TRELLIS ENCODED VECTOR QUANTIZATION FOR ROBUST SPEECH RECOGNITION

Wu Chou, Nambi Seshadri and Mazin Rahim

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill,NJ 07974, U.S.A.

ABSTRACT

In this paper, a joint data (features) and channel (bias) estimation framework for robust speech recognition is described. A trellis encoded vector quantizer is used as a pre-processor to estimate the channel bias using blind maximum likelihood sequence estimation. Sequential constraint in the feature vector sequence is explored and used in two ways, namely, a) the selection of the quantized signal constellation, b) the decoding process in joint data and channel estimation. A two state trellis encoded vector quantizer is designed for signal bias removal applications. Comparing with the conventional memoryless VQ based approach in signal bias removal, the preliminary experimental results indicate that incorporating sequential constraint in joint data and channel estimation for robust speech recognition is advantageous.

1. INTRODUCTION

Ambient noise and variations in channel conditions often cause a severe degradation in speech recognition performance. Many approaches are proposed to minimize the adverse effects of acoustic mismatch encountered in training and testing environments. Signal bias removal is an effective signal conditioning method in robust speech recognition [2], in which the acoustic space corresponding to the matched channel condition is characterized by a set of Gaussians centered at the clusters $\{\mu_i\}$. The received cepstrum vector sequence, $\bar{y}$, is used to estimate the bias term $\bar{b}$, according to a maximum likelihood formulation.

One issue related to this approach is that the centroids $\{\mu_i\}$ in SBR are based on a memoryless VQ clustering in the cepstral domain, whereas the feature vector sequence from speech is not a memoryless process and often obeys some intrinsic sequential constraint. This situation is somewhat alleviated in the hierarchical signal bias removal (HSBR) method proposed in [3][4], where the centroids, $\{\mu_i\}$, are generated from the Gaussian pdf mean vectors of HMMs. Although in HSBR the sequential constraint in feature vector sequence is implicitly carried over to the generation of the centroids $\{\mu_i\}$, the decoding process, where the bias vector $\bar{b}$ is estimated, assumes that the feature vector sequence is memoryless.

In this paper, we describe a joint data and channel estimation framework for signal conditioning methods in robust speech recognition. Based on this framework, a coded modulation approach to robust speech recognition is proposed and applied to signal bias removal applications.

2. JOINT DATA AND CHANNEL ESTIMATION

In a communication system, data transmission is often modeled by a dispersive channel with additive noise. Assuming for the moment that the channel is invariant, the received signal y_k has the form

$$y_k = \sum_{i=0}^{L} x_{k-i} h_i + n_k, \tag{1}$$

where h_i is the channel impulse response and n_k is the additive noise. The goal of signal conditioning or equalization is to determine the data sequence from the received signal sequence. Given a received sequence $\bar{y} = (y_1, y_2, \cdots, y_N)$, the maximum likelihood sequence estimator (MLSE) chooses the data sequence

$$\hat{x} = \underset{\{\bar{x} | \bar{x} = (x_1, \cdots, x_N)\}}{\operatorname{argmax}} P(\bar{y} \mid \bar{x}) \tag{2}$$

where $\bar{x} = (x_1, x_2, \cdots, x_N)$ ranges over all possible data sequences of length N and $P(\bar{y} \mid \bar{x})$ is the conditional probability of observing the received signal sequence $\bar{y}$ conditioned on the transmitted input sequence $\bar{x}$. In order to perform maximum likelihood sequence estimation, it often requires the information about the channel. For certain applications, a short training sequence (learning signal) is available, and the channel can be estimated by maximum likelihood channel estimation. Once the channel is estimated, the maximum likelihood sequence estimator can be applied to recover the data from the received signal. If channel is assumed changing, the past decoded data can be used to track the channel variations using stochastic-gradient approximation to least mean squares procedures.

However, in many cases, neither the channel nor the data is known, and there is no short learning signal sequence which can be used for a direct channel estimation. Estimating the data sequence from the received signal without knowing the transmission channel is referred to as the problem of blind sequence estimation. One of the statistical frameworks for blind sequence estimation is the joint data and channel estimation[6]. Under the joint data and channel estimation framework, the blind sequence estimation is based on the joint maximum likelihood estimate of both data sequence and the channel

$$(\bar{x}^*, \bar{h}^*) = \underset{\bar{x},\bar{h}}{\operatorname{argmax}}\, p(\bar{y} \mid \bar{x}, \bar{h}), \tag{3}$$

where the maximization is taken over all the equiprobable data sequences of length N and $\bar{h}$ is taken over all possible channel responses. For the additive white Gaussian noise channel, the joint data and channel estimation is obtained by the joint least square minimization

$$(\bar{x}^*, \bar{h}^*) = \underset{\bar{x},\bar{h}}{\operatorname{argmin}} \sum_k \mid y_k - \sum_j x_{k-j} h_j \mid^2 . \tag{4}$$

