THREE-STAGE ERROR CONCEALMENT FOR DISTRIBUTED SPEECH RECOGNITION (DSR) WITH HISTOGRAM-BASED QUANTIZATION (HQ) UNDER NOISY ENVIRONMENT

Chia-yu Wan, Yi Chen, Lin-shan Lee
Graduate Institute of Communication Engineering, National Taiwan University, Taiwan, Republic of China
chiayui@speech.ee.ntu.edu.tw, chenyi@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

In this paper, a three-stage error concealment (EC) framework based on the recently proposed Histogram-based Quantization (HQ) for Distributed Speech Recognition (DSR) is proposed, in which noisy input speech is assumed and both the transmission errors and environmental noise are considered jointly. The first stage detects the erroneous feature parameters at both the frame and subvector levels. The second stage then reconstructs the detected erroneous subvectors by MAP estimation, considering the prior speech source statistics, the channel transition probability, and the reliability of the received subvectors. The third stage then considers the uncertainty of the estimated vectors during Viterbi decoding. At each stage, the error concealment (EC) techniques properly exploit the inherent robust nature of Histogram-based Quantization (HQ). Extensive experiments with AURORA 2.0 testing environment and GPRS simulation indicated the proposed framework is able to offer significantly improved performance against a wide variety of environmental noise and transmission error conditions.

Index Terms—Speech recognition, vector quantization, robustness, error compensation

1. INTRODUCTION

The client-server framework for Distributed Speech Recognition (DSR) has been widely considered, in which speech features are extracted and compressed in the hand-held clients and the recognition performed at the server. However, the wireless networks naturally introduce transmission errors so that not all feature vectors are correctly delivered to the server, and the recognition performance is thus severely degraded.

Various error concealment (EC) techniques have been proposed in recent years in order to make DSR systems more robust against the transmission errors. They can be categorized into three groups: to reduce the transmission error rate through error detection and correction [1], to reconstruct the feature vectors by estimating the erroneous subvectors [2], and to consider the reliability of the estimated vectors at the decoding stage [3]. These methods are all very useful when the input speech is clean. But the inevitable environmental noise added to the speech entered to the clients and the transmission errors actually jointly disturb the received feature parameters in real applications. When the input speech is clean, it is possible to make up for the partial transmission errors because there are still enough correctly received feature parameters. This does not remain true if the input speech is already corrupted by environmental noise. In addition, the continuity nature or the prior statistical information of speech signals, useful in error detection with data consistency [4] or lost vectors estimation [2] may not remain useful when the input speech is noisy. This is why EC problems for noisy input speech are much more difficult.

In this paper, a three-stage EC framework based on the recently proposed Histogram-based Quantization (HQ) [5,6] is proposed, in which both the transmission errors and environmental noise are jointly considered. The first stage detects the erroneous subvectors based on the robust nature of Histogram-based Quantization (HQ) partition cells. The second stage reconstructs the erroneous subvectors using both the prior speech source statistics and the robust nature of HQ partition cells. The third stage then considers the uncertainty of the estimated subvectors during Viterbi decoding. At each stage, the robust nature of HQ is well exploited [5,6], and many problems mentioned above can be properly handled. All these advantages were verified by extensive experiments.

2. HISTOGRAM-BASED QUANTIZATION (HQ)

2.1 Basic formulation of HQ

The concept of HQ is to perform the quantization of a feature parameter x_t at time t based on the histogram or order statistics of that feature parameter within a moving segment of the most recent past T samples, $[x_{t-T+1}, \ldots, x_{t-1}, x_t] \triangleq X_{t,T}$, up to the time t being considered [5,6]. As shown in Figure 1, the values of these T parameters in $X_{t,T}$ are sorted to produce a time-varying cumulative distribution function C(y), or histogram, where $C(y_0)=b_0=0$ and $C(y_N)=b_N=1$, y_0 and y_N are respectively the minimum and maximum values within $X_{t,T}$. The N quantization levels $\{D_i = [b_{i-1}, b_i], i=1,2, \cdots,N\}$ together with their corresponding representative values $\{\bar{z}_i, i=1,2,\cdots,N\}$ defined on the vertical scale [0,1] are derived using the cumulative distribution $C_0(y)$ of a standard Gaussian N(0,1) via the Lloyd-Max algorithm. The quantization levels, $\{D_i, i=1,2,\cdots,N\}$ are then respectively transformed to the range of the feature parameter on the horizontal scale, $[y_0, y_N]$, by the histogram C(y) constructed with $X_{t,T}$, to be the N partition cells $\{[y_{i-1}, y_i], i=1,2,\cdots,N\}$ for the quantization of x_t, where $C(y_i)= b_i$. So, the partition cell $[y_{i-1}, y_i]$ on the horizontal scale is dynamic which is transformed from D_i by the time-varying histogram C(y). However the representative values $\{z_i, i=1,2, \cdots ,N\}$ for these partition cells on the horizontal scale are actually fixed, which is transformed from $\{\bar{z}_i, i=1,2,\cdots, N\}$ on the vertical scale by the standard histogram $C_0(y)$. In other words, HQ is based on a hidden codebook $\{(D_i, \bar{z}_i), i=1,2,\cdots,N\}$ on the vertical scale, which are then transformed by a dynamic histogram C(y) into time varying partition cells $[y_{i-1}, y_i]$ and by a fixed standard histogram $C_0(y)$ into the fixed representative values z_i on the horizontal scale. The quantization here is then simply mapping the present parameter x_t to a representative value z_i for the partition cell $[y_{i-1}, y_i]$,

