벡터 양자화 화자적응기법을 사용한 한국어 단어 인식

Korean Word Recognition Using Vector Quantization Speaker Adaptation

최 갑 석*

(Kap Seok Choi)

요 약

본 논문에서는 퍼지벡터양자화보다 양자화 왜곡을 더욱 저감시키기 위하여 에너지부분공간을 도입한 퍼지벡터양자화(energy subspace fuzzy vector quatization : ESFVQ)를 제안하였으며, 그것을 화자적응에 적용한 에너지부분공간 퍼지벡터양자화 화자적응기법에 의하여 미지화자의 한국어 단어를 인식하였다. 화자적응을 위한 학습과정에서 에너지 부분공간에 따른 퍼지 히스토그램으로, 사상코드북을 작성하였으며, 인식과정에서 미지화자의 음성을 ESFVQ에 의해 복호화하므로써 인식율의 향상을 도모하였다.

남성 2인과 여성 1인이 발성한 DDD 전화 지역명에 대하여 ESFVQ에 의한 양자화 왜곡 및 화자적응 단어 인식율을 측정하여 그 성능을 평가하였다. ESFVQ의 양자화 왜곡은 벡터 양자화보다 22% 감소되었으며, 퍼지 벡터 양자화보다 5% 감소되었다. 또한, ESFVQ에 의한 화자적응방법으로 인식한 결과, 화자적응을 고려하지 않은 방법보다 26%, 벡터 양자화에 의한 방법보다 11%의 향상된 인식율을 얻을 수 있었다.

ABSTRACT

This paper proposes the ESFVQ(energy subspace fuzzy vector quantization) that employs energy subspaces to reduce the quantizing distortion which is less than that of a fuzzy vector quatization. The ESFVQ is applied to a speaker adaptation method by which Korean words spoken by unknown speakers are recognized. By generating mapped codebooks with fuzzy histogram according to each energy subspace in the training procedure and by decoding a spoken word through the ESFVQ in the recognition procedure, we attempt to improve the recognition rate.

The performance of the ESFVQ is evaluated by measuring the quantizing distortion and the speaker adaptive recognition rate for DDD telephone area names uttered by 2 males and 1 female. The quatizing distortion of the ESFVQ is reduced by 22% than that of a vector quantization and by 5% than that of a fuzzy vector quantization, and the speaker adaptive recognition rate of the ESFVQ is increased by 26% than that without a speaker adaptation and by 11% than that of a vector quantization.

---

* 명지대학교 전자공학과 교수

(이 논문은 1990년도 문교부지원 한국학술진흥재단의 자유공모과제 학술연구조성비에 의하여 연구되었음.)

Ⅰ. 서 론

음성처리분야에서 불특정화자음성인식은 화자변동에 대처하기 위한 인식방법으로 특정화자음성인식에 비해 그 방법이 복잡하고 아직까지 그 성능이 낮은 실정이다[1~2].

이러한 불특정화자음성인식의 문제는 단어를 발성하는 화자가 바뀜에 따라 스펙트럼이나 조음방식이 서로 다르기 때문에 발생하게 된다[3]. 이에 대한 해결책으로 화자적응 기법을 이용한 음성인식이 연구되고 있으며, 성대와 스펙트럼과 성도의 길이를 정규화하는 방법[4], 입부의 음소로 부터 개인에 적응하는 전 음소의 스펙트럼을 추정하는 방법[5], 화자에 적응하는 표준패턴집합을 선택하는 방법[6], 벡터양자화에 의한 코드북의 사상방법[7], 사상코드북을 hidden Markov model(HMM)에 도입하는 방법[8] 등이 있다. 여기서 벡터 양자화에 의해 화자적응하는 기법은, 미지화자의 스펙트럼 공간을 벡터 양자화에 의해 집단화하여 유한 스펙트럼 공간을 생성하고 표준화자의 유한 스펙트럼 공간으로의 사상을 통하여 화자에 따라 서로 다른 스펙트럼 공간을 정규화해 주는 방법이며, 효과적인 화자적응을 위해 벡터 양자화에 의한 양자화 왜곡을 개선하려는 연구[9~14]가 계속되고 있다. 1989년 Shikano[10]는 퍼지 벡터 양자화를 스펙트로그램 정규화에 적용하여 양자화 왜곡을 저감시켰으며, 1990년 Matsuura[14]는 시간공간의 정보를 포함하면서 양자화 왜곡을 개선하기 위해 벡터 양자화에 부분공간법을 도입하여 불특정화자 음성인식에 적용한 결과 높은 인식율을 얻었다.

본 연구에서는 퍼지 벡터 양자화보다 양자화 왜곡의 저감을 더욱 정세화 하고 시간공간의 정보를 포함할 수 있는 에너지 부분공간을 도입한 퍼지 벡터 양자화(energy subspace fuzzy vector quantization : ESFVQ)를 제안하고, ESFVQ를 벡터 양자화 화자적응 기법에 적용하여 화자적응 단어 인식율을 향상시키고자 한다. 여기서 에너지 부분공간은 음성신호를 에너지 특징량에 따라 시간공간에서 분류하고 각 구간을 집단화하여 작성한다.

이 화자적응 음성인식방법을 평가하기 위해 남성 2인과 여성 1인이 발성한 DDD 전화 지역명을 대상으로 실험하였으며, 벡터 양자화, 퍼지 벡터 양자화, 및 본 연구에서 제안하는 ESFVQ에 의한 양자화 왜곡을 비교하였고, 남성과 남성 사이와 남성과 여성 사이의 화자적응 방법에 따른 인식율을 비교, 검토하였다.

Ⅱ. 에너지 부분공간 퍼지 벡터 양자화(ESFVQ)

본 장에서는 귀속도 함수를 이용하는 퍼지 집단화에 의하여 양자화 왜곡이 저감되는 것에 대하여 서술하고, 양자화 왜곡을 더욱 정세하게 개선하기 위하여 에너지 부분공간을 적용하는 퍼지 벡터 양자화에 관하여 논술한다.

