Proc. IEEE Int. Conf. Circuits, Systems (Montreal, QC), pp. 6-9, May 1984

DIGITAL CODING OF SPEECH SIGNALS

M. Berouti[1], *P. Kabal*[2,3], *P. Mermelstein*[1,2,3]

[1]*BNR* Verdun, Quebec
[2]*McGill University* Montreal, Quebec
[3]*INRS-Télécommunications* Verdun, Québec

ABSTRACT

This paper gives an overview of current work in digital speech coding. New directions in research on speech coding algorithms are discussed. At medium rates, new algorithms which use perceptually motivated fidelity criteria have significantly improved the quality of the reproduced speech. At lower rates, the application of vector quantization concepts have resulted in even lower data rates for the same speech quality. The availability of single chip signal processors has dramatically reduced the cost of implementation of practical speech coders. At the same time, considerations of issues relating to the integration of speech coders into telephone networks have resulted in new efforts at developing international standards for speech codecs.

INTRODUCTION

This paper presents a survey of recent trends in digital speech coding. Advances in coding fall into three main areas: improved coding algorithms, real-time hardware implementations, and system integration of speech codecs. Until very recently, the exploitation of advanced coding techniques in practical communication systems has been held back by the high cost of hardware implementation. At the same time, the need for coding has been somewhat mitigated by the reduced cost of providing additional transmission capacity. Recent advances in integrated circuits have dramatically changed this situation. It is the coming together of new algorithms and the means to implement them in cost effective hardware that has given speech coding work a new impetus.

The work on speech coding algorithms in the last decade has been fueled by a deeper understanding of the fundamental principles governing speech coding, both from a signal processing point of view and from a speech perception aspect. A basic requirement for high-quality coding is a parametric model for representing the speech signal, one that allows for high quality speech reproduction in the limiting case of perfect parameter information. The parameters of this model are then coded for transmission in such a way as to exploit the redundancy in the parameters. The coding process makes use of fidelity criteria corresponding to subjectively perceived degradations. In this way the quality of the reconstructed speech can be maximized at a given bit rate.

CH1993-5/84-0000-0006 $1.00

CODING ALGORITHMS

Techniques based on linear prediction have emerged as among the best methods for speech coding. The slowly varying nature of the short term energy and the spectral envelope of speech allows bit rate compression through prediction. A linear predictive coder models the speech waveform with two parts: i) an adaptive prediction filter which removes the short-term correlation from the signal, and ii) a residual signal which embodies pitch and voicing information. The corresponding decoder uses the residual signal to excite a synthesis filter (the inverse of the predictor filter) to produce reconstructed speech. At one end of the scale, simple coders such as ADPCM (adaptive differential pulse code modulation) use simple low order, continuously adaptive predictors and signal level adaptive quantizers for the residual. The adaptation is driven from the decoded speech, obviating the need to explicitly send the predictor coefficients and the gain factor. ADPCM is successfully applied in the range of rates from 64 kb/s to 24 kb/s. More complex coders use more elaborate predictors and sophisticated quantization strategies for the residual signal to achieve lower rates. For this class of coders, the adaptive predictor is optimized over a frame of data which is updated at intervals of the order of 20 ms. The predictor coefficients are sent to the decoder as side information. An example is APC (adaptive predictive coding) which is useful for bit rates down to 10 kb/s. All of the above coders may be termed waveform coders in that they attempt to preserve the input waveshape.

To achieve even lower transmission rates, the residual signal is coarsely parameterized as one of two forms, either a quasi-periodic pulse train (voiced speech) or a noise waveform (unvoiced speech). The best known example of such an analysis/synthesis coder is LPC (linear predictive coding). At 2.4 kb/s, LPC reproduces speech which is very intelligible but which has a definite synthetic quality.

Many of the same ends that are achieved by coders which operate in the time domain can be achieved using frequency domain techniques [1]. For instance, sub-band coding (SBC) filters the input speech into a number of frequency bands and codes each band separately. The digitizer for each sub-band adapts to the short-term energy in that band. This process is carried even further in adaptive transform coding (ATC) which uses a discrete cosine transform to separate the speech into 64 to 256 frequency

bands. An adaptive estimate of the short-term spectrum based on LPC principles is used to scale the quantizer for each of the bands. In addition, ATC employs an adaptive bit assignment strategy to allocate bits to the bands in such a way as to minimize the distortion. ATC is a useful technique down to rates below 10 kb/s.

