Cognition 122 (2012) 442–453

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody ☆

Erin Cvejic *, Jeesun Kim, Chris Davis

MARCS Auditory Laboratories, University of Western Sydney, Australia

ARTICLE INFO

Article history:
Received 14 March 2011
Revised 25 November 2011
Accepted 30 November 2011
Available online 21 December 2011

Keywords:
Visual prosody
Cross-modality prosody matching
Cross-speaker prosody matching
Cue distribution

ABSTRACT

Prosody can be expressed not only by modification to the timing, stress and intonation of auditory speech but also by modifying visual speech. Studies have shown that the production of visual cues to prosody is highly variable (both within and across speakers), however behavioural studies have shown that perceivers can effectively use such visual cues. The latter result suggests that people are sensitive to the type of prosody expressed despite cue variability. The current study investigated the extent to which perceivers can match visual cues to prosody from different speakers and from different face regions. Participants were presented two pairs of sentences (consisting of the same segmental content) and were required to decide which pair had the same prosody. Experiment 1 tested visual and auditory cues from the same speaker and Experiment 2 from different speakers. Experiment 3 used visual cues from the upper and the lower face of the same talker and Experiment 4 from different speakers. The results showed that perceivers could accurately match prosody even when signals were produced by different speakers. Furthermore, perceivers were able to match the prosodic cues both within and across modalities regardless of the face area presented. This ability to match prosody from very different visual cues suggests that perceivers cope with variation in the production of visual prosody by flexibly mapping specific tokens to abstract prosodic types.

1. Introduction

Prosody is a broad term used to describe systematic modifications to the way that speakers utter words in order to specify or disambiguate the meaning of an utterance without altering its segmental components. For example, speakers can indicate newness or importance by emphasizing a particular constituent and thus increase its salience relative to other words (Wagner & Watson, 2010), or phrase an utterance as a question without the use of an interrogative pronoun (i.e., an echoic question). Prosody is typically understood in terms of changes in acoustic features such as fundamental frequency (F_0), amplitude, duration and vowel space (for more details see Cooper, Eady, & Mueller, 1985; Cvejic, Kim, & Davis, 2010a; Eady & Cooper, 1986; Hay, Sato, Coren, Moran, & Diehl, 2006; Kochanski, Grabe, Coleman, & Rosner, 2005; Krahmer & Swerts, 2001; Nooteboom, 1997). There is, however, nothing in the way prosody is defined that requires it pertains only to the auditory signal. Indeed, recent research has suggested that a speakers' head and face movements (visual speech) can provide cues that affect the interpretation of an utterance in much the same way as do cues from auditory speech. The current study determined the extent of people's sensitivity to different types of visual prosodic information by examining whether they could match cues to prosody even when these derived from different people

☆ Earlier versions of this work have been presented at the 5th International Conference on Speech Prosody, Chicago, USA (2010), and the International Conference on Auditory-Visual Speech Processing, Hakone, Japan (2010), and appear in the conference proceedings.

* Corresponding author. Address: MARCS Auditory Laboratories, University Western Sydney, Locked Bag 1797, Penrith, New South Wales 2751, Australia. Tel.: +61 2 9772 6141; fax: +61 2 9772 6326.
E-mail address: e.cvejic@uws.edu.au (E. Cvejic).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.11.013

and different face regions. In what follows, we provide a background to this enterprise by briefly reviewing what is known about visual cues to prosody.

What sort of visual speech cues are likely to provide prosodic information? Some visual cues to prosody are intimately connected with speech articulation. That is, typically, articulatory movements (i.e., lips and mouth opening) or closely related movements (e.g., chin and cheek motion) are strongly correlated with aspects of the produced acoustics such as intensity variation over time that are used to signal prosody (Yehia, Rubin, & Vatikiotis-Bateson, 1998). Indeed, in order to produce a speech sound over an extended duration (a common acoustic property of narrowly focused syllables), the speaker must maintain the configuration of the articulators for this time (de Jong, 1995). Similarly, increases in amplitude are likely to be accompanied by more dynamic jaw movements that end in a lower jaw position (Edwards, Beckman, & Fletcher, 1991; Summers, 1987). Other visual cues to prosody, although also related to properties of auditory speech, are less tied to the process of speech articulation. For example, changes in a speaker's rigid head motion correlates well with changes in F_0 (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Yehia et al., 1998) and relative intensity (Hadar, Steiner, Grant, & Rose, 1983), and a considerable proportion of a speaker's eyebrow raises are accompanied by a rise in F_0 (Cavé et al., 1996; Granström & House, 2005).

Production studies that have examined such visual cues to prosody indicate that the manner in which these are realized is quite variable. For example, Dohen, Lœvenbruck, and Hill (2009) examined multiple French speakers' utterances that had narrow focus (on the subject, verb or object of the base sentences) compared to broad focus and found that in general, narrow focused syllables attracted hyperarticulation (larger mouth/jaw opening gestures). However, it appeared that speakers used different strategies to signal focus, with one speaker consistently showing larger mouth area and significantly longer gesture duration for focused vs. non-focused syllables, whereas another showed considerable variation in whether a focused syllable was marked by duration and/or enhanced mouth opening. Similarly, while all speakers hyperarticulated the prosodically marked constituent, some speakers did so to a lesser degree yet complimented this by hypoarticulating post-focal utterance content.

Dohen, Lœvenbruck, and Hill (2006) have also studied visual cues to prosody that occur outside the oral region and found that gestures that were not closely linked to speech production were more variably realized. For instance, although all speakers moved their head to some degree, only one speaker showed a significant correlation between rigid head tilts and the production of focus, a link that appeared to be non-systematic and highly variable. Moreover, only three of the five speakers raised their eyebrows on focused syllables (and these movements did not always accompany the production of a focused constituent). Consistent with this, the results of a study by Cavé et al. (1996) that examined whether changes in F_0 were accompanied by eyebrow movements showed considerable variability both within and between speakers in whether eyebrow raises were accompanied by a rise in F_0 (see also Guaïtella, Santi, Lagrue, & Cavé, 2009). The above two studies examined French speakers, however similar patterns of movement and inter-speaker variability have been shown for speakers of American English in producing words with lexical stress and phrasal focus (Scarborough, Keating, Mattys, Cho, & Alwan, 2009) and for Australian English speakers in productions that differentiate declarative statements from echoic questions and narrow focused utterances (Cvejic, Kim, Davis, & Gibert, 2010).

