Talking Machines: Theories, Models, and Designs
G. Bailly, C. Benoît, and T.R. Sawallis (Editors)
1992 Elsevier Science Publishers B.V.

Speech by eye

Dominic W. Massaro and Michael M. Cohen

Program in Experimental Psychology, University of California - Santa Cruz, Santa Cruz, CA 95064 USA

Like our colleagues Benoît, Lallouache, Mohamadi, and Abry (this volume) and Brooke (this volume), we also believe that synthetic visible speech has an obvious potential for advancing our knowledge about the visible information in speech perception and how it is utilized by human perceivers. We are using a high quality facial display originally developed by Parke (1974) and recently augmented by Pearce, Wyvill, Wyvill, and Hill (1986) and ourselves (Cohen & Massaro, 1990). In this display the face is constructed of about 900 3D-polygons controlled by about 50 parameters. There is now some simple control of visible speech, but a better model of speech articulation is being developed incorporating physical measurements from real speech and rules describing coarticulation between segments. In addition, further work is proposed to increase the available information in the animated display and to improve the quality of the speech synthesis. For example, a tongue will be added to the facial model and phoneme descriptions will be elaborated to include their systematic variation with neighboring speech segments. Psychophysical studies generating confusion matrices and standard tests of intelligibility will be used to assess the quality of the facial synthesis. Additional research is proposed to evaluate the transformation of auditory speech cues to visual cues, such as the color of the lips during articulation. A critical assumption underlying this work is the experimental, theoretical, and applied value of synthetic speech. Auditory synthetic speech has proven to be valuable in all three of these domains. Much of what we know about speech perception has come from experimental studies using synthetic speech. Synthetic speech gives the experimenter control over the stimulus in a way that is not always possible using natural speech. Although the experimental validity of synthetic speech might be questioned, it is also the case that phenomena uncovered using synthetic speech also hold up when tested using natural speech (Massaro, unpublished). Synthetic speech also permits the implementation and test of theoretical hypotheses, such as which cues are critical for various speech distinctions. The applied value of auditory synthetic speech is apparent in the multiple everyday uses for text-to-speech systems for both normal and hearing-impaired individuals.

We believe that visible synthetic speech will prove to have the same value as audible synthetic speech. In our initial studies of a /ba/-/da/ continuum using synthetic visible speech, we found very similar results for the endpoint stimuli /ba/ and /da/ compared to earlier studies using natural speech tokens (Massaro, 1987; Massaro & Cohen, 1990). Synthetic visible speech will provide a more fine-grained assessment of psychophysical and psychological questions not possible with natural speech. For example, testing subjects with synthesized syllables intermediate between several alternatives gives a more powerful measure of integration relative to the case of unambiguous natural stimuli. It is also obvious that synthetic visible speech will have a valuable role to play in alleviating some of the communication disadvantages of the deaf and hearing-impaired.

The guiding assumption of our research has been that humans use multiple sources of information in the perceptual recognition and understanding of spoken language. In this regard, speech perception resembles other forms of pattern recognition and categorization because integrating multiple sources of information appears to be a natural function of human endeavor. Integration appears to occur to some extent regardless of the goals and motivations of the perceiver. Brunswik (1955) acknowledged the multiple but ambiguous sources of influence on behavior. He stressed "the limited ecological validity or trustworthiness of cues .... To improve its (the organism's) bet, it must accumulate and combine cues" (1955, p. 207). With respect to the worth of information, Brunswik distinguished between two kinds of validity. Ecological validity defines what cues are informative about the structure of the world. Functional validity defines what cues people actually use in perceptual processing. Given this distinction it can be seen that a concern for ecological validity is not sufficient, given that some ecologically valid property of the physical world may not be used and hence not be functionally valid. Thus it is an empirical question whether the 21 classes Benoit et al. obtained from multidimensional analysis are psychologically real. It is important to determine the ecologically valid cues in the analysis of the articulation of human observers and also the functionally valid cues in perceptual tests using synthetic speech.

