It's all the same to me: Prosodic discrimination across speakers and face areas

Erin Cvejic, Jeesun Kim, & Chris Davis

MARCS Auditory Laboratories, University of Western Sydney, Milperra, Australia

e.cvejic@uws.edu.au, j.kim@uws.edu.au, chris.davis@uws.edu.au

Abstract

Visual cues to speech prosody are available from a speaker's face; however the form and/or location of such cues are likely to be inconsistent across speakers. Given this, the question arises as to whether such cues are general enough to signal the same prosody information across speakers, and if so, where and what these cues are. To investigate this, this study used visual-visual and auditory-visual matching tasks requiring participants to select pairs of stimuli that were produced with the same prosody within- and across-speakers when visual information was limited to the upper or lower face. Experiment 1 tested within-speaker prosody matching when the speaker's lower face was presented. The results showed highly accurate matching performance. Taken together with the results of our previous study which presented the upper face in the same tasks [1], these data provided a baseline for which to evaluate cross-speaker prosody matching (Experiment 2). In Experiment 2, both lower and upper face stimuli were presented. In comparison to within-speaker matching, performance was lower for cross-speaker matching but still greater than chance. Overall, the results suggest that both the upper and lower face provide general non-speaker specific as well as speaker-specific visual cues to prosody.

Index Terms: visual prosody, perception, cross-speaker, within-speaker, face area, narrow focus, echoic questions.

1. Introduction

Visual cues available from the face of a speaker can signal information not only about what has been said (phonemic content), but also how it has been said (speech prosody). While visual cues to speech content are closely linked to articulatory movements in oral regions [2], visual cues to prosody have been shown to be distributed across the face [3-6]. Furthermore, such visual cues appear to be less directly coupled to speech production and therefore may be freer to exhibit token- or speaker-specificity. Consistent with this notion is the finding that the manner in which visual prosody is expressed shows considerable individual variation [7-9] in both form and location. Given this, it might be questioned whether visual cues to prosody distributed across the face are sufficiently general enough to signal abstract prosodic information (i.e., information that can generalise beyond a token or speaker).

Studies on the perception of visual prosody have demonstrated that visual cues signalling prosody can be matched across tokens and modalities. For example, we previously examined whether people can accurately discriminate sentences differing in prosody (not lexical content) in both visual-visual and auditory-visual matching tasks when presented only the upper face [1]. The results showed that people were able to match different tokens of the same prosody, indicating some within-speaker consistency in the production of visual prosody. Moreover, people were very accurate in matching the prosody of an auditory signal with a visual token, showing that both modalities can provide common prosodic information. Of course, these results pertain to matching speech within-speakers, and although different tokens were used there may have been particular individual idiosyncrasies that facilitated (and possibly exaggerated) matching performance. That is, even across modalities, individual production characteristics might have supported the high levels of performance rather than sensitivity to the underlying prosody. Given this, the participants' accurate matching performance does not in itself clearly indicate that visual cues from the upper face provide prosodic information beyond that which is specific to a token or speaker.

Previous studies have shown that perceivers are capable of extracting abstract phonemic information from visual signals regardless of who the speaker is. For example, cross-modal priming for spoken word recognition can occur when the signals originate from two different speakers [10], and McGurk effects (the integration of incongruent auditory and visual information resulting in a 'fused' percept) can be demonstrated across speakers [11]. Such findings suggest that both auditory and visual signals from different speakers are encoded as abstract representations and thus can be integrated with each other. However, to our knowledge, no studies have examined whether people can extract abstract suprasegmental information from visual cues that can be generalised across speakers, and whether performance would vary as a function of the face area visible. As such, the current study investigated this by testing whether visual prosody information from both the upper and lower face can be matched when the signals originate from different speakers. Experiment 1 tested within-speaker prosody matching when the speaker's lower face was presented. Taken together with the results of our previous study which presented the upper face in the same tasks [1], these data provided a baseline for which to evaluate cross-speaker prosody matching (Experiment 2).

2. Experiment 1

As in [1], this experiment included two types of prosodic speech conditions: prosodic focus and sentence mode. Prosodic focus describes the situation where one word is made more perceptually salient than other words in a sentence, and is used to emphasise importance or disambiguate a particular item within the sentence (*narrow focus*) [12]. This was contrasted with *broad focused* statements which have no explicit point of informational focus. Sentence mode refers to acoustic changes made to achieve different sentence phrasings, such as statements or questions. By mimicking the syntactic content of a declarative statement, an *echoic question* can be phrased without the use of an interrogative pronoun [13].

