massachusetts institute of technology — artificial intelligence laboratory

Perceptual Evaluation of Video-Realistic Speech

Gadi Geiger, Tony Ezzat and Tomaso Poggio

AI Memo 2003-003 Feburary 2003
CBCL Memo 224

abstract

With many visual speech animation techniques now available, there is a clear need for systematic perceptual evaluation schemes. We describe here our scheme and its application to a new video-realistic (potentially indistinguishable from real recorded video) visual-speech animation system, called Mary 101.

Two types of experiments were performed: a) distinguishing visually between real and synthetic image-sequences of the same utterances, ("Turing tests") and b) gauging visual speech recognition by comparing lip-reading performance of the real and synthetic image-sequences of the same utterances ("Intelligibility tests").

Subjects that were presented randomly with either real or synthetic image-sequences could not tell the synthetic from the real sequences above chance level. The same subjects when asked to lip-read the utterances from the same image-sequences recognized speech from real image-sequences significantly better than from synthetic ones. However, performance for both, real and synthetic, were at levels suggested in the literature on lip-reading. We conclude from the two experiments that the animation of Mary 101 is adequate for providing a percept of a talking head. However, additional effort is required to improve the animation for lip-reading purposes like rehabilitation and language learning.

In addition, these two tasks could be considered as explicit and implicit perceptual discrimination tasks. In the explicit task (a), each stimulus is classified directly as a synthetic or real image-sequence by detecting a possible difference between the synthetic and the real image-sequences. The implicit perceptual discrimination task (b) consists of a comparison between visual recognition of speech of real and synthetic image-sequences. Our results suggest that implicit perceptual discrimination is a more sensitive method for discrimination between synthetic and real image-sequences than explicit perceptual discrimination.

This report describes research done at the Center for Biological & Computational Learning, which is in the Dept. of Brain & Cognitive Sciences at MIT and which is affiliated with the McGovern Institute of Brain Research and with the Artificial Intelligence Laboratory.

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, and National Science Foundation-NIH (CRCNS) Contract No. IIS-0085836.

Additional support was provided by: AT&T, Central Research Institute of Electric Power Industry, Center for e-Business (MIT), DaimlerChrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., The Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone (NTT), Oxygen, Siemens Corporate Research, Inc., Sony MOU, Sumitomo Metal Industries, Toyota Motor Corporation, WatchVision Co., Ltd., and The Whitaker Foundation.

Introduction:

Visual animation of speech, where a head is seen talking, with or without the sound of speech, has generated great interest and has been attempted by numerous researchers (Brand 1999, Bregler *et al.* 1997, Brooke and Scott 1994, Cohen and Massaro 1993, Cosatto and Graf 1998, Ezzat *et al.* 2002, Lee *et al*. 1995, Le Goff and Benoit 1996, Masuko *et al.* 1998, Parke 1974, Pearce *et al.* 1986, Vatikiotis-Bateson *et al.* 2000, Waters 1987). There are at least two major goals for video-speech animation. One is for the aesthetic pleasure, as in movies, where we get the feeling of "realistic images". The other is for help in communication, and specifically understanding uttered words.

Perceptual evaluations of the animations differ, depending on the method of animation and the goals for which the animations were intended. Examples are researchers that used synthetic model-face as the output (Brand 1999, Cohen *et al.* 1996, Le Goff *et al.* 1994), those who used faces that are look-a-likes of real ones (Bregler *et al.* 1997), or both methods (Pandzic *et al.* 1999). Perceptual evaluation of animations motivated by their aesthetic values have shown modest results (Brand 1999, Bregler *et al.* 1997, Pandzic *et al.* 1999), though it is difficult to estimate and compare the results due to the subjective nature of the measures. The animations that were used as aid for the understanding of speech embedded in noise, achieved modest intelligibility improvements --- over speech without images --- of animated talkers (LeGoff *et al.* 1994, Pandzic *et al.* 1999).

Researchers who used the animations for lip-reading purposes (Cohen *et al.* 1996) reported good results although the results were not as good as with natural images. However, these evaluations did not compare the synthetic images, which are look-a-likes of the actual real images, with that of the recorded real images. Hence the results with the animated speech could be due to the animation or to the use of model-face or both.

The topic of visual speech recognition in relation to either individual differences, levels of hearing-impairment or accuracy of visual recognition (of whole-words, syllables and phonemes) is under discussion and is very well described in the introduction to the paper by Bernstein *et al.* (2000). In particular, the authors made a strong argument for phonetic perception (see also Berenstein, 2003) rather than viseme clusters (as in Cohen *et al.*1996 and Owens and Blazek 1985). This point is also strongly supported by the notion that "the moving vocal tract simultaneously shapes the acoustics and the motion of the face" (Vatikiotis-Bateson *et al.* 1996). --- Thus there is the necessity for scrutiny of visual phonetic articulation.

The focus of the present account is not in the animation methods themselves but rather in their perceptual evaluation. To the best of our knowledge, there is no systematic perceptual evaluation that uses comparisons between synthetic and real images of the same utterances spoken by the same person. Although the animations and their evaluations are strongly dependent on the intended purpose, it is desirable as a first step to achieve a point where an observer will be unable to tell if a talking image is real or synthetic.

Our goals were: 1) to create an animation that looks like the real recorded speaking face, 2) to verify that the observers perceive it as such, and will be unable to distinguish the synthetic from the real, 3) to assess whether "working" with the animated talking face yields visual communicative efficacy comparable to that of the recorded real, and 4) to establish a method with which one can evaluate perceptually how similar is the animation to the recorded real images of the talking person. For this purpose, we use the animation of Mary 101 (Ezzat *et al.* 2002) and evaluate it.

