AUDITORY/VISUAL SPEECH IN MULTIMODAL HUMAN INTERFACES

Dominic W. Massaro and Michael M. Cohen
{massaro,mmcohen}@fuzzy.ucsc.edu

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

ABSTRACT

It has long been a hope, expectation, and prediction that speech would be the primary medium of communication between humans and machines. To date, this dream has not been realized. We predict that exploiting the multimodal nature of spoken language will facilitate the use of this medium. We begin our paper with a general framework for the analysis of speech recognition by humans and a theoretical model. We then present a system for auditory/visual speech synthesis that performs complete text-to-speech synthesis. This system should improve the quality as well as the attractiveness of speech as one of a machine's primary output communication medium. Mirroring the value of multimodal speech synthesis, multimodal channels should also enhance speech recognition by machine.

1. INTRODUCTION

Speech perception is a human skill that rivals our other impressive achievements. Even after decades of intense effort, speech recognition by machine remains far inferior to human performance. Our thesis is that 1) there are multiple sources of information supporting speech perception, 2) the perceiver evaluates each source in parallel with all of the others, and 3) all of these sources are combined or integrated to achieve perceptual recognition. Recognition of a word in a sentence is achieved via a variety of bottom-up and top-down sources of information. Top-down sources include contextual, semantic, syntactic, and phonological constraints; bottom-up sources include audible and visible features of the spoken word.

Our research is carried out within a framework of a fuzzy logical model of perception (FLMP) in which speech perception is viewed as having available multiple sources of information supporting the identification and interpretation of the language input. The assumptions central to the model are 1) each source of information is evaluated to give the degree to which that source specifies various alternatives, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated to provide an overall degree of support for each alternative, and 4) perceptual identification and interpretation follows the relative degree of support among the alternatives.

This research paradigm permits us to determine which of the many potentially functional cues are actually used [1]. 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. This paradigm has already proven to be effective in the study of audible, visible, and bimodal speech perception [1,2]. Thus, our research strategy not only addresses how different sources of information are evaluated and integrated, it can uncover what sources of information are actually used. We believe that the research paradigm confronts both the important psychophysical question of the nature of information and the process question of how the information is transformed and mapped into behavior.

2. SPEECH BY EYE AS WELL AS BY EAR

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 [1].

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 /da/, subjects often perceive the speaker to be saying /ða/ [3].

A main objective of our research is to identify the facial properties that are informative by evaluating the effectiveness of various properties in a synthetic animated face. Analogous to the valuable contribution of using auditory speech synthesis in speech perception research, visible speech synthesis permits the type of experimentation necessary to determine 1) what properties of visible speech are used, 2) how they are processed, and 3) how this information is integrated with auditory information and other contextual sources of information in speech perception. This experimental and theoretical framework has already established several facts concerning speech by eye and ear [4]. Synthetic speech has been central to our inquiry and we now describe our continued development and utilization of a realistic, high-quality, facial display.

3. VISIBLE SPEECH SYNTHESIS

Our visual synthesis software is a direct descendant of Parke's [5,6,7] parametrically controlled polygon topology synthesis technique, incorporating later improvements [8,9,10,11]. The facial synthesis is carried out on a Silicon Graphics Inc (SGI) 4D/CRIMSON-VGX workstation which allows real time manipulation of parameters and real time production of stimuli for perceptual experiments. The development system and graphical user interface (GUI) for visual speech synthesis has been described by Cohen and Massaro [10,11]. A smaller version of the visual speech software with the same functionality but without the GUI is available for use under f77 main programs for perceptual experiments including the presentation of auditory speech and collection of responses from human participants.

Figure 1. Graphical user interface for face development. Master panel in lower right has facial parameter controls, facilities for editing speech segment definitions, sentence input, speaking rate, parameter tracking, call-ups for subsidiary control panels and other miscellaneous controls. Upper right panel is text interface. Lower left panel is display output. Upper left is play control with cursors for zooming and moving face in time, and plots of control parameters (bottom), dominance functions (middle) and derived lip measures (top).

Figure 1 shows the GUI for face development. The master panel in the lower right of the screen has facial parameter controls, facilities for editing speech segment definitions, sentence input, speaking rate, parameter tracking, call-ups for subsidiary control panels and other miscellaneous controls. The facial parameter controls in the lower right quadrant are color coded by function: blue for viewpoint, red for facial shape, and yellow for those involved related to speech segment definitions. The upper right panel is a text based interface which can control the face using alphanumeric commands or files of such commands. Also in the upper right of the screen

is a menu panel for the selection of members of a set of tokens for synthesis. In this example, the menu is set to call one of 27 CV syllables whose definitions have been read in from a file. The lower left panel is the display output. This area can also be output in NTSC video. The upper left area contains the play controls with cursors for temporal zooming and displaying the face forward and backward in time, and plots of control parameters (bottom), dominance functions (middle) and derived facial measures (top). Alternate displays for the top position (not shown in the figure) include plots of digitized audio signals and speech spectrograms. The controls in this region also allow modification of segment durations which is especially useful for synchronization with natural speech. Other control panels not shown in Figure 1 include control modules for text-to-speech conversion, emotional display, sound synchronization, simultaneous laser-videodisc control, video recording, and two-dimensional parameter editing and path tracking.

