Speech by eye

Dorothy W. Massaro and Michael M. Cohen

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

Like our colleagues Renoir, Laouache, Mohamadi, and Alay (this volume) and Brodie (this volume), we also believe that synthetic visible speech has obvious potential for advancing our knowledge about the visible information in speech perception and how it is utilized by human processors. We are using a high quality facial display originally developed by Fiske (1974) and recently augmented by Lippert, Wexler, and Wexler (1980), and ourselves (Cohen & Massaro, 1980). In this display the face is constructed of about 900 3D-polygons controlled by about 50 parameters. There is now some notion of visible speech, but a better model of speech articulation is being developed incorporating physical measurements from real speech and rules describing articulation 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 face, and more natural speech will be elicited to allow their systematic variation with neighboring speech segments. Psychophysical studies regarding confusion matrices and standard tests of intelligibility will be used to assess the quality of the facial articulation. 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 synthesis has proved 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 are not found when used using natural speech. (Massaro, unpublished). Synthetic speech also permits the implementation of experiments designed to test theoretical hypotheses, such as word order, even though critical for various speech distinctions. The applied value of 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 listener continuum using synthetic visible speech, we found very similar results for the endpoint stimuli, but are few compared to earlier studies using natural speech (e.g., Massaro, 1977; Massaro & Cohen, 1980). 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 severe alternatives gives a more powerful measure of integration relative to the case of transparent natural stimuli. It is also obvious that a syllable or phrase speech will have a volume and rate, which is also a disadvantage of both.

The guiding principle 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 cognition. Integration appears to occur to some extent regardless of the goals and mental ends of the perception. Burnham (1955) acknowledged this, but his ambiguous sources of information on behavior and its influence on the physical world may not be used and hence not functionally valid. Thus it is an empirical question whether the 21 classes Bennett et al. obtained from multidimensional scaling are psychologically real. It is important to determine the ecologically valid cues in the analysis of the articulation of human observers and the perception of speech by integrating multiple sources of information.

We have developed an experimental paradigm to determine which of the many potential 12-talker cues are actually used by human observers. Identification experiments carried out with the systematic variation of properties of the speech sign, combined with the qualitative test of models on different sources of information, enable the investigator to test the psychological validity of different cues. For example, in a walking situation allow us to perceive a difference between the voice in the phrase to a cue and the noise in the phrase the cue. For more than a decade, speech scientists believed that consonant duration relative to vowel duration was the critical cue to the walking judgments (Burnham, 1955; Polk & Slattery, 1982). However, Massaro and Cohen (1977, 1985) showed that this cue, as measured by the duration ratio, was not significant when the results were analyzed in the framework of the fuzzy logic model of perception (FLMP) (Massaro, 1984). A model based on OV ratio gives a much better description of the identification results than does the assumption of independent consonant and vowel duration cues (Dem & 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 limitations, or masking. As an example, the perception of speech sentences that have been bandwidth-filtered improves from 92% to 78% normal when subjects are permuted a view of the speaker (Breevan & Plonski, 1982). The 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 perceivers' recognition of an auditory-only syllable reflects the contribution of both.

Sweeter by eye...
The process of speech comprehension is complex and involves multiple sensory modalities. In this context, the integration of visual and auditory information is crucial. The literature on speech perception has shown that the visual modality can provide important information that is not available from the auditory channel alone. For example, visual cues can help to resolve ambiguities in speech, improve the intelligibility of speech, and provide information about the speaker's identity.

In the case of voiceless speech sounds, visual information can be particularly useful. The lips and jaw movements associated with these sounds can be seen, and this information can be used to improve speech perception. Similarly, visual information about the position of the tongue and the larynx can be used to infer the identity of the speaker and the context of the speech.

It has been suggested that the integration of visual and auditory information is particularly important in the case of voiceless speech sounds. This is because the visual cues can help to resolve ambiguities in the auditory channel and improve the intelligibility of the speech. In some cases, the visual cues can even provide information that is not available from the auditory channel alone.

In conclusion, the integration of visual and auditory information is crucial for speech comprehension. The visual modality can provide important information that is not available from the auditory channel alone. This is particularly true in the case of voiceless speech sounds, where visual cues can be used to improve speech perception and resolve ambiguities in the auditory channel.

REFERENCES