The Animation:

In preparation for the animation process, we recorded audio and video of Mary 101's face while she was uttering a corpus of speech containing single words and sentences. The video recording was digitized and preprocessed for animation. The preprocessing step included phonetically aligning the audio track to associate a phoneme for each image in the corpus, normalization for head movements and removal of

eye movements. The actual visual speech synthesis was made with two successive modules: the *multi-dimensional morphable model* (MMM) and the *trajectory synthesis* module. The MMM morphs between a small set of (46) prototype mouth images to synthesize new, previously unseen mouth configurations. The trajectory synthesis module synthesizes smooth trajectories in MMM space for each utterance. The parameters of the trajectory synthesis module were trained automatically from the recorded corpus using gradient descent learning. A detailed account of the animation process can be found in Ezzat *et al.* (2002).

For the purpose of this study, the training of the parameters of the trajectory synthesis was made only on single-word utterances in the corpus. Testing of the system was performed on single words and sentences not included in the training set. The synthesis was performed on the mouth region, which was then pasted back onto original face sequences with natural head and eye movement. This animation process allows reanimation of novel utterances that were not included in the original recording.

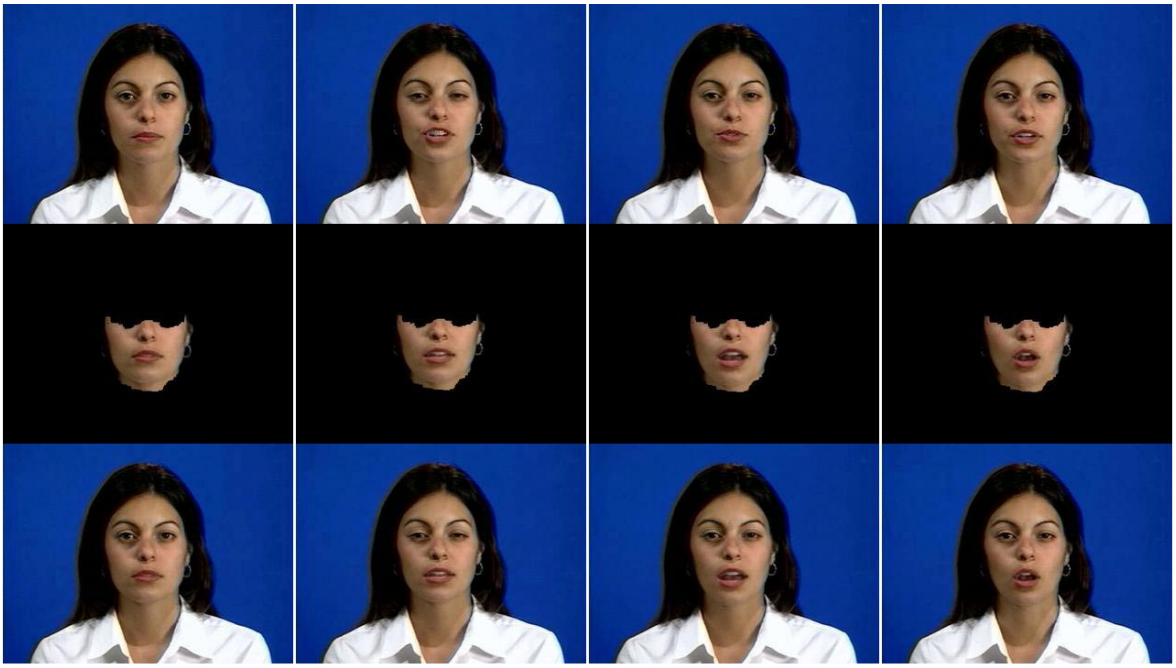

Figure 1: *Real and synthetic images of Mary 101. Top: A real image-sequence with natural head and eye movement. Middle: A sequence generated from our animation system, with the desired mouth movement and appropriate masking. Bottom: The final composited synthetic image-sequence with the desired mouth movement, but with the natural head and eye movements of the background sequence.*

The General Approach:

Once the visual animation of the utterances was made, its quality was evaluated by comparing directly the animated with the real image-sequences of the utterances. The advantage for the evaluation of this system is that the animations consisted of synthetic mouth region image-sequences, which were recomposed into the real video sequences of Mary 101. This results in identical synthetic and real image-sequences, of the same utterances, except for the mouth region (Fig. 1).

Our evaluation had two aspects: The first relates to how "real" is the animation by estimating a subject's ability to tell the animated images from the real ones. The second relates to how well one can operate

with the animated images as compared with the real ones in applications such as lip-reading. Hence, the evaluation of the animation was made in two modes:

a) Three experiments of visual detection, which can be considered as "Turing tests", were made to gauge if the synthetic image-sequences were recognized as such from the real. These we regard as explicit perceptual discrimination tasks. We named the experiments: single presentation, fast single presentation and pair presentation. In every experiment we presented each subject with the same number of synthetic and real image-sequences and observed the level of correct detection. The single and fast presentations were accompanied with the real audio input in both the synthetic and the real image-sequences. There was no audio input in the pair presentation.

b) An intelligibility experiment of lip-reading, which we regard as implicit perceptual discrimination task. In this experiment correct visual speech recognition of the utterances was recorded. The correct recognition from the real image-sequences was compared with that from the synthetic ones separately for whole words, syllables and phonemic recognition.