$$x_t \rightarrow z_i, \text{ if } b_{i-1} < C(x_t) < b_i, \text{ or } y_{i-1} < x_t < y_i, i=1, 2, \ldots, N. \quad (1)$$

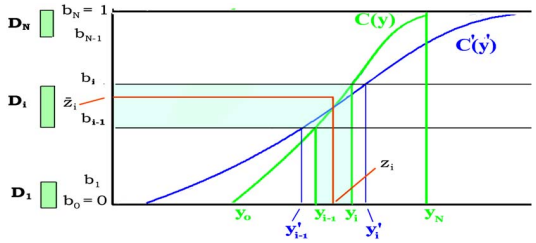

Figure 1. Histogram-based Quantization (HQ) [6]

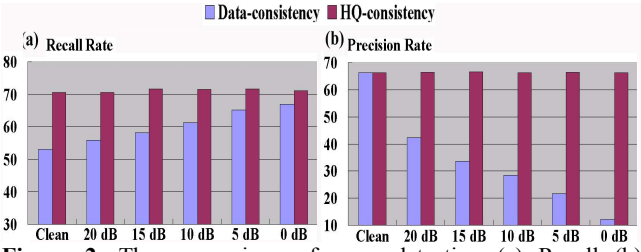

Figure 2. The comparison of error detection (a) Recall (b) Precision for the previously proposed data-consistency and the HQ-consistency proposed here

2.2 Robust nature of HQ

HQ proposed here automatically integrates two different purposes: data compression and noise robustness, as explained below. For conventional Split Vector Quantization (SVQ) in ETSI DSR standard [7], the fixed VQ codebook may not be matched to the distribution of the time-varying testing data. This mismatch inevitably increases the quantization distortion. For the proposed HQ, the actual decision boundaries are dynamically adjusted according to the local statistics. For example, as shown in Figure 1, C(y) may be changed to C′(y′) with the noise disturbances. The partition cell for the quantization level $D_i=[b_{i-1}, b_i]$ for the disturbed parameter x'_t may also be changed to $[y'_{i-1}, y'_i]$, where $C'(y'_{i-1}) = b_{i-1}$, $C'(y'_i)= b_i$, and they can be quite different from $[y_{i-1}, y_i]$. But the quantization level D_i and the corresponding representative value z_i for the disturbed parameter x'_t may remain unchanged as long as $y'_{i-1} < x'_t < y'_i$, since D_i is fixed on the vertical scale, and z_i is fixed on the horizontal scale. In other words, the quantization is based on the quantization levels D_i on 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 "absorbed" by the quantization levels D_i on the vertical scale and the dynamic histogram C(y). HQ was verified to be able to handle various noisy conditions including non-stationary noisy environments [5,6]. The Histogram Equalization (HEQ) has been proposed and popularly used to equalize the cumulative distributions (or histograms) of both the training and testing feature parameters, and shown to produce very robust features for recognition [8,9,10]. The HEQ can be viewed as the limiting case of HQ proposed here when the number of the HQ quantization levels becomes infinite. But the quantization level D_i of HQ here does bring extra robustness as explained above, which has been verified experimentally [5].

