13

Section Introduction. Talking Heads in Speech Synthesis

Dominic W. Massaro
Michael M. Cohen

This book documents that indeed "Progress in speech synthesis" is indeed being made. Just a little experience with the requirements of speech synthesis converts even the most optimistic to the realization of the tremendous endeavor that is required. Both a highly interdisciplinary approach and almost an unlimited supply of technological and human resources are necessary for starters. We can also expect that progress, although cumulative, will necessarily be gradual and too slow for many of us. We applaud the authors for their significant contributions and look forward to the continued progress of their promising research programs.

Beckman's tour de force in chapter 15 sensitizes us to the intricate relationship between basic research in speech science and speech synthesis technology. She illustrates this productive interplay of pure research and applied implementation in the case of intonation synthesis. It will be a challenge for our text-to-speech systems to arrive at the appropriate interpretation of "John doesn't drink because he's unhappy" given an analysis of the text semantics. And, of course, once we arrive at the proper interpretation, work remains to be done on both the proper acoustic [Cah90] and visual [Pel91] synthesis. She also demonstrates that the dynamics of speech articulation are better captured by higher-order properties such as timing rather than lower-order properties such as duration.

Beckman's call to more global descriptions is consistent with the framework described by Bickley, Stevens, and Williams in chapter 16. They describe procedures for the synthesis of segmental information on the basis of high-level parameters. This framework offers a productive compromise between terminal analog and articulatory speech synthesis because many of the high-level parameters are articulatory in nature: area of lip opening, average glottal area, and so forth. These high-level parameters are mapped via a set of equations into the lower-level parameters that actually control the synthesizer. Their examples of this techniques instills a degree of appreciation for this approach. We look forward to learning

more about the positive aspects of this technique as well as an objective measure of the quality of the synthesis that results.

In chapter 17, Wilhelms-Tricarico and Perkell ambitiously attack the problem of biomechanical and physiologically based speech modeling. In particular, they appear to have successfully modeled the tongue and related control processes. It is heartening to see attention in this area (see also [Pel91]), given its importance as a cue for visual speech. In terms of the control processes, we hope to see further work to determine to what extent these processes are functionally hierarchical.

Perhaps one of the most significant developments in speech synthesis has been reuniting the free-floating voice with a talking head. The history of speech research and application has reasonably viewed speech as an auditory phenomenon. If a voice (or a speech synthesizer) speaks in the forest with no audience, is there speech? Our claim is that there isn't without a talking head to accompany it. More seriously, including a talking head to text-to-speech synthesizers offers the potential of a dramatic improvement in realistic synthesis, synthesis intelligibility, and end-user acceptability.

Perceptual scientists have documented that our sensory interactions in the world are seldom via a single modality. Rather, our experience is grounded in a multisensory interplay of all of our senses with a rich set of environmental dimensions. This multidimensional scenario also exists in language communication. Psychological experiments have revealed conclusively that our perception and understanding are influenced by the visible speech in the speaker's face and the accompanying gestural actions. These experiments have shown that the speaker's face is particularly helpful when the auditory speech is degraded due to noise, bandwidth filtering, or hearing impairment [Mas87, Sum91]. Although the influence of visible speech is substantial when auditory speech is degraded, visible speech also contributes to performance, even when paired with intelligible speech sounds. The importance of visible speech is most directly observed when conflicting visible speech is presented with intelligible auditory speech. One famous example resulted from the dubbing of the auditory syllable /ba/ onto a videotape of a talker saying /ga/. A strong effect of the visible speech is observed because a person will often report perceiving (or even hearing) the syllable /da/, /va/, or /ð/, but seldom /ba/ corresponding to the actual auditory stimulus.

A peculiar characteristic of bimodal speech is the complementarity of audible and visible speech. Visible speech is usually most informative for just those distinctions that are most ambiguous auditorily. For example, place of articulation (such as the difference between /b/ and /d/) are difficult via sound but easy via sight. Voicing, on the other hand, is difficult to see visually but is easy to resolve via sound. Thus, audible and visible speech not only provide two independent sources of information, these two sources are often productively complementary. Each is strong when the other is weak.

In addition to its applied value, synthetic speech has been central to the study of speech perception by human observers. 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. Synthetic speech also permits the implementation and test of theoretical hypotheses, such as which cues are critical for various speech distinctions.

It is believed that visible synthetic speech would prove to have the same value as audible synthetic speech. Synthetic visible speech could provide a more fine-grained assessment of psychophysical and psychological questions not possible with natural speech. For example, testing people with synthesized syllables intermediate between several alternatives gives a more powerful measure of integration relative to the case of unambiguous natural stimuli. In chapter 18, Le Goff, Guiard-Marigny, and Benoit address the question of which aspects of the speaking face are informative. Using video analysis of a face with painted lips, the authors are able to track the lips and to control the lips of a wire-frame model, first developed by Parke [Par74] and made more realistic in our laboratory. Perceivers were tested with just auditory speech under varying levels of white noise, or with the addition of a natural face, just synthetic lips, or the complete synthetic face (without the tongue).