It has been suggested that the type of prosodic cues expressed on the speaker's face may vary across different face areas. That is, [6] showed that people were able to accurately identify narrow focused words and sentence mode when presented silent visual displays of a speaker's full face. When motion information from the upper face was unavailable, identification accuracy was maintained for narrow focused words but not for sentence mode, indicating that the prosodic information for sentence mode was more readily available from the upper than the lower face area. Given this, it is important to ascertain what prosodic cues are available from which part of the speaker's face. This will allow us to specify the type of prosodic cues and their distribution on the face when investigating the extent to which visual cues can signal the same prosodic information across speakers (Exp 2).

2.1. Method

2.1.1. Materials

This and the following experiment used the materials of our previous study [1]. These consisted of 10 non-expressive sentences drawn from the IEEE [14] list describing mundane events with minimal emotive content. Auditory and visual speech of two native male speakers of standard Australian English (M_{Age}= 23 years) were recorded in a well-lit, sound attenuated room using a digital video camera (25 fps), with audio recorded at 44.1 kHz, 16-bit mono with an externally connected lapel microphone.

Each sentence was recorded in three speech conditions: as a *broad focused* statement, a *narrow focused* statement, and as an *echoic question*. To elicit these conditions, a dialogue task was used that required the speakers to interact with an interlocutor, and either repeat what they heard the interlocutor say (broad focused statement), make a correction to an error made by the interlocutor (narrow focused statement, 1a/b), or question an emphasised item produced by the interlocutor (echoic question, 2a/b). An example of this dialogue is given below:

(1) **a.** The pipe ran almost the [width]$_{\text{Error}}$ of the ditch.
b. The pipe ran almost the [**length**]$_{\text{Correction}}$ of the ditch.
(2) **a.** The green light in the [**brown**]$_{\text{Emphasised}}$ box flickered.
b. The green light in the [*brown*]$_{\text{Questioned}}$ box flickered?

Two repetitions of each sentence were recorded several minutes apart. The critical item within each sentence (i.e., the word within the sentence that received narrow focus or question intonation) was kept consistent across speech conditions, speakers and repetitions. This recording procedure resulted in 120 auditory and 120 visual tokens for use as stimuli. The visual tokens were then cropped using VirtualDub [15] to generate two versions of visual stimuli; upper half videos that showed the speaker from above the tip of the nose only, and lower half videos that displayed only the lips, lower cheeks, chin and jaw of the speaker.

2.1.2. Participants

Twenty undergraduate students from the University of Western Sydney (UWS) participated in the study for course credit. None had previously participated in [1]. All were fluent speakers of English and had self-reported normal or corrected-to-normal vision and no history of reported hearing loss.

2.1.3. Procedure

Participants were tested individually in a sound-attenuated booth, with stimuli presented on a 17" LCD computer monitor. Each participant completed two experimental tasks; a visual-visual (VV) matching task, and an auditory-visual (AV) matching task (as used in [1 & 16]), in counter-balanced order.

2.1.3.1 Visual-Visual Matching Task

Stimuli were presented in a two-interval, alternate forced-choice (2AFC) discrimination task. Participants were told that they would be viewing two pairs of silent videos showing only the lower part of the speakers faces, and that their task was to select the pair in which the sentences were produced with the same prosody. The non-matching display was identical in lexical content, and differed only in prosody.

The same speaker was shown for both items within-pairs (Figure 1A). To avoid instance-specific matching, the matching items in the correct pair were always taken from a different token. Participants indicated their response as to which pair had the same prosody via a selective button press. The order of correct response pair was counter-balanced, so the correct option appeared equally in the first and second pair. DMDX [17] was used for video display, randomisation of items and collection of participant response data. In total, 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.

2.1.3.2 Auditory-Visual Matching Task

The AV matching task was similar to the VV task except it used auditory–visual stimulus pairs. Participants were told that they would be presented with two pairs of stimuli, each consisting of an auditory-only and a visual-only stimulus and that their task was to select the pair in which the visual display of prosody matched the auditory token (Fig 1B). Once again, visual information was restricted to the lower face. The initial auditory token that appeared at the start of each pair was the same, with each of the 120 auditory tokens appearing as the target once. The non-matching stimulus differed only in prosody, not lexical content. Auditory stimuli were presented binaurally via stereo headphones. Other details of stimuli presentation were similar to those of the VV matching task.

