm for stressed vs. no movement for stressless). The head moved with stress more consistently (in both isolated words and sentences), but again the mean differences between stressed and unstressed were quite small, on the order of 1 mm or less (with mean displacements of about .5–1 mm for stressless and 1–2 mm for stressed). (See Figure 5 and Tables 2 and 6 for these comparisons.) Head and eyebrow movements are thus much smaller than lip and chin movements, which are more on the order of a centimeter. Since most previous studies mentioning eyebrow and head movements with stress do not give specific measurements,[12] we cannot say whether

[12] Even studies of the cue value of synthesized head and brow movements generally do not report the absolute size of those movements, only that they were large enough to be detectable.

our talkers are typical, but Dohen et al. (2006) report that the eyebrow movements of their one French talker with frequent movements were very small, less than 2 mm for stressed movements, consistent with our results. In contrast, the movements reported here are smaller than those used by Massaro and Beskow (2002) in the stimuli for an audiovisual perception study (e.g., our eyebrow movements averaged only .5 mm, while Massaro and Beskow's synthetic movements were as large as 6.36 mm).

Our perception experiment showed that visual perception of both lexical and phrasal stress was significantly above chance, with phrasal stress more accurately perceived than lexical stress: 62.2% correct for lexical stress (vs. 50% chance), and 53.8% correct for phrasal stress (vs. 25% chance). Overall perception of lexical stress was about the same for the various conditions (reiterant vs. real words, minimal vs. non-minimal pairs, *buh* vs. *fer* reiterant syllable). Overall perception of phrasal stress was about the same for the different positions of focal accents, but broad focus was less well perceived. This overall performance on phrasal stress, 53.8% correct (vs. 25% chance), is lower than reported in some previous studies, for example, 76% in Bernstein et al. (1989) (vs. 33% chance) and 71% (vs. 25% chance) for one speaker in Dohen et al. (2005), but higher than, for example, the 43% (vs. 25% chance) reported for the other speaker in Dohen et al. (2005).

The three talkers had been selected for their different visual sentence intelligibility to Deaf speech readers. However, there were relatively few, or only small, differences in production among the talkers. T-HI made more use of head movements, and T-LO made the most and the largest articulatory differences with phrasal stress. These talker differences were not well reflected in the perception study, which showed few talker effects. T-HI was as easy to read as was T-LO for phrasal stress, suggesting that his head movements might have helped, but he was no easier to read for lexical stress, where he also moved his head. T-LO was indeed the easiest talker to read, but most clearly for the reiterant speech corpus, where our production measures did not distinguish him from the others very well. Perhaps in perceiving reiterant lexical stress (which our correlational analyses did not include), perceivers used a slightly different mix of optical dimensions, more in line with T-LO's productions. At the same time, the Lip Closing and Opening Displacement measures pattern by talkers like their intelligibility (T-LO, T-HI > T-MID), as shown in Table 7. It seems likely that in the reiterant corpus, with /b/ and /f/ syllables, lip movements may have been more salient and informative than in the sentence corpus. Finally, to the extent that the talkers differed in intelligibility, those differences were not in accord with their preliminary screening scores for sentence intelligibility, since T-LO had the lowest intelligibility in screening (and even in the later sentence perception study described in footnote 5, was still less intelligible than T-HI). That is, segmental and prosodic intelligibility are not clearly related.

There were also no large or striking differences among the perceivers in the study. Perception of lexical and phrasal stress was correlated within perceivers—that is, people who were good at reading stress in isolated words were good at reading it in sentences—but perception of stress and segmental screening were not correlated—that is, people who were good at reading sentences were not necessarily good at reading stress, and vice versa. The exceptions to this pattern were the very best perceivers, who were best at all of these perception tasks.

The results from the production and perception studies of phrasal stress were compared directly through correlation analyses. While most measures (10 of 11) were

correlated with correct perception, surprisingly, Lip Distance (i.e., mouth opening), which was one of the consistently-produced measures, was not. That is, the perceivers did not seem to rely on mouth opening in making their judgments. That left only two measures, Chin Opening Displacement and Chin Opening Velocity, that were consistently used to mark stress in production *and* were correlated with perception. The analysis of partial correlations for phrasal stress then showed that chin measures indeed were the greatest predictors of correct perception, particularly Chin Opening Displacement. In general, chin measures were more powerful than lip measures, displacement measures were more powerful than velocity measures, and opening measures were more powerful than closing measures. Thus, Chin Opening Displacement was the most powerful of these measures, accounting for the most variance in perception.