In practice, optimum solution to a joint data and channel estimation is possible only for small data symbol set and the number of possible channel responses $\bar{h}$ must also be small. Many practical sub-optimal algorithms are proposed to solve the joint data and channel estimation problem. An algorithm often used is the segmental K-means algorithm that produces the estimate of the data using the Viterbi algorithm when conditioned on an initial estimate of the channel. Then channel is re-estimated using the decoded sequence. In order to bring down the complexity in joint data and channel estimation, the signal space is typically vector quantized and the channel responses are limited to N most possible channel responses.

One way to characterize signal bias removal in robust speech recognition is from the point of view of joint data and channel estimation. The basic signal model in signal bias removal (SBR) is that the channel and the ambient noise introduce an unknown additive bias term in the cepstrum domain. The received signal is conditioned by subtracting the bias term $\bar{b}$ estimated from the received data. In this approach, neither the bias (channel) nor the data (speech in put $\bar{x}$) is known. In order to solve this joint data and channel estimation problem, the signal space represented by the cepstrum feature vectors in matched channel conditions is quantized and characterized by a set of Gaussians centered at the clusters $\{\mu_i\}$, following a process of memoryless VQ clustering of the cepstrum feature vectors. The bias term $\bar{b}$ (channel) is estimated through a two-step iterative procedure described below.

1. An estimate of the bias, $\bar{b}$, is computed for each utterance of T frames, such that

$$\bar{b} = \frac{1}{T} \sum_{t=1}^{T} (y_t - \mu_{i^*}), \tag{5}$$

where μ_{i^*} is the "nearest neighbor", according to some distance criterion, to the distorted signal spectrum y_t:

$$i^* = \arg\min_i ||y_t, \mu_i||. \tag{6}$$

2. The estimated bias, $\bar{b}$, is subtracted from the distorted signal

$$\tilde{x}_t = y_t - \bar{b} \qquad 1 \le t \le T, \tag{7}$$

In fact, this procedure is the practical segmental K-means approach to joint data and channel estimation in blind sequence estimation for signal bias removal.

3. DESIGN OF TRELLIS ENCODED VECTOR QUANTIZER

One observation made on the speech is that the sound transition in speech is continuously evolving. In fact, coarticulation is a phenomenon in human speech that the human articulator takes in between positions when transits from one sound to another. The centroids $\{\mu_i\}$ in the SBR approach are based on a memoryless VQ clustering in the cepstral domain, whereas the feature vector sequence from speech is not a memoryless process and often obeys some intrinsic sequential constraint. In addition to the sequential constraint, which is intrinsic in speech generation process, one of the important issues of using the practical quantized approach in joint data and channel estimation is the selection of the quantized signal cluster constellation. In data communication, the data sequence goes through channel coding, typically through a convolutional encoder, which imposes a very strong sequential constraint in the transmitted data sequence. Utilizing the sequential constraint in the transmitted data sequence is essential to improve the performance of the signal recovery in noisy environment. The gain obtained from selecting appropriate code constellation is called "coding gain", which is the reduction in probability of error comparing to uncoded memoryless VQ scheme.

One way to achieve "coding gain" for robust speech recognition is to treat the problem of vector quantizing the speech signal space as a problem of finding a good code

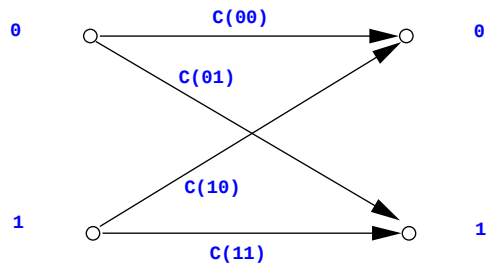

Figure 1: Diagram of a two state trellis encoded vector quantizer.

constellation in which certain sequential constraint is embedded in the transmitted data sequence. The code constellation and its sequential constraint can be specified by a trellis encoded vector quantizer, namely a finite state vector quantizer (FSVQ). The quantized signal clusters are the code words on the arcs of the FSVQ. Fig. 1 illustrates a two state FSVQ with four code words. The advantage of a trellis encoded vector quantizer is that it allows to search several vectors into the future when finding a minimum distortion path to determine the quantized value of the current received signal. This is very different from a memoryless VQ where the quantization process does not look into the trend of the future and past samples. As a consequence, the sequential constraint in the received signal sequence is not explored. Methods for designing trellis encoded vector quantizer is beyond the scope of this paper and good references can be found in [6][7][8]. However, speech feature vector sequence has no added sequential constraint from channel coding. As a consequence, the problem becomes quite special. It is to find a trellis encoded vector quantizer in which sequential constraint in the source signal (speech feature vector sequence) is properly embedded. This is a problem which is in general very difficult.

