JOINT SOURCE-CHANNEL DECODING OF SPEECH SPECTRUM PARAMETERS OVER ERASURE CHANNELS USING GAUSSIAN MIXTURE MODELS

Anand D. Subramaniam, William R. Gardner and Bhaskar D. Rao

Department of Electrical and Computer Engineering
University of California, San Diego
E-mail: anand,wgardner,brao@ece.ucsd.edu
URL: http://dsp.ucsd.edu

ABSTRACT

A joint source-channel decoding scheme that improves the performance of conventional channel decoders over erasure channels by exploiting the cross-correlation between successive speech frames is presented. Speech spectrum parameters are quantized using the scheme presented in [1]. The joint probability density function (PDF) of the spectrum parameters of successive speech frames is modelled using a Gaussian mixture model (GMM). This model is then used to process the channel decoder output over erasure channels. The performance of two decoding strategies, namely, Maximum Likelihood decoding (ML) and Minimum Mean Squared Error decoding (MMSE) is shown to provide significantly better performance than prediction based schemes.

1. INTRODUCTION

In conventional speech compression schemes, when the number of erasures introduced by the channel is more than the erasure correcting capability of the channel code, error concealment is done by employing prediction methods. In this paper, we propose a joint source-channel decoding scheme where the joint probability density function between the spectrum parameters of successive speech frames is used to aid the channel decoder. We propose two strategies, namely, Maximum Likelihood (ML) decoding and Minimum mean square error (MMSE) decoding. In ML decoding, we choose that source codepoint that maximizes the conditional probability given the previous frame and the channel decoder output. In the MMSE decoding case, we decode to the conditional mean given the previous frame and the channel decoder output. We demonstrate that the proposed schemes can perform significantly better than prediction based schemes for various channel conditions.

In order to implement the proposed joint source-channel decoding schemes, the conditional probability of a given source code-point needs to be computed. We show that the framework presented in [1] allows for an easy implementation of the proposed scheme using results from high-resolution theory. The scheme proposed in this paper can be easily extended to a broad range of channels and source-channel coding schemes.

Section 2 describes the modelling of the source using Gaussian mixture models. Section 3 provides the details of the encoding procedure and Section 4 explains the various decoding strategies. The experimental results are provided in Section 5 and the computational complexity is analyzed in Section 6.

2. SOURCE MODEL

In this work, we model the joint probability density function of the spectrum parameters of successive speech frames using a Gaussian Mixture Model (GMM). Let $\mathbf{X}$ and $\mathbf{Y}$ be d-dimensional random vectors representing the spectrum parameters of current and previous speech frames respectively.

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{m} \alpha_i \mathrm{N}_i\left(\mu_i, C_i\right) \quad (1)$$

$$\mu_i = \begin{pmatrix} \mu_i^{\mathbf{X}} \\ \mu_i^{\mathbf{Y}} \end{pmatrix} \quad (2)$$

$$C_i = \begin{pmatrix} C_i^{\mathbf{XX}} & C_i^{\mathbf{XY}} \\ C_i^{\mathbf{YX}} & C_i^{\mathbf{YY}} \end{pmatrix} \quad (3)$$

where $\mathrm{N}_i\left(\mu_i, C_i\right)$ is an individual $2d$-dimensional Gaussian with mean μ_i and covariance matrix C_i and α_i are positive scalars that sum to unity.

The parameters of the GMM can be efficiently estimated using the Expectation-Maximization (EM) algorithm [3]. The marginal density of the spectrum parameters of a speech frame can be obtained from the joint density as,

$$f_{\mathbf{X}}(\mathbf{X}) = \sum_{i=1}^{m} \alpha_i \mathrm{N}_i\left(\mu_i^{\mathbf{X}}, C_i^{\mathbf{XX}}\right) \quad (4)$$

This research is supported by Qualcomm and Nokia under the MICRO Grant No. 00-082 and 99-101 respectively.

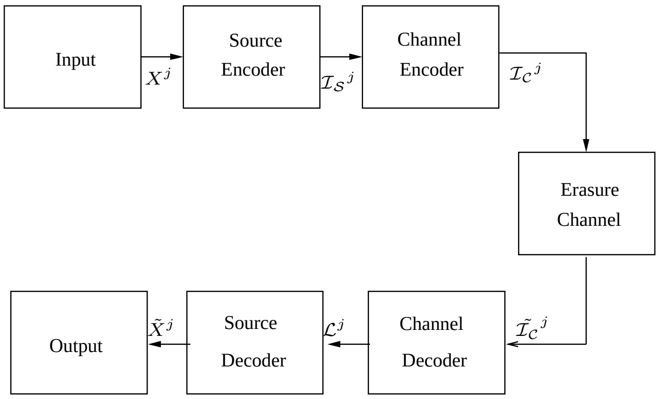

Fig 1 : Overall communication scheme

3. ENCODING

3.1. Source Encoder

We perform fixed-rate memoryless quantization on the spectrum parameters of each frame by using the method proposed in [1]. The marginal density of each frame (cf. Eq. 4) is used as the source model for the memoryless quantizer. The source encoder generates a k bit index, $\mathcal{I_S}$, for each frame.

