SPEECH SCIENCE AND TECHNOLOGY - REVIEW AND PERSPECTIVE

Professor Robert Linggard

University of Essex, Wivenhoe Park, Colchester CO4 3SQ, Essex, England

1. REVIEW

Though Speech Science is probably as old as Science itself (Aristotle wrote on the subject), Speech Technology probably began in 1938 when Homer Dudley of Bell Labs invented the Channel Vocoder. This was a speech transmission system, which, at the transmitter, analysed speech into ten frequency channels plus an excitation type (buzz or noise) and re-synthesised the speech at the receiving end using an identical ten channel filter system excited by buzz/noise. The theory of this device relied on the observation that speech was mainly a coded spectrum and its detailed wave form was not significant for perception.

The virtue of the Vocoder was that the eleven components of the analysed speech required less bandwidth to transmit them than the original speech. It was, what we would now call, a low bite-rate coding system, and was the first real attempt to take speech processing beyond amplification and filtering. A year later, in 1939, at the New York World Fair, the ten band synthesiser, fitted with manual control keys, was shown generating synthetic speech by trained, key-board operators.

Speech Science, at this time, was largely descriptive, and the phonetic alphabet and the theory of phonemes were it main theoretical base. After the invention of the Vocoder, it was hoped that, within a few years, an analyser could be developed to detect phonemes, and the synthesiser would be adapted to accept phonetic text as input. It would then be possible to transmit speech as phonetic text with extremely low bandwidth. More than fifty years later, even though speech recognition and synthesis have advanced enormously, this "Phonetic Vocoder" is still a dream.

In the 1950s and 60s, efforts to analyse speech into phonemic units met with frustration, and speech recognition research became concentrated on recognising whole words, a task which proved to be much more tractable. The invention of digital computers and the application of dynamic time warping provided some success, and the first commercial recognition systems appeared on the market in the 1970s. It has since been possible to recognise continuous speech using sub-word units based on phonemic ideals. However, the development of speech recognition has owed more to advances in stochastic processing than linguistic insight.

It had been speech synthesis research which has taken up the ideas of phonetics. In 1964, John Holmes at JSRU in England devised the first successful text-to-speech system. In this scheme, control signals are derived from phonetic text by a set of rules running on a digital computer, and are used to drive an electronic, parallel-formant synthesiser. Later work by D Klatt and J Allan at MIT improved both the intelligibility and quality of synthetic speech generated from text. The morph decomposition theory of Allan and the co-articulation rules of Klatt have made a significant contribution to speech science.

The development of electronic instrumentation and the advent of the digital computer in the 1950s gave Speech Science an impetus which transformed it from a speculative study into a modern, experimental science. The understanding of speech perception advanced rapidly through the modelling of cochlea mechanics, and speech production progressed quickly via experimental phonetics and modelling of vocal tract aerodynamics. Digital recording and processing methods facilitated more and more sophisticated psycho-acoustic experiments on humans and detailed neural recordings from the auditory pathways of animals. In many ways, Speech Technology still has to catch up with this enormous explosion of knowledge in Speech Science.

A very large amount of research in Speech Science and Technology has been carried out in the past fifty years, though it is only recently that Speech Technology has become commercially viable. The field has been fortunate in receiving huge amounts of funding, probably because the technology appeared, for a long time, to be just on the verge of breaking-through to commercial exploitation. Speech Technology is now, unquestionably, applicable in many branches of industry and commerce. Yet compared with human capabilities, speech recognition and synthesis have still a long way to go, and Speech Science still has many unanswered questions.

2. SPEECH SCIENCE

The relationship between Speech Technology and Speech Science is symbiotic. Speech Technology relies on Speech Science to provide insight into the human process of speech production and perception, and, in turn, Speech Science benefits financially from the growing commercial importance of Speech Technology.

The first rewards of electronic technology manifested themselves in the field of Phonetics. Electronic recording of speech made it possible to repeat utterances in a controlled manner and to show the extent of acoustic variation in utterances of the same phrase. Spectrograms showed clearly the existence of formants and the manner in which vowels are modified by adjacent consonants. Later, the monitoring of articulatory movements at the same time as the speech signal enabled detail study of articulation and its relationship with acoustics. More recently, the use of digital computers has made it possible for theories of articulation to be validated by modelling techniques.

In speech perception, the auditory pathways have been investigated in both animals and humans. Again, electronic instrumentation and digital computers have been essential to the rapid advances in technique and sophistication. The field of psycho-acoustics has been able to describe the human responses to a wide variety of stimuli, both spectral and temporal. In particular, the perceptual cues in speech have received detailed attention. These responses, in humans, have, by necessity, been measured by psychological testing methods. More direct responses to acoustic stimuli, at the neural level, have been observed in animals, mainly cat, guinea pig, and squirrel monkey. In some cases, such as temporal and spectral masking, the psycho-acoustic response has been validated at the neural level. However, in general, animals have been used to uncover the lower level responses, and psycho-acoustic measurements have served to investigate the higher level auditory pathways. Cochlea modelling has been used to explain observations and to validate theories.

