AUTOREGRESSIVE MODEL-BASED SPEECH PACKET-LOSS CONCEALMENT

Guoqiang Zhang and W. Bastiaan Kleijn

ACCESS Linnaeus Center, Electrical Engineering
KTH - Royal Institute of Technology
Stockholm, Sweden
{guoqiang.zhang,bastiaan.kleijn}@ee.kth.se

ABSTRACT

We study packet-loss concealment for speech based on autoregressive modeling using a rigorous minimum mean square error (MMSE) approach. The effect of the model estimation error on predicting the missing segment is studied and an upper bound on the mean square error is derived. Our experiments show that the upper bound is tight when the estimation error is less than the signal variance. We also consider the usage of perceptual weighting on prediction to improve speech quality. A rigorous argument is presented to show that perceptual weighting is not useful in this context. We create simple and practical MMSE-based systems using two signal models: a basic model capturing the short-term correlation and a more sophisticated model that also captures the long-term correlation. Subjective quality comparison tests show that the proposed MMSE-based system provides state-of-the-art performance.

Index Terms— Least mean square methods (MMSE), Packet loss concealment (PLC), Autoregressive processes

1. INTRODUCTION

With the widespread usage of the internet, voice over IP has become increasingly popular. To combat the unreliable delivery of voice packets over the internet, various *packet-loss concealment* (PLC) approaches have been proposed. PLC approaches are particularly useful when existing codecs are used without forward error correction. The basic principle behind PLC is to exploit the redundant information about the missing speech segment that is embedded in neighboring packets. Many of the approaches to PLC are formulated in a heuristic manner. However, the procedures can roughly be divided into methods that can be motivated with a maximum likelihood (ML) based criterion and methods that can be motivated with the minimum mean squared error-based (MMSE) criterion.

A large number of PLC procedures replace the missing segment with a signal that is generated by a model that is similar to that of the previous signal segment. These methods can be interpreted as ML methods. The *typical set* theorem from information theory states that asymptotically with increasing length any sequence generated by the model is equally likely [1]. A commonly used speech model is the *autoregressive* (AR) model, which is estimated using linear prediction (LP). Gündüzhan and Momtahan proposed to construct the excitation signal for the autoregressive model of the missing segment by repeating the excitation signal of the previous received speech with a periodicity that equals the pitch period [2]. In [3], Wong et al. propose to classify the missing segment into voiced, unvoiced or partially voiced types and construct the excitation signal correspondingly. The conventional PLC methods performed in the waveform domain (e.g., [4]) can also be motivated from a ML-based perspective. The missing packet is assumed to exhibit similar characteristics as the neighboring speech and signal repetition is used to cover the missing gap.

This work was funded by Swedish Research Council grant 2005-4107.

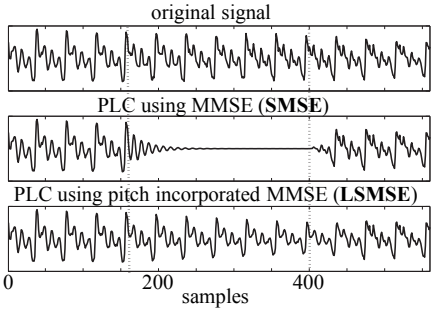

Fig. 1. PLC performance for a voiced speech segment using the proposed **SMSE** and **LSMSE** algorithms. The speech segment between the two dotted lines is assumed lost.

The literature on usage of the MMSE criterion in the PLC context is relatively limited. In [5], Rødbro et al. employ a hidden Markov model (HMM) to track the evolution of speech features such as the pitch frequency. Benefiting from the sophisticated HMM model, the feature vector of a missing packet is estimated using the MMSE criterion and the harmonic sinusoidal parameters for synthesizing the speech are then constructed from the feature vector. In [6], Kondo and Nakagawa propose the use of a high-order AR model to capture both the short-term and long-term speech correlations. The missing speech samples are then predicted recursively by running the model with zero input.

