論文95-32B-5-10

벡터 양자화에서 시간 평균 왜곡치의 수렴 특성: II. 훈련된 부호책의 검사 기법

(The Convergence Characteristics of The Time-Averaged Distortion in Vector Quantization: Part II. Applications to Testing Trained Codebooks)

金 東 植 *

(Dong Sik Kim)

요 약

벡터 양자화(vector quantization: VQ)를 위한 부호책(codebook)을 설계하기 위해서 일반적으로 훈련 집합(training set)을 이용한 군집화 알고리듬(clustering algorithm)이 사용된다. 그런데 이렇게 훈련된 부호책의 성능을 검사하기 위해서는 훈련 집합 외부의 표본 벡터들로부터 검사 집합(test set)을 구성하여 소위 훈련 벡터의 외부 왜곡치(outside-training-set-distortion: OTSD)를 얻는다. 그리고 이 값과 훈련 벡터의 내부 왜곡치(inside-training-set-distortion: ITSD)와의 차를 구하여 훈련된 부호책의 성능을 평가한다. 이러한 방법이 성립되기 위해서는 전제 조건이 필요한데, 특별한 이론적 고찰없이 사용되어 오고 있는 실정이므로 잘못된 결과나 계산의 낭비를 초래할 수 있다. 본 논문에서는 이러한 VQ 왜곡치에 대한 이론적 분석[16]을 기반으로 하여 언급된 왜곡치 차를 이론적으로 고찰하였으며, 아울러 훈련된 부호책을 간단히 검사할 수 있는 방법을 제시하였다. 이론적 고찰의 확인을 위하여 합성 데이타와 실 영상에 대해서 모의 실험도 행하였다.

Abstract

When codebooks designed by a clustering algorithm using training sets, a time-averaged distortion, which is called the inside-training-set-distortion (ITSD), is usually calculated in each iteration of the algorithm, since the input probability function is unknown in general. The algorithm stops if the ITSD no more significantly decreases. Then, in order to test the trained codebook, the outside-training-set-distortion (OTSD) is to be calculated by a time-averaged approximation using the test set. Hence codebooks that yield small values of the OTSD are regarded as good codebooks. In other words, the calculation of the OTSD is a criterion to testing a trained codebook. But, such an argument is not always true if some conditions are not satisfied. Moreover, in order to obtain an approximation of the OTSD using the test set, it is known that a large test set is requared in general. But, large test set causes heavy calculation complexity. In this paper, from the analyses in [16], it has been revealed that the enough size of the test set is only the same as that of the codebook when codebook size is large. Then a simple method to testing trained codebooks is addressed. Experimental results on synthetic data and real images supporting the analysis are also provided and discussed.

* 正會員, 서울大學校 制御計測工學科
(Dept. of Control and Instrumentation Seoul National University)

接受日字: 1994年11月10日, 수정완료일: 1995年4月27日

I. 서 론

벡터 양자화(vector quantization: VQ)는 높은 압축율이 요구되는 영상 및 음성 데이타 부호화에 사용되는 기법으로, 최근에 이를 이용한 많은 연구가 수행되고 있다[4]. VQ를 구현하는 데 있어서 중요한 문제 중 하나는 부호책(codebook) 설계이다. 실제 응용 분야에서는, 양자화하려고 하는 신호의 확률 분포 함수가 잘 알려져 있지 않기 때문에 훈련 벡터(training vector)라 불리는 표본 벡터들로부터 훈련 집합(training set: TS)을 구성하여 이를 군집화(clustering)하는, 일종의 순환(iteration) 방법으로 부호책을 설계한다[1], [4], [5], [7]. 훈련 벡터들을 미리 설정한 부호단어(codeword)의 갯수만큼 군집화하면서 매 순환 루프(loop)에서는 설계된 부호책이 최적치 또는 국부 최소치에 어느 정도 근접해 있는지를 판단하기 위해서 양자화 왜곡치(quantizer distortion)를 구한다. 그런데 이때 왜곡치는 TS에 훈련된 부호책으로 어떤 신호를 양자화할 때 구해지는 것으로, 입력 신호원의 확률 분포를 알면 수학적 평균을 통해서 구할 수 있다[7, Algorithm(Known Distribution)]. 이러한 왜곡치를 훈련 벡터의 외부 왜곡치(outside-training-set-distortion: OTSD)라 부르기로 한다. 그러나 언급한 바와 같이 확률 분포를 모르기 때문에 시간 평균 왜곡치(time-averaged distortion)를 구해서 평균 왜곡치의 근사치를 얻는다[7, Algorithm(Unknown Distribution)]. 이러한 왜곡치는 훈련 집합의 내부 왜곡치(inside-training-set-distortion: ITSD)라 불리운다. M을 TS의 크기라 하고 N을 부호책의 크기라 하면 훈련비(training ratio) β는 $\beta \triangleq M/N$으로 정의된다[4]. 최소 ITSD, 즉 실험적 최적 왜곡치(empirically optimal distortion)는 훈련비가 증가하면서 최적 양자화 왜곡치로 수렴하게 된다.[8], [10] 다시 말해서 훈련비가 큰 경우에는 ITSD 값이 작을수록 부호책 설계가 잘 되었다는 것을 의미한다. 그러나 훈련비가 크기 때문에 부호책 설계를 위한 군집화 알고리듬의 계산량은 증가하게 되므로 적절한 양의 TS를 사용하는 것이 좋다.

