request @ 10 13am pt 2 J.C. 6/6/12

California State University, Northridge
Interlibrary Loan

ILLiad TN: 197073

Borrower: RAPID:TEU

Lending String:

Patron:

Journal Title: Proceedings of the National Communications Forum.

Volume: 40 **Issue:** book 2
Month/Year: 1986 **Pages:** 691-696

Article Author: J. Picone

Article Title: Recent Trends in Digital Speech Coding

Imprint:

ILL Number: -5561918

Call #: TK5 .N37

Location: Stored

Billing: Exempt
Maxcost:

Shipping Address:
NEW: Main Library

Deliver:
Ariel:
Email: 197073
Odyssey:206.107.42.224

expected.

CONCLUSION

For a number of years, research in speech processing was ahead of hardware capabilities for economical, real-time implementation of the knowledge. Now, the explosive advances in microelectronics permit economical realization of processes of great complexity. Low bit rate speech coding, down into the data-speed range, can readily be supported by existing microprocessors. Text-to-speech synthesis, or steadily improving quality, can be supported with modest hardware on personal computers. Automatic recognition of moderate-sized, isolated-word vocabularies is similarly handled. Extension to large-vocabulary speaker-independent, connected-speech recognition will require substantial compute power, and presents a significant challenge for hardware and algorithm designers alike.

In any event, the gap between algorithm understanding and practical hardware capability has closed. And, the pressure is on speech processing research to productively harness the handsome amounts of compute power now available in small low-cost packages.

RECENT TRENDS IN DIGITAL SPEECH CODING

J. Picone
Member of Technical Staff
Exploratory Voice Capabilities Dept.

AT&T Bell Laboratories
Naperville, IL 60566

ABSTRACT

With the ever increasing emphasis of digital speech transmission and speech storage on modern telephony, there is continuing interest in digital speech coding at lower bit rates. Recent advances in device technology have made implementations of complex algorithms feasible within a single integrated circuit. An overview of recent trends in digital speech coding technology will be presented with an emphasis on quality versus bit rate and quality versus complexity considerations. Speech coding techniques ranging from wideband waveform coding systems to narrowband parametric coders will be discussed. We will conclude with a discussion of future directions in speech coding technology.

INTRODUCTION

The on-going advances in digital communication systems along with the ever-present demand for efficient bandwidth utilization have placed a greater emphasis on the need for speech transmission at various data rates below 64 kbits/s. A digital speech encoder, frequently referred to as an analyzer, can be defined as a system which encodes an analog speech signal into a digital bit stream. A decoder, known as a synthesizer, constructs an analog speech signal from a digital bit stream.

Pulse Code Modulation (PCM), a 64 kbits/s system in which the analog speech is sampled at 8 kHz and quantized to 8 bits per sample, is the most popular form of a digital speech coder used in the telephone network. PCM transmission is remarkably robust to a variety of impairments including analog background noise introduced at the transmitter, bit error rates which occur due to channel noise, noise introduced from tandem connections in which a signal will be encoded/decoded multiple times. In this paper, we will examine four techniques, or algorithms, which operate at bit rates below 64 kbits/s.

The attribute which traditionally has received the most attention in digital speech coding is voice quality. To reduce the bit rate required to transmit speech, one must inevitably remove information. A digital speech coding system strives to efficiently represent those components of the speech signal which are most relevant to our perceptions of quality.

