MORRIS, L. ROBERT.
--Audio information retrieval and
analysis in the wired city.

CARLETON UNIVER

DIVISION OF SYSTEMS ENG

checked
Jan/84

OTTAWA CANADA
K1S 5B6

FACULTY OF ENGINEERING

(2)

AUDIO INFORMATION RETRIEVAL AND ANALYSIS IN THE WIRED CITY

Contract OSR 3-0090

1973-1974

L. Robert Morris
Associate Professor of Engineering

Jean-Pierre Paillet
Associate Professor of Lingusitics

Contents

Audio Information Transmission and Retrieval in the Wired City

1. Background (Statement of Work as per Contract)

This project is specifically directed to the application of systems of audio/digital format translation to the Wired City and educational technology programs of the DOC. It will study the feasibility and cost of applying and approving existent hardware and software systems developed for the Ontario Department of Education to the Wired City and will build a demonstration using an existent computer system.

1.1 Review (A Summary of Mid-Year Report)

Speech communication requirements have (historically) profoundly influenced the development of most presently available communication facilities, the telephone network in particular. However, telephone networks now carry a heterogeneous traffic load: 3 KHz speech plus images, data and control information. Thus, while the original design premises of the system are now invalid, the possibilities of new services (and new versions of old services) promised by digital technology and communications/computer/network theory are not yet realized.

Service modes of "new" communications systems include real-time transmission (voice, images, data, and music -- broadcast or via telephone), store-and-forward (messages, word processing, and text manipulation), "canned" programmes (Muzak, etc.), audio response (telephone query, educational systems), and speech recognition (speaker and word identification).

This research report mainly concerns;

a) techniques for machine/person and person/machine interaction in the Wired City incorporating:

   i) text storage and retrieval;

   ii) audio message storage and retrieval.

b) Teaching devices incorporating speech/graphic input/output.

We will first report on those techniques considered and implemented for a stand-alone demonstration of voice information retrieval techniques.

2. Voice Storage and Response Techniques

2.1 Background of Department of Education Project

The Carleton University Language Recoder [1], [2], [3], [4] has, as one feature, voice response to optical input (after character recognition) and to manipulation of ASCII data (e.g., typed in or stored on disk).

The technique selected (in 1972) for speech parameter extraction, storage and "re-creation" was linear predictive analysis/synthesis. A major part of this report was to have been a justification of this technique and a comparison with other existing methods. During the current research report year (i.e., September 1973 - August 1974) three papers (considered by the principal investigators to be of high quality) have appeared which embody most of the ideas considered significant in our choice of "technology" for voice storage and retrieval:

i) Voice signals: bit by bit [5];

ii) Digital coding of speech waveforms: PCM, DPCM and DM quantizers [6];

iii) Automatic voice response: interfacing man with machine [7].

The appearance of these papers at this time emphasizes the relevance of the current project.

These papers cover the relevant considerations in far more detail than we had proposed. Thus we include these papers as appendices A, B, C and devote the next section to a concise examination of extant methods of digital voice analysis-transmission-synthesis.

2.2 Voice Analysis-Transmission-Synthesis Techniques: Waveform-Based

2.2.1 Waveform Digitization (PCM)

Nyquist's theorem dictates that sampling a signal bandlimited to ± W Hz is lossless if the sampling rate exceeds 2W samples/second. For a finite information transmission rate, these samples must be quantized into one of 2^B levels giving an overall information rate of 2WB bits per second. It can be shown that increasing B by unity approximately increases the quantized signal's signal-to-quantizing noise-ratio (SQNR) by about 6db [6]. Usually 10 bits is deemed sufficient to produce digitized speech with an almost imperceptible degradation due to quantizing noise. Thus, for 3.5 KHz speech (our base), 70,000 bits/second are required.

Speech transmission with $4 \leq B \leq 6$ can be made (perceptually) more acceptable by "dithering" [6]. Nonuniform quantization (companding) can produce 7-bit speech of similar quality to 11-bit uniformly quantized speech and adaptively varying the step size can produce similar improvements. These techniques exchange algorithm complexity for improved SQNR.

In differential PCM (DPCM), the derivative of the signal (or a function of the present signal sample and the previous n signal samples) is quantized. An increase in SQNR of 6db can be realized. Adaptive DPCM (ADPCM) can realize a further improvement of 6db while "predictive" ADPCM techniques can yield SQNR improvements over PCM of 20 db.

