The Quarterly Journal of Experimental Psychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pqje20

Suprasegmental lexical stress cues in visual speech can guide spoken-word recognition

Alexandra Jesse[ab] & James M. McQueen[cb]

[a] Department of Psychology, University of Massachusetts, Amherst, MA, USA

[b] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[c] Behavioural Science Institute and Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University Nijmegen, Nijmegen, The Netherlands

Accepted author version posted online: 13 Aug 2013.Published online: 18 Oct 2013.

To cite this article: Alexandra Jesse & James M. McQueen (2014) Suprasegmental lexical stress cues in visual speech can guide spoken-word recognition, The Quarterly Journal of Experimental Psychology, 67:4, 793-808, DOI: 10.1080/17470218.2013.834371

To link to this article: http://dx.doi.org/10.1080/17470218.2013.834371

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2014
Vol. 67, No. 4, 793–808, http://dx.doi.org/10.1080/17470218.2013.834371

Suprasegmental lexical stress cues in visual speech can guide spoken-word recognition

Alexandra Jesse[1,2] and James M. McQueen[3,2]

[1]Department of Psychology, University of Massachusetts, Amherst, MA, USA
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[3]Behavioural Science Institute and Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University Nijmegen, Nijmegen, The Netherlands

Visual cues to the individual segments of speech and to sentence prosody guide speech recognition. The present study tested whether visual suprasegmental cues to the stress patterns of words can also constrain recognition. Dutch listeners use acoustic suprasegmental cues to lexical stress (changes in duration, amplitude, and pitch) in spoken-word recognition. We asked here whether they can also use visual suprasegmental cues. In two categorization experiments, Dutch participants saw a speaker say fragments of word pairs that were segmentally identical but differed in their stress realization (e.g., *ˈca-vi* from *cavia* "guinea pig" vs. *ˌka-vi* from *kaviaar* "caviar"). Participants were able to distinguish between these pairs from seeing a speaker alone. Only the presence of primary stress in the fragment, not its absence, was informative. Participants were able to distinguish visually primary from secondary stress on first syllables, but only when the fragment-bearing target word carried phrase-level emphasis. Furthermore, participants distinguished fragments with primary stress on their second syllable from those with secondary stress on their first syllable (e.g., *pro-ˈjec* from *projector* "projector" vs. *ˌpro-jec* from *projectiel* "projectile"), independently of phrase-level emphasis. Seeing a speaker thus contributes to spoken-word recognition by providing suprasegmental information about the presence of primary lexical stress.

Keywords: Visual speech; Prosody; Lexical stress; Spoken-word recognition.

Speech is mostly communicated in face-to-face conversations. Speech perception is thus often a multimodal process. Seeing a speaker typically improves speech perception (Miller & Nicely, 1955; Reisberg, McLean, & Goldfield, 1987; Sumby & Pollack, 1954), as visual speech contains information that is redundant and complementary to that in auditory speech (Jesse & Massaro, 2010; Walden, Prosek, & Worthington, 1974). Visual information can influence the recognition of individual speech segments (as, e.g., the classic McGurk illusion shows; McGurk & MacDonald, 1976). Visual speech can also provide prosodic information that spans across segments. This

Correspondence should be addressed to Alexandra Jesse, Department of Psychology, University of Massachusetts, Amherst, 135 Hicks Way, Amherst, MA 01003, USA. E-mail: Alexandra Jesse: ajesse@psych.umass.edu

FUNDING

This research was supported by a grant to the first author within the German Science Foundation (DFG) focus program SPP 1234 "Phonological and phonetic competence: Between grammar, signal processing, and neural activity". We thank Bettina Braun and Lara Tagliapietra for valuable discussions, two anonymous reviewers for constructive feedback, and Anke Bergmans, Laurence Bruggeman, Lies Cuijpers, Vera Hoskam, Jessica Koppers, Marieke Pompe, Robbert van Sluijs, Jet Sueters, and Jelmer Wolterink for their help with the experiments.

suprasegmental information helps speech perception at the phrasal level. Visual speech can be used, for example, in the recognition of an utterance's intonation contour (e.g., the difference between falling and rising intonations; Fisher, 1969). In the present study, we tested whether visual speech can provide suprasegmental information also for the recognition of individual spoken words. Specifically, we investigated whether lexical suprasegmental stress information is available, and used, when only seeing a speaker.

Lexical stress

Lexical stress is the lexically defined relative emphasis of syllables on the word level. In variable stress languages, such as English, Dutch, Spanish, and Italian, stress placement is not restricted to a fixed position within a word. Stress placement can therefore be informative about the word. And indeed, lexical stress helps listeners of these languages with lexical access (Cooper, Cutler, & Wales, 2002; Cutler & van Donselaar, 2001; Reinisch, Jesse, & McQueen, 2010; Soto-Faraco, Sebastián-Gallés, & Cutler, 2001; Sulpizio & McQueen, 2012; van Donselaar, Koster, & Cutler, 2005).

The use of lexical stress cues in word recognition, however, depends on the lexical properties of a given language. Stress can be implemented segmentally—that is, by changes of individual segments depending on the stress status of a syllable. In English, for example, unstressed syllables usually contain the reduced vowel schwa. But stress can also be expressed by changes that do not affect segmental identity. These suprasegmental cues are acoustically lawful variations in duration, amplitude, fundamental frequency (F0), or spectral tilt (e.g., see Fry, 1958; Lieberman, 1960). For English and Dutch, for example, a stressed syllable tends to be longer and to be higher in amplitude and pitch than its unstressed counterpart. Languages vary in their use of these stress cues. In English, stress is almost always implemented by segmental changes. Unstressed syllables nearly always contain schwa, and schwa never occurs in stressed syllables. These stress-conditioned segmental changes often signal the only difference between pairs of words, such as between many noun–verb pairs [e.g., *(to) sus-ˈpect* versus *(a) ˈsus-pect*; the diacritic ˈ marks syllables with primary lexical stress, and the diacritic ˌ indicates secondary lexical stress on a syllable]. Minimal pairs or word beginnings that differ only suprasegmentally are possible but highly uncommon in English [e.g., *(to) for-ˈbear* versus *(a) ˈfor-bear*; Cutler, 1986]. For English listeners, therefore, segmental cues to stress are more important than suprasegmental cues (Cutler & Clifton, 1984; Fear, Cutler, & Butterfield, 1995). Nevertheless, English listeners are sensitive to suprasegmental stress variation (Cooper et al., 2002). Hearing a syllable like *ˈmu* sped up English listeners' lexical decision responses to subsequently presented printed target words that began with this syllable (Cooper et al., 2002). This facilitatory priming effect, however, was larger when the prime matched the target suprasegmentally (e.g., *ˈmu-sic* vs. *mu-ˈse-um*).

