Speech Output For Microprocessor Based Products

H.A. Cohen
La Trobe University

The first generation of speech chips utilise phoneme generation, fairly simple minded continuously variable delta modulation, linear predictive coding, or rather sophisticated data compression by removal of "nearly redundant" information. The designer wishing to add the now affordable feature of voice output to microprocessor-based products must chose between chips with different memory requirements, ease of programming, speech intelligibility, and speech character/acceptability.

HUMAN SPEECH

Speech output is now an affordable option for micro-processor based products ranging from auto-motive instruments to educational toys. The aim of this paper is to review the first generation of speech processor chips. But to appreciate the capabilities of these chips one must become aware of the acoustic characteristics of human speech.

The basic units of human speech are called phonemes. Linguists recognise just 35 phonemes, plus 6 conjoint pairs or dipthongs, as sufficient to describe the more common dialects of English. Linquists use the strange symbols of the International Phonetic Association as the names of phonemes; we follow the practice in computer speech of using typeable equivalents enclosed in slashes, such as /A/ to refer to the vowel in the "day". The term allophone is used to describe variants of a particular phoneme. For instance, the length of vowels varies markedly in various words. Another instance is supplied by the /P/ in “puff”, which is produced with a puff of air, whereas the same phoneme in the word "ape" has no such puff.

Linguists have produced several conflicting schemes for classifying phonemes. This terminology is to describe in detail the capabilities of the speech chips. The most basic classification in speech is between voiced and unvoiced sounds. A phoneme is voiced if the vocal chords vibrate during its production. The simple practical way of detecting voicing is to place a finger on the throat immediately below the Adam's apple. Vowels are voiced sounds for which the vocal tract is held (almost) rigid during production, there is no significant obstruction, (and the nose is not used). In contrast, consonants are produced in a changing or constricting the vocal tract, can involve the nasal tract, and may be either voiced or unvoiced. The manner of production of voiced sounds is similar to the production of a note by a reed instrument. Just as a musical note is largely of a fixed frequency plus the harmonics of that note. Likewise in voiced sounds most of the speech energy at any instant is concentrated in narrow frequency bands termed the first, second, and third formants. But whereas for a musical note the harmonics are simple multiples of the frequency of the fundamental, there is no simple relationship between the frequencies of the formants, due to the complex resonances of the human vocal tract. The first first formant (for an adult male) lies in the range below 900Hz, the second lies 900-2200Hz, the third 2200-5000Hz. (Female and children’s voices have formants of higher pitch). In any utterance, the trace in time of the formants is characteristic of the speaker.

In fact in 1962 Kersta used the plots formant of formant frequency aqgainst time as a “voice print” for speaker identification for legal identification.

Unvoiced sounds are produced by narrowing the vocal tract at some point, producing turbulence in the air stream; for example the */S/* in "sand". Such sounds have a wide spectrum with frequencies of 4000 to 8000Hz. Unvoiced sounds are far less characteristic of the speaker and accordingly warrant less refined techniques when synthesising utterances. Note how the /Z/ in zoo is (almost) precisely the same as a */S/*, but is voiced. These two phonemes belong to the important class of fricative consonants. Thus, so far we can see speech synthesis as adding the right amount of white noise to two or three formants. But there are further complications. Notable are the stop consonants, which include a momentary silence while the vocal tract is blocked by the tongue. Thus blocking with the tongue against the gum ridge produces with voicing a /D/, without voicing a /T/ as in "to". The silent period at the commencement of a stop consonant is quite long, 25% of the utterance "ate" is silence,

Words are the units of printed speech, but are somewhat arbitrary units in continuous speech. Thus there are three pauses (silences) within the utterance "standard". Yet as normally spoken, "he uses standard oil" contains only these same three silent periods.

In sum, a large part of continuous speech is silence or unvarying periodic or sustained white noise. But especially at the beginning of consonants there are short periods where the formant frequencies and amplitudes are changing rapidly, and the white noise is altering varying markedly. Simple minded encoding systems which treat speech as an arbitrary acoustic vibration with frequencies lying between 150 and 5000Hz say must inherently be deplorably inefficient. Although the information rate in speech is low on average, it does vary considerably.

DELTA MODULATION

The simplest means of digitising speech, once termed Pulse Code Modulation, is to sample the amplitude at regular intervals, and convert to a digital amplitude. From the work of Shannon it is known that the sampling rate must be (at least) twice the highest frequency to be retained. Sampling 10,000 times per sec (hoping to reach 5000Hz components), and using a six bit digital encoder, involves 60,000 bits per second of speech.

A delta modulator samples the incoming analogue stream at regular intervals, and outputs a digital measure of the difference (the "delta") between the incoming signal and a comparison waveform. The decoder can produce the same comparison- waveform, and hence by using the delta reproduce an approximation to the input. In the simplest variant of delta modulation, the delta is directly the difference between the sample and the previous, the comparison waveform being merely a squared off version of the input. By taking the comparison waveform as having a linear slope, the delta can measure the departure from a straight line over the sampling period. If the signal falls below the comparison, the slope of the comparison can be reversed.

