*Perceiving Prosody from the Face and Voice: Distinguishing Statements from Echoic Questions in English**

Ravindra J. Srinivasan
Dominic W. Massaro
University of California, Santa Cruz

Key words

bimodal

perception

prosody

question

synthesis

Abstract

We examined the processing of potential auditory and visual cues that differentiate statements from echoic questions. In Experiment 1, four natural speech statement-question pairs were identified by participants, and then analyzed to determine which characteristics were ecologically valid. These characteristics were tested in subsequent experiments to determine if they were also functionally valid. In Experiment 2, the characteristics of the most discriminable utterance pair were successfully extended to the other utterance pairs. For Experiment 3, an auditory continuum (varying in F0, amplitude, duration) was crossed with a visual continuum (varying in eyebrow raise, head tilt), using synthetic speech and a computer-animated head. Participants judged five levels along each of these two speech continua between a prototypical statement and prototypical question, in an expanded factorial design. Experiments 4 and 5 were unable to appreciably enhance the weak visual effect relative to the strong auditory effect (from Experiment 3). Overall, we found that both auditory and visual cues reliably conveyed statement and question intonation, were successfully synthesized, and generalized to other utterances. However, the weak visual effect relative to the robustly strong auditory effect precluded optimal integration and conclusive examination of information processing through model-fitting.

1 Introduction

This study aims to find which suprasegmental characteristics of the voice (e.g., pitch contour) and face (e.g., eyebrow raising) are functional in distinguishing statements and questions. Past studies have examined the ecological properties of the statement and

* *Acknowledgements:* The research and the writing of the paper was supported by grants from the National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), Intel Corporation, the University of California Digital Media Program, and the University of California, Santa Cruz. The authors would like to thank Michael Cohen for invaluable technical and eclectic assistance, and appreciate the reviewers for their insightful comments that helped clarify and strengthen the paper.

Address for correspondence: Dr. Dominic W. Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064 USA. Phone: 831-459-2330. Fax: 831-459-3519; e-mail: <massaro@fuzzy.ucsc.edu>. URL: <http://mambo.ucsc.edu/psl/dwm/>.

question prosody. Lieberman (1967) had participants read statements (like *Joe ate the soup.*) and echoic questions (like *Joe ate the soup?*). The echoic question has the same word order as the statement. However it is interrogative in nature, sometimes marked by an element of surprise. The recordings were analyzed for pitch information. The statements were characterized by a falling terminal fundamental frequency (F0), whereas the questions were marked by a final rise in terminal F0.

O'Shaughnessy (1979) analyzed sentences produced by four speakers and identified three critical regions: the first, medial, and last accented syllables of a question. He observed that all three of these syllables are characterized by a rising F0 contour, and concluded that the question intonation affects the F0 contour of the entire sentence, and is not limited to a rising contour at the end of the utterance. An interaction between speaking rate and intonation in French by Canadian speakers was reported by Ryalls, Dorze, Lever, Oullet, & Larfeuil (1994). They looked at the duration of matched statements ("The bird is singing.") and echoic questions ("The bird is singing?"). Although questions were spoken at a faster rate than statements, the final syllable durations were significantly longer for questions than statements. This interaction between speaker rate and sentence type has not been corroborated in English. What perceptual effect all of these potential cues have in statement-question differentiation also needs to be examined.

Several researchers have examined the characteristics of statement-question prosody that are functional in perception. Majewski and Blasdell (1969) recorded the word *farmer* spoken as a statement or as a question. The word was then synthesized and presented to participants for identification. Their conclusion was that questions and statements could potentially be distinguished from each other based on terminal F0 contour alone. Studdert-Kennedy and Hadding (1973) suggested, however, that listeners also rely on other aspects of the F0 contour in differentiating statements and questions. In their study, they imposed synthetic contours on naturally spoken utterances and manipulated the F0 of certain points on the contour. The perceptual judgments of statement or question indicated that listeners were influenced not just by the terminal F0 but the entire F0 contour.

Prosody is perceived not only in the voice, but also in visual information from the face. The visual aspects (facial expressions) associated with prosody have been explored only recently, both unimodally (alone) and multimodally (in conjunction with auditory cues). Lansing and McConkie (1999) monitored eye gaze while participants made decisions about segmental and prosodic categories for utterances presented without sound. The observers looked longer and directed more gazes toward the upper part of the speaker's face in making decisions about intonation patterns (statements vs. questions) than word segments. They could recognize prosody and segmental information based on visual cues in the upper and lower facial regions respectively. However, recognition of prosodic information from visual cues alone was more difficult than that of segmental or primary sentence stress.

Granstrom, House, and Lundeberg (1999) found that eyebrow movement (raising) can serve as an independent prosodic cue to prominence. In another study, House, Beskow, and Granstrom (2001) systematically manipulated the timing of both the eyebrow and head movements of a talking face, in a test sentence with an audiovisual speech synthesizer. For all sentences, the audio speech signal was kept the same. When participants had to indicate the most prominent word in the test sentence, both the eyebrow and head movements influenced the judgments.

Eyebrow and head movements were recently explored as feedback cues in human-machine interaction. Granstrom, House, and Swerts (2002) investigated the influence of acoustic and visual cues in signaling “negative” or “affirmative” feedback. The task was to judge a series of exchanges between a talking head (travel agent) and a human (customer) on the basis of their role as feedback signals. Surprisingly, brow raising was found to be an affirmative cue, despite usually being indicative of a question or surprise intonation. However, the brow movement was subtle in this study, and a larger raising movement might be more likely to signal a question intonation. In a similar vein, House (2002) found that visual cues such as eyebrow movement and slow vertical head tilting did not indicate interrogative intonation, but elicited rather complex responses that were more affirmative.

Massaro and Beskow (2002) studied the joint influence of F0, loudness, eye widening, and eyebrow movements on the perception of stress. Using a factorial design methodology, they manipulate eye widening, eyebrow movement, amplitude, and pitch independently of one another to determine their relative contributions to the perception of stress. Participants were asked to indicate the degree to which a given word in a sentence was stressed. Although all four independent variables had some influence on the judgments, the amplitude of the noun was the most influential factor. In quantitative tests of formal models of speech perception and language, the fuzzy logical model of perception (FLMP) gave the best description of the results. Following this strategy, the present study intends to examine possible prosodic cues in the face that differentiate statements from questions, and how they are combined with auditory cues to the same distinction.

