Eurospeech 2003, Geneva

Towards Optimal Encoding for Classification with Applications to Distributed Speech Recognition

Naveen Srinivasamurthy, Antonio Ortega and Shrikanth Narayanan

Department of EE-Systems, IMSC,
University of Southern California, Los Angeles.
[snaveen,ortega,shri]@sipi.usc.edu

Abstract

In distributed classification applications, due to computational constraints, data acquired by low complexity clients is compressed and transmitted to a remote server for classification. In this paper the design of optimal quantization for distributed classification applications is considered and evaluated in the context of a speech recognition task. The proposed encoder minimizes the detrimental effect compression has on classification performance. Specifically, the proposed methods concentrate on designing low dimension encoders. Here individual encoders *independently* quantize sub-dimensions of a high dimension vector used for classification. The main novelty of the work is the introduction of mutual information as a metric for designing compression algorithms in classification applications. Given a rate constraint, the proposed algorithm *minimizes the mutual information loss* due to compression. Alternatively it ensures that the compressed data used for classification retains maximal information about the class labels. An iterative *empirical* algorithm (similar to the Lloyd algorithm) is provided to design quantizers for this new distortion measure. Additionally, mutual information is also used to propose a rate-allocation scheme where rates are allocated to the sub-dimensions of a vector (which are independently encoded) to satisfy a given rate constraint. The results obtained indicate that mutual information is a better metric (when compared to mean square error) for optimizing encoders used in distributed classification applications. In a distributed spoken names recognition task, the proposed mutual information based rate-allocation reduces by a *factor of six* the increase in WER due to compression when compared to a heuristic rate-allocation.

1. Introduction

In distributed speech recognition (DSR) [1], low complexity clients (e.g., cellphones, PDAs) which do not have sufficient computation/memory resources to support complex recognition tasks, acquire speech and transmit it to a remote server for recognition. Instead of transmitting the speech utterance, feature frames used by the recognizer are extracted, compressed (to conserve bandwidth) and transmitted. High dimensionality of the features requires, for computational reasons, that each frame (sub-dimension) of the speech feature be *independently* quantized. The main difficulty in designing independent quantizers is that traditional distortion measures (e.g., mean square error (MSE)) used for quantization design can be separably calculated for each of the frames. However, probability of misclassification, the metric for evaluating classification performance, is a "global" measure defined only for the entire speech utterance. To address this challenge of designing quantizers for individual frames, we use the *information-theoretic* measure of mutual information [2] to define a sub-dimension based distortion measure. The distortion defined by mutual information is more suitable to evaluate the effect compression has on misclassification than was possible by MSE.

Our proposed approach attempts to minimize the classification error for a given rate with the constraint that individual frames are independently quantized. The contributions of our work include (i) a mathematically tractable distortion measure better suited for classification applications, which can be defined even for sub-dimensions of a vector used for classification and (ii) an empirical quantizer design algorithm which does not require prior knowledge of source or class *pdfs*.

Bayes VQ [3], a joint compression and classification encoder, designs a vector quantizer (VQ) to minimize the additive weighted cost of distortion and misclassification. While achieving good performance, the main drawback of this approach is that the encoder has to operate on the entire vector used for classification. In previous work [4], we extended this approach by combining the local cost (distortion) with the global cost (misclassification) to design encoders which could independently encode sub-dimension of the vector used for classification. Incorporating rate constraints in the above approaches was complicated by the fact that it required minimization of a weighted cost with two unknown Lagrange multipliers. However, in our approach inclusion of rate constraint results in a single Lagrange multiplier weighted cost which can be optimally solved [5].

The Information Bottleneck Method [6] ensures that the compressed data at a given rate retains maximal information about the class labels. It assumes a soft partitioning of the input space where every data point is probabilistically associated to a reproduction codeword. It is shown that the Kullback-Liebler (KL) distance is the "relevant" distortion for this problem. An iterative algorithm similar to the Blahut-Arimoto algorithm is provided to design the quantizers. In contrast, our proposed work provides an iterative quantizer design which assumes a hard partitioning of the input space, i.e., every input data point is uniquely assigned to only one reproduction codeword. The advantage of our quantizer design algorithm is that it is *empirical*. The proposed algorithm infers/calculates the required *pdfs* during the design process. Furthermore the Information Bottleneck Method does not address either the design of sub-dimension encoders or rate-allocation to the different sub-dimensions.