An important new addition to the facial model is a tongue which has been implemented as a shaded surface made of a polygon mesh, controlled by several parameters: tongue length, angle, width, and thickness. This tongue model is a considerable simplification compared to a real tongue which has several more degrees of freedom, but it contributes a great deal to visual speech and can be computed very quickly.

In addition to the tongue control parameters, a number of other new (relative to the earlier Parke models) parameters are used in speech control, including parameters to raise the lower lip, roll the lower lip, and translate the jaw forward and backward. Some parameters have been modified to have more global effects on the synthetic talker's face than in the original Parke model. For example, as the lips are protruded the cheeks pull inward somewhat. Another example is that raising the upper lip also raises some area of the face above.

An important improvement in our visual speech synthesis software has been the development of an new algorithm for articulator control which takes into account the phenomenon of coarticulation [10]. Coarticulation refers to changes in the articulation of a speech segment depending on preceding (backward coarticulation) and upcoming segments (forward coarticulation). An example of backward coarticulation is the difference in articulation of a final consonant in a word depending on the preceding vowel, e.g. boot vs beet. An example of forward coarticulation is the anticipatory lip rounding at the beginning of the word "stew". The substantial improvement of more recent auditory speech synthesizers, such as MITalk [12] and DECtalk [13], over the previous generation of synthesizers such as VOTRAX [14], is partly due to the inclusion of rules specifying the coarticulation among neighboring phonemes.

Our approach to the synthesis of coarticulated speech is based on the articulatory gesture model of Löfqvist [15]. A speech segment has dominance over the vocal articulators which increases and then decreases over time during articulation. Adjacent segments will have overlapping dominance functions which leads to a blending over time of the articulatory commands related to these segments. Given that articulation of a segment is implemented by several articulators, there is a dominance function for each articulator. The different articulatory dominance functions can differ in time offset, duration, and magnitude. Different time offsets, for example, between lip and glottal gestures could capture differences in voicing. The magnitude of each function can capture the relative importance of a characteristic for a segment. For example, a consonant could have a low dominance on lip rounding which would allow the intrusion of values of that characteristic from adjacent vowels. The variable and varying degree of dominance in this approach naturally captures the continuous nature of articulator positioning. This model, as implemented, provides the total guidance of the facial articulators for speech rather than simply modulating some other algorithm to correct for coarticulation. To instantiate this model it is necessary to select particular dominance and blending functions [10]. Figure 2 shows an example of our coarticulatory synthesis approach for the word "stew". The top panel of the figure shows the dominance functions for lip-protrusion for each segment. As can be seen, the consonants /s/ and /t/ have very low dominance versus the strong and temporally wide dominance function for the vowel /u/. The middle panel shows the lip-protrusion parameter over

Figure 2. Dominance functions (top panel) and parameter control functions (middle panel) for lip protrusion for the word "stew". Bottom panel shows facial display during the /s/ segment.

time. We see in this panel that protrusion comes close to its target value (indicated by the diamond) for /u/. Because of the strong dominance of the vowel, this protrusion value spreads through the preceding /s/ and /t/. The bottom panel of the figure illustrates the resulting anticipatory protrusion during the /s/ segment, with a side view of the face.

4. BIMODAL SPEECH SYNTHESIS

We have developed a multimodal speech synthesis system, building on existing auditory speech synthesis and our visible speech synthesis. Given the complexity of the high level linguistic and phonetic algorithms involved it would be a difficult task to simply attempt to synchronize the visual synthesis with a commercial product like DECtalk [13]. Our approach is to use the same higher level software to translate English text into the required segment, stress, and duration information to drive both the visual and auditory synthesis modules. We have adapted the MITalk [12] software for this higher level analysis. While the MITalk module carries out the high level analysis, separate lower level modules exist for the auditory and visual synthesis. An alternative method would be to specify all of the auditory and visual parameters controlling the two synthesizers within the same parameter space for each phoneme segment, though this would be more applicable with articulatory auditory synthesis as opposed to the terminal analog synthesis now used, since there is no one-to-one mapping of articulatory and acoustic parameters.