In all the experiments, the utterances were spoken in standard American English by Mary 101, who is a native speaker of American English. The utterances were either one-syllable words, two-syllable words, short or long sentences. The subjects participating in all the experiments had normal hearing.

General Method:

Stimuli: The image-sequences in all the experiments were run by a computer and presented on a 21" monitor (the detailed technical data are given in Appendix A.1). The actual display on the monitor was 15 cm x 11 cm (624 x 472 pixels) on which the frontal view of the head and shoulders of talking Mary 101 (Fig. 1) were presented, in color on a blue background. The distance of the viewers was set to 50 cm away from the screen. At this distance the display subtended 17 degrees x 12.5 degrees of visual arc for the viewer. The monitor's refresh rate was 60 Hz and the frame rate of the video image-sequences was 29.97 frames per second.

A real image-sequence of an utterance was taken directly from the digitized video images of Mary 101. The synthetic image-sequence of an utterance was made by cropping away the mouth region from the real image-sequence and inserting the synthetic image-sequence of that region instead (e.g. Fig.1). There were minor differences in average luminance and contrast of the mouth region between the real and the synthetic images. The average contrast difference was 4% (detailed description of the measurements are given in appendix A.2).

120 image-sequences were prepared as stimuli. They were made of 60 utterances. From each utterance two image-sequences were made, one real and one synthetic. A small number of image-sequences were randomly selected, from the entire corpus, for presentation in each experiment, with a different selection for each subject and each experiment (we use random in this writing as random within the specified category). The number of presented image-sequences varied in each experiment, as will be specified later. The order of the presented image-sequences was also randomized. However, in each experiment equal numbers of real and synthetic image-sequences were presented to every subject.

The 60 utterances comprised 40 single words and 20 sentences. Half of the single words were single-syllable words and half were two-syllable words. Average duration of the single words was about 2 seconds (range 2-3 seconds) and about 3 seconds for sentences. The corpus covered all the phonemes of the English language (a list of the utterances is given in appendix B).

In the single and the fast presentation experiments, the real audio was part of the stimuli. It was heard from two loudspeakers one at each side of the monitor. The average audio listening volume at the viewers' location was set to a comfortable listening level (55-65 dB SPL). The audio signals were well synchronized with the image-sequences, whether real or synthetic.

Participants: The participants were recruited by a circulated e-mail message to a large list of addressees. All those who responded and showed up for testing were taken as subjects. In all, we had a pool of 24 subjects from which they were assigned to the different experiments. The participants reported to have normal hearing and normal or corrected-to-normal vision. Their ages ranged from 18 to 67 years. They all had college education or were in college at the time of testing. Most were natives speakers of American English, as will be specified later for each experiment.

Procedure: The session began with an explanation of the general purpose of the experiments. The subjects were asked "to tell apart animated (synthetic) image-sequences from real ones" or in the intelligibility experiment to "tell what was said". The subjects were seated, in a dimly lit room, in front of a computer monitor, which they viewed from 50 cm distance. The display of the stimuli was controlled by a personal computer (PC), which was used also to record the subject's responses. The instructions to the subjects were different for each experiment, and will be specified accordingly. Before every stimulus presentation, an "X" on a blue background (similar to the background of Mary 101), was shown in the middle of the display. The subjects were instructed to fixate their gaze on the X whenever it appeared and when the image-sequence appeared to move the eyes freely. After a verbal prompt a stimulus was presented. The subjects were asked to respond according to the instructions. The responses were given orally and were entered into the computer by the experimenter. The options of responses will be described later for each experiment. However, in all the experiments there was also an option of "don't know (DK)" response in the instance that the subject could not make the difference between the options. Thus, the experiments were not two-way-forced-choice experiments. That gave the subjects the opportunity to render the synthetic and real image-sequences to be similar rather than force a "difference" where it was not appropriate. After the response was entered to the computer, the cycle was repeated until all the stimuli for the particular experiment had been presented.

After the conclusion of that experiment, the next experiment was performed in the same manner. At the end of the testing of all the subjects, the order of the experiments presented to each subject was arranged to give equal distribution of exposure; i.e., in each experiment there were equal numbers of subjects who were novices with no prior exposure to any of these experiments, close to equal numbers of subjects who had one prior exposure to one of the other experiments, and equal numbers of subjects who had participated in two of the other experiments. The intelligibility experiment was presented last to most subjects.

After all the experiments had been presented to the subject, the results were analyzed and shown (upon request) to the subject. The analyses of the results will be described for each experiment separately.

Distinguishing Visually between Real and Synthetic-Image Sequences:

EXPERIMENT 1: Single Presentations

In order to establish the level the synthetic image-sequences were distinguished from the real ones, we presented each image-sequence separately, and asked the subjects if it was a real or a synthetic image-sequence.

Method:
In addition to the general method described above the particular details for this experiment are described below.

Stimuli: For every subject, 16 image-sequences were randomly chosen from the corpus. Half of these were real and half synthetic, not necessarily paired to be from the same utterances. In addition, half of the image-sequences were of single-word utterances and half of sentences, evenly distributed across real and synthetic image-sequences. Due to the random process of selecting the image-sequences for presentation, 7 image-sequences (5 real and 2 synthetic) from the corpus were not presented to any of

the subjects in this experiment. Every image-sequence was accompanied by the real audio of the utterance.

Participants: There were 22 subjects (11 females and 11 males), of whom 19 are native speakers of American English. For eight of the subjects this experiment was the first in the row of experiments. Another group of eight subjects participated before in one of the other experiments and six more subjects participated in two experiments prior to this one.