2.2.2 Delta Modulation

Delta modulation (DM) involves "sampling" an input signal at a rate much higher than the Nyquist rate. The "sampling" involves noting the sign of the difference between the actual signal, $s(t)$, and a staircase approximation to that signal, $\hat{s}(t)$. The staircase is then incremented by $\mathrm{sgn}\,[s(t)-\hat{s}(t)]*\Delta$. The transmitted information consists of $b_T = \mathrm{sgn}\,[s(t)-\hat{s}(t)]$. DM is accompanied by slope overload distortion and granular noise, and the SNR increases approximately as $f_o{}^3$; i.e., about 9db per octave frequency increase. As with PCM, adaptive techniques (ADM) increase SQNR at the expense of algorithm (and thus hardware) complexity. For .2-3 KHz bandpass speech, log PCM is superior to ADM for bit rates less than 50 Kbit/sec, while ADPCM has a 12db SQNR gain over log PCM at all bit rates [6].

2.2.3 Summary - Waveform Digitization Techniques

For .2 -3 KHz bandlimited speech, the following SQNR's are noted:

$$\left.\begin{array}{r}\text{i) ADPCM - 37db}\\ \text{ii) log PCM - 25db}\\ \text{iii) ADM - 25db}\end{array}\right\}\ \text{@ 50 Kbits/sec rate}$$

The cassette made available with this report includes samples and comparisons of PCM, DPCM, and DM as well as illustrations of perceptual factors in the different techniques and the effects of transmission errors.

A computer based voice response system (10 users) has recently been described by Bell Labs [7]. The system uses 24 Kbit ADPCM to yield 3-KHz speech with a 20db SQNR. This system will be compared to Carleton's Language Recoder in section 4.1.4.

2.3 Voice Analysis-Transmission-Synthesis Techniques: Articulatory-Based Techniques

Articulatory-based techniques involve waveform analysis to compute parameters which relate to a simplified model for the <u>production of the waveform</u>. The waveform is then recreated by simulating the production system. The advantages gained from these techniques relate to the fact that the parameters of the human voice production system are relatively slowly changing; i.e., the bit rate necessary to describe the parameters required to have the model produce high quality speech is low compared to the bit rate of the speech itself. For example, 3 eight-bit formant (or resonance) positions, a 7-bit pitch period length parameter, and a 10-bit gain constant are sufficient to recreate a 12-bit, 5-KHz pitch period lasting about 10 msec. That is 41 bits/10msec = 4100 bits/second serve to produce speech with a SQNR of about 64db.

Furthermore, (and more important for future extensions) the coding virtually separates the nearly independent components present in human speech. Thus, from a stored set of coefficients, it is possible to produce speech incorporating changes in segmental quality, pitch, stress and rythm (see section 5).

2.3.1 Formant-Based Techniques

As suggested above, formant-based techniques involve transmission, for each pitch period (or, alternatively, at a fixed-frame rate), of three or more resonance values, a gain factor and a pitch period length, for voiced sounds [8]. Formant-based techniques embody (probably) the lowest bit rates possible concommittent with high quality speech for a given bandwidth. A commercial digital formant synthesizer, requiring a computer to update parameters, has been on the market for some time at a price of about $8000 [9].

2.3.2 Waveform Analysis/Synthesis Techniques

Itakura and Saito [10] and Atal and Hanauer [11] have described algorithms which involve time domain speech waveform analysis and produce "prediction coefficients" which permit computationally efficient resynthesis of the analyzed speech. For voiced speech, this technique is related to formant-based analysis/synthesis [11], [12]. Indeed, formant positions can be converted into synthesis coefficients via polynomial expansion [12]. For unvoiced sounds, predictive techniques enable synthesis without a change in synthesizer structure. The direct form of a predictive synthesizer allows synthesis with the fewest number of multiplications for a given order prediction filter. As is explained below (section 4.1.2), this method has been adopted for use in the Carleton System. The predictors can be also transformed into K-parameters which enable lower bit rates for transmission, but require more complicated hardware for synthesis. An overview of the present state of "predictor-based" analysis/synthesis is given by Markel [13], who has carried out the most extensive work in this area during the last two years. The fact that "predictors" are related to the shape of the vocal tract [14] justifies our inclusion of this technique as articulatory-based.

2.3.3 Summary - Articulatory-Based Parameter Techniques

These techniques offer low bit rates. Thus, for 4 KHz, 64db SQNR speech of approximately the same (high) quality, the following bit rates are required:

i) Formant based - 4100 bits/sec;

ii) K-parameter transmission, frame by frame - 9750 bits/sec;

iii) Hybrid (see [12]) - 8320 bits/sec;

iv) Direct pitch-synchronous predictor encoding - 19200 bits/sec; (10-16 bit coefficients, 2-16 bit gain etc. words per pitch period, 100 pitch periods/sec.)