한편 제한된 갯수의 훈련 벡터를 가지고 있는 경우, 즉 훈련비가 그리 크지 않은 경우에는 단순히 ITSD를 구하는 것만으로는 훈련된 부호책이 잘 설계되었는지 알 수 없다. 다시 말하면, ITSD 값이 작다고 해서 좋은 부호책이 설계되었다고 말할 수는 없다. 왜냐하면 극단적인 경우 $\beta=1$ 인 경우의 최소 ITSD는 0이 되지만 이는 최적 부호책과는 거리가 멀기 때문이다. 따라서 이러한 경우 흔히 OTSD를 시간 평균 근사치로 구해서 훈련된 부호책을 검사(test)한다. 다시 말해서, 부호책 설계시 사용되었던 훈련 벡터와 별도로, 설계된 부호책의 성능을 충분히 평가할 수 있는 다양한 표본 벡터들을 충분히 사용하여 부호책을 검사한다. 이러한 평가를 위한 표본 벡터의 집합을 검사 집합(test set)이라 부르기로 하고, 그 크기를 M' 그리고 $\beta' = M'/N$을 검사비라 하자. [16]에서 β'를 표본비라 하였는데 [16]에서의 표본 집합이 본 논문에서는 검사 집합으로 사용되기 때문에 그 명칭만 의미상 검사비라고 쓰기로 한다. 이때 근사화로 구한 OTSD가 적절한 수준의 왜곡치를 가지면 좋은 부호책 설계가 이루어졌다고 생각하게 된다. 그러나 검사 집합을 선정하는 데 있어서의 난이성 및 시간 평균 OTSD를 구하기 위한 많은 계산량 등이 문제가 되고 있다. 특히 TS 및 검사 집합을 선정하는 과정은 이론적 분석이 빈약하므로 체계적으로 이루어지지 않고 있다.

최근 들어서, TS의 크기가 VQ 부호책 설계 또는 VQ 왜곡치에 미치는 영향을 연구한 논문이 몇 편 소개되었다[11], [14], [15]. Cosman 등은 nonlinear regression 모델에 근거하여 OTSD와 ITSD의 차가 M에 대해 감소하는 형태의 함수를 가짐을 보였다[11]. 다시 쓰면 $AM^{-a}+B$와 같이 M에 대해 감소하는 형태가 된다. 여기서 A와 B 그리고 a는 일종의 curve fitting을 위한 계수들이다. Cohn 등은 TS의 크기가 VQ에 미치는 영향을 이론 및 실험적으로 고찰해 보았다. 그들은 분류(classification) 기법 영역에서 사용되는 학습 이론(learning theory)을 VQ의 이론적 분석에 적용하여 VQ 왜곡치의 상한 경계치(upper bound)를 유도하였다. 그러나 그들의 경계치는 실질적인 값들과 너무 차이가 나서 실용성이 결여되어 있다. 그래서 그들은 실험적 관찰을 통해서 Cosman 등의 결과와 유사한 것을 유도해 내었다. 즉 이들 역시 왜곡치의 차가 M에 반비례한다고 밝히고 있다. 앞에서 언급된 연구들은 실험적인 고찰에 중점을 두고 있는 반면에, [15]에는 최소 ITSD에 대한 상한 경계치를 TS의 크기와 부호책의 크기 그리고 입력 신호의 확률 분포 함수로 유도하였다. 이 상한 경계치는 훈련비뿐만이 아니라 입력 신호의 통계적 특성까지 고려하고 있으므로, 알고 있는 훈련비 등을 통해서 작성된 부호책의 성능을 짐작할 수 있다. 정리를 하면, [11]과 [14]에서 언급한 바와 같이 M 또는 β의 함수로 나타내지는 곡선을 실험 결과들로부터 얻어내어 이를 통해서 설계된 부호책의 성능을 알아볼 수 있으며 TS의 크기가 VQ의 성능에 미치는 영향도 알아볼 수 있을 것이다. 그러나 이러한 방법에 의한 시도에는 몇 가지 문제가 존재한다. 첫번째는, ITSD 값은 TS에 의해서 구

하므로 표본 집합인 TS에 따라 그 값이 변동된다는 사실이다. 다시 말해서 [15]에서와 같이 TS를 랜덤 표본(random sample)이라고 가정하면, ITSD는 같은 표본 공간(sample space)에 존재하는 랜덤 변수가 된다. 그러므로 이러한 값의 정합 곡선(fitting curve)을 구한다고 하는 것은 그리 큰 의미가 없다. 그러나 다행히도 [15]에서 언급된 바와 같이 부호책의 크기가 커지는 경우에는 ITSD의 randomness가 감소하게 되므로 이때는 정합 곡선에 의한 부호책의 성능 예측이 가능하리라 사료된다. 두번째는 언급된 정합 곡선을 구하기 위해서는 여러 TS의 크기에 따라서 ITSD들을 구해서 정합 곡선을 구해야 한다는 것이다. 사전에 다양한 종류의 정합 곡선을 구해 둔다고 하는 것은 상당히 많은 계산량이 요구되므로 결국 이러한 정합 곡선에 의한 방법으로 어떠한 TS로부터 만들어진 부호책의 성능 검사는 거의 불가능하다. 결국은 ITSD 또는 OTSD를 TS나 검사 집합으로부터 직접 구해서 훈련된 부호책을 검사하는 방법이 가장 타당하다.

본 논문에서는 ITSD와 OTSD를 구해서 설계된 부 호책의 성능을 검사하는 방법에 대해서 고찰해 보았다.

ITSD나 OTSD를 통해서 훈련된 부호책이 최적치에 얼마나 가까운가를 검사하기 위해서는 어떠한 선행 조건이 요구되는데 이러한 것들은 실제 적용시 특별한 주의없이 사용되어 오고 있었다. 따라서 본 논문에서는 [16]의 연구 결과를 토대로 주의해야 할 사항들을 이론적 고찰과 함께 언급하였다.

본 논문의 구성은 다음과 같다. 먼저 제 II장에서는 [16]에서 언급된 여러 VQ에 대한 정의와 더불어 부호책 검사를 위한 왜곡치를 정의하였다. 제 III장에서는 간단한 부호책 검사 기법을 소개하였다. 합성 신호와 실제 영상 데이타에 대한 모의 실험을 제 IV장에서 소개하였으며, 마지막 장에서 결론을 내렸다.

