SYMPOSIUM 6: Human and Automatic Speech Recognition

Chairman: *D.H. Klatt, Cambridge, Mass. U.S.A.*

Panel members: *V.W. Zue, S.M. Marcus, M. Liberman, R. de Mori*

In general terms, the session sought to answer two broad questions: (1) Do current speech recognition strategies have anything to tell us about speech perception, and (2) can current theories and data concerning speech perception guide and improve the performance of automatic speech recognition systems? The panel consisted primarily of scientists involved in building speech recognition systems, although several panel members have also worked on problems in speech perception. The audience, on the other hand, was biased toward greater expertise in the areas of speech perception, as evidenced by the questions asked during the discussion period.

The chairman began the session by asking a number of fundamental questions. For example, is one of the first stages of the process leading to lexical hypothesization one in which a phonetic analysis is performed, or is the acoustic input matched directly with acoustic patterns for familiar words? There are problems with either view. Performing a phonetic analysis means making decisions and discarding information that may have been useful at later stages in the process. The inevitable transcription errors that result are much harder to correct during lexical search. Furthermore, it is difficult to specify the nature of a good phonetic representation - the best one is perhaps so detailed that it is simply a recoding of the input rather than an information reduction transformation. On the other hand, if word perception is direct, one has to account for speaker differences and dialect differences when matching input with acoustic patterns for words, and one must invoke separate analysis procedures for novel words.

Additional problems face builders of speech recognition systems and models of speech perception. Are some phonetic decisions easier? If so, should one use these robust cues to narrow the search? Victor Zue described a system employing this approach. However, during the question period, he was challenged on the performance of such a system (he said no data are available as yet) and on how to overcome an error in the initial partial transcription (he said lexical redundancy may permit detection and correction of some errors). He was also asked whether such an approach is practical in continuous speech where there is less certainty as to the locations of word beginnings and endings. Zue responded that he was quite optimistic as to the feasibility of applying the approach to continuous speech.

The potential advantage of search reduction via robust cues is that the

remaining lexical candidates may be distinguished in a sort of hypothesis-and-test verification scheme where the acoustic expectations can be narrowly specified because one is assuming a particular phonetic context. However, in the discussion, it was pointed out that the advantages of verification can be achieved in a bottom-up fashion by precompiling this sort of knowledge into an acoustic decoding network for words and word sequences (e.g. LAFS). When a member of the audience asked exactly how this might be done, Mark Liberman pointed out the practical difficulties by supposing that one had a machine that stored the sentence response for every possible ten-second digitized waveform (i.e. the number of different responses would be two raised to the power of 10,000 samples/second times 8 bits/sample times ten seconds). While such a machine can be conceptualized, it will never be built within our universe.

Renato de Mori described an elaborate system for speech recognition involving many levels of representation and multiple cues leading to decisions at any level. A member of the audience challenged whether this system, or any other speech recognition device, used strategies as complex as we know to be necessary from the literature on multiple cues to phonetic contrasts (such as the voiced/voiceless distinction where Lisker has catalogued over a dozen distinct cues). The panel readily admitted that current systems do not approach the sophistication required to take advantage of this knowledge, in part because it is so difficult to program strategies that involve interacting decisions (a change at one place in the program has ramifications, often unexpected, at many other locations in the code) and in part because the various constants needed to optimize such a strategy are usually not given and require incredible effort to discover from data.

Mark Liberman stressed the importance of extracting an appropriate representation of speech in order to achieve better phonetic/lexical identification performance than has been obtained to date. He indicated that formant trajectories are good candidates, and that new strategies may result in improved formant-tracking performance, but that our knowledge in this fundamental area is still quite primitive.

Stephen Marcus described an approach to speech recognition where words consist of unordered sets of spectral changes. He emphasized how remarkably well such a system works, meaning that spectral change (and/or phonetic change) is a concise summary of the most important aspects of the acoustic pattern for a word. However, he was quick to admit that this is not the complete story, and order information is needed to distinguish many words.

A member of the audience, Adrian Fourcin, drew our attention to how the infant begins to understand language, and suggested that we might build machines that mimic this process. The panel was able to pick up on this point and stress how little we know about the role of learning (versus innateness) in speech perception, or how to implement strategies that tune themselves from experience as well as discover new rules from experience.

A member of the audience, Mac Pickett, asked whether speech recognition has progressed as a science to the point where devices could be built to be used as aids for the hearing handicapped. The panel was unanimous in concluding that the systems entering the marketplace have very limited capabilities, particularly with respect to dealing with many speakers or dealing with large vocabulary continuous speech. Only if a small vocabulary isolated word recognition capability was useful would it be worthwhile to mount an effort in this direction.

Finally, a member of the audience, Karl Eric Spens, asked whether we should worry about the potential misuse of speech recognition technology, particularly in the area of surveillance and invasion of privacy. Mark Liberman responded that the technology is too primitive, as yet, to be really concerned. However, it is clear that we, as scientists most closely tied to this technology, have a duty to inform the public of the dangers as they arise, or before!