This paper aims *i)* to study the properties of packet-loss concealment based on rigorous, model-based estimation of the missing sample sequence using the MMSE criterion and *ii)* to develop a simple and practical PLC system based on this rigorous approach. The AR signal model is employed. The model parameters must be estimated from a finite data sequence, which introduces a model estimation error. We study how this model estimation error affects the prediction of lost packet and an upper bound for the MSE bound is derived. We also consider the usage of perceptual weighting with the aim to gain better perceived quality. A rigorous argument is provided to show that the usage perceptual weighting results in a structure that is equivalent to straightforward optimal prediction.

Our work is performed in the context of two AR signal models: a basic model describing the short-term correlation and an extended model capturing also the long-term correlation. Fig. 1 shows signals reconstructed by applying optimal MMSE prediction based on these two models.

2. SIGNAL MODEL

We use autoregressive (AR) models for the speech signal. This model has been utilized extensively and successfully in speech coding and speech enhancement. In this section we first describe the basic signal model, which accounts for the short-term correlations. This model is then extended to include the so-called long-term correlations.

2.1. Basic Signal Model

This basic model captures the short-term correlations, which are largely determined by the vocal tract characteristics. Let us denote the sampled speech sequence as s_n, $n = \ldots, -1, 0, +1, \ldots$. The basic speech-signal model is then

$$s_n = \sum_{i=1}^{p} a_i s_{n-i} + w_n, \tag{1}$$

where the excitation w_n is white Gaussian noise with variance σ^2 and $\mathbf{a} = [\ a_1 \quad a_2 \quad \ldots \quad a_p\ ]^T$ is the model coefficient vector. The transfer function takes the form of $H_s(z) = \frac{1}{A(z)}$, where $A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$. (For 8 kHz sampling, the speech model order is typically set as $p = 10$.) Thus, the AR model is uniquely specified by $\{\mathbf{a};\ \sigma^2\}$. To facilitate our derivations, we use the corresponding state space model:

$$\begin{array}{l} \mathbf{x}_{n+1} = \mathbf{F}_{\mathbf{a}}\mathbf{x}_n + \mathbf{G}w_n \\ y_n = \mathbf{H}\mathbf{x}_n \end{array}, \tag{2}$$

where $\mathbf{x}_{n+1} = [s_n\ s_{n-1} \ldots\ s_{n-p+1}]^T$, and $\mathbf{G}^T = \mathbf{H} = [1\ 0\ \ldots\ 0]_{1\times p}$. The transition matrix $\mathbf{F}_{\mathbf{a}}$ is then given by

$$\mathbf{F}_{\mathbf{a}} = \begin{bmatrix} a_1 & a_2 & \ldots & a_{p-1} & a_p \\ 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}. \tag{3}$$

In the present context, the main advantage of the state space model is that it facilitates the definition of the Kalman filter, which in turn provides optimal linear prediction [7].

2.2. Extension to Long Term Correlations

It is well known that the speech signal exhibits both short-term and long-term correlations. The long-term correlations are associated with the oscillation of the vocal cords and, thus, with voiced speech. Accurate reconstruction of the long-term correlations is vital for perceptual quality.

We refine our speech generation model by incorporating the long-term correlations using a transfer function

$$H_{sl} = \frac{1}{A(z)}\frac{1}{1 - Gz^{-D}}, \tag{4}$$

where D denotes the pitch period in sample and G is a correlation factor. The extension of the state-space model to include (4) is straightforward and not reproduced here.

3. MMSE MISSING SEGMENT PREDICTION

In this section, we employ the MMSE criterion to address packet-loss concealment. The best linear prediction given the model is studied first. We then investigate how the estimation error of the model parameters affects the prediction. Finally, we study the application of perceptual weighting prediction for improving the speech quality.

3.1. MMSE Prediction

We now formulate the mean-square-error (MSE) optimal prediction for the basic model (2). We supposed that the received speech sequence is $S_{n-k}^{n-1} = \{s_i,\ n-k \le i \le n-1\}$, which is generated from the AR model $\{\mathbf{a};\ \sigma^2\}$ and that the subsequent signal samples are unknown.

