PACKET LOSS CONCEALMENT WITH NATURAL VARIATIONS USING HMM

Manohar N. Murthi[†]*, Christoffer A. Rødbro*[‡]*, Søren Vang Andersen*[‡]*, and Søren Holdt Jensen*[‡]

[†] Dept. of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124-0640 USA
mmurthi@miami.edu

[‡] Department of Communication Technology
Aalborg University
9220 Aalborg Ø, Denmark
{car,sva,shj}@kom.auc.dk

ABSTRACT

Packet loss concealment (PLC) at a receiver has a substantial effect on the speech quality in Voice over IP. Most conventional PLC systems have largely relied upon variations of signal repetition and overlap-add interpolation which can produce speech signals that do not follow the larger overall statistical trends. In this paper, we demonstrate how Hidden Markov Models can be utilized to effect PLC based on statistical signal processing. In particular, we show how HMM-based PLC yields conditional density functions that can be utilized by various statistical estimation methods that produce signal parameter estimates that produce more natural variation than conventional PLC methods, thereby providing much better speech quality.

1. INTRODUCTION

As Voice over IP proliferates, packet loss concealment (PLC) at the receiver has emerged as an important factor in determining voice Quality of Service. Receiver-based PLC techniques attempt to partially recover the speech signal content of a lost packet from its neighbors.

Although differing in terms of speech analysis/synthesis model type and some implementation details, conventional PLC techniques (e.g., [1], [2], [3] to name a few) are largely based on variations of signal and parameter repetition or interpolation. For example, conventional methods are based on various combinations of pitch synchronous extrapolation or overlap-add interpolation of pitch cycles with noise mixing; repetition or overlap-add interpolation of spectral envelope parameters with possible bandwidth expansion of poles for persistent loss; energy contour muting, and pitch lag jittering. Consequently, these conventional PLC methods can be seen as "freezing" signal statistics during variations on repetition, and "blurring" statistics during variations of interpolation. By missing the larger overall statistical trends of speech parameters such as spectral envelope, pitch, voicing, and energy amplitude evolution, a conventional PLC produces synthetic speech that can contain perceptually annoying artifacts. Therefore, PLC methods that transcend the variations of repetition and interpolation and provide more natural sounding speech are welcome.

In this paper, we present an approach to PLC based on a statistical signal processing framework offered by Hidden Markov Models [4]. Although extensively used in speech recognition and enhancement, HMMs have not been previously used in PLC. In applying HMMs to PLC, a sequence of speech parameter vectors is viewed as being produced by an HMM with continous probability density functions. Consequently, in PLC, a decoder tracks the evolution of the speech signal through the HMM. When a packet is lost, the HMM-based PLC yields conditional probability density functions that are useful for different approaches for statistically estimating the missing signal parameters. In fact, we show that HMM-based PLC provides better estimates and more natural variation than conventional PLC. Many of the results in this paper along with additional details and numerous examples can be found in our forthcoming journal paper [5].

The rest of this paper is organized as follows. In Section 2 we discuss the approach to adapting HMMs to the task of PLC. In [5], we primarily confined our discussion to MMSE-based estimation. However, in Section 3, we provide evidence that other estimation methods are more suitable for certain signal parameters, showing that HMM-based PLC produces parameter estimates that provide more natural variation than the estimates produced by conventional repetition. In Section 4 we provide some guidance on how the HMM-based PLC methods can be utilized with different codec structures.

2. METHODS

The HMM-based PLC methods work within the scenario illustrated in Figure 1. Each frame of a speech signal is coded resulting at time t in a set of perceptually relevant parameters such as spectral envelope, pitch, energy, and degree of voicing which we group together and represent by the vector ϕ_t. The parameters ϕ_t are subsequently transmitted in the form of a packet over a packet loss channel, resulting in either a

The work of M.N. Murthi was supported in part by the National Science Foundation via CAREER Award CCF-0347229.

perfect reproduction $\boldsymbol{\phi}_t$ without loss, or in $\bar{\boldsymbol{\phi}}_t$ in the case of packet loss. Therefore, the PLC system must produce an estimate $\hat{\boldsymbol{\phi}}_t$ of the missing parameters before signal synthesis at the decoder.