They found a dramatic improvement in intelligibility with the addition of visual information. Moving lips help, the addition of the synthetic face helps more, and the natural face helps even more. The synthetic face had no tongue, and a future experimental question should be how close the synthetic face will be to a real face once a tongue is added. Guiard-Marigny, Adjoudani, and Benoit added a 3D model of the jaw and found some improvement in intelligibility relative to the synthetic lips alone, as shown in chapter 19. The jaw is controlled by a single data point corresponding to a dot on the speaker's chin. However, this improvement was not quite as large as that provided by the complete synthetic head in chapter 18. Future research comparing the jaw and the synthetic head will be of important value to the development of visible speech synthesis.

Our research indicates that, like audible speech, visible speech still does not duplicate the informative aspects of a real talking face. At this stage, it is difficult to predict the trajectory of visible speech synthesis. One issue might be whether articulatory or terminal analog synthesis offers the greatest potential. Adding the dimension of visible speech might provide a boost for articulatory synthesis because the articulators only have to be made visible to include this modality as input for the perceiver, rather than needing the additional step of a transformation to the acoustic signal, a process that has not yet been totally solved. On the other hand, with terminal analog synthesis, the investigators can concentrate on achieving a realistic animation without worrying about the physical hardware of living talkers.

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. 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.

One applied value of visible speech is its potential to supplement other (degraded) sources of information. Visible speech is particularly beneficial in poor listening environments with substantial amounts of background noise. Its use is also important for hearing-impaired individuals because it allows effective spoken communication—the universal language of the community. Just as auditory speech synthesis has proved a boon to our visually impaired citizens in human machine interaction, visual speech synthesis should prove to be valuable for the hearing-impaired. Finally, synthetic visible speech had an important part of building synthetic "actors" [TT92] and played a valuable role in the exciting new sphere of virtual reality. We predict that the most progress in speech perception will be seen (no pun intended) in the continued refinement of artificial talking heads.

REFERENCES

[Cah90] J. E. Cahn. *Generating Expression in Synthesized Speech.* MIT Media Lab Technical Report (revision of 1989 MS thesis), 1990.

[Mas87] D. W. Massaro. Speech perception by ear and eye: A paradigm for psychological inquiry. Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

[Par74] F. I. Parke. *A parametric model for human faces.* University of Utah Technical Report UTEC-CSc-75-047, 1974.

[Pel91] C. Pelachaud. *Communication and Coarticulation in Facial Animation.* University of Pennsylvania, Dept. of Computer and Information Science, Report MS-CIS-91-77, 1991.

[Sum91] Q. Summerfield. Visual perception of phonetic gestures. In *Modularity and the Motor Theory of Speech Perception*, I. G. Mattingly and M. Studdert-Kennedy, eds. Lawrence Erlbaum Associates, Hillsdale, NJ, 117–137, 1991.

[TT92] N. Thalmann and D. Thalmann. *Creating and Animating the Virtual World.* Springer-Verlag, Tokyo, 1992.

14

Section Introduction. Articulatory Synthesis and Visual Speech

Juergen Schroeter

14.1 Bridging the Gap Between Speech Science and Speech Applications

When asked to write an overview of a section of a book, one is faced with the problem of what service to provide besides the obvious attempt at gluing the papers in the section together and tying them to the other sections of the book. One option is to summarize current approaches to synthesis and "hot" issues being attacked in research. Eric Moulines [Mou92] satisfied this option in an excellent way in the predecessor of this book. The fact that there is little to add to Moulines's summary allows me to focus on a more specific, yet interdisciplinary issue: I have chosen to foster closer ties between basic research in speech production and applied research in speech synthesis. I will do this by highlighting problems in speech synthesis in need of solution, as well as pointing out recent findings in speech production research that are likely to impact speech synthesis. My approach will be somewhat analogous to the one of our contribution to the special session about the role of speech production in speech recognition at the June 1994 meeting of the Acoustical Society of America in Boston (written up [RSS95]).

Consider the history of research in speech synthesis from the early days of "Pedro the Voder" demonstrated at the 1939 World's Fair in New York City, over the many contributions made at Haskins Laboratories starting with the Pattern Playback machine built by Frank Cooper (e.g., [Mat74, Lib93]), and the famous book by Fant [Fan60], to the vast contributions by Klatt (see, e.g., [Ste92, Kla87]). From these and other important contributions, it seems obvious that speech synthesis and research in speech production are closely related: if we try to make a machine talk like a human, we had better know something about human speech production. Also, to understand how humans produce speech, it might be useful to test our models by letting them synthesize speech (or certain attributes of speech), and to compare the output of our models to what we measure in natural speech. These are the lines of thought in Mary Beckman's chapter, entitled "Speech Models and Speech Synthesis."