SPEECH PROPERTIES AND METHODS FOR SYNTHESIS AND RECOGNITION

J.S. BRIDLE

Rapporteurs: Mr. R.C. Millichamp
Dr. S.K. Shrivastava

1.0 THE NATURE OF SPEECH

We are so used to using speech that is is often difficult to remember that the whole process is immensely complex and poorly understood, despite the efforts of scientists from many disciplines, such as acoustics, neurophysiology, linguistics, electrical engineering and even computer science.

Speech is a means of communication between minds. As the message passes between two minds it takes many forms. Most of the stages involved are the brains of the speaker and the listener and very little is known about the form the message takes there.

1.1 Speech Production

The complex, constantly-changing pattern of sound that carries most of the information is produced by the interaction of a wide-bandwidth source of sound (produced in the larynx, at a constriction, or by the sudden release of air pressure) and the frequency-selective action of the vocal tract, which depends on its shape. The vocal tract is shaped by the articulators, which include the tongue and the lips.

1.2 Speech Perception

The patterns of sound entering the ears are transformed by an exquisite system involving hydromechanical resonance and many stages of neural analysis. In the early stages the main action is to lay out the pattern as a function of time and frequency.

1.3 Speech Waveforms

The telephone system is based on the observation that speech communication is possible (with various limitations) if we measure the sound pressure waveform near the speaker's mouth and reproduce an approximation to it near the listener's ear. For telephone quality speech we need to reproduce the first few kilohertz of the audio spectrum, with a signal-to-noise amplitude ratio of at least 100:1. This needs tens of thousands of bits per second (say 50Kb/s), but can be reduced to about 10Kb/s with compromises in quality, by using special techniques.

1.4 Sound Pattern Analysis and Synthesis

Very much more compact representations are possible by attempting to reproduce not the waveform itself but important properties of the pattern. "Vocoders" need only a few thousand bits per second, and exploit the fact that speech is the response of a time-varying linear system to a time-varying excitation, plus other properties of speech production and perception. One particularly successful method of constructing speech-like time-frequency-energy patterns is in terms of "formants", which are peaks in the spectrum, normally corresponding to resonances of the vocal tract.

1.5 Phonemes

As children we learn that words are made up from a limited set of sounds, which occur in a different combination in each distinct word. The sounds correspond roughly to letters of the alphabet (very roughly in English). These basic sounds are the phonetician's phonemes. What could be more natural than to analyse, synthesise, recognise and transmit speech signals in terms of phonemes? Unfortunately, life is not that simple.

Speech scientists now recognise that there is no simple one-to-one correspondence between the linguistic "sounds" (Phonemes) and physically measurable sounds (which are discrete neither in time nor in acoustic properties). A basic problem, then, is the gulf between the analogue world of SIGNALS (continuous, flowing patterns of sound whose properties depend on the speaker, the context and a dozen other factors) and SYMBOLS, which computer programs might relate to meaning or conventional text. Oppenhiem has pointed out that the intrinsic difficulty is made worse by a communication gap between those experienced in signal processing and those experienced in symbol processing.

There is no reason to expect an easy solution: nature has had millions of years to adapt prodigeous processing power to the problem of communicating via a restricted channel (limited mainly by the slow-moving articulators, which had to maintain their functions in breathing and eating). We can expect most work to have been done on the "firmware" and the "protocols".

As we shall see, there has been some success in synthesising intelligible speech from symbolic specifications, including conventional text, but all practical, working automatic speech recognition systems avoid the signal-to-symbol problem.

2.0 TECHNIQUES FOR SPEECH OUTPUT FROM MACHINES

2.1 Concatenation of Stored words.

Waveform, Vocoder.

2.2 Automatic construction of synthetic speech patterns

Synthesis-by-rule from phonetic text. Synthesis from conventional text.