A different picture emerges for languages that use suprasegmental stress information to a larger degree to distinguish words. Here, suprasegmental lexical stress plays a more important role in word recognition. In Dutch, stress can be cued segmentally and suprasegmentally, but is often implemented solely suprasegmentally. Taking suprasegmental stress into account reduces the average number of words that are embedded in other words in Dutch by about a half from 1.55 to 0.74 words (Cutler & Pasveer, 2006). Similarly, stress information moves the theoretical point at which words become unique closer to their onsets, such that to recognize the average Dutch word 80% of its phonemes are needed, but only 66% when stress is also considered (van Heuven & Hagman, 1988). In Dutch, again, facilitatory priming effects can be observed when fragment primes and targets overlap suprasegmentally. But unlike in English, inhibitory priming effects can be found when bisyllabic fragment primes differ suprasegmentally from the targets (van Donselaar et al., 2005). This suggests that, in Dutch, suprasegmental stress information modulates the competition of words. Furthermore,

Dutch listeners use suprasegmental stress information to constrain lexical competition and to disambiguate words as soon as this information comes available and do not wait for disambiguating (stress-unrelated) segmental information to come available (Reinisch et al., 2010).

The question as to whether visual speech contains lexical stress information that benefits word recognition therefore has to be addressed in a language-appropriate manner. For English, it may be satisfactory to test whether native speakers can visually detect vowel reduction. For languages such as Dutch, however, it is necessary to test whether native speakers are sensitive to suprasegmental stress in visual speech.

Suprasegmental cues allow for lexical stress to be implemented to various degrees. The degree to which suprasegmental stress cues are expressed distinguishes between syllables with primary stress and secondary stress, and unstressed syllables with no vowel reduction (see, e.g., Fear et al., 1995; Plag, Kunter, & Schramm, 2011; Rietveld, Kerkhoff, & Gussenhoven, 2004; Slootweg, 1988; van Heuven, 1987). Listeners are sensitive to distinctions between these fine-grained degrees of stress (Mattys, 2000; Mattys & Samuel, 1997). Distinguishing primary from secondary lexical stress is thus more difficult than distinguishing primary from no stress (e.g., Reinisch et al., 2010). Nevertheless, Dutch listeners take advantage of being able to distinguish primary from secondary stress in their resolution of lexical competition (Reinisch et al., 2010). We therefore tested here whether Dutch native speakers are sensitive to the distinction between primary lexical stress and no stress in visual speech and also whether they are sensitive to the more fine-grained visual distinction between primary and secondary stress.

Visual lexical stress

Most research in the area of audio-visual speech perception has focused on visual contributions to the recognition of individual speech segments. It is now well established that seeing a speaker aids comprehension (e.g., see Massaro & Jesse, 2008, for an overview). A limited number of studies have shown that visual speech also provides prosodic information. But, to the best of our knowledge, only two previous studies have addressed whether visual speech contains information about lexical stress. Risberg and Lubker (1978) demonstrated that, based on seeing the speaker alone, hearing-impaired and normal-hearing young adults were equally able to distinguish bisyllabic Swedish word pairs that differed in their placement of primary stress. This study was recently replicated and extended for English. Scarborough, Keating, Mattys, Cho, and Alwan (2009) presented English minimal noun–verb stress pairs [e.g., *(a)* ˈ*sub-ject* versus *(to) sub-*ˈ*ject*] and reiterant speech versions of these and of additional nonminimal pair words in a two-alternative forced choice task. In the reiterant versions, the words were mimicked using nonsense syllables [e.g., *(to) sub-*ˈ*ject* produced as *fer-*ˈ*fer*]. All items were bisyllabic and had primary stress either on their first or on their second syllable. Participants recognized the stress pattern of 62% of these items.

Importantly, stress placement was reflected by segmental changes in both of these prior studies. As appropriate and ecologically valid for investigating stress perception in English, the materials in the Scarborough et al. (2009) study contained vowel reduction as a segmental cue to stress. Primary lexical stress can therefore be perceived visually, at least when implemented segmentally. But these findings may not apply to the recognition of languages, such as Dutch, where segmental stress cues play a more limited role in stress perception.

Suprasegmental information can, however, be perceived in visual speech, at least at the phrasal level. Visual suprasegmental information can cue a variety of prosodic properties, such as whether or not an utterance is a statement or a question (Bernstein, Eberhardt, & Demorest, 1989; Fisher, 1969; Nicholson, Baum, Cuddy, & Munhall, 2002; Srinivasan & Massaro, 2003) or the location of word and phrase boundaries (Barkhuysen, Krahmer, & Swerts, 2005; Risberg & Lubker, 1978). Visual speech can also be informative about phrase-level emphasis—that is, the relative emphasis of words within a phrase

(Bernstein et al., 1989; Dohen & Loevenbruck, 2009; Dohen, Loevenbruck, Cathiard, & Schwartz, 2004; Krahmer & Swerts, 2007; Lansing & McConkie, 1999; Nicholson et al., 2002; Risberg & Lubker, 1978; Swerts & Krahmer, 2008; Thompson, 1934). Phrase-level emphasis of words is acoustically implemented through changes in pitch movement, duration, and loudness (Beckman, 1986)—that is, through suprasegmental changes. For instance, when the question "What did you get for your birthday?" is answered, "guinea pig" would be more emphasized than the other words in the response "I got a GUINEA PIG for my birthday". Participants can detect, only from seeing the speaker, which of the words in a subject–verb–object sentence received phrase-level emphasis (Dohen et al., 2004). In Dutch, phrase-level emphasis can be better detected from seeing its effects on the upper parts of the face, such as head and eyebrow movements, than from seeing its effects on the lower face (Swerts & Krahmer, 2008). Suprasegmental prosodic information therefore seems to be available in visual speech, at least at the phrasal level.

In the present study, we examined whether word-level suprasegmental information—that is, lexical stress information—can be detected in visual speech even if the target word does not receive phrase-level emphasis, but also whether phrase-level emphasis aids the distinction between various degrees of lexical stress. Pitch accents, the most important cue to phrase-level prominence of a word in Dutch (Cohen & 't Hart, 1967), for example, tend to fall on the syllable of that word with primary stress (Bolinger, 1958), although syllables with secondary stress can also receive such accents. Phrase-level emphasis on a Dutch word affects the duration of all syllables of that word, but stressed syllables are lengthened more than unstressed syllables (Cambier-Langeveld & Turk, 1999; Eefting, 1991; Sluijter & van Heuven, 1995). Phrase-level emphasis therefore seems to increase the acoustic difference between the different stress levels. Differences in pitch and intensity between primary and secondary stressed syllables are enhanced, for example, for emphasized compared to unemphasized English words (Plag et al., 2011). Some of the suprasegmental differences between stressed and unstressed syllables can also only be found when English or Dutch target words carry phrase-level emphasis (Plag et al., 2011; Sluijter & van Heuven, 1996). Not surprisingly, Dutch listeners are better at distinguishing, for example, stressed and unstressed syllables when the target word carries a phrase-level emphasis (van Heuven, 1987). The realization of lexical stress can hence be modulated by whether or not the word receives phrase-level emphasis.

