ENHANCED STANDARD COMPLIANT DISTRIBUTED SPEECH RECOGNITION (*AURORA* ENCODER) USING RATE ALLOCATION

Naveen Srinivasamurthy, Antonio Ortega and Shrikanth Narayanan

Department of Electrical Engineering-Systems,
Integrated Media Systems Center,
University of Southern California, Los Angeles.

[snaveen,ortega,shri]@sipi.usc.edu

ABSTRACT

This paper proposes modifications to improve the recognition performance obtainable by the ETSI standard distributed speech recognition encoder, *Aurora* [1]. The proposed modifications are *standard compliant*, i.e., they require no algorithmic modifications to the *Aurora* operation. Performance improvements are achieved by more efficiently distributing the available bit budget among the seven (different) 2-dimension vector quantizers (VQs), used by *Aurora*. Improved bit-allocation to the different sub-vectors used in *Aurora* is achieved by incorporating the importance towards recognition of each of the sub-vectors into the bit-allocation algorithm. The available bits are efficiently distributed among the sub-vectors by allocating a larger fraction of the available bits to the more important sub-vectors and hence maximizing recognition accuracy. The proposed bit-allocation algorithm is based on a novel *mutual information (MI)* measure. The MI measure quantifies the information content between a sub-vector and the class label and hence is a good indicator of importance of the coefficient towards recognition. It is shown that the proposed MI based method outperforms both the standard *Aurora* encoder and an encoder designed using traditional mean square error based bit-allocation. For the TIDIGITS connected digits recognition task a 15.2% relative decrease in word error rate (WER) was possible with the proposed modified MI based *Aurora* encoder when compared to the recognition performance achieved using the standard *Aurora* encoder.

1. INTRODUCTION

In distributed speech recognition (DSR) [2], 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 vectors used by the recognizer are extracted, compressed (to conserve bandwidth) and transmitted. High dimensionality of the feature vector requires, for computational reasons, that each vector of the speech feature be quantized with scalar or split-vector quantizers. ETSI has standardized *Aurora* [1], an encoder incorporating a split VQ architecture (7 VQs/sub-vectors are used) for encoding speech feature vectors in DSR applications. The *Aurora* standard allocates eight bits for the first sub-vector and six bits for each of the other 6 sub-vectors (see Table 1).

It is well known that the lower feature coefficients are more "important" for recognition than the higher coefficients [3]. Therefore given a bit budget constraint (the constraint could be due to bandwidth/power limitations), to ensure that the degradation in recognition performance due to compression is minimized it is vital that a larger fraction of the bits be allocated to the more important coefficients. This enables coefficients important for recognition to be represented with higher fidelity than the less important coefficients. *Aurora* clearly does not make use of the coefficient importance, since it allocates the same number of bits to both the lower and higher feature sub-vectors.

One of the challenges in achieving an optimal bit-allocation is that it is not straight-forward to translate coefficient importance into actual bit requirements (for the different coefficients). Traditionally, bit-allocation techniques rely on the mean square error (MSE) metric to allocate bits to the different coefficients. This does ensure that the MSE between the original vector and the compressed vector is minimized. However, it is not clear if MSE is the right optimization criteria in recognition (classification) applications [4]. Consider the example in Figure 1. It is obvious that dimension 2 has significantly more energy than dimension 1. However, the class overlap along dimension 1 is significantly lower than in dimension 2. An MSE based bit-allocation technique would allocate *more bits* to the coefficient along dimension 2 which is the *less important* coefficient for classification. However, for classification it is more desirable to represent the coefficient along dimension 1 more accurately to ensure minimal degradation in classification performance. Hence there is a necessity for using a new distortion criteria which correlates better with the importance of the coefficient towards recognition than was possible using the MSE metric.

In this paper, we focus on the *information-theoretic* measure of mutual information (MI) [5] to define a sub-vector based distortion measure *mutual information loss (MIL)*. We propose a new bit-allocation technique which incorporates the MIL distortion and

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.

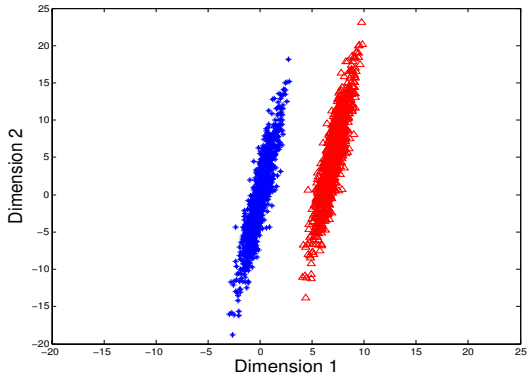

Fig. 1. A two class classification task. Observe that the signal energy is significantly higher along dimension 2. However the class separation (discriminability) is clearly better along the low signal energy dimension 1. It is clear that coefficients with higher signal energy are in general not necessarily the more important coefficients for classification.

allocates bits to different components of the feature vector to ensure that the encoded data retains as much information about the class labels as possible. The performance benefits of our proposed MIL distortion based bit-allocation is demonstrated by applying it to *Aurora*. Additionally, the proposed modifications are *standard compliant*, i.e., they do not require modifications to either the encoder or decoder operations (except that different VQ codebooks need to be used). Furthermore, we also show that MIL bit-allocation is significantly better than the traditional MSE bit-allocation for recognition applications.