We have developed an experimental paradigm to determine which of the many potentially functional cues are actually used by human observers. Identification experiments carried out with the systematic variation of properties of the speech signal combined with the quantitative test of models based on different sources of information enables the investigator to test the psychological validity of different cues. For example, a voicing distinction allows us to perceive a difference between the verb in the phrase *to use* and the noun in the phrase *the use*. For over three decades, speech scientists believed that consonant duration *relative* to vowel duration was the critical cue to the voicing judgments (Denes, 1955; Port & Dalby, 1982). However, Massaro and Cohen (1977, 1983) showed that this cue (called C/V ratio) is invalid, when the results are analyzed in the framework of the fuzzy logical model of perception (FLMP) (Massaro, 1987). A model based on C/V ratio gives a much poorer description of the identification results than does the assumption of independent consonant and vowel duration cues (Derr & Massaro, 1980). We expect this paradigm to be equally effective in the study of visible speech.

There is valuable and effective information afforded by a view of the speaker's face in speech perception and recognition by humans. Visible speech is particularly effective when the auditory speech is degraded, because of noise, bandwidth filtering, or hearing-impairment. As an example, the perception of short sentences that have been bandpass filtered improves from 23% to 79% correct when subjects are permitted a view of the speaker (Breeuwer & Plomp, 1985). This same type of improvement has been observed in hearing-impaired listeners and patients with cochlear implants (Massaro, 1987). The strong influence of visible speech is not limited to situations with degraded auditory input, however. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. If an auditory syllable /ba/ is dubbed onto a videotape of a speaker saying /ga/, subjects often perceive the speaker to be saying /da/ (Massaro & Cohen, 1990).

An interesting question concerning synthesis of visual speech is to what degree is coarticulation important. Coarticulation refers to changes in the articulation of a speech segment depending on preceding (backward or left-to-right coarticulation) and upcoming segments. Benguerel and Pichora-Fuller (1982) examined coarticulation influences on lip-reading by hearing-impaired lip-readers and normal-hearing subjects. (forward or right-to-left coarticulation). An example of forward coarticulation is the anticipatory lip rounding at the beginning of the syllable /su/. The test items in the study were /V_1CV_2/ nonsense syllables. Coarticulation influences were assessed by contrasting consonant recognition in vowel contexts that produce large coarticulatory influences relative to those that produce small influences. Consistant with the Owens and Blazek (1985) study, there was no overall difference between normal and hearing-impaired subjects. Significant coarticulation effects were noted. For example, the identity of V_2 had a significant effect on consonant recognition. Fewer consonants were recognized correctly when they were followed by /u/ than by /i/ or /æ/. By reversing the stimuli and finding the same results, they demonstrated that the effect was due to articulation differences rather than the actual position in the stimulus as presented.

The existing software for driving the visible speech display simply interpolates between segments with no coarticulation. It should be noted that the great improvement of more recent auditory speech synthesizers, such as MITalk and DECtalk, over the previous generation of synthesizers such as VOTRAX, is due to including rules specifying the coarticulation among neighboring phonemes.

In considering the coarticulation of visible speech we recognize a conflict between the natural situation and what may be optimal for information transfer during perception. We know, for example, that vowel environment influences visibility of consonants (e.g., worse performance with /u/ than with /a/). It may turn out that somewhat unnatural (e.g., slow and noncoarticulated) visual speech is most easily recognized by perceivers.

Several coarticulation models will be evaluated for implementing coarticulation in the facial model, including the binary feature model of Benguerel and Cowan (1974) and the coarticulation resistance model of Bladon and Al-Bamerni (1976) and Bladon (1977). The model of Benguerel and Cowan (1974) assumes that each segment is defined in terms of a vector of component features which are either + (present), - (absent), or 0 (undefined). For our synthesis, we may equate a feature with a facial control parameter value. Another method of programming for coarticulation effects to be investigated is the use of a connectionist or neural-network model. In such a model a network of nodes with weighted connections is used to connect a set of inputs and outputs. Such a system has recently been used for the conversion of text to phonetic features (e.g., Sejnowski & Rosenberg, 1987). We propose to use this method to relate input phonetic segments with visual articulatory measurements. Once the network has been trained, it should output different visual parameters for the same phonetic segment in different phonological environments. These parameters will be used to drive the visual synthesis, giving context sensitivity due to coarticulation among neighboring segments.