As already mentioned, the estimation of missing vectors $\boldsymbol{\phi}_t$ is based on a HMM. The HMM is trained by the Baum-Welch algorithm [4] with training vectors $\boldsymbol{\phi}_t$ extracted from the TIMIT database. No losses were introduced in the training set; the HMM should simply be a model describing the evolution in the $\boldsymbol{\phi}_t$ vectors. The HMM has a total of $N = 331$ states (split between "voiced", "unvoiced" and "silent" frames) and a single Gaussian emission pdf in each state. Thus, we shall denote the emission pdf in state n by $p(\boldsymbol{\phi}_t|s_t = n) = \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, where s_t is the state occupation at time t. Likewise, $a_{nm} = P(s_t = n|s_{t-1} = m)$ denote the state transition probabilities.

When a packet containing $\boldsymbol{\phi}_t$ is lost at time t, the PLC estimate $\hat{\boldsymbol{\phi}}_t$ is based on all correctly received past and future vectors $\boldsymbol{\phi}_1^T$, the subscript 1 and superscript T denoting "from time 1 to T". T is determined by the algorithmic look-ahead, e.g. $T = t + 2$ if a look-ahead of two frames is allowed. If $T = t - 1$, then the missing parameter at time t is estimated from the past parameters from times 1 to $t - 1$. Estimation of the missing parameter vector $\boldsymbol{\phi}_t$ is divided into three steps:

1. Given all correctly received parameter vectors $\boldsymbol{\phi}_1^T$ determine $P(s_t = n|\boldsymbol{\phi}_1^T)$, that is, the probability of being in each HMM state at time t.
2. Using these state probabilities, obtain $p(\boldsymbol{\phi}_t|\boldsymbol{\phi}_1^T)$, that is, a pdf for the missing parameter vector conditioned on the correctly received vectors.
3. Use this conditional pdf to obtain $\hat{\boldsymbol{\phi}}_t$, an estimate of the missing parameter vector.

The three steps will be examined in more detail in the following.

2.1. State probability identification

One of the standard problems associated with HMMs is that of *decoding*, that is, given an observation sequence $\boldsymbol{\phi}_1^T$ find the most probable hidden state sequence. The problem at hand is somewhat different in that, given the observation sequence $\boldsymbol{\phi}_1^T$, we want to determine the state probabilities at different time instances t, $P(s_t = n|\boldsymbol{\phi}_1^T)$. For now, we shall assume that all $\boldsymbol{\phi}_1^T$ are available and return to the problem of handling losses shortly. The first step in finding the state probabilities is to split the conditional probability into "forward" and "backward" parts,

$$P(s_t = n|\boldsymbol{\phi}_1^T) = \frac{p(s_t = n, \boldsymbol{\phi}_1^T)}{p(\boldsymbol{\phi}_1^T)} \quad (1)$$

$$= cp(s_t = n, \boldsymbol{\phi}_1^t)p(\boldsymbol{\phi}_{t+1}^T|s_t = n, \boldsymbol{\phi}_1^t). \quad (2)$$

Here $c = p(\boldsymbol{\phi}_1^T)^{-1}$ normalizes the state probabilities to sum to unity and can thus be found as,

$$c = \left(\sum_n p(s_t = n, \boldsymbol{\phi}_1^t)p(\boldsymbol{\phi}_{t+1}^T|s_t = n, \boldsymbol{\phi}_1^t) \right)^{-1} \quad (3)$$

For the other two terms in (2), we define $\alpha_t(n) = p(s_t = n, \boldsymbol{\phi}_1^t)$, and (using the first order Markov assumption), $\beta_{t+1}^T(n) = p(\boldsymbol{\phi}_{t+1}^T|s_t = n, \boldsymbol{\phi}_1^t) = p(\boldsymbol{\phi}_{t+1}^T|s_t = n)$. These can be found by the forward and backward recursions, see [4]:

$$\alpha_t(n) = \left(\sum_{m=1}^N \alpha_{t-1}(m)a_{nm} \right) p(\boldsymbol{\phi}_t|s_t = n) \quad (4)$$

$$\beta_{t+1}^T(n) = \left(\sum_{m=1}^N \beta_{t+2}^T(m)a_{nm}p(\boldsymbol{\phi}_{t+1}|s_{t+1} = m) \right) \quad (5)$$

Now, we readily obtain the state probabilities for any t by inserting (4) and (5) in (2) after some manipulations [5].

So, what if one of the observations $\boldsymbol{\phi}_{t'}$ within $\boldsymbol{\phi}_1^T$ is lost, meaning that $p(\boldsymbol{\phi}_{t'}|s_{t'} = n)$ in (4) cannot be evaluated? In [5], we demonstrate how to handle this issue, showing that a simple scaling suffices.