This study also informs an ongoing research project using a computer-generated 3-D talking head (‘Baldi') to simulate auditory and visual aspects of human speech production (Massaro, 1998). One of its applications is as an interactive language tutor which teaches language and vocabulary (Massaro, Cohen, Beskow, & Cole, 2000). Realistic and convincing prosody is important for this application. Terken and Lemeer (1988) emphasize the importance of improving prosody in synthetic speech to make it more natural and comprehensible. They compared quality judgments for natural and synthetic utterances with good and poor segmental quality. The natural stimuli were 21 sentences read by a male speaker from a newspaper article. The same sentences were then synthesized using a Dutch text-to-speech conversion system with intonation contours that obey the rules of the intonation system for Dutch ('t Hart & Collier, 1975). Participants rated the synthetic utterances significantly lower in intelligibility than the natural ones, apparently due to the lack of appropriate intonation, which sounded rather “dull.” Natural intonation was preferred to the synthetic dull intonation in speech with good segmental quality, but not in speech with poor segmental quality. An implication is that as the segmental quality of synthetic speech improves, listeners will be more demanding of the naturalness of synthetic intonation.

This study extends past research by examining potential cues from both the voice and the face, and their relative contributions to the judgment of statements/echoic questions. Statement/echoic question pairs are used in order to eliminate any potential contribution of syntactic or semantic information. It illustrates a method to measure and synthesize prosodic characteristics from natural speech, create auditory and visual synthetic continua from statement to question, and present these in perception experiments using an expanded factorial design. It also seeks to understand the processing of audio and visual cues by testing theoretical models of information processing like the FLMP (fuzzy logical model of perception), the SCM (single channel model), and the WTAV (weighted averaging model). The study intends to: (1) Replicate and supplement findings on how natural statements and questions differ in auditory (vocal) and visual (facial) characteristics, (2) Create synthetic utterances based on these measurements, (3) Examine the processing of auditory and visual prosodic cues in perceiving statements and questions, and (4) Explain the findings in terms of information processing theories and model fits. Since the FLMP has been proven to be more successful than other models in similar bimodal speech perception domains (Massaro, 1998), it is hypothesized that the FLMP will best account for and explain the processing of prosodic information better than competing models of perception.

2 Experiment 1

The first experiment presented natural statements and echoic questions to participants, and examined the auditory and visual characteristics of the statement-question pairs. The most discriminable statement-question pair (*We will weigh you.*/ *We will weigh you?*) was determined from the identification judgments. The prosodic information thus obtained was used to build and test the stimuli in subsequent experiments.

2.1 *Method*

2.1.1 *Participants*

Twenty-two undergraduate psychology students at University of California, Santa Cruz participated in this experiment for class credit. Participants ranged in age from 18 to 22, and reported having normal hearing.

2.1.2 *Stimuli*

Four English utterances in statement form and echoic question form (Table 1) were obtained from recordings of a male American English speaker, played from a laser videodisc recorded by Bernstein and Eberhardt (1986). Each of the four utterances had three versions with word stress on either the first, second, or third word. These stimuli were selected because they were clearly articulated by a radio announcer and appeared to effectively convey the prosodic information. Being recorded on a laser videodisc also made it easier to analyze the acoustic and visual properties. The program used to run the experiment and collect subject data was implemented on a Silicon Graphics 4D-/Crimson VGX workstation running under the IRIX operating system. The sentences were presented in three modalities: (1) Audio (voice) only, (2) Visual (face) only, and (3) Audiovisual (both voice and face). There were four utterances × three stress placements (first/second/third word) × two types (statement/question) × three modalities (audio/visual/bimodal) yielding a total of 72 conditions. Final (fourth) word stress placement was not part of the video because it might interact with the final rise of question intonation. Some pretesting indicated that the audio only and bimodal conditions were showing a ceiling effect, and could not be distinguished from one another. Therefore, continuous white noise (with a signal-to-noise ratio of approximately -2.5dB) was added to the experiment to enable one to look at any added contribution of the visual information to bimodal judgments.

2.1.3
Procedure

Each of the conditions was presented four times (for a total of 288 trials) in two sessions of 144 trials each. Each session lasted about 20mins, with a 5mins break between sessions. The stimuli were presented to the participants on 12-inch (30.48cm) NEC Model C12–202A color monitors; and participant responses were collected on TVI video display terminals (VDTs) and their associated keyboards. Up to four participants were run at a time, each seated in separate soundproof rooms. The items were presented in random order (within each of the 4 blocks of 72 trials) without replacement. The participants were instructed to attend to both the face and the voice, and identify the sentence as a statement or question. On a third of the trials, only voice was presented, on another third, only face, and the other third, both voice and face together. On each trial, participants identified the sentence as either a statement or a question by typing the letter ‘s' for statement or the letter ‘q' for question on a standard keyboard, with the letters ‘s' and ‘q' highlighted for easy access. After this test, a brief questionnaire was given asking what auditory and visual cues helped them identify statements and questions. They also rated what utterance best signified a statement-question pair in each condition (audio, visual, audiovisual). Data analysis was then performed on the Silicon Graphics workstation using Fortran 77 data analysis routines and the SAS statistical package (SAS Institute, Cary, NC).

2.2
Results

A two-factor analysis of variance was carried out on each of the four utterance pairs. The independent variables were the three sensory modalities (auditory, visual, bimodal), and the three stress placements (on first, second, third word). The dependent variable was the proportion of “question” responses, p(q), for each subject pooled across all trials. For each of the utterances, statements and questions were discriminated from one another significantly across all three modalities ($p < .01$) except for utterance #4 for which the visual effect was not significant (Table 1 overleaf lists the F values).

There was no significant influence of stress level on statement-question discrimination. To get a measure of how well the four statement-question pairs were discriminated in each of the modalities, d' values were computed from the p(q) responses for each subject. In accord with extant literature (Kadlec, 1999; Miller, 1996), the proportion 0 was computed as 1/2n and the value one as $(2n-1)/2n$, where n is the number of observations per condition (12 in this study, collapsing across stress placements). Table 2 lists the average of these d' values for each of the four sentence pairs under the three modalities. The higher

TABLE 1

F-ratios for the four Natural Statement/Question Pairs

Sentence pair: F(1, 21)	*Auditory*	*Visual*	*Bimodal*
1. We owe you a yo-yo. / We owe you a yo-yo?	36.63	6.83	66.04
2. Pat cooked Pete's breakfast. / Pat cooked Pete's breakfast?	26.41	8.19	67.13
3. We will weigh you. / We will weigh you?	157.80	125.78	301.67
4. Chuck caught two cats. / Chuck caught two cats?	226.18	2.09*	99.63

*Not significant at $p < .01$

the d' value of a pair, the better statements and questions were distinguished from one another. As seen in Table 2, the most well-differentiated statement-question pair was pair #3 (*We will weigh you./ We will weigh you?*). The d' values for this pair were higher than that of pair #1 and pair #2 across all three modalities. Although pair #4 had a higher auditory d' than that of pair #3, it was not distinguished visually (as evident from Table 1).

