Efficient HMM-Based Estimation of Missing Features, with Applications to Packet Loss Concealment

Bengt J. Borgström, Per H. Borgström, and Abeer Alwan

Department of Electrical Engineering, University of California, Los Angeles
jonas@ee.ucla.edu, henrik@ee.ucla.edu, alwan@ee.ucla.edu

Abstract

In this paper, we present efficient HMM-based techniques for estimating missing features. By assuming speech features to be observations of hidden Markov processes, we derive a minimum mean-square error (MMSE) solution. We increase the computational efficiency of HMM-based methods by downsampling underlying Markov models, and by enforcing symmetry in transitional probability matrices. When applied to features generally utilized in parametric speech coding, namely line spectral frequencies (LSFs), the proposed methods provide significant improvement over the baseline repetition scheme, in terms of Itakura-Saito distortion and peak SNR.

Index Terms: Missing Features, Markov Process, Packet Loss Concealment.

1. Introduction

In digital speech communication systems, transmitted data may become lost or corrupted due to channel degradation. Specifically, transmission of speech over wireless communication systems relies on error detecting codes to determine the reliability of received frames [1]. Packet-based systems, on the other hand, are subject to arrival jitter, which may induce unacceptable delay for real-time speech applications [2]. In either scenario, packet loss concealment (PLC) is applied at the receiver to reconstruct speech frames.

To reduce the impact of lost frames, some studies have applied waveform substitution or extrapolation techniques in the time domain [4]. Other studies have performed PLC for parametric coders in the parameter domain ([2],[3]). In this paper, we present novel efficient HMM-based techniques for estimating missing speech features, with applications to packet loss concealment. We assume speech parameters to be observations of hidden Markov processes, and derive generalized MMSE estimates of missing features. In this way, we aim to capture the natural progression of speech features in time by exploiting *a priori* steady-state and transitional statistics. Furthermore, we offer methods by which to increase the efficiency of the proposed framework. Specifically, we utilize downsampling of underlying Markov models, similar to [7], and the novel approach of enforcing symmetry in transitional probability matrices.

This paper is organized as follows: In Section 2, we provide derivations for MMSE HMM-based estimation. In Section 3, we present methods for increasing computational efficiency of the proposed estimation. Experimental results are provided in Section 4, followed by conclusions in Section 5.

This work as supported in part by the NSF.

2. HMM-Based Estimation

2.1. Interpreting Speech Parameters as Markov Processes

Parametric coders transmit feature vectors comprised of speech parameters which are used to synthesize speech at the receiver. Transmitted packets typically include spectral shape (such as line spectral frequencies), gain, and pitch information [6]. In this section we derive estimation techniques for a general missing feature, x_n, where n denotes time index.

Packet loss concealment (PLC) for parameter-based coders in the feature domain can be generalized as estimating x_{n+k}, conditioned on reliable features x_n and x_{n+N}, and given that $[x_{n+1}, \ldots, x_{n+N-1}]$ are missing. For some PLC applications, future features (x_{n+N}) may not be available, since this may induce unacceptable delays [4]. In this section, however, we derive missing feature estimation techniques in the general case, where both past and future samples are available. The general case solution can then be reduced to the specific solution based solely on past observations.

We interpret speech parameters as observations of Markov processes, as in [2]. In such a framework, parameter x_n is the observation of a fully connected hidden Markov model (HMM) with K states. We assume that state s_i has a continuous observation probability distribution function (pdf) with mean μ_i.

Given the Markov property, we can express the transitional probability between states at time index m as:

$$a_{ij} = P\left(s_m = i | s_{m-1} = j\right). \tag{1}$$

We define the matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$ such that $\mathbf{A}\left(i,j\right) = a_{ij}$. Furthermore, we define the matrix $\mathbf{B} \in \mathbb{R}^{K \times K}$ such that:

$$\mathbf{B}\left(i,j\right) = b_{ij}, \text{ where: } b_{ij} = P\left(s_m = i | s_{m+1} = j\right). \tag{2}$$

We assume parameter values to be outputs of a Markov process. For codec parameters with a continuous range of values, we specify HMMs with continuous observation pdfs. For such features, it may be most suitable to apply the minimum mean-square error (MMSE) estimate, defined as [5]:

$$\hat{x}_{n+k} = \sum_{i=1}^{K} \mu_i P\left(s_{n+k} = i | x_n, x_{n+N}\right). \tag{3}$$