2-1. 퍼지 ISODATA 집단화[15~18]

고전적인 crisp 집단화[15]에서는 각 벡터가 여러 집단중에서 하나의 집단에 지정되도록 집단화(clustering) 한다. 즉, crisp 집단화에서는 한 벡터가 어떤 집단에 속하는 지 아닌 지를 0과 1의 두 값으로 표현한다. 이에 반하여 퍼지 집단화에서는 한 벡터가 각 집단에 귀속된 정도를 0과 1 사이의 실수로 표현하여 애매성을 지닌 집단화를 행한다. 그 대표적인 예로 Bezdek[16~18]의 퍼지 ISODATA 알고리즘을 들 수 있다. 이 퍼지 집단화에서는 n개의 벡터 $x_k(k=1, \cdots, n)$가 c개 집단의 각 코드벡터 $v_i(i=1, \cdots, c)$에 귀속된 정도를 귀속도 μ_{ik}로 나타내었으며, 목적함수를 다음과 같이 정의하였다.

$$\min z(\mu, V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m d(x_k, v_i) \quad (1)$$

여기서, m은 애매도(fuzziness)이며, $d(x_k, v_i)$는 유클리드 거리이다.

식(1)을 최소로 하기 위하여 μ_{ik}를 고정시키고 v_i에 관하여 미분을 취하면 다음과 같은 집단의 코드벡터 v_i를 구할 수 있다.

$$v_i=\frac{\sum_{k=1}^{n}(\mu_{ik})^m x_k}{\sum_{k=1}^{n}(\mu_{ik})^m} \quad (2)$$

또한, 목적함수 식(1)에서 v_i를 고정시키고 μ_{ik}에 대하여 미분을 취하면 다음과 같이 귀속도를 구할 수 있다.

$$\mu_{ik}=\frac{\left[\frac{1}{d(x_k, v_i)}\right]^{\frac{1}{m-1}}}{\sum_{i=1}^{c}\left[\frac{1}{d(x_k, v_i)}\right]^{\frac{1}{m-1}}} \quad (3)$$

식(3)에서 애매도 m은 1.0보다 큰 실수로서, m이 ∞로 수렴하면 귀속도 μ_{ik}는 1/c이 되고, m이 1로 수렴하면 crisp 집단화를 행하여 종래의 벡터 양자화의 결과와 근사하게 된다.

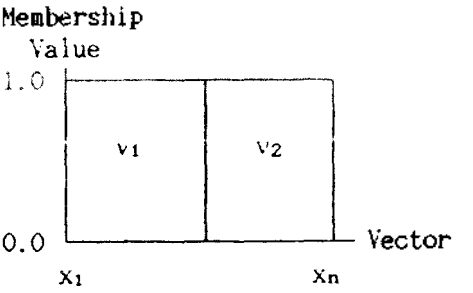
퍼지 집단화의 특성을 살펴 보기 위하여 2개의 집단에 대한 n개 벡터들의 집단화를 생각해 보자. crisp 집단화에서는 그림 1의 (a)와 같이 각 벡터 $x_k(k=1, \cdots, n)$가 어느 집단에 속했는지 아닌지를 0과 1의 두 값으로 표현하는 데 반하여 퍼지 집단화에서는 그림 1의 (b)와 같이 한 벡터 x_k가 각 집단에 귀속된 정도를 식(2)의 코드벡터와 식(3)의 귀속도를 이용하여 각 벡터 x_k의 귀속동 값을 0과 1 사이의 연속값으로 표현한다. 따라서, 퍼지 집단화에서는 귀속도 값으로 한 벡터가 각 집단에 귀속된 정도를 표현하므로 애매성을 지닌 집단화가 행해지며, 벡터를 복호화할 때 귀속도 값을 이용하므로 양자화 왜곡이 저감된다[10]. 이러한 벡터의 퍼지 집단화 방법을 퍼지 벡터 양자화라 한다.

퍼지 벡터 양자화에 의해 입력 벡터를 복호화하기 위해서는 입력 벡터 x_k에 대한 각 집단의 코드벡터와 귀속도를 구하여 합성한다. 식(2)에서 코드 벡터 대신에 입력벡터에 관하여 목적함수가 최소로 되도록 미분을 취하면 다음과 같이 귀속도 함수를 이용한 복호화 벡터 $\tilde{x}_k$을 얻을 수 있다.

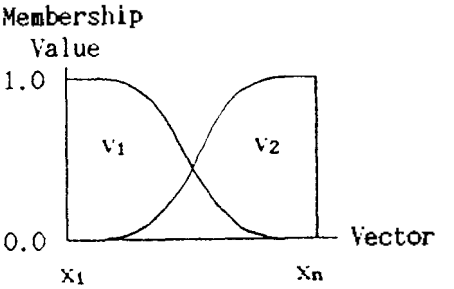
$$\tilde{x}_k=\frac{\sum_{i=1}^{c}(\mu_{ik})^m v_i}{\sum_{i=1}^{c}(\mu_{ik})^m} \quad (4)$$

2-2. 에너지 부분공간 퍼지벡터 양자화

음성 신호의 프레임 별 특징벡터를 유한 스펙트럼 벡터공간으로 한정지으려는 벡터 양자화에서는 음성 신호의 무한 스펙트럼 벡터공간에서 코드 벡터만을



(a) crisp 분할
(a) Crisp partition



(b) 퍼지 분할
(b) Fuzzy partition

그림 1. 집단의 분할
Fig. 1. Cluster partition

취해 주는 것이므로, 이 코드 벡터들이 시간공간의 정보를 포함하지 않으며, 특히 스펙트럼 특성의 과도적 변화부분에서는 양자화 왜곡이 크다. 이 때문에 벡터 양자화를 이용한 음성인식에서 인식율이 떨어지는 문제가 생긴다[10][14]. 이 문제를 해결하기 위하여 matrix 벡터 양자화[19]나 multi-section 벡터 양자화[20]를 적용하기도 하고 문헌[10]에서는 퍼지 벡터 양자화를 도입하여 벡터 양자화 왜곡이 저감되는 것을 잘 나타내고 있다. 본 연구에서는 퍼지 벡터 양자화의 도입에 의한 벡터 양자화 왜곡의 저감을 더욱 정세화하고, 벡터 양자화에 있어서 코드 벡터들의 시간공간정보의 부재로 인한 문제점 까지도 보완하기 위하여 음성신호의 에너지 특징량에 따른 부분공간(energy subspace : ES)을 작성하여 퍼지 벡터 양자화에 적용하는 에너지 부분공간 퍼지 벡터 양자화(energy subspace fuzzy vector quantization : ESFVQ) 방법을 제안한다. 에너지 부분공간은 음성

신호란 에너지 특성량에 따라 시간공간에서 분류하고 각 구간을 집단화하여 작성한 것이기 때문에 벡터 양자화 왜곡을 저감시킬 뿐만 아니라, 시간공간의 정보를 포함한다.