In a distributed spoken names recognition task at 3920 bps, rate-allocation based on our proposed technique resulted in only a 0.31% increase in the WER compared to using unquantized features. While a heuristic rate-allocation technique [7] resulted in a 1.92% increase in WER. In an eight-way classification task at 2.3 bits-per-sample our proposed quantizer had only a 2.0% increase in misclassification, while a MSE based quantizer had a 7.6% increase in misclassification.

This research has been funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152 and DARPA Grant No. N66001-02-C-6023.

2. Minimum Mutual Information Loss Encoder

Before describing our minimum mutual information loss encoder, we introduce our notation. We represent vectors in bold and the components of the vector are enclosed in $\{\}$. Random variables (RVs) are represented by uppercase letters and the value taken by the RVs by lowercase letters, i.e., $\mathbf{X} = \mathbf{x} = \{x_1, \ldots, x_N\}$ implies the vector RV $\mathbf{X}$ takes the value $\mathbf{x}$, which has N components x_i, $i = 1, \ldots, N$. $\delta(\cdot)$ denotes a statistical classifier, $\alpha(\cdot)$ and $\beta(\cdot)$ denote the encoder and decoder respectively. We use the quantizer, $Q(\cdot)$, as a shorthand for the encoder-decoder pair, i.e., $Q(\cdot) \triangleq \beta(\alpha(\cdot))$.

Let $[\mathbf{Y}, C]$ denote a continuous (NT)-dimensional RV $\mathbf{Y} = \{\mathbf{X_1}, \ldots, \mathbf{X_T}\}$, which is associated with a class label C that takes a value in $1, \ldots, L$, and $\mathbf{X_i}$ is an N-dimensional RV. In speech recognition the RV $\mathbf{Y}$ represents feature frames of a phoneme, $\mathbf{X_i}$ represents a feature frame, and C represents the phoneme $\mathbf{Y}$ belongs to. Assume to represent $\mathbf{Y}$[1] we are given a rate R, the rate constraint could be due to transmission or storage requirements. The problem we consider is that of finding the best representation $\hat{\mathbf{Y}} = \{\hat{\mathbf{X}}_\mathbf{1}, \ldots, \hat{\mathbf{X}}_\mathbf{T}\} = \{Q(\mathbf{X_1}), \ldots, Q(\mathbf{X_T})\}$ s.t. $H(\hat{\mathbf{Y}}) \leq R$, which minimizes the probability of error in classification, i.e., $P_e(\delta(\hat{\mathbf{Y}}) \neq C)$, where $H(\hat{\mathbf{Y}})$ is the entropy of the RV $\hat{\mathbf{Y}}$. The fundamental problem we are addressing is given the constraint that the classifier has to operate on compressed data with a rate limit, what is the best product quantizer that minimizes the probability that the class label is different when obtained from unquantized and quantized data? Note that we design the same quantizer $Q(\cdot)$ for all the feature frames, $\mathbf{X}_i$, $i = 1, \ldots, T$.

Unfortunately, $P_e(\cdot)$ does not have a mathematically tractable form which makes the above problem difficult to solve. In this work we consider mutual information (MI) as an approximation to $P_e(\cdot)$ that can enable designing practical quantization schemes. Unlike $P_e(\cdot)$, which is not defined for sub-vectors (the classifier can only make its decision using the entire vector), MI can be calculated even between sub-vectors $\mathbf{X_i}$ and the class labels. Since we design the same quantizer for all feature frames in what follows we drop the subscript in $\mathbf{X_i}$.

