Speech Technology in the Coming Decades

J.L. Flanagan
Murray Hill, N.J., USA

Especially in science, prognostication is at best risky – at worst, futile. Nevertheless, my assignment is to comment on speech *technology* in the coming decades, and I take up this gauntlet.

To achieve perspective, it seems prudent to look over the recent past and take note of advances that have been key in speech research, and that have significantly impacted speech technology. These may suggest the nature of accomplishments to look toward in the future. To make such assessment, some criterion of focus is naturally implied. The choice invariably is conditioned by personal experience with the field.

First, why do we do speech research? Many would say to provide greater capabilities for human communication. And, I believe this to be a moderately universal and valid motivation. How do we improve capabilities for human communication – both between humans, and between humans and machines? The possibilities branch in many directions. My choice lies with techniques for telecommunications and voice processing. Others devote effort to communication aids for the handicapped, to speech teaching and therapy, to studies of language and language acquisition, to diagnostic methods for voice disorders, and to the many areas typified in the literature of the phonetics journals. Let us agree, though, that our common motivation is betterment of human communication, and against this backdrop presume to assess – and extrapolate – contributions in speech technology. In so doing, we can try to correlate the technological needs, the advances to meet the needs, and the acquisition of fundamental understanding to support the advances.

The era of the 1940's rode the swell of the evolving electronics age, and it seems not unreasonable to commence comment here. Undoubtedly the prominent technology of this time must include the vocoder – the first practicable analysis/synthesis system for bandwidth conservation in telephony. This achievement was spurred by the desire to transmit voice over the early transatlantic cable. But, this cable (before the time of integral, submerged amplifiers) could only support a bandwidth of a couple hundred Hertz, only enough for telegraphy. The desire to have the speed and convenience of voice communications therefore gave rise to the vocoder technique for a 10-fold reduction in the bandwidth of a speech signal.

Stemming also from this motivation were the fundamental concepts of the

'carrier nature of speech', the 'source-system' model of the signal, and the 'information-bearing' properties of the short-time amplitude spectrum. And while the vocoder was never put to work on telegraph cables (because bandwidth improvements progressed more economically), it later found extensive use for voice encryption purposes. Also, its synthesizer component evolved into a human-controlled electronic speaking machine – the voder. Its analyzer component influenced the design of the sound spectrograph, a fundamental instrument commonly found in most phonetics laboratories, and the design of the visible-speech translator, a useful tool for articulatory training of hearing-impaired individuals. On the perceptual side, the need to characterize and analyze the performance of speech processing systems gave birth to the concepts of articulation testing and articulation index.

The era of 1950 continued the interest in efficient voice communications, but recognized the need for better fundamental underpinnings. The wave nature of sound propagation in the vocal tract was put on a firm basis, as was similar understanding for auditory function; i.e., for the basilar membrane. Transmission-line models – bilinear, passive circuits – were introduced to good effect as analog computers. The non-independence of speech-spectrum amplitudes, and the information properties of formants were firmly established. Engineers moved to exploit this knowledge in automatic formant trackers and in formant analyzers and synthesizers. Concomitantly, the field of electronics experienced major progress with the introduction of the transistor and solid-state circuitry – a harbinger of greater vehicles for speech technology (the digital computer and integrated electronics).

The 1960's witnessed the impact of digital computers in partnership with sampled data theory as formidable tools for speech research. Previously, great limitations were imposed on the complexity of algorithms that could be implemented in electronic circuitry, and on the speed with which new ideas could be realized for test in traditional analog electronics. Digital simulation significantly relaxed these restrictions, and allowed much greater sophistication in processing. Speaking machine programs were of immediate interest, and formant synthesizers with discrete phonetic control of segmental and supra-segmental features attracted early interest. Eventually, complete formant vocoders were implemented in the laboratory, with real-time formant tracking accomplished by dedicated computer. The traditional vocoder concepts, and the extensions to pattern-matching vocoders (now given the more prestigious term vector-quantized spectra), were also cast into digital forms. On the practical side, transistor circuitry supported the electronic artificial larynx, which was built upon fundamental understanding of vocal-cord function. And, the vocoder concept of the source-system signal model was extended to its most sophisticated level in the form of linear-predictive-coding (LPC).