2-2-1. 에너지 부분공간의 작성

음성신호의 에너지 특징량에 따른 부분공간을 작성하기 위해 대수 에너지를 파라미터로하여 음성신호를 분류한다. 분류방법은 Atal[21] 등이 제안한 유성음과 무성음구간을 분류하는 방법과, Samber [22] 등이 제안한 시작구간과 끝구간을 분류하는 방법을 기초로 분류하며, 분류구간은 음성의 시점을 포함하는 제1에너지 구간, 유성음을 포함하는 제2 에너지 구간, 무성음을 포함하는 제3에너지 구간, 및 끝점을 포함하는 제4에너지 구간으로 한다. 이와 같이 음성신호를 네 구간으로 분류한 다음, 각 구간에 따라 집단화하여 에너지 부분공간을 작성한다. 다음은 에너지 부분공간을 작성하기 위해 각 구간으로 분류하는 과정이다.

(a) 인식할 미지단어의 프레임별 대수에너지 E(n)으로 부터 최대 에너지 E_{max}를 구하고, 무성음신호 평균에너지와 유성음신호 평균에너지의 비 ξ를 E_{max}에 곱하여 대수에너지의 경계치 E_{th}를 구한다.

$$E(n)=10\log_{10}\sum_{i=1}^{L}[s(i)]^2$$
$$E_{max}=\max[E(n)]\ (n=1,\cdots,N) \qquad (5)$$
$$E_{th}=\xi\times E_{max}$$

여기서, s(i)는 음성신호이며, L은 프레임 길이이며, N은 프레임의 총 수이다.

(b) 시점부터 대수 에너지 궤적이 상승하는 부분이 E_{th}를 초과하는 점까지를 제 1 에너지 구간 e_1으로 한다.

(c) 제 1 에너지 구간의 끝점부터 대수 에너지 궤적이 상승하였다가 하강하는 부분이 E_{th} 보다 큰 점까지를 제 2 에너지 구간 e_2로 한다.

(d) 제 2 에너지 구간의 끝점부터 에너지 궤적이 하강하였다가 다시 상승하는 부분이 E_{th} 보다 작은 점까지를 제 3 에너지 구간 e_3로 한다. 단, 음성신호의 끝점에 이르면, 끝점에서 부터 역으로 대수 에너지가 E_{th}를 초과하는 점까지를 제 4 에너지 구간 e_4로 하고, 분류 과정을 마친다.

(e) 제 3 에너지 구간의 끝점부터 대수 에너지 궤적이 상승하였다가 하강하는 부분이 E_{th} 보다 큰 점까지를 제 2 에너지 구간 e_2로 한다. 그리고, 단계(d)로 간다.

그림 2에 음성신호 "안성"을 이상의 과정에 의해 에너지 특징량에 따라 분할한 각 구간을 나타내었다.

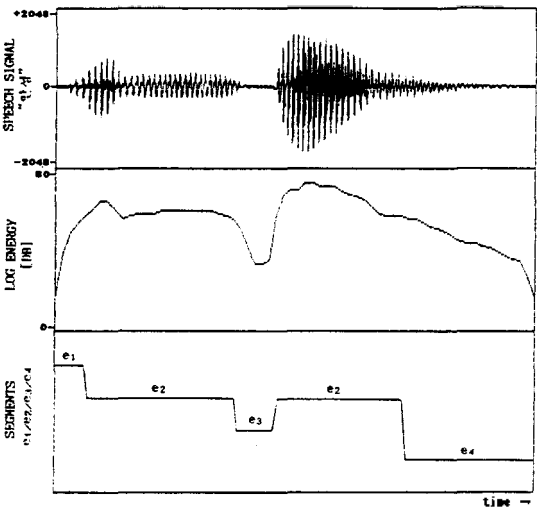


그림 2. "안성"의 에너지 부분공간을 위한 구간
Fig. 2. Segments for energy subspaces of speech "안성"

2-2-2. ESFVQ에 의한 양자화 왜곡의 개선

2-1 절에서는 퍼지 벡터 양자화에 의한 양자화 왜곡이 종래의 벡터 양자화에 의한 것보다 저감되는 것에 대하여 기술하였다. 본 절에서는 음성신호를 에너지 부분공간화하여 퍼지 벡터 양자화하므로써 양자화 왜곡이 더욱 정세하게 개선되는 것에 대하여 기술한다.

2-2-1 절에서 설명한 바와 같이 에너지 특징량에 의해 음성신호를 분류하고, 각 구간을 집단화 하여 에너지 부분공간을 작성한다. 입력 벡터열을 복호화할 때는 각 부분공간의 코드벡터에 대한 귀속도 함수를 이용하여 합성하므로, 복호화 식(4)에 의한 경우보다 더욱 정세하게 복호화된다. 각 에너지 구간에서 작성한 에너지 부분공간에 따른 퍼지 벡터

양자화의 복호화 벡터 $\tilde{x}(e_L)_k$는 퍼지 집단화의 복호화 과정에서의 코드벡터를 에너지 부분 공간에 따른 코드벡터로 대치한 식(6)에 의하여 얻어진다.

$$\tilde{x}(e_L)_k=\frac{\sum_{i=1}^{c}(\mu_{ik})^m\ v(e_L)_i}{\sum_{i=1}^{c}(\mu_{ik})^m} \qquad (L=1,\cdots 4) \quad (6)$$

여기서, μ_{ik}는 입력 벡터열에서 제 L 에너지 구간 e_L의 k번째 입력벡터 $x(e_L)_k$에 대한 제 L 에너지 구간 e_L에서 작성한 에너지 부분공간의 i번째 코드벡터 $v(e_L)_i$의 귀속도이고, m과 c는 각각 애매도와 코드북사이즈이다.