The results from various production studies indicate two sources of variation in the realization of visual cues to prosody. First, there is both within- and between-speaker variation in whether or when particular visual cues for prosody are used. Second, there are differences between visual prosody cues that are displayed in the lower and upper regions of the head and face. These differences manifest both in the strength of prosodic information signaled by a region and in the regularity that such signals are emitted. The underlying reason for such differences is that prosody related signals from the lower face occur as a direct consequence of articulation (e.g., lip and mouth movements) whereas this is less the case for those that occur in regions beyond the oral region (e.g., in the upper face). Upper face speech related motion appears to be less strongly related to the action and timing of the produced acoustic signal. For example, Yehia, Kuratate, and Vatikiotis-Bateson (2002) reported that a large amount of variance in F_0 (88% for an American English speaker; 73% for a speaker of Japanese) could be estimated from rigid head motion if head movement and F_0 were measured within each of three sentences, however if the sentences were combined, no relationship was found. Yehia et al. interpreted this as indicating that the manner in which F_0 and head motion are coupled changed from utterance to utterance. Moreover, it was found that there was considerable variation in the manner in which head motion was coupled with F_0 (i.e., the same head motion was not always coupled with changes in F_0). While not being tied directly to the speech segment production, these non-articulatory gestures may still serve a functional role. For example, a recent study has reported that eyebrow movements precede pitch rises in the acoustic signal by approximately 60 ms (Flecha-García, 2010). Such signals may be utilized by speakers as a signpost to indicate that important information is about to be presented in the auditory modality (Swerts & Krahmer, 2010).

Despite the variability of their realization, it is clear that visual speech cues can serve prosodic functions, affecting how auditory prosody is perceived (e.g., Dohen & Lœvenbruck, 2009; Foxton, Riviere, & Barone, 2010). For example, Dohen and Lœvenbruck (2009) determined the visual speech contribution to the perception of prosodic contrastive focus by using whispered utterances (whispered speech was used to make the perception of prosodic focus difficult and thus allow any visual speech effect to be apparent). Three whispered narrow focus renditions (with focus on the subject, verb or object) and a broad focused rendition were elicited from two speakers by use of a correction task (in which a speaker produced contrastive focus by correcting a constituent). These renditions were presented to perceivers in auditory-only (AO), visual-only (VO) and

auditory–visual (AV) conditions. Participants were told about the correction task and asked to identify the corrected constituent. The results showed that presenting both auditory and visual cues to prosody (AV condition) led to the focus condition being identified correctly more often (and more quickly) than the AO or VO conditions.

In addition to visual cues facilitating the perception of auditory prosody, other research indicates that such cues can directly affect the interpretation of a spoken phrase. For example, Swerts and Krahmer (2008, Experiment 2) have shown that visual markers of prosody played a role in prosody perception even in the absence of auditory prosody information. Swerts and Krahmer paired visual cues to prosody (such as a rising eyebrow movement) with monotonic acoustic renditions of a spoken Dutch sentence. On each trial only one of three content words was given visual prominence and participants were asked to indicate which word was the most prominent. The results showed that visual cues to prominence systematically influenced which word was perceived as being prominent (with prominence detection performance about 95% correct). In addition, Swerts and Krahmer manipulated whether participants saw the upper or lower face in order to ascertain the face regions from which perceivers extract the strongest cues to prominence (i.e., visual cues to prosody that occur outside vs. within the oral region). The results showed that when only the upper face was visible, the degree to which a word given visual prominence was identified as prominent was similar to that in the full face condition; when only the lower half of the face was visible, identification performance was substantially poorer. Swerts and Krahmer's results indicate that perceivers extract visual prosodic cues across face regions and that the upper face provided greater perceptual cues for prosodic focus than the lower face.

Lansing and McConkie (1999) determined the facial distribution of visual prosody cues in a slightly different way. The silent visual speech of a single speaker uttering disyllabic sentences (i.e., "Ron ran", "We won?") was presented to participants whose task was to identify either the word content (segmental task), the constituent within the utterance that received narrow focus, or the phrasal nature of the utterance (i.e., whether the sentence was phrased as a statement or question). An eye tracker was used to measure gaze behaviour during the task. For the segmental task, perceivers directed their gaze towards the lower face region. However, for the narrow focused constituent and the phrase nature tasks, the participant's gaze was distributed across the entire face, suggesting that perceivers seek this type of prosodic information from a more diverse range of face movements beyond those provided by the lower face.

Lansing and McConkie (1999) also presented participants with face displays in which the whole face moved, or in which the lower half of the face showed motion but the upper half remained static. Once again, participants were asked to identify the segments, the primary focus or intonation type. Average scores for segment identification and primary stress recognition were higher than 95% in the full dynamic face condition. In the lower face dynamic condition, task performance was maintained for identifying both segmental content and focused constituents but performance for identifying a phrase as a statement or question decreased significantly. These results indicate that the facial distribution of visual cues changes across the type of prosodic information conveyed and that the upper face provided stronger visual cues for phrase type than the lower face. With regards to visual cues for narrow focus, the lower face appears to be as good a source of cues as the upper face. Although this result is inconsistent with that of Swerts and Krahmer (2008) which showed that the upper face provided stronger cues for prosodic focus than the lower face (a difference that may be due to speaker variability in whether or when particular visual cues for prosody are used, as discussed above), what it does confirm is that perceivers extract visual cues for some prosodic types from a variety of face regions.

The above review reveals a tension between the results of production studies that indicate considerable variability in the realization of the visual cues to prosody (particularly non-articulator based ones) and those of perception studies that show people readily perceive and use such visual cues. This apparent mismatch between signal variability and constant perception highlights a basic problem in human pattern recognition (i.e., how variable form is mapped onto perception). In speech perception, this has often been characterized in terms of the problem of a lack of invariant cues to support categorical distinctions. Although there have been various proposals to account for this ability, there is a common view that this is a fundamental issue in speech perception (see McMurray & Jongman, 2011). In research on visual prosody this issue has not been considered, but the question of how perceivers cope with variability in the production of prosodic cues is equally important. To begin to answer this question, what is needed is a better understanding of how the visual cues themselves are perceived and how they relate to auditory prosodic cues.

In this regard, a recent study has examined perceiver's sensitivity to the visual prosodic cues themselves by using a prosody matching task. Cvejic, Kim, and Davis (2010b) recorded two speakers uttering a broad focused statement, a narrow focused one and an echoic question. The video recordings were cropped (similar to Swerts & Krahmer, 2008) and only the upper head and face was presented to participants in a within-modal (video–video) and cross-modal (audio–video) prosody matching task. In this task, participants were instructed to determine which of two pairs of stimuli had matching prosody. The results showed that perceivers were able to match prosody both within and across modalities at rates exceeding 80% correct. The ability of participants to correctly match the type of prosody produced across different visual tokens and different modalities showed that more than a simple feature-to-feature matching strategy was involved. Indeed, Cvejic et al. proposed that participants were able to achieve high levels of correct across token matching because they could classify the type of prosody from the visual cues (e.g., narrow or broad focus) and then use the result of this classification to decide which stimulus pair matched.