II. 부호책 검사를 위한 왜곡치

이 장에서는 어떠한 TS에 의해서 훈련된 부호책의 검사를 위한 일종의 왜곡치 차(distortion difference)를 정의하고 이에 대해 고찰해 보았다. 이 논문에서 사용된 정의들은 [16]에서 정의된 것들 및 표기를 모두 다시 사용하기로 하고, 중복되는 것들은 본 논문에서는 다시 언급하지 않고 사용하였다. 그러므로 여기서 언급되지 않은 기본적 VQ에 대한 정의는 [16]을 참조하기 바란다.

C^*를 어떠한 확률 함수 F에 대해 최적인 부호책 중의 하나라고 하면, 즉 $D_r(C^*,F)=\inf_{C\in C_N}(C,F)$ 이라 하자. 그러면 어떠한 표본점(sample point) $\omega\in\Omega$에서 구현된 TS로 훈련된 부호책 $\overline{C}^{\omega}$를 검사하기 위해서 흔히 다음과 같은 왜곡치 차를 생각하게 된다.

$$(OTSD-D^*)^{\omega}\triangleq D_r(\overline{C}^{\omega},F)-D_r(C^*,F) \quad (1)$$

식 (1)의 왜곡치 차 값이 줄어들게 된다면 $\overline{C}^{\omega}$는 C^* 또는 다른 최적 부호책으로 근접한다는 사실을 의미한다. 그러므로 이 값을 계산하여 부호책 검사의 기준으로 삼으면 되지만 최소 양자화 왜곡치 $D_r(C^*,F)$를 계산해 내는 것은 거의 불가능하다. 따라서 훈련된 부호책을 검사하기 위하여 다음과 같은 왜곡치 차를 정의하였다.

정 의 1 (왜곡치 차): 각 표본점 $\omega\in\Omega$에 대해 다음과 같은 왜곡치 차를 정의하자.

$$(OTSD-ITSD)^{\omega}\triangleq D_r(\overline{C}^{\omega},F)-D_r(\overline{C}^{*},F_M^{\omega}) \quad (2)$$

이러한 왜곡치 차는 훈련 집합의 외부와 훈련 집합의 내부 왜곡치 차라고 하여[11],[14] 부호책 검사에 흔히 사용되고 있다. 본 논문에서는 이 왜곡치 차가 가지는 의미와 계산하는 방법 등을 [16]의 이론적 분석과 더불어 소개하려 한다.

훈련된 부호책의 성능은 일반적으로 TS에 속하지 않는 다른 표본 집합, 즉 검사 집합을 그 부호책으로 부호화해서 그때의 양자화 왜곡치를 가지고 평가를 하게 된다. 이때 검사 집합은 실제 모르는 F의 통계적 특성을 충분히 나타낼 수 있도록 충분히 커야 한다. 다시 말해서 어떠한 표본 점에 대해 훈련된 부호책의 OTSD를 구한 다음(OTSD - ITSD)$^{\omega}$를 구한다.여기서 ITSD는 부호책을 작성하는 과정에서 구해진다[7]. 만일 어떠한 부호책이 작은 (OTSD - ITSD)$^{\omega}$ 값을 나타낸다고 하면 우리는 훈련된 부호책이 훈련 집합 외부의 표본 집합에 대해서, 즉 보다 정확히 말해서 입력 신호의 확률 함수 F에 ``좋다''고 말할 수 있다. [11]과 [14]에서 언급된 바와 같이 이 왜곡치 차는 고정된 N 값에 대해서 M이 증가하면서 감소하게 된다. 그러나 이러한 왜곡치 차는 훈련된 부호책의 최적화도(optimality)를 항상 의미하는 것은 아니다. 즉 어떠한 조건이 만족되어야 하는데, 그 조건에 대해서 언급하면 다음과 같다. 먼저 최소 ITSD, 즉 실험적 최적 왜곡치를 다음과 같이 정의하였다.

정 의 2(실험적 최적 왜곡치): 어떠한 표본점 $\omega\in\Omega$를 가지는 실험적 분포 함수(empirical

distribition function) F_M^{ω}에 대해 최소 ITSD, 즉 실험적 최적 왜곡치는 다음과 같이 정의된다.

$$\inf D(C,F_M^{\omega}) \tag{3}$$

이 실험적 최적 왜곡치는 TS를 구성하는 랜덤 벡터의 함수이므로 이 또한 랜덤 변수가 됨을 알 수 있다.

보조정리 1: 실험 분포 함수를 $\omega\in\Omega$에 대해 $F_M^{\omega}(x) \triangleq M^{-1}\sum_{j=1}^{M} I_{(-\infty,x)}(X_j^{\omega})$로 정의하자. 이 식에서 $-\infty=(-\infty\cdots-\infty)^T$이다. 그러면 $\beta\to\infty$함에 따라 $\inf_{C\in C_N} D_r(C,F_M^{\omega})$ 는 $\inf_{C\in C_N} D_r(C,F)$에 확률 1로 수렴한다[10, Theorem 1].

보조정리 1의 증명: Glivenko-Cantelli 정리[6]에 의하면, $\lim_{M\to\infty} F_M^{\omega}(x)=F(x)$ a.s.(almost surely) 이다. 그러므로 $|x-y|^r$는, $E\|X\|^r<\infty$ 조건에서 거의 모든 ω의 $\{F_M^{\omega}\}$ 에 대해(almost every ω with respect $\{F_M^{\omega}\}$) 균일하게 적분 가능하다(uniformly integrable)[10].

그러므로 Abaya와 Wise의 정리[10, Theorem 1]로부터 $\inf_{C\in C_N} D_r(C,F_M)$이 $\inf_{C\in C_N} D_r(C,F)$에 $\beta\to\infty$ 함에 따라 확률 1로 수렴한다.

추 론 1: 보조정리 1과 [10]으로 부터 $\beta\to\infty$함에 따라 $D_r((C^*_{(M)})^{\omega},F)$는 $\inf_{C\in C_N} D_r(C,F)$에 확률 1로 수렴한다. 이 식에서 $(C^*_{(M)})^{\omega}$는 $D_r(C,F_M^{\omega})$를 최소화시킨다.