3 THREE-STAGE ERROR CONCEALMENT (EC)

3.1 Error detection

In the ETSI DSR standards [7], every two quantized frames are grouped together and protected with 4-bit cyclic redundancy check (CRC). In this way, the entire frame-pair is labeled erroneous even if only a single bit error occurs in the frame-pair packet. A more efficient way is to make use of the feature characteristics at the subvector level for error detection. The data consistency test checks the continuity of the parameters in two neighboring subvectors [4]. When the difference between two consecutive values of a feature parameter in a subvector exceeds a pre-determined threshold, the subvector is classified as inconsistent. The thresholds are obtained from the statistics of the training corpus. If the statistics of the test data were time-varying and different from those of the training corpus, this approach becomes less reliable. With environmental noise, the parameters are likely to be classified as inconsistent even if they are correctly received.

The consistency test in the HQ framework proposed here is as follows. For a two-dimensional HQ, $r_{t,n}^{(1)}$ and $r_{t,n}^{(2)}$ are the two parameters in the n-th subvector of the present received frame at time t, respectively with histograms $C_1(r_1)$ and $C_2(r_2)$ for the progressively moving segment of past T values, $C_0[\bullet,\bullet]$ is the two-dimensional histogram for a standard Gaussian, and $HQ(r_{t,n}^{(1)}, r_{t,n}^{(2)})$ represents the partition cell for the subvector $(r_{t,n}^{(1)}, r_{t,n}^{(2)})$ assigned by HQ. The subvector $(r_{t,n}^{(1)}, r_{t,n}^{(2)})$ is classified as consistent if

$$HQ(C_0^{-1}[C_1(r_{t,n}^{(1)}), C_2(r_{t,n}^{(2)})]) = HQ(r_{t,n}^{(1)}, r_{t,n}^{(2)}). \quad (2)$$

In other words, if these two parameters are correctly received, the order-statistics for them should be similar to the order-statistics for the original values before quantization and therefore similarly quantized into the same HQ partition cell.

We compared the error detection ability of the data consistency check [4] and the HQ-consistency check proposed above under different SNR values for the AURORA 2.0 testing environment [11]. The recall and precision rates are shown in Figure 2(a) and (b). For lower SNR cases, the noise apparently seriously affects the data consistency as verified by the precision degradation in Figure 2(b) (from 66% down to 12% at 0 dB). With the proposed HQ-consistency approach, the precision is much more stable at all SNR values, and both the recall and precision are consistently higher.

3.2 Estimation of erroneous feature vectors

The erroneous subvector estimation under the HQ framework is based on the MAP criterion, which estimates the erroneous n-th subvector at time t, $\hat{S}_{t,n}$, conditioned on the present and previously received n-th subvectors $R_{t,n}$ and $R_{t,n-1}$, using the prior source information (estimated from the clean speech) $P(S_{t,n}(i) \mid R_{t-1,n})$, and the channel transition probability $P(R_{t,n} \mid S_{t,n}(i))$,

$$\begin{aligned} \hat{S}_{t,n} &= \arg\max_{S_{t,n}(i)} \{P(S_{t,n}(i) \mid R_{t,n}, R_{t-1,n})\} \\ &\approx \arg\max_{S_{t,n}(i)} \{P(S_{t,n}(i) \mid R_{t-1,n}) P(R_{t,n} \mid S_{t,n}(i))\}, \end{aligned} \quad (3)$$

where $S_{t,n}(i)$ is the i-th HQ codeword for the n-th transmitted subvector at time t, $S_{t,n}$, and the maximization is over all possible codewords. In order to rely more on the prior source information $P(S_{t,n}(i) \mid R_{t-1,n})$ than the channel transition probability $P(R_{t,n} \mid S_{t,n}(i))$ in equation (3) when the channel condition is less reliable, the channel transition probability in equation (3) is calculated according to the estimated bit error rate (BER) of the current frame

$$P(R_{t,n} \mid S_{t,n}(i)) = BER^{d(S_{t,n}(i),R_{t,n})} * (1-BER)^{M-d(S_{t,n}(i),R_{t,n})}, \quad (4)$$

where M is the number of bits in the received subvector $R_{t,n}$,

$d(\bullet,\bullet)$ represents the Hamming distance between two symbols, and BER is the number of inconsistent subvectors over the total bits within a frame (The approximation is reasonable because in our simulation the number of bit errors is mostly 1 in an erroneous symbol). Note that the BER value is estimated for each frame and the same BER is used for all the subvectors in the frame. The estimated BER is always below 0.5, so equation (4) made $P(R_{t,n} \mid S_{t,n}(i))$ more uniformly distributed for all possible codewords $S_{t,n}(i)$ (i.e. difference in $P(R_{t,n} \mid S_{t,n}(i))$ is smaller for different Hamming distance d) when $R_{t,n}$ is less reliable (larger BER), and significantly differentiated when $R_{t,n}$ is more reliable (smaller BER).