3.0 APPROACHES TO AUTOMATIC SPEECH RECOGNITION (ASR)

3.1 Problems in automatic speech recognition.

Continuity, Variability, Ambiguity, Complexity.

3.2 Isolated word recognition using whole word templates

The most popular technique for ASR solves the above problems by re-defining ASR so that the problems are by-passed or minimised. Instead of trying to recognise anything, said by anyone, in a normal speaking style, the designers of the first commercially-available speech recognition machines insisted that: the set of words would be limited to a few dozen; the words be uttered with distinct gaps of silence between them; the user of the machine must provide examples of all the words in an "enrollment" or "training" phase before the machine attempts recognition.

3.2.1 time-frequency-energy patterns

3.2.1 time-frequency-energy patterns - The first step is to turn each utterance into a pattern, using some form of short-term spectrum analysis. It is also neccessary to determine what part of the pattern of sound picked up by the microphone corresponds to the utterance to be analysed. This figure-ground separation, or endpoint detection, is often very difficult.

3.2.2 time-flexible matching - One of the most serious types of variability amoung different utterances of the same word by the same speaker is variations in the timescale. The most successfull isolated word recognition machines incorporate a powerful method of comparing word-patterns which copes with unknown non-linear differences in timescale. This method, which is based on Dynamic Programming, is related to algorithms familiar in Computer Science for comparing strings.

3.3 Connected Word Recognition Using Whole Word Template Matching

It is possible to remove the restriction that the speaker leave gaps between words. This allows faster data entry and needs less skill, but all the other limitations still apply. Perhaps the obvious approach is to divide the input pattern into words, then recognise each word as before. Unfortunately continuity beats us (consider "three eight" spoken fluently).

Alternatively, we could try all sequences of words, synthesise the corresponding patterns, compare them with the unknown speech pattern in a way that could cope with unknown differences in timescale, and choose the text that produced the pattern that matched the input best. This is likely to give good answers, but would take an impractical amount of computation.

The currently favoured approach, which is the basis of several commercially available connected word recognisers, does indeed find the sequence of templates which, as a single complete pattern, matches the whole of the input best. However, this is done without trying all possibilities, and very efficient algorithms exist, some of which have been used as the basis for real-time connected word recognition machines. These efficient connected word recognition algorithms are again based on dynamic programming, and the solution to the template sequence problem can in fact be integrated into the solution of the timescale variability problem.

Some versions of the integrated connected word recognition algorithm can be constrained so as to consider only those sequences of words that conform to the rules of a given simple formal grammar. This can reduce the amount of computation and increases the chances that the best-fitting sequence of templates will actually correspond to the words that are spoken (assuming that the speaker obeys the rules of the grammar).

3.4 The Current Generation of Connected Word Recognition Equipment.

Principles, Capabilities, Limitations, Likely developments.

DISCUSSION

The speaker was asked how soon realistic continuous speech recognition would be possible, and what speech recognition systems there were at Cheltenham.

Mr. Bridle replied that they have a continuous speech recognition system which is manufactured under contract by Logica. Other systems which are of current interest are ones capable of handling several speakers, of which there is a Laboratory model capable of recognising isolated words from a vocabulary of approximately twenty words. There is also a need for more robust recognition systems which are capable of working in an environment of background noise; the military have several possible applications of this. In the future a much larger vocabulary, possibly of a few thousand words, is desirable. IBM are developing a business letter dictating machine with a vocabulary of five thousand words, which they hope will be in operation soon. To advance further however will require significant breakthroughs, particularly in the area of signal/symbol conversion. One possibly is to use higher level information, such as what the person is trying to do.

Professor Randell asked Mr. Bridle what his opinion was of the claims made by the Japanese for their Fifth Generation Project, and if the Japanese language helped in speech recognition.

The speaker said that the Japanese have done some good work, but they had a lot of incentive because their written language was so cumbersome. The Japanese language has fewer syllables than European languages have, which may make continuous syllable recognition possible.