NEW IDEAS FOR CODING ALGORITHMS

The development of speech coding algorithms has seen a steady growth in the last five years. The survey paper by Flanagan et al [2] gives a comprehensive view of the state of the art up to the early part of 1979. Fig. 1 shows a schematic view of the tradeoffs between transmission rate and speech quality for a number of coding techniques showing the change in status in recent times. The vertical scale gives quality assessments for narrowband speech. Toll quality speech is that equivalent to 7 bit companded PCM as used in telephony. Communications quality speech represents speech that is completely intelligible but carries noticeable degradation (equivalent to 5–7 bit companded PCM). The synthetic quality rating includes the artificial quality typified by LPC coding. An additional category is commentary grade, which applies to wideband speech (7 kHz or more bandwidth). At rates of 32 kb/s and above, the progress has not been directed at improving speech quality; for in 1979 algorithms with toll quality were already available. Instead, much of the work on algorithms at these rates has been concerned with implementational and system integration issues. These issues are discussed in a separate section below.

Wideband speech coders, by providing increased high frequency information, give a subjectively pleasing sense of immediacy and clarity which contrasts with the distant and muffled quality inherent in the narrowband speech signals used in conventional telephony. This type of service is important in the distribution of program material and in new services such as audio and video teleconferencing. The challenge has been to code wideband speech at rates comparable to conventional PCM (64 kb/s) for speech sampled at 16 kHz. Several techniques have been suggested as candidates for a new CCITT standard covering wideband speech [3]. In one, a two band sub-band coder is used. ADPCM and PCM are used to code the individual bands. In another, ADPCM is applied using an adaptive pole/zero predictor structure with quantizer noise shaping (see below).

At intermediate rates, efforts to improve quality have also paid off. Linear prediction with multi-pulse excitation shows great promise at rates from 4.8 to 16 kb/s [4]. In fact, it tends to fill in a gap between LPC and conventional waveform coders which operate at higher rates. In multipulse coders, the residual signal is modelled with a given number of pulses per time segment. The amplitudes and locations of the pulses are adjusted to give the best quality reproduced speech. The success of this technique can to a large extent be attributed to the use of a perceptually motivated objective measure in optimizing the parameters of the residual waveform. Further progress can be expected with the use of such quality criteria as we develop a deeper understanding of the information in the speech signal that is essential to preserving its naturalness.

Some of these same ideas have been used in APC. In APC, they take the form of noise spectral shaping [5]. The noise inherent in coarse quantization of the residual can have its spectrum shaped so that the speech spectrum tends to mask the distortion. APC also uses a variable length code words to represent the quantization intervals. In this form of entropy coding, short code words are used for the most frequently occurring (low amplitude) levels while longer code words are used for the others. This results in a reduced average bit rate but requires that the bit stream be buffered before transmission over a fixed rate channel. By incorporating these ideas, various forms of APC can achieve toll quality at 16 kb/s.

At even lower rates, the limitations of the LPC model have not resulted in significantly improved quality, but new techniques have allowed LPC based schemes to operate at very low bit rates (150 to 800 b/s). The vector of LPC coefficients specifies a spectral envelope for the speech. Simple scalar quantizers code each component in the vector independently. In contrast, vector quantization techniques exploit the fact that the vectors of LPC coefficients which occur for natural speech, are not uniformly distributed over the space of possible vectors [6]. The quantization criteria used are objecive models of subjective requirements, in this case a perceptually motivated spectral distortion measure. This technique has allowed very efficient coding of the LPC coefficients (10 bits for a vector of 10 coefficients). Efforts are being directed into finding simplifications which give the benefits of vector quantization while reducing the computational requirements.

CODEC HARDWARE REALIZATION

A few years ago, real time hardware implementation in hardware of all but the simplest forms of speech coding was impractical. Coders were developed using computer simulation (not real-time) or high speed array processors (real-time). The advent of single chip digital signal processing (DSP) elements has dramatically changed the situation and has lead to a rapid development of real-time coders.

A typical DSP chip is organized to perform fast multiplication and addition. The distinguishing feature which differentiates DSP chips from microprocessor chips is the allocation of a significant portion of the chip area to a high speed multiplier. DSP chips have architectures organized so as to perform repetitive operations on a stream of data very fast. For example a dot product of two vectors, a basic step in any filtering operation, can be implemented such that a new pair of data elements is processed every 400 ns (TI TMS32010). More complex code, accessing non-consecutive locations, results in a lower throughput.