Head and Brow Displacements were found to be small, but both measures were significantly correlated with correct perception of stressed and unstressed words in sentences, with Brow Displacement accounting for 12.7% of variance and Head Displacement accounting for 16.9% of variance in perception of stressed words. Clearly perceivers pay some attention to head and eyebrow movements. These contributions to variance are much smaller than those of some other measures, most notably the 40% variance accounted for by Chin Opening Displacement. However, when the results of the partial correlation analysis are taken into account, in which some variables' contributions are seen to be subsumed under other variables, the contributions of head and brow movements to perception are not negligible. Since Head and Brow Displacements are not correlated with any other measures, including each other, they provide independent information about stress, and Brow Displacement also shows a significant contribution in addition to Chin Opening Displacement in a stepwise multiple regression model predicting correct perception.

The production measure that Dohen et al. (2006) found to vary most with stress in French, Lip Protrusion, was not analyzed in our study. Thus we cannot tell if Lip Protrusion might have varied with stress and been a useful cue to our perceivers, perhaps making a contribution independent of the vertical chin and lip opening measures that we did analyze. However, given the limited use of lip rounding in the phonemes of California English, it seems unlikely that this measure would be anywhere near as important in English as in French. Dohen et al.'s studies have also highlighted the importance of durational measures. Our study does not consider duration, per se.[13] To some extent our peak velocity measures stand in lieu of duration measures, and our correlation analyses found velocity measures to be relatively unimportant

[13] For the sake of comparison, acoustic duration measurements of the stressed syllable of each name and of the monosyllabic word preceding each name were made from the audio-recordings of a subset of the phrasal corpus. Analysis indicated that the syllables in phrasally stressed names were longer than those in unstressed names, as expected. Words in prefocal position were longer than those preceding unfocused names or in broad focus sentences for one talker, T-MID, but T-LO and T-HI showed no durational differences across focus conditions. Thus, there is some evidence that our English talkers may vary like Dohen et al.'s French talkers, with some marking stress with prefocal lengthening (as well as focal properties) and others not. However, T-MID was least-well perceived in our study, whereas the talkers using prefocal lengthening were better perceived in Dohen and Loevenbruck (2005).

(once displacement is taken into account) in perception. But it is possible that actual duration measures played a role for our perceivers.

Evidence from three parts of the study, then, points toward movements of the chin as important information for the visual perception of prosody. Chin Opening Displacement and Velocity were two of the three measures that best differentiated stressed versus stressless syllables. Chin Opening and Closing Displacement measures accounted for the most variance in perception in individual correlations, and Chin Opening Displacement accounted for the most independent variance in partial correlations. That is, talkers made larger and faster chin movements with stress, and perceivers made use of those differences, paying most attention to how far the chin moved when opening.

This is not to say 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 cues, correlated and uncorrelated. This study has revealed a number of such potential cues in the optical speech signal. Several measures related to chin movement provide redundant information about stress, and head and eyebrow movements can provide independent (and, in the case of brows, additional) information. But among these various cues, Chin Opening Displacement is one of the most consistently present markers of stress, and the most informative to perceivers above and beyond the other cues we considered.

Our study is the first to suggest which of a large set of facial measures are most useful for the perception of stress. The most similar previous study, Dohen et al. (2004c), compared items in which French phrasal stress had been well perceived to determine which of a set of production measures most differentiated those particular items. These informal comparisons suggested that Duration, Lip Area, and Jaw Opening (really chin opening, since measured from a video-recording, i.e., from a fleshpoint rather than from the mandible) were the most important measures. Our results were similar to the extent that we also identified chin opening movements as important, but our study could localize this importance specifically to Chin Opening Displacement. Our results differ to the extent that our measures most related to duration—peak velocity measures—and to lip area—Lip Distance—while marking stress in production, were relatively unimportant in perception. It remains possible, however, that other measures, including Dohen et al.'s, could also have been important to our perceivers.

