MULTIPLEXED PREDICTIVE CODING OF SPEECH

Søren Vang Andersen

Department of Speech, Music, and Hearing
KTH (Royal Institute of Technology)
D. Kristinas v. 31 — SE-100 44 Stockholm — Sweden
Email: sva@speech.kth.se

Gernot Kubin

Institute of Communications and Wave Propagation
Graz University of Technology
Inffeldgasse 16c — A-8010 Graz — Austria
Email: G.Kubin@ieee.org

ABSTRACT

In this paper we present a novel method for predictive coding with application to transmission of speech over packet-switched networks. Our method uses multiplexing to distribute a part of the information about a segment of each speech signal in several data packets while keeping the data packet rate and payload for that part of the information unchanged. We investigate three multiplexing schemes: a packet hopping, a Hadamard multiplexing, and an extension of the Hadamard multiplexing that exploits a nonlinear preprocessing and estimation method. We show by means of formal AB-preference tests that multiplexed predictive coding can lead to coders that are more robust to packet losses than scalar quantization and packet loss concealment according to the G.711 standard.

1. INTRODUCTION

Transmission of speech and audio signals over packet-switched networks has recently become a topic of significant interest, the most prominent example being Internet telephony.

The internet protocol makes efficient use of network resources only when allowing IP routers to drop packets. Such packet loss may cause severe impairments to the speech or audio quality. To mitigate these effects, advanced protocol mechanisms have been proposed as well as two classes of signal processing methods: A first class adds redundancy and significant delay at the transmitter side, e.g., with loss-resilient codes[1] or with multiple-description source/channel coding[2, 3, 4, 5]. A second class attempts perceptual concealment of the packet loss through signal interpolation at the receiver side[6, 7]. We propose a new class of signal processing methods which modify the transmitter without increasing the network payload data rate while minimizing the perceptual effect of packet losses at the receiver side.

We develop our method in the framework of a heterogenous networking scenario where some IP traffic may share a connection between two gateways. In this context, multiplexing can be used to distribute a part of the information about one sound segment in several data packets while keeping the data packet rate and payload for that part of the information unchanged. A lost packet leads to a partial loss, i.e., the degradation is smeared over several sound segments, instead of a total loss of a single sound segment. We show by experiments in this paper that the modified loss characteristic obtained by multiplexing can lead to degradations that are less objectionable to the listener than those originating from a packet loss concealment method. The main drawback of multiplexing is that it depends on encoding and packetization of multiple sound segments in parallel.

While multiplexing seems applicable to a wide range of coding algorithms including filter-bank coders, transform coders, and predictive coders, we focus in this paper on the application of multiplexing to scalar predictive speech coders with perceptual weighting. In such coders, multiplexing is advantageously employed in the encoding of the prediction residual signal, and combined with another method such as loss-resilient coding for the robust transmission of side information. The prediction synthesis filter in the receiver performs appropriate perceptual weighting of the transmission errors due to packet losses. The restriction to scalar quantization is made to keep complexity moderate for the entire system.