As pointed earlier, the goal of using a trellis encoded vector quantizer in signal bias removal application is to incorporating certain sequential constraint in both the selection of the signal centroids $\{\mu_i\}$ and the decoding process where the bias $\bar{b}$ is estimated. Unlike in data communication, the sequential constraint imposed on the speech feature vector sequence must be consistent with the speech generation process. In speech recognition, the speech generation process is characterized by the distribution and the structure of the acoustic model. This requires that sequential constraint, if it is imposed, has to be consistent with the constraint specified in the acoustic model units used in recognition.

In our approach, a two state trellis encoded vector quantizer is used. The code constellation consists of four code words (signal clusters) $\{C_{00}, C_{01}, C_{10}, C_{11}\}$. C_{00} is generated by clustering all the mean vectors of Gaussian pdfs in the HMMs for silence and background speech events; C_{01} is generated by clustering all the mean vectors of Gaussian pdfs in the head state of HMMs for keywords; C_{10} is generated by clustering all mean vectors of Gaussian pdfs in the tail state of HMMs for the keywords; C_{11} is generated by clustering all mean vectors of Gaussian pdfs in the HMMs for all keywords. Fig.1 illustrates the structure of this trellis encoded vector quantizer. This two state trellis encoded vector quantizer is a 1-bit quantizer. The transition between codewords bears no direct relationship with the closeness of the corresponding codewords. Such a code constellation design is consistent with the sequential constraint imposed by the acoustic modeling. In fact, the sequential constraint of the feature vector sequence is embedded in the transition between codewords.

4. FSVQ BASED SBR

Signal bias is estimated by a Viterbi search through the trellis determined by the trellis encoded vector quantizer. The optimal path with minimum sequence distortion is specified by

$$\bar{s}^* = \underset{\{\bar{s}|\bar{s}=(s_1,s_2,\cdots,s_T)\}}{\text{argmax}} \sum_{k=1}^{T} \| y_k - \alpha(y_k, s_k) \| \quad (8)$$

$$= \underset{\{\bar{s}|\bar{s}=(s_1,s_2,\cdots,s_T)\}}{\text{argmax}} \sum_{k=1}^{T} \| y_k - \mu_k(y_k, s_k) \|, \quad (9)$$

where $\bar{s} = (s_1, s_1, \cdots, s_T)$ is taken over all state sequences, $\alpha(y_k, s_k) = \mu_k(y_k, s_k)$ is the encoding function in the trellis encoded vector quantizer. The bias term $\bar{b}$ is estimated by

$$\bar{b}^* = \underset{\bar{b}}{\text{argmax}} \sum_{k=1}^{T} \| y_k - \mu_k(y_k, s_k^*) - \bar{b} \| . \quad (10)$$

which is evaluated along the optimal path with minimum distortion $\bar{s}^* = (s_1^*, s_2^*, \cdots, s_T^*)$. For Gaussian noise, it reduces to

$$\bar{b} = \frac{1}{T} \sum_{k=1}^{T} (y_k - \mu_k(y_k, s_k^*)), \quad (11)$$

which is the original SBR formulation.

Multiple signal cluster dependent biases $\{\bar{b}_i\}$ can also be estimated which is given by

$$\bar{b}_i = \frac{1}{T_i} \sum (y_{t_i} - \mu_{t_i}(y_{t_i}, s_{t_i}^*)) \mid_{\mu_k(y_{t_i}, s_{t_i}^*) = C_i}, \quad (12)$$

where T_i is the number of received signal samples being mapped to C_i. Similar to HSBR, these multiple biases can be integrated to form a time dependent bias. In fact, the

time dependent bias $\bar{b}_t$ for each received signal can be based on the conditional expectation

$$E(\bar{b}_t \mid y_t) = \sum_{j=1}^{M} \bar{b}_j \frac{f(y_t \mid Q(x_t) = C_j) P(Q(x_t) = C_j)}{\sum_{i=1}^{M} f(y_t \mid Q(x_t) = C_i) P(Q(x_t) = C_i)}, \tag{13}$$

where $Q(x_t)$ is the quantization function in the signal space of matched conditions, $\{C_i \mid i = 1, \cdots, M\}$ is the signal cluster constellation. If the prior probability of the vector quantization $P(Q(x_t))$ is assumed to be uniform, equation (13) becomes

$$E(\bar{b}_t \mid y_t) = \sum_{j=1}^{M} \bar{b}_j \frac{f((y_t \mid Q(x_t) = C_j)}{\sum_{i=1}^{M} f(y_t \mid Q(x_t) = C_i)}. \tag{14}$$