The findings of this study may have application to facial speech synthesis of continuous English speech (talking heads). Data from our sentence corpus suggests that, as a rule of thumb, to convey intelligible phrasal stress the chin should open on average by at least a centimeter, and open less than that for unstressed items. (Of course the degree of opening should vary with vowel height as well as with stress; the 1 cm figure applies to high vowels, while stressed low vowels open about 1.5 cm.) Movements of the head and/or eyebrows, even if small, will provide additional, independent, information about stress. At the same time, since in our data the largest displacements and the largest differences observed were for Lip Opening and Closing Displacements—that is, the lips move more than the chin overall, and even more for stress—then lip movements, even though they made no significant independent contribution to correct perception, might be important for naturalness. A rule of thumb would be to make the lips move on average at least 1.6 cm for stress.

We have already noted that the head and eyebrow movements observed in this study are smaller than the synthetic movements tested in Massaro and Beskow's (2002) audiovisual perception study. They found that even their most exaggerated movements had only a limited influence on perception, when visual cues conflicted with auditory cues: auditory cues dominated perception, and the extreme visual cues were barely sufficient to switch the percept category. It remains to be determined whether the size of movements reported here can contribute to congruent-cue, as opposed to conflicting-cue, audiovisual perception of English stress. But we can suggest that an audiovisual perception study whose goal is to understand the nature of audiovisual cue integration might profitably vary not eyebrow and head movement, but rather the visual cue that seems most important to visual-only perceivers, namely Chin Opening Displacement. That cue is likely to be more robust in competition with auditory cues.

In conclusion, the question addressed by our study has been how information about pitch-accented syllables in isolated words ("lexical stress") and in sentences ("phrasal stress") is conveyed by a talker's face to visual perceivers. We identified some aspects of production that vary most systematically with stress: Lip Distance, displacements and velocities of the opening and closing movements of the chin and lips, and to a lesser extent Head Displacement. We then tested visual perception of the utterances whose production was studied, and found that phrasal stress was more accurately perceived than lexical stress. Though the three talkers had been chosen to vary in their visual segmental intelligibility, we found only small differences in their prosodic intelligibility, and these small differences were not as expected based on their segmental intelligibility. Finally, we related perception of phrasal stress to the production measures. While most measures were correlated with perception performance, the chin measures, especially Chin Opening Displacement, contributed the most to correct perception independently of the other measures. Head and eyebrow movements also made an independent contribution. Thus, our results indicate that visual perceivers, when they perceive stress, attend mainly to mouth opening movements.

References

ALEXANDRE, M.-A., & GÉRARD, C. (2002, April). Perception of emphatic stress: Multiple regression analyses. Paper presented at Temporal Integration in Perception of Speech, Aix-en-Provence, France.

AUER, E. T., JR., BERNSTEIN, L. E., & COULTER, D. C. (1998). Temporal and spatio-temporal vibrotactile displays for voice fundamental frequency: An initial evaluation of a new vibrotactile speech perception aid with normal-hearing and hearing-impaired individuals. *Journal of the Acoustical Society of America*, **104**, 2477–2489.

BECKMAN, M., & ELAM, G. A. (1997). *Guidelines for ToBI labelling* (Version 3). Unpublished manuscript, Ohio State University, USA.

BECKMAN, M. E. (1986). *Stress and non-stress accent*. Dordrecht: Foris.

BECKMAN, M. E., & EDWARDS, J. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Ed.), *Papers in laboratory phonology III: Phonological structure and phonetic form* (pp.7–33). Cambridge: Cambridge University Press.

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 [abstract]. *Journal of the Acoustical Society of America*, **107**, 2887.

BERNSTEIN, L. E., EBERHARDT, S. P., & DEMOREST, M. E. (1989). Single-channel vibrotactile supplements to visual perception of intonation and stress. *Journal of the Acoustical Society of America*, **85**, 397–405.

BOLINGER, D. L. (1958). A theory of pitch accent in English. *Word*, **14**, 109–149.

CAMPBELL, W. N. (1993). Automatic detection of prosodic boundaries in speech. *Speech Communication*, **13**, 343–354.

CAVÉ, C., GUAITELLA, I., BERTRAND, R., SANTI, S., HARLAY, F., & ESPESSER, R. (1996). About the relationship between eyebrow movements and F0 variation. *ICSLP 96*, **4**, 2175–2179.

CHO, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/ in English. *Journal of the Acoustical Society of America*, **117**, 3867–3878.

CHO, T. (2006). Manifestation of prosodic structure in articulation: Evidence from lip kinematics in English. In L. M. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Laboratory phonology 8: Varieties of phonological competence* (pp.519–548). Berlin/New York: Mouton de Gruyter.