3.2. Channel Encoder

The channel encoder produces a n-bit channel codeword, $\mathcal{I_C}$, for each input k-bit source index $\mathcal{I_S}$. The erasure correcting capability of the channel code is given by $e_{code} = n - k$, i.e., the channel code can correct upto e_{code} erasures. One example of a channel code that can be employed for this purpose is a Reed-Solomon code.

4. DECODING

The channel codeword, $\mathcal{I_C}$, is sent over a binary erasure channel with erasure probability p. We define the quantity ”excess erasure”, e_{ex}, as

$$e_{ex} = np - e_{code} \tag{5}$$

e_{ex} is a measure of the remaining erasures after channel decoding. In this paper, we design source decoders that decode in the presence of e_{ex} erasures.

4.1. Channel Decoder

Let e^j be the number of erasures introduced by the channel in the j^{th} speech frame ($e^j \leq n$). If $e^j \leq e_{code}$, the channel output can be perfectly decoded by the channel decoder. When $e^j > e_{code}$, the channel decoder outputs a list $\mathcal{L}_j$ containing $N_j = 2^{e^j_{ex}}$ entries, where $e^j_{ex} = e^j - e_{code}$ ($e^j_{ex} < k$). This list contains the transmitted source code-point as one of its entries.

4.2. Source Decoder

The source decoder comes into play only when the number of erasures introduced by the channel is more than the erasure correcting capability of the channel code, i.e., $e^j > e_{code}$. Under these circumstances, the source decoder tries to utilize the joint statistics of the spectrum parameters of successive frames to process the channel decoder output $\mathcal{L}_j$. It must be noted that the joint source-channel decoding schemes presented in this paper can be employed in any setting where the channel decoder outputs a list of possible code-points. One example of a plausible application area is the list decoding of Reed-solomon codes over additive white Gaussian channels.

4.2.1. Prediction

In this approach, when the channel decoder is unable to perfectly decode the channel output, we predict the spectrum parameters of the current frame based on the decoded parameters of the previous frame. We present two options: Linear Prediction (LP) and Gaussian mixture model prediction (GMM-P).

- Linear Prediction

  In this case, we use the optimal linear prediction estimate of the spectrum parameters of the current frame based on the spectrum parameters of the previous frame.

$$\begin{aligned} \tilde{X}^j_{LP} &= \mathbf{A} \cdot (\tilde{X}^{j-1}_{LP} - \mu^{\mathbf{X}}) + \mu^{\mathbf{X}} && (6) \\ \mathbf{A} &= C^{\mathbf{XY}} (C^{\mathbf{YY}})^{-1} && (7) \end{aligned}$$

- GMM Prediction

  In this case, we use the conditional mean of the current frame based on the previous frame. The conditional density of the current frame based on the frame is given by [2],

$$\begin{aligned} f_{\mathbf{X}|\mathbf{Y}}(\mathbf{X} \mid \mathbf{Y}) &= \sum_{i=1}^{m} \beta_i(\mathbf{Y}) \mathrm{N}_i \left(m_i(\mathbf{Y}), \Sigma_i \right) && (8) \\ m_i(\mathbf{Y}) &= \mu_i^{\mathbf{X}} + C_i^{\mathbf{XY}} \left(C_i^{\mathbf{YY}} \right)^{-1} \left(\mathbf{Y} - \mu_i^{\mathbf{Y}} \right) \\ \Sigma_i &= C_i^{\mathbf{XX}} - C_i^{\mathbf{XY}} (C_i^{\mathbf{YY}})^{-1} C_i^{\mathbf{YX}} \\ \beta_i(\mathbf{Y}) &= \frac{\alpha_i f_i(\mathbf{Y})}{\sum_{j=1}^{m} \alpha_j f_j(\mathbf{Y})} \\ f_i(\mathbf{Y}) &= \mathrm{N} \left(\mu_i^{\mathbf{Y}}, C_i^{\mathbf{YY}} \right) \end{aligned}$$

  Note that the conditional covariance is independent of $\mathbf{Y}$ and can be pre-computed. Let $\tilde{X}^j_{PRED}$ be the