While initial performance of our synthesis strategy is encouraging, some problems remain to be totally solved regarding the exact synchronization of events even with the same durations for the visual and auditory segments. For example, auditory voice onset in /b/ occurs at the appropriate time relative to mouth opening. An important source of variability in speech comes from differences in speaking rate. Given the complex changes which occur with changes in speaking rate [16], we plan to analyze our synthetic visual speech to assess whether our algorithm correctly represents these changes. In our text-to-speech translation algorithm, we now

alter segment duration based on the stress from the phonological, morphological, and syntactic analysis. These segment durations also are used in the visual synthesis, but we do not yet use the stress to alter the intensity of the visual articulation. One approach that we have used is to vary the power of the dominance functions controlling articulation. With lower power, articulation is more gradual and does not come as close to achieving the segment targets. Finally, we plan to refine paralinguistics in our synthetic visual and auditory speech. Currently, we utilize word and phrase boundary information from the MITalk module to control blinking, eye movements, and head nodding. We plan additional work to convey other paralinguistic information such as emotion.

Because lipreaders are faced with a variety of talkers, both in training and in everyday communication, it is important to both consider how this variability affects perception and to model this variability in our visual speech synthesis. Regarding perception, the use of a variety of talkers is necessary to achieve a true picture of which cues are used. Concentration on too small a sample of natural (or synthetic) faces can lead to a lack of generality and ecological validity. Regarding visual speech synthesis, one should be able to simulate the same variety of talkers that lipreaders face in the real world. In order to create a variety of faces, we have used texture mapping from real faces in conjunction with adjustment of facial shape parameters for best fit to the source faces. For example, the size of the head, the relative sizes of the jaw, cheeks, neck, and nose, the relative heights of mouth, nose, eyes and forehead in the face can all be adjusted. Of special interest for lipreading, we can adjust characteristics such a overall mouth size and lip thickness. Figure 3 illustrates a sample of our texture mapped faces. These illustrate differences in sex, race, build, and facial hair. Although we can simulate visible appearance with texture mapping, a challenging goal of simulating individual voices remains.

Figure 3. Examples of texture mapping used to create a variety of faces. Note the differences in head shape, facial hair, mouth size, and lip thickness.

5. SUMMARY

Although our inquiry addresses different problem domains in cognitive science, speech science, and engineering, their simultaneous study affords potential developments not feasible in separate investigations. For example, the study of how humans perceive visible speech is critically dependent on manipulating synthetic visible speech. Development of an adequate synthesis system, however, must be assessed against human production and perception of speech. Our investigation seeks to understand articulation of speech and its acoustic and visual consequences in order to develop realistic speech synthesis. In addition, we continually test the psychological reality of our synthesis by studies of human perception of the synthetic speech and to comparisons to natural speech. The general hypotheses of this research are that 1) a synthetic

talker is an important challenge to computer animation and offers a potentially valuable medium for communication among both normal and disabled individuals, human-computer interaction, and virtual worlds. 2) bimodal synthetic speech provides a valuable experimental tool for our understanding of speech perception by ear and by eye, 3) visual speech information offers an additional source of information for both normal and hearing-impaired individuals, and 4) the research has immediate and direct application to improving the communication alternatives for deaf and hearing-impaired individuals.

6. REFERENCES

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

[2] Massaro, D. W. (1989) A *precis* of *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. *Behavioral and Brain Sciences, 12*, 741-794.

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

[4] Massaro, D. W. (1994) Psychological aspects of speech perception: Implications for research and theory. In M. Gernsbacher (Ed.) *Handbook of Psycholinguistics*. New York: Academic Press.

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

[6] Parke, F.I. (1975) A model for human faces that allows speech synchronized animation. *Journal of Computers and Graphics, 1(1),* 1-4.

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

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

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

[10] Cohen, M. M., & Massaro, D. W. (1993) Modeling coarticulation in synthetic visual speech. In M. Thalmann & D. Thalmann (Eds.) *Computer Animation '93.* Tokyo: Springer-Verlag.

[11] Cohen, M. M., & Massaro, D. W. (1994) Development and Experimentation with Synthetic Visible Speech *Behavioral Research Methods and Instrumentation, 26*, 260-265. press.

[12] Allen, J., Hunnicutt, M. S., and Klatt, D. (1987) *From text to speech: The MITalk system* Cambridge, MA: Cambridge University Press.

[13] DECtalk (1985). *Programmers Reference Manual*. Maynard, MA: Digital Equipment Corporation.

[14] VOTRAX (1981) *User's Manual* Votrax, Div. of Federal Screw Works.

[15] Löfqvist, A. (1990) Speech as audible gestures. In W.J. Hardcastle and A. Marchal (Eds.) *Speech Production and Speech Modeling*. Dordrecht: Kluwer Academic Publishers, 289-322.

[16] Klatt, D. (1979) Synthesis by rule of segmental durations in English sentences. in B. Lindblom and S. Ohman (Eds.) *Frontiers of Speech Communication Research.* London: Academic Press.

7. ACKNOWLEDGMENT

The research reported in this paper and the writing of the paper were supported, in part, by a grant from the Public Health Service (PHS R01 NS 20314).