Behavior Research Methods, Instruments, & Computers
1990, 22 (2), 260-263

Synthesis of visible speech

MICHAEL M. COHEN and DOMINIC W. MASSARO
University of California, Santa Cruz, California

We have implemented a facial animation system to carry out visible speech synthesis. Using this system, it is possible to manipulate control parameters to synthesize a sequence of speech articulations. In addition, it is possible to synthesize novel articulations, such as one that is half way between /ba/ and /da/.

Given the importance of visible information in face-to-face communication, visible speech synthesis is being developed to control and manipulate visible speech. Experiments have shown that this visible speech is particularly important when the auditory speech is degraded, because of noise, bandwidth filtering, or hearing impairment (Massaro, 1987). The strong influence of visible speech is not limited to situations with degraded auditory input, however; it occurs even when visible speech is paired with perfectly intelligible speech sounds. The influence of visible speech is easily experienced in a demonstration of a McGurk effect (McGurk & MacDonald, 1976). A videotape of a person making a visible labial articulation /pa-pa/ is dubbed with the alveolar nasal speech sounds /na-na/. This dubbed speech event gives a situation in which intelligible auditory speech is paired with a contradictory visual articulation. A strong effect of the visual source of information is observed, with the viewer often reporting hearing the labial nasal /ma-ma/. It can be shown that the viewer's experience is influenced by both the audible and the visible speech.

The value of visible speech in speech perception warrants that visible speech should be studied, just as auditory speech has been studied. Although some progress has been made using natural speech as stimuli, the synthesis of a speaker's face permits a better controlled and more systematic analysis of the perceptual process. Given that synthetic auditory speech has proven valuable for the study of auditory speech perception, visible speech synthesis should be a valuable tool for the study of visual and bimodal (auditory-visual) speech perception. In addition to enabling exact control over the speech stimulus, synthetic speech allows the creation of novel speech segments that are not easy to produce naturally. Presenting novel stimuli to subjects in psychophysical tasks provides important information about the processes involved in perception.

The research reported in this paper and the writing of the paper were supported, in part, by grants from the Public Health Service (PHS R01 NS 20314), the National Science Foundation (BNS 8812728), the graduate division of the University of California, Santa Cruz, and a James McKeen Cattell Fellowship to Dominic Massaro. Correspondence should be addressed to Michael M. Cohen, Program in Experimental Psychology, University of California, Santa Cruz, CA 95064.

Several forms of simulated facial display have been used for speech studies. Relatively simple Lissajous figures have been displayed on an oscilloscope to simulate lip movement, using analog control voltages to vary the height and width of the simulated lips (Erber & De Filippo, 1978). A model for lip shape was developed that allowed computation of coarticulatory effects for CVCVC segments (Montgomery, 1980). The lip-shape display was done on a vector graphic device using about 130 vectors at a rate of about 4 times real time. A real-time vector display system for displaying simple two-dimensional faces has also been reported (Brooke & Summerfield, 1983).

To generate more realistic full facial displays, two general strategies have been employed: musculoskeletal models (Platt & Badler, 1981) and parametrically controlled polygon topology (Parke, 1974). The latter strategy was used in a fairly realistic animation by modeling the facial surface as a polyhedral object composed of about 900 small surfaces arranged in three dimensions and joined together at the edges (Parke, 1974, 1975, 1982). The left panel of Figure 1 shows a framework rendering of this model. To achieve a natural appearance, the surface is smooth shaded, using Gouraud's (1971) method (shown in the right panel of Figure 1). The face is animated by altering the location of various points in the grid under the control of 50 parameters, 11 of which are used for speech animation. Control parameters used for several demonstration sentences were selected and refined by the investigator, who studied his own articulation frame by frame and estimated the control parameter values (Parke, 1974).