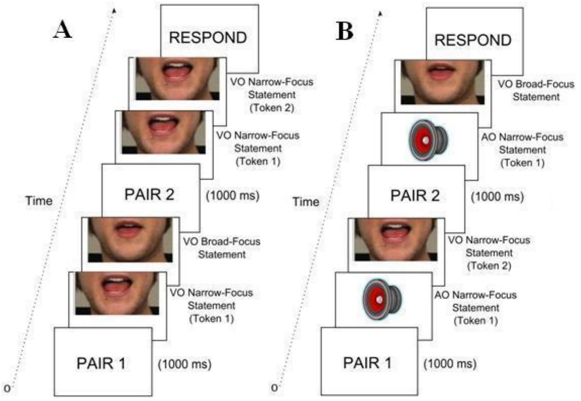
A
RESPOND
VO Narrow-Focus Statement (Token 2)
VO Narrow-Focus Statement (Token 1)
PAIR 2
(1000 ms)
Time
VO Broad-Focus Statement
VO Narrow-Focus Statement (Token 1)
PAIR 1
(1000 ms)
0
B
RESPOND
VO Broad-Focus Statement
AO Narrow-Focus Statement (Token 1)
PAIR 2
(1000 ms)
Time
VO Narrow-Focus Statement (Token 2)
AO Narrow-Focus Statement (Token 1)
PAIR 1
(1000 ms)
0

Figure 1. *Schematic representation of the 2AFC task used in the within-speaker (A) VV and (B) AV matching task. In both (A) and (B), the same item appeared first for both pairs, and was the standard that the matching judgment was to be made on. The matching item was always taken from a different recorded token.*

3. Results and Discussion

The results were analysed together with the data previously obtained from the presentation of the upper face [1] to allow for a full range of comparisons. Table 1 shows the mean percent of correct responses for the VV and AV tasks of the two studies. As can be seen, performance was considerably greater than chance for all conditions, confirmed by a series of significant one-sample *t*-tests.

For VV matching 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 (upper vs. lower half), with prosodic speech condition (narrow focus; echoic question) as the within-subjects factor. No significant main effect was found for prosodic speech condition (*F* (1,29) = 1.58, *p* = .22) or visible face area (*F* (1,29) < 1). However, the interaction was significant (*F* (1,29) = 10.67, *p* = .003), i.e., performance for discriminating narrow focus improved in the lower face condition, whereas performance for discriminating echoic questions was better with the presentation of the upper face. This result is consistent with the suggestion of [6].

For AV matching performance, a 2 (upper vs. lower face) ×3 (broad focus; narrow focus; echoic question) mixed repeated measures ANOVA was conducted. The main effect of prosodic speech condition was significant (*F* (2,25) = 8.70*, p* = .001), however the main effect of visible face area and the interaction were not significant (*F* < 1). Post-hoc comparisons showed that performance when matching auditory to visual tokens was significantly better for narrow focus items than both broad focus and echoic question items.

Table 1. *Mean percent correct responses in the within-speaker VV & AV tasks as a function of visible face area in each of the prosodic speech conditions. Data in bold is from* [1], *** indicates p* < .001.

Visible Face Area	Prosodic Speech Condition	Mean Correct (%)	Std. Error	*t*-test vs. chance (50%)
VV Task				
UpperHalf	*Narrow F*	**82.7**	**2.97**	*11.03***
(df = 10*)*	*Echoic Q*	**87.7**	**3.26**	*11.58***
Lower Half	Narrow F	91.7	2.87	20.92**
(df=19)	Echoic Q	80.5	2.84	10.28**
AV Task				
UpperHalf	*Broad F*	**88.9**	**2.14**	*18.15***
(df=10*)*	*Narrow F*	**94.8**	**1.61**	*27.79***
	Echoic Q	**88.8**	**2.39**	*16.25***
Lower Half	Broad F	87.3	3.12	14.70**
(df=16)	Narrow F	91.4	2.69	17.71**
	Echoic Q	84.4	2.85	12.34**

In sum, for within-speaker stimuli, participants were able to match visual speech to other visual or auditory speech tokens based on prosody, regardless of which face area was visible. As different tokens were used for matching items, the result supports the notion that visual prosody is realised consistently within-speakers over multiple repetitions.

4. Experiment 2

This experiment aimed to ascertain to what extent visual cues can signal the same prosodic information across speakers. The details of the study were similar to those of Exp 1 and [1].

4.1. Method

4.1.1. Participants

Thirty-two undergraduate students from UWS participated in return for course credit. None of these participants had taken part in the previous study, and all reported normal or corrected-to-normal vision and no history of reported hearing loss.

4.1.2. Materials & Procedure

The materials and procedure were the same as outlined for Exp 1 except that both upper and lower face stimuli were used and the paired stimuli for matching consisted of two tokens from different speakers.