2. MULTIPLEXED PREDICTIVE CODING

We consider a collection of K scalar adaptive predictive speech coders with noise feedback coding of the kind proposed by Atal and Schroeder[8]. In this system we replace the traditional single-input single-output scalar quantizer with what we call a multiplexed quantizer. The multiplexed quantizer takes at each sampling instant n the K inputs q_n^1 to q_n^K from all K predictive encoders and outputs quantized representations $\hat{q}_n^1$ to $\hat{q}_n^K$ back to the individual predictive encoders. In doing this, the multiplexed quantizer generates K quantization indices i_n^1 to i_n^K each in the range 1 to 2^b where b is the number of bits allocated for each index. These indices are packetized and transmitted over the packet-switched network in K independent packet streams. At the decoder side, the received indices $\tilde{i}_n^1$ to $\tilde{i}_n^K$ are available. These indices differ from the indices i_n^1 to i_n^K in the encoder whenever an index has been lost with a data packet on the network, in which case the index value is replaced by zero. The demultiplexer resolves K representations $\tilde{q}_n^1$ to $\tilde{q}_n^K$ of the quantized information from the available indices $\tilde{i}_n^1$ to $\tilde{i}_n^K$. These representations are subsequently input to the K predictive decoders. Our encoding and decoding system is shown in Figure 1. Not included in this figure is the LPC analysis and side information quantization which leads to coefficients for the predictors $\hat{P}_n^k(z)$, noise-feedback filters $F_n^k(z)$, and scaling factors $\hat{\sigma}_n^k$ for $k = 1..K$. The side information is conveniently encoded with a loss-resilient coding such as the Reed-Solomon code[9] to enable transmission of this information in a robust manner using the same K packet streams as the multiplexed quantized information.

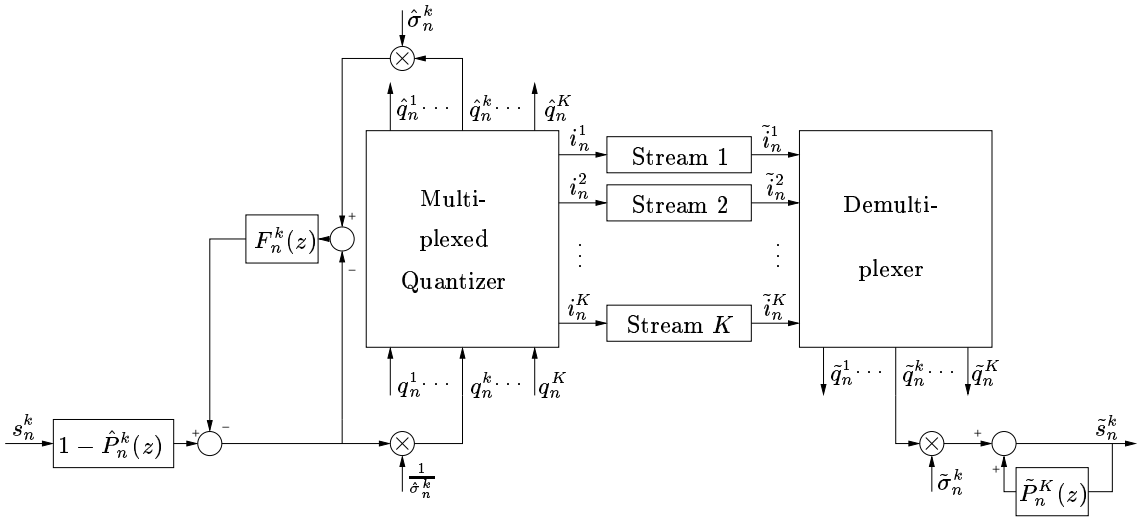

Figure 1: The multiplexed encoding and decoding systems. Only the $k'th$ predictive encoder and decoder are shown.

2.1. Packet Hopping

The simplest system that we can think of for the multiplexed quantizer block is a system in which indices from scalar quantizers are hopped, i.e., cycled from one packet stream to the next as the sample instant n increments. One version of this system is specified by the equations

$$i_n^k = Q\left(q_n^{(n+k)\mathrm{mod}(K)+1}\right), \tag{1}$$

$$\hat{q}_n^{(n+k)\mathrm{mod}(K)+1} = Q^{-1}\left(i_n^k\right), \tag{2}$$

and

$$\tilde{q}_n^{(n+k)\mathrm{mod}(K)+1} = Q^{-1}\left(\tilde{i}_n^k\right), \tag{3}$$

for $k = 1..K$. Here $Q(\cdot)$ and $Q^{-1}(\cdot)$ are the mappings from quantizer input to quantization index and from quantization index to the quantized representation of the quantizer input, respectively. For an adequate response to packet losses $Q^{-1}(0) = 0$. The notation $(\cdot)\mathrm{mod}(K)$ denotes the modulo K operation.