2.2. Deriving State-Specific Probabilities

In this section we provide derivations for state-specific conditional probabilities $P\left(s_{n+k} = i | x_n, x_{n+N}\right)$, which are required by the MMSE solution. Using a Bayesian approach, the conditional probabilities in Eq. 3 can be expressed as:

$$P\left(s_{n+k}=i|x_n,x_{n+N}\right)=\frac{P\left(s_{n+k}=i,x_n,x_{n+N}\right)}{P\left(x_n,x_{n+N}\right)} \quad (4)$$
$$=\frac{P\left(s_{n+k}=i,x_n\right)P\left(x_{n+N}|s_{n+k}=i,x_n\right)}{P\left(x_n,x_{n+N}\right)}$$

Using the Markov property previously assumed, the left-hand probability in the numerator of Eq. 4 can be approximated as:

$$P\left(s_{n+k}=i,x_n\right) \quad (5)$$
$$=\begin{cases}\sum_{j=1}^{K}P\left(s_{n+k}=i|s_{n+k-1}=j\right)P\left(s_{n+k-1}=j,x_n\right), & \text{for } k>0,\\ P\left(s_n=i|x_n\right)P\left(x_n\right), \text{ for } k=0\end{cases}$$

To infer the hidden state at time n we minimize the distortion between the observed reliable feature and the underlying model:

$$P\left(s_n=i|x_n\right)=\delta_{i,q_n}, \text{ where } q_n=\underset{j}{\operatorname{argmin}}\left\|x_n-\mu_j\right\|^2. \quad (6)$$

Thus, Eq. 5 can be simplified as:

$$P\left(s_{n+k}=i,x_n\right)=P\left(x_n\right)\left[\mathbf{A}^{(k)}\mathbf{e}_{q_n}\right]_{(i)}, \quad (7)$$

where the vector $\mathbf{e}_j\in\mathbb{R}^K$ is comprised of zeros, except for a one at the j^{th} element. As in the previous equation, we will use the notation $[\mathbf{m}]_{(j)}$ to refer to the j^{th} element of vector $\mathbf{m}$. The right-hand probability in the numerator of Eq. 4 can be approximated as:

$$P\left(x_{n+N}|s_{n+k}=i,x_n\right)\approx P\left(x_{n+N}|s_{n+k}=i\right) \quad (8)$$
$$=\frac{P\left(s_{n+k}=i,x_{n+N}\right)}{P\left(s_{n+k}=i\right)}$$

Note that the denominator of Eq. 8 is the steady-state probability of state i, denoted by $\boldsymbol{\pi}\left(i\right)$. Steady-state statistics can be determined as the elements of the eigenvector of $\mathbf{A}$ or $\mathbf{B}$ corresponding to the unit eigenvalue:

$$\mathbf{A}\boldsymbol{\pi}=\boldsymbol{\pi}, \text{ and } \mathbf{B}\boldsymbol{\pi}=\boldsymbol{\pi}. \quad (9)$$

Furthermore, the numerator of Eq. 8 can be simplified by assuming the Markov property:

$$P\left(s_{n+k}=i,x_{n+N}\right)= \quad (10)$$
$$\begin{cases}\sum_{j=1}^{K}P\left(s_{n+k}=i|s_{n+k+1}=j\right)P\left(s_{n+k+1}=j,x_{n+N}\right), & \text{for } k<N\\ \delta_{i,q_{n+N}}P\left(x_{n+N}\right), \text{ for } k=N\end{cases}$$

Thus Eq. 8 can simplified as:

$$P\left(x_{n+N}|s_{n+k}=i\right)=\frac{P\left(x_{n+N}\right)}{\boldsymbol{\pi}\left(i\right)}\left[\mathbf{B}^{(N-k)}\mathbf{e}_{q_{n+N}}\right]_{(i)}. \quad (11)$$

By substituting Eqs. 7 and 11 into Eq. 4, the underlying state for time index $n+k$ is inferred via:

$$P\left(s_{n+k}=i|x_n,x_{n+N}\right) \quad (12)$$
$$=\frac{P\left(x_n\right)P\left(x_{n+N}\right)}{P\left(x_n,x_{n+N}\right)}\frac{1}{\boldsymbol{\pi}\left(i\right)}\left[\mathbf{A}^{(k)}\mathbf{e}_{q_n}\right]_{(i)}\left[\mathbf{B}^{(N-k)}\mathbf{e}_{q_{n+N}}\right]_{(i)}.$$