그러므로 어떠한 군집화 알고리듬이 부호책을 작성하는 데 있어서 실험적 최적 부호책(empirically optimal codebook)과 근접하는 부호책을 설계한다고 하면, 보조정리 1과 추론 1로부터 확률 1로의 수렴 $(\mathrm{OTSD}-\mathrm{ITSD})^{\omega}\to 0$은 근사적으로 $(\mathrm{OTSD}-D^*)^{\omega}\to 0$을 의미한다[10]. 결국 설계된 부호책이 ITSD를 최소화시킨다고 하면 왜곡치 차 $(\mathrm{OTSD}-\mathrm{ITSD})^{\omega}$는 훈련된 부호책의 성능 검사에 사용될 수 있다.

그러면 이제 일반적인 군집화에 의한 부호책 설계 알고리듬에 대한 근사식을 유도해 보기로 하자. 어떤 주어진 TS T_M^{ω}에 대해 실험적 최적 부호책 $(C^*_{(M)})^{\omega}$는 T_M^{ω}이 유한 집합이므로 존재한다[1]. 그러나 그러한 부호책을 찾는 것은 매우 비효율적이며 엄청난 계산량으로 인하여 거의 실현 가능성이 없다[1]. 그래서 분류화 분야(classification area)에서는 군집화 알고리듬이 일찍이 도입되었으며[1], VQ 분야에서도 이러한 기법이 도입되어 부호책 설계에 사용되고 있다.[5],[7] [3]과 [7], 그리고 [13]에 소개된 알고리듬들도 실험적 최적 부호책을 찾는 것이 목표로 되어 있다. 사실, GLA(generalized Lloyd algorithm)[7]는 초기 부호책의 선정에 따라 최적 부호책을 찾을 수 있다. 그렇지 않으면 국부적 최소치로 빠지는데 약간의 알고리듬의 변형으로 이러한 문제를 다소간 완화할 수 있다[4],[7],[12]. 그러나 이러한 노력에 의해서 감소되는 양자화 왜곡치의 양은 매우 적으므로, 우리는 군집화 알고리듬에서 최소화된 양자화 왜곡치가 실험적 최소 왜곡치에 근접한 값이라고 생각할 수 있다. 그러므로 다음과 같은 근사식을 얻을 수 있다.

$$D_r(\overline{C}^{\omega},F_M^{\omega})\cong \inf D_r(C,F_M^{\omega}) \tag{4}$$

결국은 이러한 근사식을 가정함으로 $(\mathrm{OTSD}-D^*)^{\omega}$를 구하는 대신, (이 식에서 $D^*=D(C^*,F)$는 일반적으로 모른다.) 왜곡치 차 $(\mathrm{OTSD}-\mathrm{ITSD})^{\omega}$를 구해서 $\overline{C}^{\omega}$가 잘 설계되었는지를 검사할 수 있다.

Ⅲ. 훈련된 부호책 검사 기법

본 장에서는 정의 1에 정의된 왜곡치 차를 [16]의 이론적 분석을 적용하여 고찰해 보았으며, 훈련된 부호책을 검사하는 간단한 방법을 소개하였다.

어떤 부호책 $C\in C_N$이 주어져 있을 때, [16, 추론 1]로부터 다음과 같은 왜곡치 차

$$\sum_{i=1}^{N}\sigma_{r,i}(C,F)-\sum_{i=1}^{N}\sigma_{r,i}(C,F_M^{\omega})=D_r(C,F)-(C,F_M^{\omega}) \tag{5}$$

는 검사비 β가 증가하면 확률 1로, 0에 수렴하게 된다. 이때 N은 고정되어 있다. 즉 식 (5)는 보기에는 정리 1의 (OTSD-ITSD) 형태로 되어 있지만 부호책 C의 최적성과는 전혀 무관하게 0으로 수렴한다. 즉 앞장에서 언급한 바와 같이 어떤 부호책이 식 (4)를 만족하지 못한다면 왜곡치 차가 작게 나온다고 해도 그 부호책이 잘 설계되었다고 말할 수는 없다. 식 (5)는 단순히 실험적 확률 분포 F_M^{ω}가 얼마만큼 참값인 F에 근접되어 있는가를 나타내는 수치가 될 것이다. 또한 [16,정리 2]로 부터, 어떠한 특정한 표본점 ω에 대해서 그리 크지 않은 검사비, 예를 들어 $\beta=1$라 하더라도 N이 증가하면 역시 평균 자승 개념으로(in mean square) 0으로 수렴하게 된다. 다시 말하면 추정치(estimator) $D_r(C,F_M^{\omega})$는 부호책의 크기가 큰 경우 VQ 왜곡치인 $D_r(C,F)$를 근사화할 수 있다.[16] 결국 검사 벡터로 N개만 있으면 OTSD를 근사화로 구할 수 있다는 것이 된다.

이제 이러한 개념을 도입하여 식 (2)의 양자화 왜곡치 차를 구해서, 적절한 성능을 가지는 부호책 설계를 위한 TS의 크기 결정에 대해 이야기해 보겠다. 어떠한 특정한 표본점 $\omega_1 \in \Omega$로 구성된 훈련 집합 $T_M^{\omega_1}$을 군집화하여 부호책 $\overline{C}^{\omega_1}$을 설계하였다고 하고 이는 식 (4)를 만족한다고 하자. $\overline{C}^{\omega_1}$을 검사하기 위해 새로운 N개의 표본 벡터를 F로부터 검사 집합을 만들기 위해 구성하면 검사비는 $\beta' = 1$ 이 된다. 다음에 이러한 N 검사 벡터만을 가지고 [16, 정리 2]에 근거하여 식 (5)로부터 OTSD의 근사치를 구할 수 있다. 즉 큰 값 N에 대해 왜곡치 차 $(OTSD-ITSD)^{\omega_1} = D_r(\overline{C}^{\omega_1}, F) - D_r(\overline{C}^{\omega_1}, F_M^{\omega_1})$는 다음과 같이 쓸 수 있다.