식(6)과 같이 ESFVQ에 의해 입력 벡터를 복호화하면 각 에너지 구간에 따른 부분공간을 이용하기 때문에 전체 공간을 이용하는 퍼지 벡터 양자화보다 양자화 왜곡을 저감시킨다. 여기서 귀속도를 구하기 위한 거리척도는 다음 식과 같이 Itakura가 제안한 likelihood ratio[23]를 이용하여 측정한다.

$$d(x,\ v)=\frac{a_x^T R_v a_x}{a_v^T R_v a_v}-1.0 \qquad (7)$$

여기서, a_x는 미지벡터 x의 선형예측계수벡터, a_v는 코드벡터 v의 선형 예측계수벡터, R_v는 코드벡터의 자기상관계수벡터이다.

Ⅲ. ESFVQ에 의한 화자적응 단어인식

화자적응을 위해 음성신호의 스펙트럼을 사상하는 것은 미지화자 음성의 스펙트럼 공간을 표준화자 음성의 스펙트럼 공간으로 정규화하는 방법이다. 벡터 양자화에 의한 화자적응 방법[7]에서는 두 화자의 스펙트럼 공간을 유한 스펙트럼 공간으로 벡터 양자화하여 화자정규화하였으며, 화자정규화의 고정도화를 위하여 벡터 양자화 대신 양자화 왜곡을 저감시킬 수 있는 퍼지 벡터 양자화를 도입하였다[10~11].

본 연구에서는 퍼지 벡터 양자화보다 양자화 왜곡을 더욱 저감 시키며 시간공간의 정보를 포함할 수 있는 ESFVQ를 화자정규화에 적용하여 단어인식율을 향상시키고자 한다.

3-1. 퍼지 히스토그램

벡터양자화에 의한 화자정규화에서 미지화자와 표준화자의 유한 스펙트럼공간을 사상할 때 코드벡터 대응확률로 사용하는 히스토그램 누적 방법에서는 DTW의 최적경로에 따라 미지화자의 코드벡터와 표준화자의 코드벡터가 대응될 때마다 1을 히스토그램에 누적하여 대응확률을 구하는 데 반하여 에너지 부분공간에 따른 퍼지 히스토그램 누적 방법에서는 DTW의 최적경로에 따라 대응하는 미지화자와 표준화자의 코드벡터 사이의 귀속도를 누적하여 주므로써 종래의 히스토그램 누적 방법에서 보다 정확하게 대응확률을 구할 수 있다. 다음 식(8)은 DTW의 최적경로에 따라 미지화자와 표준화자의 각 에너지 부분공간에 따른 코드벡터가 대응되었을 경우 귀속도를 구하는 식이다.

$$\mu(e_L)_{ij}=\frac{\left(\frac{1}{d(v(e_L)_i^{(A)},\ v(e_L)_j^{(B)})}\right)^{\frac{1}{m-1}}}{\sum_{i=1}^{c}\left(\frac{1}{d(v(e_L)_i^{(A)},\ v(e_L)_j^{(B)})}\right)^{\frac{1}{m-1}}} \qquad (8)$$

여기서, $v(e_L)_i^{(A)}$는 미지화자 A가 발성한 음성에서 제 L 에너지 부분공간의 i번째 코드벡터이며, $v(e_L)_j^{(B)}$는 표준화자 B가 발성한 음성에서 제 L 에너지 부분공간의 j번째 코드벡터이다.

이와 같이 미지화자의 코드벡터와 표준화자의 코드벡터가 대응할 때 이두 코드벡터 사이의 귀속도를 식(8)과 같이 구하고 다음 식과 같이 에너지 부분공간에 따른 퍼지 히스토그램을 누적한다.

$$h(e_L)_{ij}=h(e_L)_{ij}+\mu(e_L)_{ij} \qquad (9)$$

여기서, $h(e_L)_{ij}$는 $v(e_L)_i^{(A)}$와 $v(e_L)_j^{(B)}$ 사이의 퍼지 히스토그램이다.

3-2. 단어인식 과정

ESFVQ를 도입한 화자정규화에서는 에너지 부분공간 사이의 퍼지 히스토그램을 대응확률로 이용하여 에너지 부분공간 사상코드북을 작성하며, 미지화

자의 코드북을 사상코드북으로 대치하고 ESFVQ에 의해 복호화하므로써 화자정규화한다. 그림 3은 미지화자 A의 에너지 부분공간으로 부터 표준화자 B의 에너지 부분공간으로의 에너지 부분공간 사상코드북을 보이고 있다.

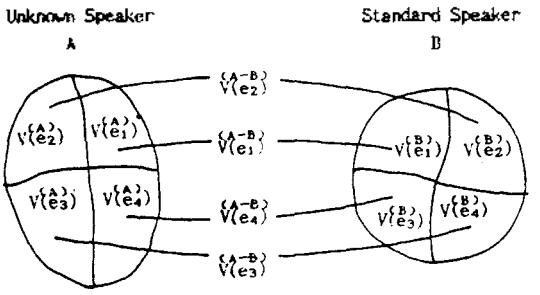


그림 3. 에너지 부분공간 사상코드북
Fig. 3. Mapped codebooks of energy subspaces

3-2-1. 학습과정

그림 3의 에너지 부분공간 사상코드북은 에너지 부분공간에 따른 퍼지 히스토램을 가중치로 하여 표준화자의 에너지 부분공간의 코드북과 선형결합하므로써 얻는다. 다음식은 미지화자 A로 부터 표준화자 B로의 에너지 부분공간 사상코드북을 얻는 식이다.