We now consider the MSE-optimal prediction (or estimation) of s_{n+m} or y_{n+m+1}, $m \ge 0$, from the real model and history data, $\Theta = \{\mathbf{a};\ S_{n-k}^{n-1}\}$. Applying the state space model (2), $\mathbf{x}_{n+m+1}$ can be expressed as

$$\mathbf{x}_{n+m+1} = \mathbf{F}_{\mathbf{a}}^{m+1}\mathbf{x}_n + \sum_{i=0}^{m} \mathbf{F}_{\mathbf{a}}^{i}\mathbf{G}w_{n+m-i}. \tag{5}$$

It follows immediately from *Kalman* filter theory [7] that the optimal MSE estimator $\hat{\mathbf{x}}_{n+m+1}$ of $\mathbf{x}_{n+m+1}$ is

$$\hat{\mathbf{x}}_{n+m+1} = \mathbf{F}_{\mathbf{a}}^{m+1}\mathbf{x}_n. \tag{6}$$

Let us denote the prediction error by $\tilde{\mathbf{x}}_{n+m+1|n} = \mathbf{x}_{n+m+1} - \hat{\mathbf{x}}_{n+m+1}$. The covariance matrix of the estimation error, $\mathbf{P}_{n+m+1|n} = \mathbb{E}[\tilde{\mathbf{x}}_{n+m+1|n}\tilde{\mathbf{x}}_{n+m+1|n}^T]$, is then given by

$$\mathbf{P}_{n+m+1|n} = \sigma^2 \sum_{i=0}^{m} \mathbf{F}_{\mathbf{a}}^{i}\mathbf{G}\mathbf{G}^T\mathbf{F}_{\mathbf{a}}^{iT}. \tag{7}$$

Thus, the corresponding prediction and prediction error for s_{n+m} are

$$\hat{s}_{n+m}(\mathbf{a}) = \mathbf{F}_{\mathbf{a}}^{m+1}(1,:)\mathbf{x}_n, \tag{8}$$

$$\varepsilon_{n+m|n-1}(\mathbf{a}) = \sigma^2 \sum_{i=0}^{m} (\mathbf{F}_{\mathbf{a}}^{i}(1,1))^2, \tag{9}$$

where $\mathbf{F}_{\mathbf{a}}^{m+1}(1,:)$ is the first row of $\mathbf{F}_{\mathbf{a}}^{m+1}$. The analysis of the prediction of the missing segment is analogous for the advanced signal model (4).

The estimation process for the advanced model is identical for voiced and unvoiced speech. For unvoiced speech, the factor G is close to zero. As a result, the MMSE-based recovered signal decays sufficiently fast without inducing artificial annoying sounds. In contrast, for voiced speech the recovered signal decays relatively slowly. Comparing to [6] which applied a high order (e.g., 128) AR model for tracking the long term speech correlation, the model we exploit adapts to the estimated pitch and has fewer parameters to be estimated. Thus, the required signal length for parameter estimation is reduced.

Note that only the speech preceding the missing segment is utilized for reconstruction. This makes the proposed algorithms more flexible as the speech after the missing segment is not required compared with existing methods (e.g., [6]). To distinguish the algorithms for the two models, we denote the method based on the basic model as **SMSE** and the method based on the extended model as **LSMSE**.

3.2. Effect of Model Estimation Error on Prediction

In a practical system, the model parameters are not available but must be estimated from the signal. This motivates us to study the effect of the model estimation error on the prediction (estimation) of the missing segment.

It is known from [8] that estimating the model vector $\mathbf{a}$ from the received data S_{n-k}^{n-1} by applying *Yule-Walker* estimation results in a random vector $\mathbf{a}_d$ satisfying

$$\mathbf{a}_d \sim \mathcal{AN}(\mathbf{a}, \frac{\sigma^2}{k}\mathbf{R}_p^{-1}), \tag{10}$$

where $\mathbf{R}_p$ is the covariance matrix $\left[r_{(i-j)} = \mathbb{E}[s_i s_j]\right]_{i,j=1}^{p}$. Equation (10) shows that as $k \to \infty$, $\mathbf{a}_d$ is asymptotically unbiased and has a *normal* distribution with covariance matrix $\frac{\sigma^2}{k} R_p^{-1}$. To predict s_{n+m} from $\{\mathbf{a}_d;\ S_{n-k}^{n-1}\}$ we can use (8):

$$\hat{s}_{n+m}(\mathbf{a}_d) = \mathbf{F}_{\mathbf{a}_d}^{m+1}(1,:)\mathbf{x}_n. \tag{11}$$