2.2. Forming the pdf

Once the state probabilities are found, formation of the missing parameter vector conditional pdf is straight-forward:

$$p(\boldsymbol{\phi}_t|\boldsymbol{\phi}_1^T) = \sum_{n=1}^N P(s_t = n|\boldsymbol{\phi}_1^T)p(\boldsymbol{\phi}_t|s_t = n, \boldsymbol{\phi}_1^T) \quad (6)$$

$$= \sum_{n=1}^N P(s_t = n|\boldsymbol{\phi}_1^T)p(\boldsymbol{\phi}_t|s_t = n), \quad (7)$$

where we used the Markov assumption,

$$p(\boldsymbol{\phi}_t|s_t = n, \boldsymbol{\phi}_1^T) = p(\boldsymbol{\phi}_t|s_t = n) \quad (8)$$

i.e. given the state, the emission pdf is independent of the emissions at all other time instances. As described above, in the setup at hand $p(\boldsymbol{\phi}_t|s_t = n) = \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, so that (7) forms a Gaussian mixture model (GMM) with mixture weights $w_n = P(s_t = n|\boldsymbol{\phi}_1^T)$.

In practice with a reasonable number of states, the HMM is not capable of sufficiently decorrelating the state emissions at subsequent time indexes; that is, (8) is not fulfilled. To alleviate this problem, we assign an auxiliary Gaussian pdf to each HMM state specifically modeling the interframe parameter dependence, see [5] for details. The resulting conditional pdf is still a GMM with weights $w_n = P(s_t = n|\boldsymbol{\phi}_1^T)$, however, the means and covariances are modified:

$$\tilde{p}(\boldsymbol{\phi}_t|\boldsymbol{\phi}_1^T) = \sum_{n=1}^N P(s_t = n|\boldsymbol{\phi}_1^T)\mathcal{N}(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\Sigma}}_n) \quad (9)$$

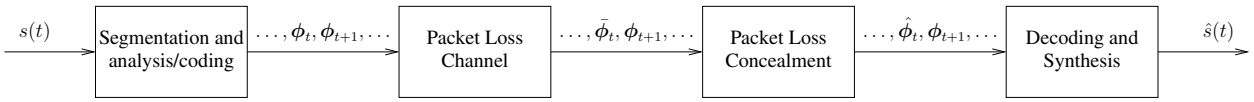

Fig. 1. Block diagram for the packet loss scenario with packet at time t being lost.

2.3. Estimation of ϕ_t from the GMM

Given the GMM in (9) various strategies can be applied for estimating $\boldsymbol{\phi}_t$. One possibility is that of minimum mean squared error estimation (MMSE), which was also the choice of [5]. It is straight forward to show that the MMSE estimator is:

$$\hat{\boldsymbol{\phi}}_t^{(N_2)} = \sum_{n=1}^{N} P(s_t = n|\boldsymbol{\phi}_1^T)\tilde{\boldsymbol{\mu}}_n, \tag{10}$$

where N_2 denotes "Norm 2", the error norm of which the expectation is minimized.

However, as will be illustrated in Section 3, MMSE is not necessarily the best choice for the estimation of some of the parameters (specifically the pitch frequency). An alternative is to minimize the expected error 1-norm, to which end we define the cost-function:

$$J(\hat{\boldsymbol{\phi}}_t^{(N_1)}) = E_{\boldsymbol{\phi}_t}[\|\boldsymbol{\phi}_t - \hat{\boldsymbol{\phi}}_t^{(N_1)}\|_1]. \tag{11}$$

To keep notation reasonably simple, and since the N_1 estimator will be applied to a single parameter anyway, at this point we shall replace the full parameter vector $\boldsymbol{\phi}_t$ with its m'th element, $\phi_{t,m}$. Using the definition of expectation, (11) then becomes:

$$J(\hat{\phi}_{t,m}^{(N_1)}) = E_{\phi_{t,m}}[|\phi_{t,m} - \hat{\phi}_{t,m}^{(N_1)}|] \tag{12}$$

$$= \int_{-\infty}^{\infty} |\phi_{t,m} - \hat{\phi}_{t,m}^{(N_1)}| p(\phi_{t,m}|\boldsymbol{\phi}_1^T) d\phi_{t,m}. \tag{13}$$