TABLE 2

d' Values for the four Natural Statement/Question Pairs

Sentence pair	*Auditory*	*Visual*	*Bimodal*
1. We owe you a yo-yo./ We owe you a yo-yo?	1.52	0.42	1.79
2. Pat cooked Pete's breakfast./ Pat cooked Pete's breakfast?	1.47	0.50	1.82
3. We will weigh you./ We will weigh you?	2.14	1.35	2.47
4. Chuck caught two cats./ Chuck caught two cats?	2.64	0.20	2.38

Table 3 shows the p(q) responses for the most well-discriminated natural statement/question pair #3.

TABLE 3

Proportion of "question" responses for the Natural Pair (We will weigh you./We will weigh you?)

Natural speech (with noise) proportion of "question" response	
Visual Statement	.155
Visual Question	.557
Auditory Statement	.064
Auditory Question	.705
Bimodal Statement	.042
Bimodal Question	.773

The above observations were corroborated by the questionnaires administered, in which participants indicated that utterance pair #3 was most discriminable and that the auditory cues conveyed significantly more prosodic information than the visual cues. The participants reported using the acoustic cues (pitch, duration, and amplitude) and visual cues (eyebrow raising, head tilting) to differentiate questions from statements. Our goal was to pick the statement/question pair that was most discriminable (and thus conveyed the greatest prosodic information) to use as a prototype for synthesis. Therefore the utterance #3 (*We will weigh you./ We will weigh you?*) and in specific the pair with the stress on the first word (chosen arbitrarily since stress did not interact significantly with statement/question discrimination) was used to create the synthetic stimuli for the following experiments. The acoustic characteristics of this particular utterance pair were examined using a spectrograph analysis tool called *Wavesurfer* (Sjolander & Beskow, 1999), which is a tool for recording, playing, editing, viewing, printing, and labeling audio data. It enables one to look at the pitch contour and duration (among other speech characteristics) of an utterance. An SGI (Silicon Graphics Interface) program was used to look at the amplitude of the pair. The following characteristics were noted: (1) The statement was characterized by a gradual decline in terminal F0 contour (from 97Hz–64Hz), a shorter final syllable duration of 200ms (overall utterance duration of 1192ms), and a sharp drop in amplitude (80%) on the final syllable, (2) The question was characterized by an entirely different contour with a high overall rise (from 86Hz to 170Hz) and slight terminal fall in F0 (from 170Hz to 148Hz), a longer final syllable duration of 280ms (overall utterance duration of 1289ms), and a smaller drop in amplitude (40%) on the final syllable.

The visual cues for this chosen utterance pair were examined using an SGI program called RIM, which enables one to capture video stills, mark points, and take measurements. It was used to measure the eyebrow raise and head tilt for the statement and question. It was found that (1) the statement was associated with little or no eyebrow raise, and insignificant head movement and (2) the question was accompanied by a significant eyebrow raise (20 units or 3.18mm) and head tilt (4°). The question cues extended dynamically across the length of the utterance. The eyebrow raise and head tilt initially increased in a monotonically decelerating fashion, peaked around the end of the second word (about 400ms into the utterance), and then persisted for the remainder of the utterance. These auditory and visual measurements were used to construct synthetic versions of statements and questions to test in subsequent experiments.

3 Experiment 2

This experiment used synthetic speech and facial animation to simulate the auditory and visual prosodic information from our unique test utterance. The other three utterances examined in our natural experiment (Experiment 1) were synthesized with the test utterance features. Even though the auditory cues were saliently informative in all four utterances (Table 1), the visual cues were not. If our visual cues (from the test utterance #3) are informative and robust, then these new sentences should be more discriminable than their natural counterparts. Positive results would enable one to generalize the effectiveness of the auditory and visual prosodic cues to some degree to the class of statements and echoic questions.

3.1 Method

3.1.1 Participants

Sixteen undergraduate psychology students at University of California, Santa Cruz, participated in this experiment for course credit.

3.1.2 Stimuli

The stimuli were the same four statement/question utterance pairs (from Experiment 1) synthesized based on the auditory and visual prosodic cues from the test utterance (statement/question pair #3). The pitch contour of the statement and question was obtained from Wavesurfer (Sjolander & Beskow, 1999). A speech software tool (that employs sable tags) called *MarkupGUI* (Woulters, Rundle, & Macon, 1999) was used to modify the acoustic (pitch contour, amplitude, duration) and visual (eyebrow, head tilt) parameters.

3.1.3 Procedure

The procedure was similar to that of the previous Experiment 1 except that each of the conditions was presented 16 times in two sessions (8 times per session). Each session lasted about 20mins, with a 5mins break between sessions. The total number of trials presented to each subject was 384 (192 trials per session).

3.2 Results

The independent variables were the synthesized utterances (four statement/question pairs) and the three sensory modalities (auditory, visual, and bimodal). The dependent variable was the p(q) responses. There was a significant effect of statement/question type, $F(1, 15) = 498.15, p < .01$, and this effect did not interact with utterance. To give a more detail measure of discrimination, d' values were computed for each subject and then averaged (16 observations were used to transform the 0 and 1 values).

TABLE 4

d' Values for the four Synthesized Statement/Question Pairs

Sentence pair	*Auditory*	*Visual*	*Bimodal*
1. We owe you a yo-yo. / We owe you a yo-yo?	3.44	1.88	3.45
2. Pat cooked Pete's breakfast. / Pat cooked Pete's breakfast?	3.10	1.93	3.09
3. We will weigh you. / We will weigh you?	3.34	1.67	3.30
4. Chuck caught two cats. / Chuck caught two cats?	2.77	1.93	3.27

Table 4 lists the d' values for each of the four synthetic sentence pairs under the three modalities. In order to evaluate the visual results of the synthetic utterances with respect to the natural ones, a single-factor analysis of variance was carried out comparing the synthetic visual d' values of Experiment 2 to the natural d' values of Experiment 1.

TABLE 5

Synthetic auditory continua

	(Statement)				*(Question)*
	Level 1	*Level 2*	*Level 3*	*Level 4*	*Level 5*
For Experiments 2, 3, and 4:					
Amplitude (% change)	–60	–50	–40	–30	–20
Speech Rate (% change)	0	–5	–10	–15	–20
Revised Experiment 5:					
Amplitude (% change)	–50	–45	–40	–35	–30
Speech Rate (% change)	–5	–8	–10	–13	–15

Table 5 shows the F-ratios of these comparisons. The synthesized utterance #3 was discriminated (between statement/question) as well as its natural counterpart, $F(1, 36)= 1.47, p > .01$, indicating that the visual synthesis of prosodic cues was successful and effective. With the visual cues of utterance #3 applied to them, the other three synthetic utterances were all discriminated significantly better than their natural visual-only counterparts, utterance # 1: $F(1, 36)= 26.30, p < .01$; utterance # 2: $F(1, 36)= 17.80$, $p < .01$; utterance # :4 $F(1, 36)= 34.83, p < .01$. Thus, the visual cues were effective and generalized to the new synthetic utterances. This does not imply that these particular visual cues necessarily apply to the prosodic utterances of other synthetic talking heads. These visual and auditory prosodic cues were varied independently of one another in Experiment 3 to study unimodal and bimodal information processing.