Procedure: The instructions to the subjects were as follows: "You will be presented with an image-sequence of a talking woman, she will say either one word or a sentence. Please tell if the image-sequence you saw was real or synthetic." It was also explained that all the image-sequences would be accompanied by the real audio recording of the utterances. As mentioned, the subjects were asked to fix their gaze on the X. After a verbal prompt, a single image-sequence of an utterance was displayed followed by the reappearance of the X. The subjects were asked to tell if the image-sequence was real, synthetic or, if unable to decide, to say, "don't know". The subjects took their time to respond. The response was entered into the computer by the experimenter, clicking with the mouse on the appropriate field shown in the corner of the monitor. The next cycle followed. There was no mention of the number of image-sequences to be presented or the ratio of real to synthetic image-sequences (although it is reasonable to assume that the subjects thought it might be half of each). At the end of presenting all the 16 stimuli, a new experiment was prepared and the file with the collected responses was kept for evaluation.

Results and Discussion:

Main results: The average of correctly identifying the image-sequences to be either real or synthetic (for all the 22 subjects on all the 16 image-sequences presented to each) was 54.26%. As indicated in Table 1, that number is not significantly different from chance level (50%), as was verified by a t-test ($p<0.3$). The results were similar when utterances of single words or sentences were considered separately. (There were 3 subjects who correctly identified the image-sequences at or above 75%; none were below 25%.).

From the real image-sequences presented 74.43% were detected as such, compared with 34.09% of the synthetic image-sequences, which means that 65.91% of the synthetic image-sequences were detected incorrectly. A summary of the results is given in Table 1.

Table 1: Responses to Single Presentations

Means, Standard Deviations, and Significance from Chance

Prior exposure				None			One			Two		
n	All 22 subjects			8			8			6		
	Mean	SD	p<	Mean	SD	p<	Mean	SD	p<	Mean	SD	p<
All utterances												
% correct	54.26	15.72	0.3	52.34	12.91	0.7	55.47	18.43	0.5	55.21	17.86	0.6
% of DK responses	12.50	16.70		10.16	14.92		11.72	20.98		16.67	14.61	
Single words												
% correct	52.84	16.78	0.5	50.00	11.57	1	53.13	18.60	0.7	56.25	22.01	0.6
% of DK responses	12.50	19.29		7.81	13.26		12.50	22.16		18.75	23.39	
Sentences												
% correct	55.68	21.73	0.3	54.69	16.28	0.5	57.81	21.06	0.4	54.17	31.29	0.8
% of DK responses	12.50	17.25		12.50	17.68		10.94	20.53		14.58	14.61	

These results suggest that when image-sequences are presented to the subjects, on average, they are unable to tell whether the presented image-sequence is synthetic or real.

Additional details: On average, 12.5% of the responses were DK, where the subjects could not decide if the image-sequence was either real or synthetic. These responses of "unable to tell" were considered as not recognized to be positively synthetic or real. Hence, we did not count them as correct responses. This can be justified with relation to the aim of the experiment, which is to gauge the level of positively detecting the animated from the real, and the DK responses do not tell them apart.

In addition, 8 of the subjects had no prior exposure to the image-sequences, 8 had one similar experiment, which they performed prior to this one, and 6 subjects had two experiments prior to this one. The average correct identification for each group was 52.34%, 55.47% and 55.21%, respectively. None were significantly different from chance level, and there was no significant difference between the groups. As there were no significant advantages (or disadvantages) to prior exposure to the other experiments, we regarded the 22 subjects as one group.

With regard to individual utterances among multiple presentations of synthetic image-sequences, three single words and one sentence were recognized as synthetic in all cases. This suggested that most synthetic image-sequences were evenly distributed with regard to the ability to recognize them as such. This experiment measures best the impression one gets from the presented image-sequence, whether it is real or animated.

EXPERIMENT 2: Fast Single Presentations

The previous experiment forced subjects to treat each image-sequence on its own. Perhaps the differences between real and synthetic image-sequences will be more evident if subjects could compare one against the other in rapid succession. This is the motivation for the second experiment.

Method: This experiment is basically the same as the previous one. The difference is that in this experiment the image-sequences followed each other with only one-second intervals between them. Subjects were asked to respond during the presentation of the image-sequence and the following interval, thus requiring fast responses. The other differences are detailed below.

Stimuli: Eighteen image-sequences were randomly chosen from the corpus for each subject, half of which were real and half synthetic, and not necessarily paired to be from the same utterances. Eight image-sequences were of single-word utterances and 10 of sentences, in equal numbers of real and synthetic. As in Experiment 1, the real audio accompanied all the image-sequences.

Participants: Twenty-one subjects participated in this experiment (11 females and 10 males), 18 were native speakers of American English. For 8 subjects, this was the first encounter with such an experiment. Seven participated in one similar experiment before and 6 participated in two previous experiments.

Procedure: The procedure was similar to that in Experiment 1. Instead of presenting each image-sequence and pausing for the subject's response after each presentation, here the image-sequences were presented in blocks of six. In each block the image-sequences were presented one after the other with one-second intervals between them. During the intervals the X was displayed. Before the beginning of the presentation the subject was instructed to respond during the presentation of the image-sequence and immediately following interval, before the presentation of the next image-sequence started. The subject's response choices were either real, synthetic, or don't know. The experimenter entered them into the computer immediately after the response was made by clicking the mouse on the appropriate field, as in Experiment 1.

Results and discussion: The average correct identification of the 21 subjects was 52.12%, which is not significantly different ($p<0.5$) from chance level (50%). (There were no subjects with correct identification above 75% or below 25%.) As in Experiment 1, correct identification of the real image-sequences (66.67%) was higher than that of the synthetic ones (37.57%).