By the early 1970's, with broadband transmission technologies such as coaxial cable a reality, light guide showing great promise, and digital machines increasing in capability, the needs in speech technology largely shifted

away from band-conservation and toward human/machine communications. Giving machines the ability to speak stored information to a human, and to respond to human-spoken commands (even to confirm the identity of the talker) became central foci of research. Initially, the accumulated understanding from the vocoder art, and its direct derivatives, supported these efforts. But the sophistication of the machines permitted much more. Complete systems for speech synthesis from printed text were demonstrated and tested, for information retrieval purposes and as reading machines for the visually handicapped. Waveform coding methods such as adaptive differential-PCM (ADPCM) were devised for transmission economies, but used initially for multi-line computer voice answerback systems. Isolated word recognition systems of high performance, and talker verification systems of high accuracy filled in the developing picture. Fundamental studies to support more ambitious undertakings did not languish either. Detailed computer models of vocal-cord and vocal-tract function were established for speech synthesis. Sub-languages, having usefully-large vocabulary size and quantitatively-delineated grammar, were designed and programmed for automatic syntax analysis in speech recognition systems. But in all of this work, the central tool, the laboratory digital computer and its elaborate peripherals, remained large, expensive and oftimes not fast enough for real-time simulations.

Around 1980 this picture changed dramatically, with explosive advances in microelectronics. Already in the early 80's we have single-chip computers that are more powerful than the dedicated laboratory computers of the 70's. Integrated speech synthesizers are pervasive, and even provide convenient test beds for phonetics laboratories. Chip-set speech recognizers are appearing, and most of the designs can be made compatible with the communication protocol of existing microcomputers.

As we approach the mid 80's, activities in speech technology are still dominated largely by the needs of human/machine communication. The advances necessary to meet these needs are in the areas of higher-quality synthetic voice, automatic recognition of connected speech, and simultaneous speech and talker recognition. By the end of the decade, 1990, it seems reasonable to expect significant advances in each sector. We will have text-to-voice converters that will deal reliably with virtually unlimited vocabulary and will produce intelligible, natural-sounding output. We will not be able to specify and duplicate the subtleties of dialect and accent, but we will be substantially past the stage of the inept automaton. Similarly, recognizer-/synthesizer systems will be able to carry on intelligent, interactive conversations with humans. Not fluently, nor with all talkers on all subjects, but constrained to vocabularies, grammars and topic areas that are nevertheless comfortably large. The applications outside telecommunications, such as aids for handicapped and speech teaching, are apparent.

The fundamental studies to support these advances are numerous and do not differ much from the objectives of the recent past. Accurate, agile,

models of articulation are needed. Letter-to-sound conversion – utilizing the proper marriage of stored pronouncing dictionaries and grapheme/phoneme rules – will be refined. Connected speech recognition will utilize stored word templates initially. But, as aspirations expand to vocabularies on the order of 1,000 words, or more, feature labelling and statistical modelling may prove more attractive.

While the need for bandwidth conservation, of late, has been in the background, it will not remain so. Telecommunication transmission systems are rapidly evolving to digital techniques. And while light guide will provide enormous bandwidths, specific considerations of access and switching make band conservation attractive, and in some cases necessary. Already 32K bits/sec ADPCM is on the threshold of use to achieve 2:1 savings in transmission capacity. Other techniques, such as sub-band coding (SBC), perceptually-weighted multi-pulse LPC, and adaptive transform coding (ATC), are in advanced stages of research and address the transmission ranges below 32K bits/sec. A strong, emerging need is to encode and transmit high-quality speech at data speeds, 9.6K bits/sec and lower. This low bit-rate coding makes speech signals adaptable to networks that also handle data and low-rate non-speech signals. The end of this decade will see high-quality speech transmission at rates in the range of 9.6K. High-quality at much lower rates, for example 2.4K, is possible, but probably will be longer in coming. Fundamental understanding, that significantly surpasses the traditional source-system signal model, must first be acquired.

If we look toward the 2000 era, predictions indeed become risky. But, I believe this time scale can see the beginning of a unified 'ultimate' solution to the synthesis, recognition and coding problems. At this moment, I believe this solution must be based upon a speech signal model that much surpasses the traditional vocoder source-system model – one that allows exquisite, dynamic representation of the details of laryngeal and tract functions and which is controlled adaptively to 'mimic' an unknown input (either to duplicate it for synthesis, to categorize it for recognition, or to parameterize it for coding and transmission). The adaptation algorithms obviously must contain voluminous built-in information about speech constraints and conventions, and about the mechanism of speech perception. Some of this insight is being accumulated. But the complete solution is some years away. In part, it will depend upon quantum advances in computer capability. For practical application, inexpensive processors with arithmetic capability in excess of 100 Mips will be needed to support this complexity.