Perception of Synthetic Visual Speech

Michael M. Cohen, Rachel L. Walker, and Dominic W. Massaro

University of California—Santa Cruz

Abstract. We report here on an experiment comparing visual recognition of monosyllabic words produced either by our computer-animated talker or a human talker. Recognition of the synthetic talker is reasonably close to that of the human talker, but a significant distance remains to be covered and we discuss improvements to the synthetic phoneme specifications. In an additional experiment using the same paradigm, we compare perception of our animated talker with a similarly generated point-light display, finding significantly worse performance for the latter for a number of viseme classes. We conclude with some ideas for future progress and briefly describe our new animated tongue.

Keywords. Visible speech synthesis, coarticulation, speechreading, point-light displays, text-to-speech

1 Introduction

Much of what we know about speech perception has come from experimental studies using synthetic speech. Although some research questions can be answered in part with natural speech stimuli, our overall progress in analyzing human speech perception has been critically dependent on the use of synthetic speech. Extending this approach to the visual side of speech, we have developed a high quality visual speech synthesizer—a computer-animated talking face—incorporating coarticulation based on a model of speech production using rules describing the relative dominance of speech segments (Cohen & Massaro, 1993). Our goals for this technology include gaining an understanding of the visual information that is used in speechreading, how this information is combined with auditory information, how such information may be used in automatic speech recognition (ASR) systems, and its use as an improved channel for man/machine communication. An essential component of the development process is an evaluation of the synthesis quality. This analysis of the facial synthesis may be seen as a validation process. By validation, we mean a measure of the degree to which our synthetic faces mimic the behavior of real faces. Confusion matrices and standard tests of intelligibility are being utilized to assess the quality of the facial synthesis relative to the natural face. These same results will also highlight those characteristics of the talking face that could be made more informative.

2 Visual Speech Synthesis Techniques

Two general strategies for generating highly realistic full facial displays have been employed: musculoskeletal models and parametrically controlled polygon topology. Using the first basic strategy, human faces have been made by constructing a computational model for the muscle and bone structures of the face (e.g. Platt & Badler, 1981; Waters, 1987; Waters & Terzopoulous, 1991). At the foundation of this type of model is an approximation of the skull and jaw including the jaw pivot. Simulated muscle tissues and their insertions are placed over the skull. This requires complex elastic models for the compressible tissues. A covering surface layer changes according to the underlying structures. The dynamic information for such a model is defined by a set of contraction-relaxation muscle commands. Platt and Badler (1981) use Ekman and Friesen's (1977) "Facial Action Coding System" to control the facial model. These codes are based on about 50 facial actions (action units or AU's) defined by combinations of facial muscle actions.

Using the second basic strategy, Parke (1974, 1975, 1982, 1991) modeled the facial surface as a polyhedral object composed of about 900 small surfaces arranged in 3D, joined together at the edges and smooth shaded. The face was animated by altering the location of various points in the grid under the control of 50 parameters, about 10 of which were used for speech animation, such as jaw rotation, mouth width, lip protrusion, and lower lip "f" tuck. Parke (1974) selected and refined the control parameters used for several demonstration sentences by studying his own articulation frame by frame and estimating the control parameter values.

One advantage of the polygon topology strategy is that calculations of the changing surface shapes in the polygon models can be carried out much faster than those for the muscle and tissue simulations. It also may be easier to achieve the desired facial shapes directly rather than in terms of the constituent muscle actions. This difference in synthesis methods is parallel to the difference between articulatory (e.g. Flanagan, Ishizaka, & Shipley, 1975) and terminal-analogue formant (Klatt, 1980) synthesizers for auditory speech. Auditory articulatory synthesizers require several orders more computation than do terminal-analogue ones.

Our current software (Cohen & Massaro, 1993, 1994) is a descendant of the Parke software, incorporating additional and modified control parameters, a tongue (which was lacking in Parke's model), and a new visual speech synthesis control strategy. Consisting of about 20,000 lines of C code, the visual synthesis program runs in real-time on an SGI Crimson-Reality Engine. An important improvement in our visual speech synthesis software has been the development of an new algorithm for articulator control which takes coarticulation into account. Our approach to the synthesis of coarticulated speech is based on an articulatory gesture model described by Löfqvist (1990). In this model, a speech segment has dominance over the vocal articulators which increases and then decreases over time during articulation. Adjacent segments have overlapping dominance functions which leads to a blending over time of the articulatory commands related to these segments. We have instantiated this model in our synthesis algorithm using negative exponential functions for dominance. Given that articulation of a segment is implemented by many articulators, there is a separate dominance function for each control parameter. Each phoneme is specified in terms of the speech control parameter target values and the dominance function characteristics of magnitude, time offset, and leading attack and trailing decay rates.

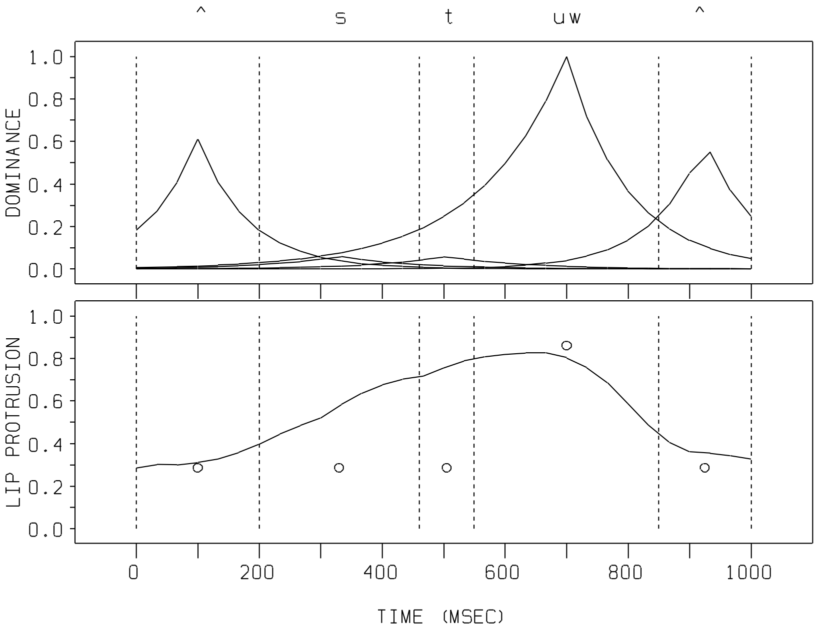

Fig. 1. Dominance functions (top panel) for the speech segments and parameter targets (circles) and resulting parameter control function (bottom panel) for lip protrusion for the word "stew" using the Cohen and Massaro (1993) speech synthesis algorithm.