In the present study, we therefore examined whether the perception of visual suprasegmental stress is affected by whether or not the target word is emphasized within the phrase. As phrase-level accent can be visually perceived and, at least in English, is better perceived than lexical stress (Scarborough et al., 2009), this manipulation also tests whether Dutch participants do indeed perceive visual prosodic cues to lexical stress or simply visual prosodic cues to phrase-level accent. If the former is true, then participants should be able to detect lexical stress on syllables even when the target word is not emphasized at the phrasal level.

In summary, we investigated whether suprasegmental cues to lexical stress can be detected in visual speech. More specifically, we tested whether, in Dutch, syllables with primary lexical stress can be distinguished visually from unstressed and from secondary-stressed syllables. Video recordings were made of a native speaker of Dutch saying Dutch words with matched onsets (the first two syllables) that formed minimal pairs differing only in stress. Fragments of the first two syllables of these videos, without audio, were presented to native Dutch participants in a two-alternative forced-choice identification task. In Experiment 1, all fragments were taken from words that carried phrase-level emphasis.

One set of pairs of word fragments had primary stress either on the first syllable of the fragment or on the (absent) third syllable; the latter had secondary stress on the first syllable [*primary–secondary stress contrast on the first syllable*; e.g., ˈ*ca-vi-(ˌa)* vs. ˌ*ka-vi-(ˈaar)*, "guinea pig"–"caviar"]. This set of items allowed us to test whether primary stress

can be distinguished from secondary stress in visual speech. To test whether syllables with primary stress can be distinguished from unstressed syllables, another set of pairs of word fragments had primary stress either on the second syllable or on the (absent) third syllable [*primary–unstressed stress contrast on the second syllable;* e.g., *pro-ˈjec-(tor)* vs. *ˌpro-jec-(ˈtiel)*; "projector"–"projectile"]. Note that fragments from the latter word pairs could also be distinguished based on the difference in their first syllables. This is because the words with primary stress on the third syllable all had secondary stress on their first syllable (as is usually the case for Dutch words; Booij, 1995). Nevertheless, if participants can identify these fragments correctly, they must do so using visual cues to suprasegmental lexical stress (i.e., in one or other or both syllables). Demonstrating that listeners can use cues to the suprasegmental structure of spoken words from the visual modality would place constraints on the further development of models of speech recognition, which currently do not include a role for such cues in lexical access.

EXPERIMENT 1

Method

Participants

Twenty university students were paid for their participation. Their average age was 21.75 years. All participants were native Dutch speakers, reported no language or hearing deficit, and had normal or corrected-to-normal vision.

Materials

Twenty-nine Dutch word pairs were selected as targets (see Table 1). Thirteen word pairs had primary stress on the first or the third syllable [*primary–secondary stress contrast on the first syllable*; e.g., *ˈca-vi-(ˌa)* vs. *ˌka-vi-(ˈaar)*]. Their initial two-syllable fragments had, respectively, either primary or secondary stress on the first syllable. Sixteen other word pairs had primary lexical stress on the second or the third syllable [*primary–unstressed stress contrast on the second syllable*; e.g., *pro-ˈjec-(tor)* vs. *ˌpro-jec-(ˈtiel)*]. The initial two-syllable fragments of these pairs had, respectively, either primary or no stress on the second syllable and either no stress or secondary stress on the first syllable. All of these stress contrast pairs are listed in the Appendix. Four additional word pairs from an additional stress contrast set were included [e.g., *ˈoc-to-(pus)* vs. *ok-ˈto- (ber)*]. Their fragments had primary lexical stress on the first or the second syllable. Given the limited number of suitable pairs available for this contrast set, these items were not included in any of the analyses. Note also that matched fragments of longer words were used rather than complete words because there are only a few Dutch minimal pairs of words that differ only in stress (Cutler & van Donselaar, 2001).

All words within a pair were semantically unrelated to each other. Words within a pair differed in their stress pattern, but overlapped visually in the segments up to the first phoneme of their third syllable. All but three target pairs shared the same first phoneme in the third syllable; these three target pairs shared the same viseme (/b/ and /m/; /x/ and /k/; /s/ and /t/[1]; van Son, Huiskamp, Bosman, & Smoorenburg, 1994).

Two-syllable fragments from each of six additional stress pairs were selected as practice items so that a third of the six practice fragments had primary stress on the first, second, or third syllable, respectively. The same practice fragments were presented to all participants. Three additional practice items were selected to serve as fillers on the first three trials of the main part of the experiment. These items also had primary stress on the first, second, or third syllable, respectively.

A female native speaker of Dutch was video-recorded saying all words embedded sentence-finally in semantically meaningful but low cloze probability sentences (e.g., "Het nieuwe schip droeg de naam projector/projectiel"; "The new

[1] Only good lip-readers can distinguish /s/ and /t/ visually (van Son et al., 1994). Although this did not change the general pattern of results, we excluded this item pair, e-ˈro-(sie) versus ˌe-ro-(ˈtiek), from all analyses in both experiments.

Table 1. *Design of Experiments 1 and 2*

Stress contrast	*Primary stress presence in fragment*	*Example*	*Stress status*	
			First syllable	*Second syllable*
Primary–secondary contrast on first syllable	Present	*ˈca-vi-(ˌa)* ["guinea pig"]	Primary	Unstressed
	Absent	*ˌka-vi-(ˈaar)* ["caviar"]	Secondary	Unstressed
Primary-unstressed contrast on second syllable	Present	*pro-ˈjec-(tor)* ["projector"]	Unstressed	Primary
	Absent	*ˌpro-jec-(ˈtiel)* ["projectile"]	Secondary	Unstressed

Note: Fragments received sentence-level emphasis in Experiment 1 but not in Experiment 2.

ship had the name projector/projectile"). The same carrier sentence was used for both items within a word pair. The speaker was not instructed to realize the sentences in any systematic way but nevertheless gave all the critical words phrase-final emphasis. Videos showed a close-up of the speaker's face and shoulders. Videos were digitized to uncompressed avi files (PAL format, 720 × 576). The audio sampling rate was 48 kHz. The acoustic onset and offset of the first two syllables of each word pair were annotated. Videos were cut such that they started with the frame containing the auditory onset of the fragment and ended with the last frame containing auditory information about the fragment. Final videos were converted to mpg1 format.

Design and procedure

Participants were tested individually in a sound-attenuated booth. NESU experimental software controlled the experiment on a PC. Videos were presented as 720 × 576-pixel displays centred horizontally on a black background in the upper part of a 17″ CRT computer monitor. The audio channel was muted.