The above mentioned auditory and visual measurements were used to create synthetic versions of the ideal statement and ideal question. A five-level continuum was then made, going in equal steps from the ideal statement to the ideal question. The auditory continuum becomes more question-like with changing pitch contour (of the entire sentence), and increasing amplitude and duration (of the final syllable). The visual continuum becomes more question-like with increasing eyebrow raise and head tilt. These synthetic stimuli were then tested to see how discriminable they are, and how the auditory and visual cues are integrated in perception of prosody.

4 Models of perception

Experiments 3, 4, and 5 intend to look at how prosody is processed given two sources of information from the face and voice, respectively. In addition to traditional statistical analyses, the results will be used to test among quantitative models of perception.

The single channel model (SCM) is a nonintegration model, according to which only a single channel of information is used at any one time. So even when there are several inputs, the SCM predicts that only one of the multiple sources of information influences the response on any given trial. This idea is akin to selective attention theories, according to which only a single channel of information can be processed at any one time. In bimodal speech, the auditory and visual modalities are two channels of information. SCM would predict that only one of the auditory and visual inputs is functional on any given bimodal trial.

The SCM is in opposition to the integration theories. Integration models theorize that the perceptual experience is influenced by both auditory and visual information, which are evaluated and somehow used together in the pattern recognition process. We consider two integration models. The fuzzy logical model of perception (FLMP) has been a successful predictive model for similar perception tasks (Massaro & Cohen, 1993). The model supposes three stages of processing: (1) each source of information is evaluated to determine the continuous degree to which it matches stored prototypes; (2) the sources are integrated according to a multiplicative formula to calculate the overall degree of support for each alternative; and (3) a decision is made based on the relative goodness of fit with each prototype. The FLMP predicts that both auditory and visual modalities will influence the perception of prosody, and that the influence of one modality will be greater to the extent that the other is ambiguous.

In the weighted averaging model of perception (WTAV) the sources are averaged according to weight assigned to each modality. Though qualitatively different from the SCM (a nonintegration model), this WTAV is mathematically equivalent to the single channel model, and makes identical quantitative predictions (Massaro, 1987, 1998). This equivalence between the SCM and WTAV is sobering in that two very different models can make equivalent predictions. Given this result, it should not be too surprising that the SCM and FLMP also make fairly similar predictions (Massaro, 1998). They both predict main effects of the auditory and visual sources of information, whereas the SCM predicts no interaction between the sources and the FLMP predicts that the influence of one source will be largest when the other source is ambiguous. These two models are particularly difficult to distinguish if one of the sources of information has a relatively small influence and most of the responses tend to be in the middle of the factorial plot, as shown in Figure 2.

5 Experiment 3

5.1 *Method*

5.1.1 *Participants*

Forty-three undergraduate psychology students at University of California, Santa Cruz, participated in the Experiment 3 for course credit.

5.1.2 *Stimuli*

The stimuli consisted of the sentence *We will weigh you*, which was varied in prosodic information along an auditory and visual continuum of five levels from ideal statement to ideal question.

The auditory continuum consisted of changing pitch contour (shown in Fig. 1) and increasing amplitude and duration on the final syllable of the utterance (indicated in Table 5). A C program then interpolated and generated the five-level pitch contour continuum from the statement to the question endpoint.

Figure 1

Original five-level F0 continuum for Experiments 2, 3, and 4

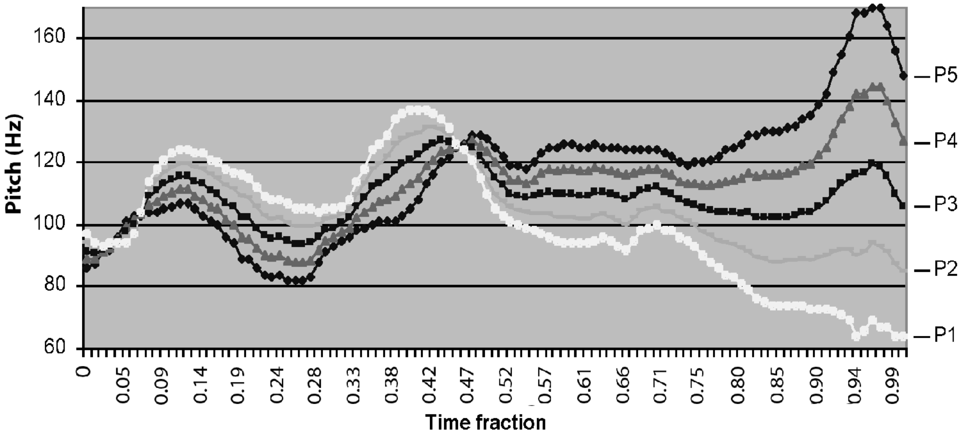


TABLE 6

Synthetic visual continua

	(Statement)				*(Question)*
	Level 1	*Level 2*	*Level 3*	*Level 4*	*Level 5*
For Experiments 2, 3:					
Eyebrow Raise (units) (where 20 units = 3.18 mm)	0	5	10	15	20
Head Tilt (degrees)	0	1	2	3	4
Revised Experiments 4, 5:					
Eyebrow Raise (units)	0	10	20	30	40
Head Tilt (degrees)	0	2	4	6	8

The visual continuum consisted of increasing eyebrow raising (0–20 units or 3.18 mm) and head tilting (0–4°) during the initial part of the utterance (see Table 6). Sable markup tags were used to generate the visual continuum for the synthetic talking head (‘Baldi'). These marked up auditory and visual continua constituted the MSS (Marked synthetic speech). The original default Festival synthetic speech (FSS) versions of the utterance were also presented. The FSS echoic question version resulted from affixing an interrogative character at the end of the statement, with the intonation determined by the default settings of the Festival text-to-speech synthesis system. These stimuli were included to see how well the experimentally marked up cues (MSS) performed compared to the unmarked default FSS representations.[1]

5.1.3 *Procedure*

In Experiment 3, there were a total of 41 test conditions. In addition to the six FSS conditions (Festival default statement/question pair × 3 modalities), another 35 MSS conditions were generated from an expanded factorial design. Crossing the five auditory and five visual levels yielded 25 bimodal conditions. Five of these bimodal conditions were consistent in terms of pairing an auditory level with the same level of the visual. The inconsistent bimodal conditions are a result of combining different levels of the auditory and visual dimensions. There were 10 unimodal conditions (5 audio only, and 5 visual only). Each of the conditions was presented eight times in two sessions (4 times per session). Each session lasted about 20mins, with a 5min break between sessions. The total number of trials presented to each subject was 328 (164 trials per session), randomized within each block of 41 conditions. Up to four participants were run at a time, each seated in front of 15-in. monitors in separate soundproof rooms. The audio was set constant at a comfortable intensity level, and Baldi's face was a constant size of 300 by 400 pixels animated at 30 frames per second. The synthetic face was displayed in the center of the screen and subtended a visual angle of about 10°. The trials were presented in random order without replacement, and participants were instructed to attend to both the auditory and visual stimuli. On each trial, participants identified each item as either a statement or a question by clicking the “Statement” button or the “Question” button. At the end of the experiment, the participants filled out a brief questionnaire describing what auditory and visual cues helped them identify statements and questions.