2.2. Hadamard Multiplexing

The packet hopping described in Section 2.1 can be expressed as a particular orthogonal transformation of the input to the multiplexed quantizer followed by a quantization of the transform output. Define column vectors with elements equal to the K scalar input, output, or index values for the multiplexed quantizer and demultiplexer, such that e.g.,

$$\boldsymbol{q}_n \equiv \left[q_n^1\ q_n^2\ \cdots\ q_n^K\right]^T.$$

Then the multiplexed quantizer is defined by the equations

$$\begin{aligned} \boldsymbol{c}_n &= \boldsymbol{M}_n\boldsymbol{q}_n, \\ \boldsymbol{i}_n &= Q(\boldsymbol{c}_n), \\ \hat{\boldsymbol{c}}_n &= Q^{-1}(\boldsymbol{i}_n), \end{aligned}$$

and

$$\hat{\boldsymbol{q}}_n = \boldsymbol{M}_n^T\hat{\boldsymbol{c}}_n.$$

The demultiplexer on the receiver side is defined by

$$\tilde{\boldsymbol{c}}_n = Q^{-1}(\tilde{\boldsymbol{i}}_n),$$

and

$$\tilde{\boldsymbol{q}}_n = \boldsymbol{M}_n^T\tilde{\boldsymbol{c}}_n.$$

The equivalence of these equations with the packet hopping described by Equations 1 to 3 is obtained by letting the transform matrix $\boldsymbol{M}_n$ equate an adequate time varying row or column permutation of an identity matrix.

With this formulation of the multiplexing, it is relevant to investigate other transform matrices than the row or column permuted identity matrix. A simple, yet relevant, transform for this purpose is the normalized Hadamard transform[10]. We expect the Hadamard multiplexing to hold advantages over the packet hopping method. These advantages are explained in the following.

The variance scaled prediction errors from the multiple predictive coders can well be assumed to be independent and identically distributed. Therefore, one advantage of the Hadamard transform is that the elements of the transformed vector $\boldsymbol{c}_n$, i.e., the inputs to the quantizers become closer to Gaussian. This is a result of the central limit theorem[11]. More Gaussian quantizer inputs result in less outliers and thereby less overload distortion in the coded prediction errors. Another advantage is that whenever less than K packets are lost in the network there are no full erasures of any sample in any of the quantized prediction error signals. This advantage can be exploited when the Hadamard transform is combined with a nonlinear preprocessing and estimation scheme as described in the next section.

2.3. Nonlinear Preprocessing and Estimation

Let us assume the elements of $\boldsymbol{q}_n$ to be uncorrelated and neglect the impact of quantization noise. Then the matrix $\boldsymbol{M}_n^T$ is the linear minimum mean-squared error estimator for $\boldsymbol{q}_n$ given the coefficient vector $\tilde{\boldsymbol{c}}_n$. This estimator is the mean-square optimum for Gaussian $\boldsymbol{q}_n$. However, the

gain-scaled linear prediction errors for voiced speech signals are known to be non-Gaussian. Thus, a nonlinear estimator can result in lower mean-squared error. Indeed, we observed in preliminary experiments that nonlinear estimation could lead to a significant decrease of the mean-squared error. However, the nonlinear estimation led to very high computational complexity. What we instead propose in this paper is an alternative method in which a well defined nonlinearity is applied to the input of the multiplexed quantizer. Knowledge of this nonlinearity can then subsequently be exploited to improve the reconstructed quantized prediction errors in the case of packet losses.

The general method that we propose is to zero K_{zi} of the K inputs to the multiplexed quantizer prior to applying the transform $\boldsymbol{M}_n$. Advantageously, the K_{zi} inputs with lowest amplitudes are set to zero. In our method no information about the position of zero valued elements in $\hat{\boldsymbol{q}}_n$ is conveyed to the decoder, only the knowledge that K_{zi} of the elements were zero is exploited. The method is described as follows.