CHOI, J-Y., HASEGAWA-JOHNSON, M., & COLE, J. (2005). Finding intonational boundaries using acoustic cues related to the voice source. *Journal of the Acoustical Society of America*, **118**, 2579–2587.

CONDON, W. (1976). An analysis of behavioral organization. *Sign Language Studies*, **13**, 285–318.

de JONG, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, **97**, 491–504.

DOHEN, M., LOEVENBRUCK, H., CATHIARD, M.-A., & SCHWARTZ, J.-L. (2004a). Visual perception of contrastive focus in reiterant French speech. *Speech Communication*, **44**, 155–172.

DOHEN, M., LOEVENBRUCK, H., CATHIARD, M.-A., & SCHWARTZ, J.-L. (2004b). Identification of the possible visible correlates of contrastive focus in French. In B. Bel and I. Marlien (Eds.), *Speech prosody 2004* (pp.77–80). Retrieved February 2009, from http://www.isca-speech.org/archive/sp2004

DOHEN, M., LOEVENBRUCK, H., CATHIARD, M.-A., & SCHWARTZ, J.-L. (2004c). Can we see focus? A visual perception study of contrastive focus in French. In B. Bel & I. Marlien (Eds.), *Speech prosody 2004* (pp.73–76). Retrieved February 2009, from http://www.isca-speech.org/archive/sp2004

DOHEN, M., & LOEVENBRUCK, H. (2005). Audiovisual production and perception of contrastive focus in French: A multispeaker study. In *Interspeech 2005* (pp.2413–2416). Retrieved February 2009, from http://www.isca-speech.org/archive/interspeech_2005

DOHEN, M., LOEVENBRUCK, H., & HILL, H. (2005). A multi-measurement approach to the identification of the audiovisual facial correlates of contrastive focus in French. In E. Vatikiotis-Bateson (Ed.), *Auditory-visual processing workshop* (pp.115–116). Retrieved February 2009, from http://www.isca-speech.org/archive/avsp05

DOHEN, M., LOEVENBRUCK, H., & HILL, H. (2006). Visual correlates of prosodic contrastive focus in French: Description and inter-speaker variability. In R. Hoffmann & H. Mixdorff (Eds.), *Speech prosody 2006* (pp.221–224). Retrieved February 11, 2008, from http://www.isca-speech.org/archive/sp2006

EKMAN, P. (1979). About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, & D. Ploog (Eds.), *Human ethology: Claims and limits of a new discipline* (pp.169–202). Cambridge University Press.

ERICKSON, D. (2002). Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, **59**, 134–149.

ERICKSON, D., FUJIMURA, O., & PARDO, B. (1998). Articulatory correlates of prosodic control: emotion and emphasis. *Language and Speech*, **41**, 395–413.

FANT, G., KRUCKENBERG, A., & LILJENCRANTS, J. (2000). Acoustic-phonetic analysis of prominence in Swedish. In A. Botinis (Ed.), *Intonation* (pp.55–86). Kluwer Academic Publishers.

FRY, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, **1**, 126–152.

GRANSTRÖM, B., HOUSE, D., & LUNDEBERG, M. (1999). Prosodic cues to multimodal speech perception. In *Proceedings of the 14th International Congress on Phonetic Sciences* (pp.655–658). San Francisco.

GRANT, K. W., ARDELL, L. A., KUHL, P. K., & SPARKS, D. W. (1986). The transmission of prosodic information via an electrotactile speech reading aid. *Ear and Hearing*, **7**, 328–335.

HADAR, U., STEINER, T. J., GRANT, E. C., & CLIFFORD ROSE, F. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, **26**, 117–129.

HARRINGTON, J., FLETCHER, J., & BECKMAN, M. (2000). Manner and place conflicts in the articulation of accent in Australian English. In M. B. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V* (pp. 40–51). Cambridge: Cambridge University Press.

HAYES, B. (1995). *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.

HELDNER, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics*, **31**, 39–62.

HERMENT-DUJARDIN, S., & HIRST, D. (2002). Emphasis in English: A perceptual study based on modified synthetic speech. In B. Bel & I. Marlien (Eds.), *Speech prosody 2004* (pp.379–382). Retrieved February 2009, from http://www.isca-speech.org/archive/sp2004

HIRST, D., & DI CRISTO, A. (1998). A survey of intonation systems. In D. Hirst & A. Di Cristo (Eds.), *Intonation systems: A survey of twenty languages* (pp.1–44). Cambridge: Cambridge University Press.