The first term, which we refer to as κ, is independent of state i, and is simply used to normalize the probability distribution:

$$\kappa=\left(\sum_{i=1}^{K}\frac{1}{\boldsymbol{\pi}\left(i\right)}\left[\mathbf{A}^{(k)}\mathbf{e}_{q_n}\right]_{(i)}\left[\mathbf{B}^{(N-k)}\mathbf{e}_{q_{n+N}}\right]_{(i)}\right)^{-1} \quad (13)$$

The value κ need not be explicitly determined during estimation of x_{n+k}. Instead, κ cancels from the solution by assuring that:

$$\sum_{i=1}^{K}P\left(s_{n+k}=i|x_n,x_{n+N}\right)=1. \quad (14)$$

We can infer the underlying state of a missing feature for the case where future reliable features are not available by marginalizing the observation x_{n+N}:

$$P\left(s_{n+k}=i|x_n\right)=\int_{x_{n+N}}P\left(s_{n+k}=i|x_n,x_{n+N}\right)\partial x_{n+N} \quad (15)$$
$$=\hat{\kappa}\left[\mathbf{A}^{(k)}\mathbf{e}_{q_n}\right]_{(i)},$$

$$\text{where: } \hat{\kappa}=\left(\sum_{i=1}^{K}\left[\mathbf{A}^{(k)}\mathbf{e}_{q_n}\right]_{(i)}\right)^{-1} \quad (16)$$

Depending on the availability of future observations, $\hat{x}_{n+k}$ is determined by substituting either Eq. 12 or 15 into Eq. 3.

Note that [2] offers a similar derivation for estimation of missing features. However, several important differences exist between the two approaches. In [2], determining state-specific probabilities of the reliable "boundary" feature requires HMM-based decoding of a series of preceding features. In the proposed method, such probabilities are determined simply by Eq. 6. Furthermore, in the proposed work, Eq. 12 includes the steady-state probability in the denominator, which is missing from the corresponding equation in [2]. Finally, in this paper, we provide the solution for the scenario in which future observations are not available via marginalization of a generalized solution conditioned on past and future observations.

3. Reducing the Complexity of HMM-Based Estimation

3.1. Markov Model Downsampling

From Eq. 12, it can be observed that the size of underlying signal models, K, has a large effect on the resulting complexity of the proposed algorithm. As we previously explored in [7], we propose to downsample the configuration of underlying Markov models to increase the computational efficiency of HMM-based estimation. Underlying models can either be trained at multiple resolutions, or lower resolution statistics can be extracted from a higher order model. Details are provided in [7].

3.2. Enforcing Transition Matrix Symmetry

From Section 2, the HMM-based estimation in the general missing feature framework is given by:

$$P\left(s_{n+k}=i|x_n,x_{n+N}\right)=\frac{\kappa}{\boldsymbol{\pi}\left(i\right)}\left[\mathbf{A}^{(k)}\mathbf{e}_{q_n}\right]_{(i)}\left[\mathbf{B}^{(N-k)}\mathbf{e}_{q_{n+N}}\right]_{(i)}. \quad (17)$$

Note that Eq. 17 requires numerous self-multiplications of transitional matrices $\mathbf{A}$ and $\mathbf{B}$. For large underlying Markov models (i.e. large K), this may prove computationally expensive. However, certain statistical patterns of transitional matrices can be exploited to reduce the complexity of Eq. 17.

Suppose $\mathbf{A}$ is symmetric. Its singular value decomposition (SVD) then reveals equivalent input and output bases: $\mathbf{A}=\boldsymbol{\Phi}_a\boldsymbol{\Lambda}_a\boldsymbol{\Phi}_a^*$, where $\boldsymbol{\Lambda}_a$ is diagonal and $\boldsymbol{\Phi}_a$ is orthonormal. The self-multiplication of $\mathbf{A}$ can then be expressed as:

$$\mathbf{A}^k = (\mathbf{\Phi}_a \mathbf{\Lambda}_a \mathbf{\Phi}_a^*)^k = \mathbf{\Phi}_a \mathbf{\Lambda}_a^k \mathbf{\Phi}_a^*, \tag{18}$$

reducing the required computation due to the diagonal nature of $\mathbf{\Lambda}_a$. If $\mathbf{A}$ and $\mathbf{B}$ are symmetric matrices, Eq. 17 becomes:

$$P\left(s_{n+k} = i | x_n, x_{n+N}\right) = \frac{\kappa}{\boldsymbol{\pi}\left(i\right)} \left[\mathbf{\Phi}_a \mathbf{\Lambda}_a^k \mathbf{\Phi}_a^* \mathbf{e}_{q_n}\right]_{(i)} \left[\mathbf{\Phi}_b \mathbf{\Lambda}_b^{N-k} \mathbf{\Phi}_b^* \mathbf{e}_{q_{n+N}}\right]_{(i)}. \tag{19}$$

