Speech Coding for Wireless Communication

Paul Mermelstein

INRS-Telecommunications[1]

Abstract

Personal communication systems will be judged primarily on the quality of speech communication they provide. The voice-call capacity depends directly on the bit-rate required to achieve the quality objectives of the specific service. To provide speech quality on wireless systems that is comparable tc that attained on today's wireline systems, at least 16 kb/s are required. New algorithms are being considered for a standard to transmit speech at 8 kb/s in the presence of modest transmission errors. Mobile systems with slightly lower quality requirements in service today employ 8 kb/s (IS-54 and IS-95) and 13 kb/s (GSM), respectively. Efforts are under way to halve these values and thereby double the capacity. The paper reviews current research into achieving higher speech quality at lower bit rates under a variety of environmental and transmission conditions. It provides a short-range perspective on where significant progress is expected in the short-term.

Introduction

The objectives of this presentation are to discuss the current needs in the area of speech coding for wireless communications and identify some research directions directed at meeting those needs. The discussion will be restricted to cellular and microcellular systems where capacity in terms of speech calls per cell per unit bandwidth is directly related to the bit rate utilized by each conversation. Digital broadcast systems do not suffer from the same bandwidth limitations and the larger bit-rates they utilize are justified by the advantages of providing high-quality wideband audio, possibly in stereophonic form. Personal communication systems provide telephone bandwidth speech to terminals indoors and outdoors and these systems are experiencing rapid growth due to the advantages of terminal mobility and portability. The imminent prosects of greater spectrum availability in the 1.8 GHz region in the near future suggests a continuing explosion of wireless telephony services, possibly replacing wireline services to a large exent. Acceptance by customers depends greatly on being able to provide end-to-end audio quality that is comparable to that of wireline services today.

1. This work was funded by the Bell/BNR/NSERC Industrial Research Chair in Personal Communications, 16 Place du Commerce, Verdun, Quebec H3E 1H6, Canada.

E-mail: mermelstein@inrs-telecom.uquebec.ca

Applications and Requirements

Although speech coding research has a long history, its applications have been generally limited to environments where the costs of signal processing for compression have been outweighed by the cost advantages of increased multiplexing of transmissions within the same bandwidth on the same link. Thus satellite and underwater cable links have long utilized less than 64 kb/s transmissions for the 300-3300 kHz telephone signal. However, shorter-haul wireline channel banks and even microwave transmission links have not followed the same approach. With the digital revolution in cellular communication, speech coding has received renewed impetus. The transition from analog frequency- modulated transmission used in the North American AMPS system to the time-division multiple access (TDMA) systems has required the development of high-quality speech coding techniques near 8 kb/s. Algorithms have been previously standardized by the International Telecommunications Union (ITU) for speech transmission at 32 kb/s (ITU-T G.721 standard) and more recently at 16 kb/s (ITU-T G.728 standard [Chen et al, 1992]). In wireline transmission the expected transmission error rates are quite low and the complexity requirements are less exacting. In contrast, for wireless transmission to and from rapidly moving terminals, robustness to fades and low complexity to minimize power dissipation become more important considerations.

The quality requirements of a newly introduced service are generally determined by the expectations of customers based on experience with similar services in the past. In mobile applications the relevant quality benchmark was that of the previous analog service. For the more general personal communications services, the quality benchmark is expected to be that of wireline speech services today, otherwise known as toll quality. This requires that the

1058-6393/95 $4.00 © 1995 IEEE

otal subjective listening experience be comparable to that attained on 64 kb/s transmissions with very low error rates and delays of the order of a few samples at the sampling frequency of 8 kHz. The ability to achieve this goal, while dependent most crucially on advances in speech coding, is also affected rather significantly by the fading characteristics of the multiple access channel experienced duration longer than about 200 ms are particularly damaging to the reconstructed signal. Retransmission of erroneously received speech frames would require delays of the order of the fade duration. Since delays of that order cannot be introduced into the speech path due to their impact on the audibility of echos generated at imperfectly balanced 2 wire-4 wire junctions in the wireline systems, the effects of long fades must be mitigated by techniques such as frequency hopping or bandwidth spreading by code division (CDMA).

