Histogram-based Quantization (HQ) for Robust and Scalable Distributed Speech Recognition

Chia-yu Wan and Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University, Taiwan, Republic of China

Abstract

The performance of conventional distance-based vector quantization (VQ) for distributed speech recognition (DSR) is inevitably degraded by the environmental noise and quantization distortion. The pre-trained codebook is less scalable and may not be matched with the testing speech. A new concept of Histogram-based Quantization (HQ) is proposed in this paper, in which the quantization levels are dynamically defined by the histogram or order statistics of a segment of local most recent past samples of the parameter to be quantized. All problems with a pre-trained codebook is automatically solved because the pre-trained codebook is not used at all. The computation requirement is low because no distance measure is needed. The approach is robust because most disturbances (including very non-stationary types) can be absorbed by the dynamic histogram. Extensive experiments with AURORA 2.0 testing environment indicated the new approach is highly robust and scalable, suitable for future personalized and context-aware DSR environment.

1. Introduction

The various applications of the automatic speech recognition (ASR) technologies have been highly anticipated by many people[1]. The client-server architecture for distributed speech recognition (DSR) is very attractive. One of the approaches very often considered is to extract recognition-oriented features locally at the client, then compress and transmit them across the network to the server for recognition. The subject of this paper is for this approach.

Various schemes for compression of ASR features have been proposed in recent years. Distance-based vector quantization (VQ) has been found very useful. Very good results were obtained for clean speech and/or matched vector quantization (VQ) codebook conditions [2][3]. But environmental noise and quantization distortion may jointly degrade the recognition performance. Quantization process may further increase the distance between clean and noisy features, and the environmental noise may move the feature vectors to a different quantization cell too. So the performance can be seriously deteriorated under lower SNR conditions. In addition, the quantization codebook is usually optimized based on a training corpus. The mismatch between the VQ codebook and the testing feature vectors certainly further increase the distortion.

Differential encoding of transformed coefficients have been shown to be very helpful in making quantized features less sensitive to environmental variations [4]. Quantization in the transformed domain can efficiently improve the desired robustness for feature vectors under environmental disturbances [5][6][7]. Use of robust feature extraction approaches before quantization is also helpful in reducing the mismatch and VQ distortion, especially at lower SNR conditions [6].

In this paper, a new concept of Histogram-based Quantization (HQ) is proposed, in which the quantization levels are dynamically defined by the histogram or order statistics of a segment of local most recent past samples of the parameter to be quantized. All problems with a pre-trained codebook is automatically solved because the pre-trained codebook is not used at all. The approach is robust because most disturbances (including very non-stationary types) can be absorbed by the dynamic histogram. It is also highly scalable because it is simply based on a look-up table of local order statistics. All these advantages were verified by extensive experiments to be presented below, after the basic formulation.

2. Histogram-based Quantization (HQ)

2.1 Basic formulation

The concept of Histogram-based Quantization (HQ) is to perform the quantization process based on the histogram or order-statistics of a feature parameter within a progressively moving segment of T frames, \([x_{t-T+1}, x_{t-T+2}, ..., x_{t}, x_{t+1}]]\), up to the time index \(t\) being considered. As shown in Figure 1, the values of these T parameters in \(X_{t-T}\) are sorted to produce a cumulative distribution function \(C(y)\), or histogram, where \(y_0\) and \(y_N\) are respectively the minimum and maximum values within \(X_{t-T}\). The quantization levels, \(\{D_i\}_{i=1,...,N}\), on the other hand, are actually defined on the vertical scale of \(C(y)\) between 0 and 1, although these levels are practically transformed back to the range of the values of the parameters in \(X_{t-T}\) on the horizontal scale by \(C^{-1}(y)\). For example, the quantization level \(D_i\) with upper and lower boundaries \(b_i\) and \(b_{i-1}\) on the vertical scale, are for those values between \(y_i\) and \(y_{i-1}\), where \(C(y_i) = b_i\) and \(C(y_{i-1}) = b_{i-1}\). So \(\{D_i\}_{i=1,...,N}\) are all the quantization levels constructed from the T parameters in \(X_{t-T}\) accumulated up to time \(t\) to be used for the quantization of the parameter \(x_t\) at time \(t\), where \(N\) is the total number of quantization levels. In other words, the quantization level \(q_i\) for the present parameter \(x_i\) being considered is then

\[
q_i = D_i, \quad \text{if} \quad b_{i-1} < C(x_i) < b_i,
\]

or \(y_{i-1} < x_i < y_i, i=1,2,...,N\).