Suppose that the number of lost packet streams is K_{lps} and the lost packet streams are indexed by an integer set $\boldsymbol{k}_{lps}$. Then we may formulate a set of equations relating the received coefficients $\tilde{\boldsymbol{c}}_n$ with the encoded scaled prediction errors $\hat{\boldsymbol{q}}_n$.

$$\hat{\boldsymbol{q}}_n = \boldsymbol{M}_n^T \tilde{\boldsymbol{c}}_n + \boldsymbol{M}_n^T(:, \boldsymbol{k}_{lps})\boldsymbol{\alpha}. \quad (4)$$

In this equation $\boldsymbol{\alpha}$ is a K_{lps} dimensional unknown vector. We have used a matlab-style notation $\boldsymbol{M}_n^T(:, \boldsymbol{k}_{lps})$ to denote a matrix consisting of the columns of $\boldsymbol{M}_n^T$ that are indexed by $\boldsymbol{k}_{lps}$.

Now assume that $\hat{\boldsymbol{q}}_n$ had K_{lps} zero-valued elements indexed by the set $\boldsymbol{k}_{zi}$: $\hat{\boldsymbol{q}}_n(\boldsymbol{k}_{zi}) = \boldsymbol{0}$, then

$$\boldsymbol{\alpha} = -\boldsymbol{M}_n^T(\boldsymbol{k}_{zi}, \boldsymbol{k}_{lps})^{-1} \boldsymbol{M}_n^T(\boldsymbol{k}_{zi}, :)\tilde{\boldsymbol{c}}_n,$$

provided that $\boldsymbol{M}_n^T(\boldsymbol{k}_{zi}, \boldsymbol{k}_{lps})$ has full rank. Furthermore,

$$\boldsymbol{\alpha} = \hat{\boldsymbol{c}}_n(\boldsymbol{k}_{lps}). \quad (5)$$

The indexing for the zero-valued elements of $\hat{\boldsymbol{q}}_n$ is not known by the decoder, however there are

$$C_1 = \begin{pmatrix} K \\ K_{lps} \end{pmatrix}$$

ways in which the decoder can assume K_{lps} elements of $\hat{\boldsymbol{q}}_n$ to be zero. Of these

$$C_2 = \begin{pmatrix} K_{zi} \\ K_{lps} \end{pmatrix}$$

will be true assumptions. Whenever $K_{zi} > K_{lps}$ there are multiple true assumptions and all true assumptions will result in the same value for $\boldsymbol{\alpha}$, i.e., the one given in Equation 5. Thus, the method applicable in the decoder is to calculate $\boldsymbol{\alpha}$ for all C_1 possible choices of $\boldsymbol{k}_{zi}$ and select the $\boldsymbol{\alpha}$ vector that occurred C_2 times. Hereafter $\tilde{\boldsymbol{q}}_n$ is obtained as the right-hand side of Equation 4.

Rank deficiency of the matrix $\boldsymbol{M}_n^T(\boldsymbol{k}_{zi}, \boldsymbol{k}_{lps})$ limits the use of this method. For example, when $\boldsymbol{M}_n$ equals the permuted identity matrix that follows from the packet hopping our nonlinear preprocessing and estimation does not apply. In contrast, when $\boldsymbol{M}_n$ is the normalized Hadamard transform, the method applies with no complications for $K_{lps} = 1$. For $K_{lps} > 1$, rank deficiency can occur for some of the possible choices of $\boldsymbol{k}_{zi}$. In this case heuristics must be introduced in the selection of $\boldsymbol{\alpha}$.