$$v(e_L)_i^{(A-B)} = \frac{\sum_{j=1}^{c} h(e_L)_{ij}\, v(e_L)_j^{(B)}}{\sum_{j=1}^{c} h(e_L)_{ij}} \qquad (10)$$

여기서,

$v(e_L)_i^{(A)}$는 미지화자 A의 제 L 에너지 부분공간의 i번째 코드벡터, $v(e_L)_j^{(B)}$는 표준화자 B의 제 L 에너지 부분공간의 j번째 코드벡터, $v(e_L)_i^{(A-B)}$는 미지화자 A의 코드북을 표준화자 B의 코드북으로 대응시키는 각 에너지 부분공간에 따른 사상코드북의 i번째 코드벡터, $h(e_L)_{ij}$는 $v(e_L)_i^{(A)}$와 $v(e_L)_j^{(B)}$ 사이의 퍼지 히스토그램, c는 코드북사이즈이다.

다음식은 식(10)의 에너지 사상코드북을 작성하기 위한 학습과정이다.

(a) 미지화자의 학습단어의 에너지 부분공간을 작성하여 에너지 부분공간에 따른 코드북을 생성한다.

(b) 에너지 부분공간에 따른 미지화자의 코드북과 표준화자 코드북 사이의 거리행렬을 구한다.

(c) 표준화자의 에너지 부분공간에 따른 코드북으로 벡터양자화한 표준패턴과 미지화자의 에너지 부분공간에 따른 코드북에 의한 시험패턴의 DTW를 행하여 동일한 학습단어간에 최적경로를 구한다.

(d) 최적 경고에 따라 대응하는 코드 벡터들의 에너지 부분공간에 따른 퍼지 히스토그램을 구한다.

(e) 식(10)에 의해 에너지 부분공간 사상코드북을 구한다.

(f) 미지화자의 에너지 부분공간에 따른 코드북을 (e)과정의 에너지 부분공간 사상코드북으로 대치하여 (b)에서 (e)과정을 반복한다. 단, (c)과정에서 DTW에 의해 구해지는 왜곡척도가 수렴하면 학습을 끝낸다.

그림 4는 에너지 부분공간에 따른 퍼지 히스토그램에 의해 각 부분공간의 사상코드북을 작성하는 학습과정의 구성도이다.

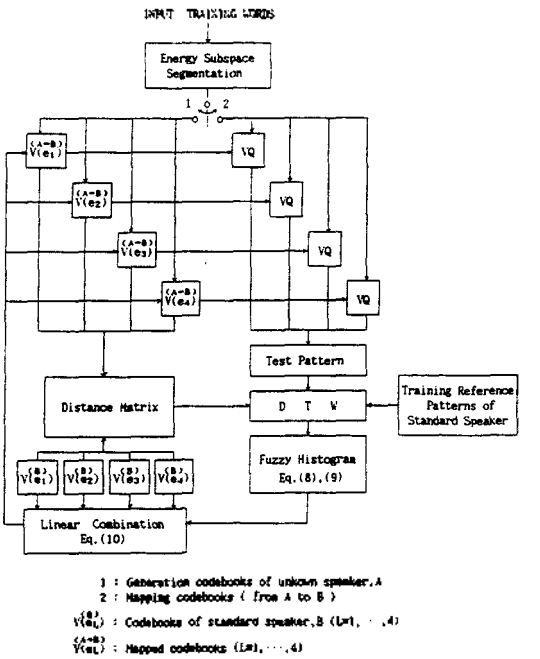


그림 4. 에너지 부분공간의 퍼지 히스토그램에 의한 사상코드북 학습과정
Fig. 4. Training procedure for mapped codebooks using fuzzy histogram of energy subspaces

3-2-2. 인식과정

식(3)과 같이 에너지 부분공간 코드북이 작성되면 미지화자의 에너지 부분공간에 따른 코드북을 에너지 부분공간 사상코드북으로 대치하여 ESFVQ에 의해 복호화함으로써 화자정규화를 수행한다. 다음 식은 ESFVQ에 의해 복호화하여 입력벡터를 화자정규화하는 식이다.

$$x(e_L)_k^{(A-B)}=\frac{\sum_{i=1}^{C}(\mu_{ik})^m\, v(e_L)_i^{(A-B)}}{\sum_{i=1}^{C}(\mu_{ik})^m}\quad (L=1,\ \cdots,\ 11)$$

여기서 ESFVQ에 의한 화자정규화는 두 화자의 에너지 부분공간의 코드북 사이에서 퍼지 히스토그램으로 사상코드북을 작성하므로 전체공간에서 사상코드북을 작성하여 정규화하는 것보다 고정도화된 화자정규화가 가능하다.

ESFVQ에 의한 화자적응 단어 인식은 미지화자의 학습단어가 입력되면 먼저 2-2절에서 설명한 바와 같이 에너지 부분공간을 작성하며 3-2-1절에서의 식(10)에 의해 에너지 부분공간 사상코드북을 작성하는 학습과정을 행하고, 학습과정에서 작성된 에너지 부분공간 사상코드북으로 입력된 미지화자의 미지단어를 식(11)에 의해 복호화하여 화자정규화한 후, DTW와 KNN 결정규칙에 의해 표준화자의 다수 표준패턴과의 최소왜곡을 선택하여 단어 인식을 수행한다. 다음은 ESFVQ에 의해 화자정규화하고 미지화자의 입력음성을 인식하는 인식과정이다.

(a) 미지화자의 음성이 입력되면 에너지 부분공간을 작성하고 에너지 부분공간에 따른 귀속도 값을 구한다.

(b) 학습과정에서 작성한 에너지 부분공간 사상코드북을 미지화자와 에너지 부분공간에 따른 코드북과 대치하고 식(11)에 의해 복호화하여 화자정규화한다.

(c) 화자정규화된 미지화자의 입력단어를 DTW와 KNN 결정규칙 [13][14]에 의해 인식한다.

