THE SYMBIOSIS OF DSP AND SPEECH RECOGNITION
OR
AN OUTSIDER'S VIEW OF THE INSIDE

James F. Kaiser

Duke University
Department of Electrical and Computer Engineering
Durham, NC 27708-0291

ABSTRACT

From an historical review of how we got to where we are now, we discuss the interrelationship between our system design objectives and goals, our modeling of the speech signal and its generation and parameterization, and the broadly developing DSP methodology. We take a critical look at some of the underlying assumptions in our modeling to see if they may be limiting the performance that can be obtained with ASR systems. We close with some open questions and challenges for new work.

1. INTRODUCTION

Automatic speech recognition systems have now become an integral part of our everyday lives. Intense effort is being spent by engineers and scientists worldwide to continuously improve the performance and extend the range of these systems. An assessment of where we are and descriptions of new ways to improve these systems is the main focus of this workshop. It is therefore worthwhile for us to take a step to the side and look very broadly at how we got to where we are now and what fundamental assumptions are implied.

Mankind for many hundreds of years has been dependent on and fascinated by speech as the primary means of communication between individuals. Flanagan [1] gives a very readable account of the early work and foundation of speech analysis and synthesis. To enable communication at great distances the development of the telephone became an important necessity. From the early 1920's on research and development of the telephone was actively carried out in the engineering laboratories of AT&T, later becoming Bell Telephone Laboratories.

The extensive work in these early years at AT&T is described in detail by Harvey Fletcher in his book [2] which merits serious study even today for its breadth and depth of coverage of the many aspects of speech production and the mechanics of hearing. Here one can see clearly how the research emphasis is concentrated on trying to understand the physical nature of the speech acoustic wave that is transduced by a microphone to an electric signal for efficient transmission through a telephone network to a handset on the receiving end. A cogent understanding of how the ear functioned then also became an absolute necessity in order to effectively determine the bandwidth requirements and performance of the electric transmission network that was being rapidly installed worldwide.

This speech signal then is the input for all automatic speech recognition (ASR) systems; efficiently characterizing, parameterizing, and quantifying the speech signal is the most critical step prior to design of an ASR system. What are the essential features to look for in the signal and to parameterize? This was and still is a key question. Its answer is required for both speech recognition and speaker identification as well as for a guide to speech synthesis, speech compression, and speech coding work.

Let us now look at how the work proceeded.

2. THE AGENDA PROBLEM

Research in speech characterization was carried out primarily by scientists who were, by training, electrical engineers. The people who designed and built the speech transmission systems were also electrical engineers who were expert in handling circuits and systems. However the goals and considerations of the researchers on one hand and the system builders on the other were somewhat different.

For the system builders, what I term the engineering approach (Agenda I), the goal is to build an efficient, economically affordable, maintainable, and viable system to reproduce the speech wave at the receiving end to the prescribed accuracy. It is irrelevant whether the model used for speech characterization is physically correct or not; it must simply do the job and meet the specifications.

On the other hand the researcher in speech modeling has the goal of getting the physics of speech production correct first and then worrying about how to mathematically characterize the physics to model the essential features of production. This is called the scientific approach (Agenda II). Implementation, constructability, cost, simplicity, and countless other system design considerations are not relevant to the search for understanding the physics of speech production and hearing.

In the early years, the 1920's, research and engineering were carried out side by side often by the same people. As the telephone system grew, the two functions slowly separated in the 1930's becoming nearly disjoint in the 1950's. This enabled great strides to be made on both agendas. What were the results and what were the major signal processing implications?

3. THE MODELING PROBLEM

The speech signal as far as the telephone network is concerned is solely represented by the electric signal from the telephone transmitter or microphone. It is this electric signal that we strive to measure, understand, and characterize. So it was very logical for the electrical engineer to view this signal from a frequency-domain point-of-view especially since speech is 60–65% voiced giving a quasi-periodic output signal. Spectrum analyzers in many forms including banks of bandpass filters were the primary means of viewing and characterizing the speech signal. Attention focused on the peaks in the spectrum, the so-called formants, and the different voiced phonemes were then defined and categorized in terms of the relative positions of these formants. This was recognized quite early and dominated the characterization work; see [3, 4, 5]. From this point, viewing the vocal tract as a quasi-linear passive system having as its input the glottal flow wave and as its output the acoustic pressure wave, the speech signal, at the mouth exit, one could now construct meaningful electric circuit analogies for the operation of the vocal system. The seminal work of Flanagan [6] and that of Fant [7] describe this approach in fine detail. This approach has dominated the speech scientist's and engineer's work on the study of the vocal tract and on the construction of speech synthesizers.

Further, if the vocal tract could satisfactorily be represented this way then the obvious way to view the operation of the ear was as a spectrum analyzer. Hence speech recognizers and speaker identification methodologies centered around detailed short-time spectrum analysis. This approach was further cemented in place by the ubiquitous use of the fast Fourier transform (FFT) algorithm developed by Cooley and Tukey in the mid 1960's for efficient computation of signal spectra. Concomitant with the FFT was the rapidly developing availability of more powerful digital computers.

