한국음향학회지 제22권 제4호 pp. 299~307 (2003)

DHMM 음성 인식 시스템을 위한 양자화 기반의 화자 정규화

Quantization Based Speaker Normalization for DHMM Speech Recognition System

신 옥 근*
(Ok-Keun Shin*)

한국해양대학교 · 자동차정보공학부*

(접수일자: 2003년 1월 8일; 수정일자: 2003년 5월 6일; 채택일자: 2003년 5월 19일)

화자독립 음성인식기에서 화자사이의 성도 길이의 영향을 최소화시켜 인식 성능을 개선하는 화자 정규화에 대한 많은 연구가 있어 왔다. 본 연구에서는 벡터양자화기를 이용하여 화자 검증이 가능하다는 사실에 착안하여 벡터 양자화기를 이용한 비교적 간단한 선형 워핑 화자정규화 방법을 제안한다. 제안하는 방법에서는 먼저 정규화에 이용될 최적의 코드 북을 생성한 다음, 이 코드 북을 이용하여 화자의 선형 워핑계수를 추출하고 추출된 워핑계수는 멜 켑스트럼 추출시에 사용되는 멜 스케일 필터뱅크를 워핑하기 위해 이용된다. 본고에서 제안한 워핑계수 추출 및 적용 방법의 성능을 확인하기 위해 이산 HMM을 이용한 13가지의 단음절 한글 숫자음 인식기를 이용하여 인식실험을 수행하였으며, 실험 결과 약 29%의 오인식률 감소를 보여 제안하는 화자 정규화방법이 다른 라인서치 워핑계수추출 방법보다 간단한 동시에 효용가치가 있음을 확인하였다.

핵심용어: 음성인식, 화자정규화, 워핑, 벡터양자화
투고분야: 음성처리 분야 (2.4)

There have been many studies on speaker normalization which aims to minimize the effects of speaker's vocal tract length on the recognition performance of the speaker independent speech recognition system. In this paper, we propose a simple vector quantizer based linear warping speaker normalization method based on the observation that the vector quantizer can be successfully used for speaker verification. For this purpose, we firstly generate an optimal codebook which will be used as the basis of the speaker normalization, and then the warping factor of the unknown speaker will be extracted by comparing the feature vectors and the codebook. Finally, the extracted warping factor is used to linearly warp the Mel scale filter bank adopted in the course of MFCC calculation. To test the performance of the proposed method, a series of recognition experiments are conducted on discrete HMM with thirteen mono-syllabic Korean number utterances. The results showed that about 29% of word error rate can be reduced, and that the proposed warping factor extraction method is useful due to its simplicity compared to other line search warping methods.

Keywords: *Speech recognition, Speaker normalization, Warping, Vector quantization*
ASK subject classification: *Speech signal processing (2.4)*

I. 서 론

일반적으로 화자종속 음성인식은 화자 독립 음성인식에 비해 성능이 우수한 반면 인식 대상 화자에 대한 많은 양의 학습 데이터를 필요로 할 뿐만 아니라 훈련된 화자가 아니면 성능이 저하되는 단점이 있다. 이러한 화자종속 음성인식기의 단점을 극복하는 동시에 화자독립 인식기의 성능을 화자 종속인식기의 성능과 비슷한 수준으로 향상시키기 위해 많은 연구가 이루어져 왔다. 이러한 노력은 크게 화자 적응 (speaker adaptation), 혹은 모델 맵핑 (model mapping)이라 불리는 방식과 화자 정규화

책임저자: 신옥근 (okshin@mail.hhu.ac.kr)
606-791 부산광역시 영도구 동삼동 1
한국해양대학교 자동화정보공학부
(전화: 051-410-4572; 팩스: 051-404-3986)

(speaker normalization), 또는 스펙트럼 맵핑 (spectral mapping)이라 불리는 방식의 두 가지로 분류할 수 있다. 전자는 주로 통계적인 방법을 이용하여 인식기의 파라미터를 화자의 특성에 맞게 조정함으로써 인식 성능을 개선하고자 하는 방법인데 MAP (Maximum A Posteriori)[1], MLLR (Maximum Likelihood Linear Regression)[2,3], 등이 연구되어 왔으며, MLLR 방식이 비교적 적은 양의 적응데이터로 우수한 성능을 갖는 것으로 알려져 있다[3].