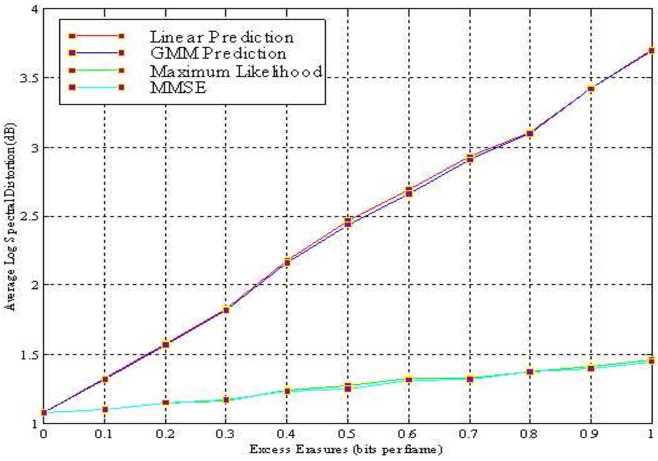

Fig. 2: Performance

decoded prediction estimate. This can be computed from the conditional density as,

$$\begin{aligned}\tilde{X}^j_{GMMP} &= \mathrm{E}\left(\mathbf{X} \mid \mathbf{Y} = \tilde{X}^{j-1}_{GMMP}\right) \qquad (9)\\ &= \sum_{i=1}^{m} \beta_i(\tilde{X}^{j-1}_{GMMP}) \cdot m_i(\tilde{X}^{j-1}_{GMMP})\end{aligned}$$

It can be easily seen that this estimate performs better than linear prediction estimate for minimizing mean-squared error criterion since the GMM captures the joint statistics between successive frames better than single Gaussians.

4.2.2. *Maximum Likelihood*

In this approach, we decode to that source code-point that has maximum probability of occurrence given the previous decoded frame and the channel decoder output, i.e.,

$$\tilde{X}^j_{ML} = \arg\max_{X \in \mathcal{L}_j} P(\mathbf{X} = X \mid \mathbf{Y} = \tilde{X}^{j-1}_{ML}) \qquad (10)$$

In order to compute the probability of occurrence of a given source code-point X, we need to integrate the conditional density over the voronoi cell of X whose volume is $V(X)$. We use the high-resolution approximation and approximate this probability as,

$$P(X \mid \tilde{X}^{j-1}_{ML}) = f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{X} = X \mid \mathbf{Y} = \tilde{X}^{j-1}_{ML}\right) \cdot V(X) \quad (11)$$

In high resolution, the volume of a voronoi cell can be approximated by,

$$V(X) = \frac{1}{2^k \cdot \Lambda(X)} \qquad (12)$$

Table 1. Performance Results

e_{ex}	L-P	GMM-P	M-L	MMSE
0.0	1.0749	1.0749	1.0749	1.0749
0.2	1.5792	1.5682	1.1463	1.1485
0.4	2.1755	2.1617	1.2370	1.2311
0.6	2.6915	2.6609	1.3244	1.3123
0.8	3.1068	3.0978	1.3700	1.3697
1.0	3.7007	3.6991	1.4603	1.4451

where $\Lambda(X)$ is the code-point density of the memoryless quantizer. As shown in the Appendix, the point density of the fixed-rate memoryless quantizer is given in closed form as,

$$\begin{aligned}\Lambda(X) &= \frac{\sum_{i=1}^{m} g_i(X)^{1/3}(\alpha_i\lambda_i)^{d/(d+2)}\lambda_i^{-d/3}}{\sum_{j=1}^{m}(\alpha_j\lambda_j)^{d/(d+2)}} \qquad (13)\\ g_i(X) &= \mathrm{N}\left(\mu_i^{\mathbf{X}}, C_i^{\mathbf{XX}}\right) \qquad (14)\end{aligned}$$

where λ_i is the geometric mean of the singular values of $C_i^{\mathbf{XX}}$.

4.2.3. *Minimum Mean Square Error*

The MMSE estimate is the conditional mean given the previous frame and the channel decoder output. Specifically, we use the source model to compute the probability of each possible codeword and then compute the conditional mean.

$$\begin{aligned}\tilde{X}^j_{MMSE} &= \mathrm{E}\left(\mathbf{X} \mid \mathbf{Y} = \tilde{X}^{j-1}_{MMSE}, \mathcal{L}_j\right) \qquad (15)\\ &= \frac{\sum_{X \in L_j} P(X \mid \tilde{X}^{j-1}_{MMSE}, L_j) \cdot X}{\sum_{X \in \mathcal{L}_j} P(X \mid \tilde{X}^{j-1}_{MMSE}, L_j)} \qquad (16)\end{aligned}$$

5. EXPERIMENTAL RESULTS

The performance of the proposed scheme was tested in the application of speech spectrum quantization. Speech was broken into frames of duration 25 msec and a ten dimensional ($d = 10$) vector of Line spectrum pairs (LSP) was extracted from each frame. The joint PDF of the spectrum parameters of successive speech frames was estimated using a Gaussian mixture model consisting of $m = 32$ clusters. The training database consisted of 100,000 frames of speech. The proposed scheme was tested on an independent database consisting of 10,000 frames of speech.