Figure 2 outlines the composition of the stimuli in Exp 2. The first token (produced by one speaker) was the same for both pairs in a trial but the second token (produced by a different speaker) differed between the pairs; one pair with matching (same), and one with non-matching (different) prosody. Participants were told to select the pair that had the same prosody. As in Exp 1, the 2AFC procedure was used.

Participants were randomly allocated to a visible face area condition (upper/lower face), and completed both VV and AV matching tasks with the cross-speaker stimuli. Tasks were completed in a counter-balanced order.

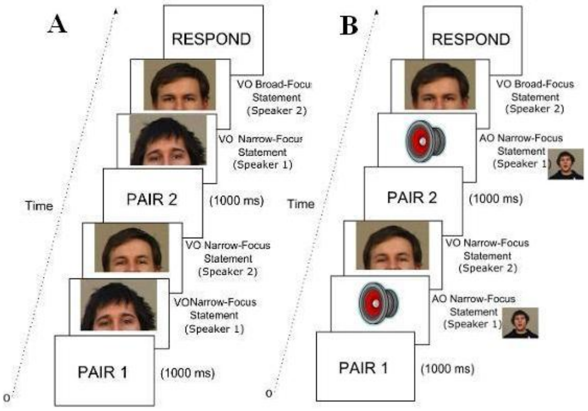

Figure 2. *Schematic representation of the 2AFC task used in the cross-speaker (A) VV and (B) AV matching tasks. The non-matching video in both (A) and (B) were always the same sentence produced with different prosody. Both upper and lower face stimuli were used.*

5. Results and Discussion

The mean percent of correct responses for both tasks using cross-speaker stimuli are presented in Table 2. Once again, participants were able to complete the task across all conditions at levels much greater than chance, confirmed by a series of significant one-sample *t*-tests.

Table 2. *Mean percent correct responses in the cross-speaker VV & AV tasks as a function of visible face area when the items within pairs came from different speakers. (df* = 15*), ** indicates p* < .001.

Visible Face Area	Prosodic Speech Condition	Mean Correct (%)	Std. Error	*t*-test vs. chance (50%)
VV Task				
Upper Half	Narrow F	70.9	4.33	4.83**
	Echoic Q	78.4	3.38	8.42**
Lower Half	Narrow F	85.9	2.89	12.42**
	Echoic Q	70.0	3.03	6.61**
AV Task				
Upper Half	Broad F	79.8	4.85	6.15**
	Narrow F	85.9	4.49	8.01**
	Echoic Q	81.4	3.93	8.00**
Lower Half	Broad F	81.6	2.39	13.19**
	Narrow F	94.2	1.22	36.16**
	Echoic Q	84.8	2.30	15.17**

A 2×2 mixed repeated measures ANOVA was conducted for VV task performance, with visible face area as the between-subjects factor and prosodic speech condition as the within-subjects factor. A significant main effect was found for prosodic speech condition (F (1,30) = 4.73, p = .038), but not for visible face area (F < 1). The significant prosody by face area interaction found with the within-speaker results was maintained even when visual signals were from different speakers (F (1,30) = 36.51, p < .001).

The cross-speaker AV task performance was analysed with a 2×3 mixed repeated measures ANOVA with visible face area as the between-subjects factor and prosodic speech condition as the within-subjects factor. The main effect of prosodic speech condition was significant (F (2,29) = 11.05, p < .001), however the main effect of visible face area (F (1,30) = 1.46, p >.10) and the interaction (F < 1) were not significant.

The results for cross-speaker prosody matching (Exp 2) were compared to the results obtained for within-speaker matching (Exp 1). A 2×2×2 ANOVA was conducted for VV task performance and a 2×2×3 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 = .002), but not for the AV task (F_{AV} (1,56) = 3.52, p = .066). Overall, performance across both tasks was greater for within-speaker prosody matching suggesting that although prosodic cues from both the upper and lower face include general non-speaker specific information, there is a considerable speaker-specific component. The main effect of visible face area was not significant for either task (Fs < 1), suggesting that both the upper and lower face can provide visual cues to prosody. For both the tasks, the main effect of prosody was significant (F_{VV} (1,59) = 5.50, p = .022; F_{AV} (2,112) = 11.67, p < .001), i.e., narrow focus seems to be easier to visually discriminate than broad focused statements and echoic questions.

No interactions (across both tasks) were significant (Fs < 1.5) except the prosody by visible face area interaction for the VV task (F_{VV} (1,59) = 40.15, p < .001), showing that upper and lower parts of the face convey different information dependant on the prosodic speech condition.

6. General Discussion

The aim of the current study was to investigate the distribution of visual cues for prosody on the face of a speaker and determine if these cues are general enough to signal the same prosodic information across different speakers.