The idea that good performance in the matching task is based upon the categorization of visual prosody cues sug-

gests that this task may be useful in probing the extent to which a prosodic category can be determined from different inputs. Indeed, the matching paradigm of Cvejic et al. (2010b) provides a well controlled experimental situation to assess the extent to which perceivers can cope with variation in visual prosodic cues. To use this technique to more fully investigate the degree to which variable input can be overcome, experiments need to test people's ability to match visual prosody when cues manifest in very different ways. For instance, the within-modal matching task can be used to investigate whether people can determine prosodic counterparts across different face regions by testing whether people can reliably match cues to prosody that are signaled by the upper and lower face (as mentioned above, this division is important as it picks out those cues that stem more directly from articulation from those that do not). Furthermore, the cross-modal matching task can also reveal whether perceivers can ascertain the underlying prosody type regardless of who is speaking by testing matching ability across different speakers.

In sum, production studies have shown that the way visual prosody is realized is variable. In this experimental series, we aim to determine whether people can match the visual cues for different types of prosody across different face regions and across different people; such ability would provide a basis for using visual cues to prosody even though they are variable. To this end, Experiment 1 examined whether perceivers were able to match within-modal (visual to visual) and cross-modal (auditory to visual) cues from the lower face region (this was not assessed in Cvejic et al., 2010b and provides an opportunity to assess whether cues to visual prosody that are linked to articulation show a different pattern across prosodic types). Experiment 2 examined people's ability to match cues across different speakers for the same face region. Experiment 3 tested if perceivers could successfully match across face areas within a speaker and Experiment 4 examined matching both across face areas and across different speakers.

2. Experiment 1

The aim of Experiment 1 was to ascertain the extent to which perceivers were able to use visual cues from the lower face region for matching within (i.e., visual to visual) and across modalities (i.e., auditory to visual). These results, taken together with the results of Cvejic et al. (2010b), will provide a baseline from which to evaluate cross-speaker matching (Experiment 2).

Note that the purpose of the within-modal (VV) matching task was to test whether the differences between the visual cues used to contrast prosodic types (i.e., broad vs. narrow focus; statements vs. questions) from selective face areas were perceptually salient. Indeed, successful performance on this task could be simply explained as perceivers being sensitive to the overall presence vs. absence of prosody-based visual signals (i.e., choosing the pair where there was some sort of distinctive motion in both stimuli), and does not necessarily indicate that perceivers base their performance on having recognized the prosodic type per se. In contrast, the cross-modal (AV) matching task requires perceivers to interpret the prosodic information from one modality, and find suitable correlates in another modality. Thus, successful performance in the AV task would seem to indicate that not only are there perceptually salient differences between the visual cues used to contrast prosodic types, but also that such differences can be linked to those types. Furthermore, no auditory–auditory matching task was included, as linguistic prosodic contrasts are typically well perceived from the auditory modality alone (see Dohen & Lœvenbruck, 2009) and as such the results may have been near ceiling. Indeed, in an independent study we recently conducted (that used one of the current speakers) we found that participants were able to achieve excellent performance on the task of classifying auditory prosody as an echoic question or narrow or broad focus (Cvejic, Kim, & Davis, submitted for publication).

2.1. Method

2.1.1. Participants

Twenty undergraduate students (M_{Age} = 21.3 years) from the University of Western Sydney (UWS) participated in the study for course credit. All participants were fluent speakers of English, and had self-reported normal or corrected-to-normal hearing and vision. No participants took part in the study reported in Cvejic et al. (2010b).

2.1.2. Materials

The materials used in this and the following experiments are the same as in Cvejic et al. (Experiments 1 and 3, 2010b). These stimuli consisted of ten non-expressive sentences drawn from the IEEE (1969) Harvard Sentence List (e.g., "The pipe ran almost the length of the ditch"). Auditory and visual speech tokens from two native male speakers of Australian English (M_{Age} = 23 years) were recorded in a well lit, sound-attenuated room using a Sony TRV19E digital video camera (25 fps), with audio recorded simultaneously at 44.1 kHz, 16-bit mono with an externally connected Sony lapel microphone.

Each sentence was recorded as a *broad focused* statement, a *narrow focused* statement and as an *echoic question*. A dialogue exchange task was used to elicit these conditions in which the speaker interacted with an interlocutor by repeating what the interlocutor said (broad focused statement), making a correction to an error made by the interlocutor (narrow focused statement), or questioning an emphasized item that was produced by the interlocutor (echoic question). The critical item within each utterance (i.e., the word that received narrow focus or question intonation) was a content word, began with a consonant, and was not located in phrase-initial or phrase-final position, with the position within the utterance varying across the ten sentences. The same critical item was maintained across speech conditions and speakers. Two repetitions of each utterance were recorded several minutes apart. This recording procedure resulted in 120 auditory and 120 visual speech tokens for use as stimuli. The visual tokens were then processed using custom designed scripts in VirtualDub (Lee, 2008) to generate two versions of the visual stimuli; upper half videos that showed the speaker from above the tip of the nose, and lower face videos that displayed only the lips, cheeks, chin and jaw of the speaker.

2.1.3. Procedure

The experiments were run in DMDX (Forster & Forster, 2003) on a desktop computer connected to a 17″ LCD Monitor. Participants were tested individually in a double-walled, sound attenuated booth. Each participant completed two experimental tasks: a visual–visual (VV) matching task and an auditory–visual (AV) matching task (as used in Cvejic et al., 2010b; Davis & Kim, 2006) in a counter-balanced order. Stimuli were presented in a two-interval, alternate forced choice (2AFC) discrimination task, in which each interval included a pair of stimuli to be compared and the participant's task was to select the pair of the same prosodic type. Participants were informed of the three prosodic conditions used (in straightforward language that made the distinctions clear), and that the sentences they would be judging differed only in prosody, not segmental content. The matching items within pairs were always taken from a different recorded token. All items within-pairs were produced by the same speaker.

Participants indicated their response as to which pair was produced with the same prosody via a selective button press. No feedback was given to participants about the correctness of their responses. Video display, item randomization and collection of response data was controlled by DMDX (Forster & Forster, 2003).

Fig. 1 provides an overview of the sequence of displays used in the tasks. In the VV matching task, a total of 40 matching responses were involved across two prosodic speech conditions (i.e., narrow focus and echoic questions), with broad focused renditions always acting as the non-matching item within pairs. The first video within each pair was identical and was the standard with which the other stimulus items within pairs were to be judged. The order of correct response pair was counter-balanced, so the correct option appeared equally in the first and second interval. The AV matching task was similar to the VV task, except that it used auditory–visual stimuli pairs and non-matching items within-pairs were the same sentence produced as one of the alternate prosodic types (e.g., the non-matching items for half of the narrow focus trials were broad focused renditions and echoic question renditions for the other half). The initial auditory token that appeared at the start of each pair was the same, with each of the 120 recorded auditory tokens appearing as the target once. Stimuli were randomly presented in four blocks of 30 items, with each block containing one of each sentence in all three prosodic speech conditions from an individual speaker. Auditory stimuli were presented binaurally via Senheiser HD650 stereo headphones.