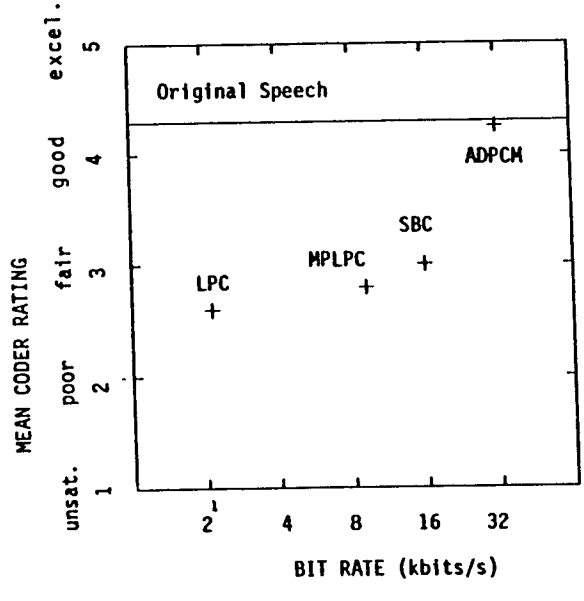

FIGURE 2-1

MEAN CODER RATING BY BIT RATE

ADPCM: Adaptive Differential PCM
SBC: Sub-Band Coding
MPLPC: Multi-Pulse Linear Predictive Coding
LPC: Linear Predictive Coding

One means of comparing the technologies described in this paper is given in Fig. 2-1. Speech quality, as judged by untrained listeners via a Mean-Opinion-Score (MOS) Test[1] is displayed versus bit rate. The MOS test is a subjective test in which listeners are asked to judge the overall speech quality of an utterance on a scale of 1 to 5. A score of 1 denotes speech quality which is extremely poor and unintelligible. A score of 5 denotes speech quality which is indistinguishable from the original analog recording. Note that PCM telephone bandwidth speech is closest to the original analog speech.

The algorithms represented in Fig. 2-1 were chosen for two reasons. While we will no doubt neglect certain technologies in our overview of speech coding, these algorithms are representative of the speech quality achievable at their respective bit rates. Also, these algorithms have been implemented in a single VLSI chip, called a Digital Signal Processing (DSP) chip, and are commercially available today. The scores displayed in Fig. 2-1 represent evaluations of actual hardware implementations.

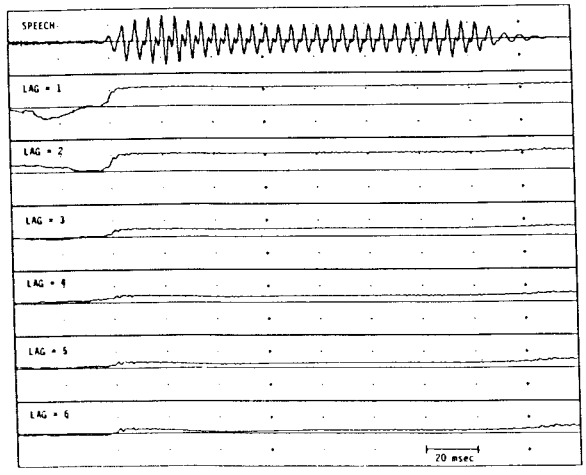

FIGURE 2-2
8 KHZ SPEECH SIGNAL AND CORRELATED SIGNALS AT DIFFERENT TIME LAGS

ADAPTIVE DIFFERENTIAL PULSE CODE MODULATION (ADPCM)

Pulse Code Modulation (PCM) belongs to a class of speech coders known as waveform coders. In PCM, an analog signal is sampled uniformly in amplitude and in time. A telephone bandwidth speech signal has a low-pass characteristic with an effective bandwidth extending from 200 to 3300 Hz, allowing it to be sampled at 8 kHz. Because the speech signal has a slowly varying spectral structure[2], there is significant sample-to-sample correlation. This is depicted in Fig. 2-2. A speech signal, sampled at 8 kHz, is plotted, along with signals which represent the correlation between speech samples separated by a short time interval. In this plot, delays ranging from 1 to 6 sample periods were used.