The results showed that perceivers are sensitive to visual cues for prosody from both the upper and lower areas of the face. Furthermore, the lower face (e.g., mouth and jaw movements) conveys prosodic focus more efficiently (perhaps in terms of duration and amplitude, [1]) whereas the upper face (e.g., eyebrow movements) was more informative for ascertaining sentence mode (for which change in fundamental frequency might be important, [13]). Despite variation in visual cues across speakers [7], these cues appear to have been able to be processed to represent abstract, non-episodic visual speech events [11] and thus be generalised across tokens, modalities and speakers.

The current study is the first to our knowledge to demonstrate that visual cues from different speakers can be effectively matched based on the prosodic information conveyed. To generalise the current findings, further studies are required that include visual cues from multiple speakers. Also, conducting a study that directly examines the production of visual speech will be necessary to determine whether similarities can be found and quantified among the visual cues that represent the same type of prosodic information.

Acknowledgements:

The authors thank Bronson Harry and Catherine Gasparini for their assistance with recording the stimuli and conducting the perceptual experiments, and 3 anonymous reviewers for their useful feedback. The first author thanks MARCS Auditory Laboratories for their generous financial support. The second and third authors acknowledge support from the Australian Research Council (DP0666857 & TS0669874).

7. References

[1] Cvejic, E., Kim, J., and Davis, C., “Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion”, Speech Commun., 2010. doi: 10.1016/j.specom.2010.02.006

[2] Summerfield, Q., “Lip-reading and audiovisual speech perception”, Phil. Trans.: Bio. Sci., 335: 71-78, 1992.

[3] Yehia, H. C., Kuratate, T. and Vatikiotis-Bateson, E., “Linking facial animation, head motion and speech acoustics”, J. Phonetics., 30: 555-568, 2002.

[4] Munhall, K .G., Jones, J. A., Callan, D. E., Kuratate, T. and Vatikiotis-Bateson, E., “Visual prosody and speech intelligibility”, Psych. Science, 15: 133-137, 2004.

[5] Swerts, M. and Krahmer, E., “Facial expression and prosodic prominence: Effects of modality and facial area”, J. Phonetics., 36: 219-238, 2008.

[6] Lansing, C. R. and McConkie, G. W., “Attention to facial regions in segmental and prosodic visual speech perception tasks”, J. Speech, Lang. & Hearing Res., 42: 526-539, 1999.

[7] Dohen, M., Lœvenbruck, H. and Hill, H., “Recognizing prosody from the lips: Is it possible to extract prosodic focus from lip features?”, in A.W.-C. Liew and S. Wang [Eds], Visual Speech Recognition: Lip Segmentation and Mapping, 416-438, IGI Global, 2009.

[8] Dohen, M. and Lœvenbruck, H., “Interaction of audition and vision for the perception of prosodic contrastive focus”, Lang. & Speech, 52: 177-206, 2009.

[9] Scarborough, R., Keating, P., Mattys, S. L., Cho, T. and Alwan, A., “Optical phonetics and visual perception of lexical and phrasal stress in English”, Lang. & Speech, 52): 135-175, 2009.

[10] Buchwald, A. B., Winters, S. J. and Pisoni, D.B., “Visual speech primes open-set recognition of spoken words”, Lang. & Cog. Processes, 24: 580-610, 2009.

[11] Green, K. P., Kuhl, P. K., Meltzoff, A. N. and Stevens, E. B., “Integrating speech information across talkers, genders, and sensory modality: Female faces and male voices in the McGurk effect”, Perc. & Psychophysics, 50: 524-536, 1991.

[12] Krahmer, E. and Swerts, M., “On the alleged existence of contrastive accents”, Speech Commun., 34: 391-405, 2001.

[13] Eady, S. J. and Cooper, W. E., “Speech intonation and focus in matched statements and questions”, J. of the Acoust. Soc. of America, 80: 402-415, 1986.

[14] IEEE Subcommittee on Subjective Measurements, “IEEE recommended practices for speech quality measurements”, IEEE Trans. on Audio & Electroacoustics, 17: 227-246, 1969.

[15] Lee, A., VirtualDub, Version 1.9.3 [Software]: http://www.virtualdub.org, 2009.

[16] Davis, C. and Kim, J., “Audio-visual speech perception off the top of the head”, Cognition, 100: B21-B31, 2006.

[17] Forster, K. I. and Forster, J. C., “DMDX: A windows display program with millisecond accuracy”, Behav. Res. Methods: Instruments & Computers, 35: 116-124, 2003.