Linguistics has been transformed from a descriptive, to a computational, science. Theories of syntax and grand grammatical schemes can now be tested using computer programmes. At a more practical level, lexicons and dictionaries have been put into computer readable form. Spelling rules have been devised, and Allan's morph theory used to analyse words into their constituents. Much linguistic work, originally done for text processing, has now been applied to Speech Technology. The field of language understanding is now being pursued for text and speech systems, and constitutes the ultimate objective of both fields.

3. SPEECH TECHNOLOGY

The original objective of Speech Technology, to create low bandwidth transmission systems, is still valid. In the fifty years since the Channel Vocoder, transmission rates have fallen from 64 kBits/sec through 32, 16, 8 and now 4 kBits/sec. The next target of 2 kBits/sec is already within sight. The two spin-off applications, speech recognition and speech synthesis, are also still subjects of research, even though commercial recogniser and synthesisers are now widely available.

In the latest recognition research, HMMs have replaced DTW as the dominant method. For large vocabulary systems, whole word recognition is not practical, and sub-word units have become the elementary particles of recognition. Though these units are based on the ideal of phonemes, they allow for the full contextual variation found in practice. Thus there are diphones, triphones and "phonemes-in-context". In order to accommodate the full range of variation about 1000 to 2000 of these basic units are needed. The technique of training sub-word models is very sophisticated and requires very large amounts of training data and processing time.

Speech synthesis is now very advanced, and special chips are available which perform the synthesis when driven by appropriate control signals. Given the good intelligibility of most synthesisers, speech quality has now become the main research issue. In this respect, the field is still very active and has returned to earlier ideas of using segments of real speech. However, the segments are now at sub-word level and their assembly into whole utterances is very complicated indeed. Also, new theories of phonology have lead to better rules in the construction of words from phonemic units.

In addition to the three traditional technologies of speech coding, recognition and synthesis, a commercial need has arisen for speaker recognition and verification. The research here is very active, and speaker recognition may be one field where the computer will out-perform humans. At present, it is possible to recognise hundreds of different speakers with fairly good reliability. As yet, it is still not known which parameters will prove to be the most robust in speaker recognition, and there are still plenty of opportunities for the technology to be improved. Commercial systems are somewhat pragmatic, and there is a need for Speech Science to investigate the fundamental issues.

4. THE FUTURE

The division between Speech Science and Speech Technology is not as definitive as the above review would suggest. The two fields overlap to a considerable extent and many individuals work in both fields simultaneously. This is as it should be and even closer co-operation will be necessary in the future. The great hope is that Speech Science will discover exactly how humans produce and perceive speech, and that Speech technology will use this knowledge to construct better and more efficient systems - better, perhaps, than humans themselves.

In the past, the pressure to produce commercially viable systems has caused technology to go its own way. For example, the great advances in Speech Recognition technology have been due almost entirely to innovations in stochastic processing rather than the insights of speech perception. Even though sub-word units are based on phonemic ideals, the increase in size of the basic set of units from the 40 phonemes, familiar to phoneticians, to the 1500 triphones used in successful recognition, suggests that a gap has opened up between theory and practice.

In text-to-speech synthesis, though initial systems used ideas from phonetic theory, the lack of good voice quality has led, again, to the use of sub-word units which, in number and kind, are only vaguely related to phonemes. In addition, the sub-word units of recognition are not the same as those of synthesis. Clearly, there is a need to close this gap with a theory of speech production and perception which takes into account the large acoustic variations between utterances of the "same" phoneme.

One of the greatest discrepancies between human and machine recognition of speech, is that for humans voice pitch carries important information, whilst in most speech recognisers pitch effects are smoothed out. Again, there is need to account for the use of pitch information in the theory of speech perception, and to use this knowledge to improve the performance of speech recognisers.

However, it is at the level of language understanding that the greatest difficulties and the greatest challenges occur. It is well known that speech is highly structured and that this structure constricts the search space at all levels of recognition and understanding. At the lowest levels, this constriction is due to the limitations of the human vocal system, at higher levels it is due to the limits of phonology, language, semantics, and context. There is as yet no theory which describes this structure at all levels, and how the human brain learns to use this limited search space to perform extremely efficient communication.

This review has attempted to step back and look at the whole field of Speech Science and Technology, to examine its triumphs and to assess its current failings. The hope is that by careful analysis, the trends and tendencies of the past may be projected into future. An essential message becomes clear; it is that the failures and successes of Speech Technology are a good indicator of the adequacy of the theories of Speech Science; and also, that though Speech Technology may go its own way for some time, it must ultimately come back to the theories of Speech Science.