In practice, we must use a *realization* of (11) and it is important to define a bound on the variance in the prediction resulting from the estimation process for $\mathbf{a}$:

Proposition 3.1 *For the prediction of s_{n+m} as specified by (11), the associated mean square error is asymptotically bounded by*

$$\varepsilon_{n+m|n-1}(\mathbf{a}_d) \leq \sigma^2 \sum_{i=0}^{m} (1 + \frac{p+2i}{k})(\mathbf{F}_{\mathbf{a}}^{i}(1,1))^2. \tag{12}$$

as the signal length k for model estimation increases. The bound is asymptotically tight everywhere for $k \to \infty$.

The proof is not trivial and will be provided elsewhere. The proof uses that the Taylor expansion [8] of $\mathbf{F}_{\mathbf{a}_d}^{m+1}(1,:)$ in (11) is asymptotically normal distributed with mean $\mathbf{F}_{\mathbf{a}}^{m+1}(1,:)$. The MSE for $\hat{s}_{n+m}(\mathbf{a}_d)$ is then analyzed. $\mathbf{x}_n$ and $\mathbf{a}_d$ are assumed uncorrelated in the derivation.

It is known that for a stable system as specified by (2), all the eigenvalues of $\mathbf{F}_{\mathbf{a}}$ are within the unit circle, i.e., $|\lambda_i(\mathbf{F}_{\mathbf{a}})| < 1$, for all i. Then as $m \to \infty$, the eigenvalue λ_i^m of $\mathbf{F}_{\mathbf{a}}^m$ would

$$\lambda_i^m \to 0 \text{ for all } i.$$

Denote $\lambda_{max} = \max_{i=1}^{p} |\lambda_i(\mathbf{F}_{\mathbf{a}})|$. It can be shown that $\mathbf{F}_{\mathbf{a}}^{m}(1,1))^2 = O(|\lambda_{max}|^{2m})$. Thus, as $m \to \infty$, the right side of (12) results in

$$\lim_{m\to\infty} \left[\sigma^2 \sum_{i=0}^{m} (1 + \frac{p+2i}{k})(\mathbf{F}_{\mathbf{a}}^{i}(1,1))^2\right] = C, \tag{13}$$

where C is a constant. The convergence speed is related with the prediction vector $\mathbf{a}$. As the correlation between samples is getting weak,i.e., $|\mathbf{a}|$ is getting small, it would converge fast. On the other hand, (9) takes the form of

$$\lim_{m\to\infty} \left[\sigma^2 \sum_{i=0}^{m} (\mathbf{F}_{\mathbf{a}}^{i}(1,1))^2\right] = r_0. \tag{14}$$

Observing (13) and (14), it is immediate that for finite signal length utilized for model estimation, i.e., $k < \infty$, $C > r_0$ and as $k \to \infty$, C lands on r_0.

3.3. Why perceptual weighting does not affect Prediction

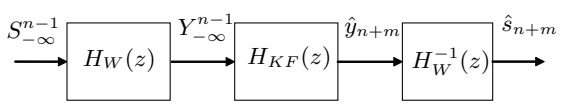

Fig. 2. the Kalman filter in tandem with perceptual weighted pre- and post- filters.

The weighted Kalman filter is successfully employed in [9] for speech enhancement, with the filter structure shown in Fig. 2. The weighting filter, which deemphasizes the formant structure of the speech segment, takes the form

$$H_W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \qquad 1 \geq \gamma_1 \geq \gamma_2. \tag{15}$$

When using a weighted Kalman filter, the mean squared error is redistributed towards high energy spectrum regions. This leads to better perceived quality. In this section, we investigate perceptual weighting in the context of packet-loss concealment. We show that no perceptual improvement can be achieved by perceptual weighting.