At the beginning of each trial, participants were presented with two response alternatives (e.g., "projector", "projectiel") at the bottom of the screen (Arial, font size 48). These response alternatives were printed in white, with their first two syllables marked in red to indicate the length of the to-be-presented fragment. Participants were instructed that they would see a speaker say the beginning of one of the two displayed words and that these beginnings would always be marked in red on the screen. They were told that the beginnings of each word pair would be pronounced the same, but that they would sometimes be spelled differently. The participants' task was to indicate by pressing one of two corresponding coloured keys on the computer keyboard whether what the speaker had said was the beginning of the word displayed on the right or of the word displayed on the left. Stress was not mentioned in the instructions and was not marked in the response alternatives. The response alternatives were presented on the screen for 40 ms, disappeared for another 40 ms, and were then shown again until the end of a trial. The alternatives thus flashed to signal the beginning of a trial. A period of 2500 ms after the second presentation of the response alternatives, a fixation cross was shown for 250 ms, centred where the video would appear immediately afterwards. After the offset of the video presentation, participants were given 3000 ms to respond. Participants were instructed to respond as fast and as accurately as possible. No feedback was given. The next trial was presented 200 ms after a response or 100 ms after the time out.

The experiment always began with a practice block, consisting of the same six trials presented in a newly randomized order for each participant. The main experiment then started with three filler trials that were identical across participants. The order of the subsequently presented target trials was newly randomized for every group of four participants. Two participants within each group gave responses to one item in a given pair (with the correct response alternative shown on either side across the two participants); the other

two participants gave responses to the other item in that pair. Overall, half of the correct responses were displayed on each side for each participant. Also, half of all target trials had primary stress on the fragment; for the other half this was not the case. Each participant was presented with only one of the items from a pair, but the presentation of pairs to participants was counterbalanced.

Analyses

Linear mixed effect models were implemented using the lmer function in the lme4 package (Bates & Sarkar, 2009) of the R statistical program (Version 2.8.0; R Development Core Team, 2007). Given the categorical nature of the dependent variable, a binomial logit linking function between responses and predictors was included in the models. Models were fitted using the residual maximum likelihood criterion. Stress presence (primary stress placed on the fragment or not) was evaluated as a fixed factor. The levels of this factor were contrast-coded as –0.5 and 0.5, respectively.

Additionally, number of phonemes and duration of the fragments, number of syllables of the target words, CELEX word form frequency, and CELEX lemma frequency (Baayen, Piepenbrock, & van Rijn, 1995) were evaluated as numerical control variables. None of these control variables had an effect of performance in Experiment 1 or 2; they are therefore not discussed further.

All models contained the design's maximal random effect structure (Barr, Levy, Scheeper, & Tily, 2013). This structure allowed for specific adjustments to the regression weights based on the subject's or item's means and by-subject slope adjustments for stress presence. The latter random effect was not included in the comparisons of performance relative to chance level, as these comparisons were conducted separately for each level of the stress presence factor.

Results

The primary–secondary contrast pair "ave" ("averechts" and "averij") was removed from all data analyses, since "averechts" had been produced with a different stress pattern from that intended. Figure 1 shows the mean percentage of correct responses as

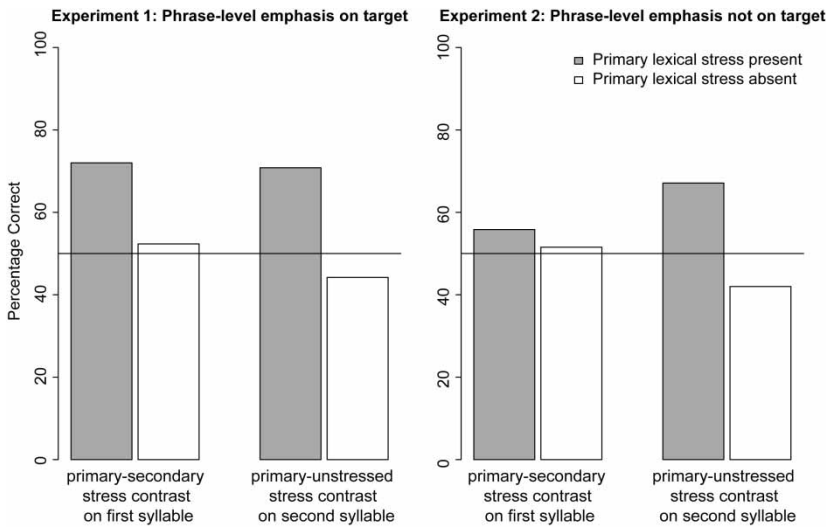

Figure 1. *Mean percentages of correct responses for each stress contrast by stress presence condition for Experiments 1 and 2. The horizontal line indicates the chance level of 50%.* Primary–secondary stress contrast on the first syllable *refers to fragments such as* ˈca-vi-(ˌa) vs. ˌka-vi-(ˈaar). Primary-unstressed contrast on the second syllable *refers to fragments such as* pro-ˈjec-(tor) vs. ˌpro-jec-(ˈtiel). *These item pairs also differed in the stress of their first syllable, which was unstressed or carried secondary stress.*

a function of stress presence for each stress contrast type. Participants were able to detect stress when it was present (primary–secondary contrast on first syllable: $M = 72\%$; primary-unstressed contrast on second syllable: $M = 70.8\%$) but not when it was absent (primary–secondary contrast on first syllable: $M = 52.3\%$; primary-unstressed contrast on second syllable: $M = 44.2\%$).

Performance was better for both types of stress contrasts when the presented fragment contained stress than when it did not (primary–secondary contrast on first syllable: $\beta = -0.88$, $SE = 0.32$, $p < .01$; primary-unstressed contrast on second syllable: $\beta = -1.33$, $SE = 0.45$, $p < .01$). Intercept-only models for each stress presence condition for each stress contrast showed that the recognition of the presence of stress was significantly better than chance (primary–secondary contrast on first syllable: $\beta = 0.97$, $SE = 0.22$, $p < .0001$; primary-unstressed contrast on second syllable: $\beta = 1.07$, $SE = 0.32$, $p < .001$), but not the recognition of absence of primary stress in the fragment (primary–secondary contrast on first syllable: $\beta = 0.10$, $SE = 0.23$, $p = .66$; primary-unstressed contrast on second syllable: $\beta = -0.30$, $SE = 0.31$, $p = .34$).

Discussion

The results showed that suprasegmental lexical stress can be detected from visual speech in Dutch. Segmental changes due to vowel reduction in unstressed syllables (Scarborough et al., 2009) are hence not necessary for the visual perception of a word's lexical stress pattern. Suprasegmental primary stress was detected in visual speech when stress was present in the fragment. Its absence, however, was not informative. Performance was at chance level when the word fragments did not contain primary stress. Given this at-chance performance, the other, positive results cannot be explained by an overall bias to give a stress response. Furthermore, measures of word frequency, number of phonemes, and overall duration of the video did not predict identification performance.

The present results also provide evidence for the first time that primary and secondary suprasegmental stress can be distinguished visually. Participants were able to detect primary stress on the first syllable when the alternative word choice contained secondary stress on that syllable (the primary–secondary contrast on the first syllable condition). Correct identification rates were about the same as those for the word pairs with either primary stress on the second syllable or no stress on that syllable (the primary-unstressed stress contrast). Although we cannot rule out the possibility that correct identification in this latter condition reflects participants' use of visual stress cues in the first syllable (a contrast between secondary-stressed and unstressed syllables), or indeed the contrast between the two syllables of a given fragment, the similarity of the performance levels across the two sets of pairs suggests that participants in both conditions were relying mainly on visual cues to primary lexical stress.