Note that the lower the bit rate, the more complicated the synthesizer for a given bandwidth and SQNR.*

3. Real-Time Voice Analysis-Transmission-Synthesis Techniques

Although we are primarily interested in audio information retrieval and transmission as opposed to real-time voice communication, it is interesting to note that the techniques described above (detailed in the references), are at the moment (summer 1974), in the process of being implemented so as to yield low bit-rate, high quality real-time voice analysis-transmission-synthesis. Algorithms developed and tested via software simulation at

a) Speech Communications Research Laboratory, Santa Barbara, California.

b) GTE Sylvania, Massachusetts.

c) Bolt, Beranek and Newman, Massachusetts.

are in the process of being implemented in hardware (by different manufacturers) with the support (mainly) of the U.S. armed forces.

* The cassette made available with this report includes samples of formant and predictive coded speech.

These algorithms not only are high quality, low bit-rate, but are amenable to coding for secrecy. Thus, just as the FFT algorithm gave rise to, first, low speed real-time software implementations and, then, high speed FFT hardware boxes, we may expect the current speech analysis/synthesis research and software implementations to give us hardware "VOCODERS", probably within 2-3 years.

Finally, one could extrapolate to a "Wired City" where all speech information transmission is carried out at the "parameter" level with coders at the transmitter and receiver (e.g., an analyser in the microphone and a synthesizer in the earphone).

4. The Carleton University Language Recoder

The language recoder is designed to translate from one form of language to another. For example, the two primary implementations of the recoder are:

i) Text In, Spoken Speech Out
Character strings derived either from optical character recognition, typewriter input, or stored data are converted to spoken speech.

ii) Speech In, Graphics or Tactile Response Out
Speech is analyzed and either visual or tactile images are produced.

Our current research in the area of item ii) involves only graphical output and will be described in section 6.

4.1 The Reading Machine/Audio Text Editor

4.1.1 Design and Implementations

Low speed, high accuracy, low cost optical character recognition is now practical [19]. The output of such a device -- a string of ASCII characters -- can be input to a system designed to translate characters into speech. Grouped characters yield spoken words or, for a finite vocabulary (see section 5), spoken speech X% of the time and spelled speech (100-X) % of the time.

The first version of the language recoder, Carleton Audio Text Editor (CATE), has an 1100 word vocabulary with X = 87 for non technical text. The speech is produced by a specially designed 41-user hardware synthesizer. The synthesizer uses 10 prediction coefficients (16 bits each) per pitch period, with 10 bits per speech sample output, and can service 41 users simultaneously. The design of the synthesizer was preceeded by real-time software simulation [16] and the construction of a real-time single user hardware synthesizer [12]. It is a characteristic of digital hardware that the hardware must give precisely the same results as the simulation since all that is involved is integer arithmetic operations. The synthesizer is serviced by a PDP 11/45 computer with dual RK05 removeable 1.2 million word disc packs. Storage of the initial 1100 word vocabulary occupies approximately one half the available disc storage. The problems of storage and retrieval of speech synthesis parameters for 41 users was examined at length by D. Sproule in a Carleton 1974 M.Eng. thesis [17]. Interfacing is via multi-user serial communications multiplexers.

4.1.2 Cost

The main hardware costs for the design system are as follows:

PDP 11/40 with dual RK05 discs and 48 K memory:	$40,000
Communications @ $500 user:	$20,000
Synthesizer	$10,000
	$70,000

This represents a per user cost of $1,700. However, the system is designed to provide the user with computing facilities as well as audio response to all I/O routed through the computing facility. That is, the user will see a time share computer system which intercepts all I/O and provides audio response. Speech and data can be transmitted simultaneously over telephone lines using a GTE Lenkurt 5249A speech plus data panel.

4.1.3 Status

As of August 1, 1974, the first implementation of the system is nearing completion. At present, a GT-40 (PDP 11/10 + Graphics) simulates 41 users typing statistically correct English, and feeds the PDP 11/45 system with the characters generated. The system fetches "words" from disc for all the users and synthesizes output for any one of the 41 users using software. The CPU overhead required for software synthesis for one user is more than that required to feed the hardware synthesizer which then synthesizes simultaneously for all users. The hardware synthesizer is nearing completion and its primary components have already undergone extensive computer testing.

The first "complete" hardware version will offer audio response to typed input of 41 users, while the second version will add basic editing facilities. A "final" version will intercept all ASCII I/0 connected with normal use of a multi-user time share computer facility. Note that all versions above will offer 87% pronounced, 13% spelled speech. Development is proceeding simultaneously in the linguistic direction as per section 5. That is, the speech offered in each of the facilities above will be improved to incorporate those manipulative techniques necessary to add suffixes, prefixes, emphasis etc. until the final result is "sentences" rather than strung together pronounced and spelled words.