Returning now to the papers in the present volume, we note that both Brooke and Benoit et al. refer to the complementarity of audible and visible speech. For example, Benoit et al. state that "the partial complementarity between the auditory and they put made' is basic in the integration of visual and auditory cues..." (p. 8). Visible speech appears to be informative for exactly those distinctions that are difficult to distinguish on the basis of auditory information. Although this observation is accurate and important, we should not be mislead into thinking that two sources are better than one because if the signal is not recognized via one source, it will be via the other. Similarly, complementarity is not necessary to achieve the benefits of integration. In speech perception, two sources are better than one because they are integrated or combined in an optimal manner (Massaro & Cohen, 1990). This observation has important implications for strategies in speech analysis and synthesis by machine. For analysis, it is important to integrate multiple sources of continuously-valued information. For synthesis, it might be advantageous to provide highly redundant cues across the two modalities.

Benoit et al. observe that vision is useless in optimal acoustic conditions. However, this point of view implies an inherent asymmetry between the two modalities. We might also argue that audition is useless in optimal visual conditions — even though normal visible speech does not provide sufficient information. Also, visual information might make a helpful contribution when processing load of the perceiver is increased — even though the acoustic information would normally be sufficient for accurate recognition.

The two modalities are be viewed as two sources of information, but there are also several sources of information within each of these modalities. An important question addressed by Brooke is the potential value of each of several sources. He only considers these sources one at a time, however, and the value of combining the sources is not addressed (see Brooke and Templeton, 1990).

Another approach in using multiple features for recognition of visual speech was developed by Finn (1986). She measured 14 parameters of the mouth, 30 times during a 1 second interval, for 23 consonants in a /aCa/ environment. The basic approach used in recognition was to consider each item as located in a multidimensional space and consider the distance of a new item from each possible prototype in memory. A reduced set of 5 parameters (derived from 3-D dots on the face) was found to provide optimal performance with differential weighting on the parameters. The parameters used were: 1) nose to center, outside border of upper lip, 2) chin to center, outside border of upper lip, 3) inner corner of open mouth to inner corner of rounded lips, 4) inner corner of rounded lips to center of lips, and 5) nose to chin. The performance of the system was 74% correct.

Benoit et al. also carried out a series of multidimensional analyses of their syllables from a single speaker. In addition to the features used by Brooke, they analysed protrusions of the lips (having available a three-dimensional recording) of the syllable. They found that accurate classification depended on not only the vertical separation of the lips and the internal areas encompassed by the lips, but also the width (protrusion) of the lips. These results provide valuable normative data on the visual characteristics of the segments of French.

Our proposal that multiple features must be analyzed is particularly germane because Brooke compares his limited feature approach to a connectionist model using a multi-layer perceptron (MLP). It is our contention that a fairer comparison would be the FLMP or Finn's (1986) analysis which utilize the multiple sources of information available. The input for Brooke's MLP was a compressed 16 by 12 pixel representation of the image. However, the MLP could also use the features as inputs. There are three features and given that there were 192 (16 x 12) input units in Brooke's image-based MLP, 64 units could be used to code each of the three features. This feature-based MLP would be directly comparable to the image-based MLP because the same number of units are used. Our suggestion of comparing the two MLPs is simply another instance of assessing template versus feature analysis. Of course, an important question is to determine which features, if any, are the most informative.

Benoit et al. provide some evidence that visual information about rounding can be processed by a human perceiver before auditory information is processed about the same contrast. This result illustrates that although sound and sight come from the same set of speech movements, the two sources of information are not necessarily equivalent, and can arrive at different times for example. (This independence of auditory and visual speech is apparent in a very different context. A hearing-impaired child at a given stage of speech development might be more intelligible if he or she simply mouths the speech rather than adding sound to it. The sound can be so inappropriate that it actually misleads the perceiver rather than providing a helpful source of information.) We have similar evidence based on the identification of auditory, visual, and bimodal syllables /ba/ and /da/ (Massaro 1987, Chapter 6). Subjects identified a visual /ba/ more quickly than an auditory /ba/, whereas they were slower in identifying a visual /da/ than an auditory /da/. Also of interest was that the subjects were faster to the bimodal syllables than to the faster of the unimodal syllables. This bimodal advantage was completely accounted for by the statistical advantage of having two sources of information relative to just one

REFERENCES

Benguerel, A. P., & Cowan, H. A. (1974). 'Coarticulation of upper lip protrusion in French', *Phonetica*, 30, 41-55.

Benguerel, A. P., & Pichora-Fuller, M. K. (1982). 'Coarticulation effects in lipreading', *Journal of Speech and Hearing Research*, 25, 600-607.