We denote the output of the filter $H_W(z)$ as $\{y_n\}$. It can be modeled with an ARMA model with transfer function $H(z) = \frac{A(z/\gamma_1)}{A(z)A(z/\gamma_2)}$. Ignoring the Kalman filter of Fig. 2 and the application of the inverse filter $H_W^{-1}(z)$ to $\{y_n\}$ gives exactly $\{s_n\}$. Thus, s_{n+m} can also be expressed as

$$s_{n+m} = \sum_{i=-\infty}^{n+m} c_{n+m-i} y_i. \tag{16}$$

Suppose the received speech sequence is $S_{-\infty}^{n-1}$ or equivalently $Y_{-\infty}^{n-1}$ (as $H_W(z)$ is linear and invertible). Then the optimal prediction for y_{n+m} is given as

$$\hat{y}_{n+m|n-1} = \mathbb{E}[y_{n+m}|Y_{-\infty}^{n-1}] = \mathbb{E}[y_{n+m}|S_{-\infty}^{n-1}]. \tag{17}$$

Consequently,

$$\hat{s}_{n+m} = \sum_{i=-\infty}^{n-1} c_{n+m-i} y_i + \sum_{i=n}^{n+m} c_{n+m-i} \hat{y}_{i|n-1}. \tag{18}$$

Note now that $\hat{s}_{n+m}$ in (18) is the optimal prediction

$$\hat{s}_{n+m} = \mathbb{E}[s_{n+m}|S_{-\infty}^{n-1}].$$

This shows that exploiting perceptual weighting cannot improve perceived speech quality. The key point for the successful application of weighting in speech enhancement is that the component y_i of s_{n+m}, $i < n+m$, is estimated from the noisy speech up to i due to the causality constraint. Thus, the estimation is not optimal in the enhancement case, as the noisy speech after i is not exploited in estimation.

4. RESULTS

The upper MSE bound of Proposition 3.1 is valid everywhere, and is asymptotically tight as the data length for the model estimation increases. We verify the proposition for data lengths that occur in practical communication systems (of the order of several hundred samples). In addition, we report the performance of the proposed **SMSE** and **LSMSE** algorithms as obtained in a subjective listening test.

4.1. Experimental MSE Bound Verification

The verification was performed on a synthetic speech sequence generated with 10'th order AR model with model parameters estimated from a real speech sequence sampled at 8 kHz. The variance of the stationary process is $r_0 \doteq 1142$. We examined the upper bounds for data lengths $k = 160$ and $k = 640$ (corresponding to 20 ms and 80 ms, respectively). Fig. 3 displays both the measured MSE curve and the theoretical upper bounds for the MSE for the missing segment for the indicated data lengths k. For $k = \infty$ the model parameters are estimated without error.

Fig. 3 shows that the bound of Proposition 3.1 is essentially tight when the MSE is less than the signal variance. That the error does not exceed the variance of the signal follows from the stability of the models estimated using the *Yule-Walker* estimation procedure. This region increases in duration for increasing data length k. The figure also shows that, for common data lengths, the prediction error is significantly larger than the assumption of perfect estimation of the prediction parameters $\mathbf{a}$ would indicate. When applied to other AR parameter sets, they exhibit the same trends.

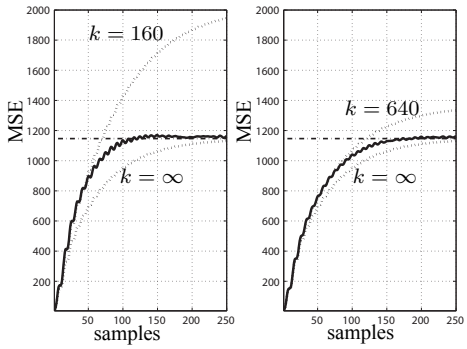

Fig. 3. Dotted curve denotes the theoretical MSE bound and solid curve denotes the simulated MSE for $k = 160$ and $k = 640$, respectively. The dash-dotted line denotes the variance r_0 for reference.