The estimated input signal given the received signal y_t is

$$\tilde{x}_t = E(x_t \mid y_t) = y_t - E(\bar{b}_t \mid y_t) \tag{15}$$

which is the optimal minimum mean square error bias compensation given the signal bias model $y_t = x_t + \bar{b}_t$. In fact, the fuzzy distance used in HSBR is one of such special cases of (15).

5. PRELIMINARY EXPERIMENTAL RESULTS

The proposed trellis encoded vector quantization is applied to signal bias removal applications. It was experimented on a connected digit database collected from the wireless environment. The database contains speech waveforms from various wireless transmission medias, such as TDMA (SI-54), GSM and AMPS (analog). The back ground noise level in the speech was quite high. The acoustic model used in the experiments was a context dependent model with 274 context dependent model units. The model was obtained from ML training.

The proposed approach was compared with the original HSBR approach on the same model. Both FSVQ based SBR and the original HSBR applied the frame dependent bias for signal conditioning. The frame dependent bias was obtained by using the same type of fuzzy distance described in HSBR. It was a conditional expectation taken over a set of estimated biases. The differences between the two approaches were in their signal cluster constellations and in the way that the signal cluster biases were estimated. Biases in FSVQ based SBR were estimated based on a two state trellis encoded VQ described in the previous section. There are about 11.5K utterances in the training set and 9K utterances in the test set. The experimental results are tabulated in Table 1. The proposed FSVQ based SBR provided an additional 13% over the HSBR method and the word error rate reduction of the FSVQ based SBR over the baseline model was 29%.

Conditioning	Wd_Err_rate	Wd_Err_reduction
Baseline	3.8%	N/A
HSBR	3.1%	18%
FSVQ	2.7%	29%

6. SUMMARY

In this paper, we described a joint data and channel estimation framework for robust speech recognition. A trellis encoded vector quantizer is used as a pre-processor to estimate the bias using blind maximum likelihood sequence estimation. In this approach, sequential constraint in the feature vector sequence is explored and used in two ways, namely, a) the selection of the quantized signal constellation, b) the decoding process in joint data and channel estimation. A two state trellis encoded vector quantizer is designed for signal bias removal applications. Comparing with the conventional memoryless VQ based approach in signal bias removal, the preliminary experimental results indicate that incorporating sequential constraint in joint data and channel estimation for robust speech recognition is advantageous. In addition, this approach is quite easy to implement and the CPU overhead is almost negligible, because the search trellis is extremely small.

REFERENCES

[1] Liu, F.-H., Stern, R., Acero, R., and Moreno, P. (1994) "Environment Normalization for Robust Speech Recognition using Direct Cepstral Normalization," *Proc. ICASSP' 94*, **II,** pp. 61-64.

[2] Rahim, M. and Juang, B-H. (1994) "Signal Bias Removal for Robust Telephone Speech Recognition in Adverse Environments," *Proc. ICASSP' 94*, **II.**

[3] Rahim, M., Juang, B-H. Chou, W. and Buhrke, E., "Signal Conditioning Techniques for Robust Speech Recognition," IEEE Signal Processing Letters.

[4] Chou, W., Rahim, M., and Buhrke, E., "Signal Conditioned Minimum Error Rate Training", *Proc. Eurospeech'95*, **I**, pp. 495–498.

[5] W. Chou, B.H. Juang and C.H. Lee, "Segmental GPD training of a HMM based Speech Recognizer", *Proc. ICASSP92* pp. 473–476, 1991.

[6] Seshadri, N., "Joint Data and Channel Estimation Using Blind Trellis Search Techniques", *IEEE Trans. on Communications,* **Vol. 42**, No.2/3/4, pp. 52-59.

[7] Juang, B.-H., "Design and Performance of Trellis Vector Quantizers for Speech Signals", *IEEE Trans. on ASSP*, Vol 36, No. 9, Sept 1988, pp.549–552.

[8] Gersho, A. and Gray, R. M., "Vector Quantization and Signal Compression", Kluwer Academic Publishers, 1991.