Bladon, R. A. (1977). 'Some control components of a speech production model'. Paper presented to the Congress of the International Society of Phonetic Sciences, Miami Beach, Florida.

Bladon, R. A., & Al-Bamerni, A. (1976). 'Coarticulation resistance of English /l/', *Journal of Phonetics*, 4, 135-150.

Breeuwer, M., & Plomp, R. (1985). 'Speech-reading supplemented with formant-frequency information for voiced speech', *Journal of the Acoustical Society of America*, 77, 314-317.

Brooke, N. M., & Templeton, P. D. (1990). 'Classification of lip-shapes and their association with acoustic speech events', *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, France, 245-248.

Brunswik, E. (1955). 'Representative design and probabilistic theory in a functional psychology', *Psychological Review*, 62, 193-217.

Cohen, M. M., & Massaro, D. W. (1990). 'Synthesis of visible speech', *Behavioral Research Methods and Instrumentation*, 22(2), 260-263.

Denes, P. (1955). 'Effect of duration on the perception of voicing', *Journal of the Acoustical Society of America*, 27, 761-764.

Derr, M. A., & Massaro, D. W. (1980). 'The contribution of vowel duration, F0 contour, and frication duration as cues to the /juz/-/jus/ distinction', *Perception & Psychophysics*, 27, 51-59.

Finn, K. E. (1986). 'An investigation of visible lip information to be used in automated speech recognition', *Ph.D. Dissertation*, Washington DC: Georgetown University.

Massaro, D. W. (1987). 'Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry', *Hillsdale, NJ: Lawrence Erlbaum Associates.*

Massaro, D. W., & Cohen, M. M. (1977). 'The contribution of voice-onset time and fundamental frequency as cues to the /zi/-/si/ distinction', *Perception & Psychophysics*, 22, 373-382.

Massaro, D. W., & Cohen, M. M. (1983). 'Consonant/vowel ratio: An improbable cue in speech', *Perception & Psychophysics*, 33, 502-505.

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

Owens, E., & Blazek, B. (1985). 'Visemes observed by hearing-impaired and normal-hearing adult viewers', *Journal of Speech and Hearing Research*, 28, 381-393.

Parke, F.I. (1974). 'A parametric model for human faces', *Technical Report UTEC-CSc-75-047* Salt Lake City: University of Utah.

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

Port, R. F., & Dalby, J. (1982). 'Consonant/vowel ratio as a cue for voicing in English', *Perception & Psychophysics*, 32, 141-152.

Sejnowski, T. J., & Rosenberg, C. R. (1986). 'NETtalk: A parallel network that learns to read aloud', *The John Hopkins University Electrical Engineering and Computer Science Technical Report*, JHU/EECS-86/01.

Finn, K. E. (1986). 'An investigation of visible lip information to be used in automated speech recognition', *Ph.D. Dissertation*, Washington DC: Georgetown University.

Massaro, D. W. (1987). 'Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry', *Hillsdale, NJ: Lawrence Erlbaum Associates*.

Massaro, D. W., & Cohen, M. M. (1977). 'The contribution of voice-onset time and fundamental frequency as cues to the /zi/-/si/ distinction', *Perception & Psychophysics*, 22, 373-382.

Massaro, D. W., & Cohen, M. M. (1983). "Consonant/vowel ratio: An improbable cue in speech', *Perception & Psychophysics*, 33, 501-505.

Massaro, D. W., & Cohen, M. M. (1990). "Perception of synthesized audible and visible speech', *Psychological Science*, 1, 55-63.

Owens, E., & Blazek, B. (1985). 'Visemes observed by hearing-impaired and normal-hearing adult viewers', *Journal of Speech and Hearing Research*, 28, 381-393.

Parke, F.I. (1974). 'A parametric model for human faces', *Technical Report UTEC-CSc-75-047* Salt Lake City: University of Utah

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

Port, R. F., & Dalby, J. (1982). 'Consonant/vowel ratio as a cue for voicing in English', *Perception & Psychophysics*, 32, 141-152.

Sejnowski, T. J., & Rosenberg, C. R. (1986). 'NETtalk: A parallel network that learns to read aloud', *The Johns Hopkins University Electrical Engineering and Computer Science Technical Report*, JHU/EECS-86/01.