Note the following:

a) i) The first version will suffice for the purpose of teaching touch typing to the blind.

ii) The second version will suffice to enable the blind to use touch typing to produce 'perfect' correspondence.

iii) The third system will enable the blind to use all facilities of a normal computer time sharing system.

b) All versions will use the same hardware: synthesizer and computer. Improvements will generally require new software and, perhaps, additional CPU power for linguistic manipulation. The PDP 11 computer allows for multiple CPU configurations.

4.1.4 Comparisons with Other Schemes

The current implementation represents a trade-off between synthesizer complexity and bit rate. Twelve 16-bit words are required per pitch period for an overall bit rate of 19.2 Kbits/second. This system produces 10-bit 4-KHz speech (about 52db SQNR).* The synthesizer is also designed to accommodate the hybrid synthesis scheme outlined by Morris in [12]. This provides for a bit rate of 5600 bits/second for voiced sounds and 19200 bits/second for unvoiced sounds for an overall bit rate of (0.8 x 5600 + 0.2 x 19200) 8320 bits/second based on average frequency of voiced and unvoiced sounds. More important is the fact that only articulatory-based

* The Bell Labs 10 user ADPCM system recently described requires 24 Kbit to produce 4 bit, 20db SQNR 3 KHz speech [7]. It is not amenable to linguistic manipulative techniques.

systems can provide for the linguistic manipulation necessary for connected speech production.

4.2 Applications to the Wired City

The general applications of low bit rate, high quality audio information retrieval to the Wired City were discussed at length in the mid-year report. Together with the developments noted in section 3, (re: real-time analysis-transmission-synthesis) the possibilities now exist for:

a) low bandwidth real-time digital speech analysis-transmission-synthesis with easy coding capabilities;

b) economical, high quality, audio information retrieval from on-line information storage systems, via telephone lines.

We emphasize that the capabilities exist now for single user fairly high quality "sentence" audio information retrieval. The current Carleton linguistic effort is towards expanding the current economical high quality multi-user "word" machine to a "sentence" machine via economical (in machine time) algorithms and to improving the "fairly high quality" to "very high quality".

5. Linguistic Aspects of Audio Information Retrieval

Linguistic analysis provides the basis for improvement of audio information retrieval systems in four main areas:

i) The knowledge of linguistic types of redundancy in the speech signal allows the development of synthesis algorithms which further increase the reduction of storage and/or transmission bit-rate;

ii) Linguistic rules of morphology, syntax, segmental and suprasegmental phonology, and eventually semantics are the basis for manipulation of stored parameter values for the production of new message items.

iii) The transcoding from written to spoken speech (and, eventually, the converse transcription) requires prohibitive amounts of storage if it is not based on the linguistic regularities of the written language.

iv) Linguistic information (at all the levels listed in ii)) is essential for any scheme of connected speech recognition and understanding.

We take up the consideration of these four areas in the order given. A differently oriented survey can be found in [27].

5.1 Speech Synthesis by Rule

The designation "speech synthesis by rule" is commonly applied to any scheme of synthesis based on the use of parameter values obtained not by direct analysis of samples of human speech, but rather by computation according to linguistic regularities of the speech signal. Consequently, any scheme of speech synthesis by rule will be comprised of two sections, the second of which is an algorithm of speech synthesis (see section 2.3) and the first of which is an algorithm for the linguistically-based computation of the parameter values needed, from an input which normally consists of a string of linguistic symbols. Assuming a "reading-rate" of up to five syllables per second, and an average of three five-bit characters of printed text per syllable, the ultimate bit-rate achievable is of the order of 5 x 3 x 5 = 75 bits per second of spoken speech synthesized from printed text. However, a scheme of speech synthesis by rule which would realize such a rate requires incorporation of information from all linguistic levels (including the semantic and the "pragmatic" levels). Several intermediate schemes are possible, which rely on knowledge from the lower levels of linguistic regularity.

