AUTOMATIC SPEECH RECOGNITION TECHNOLOGY DEVELOPMENT AT ITT DEFENSE COMMUNICATIONS DIVISION

DR. GEORGE M. WHITE, PH.D.

ITT DEFENSE COMMUNICATIONS DIVISION
SAN DIEGO, CALIFORNIA

INTRODUCTION

ITT Defense Communications Division supports the needs of the Government and Department of Defense in voice processing through a wide range of research and development activities. ITTDCD anticipates significant increases in the interest of government agencies for voice processing equipment in the next few years. This increased interest will be promoted through the maturing of signal processing technology. At least an order of magnitude improvement in the performance/cost ratio can be expected within five years due to advances in microelectronics, i.e. the commercial devices that today sell for $10,000 to $20,000 could be manufactured for less than $500 for many vocoding, speech recognition, and speaker verification devices. This lower cost is expected to open up new application areas within the government and defense community.

Considering speech recognition devices, for the task of recognizing a small vocabulary of isolated utterances, spoken by a small number of known cooperative speakers, in a relatively noise-free environment, over a high-quality microphone, the problems are practical rather than theoretical. There are many algorithms that achieve accuracies in excess of 99.0%. However, the relatively high cost of practical devices (typically more than $5,000) is prohibitive for most applications. The cost of such devices could soon be lowered dramatically by technological advances in LSI and charge transfer devices. For this reason, ITT has considerable interest in charge transfer devices and their application to Fourier analysis and bandpass filtering and Itakura LPC analysis. It is felt that recognition systems of the above type could be built for less than $500 per unit (excluding mask-making costs) that would achieve better than 99.0% correct recognition scores for 50 word vocabularies of polysyllabic utterances that differ in more than one syllable. Of course, the actual costs will be strongly dependent on production volume and the above estimate is based on high volume.

For all but the simplest form of recognition mentioned above, the problems are both theoretical and practical. ITT is actively

[redacted]

pursuing research on the practical and theoretical problems in the areas of speech vocoding (bandwidth compression), speaker verification and speech recognition.

I. PRE-FY78 TECHNICAL REVIEW

In FY77, ITTDCD completed development of the ITT processor: a high speed programmable signal processor. The ITT processor has been programmed to function as an LPC vocoder. It executes an LPC-10 analysis for the encoding operation, the characteristics of which are described below.

LPC-10 Characteristics

Predictor Order	10
Sampling Rate	8 KHz
Bit Rate	2400 bps
Frame	22.5 msec (54 bits per frame)
Analyzer	Semi-Pitch Synchronous

Low Pass Filter:	4th Order Butterworth
Pitch:	AMDF function with Dynamic Programming (DYPTRACK) smoothing (50 Hz to 400 Hz, 60 Values).
Voicing:	2 decisions per frame based on Low Band Energy, zero crossing count and reflection coefficients RD1 and RC2.
Preemphasis:	$Z_n - \frac{15}{16} Z_{n-1}$
Matrix Load:	Covariance (Modified ATAL)
Matrix Invert:	Modified Cholesky Decomposition
Coding of RCs:	Log Area Ratio for RC1 and RC2 and linear for others.

(The Synthesizer: Uses Interpolation and is Pitch Synchronous)

The unit has two processors and two memories (a data memory and a program memory). The LPC-10 analysis code uses only 1238 words of program memory and 2900 words for data memory. It is very fast: 10 LPC (reflection) coefficients are generated in 2 msec, and pitch tracking and voicing analysis are performed in 4 msec for each 22.5 msec window. The processor itself weighs 50 pounds, consumes 180 watts of power, and uses about 180 TTL chips.

II-A. POST FY77 CAPABILITIES AND PLANS

ITT Defense Communications Division has great interest and capability in the automatic speech recognition (ASR) area. ITTDCD is actively seeking government study contracts in ASR, and is also investing several hundred thousand dollars in salary a year of internal funds on research and development in this area. ITTDCD plans to produce an ASR demonstration system early in FY78.

