Book Review

Principles of Computer Speech

Ian H. Witten

Academic Press, London, 1982.
ISBN 0-12-760670-9

To understand the field of computer speech fully requires an interdisciplinary approach which probably spans acoustics, computer science, electrical engineering, linguistics, phonetics, phonology, and psychology. The most one can hope for from a single author is a comprehensive overview with pointers to sources with more depth and a clear presentation of some fundamental issues. These tasks are made much more difficult by a field that is literally exploding. On the whole, I think the author of *Principles of Computer Speech,* Ian H. Witten, has done a very credible job of reviewing the state of the art in 1980. His treatment is somewhat uneven and (obviously!) does not cover some important work done since the book was written. This review outlines a few criticisms, but mainly it attempts to help the reader of *Principles of Computer Speech* know where to look to fill in the gaps. On the whole, I do recommend the book as an introduction to the topic. It does cover most of the relevant fields and is clearly written.

The author's expertise seems to be more in the area of engineering than linguistics and more in linguistics than in psychology. The reader needs to keep this and a few consequent corollaries in mind while reading the book. For instance, the author (as is common practice!) talks about algorithms whose purpose is to determine fundamental frequency as "pitch trackers". This ***is*** common parlance, but is nonetheless misleading. Pitch is psychological; that is, a perceptual quantity that is primarily related to fundamental frequency but is also related to a number of other things.

One example of the sometimes tenuous relationship between pitch and fundamental frequency may illustrate the point. Because of the anatomical linkage between the tongue and the pharynx, high vowels like "ee" in "beet" and "uu" in "dude" are more easily made with higher fundamental frequency, at least in English. If one does ***not*** make this correction in synthetic speech, "ee" and "uu" may sound "lower pitched" than an "ah" vowel with exactly the same fundamental frequency. Thus, pitch (a perceptual event) depends upon expectations based on experience. One could easily infer from *Principles of Computer Speech* that pitch and F0 are synonymous.

In the index, we find the curious entry "Loudness, See Amplitude." Here again, it seems, there is some confusion. Amplitude is a physical attribute that is the main correlate of perceived loudness, but by no means the only one, particularly for speech. At least there are the effects of frequency, contrast, and expectation. These effects are more than laboratory demonstrations with strange stimuli. They play a part in the perception of synthetic (and natural) speech. For example, there are typical relations in amplitude between various phonemes; for example, the "ah" vowel in "father" is typically of greater amplitude than the "u" in "soon". If one matches these vowels in amplitude, the "u" in "soon" will sound much too loud.

While there is a chapter devoted to articulation and how sounds are produced, the real point of synthetic speech is to obtain sounds that are *perceived* as speech. In order to do this, it helps to have some understanding of how the ear-brain system works. This forms a foundation for deciding on the proper granularity of various parameters, although some trial and error will still be required. The problem may be deeper than simply not having a special chapter on hearing in the book. I believe that there is an underlying assumption here that mathematical formulae are a sufficient basis for making decisions about how to produce synthetic speech.

This philosophy is common in engineering circles and was perhaps best (worst) typified at a recent ICASSP meeting where someone gave a paper that "proved" that their noise reduction technique was superior to the Weiner technique even though (at least) this listener could clearly make out the speech whose noise had been damped by a Weiner technique while the author's technique made the speech totally unintelligible!!

We would all like to advance science to the point where mathematical models would be better representations of what actually happens in perception, but the fact is that we are not yet there. Where psychophysics has advanced and we have better representations of the relationship between objective and subjective scales (as in critical bands and the bark scale), the author has failed to inform the reader of these advances.

While the prose parts of the text are very clearly written, I find the diagrams and mathematics somewhat harder to follow. The mathematical treatments are not in depth enough for me to simply "use" the formulae. Nor are there enough steps in most of the derivations to allow me to follow the logic closely. To someone with the right engineering background, they may well be useful as reminders of well-known results. I think that for the typical linguistics or computer science reader, even with a few years of calculus, the

treatment will not be sufficient to allow them to gain any real insight into the theory over and above what is presented in the (clearly written) text.