The basic idea here is to exploit the correlation between consecutive frames in speech signals to estimate the lost subvectors. Therefore a relatively robust quantization process such as HQ is very helpful, because with a less robust quantization process, the environmental noise may move the feature vectors to a different partition cell and the subvector transition relationship in speech signals may be disturbed. Table 1 lists the entropy measure obtained from the conditional probabilities, $H(S_{t,n} \mid S_{t-1,n})$, for the symbols obtained from the SVQ approach [7] and the HQ approach proposed here for the AURORA 2.0 [11]. It can be found that this entropy measure is almost always lower for HQ, which means with HQ the estimate of lost subvectors can be made better.

3.3 Compensation in Viterbi decoding

The distribution of the posterior probability $P(S_{t,n}(i) \mid R_{t,n}, R_{t-1,n})$ in equation (3) characterizes the uncertainty of the estimated features. If this distribution is assumed to be Gaussian, the uncertainty decoding of the estimated features can be easily performed within the HQ framework by increasing the variance of each Gaussian mixture in the HMM as proposed and verified to be useful previously [6].

3.4 Three-stage EC under the HQ framework

The three stages of EC presented above can be easily integrated. In the first stage, the received frame-pairs are first checked with CRC to detect errors at the frame level. The erroneous frame-pairs are then further checked at the subvector level by the HQ consistency test as mentioned in section 3.1. In the second stage, the erroneous subvectors detected at the first stage are estimated and reconstructed as presented in section 3.2. This estimation is conditioned on the estimated BER of the received frames. In the third stage, uncertainty decoding in the Viterbi search process makes the HMMs less discriminative for subvectors with higher uncertainty.

4 EXPERIMENTAL CONDITIONS

The experiments reported here were performed on the AURORA 2.0 testing environment [11]. To evaluate the robustness against mismatched conditions, the clean-speech training condition were tested with testing sets A, B and C for SNR ranging from 20dB to 0 dB. The MFCC extraction uses the WI007 front-end.

General Packet Radio Service (GPRS) was chosen as an example for the wireless channels in the experiments, which was developed by ETSI to enhance the GSM system based on the packet switching framework. It includes four different error control coding schemes, CS1~CS4, each with a different code rate. The GPRS simulation software used here was developed by the Wireless-Communication-Lab of National Taiwan University, in which all complicated transmission phenomena have been carefully taken care of, such as the propagation model, the multi-path fading, the Doppler spread, etc. The experimental results presented below are based on the following simulation configurations: typical urban (TU, an environment with more severe fading), transmission SNR of 10 dB, client traveling with speed of 0/3/50/100/250 km/hr, single antenna, hard decision at the receiver, and CS4 coding scheme (i.e., without any protection).

$H(S_{t,n} \mid S_{t-1,n})$	c1,c2	c3,c4	c5,c6	c7,c8	c9,c10	c11,c12	logE
SVQ	4.32	4.57	4.58	4.61	4.55	4.49	1.85
HQ	3.39	3.87	4.23	4.42	4.47	4.51	1.31

Table 1. Conditional entropy of SVQ and the proposed HQ

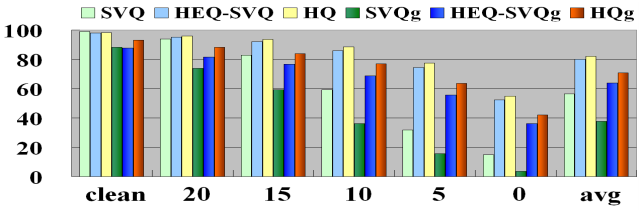

Figure 3. Comparison of SVQ, HEQ-SVQ and HQ, without and with GPRS transmission errors (SVQ$_g$, HEQ-SVQ$_g$, HQ$_g$), averaged over all types of noise, but separated for each SNR value.

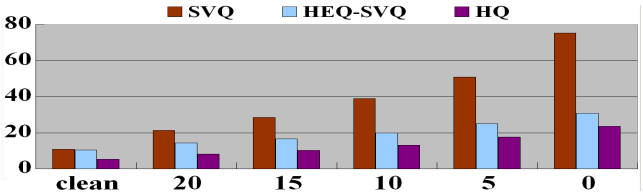

Figure 4. Comparison of SVQ, HEQ-SVQ and HQ with the percentage of words which were correctly recognized if without transmission errors, but incorrectly recognized after transmission.