5.2 *Results*

TABLE 7

F-ratios for effects for statement/question type

		Unimodal effect	*Bimodal effect*
Auditory:	F(4, 168)	213.97	190.80
Visual:	F(4, 168)	47.19	11.86
Interaction:	F(16, 672)		2.38

*All significant at $p < .01$

[1] Links to sample experimental stimuli (synthetic bimodel continuum) can be found at <http://www.asel.udel.edu/lgsp/TOC/>.

A two-factor analysis of variance was carried out with the independent variables the three sensory modalities (auditory, visual, bimodal), and the five levels of the prosodic continuum (from most "statement"-like to most "question"-like). The dependent variable was the proportion of "question" responses, p(q), for each of the 41 conditions pooled across all trials.

For the default Festival speech, p(q) was very low for the questions, and was not significantly distinguishable from that for the statements. Table 7 shows the analysis of variance results for the expanded factorial design with the modified synthetic speech. For statement/question type, we see that the auditory and visual effects (in both unimodal and bimodal conditions), and the auditory-visual interaction were significant. These results are consistent with what participants reported in the questionnaire. They reported using the manipulated acoustic cues (pitch, duration, and amplitude) and visual cues (eyebrow raising, head tilting). They also reported ambiguity given the conflicting bimodal stimuli, and more influence of the auditory speech, consistent with identification results. The identification results or mean p(q) responses for the statements and questions are listed by modality in Table 8. Figure 2 (overleaf) plots the average identification results across all participants, showing the proportion of question judgments for the unimodal auditory, unimodal visual, and bimodal conditions.

TABLE 8

Proportion of "question" responses for the Synthetic Endpoints on unimodal and consistent bimodal trials (We will weigh you./ We will weigh you?)

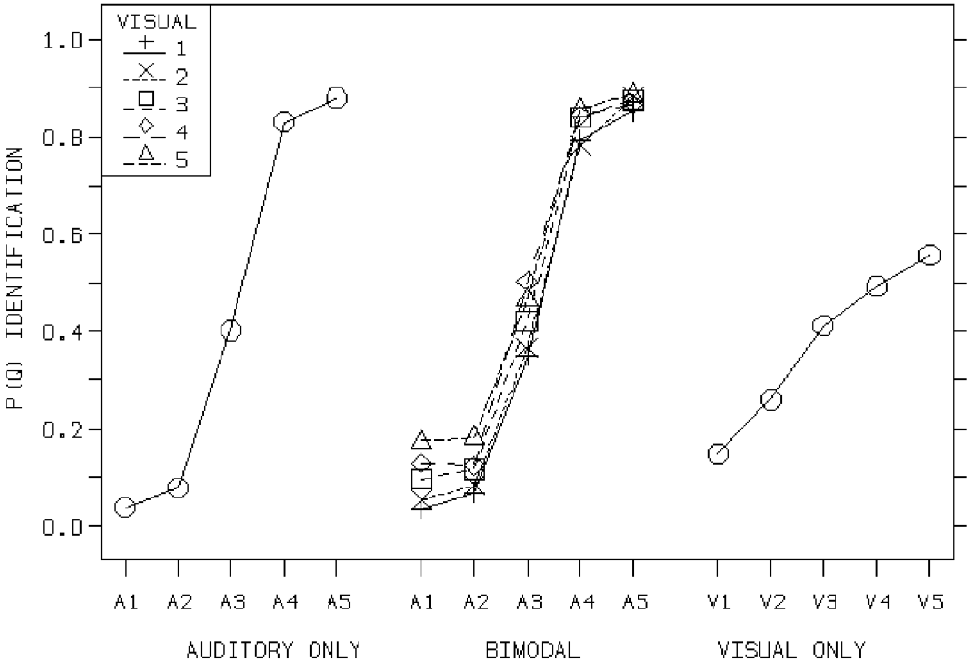
	Experiment 3	*Experiment 3*	*Experiment 4*	*Experiment 5*
	(Default)	*(Modified)*	*(Revised Visual)*	*(Revised Auditory)*
Visual Statement	.158	.156	.058	.054
Visual Question	.174	.563	.651	.548
Auditory Statement	.129	.039	.094	.030
Auditory Question	.132	.883	.935	.935
Bimodal Statement	.118	.036	.044	.018
Bimodal Question	.123	.890	.947	.905

Individuals may vary greatly in the relative influence of the audible and visible speech in bimodal perception. An index of the influence of a modality is given by the difference in average probability of a "question" response to the two endpoint stimuli from that modality presented unimodally. This marginal-range difference was calculated for each participant for both the audible and visible continua endpoints, giving an auditory effect and visual effect for each of the 43 participants in Experiment 3. The visual effect as a function of the size of the auditory effect for all the 43 participants are plotted in Figure 3.

Overall one notes that first, the size of the auditory and visual effects varied significantly across participants, as indicated by the points spread across the graph. Second, a few participants showed a zero visual effect and one a zero auditory effect, although

Figure 2

Observed proportion of question judgments as a function of the levels of the auditory and visual prosodic cues for Experiment 3. The left panel shows performance for just the unimodal auditory and the right panel for just the unimodal visual. The middle panel gives performance for the bimodal factorial combination of the auditory and visual. Average results across 43 participants are shown



most were influenced by both visible and audible speech. Third, the audible speech had a greater influence than visible speech, indicated by the fact that many observers are clustered at the lower right-hand corner of the graph.

It should be stressed that the large individual differences seen in Figure 3 do not make the present findings any less reliable or conclusive. Similar results have been repeatedly found in studies of the perception of bimodal segmental information (Massaro, 1998). Participants simply tended to differ in terms of the overall and relative influence of the auditory and visual information, not in terms of how the information from the two modalities is integrated. Similarly, the weaker influence of the visible speech only means that the auditory speech is more informative. It does not mean that perceivers do not use visual information to determine question versus statement. There is a significant influence of both modalities in the above experiments, even though the auditory modality is more informative than the visual. Although we were interested in using these data to distinguish between the SCM and FLMP, the descriptions of the two models gave about equally good fits. The failure to distinguish between the models was probably due to the relatively weak visual effect combined with fairly ambiguous response probabilities.