Seeing a speaker can therefore help word recognition by providing information about lexical stress. In Experiment 1, however, the fragments were taken from words that had received phrase-level emphasis because they were utterance final, and the speaker naturally emphasized them. In Experiment 2, the fragments were taken from words without phrase-level emphasis. This allowed us to ask whether what participants detected in Experiment 1 were indeed visual lexical stress cues or rather were visual cues to phrase-level emphasis. If the latter were the case, then participants should perform at chance level in Experiment 2, both for fragments containing primary lexical stress and for those without primary stress cues. Alternatively, phrase-level emphasis could aid the perception of stress, especially when the distinction between different degrees of stress is subtler, as is the case for primary compared to secondary stress (i.e., for the pairs in the primary–secondary stress on the first syllable set). If phrase emphasis aids stress perception, stress should be detected better in Experiment 1 than in Experiment 2.

EXPERIMENT 2

Method

Participants

Twenty new native Dutch speakers from the same population as that in Experiment 1 were paid for their participation. Their average age was 19.7 years.

Materials

The stress pairs used in Experiment 1 were re-recorded spoken by the same female speaker of Dutch. Target words were still recorded in sentence-final position. In contrast to the Experiment 1 materials, phrase-level emphasis was placed on parts of the carrier sentences that preceded the target words. This was achieved by creating context questions that repeated all parts of the target-bearing sentence except the word that was to receive the emphasis in the target-bearing sentence. For example, to elicit a phrase-level emphasis on the word "galgje" ("hangman"), the question "Waarvoor koos je het woord cavia/kaviaar?" ("For what are you choosing the word guinea pig/caviar?") was presented to the speaker before she recorded the sentence "Voor *galgje* koos ik het word cavia/kaviaar" ("For hangman I chose the word guinea pig/caviar"). A linguistically trained native Dutch speaker verified that the accents had been assigned correctly in all sentences and that each target word had not received a phrase-level emphasis.

Design and procedure

The design and procedure were the same as those for Experiment 1. The fragment "ave" of "averij" was included in the trial order randomization; "ave" from "averechts", however, was always presented as the last trial in the experiment, since it had been recorded with an incorrect stress pattern. As in Experiment 1, this stress pair was excluded from all analyses.

Results

Figure 1 shows the mean percentage of correct responses by stress contrast and stress presence. Performance was at chance level when the presented fragment did not contain primary stress (primary–secondary contrast on first syllable: $M = 51.5\%$; primary-unstressed contrast on second syllable: $M = 42\%$). Participants were able to distinguish primary stress from no stress (primary-unstressed contrast on second syllable: $M = 67.1\%$), but not from secondary stress (primary–secondary contrast on first syllable: $M = 55.8\%$).

The analysis methods were identical to those of Experiment 1. For pairs in the primary-unstressed contrast on the second syllable condition, recognition of fragments was better when they contained primary stress than when they did not ($\beta = -1.08$, $SE = 0.27$, $p < .0001$). This difference was not found for the pairs with either primary or secondary stress on their first syllable ($\beta = -0.17$, $SE = 0.28$, $p = .54$). For both types of stress contrasts, participants performed at chance when no primary stress was present in the fragments (primary–secondary contrast on first syllable: $\beta = 0.06$, $SE = 0.18$, $p = .73$; primary-unstressed contrast on second syllable: $\beta = -0.29$, $SE = 0.23$, $p = .2$). Participants were also not able to recognize the presence of primary stress in fragments when it had to be distinguished from secondary stress on the same syllable ($\beta = 0.26$, $SE = 0.24$, $p = .29$). Participants were, however, able to recognize primary lexical stress on the second syllable when it had to be distinguished from fragments that had no stress on the second syllable but had secondary stress on the first syllable ($\beta = 0.75$, $SE = 0.21$, $p < .001$).

To assess the effect of sentence-level emphasis on the perception of visual lexical stress, we conducted planned comparisons across the two experiments on the conditions with primary stress present, separately for the two stress contrasts. For both contrasts, the model included experiment as a fixed, contrast-coded factor. Adjustments to the intercept by item and by subject were modelled as well as by-subject slope adjustments for the experiment factor. For the primary–secondary contrast on the first syllable pairs, performance was significantly worse ($\beta = -0.72$, $SE = 0.28$, $p < .01$) when the fragment came from a target word that had not received sentence-level emphasis

(Experiment 2) than when it came from one that had (Experiment 1). For the pairs in the primary-unstressed contrast on the second syllable condition, no difference was found across experiments ($\beta = -0.33$, $SE = 0.41$, $p = .42$). Phrase-level emphasis thus helped participants distinguish primary stress from secondary stress, but did not help them distinguish primary stress from no stress (or no stress from secondary stress).

Discussion

The results of Experiment 2 replicated the finding from Experiment 1 that the presence of primary suprasegmental lexical stress but not its absence is informative in visual speech. However, only the presence of stress for pairs in the primary-unstressed contrast on the second syllable set could be correctly detected independently of whether or not the word received phrase-level emphasis. That is, even without phrase-level accent, suprasegmental primary stress could be distinguished from no stress. (Note once again, however, that these pairs also differ in the stress patterns of their first syllables, which either are unstressed or carry secondary stress). In contrast, primary suprasegmental stress could be distinguished from secondary stress only when the stress-bearing words received phrase-level accents (the primary–secondary contrast pairs in Experiment 1). Stress could not be successfully detected for the stressed fragments of the primary–secondary pairs in Experiment 2. Taken together, these findings suggest that phrase-level accents help participants distinguish the various degrees of lexical stress in visual speech but also that sentence-level emphasis is not necessary for perceivers to detect that a visual syllable is stressed.

GENERAL DISCUSSION

Lexical stress helps with spoken-word recognition in variable-stress languages such as Dutch. Listeners of variable-stress languages take lexical stress information into account to resolve the competition among possible candidate words (e.g., van Donselaar et al., 2005). But listeners often not only hear but also see a speaker talk. In such situations, listeners typically process and combine speech information from both modalities. Seeing a speaker can generally help with recognizing speech (e.g., Reisberg et al., 1987) by providing information about the speech segments (e.g., Jesse & Massaro, 2010) and by providing information about suprasegmental modulation at the phrase level (e.g., Krahmer & Swerts, 2007). In the present study, we showed for the first time that seeing a speaker also supplies suprasegmental information on the word level—namely, information about the lexical stress patterns of spoken words.