Different speech utterances could contain the same feature frames (possibly in different temporal locations) and still belong to different classes. So it is obvious that given a feature frame the class labels need not be the same, i.e., every feature frame $\mathbf{X} = \mathbf{x}$ is associated with a conditional *pdf* $p(c|\mathbf{x})$. Hence MI between the RV $\mathbf{X}$ and the RV C (class label) is given by

$$I(\mathbf{X}; C) = \int_{\mathbf{x}} f(\mathbf{x}) \sum_c p(c|\mathbf{x}) log \left(\frac{p(c|\mathbf{x})}{p(c)} \right) d\mathbf{x} \quad (1)$$

It is well known that

$$H(C|\mathbf{X}) = H(C) - I(\mathbf{X}; C) \quad (2)$$

where $H(C)$ is the original entropy of the class labels and $H(C|\mathbf{X})$ is the entropy of the class labels after observing $\mathbf{X}$. Therefore the MI between C and $\mathbf{X}$ is the amount by which uncertainty in class labels is reduced by observing $\mathbf{X}$. Obviously the larger $I(\mathbf{X}; C)$ is, the more relevant (useful) $\mathbf{X}$ is for the classification task. This intuition has been used previously in speech recognition to propose a maximum mutual information speech recognizer design technique [8] as an alternative to maximum likelihood techniques. Our work also makes use of this above intuition to design a *minimum mutual information loss (MMIL) quantizer*, $Q_{MI}(\cdot)$, which minimizes the loss in MI, $I(\mathbf{X}; C) - I(\hat{\mathbf{X}}; C)$, due to compression, where $\hat{\mathbf{X}} = Q_{MI}(\mathbf{X})$. The loss in MI due to compression corresponds to an increase in class uncertainty (see Eq (7) below), hence the MMIL quantizer is well suited for compressing data in classification applications. The MMIL quantizer treats loss in MI as the distortion incurred during quantization similar to a minimum mean square error (MMSE) quantizer treating Euclidean distance as the distortion incurred.

The MI between the quantized data $\hat{\mathbf{X}}$ and C is

$$I(\hat{\mathbf{X}}; C) = \sum_c \sum_{\hat{\mathbf{x}}} p(c, \hat{\mathbf{x}}) log \left(\frac{p(c, \hat{\mathbf{x}})}{p(c) p(\hat{\mathbf{x}})} \right) \leq I(\mathbf{X}; C) \quad (3)$$

where $I(\hat{\mathbf{X}}; C) \leq I(\mathbf{X}; C)$ by the data processing inequality.

The *optimal MMIL quantizer*, $Q^*_{MI}(\cdot)$, subject to a rate constraint, is obtained by a constrained minimization

$$Q^*_{MI}() = argmin_{Q_{MI} : (I(\mathbf{X}; C) - I(Q_{MI}(\mathbf{X}); C)) \leq D} [I(\mathbf{X}; Q_{MI}(\mathbf{X}))]$$

Based on standard Lagrangian techniques this constrained minimization can be converted into an unconstrained minimization, i.e.,

$$\begin{aligned} Q^*_{MI}(\cdot) \quad &= \quad argmin_{Q_{MI}} I(\mathbf{X}; Q_{MI}(\mathbf{X})) \\ &\quad\quad + \lambda(I(\mathbf{X}; C) - I(Q_{MI}(\mathbf{X}); C)) \end{aligned} \quad (4)$$

where λ is the Lagrange multiplier which controls the trade-off between rate and distortion (i.e., loss in MI).

2.1. MMIL Encoder: Quantizer Design

One of our main motivations is to provide an *empirical* algorithm which can be used in practical applications to design quantizers directly from sample training data. To enable this we show how the distortion in Eq (4) can be estimated directly from data samples used to design the quantizer. First from the Markov chain $C \leftrightarrow \mathbf{X} \leftrightarrow \hat{\mathbf{X}}$, we have

$$p(\hat{\mathbf{x}}|c) = \int_{\mathbf{x}} p(\hat{\mathbf{x}}|\mathbf{x}) p(\mathbf{x}|c) d\mathbf{x} \quad (5)$$

and

$$p(c|\hat{\mathbf{x}}) = \int_{\mathbf{x}} p(c|\mathbf{x}) p(\mathbf{x}|\hat{\mathbf{x}}) d\mathbf{x} \quad (6)$$