2.2. Results and discussion

The mean percent of correct responses for the VV and AV tasks of both studies are shown in Table 1. The results are presented together with data previously obtained in Cvejic et al. (2010b) to allow for full comparisons between upper and lower face stimuli. As can be seen, performance was considerably greater than chance across all prosodic speech conditions in both tasks, as confirmed by a series of significant one-sample *t*-tests.

Further analyses were conducted to compare the results from the upper and the lower face conditions. For VV match-

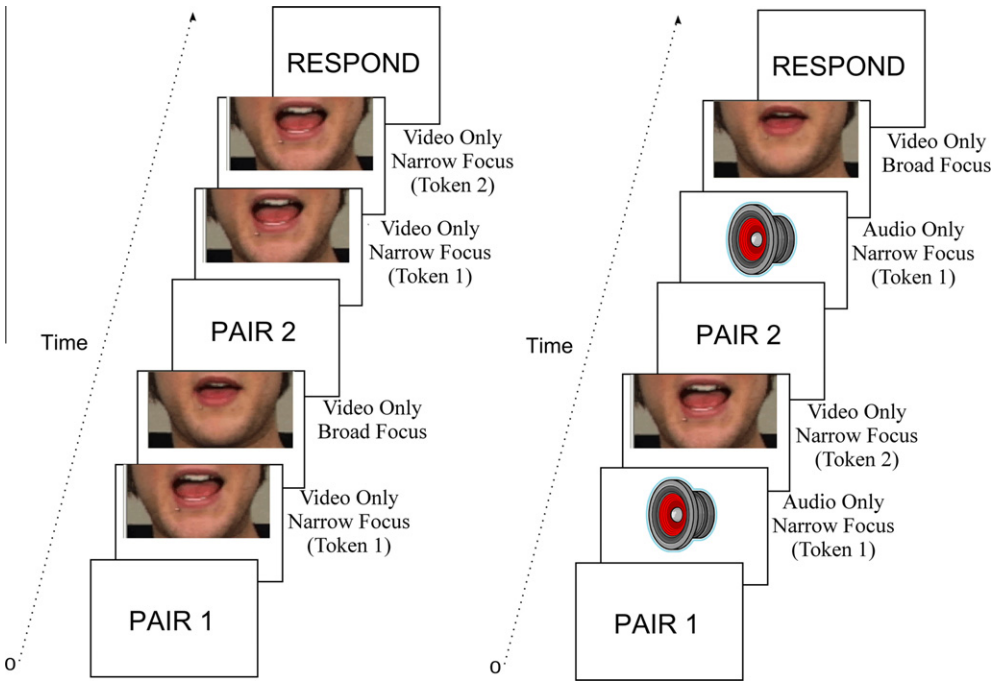

Fig. 1. Schematic representation of the 2AFC visual–visual (left) and auditory–visual (right) matching tasks used in Experiment 1. The same item appeared first for both pairs, and was the standard that the matching judgment was to be made on. The matching item within pairs was always taken from a different recorded token, and non-matching items were the same sentence produced as a different prosodic type.

Table 1
Mean percent of correct responses in the within-speaker VV and AV matching tasks as a function of visible face area in each prosodic speech condition. Data in italics obtained from Cvejic et al. (2010b).

Visible face area	Prosodic speech condition	Mean correct (%)	Standard error of mean	*t*-test vs. Chance (50%)
Visual–visual (VV) matching task				
Upper half (df = 10)	*Narrow focus*	*82.7*	*2.97*	*11.03***
	Echoic question	*87.7*	*3.26*	*11.58***
Lower half (*df* = 19)	Narrow focus	91.7	2.87	20.92***
	Echoic question	80.5	2.84	10.28***
Auditory–visual (AV) matching task				
Upper half (df = 10)	*Broad focus*	*88.9*	*2.14*	*18.15***
	Narrow focus	*94.8*	*1.61*	*27.79***
	Echoic question	*88.9*	*2.39*	*16.25***
Lower half (*df* = 16)[a]	Broad focus	87.4	3.12	14.70***
	Narrow focus	91.4	2.69	17.71***
	Echoic question	84.4	2.85	12.34***

Degrees of freedom (*df*) are indicated in brackets.
*** $p < .001$.
[a] For the AV task, three participants were not included in the analysis as they recorded matching accuracy of 0% due to a technical error.

ing performance, a 2 × 2 mixed repeated measures ANOVA was conducted to determine if task performance (percent correct responses) varied as a function of the visible face area, with prosodic speech condition (narrow focus; echoic question) as the within-subjects factor, and face area (upper vs. lower half) as a between-subjects factor. No significant main effect was found for prosody condition, $F(1,29) = 1.58$, $p > .05$, or visible face area, $F(1,29) = 0.08$, $p > .05$. However, the interaction was significant, $F(1,29) = 10.67$, $p < .05$, $\eta_p^2 = .269$, reflecting the pattern that displays of the lower face produced better discrimination of focus whereas phrasing was better discriminated from the upper face displays (consistent with Lansing & McConkie, 1999). A series of post-hoc *t*-tests (with a Bonferroni adjusted α of .025 for multiple comparisons) revealed that narrow focused items were discriminated with significantly greater accuracy from the lower face compared to upper face presentation condition, $t(29) = 2.60$, $p < .025$, but the difference in discriminating echoic questions was not statistically significant, $t(29) = 1.54$, $p > .025$.

For AV matching performance, a 2 (upper vs. lower face) × 3 (broad focus; narrow focus; echoic question) mixed repeated measures ANOVA was conducted, with visible face area as the between-subjects factor, and prosodic condition as a within-subjects factor. The main effect of prosodic speech condition was significant, $F(2,52) = 9.62$, $p < .05$, $\eta_p^2 = .270$, this difference appears to be driven by participants being better able to discriminate narrow focus renditions across both upper and lower face presentations. The main effect of visible face area, $F(1,26) = 0.97$, $p > .05$, and the interaction, $F(2,52) = 0.46$, $p > .05$, did not reach statistical significance. Unlike VV matching results, no significant interaction between face area and prosodic conditions was found. This result could be due to differences in how well prosody was specified by the initial item of a pair. For visual presentation, it seems the lower face provides a better cue to narrow focus, so matching performance is very good for narrow focus items when this information has been clearly presented by the initial item of the lower face VV trials (compared to a less clear specification in the upper face VV displays). A similar argument applies for the echoic question items (only here, it is the upper face that provides the clearest information to the relevant prosodic condition). This interaction between face area and prosodic condition was not found in the AV trials because the auditory specification of prosodic type is the same regardless of whether it is followed by a lower or upper face item.

In sum, reliable matching of visual speech to other visual or auditory speech tokens based on prosodic differences alone was observed regardless of whether the upper or lower face area was presented. This result is consistent with the proposal that perceivers are able to resolve the type of prosody from any of multiple visual cues, and when a particular cue is not available, the underlying prosody can still be determined from those cues that remain. The current within face region and within speaker results provide a baseline measure of performance from which to further examine this proposal by investigating people's ability to match prosodic counterparts across face region and across different speakers.