The emphasis in adapting speech coding algorithms for real-time implementation is on the computational aspects. Algorithms developed using floating point computations must be converted to use fixed point arithmetic.

In some cases, double precision operations are needed to preserve accuracy. In addition, due to limitations of chip memory, computational reorganization is often called for in order to produce an algorithm which is implementable. Coding algorithms are varied, but only small parts can usually be cast in the form of stream processing that DSP chips handle so well. This means that operations counts for simulations can only give a first order estimate of the feasibility of real time operation of a given algorithm. Still the raw speed of the new DSP chips is such that with careful algorithm design, a great deal of processing may be carried out in real-time.

The changes brought about by DSP chips are dramatic enough that algorithms which a few years ago required a rack full of equipment can now be implemented on circuitry based around a single DSP chip. A number of different algorithms have been implemented in real-time hardware using DSP chips. A standard form of LPC, known as LPC-10, has been implemented on a single TMS32010 chip [7]. Simpler algorithms such as ADPCM at 32 kb/s have also been implemented on DSP chips such as the NEC 7720.

Consider a flexible experimental coder workbench that has been developed at BNR to illustrate the capabilities of today's technology for speech processing. The entire coder (exclusive of a power supply), sits on a single 11.5 cm by 24 cm card. This card contains the necessary analog interface circuitry, line transformer, A/D and D/A converters (μ-law converter with filtering), serial ports for the input and output data streams, programmable clock circuitry, 4K of ROM or RAM, and a single DSP chip (TMS32010). This arrangement has been used to implement a full duplex (coder and decoder operating simultaneously) 16 kb/s APC coder using the algorithm described above.

CODING IN THE TELEPHONE NETWORK

An indication of the maturity of the simpler coding algorithms such as ADPCM has been the development of a CCITT international telephony standard for 32 kb/s transmission [8]. Such coders have not been widely deployed until now due to the expense of hardware implementation and the degradations introduced, in the absence of standards, when different algorithms occur in tandem in a particular end-to-end connection. Agreement on a standard algorithm will likely justify the development of special-purpose chips for 32 kb/s coding.

The requirements on speech coding in either public switched networks or private networks are varied [9]. Excellent speech quality and relatively modest coding complexity are but minimal requirements for practical applications. Since telephone networks today carry a multitude of signals other than speech, an ability to encode these signals with adequate fidelity is required. Without this transparency, constraints on routing of signals other than speech would be required, a step that would significantly complicate network operations. In telephony, the most important non-speech signals to be transmitted over speech band coders are voiceband data signals and signalling tones.

The coding of voiceband data signals presents many requirements that tend to be in conflict with those for speech signals. The short-term data spectrum is relatively constant, particularly when the transmitted bits are scrambled in the coding process, while the speech spectrum changes significantly with each speech sound. Also, the short-term energy of the data signal is relatively constant and its accurate reproduction is necessary to avoid data errors, especially when some form of amplitude encoding is employed. By contrast, the dynamic range of the speech signal is much wider and the ear is not very sensitive to modest changes in signal gain. Successful algorithms employ a soft mode switch to control adaptation speeds to allow them to accommodate both speech and data signals for modems supporting up to 4800 b/s data rate [10].

Speech channels are also called upon to carry tones. These may be tones generated by simple FSK data modems or by in-band signalling schemes such as those using DTMF tone pairs. The narrowband nature of these tones can cause problems for coders designed for speech. The differences in coding requirements can manifest themselves if the channel introduces transmission errors. Coders employing backward adaptive predictors require that the predictors at the encoder and decoder converge rapidly after transmission errors. Such convergence is much more difficult to guarantee for tones, particularly when relatively simple gradient adaptation techniques are employed for predictor adaptation.

In switched networks, coders may be tandemed with other transmission facilities. Indeed a connection may in some circumstances consist of several links, each of which uses reduced-rate speech coding. For digital links special precautions may be taken so that in the absence of transmission errors, quantization noise does not accumulate over successive coding stages. For instance, the CCITT proposal for 32 kb/s coders incorporates modifications which allow transcoding between PCM links and ADPCM links without introducing further degradation.