From Eq (2), the MI loss (distortion) is

$$I(\mathbf{X}; C) - I(\hat{\mathbf{X}}; C) = H(C|\hat{\mathbf{X}}) - H(C|\mathbf{X}) \quad (7)$$

where

$$H(C|\mathbf{X}) = -\int_{\mathbf{x}} f(\mathbf{x}) \sum_c p(c|\mathbf{x}) log \left(p(c|\mathbf{x}) \right) d\mathbf{x} \quad (8)$$

and

$$H(C|\hat{\mathbf{X}}) = -\sum_c p(c) \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|c) log \left(p(c|\hat{\mathbf{x}}) \right) \quad (9)$$

Substituting Eq (5) in Eq (9) we get

$$H(C|\hat{\mathbf{X}}) = -\sum_c p(c) \int_{\mathbf{x}} p(\mathbf{x}|c) \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\mathbf{x}) log \left(p(c|\hat{\mathbf{x}}) \right) d\mathbf{x} \quad (10)$$

Note that $p(\hat{\mathbf{x}}|\mathbf{x}) = 1$ if $Q(\mathbf{x}) = \hat{\mathbf{x}}$ and 0 otherwise. Therefore define

$$q(c|\hat{\mathbf{x}}) = \begin{cases} p(c|\hat{\mathbf{x}}) & \text{if } Q(\mathbf{x}) = \hat{\mathbf{x}} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Substitute Eqs (8), (10) and (11) in Eq (7) (by continuity arguments we assume $0log(0) = 0$). After rearranging, we get

$$I(\mathbf{X}; C) - I(\hat{\mathbf{X}}; C) = \int_{\mathbf{x}} f(\mathbf{x}) \sum_c p(c|\mathbf{x}) log \left(\frac{p(c|\mathbf{x})}{q(c|\hat{\mathbf{x}})} \right) d\mathbf{x} \quad (12)$$

[1] The rate required to represent a continuous RV $\mathbf{Y}$ with exact precision is ∞

If $d(\mathbf{x}, \hat{\mathbf{x}})$ represents the distortion between $\mathbf{x}$ and $\hat{\mathbf{x}}$, then $E[d(\mathbf{x}, \hat{\mathbf{x}})] = \int_{\mathbf{x}} f(\mathbf{x}) d(\mathbf{x}, \hat{\mathbf{x}}) d\mathbf{x}$. This implies that the distortion measure for the MMIL quantizer is $\sum_c p(c|\mathbf{x}) log \left(\frac{p(c|\mathbf{x})}{q(c|\hat{\mathbf{x}})} \right)$, i.e., the distortion is the KL distance between the a-priori conditional class *pdfs* before quantization, $p(c|\mathbf{x})$, and the a-posteriori conditional class *pdfs* after quantization, $q(c|\hat{\mathbf{x}})$. Therefore the MMIL quantizer attempts to choose those codewords $\hat{\mathbf{X}}$ which best preserve the a-priori conditional class *pdfs*. It is obvious that for statistical classifiers (e.g., HMMs) which choose the class label based on a maximum likelihood decision using $p(c|\mathbf{x})$, the MMIL quantizer will have less detrimental effect on the probability of misclassification when compared to traditional mean square error (MSE) based quantizers which do not explicitly consider the conditional class *pdfs* during quantization design. If the source *pdfs* are not known fine quantization is used to find an empirical estimate $\tilde{p}(c|\mathbf{x})$ of $p(c|\mathbf{x})$ from labeled training data [7]. Eq (6) is then used to find $p(c|\hat{\mathbf{x}})$, where $\tilde{p}(c|\mathbf{x})$ is used instead of $p(c|\mathbf{x})$.