For the TIDIGITS digits recognition task using our proposed MIL bit-allocation, a 15.2% relative reduction in WER was achieved compared to the standard *Aurora* encoder. A 12% relative reduction was achieved compared to a MSE bit-allocation. Similar recognition performance improvements can be expected for more complex recognition tasks.

The paper is organized as follows. Section 2 describes the *Aurora* encoder and the MI based bit-allocation algorithm used to improve it. Experiments and results are presented in Section 3. Finally conclusions and future work is given in Section 4.

2. *AURORA* OVERVIEW

2.1. *Aurora* Operation

In the *Aurora* standard the speech features used for recognition are the first 12 MFCCs: c1-c12; the zeroth cepstral coefficient (c0) and the energy in the frame (E). The 14-dim feature vector is split into seven 2-dim sub-vectors. Each of the sub-vectors is encoded with a different 2-dim VQ. The standard computes a feature vector every 10ms and allocates 44 bits to each feature vector to achieve a total bitrate of 4.4 kbps[1]. The number of bits allocated to the different sub-vectors are shown in Table 1. Notice that 8 bits are allocated to the (c0,E) sub-vector and 6 bits are allocated to each of the other 6 sub-vectors. It is not clear if the bit-allocation in Table 1 is optimal. For example, it is well known that the sub-vector (c1,c2) is more important for recognition than the sub-vector (c11,c12) [3]. Hence allocating the same number of bits for both these sub-vectors ignores the different contributions of the sub-vectors towards recognition performance. Hence it can be expected that optimizing the bit-allocation by incorporating the importance of the sub-vectors can yield improved recognition performance.

[1] 4 additional bits are used for channel coding

Sub-vector	Bits allocated
c0,E	8
c1,c2	6
c3,c4	6
c5,c6	6
c7,c8	6
c9,c10	6
c11,c12	6

Table 1. The sub-vectors used by *Aurora* encoder. The number of bits allocated to each sub-vector is not necessarily reflective of the importance of the sub-vector towards recognition. While it is well know that lower sub-vectors contribute more towards recognition, they have been allocated the same number of bits as higher sub-vectors.

2.2. Modified *Aurora* with Mutual Information Based Rate Allocation

It is possible to improve the performance of *Aurora* (or other DSR encoders) (i) by using a modified encoder operation based on VQs optimized for a distortion measure other than MSE [6], or (ii) by the use of entropy encoders to reduce the bitrate. However these modifications will not be standard compliant. In this paper we concentrate on the bit-allocation problem. The advantage of choosing bit-allocation is that it only requires that the VQ codebooks be changed. Therefore there will be no algorithmic change to either the encoder or the decoder. The encoder can indicate to the decoder the codebook it has used, thus enabling the decoder to choose the right codebook while decoding the transmitted data.

Traditionally, bit-allocation has been performed to minimize the MSE metric. However as mentioned before MSE is not the right distortion criteria for use in classification tasks. The objective of the recognizer is to identify the class (phonemes, words, sentences) from the (speech) feature vectors. Each of the feature vectors carries information about the class it belong to, and it is this information that is required by the recognizer. Hence it seems intuitive that the goal of the encoder should be to compress data such that the compressed data retains maximal information about the class labels.

Mutual information (MI) between the class labels and feature vectors is given by

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

Fine quantization was used to find an empirical estimate $\tilde{p}(c|\mathbf{x})$ of $p(c|\mathbf{x})$ from labeled training data [7]. Fine quantization was so chosen to ensure that recognition with the finely quantized data

introduced negligible degradation to the recognition performance. MI provides a quantitative method of identifying the information carried by the feature vectors about the class labels. Table 2 shows the MI between the seven sub-vectors in *Aurora* and phoneme class labels. We can observe that lower cepstral coefficients carry significantly more information than the higher cepstral coefficients. This fact can be used by bit-allocation algorithms to optimally allocate bits to the different sub-vectors so as to maximize the mutual information between the quantized data and the class labels for a given bit budget.

Sub-vector	Mutual information (in bits)
c0,E	0.94
c1,c2	1.30
c3,c4	0.86
c5,c6	0.57
c7,c8	0.47
c9,c10	0.44
c11,c12	0.33

Table 2. The mutual information between the feature sub-vectors and the phoneme classes. Observe that the lower sub-vectors carry significantly more information about the phoneme classes when compared to the higher sub-vectors.