The above considerations have led to the definition of speech coding algorithms for TDMA systems that encode a single direction on a speech conversation into 16-25 kb/s including channel protection and overhead for equalization and signalling requirements. The corresponding raw speech coding rates are 8 kb/s for the North American IS-54 system and 13 kb/s for the Eukropean GSM system. The IS-95 CDMA system uses variable-rate coding with a peak rate of 9.6 kb/s per speech transmission. It exploits the fact that roughly half the time a conversational speech source is inactive by reducing the coding rate at those times to 1.2 kb/s. A similar approach is taken in the proposed E-TDMA system [Kay, 1992], where groups of TDMA channels are pooled to allow demand-based slot assignment whenever a conversation changes from inactive to active speech. These first generation digital speech coding algorithms yield a speech quality considered acceptable for cellular applications, namely one where the speech is highly intelligible but where both quantization noise and the effects of fading are audible.

The current primary research objectives in speech coding for wireless are twofold:
1.To improve the quality of the near 8 kb/s systems to toll,
2.To reduce the coding rate near 4 kb/s while maintaining the near 8 kb/s quality.
The first requirement is needed to achieve high acceptance in indoor microcellular environment where the alternative is a wireline phone. The second requirement would double the capacity of currently used cellular systems without increasing the spectrum space needed.

Evaluation of such codecs involves testing the quality in the absence of transmission errors, typically measured by collecting mean opinion score (MOS) ratings of groups of listeners of the same source materials passed through test and reference codecs. Quality assessment of signals decoded in the presence of specified levels of transmission errors ensures that the degradation within the operating region remains within specified limits. Since the latter evaluation involves the channel coder as well, and the channel coder is usually designed to fit the fading environment of the multiple access technique, the speech and channel coder combination that meets the degradation limits for one access environment need not be the most suitable for a different access environment. Additional performance criteria include signal delay, robustness to background noise and robustness to speaker variations as well as the presence of multiple speakers. Finally, robustness to tandeming or repeated encoding-decoding operations ensures good quality in mobile to mobile transmission situations.

Low power microcellular systems today employ the 32 kb/s ADPCM standard G.721 [Daumer et al., 1984]. Its advantages are low delay and low complexity. However, since it was designed for the low bit-error rates of the wireline environment, it is not well suited to transmission conditions with high error rates. While the quality is indeed comparable to wireline under good transmission conditions, the codec is difficult to protect against transmission errors unless much more processing is carried out to derive parameters such as pitch and spectrum envelope which lower-rate codecs extract as part of the encoding process. In most cases only error-detection is employed followed by muting the output in case of errors in the received block. However, the encoder and decoder states may differ after transmission errors and the effects of errors will persist until the decoder can retrack the encoder. As the power dissipation of codecs realized in VLSI decreases, we can expect lower rate codecs of increased complexity to be employed in the low power environments as well.

Linear Prediction Coding of Speech Signals

Almost all modern low bit-rate coding techniques fit into the general linear prediction coding framework introduced by Atal and Schroeder [1979]. This framework models the signal as resulting from the excitation of a quasi-stationary linear filter by a time-varying and partially periodic source. The filter parameters vary relatively slowly and can be assumed fixed for short intervals of the order of 10-20 ms. By inverse filtering a segment of signal with a linear filter of order about 10 optimized for that segment, a residual signal is obtained whose main characteristic for voiced sounds is a dominant periodicity or pitch in the frequency range 50 to 500 Hz. Inverse filtering by a corresponding long-term filter of period between 20 and 2 ms results in a quasi-random signal. An approximation to this residual reconstructed at the receiver can be used to excite appropriate long-term (pitch) and short-term (linear prediction) filters to yield a reconstructed speech signal. To allow reconstruction with adequate fidelity, we need to transmit at least the short-term filter parameters, the long-

term filter delay and gain values, and some scalar or vector quantized form of the residual signal. Typically 25 to 30 bits are needed to quantize effectively the parameters of a 10th order LPC filter, most frequently represented in terms of line spectral pair (LSP) frequencies [Sugamura and Itakura, 1981]. Scalar quantization of the residual does not generally achieve a sufficiently low bit-rate, therefore some generalized form of vector quantization is used. The complexity of the coder tends to be dominated by the residual quantization operation at the transmitter. The GSM system employs a particularly simple form of quantization, namely regular pulse excitation (RPE). This technique employs groups of regularly spaced pulse sequences of equal magnitude pulses with adjustable starting positions [Kroon et al., 1986]. The regular pulse groups provide broadband excitation across the 4 kHz band and one amplitude parameter is used to ensure that each 5 ms segment has the correct amount of residual energy. The resulting residual spectra, while overall flat, reveal significant power variation with frequency which may give rise to audible artifacts.