화자정규화는 입력되는 음성 신호로부터 특징 벡터를 추출하는 과정에서 화자 사이의 음향학적 특징의 변이를 최소화시킴으로써 인식성능을 향상시키는 방법이다. 화자 사이의 변이의 원인은 문화적인 배경에 따른 억양, 감정 상태 등 외부적인 요인과 성도의 길이, 또는 모양의 차이에 기인하는 내재적인 요인으로 나눌 수 있는데[4,5] 화자정규화는 대부분 화자의 성도의 길이에 따른 특징벡터의 변이를 최소화함으로써 인식 성능을 향상시키고자 하는 것이며, 따라서 VTLN (Vocal Tract Length Normalization)으로 불리기도 한다. 이 방법은 다시 두 가지로 분류할 수 있다. 하나는 가능한 모든 워핑계수 (warping factor)를 적용하여 특징벡터들을 추출한 다음, 이들로부터 직접적인 인식시험을 통하여 최적의 인식율을 보이는 워핑계수를 찾아내는 소위 라인 서치 (line search) 방식[3,6]이며, 또 다른 하나는 화자의 대표적인 음향학적 특징 (acoustic features)인 포만트들 (formants)로부터 워핑계수를 찾아내는 방식[4,7]이다. Gouvea[4]는 세 개의 포만트 F1, F2, F3, 그리고 4 kHz 이상의 고주파 성분에 대한 무게 중심 (center of mass) 등으로부터 구한 주파수 워핑 계수를 이용하여 약 22%의 WER (word error rate)를 감소시킬 수 있었다.

화자의 음향학적 특징을 이용하는 방법의 단점은 첫째, 음향학적인 특징을 정확하게 추출하는 것이 쉽지 않고, 둘째 포만트 등의 음향학적인 특징들은 문맥 (context)에 따라 쉽게 변화하기 때문에 화자의 대표적인 워핑계수를 결정하기 어렵다는 것이다[5]. 라인 서치 방식의 대표적인 단점은 계산량이 너무 많다는 것이다. 이 방식의 대표적인 방법은 Lee 등[6]이 제안한 선형 워핑 방법인데, 각 화자의 음성 신호에 가능한 모든 워핑계수를 적용하여 추출한 특징벡터들로 HMM 음성인식기를 반복적으로 학습시킨 다음 인식 시험을 거쳐 최적의 워핑계수를 찾아내고 이 계수를 멜 스케일 필터에 적용하였다. 그 결과 일반적으로 인식율이 낮은 것으로 알려진 아주 짧은 발화에 대해서도 최소 20% 이상의 WER 감소를 가질 수 있음을 보였다.

한편, 화자 검증 (speaker verification)의 한 가지 방법으로 벡터 양자화를 이용하는 방법들이 연구되어 왔다[8-10]. 화자검증의 일반적인 원리는 발화로부터 화자의 발성구조의 특징을 표현하는 파라메터를 추출하고 이들의 유사도를 비교하여 화자를 검증하는 것이다[10]. 화자검증은 비교적 적은 양의 발화 데이터로 화자를 구별, 검증해야 하는 문제의 특성상 많은 양의 데이터를 필요로 하는 통계적인 방법에 비해 적은 양의 데이터로 화자를 검증할 수 있는 벡터 양자화 방법이 효율적일 때가 많다.