While the book offers a good discussion of many of the major issues in speech coding and synthesis, the reader should supplement his/her study by referring to more recent works. For instance, a number of commercial synthesizers that do a fairly credible job of unlimited text-to-speech conversion have come on the market since the book was written. These are based primarily on the work of Gunnar Fant and/or Dennis Klatt. The interested reader can typically see demonstrations of such devices at meetings of AVIOS (American Voice Input Output Society), ASA (Acoustical Society of America), or ICASSP (International Conference on Acoustics, Speech, and Signal Processing).

In addition, some advances have been made in speech coding techniques, including Time Domain Harmonic Scaling and "Multipulse" (Malah 1979; Atal 1983; Singhal and Atal 1984). Applications of speech synthesis are expanding also. Speech synthesis is being used in special telephones, more widely in manufacturing, for remote sensing, and for access to computers via touch-tone phones.

A number of relevant human factors studies have also been carried out very recently. The development of IBM's Audio Distribution System was heavily influenced by human factor considerations (Richards and Boies 1981; Thomas 1983; Gould and Boies 1984). In addition, numerous studies on the perception of synthetic speech have been carried out in David Pisoni's lab at Indiana (Pisoni 1982; Nusbaum, Schwab, and Pisoni 1984) as well as in other places (Jenkins and Franklin 1982; McPeters and Tharp 1984; Thomas, Rosson, Chodorow, Kluender, and Carr 1984).

There appear to be very large individual differences in the ability to perceive synthetic speech. The intelligibility also varies tremendously as a function of how predictable the materials are. (A fact that demonstrators of synthetic speech materials use to their advantage!!) In addition, it appears that even when synthetic speech is perceived correctly, it requires more attentional demands than natural speech. Adaptation effects are quite substantial. There are both specific and general learning effects. All these factors mean that if one is interested in using synthetic speech in a real application, a little "demo" of the application that seems nice to the developers falls far short of the testing required to see whether or not the application will be useful in the real world.

John Thomas
IBM Thomas J. Watson Research Center

References

Atal, B.S. 1983 Efficient Coding of LPC Parameters by Temporal Decomposition. *IEEE Proceedings of ICASSP Meeting:* 81-84.

Gould, J.D. and Boies, S.J. 1984 Speech Filing – An Office System for Principals. *IBM Systems Journal* 23(1): 65-81.

Jenkins, J.J. and Franklin, L.D. 1982 Recall of Passages of Synthetic Speech. *Bulletin of the Psychonomic Society* 20(4): 203-206.

Malah, D. 1979 Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-27(2): 121-133.

Nusbaum, H.C.; Schwab, E.C.; and Pisoni, D.B. 1983 Perceptual Evaluation of Synthetic Speech: Some Constraints on the Use of Voice Response Systems. Presented at the American Voice Input Output Society, Chicago, Illinois (September).

Peters, D.L. and Tharp, A.L. 1984 The Influence of Rule-Generated Stress on Computer-Synthesized Speech. *International Journal of Man-Machine Studies* 20: 215-226.

Pisoni, D.B. 1982 Perception of Speech: The Human Listener as a Cognitive Interface. *Speech Technology* 1(2): 10-23.

Richards, J.T. and Boies, S.J. 1981 The IBM Audio Distribution System. *Proceedings, IEEE MIDCON Conference,* Chicago, Illinois. IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854.

Singhal, S. and Atal, B.S. 1984 Improving Performance of Multipulse LPC Coders at Low Bit Rates. *Proceeding, IEEE ICASSP Meeting:* 1.3.1–1.3.4.

Thomas, J.C. 1983 Office Communication Studies: Effects of Communication Behavior on the Perception of Described Persons. *Office Systems Research Journal* 1(2): 75-88.

Thomas, J.C.; Rosson, M.G.; Chodorow, M.; Kluender, K.; and Carr, T. 1984 Human Factors and Synthetic Speech. Presented at INTERACT '84, London, England (September 7).