The most general technique for selecting the best match to the input signal segment with respect to the error criterion is a search through a large codebook of stochastic signals [Schroeder and Atal, 1985]. Since the desirable characteristics of the codebook entries are not generally known explicitly, it is frequently effective to train the codebook by first populating it with random entries and discarding those entries that turn out to be used rarely in the process of coding a wide variety of speech signals. The resulting search has to be carried out in a closed-loop form as shown in Fig. 1.The complex error surface manifests numerous secondary minima whose structure is not well understood. Open-loop approximation of course reduce the complexity of the search process, but degrade the resulting performance as well. Thus reducing the complexity of the codebook search remains the main impediment to use of larger codebooks and thereby achieving higher quality.

Various objective error criteria have been introduced in an attempt to approximate the subjective auditory criteria of the human ear. Thus the mathematically attractive minimum mean square error minimization one may be tempted to use to search for a codebook entry that best matches the segment of residual signal at hand is replaced by mean square minimization after perceptual filtering the difference signal between the input speech and the reconstructed approximation. The time-variant perceptual filter attempts to distribute the quantization noise so that it is best masked by the speech signal at all times. The quasi-stationary treatment of the speech signals permits ready exploitation of the frequency masking characteristics. However, temporal masking has generally not been introduced into the error to be minimized. Even in frequency masking one generally employs a linear approximation to a filter that is known to be nonlinear [Zwicker et al, 1990].

Channel Error Protection

The quality of the reconstructed speech is sensitive to differing extents to errors introduced by the transmission process into the received speech coding parameters. To make best use of a limited number of available error protection bits, we tend to protect with both error correction and detection capability those transmitted parameters in which errors would cause the most significant quality degradation. Both GSM and IS-54 use one-half rate convolutional coding for this purpose. The gain and parameter values of both the pitch and LPC filters tend to be protected, but the information concerning the quantized residual may carry error detection only. In many codecs only a most significant subset of the bits carries error protection to improve further the efficiency of the channel codec.

In CDMA systems the channel error coding can form part of the bandwidth spreading process. This reduces the effective overhead for channel protection and allows all parameters of the speech signal to be protected equally.

An error masking process can be invoked to mitigate the effects of channel errors of short duration that have been detected but not corrected. These techniques attempt to exploit the short-time stationarity of the signal by extrapolating the filter paramters from the past. While the subjective improvement resulting from these techniques is significant for durations of less than about 200 ms, for longer fades generating an erroneous signal appears worse than allowing the reconstructed signal energy to decay to zero.

The overall speech quality of a cellular service tends to be judged by that experienced under worst case transmission conditions. In TDMA systems random bit error rates can be as high as 3% after channel decoding. In CDMA systems, because the grade of service requirement is set somewhat higher in terms of channel quality - BER less than 0.1% more than 99% of the time - the peak error rate would not generally exceed 1%. It is the long burst errors, however, that tend to be most damaging to speech quality. Some GSM systems limit the maximum burst error duration by use of frequency hopping. Since the worst conditions will generally be experienced just prior to handoff to a neighboring cell, rapid handoff capabilities can have a significant effect on overall speech quality.

Proposed standard at 8 kb/s

The ITU is currently considering standardization of a speech codec near 8 kb/s for personal communications applications. Two codecs, one following an algebraic

coding approach from the University of Sherbrooke, Canada [Lefebvre et al., 1993], and one using multiple codebooks from NTT, Japan [Kataoka et al., 1993], have demonstrated toll quality under most evaluations conditions. These codecs employ frames of 10 ms or less to result in an algorithmic coding delay of less than 16 ms. More recently an effort has started to combine the most valuable features of each to produce a codec meeting all the quality requirements.