The decision boundaries for the quantization levels, \(\{b_i\}_{i=0,1,...,N}\) on the vertical scale, where \(b_0=0\), \(b_N=1\), can be either uniformly or non-uniformly distributed on the vertical scale. But the relationships between the vertical and horizontal range of a quantization level \(D_i\) \([b_i, b_{i-1}]\) and \([y_i, y_{i-1}]\) are simply dynamically defined by \(C(y)\). This is referred to as Histogram-based Quantization (HQ) here in this paper.
The first nice feature of this new quantization scheme is that no codebook training is needed at all, but instead the quantization is based on the order-statistics of the most recent past parameter values. Therefore it is dynamic based on local statistics only, thus there is no such problem as mismatch between training corpus and testing data. The computation requirement is also minimum, because all needed is to sort the T most recent past values. The quantization level \( q_i \) for \( x_i \) is then readily available by a table look-up. Note that we do not need to sort the T parameters each time to obtain the new order statistics [8]. Each time we only need to delete one parameter and insert the new one into the previous order statistics to obtain the new order statistics to perform the quantization for the new parameter. There is no processing delay either, because quantization is based on previous values.

The second nice feature of this new quantization scheme is the inherent robust nature. When a segment of parameters \( X_{t,T} \) are corrupted by small disturbances, the order statistics may remain similar although all individual values can be changed. For example, as shown in Figure 1, with the disturbances, \( y_0 \) may be changed to \( y'_0 \), \( y_N \) to \( y'_N \) and \( C(y) \) changed to \( C'(y') \). The quantization for the new disturbed parameter \( x_i' \) is now

\[
q_i = D_i, \text{ if } b_{i-1} < C'(x'_i) < b_i,
\]

or

\[
q_i = D_i', \text{ if } b_{i-1} < C'(x'_i) < b_i',
\]

where \( y'_i \) and \( y'_i \) are the boundaries after the disturbances, i.e. \( C'(y'_i) = b_{i-1}, C'(y'_i) = b_i \). It can be imagined that with a good probability the quantized level \( D_i \) may remain unchanged. In other words, the quantization is based on the levels \( D_i \) in the vertical scale and the histogram \( C(y) \), therefore less sensitive to the disturbances on the horizontal scale, or the disturbances on the horizontal scale is kind of "ABSORBED" by the histogram \( C(y) \). Such robustness is based on local statistics within the most recent past values, therefore it is dynamic and can even handle non-stationary disturbances as well to a good degree. This will be verified later on in the experiment.

### 2.2 Non-uniform Histogram-based Quantization (NHQ)

It was mentioned previously the boundaries \( \{y_i, i=0,1, ..., N\} \) on the vertical scale for the quantization levels can be either uniformly or non-uniformly distributed. However it is reasonable to believe that some non-uniform distribution based on some assumptions on the statistics of the values for the parameters in \( X_{t,T} \) may be helpful. For example, we may assume the distribution of the value for the parameters in \( X_{t,T} \) is zero-mean Gaussian (or any other reasonable statistics). We can then derive the boundaries for the quantization levels on the horizontal scale, \( \{y_i, i=0,1, ..., N\} \), using a well-known quantization algorithm such as the Lloyd-Max algorithm[9], and then transform the obtained boundaries, \( \{y_i, i=0,1, ..., N\} \) on the horizontal scale back to the boundaries on the vertical scale, \( \{b_i, i=0,1, ..., N\} \), based on \( C(y) \) of the assumed statistics (e.g. zero-mean Gaussian). In this way, the boundaries \( \{b_i, i=0,1, ..., N\} \) on the vertical scale are non-uniform distributed, but carry the statistics helpful to quantization. This is referred to as Non-uniform Histogram-based Quantization (NHQ) here in this paper.