5 EXPERIMENTAL RESULTS

We first compared the robustness of SVQ in ETSI standard and HQ proposed here against environmental noise and transmission errors (at speed 0 km/hr). Figure 3 shows the averaged results over all different types of noise but separated for different SNR values. The results for the standard SVQ (4.4kbps), SVQ compensated by HEQ (HEQ-SVQ) and HQ (3.9kbps) are the first three bars, and the next three bars are those with GPRS transmission errors (SVQ$_g$, HEQ-SVQ$_g$, HQ$_g$). For SVQ, the performance degradation with GPRS is larger when SNR is lower even with HEQ (2nd bar compared to 4-th bar, e.g. 99% to 87% for clean data, 92% to 76% for 15dB SNR, and 86% to 69% for 10dB SNR). Apparently, features corrupted by noise are more sensitive to transmission errors. The improvements HQ offered over HEQ-SVQ are consistent at all SNR values with transmission errors (6-th bar to 5-th bar) or without transmission errors (3rd bar to 2nd bar), and especially significant with transmission errors. For example, in the case of 10dB SNR with GPRS, HQ offered an accuracy of 77% while the number was 35% and 69% for SVQ and HEQ-SVQ. This verified that HQ is robust against both environmental noise and transmission errors.

To analyze the degradation of recognition accuracy caused by the transmission errors, we examine the percentage of words which were correctly recognized if without transmission errors, but incorrectly recognized after transmission. The comparisons of this

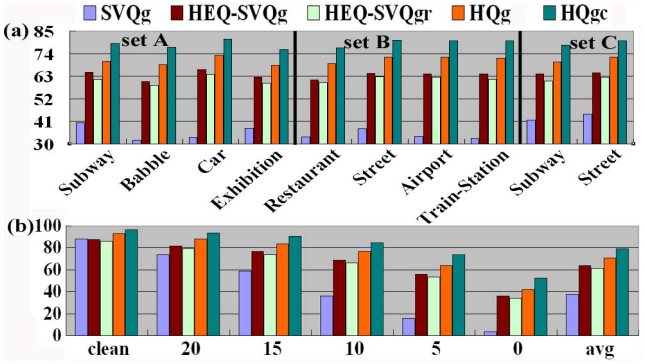

Figure 5. Comparison of SVQ, HEQ-SVQ without and with repetition (HEQ-SVQ$_{gr}$), HQ without and with EC techniques (HQ$_{gc}$): (a) averaged over all SNR values, but separated for different noise types in sets A, B, C, and (b) averaged over all types of noise, but separated for each SNR value.

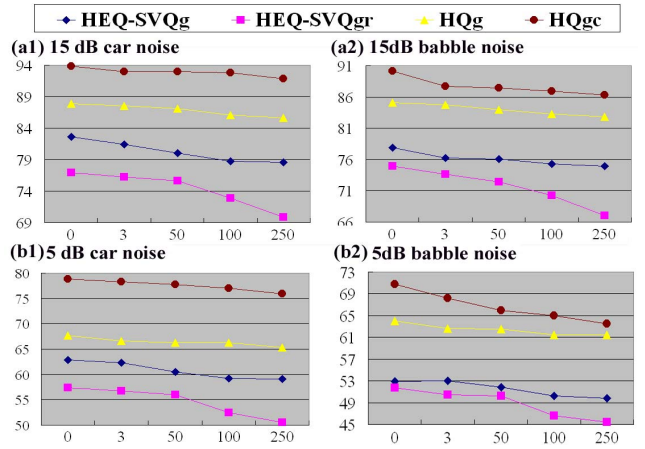

Figure 6. Comparison of HEQ-SVQ without and with repetition, HQ without and with EC, all with GPRS at travel speeds 0/3/50/100/250 km/hr: (a1)/(a2) for car/babble noise at 15 dB SNR, and (b1)/(b2) for car/babble noise at 5 dB SNR.

percentage for SVQ, HEQ-SVQ and HQ under different input speech SNR conditions with exactly the same GPRS condition as mentioned above are shown in Figure 4. The rapid increase of this percentage for SVQ indicated that the noise corrupted SVQ symbols were very sensitive to the transmission errors. HEQ-SVQ was much better, while HQ was the best in all cases.