In Adaptive Differential Pulse Code Modulation (ADPCM)[3,4], the difference signal formed by differencing adjacent speech samples is quantized by an adaptive quantizer which attempts to adjust to the root-mean-square (rms) level of the signal. The block diagram of an ADPCM encoder/decoder is shown in Fig. 2-3. Because the quantizer is adaptively adjusting to the rms level of the difference signal, a fewer number of levels in the quantizer are required, thereby allowing a reduction in the bit rate. ADPCM offers "transparent" telephone quality speech at a bit rate of about 32 kbits/s.[1]

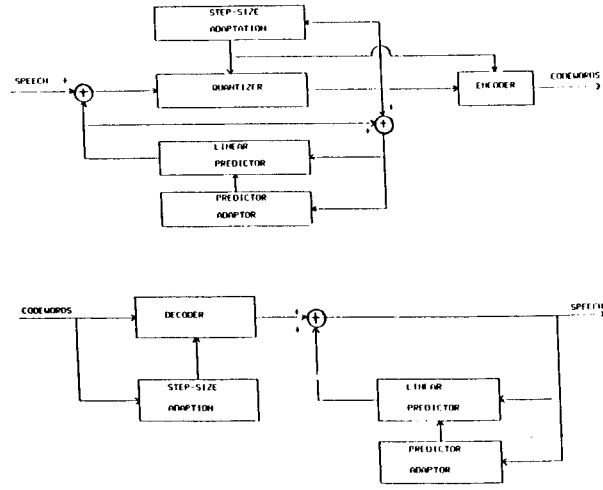

FIGURE 2-3
ADPCM ENCODER/DECODER

The function of the predictor block is to further reduce the level of the difference signal by attempting to predict the current sample from a number of previous samples. This is known as linear prediction. If this prediction process is based on the past 5 msec, it is generally referred to as short-time prediction. The predictor is updated frequently to allow the encoder to track the slowly-varying spectral structure of the speech signal.

CCITT has standardized a particular version of an ADPCM coder.[4] There are two key features of the CCITT 32 kbits/s encoder; a parallel predictor structure which allows undistorted transmission of 2400 bps modem data, and feedback adaptation of predictor parameters, which has the advantage that no side information about the encoder is required to be transmitted. Though feedback adaptation is not as robust as feed-forward adaptation in the presence

of channel errors, it gives acceptable performance over most digital telecommunications channels.

SUB-BAND CODING

Upon examining the behavior of a speech signal in the frequency domain, it was observed that different frequency regions display different characteristics.[5] In Fig. 2-4, a speech signal has been filtered by a series of filters with frequency ranges extending from 0 to 500 Hz, 500 Hz to 1000 Hz, 1000 Hz to 2000 Hz, and 2000 Hz to 3000 Hz. This is known as a filter bank[6], and is a crude model of sensitivity of the human ear to frequency. Sub-Band Coding (SBC) is a waveform coding technique which exploits the different characteristics of each filter output by processing individually each filter output through the ADPCM algorithm described above.

A block diagram of the system is shown in Fig. 2-5. The speech signal is filtered through a carefully-designed filter bank[6], typically implemented using Quadrature Mirror Filters (QMF).[7] The outputs of the filter bank are reduced in sampling rate and encoded using ADPCM coders which have been optimized for the statistics of the particular frequency band in which they operate. By partitioning the speech signal into independent frequency bands, this technique can dynamically adapt to coarse spectral variations (formant structure) in the speech signal. In addition, the number of bits allotted to each frequency band, or bin, can be optimized to match the sensitivity of the human ear, a process which is known as perceptual-weighting.[8]

SBC provides good quality speech at bit rates in the range of 16 to 24 kbits/s. Even though SBC is considered to be a reasonably robust algorithm, its two major limitations for the transmission environment are that its performance under tandem encodings degrades more rapidly than 32 kbits/s ADPCM, and it will significantly distort high speed modem data.

MULTI-PULSE LINEAR PREDICTIVE CODING (MPLPC)

The articulatory system is composed of components such as the tongue, lips, and palate which, due to their finite masses, move with a certain sluggishness during speech production. This sluggishness of motion allows the speech signal to be modeled as a quasi-stationary signal; that is, over a 10 or 20 msec interval, the general structure of the speech signal can be considered stationary. This gives rise to a slowly varying spectral structure which can be adequately modeled with a linear predictive filter.[9] In Fig. 2-6, a waveform is plotted along with the error in prediction, that is the error which results from modeling the speech signal using a tenth order predictor.