Transitional matrices of LSF parameters show strong symmetric patterns. However, because such matrices are data-generated, they are not perfectly symmetric, and we wish to add a perturbation matrix $\mathbf{\Delta}$ to $\mathbf{A}$ such that:

$$\bar{\mathbf{A}} = \mathbf{A} + \mathbf{\Delta} = (\mathbf{A} + \mathbf{\Delta})^T = \bar{\mathbf{A}}^T. \tag{20}$$

Because $\mathbf{A}$ is a transition probability matrix, its columns each sum to unity, i.e. $\mathbf{1}^T\mathbf{A} = \mathbf{1}^T$, where $\mathbf{1}$ is an appropriately-sized unity vector. Additionally, $\mathbf{\Delta}$ should be as small as possible to minimize its statistical effect on $\mathbf{A}$. Thus, we have:

$$\overset{*}{\mathbf{\Delta}} = \underset{\mathbf{\Delta}}{\operatorname{argmin}} \|\mathbf{\Delta}\|_{FW} \text{ s.t. } (i)\, \mathbf{1}^T\mathbf{\Delta} = \mathbf{0}^T \quad (ii)\, \mathbf{A} + \mathbf{\Delta} = (\mathbf{A} + \mathbf{\Delta})^T. \tag{21}$$

Here, $_{FW}$ indicates a weighted Frobenius norm where each entry in Δ can be weighted differently to avoid negative entries in $\bar{\mathbf{A}}$. We weight Δ_{ij} by $\frac{1}{\mathbf{A_{ij}}+\mathbf{A_{ji}}}$, thereby restricting large changes in near-zero entries of $\mathbf{A}$. Eqn. (21) can be posed as:

$$\overset{*}{\mathbf{d}} = \underset{\mathbf{d}}{\operatorname{argmin}} \|\mathbf{d}\|_{FW} \text{ s.t. } (i)\, \mathbf{A}_{eq}\mathbf{d} = \mathbf{b}_{eq}, \tag{22}$$

where $\mathbf{d} = \mathbf{vec}(\mathbf{\Delta})$, and where $\mathbf{A}_{eq}$ and $\mathbf{b}_{eq}$ represent a set of linear equality constraints on $\mathbf{d}$ imposed by (i) and (ii) from Eq. 21. The first constraint in Eqn. 21 results in a set of n equality constraints, and the second yields another $\frac{n^2-n}{2}$ constraints. Thus, $\mathbf{A}_{eq} \in \mathbb{R}^{\frac{n^2+n}{2} \times n^2}$ represents an underdetermined linear system, and $\mathbf{A}_{eq}\mathbf{d} = \mathbf{b}_{eq}$ has an infinite set of solutions. The weighted least-norm solution to this underdetermined set of equations is well-known:

$$\overset{*}{\mathbf{d}} = \mathbf{W}^{-1}\mathbf{A}_{eq}^T(\mathbf{A}_{eq}\mathbf{W}^{-1}\mathbf{A}_{eq}^T)^{-1}\mathbf{b}_{eq}, \tag{23}$$

where $\mathbf{W}$ is a weighting matrix and $\overset{*}{\mathbf{\Delta}}$ is obtained by reshaping $\overset{*}{\mathbf{d}}$. A corresponding perturbation matrix for $\mathbf{B}$ can be found similarly. It is important to note that the singular value decompositions and optimization techniques proposed in this section are performed offline, and required complexity of these operations is therefore not an issue.

3.3. Complexity Analysis

Sections 3.1 and 3.2 present methods by which to reduce the complexity of the original HMM-based estimation method of Eq. 12, without noticeable degradation in performance. In this section, we provide quantitative complexity analysis, and show the efficient method of Eq. 19 to result in significant reductions in required computation for error burst lengths of ≥ 2.