3. CODING EXPERIMENT

We conducted a coding experiment in which 12 speech files, each containing two utterances, were jointly encoded by multiplexed predictive coders ($K = 12$). Each predictive coder had a 10th order linear predictive filter updated every 20 ms using a 30 ms Hann window and the autocorrelation method to calculate the linear predictive coefficients. The 10 coefficients were uniformly quantized in the log area ratio domain using 5, 4, 3, 3, 2, 2, 2, and 1 bits respectively. A long-term section was included in the prediction filter. This section had one non-zero coefficient at an optimized lag in the range 20 to 147. The long-term lag and coefficient was determined every 20 ms using the covariance method and quantized using 5 and 7 bits respectively. Finally the residual standard deviation σ_n^k was determined every 20 ms and uniformly quantized in the logarithmic domain using 6 bits. In total 40 bits were allocated for side information every 20 ms.

The side information from 20 ms of all 12 speech signals was organized in 3 messages of 160 bits each. For these 3 messages a Reed-Solomon code can be designed to generate 12 messages of 160 bits each from which the complete side information can be recovered upon reception of any 3 of these 12 messages[9]. The 12 messages could constitute the first part of the information in packets on each of the 12 packet streams. When, in addition, the multiplexed prediction errors were quantized using 3 or 4 bits per sample and appended to the messages in these packets then this resulted in an encoding with a total bit rate of 32 or 40 kbps per speech signal and with a packet rate of one every 20 ms in each of the packet streams.

The noise feedback filter $F_n^k(z)$ was derived from unquantized linear predictive coefficients using a bandwidth expansion factor of 0.6. We applied a postfilter following the design of Chen and Gersho[12]. The parameters of the postfilter were adapted to the number of lost packet streams. In the postfilter, the numerator and denominator coefficient of the pitch-sharpening section, and the bandwidth expansion factor of the short-term denominator were increased linearly with the number of lost packet streams. Simultaneously, the bandwidth expansion factor of the short-term numerator was decreased linearly.

In the experiment, the nonlinear preprocessing and estimation method was applied on dimension 4 subsets of the 12 coders and with $K_{zi} = 2$ for each of these subsets. This multiplexing was supplemented with packet hopping of the dimension 4 coefficient sets.

Speech files were encoded at both 32 and 40 kbps and decoded after that a percentage of the data packets had been randomly dropped. Random packet loss rates between 0% and 40% were simulated. As a reference system we used a 64 kbps μ-law quantization and packet loss concealment (PLC) according to the ITU-G.711 standard[7]. The reference system was simulated for the same speech files and packet losses as the multiplexed predictive coders.

Packet loss rate	0%	10%	20%	30%	40%
Packet Hopping	18.9	12.3	9.0	7.1	5.5
Hadamard Mult.	21.7	13.0	9.4	7.3	5.7
Nonlinear Method	16.5	15.1	11.6	8.5	6.6

Table 1: Seg-SNR measures for the multiplexed predictive coders at 32 kbps. Packet hopping, Hadamard multiplexing, and nonlinear preprocessing and estimation are compared.

Packet loss rate	0%	10%	20%	30%	40%
Packet Hopping	23.7	13.4	9.5	7.5	5.8
Hadamard Mult.	27.0	13.7	9.8	7.6	6.0
Nonlinear Method	19.1	17.2	12.7	9.4	7.0

Table 2: Seg-SNR measures for the multiplexed predictive coders at 40 kbps. Packet hopping, Hadamard multiplexing, and nonlinear preprocessing and estimation are compared.

4. RESULTS

Segmental signal-to-noise ratio (Seg-SNR) measures were obtained by averaging over all 12 speech files and all 20 ms segments with a standard deviation larger than the average standard deviation within each speech file minus 40 dB. The Seg-SNR measures were obtained for the decoded signals prior to postfiltering. The results are given in Tables 1 and 2. We see that when the packet loss rate is nonzero the Hadamard multiplexing, and especially the nonlinear preprocessing and estimation, resulted in Seg-SNR measures that were consistently higher than those obtained by the packet hopping. As a maximum, an improvement of 3.8 dB was observed for the nonlinear method over the packet hopping. This occurred for coding at 40 kbps with a packet loss rate of 10%. We also noticed that without packet losses the Hadamard multiplexing had significantly higher Seg-SNR measures than the packet hopping, which without packet losses has the performance of a standard adaptive predictive coder. This was consistent with our expectations according to Section 2.2.