In Dutch, lexical stress is primarily implemented as suprasegmental changes. We showed here, in two experiments, that Dutch speakers were able to obtain suprasegmental lexical stress information from seeing a talking face. The presence of primary lexical stress was reliably detected from seeing a speaker. Participants seemed to be able to distinguish primary lexical stress and no stress, but were also sensitive to the more fine-grained distinction between primary and secondary lexical stress, at least when the target word also carried phrase-level emphasis. Participants could not reliably detect secondary lexical stress or the absence of primary stress. This suggests that it was indeed the presence of primary lexical stress in the condition with the primary-unstressed contrast on the second syllable that cued the identity of the fragments with primary lexical stress on the second syllable. Fragments in this condition either had unstressed first syllables and primary lexical stress on the second syllables, or had secondary stress on their first syllables and unstressed second syllables. It is hence possible that participants identified fragments of the former kind because they had primary lexical stress on their second syllable, or because they had no stress on their first syllable. This latter possibility seems unlikely though, as participants should then also have been able to reliably identify fragments with secondary stress on the first syllable based on their unstressed second syllables. Participants are at chance performance when asked to identify these words. It is, however, possible that participants

relied on the relative emphasis on the first and the second syllable when identifying any fragments with primary lexical stress, rather than detecting only cues to primary lexical stress. We cannot rule out this latter possibility, but note that it is nevertheless consistent with our main conclusion that participants can reliably identify stress patterns from visual speech if they involve primary lexical stress.

Our finding that words can be identified in visual speech based on the presence of primary lexical stress but not based on its absence dovetails nicely with similar results observed for the auditory perception of lexical stress. Reinisch et al. (2010) found, for example, that words with noninitial stress competed less strongly for recognition than words with initial stress. *Kaviaar*, for example, competed less for recognition when the listener heard the beginning ˈ*ca-vi* of ˈ*ca-vi-ˌa*, than *cavia* did when the listener heard the beginning ˌ*ka-vi* of ˌ*ka-vi-*ˈ*aar*. Primary lexical stress information in a presented fragment was used to reduce lexical competition of noninitial stress candidate words more efficiently than the absence of lexical stress information in a presented fragment was used to reduce the competition of initial stress candidate words. Similarly, gating studies have shown that stressed syllables are more helpful in determining the spoken target word earlier than unstressed syllables (McAllister, 1991).

The result that only the presence of lexical stress and not its absence was detected by Dutch participants in the talking face is unlikely to reflect a bias to detect stress, since performance was at chance level when stress was absent in the signal. Rather, the presence of stress may be more reliably detected than the absence of stress because syllables with lexical stress tend to be produced with more precise articulation (e.g., Brown, 1977). Consequently, syllables with stress are more reliably recognized (Lieberman, 1963). For example, listeners detect clicks and mispronunciations more accurately in stressed than in unstressed syllables (Bond, 1971; Cole, Jakimik, & Cooper, 1978). The presence of stress is hence more informative than the absence of stress when perceived by ear and by eye. By providing information about the presence of primary lexical stress, visual speech thus has the potential to help with spoken-word recognition.

Phrase-level emphasis increases acoustic differences between different stress levels (e.g., Plag et al., 2011; Sluijter & van Heuven, 1996) and helps the listener in making auditory stress distinctions (van Heuven, 1987). In the present study, we found that phrase-level emphasis on the target word helped to distinguish primary from secondary lexical stress in visual speech. For word pairs that differed in whether primary or secondary lexical stress was placed on their first syllable, primary stress was only reliably detected when the target word had received phrase-level emphasis (Experiment 1). Without phrase-level emphasis, performance was at chance level no matter whether the fragment contained primary or secondary lexical stress (Experiment 2). Phrase-level accents were not needed to distinguish primary stress from no stress. This also suggests that what participants were able to perceive visually was indeed lexical stress and not phrase-level emphasis. Phrase-level accents also did not aid the perceiver in making a distinction between primary and no lexical stress. A likely explanation is that the contrast between primary and no stress was already salient and hence that cues to this distinction were available, even without the phrase-level emphasis. If participants performed as well as they could ever possibly do in the condition without phrase-level emphasis, then adding this emphasis could not improve scores further. Phrase-level emphasis is thus needed only to make more fine-grained distinctions about the degree of lexical stress in visual speech. Phrase-level accents increase visual differences between different stress levels and can thus help us perceive a stress distinction better when we see a speaker talk.

How can phrase-level emphasis help with the visual detection of lexical stress? Syllables with primary lexical stress seem to be indicated by larger opening of the lips and larger movement of the chin than syllables with no stress (Scarborough et al., 2009). Syllables with primary lexical stress are also produced with faster chin

movements than syllables with no stress. This production analysis was, however, conducted for English, where, unlike in Dutch, unstressed syllables are almost always produced with a schwa rather than with a full vowel. Furthermore, only productions of syllables with primary lexical stress and no stress were compared. Phrasal-level emphasis affects the extent of head and eyebrow movements, but has also articulatory effects on lip and chin (Scarborough et al., 2009). Phrase-level emphasis is generally known to result in an increase in articulatory effort (e.g., Cho, 2005; de Jong, 1995; Erickson, 1998; Fowler, 1995; Harrington, Fletcher, & Roberts, 1995; Kelso, Vatikiotis-Bateson, Saltzman, & Kay, 1985; Summers, 1987). For emphasized (monosyllabic English) words, the lip and jaw movements are produced, for example, with larger and longer openings and with higher peak velocities than they are for unemphasized words (Fowler, 1995). Seeing articulatory movements is sufficient to detect phrase-level emphasis (Dohen et al., 2004; Lansing & McConkie, 1999; Swerts & Krahmer, 2008). The effects of phrase-level emphasis on articulatory movements could strengthen the articulatory and hence the visible differences between different degrees of stress.

The present study was not designed to address which suprasegmental cues to lexical stress are present on the speaker's face. The recordings were optimized for studying perception and not production. We hence did not make recordings with markers on the speaker's face that would have allowed us to track their motion for production analyses. These markers could have led to artefacts in our perceptual studies. Instead, our study shows that visual stress cues can be detected from unmarked video recordings. An analysis of these recordings is unlikely to be informative, given the low temporal resolution (40-ms frames) and that only one token of each word pair was recorded in each experiment. Furthermore, such an analysis could only provide correlates of visual stress and could not identify which of these correlates the perceivers actually used. For this purpose, an artificial talking head with which potential stress cues could be manipulated systematically would have to be used (for a similar study on intonation, see, e.g., Srinivasan & Massaro, 2003). Future studies should clarify what the visible articulatory differences in Dutch are between the different degrees of lexical stress and how phrase-level emphasis impacts these differences to help with the perception of spoken words.