4.2. Subjective Quality Comparison

We tested the performance of our MMSE predictor using four sentences from female speakers and four from male speakers with 16kHz sampling frequency were selected from the TIMIT database [10]. The sentences were down-sampled to 8kHz and represented in 16-bit linear PCM form. The packet length was set to 10 msec. The analysis window length for extracting AR vectors $\mathbf{a}$ was 20 msec, for the two packet lengths before the missing one. A Hamming window was applied in the LPC analysis. When part of the real speech in the analysis window was not available due to packet loss, it was replaced by reconstructed speech. To smooth the transition from the reconstructed speech to the received speech, the prediction operation was continued with $L = 40$ into the received speech and overlap-adding operation was then performed with the linear weighting strategy. The pitch period in **LSMSE** method was estimated by applying a pitch tracking algorithm similar to that of [11]. The factor G in (4) is the associated normalized correlation value. Packets were discarded randomly and independently.

The DCR subjective quality test [12] was conducted to evaluate our algorithms. The compared PLC schemes were the ITU-T G.711 Appendix I algorithm (denoted as **g711a1**) and the forward prediction algorithm presented in [6] (denoted as **HighOrder**). The listeners were instructed to rate each processed sample compared to the original one to a 5-graded quality score: 5 - *degradation is inaudible*, 4 - *degradation is audible but not annoying*, 3 - *degradation is slightly annoying*, 2 - *degradation is annoying*, and 1 - *degradation is very annoying*. The Mean Opinion Score (MOS) is displayed in Fig. 4. Eleven listeners participated in the test. The test results indicate that our **LSMSE** algorithm gives comparable result to **g711a1** and outperforms the **HighOrder** algorithm. The test confirms that considering only the short-term correlations is insufficient, rendering the worst performance.

5. CONCLUSION

We conclude that formal mean-square-error (MSE) based estimation methods can provide state-of-the-art packet-loss concealment performance. The resulting **LSMSE** algorithm has low complexity and can be integrated in existing codecs easily. We provided an upper bound for the MSE that was shown to be tight when the MSE is larger than the signal variance even for practical data lengths (e.g., k=160). We conclude from our results that the estimation error for the predictor parameters contributes significantly to the estimation error for the missing packet, thus emphasizing the importance of relatively long estimation windows. We showed that perceptual weighting can not improve speech quality in the context of MSE based packet loss concealment.

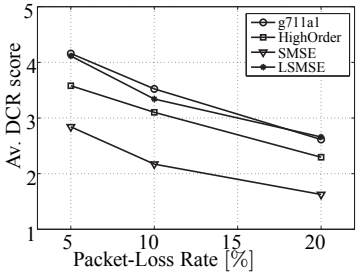

Fig. 4. Subjective PLC performance using the DCR test.

6. REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Thoery*. Wiley-Interscience, 2006.

[2] E. Gündüzhan and K. Momtahan, "A linear prediction based packet loss concealment algorithm for PCM coded speech," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 778–785, November 2001.

[3] J.-F. Wang, J.-C. Wang, J.-F. Yang, and J.-J. Wang, "A voicing-driven packet loss recovery algorithm for analysis-by-synthesis predictive speech coders over internet," *IEEE Trans. Multimedia*, vol. 3, pp. 98–107, March 2001.

[4] ITU-T Recommendation G.711 Appendix I, "A high quality low complexity algorithm for packet loss concealment with G.711," 1999.

[5] C. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, "Hidden Markov model-based packet loss concealment for voice over ip," *IEEE Trans. Audio Speech Language Process.*, vol. 14, pp. 1096–1623, September 2006.

[6] K. Kondo and K. Nakagawa, "A speech packet loss concealment method using linear prediction," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 2, pp. 806–813, February 2006.

[7] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*. New Jersey: Prentice Hall, 2000.

[8] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. Springer-Verlag, 1991.

[9] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 14, pp. 764–773, May 2006.

[10] DARPA-TIMIT, "Acoustic-phonetic continuous speech corpus," 1990.

[11] ITU-T Recommendation G.729, "Coding of speech at 8kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)," March 1996.

[12] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," 1996.