In what is often called "adaptive" delta modulation the comparison waveform is made to vary with time. Thus in two slope delta modulation, the comparison waveform either increases or decreases with one of two slopes. When the incoming signal greatly exceeds the comparison, the greater slope is adopted. In continuously variable slope delta modulation (CVSD), the* *slope of the comparison (and output) amplitude is altered for each difference sampled. Thus in CVSD modulation the comparison wave is linear only between each sample point.

There are several CVSD chips available. Firstly, There is the Harris HC-55516 chip. This is claimed to 'produce' intelligible speech for 9K bits per Second. Other chips include the MC3417 and 3418 from Motorola.

Delta Modulation systems are notable for the minimal programming effort. In the development of the product the utterances are spoken directly into memory via the modulator, for storage on a suitable medium prior to ROM production. Production of the digital data stream for demodulation is so simple that it would be practical for a product to have a speech output programmable by the user. For instance, a greatly enhanced "Speak and Spell" might have word lists and the spoken output supplied by a teacher for class use. This user data would need to be stored in non-volatile form as in electrically erasable PROMs, or in CMOS ram battery backed. This particular product would be rather too expensive at 1982 memory prices. With today's technology, user programmability of speech responses is more practical for microcomputers equipped with mass storage.

TEXAS INSTRUMENTS LPC CHIP SETS

The Texas Instruments "Speak and Spell" was the first mass marketed product that offered voice output. "Speak and Spell" featured a calculator chip, together with a linear predictive coding (LPC) chip, the TMS 5200, and two 16K roms. The components of "Speak and Spell" are not marketed, but TI has announced the imminent release of chips that incorporate the same LPC algorithms, and are compatible with conventional microprocessor systems.

For a human to talk, the brain must send out to the components of the vocal apparatus a series of instructions, specifying whether there be voicing or not, and also specifying the shape of the tract. For the brain, speech is therefore a set of such control words. In the TI LPC system, the configuration of what is certainly a model of the vocal tract is likewise determined by a set of just 10 coefficients, with a pitch/voicing parameter In the TI system, the LPC chip has applied to 1t either a periodic or white noise source, depending on the pitch/voicing parameter. This source constitutes the input stream to the LPC decoder the output being after conversion to analogue and filtering the computed speech. To decode an input stream of numbers using the LPC algorithm **one** proceeds as follows: output n is a linear function of inputs n, n-1,n-2,n-3...n-k where for the **TI** system k is 10. Every one fiftieth of a second, new set of ten coefficients for the LPC algorithm is loaded into the LPC chip, while a new pitch/voice parameter is collected. As a complete ***set*** of these parameters involves just 48 bits, tie apparent bit rate of operation of the TI system is 2400 bits per second of speech. By including one more bit, a repetition bit, indicating that no new parameters are to be used, the bit rate is reduced to around 1100 bits per second of speech.

To prepare; speech data for the system so far described involves the processing of spoken speech utterances into the set of coefficients. Texas Instruments insist on supplying the original spoken utterances from written specifications. This is apparently done not -just- to retain for TI 100% copyright to the speech data, but due to the need for tinkering with the processing. The direct use of such chips as the TMS 5200 speech synthesiser chip seems only appropriate to the large volume manufacturer.

National DIGITALKER

The DIGITALKER is a single chip speech decoder that decodes compressed speech to produce a wave form that to a human sounds much the same as original, although on a CRO the waveform is quite different. The speech compression algorithms devised by Mozer of UCLA are clearly quite ingenious. Starting from pulse modulated speech. superfluous phase information is eliminated, a¯' small amplitudes reduced to silences, s eliminating 75% of the apparent data. Then near similar adjacent sounds are made identical, so the¯ by use of a repetition bit further compression **LE** achieved. Male voices can be synthesised at a bit rate of 1000 bits per second of speech, which is about the same as the bit rate in Texas Instruments LPC chip sets. The National chip is a sell contained processor, with its own data bus, **C7** which 128K of compressed speech can reside.

National do NOT supply a development system for encoding the spoken speech samples in house. The manufacturer is required to supply the words (cr phrases) to be encoded on acoustic tape. However National do supply in a prototyping kit a rom holding an "automotive vocabulary". It .is certain that National will produce other chips of precoded speech, presumably all with an American accent. suited for telecommunication and other applications. Thus if a would-be-manufacturer is part of an obvious market that has already been probed, it will be possible to acquire prepackaged speech.

VOTRAX

Votrax, a division of the Federal Screw Works, ***has*** for some years been manufacturing voice output peripherals that function by phoneme generation. (Most recently for the hobbyist TRS-80). In late 1980 it announced the impending release of a single chip 22 pin CMOS DIP phoneme generator, the SC-01 Speech Synthesiser. This device can produce a nominal 64 phonemes (including 3 different duration silences). The chip generates each phoneme for a definite duration, for example the /ZH/ phoneme as in azure, or leisure, has a duration of 90 ms. However, the vowels are available in various lengths, for example the indefinite /EH/ phoneme comes in three different durations.