These results suggest that even when the presentations of the image-sequences follow one another rapidly, the subjects, on average, are unable to distinguish the synthetic from the real image-sequence.

Table 2: Responses to Fast Single Presentations

Means, Standard Deviations, and Significance from Chance

Prior exposure				None			One			Two		
n	All 21 subjects			8			7			6		
	Mean	SD	p<	Mean	SD	p<	Mean	SD	p<	Mean	SD	p<
All utterances												
% correct	52.12	15.66	0.5	54.17	12.86	0.4	44.44	20.03	0.60	58.33	11.52	0.2
% of DK responses	2.65			1.39			6.35			0.00		
Single words												
% correct	50.00	14.79	1	56.25	9.45	0.2	42.86	17.47	0.40	50.00	15.81	1
% of DK responses	1.19			0.00			3.57			0.00		
Sentences												
% correct	53.81	20.61	0.2	52.50	19.09	0.8	45.71	24.40	0.70	65.00	15.17	0.1
% of DK responses	4.76			3.13			10.71			0.00		

As seen in Table 2, the average correct identification between real and synthetic image-sequences is 50% for single word utterances and 53.81% for sentences. Neither is significantly different from chance level ($p<1$; $p<0.2$ respectively), and not significantly different from each other ($p<0.2$). One possible explanation for this slight difference is that the subjects had longer response time when sentences were presented than when single words were presented. There were, on average, only 2.6% of the cases where the subjects did not know (DK responses). As before, this was considered as an incorrect response. This low DK response level could suggest that fast presentation makes the experiment more similar to a two-way forced-choice experiment. As in Experiment 1, the prior exposure to similar experiments did not affect performance systematically or significantly.

EXPERIMENT 3: Pair Presentations

In the previous experiments each image-sequence was randomly selected, for presentation to the subject, from the entire corpus. This made direct comparisons of real and synthetic image-sequences of the same utterances impossible. It was possible that presenting the synthetic and real image-sequences, of the same utterances, one immediately after the other would have made distinguishing between them easier. The following experiment address that concern.

Method: Pairs of real and synthetic image-sequences of the same utterances were presented as stimuli, one immediately after the other in a randomized order. The subject's task was to tell the order of the presented real and synthetic image sequences. Otherwise this experiment is similar to the previous ones.

Stimuli: Sixteen utterances were randomly chosen from the corpus. This gave 32 image-sequences, one real and one synthetic from each utterance. Half of the utterances were single words and half sentences. The order of presentation within each pair was also randomly set. There was no audio input in this experiment.

Participants: Participants in this experiment were 22 subjects (10 females and 12 males) of whom 19 were native speakers of American English. For 8 subjects this was the first presentation of such an experiment. Another 8 subjects participated beforehand in one similar experiment and another 6 subjects participated before in two similar experiments.

Procedure: The instructions to the subject were: "You will be presented with two image-sequences, of the same utterance, one after the other with an interval of 1 second between the end of the first and the beginning of the second. One image-sequence is real and the other synthetic appearing in random order. Please tell the order of the presentation: real-synthetic, synthetic-real, or don't know (DK). Take your time to answer it."

In the interval between the image-sequences the X reappeared and the subject was asked to fixate it. The subject's response was given verbally and entered to the computer, by the experimenter, as in the previous experiments.

Results and Discussion:

Main results: The average score of correctly identifying the order of the real and the synthetic image-sequences, for the 22 subjects, was 46.59%, which is not significantly different ($p<0.5$) from chance level (50%). The correct identification was similar for single word and sentence utterances (45.45% and 47.73% respectively). (There were 3 subjects with above 75% correct identification and one below 25%.) These results suggest that even when pairs of real and synthetic image-sequences of the same utterances were presented directly one after the other the subjects were unable to tell the synthetic from the real.

Table 3: Responses to Pair Presentations

Means, Standard Deviations, and Significance from Chance

Prior exposure				None			One			Two		
n	All 22 subjects			8			8			6		
	Mean	SD	p<	Mean	SD	p<	Mean	SD	p<	Mean	SD	p<
All utterances												
% correct	46.59	21.28	0.5	39.84	21.38	0.3	46.09	20.58	0.70	56.25	22.01	0.6
% of DK responses	28.69			36.72			32.81			12.50		
Single words												
% correct	45.45	27.43	0.5	37.50	21.13	0.2	45.31	27.50	0.30	56.25	35.13	0.7
% of DK responses	28.41			42.19			28.13			10.42		
Sentences												
% correct	47.73	21.70	0.7	42.19	26.67	0.5	46.88	17.36	0.60	56.25	20.54	0.6
% of DK responses	28.98			31.25			37.50			14.58		

Additional results: From all the responses the average responses of DK was 28.68%. As mentioned before, in these cases the subjects were unable to tell the synthetic from the real image-sequences. Hence they were counted as incorrect responses.

From the 22 subjects 8 had no prior exposure to the image-sequences, 8 subjects participated in one of the experiments above prior to the current one and 6 subjects in the two experiments above. There was a slight, not significant, increase of correct response with exposure. The novices scored 39.84% correct, the subject with one prior exposure scored 46.09% and the subjects who participated in two previous experiments scored 56.25% correct. Cross t-tests for all possible combination were not significantly different with $p<0.2$ for the smallest value.

On Experiments 1-3, which can be thought of as "Turing tests" (or variations of it), the subjects, on average, were unable to visually distinguish between the animated and the real image-sequences.