Recently, this software and facial topology have been translated from the original ALGOL to C and given new speech- and expression-control routines (Pearce, Wyvill, Wyvill, & Hill, 1986). In this system, a user can type a string of phonemes, which are then converted to control parameters and changed over time to produce the desired animation sequence. Each phoneme is defined in a table according to values for segment duration, segment type (stop, vowel, liquid, etc.), and 11 control parameters. The parameters that are used are jaw rotation, mouth x scale, mouth z offset, lip corner x width, mouth corner z offset, mouth corner x offset, mouth corner y offset,

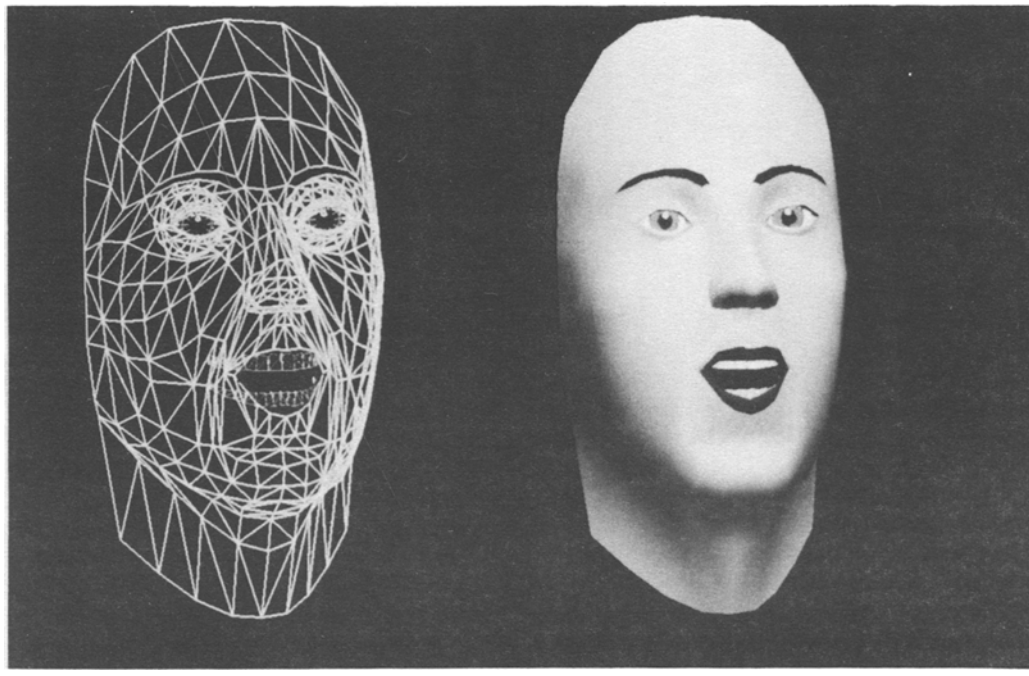

Figure 1. Framework (left) and Gouraud shaded (right) renderings of polygon facial model.