Visual speech thus provides information about the lexical stress of words and can help listeners recognize words in variable-stress languages such as Dutch. This is important as Dutch listeners rely on lexical stress to resolve lexical competition (Cutler & van Donselaar, 2001; Reinisch et al., 2010; van Donselaar et al., 2005). Visual stress information, similarly to auditory stress information across variable-stress languages (Cooper et al., 2002; Reinisch et al., 2010; Soto-Faraco et al., 2001; Sulpizio & McQueen, 2012), should modify the competition process among word candidates. Visual stress information should help with the recognition of a word by providing additional support for it and against its stress-mismatching competitors. This should lead to faster and more accurate recognition. There is, however, currently no formal account of how audio-visual prosody helps with word recognition. The processing of audio-visual prosody has been implemented in the fuzzy logical model of perception (FLMP; Massaro, 1998) such that visual and auditory signals are first evaluated for prosodic information independently from one another. The integration of auditory and visual prosodic information then follows the same algorithm as the integration of segmental information (see e.g., Chen & Massaro, 2008; Srinivasan & Massaro, 2003). How this information affects lexical processing, however, has not been specified. One possibility is that this prosodic analysis works in parallel to a segmental analysis of the speech signals (for similar proposals for the processing of auditory prosodic boundary information, see Cho, McQueen, & Cox, 2007; Christophe, Peperkamp, Pallier, Block, & Mehler, 2004; Salverda, Dahan, & McQueen, 2003). In a model like the FLMP (see Jesse, 2005), but also in Shortlist B (Norris & McQueen, 2008), for example, the segmental and the prosodic analysers could independently inform

lexical processing about the support for lexical candidates provided by auditory and visual signals. Existing models of audio-visual speech perception and of auditory word recognition thus need to be expanded to account for how visual cues to lexical stress constrain recognition. In particular, the present results suggest that models of speech recognition need to be expanded not only to include the uptake of suprasegmental as well as segmental information (see, e.g., Cho et al., 2007) and to include the uptake of visual as well as auditory information (see, e.g., Massaro, 1998), but also to include the combination of the two: the uptake of visual cues to suprasegmental structures.

In summary, three main conclusions can be drawn. Most importantly, visual speech provides information about suprasegmental lexical stress. In particular, primary lexical stress can be reliably recognized even when participants only see the speaker. Secondly, visual speech provides information only about the presence of primary stress and not about the presence of secondary stress or the absence of any stress. Thirdly, primary lexical stress can be distinguished visually from secondary lexical stress, at least when the word receives phrase-level emphasis. But phrase-level emphasis is not necessary for distinguishing a visual syllable with primary stress from one with no stress. Seeing a speaker can therefore guide spoken-word recognition by providing the listener with suprasegmental information about the stress patterns of words.

Original manuscript received 5 May 2012
Accepted revision received 24 July 2013
First published online 21 October 2013

REFERENCES

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). *The CELEX lexical database* [CD-ROM]. Philadelphia, PA: University of Pennsylvania, Linguistic Data Consortium.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2005). The interplay between the auditory and visual modality for end-of-utterance detection. *Journal of the Acoustical Society of America*, *123*, 354–365. doi:10.1121/1.2816561

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D. M., & Sarkar, D. (2009). lme4: Linear mixed-effects models using s4 classes (R package version 0.999375-27) [Computer software]. Retrieved from http://cran.r-project.org/web/packages/lme4/index.html

Beckman, M. E. (1986). *Stress and non-stress accent*. Dordrecht, The Netherlands: Foris. doi:10.1515/9783110874020

Bernstein, L. E., Eberhard, 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. doi:10.1121/1.397690

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

Bond, Z. S. (1971). Units in speech perception. *Ohio State University Working Papers in Linguistics*, *9*, 1–112.

Booij, G. (1995). *The phonology of Dutch*. Oxford: Oxford University Press.

Brown, G. (1977). *Listening to spoken English*. London: Longman.

Cambier-Langeveld, T., & Turk, A. E. (1999). A cross-linguistic study of accentual lengthening: Dutch vs. English. *Journal of Phonetics*, *27*, 255–280. doi:10.1006/jpho.1999.0096

Chen, T., & Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *Journal of the Acoustical Society of America*, *123*, 2356–2366. doi:10.1121/1.2839004

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. doi:10.1121/1.1861893

Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically-driven detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*, *35*, 210–243. doi:10.1016/j.wocn.2006.03.003

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access. I. Adult data. *Journal of Memory and Language*, *51*, 523–547. doi:10.1016/j.jml.2004.07.001

Cohen, A., & 't Hart, J. (1967). On the anatomy of intonation. *Lingua*, *19*, 177–192. doi:10.1016/0024-3841(69)90118-1

Cole, R., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *Journal of the Acoustical Society of America*, *64*, 44–56. doi:10.1121/1.381955

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech*, *45*, 207–228. doi:10.1177/00238309020450030101

Cutler, A. (1986). Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech*, *29*, 201–220.

Cutler, A., & Clifton, C. E. (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 183–196). Hillsdale, NJ: Erlbaum.

Cutler, A., & van Donselaar, W. (2001). Voornaam is not a homophone: Lexical prosody and lexical access in Dutch. *Language and Speech*, *44*, 171–195. doi:10.1177/00238309010440020301

Cutler, A., & Pasveer, D. (2006). Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. In R. Hoffman & H. Mixdorff (Eds.), *Proceedings of the Third International Conference on Speech Prosody* (pp. 250–254). Dresden, Germany: TUD Press.

Dohen, M., & Loevenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech*, *52*, 177–206. doi:10.1177/0023830909103166

Dohen, M., Loevenbruck, H., Cathiard, M.-A., & Schwartz, J.-L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication*, *44*, 155–172. doi:10.1016/j.specom.2004.10.009

van Donselaar, W., Koster, M., & Cutler, A. (2005). Exploring the role of lexical stress in lexical recognition. *The Quarterly Journal of Experimental Psychology*, *58A*, 251–273.

Eefting, W. (1991). The effect of "information value" and "accentuation" on the duration of Dutch words, syllables, and segments. *Journal of the Acoustical Society of America*, *89*, 411–424. doi:10.1121/1.400475

Erickson, D. (1998). Effects of contrastive emphasis on jaw opening. *Phonetica*, *55*, 147–169. doi:10.1159/000028429

Fear, B. D., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America*, *97*, 1893–1904. doi:10.1121/1.412063

Fisher, C. G. (1969). The visibility of terminal pitch contour. *Journal of Speech and Hearing Research*, *12*, 379–382.

Fowler, C. (1995). Acoustic and kinematic correlates of contrastive stress accent in spoken English. In F. Bell-Berti & L. J. Raphael (Eds.), *Producing speech: Contemporary issues. For Katherine Safford Harris* (pp. 355–373). New York, NY: AIP Press.

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

Harrington, J., Fletcher, J., & Roberts, C. (1995). Coarticulation and the accented/unaccented distinction: Evidence from jaw movement data. *Journal of Phonetics*, *23*, 305–322. doi:10.1016/S0095-4470(95)80163-4

van Heuven, V. J. (1988). Effects of stress and accent on the human recognition of word fragments in spoken context: Gating and shadowing. In W. A. Ainsworth & J. N. Holmes (Eds.), *Proceedings of Speech '88, 7th FASE symposium* (pp. 811–818). Edinburgh: Institute of Acoustics.

van Heuven, V. J., & Hagman, P. J. (1988). Lexical statistics and spoken word recognition in Dutch. In P. Coopmans & A. Hulk (Eds.), *Linguistics in the Netherlands 1988* (pp. 59–69). Dordrecht, the Netherlands: Foris.