An example of the system's operation is shown in the top panel of Figure 1, illustrating the lip protrusion dominance functions for the word "stew". As can be seen, the /s/ and /t/ segments have very low dominance with respect to lip protrusion compared to /u/. Also the dominance of /u/ extends far forward in time. The lower panel gives the resulting lip protrusion trace. One can see how the lip protrusion extends forward in time from the vowel. Note that the figure only illustrates the dynamics for lip protrusion. Other control parameters, e.g. tongue angle, for /t/ and /u/ have equal dominance. For /t/, this scheme allows the tongue tip to reach the proper location that would correspond to being against the alveolar ridge, while for /u/ the tongue tip will lie behind the lower teeth. As part of our higher level control strategy we have integrated a number of text-to-speech modules including MITalk (Allen, Hunnicutt & Klatt, 1987) and the AT&T TTS system. These provide the segments, durations, and suprasegmental information to the visual synthesis algorithms and also provide the auditory speech, which can be played in synchrony with the visual speech.

3 Experiment 1: Natural versus Synthetic Speech

One of the goals of our synthesis is to have a talking head that articulates as clearly (or even more clearly) than human talkers. We want our talking head to be easy to speechread. One test of speaking quality is to compare the recognition performance of our synthetic talker with a natural human talker. In Experiment 1, we presented silently for identification monosyllabic English words (e.g. *sing, bin, dung, dip, seethe*) produced either by a natural speaker (Bernstein & Eberhardt, 1986) or our synthetic talker randomly intermixed. The synthetic stimuli used a specific set of parameter values and dominance functions for each phoneme and our scheme for coarticulation. The MITalk text-to-speech module was utilized to give the phonetic representation for each word and the relative durations of the speech segments.. Other characteristics such as speaking rate and average acoustic amplitude were equated for the two talkers. By comparing the overall proportion correct and analyzing the perceptual confusions, we can determine how closely the synthetic visual speech matches the natural visual speech. Because of the data-limited property of visible speech, we group the consonants into viseme categories, based on the work of Walden et al. (1977) and Massaro et. al. (1993). In addition, because of the difficulty of speechreading, we also expect confusions even between viseme categories for both the natural and synthetic visual speech. The questions to be answered are what is the amount of confusions and how similar are the patterns of confusion for the two talkers.

3.1 Method

Twelve college students who were native American English speakers served as subjects, in two 40 minute sessions each day for two days. Up to four at a time were tested in individual sound attenuated rooms under control of the SGI-Crimson computer, with video from the laserdisk (the human talker) or the computer being presented over 13" color monitors. On each trial they were first presented with a silent word from one of the two faces and then typed in their answer on a terminal keyboard. Only actual monosyllabic English words were accepted as valid answers from a list of about 12,000 derived mainly from the Oxford English dictionary. After all subjects had responded, they received feedback by a second presentation of the word, this time with auditory speech (natural or synthetic) and in written form on the left side of the video monitor.

There were 264 test words, and each word was tested with both synthetic and natural speech, for a total of 2 times 264 = 528 test trials. For the counterbalancing of the test words and presentation modes, the subjects were split into two groups. Each group received the same random order of words but with the assignment of the two faces reversed. Five unscored practice trials using additional words preceded each experimental session of 132 test words.

3.2 Analysis

A number of analyses were automatically carried out on the subjects' responses. These included the derivation of confusion matrices for initial and final consonants, initial and final consonant visemes, vowels, consonant clusters, and the proportion of correct consonants, consonant visemes, vowels, and words. There is insufficient space here to provide all of these results—these will be described in another paper. This report includes a representative sample of the analyses, which are illustrative of the general pattern of results.

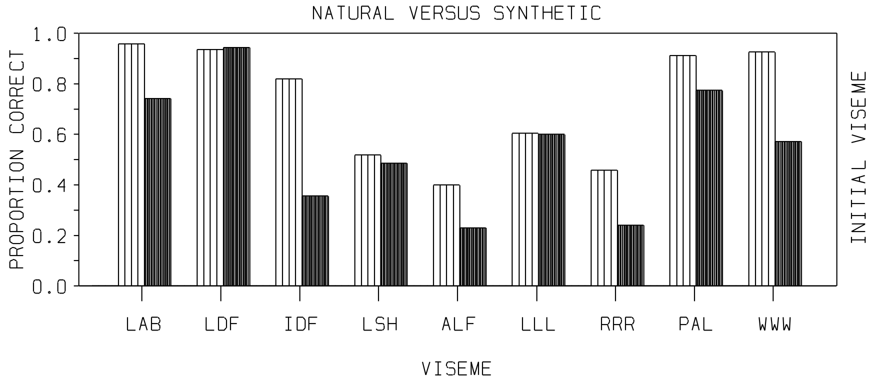

Fig. 2. Proportion of correct responses for initial consonant visemes for natural speech (striped bars) and synthetic speech (black bars) as a function of viseme class. The viseme classes are LAB (labial: b, p, m), LDF (labiodental fricatives: f, v), IDF (interdental fricatives: θ, ð), LSH (lingual stops and h: d, t, n, g, k, ng, h), ALF (alveolar fricatives: s, z), LLL (l), RRR (r), PAL (palato-alveolars: č, ǰ, š, ž), and WWW (w).