$$(OTSD-ITSD)^{\omega_1} \cong D_r(\overline{C}^{\omega_1}, F_N^{\omega_2}) - D_r(\overline{C}^{\omega_1}, F_M^{\omega_1}) \qquad (6)$$

이 식에서 검사 집합의 표본점 $\omega_2 \in \Omega$는 $\omega_2 \neq \omega_1$이다. 언급된 바와 같이 ITSD는 부호책 설계 알고리듬이 돌아가는 과정에서 구해지며, OTSD는 작은 크기를 가지는 검사 집합으로부터 계산될 수 있으므로 훈련된 부호책을 쉽게 검사할 수 있다. 이렇게 계산된 왜곡치 차로 부호책의 성능 평가가 가능하며 어떠한 임계치 ξ를 설정하여 $\xi \geq (OTSD-ITSD)^{\omega_1}$를 만족하는 최소 TS의 크기를 찾을 수 있어서 부호책 설계 계산량을 줄일 수 있다.

언급된 바와 같이 검사 집합의 원소는 N개만 있으면 충분하지만 반드시 TS와는 표본점이 다른, 즉 군집화에 사용된 표본 벡터를 다시 사용해서는 안된다. 이러한 내용을 보다 자세히 고찰해 보면 다음과 같다. 검사 집합을 얻는 방법을 다음과 같이 두 가지로 하였을 때에 대해 논해 보았다. 먼저 경우는 ``훈련 집합에서 임으로 추출, 재 사용"이며 다음은 ``훈련 집합의 일부를 사용하기 전에 남겨두는 방법"이다.

훈련 집합에서 임으로 추출, 재 사용

만일 N 개의 검사용 표본 벡터가 없으면, 부호책의 훈련에 사용된 훈련 벡터들을 다시 사용하는 방법을 생각할 수 있다. 그러나 결론적으로 말해서 이렇게 군집화에 사용된 벡터를 재사용하게 되면 OTSD의 참 값을 구할 수 없다. 이를 보다 구체적으로 논하면 다음과 같다.

어떠한 표본점 $\overline{\omega} \in \Omega$에 대해서 TS $T_M^{\overline{\omega}}$를 새로운 표본 공간으로 정의하자. 이 식에서 $\overline{M} \geq N$이다. 또한 $Z_j, j = 1, 2, \cdots, M'$를 i.i.d.(independent, identically distributed) 이산 랜덤 변수(discrete random variable)로 각 변수는 확률 함수로 $F_M^{\overline{\omega}}$를 가진다고 하자. 그러면 랜덤 변수 $\{Z_j\}$로부터 만들어진 재표본 집합(resampled set)을 $T_M^{\zeta} \triangleq \{Z_1, Z_2, \cdots, Z_M\}^{\zeta}$라 표기하면 실험적 확률 함수를 G_M^{ζ}로 정의할 수 있다. 이 때 표본점은 $\zeta \in T_M^{\overline{\omega}}$이다. 그러므로 이때의 시간 평균 왜곡치는 다음과 같다.

$$D_r(C, G_M^{\zeta}) = \frac{1}{M} \sum_{j=1}^{M'} \| Z_j^{\zeta} - Qc(Z_j^{\zeta}) \|^r \qquad (7)$$

이 식에서 $E\|X\|^r < \infty$ 라 가정한다.

추 론 2: 각각의 검사비 $\beta' \geq 1$ 에 대해서 어떠한 부호책 $C \in C_N$에 대해 다음과 같은 식이 성립된다.

$$ED_r(C, G_M) = D_r(C, F_M^{\overline{\omega}}) \qquad (8)$$

추론 2의 증명: 증명은 부록을 참조하기 바람.

추론 2는 왜곡치 $D_r(C, G_M^{\zeta})$가 $D_r(C, F_M^{\overline{\omega}})$의 바이어스되어 있지 않은 추정치임을 증명하고 있다. 또한 $\beta' \to \infty$ 함에 따라 확률 1로 $D_r(C, G_M^{\zeta}) \to D_r(C, F_M^{\overline{\omega}})$가 성립된다고 할 수 있다. 이에 대한 증명은 [16]의 부록 A와 부록 B에 소개된 것과 같은 방법으로 가능하다.

이제 식 (7)에 있는 왜곡치가 $D_r(C, F)$값에 대해 바이어스된 값이 될 수 있음을 하나의 예를 통해서 보였다. 먼저 왜곡치 측정(distortion measure)은 유크리디언 거리(Euclidian distance)로 r=2 이라 하자. 만일 TS $T_M^{\overline{\omega}}$으로 만든 부호책을 $\overline{C}^{\overline{\omega}}$ 로 표기하면, 부호책은 $D_2(\overline{C}^{\overline{\omega}}, F^{\overline{\omega}_M}) \cong \inf_{C \in C_N} D_2(C, F_M^{\overline{\omega}})$ 로 가정할 수 있다. 이러한 가정은 식 (4)와 같은 맥락에서 세워진 것이다. 그러면 M'개의 검사 벡터를 확률 함수 (probability mass function) $F_M^{\overline{\omega}}$에 따라 추출하면, 추론 2에 의해서 $ED_2(\overline{C}^{\overline{\omega}}, G_M) = D_2(\overline{C}^{\overline{\omega}}, F_M^{\overline{\omega}})$가 성립되며, 이 식에서 G_M^{ζ}는 검사 집합 T_M^{ζ}로 만들어진 시험적 확률 함수이다. 그러므로 $D_2(\overline{C}^{\overline{\omega}}, G_M^{\zeta})$는 $D_2(\overline{C}^{\overline{\omega}}, F_M^{\overline{\omega}})$의 바이어스되지 않은 추정치이다. 여기서 $\lim_{N \to \infty} \inf_{C \in C_N} D_2(C, F) = J_k \|f\|_k$가 성립된다고 가정하자[9]. 이 식에서 J_k는 벡터 차수 k의 함수이며, $\rho = k/(k+2)$이다. 그러면 $\overline{\omega}$에 따라서 부호책의 크기가 클 때, 다음과 같은 점근적 상한 경계치를 가진다[15].

$$D_2(\bar{C}^{w}, F_M^{w}) < N^{-2/k} J_k \| f \|_p \qquad (9)$$

결국 큰 N에 대해 $D_2(\bar{C}^{w}, F_M^{w}) < N^{-2/k} J_k \| f \|_p \le D_2(\bar{C}^{w}, F)$ 이 유도되며, 그러므로 $D_2(C, G_M^{r})$ 는 OTSD $D_2(\bar{C}^{w}, F)$ 의 바이어스가 되어 있는 추정치가 된다.