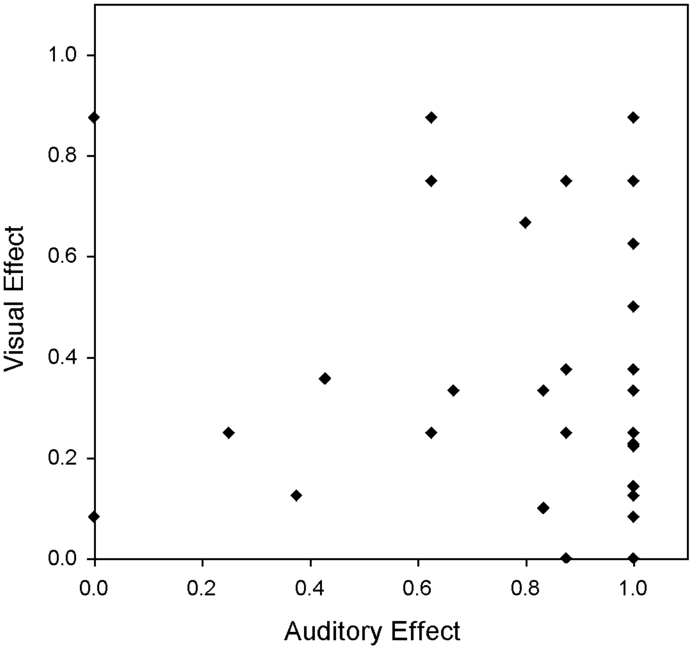


Figure 3

The visual effect plotted as a function of the auditory effect for the 43 participants in Experiment 3, showing the relative influence of the auditory and visual cues on proportion of question identification (Note: Several points are identical and hence overlap)

6 Experiment 4

The results of Experiment 3 indicated a larger influence of the audio than the visual component in bimodal judgments. To determine if synthetic facial cues are necessarily substantially less influential than auditory ones, Experiment 3 was modified (to create Experiment 4) by enhancing the visual cues to test whether the facial information could be more effectively engaged in the task. Increasing the visual effect relative to the auditory would provide a more definitive range of test data that can help distinguish between competing information processing models (Massaro, 1998). The auditory speech was kept the same as in Experiment 3.

6.1 *Method*

6.1.1 *Participants*

Seventeen undergraduate psychology students at University of California, Santa Cruz, participated in this experiment for course credit.

6.1.2 *Stimuli*

The auditory stimuli were the same synthetic utterances used in Experiment 3. They consisted of the sentence *we will weigh you* that was varied in prosody along a continuum of five levels from most statement-like to most question-like. The only difference was that the visual cues were doubled in magnitude. So the revised visual continuum (see

Table 6) consisted of increasing eyebrow raising (0–40 units or 6.36mm) and head tilting (0–8°) on the initial part of the utterance.

6.1.3 Procedure

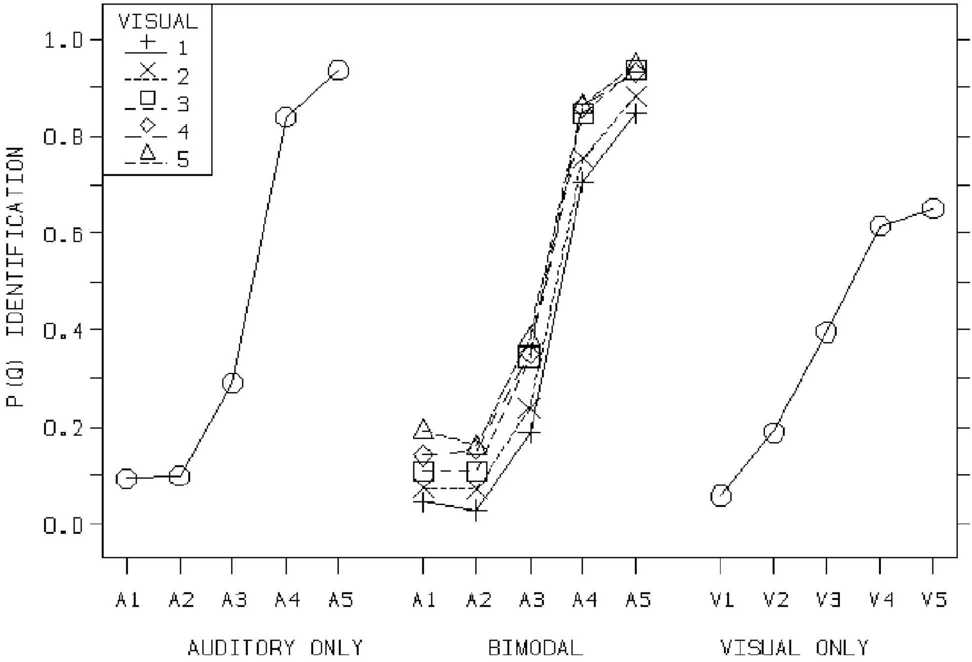
The procedure was similar to that of the previous Experiment 3 except that each of the conditions was presented 20 times in two sessions (10 times per session). Each session lasted about an hour long, and was conducted on two successive days. The total number of trials presented to each subject was 700 (350 trials per session). The FSS (default Festival speech) conditions were not presented because their weak effects were demonstrated in Experiment 3.

6.2 Results

The results were similar to that of the previous experiment. Figure 4 plots the average identification results across all participants, showing the proportion of question judgments for the unimodal auditory, unimodal visual, and bimodal conditions.

Figure 4

Observed proportion of question judgments as a function of the levels of the auditory and visual prosodic cues for Experiment 4. The left panel shows performance for just the unimodal auditory and the right panel for just the unimodal visual. The middle panel gives performance for the bimodal factorial combination of the auditory and visual. Average results across 17 participants are shown



A single-factor ANOVA was performed to compare the visual only performance on Experiment 3 (43 participants) to that on Experiment 4 (17 participants). The dependent variable was the visual effect. Although enhancing the visual cues produced more influential visible speech compared to Experiment 3, $F(1,58)= 6.41, p< .01$, the size of the influence was still relatively small. The pattern of results was similar to that of Experiment 3, as can be seen in Figure 5. Once again, the data did not distinguish between the SCM and FLMP.