본고에서는 벡터 양자화를 이용하여 화자의 특징을 추출할 수 있으며, 이를 이용하여 화자를 검증할 수 있다는 점에 착안하여 벡터 양자화를 화자 정규화에 이용하는 비교적 간단한 화자 정규화방법을 제안한다. 대부분의 라인 서치 방법에서는 가능한 모든 워핑계수를 적용하여 워핑시킨 특징벡터들을 이용하여 직접적이고 반복적인 인식시험을 거쳐서 최적의 워핑계수를 구하는 것에 비해 제안하는 방법은 벡터 양자화기만을 이용하여 워핑계수를 추출해 낼 수 있으므로 계산량을 줄일 수 있을 뿐 아니라 간단한 방법으로 워핑계수를 추출해 낼 수 있는 장점이 있다. 본 논문의 구성은 다음과 같다. 제 II장에서는 본 연구에서 제안하는 양자화기를 이용한 워핑에 대해 설명한 다음, 제 III장에서는 화자 정규화에 이용될 최적의 양자화기의 생성 방법에 대해 설명한다. 제 IV장에서는 제안한 방법의 성능을 실험을 통해 확인하고 실험 결과에 대해 고찰한 다음 제 V장의 결론으로 끝맺는다.

II. 벡터 양자기를 이용한 워핑

본 연구에서는 MFCC (Mel Frequency Cepstrum Coefficient)를 이용하는 음성 인식기에서 벡타 양자화기를 이용하여 워핑계수를 추정한 다음, 추정된 워핑계수에 따라 Mel 필터를 압축 또는 확장함으로써 정규화된 특징벡터를 추출하는 방법을 제안한다. 이 방법에서는 추출하고 적용하기에 간편한 선형 워핑 (linear warping)방법을 이용하는데, 이미 여러 연구[4,6]에서 비선형 워핑 방법과 선형 워핑방법 사이의 성능 차이는 거의 없음이 알려져 있다. 추출된 워핑계수를 적용하는 방법은 주파수 영역의 스펙트럼에 적용하는 방법과 멜 스케일 필터에 적용하는 방법이 있다[5]. 전자의 방법으로 워핑하는 것은 이산 주파수의 밴드가 정수로 표현되어야 하기 때문에 적용하기 쉽지 않은 반면에 후자의 방법에서는 멜 필터의 중심 주파수와 대역폭을 스케일링하면 되므로 전자에 비해

쉽게 적용할 수 있는 장점이 있다. 본 연구에서는 멜 필터를 워핑하는 방법을 선택한다.

본고에서 제안하는 화자 정규화는 크게 최적의 양자화기 생성과 이 양자화기를 이용한 워핑계수 추출 등의 두 가지 부분으로 이루어진다. 본 연구에서 최적의 양자화기란 일반적인 의미에서의 최적의 양자화기일 뿐만 아니라 최적으로 워핑된 특징벡터들로 만들어진 양자화기를 의미한다. 이해를 돕기 위해 최적의 양자화기가 주어졌다고 가정하고 이 양자화기를 이용한 워핑에 대해 먼저 설명한 다음, III장에서 최적의 양자화기 생성에 대해 기술한다.

최적의 양자화기가 주어졌을 때 임의의 화자의 발화로부터 워핑계수를 추출하는 방법은 다음과 같다. 주어진 화자의 발화로부터 특징벡터를 추출할 때, 고려할 수 있는 모든 워핑계수를 적용하여 일련의 워핑된 특징벡터들을 추출한 다음, 이들을 최적의 양자화기로 양자화한다. 이 과정에서 얻어지는 화자별 양자화 오차들 중에서 최소의 오차를 갖게 하는 워핑계수를 이 화자의 워핑계수로 결정한다. 이 과정은 다음 식 (1)과 같이 표현할 수 있다.

$$\alpha_i^* = \arg\min_{\alpha} \sum_{x_i \in X_i} d(x_i^{\alpha}, C^*(x_i^{\alpha})) \quad (1)$$

여기서 α_i^*는 화자 i의 워핑계수, x_i는 화자 i의 프레임별 파워 스펙트럼, X_i는 화자 i의 학습용 발화의 파워 스펙트럼의 집합, x_i^{α}는 파워 스펙트럼 x_i를 워핑계수 α로 워핑하여 구한 특징벡터이며, C^*는 최적의 코드 북, $C^*(x_i^{\alpha})$는 x_i^{α}를 최적의 코드 북 C^*로 디코딩한 특징벡터, 그리고 d는 두 인수 사이의 유클리드 거리이다.