Recent research at INRS-Telecom

A parallel research effort at INRS-Telecommunications has attempted to reduce the complexity of the codebook search as well as model more effectively the long-term filter. To this end we have designed a multi-band filter as shown in Fig 2. .The long-term or pitch synthesis filter is separated into three parallel filters with different loop gains operating on separate band limited components of the residual. The three bands occupy the spectrum regions 0-1 kHz, 1-2 kHz and 2-4 kHz. The three pitch filters employ a common delay value, that found optimal in analyzing the periodicity of the full-band inverse-filtered residual signal. The multi-band filter structure allows better reconstruction of the harmonic structure observed for many speech segments, namely the magnitude of the voicing harmonics appears to decrease with increased frequency. The resulting filter structure achieves a significant improvement in the total prediction gain of the pitch filter.

The same three-band structure can be used to decompose the codebook search process into three search operations one for each band. The residual signal is coded in 5 ms segments, larger than the 1.25 ms segments used in the LD-CELP standard for 16 kb/s coding but considered to provide adequate temporal resolution for gain variations. We observe that a large codebook needs to be retained only for the lowest band [Mermelstein et al., 1994] Only in that band is providing precise time resolution in the variation of the excitation energy most important. Reducing the sizes of the higher-frequency codebooks has little effect on the quality of the generated speech. Since the low-frequency codebook can be subsampled to 0.5 us from the original 0.125 us, the search complexity for the first codebook can be as low as 1/4 of the full-band codebook. In addition, specifying the energy levels of the residual in the three bands independently allows a much improved match to the original residual which frequently deviates from a flat spectrum form even though that is what the optimum linear filter was meant to generate.

The resulting codec is found to produce excellent speech quality at 7 kb/s. A more formal subjective quality evaluation is in progress. The results indicate that judicious use of parameters in both the time and frequency domain can go a long way towards exploiting the complementary characteristics of the hearing system, namely good time.resolution at low frequencies and good frequency resolution at the higher frequencies.

Conclusions

Advances in speech coding research are occurring rapidly. Toll quality codecs with reduced complexity for speech coding near 8 kb/s are expected to be standardized within the next two years. Availability in VLSI chips should follow quickly thereafter. Similar quality codecs at 4 kb/s require additional research and may not become widely available for another five years. However, applications that do not require the same high robustnes to heavy rates of channel errors because of a more benign transmission environment may be able to exploit such codecs earlier. The higher quality and reduced bit-rates promise to enhance the capacity and performance of a new generation of personal communication systems expected to be introduced within the next five years.

References

Atal, B.S., and M.R. Schroeder (1979) "Predictive coding of speech signals and subjective error criteria," IEEE Trans. Speech, Signal Proc. 27, 247-254.

Chen, J.H., R.V. Cox, Y.C. Lin, N. Jayant, M.J. Melchner (1992),"A Low-Delay CELP Coder for the CCITT 16 kb/s Speech Coding Standard", IEEE Jour. Sel. Areas Comm. 10, 830-849.

Daumer, W.R., P. Mermelstein, X. Maitre, I. Tokizawa,"Overview of the ADPCM Algorithm", Conf. Record Globecom '84, Atlanta, GA., 23.1.1 - 23.1.4

Gersho, A. (1994), "Advances in Speech Coding and Audio Compression," Proc.IEEE 82, 900-918.

Kataoka, A.,T. Moriya and S. Hayashi (1993) "An 8 kb/s speech coder based on conjugate structure CELP," Proc. IEEE Int. Conf. Acoust. Speech, Sig. Proc., Minneapolis, 11-592-595.

Kay, S. (1992),"Extended-TDMA, A High Capacity Evolution of US Digital Cellular", First Int. Conf. Pers. Comm., Dallas, Texas, 182-184.

Kroon, P., E.F. Deprettere and R.J.Sluyter (1986), "Regular pulse excitation: A novel approach to effective and efficient multipulse coding of speech," IEEE Trans. Acoust. Speech, Sig. Proc. ASSP-34, 1054-1063.

Lefebvre, R., R. Salami, C. Laflamme and J.P.Adoul (1993), "8 kbit/s coding of speech with 6 ms frame lengths, " Proc. IEEE Int. Conf. Speech, Sig. Proc., Minneapolis, 612-615.

Mermelstein, P., P. Zheng and M. Saikaly (1994), "Multiband residual coding of Celp codecs at 8 kb/s," Proc. IEEE Int. Conf. Acoust. Speech, Sig. Proc., Adelaide, Australia, II-117-120.

Schroeder, M.R. and B.S. Atal (1985), "Code Excited Linear Prediction (CELP) - High quality speech at very low bit rates," Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc., 937-940.