Fixed-rate memoryless quantization is performed on the LSP parameters with $k = 24$ bits per frame which provides an average encoder log spectral distortion of 1.0749 dB. Since the number of bits used by the source encoder per

Table 2. Computational Complexity

Scheme	N_{tot} (flops per frame)
LP	$2d^2 + d$
GMM-P	$m(N_{exp} + 2d^2 + d + 3) - 2$
M-L	$2^{e^j_{ex}}[(2m+1)N_{exp} + m(2d^2+d) + 5m - 2]$
MMSE	$2^{e^j_{ex}}[(2m+1)N_{exp} + m(2d^2+d) + 5m + d]$

dimension is 2.4, the high resolution approximations employed in this paper are reasonable [4].

The performance results for all the decoding schemes is compared in Table 1 and Fig. 2. The average end-to-end log spectral distortion is plotted against the average excess erasures seen by the source decoder. A value of 0.1 means that on an average, the source decoder sees an erasure rate of one bit in ten source frames. The simulations reveal that GMM-P perrformance is only marginally better than LP performance. Furthermore, both ML and MMSE joint source channel decoding schemes perform significantly better than the prediction schemes although there is no discernable difference between the performance of these joint source-channel decoding schemes.

6. COMPUTATIONAL COMPLEXITY

In this section, we analyze the complexity of the proposed schemes. The complexity of the joint source-channel decoding schemes scales exponentially with the number of erasures while the complexity of the prediction based schemes is independent of the number of erasures. Let N_{exp} be the number of flops used to perform one exponentiation. Let e^j_{ex} be the number of excess erasures in the j^{th} frame. Table 2 compares the complexity of the proposed schemes.

As shown in the simulations, the added complexity of GMM prediction is not commensurate with the incremental improvement in the decoder performance as compared with Linear prediction. Furthermore, the significantly improved performance of the M-L and MMSE approaches may be worth the additional complexity in many applications since the complexity of these approaches is only a fraction of the overall source-channel encoder-decoder complexity.

7. CONCLUSION

A joint source-channel decoding scheme for erasure channels based on Gaussian mixture models is proposed. The proposed scheme provides significantly better performance in comparison to prediction based schemes. The proposed scheme can be extended to other channels where the channel decoder is capable of providing a short-list of possible codewords to the source decoder.

8. APPENDIX

In the fixed rate case, the cluster bit-allocation is given by [1],

$$2^{b_i} = 2^{b_{tot}} \frac{(\alpha_i \lambda_i)^{d/(d+2)}}{\sum_{j=1}^{m} (\alpha_j \lambda_j)^{d/(d+2)}} \quad (17)$$

Since each cluster quantizer is essentially a product of individual scalar quantizers, the point density function for a cluster i is given by,

$$\Lambda_i(X) = \frac{g_i(X)^{1/3}}{\int g_i(X)^{1/3} dX} \quad (18)$$

$$g_i(X) = \mathrm{N}\left(\mu_i^{\mathbf{X}}, C_i^{\mathbf{XX}}\right) \quad (19)$$

The total point density of the quantizer is given by,

$$\Lambda(X) = \sum_{i=1}^{m} \frac{2^{b_i}}{2^{b_{tot}}} \Lambda_i(X) \quad (20)$$

$$= \sum_{i=1}^{m} \frac{(\alpha_i \lambda_i)^{d/(d+2)}}{\sum_{j=1}^{m} (\alpha_j \lambda_j)^{d/(d+2)}} \cdot \Lambda_i(X) \quad (21)$$

After performing a few manipulations we get,

$$\Lambda(X) = \frac{\sum_{i=1}^{m} g_i(X)^{1/3} (\alpha_i \lambda_i)^{d/(d+2)} \lambda_i^{-d/3}}{\sum_{j=1}^{m} (\alpha_j \lambda_j)^{d/(d+2)}} \quad (22)$$

9. REFERENCES

[1] Anand D. Subramaniam and Bhaskar D. Rao, “Speech LSF quantization with rate-independent complexity, bit-scalability and learning.”, International Conf. on Acoustics, Speech and Signal Processing,pp. 705-8, 2001.

[2] Anand D. Subramaniam, William R. Gardner and Bhaskar D. Rao, “Low complexity recursive coding of spectrum parameters.”, International Conf. on Acoustics, Speech and Signal Processing, 2002.

[3] R.A. Redner and H.F. Walker, “Mixture densities, Maximum Likelihood and The EM Algorithm”, SIAM Rev., Apr 1984.

[4] R.M. Gray and D.L. Neuhoff, “Quantization”, IEEE Transactions on Information Theory, vol.44, (no.6), IEEE, Oct. 1998.