5.1.1 Segmental Synthesis

It is possible in first approximation to separate the treatment of segmental material ("phonemes") from that of suprasegmental information (pitch, rhythm, stress patterns) and to generate the parameter values for synthesis from an input consisting of a string of phonetic segments (given in some appropriate alphabet) and added markers giving the suprasegmental information (stress marks, pitch contours, timing).

i) The computation of parameter values is based essentially on obtaining continuous formant trajectories (or the trajectories of equivalent parameters) from a discontinuous string of nominal values characterizing each phonetic segment. In the simpler schemes, segments can be characterized directly by nominal values of the formants. While these nominal values have a straightforward interpretation in the case of steady-state sounds (e.g., vowels or fricatives) their use in the case of transient phonetic events (e.g., stops) is based on the locus theory of consonants [20], [21]. It is possible to obtain a fairly intelligible approximation of human speech by this simple method, provided great care is applied in the compilation of the table of nominal values [24].

ii) More elaborate schemes rely on a simulation of the human articulatory apparatus. Each phonetic symbol is then seen as indicating a target articulatory position or even, as in [25], a higher order neurophysiological correlate, and the shapes of the vocal tract throughout the utterance are interpolated as the response of a damped system representing the dynamics of the various articulators. It is then possible to compute the continuous by varying frequencies of the first few formants (which are the main resonance frequencies of the varying vocal tract), or to derive linear predictive synthesis parameters, or even to solve directly for the output waveform. The main advantage of such articulatory-based schemes is that they allow in principle for more flexibility in the timing of speech, which is computed in each particular case, rather than given in a general table as in the simpler schemes.

This naturally requires much more computational power than simple formant-interpolation techniques. Moreover, in the present state of the art, vocal tract simulation does not produce results as good as those obtained by the simpler method. This is mainly due to the enormous amount of data required to determine the values of the many parameters characterizing the dynamics of the vocal tract (see e.g., [31], for the importance of aerodynamic factors). Continuing work at Bell Labs [22], [23], has already produced impressive results, based on experimental simulation of paragraph-length utterances. This work also involves synthesis by rule of suprasegmental features.

5.1.2 Suprasegmental Synthesis

While segmental synthesis can be undertaken on the basis of a simple phone-string input, the computation of suprasegmental features requires information from the syntactic and possibly semantic levels of language.

Most linguistic studies on intonation (i.e., roughly, patterns of fundamental frequency variation) and stress (prominence of syllables marked by pitch breaks, amplitude variations, and lengthening), e.g., [29], indicate that these features of speech are correlated either with syntactic grouping of words into phrases or directly with certain features of meaning.

i) While it is premature to attempt to take meaning directly into account for speech synthesis, it is possible to incorporate some semantic information into it. For instance, the syntactic algorithm of Teranishi and Umeda, [25], distinguishes between several classes of words which have different stress properties, correlated with their semantic importance.

ii) Syntactic analysis algorithms come in very many different shapes and formats. The most useful for speech synthesis are rather simple. Their main output is a parsing (if possible unambiguous) of the input (alphabetic text) into phrases, an assignment of a hierarchy of stress marks, and a decision as to the basic intonation pattern to be used. Associated with the algorithm is a dictionary providing both syntactic information (grammatical categories) and a phonetic description of each item. Thus the output of the syntactic analysis can be put in the form of a string of segmental symbols and suprasegmental markers for input to an algorithm of formant-based or articulatory-based synthesis. The most outstanding realization in this area is the above mentioned work by Coker, Umeda and Browman at Bell Labs [24].

5.2 Manipulation of Speech Items

The remarkable work of the past few years on synthesis by rule has overshadowed certain less ambitious possibilities offered by certain schemes of synthesis (see section 2.3) which allow for the manipulation of suprasegmental parameters on the basis of a constant store of segmental strings. Given a basic vocabulary stored in the form of segmental synthesis parameters (e.g., linear predictive parameters obtained by analysis of spoken utterances) it is possible to construct intelligible messages by stringing out some of these stored words and supplying suprasegmental information [28], [32]. While the automatic reading of a printed text requires all the complex apparatus detailed in section 5.1, the scheme outlined here would be sufficient for many voice response systems, where the information is stored in coded form and requires a rather restricted vocabulary. What is needed in this case is an algorithm for *syntactic* synthesis, which need not cover all the syntactic patterns of the language (as is needed if one is to *read* from unrestricted text) and may even in many cases reduce to a collection of fixed patterns, into which the required words are fitted according to the information to be transmitted: this would be the case, e.g., for a telephone information service, a stock-market quotation system, an airline answering service, etc.). The syntactic synthesis algorithm provides the grouping of words into phrases, and the dictionary supplies the segmental description of the words. The resulting synthesized messages, although distinctly "machine-like", can be perfectly intelligible.

5.3 Graphemic to Phonemic Transcoding

The *word* is not the most economical unit for storage in a dictionary. The number of lexical items of English runs to several hundred thousand (most of which, fortunately, having low frequency of occurrence) and possibilities for inflexion, however reduced in English, multiply this number again by a factor of two or three. It is thus quite impractical to have an unrestricted dictionary of word-forms for the purposes of synthesis.