The computational complexity associated with matrix multiplication of full matrices of size $M \times M$ is known to be $O\left(M^3\right)$. If either of the matrices is known to be diagonal, this sparse structure can be exploited, and the complexity reduces to $O\left(M^2\right)$. If both matrices are diagonal, the complexity reduces further to $O\left(M\right)$. Using this, the number of required multiplications for the standard method of Eq. 12 and reduced complexity method of Eq. 19 are plotted in Figure 1 as a function of the duration of error burst. Note that R refers to the resolution of the downsampled HMM. It is clear from Figure 1 that model downsampling and enforcing matrix symmetry offers significantly reduced complexity for error durations of ≥ 2, making the algorithm ideal for bursty channels. In fact, PLC schemes can be implemented wherein Eq. 19 is applied only for missing features occurring at least 2 frames into an error burst.

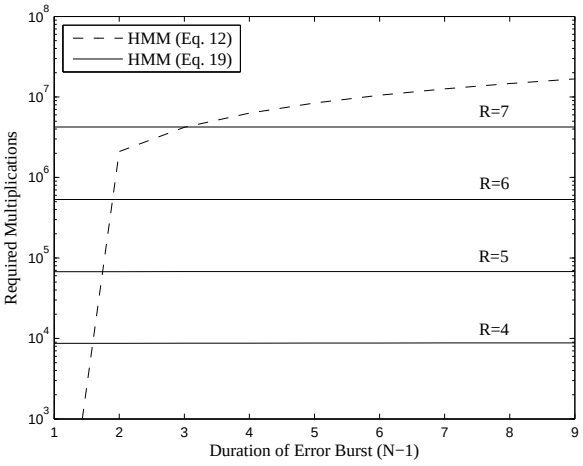

Figure 1: Induced Complexity of HMM-based Estimation of Missing Features for Original Method (Eq. 12) and Reduced Complexity Method (Eq. 19)

4. Experimental Results

The proposed estimation techniques were applied to speech features generally utilized by parametric speech coders, namely line spectral frequencies (LSFs). We simulated a bursty channel for which we used a two-state model wherein state 0 incurred no loss, and state 1 incurred a dropped packet with probability 1. The probability of self-transition within state 1 was twice that of the probability of transition from 0 to 1. Note that within the assumed model, the average burst duration is equal to:

$$\text{Ave. Burst Duration} = 2\left(\text{Error Rate}\right) + 1. \tag{24}$$

The proposed HMM-based framework was applied to 100 randomly selected utterances from the TIMIT database, separate from the training set, and using the previously mentioned channel model. Figure 2 provides an illustrative example of the reconstruction of missing values for the 1^{st} LSF. "REP" refers to the baseline scheme wherein reliable features are repeated (as in [3]), whereas "HMM" refers to the proposed HMM-based framework. It can be observed that proposed HMM-based estimation generally provides more accurate reconstructions, as well as smoother transitions, relative to the baseline. The quantitative quality of reconstructed LSF feature trajectories was assessed by peak-SNR (PSNR) and Itakura-Saito distortion (ISD). Figure 3 and Table 1 provide results for R=7. Recall that the difference between Eq. 15 and 12 is that the latter requires future observations, whereas the former does not. It can be observed that both HMM-based methods provide significant improvements relative to the baseline, in terms of PSNR and ISD.

Figure 4 provides results for estimation of missing LSF

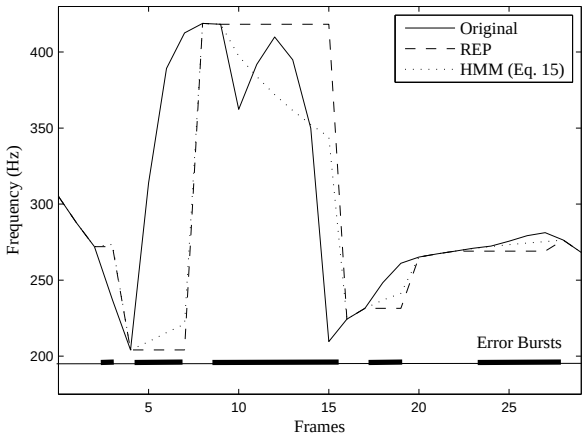

Figure 2: Reconstructed Trajectories for the 1^{st} LSF in the Presence of Error Bursts. "REP" refers to the baseline repetition scheme, whereas "HMM" refers to HMM-based estimation with R=7. Error burst are denoted by horizontal bars.