Algorithm 1 *Minimum mutual information loss vector quantizer design*

```
**Step 0:** Calculate $\tilde{p}(c|\mathbf{x})$ for $c = 1, \ldots, L$
**Step 1:** Initialize all codewords $\hat{\mathbf{x}}_i$, $i = 1, \ldots, I$, $d(0) = \infty, k = 1$. Let $\epsilon$ be a small positive constant.
**Step 2:** Find $p(c|\hat{\mathbf{x}}_i)$, $c = 1, \ldots, L$, $i = 1, \ldots, I$; using Eq (6)
**Step 3:** $\alpha(\mathbf{x}) = argmin_i \sum_c p(c|\mathbf{x}) log \left( \frac{p(c|\mathbf{x})}{q(c|\hat{\mathbf{x}}_i)} \right)$
**Step 4:** $\hat{\mathbf{x}}_i = \beta(i) = E[\mathbf{x}|\alpha(\mathbf{x}) = i]$
**Step 5:** $d(k) = \int_{\mathbf{x}} f(\mathbf{x}) \sum_c p(c|\mathbf{x}) log \left( \frac{p(c|\mathbf{x})}{q(c|\beta(\alpha(\mathbf{x})))} \right) d\mathbf{x}$
**Step 6:** if $(d(k-1) - d(k))/d(k) < \epsilon$ STOP, else $k = k + 1$, go to **Step 2**
```

Here in **Step 4** we make use of the result that for encoders designed using KL distance as the distortion measure, the optimal decoder is the Lloyd decoder [9].

The flexibility of our encoder design is that since we are using loss in MI as the distortion measure rather than $P_e(\cdot)$, it can easily be applied to design independent quantizers for each of the components of the vector X. If the entire vector has N components i.e., $\mathbf{X} = X_1, \ldots, X_N$, then the optimal quantizer $Q_{MI}^{j*}(\cdot)$ for the j^{th} component is designed by minimizing $I(X_j; C) - I(Q_{MI}^j(X_j); C)$.

Given that the designed quantizers need to satisfy a rate constraint, the standard entropy constrained quantization design technique [5] can be adopted, with MI loss as the distortion. The entropy constrained encoder $\alpha_n(\cdot)$ for the n^{th} component is

$$\alpha_n(x_n) = argmin_i \sum_c p(c|x_n) log \left(\frac{p(c|x_n)}{q(c|\hat{x}_{n_i})} \right) + \lambda l_{n_i} \tag{13}$$

where l_{n_i} is the number of bits used to represent the i^{th} codeword $\hat{x}_{n_i}$ in the n^{th} dimension. As a simplification we use $l_{n_i} = -log_2(p(\hat{x}_{n_i}))$. The decoder is the Lloyd decoder $\beta_n(i) = E[x_n|\alpha_n(x_n) = i]$.

2.2. MMIL Encoder: Rate-Allocation

Rate-allocation (or bit-allocation) plays an important role during designing independent encoders for components of a vector. The GBFOS algorithm [10] has been used for rate-allocation in MMSE encoders. It relies on the calculation of rate vs. distortion points for each of the components and then selects the combination of points which satisfy the rate constraint and yield the minimum distortion. For MMIL quantizers the distortion is MI loss, hence several *rate vs. mutual information loss* points are calculated for each of the components. Then if the available rate is R, these calculated points are used by the GBFOS algorithm to allocate rates, $R_n, n = 1, \ldots, N; \sum_n R_n \leq R$, to each of the components. During quantizer design, λ in Eq (13) is modified until $H(\hat{X}_n) \leq R_n$, standard bisection techniques can be used to find the "best" λ.

3. Experiments and Results

To evaluate our techniques we generated a mixture of eight 2D Gaussian sources. The means and the optimal Bayes classification boundaries for this mixture source are shown in Figure 1. 10,000 samples from each class were generated, each dimension was independently quantized and then used for classification. A classification error occurs if a sample from class i is classified as belonging to class $j \neq i$. The experiments were repeated 100 times and results reported are average results. The baseline classification error using unquantized data was 27.4%.