5.3.1 Morpheme Analysis

However, most words in the inventory of English are derived words: i.e., they are composed of parts (morphemes) which recur in many words and whose total number is much more restricted (possibly a few thousand).

It is thus possible to rely on algorithms to isolate the component morphemes of words (in their written forms) and recompose the phonetic form of the words from the phonetic forms of the morphemes. J. Allen [26] has carried out the most successful work on this problem.

5.3.2 Pronunciation Rules

It is possible further to reduce the size of the dictionary by excluding from it all the words whose pronunciation is predictable. Pronunciation rules are, in a way, the reverse of spelling rules; they are (unconsciously) used by most speakers of English when encountering an unknown word. Certain combinations of letters (e.g., -th-, -ps-, -ph-) have regular pronunciations from which the pronunciation of whole words or morphemes can be reconstructed. The dictionary is then reduced to a list of exceptions to the pronunciation rules. Here also, the best reference is to the work of J. Allen [26].

5.4 Speech Recognition (Understanding)

5.4.1 Limited Vocabulary

Many schemes have been proposed for the recognition of limited vocabularies. In most of them, the main problem is that of tuning the system to the speaker. Some proposals solve this problem by using recognition criteria which are largely independent of the speaker, but depend on the very small size of the vocabulary (e.g., the set of 10 digits). As a matter of fact, the problem of automatic tuning to a speaker is far from being solved, and one might even say that the phonetic and linguistic theoretical basis for a solution is nonexistent.

5.4.2 Connected Speech

In the case of connected speech, the problem is compounded by the fact that words receive notoriously varying emphasis, and that the quality of various phonetic segments depends very much on their position with respect to stress points, phrase boundaries, etc.

The recognition of speech items within the context of connected speech thus requires automatic tuning not only to the speaker, but also to the (syntactic and semantic) context. This was not feasible at all until recently. Several projects are now underway to try out various solutions to the problem of context-tuning. Most of them are based on the principle of recognizing "expectable" words within the running speech, and working from these words to the rest of the utterance. Words are "expectable" if they fall within the usual vocabulary of the (known) subject-matter of the speech. Such words are likely to have prominence in the speech, and are expected to occupy certain syntactic positions. These expectations can be variously incorporated in algorithms which make hypotheses about the syntactic structure of the speech and check the phonetic features predicted from these hypotheses against the actual input. This type of approach, essentially based on knowledge of subject-matter and on the selection of (semantically and pragmatically) important words to work "down" to phonetic detail justify the more ambitious, but more accurate label of Speech Understanding.

6. Educational Technology

We have noted that one implementation of the Language Recoder is speech in, graphics out.

Further, our experience in developing such systems have led to the implementation of a Mobile Instructional Computer (MIC) for production of visual animations in classroom demonstrations. Both these realizations are, we believe, advanced examples of applications of computers to educational technology.

6.1 Graphical Aids to the Hearing Handicapped

Lack of audio feedback impairs language acquisition in the hearing handicapped. Numerous attempts have been made to process speech so as to produce graphical displays which compensate somewhat for this lack of audio feedback.

6.1.1 Pitch Trajectories

Voiced sounds in speech are produced by quasi-periodic excitation of a slowly varying vocal tract configuration. The reciprocal of the period of excitation is termed "pitch" and lack of control of pitch is responsible for one aspect of the poor quality of "deaf" speech. If a target pitch trajectory could be displayed for a particular sound or sound sequence, and the actual pitch trajectory extracted (in real-time) and displayed for comparison, then compensating visual feedback is available. Until recently, real-time accurate pitch extraction was impractical (see [33] for a review). In the last two years at least three practical techniques have been developed.

The first, [34], [35], involves measurement of the varying impedance of the glottis by ultrasonic techniques, and involves strapping electrodes on the neck. Two others involve extraction and processing of error waveform via linear predictive techniques. As originally suggested by Atal [17], the techniques involves matrix inversions and was refined by Markel [36]. This algorithm requires a high speed computer. J.N. Maksym, of Carleton University, showed that the matrix techniques could be replaced by adaptive gradient algorithms which could be economically implemented in digital or analog hardware [33].

As of August 1974, pitch displays have generally been available mainly to those possessing a Fourcin "Larynograph", and Fourcin et al. have had much success in field applications in instructional environments. At present, a comparison of the various algorithms is underway at Carleton (under the DOE grant) with a view to making the displays economically available, either as a very cheap hardware device or as part of a package on a general purpose computer. It is this latter implementation which represents a novel use of state of the art computer/display technology for education.