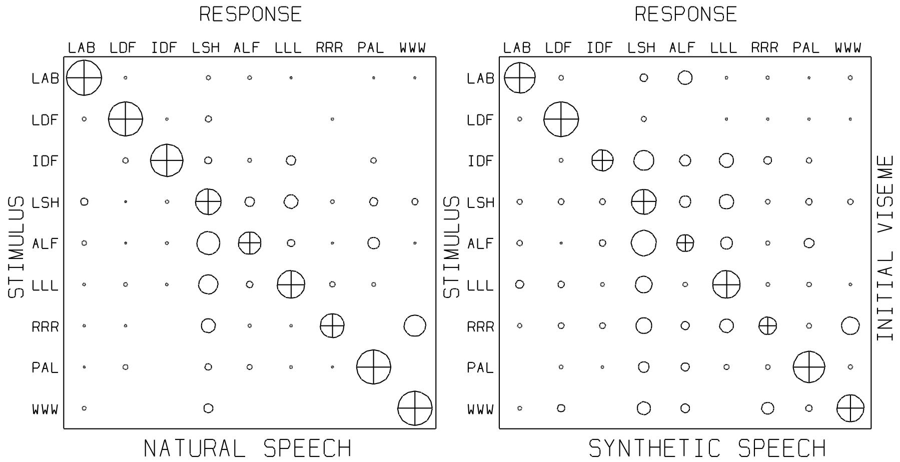

Fig. 3. Stimulus-response confusions for initial consonant visemes for natural (left panel) and synthetic (right panel) talkers. The area of each circle is proportional to the response probability. See Figure 2 for an explanation of the visemes.

The initial and final consonants were grouped into the 9 viseme classes of Walden et al. (1977). The viseme classes are LAB (labial: b, p, m), LDF (labiodental fricatives: f, v), IDF (interdental fricatives: θ, ð), LSH (lingual stops and h: d, t, n, g, k, ng, h), ALF (alveolar fricatives: s, z), LLL (l), RRR (r), PAL (palato-alveolars: č, ǰ, š, ž), and WWW (w). Figure 2 shows the proportion of correct initial consonant viseme (including correct initial C of a CCV) identifications for the two faces for each viseme class. Overall performance averaged 63.9% correct. There was a broad range of performance across the 9 viseme classes, ranging from .315 for ALF and .350 for RRR to .851 for LAB and .942 for LDF, $F(1,11) = 27.63$, $p < .001$. Performance given the natural face (.727) was superior to that for the synthetic face (.550), $F(1,11) = 35.58$, $p < .001$. For some distinctions (LDF, LSH, and LLL), however, performance given the two faces is about equivalent. Figure 3 shows the the pattern of viseme results in more detail, giving the proportion of each viseme response for the initial viseme of each word for the two faces. In this figure, the correct responses fall on the main diagonal. The overall advantage of natural speech is also apparent in this figure. There are many more off-diagonal responses for the synthetic than for the natural speech. A particularly striking limitation of the synthetic speech is the interdental fricatives, which are often identified as /d/, /z/, /l/, and /r/. Figure 4 shows this interdental fricative during articulation for the synthetic and natural speech. It is rather difficult to see what properties of the speech could account for this large difference.

Fig. 4. Interdental fricative during articulation for the natural and synthetic talkers.

Although the natural LAB is identified almost perfectly, the synthetic LAB is sometimes perceived as ALF. In general, similar confusions are made for the synthetic as for the natural face, even though performance on the synthetic face is generally poorer. As an example, ALF is called LSH, LLL, and PAL but more often for synthetic than for natural speech. The synthetic RRR is sometimes called LLL, although this seldom occurs for natural speech.

Figure 5 shows the proportion of correct final viseme identifications for the two faces for each viseme class. There are only 8 viseme classes because /w/ does not occur in final position in English. Overall performance averaged 57.3% correct. Again, there was a broad range of performance across the 9 viseme classes, ranging from .201 for ALF and .146 for RRR to .877 for LAB, .858 for LDF, and .814

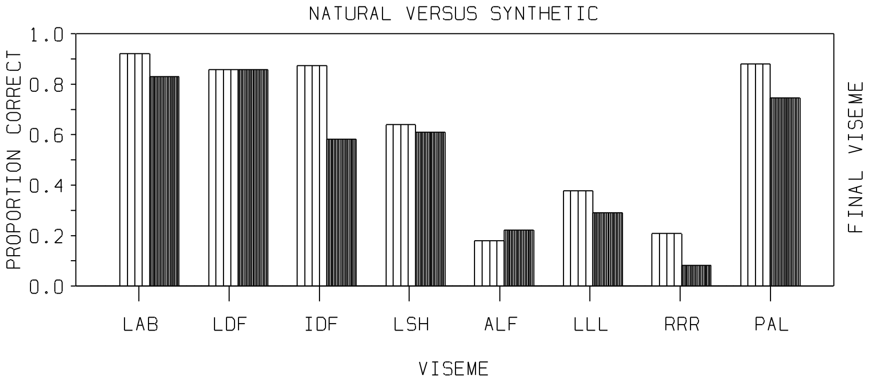

Fig. 5. Proportion of correct responses for final consonant visemes for natural speech (striped bars) and synthetic speech (black bars) as a function of viseme class. See Figure 2 for an explanation of the visemes.