Let $\mathbf{X} = [\mathbf{X_1}, \ldots, \mathbf{X_N}]$ be a feature vector, where $\mathbf{X_i}$ is the i^{th} sub-vector of $\mathbf{X}$. The generalized Breiman, Friedman, Olshen, and Stone (GBFOS) algorithm [8] allocates bits $B_i, i = 1, \ldots, N; s.t. \sum_i B_i \leq B$, to each of the components, where B is the total bit budget. In traditional MSE bit-allocation, rate vs distortion (MSE) points are calculated for each of the components. Then the combination of points which satisfy the bit budget while yielding the minimum MSE are selected. For MIL quantizers the distortion is MI loss, hence several *rate vs mutual information loss* points are calculated for each of the components. Then given the bit budget B, these calculated points are used by the GBFOS algorithm to allocate bits, $B_i, i = 1, \ldots, N; s.t. \sum_i B_i \leq B$, to each of the components, so that the MI loss $MIL = \sum_{i=1}^{N} I(\mathbf{X_i}; C) - I(\hat{\mathbf{X}}_\mathbf{i}; C)$, is minimized, where $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_\mathbf{1}, \ldots, \hat{\mathbf{X}}_\mathbf{n}]$ is the quantized vector and

$$I(\mathbf{X_i}; C) - I(\hat{\mathbf{X}}_i; C) = \int_{\mathbf{x}} f(\mathbf{x_i}) \sum_c p(c|\mathbf{x_i}) log \left(\frac{p(c|\mathbf{x_i})}{p(c|\hat{\mathbf{x}}_\mathbf{i})} \right) d\mathbf{x} \tag{2}$$

where $p(c|\mathbf{x_i})$ is the probability of the class label given the i^{th} feature sub-vector, $p(c|\hat{\mathbf{x}}_\mathbf{i})$ is the probability of the class label given the i^{th} quantized feature sub-vector and $f(\mathbf{x_i})$ is the *pdf* of the i^{th} feature sub-vector. From the Markov chain $C \leftrightarrow \mathbf{X_i} \leftrightarrow \hat{\mathbf{X}}_\mathbf{i}$, we can find

$$p(c|\hat{\mathbf{x}}_\mathbf{i}) = \int_{\mathbf{x_i}} p(c|\mathbf{x_i}) p(\mathbf{x_i}|\hat{\mathbf{x}}_\mathbf{i}) d\mathbf{x_i} \tag{3}$$

The modified GBFOS algorithm is summarized below.

Algorithm 1 *(Modified GBFOS Algorithm)*
Step 1 : *For $n = 1, \ldots, N$, set $B_n = q$. This is the initial bit allocation.*
Step 2 : *Calculate, for $n = 1, \ldots, N$, for $i = 1, \ldots, B_n$,*

$$\begin{aligned} S_n(B_n, B_n - i) &= \frac{\Delta MIL_{overall}}{\Delta B_{overall}} \\ &= -\frac{MIL_n(B_n) - MIL_n(B_n - i)}{i} \end{aligned} \tag{4}$$

Step 3 : *For each $n = 1, \ldots, N$, determine i for which $S_n(B_n, B_n - i)$is minimized.*
Step 4 : *Determine the component for which $S_n(B_n, B_n - i)$ is the lowest. Assume it is component l. (If minimum $S_n(B_n, B_n - i)$ is not unique, then select all components with this value). Set $B_l = B_l - i$.*
Step 5 : *Calculate $B_{alloc} = \sum_n B_n$. Check if $B_{alloc} \leq B$; if so stop.*
Step 6 : *Repeat Steps (2), (3), (4) and (5).*

The bit-allocations obtained by our method is shown in Table 3. Observe that our bit-allocation allocates 7 bits for the sub-vector (c1,c2), while only 4 bits are allocated for the (less important) sub-vector (c11,c12). The Table also shows the bit-allocation obtained by using the MSE bit-allocation algorithm.

Sub-vector	Bits allocated		
	Aurora	MSE	MIL
c0,E	8	8	11
c1,c2	6	9	7
c3,c4	6	7	6
c5,c6	6	6	6
c7,c8	6	5	5
c9,c10	6	5	5
c11,c12	6	4	4
Total	44	44	44

Table 3. The bit-allocations for the different encoders. Notice that in each case the total number of bits allocated are 44. Also notice both the MSE and MIL bit-allocations tend to reduce the bits for the higher sub-vectors and allocate these to the lower sub-vectors. The MSE bit-allocation optimizes the allocation to ensure that MSE is minimized. However, the MIL bit-allocation ensures that the information contained in the quantized data about the phoneme classes is maximized.