ITT Defense Communications Division has engineering offices and personnel at Nutley, N.J. and San Diego, California. Both of these installations have PDP-11 computers and associated peripherals which are utilized for voice processing research and development. Programs developed at one location are transferred to the other so that both facilities have a total system capability for pursuing research activities at any given time. It is planned that activities in the area of automatic speech recognition will be supported by both ITT facilities, with a majority of the work performed at San Diego.

The equipment, facilities, and personnel presently allocated to automatic speech recognition are the following:

In San Diego, California, ITTDCD has a 2,000 square foot office with 8 scientists and engineers (3 Ph.D's, 5 senior engineers and programmers) and is acquiring a PDP-11/60 with 96K core, and 14 and 88 megabyte disks, tape drive, 3 interactive "smart" CRT terminals, a graphics display unit with hard copy, and UNIX operating system. The principle activity of the San Diego office is ASR research. The personnel in San Diego includes: Dr. George White, formerly of Xerox (7 years research in ASR); Dr. James Dunn (10 years experience in "voice processing"); Mr. Robert Wohlford, formerly with NSA (10 years experience in ASR research); Dr. A. Richard Smith (Ph.D. in computer science from Carnegie-Mellon University whose thesis is on speech recognition); Mr. Russell Lemon, formerly with the USAF (10 years experience in modems and digital voice processing); Mr. John Lowry (formerly with RAND), a recent Masters level graduate from Carnegie-Mellon University (at CMU he worked on CMU's speech recognition project); Mr. George Vensko, formerly with Technology Service Incorporated (6 years experience in signal processing/speech compression), and Mr.

Douglas Landauer, formerly with Pattern Analysis and Recognition Corporation (two years experience programming in signal processing areas).

In Nutley, New Jersey, a comparable computer facility exists with a PDP-11/55 plus peripherals. The group in Nutley is headed by Dr. Marvin Sambur, formerly of Bell Labs (Dr. Sambur has more than 5 years experience in ASR research and is one of the better known personalities in the ASR field. Directly supporting Dr. Sambur in the ASR work will be: Mr. Paul Gilmour (5 years experience in voice processing with his Masters thesis concerned with Formant Tracking); Dr. Walter Fan (extensive experience in developing software utilizing ITTDCD's disciplines of Top Down design and test and structured programming); and Mr. Anthony Russo (headed ITTDCD's team which developed a hardware version for the Navy of Itakura's LPC synthesizer).

II-B. POST FY77 DETAILED RESEARCH PLANS

The following are brief descriptions of several of the areas ITTDCD will be investigating in FY78.

A. DYNAMIC PROGRAMMING

ITT's dynamic programming research has several different aspects. One aspect concerns constraints on the amount of nonlinear time warping that can be performed by dynamic programming. The goal is to insure that the degree of time warping is commensurate with experimentally observed time axis deformations.

It is known that different pronunciations of the same word result in different segmental durations, and that a nonlinear time alignment strategy must be used to match such utterances against standardized templates. However, the degree of temporal variability that should be permitted is not known. For example, some phonemes are characterized by their time rate of change while others are not; e.g., stop consonants are and most vowels are not. It would seem that a time alignment strategy that reflects this fact ought to be better than one that treats an entire utterance with constant constraints on non-linearity regardless of the types of phonemes found in the utterance. Perhaps piece-wise linear matching would be adequate or perhaps strict linearity would be inadequate even for segments as small as phonemes. Research into this problem has been partitioned into the following tasks:

- SPEECH DATA BASE ANALYSIS - to measure the extent to which subword segments exhibit differing degrees of temporal deformation;

- ALGORITHM DEVELOPMENT AND TESTING - to discover computationally efficient means of encoding and utilizing information about the amount of temporal deformation for different speech segments during the classification process:

- AUTOMATIC TEMPLATE GENERATOR DEVELOPMENT - to create a procedure which automatically generates templates that incorporate segment controlled deformation parameters, as well as the usual spectral information;

- PERFORM RECOGNITION EXPERIMENTS - to quantify the improvement in recognition accuracy gained from using segment controlled deformation parameters.