훈련 집합의 일부를 사용하기 전에 남겨 두는 방법

언급된 바와 같이 검사 집합을 TS 자체에서 뽑아낸다고 하면 OTSD $D(C,F)$를 주어진 부호책 C에 대해 구할 수는 없다. 그러므로 정확한 값을 구하기 위해서 다음과 같은 간단한 방법을 소개하였다.

먼저 검사 집합을 훈련 집합 T_M^{w} 으로부터 군집화하기 전에 그 일부분을 뽑아 둔다. 그리고 이 집합을 T_N^{w} 로 표기하면, $T_N^{w} \subset T_M^{w}$이다. 그리고 나서 부호책 $\bar{C}^{w}$ 를 다시 구성된 TS, 즉 $T_{M-N}^{w} \triangleq T_M^{w} - T_N^{w}$ 으로부터 군집화 알고리듬을 사용 설계한다. 그러므로 훈련비(training ratio)는 $\beta = \beta - 1$로 줄어든다. 결국 왜곡치 차는 $(OTSD - ITSD)^{w} = D_r(\bar{C}^{w}, F_N^{w}) - D_r(\bar{C}^{w}, F_{M-N}^{w})$ 로부터 계산을 할 수 있다. 이 식에서 F_N^{w} 와 F_{M-N}^{w} 는, T_N^{w} 과 T_{M-N}^{w}으로부터 각각 만들어진 실험적 확률분포이다. 다시 말하면 부호책의 크기가 큰 경우, N 개의 표본 벡터를 TS로부터 따로 뽑아내어 $(OTSD - ITSD)^{w}$를 계산할 수 있다. 그러나 훈련비는 $\beta = \beta - 1$로 줄어들게 되는데, 다행히도 이 정도는 β가 어느 정도 큰 경우에는 훈련된 부호책 성능에 그리 큰 영향을 미치지는 못한다.

Ⅳ. 모의 실험 결과

먼저 식(6)에 언급된 근사식에 대한 모의 실험을 i.i.d. 가우시언(Gaussian) 신호원에 대해 실행하였다. 이 실험 결과를 그림 1에 도시하였다. 이 실험에서 검사 집합의 크기는 $M' = 16384$인 경우와 $M' = 256$인 경우 두 가지인데, 보는 바와 같이 그 크기는 왜곡치 차를 구하는 데 있어서 중요한 요소가 되지 못한다. 이 실험에서 $k=16$ 그리고 $N=256$이다.

추론 2에 언급된 바이어스된 추정치에 대한 실험 결과를 그림 2에 나타내었다. 이 그림에서 "sample distortion"이라고 언급된 왜곡치는 여러 표본점 ξ에 대해 $\beta = 1$일 때 구해진 $D_2(\bar{C}^{w}, G_N^{\xi})$값들이며, 근사적으로 $D_2(\bar{C}^{w}, F_M^{w})$와 같음을 알 수 있다. 이 값은 그림 2에서 "55.30"으로 표기되어 있다. 그에 반해서 구하려고 하는 참값 $D_2(\bar{C}^{w}, F)$는 "59.11"로 표기되어 있는데, 이는 $D_2(\bar{C}^{w}, G_N^{\xi})$와 많은 차이를 보이고 있다. 즉 바이어스되어 있으므로 참값을 구할 수 없다는 결과를 보여주고 있다. 이 실험에서 $M' = 16384$ 이고 $\beta = 64$ 이다.

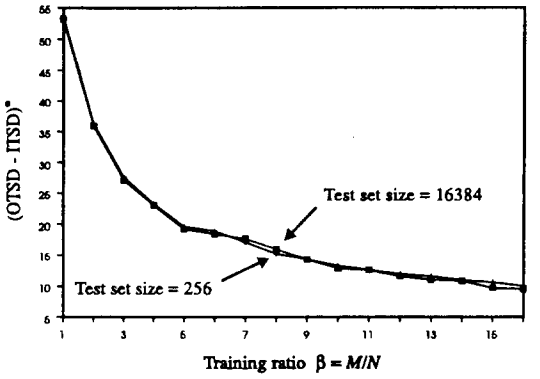

그림 1. 검사비와 왜곡치 차 $(OTSD - SD)^{w}$
Fig. 1. $(OTSD - SD)^{w}$ and test set size.

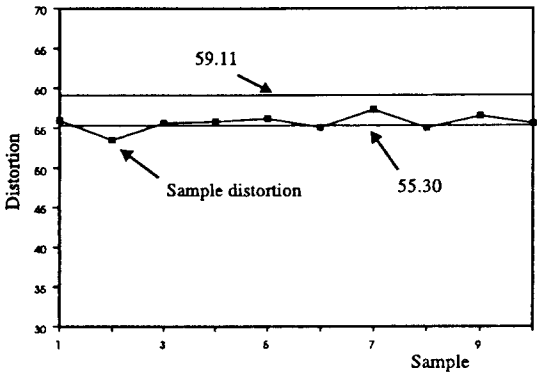

그림 2. 재표본 검사 집합과 시간 평균 왜곡치
Fig. 2. Time-averaged distortions for various resampled test set.