The rapid proliferation of digital computers into academia as well as industry saw speech research and development now being carried out in many locations; it was no longer the province primarily of the telephone companies. A second major benefit of the digital computer was that it could be programmed to serve as a very powerful signal and system simulation tool. New speech processing systems could now be rapidly tested and developed by simulation without going to the time and cost expense of designing and building experimental hardware only to find out that there was a major flaw in the concept. Designs could be changed rapidly on the computer and honed to perfection.

The new field of digital signal processing (DSP) encompassed this rapidly emerging area of system simulation and signal analysis. In point of fact much of the theory and methodology of DSP was strongly driven by application problems in the speech and in the geophysical prospecting areas. Nebeker [8] describes the revolution in DSP from 1948–1998 in his very informative book well worth the reading.

The universities were now educating new students in the science and technologies of speech science. New speech products could now be considered by industry such as text-to-speech systems and speech-to-text systems. The problems were formidable but work began in earnest. A new revolution was about to take place.

4. THE DIGITAL REVOLUTION

The digital integrated circuit began to appear on the scene at the end of the 1960's. The rapid development of the scale and complexity of available integrated circuits was phenomenal. System designers could now consider complete digital implementations of their systems. The scales of integration were increasing, the circuit speeds were increasing, the power requirements were dropping, the number of part types were increasing, and above all the part costs were decreasing, all to the shear delight of the system designer. The appearance of TI's Speak & Spell toy showed the world that workable speech products could be produced and marketed in the well-under \$100 range; and this was in the 1970's!

The next major events were the development of the DSP chip and the microprocessor chips. Along with cheap memory, the designer now had most all the parts out of which to implement almost any conceivable speech processing and speech recognition system imaginable. This has ushered in strong competition between many companies to devise needed and marketable speech application products of which ASR devices undoubtedly head the list.

This digital revolution and the early limited success of a number of application products along with the general tightening of the economies worldwide have had the additional effect of concentrating almost all the work to be of the Agenda I type. Pure research seems to have given way to the so-called applied research. Our current models for speech production seem to be satisfactory for most all of our needs.

But is this really the case? Are our models really sufficient?

5. OPEN QUESTIONS

Do we really understand the speech production process? Do we really understand how the ear works? Will better answers to these questions help us to design and build improved ASR systems? Hyde [9] notes that "feature extraction is better than pattern matching in recognition at the acoustic level" and "the human-speech communication process is remarkably resistant to very severe corrupting influences." Why? Is the ear really behaving primarily as a spectrum analyzer or is this idea simply a consequence of the mathematical tool, the FFT, that we choose to use for analysis of the speech wave? On the plus side we clearly have the fact that the systems we build using these models really work to a strong degree. Agenda I seems to be serving us just fine. Hyde's conclusion "that significant advances in speech recognition are not likely to come from researches into signal analysis, adaptive pattern matching, or computer implementation, but from studies of speech perception and generation, phonetics, linguistics, and psychology" made in 1968 still has a strong element of reality in it although adaptive pattern matching schemes and digital implementations have solved many problems. Our present systems are still quite complex but work.

There is the additional consideration that in trying to build ASR systems using the integrated circuit technology of today, it is the design constraints and properties of this technology that really dictate how we go about the design. We do not have integrated circuit parts that faithfully replicate even in a crude sense the function of the cochlea. We don't even come close. Nor should we as this argument suggests.

So is there really a problem? I believe there still is. Then what is the evidence? On the speech production side we note that the most complete models of the source-filter theory and vocal fold oscillation behavior still do not explain or begin to describe the exact nature of the airflow above the glottis during phonation let alone in the regions further downstream. These models look only at the hard boundaries of the vocal tract and then only from the standpoint of approximate cross-sectional areas; they even change the geometry to that of a straight tube. This step seriously compromises the flow field eliminating important secondary flows.

Hamming noted that "without measurement it is difficult to have a science." From a large number of detailed airflow measurements on the vocal tract Teager in [10, 11] gives very compelling evidence that acoustic components of the speech can be and are actively generated downstream from the glottis itself and that both amplitude and frequency modulations abound in the process. His chapter in the Daniloff book makes for very interesting reading by the serious researcher; it distills some of his careful thinking and observations made in the more than twenty years he worked on the modeling problem before his untimely death in 1990. His work and writings have opened many eyes as we are now beginning to see work on careful studies of the actual airflow behavior during phonation. The results indicate that much more careful work is needed. The implications on feature definition can be profound.

That we do not understand the speech production process adequately is also shown when we note that after working for more than fifty years on vocal tract modeling we still have no parameterization that can be fed into a synthesizer to generate faithfully, or even approximately, a particular person's speech. Note the news article by Guernsey [13] on software for copying any human voice. The proposed methodology simply requires the human whose voice is to be copied, to read 10–40 hours of text into the machine. Then the machine analyzes and categorizes all the sounds into a large dictionary of phonemes from which it can then "synthesize" by concatenation any arbitrary utterance. This is clearly an Agenda I solution.