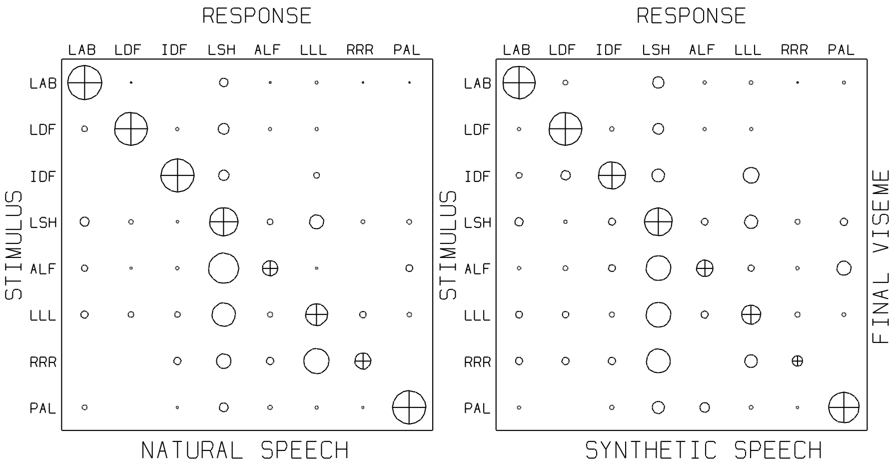

Fig. 6. Stimulus-response confusions for final consonant visemes for natural (left panel) and synthetic (right panel) talkers. The area of each circle is proportional to the response probability. See Figure 2 for an explanation of the visemes.

for PAL, $F(1,11) = 79.85$, $p < .001$. As with the initial visemes, performance given the two faces is about equivalent for some distinctions (LDF, LSH, and ALF), Overall performance given the natural face (.618) was superior to that for the synthetic face (.528), $F(1,11)=17.88$, $p=.002$, but by a somewhat smaller margin than for the initial viseme.

Analysis of performance on vowel visemes shows a larger advantage for the natural face (.717) over the synthetic (.369), $F(1,11)=128.60$, $p<.001$, than was observed for either consonant viseme position. The vowel viseme classes used, based on Montgomery and Jackson (1983), were HFR (high front: i, I), LBC (lower back: a, , Λ), NBL (non-low back lax: ʊ, ɜ), UW (u), NHF (non-high front: ɛ, æ, eI, aI), OY (I), AU (aʊ), and OU (oʊ). Figure 7 shows the

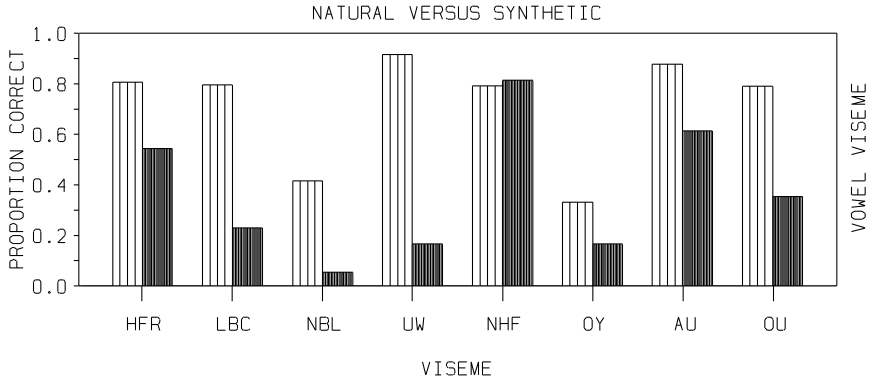

Fig. 7. Proportion of correct responses for vowel visemes for natural speech (striped bars) and synthetic speech (black bars) as a function of viseme class. The vowel viseme classes are HFR (high front: i, I), LBC (lower back: a, , Λ), NBL (non-low back lax: ʊ, ɜ), UW (u), NHF (non-high front: ɛ, æ, eI, aI), OY (I), AU (aʊ), and OU (oʊ).

proportion correct vowel viseme responses and Figure 8 shows the confusions. As can be seen in Figure 7, performance was better for every viseme except NHF. As can be seen in Figure 8, however, this case may have been due to a rather high rate of false-alarms to the synthetic NHF.

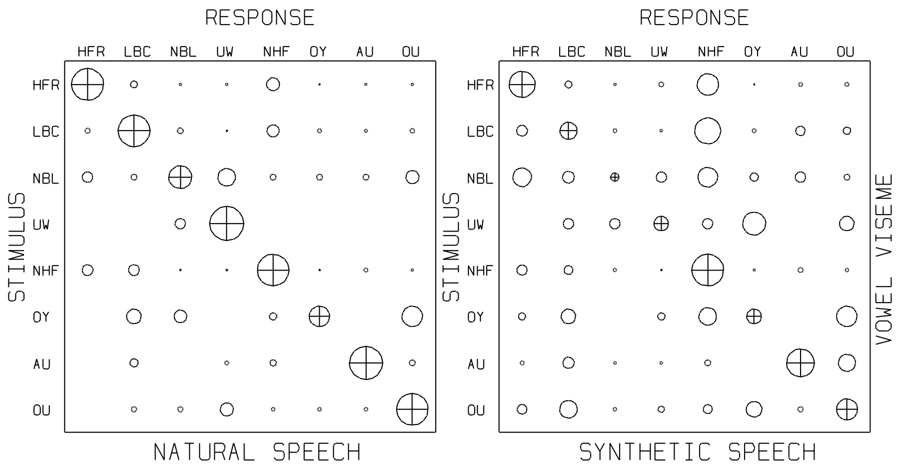

Fig. 8. Stimulus-response confusions for vowel visemes for natural (left panel) and synthetic (right panel) talkers. The area of each circle is proportional to the response probability. See Figure 7 for an explanation of the visemes.