즉 미지화자의 에너지 부분공간 코드북 대신에 학습과정에서 작성한 에너지 부분공간 사상코드북을 이용하여 화자 정규화를 행하고 표준화자의 표준패턴에 대하여 인식을 수행한다. 그림 5는 이러한 화자적응 단어인식과정의 구성도이다.

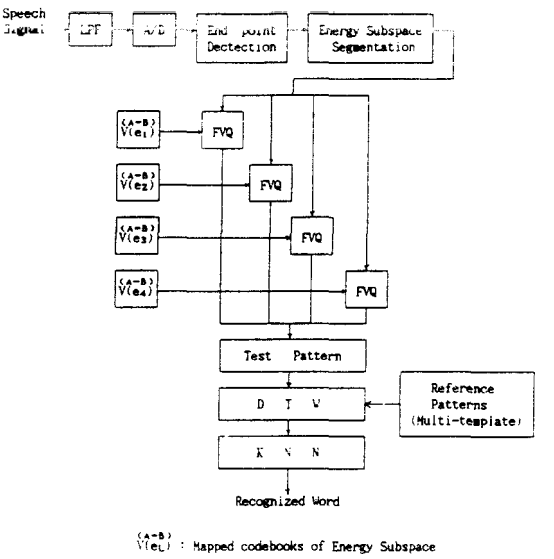


그림 5. ESFVQ에 의한 화자적응 단어인식의 구성도.
Fig.5. Block diagram of speaker-adaptive word recognition using ESFVQ.

Ⅳ. 실험결과 및 고찰

4-1. 음성 자료

실험에서 사용된 음성은, 방음처리가 되지 않은 실내에서 남성 2인과 여성 1인의 화자에 의해 자연스럽게 4회씩 발성한 DDD 전화 지역명(3인×5회×148 개=2220 개) 들 이다. 이 중에서 표준화자와 미지화자의 학습단어는 각 화자가 DDD 전화 지역명을 1회 발성한 것(154 개)을 사용하였으며, 학습과정에서 사용하는 DTW는 기울기가 1이고 대칭 형태인 경로를 이용하였다. 나머지 3회씩 발성한 것(3회×148개=444 개)으로는 KNN 결정규칙에 이용하기 위해 각 화자마다 세 개의 표준패턴을 작성하였으며, 이 세개의 표준패턴을 각 미지화자의 시험패턴으로 사용하였다. 이러한 음성들을 차단주파수가 3.4 kHz인 저역통과 필터에 통과시킨 후, 샘플링 주파수가 10 kHz인 AD 콘버터(12-bit resolution)로 샘플링 하였다.

음성신호의 프레임 구간은 20.0ms(200 샘플)로 하였으며 이동구간은 10.0ms(100 샘플)로 하여 5

0%의 감소효과를 보이고 있다. 벡터 양자화를 위한 코드북을 벡터 특징벡터는 1차 선형예측계수벡터와 자기상관계수벡터로 하였으며, 에너지 부분공간을 분류하기 위한 에너지 특징량으로는 0차 자기상관 계수를 사용하였다. 여기서 각 화자의 에너지 부분공간에 따른 코드북은 LBG 알고리즘[25]에 의하여 생성하였다.

4-2. 양자화 왜곡의 비교

ESFVQ에 의해 복호화할 때 코드북 사이즈가 적은 데 비해 양자화 왜곡을 저감시킬 수 있는 애매도 m을 결정하기 위해 애매도 m과 코드북사이즈를 변화시키면서 여성이 발성한 28개의 DDD 전화 지역명(직할시 이상 6개와 경기도 내의 22개 지역명의 합)을 대상으로 퍼지 벡터 양자화에 의한 양자화 왜곡을 식(7)에 의하여 측정하였으며 그림 6에 나타내었다.

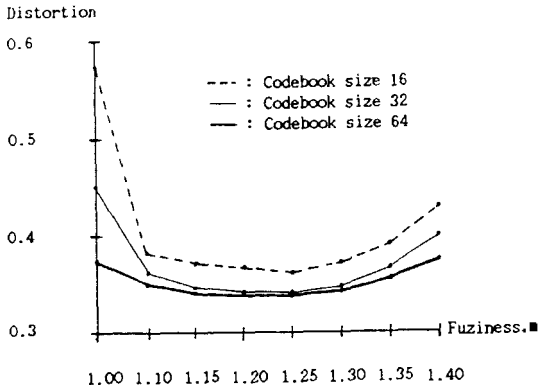


그림 6. 애매도 m에 따른 퍼지 벡터 양자화의 왜곡
Fig. 6. Fuzzy VQ distortion versus fuzziness, m

그림 6에서 m=1.0일 때의 양자화 왜곡은 2-1절에서 서술한 바와 같이 종래의 벡터 양자화에 의한 왜곡으로 측정하였다. 그림 6의 곡선에서 양자화 왜곡이 가장 작을 때의 애매도 m은 1.25일 때 임을 알 수 있으며, 코드북 사이즈가 16 일 때 m=1.25일 때의 양자화 왜곡은 m=1.0 일 때의 양자화 왜곡보다 37% 감소되었으며, 코드북사이즈가 32일 때는 25%, 코드북사이즈가 64일 때는 9% 감소되었다.

또한, ESFVQ에서 애매도 m을 1.25로 할 때 식(6)에 의한 복호화 벡터의 양자화 왜곡이 벡터 양자화나 에너지 부분 공간을 사용하지 않는 퍼지 벡터 양자화보다 개선됨을 보이기 위해 코드북사이즈의 크기를 변화시키면서 각 방법에 따른 양자화 왜곡을 측정하였으며, 그림 7에 나타내었다.