이제 실 영상 데이타에 적용하는 문제를 생각해 보자. [16] 및 본 논문에서 입력 신호원에 대한 가정은 TS 또는 검사 집합이 랜덤 표본(random sample)으로부터 구성된다는 것이다[2]. 즉 i.i.d. 랜덤 백터들로부터 어떠한 표본점에 대해 구성된다. 그러나 영상으로부터 얻어지는 데이타는 인접한 백터 간에 공간 영역(spatial domain)에서 높은 상관도(correlation)를 보이고 있음을 잘 알고 있다. 그러므로 독립인 검사 벡터를 얻기 위해서는 인접한 벡터 선택

을 피하여 가상적으로 독립인 벡터 집합을 만들어야 한다. [16]의 제 II장에서 언급된 바와 같이 GLA와 같은 군집화 알고리듬들은 그 성능이 입력되는 벡터들의 순서와는 무관하다. 그러므로 각 표본점마다, 영상 데이타에서 이웃해 있는 벡터들의 순서와는 무관하다. 그러므로 각 표본점마다, 영상 데이타에서 이웃해 있는 벡터 간의 순서에 무관하여, 즉 임의로 불규칙하게 벡터를 뽑는다고 하면 마치 영상 데이타로부터 서로 독립인 벡터를 추출한 것과 같다고 간주할 수 있다. 그러므로 이렇게 해서 구성된 집합으로 OTSD를 구하여 왜곡치 차 $(DTSD-ITSD)^{\alpha}$를 계산할 수 있다.

여러 영상 데이타에 대해 OTSD를 시간 평균을 통해서 계산한 결과를 그림 3에 도시하였다. 여러 가지 검사비에 대해서 실험을 하였는데, 각 검사 집합의 구성은 각 영상에서 임의의 위치에 있는 벡터를 필요한 갯수만큼 선정하였다. 즉 균일 분포를 가지는 랜덤 신호 발생을 통해서 벡터의 위치를 불규칙하게 선정하여 각 검사 집합을 구성하였다. 검사하기 위한 부호책은 LENA 영상을 사용하여 훈련되었으며, $N=256$, $\beta=64$, 그리고 $k=16$이다. 이 때 계산된 ITSD는 PSNR로 31.51 dB이다. 그림 3에서 보는 바와 같이 최소한 N개의 검사 벡터만 있으면 OTSD의 근사치를 구할 수 있음을 알 수 있다. 만일 어떤 부호책의 성능을 검사하기 위해서 여러 장의 TS 외부의 영상을 사용했던 경우에, 본 논문의 결과에 따른다면 여러 장의 영상으로부터 구성된 방대한 크기의 검사 집합 대신에 각 영상에서 몇 개의 벡터만을 임의로 추출하여 총 N개의 벡터만을 가지는 검사 집합을 구성해도 이전과 같은 부호책 검사가 가능하다.

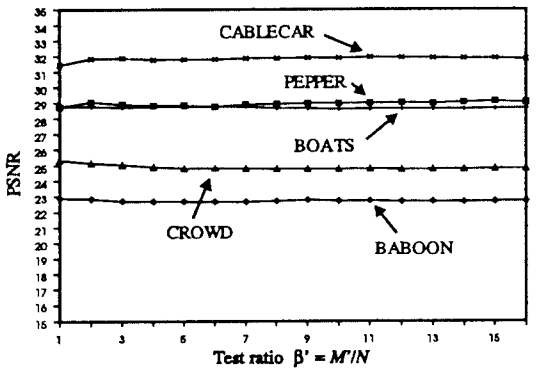

그림 3. 여러가지 실 영상에 대한 OTSD와 검사비 β'

Fig. 3. OTSD's and the test ratio β' for various real images.

한편 KNN(Kohonen neural network)과 같은 알고리듬은 입력되는 벡터 열의 순서에 그 성능이 민감하게 변한다[13-15]. 그러나 상관도가 존재하는 신호열을 불규칙하게 섞어 주면 알고리듬의 수렴 특성이 더 빨라지게 되어 오히려 앞에서 언급된 방법의 적용이 보다 용이하게 된다[15]. 아울러 본 논문에서 근사화를 위한 충분한 부호책의 크기는 대충 16 또는 32 이상이면 되는 것으로 실험적으로 관측되었다.

V. 결 론

본 논문에서는 막연히 많은 TS의 외부 벡터를 사용해서 훈련된 부호책을 검증하던 방법을 먼저 이론적인 분석 [16]을 중심으로 고찰하였으며, 아울러 간단히 부호책을 검사하는 방법을 제시하였다. OTSD와 ITSD의 왜곡치 차를 이용해서 훈련된 부호책을 검사할 때에는 먼저 ITSD를 구하는 단계, 즉 군집화 과정에서 설계된 부호책은 실험적 최적 왜곡치 또는 그와 근접한 왜곡치를 가져야 한다는 전제조건이 필요하다. 또한 OTSD를 시간 평균을 통해서 근사치를 얻는 경우 검사 집합의 크기가 크다고 해서 좋은 것은 아니며, 설계자가 원하는 확률적 특성을 모두 가지고 있으면서 그 크기는 부호책의 크기만큼만 되면 OTSD를 구할 수 있다는 사실을 밝혔다. 단 이때 부호책의 크기는 어느 정도는 커야 한다.

부 록

이 증명은 [16, 추론 2]의 증명과 유사하다.