4 Improving Phoneme Specifications

Given the results of Experiment 1, it is clear that there is room for a great deal of improvement in our synthetic speech. Improvement of our visual speech synthesis entails revising the phoneme specifications on the basis of several sources of information including: 1) phonetic descriptions of articulation (e.g. Ladefoged, 1975), 2) visual comparisons of the synthesized speech with a human talker, and 3) examination of recognition errors. The recognition data are particularly useful for spotting problem areas.

To illustrate the phoneme revision procedure, we describe the assessment of improvements needed for each viseme class and the types of modifications made to the articulatory parameter settings and dominance functions in response to this analysis. Looking at the consonant confusion data, we noted that the labial stops /p, b, m/ were often mistaken for clusters of /s/ followed by a labial stop, which is not a confusion that appears with the human talker. /s/ involves separation of the lips, while the labial stops have lip closure. To eliminate the confusion with an initial /s/, we emphasized the labial articulation by adjusting the settings for lip position parameters to increase lip contact and moved the dominance time centers for these parameters earlier in the segments to ensure an early onset of closure. The synthetic labiodental fricatives /f, v/ performed well in comparison to natural speech, so no significant adjustments were made for these segments.

The synthetic interdental fricatives /θ, ð/ were often mistaken for /h, l/ and alveolar and velar stops, yet these confusions occurred only occasionally with the human talker. The confusion data indicated that the interdental articulation was not visible enough for the synthetic /θ, ð/. Adjustments were made to the tongue length and angle parameter settings as well as to ones effecting mouth aperture in order to better produce an articulation between tongue and teeth. In making these modifications, we noted that finer control over different parts of the tongue would aid in the synthesis of these segments. A similar observation was made for /l/. Confusions for the synthetic /l/ were generally of the same type as those for /l/ in natural speech, although a greater amount of confusions occurred for synthetic speech. Parameters of tongue position for /l/ were fine-tuned, but visual comparison with the human talker indicated that more detailed control of tongue shape and shading may be needed to improve production of this segment.

For the lingual stops and h /t, d, n, k, g, ng, h/, confusion data was examined separately for each place of articulation. We observed that outside of their viseme class, the alveolar stops /t, d, n/ were most often confused with /s, l/, indicating that that the place of articulation was being recognized but the manner of articulation—a stop—was not. Phonetic studies of alveolar stops (e.g. Kantner & West, 1960) describe these segments as involving a slight raising of the upper lip and a slight lowering of the lower lip. The jaw also lowers. These aspects of the segments were enhanced by adjusting the target parameter settings for lip and jaw height. For the velar stops /k, g, ng/, confusion data revealed that outside of their viseme class these segments are most systematically mistaken for /l/. From careful comparison with visual recordings of the human talker, we determined that more spreading at the corners of the mouth was needed, as well as some extension of the chin and cheeks and a greater tongue angle. Target parameter settings were adjusted accordingly, and dominance functions were strengthened for the appropriate facial gestures. The last member of the LSH consonant viseme group, /h/, was often mistaken for an alveolar consonant. We based the revisions of settings for this phoneme on the phonetic observation that rather than having one set position for the visible articulation of /h/, the mouth should assume the position of the following vowel (Kantner & West, 1960). To realize this kind of dependent articulation, the strength of jaw rotation dominance function was decreased, so that the mouth aperture for /h/ would more closely assume that of the following vowel. Also, the tongue angle was set closer to that for the vowels.

Looking at the confusion data for the alveolar fricatives /s, z/, we noted that these segments were often confused with the alveolar stops. However, this occurred also for the human talker. More problematic were confusions with interdental fricatives, /l/ and velar stops. From phonetic research (Ladefoged & Maddieson, 1986) and the human talker we know that the fricatives should have the teeth very close together, and that the lower lip may be used to direct the airstream toward the upper teeth. These characteristics were emphasized in the new parameter set.

The rounded consonant /w/ was often mistaken for /r/, which has some but less lip rounding, or it was mistaken for a cluster containing an unrounded consonant plus a rounded consonant, such as /kw, tw/. These confusions indicated that /w/ needed more lip rounding, so we increased the relevant parameter settings. The phoneme /r/ was often mistaken for /w, l, θ, ð/ and clusters containing /r/. With the human talker /r/ was also confused with /w/, so confusions with /w/ are not very serious for the synthetic /r/. The more serious confusions involved mistaking /r/ for an unrounded consonant. We found that especially with front or central unrounded vowels, the initial /r/ was insufficiently rounded, so we boosted settings for parameters contributing to lip protrusion and a circular mouth aperture—both characteristics of rounding in /r/. Final /r/, however, should not be as rounded as initial /r/. To solve this problem, we modified the AT&T TTS to output separate allophones for initial and final /r/ and gave these two separate definitions—an approach that might be extended to some other phonemes.

The palato-alveolar group /š, ž, č, ǰ/ were often confused with /s/, /t/, and in the case of the affricates /č, ǰ/ there were also confusions with an /st/ cluster. The palato-alveolar phonemes involve lip rounding, in contrast to the alveolar sounds segments /s, t/, so rounding of the mouth aperture in the palato-alveolar segments and protrusion of the lips were boosted. To eliminate confusions with an /st/ cluster, the stop-fricative sequence in the affricates was enhanced by adjusting dominance time centers of parameters to push lip rounding for the fricative portion later in the segment.

We turn now to the vowels. For the viseme class consisting of the high front vowels /i, I/, the confusion data revealed that these segments were most often mistaken for the front mid vowels /eI, ε/. Since /i, I/ have a visibly greater lip spreading than /eI, ε/, this feature of the phonemes was increased. The nonhigh front vowels form a second viseme class /eI, ε, æ, aI/. Overall synthetic /eI/ and /ε/

were getting good recognition with minimal confusions outside of their viseme class, so these phonemes were left basically intact. On the other hand /æ/ and /eI/ were sometimes being mistaken for higher front vowels. For /æ/ the target jaw position was adjusted to increase the vertical dimension of the mouth aperture and thereby enhance the visibility of the low height of this vowel. /aI/ is a diphthong comprised of /a+i/. Wozniak & Jackson (1979) find that dipthongs have better recognition than monophthongs, so where possible we emphasized the movement in diphthongs in the vowel phoneme revisions. For /aI/ the visibility of the movement from a low central to a high front position was emphasized by refining the position of the component vowel articulations.