HOUSE, D., BESKOW, J., & GRANSTRÖM, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. In by P. Dalsgaard, B. Lindberg, H. Benner, & Z. Tan (Eds.), *Proceedings of Eurospeech 2001* (pp.387–390). Retrieved February, 2009, from http://www.isca-speech.org/archive/eurospeech_2001

JENSEN, C. (2003). Perception of prominence in Standard British English. In *Proceedings of the 15th International Congress on Phonetic Sciences* (pp.1815–1819). Barcelona.

JIANG, J., ALWAN, A., KEATING, P. A., AUER, E. T., & BERNSTEIN, L. E. (2002). On the relationship between face movements, tongue movements and speech acoustics. *EURASIP Journal on Applied Signal Processing*, **11**, 1174–1188.

JIANG, J., AUER, E. T., ALWAN, A., KEATING, P. A., & BERNSTEIN, L. E. (2007). Similarity structure in visual speech perception and optical phonetic signals. *Perception and Psychophysics*, **69**, 1070–1083.

KEATING, P. A., CHO, T., BARONI, M., MATTYS, S., BERNSTEIN, L. E., CHANEY, B., ET AL. (2000). Articulation of word and sentence stress [abstract]. *Journal of the Acoustical Society of America*, **108**, 2466.

KENDON, A. (1978). Differential perception and attentional frame: Two problems for investigation. *Semiotica*, **24**, 305–315.

KOCHANSKI, G., GRABE, E., COLEMAN, J., & ROSNER, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, **118**, 1038–1054.

KRAHMER, E., RUTTKAY, Z., SWERTS, M., & WESSELINK, W. (2002). Pitch, eyebrows and the perception of focus. In B. Bel & I. Marlien (Eds.), *Speech prosody 2004* (pp.443–446). Retrieved February 2009, from http://www.isca-speech.org/archive/sp2004

KRAHMER, E., & SWERTS, M. (2006). Hearing and seeing beats. In R. Hoffmann & H. Mixdorff (Eds.), *Speech prosody 2006*. Retrieved February 2009, from http://www.isca-speech.org/archive/sp2006

LADD, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

LANSING, C. R., & McCONKIE, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, **42**, 526–539.

LEHISTE, I. (1970). *Suprasegmentals*. Cambridge, MA: The MIT Press.

MASSARO, D. W., & BESKOW, J. (2002). Multimodal speech perception: A paradigm for speech science. In B. Granström, D. House, & I. Karlsson (Eds.), *Multimodality in language and speech systems* (pp.45–71). Dordrecht: Kluwer Academic Publishers.

MUNHALL, K. G., JONES, J. A., CALLAN, D. E., KURATATE, T., & VATIKIOTIS-BATESON, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, **15**, 133–138.

PENTLAND, A. P., & DARELL, T. (1994). Visual perception of human bodies and faces for multi-modal interfaces. *ICSLP 94*, Yokohama, 543–546.

RIETVELD, T., & GUSSENHOVEN, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics*, **13**, 299–308.

RISBERG, A., & LUBKER, J. (1978). Prosody and speechreading. *Speech Transmission Laboratory-Quarterly Progress Report, Status Report*, **4**, 1–16.

SILVERMAN, K., BECKMAN, M., PITRELLI, J., OSTENDORF, M., WIGHTMAN, C., PRICE, P., ET AL. (1992). ToBI: A standard for labeling English prosody. *Proceedings of the 1992 International Conference on Spoken Language Processing*, **2**: 867–870.

SLUIJTER, A. M. C., van HEUVEN, V. J., & PACILLY, J. J. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, **101**, 503–513.

SRINIVASAN, R. J., & MASSARO, D. W. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech*, **46**, 1–22.

SWERTS, M., & KRAHMER, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, **36**, 219–238.

THOMPSON, D. M. (1934). On the detection of emphasis in spoken sentences by means of visual, tactual, and visual-tactual cues. *Journal of General Psychology*, **11**, 160–172.

van KUIJIK, D., & BOVES, L. (1999). Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication*, **27**, 95–111.

VATIKIOTIS-BATESON, E., EIGSTI, I.-M., YANO, S., & MUNHALL, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception and Psychophysics*, **60**, 926–940.

YEHIA, H. C., KURATATE, T., & VATIKIOTIS-BATESON, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, **30**, 555–568.