6.1.2 Vocal Tract Cross Sections

Speech has its origins in a set of motor commands which cause the vocal tract, tongue, velum, jaws etc. to assume a certain configuration. For vowels, the shape of the vocal tract gives rise to resonances which result in formants, the main perceptual attribute of vowels. Although more than one vocal tract configuration can produce a given vowel, it has been demonstrated that normal speakers assume a "prototype" configuration for a given vowel [37]. If this prototype "target" configuration could be displayed superimposed upon the actual vocal tract configuration as a sustained vowel is uttered, then visual articulatory feedback would be present to substitute for the absent auditory feedback in a hearing impaired child. During the past five years, the theory of vocal tract cross section recovery has been explored [38], [39], and it is now possible to extract a good approximation in near real-time on a general purpose minicomputer. Crichton and Fallside [40] have implemented this algorithm on a dual computer system (PDP-8, Interdata 80) and have had success in using the display for training.

6.1.3 Other Displays

Other displays found useful in deaf training are formant trajectories, and envelope vs. time.

6.1.4 Summary

Several graphical displays, obtained via speech analysis, have been found useful as aids in education of the hearing impaired. These displays have been implemented in hardware and software and are available to relatively few educators. At Carleton we are taking advantage of the recent availability of an economical minicomputer/graphics processor to implement these primary displays in a single package involving no specialized hardware. Thus:

a) it will be available to anyone purchasing the required general purpose computer/graphics processor;

b) because the algorithms are implemented in software, they can be updated and new algorithms introduced.

<u>This, we believe, is the key to cost efficient application of educational technology: general purpose machines capable of infinite extensions of algorithms via software "produceable" by anyone and "playable" by all</u>.

Historically, this <u>was</u> the case. The slide projector and the tape recorder are the prime examples of machines which satisfy the requirement of universal "playability" and software "produceability". The SONY Videocassette machine is fast approaching this status in certain geographical areas (e.g., the Carleton campus). However, recently -- as evidenced by a perusal of goods available at the Canadian Educational Showplace in Toronto -- the trend has been towards specialized incompatible hardware which are either unifunctional or can only "play" certain software.

Thus, at Carleton, we are attempting experiments which may determine the viability of general purpose computer/graphics processors as universal educational tools used in much the same manner as a slide projector or tape recorder.

7. Summary

This contract report has included a state-of-the-art survey of methods of speech analysis-transmission-synthesis, and linguistic and signal theoretic manipulation of stored data so as to produce spoken "audio information retrieval" of the data. The contract has (in small part) financed a prototype system which, at this time (August 1974) is -- in design details and capability -- more advanced than any reported in the literature, or known to us. (The principal investigators have spoken to most of the authors noted in the references). The system is nearing completion of the first phases and is designed so that it could be linked to the present Wired City Laboratory for tests of the general concepts outlined. Demonstrations of the system(s) will be arranged.

A by-product of the DOE project has been the acquisition of competence and experience in application of new technology to special education. We believe such techniques have wider applications in general education. At minimum, the portable computer/graphics processor is capable of producing sophisticated, informative and interactive extensions of the traditional black-board diagram.

References

[1] Department of Education grant application, 1972-73, March 1972.

[2] Department of Education grant application, 1973-74, March 1973.

[3] Audio Information Transmission and Retrieval in Wired City, pp. 7-9, January 1971.

[4] Department of Education grant application, 1974-75, March 1974.

[5] J.W. Bayless, S.J. Campanella, and A.J. Goldberg, "Voice signals: bit-by-bit", IEEE Spectrum, vol. 10, no. 10, pp. 28-34, October 1973.

[6] N.S. Jayant, "Digital coding of speech waveforms: PCM, DPCM and DM quantizers," Proceedings of the IEEE, vol. 65, no. 5, pp. 611-632, May 1974.

[7] L.H. Rosenthal, L.R. Rabiner, R.W. Schafer, P. Cummiskey and J.L. Flanagan, "Automatic voice response: interfacing man with machine", IEEE Spectrum, vol. 11, no. 7, pp. 61-68, July 1974.

[8] L.R. Rabiner, "Digital formant synthesizer for speech synthesis studies", J. Acoust. Soc. Am., vol. 43, pp. 822-828, April 1968.

[9] L.R. Rabiner, L.B. Jackson, R.W. Schafer and C.H. Coker, "A hardware realization of a digital formant speech synthesizer", IEEE Transactions on Communications Technology, vol. COM-19, pp. 1016-1020, December 1971.

[10] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies", Electron. Commun. Japan, vol. 53-A, pp. 36-43, 1970.

[11] B.S. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Am., vol. 50, pp. 637-655, August 1971.