Sugamura, N. and F. Itakura (1981), "Line spectrum representation of linearpredictor coefficients of speech signals and its statistical properties,"Trans, Inst. Electron, Commun.Eng., Japan, J64-A, 323-340.

Zwicker, E. and H. Fastl (1990), Psychoacoustics. Facts and Models, Springer Verlag.

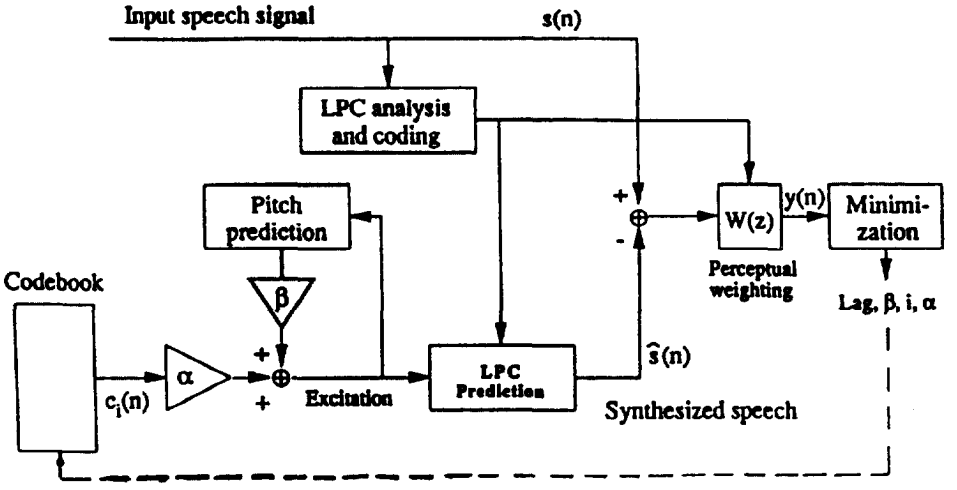

Fig.1. Code excited linear prediction encoder

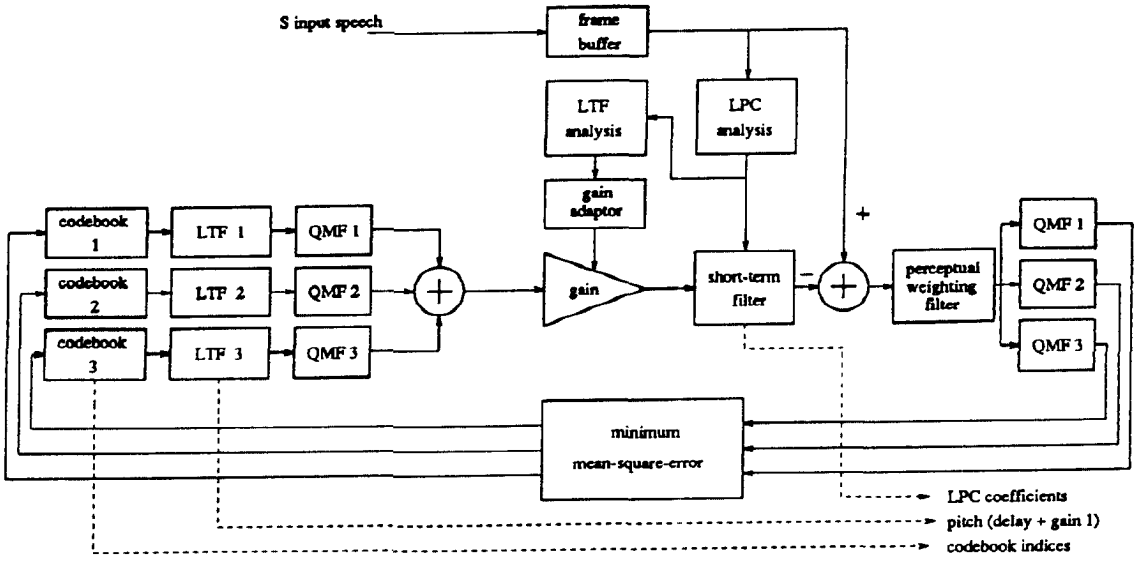

Fig.2. Celp encoder with subband codebooks and multiband pitchfilter