Comparing Visual Speech Perception of Real and Synthetic Image Sequences:

EXPERIMENT 4: Lip-Reading

The previous experiments demonstrated that the synthetic image-sequences could not be distinguished visually from the real ones. However, it remains to be seen if performing a visual-speech recognition task with the real and synthetic image-sequences could discern the differences between them.

Method: The main difference between this experiment and the ones above was the task. In this experiment the subjects were to tell what Mary 101 said reading her lips rather than judging if the image-sequences were real or animated. There was no audio input and there were no top-down psycholinguistic cues given before presenting each utterance, although in the utterances of sentences the internal structure of the sentences provided some of these cues.

Stimuli: Each subject was presented with the real and the synthetic image-sequences of the 16 utterances, which were randomly chosen from the entire corpus. Eleven utterances were single words and five were sentences. Each utterance was presented twice to each subject, once as a real image-sequence and once as a synthetic. The image-sequences were presented to each subject in random order to avoid direct priming by the same utterances. On average, the same number of either real or synthetic image-sequences was presented first in an utterance.

Participants: Eighteen subjects (8 females and 10 males) participated in this experiment, of whom 16 were native speakers of American English. All had participated in one or more of the experiments above. All were proficient in English.

Procedure: The subject was instructed to "read the lips" of Mary 101 and at the end of the stimulus presentation tell verbally what the utterance, or any part of it, was. As in the previous experiments, at first the subject was asked to fixate on the X and move the eyes freely once the image-sequence appeared. After a verbal warning "ready" by the experimenter the image-sequence was shown. At the end of the image-sequence the X reappeared. After the subject's verbal response the experimenter entered the response to the computer by using its keyboard. Hereafter a new cycle started. The subject was encouraged to respond to the whole utterance, however a response to any part of it was also considered valid. When the subject had no clue the appropriate sign was entered. The evaluations of the correct responses were made separately on three different levels: the number of correct responses to whole words, to syllables and to phonemes. Once all the subjects were tested the average responses were calculated.

Results and discussion: The 18 subjects were presented with 32 image-sequences each. Due to the different lengths of the sentences and words and due to the random selection of the utterances to be presented to each subject, the total amount of words, syllables and phonemes was different for each subject. However, as seen from Table 4, the average number of words presented to each subject was 65.56, which comprised 85.9 syllables or 224.6 phonemes. The subjects responded, on average, to 81.42% of all the image-sequences (real and synthetic) presented. Most of the responses were incorrect. The high rate of responses indicated familiarity with the mode of Mary 101's speech, whether presented as real or synthetic image-sequences.

Average correct recognition of whole-words from the real and synthetic image-sequences together was 10.74%, that of syllables 12.4% and 25.6% of the phonemes, all significantly different from 0 (the significance level is not shown in the table). That level is similar to the range of correct phoneme recognition reported in the literature (Bernstein *et al*. 2000) but it is lower for words presented in sentences. The individual differences were in the range 1.56% to 29.17% for whole-words and 7.14% to 38.24% for phonemes.

Table 4: Visual Speech Recognition

Correct Recognition of Whole Words, Syllables, and Phonemes

	# of	average # of presented			% of	% correct response to		
18 subjects	**stimuli**	**words**	**syllables**	**phonemes**	**responses**	**words**	**syllables**	**phonemes**
All utterances								
synthetic and real	32	65.56	85.89	224.56	81.42	10.74	12.40	25.60
synthetic only (S)					72.92	6.96	8.52	21.19
real only (R)					89.93	14.52	16.29	30.01
t (difference S-R)						-3.595	-4.004	-5.102
p<						0.01	0.001	0.001
Single words								
synthetic and real	22	22	32.93	90.93	83.59	10.35	14.42	33.31
synthetic only (S)					76.26	6.06	9.45	28.07
real only (R)					90.91	14.65	19.40	38.55
t (difference S-R)						-3.308	-3.989	-4.196
p<						0.01	0.001	0.001
Sentences								
synthetic and real	10	44.27	53.60	135.60	76.67	11.00	11.17	20.45
synthetic only (S)					65.56	7.49	8.04	16.52
real only (R)					87.78	14.51	14.30	24.38
t (difference S-R)						-2.278	-2.496	-3.132
p<						0.05	0.05	0.01

The differential responses to real and synthetic image-sequences are that which interests us most in this account. For all the utterances presented, the subjects responded on average to 72.92% of the synthetic image-sequences as compared with 89.93% to the real, most of the responses were incorrect in both cases. That suggests higher familiarity with the real image-sequences than with the synthetic.

As seen in Table 4, on average the correct recognition of words, syllables and phonemes were significantly higher when real image-sequences were presented than when synthetic ones were. This holds true also when single-words or sentences were considered separately. This indicates that the animated image-sequences were not as efficient for lip-reading as the real recording.

The individual differences of correct responses for real image-sequences were in the range of 0% to 34.5% for words and 11.11% to 44.45% for phonemes, and those for synthetic image-sequences were 0 to 27.78% for words and 3.17% to 37.6% for phonemes. Although most subjects recognized visual-speech better in the real image-sequences, three subjects were better with the synthetic ones at all levels of word, syllables and phonemes.

The ratio of correct responses of phonemes to whole words is higher for the synthetic image-sequences (21.19/6.96=3.04) than for the real ones (30.01/14.52=2.07). That holds also for single-words utterances alone (28.07/6.06=4.63; vs. 38.55/14.65=2.64) as well as for utterances of sentences (16.52/7.49=2.21; vs. 24.38/14.51=1.68). That might suggest that the dynamics of combining phonemes to words is compromised more in the animation process than the appearance of the phonemes. However, that could not be the only factor, otherwise phoneme recognition in the real and synthetic image-sequences would have been the same.