이러한 방법으로 구한 워핑된 특징 벡터들을 이용하여 음성인식기를 학습시키고, 인식단계에서도 학습시에 사용한 것과 같은 양자화기를 이용하여 워핑계수를 구한 다음 이 계수로 워핑한 특징벡터를 이용하여 인식한다.

그림 1에 최적의 코드 북과 미지의 화자 i의 발화가 주어졌을 때 최적의 워핑계수 α_i^*를 추출하여 화자 정규화를 수행한 다음 인식하는 과정을 도시하였다.

이상의 방법으로 구한 워핑계수가 최적, 혹은 최적에 가까운 워핑계수임을 보이는 것이 필요하다. 그러나 같은 화자라 해도 발화하는 음소에 따라 입모양이 달라지고 그에 따라 워핑계수가 많은 변화를 보이기 때문에 최적의 워핑계수를 정의하는 것부터 쉽지 않은 일이다[6]. 본고에서는 최적성을 증명하기보다는 다음과 같이 간단히 추론하기로 한다. 즉, 이 장의 서두에서 언급한 것처럼 주어진 양자화기는 최적의 워핑계수로 워핑한 특징벡터로부터 만들어졌다고 가정하였다. 따라서 최적으로 워핑된 시험용 벡터는 그렇지 않은 벡터들보다 작은 양자화 오차를 가질 것이다. 이러한 추론의 정당성은 음성인식 실험

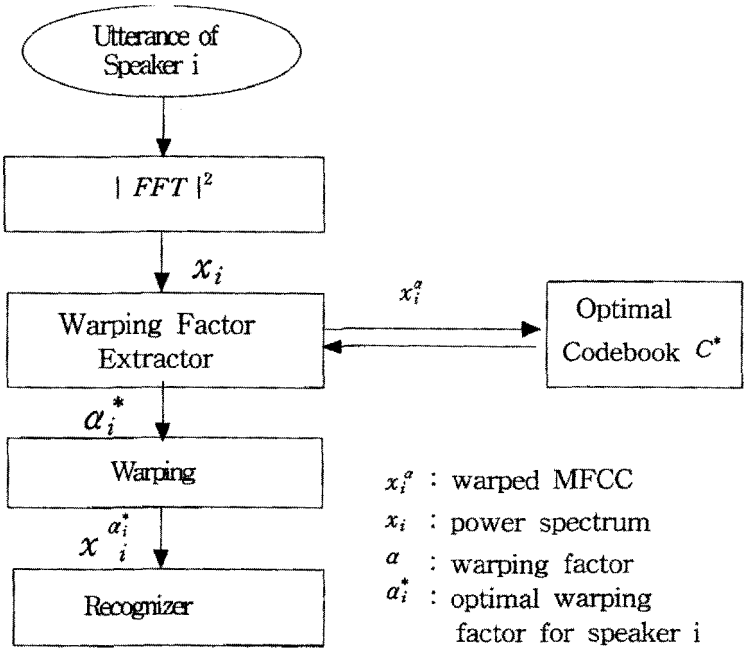


그림 1. 양자화기를 이용한 워핑계수의 추출
Fig. 1. Warping factor extraction based on vector quantizer.