Figure 5 shows the results for GPRS at speed 0km/hr, where the five bars in order in each set are respectively SVQ, HEQ-SVQ, HEQ-SVQ with repetition (HEQ-SVQ$_{gr}$, the ETSI error mitigation strategy), HQ, and HQ with the three-stage EC techniques (HQ$_{gc}$). Figure 5(a) are those averaged over all SNR values but separated for different noise types in sets A, B, C, and (b) are those averaged over all types of noise but separated for different SNR values. The ETSI repetition technique actually degraded the performance of HEQ-SVQ (3rd bar vs. 2nd bar) because the whole feature vectors including the correct subvectors are replaced by the very possibly inaccurate estimations. HQ without any EC techniques (the 4-th bar) actually outperformed the first three bars for all noise types and all SNR values. Applying the proposed three-stage EC techniques further improved the performance significantly for all noise types and all SNR values. This verified that the three-stage EC framework is robust against both environmental noise and transmission errors.

The next sets of experiments compared in Figure 5 are HEQ-SVQ, HEQ-SVQ with repetition of ETSI, HQ, and HQ with the three-stage EC, all with GPRS at traveling speeds 0/3/50/100/250 km/hr, for car/babble input speech noise at 15 dB and 5 dB SNR in Figure 6 (a1)/(a2) and (b1)/(b2) respectively. The superiority of the proposed HQ with EC (HQ$_{gc}$) are quite clear as verified by the highest curves in all cases. As an example, for 15 dB car noise at 100km/hr as shown in Figure 6(a1), the performance of HEQ-SVQ degraded seriously (79%), HQ is much better (86%), applying ETSI repetition on HEQ-SVQ does not help (73%), while the three-stage EC offered very good improvements (93.5%). Note that here in Figure 6(a1) the HEQ-SVQ features with noise disturbances are more sensitive to transmission errors (83% at 0km/hr and 79% at 100km/hr), while HQ features are more robust (88% at 0km/hr and 86% at 100km/hr). For 5 dB car noises as shown in Figure 6(b1), the performance of HEQ-SVQ degraded (e.g. 59% at 100 km/hr), HQ is much better (e.g. 67% at 100 km/hr), and the three-stage EC further improved the performance significantly (e.g. 77.5% at 100 km/hr). Similar situations can also be found in all cases in Figure 6.

6 CONCLUSIONS

A three-stage error concealment (EC) framework based on the Histogram-based Quantization (HQ) for Distributed Speech Recognition (DSR) is proposed. Improved recognition performance was obtained consistently for a wide variety of environmental noise and transmission error conditions.

7 REFERENCES

[1] Constantinos Boulis, Mari Ostendorf, Eve A. Riskin, Scott Otterson, "Graceful Degradation of Speech Recognition Performance over Packet-Erasure Networks," IEEE Transactions on Speech and Audio Processing, Nov. 2002.

[2] Ben Milner, Alastair James, "Robust Speech Recognition over Mobile and IP Networks in Burst-Like Packet Loss," IEEE Transactions on Speech and Audio Processing, Jan. 2006.

[3] Alexis Bernard, Abeer Alwan, "Low-bitrate Distributed Speech Recognition for Packet-based and Wireless Communication," IEEE Transactions on Speech and Audio Processing, Nov. 2002.

[4] Zheng-Hua Tan, Paul Dalsgaard, Borge Lindberg, "A Subvector Based Error Concealment Algorithm for Speech Recognition over Mobile Networks," ICASSP 2004.

[5] Chia-yu Wan, Lin-shan Lee, "Histogram-based Quantization (HQ) for Robust and Scalable Distributed Speech Recognition," Eurospeech 2005.

[6] Chia-yu Wan, Lin-shan Lee, "Joint Uncertainty Decoding (JUD) with Histogram-based Quantization (HQ) for Robust and/or Distributed Speech Recognition," ICASSP 2006.

[7] "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression algorithms," ETSI ES 202 050 v1.1.3, ETSI standard, 2003.

[8] Sirko Molau, Michael Pitz and Herman Ney, "Histogram Based Normalization in the Acoustic Feature Space, " ASRU 2001.

[9] George Saon, Satya Dharanipragada, Daniel Povey, "Feature Space Gaussianization," ICASSP 2004.

[10] Sirko Molau, Florian Hilger, Daniel Keysers, Hermann Ney, "Enhanced Histogram Normalization in the Acoustic Feature Space," ICSLP 2002.

[11] David Pearce, Hans-Gunter Hirsch, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," ISCA ITRW ASR2000, Paris, France, September 18-20, 2000.