Jesse, A. (2005). *Towards a lexical fuzzy logical model of perception: The time-course of information in lexical identification of face-to-face speech* (Unpublished doctoral dissertation). University of California, Santa Cruz, CA.

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics*, *72*, 209–225. doi:10.3758/APP.72.1.209

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, *97*, 491–504. doi:10.1121/1.412275

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E., & Kay, B. A. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America*, *77*, 266–280. doi:10.1121/1.392268

Krahmer, E. J., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*, 396–414. doi:10.1016/j.jml.2007.06.005

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

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, *33*, 451–454. doi:10.1121/1.1908095

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*, 172–187.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioural principle*. Cambridge, MA: MIT Press.

Massaro, D. W., & Jesse, A. (2008). Audiovisual speech perception and word recognition. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 19–35). Oxford: Oxford University Press.

Mattys, S. L. (2000). The perception of primary and secondary stress in English. *Perception & Psychophysics*, *62*(2), 253–265. doi:10.3758/BF03205547

Mattys, S. L., & Samuel, A. G. (1997). How lexical stress affects speech segmentation and interactivity: Evidence from the migration paradigm. *Journal of Memory and Language*, *36*, 87–116. doi:10.1006/jmla.1996.2472

McAllister, J. (1991). The processing of lexically stressed syllables in read and spontaneous speech. *Language and Speech*, *34*, 1–26.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. doi:10.1038/264746a0

Miller, G. A., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, *27*, 338–352. doi:10.1121/1.1907526

NESU [Computer software]. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics. Retrieved from http://www.mpi.nl/world/tools/nesu.html

Nicholson, K. G., Baum, S., Cuddy, L . L., & Munhall, K. G. (2002). A case of impaired auditory and visual speech prosody perception after right hemisphere damage. *Neurocase*, *8*, 314–322. doi:10.1076/neur.8.3.314.16195

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357–395. doi:10.1037/0033-295X.115.2.357

Plag, I., Kunter, G., & Schramm, M. (2011). Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics*, *39*, 362–374. doi:10.1016/j.wocn.2011.03.004

R Development Core Team. (2007). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *The Quarterly Journal of Experimental Psychology*, *64*(4), 772–783. doi:10.1080/17470210903104412

Reisberg, D., McLean, J. M., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rietveld, T., Kerkhoff, J., & Gussenhoven, C. (2004). Word prosodic structure and vowel duration in Dutch. *Journal of Phonetics*, *32*, 349–371. doi:10.1016/j.wocn.2003.08.002

Risberg, A. M., & Lubker, J. (1978). Prosody and speech-reading. *Speech Transmission Laboratory Quarterly Progress Status Report*, *4*, 1–16.

Salverda, A.-P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*, 51–89. doi:10.1016/S0010-0277(03)00139-2

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech*, *52*, 135–175. doi:10.1177/0023830909103165

Slootweg, A. (1988). Metrical prominence and syllable duration. In P. Coopmans & A. Hulk (Eds.), *Linguistics in the Netherlands 1988* (pp. 139–148). Dordrecht: Foris.

Sluijter, A. M., & van Heuven, V. J. (1995). Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica*, *52*, 71–89. doi:10.1159/000262061

Sluijter, A. M., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, *100*, 2471–2485. doi:10.1121/1.417955

van Son, N., Huiskamp, T. M. I., Bosman, A. J., & Smoorenburg, G. F. (1994). Viseme classifications of Dutch consonants and vowels. *Journal of the*

Acoustical Society of America, 96, 1341–1355. doi:10.1121/1.411324

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language*, *45*, 412–432. doi:10.1006/jmla.2000.2783

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. doi:10.1177/00238309030460010201

Sulpizio, S., & McQueen, J. M. (2012). Italians use abstract knowledge about lexical stress during spoken-word recognition. *Journal of Memory and Language*, *66*, 177–193. doi:10.1016/j.jml.2011.08.001

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215. doi:10.1121/1.1907309

Summers, W. V. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *Journal of the Acoustical Society of America*, *82*, 847–863. doi:10.1121/1.395284

Swerts, M., & Krahmer, E. J. (2008). Facial expressions and prosodic prominence: Comparing modalities and facial areas. *Journal of Phonetics*, *36*, 219–238. doi:10.1016/j.wocn.2007.05.001

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. doi:10.1080/00221309.1934.9917823

Walden, B. E., Prosek, R. A., & Worthington, D. W. (1974). Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech & Hearing Research*, *17*, 270–278.

APPENDIX

Stress contrast pairs

Pairs with primary–secondary contrast on the first syllable: ˈal-ge-(ˌbra) vs. ˌal-ge-(ˈmeen); [ˈa-ve-(ˌrechts) vs. ˌa-ve-(ˈrij)]*; ˈca-vi-(ˌa) vs. ˌka-vi-(ˈaar); ˈcen-ti-(ˌme-ter) vs. ˌsen-ti-(ˈment); ˈdi-a-(ˌme-ter) vs. ˌdi-a-(ˈmant); ˈdo-mi-(ˌnee) vs. ˌdo-mi-(ˈnant); ˈdu-bi-(ˌo) vs. ˌdu-bi-(ˈeus); ˈe-thi-(ˌcus) vs. ˌe-ti-(ˈket); ˈge-ni-(ˌtief) vs. ˌge-ni-(ˈtaal); ˈhos-pi-(ˌtaal) vs. ˌhos-pi-(ˈtant); ˈin-di-(ˌgo) vs. ˌin-di-(ˈges-tie); ˈo-pe-(ˌra) vs. ˌo-pe-(ˈra-tie); ˈra-di-(ˌus) vs. ˌra-di-(ˈa-tor); ˈspi-ri-(ˌtus) vs. ˌspi-ri-(ˈtist).

Pairs with primary-unstressed contrast on the second syllable: ca-ˈbi-(ne) vs. ˌca-bi-(ˈnet); co-ˈmmu-(nie) vs. ˌco-mmu-(ˈnist); di-ˈplo-(ma) vs. ˌdi-plo-(ˈmaat); di-ˈsci-(pel) vs. ˌdi-sci-(ˈpli-ne); dy-ˈna-(mo) vs. ˌdy-na-(ˈmiet); [e-ˈro-(sie) vs. ˌe-ro-(ˈtiek)]*; for-ˈmu-(le) vs. ˌfor-mu-(ˈlier); kas-ˈtan-(je) vs. ˌcas-tagn-(ˈett-en); ko-ˈlo-(nie) vs. ˌko-lo-(ˈnel); ma-ˈri-(ne) vs. ˌma-ri-(ˈna-de); pa-ˈra-(de) vs. ˌpa-ra-(ˈdijs); po-ˈli-(tie) vs. ˌpo-li-(ˈtiek); pro-ˈjec-(tor) vs. ˌpro-jec-(ˈtiel); re-ˈli-(gie) vs. ˌre-li-(ˈkwie); sa-ˈla-(mi) vs. ˌsa-la-(ˈman-der).

Note. * = excluded from analyses.