Task 1:

The first task is the study of best match paths through dynamic programming matrices. Dynamic programming matrices contain speech sound similarity scores between time windows of two utterances where the rows and columns of the matrix represent the time in the known and unknown utterances being compared. It is generally observed in dynamic programming matrices that there are broad rectangular regions of good similarity scores connected by narrow paths of good scores.

The goal of the study is to quantify the deviation from linearity for narrow paths and broad rectangular regions.

As a result of this study we will be able to tell how flexible different types of segments can be. We may discover that utterances can be represented by a mixture of flexible and inflexible segments. The inflexible segments would be those that are adequately modeled as straight lines of slope 1. If this is the case, it might not only allow more accurate recognition results, it might permit more compact storage of utterance templates.

Task 2:

The second task of this research is the development of algorithms for representing and using variable template flexibility in dynamic programming pattern matching. In particular, we will investigate the use of A and B values in dynamic programming calculation using the following equation:

$$D_{ij} = S_{ij} + MIN\ (A*D_{i-1j}, B*D_{ij-1}, D_{i-1j-1})$$

Note that when A and B are larger than 1, this forces selection of D_{i-1j-1} which represents movement along a path of slope 1. The proper values of A and B would be determined experimentally and carried along in templates with each time frame to tell the dynamic programming classifier which values of A and B it should use.

Task 3:

The first step in Task 3 is to generate templates that contain average parameters of several exemplars which are time-aligned with dynamic programming. The second step is to encode a deformation parameter expressing the allowable temporal variation for each segment.

Task 4:

The fourth task is the performance of recognition experiments. An attempt will be made to vary the basis vectors and similarity functions along with the constraints on segmental deformation in order to study their interaction.

B. RECOGNITION OF SPEECH DEGRADED BY NOISE

The objective is the determination of an optimum method for reducing the deleterious effects of noise on the accuracy of automatic word recognition systems. Three alternate approaches will be studied.

Task 1:

The first approach involves an investigation of various recognition feature sets (basis vectors) to determine the feature set that provides the highest recognition accuracy under noisy conditions. The feature sets to be examined are:

- Linear predictive coefficients (LPC)
- Vocal tract area functions
- Autocorrelation coefficients
- Cepstral coefficients
- LPC derived Pseudo Formants

The first four feature sets are defined in the recent book by Markel and Grey.[1] The LPC derived pseudo formants are obtained from the LPC coefficients by setting the magnitude of the last coefficient to unity and solving for the pole frequencies of the resulting LPC transfer function.

To evaluation of these feature sets will determine the best recognition set for a wideband quiet environment, for a telephone bandwidth quiet environment, and for a noisy telephone bandwidth environment. The overall optimum feature set will then be selected. Incidentally, as a by-product of this evaluation, an optimum feature set for speaker independent recognition will also be determined.

Task 2:

The second approach involves the use of a noise cancelling filter applied at the input stage of ITT's Kernel recognition system. The method assumes a means for determining a noise signal $w_1(n)$ that is highly correlated with the actual additive noise signal $w(n)$ and uncorrelated with the speech signal $s(n)$. If $w_1(n)$ can be determined, then it can be shown[2] that an adaptive filter can be constructed whose output is a maximum likelihood estimate of $w(n)$, and the signal $z(n)$ is then a maximum likelihood estimate of the clean speech signal $s(n)$. Thus a proper selection of $w_1(n)$ will lead to a filter that effectively removes the additive noise component.

Various schemes for generating $w_1(n)$ will be considered in the study.

These schemes include:

- Setting $w_1(n)$ equal to the average background noise during periods when it is known that no speech is present;
- Setting $w_1(n)$ equal to the updated average signal during periods classified as silence;
- Setting $w_1(n)$ equal to the LPC residual error.