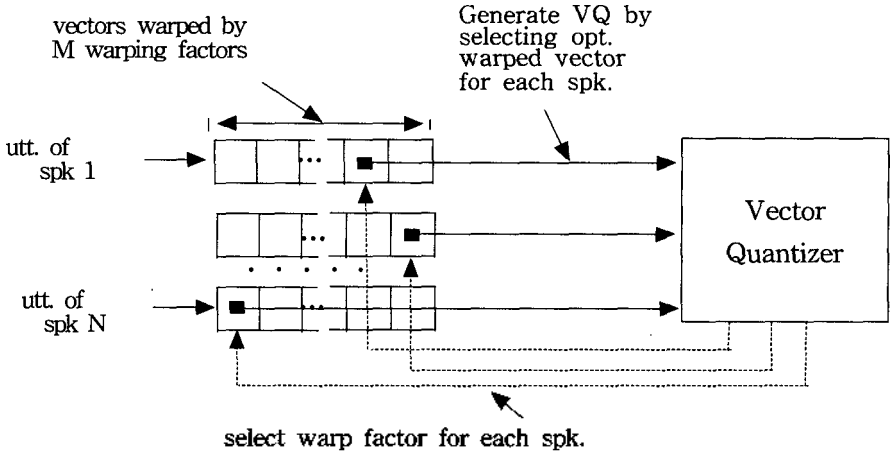


그림 2. 최적의 양자화기 생성
Fig. 2. Generation of optimal vector quantizer.

결과와 추정된 워핑계수의 남녀별 분포를 통해 확인할 수 있다.

III. 벡터 양자화기의 최적화

최적의 양자화기를 생성하는 방법은 직접적인 방법과 간접적인 방법의 두 가지로 나누어 생각할 수 있다. 먼저 직접적인 방법은 학습용 발화로부터 포만트 정보 등을 이용하여[6,11] 직접적으로 워핑계수를 추출한 다음, 워핑된 특징벡터들을 추출하고 이들로부터 양자화기를 생성하는 방법이다. 직접적인 방법으로 워핑계수를 구하는 문제에 관해서는 많은 연구가 수행되어 왔으므로 본고에서는 고려하지 않는다.

제안하는 최적의 양자화기 생성 방법은 간접적인 방법으로 음향학, 혹은 음성학적인 사전 지식없이 화자들의 발화만을 이용하여 최적의 양자화기를 생성하는 것이며, 기본적인 개념은 앞에서 설명한 워핑계수 추출 과정을 반복적으로 적용하는 것이다. 이 방법은 그림 2와 같이 도시할 수 있다. 주어진 N화자들의 발화 각각을 M가지의 워핑계수로 워핑한 특징 벡터를 추출하는데, 이들 워핑계수의 집합에는 1, 즉 워핑하지 않는 경우도 포함시킨다.

먼저 워핑되지 않은 특징벡터들을 이용하여 초기 양자화기를 만든 다음, 각 화자별로 워핑된 M개의 특징벡터들을 차례로 양자화한다. 양자화 결과, 화자별로 가장 작은 오차를 가지게 하는 워핑계수와 이에 상응하는 특징벡터를 구한 다음, 이 특징벡터들을 이용해서 새로운 양자화기를 구성한다. 새로 생성된 양자화기로 이전의 양자화기를 대체한 다음, 화자별로 최소의 양자화 오차를 갖게 하는 워핑계수와 워핑된 특징벡터를 찾는 과정을 반복하고, 더 이상 새로운 양자화기가 만들어지지 않으면 최적의 양자화기가 만들어진 것으로 간주하고 멈춘다. 이 과정에서 워핑계수의 범위는 Lee 등[4]과 Welling 등[12]이 사용한 방법과 비슷하게 0.88 ~ 1.12 사이의 값으로 설정하며, 이 범위는 인간 성도의 길이가 짧게는 약 13 cm (여성의 경우)에서 길게는 약 18 cm (남성의 경우)까지 약 25% 범위에서 변화할 수 있다는 것을 감안한 것이다. 아래에 워핑계수 사이의 간격을 0.01로 이산화한 벡터 양자화기의 최적화 알고리즘을 보인다.