In addition, phoneme recognition in single word presentation was significantly higher ($p<0.001$, not shown in the table), than that in sentence presentation. However, word recognition was similar in both presentations. These relations are similar for both, synthetic and real image-sequences when considered

separately. Whatever the interpretation, a bottom up – constructing the seen words from the phonemic parts – or a direct resemblance of the seen words, the dynamics of the synthetic and the real image-sequences is similar.

In most cases when the participants tried to figure out what was uttered, they moved their own lips trying to mimic the mouth movements that they saw on the screen without uttering any sound. Also, when the participants were casually asked if they could tell if there was any difference between the image-sequences, or if they were real or synthetic, they could not. In most cases they thought all the image-sequences were real.

General Discussion:

The way Mary 101 speech was recorded and animated has given us the opportunity to compare the perception of real and synthetic image-sequences of the same utterances that were not used to train the system for the animation. Also, as only the mouth region was animated, the rest of the talking face was real. In this way only the animated part was different from the real and not the whole face, leaving eye-movements and facial expressions identical in both real and synthetic image-sequences of an utterance. The advantage is that the comparison is of visual speech alone and of no other elements of the images. This aspect is different from previous evaluations of animated speech, where either model-faces or comparison of other images were made. The other advantage is that our experimental method permitted versions of Turing tests of visual-speech animation.

Some of the subjects, after performing Experiments 1 to 3 in the manner described above, were asked to repeat the experiments focusing their attention to the mouth region only. The individual results were very similar in both cases, suggesting that focusing the attention on the mouth region did not make it easier to distinguish between synthetic and real image-sequences.

The first three experiments (Experiments 1-3), each a version of the Turing test, were designed to gauge the level of recognizing the animated from the real image-sequences of the utterances. Each experiment showed that, on average, the participants could not tell the synthetic from the real. That is, in all the three experiments the animated looked as natural as the real. These results held whether the image-sequences were presented singly with time for scrutiny (experiment 1), or presented singly with little time to respond (Experiment 2), or presented in pairs for comparison of the real and the synthetic image-sequence of the same utterance (Experiment 3), suggesting that the synthetic appeared as real to the viewer.

We refrained from using a simultaneous side-by-side presentation of synthetic and real image-sequences of the same utterances. As we understood it, the subjects would have shifted their gaze from one image to the next while the utterance was progressing. As a result, the subjects could have compared local features but not the impression the moving mouth would have given throughout the utterances. In addition, it appears that adding the real audio to the stimuli (Experiments 1 and 2) did not enhance the visual detection of the synthetic from the real, as the detection levels without audio input (Experiment 3) were not significantly different from the former ones.

This evaluation pertains only to the quality of the animation as related to the similarity of the synthetic and the real image-sequences. It does not address notions like appeal, liking of the image and other attributes, which may be associated with the animated images. In the explicit perceptual discrimination task (Experiments 1-3), the viewers could not distinguish the synthetic image-sequences from the real ones, when asked to do so directly. In this sense, the animation of Mary 101 achieved the goal of passing a Turing test. However, the same viewers in an implicit perceptual discrimination task, i.e. the intelligibility Experiment 4, had significantly higher speech recognition of the real image-sequences than of the synthetic ones (while they could not tell if the images were synthetic or real). This suggests that the implicit method, using the intelligibility task, is a more sensitive discrimination tool that the explicit method using a direct discrimination task. This suggests also that the animation of Mary 101, as it stands now, is not as good as we would like it to be especially for the purpose of rehabilitation and language

learning. We are currently concentrating on the analysis of individual phonemic recognition and the dynamics of phoneme transition, to improve the animation in order to achieve the goal of equal intelligibility performance with real and synthetic image-sequences.

Acknowledgments:

Many thanks are due to Luis Pérez-Breva, Pawan Sinha, Sayan Mukherjee, Mary Pat Fitzgerald, and the anonymous participating subjects who helped us in accomplishing this account.

References:

Brand, M. (1999). "Voice puppetry." In: Proc. SIGGRAPH 1999, ACM Press/ACM SIGGRAPH, Los Angeles, 21-28.

Bregler,C., Covell, M. and Slaney, M. (1997). "VideoRewrite: Driving visual speech with audio." In: Proc. SIGGRAPH 1997, ACM Press / ACM SIGGRAPH, Los Angeles, 353-360.

Bernstein, L.E., Demorest, M.E. and Tucker, P.E. (2000). "Speech perception without hearing," *Perception & Psychophysics,* Vol. 62(2), 233-252.

Berenstein, L. (2003). "Visual speech perception." In: Audiovisual Speech Processing, E. Vatiokis-Bateson, G. Bailly & P. Perrier (Eds.), to appear.

Brooke, N. and Scott, S (1994). "Computer graphics animation of talking faces based on stochastic models." In: Int'l. Symposium on Speech, Image Processing and Neural Networks, Hong Kong, 73-76.

Cohen, M.M. and Massaro, D.W. (1993). "Modeling co-articulation in synthetic visual speech." In: Models and Techniques in Computer Animation, N.M. Thalmann and D. Thalmann (Eds.), Tokyo: Springer Verlag, 139-156.

Cohen, M.M., Walker, R.L. and Massaro, D.W. (1996). "Perception of synthetic visual speech." In: Speech reading by humans and Machines, D.G. Stroke and M.E. Hennecke (Eds.), New York: Springer 153-168.