$m_i^{(\omega)} \triangleq |S_i \cap \{X_j^{(\omega)}\}|$라 정의할 때, $m_i^{(\omega)} \neq 0$인 양자화 영역(quantizer region)에 대해 G_M으로부터 구성되는 개체 왜곡치는 다음과 같다.

$$\sigma_{r,i}(C, G_M) = \frac{1}{M} \sum_{j \in A_i} \| Z_j - y_i \|^r \qquad \text{(A1)}$$

이 식에서 그 원소가 랜덤인 인덱스 집합은 $A_i = \{j: Z_j \in S_i\}$이다. 랜덤 벡터 $Z_j, j=1, \cdots, N$에 수학적 평균을 취하면 식 (A1)은 다음과 같이 된다.

$$E\sigma_{r,i}(C, G_M) = \int \cdots \int \frac{1}{M} \sum_{j \in A_i} \| Z_j - y_i \|^r d(F_M^{(\omega)})^M \qquad \text{(A2)}$$

여기서 $d(F_M^{(\omega)})^M = dF_M^{(\omega)}(z_1) \cdots dF_M^{(\omega)}(z_M)$이며 매개변수(parameter) z_j들에 대해 $A_i = \{j: z_j \in S_i\}$이다. 그리고 [16]에서와 같은 방식으로 적분 구간을 나누

면 식 (A2)는 다음과 같이 다시 쓸 수 있다.

$$\sum_{i=1}^{NM}\int\cdots\int_{B_i}\frac{1}{M}\sum_{j\in\bar{A}_i}\|z_j-y_i\|^r d(F_M^n)^M \qquad (A3)$$

이 식에서 비공집합의 인덱스 집합을 $\bar{H}\triangleq\{i:m_i^{(n)}\neq 0\}$로 표기할 때, $\bar{N}\triangleq|\bar{H}|$ 이며 적분 구간의 Cartesian product는 $B_i\triangleq\Pi_{j=1}^{M}S_{j,\nu}$, $\nu=1,2,\cdots,\bar{N}^M$으로 정의된다. 인덱스집합은 $\ell_j^{(\nu)}\in\bar{H}$이다. 매개 변수들의 순서를 바꾸면 식 (A3)은 다음과 같이 된다.

$$\sum_{i=1}^{\bar{N}^M}\Pi\bar{P}_{\ell_j^{(\nu)}}\int_{S_i}\cdots\int_{S_i}\frac{1}{M}\sum_{j=1}^{M}\|z_j-y_i\|^r\,\bar{P}_i^{M-1}dF_M^n(\bar{z}_1)\cdots dF_M^n(\bar{z}_{mn})$$
$$=\sum_{i=1}^{\bar{N}^M}\Pi\bar{P}_{\ell_j^{(\nu)}}\frac{\bar{m}_i^{(\nu)}}{M}\int_{S_i}\|z-y_i\|^r\,\bar{P}_i^{-1}dF_M^n(Z), \qquad (A4)$$

이 식에서, $\bar{m}_i=|\bar{A}_i|$, $\sum_{i=1}^{\bar{N}}\bar{m}_i=M$. 그리고 $P_i=\int_{S_i}dF_M^n$, $i=1,\cdots,N$ 이다. 랜덤 벡터 $\bar{m}_i$가 다진 분포(multinomial distribution)[2]를 가지므로 다음 식이 성립된다.

$$\sum_{i=1}^{\bar{N}^M}\Pi\bar{P}_{\ell_j^{(\nu)}}\frac{\bar{m}_{j^{(\nu)}}}{M}=\bar{P}_i \qquad (A5)$$

그러므로 식 (A5)를 식 (A4)에 대입하면 각 $i, m_i^{(n)}\neq 0$에서 다음과 같은 결과를 얻는다.

$$E\sigma_{r,i}(C,G_M)=\int_{S_i}\|z-y_i\|^r dF_M^n(z)=\sigma_{r,i}(C,F_M^n) \qquad (A6)$$

이는 추론 1의 증명이 된다.

참 고 문 헌

[1] M.R. Anderberg, *Clustering Analysis for Applications*. New York: Academic Press, 1973.

[2] M.H. DeGroot, *Probability and Statistics*. New York: Addison-Wesley, 1984, 2nd. ed.

[3] T. Kohonen, *Self-Organizing and Associative Memory*. New York: Springer-Verlag, 1987.

[4] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers, 1992.

[5] R. Hecht-Nielsen, *Neurocomputing*. New York: Addison-Wesley, 1990.

[6] P. Gaenssler and W. Stute, "Empirical processes: A survey of results for independent and identically distributed random variables," *Annals of Probability*, vol.7, no.2, pp. 193-243, 1979.

[7] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, Jan. 1980.

[8] J.A. Bucklew and G.L. Wise, "Multidimensional asymptotic quantization theory with rth power distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-28, no.2, pp. 239-247, March 1982.

[10] E.A. Abaya and G.L. Wise, "Convergence of vector quantizers with applications to optimal quantization," *SIAM J. Appl. Math.*, vol.44, pp. 183-189, 1984.

[11] P. Cosman, K. Perlmutter, S. Perlmutter, R.A. Olshen, and R.M. Gray, "Training sequence size and vector quantizer performance," in *Proc. 25th Asilomar Conference on Signals, Systems, and Computers*, Nov. 1991, pp. 434-438.

[12] D.S. Kim and S.U. Lee, "A classified vector quantizer based on the minimum-distance partitioning," in *Proc. SPIE. Visual Commun. Image Processing*, Boston, Nov. 1991, pp 190-201.

[13] E. Yair, K. Zeger, and A.Gersho, "Competitive learning and soft competition for vector quantizer design," *IEEE Trans. Signal Processing*, vol. SP-40, no.2, pp. 294-309, Feb. 1992.

[14] D. Cohn, E.A. Riskin, and R. Ladner, "Theory and practice of vector quantizers trained on small training set," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-16, no.1, pp. 54-65, Jan. 1994.

[15] D.S. Kim, "Performance of vector quantizers and its relation to training-set size," Ph.D. dissertation, Dep. of

Control and Instrumentation, Seoul National Univ. Seoul, Korea, Feb. 1994.
[16] 김동식, "벡터 양자화에서 시간 평균 왜곡치의 수렴 특성: I. 대수 법칙에 근거한 이론," 전자공학회지, 심사 중, 1994

저 자 소 개

金 東 植(正會員)

1963年 10月 16日生. 1986年 2月 서울대학교 공과대학 제어계측공학과 졸업. 1988년 2월, 1994년 2월 서울대학교 대학원 제어계측공학과에서 각각 석사 학위와 박사 학위 받음. 주관심 분야는 벡터 양자화, 영상 데이타 감축, VLSI 신호처리, 패턴 인식 등임.