Informal listening revealed that the effect of packet losses in the multiplexed predictive coders is to introduce a stationary sounding noise source in the decoded signal. The adaptive postfilter was able to lower the perceived loudness of this noise significantly. Perceptually, the noise appeared as a very different type of distortion than the tonal or transient like noise, which is the result of packet losses when G.711 with PLC is used. During informal listening, we observed no large difference between the packet hopping and the Hadamard multiplexing. However, the substantial gain in Seg-SNR for the nonlinear preprocessing and estimation method was clearly audible, especially for packet loss rates of 10% and 20%. This improvement was obtained at the cost of a slight degradation of perceived quality when no packets were lost.

We conducted a formal AB-preference test of the 40 kbps multiplexed predictive coder with packet hopping versus the reference system. In this test 7 listeners were subjected to 24 utterances processed by the two systems in randomized order. For each utterance the listeners made a preference decision. The results are given in Table 3. We see from these results that multiplexed predictive coding was preferred over G.711PLC for packet loss rates in the range from 10% to 40%.

Packet loss rate	0%	10%	20%	30%	40%
Preference to MPC	35%	63%	77%	92%	97%
Preference to G.711	65%	37%	23%	8%	3%

Table 3: Results from AB-preference test of the 40 kbps multiplexed predictive coder (MPC) with packet hopping versus the 64 kbps G.711PLC system.

5. CONCLUSION

In this paper we have described methods for multiplexed predictive coding and shown by means of formal AB-preference tests that such methods can lead to coders that are more robust to packet losses than scalar quantization and packet loss concealment according to the G.711 standard. In addition, the multiplexed predictive coders can be designed to operate at significantly lower bit rates than the G.711 standard.

6. REFERENCES

[1] M. Luby, M. Mitzenmacher, A. Shokrollahi, D. Spielman, and V. Stemann, "Practical loss-resilient codes," in *Proc. 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, January 1998.

[2] G. Kubin and W. B. Kleijn, "Multiple-description coding (mdc) of speech with an invertible auditory model," in *IEEE Speech Coding Workshop*, Porvoo, Finland, 1999.

[3] C. C. Lee, "Diversity control among multiple coders: A simple approach to multiple descriptions," in *IEEE Speech Coding Workshop*, Wisconsin, Sept. 2000, pp. 69–71.

[4] O. Ghitza and P. Kroon, "Dichotic presentation of interleaving critical-band envelopes: an application to multi-descriptive coding," in *IEEE Speech Coding Workshop*, Wisconsin, Sept. 2000, pp. 72–74.

[5] A. K. Anandakumar, A. V. McCree, and V. Viswanathan, "Efficient, celp-based diversity schemes for voip," in *ICASSP*, Istanbul, 2000.

[6] J. C. De Martin, T. Unno, and V. Viswanathan, "Improved frame erasure concealment for CELP-based coders," in *ICASSP*, Istanbul, 2000, pp. III–1483–1486.

[7] "Pulse code modulation (PCM) of voice frequencies. Appendix I: A high quality low-complexity algorithm for packet loss concealment with G.711," 1999.

[8] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 27, no. 3, pp. 247–254, 1979.

[9] S. B. Wicker and V. K. Bhargava, *Reed-Solomon Codes and their Applications*, IEEE Press, New York, 1994.

[10] A. V. Geramita and J. Seberry, *Orthogonal Designs; Quadratic Forms and Hadamard Matrices*, Marcel Dekker, 1979.

[11] Z. Peyton and J. Peebles, *Probability, Random Variables, and Random Signal Principles*, Mc Graw-Hill, 1993.

[12] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 59–71, 1995.