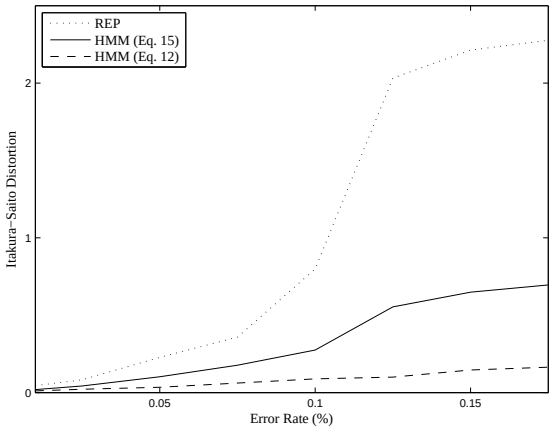

Figure 3: Itakura-Saito Distortion for Estimation of Missing LSF Features as a Function of Error Rate. "REP" refers to the baseline repetition scheme, whereas "HMM" refers to the proposed HMM-based framework with R=7.

Table 1: Improvements in Peak-SNR (dB), relative to feature repetition, for Estimation of Missing LSF Features as a Function of Error Rate, for R=7. Results are Averaged Across 10 Individual LSFs.

Error Rate (%)	5	10	15	20
HMM (Eq. 15)	0.75	1.06	0.96	0.84
HMM (Eq. 12)	3.62	3.62	3.64	3.46

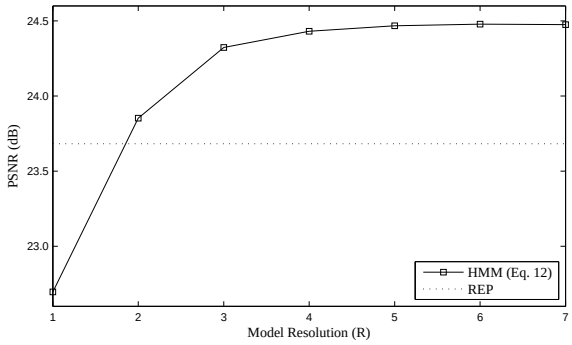

Figure 4: The Effect of Model Downsampling on PSNR for HMM-based Estimation with Eq. 12 with a 5% Error Rate

vectors using reduced complexity estimation developed in Section 3. Specifically, the figure illustrates the effect of HMM downsampling on PSNR, for a 5.0% error rate. It can be observed that performance of the proposed estimation technique converges for $R \approx 4$, corresponding to 16 states. Note that this corresponds to a model size that is significantly smaller than those used previously in [2]. Furthermore, it was observed that enforcing transition matrix symmetry had a negligible effect on estimation performance in terms of ISD.

5. Conclusions

In this paper, we present efficient HMM-based estimation techniques for missing speech features, with applications to parametric coding. By assuming features to be observations of hidden Markov processes, we derive the minimum mean-square error solutions for estimating unreliable features. We explore computationally efficient approximations to the derived solutions by downsampling underlying Markov models, and by enforcing symmetry in transitional probability matrices. When applied to features generally utilized by parametric coding, the proposed estimation methods outperform baseline repetition scheme in terms of both PSNR and ISD.

6. References

[1] T. Fingscheidt and P. Vary, *Softbit Speech Decoding: a New Approach to Error Concealment*, IEEE Trans. on Speech and Audio Processing, Vol. 9, No. 3, 2001.

[2] C. A. Rodbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, *Hidden Markov Model Based Loss Concealment for Voice Over IP*, Transactions for Audio, Speech, and Language Processing, Vol. 14, No. 5, pp. 1609-1623, 2006.

[3] R. Martin, C. Hoelper, and I. Wittke, *Estimation of Missing LSF Parameters Using Gaussian Mixture Models*, ICASSP, pp. 729-732, 2001.

[4] J.-H Chen, *Packet Loss Concealment for Predisctive Speech Coding Based on Extrapolation of Speech Waveform*, Proc. of ACSSC, pp. 2088-2092, 2007.

[5] L. R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, vol. 77, No. 2, pp. 257-286, 1989.

[6] A. M. Kondoz, *Digital Speech: Coding for Low Bitrate Communication Systems*, Wiley, 2004.

[7] B. J. Borgstrom and A. Alwan, *HMM-Based Reconstruction of Unreliable Spectrographic Data for Noise Robust Speech Recognition*, IEEE Trans. on Audio, Speech and Language Processing, to appear.