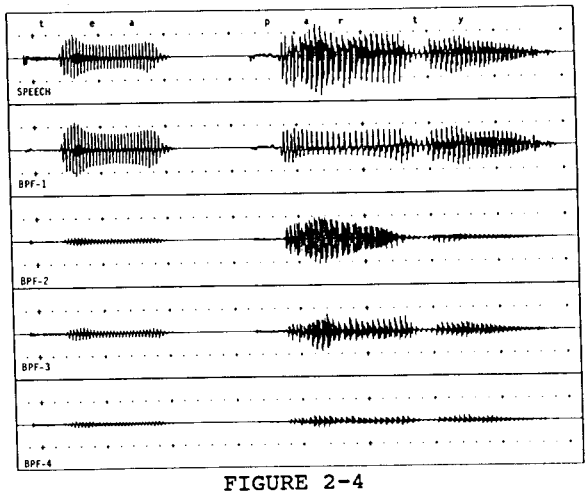

FIGURE 2-4
SPEECH SIGNAL FILTERED THROUGH A FILTER BANK

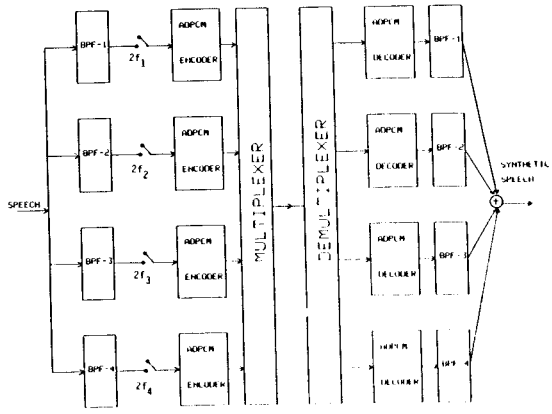

FIGURE 2-5
SUB-BAND CODER

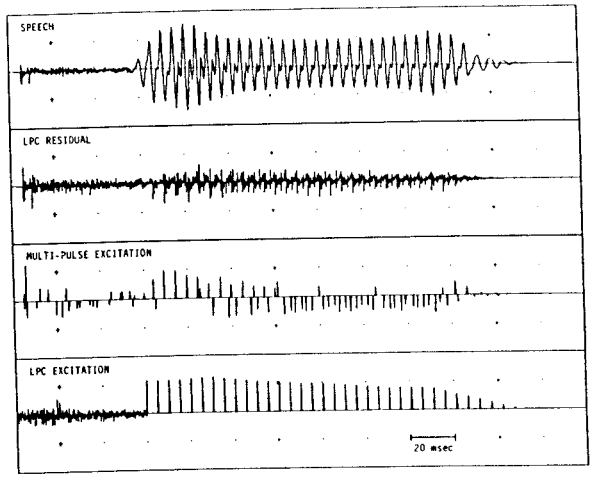

FIGURE 2-6
SPEECH WAVEFORM PLUS PREDICTION ERROR

Observe that the error signal has a pulse-like nature during voiced sounds (the periodic section of the waveform). The pulses in this signal correspond to the moment at which the linear predictor does not adequately model the signal. Other sections of the prediction residual are noisy, corresponding to the unvoiced areas (noisy sections) of the speech waveform. Efficient methods of modeling this complicated residual structure has been the subject of much research into speech coding.

Multi-Pulse Linear Predictive Coding (MPLPC) attempts to model this signal as a sequence of pulses. A typical MPLPC excitation is shown in Fig. 2-6. In this system, shown in Fig. 2-7, a pulsed excitation is constructed which creates the best match between the synthetic speech and the original speech, a process known as error minimization. Obviously, if pulses were allowed to be located at every sample, the MPLPC system would produce perfect speech quality, with no savings in bit rate. Typically, to create a speech coder which operates at 9.6 kbits/s, 4 pulses are allocated every 5 msec.[10]