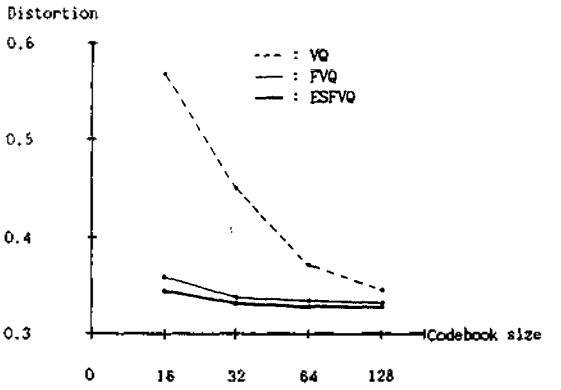


그림 7. 각 방법에 따른 양자화 왜곡(m=1.25)
Fig.7. Quantizing distortion for each method(m=1.25)

그림 7에서 16부터 128까지의 코드북사이즈에 대한 퍼지 벡터 양자화의 왜곡은 벡터 양자화의 왜곡보다 평균 18.5% 감소되었으며, ESFVQ에 의한 왜곡은 벡터 양자화의 왜곡 보다 평균 22.5% 감소되어 퍼지 벡터 양자화보다 평균 4.0% 개선되었다. 특히 코드북사이즈가 32일 때 애매도 m이 1.25인 ESFVQ에 의한 양자화 왜곡은, 애매도가 같으나 코드북사이즈가 64일 때의 퍼지 벡터 양자화 방법보다 4.8% 감소된 양자화 왜곡을 보이므로 ESFVQ에 의해 음성신호를 복호화하면 코드북사이즈를 적게하면서 양자화 왜곡을 개선할 수 있음을 확인하였다. ESFVQ에서는 퍼지 벡터 양자화와 마찬가지로 귀속도를 이용하여 복호화벡터를 합성할 뿐만 아니라 시간공간에서 에너지 특징량에 따른 부분공간을 작성하기 때문이라 생각된다.

4-3. 화자 적응 인식율 결과 및 고찰

본 절에서는 3장에서 서술한 ESFVQ에 의한 화자적응 단어 인식 방법을 평가하기 위해 종래의 벡터 양자화에 의한 화자적응 방법[13]의 인식율과 비교하였다. 여기서 사용하는 애매도 m과 코드북사이즈

는 4.2절의 비교 실험에서 실정한 바와 같이 각각 1.25와 32로 하였다. 표 1에는 각 방법에 따른 인식율을 나타내었다.

표 1. 인식율의 비교
Table 1. Comparosin of recognition rates

METHODS	RECOGNITION[%]		
	MALE1·MALE2	FEMALE1 MALE2	AVERAGE
Without adaptation	58.3	43.3	50.8
VQ+ Histogram	70.2	59.6	64.9
ESFVQ+Fuzzy Histogram	81.1	70.2	76.1
Speaker dependent	97.8	95.4	96.6

여기서,

Without adaptation은 화자적응 학습을 하지 않은 표준 화자의 코드북으로 미지화자의 음성을 인식하는 방법, VQ+Histogram은 히스토그램에 의한 학습과정으로 사상코드북을 작성하고 벡터 양자화에 의해 화자적응 단어인식하는 방법, ESFVQ+Fuzzy Histogram은 에너지 부분 공간에 따른 퍼지 히스토그램에 의한 학습과정으로 사상코드북을 작성하고 ESFVQ에 의해 화자적응 단어인식하는 방법, Speaker dependent는 표준화자의 특정화자 단어인식 방법이다.

화자적응을 하지 않는 Without adaptatin 방법에서는 남성간 58.3%의 인식율을 얻었으며 남녀간에는 43.3%의 인식율을 얻었다. 여기서, 남녀간의 인식율이 남성간의 인식율 보다 낮은 것은 화자정규화 하지 않은 경우 같은 음성이라도 남녀간의 피치주기나 공진주파수의 차이가 크기 때문에 생기는 결과로 생각된다.

종래의 VQ+Histogram에 의한 화자적응 단어인식 방법에서는 남성간 70.2%의 인식율과 남녀간 59.6%의 인식율을 얻으므로써 Without adaptation 방법보다 평균 인식율이 14.1% 개선되었으며, 본 연구에서 제안한 ESFVQ+Fuzzy Histogram에 의한 방법에서는 남성간 81.1%의 인식율과 남녀간 70.2%의 인식율을 얻으므로써 Without adaptation 방법보다 평균 25.7%, VQ+Histogram에 의한 방법에서 VQ+Histogram 방법보다 인식율이 높은것은, 학습과정에서 에너지 부분공간에 따른 Fuzzy Histogram에 의한 대응확률은 부분공간을 사용하지 않은 Histogtam에 의한 것보다 정확하게 추정할 수 있으며, 인식과정의 DTW와 KNN 결정규칙에서는 평균 왜곡척도가 최소인 것을 선택하여 인식하기 때문에 VQ에 의해 복호화하여 인식하면 에너지 부분공간이 시간 공간에서 분류된 에너지 부분공간을 이용하여 양자화 왜곡이 더욱 작게된 ESFVQ에 의해 복호화할 때보다 오인식이 많은 것으로 생각된다.

V. 결 론

종래의 벡터 양자화를 이용하여 화자적응하는 방법에 본 연구에서 제안하는 ESFVQ와 퍼지 히스토그램을 도입하여 화자적응 단어인식을 수행하였다. 학습과정에서는 에너지 부분공간에 따른 퍼지 히스토그램을 사용하여 사상코드북을 작성하였으며, ESFVQ에 의한 화자적응 단어인식에서는 미지화자의 입력단어를 학습과정에서 작성된 사상코드북에 의해 복호화 하였다.

본 연구에서 제안한 ESFVQ에 의한 양자화 왜곡은 종래의 벡터 양자화에 의한 것에 비해 22.5% 감소시킬 수 있었으며, 코드북사이즈를 32개로 하여 ESFVQ 방법으로 입력벡터를 복호화한 경우 코드북사이즈를 64개로 하여 퍼지 벡터 양자화 방법으로 입력벡터를 복호화했을 때보다 양자화 왜곡을 4.8% 감소시키므로써 적은 코드북사이즈로 양자화 왜곡을 개선할 수 있는 효과를 확인하였다.