Cosatto, E. and Graf, H. (1998). "Sample-base synthesis of photorealistic talking heads." In: Proceedings of Computer Animation '98, 103-110.

Ezzat, T., Geiger, G. and Poggio, T. (2002). "Trainable videorealistic speech animation." In: Proceedings of SIGGRAPH 2002, ACM Press / ACM SIGGRAPH.

Lee, Y., Terzopoulos, D.and Waters, K. (1995). "Realistic modeling for facial animation." In: Proc. SIGGRAPH 1995, ACM Press / ACM SIGGRAPH, Los Angeles, 55-62.

Le Goff, B. and Benoit, C. (1996). "A text-to-audiovisual-speech synthesizer for French." In*:* Proceeding of the International Conference on Spoken Language Processing (ICSLP), 55-62.

Le Goff, B., Guiard-Marigny, T., Cohen, M. and Benoit, C. (1994). "Real-time analysis-synthesis and intelligibility of talking faces." In*:* 2nd International Conference on Speech Synthesis*,* New Paltz, NY, 53-56.

Masuko, T., Kobayashi, T., Tamura, M., Masubuchi, J. and Tokuda, K. (1998). "Text-to-visual speech synthesis based on parameter generation from hmm." In: *ICASSP*, 3745-3748.

Owens, E. and Blazek, B. (1985). "Visemes observed by hearing-impaired and normal-hearing adult viewers," *J. Speech and Hearing Res., Vol. 28, 381-393.*

Pandzic, I.S., Ostermann, J. and Millen, D. (1999). "User evaluation: Synthetic talking faces for interactive services," *The Visual Computer, 15, 330-340.*

Parke, F.I. (1974). "*A Parametric Model of Human Faces." Ph.D. thesis, University of Utah.*

Pearce, A., Wyvill, G. and Hill, D. (1986). "Speech and expression: A computer solution to face animation." In: *Graphics Interface*, 136-140.

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. C., & Terzopoulos, D. (1996). "The dynamics of audiovisual behavior in speech." In: Speechreading by humans and machines (NATO-ASI Series F), D. Stork & M. Hennecke (Eds.), Berlin: Springer-Verlag, 150, 221-232.

Vatikiotis-Bateson, E., Kroos, C., Kuratate1, T., Munhall, K. G., Rubin, P., & Yehia, H. C. (2000). "Building talking heads: Production based synthesis of audiovisual speech." In: Humanoids 2000 First IEEE-RAS International Conference on Humanoid Robots, Cambridge, MA: IEEE-RAS.

Waters, K. (1987). "A muscle model for animating three-dimensional facial expression." In: Computer Graphics (Proceeding of ACM SIGGRAPH 87), Vol. 21(4) ACM, 17-24.

Appendix A:

1. Technical data on the display method

The image-sequences in all the experiments were run by a Dell Dimension 8100 computer system with an Intel Pentium 4 processor at 1.4 GHz. running Windows 2000, it had 512 Mbytes of RAM, and 80 GB hard drive. The images were presented on a 21 inch Dell P1110 Trinitron monitor with screen resolution of 1600 by 1200 pixels, with a 32 Mbyte DDR Nvidia Geforce 2 GTS video card. The actual display of the images, on the monitor, was about 15 cm x 11 cm, with matched resolution to that of the monitor. The monitor's refresh rate was 60 Hz and the frame rate of the video image-sequences was 29.97 frames per second. All the video sequences were encoded in Quicktime 5 format, using Sorenson codec with the image quality setting set to highest level. The encoding was done using the Adobe Premier 6.0 program by importing the synthesized image frames and audio, and exporting the movie sequences. The final sequences were displayed using the Quicktime 5 player.

The audio was provided through a Soundblaster Live Digital sound card. The speakers were Altec Lancing ACS-340 speakers with subwoofer.

2. Contrast differences between the real and synthetic image-sequences.

From the entire corpus 8 utterances were chosen as representatives. In each of the 16 image-sequences of these utterances, 6 patches of 10 X 10 pixels were selected. They were in the same locations for all the image-sequences. Four patches were around the mouth of Mary 101, one on the upper lip and one on the lower lip. The average grey-level was noted for each patch separately every 200 ms, for the duration of the sequence.

In order to calculate the general differences in grey-levels between real and synthetic-image sequences we compared the grey-levels of patches of the same location in the real and in the synthetic image-sequences at each time. The largest average difference for the same patch location and for the same time was 6.4% of the total grey-level values. Most were higher for the synthetic images. From these values we were able to calculate the (Maxwell) contrast between the region around the mouth and the lips, which was averaged over the duration of the each image-sequences. The average contrast was about 20%. The largest average contrast difference between the real and synthetic image-sequence was 4%, mostly higher for the synthetic image-sequences.

Appendix B: A list of the utterances used in this study.

air	amaze	I'm not happy with them
all	assume	Others looked sad
badge	bracket	We'll come back to you later
bounce	caress	He opposes the Americans
check	dagger	The meeting was frank
dam	endorse	Have a good evening
dwarf	gateway	You can't stop science
flesh	iceberg	Eat the fish later
growth	landmark	Thank you Diane
jam	mammal	53 percent
lack	motive	What can he do
lounge	oilfield	Ken is at the courthouse
name	pillow	Not at all Tom
pain	saddle	More news in a moment
risk	tabloid	Its a matter of money
safe	teacher	That's all for tonight
shell	Thursday	It could take months
tribe	vision	His name is Morgan
wade	weather	Thank you very much
yes	zebra	We had a good meeting