---

1. J. Markel and A. Grey, "Linear Prediction of Speech," Springer Verlag, 1977.

2. Widrow, et al. "Adoptive Noise Cancelling: Principles and Applications." Proceedings of the IEEE, Vol. 63. No. 12, December 1975.

Task 3:

The third task involves the investigation of a noise-reduced LPC parameter set recently proposed by Sambur.[3] This parameter set is determined by subtracting a term proportional to the residual signal power from the diagonal of the autocorrelation matrix used to determine the standard LPC set. The new parameters have been shown to provide a more accurate representation of the speech spectrum in a noisy environment. These parameters should provide a superior feature set for recognition purposes.

C. SPEAKER INDEPENDENT RECOGNITION

In order to develop word recognition algorithms that will be insensitive to the characteristics of individual speakers, signal parameters will be investigated that carry very little information about the speaker. Three such parameter sets are LPC-derived pseudo formants, orthogonal LPC parameters and vocal tract area functions. LPC-derived pseudo formants are obtained from the LPC coefficients by setting the magnitude of the last coefficient to unity and solving for the pole frequencies of the resulting LPC transfer function. Assuming that the LPC-derived transfer function is:

$$H_n(z) = \frac{1}{1-a_1 z^{-1} - a_2 z^{-2} - \ldots a_n z^{-n}}$$

then the pseudo formants are obtained by setting $a_n=1$ and solving for the a_i values and then finding the corresponding poles. The poles are now on the unit circle and have zero bandwidth. This result is a natural consequence of the fact that the last predictive coefficient is a product of all pole moduli of the vocal tract filter. By making the product unity, the individual pole modulus becomes unity, signifying that all poles are located on the unit circle. The pseudo formants are closely related to the actual formants, but unlike the formants, the pseudo formants vary smoothly across an analyzed utterance and can be easily labelled. Due to the normalization of the speaker sensitive bandwidth information, pseudo formants should be more effective than formants in providing a speaker independent representation of a speech word. The ability of pseudo formants to provide speaker independent recognition will be thoroughly examined and compared to other recognition features in ITTDCD's internal recognition system.

---

3. To be presented at Acoustical Society of America's meeting in Miami in December, 1977.

A recent experimental study has shown that by an appropriate eigenvector analysis of the linear prediction parameters, a set of orthogonal parameters are obtained that can be used to achieve a high-quality synthesis of the original utterances. The interesting aspect of these orthogonal parameters is that only a small subset of the parameters demonstrate any significant variation across the analyzed utterance. The remaining orthogonal parameters are essentially constant and, for purposes of synthesis, are completely specified by their measured mean values across the utterance. In a later experimental study it was shown that these remaining orthogonal LPC parameters were associated with the speaker identity and characteristics of the channel. Thus it may be assumed that the orthogonal parameters with the most variation were conveying information about the identity of the spoken words and very little information about the speaker. The speaker independent recognition potential of the higher order orthogonal LPC parameters will be investigated.

Vocal tract area functions are another attractive set of speaker independent recognition features that will be examined. The advantages of the vocal tract area function for use as speaker independent recognition parameters are explained by the fact that the vocal track length can be estimated from analysis of LPC parameters, and then the vocal tract length can be normalized to a standard length.

D. LARGE VOCABULARY RECOGNITION

The representation, storage and retrieval of speech reference data has been, and continues to be, a major problem in automatic speech recognition of large vocabularies. Computational limitations of general purpose computers have led to the emphasis of reference data coding in terms of rules of syntax and phonological rules. A compilation of phonological and syntactic rules is a monumental task that is, at present, far from complete. On the other hand, storage of reference data as single utterance template data is more easily implemented but more difficult to use because of the large amount of memory that must be searched. The solution proposed here is to make pure template information more useful through better information retrieval search strategies for large data bases.