### 2.3 Histogram-based Vector Quantization (HVQ)

The above Histogram-based Quantization (HQ) can be easily extended to vector quantization. Considering the Split Vector Quantization (SVQ) popularly used for distributed speech recognition (DSR) as an example, two MFCC parameters, such as \( c_1 \) and \( c_2 \), can be quantized together by a two-dimensional VQ codebook. Extended from the Non-uniform Histogram-based Quantization (NHQ) mentioned above in section 2.2, progressively moving segment of T values of \( c_i \) up to time \( t \), \( X_t^{(1)} \), gives a histogram \( C_1(y_1) \) for \( c_1 \), an a similar segment of T values of \( c_2 \) up to time \( t \), \( X_t^{(2)} \), gives another histogram \( C_2(y_2) \) for \( c_2 \), and they jointly produce a two-dimensional distribution for \( (c_1, c_2) \). Using the Non-uniform Histogram-based Quantization (NHQ) concept presented above, LBG VQ algorithm can be performed on the \( (c_1, c_2) \) plane based on some assumed statistics, such as a bi-variable zero-mean Gaussian (or any other reasonable statistics). For the present parameter pair \( (c_1, c_2) \) at time \( t \), the inverse function \( C^{-1}_c[\bullet, \bullet] \) of the two-dimensional histogram \( C_2[\bullet, \bullet] \) for the assumed two-dimensional statistics used for performing LBG VQ produces the pair \( (x_1^{(1)}, x_2^{(2)}) \), where \( C_2^{-1}[C_1(c_1), C_2(c_2)] = (x_1^{(1)}, x_2^{(2)}) \), whose location on the \( (c_1, c_2) \) plane mapped to the two-dimensional LBG VQ partition cells previously obtained then defines the quantization cell for \( (c_1, c_2) \). In this way, two-dimensional VQ can be easily performed on \( (c_1, c_2) \) based on the local past order statistics of \( X_{t,T}^{(1)} \) and \( X_{t,T}^{(2)} \) respectively just as in one-dimensional case. This is referred to as Histogram-based Vector Quantization (HVQ) here in this paper. This HVQ can be easily extended to higher dimension without increasing the computation complexity, because it is not needed to compute the distance measure. All that needed is the order statistics of the parameters.

With the HVQ concept in mind, the special feature of this new quantization approach can be better visualized as shown in Figure 2. The distribution of \( (c_1, c_2) \) of the testing data may be quite different from that of the VQ training corpus which causes the primary difficulties in the conventional distance-based quantization approaches. In the proposed HQ approach,
however, we no longer rely on the distance on the \((c_1, c_2)\) plane, but instead we let the quantization codebook (or look-up table) move with the testing data, because the quantization is now based on the distribution or histogram on the vertical scale. As can be found in Figure 2, the shift of vectors \((c_1, c_2)\) due to various reasons becomes almost irrelevant to the quantization process.

### 2.4 Scalability

To fully realize the concept of distributed speech recognition in real-world application, adjustable transmission bit rate considering the varying channel condition or network bandwidth is necessary. In the conventional distance-based quantization a codebook has to be stored, which is trained under pre-defined codebook size with fixed bit rate. If the bit rate needs to be changed, retraining the codebook with different codebook size and bit allocation among different parameters is required. It may not be practical to keep a whole set of many codebooks with different codebook size in the client and select the codebook in real-time on demand. Additional bandwidth is also needed for identifying the selected codebook. So the scenario is in general quite difficult for the conventional distance-based quantization to scale the bit rate considering the time-varying channel conditions.

The proposed HQ approach is, on the other hand, completely scalable, because HQ does not need any pre-trained codebook to be stored in the client. Only one look-up table for local order statistics with sufficient quantization resolution kept in the client will be enough. For example, assume a 64-level quantization look-up table is stored in the client. If we need to reduce the number of bits from 6 to 4 bits under poor channel conditions, we can simply transmit the 4 most significant bits and delete the 2 less significant bits. Similarly, with the proposed HQ it is also easy to use different channel coding schemes to protect different bits with different significance.

### 3. Experimental conditions

The experiments reported in this paper were performed on the AURORA 2.0 testing environment. The training/testing corpus consist of clean and noisy connected English digits. Ten different types of noise are considered, as representatives of real-world noises. There are two training conditions (clean-condition training and multi-condition training) and three sets of testing conditions (sets A, B, and C) defined in AURORA 2.0. In each case SNR from 20dB to 0dB were tested. The MFCC extraction follows the W007 front-end, which gives 13 coefficients \((C1\sim C12\text{ and log energy})\) plus the delta and delta-delta features for recognition.