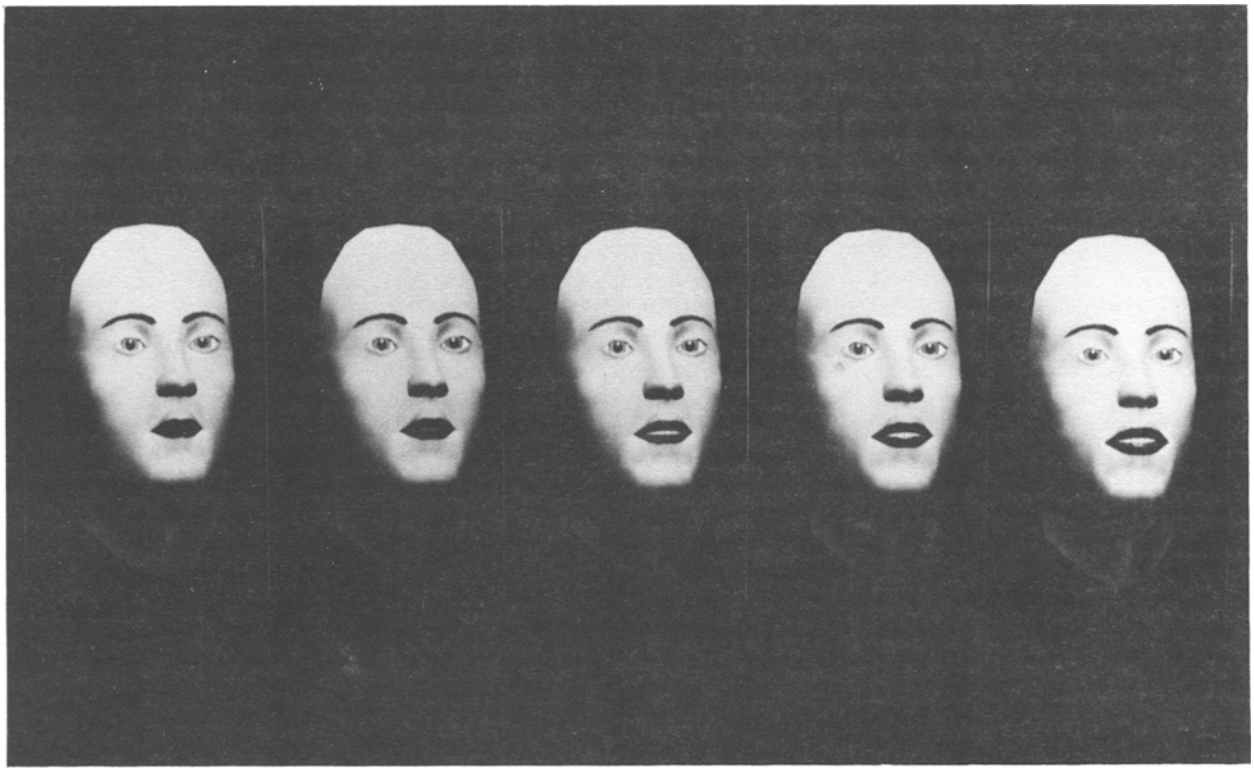

Figure 2. The facial model at the moment of stop closure for each of the five levels of visible speech between /ba/ and /da/.

Table 1
Visual Synthesis Parameters for the Default Position, Five Stops, and /a/

Parameter	Default	/b/	2	3	4	/d/	/a/
jaw rotation	3.00	0.00	0.75	1.50	2.25	3.00	10.00
mouth x scale	1.00	1.00	1.05	1.10	1.15	1.20	1.00
mouth z offset	0.00	−1.00	−0.75	−0 50	−0.25	0.00	2.00
lip corner x width	0.00	0.00	1.25	2.50	3.75	5.00	20.00
mouth corner z offset	0 00	−15 00	−15 00	−15.00	−15.00	−15.00	0.00
mouth corner x offset	0.00	2.00	4.50	7.00	9 50	12.00	0.00
mouth corner y offset	0.00	0.00	0.75	1.50	2.25	3.00	−5 00
lower lip "f" tuck	0 00	−5.00	−5.00	−5.00	−5.00	−5.00	0.00
upper lip raise	0 00	2.00	4 75	7.50	10.25	13.00	2 00

lower lip "f" tuck, upper lip raise, and x and z teeth offset. The revised software of Pearce et al. (1986) was implemented by us on a Silicon Graphics Inc. IRIS 3030 computer. We have adapted the software to synthesize new intermediate test phonemes and written several output processors (pipes) for rendering the polygonal image information in different ways. One pipe produces wire frame images, a second produces Gouraud shaded images with a diffuse illumination model, a third also includes specular illumination (white highlights), and a fourth pipe is being developed, which uses tesselation (recursive polygon subdivision) for improved skin-texture appearance as well as randomly determined hair. The diffuse pipe currently being used requires about 1 min to render each frame, while the diffuse + specular rendering takes 3 min. This time is a considerable improvement over a speed of about 15 min per frame previously reported for a diffuse illumination model (Pearce et al., 1986). To create an animation sequence, each frame is recorded using a broadcast quality BETACAM video recorder under control of the IRIS.

In prototypical experiments on auditory-speech perception, some property of the speech stimulus is varied in small steps to give a continuum of speech sounds between two alternatives. For example, using software synthesis (Klatt, 1980), the onsets of the second and third formants can be varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, we can systematically vary parameters of the facial model to give a continuum between visual /ba/ and /da/. Figure 2 gives pictures of the facial model at the time of maximum stop closure for each of the five levels between /ba/ and /da/. Table 1 gives the parameter target values used in the visual synthesis for the default resting parameter values, the consonant portion of each visual stimulus, and the values for the vowel /a/. The top panel of Figure 3 shows how the