The primary vehicle for speeding up the search process is to have several data bases of differing size and to allow the smaller ones, which are more quickly searched, to control the search of the larger ones. The question of what to put in the smaller dictionaries may be answered in many ways. It is suggested that compressed speech, with differing degrees of compression, be used to fill up the smaller dictionaries. Speech subunits, such as phonemes, may also be used to achieve a more compact data representation as well as well

known parametric speech compression techniques. Storage of template data in nets and trees as used in the HARPY system at CMU will also be investigated.

E. WORD SPOTTING AND CONTINUOUS SPEECH RECOGNITION

The recognition of continuous speech and the ability to "spot words" in a stream of continuous speech is not an unsolvable problem for carefully pronounced speech with a good signal-to-noise ratio. ITT is not planning to research the issues of semantic information processing nor syntactic analysis to support acoustic analysis. However, ITT believes that research into temporal variation and allophonic spectral variations will yield sufficient information to allow useful continuous speech recognition and word spotting systems to be built for cooperative speakers. The inherent temporal variability found in continuous speech utterance can be satisfactorily modeled by new techniques of dynamic programming. The inherent spectral variations caused by coarticulation can be satisfactorily modeled with principle components analysis. Dynamic programming research was mentioned above. As for analysis of spectral variation using the principle components technique, ITT plans to study this area and combine the results with dynamic programming results. A large amount of data will be analyzed to test the power of the resulting techniques.

F. ISOLATED WORD RECOGNITION

The objective is to develop an isolated word recognition system that will serve as a kernel system for research in the other areas mentioned above. The specifications for this system are listed in Table 1. The goal is to implement the system on ITT's fast processor. Dynamic programming will be used in the classifier stage. Perhaps the most interesting aspect of this system is that it could potentially be implemented on a dozen chips costing less than $100 for parts if use is made of a single chip band pass filter bank.

III. GOVERNMENT SPONSORED WORK

ITTDCD has gross revenues in excess of $60 million a year from Government contracts in the area of data and voice transmission and data handling. ITTDCD has contracts with the Navy and NSA for research and development in low bit rate and/or secure voice communication research and development.

TABLE 1

PERFORMANCE SPECIFICATIONS FOR ISOLATED WORD RECOGNITION SYSTEM

(1)	TYPE OF SPEECH	ISOLATED UTTERANCES, COOPERATIVE SPEAKER
(2)	VOCABULARY SIZE	UP TO 50 POLYSYLLABIC WORDS
(3)	ACCURACY	BETTER THAN 99.0%
(4)	TYPE OF SPEAKER	SYSTEM ADAPTATION REQUIRED: I.E. A TRAINING PERIOD IS REQUIRED - NO RESTRICTION ON USERS LANGUAGE OR DIALECT
(5)	RECOGNITION SPEED (USING EXISTING RAM PROCESSOR)	10 SEC. MAX 1 SEC. DESIGN GOAL
(6)	SPEECH INPUT DEVICE	A TELEPHONE HANDSET (BUT NOT OVER DIAL UP PHONE LINE)

BIOGRAPHICAL SKETCH

George M. White

George M. WHITE, Manager of San Diego Laboratory, ITT Defense Communications Division.

George M. White graduated magna cum laude in physics from Michigan State University in 1964. He received Danforth and Woodrow Wilson Fellowships for graduate studies at the University of Oregon and the University of California at Santa Cruz. He received his Ph.D. degree in theoretical chemistry from the University of Oregon in 1968. He spent two years, 1968 to 1970, at Stanford University Artificial Project on a National Institutes of Health post-doctoral study fellowship. In 1970, he joined the Xerox Palo Alto Research Center where for the next seven years he managed Xerox Corporation's automatic speech recognition project; designed many internally constructed recognition systems, some of which achieved notably high recognition scores for English; monitored speech processing technology for Xerox Corporation; advised Xerox venture capital groups on corporate acquisitions of speech technology firms. In 1977, he joined ITT Defense Communications Division where he is Manager of the ITTDCD West Coast office, which carries on research and development projects in speaker verification, speech recognition and speech compression.