또한, 남성 2인과 여성 1인이 발성한 음성을 ESFVQ 에 의해 화자적응 단어인식을 수행한 결과, 남성간에는 81.1%, 남녀간에 평균 70.2%의 인식율을 얻었으며, 화자적응 방법을 고려하지 않은 단어인식율 보다 평균 25.7%, 벡터 양자화에 의한 방법보다 평균 11.2%의 향상된 인식율을 얻을 수 있었다. 이러한 결과로 부터 다음과 같은 결론을 얻을 수 있었다.

1. 에너지 부분공간을 이용하는 ESFVQ에 의하여 입력벡터를 양자화하면 종래의 벡터 양자화나 퍼지 벡터 양자화보다 양자화 왜곡이 개선된다.

고 ESFVQ의 퍼지 히스토그램을 이용한 묘사 화자 적응 단어인식 방법으로 미지화자의 음성을 인식하면 종래의 벡터 양자화에 의한 방법보다 인식율이 향상된다.

본 연구에서는 부분공간의 작성을 용이하게 하기 위해 예비적 부강방을 사용하였으나, 공통적인 음향적 특성을 보다 많이 포함할 수 있도록 작성방법을 개선하면 양자화 왜곡을 더욱 감소시킬 수 있으며, 화자적응이 효과적으로 수행되어 인식율도 더욱 향상될 것으로 기대된다.

참 고 문 헌

1. V.Steinbiss, et al., "A 10,000 word continuous speech recognition system", Proc. ICASSP 90, S2.5, 1990.
2. K.F.Lee, H.W.Hon, "Large vocabulary speaker independent continuous speech recognition usingf HMM", Proc. ICASSP 88, S3.7, 1988.
3. K. Shikano, et al., "Recent trends on speech recognition", The Journal of the Institute of Electronics, Information and Communication Eng., Vol.71, No.11, pp. 1170~1181, 1988.
4. H.Matsumoto, H. Wakita, "Vowel normalization by frequency warped spectral Matching", Speech Comm., 5,2, pp. 239~251, Jun. 1986.
5. S.Furui, "A training procedure for isolated word recognition systems", IEEE Trans., Acoust. Speech, Sigal Processing, Vol. ASSP-28, pp. 128~136, Apr. 1980.
6. R.Mizoguchi, M.Kinoshita, O.Kakusho, "Word recognition system for unspecified speakers based on interrelated phoneme templates", The Trans. of the Inst.of Electronics and Comm. Eng.of Japan, Vol.J67-A, No.6, Jun. 1984.
7. K.Shikano, K.F.Lee, R.Reddy, "Speaker adaptation through vector quantization", Proce. ICASSP 86, 19.5, 1986.
8. R.Schwartz, Y-1.Chow, F.Kubala, "Rapid speaker adaptation using a probabilistic spectral mapping", Proc. ICASSP 87, 15.3, 1987.
9. M.W.Feng, "Improved speaker adaptation using text dependent spectral mapping", Proc. ICASSP 88, S3.9, 1988.
10. S.Nakamura, K.Shikano, "spectrogram normalization using fuzzy vector quantization", The Journal of the Acoustic Society of Japan, Vol.45, No.2, 1989.
11. S.Nakamura, K.Shikano, "VQ-based speaker adaptation applied to HMM phoneme recognition", The Journal of the Acoustic Society of Japan, Vol.45, No.12, 1989.
12. S.Nakamura, K.Shikano, "A comparative study of spectral mapping for speaker adaptation", Proc. ICASSP 90, S3.7, 1990.
13. 최갑석, 이기영, "사상코드북을 이용한 화자적응 한국어 숫자음 인식에 관한 연구", 한국음향학회, 제 9권 5호, pp. 13~49, 1990.
14. H.Matsuura, J.Iwasaki, T.Nitta, "Speaker independent word recognition based on SM-HMM", Speaker independent word recognition based on SM-HMM", 信學技報, Vol.89, No.43, SP90-68, 1990.
15. G.Ball, D.Hall, "A clustering technique for summarizing multivariate data", Behav. Soc., Vol.12, pp. 153~155, 1967.
16. J.C.Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithm", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. PAMI-2, pp. 1~8, 1980.
17. H, Tseng, M.J.Sabin, E.A.Lee, "Fuzzy vector quantization applied to hidden Markov modeling", Proc. ICASSP 87, 15.5, 1987.
18. Yoh-Han Pao, Adaptive Pattern Recognition and Newral Networks, Addison-Wesley Pub. Com., Inc., pp. 57~82, 1989.
19. D.K.Burton, "Applying matrix quantization to isolated word recognition", Proc. 1984 IEEE Int. Conf. Acoust., Speech Signal Processing, 1.8, 1984.
20. D.K.Burton, J.E.Shore, J.T.Buck, "Isolated-word speech recognition using multisection vector quantization codenooks", IEEE Trans. Acoust. Speech, Signal Processing, Vol.ASSP-33, No.4, pp. 837~849, Aug. 1985.
21. B.S.Atal, L.R.Rabiner, "A pattern recognition approach to voice-unvoice-silence classification with applications to speech recognition", IEEE Trans. Acoust. Speech, Signal Processing, Vol.ASSP-24, No.3, Jun. 1976.
22. M.R.Sambur, L.R.Rabiner, "A speaker-independent digit-recognition system", B.S.T.J., Vol.54, No.1, Jan. 1975.
23. F.Itakura, "Minimum prediction residual principles applied to speech recognition", IEEE Trans., Acoust.Speech, Signal Processing, Vol.ASSP-23, pp. 52~72, Feb. 1975.

24. J.G.Wilpon, L.R. Rabiner, "A modified K means clustering algorithm for use in isolated word recognition", IEEE Trans. Acoust,Speech, Signal Processing, Vol.ASSP 33, Jun. 1985.

▲최 갑 석(Kap Seok Choi) 1930년 9 월30일생

1951년 9 월~1955년 3 월 : 서울대학교공과대학 통신공학과 졸업

1975년 2 월25일 : 명지대학교 공학박사

1980년 8 월~1981년 8 월 : 일본 동경대학 공학부 전자공학과 연수

1972년 3 월~현재 : 명지대학교 전자공학과 교수