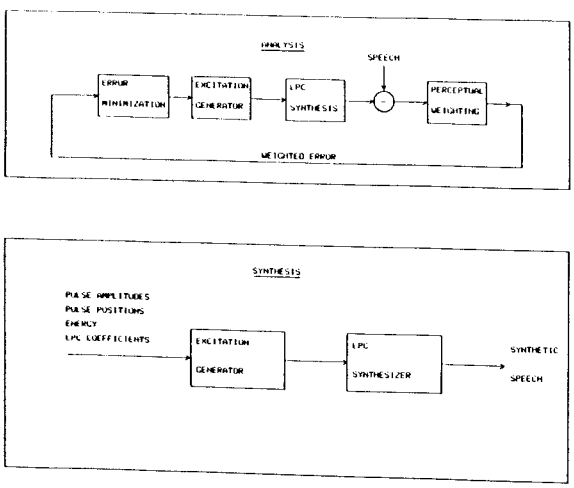

FIGURE 2-7
MULTI-PULSE LINEAR PREDICTIVE CODER

The pulses are located sequentially using a matched filter type process. First, an optimum pulse location is defined by searching for the pulse location which produces a minimum error. After this pulse is located, its amplitude is computed. A local synthesizer creates the synthetic speech corresponding to this pulse, and this synthetic speech is subtracted from the original speech signal. The process is repeated for the next pulse, using this difference signal. Eventually, when all pulses are located, the residual error should be small. This process is called analysis-by-synthesis. The coder, though it employs a parametric model of the spectrum, also employs waveform matching to locate pulses, and therefore belongs to a class of coders known as hybrid coders.

Though this coder is significantly more complex than the previous two, requiring on the order of four MFLOPS, it has also been implemented on a single DSP. It is considerably robust to background noise, multi-speaker noise, etc. The speech quality at 9.6 kbits/s, though quite good, is not toll quality. The coder tends to add distortion in the form of speech correlated noise.

LINEAR PREDICTIVE CODING (LPC)

The three previous coders all employed some type of waveform matching process to code the speech signal. In some sense, they attempt to produce the minute details of the speech waveform. This attention to detail has the side affect of requiring a high bit rate for transmission. Parametric speech coders attempt to model only very general properties of the speech signal. The most popular of these methods, the Linear Predictive Coder (LPC), makes hard decisions about the nature of the speech signal, a process known as source coding.[12]

In the LPC coder, shown in Fig. 2-8, speech is approximated by a very simple model. The excitation signal is either a pulse train spaced at periods known as the pitch frequency, or a white noise sequence. This is known as a voicing decision. Speech production theory suggests that the periodic excitation should be suitable for producing voiced sounds while the white noise excitation should be appropriate for synthesizing unvoiced sounds. Estimation of the pitch and voicing parameters is still an elusive process today, though many good algorithms exist.[13] An example of the excitation produced by the LPC vocoder is shown in Fig. 2-6.

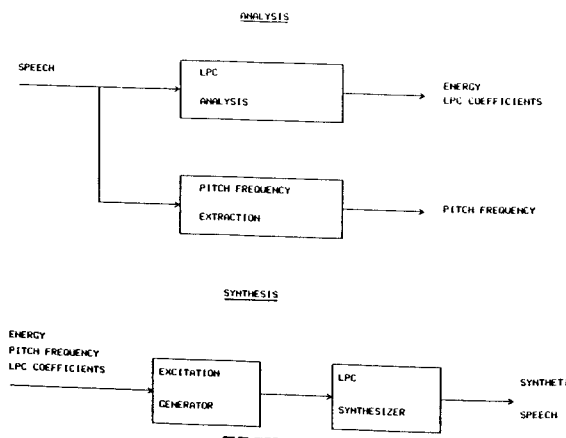

FIGURE 2-8
LINEAR PREDICTIVE CODER

To synthesize speech, the excitation signal is filtered by a time-varying all-pole filter known as an LPC filter. The coefficients of this

filter are extracted from the speech signal at a rate on the order of 50 times a second. This filter attempts to model the shape of the vocal tract, and the acoustic effects of the position of the lips. It is the same type of predictor used in the ADPCM and MPLPC systems described above.