Here, $p(\phi_{t,m}|\boldsymbol{\phi}_1^T) = \sum_{n=1}^{N} P(s_t = n|\boldsymbol{\phi}_1^T)\mathcal{N}(\tilde{\mu}_{n,m}, \tilde{\sigma}_{n,m}^2)$ is the pdf of (9) marginalized w.r.t $\phi_{t,m}$. We have not been able to find a closed form minimizer for this cost-function. However, note that J is a convex function (a sum of convex functions is convex, generalizing to infinite sums), meaning that proper iterative optimization procedures are guaranteed to converge to the global minimizer. The 1st and 2nd order derivatives needed for Newton-type iterations are,

$$J'(\hat{\phi}_{t,m}^{(N_1)}) = \sum_{n=1}^{N} P(s_t = n|\boldsymbol{\phi}_1^T)\,\mathrm{erf}\left(\frac{\hat{\phi}_{t,m}^{(N_1)} - \tilde{\mu}_{n,m}}{\sqrt{2}\tilde{\sigma}_{n,m}}\right) \tag{14}$$

$$J''(\hat{\phi}_{t,m}^{(N_1)}) = 2\sum_{n=1}^{N} P(s_t = n|\boldsymbol{\phi}_1^T)\mathcal{N}(\tilde{\mu}_{n,m}, \tilde{\sigma}_{m,n}^2), \tag{15}$$

found through straight-forward but tedious calculus. When the MMSE estimate was used as the initial guess, simulations showed convergence after 6 iterations.

A final estimation concept to be used in this paper is that of maximum likelihood (ML). In general, the GMM of (9) may have multiple local maxima, rendering the ML-principle infeasible. However, for the single parameter problem, we can numerically search for the maximum of the marginalized conditional pdf, $p(\phi_{t,m}|\boldsymbol{\phi}_1^T)$. Since $p(\phi_{t,m}|\boldsymbol{\phi}_1^T)$ may have multiple local maxima, the global maximum is simply found through a sweep over $\phi_{t,m}$.

3. EXPERIMENTAL RESULTS

In [5], we show that the HMM-based MMSE approach as described above produces more accurate parameter substitutes as compared to conventional repetition/interpolation based methods. For example, on the average the spectral distortion (SD) between the true and estimated LSFs was improved by more than 0.5 dB in lost frames for a 20 % packet loss rate. Also, perceptual improvements were demonstrated through listening tests. Numerous examples and results are presented in [5], and we choose not to present them here. At this point, instead of estimation accuracy, we shall focus on the parameter *variation* introduced through lost sequences. We demonstrate that the HMM-based PLC produces more natural variation than conventional repetition methods.

For simulation, we used the sinusoidal coder in [5] simulated over a Gilbert packet loss model configured for an overall packet loss probability of 0.2 with the loss probability being two times larger after a loss than after a non-lost packet. This results in somewhat bursty losses.

The evaluation of the parameter variation will be based on histograms of the frame-to-frame change in parameters. For the pitch frequency ω_0 we use the ratio between adjacent frames, i.e. $\Delta\omega_{0,t} = \frac{\omega_{0,t}}{\omega_{0,t-1}}$ and a histogram bin size of 0.05. The pitch ratio is chosen because it is perceptually more relevant than the difference $\omega_{0,t} - \omega_{0,t-1}$, cf. the common practice of quantizing pitch in the log-domain. For the voicing cut-off frequency ω_c we use $\Delta\omega_{c,t} = \omega_{c,t} - \omega_{c,t-1}$ and a bin size of 0.2. This is in line with 4 bit linear quantization ($0.2 \approx \frac{\pi}{2^4}$). Gain changes ΔG_t are measured in dB with a bin size of 3 dB. To evaluate the *similarity* between the correct normalized histogram $\mathcal{H}_{true}$ and the normalized histograms resulting from PLC estimates $\mathcal{H}_{plc}$ we use a Jaccard-like similarity measure, corresponding to the ratio between

the "intersection" and the "union" of the two histograms:

$$S(\mathcal{H}_{true}, \mathcal{H}_{plc}) = \frac{\sum_{j=1}^{B} \min(\mathcal{H}_{plc}(j), \mathcal{H}_{true}(j))}{\sum_{i=1}^{B} \max(\mathcal{H}_{true}(i), \mathcal{H}_{plc}(i))}, \quad (16)$$

where B denotes the number of bins determined by the bin size and parameter range. If the histograms are identical, then the Similarity attains its maximal value of $S(\mathcal{H}_{true}, \mathcal{H}_{plc}) = 1$. If the histograms are very dissimilar, then $S(\mathcal{H}_{true}, \mathcal{H}_{plc})$ tends towards 0.