Optimal Codebook Generation

Given the set of training utterances of all the speakers,

1. Generate an initial codebook C^0, with all the speakers' unwarped training feature vectors $x_i^{\alpha^0}$, where α^0 is 1 (i.e., unwarped), $i \in I$, and I is the set of all speakers.
2. Let k = 1, where k is the iteration number.
3. For all speakers $i \in I$, and $x_i \in X_i$, where X_i is the set of all power spectra of speaker i,
   3.1. For each warping factor α^k ($0.88 \le \alpha^k \le 1.12$; step 0.01),
      3.1.1. Get the warped feature vectors $x_i^{\alpha^k}$ which are obtained by warping the power spectrum of speaker i's utterances with the factor α^k.

3.1.2. Decode $x_i^{\alpha^k}$ with codebook C^{k-1}.

3.2. Find the minimum Euclidean distance d_i^* and the optimal warping factor α_i^* at k-th iteration as follows:

$$d_i^k = \min_{\alpha^k} \sum_{x_i \in X_i} d(x_i^{\alpha^k}, C^{k-1}(x_i^{\alpha^k}))$$

$$\begin{cases} d_i^* = d_i^k, & \text{if } d_i^k \le d_i^{k-1} \\ d_i^* = d_i^{k-1}, & \textit{otherwise} \end{cases}$$

$$\begin{cases} \alpha_i^* = \arg\min_{\alpha^k} \sum_{x_i \in X_i} d(x_i^{\alpha^k}, C^{k-1}(x_i^{\alpha^k})), \\ \qquad \text{if } d_i^k \le d_i^{k-1} \\ \alpha_i^* = \alpha_i^{k-1}, \quad \textit{otherwise.} \end{cases}$$

4. Reconstruct the codebook with the set of optimally warped feature vectors obtained in step 3 and replace the old codebook with the new one.
5. If the steady state is reached, then stop. Else increment k and goto step 3.

이상의 방법으로 구한 양자화기가 최적의 양자화기인지를 정량적으로 증명하는 것 역시 쉽지 않으나 다음과 같이 생각해 볼 수 있다. 즉, 위에서 설명한 과정 중의 어떤 양자화기에 대하여 특정한 계수로 워핑한 특징벡터가 최소의 양자화 오차를 갖는다면 이 양자화기에 대해서는 언급한 특징벡터가 최적으로 워핑되었다고 할 수 있다. 그리고 현재의 이 양자화기는 그전 단계에서 최소의 오차를 가지도록 워핑된 특징벡터들로부터 만들어진 것이다. 따라서 각 단계를 거칠 때마다 모든 화자에 대한 양자화 오차의 합은 작아지게 될 것이고 궁극적으로 양자화 오차는 특정한 값에 수렴하게 된다. 이 때의 양자화기는 어떤 워핑계수를 적용한 특징벡터로 양자화하더라도 더 작은 양자화 오차를 가질 수 없으므로 최소의 오차를 갖는다. 그러나 이 최소 오차가 국부적인 최소값 (local minimum)인지, 아니면 전체적인 최소값 (global minimum)인지는 밝혀지지 않았다. 따라서 제안한 방법으로 구한 양자화기는 엄밀한 의미에서 최적의 양자화기라고 할 수 없지만 편의상 최적의 양자화기라고 부르기로 하고 실험을 통해 검증하기로 한다.