The speech quality produced by this model can be classified as intelligible, but poor, having a certain characteristic buzziness, or hollowness. However, the system can operate at data rates in the range of 2.4 kbits/s. LPC synthesizers have been used in commercial products such as Speak and Spell.[14] The government has standardized a version of an LPC vocoder which operates at 2.4 kbits/s and is known as LPC-10E.[15] Because the LPC vocoder is a source model, it is somewhat more flexible than the other waveform coding-based models. On the other hand, this system models a single speaker, and is not capable of handling multiple speakers at 2.4 kbits/s.

FUTURE TRENDS IN DIGITAL SPEECH CODING

We have reviewed the current state of the art in speech coding technology. We have presented four algorithms available in simple hardware at various data rates. Why has the speech coding business not exploded? A major factor seems to be that the voice quality/bit rate/cost per channel tradeoffs do not yet seem to be attractive.

There are two significant trends in speech coding research today. First, with the advent of such narrow channel applications as mobile telephone and secure communications (encrypted data channels), applications which seem to be capable of accepting present day coder costs per channel, there is much interest in digital speech coding at 4.8 kbits/s. Current trends seem to indicate that it is reasonable to believe that speech coding will provide quality which is significantly better than a vocoder, yet not transparent, at this data rate. Second, there is interest in true "transparent" quality speech at 9.6 or 16 kbits/s for transmission applications in ISDN and voice-mail type applications. Here the emphasis will be on a low cost per channel, perhaps using custom VLSI chips, or low-level DSP chips. While our ability to transmit speech at low data rates will continue to improve in the laboratory, advances in VLSI hardware will play the key role defining the growth of the speech coding business.

REFERENCES

[1] W. R. Daumer, "Subjective Evaluation Of Several Efficient Speech Coders," IEEE Trans. Commun. (special issue on bit rate reduction), Vol. COM-30, No. 4, pp. 655-662, April 1982.

[2] J. L. Flanagan, Speech Analysis, Synthesis, and Perception," Second Edition, Springer Verlag, New York, 1972.

[3] N. S. Jayant, "Digital Coding Of Speech Waveforms: PCM, DPCM, And DM Quantizers," Proc. IEEE, Vol. 62, pp. 611-632, May 1974.

[4] C.C.I.T.T. Recommendation G. 721 (32 kbits/s Adaptive Differential Pulse Code Modulation), C.C.I.T.T. Red Book, 1984.

[5] M. R. Shroeder, "Linear Predictive Coding Of Speech: Review And Current Directions," IEEE Commun. Mag., Vol. 23, No. 8, pp. 54-61, August 1985.

[6] R. E. Crochiere, R. V. Cox, and J. J. Johnston, "Real-Time Speech Coding," IEEE Trans. Commun., Vol. COM-30, No. 4, pp. 621-634, April 1982.

[7] D. Esteban and C. Galand, "Application Of Quadrature Mirror Filters To Split Band Voice Coding Schemes," Proc. 1977 IEEE Int. Conf. Acoust., Speech, Signal Processing, Hartford, CT, pp. 191-195, May 1977.

[8] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing Digital Speech Coders By Exploiting Masking Properties Of The Human Ear," J. Acoust. Soc. Am., Vol. 66, pp. 1647-1652, Dec. 1979.

[9] B. S. Atal and S. L. Hanauer, "Speech Analysis And Synthesis By Linear Prediction Of The Speech Wave," J. Acoust. Soc. Am., Vol. 50, pp. 637-655, 1971.

[10] B. S. Atal and J. R. Remde, "A New Model Of LPC Excitation For Producing Natural Sounding Speech At Low Bit Rates," Proc. 1982 IEEE Int. Conf. on Acoustics, Speech, Signal Processing, Paris, France, pp. 614-617, 1982.