The SC-01 is configured for installation in a typical microcomputer architecture. The phoneme to be generated is selected by a six bit code, while a further two bits select the pitch of the output. A single strobe initiates the phoneme production, while when the chip is ready for new phoneme data its signal may serve as an interrupt to the microprocessor. Note that the low data rate means that the processor can be primarily devoted to other functions. The analogue (speech) output is suitable for connection to Class A or B transistor amplifiers.

Preparation of the digital speech data for decoding in the SC-10 involves formulating the utterances in terms of the Votrax (extended) phonemes. Votrax maintains a library of phonetically programmed records, which is also available as part of a microcomputer-based development system,

This device operates on an (average) bit rate of 70 bits per second for continuous speech.

OTHER DEVICES

Mimic electronics market a compact speech digitiser "The Speech DATA-BOY", which readily interfaces to microcomputer, and incorporates a microphone and speaker jack. Incoming speech is reduced to a single bit stream. The same device will recode a bit stream to reproduce the speech. The device operates at about 8000 bits per second. However, the speech produced has a background hiss akin to CB radio.

Street Electronics Corporation (SEC) of Anaheim has this year announced the forthcoming production of the ECHO II speech Synthesiser for the Apple microcomputer. This device is a formant synthesiser, based on the TI TMS 5200 LPC processor. Data for 43 formants are incorporated, and each phoneme can be produced at any of 16 pitch levels, 8 lengths, 8 volume levels. In principle, this synthesiser can produce the correct pattern of stress in a synthesised utterance, which can be male, female, a child, a whisper.

In February Texas Instruments announced the — forthcoming availability of an allophone generator based on the TMS 5200. It is noteable that the generator can produce 128 allophones, twice as . as the Votrax single chip. A voice output gadget for the TI 99/4 personal computer will incorporate the generator. Votrax have been marketing a similar gadget providing voice output for the Tandy TRS-80, with a menu of just 64 allophones. Note that the TI and SEC devices do not offer any pitch or loudness, control. In the case of the Votrax product the voice output is highly mechanical, non-human, in character.

This paper has been concerned with specialist speech chips. However the latest signal processor chips, such as the Intel 2920 may be configured by programming as phoneme generators. In the case of the 2920 the only difficulty is the minimal amount of on-board ROM. The coding of phonemes on these signal processors is a highly specialised task.

REFERENCES

J. Anderson, "DATA-BOY Speech Processor", Dr Dobb's Journal, No. 28, Aug-Sept 1978, pp 30-34.

N. Bodley, "A low cost speech synthesiser on a chip", Elect. Des. 15, July 19, 1979, p 32.

L. Brantingham, "Speech Synthesis with Linear Predictive Coding", Interface Age, June 1979, pp 72-75.

C.W. Behrens, "Synthetic Speech/Voice Recognition", Appliance Manufacturer. November 1978, pp 44.

T.A. Gargagliano and K. Fons, "Text translator builds vocabulary for speech chip", Electronics, Feb 10, 1981, pp 118-121.

T. Gargagliano and K. Fons, "The TRS-80 Speaks: Using Basic to Drive a Speech Synthesiser", Byte, Vol. 4, No. 10, October 1979, pp 113-122.

Heuristics Inc, "Speechlab Laboratory Manual", published 1977 by Heuristics Inc, Los Altos, Calif.

D. Jones, "Delta Modulation for Voice Transmission", Application Note 607, Harris Corp, 1979.

I. Kameny, "Comparison of the Formant Spaces of Retroflexed and Non-retroflexed Vowels", IEEE Trans Acoustics, Speech, and Signal Processing, Vol 23 No 1, Feb 1975, pp 38.

L.G. Kersta, "Voiceprint Identification", Nature Vol 196, 1962, pp 1253-1257.

C.J. Kikkert, "A Comparison of Code Modulation Systems", Proc IREE, March 1975, pp 44-48.

B. LeBoss, "Speech I/O is making itself heard", Electronics, May 22, 1980, pp 95-105.

K.S. Lin, G.A. Frantz, and K. Goudie, "Software rules give personal computer real word power", Electronics, Feb 10, 1981, pp 122-125.

J. Makhoul, "Spectral Analysis of Speech by Linear Prediction", IEEE Trans Audio Electro-Acoustics, Vol 21 No 3, 1973, pp 140-148.

B.A. Sherwood, "The Computer Speaks", IEEE Spectrum, August 1979, pp 18.

E.R. Teja, "Voice Input and Output", Elect. Design News, Nov 20, 1979, pp 159-167.

D.W. Weinrich, "Speech-synthesis chip borrows human intonation", Electronics, April 10, 1980, pp 113-118.

R. Wiggins and L. Brantingham, "Three chip system synthesises human speech", Electronics, Aug 31, 1978.