3. Experiment 2

Experiment 2 examined perceivers' ability to match visual speech tokens within and across modalities when the signals were produced by different speakers (with the same face area shown across speakers). If perceivers can categorize the type of prosody (e.g., narrow or broad focus) from the visual cues regardless of who produced the token, then it is expected that they should be able to successfully perform the VV and AV matching tasks. That is, if task performance is based on matching information at the level of abstract form (category type), then accurate prosody matching can be achieved despite any individual variation in how the prosodic cues were realized.

3.1. Method

3.1.1. Participants

Thirty-two undergraduate students (M_{Age} = 22 years) from UWS participated in the experiment in return for course credit. All were fluent speakers of English. None of

these participants had previously taken part in any other experiment reported in the current study, and all reported normal or corrected to normal vision with no history of hearing loss.

3.1.2. Materials and procedure

The same procedure was used as in Experiment 1 and Cvejic et al. (2010b). In this experiment, the paired stimuli for matching consisted of two tokens from different speakers. Fig. 2 outlines the composition of the 2AFC tasks used. Participants were randomly allocated to a visible face area condition (upper or lower face, 16 in each condition). Participants took part in both VV and AV matching tasks in one of the presentation conditions (i.e., either upper or lower face presentations), and completed the tasks in a counter-balanced order. All other material and procedural details are the same as Experiment 1.

3.2. Results and discussion

The mean percent of correct responses for both VV and AV tasks are shown in Table 2. A series of significant one-sample *t*-tests of the results showed that despite the within-pair signals originating from different speakers, participants were able to perform the task at levels well above chance for all conditions.

The results for cross-speaker prosody matching (Experiment 2) were compared to the results obtained for within-speaker matching (Experiment 1 and Cvejic et al., 2010b). A $2 \times 2 \times 2$ ANOVA was conducted for VV task performance, and a $2 \times 2 \times 3$ ANOVA for AV performance, each with speaker congruency (within- vs. cross-speaker) and visible face area (upper vs. lower face) as between-subjects factors, and prosodic speech condition as the within-subjects factor. The main effect of speaker congruency was significant for the VV task, $F_{VV}(1,59) = 11.00$, $p < .05$, $\eta_p^2 = .157$, but not for the AV task, $F_{AV}(1,56) = 3.52$, $p > .05$. Overall, performance across both tasks was greater when the matching speech tokens were produced by the same speaker, suggesting that although non-speaker specific suprasegmental content can be extracted from visible movements, there is also a speaker-specific component.

The main effect of visible face area was not significant for either task, $F_{VV}(1,59) = 0.55$, $p > .05$, $F_{AV}(1,56) = 0.08$, $p > .05$, suggesting that both the upper and lower face provide equally effective prosodic cues. For both tasks, the main effect of prosody was significant, $F_{VV}(1,59) = 5.50$, $p < .05$, $\eta_p^2 = .085$; $F_{AV}(2,112) = 11.67$, $p < .05$, $\eta_p^2 = .172$, which appears to be driven by narrow focus being easier to visually discriminate than broad focused statements and echoic questions. The prosody by visible face area interaction for the VV task was significant, $F_{VV}(1,59) = 40.15$, $p < .05$, $\eta_p^2 = .405$, however this interaction for the AV task was not significant, $F_{AV}(2,112) = 0.44$, $p > .05$. Likewise, the prosody by speaker congruence interactions, $F_{VV}(1,59) = 0.12$, $p > .05$; $F_{AV}(2,112) = 1.06$, $p > .05$, and the three-way interactions, $F_{VV}(1,59) = 1.32$, $p > .05$; $F_{AV}(2,112) = 0.81$, $p > .05$, failed to reach significance.

The results showed that despite visual cues coming from two different speakers, perceivers were able to visu-

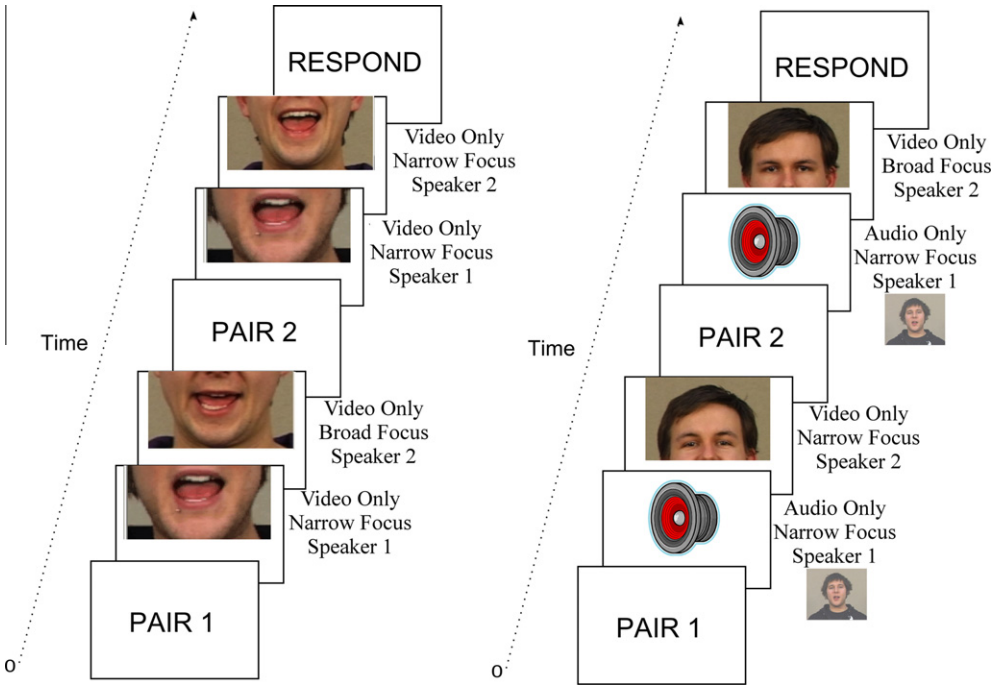

Fig. 2. Schematic representation of the 2AFC visual–visual (left) and auditory–visual (right) matching tasks used in Experiment 2. The same item appeared first for both pairs produced by one speaker, and was the standard from which the matching judgment was to be made on. The second item within pairs was produced by a different speaker, with the non-matching item being the same sentence produced with a different prosody. Participants completed the task with either upper or lower face stimuli.

Table 2
Mean percent of correct responses in the cross-speaker VV and AV matching tasks as a function of visible face area in each prosodic speech condition, when the items within pairs were produced by different speakers.