As mentioned in Section 2.3, MMSE estimation is not always the best choice for estimating the pitch in lost frames. The reason for this can be deduced from Figure 2. Here, in the upper plot we see how the true frame-to-frame pitch variations contain pitch double-, triple-, and halving errors (stemming from the pitch detector), the same phenomena being present in the HMM training data. The lower plot shows the impact on estimation: the estimation procedure very rarely predicts pitch multiples, instead the slight possibility of such results in a widening of the histogram "main lobe". For example, consider a signal region where the pitch is stable at 200 Hz; then, the PLC should also produce a pitch close to 200 Hz. However, due to the chance of a pitch doubling error in the missing frame, the MMSE estimator may produce an estimate of say 220 Hz. Therefore, a ML approach disregarding the chance of pitch multiples should result in less (and more natural) pitch variation. 1-norm minimization is expected to fall in between MMSE and ML. In Table 1 the histogram similarity measure as defined in (16) is shown when using the MMSE, N_1, and ML-estimators. Also, the case of parameter repetition is included for reference. We see that for all three parameters, the HMM-based PLC estimators produce a histogram much closer to that of true speech than does parameter repetition for the variation of the pitch, voicing, and gain parameters. Also as expected, for the pitch, ML outperforms the N_1 approach, in turn being better than MMSE. On the other hand, the three HMM-based estimators produce nearly identical results (difference on 3rd or 4th decimal only) when applied to the gain and voicing. The reason is that the gain and voicing does not exhibit very abrupt change phenomena as does the pitch, so that the marginal conditional pdf $p(\phi_{t,m}|\boldsymbol{\phi}_1^T)$ for these parameters has no "side-lobes".

4. CONCLUDING REMARKS

Our results demonstrate that HMM-based PLC estimates produce more natural variation than conventional repetition-based methods. This along with our positive results in [5] clearly demonstrates the efficacy of HMM-based PLC. A natural question is how these methods can be utilized with conventional existing codecs such as CELP-based methods. The HMM-based PLC can clearly be used to provide lsf, pitch, voicing, and gain information that can be used for guiding an existing CELP-based PLC. In particular, the parameter estimates created by the HMM-based PLC can be used to orient the time-domain pitch synchronous operations that are used to synthesize a substitute signal that is phase-matched to the true known speech signal. Thus, the HMM-based PLC can be used to complement the existing codec PLC systems.

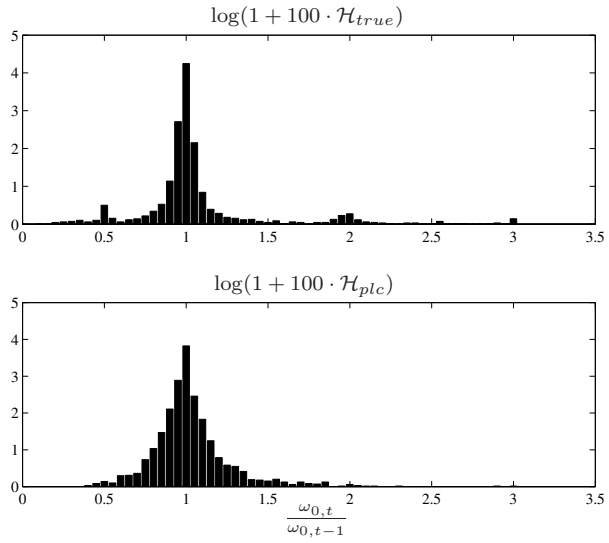

Fig. 2. The correct and PLC histograms for $\Delta\omega_{0,t}$, modified to enhance illustration. The PLC histogram is for the MMSE based estimation as described in Section 2.3.

	Rep.	MMSE	N_1	ML
$\Delta\omega_{0,t}$	0.52	0.58	0.59	0.61
$\Delta\omega_{c,t}$	0.25	0.66	0.66	0.66
ΔG_t	0.22	0.63	0.63	0.63

Table 1. Similarity measure (16) between histograms for the correct and the PLC produced frame-to-frame variations. The HMM-based estimators always provide more natural variation than repetition.

5. REFERENCES

[1] J. Lindblom, *Coding Speech for Packet Networks*, Ph.D. thesis, Chalmers University of Technology, November 2003, App. G.

[2] J. Wang and J. D. Gibson, "Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, vol. 2, pp. 745–748.

[3] J. C. De Martin, T. Unno, and V. Viswanathan, "Improved frame erasure concealment for CELP-based coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 1483–1486.

[4] X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing*, chapter 8, Prentice Hall PTR, 2001.

[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. Speech Audio Processing*, Sep. 2006.