[11] B. S. Atal and M. R. Shroeder, "Predictive Coding Of Speech Signals And Subjective Error Criteria," IEEE Trans. Acoust., Speech, and Signal Proc., Vol. ASSP-27, No. 4, pp. 247-254, June 1979.

[12] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis Of Speech From A Dynamic Model Of The Vocal Cords And Vocal Tract," Bell Syst. Tech. J., Vol. 54, No. 3, pp. 485-506, March 1975.

[13] W. Hess, Pitch Determination Of Speech Signals, Springer Verlag,

New York, 1983.

[14] G. A. Frantz and R. H. Wiggins, "Design Case History: Speak And Spell Learns To Talk," IEEE Spectrum, Vol. 19, No. 2, pp. 45-49, February 1982.

[15] T. E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," Speech Technology, Vol. 1, No. 2, pp. 40-49, April 1982.

AN OVERVIEW OF TEXT-TO-SPEECH TECHNOLOGY

Murray F. Spiegel
Member of Technical Staff
Speech Technology Studies Group

Bell Communication Research
435 South Street
Morristown, NJ 07970

ABSTRACT

This paper explains the range of methods used in converting computer text to intelligible, if not natural sounding, speech. The components in most commercial and research systems, likely telecommunication applications, and future research trends will be discussed.

INTRODUCTION

Speech synthesis, or text-to-speech, is the most flexible of all sound output technologies, since it does not require special audio recording or special input processing.

Text-to-speech synthesis systems require concatenation of speech segments. A high-level view of the steps in the text-to-speech process is:

Analysis of text input. One result of the text analysis is the determination of the sequence in which the stored units of speech should be produced. Systems based on subword units (larger than the phoneme) must translate the phoneme string into the larger units. Another result obtained at this stage concerns suprasegmental factors. The controls necessary for proper stress, pauses, and intonation patterns are selected.

Retrieval of stored units. Although it is possible to store a high-fidelity recording of the speech elements, there are coarticulation effects that cause poor speech quality when waveforms of short-speech units are concatenated. Because the output is a concatenation of these units, some smoothing of the speech parameters is performed across the time boundaries between adjacent units. The intent of this smoothing is to eliminate some of the choppiness that results when elements are concatenated. Also, at this stage the speech parameters are adjusted to reflect the stress, timing, and intonation requested by the analysis stage.

Synthesis output. Finally, the adjusted speech parameters are sent to a speech-synthesis circuit, usually a chip, that is responsible for final speech output.

MAJOR VARIABLES IN SPEECH SYNTHESIS

The flexibility and success of a speech-synthesis system depends on the choice of the sound element for concatenation, the coding technique, and the kinds of manipulations or rules that are performed on the sound elements to provide natural-sounding continuous speech. In addition, a text-to-speech system requires a set of rules for converting letters into sounds for accurate word and sentence pronunciation.

CHOICE OF THE SOUND ELEMENT

Speech sounds of any duration may be concatenated: phonemes, allophones (phoneme variations due to context), dyads or diphones (the transitions between pairs of phonemes), demi-syllables (half-syllables), syllables, morphs (word roots, prefixes, and suffixes), words, or phrases. To a large extent, as the unit of speech stored becomes smaller, the storage requirements for arbitrarily long messages becomes smaller also. Whereas, 50,000 to 200,000 words would reproduce most of an adult's language, about 100 allophones ideally accomplishes the same thing.

Simple concatenation of any sized units comes with the penalty of quality loss because: a) simply concatenating words or phrases of speech does not allow for a natural-intonation pattern, b) concatenation of smaller units suffers in speech quality at the transitions even with extensive rules for "smoothing."

CHOICE OF SYNTHESIS ALGORITHM

Synthesis algorithms specify how to produce continuous, natural-sounding speech using a small inventory of speech segments plus a set of rules for retrieving units, combining them, and modifying them. The rules for modifying segments are governed by two types of effects: 1) segmental effects, such as coarticulation and allophonic variation, which operate on a shorter time scale than do b) suprasegmental issues, such as phrasal stress and intonation patterns.