Visible face area	Prosodic speech condition	Mean correct (%)	Standard error of mean	*t*-test vs. Chance (50%)
Cross-speaker VV matching task				
Upper half	Narrow focus	70.9	4.33	4.83**
	Echoic question	78.4	3.38	8.42**
Lower half	Narrow focus	85.9	2.89	12.42**
	Echoic question	70.0	3.03	6.61**
Cross-speaker AV matching task				
Upper half	Broad focus	79.8	4.85	6.15**
	Narrow focus	85.9	4.49	8.01**
	Echoic question	81.4	3.93	8.00**
Lower half	Broad focus	81.6	2.39	13.19**
	Narrow focus	94.2	1.22	36.16**
	Echoic question	84.8	2.30	15.17**

df = 15.
** $p < .001$.

ally match prosodic content when information was restricted to either the upper or lower face. This finding is consistent with previous evidence that visual speech cues can be processed at an abstract level. For example, observing a silently spoken word facilitates lexical decisions on the same word subsequently presented in either written or spoken form (Kim, Davis, & Krins, 2004). This priming effect occurs even when auditory and visual speech tokens are produced by different speakers (Buchwald, Winters, & Pisoni, 2008). Similarly, McGurk effects (i.e., the integration of incongruent auditory and visual information resulting in a "fused" percept, McGurk & MacDonald, 1976) are observed when auditory and visual signals originate from different speakers (Green, Kuhl, Meltzoff, & Stevens, 1991), even a male face paired with a female voice. These studies demonstrate that equivalent phonemic information is extracted from the visual speech signal regardless of the speaker that produced the signal. Indeed, in the current task, perceivers performed just as well when matching the auditory prosody of one speaker to the visual prosody of another. These results suggest that information for visual prosody, much like phonemic information, is processed in terms of abstract visual speech events, allowing for generalization across tokens, modalities and speakers. Furthermore, participants can match visual cues to prosody regardless of whether they were from the upper or lower face, suggesting that multiple (and potentially redundant) visual cues to prosody are distributed across face areas, and that perceivers must be sensitive to prosodic counterparts across different face regions. This suggestion concerning the flexibility of perceivers' use of visual prosody is further tested in Experiments 3 and 4.

4. Experiments 3 and 4

In these experiments, perceivers were shown visual-only tokens of the upper face and the task was to match these to the prosody displayed from the lower face (and vice versa). In Experiment 3, the upper and lower faces stimuli were of the same speaker (but different tokens), whereas in Experiment 4, the tokens were from different speakers.

4.1. Method

4.1.1. Participants

Forty undergraduate students (M_{Age} = 21.5 years) from UWS participated in the experiment for course credit. All were fluent speakers of English, and had self-reported normal or corrected-to-normal vision and no history of hearing loss. None had taken part in the earlier experiments of the study. Participants took part in both Experiment 3 and 4 in a counter-balanced order (i.e., 20 completed Experiment 3 first; the remaining 20 completed Experiment 4 first).

4.1.2. Materials and procedure

Experiments 3 and 4 used the same materials outlined in the VV task of Experiment 2. Fig. 3 provides an overview of the sequence of displays used in the task of each experiment. In Experiment 3, items within-pairs were produced by the same speaker, but displayed one half of the face in the first item, then the opposite half of the face in the second video. Experiment 4 was basically the same as Experiment 3 except items within-pairs were produced by different speakers.

For each experiment, two different versions of the experimental task were created, so that the upper and lower face stimulus of each item appeared as the target only once across both versions. Participants completed only one version of the task for each experiment (n = 20 in each version), and were never exposed to the same face area producing the same sentence by either speaker more than once. In total, each version required 40 matching responses across two prosodic speech conditions (i.e., narrow focus and echoic questions), with the broad focused rendition always acting as the non-matching item within pairs. Half of the trials displayed the upper face followed by the lower face within pairs, while the remaining half displayed the lower face before the upper face. All other material and procedural details are the same as the VV task of Experiment 2.

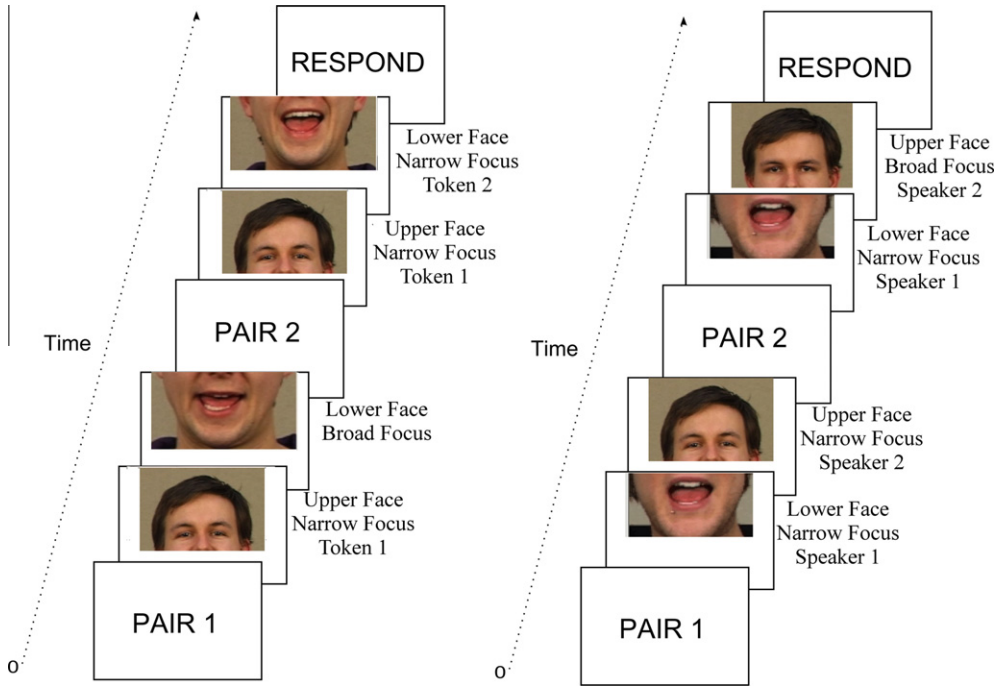

Fig. 3. Schematic representation of the 2AFC tasks used in Experiment 3 (left) and Experiment 4 (right). In each item either the upper face or lower face stimuli were displayed first, followed by the opposite face area within each pair. In Experiment 3, items within-pairs were produced by the same speaker, with the matching video always taken from a different recorded token. In Experiment 4, items within-pairs were produced by different speakers. The non-matching video in both tasks were always the broad focused rendition of the same sentence. Order of presentation was randomized by the presentation software.

4.2. Results and discussion

Table 3 shows the mean percent of correct responses for the 2AFC visual–visual matching task across face areas of the same speaker (Experiment 3) or different speakers (Experiment 4). A series of one-sample *t*-tests indicated that performance was significantly greater than that expected by chance for both experiments.

