SPEECH INTERFACES: SESSION INTRODUCTION

Douglas L. Hogan
13625 Middlevale Lane
Silver Spring, MD 20906

ABSTRACT

The speech interface is the natural one for the human user and is beginning to be used in a limited way in many applications. Some of these applications are experimental; still others have achieved the status of cost-effective utility. A brief summary of the current state-of-the-art of speech input and output are presented. The two papers in the session represent specific examples of current work. Some comments on the need for linguistically oriented development conclude the paper.

I INTRODUCTION

Over the past four decades it has often been felt that the solution to the problem of "machine recognition of speech" is ".. just around the corner." When the sound spectrograph was invented (a little less than forty years ago) engineers, acousticians, phoneticists, and linguists were certain that the mysteries of speech were about to be unveiled. When powerful computers could be brought to bear (say - twenty years ago) there was a renewed feeling that such tools would provide the means to a near term solution. When artificial intelligence was the buzzword (a little over ten years ago) it was clear that now the solution of the recognition problem was at hand. Where are we today? A number of modest, and modestly priced, speech recognition systems are on the market and in use. This has come about because technology has permitted some brute force methods to be used and because simple applications have been found to be cost effective.

In speech output systems a similar pattern has emerged. Crude synthesizers such as the Haskins pattern playback of thirty years ago were capable of evoking "correct" responses from listeners. Twenty-five years ago it was thought that reading machines for the blind could be constructed by concatenating words. Twenty years ago formant synthesizers sounded extremely natural when their control was a "copy" of a natural utterance. Modern synthesizers are one one-thousandth the size and cost; they still only sound natural when a human utterance is analyzed and then resynthesized as a complete entity. Concatenating words is still no better, though cheaper, than it was twenty years ago.

II CURRENT TECHNOLOGY

A. Speech Input

There are now several speech recognition systems on the market which are intended to recognize isolated words and which have been trained for an individual speaker. The vocabulary sizes are on the order of 100 words are phrases. Accuracy is always quoted at "99+%." These recognizers use a form of template matching within a space which has the dimensions of features versus time. The "true" accuracy is a function of the vocabulary size, the degree of cooperativeness of the speaker, and the innate dissimilarity of the vocab ulary. Since the systems are recognizing known words by known speakers the major source of varia bility in successive words is the time axis. The same word may (and will) be spoken at different speaking rates. Unfortunately, different speaking rates do not result in a linear speed change in all parts of a word; the voiced portions of the word, loosely speaking the vowels, respond more to speed change; the unvoiced portions of the word, loosely the consonants, respond less to speed change. As a result, a nonlinear time adjustment is desired when matching templates. This sort of time adjustment is carried out with a mathematical process known as dynamic programming which permits exploration of all plausible non-linear matches at the expense of (approximately) squaring the compu tational complexity in contrast to the combina torial computational growth that would otherwise be required. The medium and high performance speech recognizers usually contain some form of dynamic programming. In some cases more than one level of dynamic programming is used to provide for recognition of short sequences of words.

The actual use of these recognizers has developed a number of consequences. Many of them, including the first paper in this session involve the use of speech recognition during hands-and-eyes busy operations. These applications will almost always be interactive in nature; the system response may be visual or aural. Prompt response saying what the system "heard" is crucial for improving the speaker's performance. A cooperative speaker clearly adapts to the system. To date, many applications are found where a restricted interactive speech dialog is useful and economical. At this time the speech recognition

mechanism is relatively inexpensive; the expensive component is the initial cost of developing the dialog for the application and interfacing the recognition element to the host computer system.

At the present time recognition is not accomplished in units smaller than the word. It has been hoped that it might be possible to segment speech into phonemes. These would be recognized, albeit with some errors; the strings of phonemes would then be matched with a lexicon. To date, adequate segmentation for this sort of approach has not been achieved. In fact, in continuous fluent speech good word boundaries are not readily found by any algorithmic means.

B. Speech Output

There are relatively few speech synthesizers in the pure sense of the word. There are many speech output devices which produce speech as the inverse of a previously formed analysis process. The analysis may have been performed by encoding techniques in the time domain; alternatively, it may be the result of some form of extracting a vocal source or excitation function and a vocal tract description. When the analysis is performed on a whole phrase the prosodic features of the indivdual uttering the phrase are preserved; the speech sounds natural. When individual words produced by such an analysis-synthesis process are concatenated the speech does not sound natural.

In any event, the process described above does not allow for the open ended case, synthesis of unrestricted text. This process requires that a number of steps be carried out in a satisfactory way. First, orthographic text must be interpreted; e.g. we read "NFL" as a sequence of three words but we pronounce the word "FORTRAN", we automatically expand out the abreviation "St.", etc. Second, the orthography must be converted to pronunciation, a distinctly non-trivial task in English. This is normally accomplished by a set of rules together with a table of exceptions to those rules. Although pronouncing dictionaries do exist in machine form, they are still too large for random access memory technology, although this will not be true in the reasonably near future. Proper nouns, especially names of people and places, will often not be amenable to the rules for normal English. Third, the pronunciation of the word must be mapped into sequences drawn from an inventory of smaller units. At various times these units have been allophones, phonemes, diphones (phoneme pairs), demisyllables, and syllables. The units are connected with procedures which range from concatenation to smooth interpolation. Finally, it is necessary to develop satisfactory prosody for a whole phrase or sentence. This is normally interpreted as providing the information about inflection, timing, and stress. This final step is the one in which the greatest difficulty exists at the present time and which presents the strongest bar to natural sounding speech. The second paper in this session deals with the development of stress rules for prosody, one component of the overall problem.

III LINGUISTIC NEEDS IN SPEECH INTERFACES

A. Current Research

Most of the current high end work in speech recognition attempts to constrain the allowable sequence of words by the application of some kind of grammar. This may be a very artificial grammar, for example the interaction with an airline reservation system. Other research efforts attempt to develop models of the language through an information theoretic analysis. Coming full circle we find words being analyzed as a Markov process; Markov, of course, was analyzing language when he developed this "mathematically defined" process.

Normalizing recognition to the speaker is being approached in two ways. The first, currently being explored at the word recognition level consists of developing enough samples of each word from many speakers so that clustering techniques will permit the speaker space to be spanned with a dozen or so examples. The second approach attempts to enroll a speaker in a recognition system by speaking "enough" text so that the system is able to develop a model of that person's speech.

In research on speech synthesis considerable attention is now being given to try, by analysis, to determine rules for prosody. Application of these rules requires grammatical analysis of the text which is to be converted to speech.

B. The Future

As both of the speech interface tasks become more and more open-ended it is clear that satisfactory performance will require very substantial aid from linguistic reseacrh. In the case of recognition this is necessary to reduce the number of hypotheses that must be explored at any given point in a stream of unknown words. In the case of text-to-speech, understanding of what is being said will contribute to producing more natural and acceptable speech.

IV FURTHER READING

The reference below surveys the current state-of-the art more deeply than can be presented here. It also calls out the need for increased application of linguistic information to speech interface development as well as providing an extensive set of references for those of you who would like to dig deeper.

Flanagan, James L., Talking with Computers: Synthesis and Recognition of Speech by Machines, IEEE Trans. on Biomed. Eng., BME-29, No.4, pp 223-232 (April 1982).