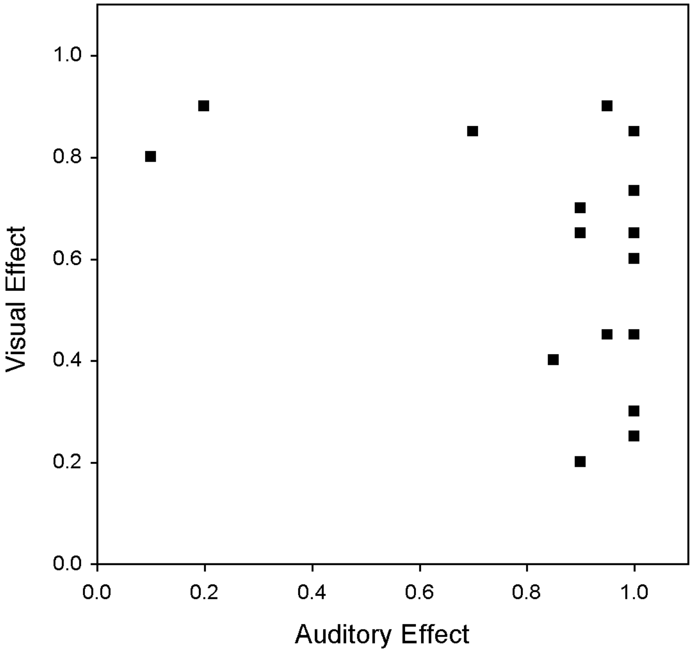


Figure 5

The visual effect plotted as a function of the auditory effect for the 17 participants in Experiment 4, showing the relative influence of the auditory and visual cues on proportion of question identification (Note: Several points are identical and hence overlap)

7 Experiment 5

Given that the visual manipulation in Experiment 4 was not effective enough, Experiment 5 modified the auditory continuum to make it more ambiguous, so that the visual signal could have a better chance to help disambiguate the signal in bimodal judgments. This implementation was accomplished by creating a new five-level continuum between the original second and fourth levels of the auditory stimuli.

7.1 *Method*

7.1.1 *Participants*

21 undergraduate psychology students at University of California, Santa Cruz, participated in this experiment for course credit.

7.1.2
Stimuli

The utterances used were the same as in Experiment 4. The visual continuum was kept the same as in Experiment 4. The auditory continuum was changed so that it had a narrower range between the second level to the fourth level on the statement-question continuum in Experiment 4 (see Table 5).

7.1.3
Procedure

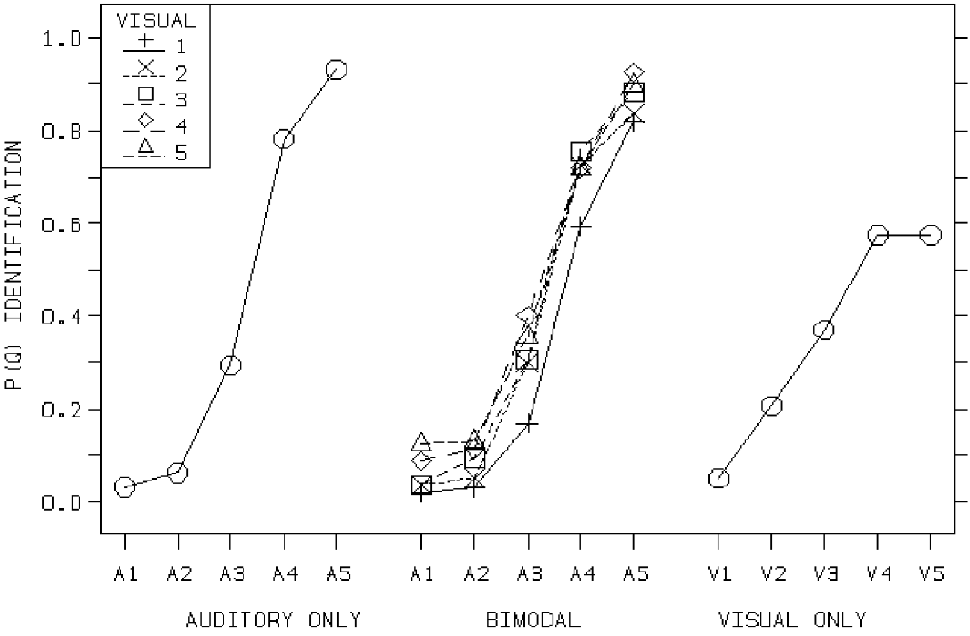
The procedure was similar to that of the previous Experiment 4 except that each of the conditions was presented eight times in two sessions (4 times per session). Each session lasted about 25 mins, with a 5 mins break between sessions. The total number of trials presented to each subject was 280 (140 trials per session).

7.1.4
Results

The pattern of results was similar to that of the previous Experiments 3 and 4. Figure 6 plots the average identification results across all participants, showing the proportion of question judgments for the unimodal auditory, unimodal visual, and bimodal conditions.

Figure 6

Observed proportion of question judgments as a function of the levels of the auditory and visual prosodic cues for Experiment 5. The left panel shows performance for just the unimodal auditory and the right panel for just the unimodal visual. The middle panel gives performance for the bimodal factorial combination of the auditory and visual. Average results across 21 participants are shown



Attenuating the differences in the auditory stimuli in Experiment 5 did not make the visible speech significantly more influential relative to Experiment 4. As can be seen in Figure 7, this manipulation was evidently not powerful enough to alter the relative influence of the audible and visible speech. As in Experiments 3 and 4, the models gave equally good descriptions of the results.

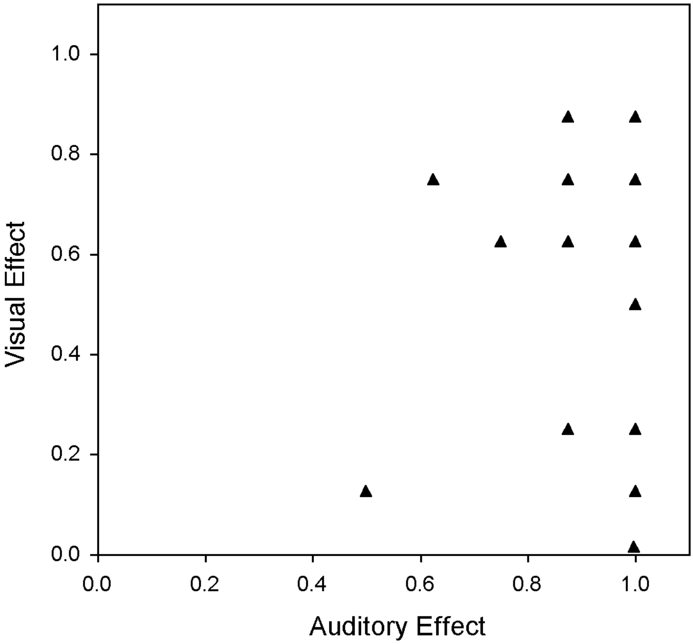


Figure 7

The visual effect plotted as a function of the auditory effect for the 21 participants in Experiment 5, showing the relative influence of the auditory and visual cues on proportion of question identification (Note: Several points are identical and hence overlap)

8 Discussion

There are functional prosodic cues in the face and the voice that can be marked to effectively distinguish statements from questions in synthetic speech. The present results show that statements and questions were discriminated auditorily (on the basis of the F0 contour, amplitude, and duration), and visually (based on the eyebrow raise and head tilt). The findings suggest that there were clear auditory (vocal) and visual (facial) prosodic cues that conveyed statement and question intonation. Although these cues were successfully synthesized, perceived, and applied to the other utterances in this study, a caveat is that these cues might not necessarily generalize to the prosodic utterances of other synthetic talking heads.