The data of Experiment 3 and 4 were compared using a $2 \times 2 \times 2$ repeated measures ANOVA. Speaker congruency within-pairs (congruent; incongruent), presentation order (upper first; lower first) and prosodic contrast (narrow focus; echoic question) were all treated as within-subjects factors. In general, performance was better when items within-pairs were produced by the same speaker, however no main effect was observed for speaker congruency, $F(1,39) = 1.64$, $p > .05$. The main effect of prosody was significant, $F(1,39) = 44.85$, $p < .001$, $\eta_p^2 = .535$, an effect driven by the participants superior performance in discriminating prosodic focus in comparison to prosodic phrasing. The main effect of presentation order was also significant, $F(1,39) = 14.76$, $p < .001$, $\eta_p^2 = .275$. As expected, the prosody by presentation order interaction was significant, $F(1,39) = 91.05$, $p < .001$, $\eta_p^2 = .700$, as was the speaker congruence by prosody interaction, $F(1,39) = 5.67$, $p = .022$, $\eta_p^2 = .127$. The speaker congruence by presentation order interaction, $F(1,39) = 2.84$, $p > .05$ and three-way interaction, $F(1,39) = 0.26$, $p > .05$, did not reach significance.

In the above analyses, it is clear that presentation of the lower face before the upper face resulted in better matching accuracy for focus, whereas the opposite presentation order (i.e., the upper face followed by the lower face) yielded better results for judgements of phrasing, regardless of speaker congruency. In explaining this significant interaction between presentation order and prosodic contrast, it is useful to consider the relative effectiveness of prosodic cues from the lower and upper face regions. The results of Experiments 1 and 2 (along with the comparison to the upper face matching data of Cvejic et al., 2010b) indicate that compared to the upper face, the lower face provides more effective cues for determining whether a constituent has been focused or not, whereas the upper face appears to provide more effective cues concerning phrasing. Given this, it may be that when the face area containing more robust, salient visual cues was initially presented, matching performance was facilitated because the prosodic type (category) resolved from the salient cues guides the perceiver as to which type of cues they should seek when viewing the second item within the pair, increasing the perceiver's sensitivity to subtle, non-salient cues. In contrast, when the initially presented face area included less salient cues, the perceiver (without any guide) might be relatively less sensitive to subsequently presented cues in the second interval. Effects of presentation order have been observed in a variety of psychophysical studies with it being a common idea that participants employ categorical coding to compare stimuli (Repp & Crowder, 1990).

5. General discussion and conclusions

The current study examined perceivers' sensitivity to visual cues to prosody from upper and lower face regions.

Table 3
Mean percent of correct responses in the 2AFC visual–visual matching task across face areas using within- and cross-speaker stimuli, presented as a function of presentation order (upper to lower; lower to upper face), separated by prosodic speech condition.

Presentation order	Prosodic speech condition	Mean correct (%)	Standard error of mean	*t*-test vs. Chance (50%)
Within-speaker stimuli (Experiment 3)				
Upper to lower	Narrow focus	78.0	2.56	10.93**
	Echoic question	86.3	2.42	14.98**
Lower to upper	Narrow focus	88.0	2.21	17.17**
	Echoic question	70.8	2.22	9.35**
Cross-speaker stimuli (Experiment 4)				
Upper to lower	Narrow focus	79.3	3.07	9.54**
	Echoic question	82.3	2.44	13.21**
lower to upper	Narrow focus	86.5	2.25	16.21**
	Echoic question	61.0	3.05	3.60**

df = 39.
** $p < .001$.

The study was motivated by the apparent mismatch between behavioural studies indicating observers use visual prosody very effectively and the observed variability in the production of such cues across speakers. To account for the effectiveness of visual cues, we proposed that perceivers are able to resolve the type of prosody from any of multiple visual cues, so that if a particular cue is not available, other cues will still permit the underlying prosody to be determined. To test whether perceivers were able to determine prosody from different visual cues, the experiments tested whether people could match upper and lower face cues as well as auditory to visual cues (not only from the same speaker but also from different speakers). The results showed that despite differences in the form and temporal structure of prosodic cues across modalities and face areas, perceivers could reliably match both auditory and visual items to visual tokens of the same prosodic type across speakers, regardless of which face area is presented. These results confirm that perceivers can use visual prosody very effectively (Cvejic et al., 2010b; Lansing & McConkie, 1999). More importantly, the results show that perceivers can match prosody from visual cues provided by the upper and lower face and across different speakers. This ability to match very different cues suggests that matching was performed at an abstract level (i.e., the perceiver used cues to determine the prosodic category and matched at this level).

These results and our interpretation that the ability to perform the prosody matching task (given very different visual cues) was based on matching at an abstract level, raise a set of issues about the different sources of visual cues to prosody and what is precisely meant by matching at an abstract level. Description of visual prosody and theories concerning how variable visual cues might map to prosodic categories are still in their infancy. Given this, we have developed the discussion of these issues based upon two recent theoretical accounts proposed within the auditory domain. The first issue we consider concerns the nature of different types of prosodic cue and the second is about how variable signals might be mapped to categories (and how these categories are specified).

The finding that perceivers can reliably match prosody using visual cues based on very different signals indicates that visual prosody may be derived from more than one source. Recently, Watson (2010) has developed a multisource account of acoustic prominence that appears relevant to this issue. This view proposes that prosody (at least in terms of prominence) is the product of a number of different cognitive processes that give rise to different realizations. This proposal allows a distinction to be drawn between sources. For example, Watson suggested that although changes in intensity, F_0 and duration are all linked in some way to important or focused information, duration may sometimes be a less reliable cue as it is related to speaker-centric production processes. That is, it was argued that different acoustic factors will influence prominence only in as much as they mark relevant information for the listener.

It is this latter suggestion that appears relevant to the current results, as it seems that the perception of what is relevant information determines the extent to which a visual cue was influential. More specifically, consider the interaction between face region and prosody type in visual matching performance (Experiment 1). The level of AV matching showed that both the lower and upper face stimuli could be matched to narrow focus equally well (both above 90% correct). However, this was only the case when the participant knew to look for narrow focus (when this was indicated by the auditory signal). That is, when presented with visual cues from the upper face (in the VV task), performance was worse (around 80%) whereas performance remained above 90% when visual cues from the lower face were presented. This may be because the lower face provides distinct cues related to prosodic focus (e.g., cues for change in amplitude or duration, Kochanski et al., 2005) whereas the upper face provides more reliable visual cues (head and eyebrow movements) to distinguish echoic questions from statements (as these cues are linked with variations in F_0, Cavé et al., 1996; Yehia et al., 1998, an acoustic feature that can differentiate statements from echoic questions, see Eady & Cooper, 1986).

Having considered the idea that there are multiple visual cues to prosody and that the perception of what is relevant information may determine how a visual cue is weighted, we now consider the related issue of how such diverse source cues might be mapped to prosodic catego-

ries. Once again, we base our discussion on ideas derived from studies of auditory speech processing as these have a long history of development and refinement. It should also be noted that the models we consider have been proposed with regard to speech recognition (i.e., distinguishing phonemes) and not prosody recognition. Given this, the discussion will focus on the setting out of alternative models and whether they are suited for describing the current results rather than attempting to specify particular mechanisms.