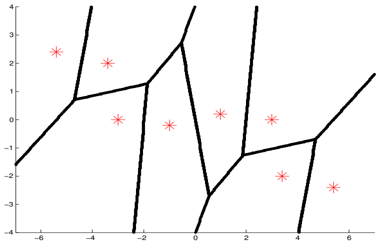

Figure 1: The mixture of eight 2D Gaussian sources used to evaluate our proposed techniques. Misclassification was 27.4% even when unquantized features were used, indicating significant overlap between the different sources.

ID	Design	Rate-allocation	Encoding
Q1	MSE	MSE	MSE
Q2	MSE	MI	MSE
Q3	MI	MI	MSE
Q4	MI	MI	MI

Table 1: In the different quantization techniques, we varied the design algorithm, the rate-allocation technique and the encoding scheme. Here MSE refers to mean square error distortion and MI refers to our proposed mutual information loss distortion. Note that we progressively increase the significance of the role MI plays in the quantizer operation, by first using it only for rate-allocation, then for both rate-allocation and quantizer design and finally for all three operations.

The different quantization techniques (Q1-Q4) evaluated are listed in Table 1. To illustrate the effect of MI on the different aspects of the quantizer, we progressively increase the significance the role MI plays in the quantizer operation. Q4 represents the quantization technique incorporating MI in all three phases, design, rate-allocation and encoding. This represents our best system. Figure 2 shows the results obtained when the different quantizers were used for encoding the data before classification. We plot the increase in misclassification vs. bits-per-sample. The increase in misclassification is with respect to the baseline performance using unquantized data. Observe that at high bitrate ($>$ 3.5 bits-per-sample), the performance of all systems is almost the same, although Q4 achieves the best results of the four techniques. However at low bit rates the sub-optimality of MSE for classification becomes clear. When the GBFOS algorithm uses rate vs. mutual information loss points instead of rate vs. MSE points to allocate rates to the different dimension of the vector we observe that the performance improves or stays the same (Q2 is better than Q1). Once MI is also included for quantizer design (Q3), we observe substantial improvement in performance. At 2.3 bits-per-sample the increase in misclassification due to Q1 is 7.6%, while due to Q3 is only

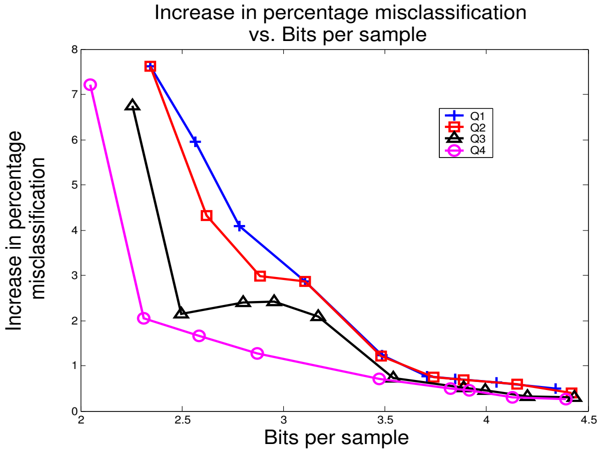

Figure 2: The results obtained by the different quantization schemes. Observe that as MI is increasingly used in the quantizer, the rate-classification performance always becomes better.

2.2%. Notice that this improvement requires no extra cost during encoding, i.e., Q3 still uses the non-optimal MSE encoding, thus having no extra run time computational increase. However we observe that between 2.5 and 3.5 bits-per-sample, the performance of Q3 is not monotonic. This is due to the fact that a quantizer designed to minimize loss in MI, is used by an encoding scheme which minimizes MSE. To eliminate this mismatch MI can be used as the criterion during actual encoding (Q4). Observe that Q4 outperforms Q1,Q2 and Q3 at all bitrates, and additionally eliminates the non-monotonicity of Q3 (however the run time encoding complexity increases).