[12] L.R. Morris and M.J. Gough, "An economical hardware realization of a digital linear predictive speech synthesizer", IEEE Trans. on Communications, vol. COM-22, no. 1, pp. 82-88, January 1973.

[13] J. Markel, "Linear prediction of speech - theory and practice", SCRL Monograph No. 10, September 1973.

[14] H. Wakita, "Estimation of the vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion techniques", SCRL Monograph No. 9, July 1972.

[15] H. Bar Lev, "The Transicon", Proceedings, 1973 Carnahan Conference on Electronic Prosthetics, Lexington, Kentucky, November 1973.

[16] L.R. Morris and J.P. Paillet, "Real-time software speech synthesis on minicomputers", Proceedings of 1972 International Conference on Speech Communication and Processing, pp. 166-169, Boston, April 1972.

[17] D.E. Sproule, "Response time prediction and optimization for stored data in speech synthesis", M.Eng. thesis, Department of Systems Engineering, Carleton University, June 1974.

[18] Flanagan, James L. & L.R. Rabiner, 1973, Speech Synthesis, Stroudsburg, Pa., Dowden Hutchinson & Ross, Inc.

[19] Preprints of the Speech Communication seminar, Stockholm, August 1974.

[20] Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., Gerstman L.J., "Some experiments on the perception of synthetic speech sounds", JASA 24, 1952. Reprinted in [18] pp. 67-76.

[21] Fant, G., "The acoustics of speech", Proc. Third Int. Congr. Acoust., 1959. Reprinted in [18] pp. 77-90.

[22] Coker, C.H., "Speech synthesis with a parametric articulatory model", Speech Symposium, Kyoto, 1968. Reprinted in [18] pp. 135-139.

[23] Holmes, J.N., Mattingly, I.G., Shearme, J.N., "Speech synthesis by rule", Language and Speech 7, 1964. Reprinted in [18] pp. 351-368.

[24] Coker, C.H., Umeda, N., Browman, C.P., "Automatic synthesis from ordinary English text", in [18] pp. 400-411.

[25] Teranishi, R., & Umeda, N., "Use of pronouncing dictionary in speech synthesis experiments", Proc. Sixth Int. Congr. Acoust., Tokyo, 1968. Reprinted in [18] pp. 412-415.

[26] Allen, J., "Speech synthesis from unrestricted text", in [18] pp. 416-428.

[27] Flanagan, J.L., Coker, C.H., Rabiner, L.R., Schafer, R.W., Umeda, N., "Synthetic voices for computers", IEEE Spectrum 7, (1970). Reprinted in [18] pp. 432-455.

[28] Rabiner, L.R., Schafer, R.W., & Flanagan, J.L., "Computer synthesis by concatenation of formant coded words", Bell System Technical Journal 50 (1971). Reprinted in [18] pp. 479-496.

[29] Lawrence, W., "The phoneme, the syllable, and the parameter track" in [19] pp. 173-178.

[30] Hiki, S., and Oizumi, J., "Speech synthesis by rule from neurophysiological parameters" in [19] pp. 219-226.

[31] Scully, C., "A synthesizer study of aerodynamic factors in speech segment durations" in [19] pp. 227-234.

[32] Olive, J.P., "Speech synthesis by rule" in [19] pp. 255-260.

[33] J.N. Maksym, "Real-time pitch extraction by adaptive prediction of the speech waveform", IEEE Transactions on Audio and Electroacoustics, vol. AU-21, pp. 149-154, June 1973.

[34] A.J. Fourcin, and E. Abberton, "First applications of a new laryngograph", Medical and Biological Illustration, vol. 21, no. 3, pp. 172-182, July 1971.

[35] A.J. Fourcin, and E. Abberton, "A visual display for teaching intonation and rhythm", English Language Teaching Document (73/5). English Teaching Information Centre, London, England.

[36] J. Markel, "The SIFT algorithm for fundamental frequency estimation", IEEE Transactions on Audio and Electroacoustics, vol. AU-20, pp. 367-377, December 1972.

[37] G. Fant, Acoustic Theory of speech production, Mouton, 1960.

[38] E. Stansfield, "An articulatory model for speech recognition", Ph.D. Thesis, University of London, 1971.

[39] H. Wakita, "Direct estimation of vocal tract shape by inverse filtering of the acoustic speech waveform", IEEE Trans. on Audio and Electroacoustics, vol. AU-21, pp. 417-427, October 1973.

[40] R.G. Crichton and F. Fallside, "The linear prediction model of speech production with applications to deaf speech training", Proceedings, Stockholm Speech Seminar, August 1974.