One approach to the issue of source variability is to regard it as a problem that has to be overcome. Under this (invariance) approach, sources of variability are removed by processes of normalization or compensation (the latter being a more general term that typically includes processes dealing with coarticulation). The basic assumption here is that a few invariant underlying signal properties can be revealed by recoding the signal by grouping overlapping cues or gestures, or by exploiting mechanisms of contrast (that either use other signal events or long-term expectations about cues). A problem with this approach is that it is not clear that visual cues to prosody can be defined in terms of a small number of invariant properties. This is because not only is the relationship between the auditory and visual signals variable (both in whether they occur at all and how and when they occur), but visual cues also appear to vary with respect to each other (i.e., the visible movements of the upper and lower face do not occur simultaneously or systematically, Cavé et al., 1996; McClave, 1998).

The exemplar approach provides a different scheme for how variable form is mapped onto categories. Here, it is assumed that the input is encoded in detail by using all available cues and context dependency is overcome because perceivers store multiple exemplars and categorization decisions are made by comparing the incoming input to these clouds of stored exemplars. Such an approach provides a natural way of dealing with the effects of context and talker variability (without compensation per se), but it is unclear how it can deal with completely novel input, such as having to match cues derived from one face region to another.

Intermediate between the above two approaches are models that McMurray and Jongman (2011) call cue-integration approaches. The idea of these models is that if sufficient cues are encoded, then in combination, the variability of any one cue can be overcome (possibly without the need for compensation). Examples of this type of model include the FLMP (Oden & Massaro, 1978) and TRACE models (Elman & McClelland, 1986), with the former having been used to successfully model how prosody cues (duration and pitch) are integrated to influence syntactic identification (Beach, 1991). The most recent model of this type, called the "computing cues relative to expectations" (C-CuRE) model by McMurray and colleagues, combines aspects of the invariance approach (compensation) with aspects of exemplar approaches (retaining every cue, e.g., McMurray, Cole, & Munson, 2011; McMurray & Jongman, 2011). That is, like exemplar models, C-CuRE maintains a continuous representation of cue values that include variations due to the talker, context and coarticulation. However, unlike exemplar models, it uses this variation to build categories relating to such variables as context and talker and these in turn are used to interpret the informational content of the speech signal. Such a model offers a more principled way of taking into account instance-based variation (by partitioning out of sources of variance prior to cue-integration). In our view, it is this feature of C-CuRE that seems an attractive framework for the integration of auditory and visual prosody cues while taking into account their variability.

Acknowledgements

The authors thank Bronson Harry for his assistance with recording the stimuli, Catherine Gasparini and Michael Fitzpatrick for their help conducting the perceptual experiments, and three anonymous reviewers for their constructive feedback on an earlier version of the manuscript. The second and third authors acknowledge support from the Australian Research Council (DP0666857 and TS0669874).

References

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language, 30*, 644–663.

Buchwald, A. B., Winters, S. J., & Pisoni, D. B. (2008). Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes, 24*, 580–610.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and F_0 variations. In *International conference on spoken language processing* (pp. 2175–2178).

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America, 77*, 2142–2156.

Cvejic, E., Kim, J., & Davis, C. (submitted for publication). Effects of seeing the interlocutor on the production of prosodic contrasts. *Journal of the Acoustical Society of America.*

Cvejic, E., Kim, J., & Davis, C. (2010a). Modification of prosodic cues when an interlocutor cannot be seen: The effect of visual feedback on acoustic prosody production. In *Proceedings of the 20th international congress on acoustics*, ICA, 2010, pp. 1–7.

Cvejic, E., Kim, J., & Davis, C. (2010b). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication, 52*, 555–564.

Cvejic, E., Kim, J., Davis, C., & Gibert, G. (2010). Prosody for the eyes: Quantifying visual prosody using guided principal component analysis. *Proceedings of Interspeech, 2010*, 1433–1436.

Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition, 100*, B21–B31.

Dohen, M., & Lœvenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech, 52*, 177–206.

Dohen, M., Lœvenbruck, H., & Hill, H. (2006). Visual correlates of prosodic contrastive focus in French: Description and inter-speaker variability. *Proceedings of Speech Prosody, 2006*, 221–224.

Dohen, M., Lœvenbruck, H., & Hill, H. (2009). Recognizing prosody from the lips: Is it possible to extract prosodic focus from lip features? In A. W.-C. Liew & S. Wang (Eds.), *Visual speech recognition: Lip segmentation and mapping* (pp. 416–438). London: IGI Global.

Eady, S. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America, 80*, 402–415.

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America, 89*, 369–382.

Elman, J., & McClelland, J. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360–380). Hillsdale, NJ: Erlbaum.

Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication, 52*, 542–554.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods: Instruments & Computers, 35*, 116–124.

Foxton, J. M., Riviere, L.-D., & Barone, P. (2010). Cross-modal facilitation in speech prosody. *Cognition, 115*, 71–78.

Granström, B., & House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication, 46*, 473–484.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, genders, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50*, 524–536.

Guaïtella, I., Santi, S., Lagrue, B., & Cavé, C. (2009). Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation. *Language and Speech, 52*, 207–222.

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech, 26*, 117–129.

Hay, J. F., Sato, M., Coren, A. E., Moran, C. L., & Diehl, R. L. (2006). Enhanced contrasts for vowels in utterance focus: A cross-language study. *Journal of the Acoustical Society of America, 119*, 3022–3033.

IEEE (1969). IEEE recommended practices for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*, 2225–2246.

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.

Kim, J., Davis, C., & Krins, P. (2004). A modal processing of visual speech as revealed by priming. *Cognition, 93*, B39–B47.

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., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication, 34*, 391–405.

Lansing, C., & McConkie, G. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language and Hearing Research, 42*, 529–539.

Lee, A. (2008). Virtual Dub (Version 1.8.6) [Software]. <http://www.virtualdub.org>.

McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research, 27*, 69–89.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

McMurray, B., Cole, J. S., & Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. In G. N. Clements & R. Ridouane (Eds.), *Where do phonological features come from?: Cognitive, physical and developmental bases of distinctive speech categories* (pp. 197–236). The Netherlands: John Benjamins Publishing Company.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*, 219–246.

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*, 133–137.

Nooteboom, S. (1997). The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of Phonetic Science*. London: Blackwell.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85*, 172–191.

Repp, B. H., & Crowder, R. G. (1990). Stimulus order effects in vowel discrimination. *Journal of the Acoustical Society of America, 88*, 2080–2090.

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

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.

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

Swerts, M., & Krahmer, E. (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics, 38*, 197–206.

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes, 25*, 905–945.

Watson, D. G. (2010). The many roads to prominence: Understanding emphasis in conversation. In B. H. Ross (Ed.). *Psychology of learning and motivation* (Vol. 52, pp. 163–183). The Netherlands: Academic Press.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*, 555–568.

Yehia, H. C., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*, 23–43.