On the hearing side there, too, are many rumblings that things are not as understood as one might be led to believe. Is the ear primarily a spectrum analyzer? Capranica has some reservations that he articulates well in [12]. He notes that "the auditory system is uniquely endowed with an ability to process rapid signal variations in time" and that "more attention should be paid to the encoding of temporal waveforms directly in the time domain rather than invoking a transformation into the frequency domain." The ear is basically designed to be a very efficient transient detector necessary for the survival of the human. It did not evolve to detect efficiently simple sinusoids as its main objective. Again, the active ASR researcher may benefit from a careful reading of Capranica.

I have described only briefly a few of the very disconcerting open questions and relevant work. Where do we go from here?

6. THE CHALLENGE

Recall the words of von Kempelen written in 1791: "The invention of a talking machine, and its operation in accordance with a well-considered plan, would be one of the boldest schemes to occur to the human intellect." Here we are 210 years and several billion research dollars later and still we do not understand the voice and its recognition in spite of the very powerful tools at our command.

Henry Petroski in [14] notes "As long as one does not question the validity or recognize the restrictiveness of basic assumptions, one can overlook the fact that they are limiting one's interpretation of results." Much good science remains to be done. We now have at our command very powerful investigative signal processing tools to aid us in this task. But we should make sure that we understand the assumptions that underlie these tools so that we, too, will know what limitations they may be placing on what we can observe and measure with them as we probe and push forward the frontiers of our field.

I will close with a few quotes by well-known people to ponder.

- Sometimes it is not how much you see, but how deeply you look. Jeff Rennicke, *The Grand Staircase*.
- The great tragedy of science is the slaying of a beautiful hypothesis with an ugly fact. T. H. Huxley.
- More things are known than are true. J. R. Pierce.
- Beware of finding what you are looking for. R. W. Hamming.
- Nature is not embarrassed by difficulties of analysis. Augustin Fresnel.

And finally from Colin Fletcher [15]:

> You cannot escape the age you live in: you are a product of it. You have to stand back from time to time and get your perspective right. But then you have to come back and resume the task of contributing in your own way to your own age.

Go forth and do great science!

7. REFERENCES

[1] J. L. Flanagan, "Voices of Men and Machines", *J. Acoust. Soc. Amer.*, vol. 51, pp. 1375–1387, 1972; reprinted in *Speech Synthesis*, J. L. Flanagan and L. R. Rabiner, Editors, Stroudsburg, PA: Dowden, Hutchinson & Ross, Inc., PA, pp. 9–21, 1973 and in *Speech Analysis*, R. W. Schafer and J. D. Markel, Editors, New York: IEEE Press, pp. 4–16, 1979.

[2] H. Fletcher, *Speech and Hearing in Communication*, J. B. Allen, Editor, ASA, 1995; reprint of 1949 edition which is an integrated version of Fletcher's *Speech and Hearing*, 1929.

[3] T. Chiba and M. Kajiyama, *The Vowel, Its Nature and Structure*, Tokyo: Tokyo-Kaiseikan Pub. Co., 1941.

[4] H. K. Dunn, "The Calculation of Vowel Resonances, and an Electrical Vocal Tract", *J. Acoust. Soc. Amer.*, vol. 22, pp. 740–753, Nov. 1950.

[5] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels", *J. Acoust. Soc. Amer.*, vol. 24, pp. 175–184, 1952.

[6] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, New York: Springer-Verlag, 1965.

[7] G. Fant, *Speech Sounds & Features*, Cambridge, MA: The MIT Press, 1972.

[8] F. Nebeker, *Signal Processing: The Emergence of a Discipline 1948 to 1998*, New Brunswick, NJ: IEEE History Center, 1998.

[9] S. R. Hyde, "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature", chapter in *Human Communication: A Unified View*, E. E. David, Jr. and P. B. Denes, Editors, New York: McGraw Hill, pp. 399–438, 1972; also in *Automatic Speech and Speaker Recognition*, N. R. Dixon and T. B. Martin, Editors, New York: IEEE Press, pp. 16–55, 1979.

[10] H. M. Teager and S. M. Teager, "A Phenomenological Model for Vowel Production in the Vocal Tract", in R. G. Daniloff (Editor) *Speech Sciences: Recent Advances*, San Diego, CA: College-Hill Press, pp. 73–109, 1983.

[11] H. M. Teager and S. M. Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract", in *Speech Production and Speech Modelling*, Edited by William J. Hardcastle and Alain Marchal, NATO Advanced Study Institute Series D, Vol. 55, Bonas, France, July 17–29, 1989, Boston, MA: Kluwer Academic Publishers, pp. 241–261, 1990.

[12] R. R. Capranica, "The untuning of the tuning curve: is it time?", *The Neurosciences*, vol. 4, pp. 401–408, 1992.

[13] L. Guernsey, "Software is Called Capable of Copying Any Human Voice", *N. Y. Times*, New York, p. 1, July 31, 2001.

[14] H. Petroski, *Invention by Design*, Cambridge, MA: Harvard U. Press, 1998.

[15] C. Fletcher, *The Man Who Walked Through Time*, New York: Alfred Knopf, 1973.