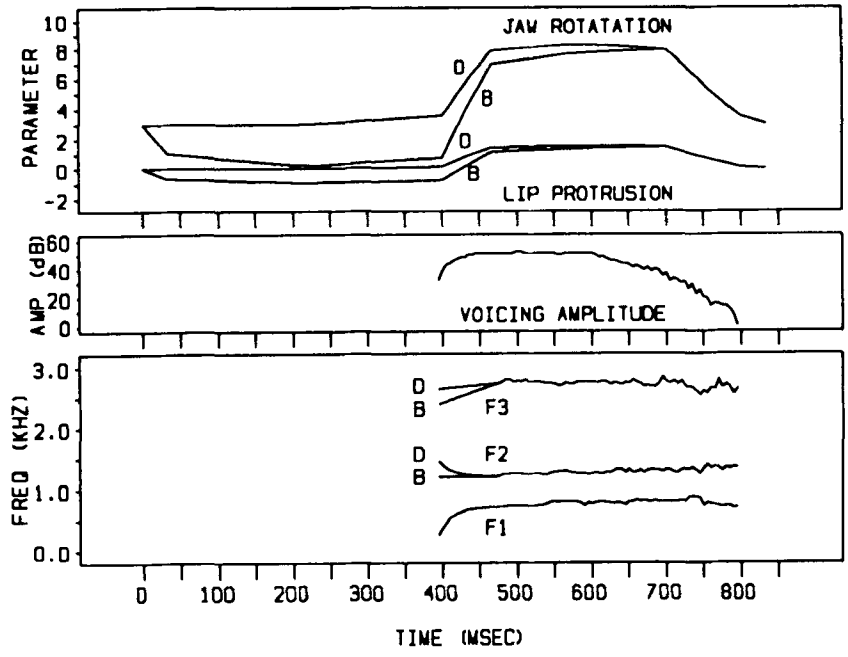

Figure 3. Visual and auditory parameter values over time for auditory-visual /ba/ and /da/. Bottom and middle panels show auditory formants F1, F2, and F3 and voicing amplitude AV. Top panel shows visible jaw rotation and lip protrusion.

visual synthesis parameters change over time for the first (/ba/) and last (/da/) visual levels. For clarity, only two of the visual parameters are shown: jaw rotation (the larger parameter means more open), and lip protrusion (mouth z offset in Table 1; the smaller number means more protrusion). Not shown in Figure 3, the face with the default parameter values is recorded for 2,000 msec preceding and 2,000 msec following the time shown.

Following the synthesis and frame-by-frame recording, the BETACAM tape can be used to create experimental tapes for editing. A tone marker is dubbed onto the audio channel of the tape at the start of each syllable, to allow exact alignment of an auditory speech stimulus that can be presented on bimodal trials with both audible and visible speech. The marker tone on the videotape is sensed by a Schmidt trigger on a PDP-11/34A computer, which also presents the auditory stimuli from digitized representations on the computer's disk. The temporal relationship between the auditory and visual parts of two of the bimodal stimuli is illustrated in Figure 3.

Experiments have shown that this animated visible speech influences speech perception (Massaro & Cohen, 1990). A demonstration of the powerful influence of visible speech, the synthetic visible speech, and an experiment on bimodal speech perception are available on VHS videotape. Requests should be sent to Dominic W. Massaro.

REFERENCES

BROOKE, N. M., & SUMMERFIELD, A. Q. (1983). Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics*, **11**, 63-76.

ERBER, N. P., & DE FILIPPO, C. L. (1978). Voice-mouth synthesis of /pa, ba, ma/. *Journal of the Acoustical Society of America*, **64**, 1015-1019.

GOURAUD, H. (1971). Continuous shading of curved surfaces. *IEEE Transactions on Computers*, **C-20**, 623-628.

KLATT, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-995.

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

MASSARO, D. W., & COHEN, M. M. (1990). Perception of synthesized audible and visible speech. *Psychological Science*, **1**, 55-63.

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

MONTGOMERY, A. A. (1980). Development of a model for generating synthetic animated lip shapes. *Journal of the Acoustical Society of America*, **68**(Suppl. 1), S58 (Abstract No. FF13).

PARKE, F. I. (1974). *A parametric model for human faces*. (Tech Rep. No. UTEC-CSc-75-047). Salt Lake City: University of Utah, Department of Computer Science.

PARKE, F. I. (1975). A model for human faces that allows speech synchronized animation. *Computer & Graphics Journal*, **1**, 1-4.

PARKE, F. I. (1982). Parameterized models for facial animation. *IEEE Computer Graphics*, **2**(9), 61-68.

PEARCE, A., WYVILL, B., WYVILL, G., & HILL, D. (1986). Speech and expression: A computer solution to face animation. *Graphics Interface '86*.

PLATT, S. M., & BADLER, N. I. (1981). Animating facial expressions. *Computer Graphics*, **15**(3), 245-252.