The lower back and central vowels form a viseme class comprised of /a, , Λ/. The most common confusions for /a/ and /Λ/ were for vowels in the nonhigh front viseme class, indicating that the nonfront quality of the vowels was not visible enough. To emphasize backness in /a/ the tongue angle was adjusted, although we noted that control of tongue shading and independent manipulation of the tongue tip versus body may be needed for more significant improvement. The central vowel /Λ/ was distinguished from the front vowels by decreasing the lip spreading and producing a more neutral mouth position. The most problematic confusions for / / were /aU, oU/, both diphthongs terminating in a high back position. Accordingly, we modified settings for / / to emphasize the mid-low position by lowering the jaw. Since rounded / / was also sometimes confused with unrounded /a/, rounding was also boosted slightly in / /.

The mid back rounded vowel /oU/ was most often confused with the lower back vowels /a, /. Visual comparison of synthetic /oU/ versus the /oU/ of the human talker revealed that /oU/ needs more anticipatory rounding and the rounding needed to have a narrower and taller oval shape, so adjustments were made to target settings and the anticipatory dominance functions of rounding parameters to achieve this. Data on the high rounded vowels /U/ and /u/ was limited, as there were not many instances of these phonemes in the words used in the experiment. However, comparison of the synthetic speech with the human talker indicated that both phonemes needed more lip rounding, so this was increased. No direct adjustment was made to the diphthongs /aU, I/, as these phonemes are synthesized by the combination of two component phonemes which were independently adjusted.

After completing adjustments to each of the vowel phonemes based on the above kinds of considerations, the target settings for pairs of parameters for the vowels were plotted in several two-dimensional spaces (e.g. jaw-rotation and mouth width) and inspected for basic correspondence to the physical distribution of the vowels according to phonetic charts. Fine-tuning of the parameter settings were then made as needed. Special attention was paid to the parameters of horizontal mouth size, protrusion of the mouth, and jaw rotation, which roughly match the dimensions of lip spreading, lip rounding, and vertical lip separation, respectively. These three dimensions were identified by Jackson, Montgomery, & Binnie (1976) as important visual cues in vowel perception. We are currently evaluating these new parameter values for the synthetic speech and will report on thes experiments in the near future.

5 Experiment 2: Point-Light Visual Speech

In addition to the experimental results comparing natural and synthetic speech, the facial synthesis technology is being used to assess the use of kinematic information in visual speech perception. Recently, Rosenblum and Saldana (1995) using hybrid visual-auditory tests have argued that a point-light display using 28 lights attached to the face is effective as a visual speech stimulus. Why should point-light displays be effective stimuli for speechreading? First, it is well-known that point-light stimuli are effective stimuli for displaying biological motion, such as human walking, running, dancing, cycling, etc. If spots of light are placed on the joints of a human, a static presentation of just the points of light is not seen as such. If the human now moves, however, then the appropriate event is perceived. Johansson (1973) originally explained these findings in terms of a kinetic-geometric model for visual vector analysis. The visual system putatively interprets two distal points of light that move together as end points of a distinct line segment. In this way, points of light are actually being interpreted as line segments and the coordinated motion can give rise to the perception of meaningful events.

This same logic has been applied to the perception of visible speech (Rosenblum & Saldana, in press). These investigators operate under the assumption that speech perception is based on the recovery of stimulus information, which is lawfully related to a coherent auditory event. Thus, optic and acoustic information both specify a speech event, and thus both are functional in perception. Furthermore, the information is necessarily higher-order and structured across time. Point light displays that have the points of light on the important articulators should therefore be capable of specifying the higher-order information supporting speechreading. Thus, the point-light displays test the assumptions that the information supporting speechreading is necessarily higher-order relationships among the moving points of light over time. Some additional support for the use of kinematic information in visual speech perception comes from the work of Goldschen (1993) who analyzed 35 static and dynamic features of the oral cavity and their information value. Using principal components analysis, the 13 most informative features were for the most part dynamic ones. To further assess the value of such kinematic information, in Experiment 2 we used the word recognition paradigm to compare our synthetic face to a point-light version of the face. This test provides a more extensive assessment of how much information can be transmitted by point-light displays, as well as a better control to equate the stimuli for the two conditions. The original study 1) only used /ba/ and /va/ tokens, and 2) the facial and point-light displays were made separately under differing conditions, which makes it difficult to say whether the stimuli had equivalent articulations.

5.1 Method

Twelve college students served as subjects. The experimental procedure was identical to that used in Experiment 1, except that the synthetic point-light face replaced the natural one. The point-light display was made by putting tiny spheres on 28 of the polygon vertices of the face, positioned on the face, lips, teeth, and tongue identically to those used by Rosenblum and Saldana. Figure 9 shows the regular synthetic face, the synthetic face with sphere markers, and point-light face displays. The markers were colored white while the face itself was colored black. Although the face was invisible against the black background, the surfaces were displayed normally so that, for example, the points on the tongue and teeth disappeared when the mouth closed.

Fig. 9. Synthetic face, synthetic face with sphere markers, and point-light face displays.