### 4. Experimental results

#### 4.1 General observations

The recognition accuracies for baseline experiments with original MFCC features without quantization and compression, compared to those with standard SVQ at 4.4kbps[10] and recently proposed 2DDCT transform encoding at 1.45kbps[7] are listed in the first three rows (a)(b)(c) of Table 1 for clean-condition training and testing sets A, B, C respectively. Because they are averages over all SNR values from 20dB down to 0dB, the results here are not very high. The results for the proposed Histogram-based Vector Quantization (HVQ) are then listed in rows (d)-(h) of Table 1 for bit rates 1.9kbps all the way up to 3.9kbps. In each case of bit rate the best choice of quantization configuration was used, including 3-dimensional HVQ for rows (d)(e) and 2-dimensional HVQ for rows (f)(g)(h). It seemed that HVQ was able to achieve the best choice of bit rate (in rows (f)), because higher performance was not necessarily achievable if higher bit rates in (rows (g)(h)) were used. Note that the performance in all cases in rows (d)-(h) were consistently significantly better than rows (a)-(c). For example, the overall accuracy in row (f) (82.08%) represented relative error rate reduction of 55%, 58%, or 54% as compared to those with 2DDCT at 1.45kbps (59.89% in row(c)), SVQ at 4.4kbps (56.51% in row (b)), or the original MFCC (61.08% in row(a)). This was apparently due to the robust nature of HQ as discussed previously. Note that the original MFCC without quantization also performed relatively poor under such noisy environment as tested here, because the MFCC parameters were also seriously disturbed by the various types of noise. The HQ or HVQ approaches as proposed here, however, reconstructed the feature parameters based on the histogram, which apparently automatically ABSORBED many of the disturbances, therefore offered much better recognition accuracy.

The results in Table 1 are averaged over all SNR values and all noise types in sets A, B, C. Figure 3 then shows the similar accuracies averaged over all noise types, but separated for each SNR values, and Figure 4 shows cases (g)(h) left out due to space limitations. The other way around, accuracies averaged over all SNR values, but separated for each noise type in sets A, B, C. From Figure 3, we can found that except in the clean speech case, in which HVQ leads to a slight accuracy degradation (probably due to the minor uncertainty introduced when performing HVQ and mapping the parameter histogram to an assumed zero-mean Gaussian distribution), HVQ consistently performed very well for all SNR values. Under very poor SNR conditions, the noisy disturbances may be very serious, but still well absorbed by the histogram used here. For example, in the case of 5dB SNR, HVQ offered an accuracy of higher than 75% while the number was 30% or 40% for the other conventional approaches. HVQ offered an accuracy of higher than 50% even at 0dB of SNR. Similar results can be observed in Figure 4, in which HVQ consistently performed much better than those in rows (a)-(c). HVQ can even handle non-stationary disturbances as well to a good degree, apparently due to its dynamic nature. For example, the accuracy of 2.7kbps HVQ achieves 75% –85% for babble/restaurant/airport/train-station noise case, while the number was around 50% for the other conventional approaches.

![Table 1](image-url)
coarse quantization is enough to represent the desired information, while extra bits (or finer quantization) actually only carry irrelevant noise variation. On the other hand, 3.9kbps shows better performance than 2.7kbps/3.3kbps when SNR is 15dB, 20dB and clean test data. So more bits can be used to describe finer speech information under cleaner environment. In other words, the scalability of HQ discussed here can also be applied with respect to varying noisy conditions, in addition to bit rate constraints. Typical examples of the effect of scalable quantization (compared with the same quantization configuration) from 2.3kbps to 2.7kbps with 3-dimensional HVQ and 3.3kbps to 3.9kbps with 2-dimensional HVQ for clean speech test are listed in Table2.

For two-dimensional HVQ case, going from 3.3kbps to 3.9kbps actually achieved an 15.1% relative error rate reduction even at the very high accuracy of clean speech case. Similar situations were obtained for the 3-dimensional HVQ cases as well.

5. Conclusions

A new approach of Histogram-based Quantization (HQ) is proposed in this paper for robust and scalable distributed speech recognition (DSR). HQ has shown to be robust for all types of noise and all SNR conditions. The HQ configuration has been shown to be easily scalable based on the bandwidth or noisy conditions. For future personalized and context-aware DSR environment, the proposed HQ can be adapted to network and terminal capabilities, with recognition performance optimized based on environmental conditions.

6. References