In public networks, echo control is required if the one-way signal delay approaches 20 ms. Low delay coders are desirable if the economic benefits of reduced rate coding are not to be eliminated by the cost of installing echo control devices (suppressors or cancellers). Any limit on coder delay must take into account the possibility of several tandem coders in a connection. This suggests that codecs with end-to-end delays exceeding roughly 10 ms will not be readily deployed except in applications where echo control is already required, such as satellite links, or where the tandeming requirements are not very stringent, such as for special private networks.

CONCLUSIONS

Speech coding work has seen great changes in emphasis as a result of the narrowing of the gap between research in algorithms and what is realizable in practical hardware.

This trend will probably continue—developments in integrated circuits bode well for speech coding. One can expect DSP chips with more on-chip memory, faster effective execution times, and even floating point arithmetic capabilities. These new chips will allow the application of even the more complex coding algorithms in cost-effective communications systems. The improved support systems for software preparation will likely reduce the time and cost of producing hardware/software systems for advanced coding applications. In addition, systems integration issues will continue to be important as coders find their way into existing networks.

New algorithmic advances are expected in the 4.8 to 9.6 kb/s range, allowing the lower limit for toll quality coding to drop even further. We can expect to see transmission systems employing codecs in the 10–16 kb/s range and speech storage systems using 4–10 kb/s coded speech, all providing excellent quality speech.

REFERENCES

1. J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech", *IEEE Trans. ASSP*, vol. ASSP-29, pp. 512–530, Oct. 1979.
2. J. L. Flanagan, M. R. Schroeder, B. S. Atal, R. E. Crochiere, N. S. Jayant, and J. M. Tribolet, "Speech Coding", *IEEE Trans. Commun.*, vol. COM-27, pp. 710–736, April 1979.
3. P. Combescure, A. LeGuyader, and M. Haghiri, "ADPCM algorithms applied to wideband speech encoding", *Proc. ICASSP '82*, Paris, France, pp. 1976–1979, May 1982.
4. B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates", *Proc. ICASSP '82*, Paris, France, pp. 614–617, May 1982.
5. B. S. Atal, "Predictive coding of speech at low bit rates", *IEEE Trans. Commun.*, vol. COM-30, pp. 600–614, April 1982.
6. A. Buzo, A. H. Gray, R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization", *IEEE Trans. ASSP*, vol. ASSP-28, pp. 562–574, Oct. 1980.
7. B. Secrest, M. Arjmand, and M. Ni, "Speech analysis and synthesis become practical on μC chip", *Electronic Design*, pp. 129–136, May 27, 1982.
8. W. Daumer et al, "Overview of the ADPCM coding algorithm", to be presented at Globecom '84, Atlanta, Ga.
9. P. Mermelstein and G. Williams, "Network performance issues in 32 kb/s coding of speech and voiceband data", *Proc. Globecom '82*, Miami, Fla., pp. A8.2.1–A.2.5, Dec. 1982.
10. D. W. Petr, "32 kb/s ADPCM-DLQ coding for network applications", *Proc. Globecom '82*, Miami, Fla., pp. A8.3.1–A8.3.5, Dec. 1982.

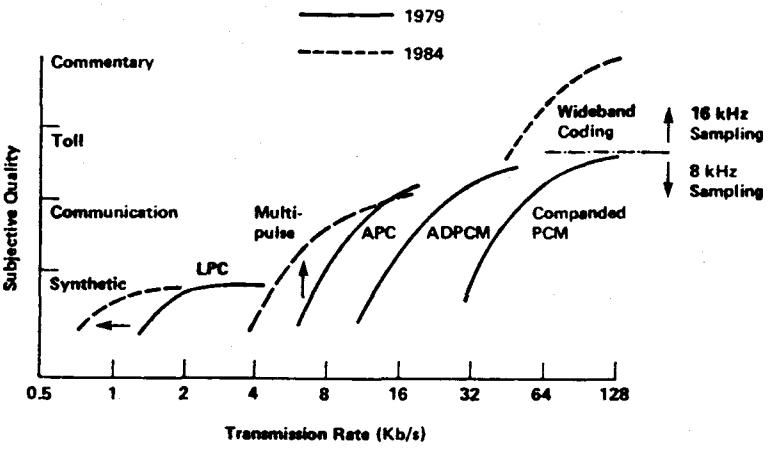

Fig. 1 Speech Quality as a Function of Transmission Rate