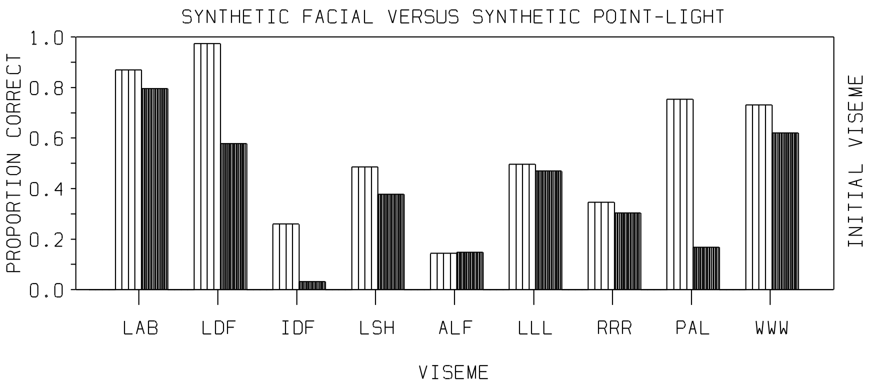

Fig. 10. Proportion of correct responses for initial consonant visemes for synthetic facial (striped bars) and point-light speech (black bars) as a function of viseme class. See Figure 2 for an explanation of the visemes.

5.2 Analysis

The data from Experiment 2 was analyzed identically to that in Experiment 1. Figures 10 and 11 give the proportion correct and confusions for the initial consonant visemes for the synthetic facial and synthetic point-light talkers. As can be seen, performance on initial consonant visemes was clearly worse for the point-light display compared to the synthetic facial display overall, (.564 vs .389) $F(1,11)=92.92$, $p<.001$, especially for the LDF, IDF, and PAL classes, with a significant interaction of face by viseme, $F(8,88)=11.88$, $p<.001$. Figures 12 and

13 similarly give the proportion correct and confusions for the final consonant visemes for the two talkers. As can be seen, the point-light display (.273) was worse than the synthetic face (.516) on every viseme, F(1,11)=281.54, p<.001.

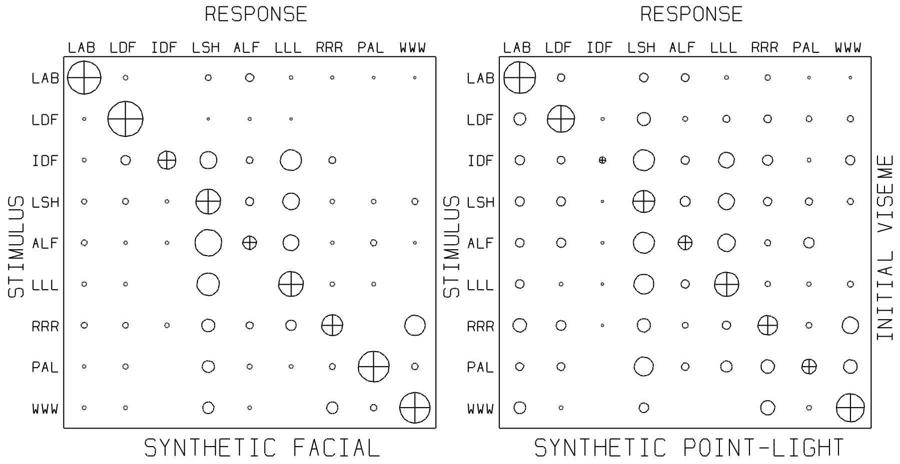

Fig. 11. Stimulus-response confusions for initial consonant visemes for synthetic facial (left panel) and synthetic point-light (right panel) talkers. The area of each circle is proportional to the response probability.

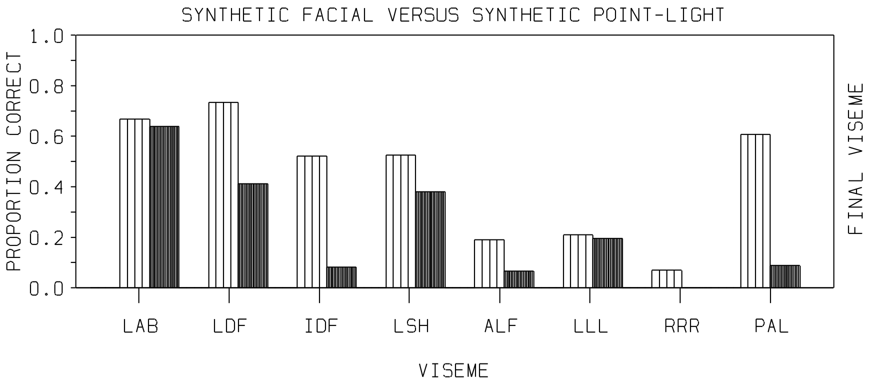

Fig. 12. Proportion of correct responses for final consonant visemes for synthetic facial (striped bars) and point-light speech (black bars) as a function of viseme class. See Figure 2 for an explanation of the visemes.

Particularly poor in final position were the LDF, IDF, ALF, RRR, and PAL classes, with a significant interaction of face by viseme, F(8,88)=17.78, p<.001.

Analysis of performance on vowels visemes shows a small but significant advantage for the synthetic facial (.353) over the point-light (.299) talker, F(1,11)=8.59, p=.013. Figure 14 shows the proportion correct vowel viseme and Figure 15 shows the confusions, which are a bit more randomly distributed for the point-light talker.