3. EXPERIMENTS AND RESULTS

The experiments were carried out on the TIDIGITS corpus using HTK 3.2 speech recognizer, with MFCCs extracted using the *Aurora* front end. Only the coefficients c1 to c12 and E were used during recognition (the zeroth coefficient (c0) was not used[2]). Only the utterances from the male and female speakers were used. The database consists of variable length connected digit utterances (1 to 7 digits per utterance). The models on the server were initially trained using clean speech from the "train" part of the database

[2]The speech recognizer used HTK 3.2, does not support the simultaneous use of both c0 and E, hence c0 was not used during recognition

(8623 utterances). Each digit was modeled using a 10 stage HMM with 16 GMMs per state. A silence model was used before and after the digit utterance to take care of the pre and post utterance silence. In addition a short pause model was used to account for inter-digit short pauses. The testing (using utterances from the "test" part of the database (8700 utterances)) was carried out using (i) unquantized features (ii) *Aurora* encoded features (iii) MSE bit-allocated *Aurora* encoded features (Aurora-MSE) (iv) proposed MIL bit-allocated *Aurora* encoded features (Aurora-MIL). Only the MFCCs were computed and encoded. The Δ and $\Delta\Delta$ coefficients were computed from the encoded MFCCs.

The MI between the features (both clean and encoded) were computed using the TIMIT database. The class labels were the 45 phonemes as defined in TIMIT. This computed information was used to perform bit-allocation for the TIDIGITS recognition task. It can be expected that better performance can be achieved if the MI was computed using the TIDIGITS database, with the class labels being the different digits. However, we choose not to do so to indicate the generality of our proposed approach, i.e., the proposed approach is not task dependent.

First, observe from Table 4 that the MI between the quantized data and the phone classes is 3.86 bits for the Aurora-MIL encoder while it is 3.76 bits and 3.78 bits for the standard *Aurora* encoder and the Aurora-MSE encoder respectively. It can be expected that this increased MI in the compressed data will translate into superior recognition performance. Looking at the recognition results in Table 4 obtained for the different cases, we observe that this is indeed true. Observe that with the *Aurora* encoder there is a 53.5% degradation in WER compared to using clean speech. However when MIL bit-allocation was incorporated into *Aurora* the degradation dropped to 30.2%, i.e., a 23.5% relative or 0.2% absolute reduction in WER. Also observe that the MIL bit-allocation clearly outperforms the MSE bit-allocation, indicating that the MIL measure is better than the MSE metric for recognition applications. It can be expected that similar advantages can be achieved for other recognition tasks and scenarios when noisy speech is used during recognition.

Encoder	Percentage WER	Percentage degradation	Total MI (in bits)
Clean	0.86	-	4.91
Aurora	1.32	53.5%	3.76
Aurora-MSE	1.27	47.7%	3.78
Aurora-MIL	1.12	30.2%	3.86

Table 4. WER for TIDIGITS recognition task for the different encoders. Clearly the proposed MIL bit-allocated Aurora-MIL encoder significantly outperforms both the other encoders, achieving 23.3% relative WER reduction over the standard *Aurora* encoder. Also notice as expected the MI between the encoded data and the phoneme classes is maximum for Aurora-MIL when compared to the other encoders.

4. CONCLUSIONS AND FUTURE WORK

We proposed a mutual information loss based bit-allocation to improve the performance of *Aurora*. The proposed modifications were standard compliant, requiring no algorithmic changes to either the encoder or decoder operations. It was shown that the proposed MIL bit-allocated encoder outperformed both the standard *Aurora* encoder and a traditional MSE based bit-allocation. One of the contributions of this paper was demonstrating that the use of *application tailored* distortion measures can yield significant performance improvements when compared to using traditional MSE metrics.

While in this paper we concentrated on standard compliant modifications, further work will involve incorporating mutual information into both the design and encoding operations to enable further performance improvements. Also, the proposed techniques will be evaluated for other recognition tasks, involving continuous speech recognition and recognition tasks with noisy databases. Finally, to ensure standard compliance, we did not investigate grouping of coefficients to form sub-vectors. Further work will involve selection of both size and components to be used to construct optimal sub-vectors.

5. REFERENCES

[1] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," Tech. Rep. Standard ES 201 108, European Telecommunications Standards Institute (ETSI), April 11 2000.

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

[3] E. Bocchieri and J. Wilpon, "Discriminative feature selection for speech recognition," in *Computer Speech and Language*, vol. 7, pp. 229–246, 1993.

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

[6] N. Srinivasamurthy, A. Ortega, and S. Narayanan, "Towards optimal encoding for classification with applications to distributed speech recognition," in *Eurospeech 2003*, (Geneva, Switzerland), September 2003.

[7] N. Srinivasamurthy, *Compression Algorithms for Distributed Classification with Applications to Distributed Speech Recognition*. PhD thesis, Dept. of Electrical Engineering-Systems, University of Southern California, 2003.

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