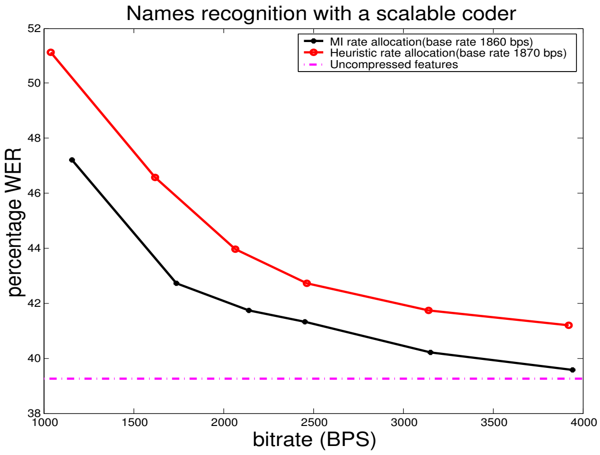

Figure 3: WER in spoken names recognition task. Significant performance improvements is achieved at all bitrates by the MI rate-allocation compared to the heuristic rate-allocation. At 3920 bps, the MI rate-allocation scheme results in only a 0.31% increase in WER when compared to using unquantized MFCCs.

The significant performance improvements achieved for this eight class task motivated us to apply the MMIL technique to the problem of encoding speech in a distributed speech recognition (DSR) task. Our scalable DSR encoder [1] uses uniform scalar quantizers, hence we concentrate on rate-allocation. The rate-allocation to each of the mel frequency cepstral coefficients (MFCCs) was based on the importance of each MFCC for recognition [7]. It is unclear how to map importance to rate-allocation. However loss in MI is a better metric for allocating the rates, since the loss in MI indicates the increase in class uncertainty. The results obtained when rate-allocation to the MFCCs was done with MI is shown in Figure 3. The recognition task was a two stage spoken names recognition (see [1] for details). Observe that at all bitrates the MI rate-allocated encoder significantly outperforms the heuristically rate-allocated encoder. At 3920 bps the MI rate-allocation reduces the increase in WER due to compression from 1.92% for the heuristic rate-allocation to 0.31%, more than a six fold decrease. In terms of bitrate, the MI rate-allocation achieves at 2500 bps the same recognition performance as the heuristic rate-allocation does at 3920 bps, a 36% reduction in bitrate with no penalty in recognition performance.

4. Conclusions

We proposed mutual information, an alternative to MSE, as a more appropriate distortion criteria for encoder design in classification (recognition) applications. We showed that KL distance is the right optimization measure to ensure that the quantizer design minimizes loss in MI between the data samples and class labels. We provided a practical empirical quantizer design technique. We also showed that rate-allocation based on MI can result in significant performance improvements in distributed speech recognition.

5. References

[1] N. Srinivasamurthy, A. Ortega, and S. Narayanan, "Efficient scalable encoding for distributed speech recognition," *IEEE Transactions on Speech and Audio Processing*, Submitted Jan 2003.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley Series in Telecommunications. John Wiley & Sons, 1991.

[3] K. Oehler and R. Gray, "Combining image compression and classification using vector quantization," *IEEE Trans. Pattern Analysis and Machine Learning*, vol. 17, pp. 461–473, May 1995.

[4] N. Srinivasamurthy and A. Ortega, "Reduced complexity quantization under classification constraints," in *Proc. IEEE Data Compression Conference (DCC)* (J. A. Storer and M. Cohn, eds.), pp. 402–411, 2002.

[5] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy constrained vector quantization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-37, pp. 31–42, January 1989.

[6] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

[7] N. Srinivasamurthy, *Efficient Compression Algorithms for Distributed Classfication Applications*. PhD thesis, Dept. of Electrical Engineering-Systems, University of Southern California, 2003.

[8] Y. Normandin, R. Cardin, and R. D. Mori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, 1994.

[9] J. Shore and R. M. Gray, "Minimum cross-entropy pattern classification and cluster analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-4, pp. 11–17, January 1982.

[10] E. A. Riskin, "Optimum bit allocation via the generalized BFOS algorithm," *IEEE Transactions on Information Theory*, vol. 37, pp. 400–402, March 1991.