To summarize the results of Experiment 2, it is evident that with equivalent display geometries, the point-light display does provide some valuable information for speech reading, replicating the results of Rosenblum et al., although performance was significantly worse than the full synthetic facial display. Thus, it appears that kinematic properties are informative for speechreading, although we

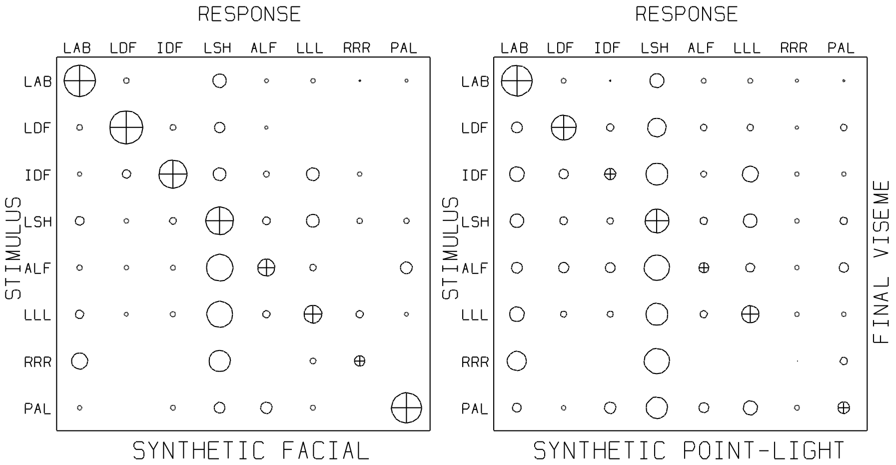

Fig. 13. Stimulus-response confusions for final consonant visemes for synthetic facial (left panel) and synthetic point-light (right panel) talkers. The area of each circle is proportional to the response probability. See Figure 2 for an explanation of the visemes.

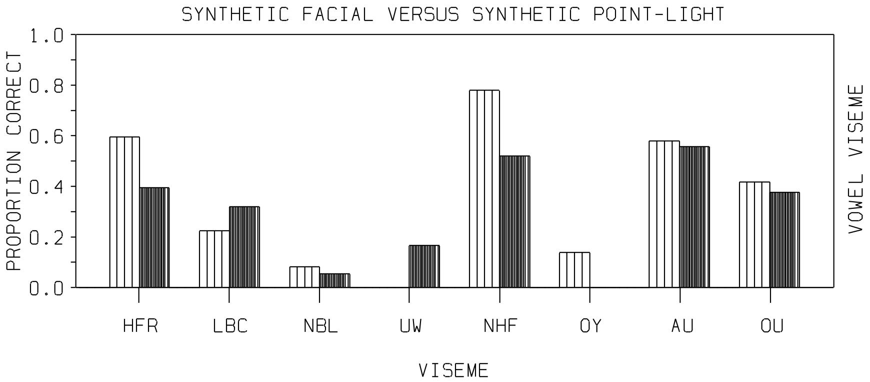

Fig. 14. Proportion of correct responses for vowel visemes for synthetic facial (striped bars) and point-light speech (black bars) as a function of viseme class. See Figure 7 for an explanation of the visemes.

do not believe they are sufficient. Some might argue that this performance difference occurred because of the overall inferiority of the synthetic face relative to a natural face, or that the markers (albeit in the same locations as prior point-light experiments) were not optimally placed. Further empirical tests with improved displays may answer these concerns.

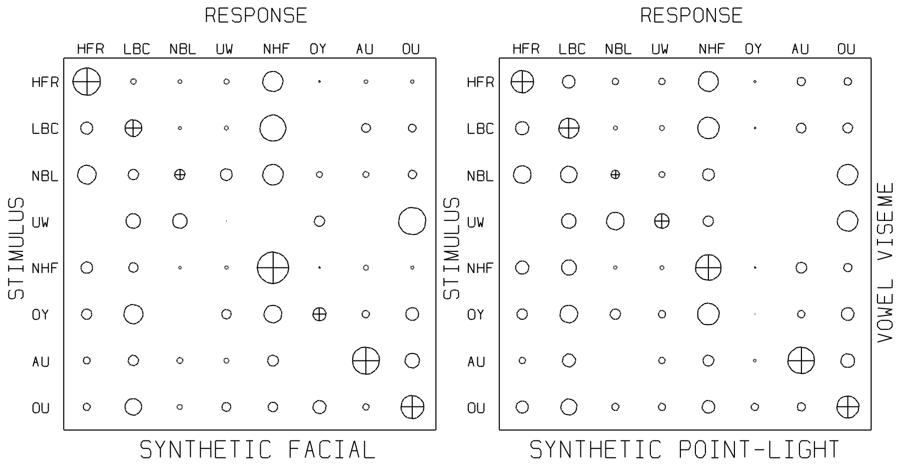

Fig. 15. Stimulus-response confusions for vowel visemes for synthetic facial (left panel) and synthetic point-light (right panel) talkers. The area of each circle is proportional to the response probability.

6 Future Improvements

Overall, visual recognition of the synthetic talker is reasonably close to that of the human talker, but a significant distance remains to be covered. We believe that for both palato-alveolars and /w/, the current lips do not provide sufficient rounding of the upper lip. This is being addressed in a revision of the lip model. The interdentals, alveolar fricatives and some vowels might have also been a problem due in part to our original tongue model. This model is a rigid grooved paddle, having only three control parameters: tongue length, tongue angle, and tongue width. In natural speech certain consonants cause cupping or grooving of the tongue which also coarticulates to following vowels. As examples, it has been documented that the fricative /s/ can cause a deep groove to occur (e.g. Hardcastle, 1976; Ladefoged & Maddieson, 1986), and for /ð, θ/ we observe that there is a cupping action. In addition, tongue articulations generally involve raising a specific part of the tongue, such as the tongue tip or body. These characteristics can now be simulated in our new tongue model. This new model is based on b-spline curves controlled by seven sagittal and seven coronal parameters including tip, body, and overall thickness, tip and top height, tip, body and top advancement, width, grooving, edge height and thickness, and tip shape. Currently we are tuning the use of these parameters, based in part on automatic parameter adjustment to best fit sagittal Flash-MRI recordings. In addition we are working on rendering real-time shadows on the tongue which may be perceptually important.

Acknowledgment

The writing of this paper was supported, in part, by Public Health Service Grant R01 DC 00236 to Dominic Massaro and a Social Sciences and Humanities Research Council of Canada (SSHRC) Fellowship 752-93-2397 to Rachel Walker.