The present study revealed a much larger influence of the auditory cues than visual cues in the judgment of statement versus question. These results were consistent with those reported for Swedish by House (2002), which found that visual cues such as eyebrow movement and slow vertical head tilting did not strongly signal interrogative intonation. This held true despite our attempt to enhance the visual cues (Experiment 4) and attempting to make the auditory information more ambiguous (Experiment 5). This outcome limited how much could be learned about how the auditory and visual cues are processed. Normally, model tests are carried out to determine how the cues are evaluated and integrated to achieve identification. The robustly strong auditory cues and relatively much weaker visual cues, however, produce a situation in which very different models make similar predictions. For example, we found that the SCM and the FLMP (Massaro, 1998) gave roughly equivalent predictions for the results of 3, 4, and 5. It should be valuable to explore further means to strengthen the visual effect relative to the auditory effect in order to quantitatively examine information processing through model testing.

Although we expected the FLMP to predict the integration of auditory and visual information better than competing models of perception, the experiments proved otherwise. The FLMP did not prove significantly better than the WTAV/SCM. The WTAV/SCM could not be eliminated as a competing explanation to the FLMP. In accordance with the nonintegrative SCM, it is possible that only a single modality was being used at a time on any given trial (the more influential auditory channel being selected with a greater probability than the visual channel) in this bimodal perception task. These results also demonstrate that the FLMP is not superpowerful (Massaro, 1998, Chap. 11).

The lack of an advantage for the FLMP could be due to the nature of this task, which involves integrating cues from several modalities across the length of a sentence. Since the information is not confined to a single small event (like a syllable or word), it might be harder to engage integration in the optimal way predicted by the FLMP. Integrating the auditory and visual information across a sentence would probably require that the visual information from one part of the sentence be integrated with the auditory information from another part. Persons have trouble, however, optimally integrating information occurring during different time periods (Massaro, 1998, Chap. 3). The FLMP also tends to do better explaining the more automatic perceptual tasks (Massaro & Cohen, 1983) than more cognitive decision-making processes (Massaro, 1994). To assess whether the extended length of the sentence was responsible for nonoptimal integration, a shorter test stimulus (e.g.: "Sunny. / Sunny?") might be used. A short utterance might make statement/question identification a more automatic perceptual task, and less of a cognitive decision-making process. This task might engage an optimal bimodal integration process.

Several important aspects remain to be examined with regard to the perception of statement and question prosody. The present study tested only a few auditory (pitch contour, amplitude, duration) and visual (eyebrow raising, head tilting) cues together. Additional auditory and visual cues could be obtained from a larger corpus of statements and questions recorded by natural speakers. These cues could then be synthesized and manipulated independently to measure the effects of each individually. More generally, it might be useful to look at utterances of different lengths and different kinds (yes/no and *wh* questions) to see whether similar auditory and visual prosodic characteristics are observed, and what generalizations can be made about perceiving statement and question prosody.

Received: September 11, 2002; revised manuscript received: February 03, 2003; accepted: May 27, 2003

References

BERNSTEIN, L. E., & EBERHARDT, S. P. (1986). *Johns Hopkins lip-reading corpus video-disc set*. Baltimore, MD: Johns Hopkins University.

GRANSTROM, B., HOUSE, D., & LUNDEBERG, M. (1999). Prosodic cues in multimodal speech perception. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 99)*, 655–658.

GRANSTROM, B., HOUSE, D., & SWERTS, M. G. (2002). Multimodal feedback cues in human-machine interactions. In B. Bel & I. Marlien (Eds.), *Proceedings of the Speech Prosody 2002 Conference*, 347–350.

't HART, J., & COLLIER, R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics*, 235–255.

HOUSE, D. (2002). Intonation and visual cues in the perception of interrogative mode in Swedish. *Proceedings of ICSLP* 2002, 1957–1960.

HOUSE, D., BESKOW, J., & GRANSTROM, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. *Proceedings of Eurospeech* 2001, 387–390.

KADLEC, H. (1999). Statistical properties of d' and beta estimates of signal detection theory. *Psychological Methods*, 22–43.

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*, 526–539.

LIEBERMAN, P. (1967). *Intonation, perception and languages.* MIT Press: Cambridge, MA.

MAJEWSKI, W., & BLASDELL, R. (1969). Influence of fundamental frequency cues on the perception of some synthetic intonation contours. *Journal of the Acoustical Society of America*, 450–457.

MASSARO, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

MASSARO, D. W. (1994). A pattern recognition account of decision making. *Memory and Cognition*, 616–627.

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

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

MASSARO, D. W., & COHEN, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 753–771.

MASSARO, D. W., & COHEN, M. M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, 115–124.

MASSARO, D. W., COHEN, M. M., BESKOW, J., COLE, R. A. (2000). Developing and evaluating conversational agents. In J. Cassell, J. Sullivan, S. Prevost, E. Churchill (Eds.), *Embodied conversational agents* (pp. 287–318). MIT Press: Cambridge, MA.

MILLER, J. (1996). The sampling distribution of d'. *Perception & Psychophysics*, 65–72.

O'SHAUGHNESSY, D. (1979). Linguistic features in fundamental frequency patterns. *Journal of Phonetics*, 119–145.

RYALLS, J., DORZE, G., LEVER, N., OULLET, L., & LARFEUIL, C. (1994). The effects of age and sex on speech intonation and duration for matched statements and questions in French. *Journal of the Acoustical Society of America*, 2274–2276.

SJOLANDER, K., & BESKOW, J. (1999). *WaveSurfer—an open source speech tool.* Stockholm, Sweden: Center for Speech Technology (CTT) at KTH. <http://www.speech.kth.se/wavesurfer>.

STUDDERT-KENNEDY, M., & HADDING, K. (1973). Auditory and linguistic processes in the perception of intonation. *Language and Speech*, 293–313.

TAYLOR, P. A., BLACK, A., & CALEY, C. (1998). The architecture of the Festival speech synthesis system. *Proceedings of Third ESCA Workshop in Speech Synthesis*, 147–151.

TERKEN, J., & LEMEER, G. (1988). Effects of segmental quality and intonation on quality judgments for texts and utterances. *Journal of Phonetics*, 453–457.

WOUTERS, J., RUNDLE, B., & MACON, M. (1999). Authoring tools for speech synthesis using the sable markup standard. *Proceedings of Eurospeech*, 963–966.