기존의 라인 서치 방식에 비하여 제안하는 방법이 갖는 장점은 다음과 같다. 첫째, 정규화된 특징벡터로 학습된 인식기를 얻기 위해 인식기 전체를 반복 학습시킬 필요 없이 양자화기만 학습시키면 되므로 상대적으로 간단한 방법으로 학습시킬 수 있다. 둘째, 제안하는 양자화기는 인식기와는 독립적인 전처리 과정이므로 양자화기를 학습시킬 때 인식대상 어휘를 골고루 포함시키는 동시에 다양한 워핑계수를 갖는 화자들의 발화를 이용하였다면 이 양자화기는 범용의 워핑계수 추출기로 사용될 수 있다. 셋째, 포만트를 이용해서 워핑계수를 추출하는 경우, 유사 포만트 (spurious peaks), 인접한 포만트의 혼합 (Blending), 높은 톤의 발화 (High pitched speech) 등의 문제로 인하여 안정적인 포만트를 추출하는 것이 쉽지 않다[4]. 양자화기를 이용하는 방법에서는 이러한 문제가 생기지 않는다. 뿐만 아니라 대부분의 워핑계수 추출 알고리즘에서는 표준 스펙트럼을 선정해야 하는데 제안하는 방법에서는 표준 스펙트럼을 선정할 필요가 없다. 마지막으로 워핑계수 추출을 위한 특별히 정해진 시험용 발화를 필요로 하지 않는다. 따라서 화자의 발화 방법이 바뀌거나 새로운 화자가 발화를 시작할 경우 언제든지 발화의 일부를 워핑계수 추출기 (양자화기)에 입력하여 계수를 추출하고 적용할 수 있다.

이 방법의 한가지 단점은 연속음성을 대상으로 하는 경우와 같이 코드북의 크기가 크고 학습 데이타가 많아야 할 경우, 양자화기의 반복적인 학습에 기인하는 과다한 계산량이다. 이것은 양자화기의 코드 북의 크기를 처음부터 최종적인 크기로 잡아서 학습시키는 대신, 초기에는 비교적 작은 크기의 코드 북을 이용하여 학습 데이터들의 워핑계수를 대략적으로 구한 다음, 코드북의 크기를 점차 늘려서 학습해 나감으로써 계산량을 크게 감소시킬 수 있다.

IV. 인식 실험 및 고찰

II 장과 III장에서 기술한 화자 정규화 방법의 효용성과 성능을 시험하기 위하여 이산 HMM (DHMM)을 이용한 단음절 숫자음 인식 실험을 수행하였다. 먼저 인식 실험에 이용한 인식기와 실험 데이터에 대하여 기술한 다음, 화자 정규화를 거친 특징벡터들에 대한 인식 성능과 CMS-MFCC (cepstrum mean subtracted MFCC)를 특징벡터로 하였을 때의 성능을 비교 분석한다.

4.1. 인식 데이터와 실험환경

본 연구에서 사용한 음성인식기는 음절별로 5개의 상태 (state)를 갖는 이산 HMM (Discrete HMM)이며, 우리말 단음절 숫자음 영 ~ 구, 십, 백, 천 등 모두 13가지의

음절을 남 여 화자 각각 25명이 3번씩 발화한 1950 음절의 데이터로 실험을 수행하였다. 이 데이터들은 모두 조용한 사무실 환경에서 16비트, 11.025 K/sec 단위로 샘플링하여 녹음한 것이며 특별한 잡음제거의 과정은 거치지 않았으나 음절의 시작점과 끝점은 수작업으로 검출하였다. 모든 실험용 음절을 절반으로 나누어 학습용 975음절과 시험용 975음절로 구성하였으며 화자별 발화를 균등하게 배분하여 학습용과 시험용으로 구분하였다. 양자화기의 학습에 사용한 음절 데이타와 인식모델의 학습을 위해 사용된 음절 데이타들은 서로 같다.

한편 일반적인 워핑계수 추출이나 화자 독립 음성인식 실험의 경우, 학습에 참여시키지 않은 화자의 데이터를 인식에 포함시키는 것이 일반적이지만 본 실험에서는 발화 데이터가 충분하지 못하여 학습과 인식에 사용한 데이터는 모두 같은 화자의 발화들을 이용하였으나 같은 발화 데이터를 학습과 인식에 중복해서 사용하지는 않았다.

4.2. 전처리 및 벡터 양자화기

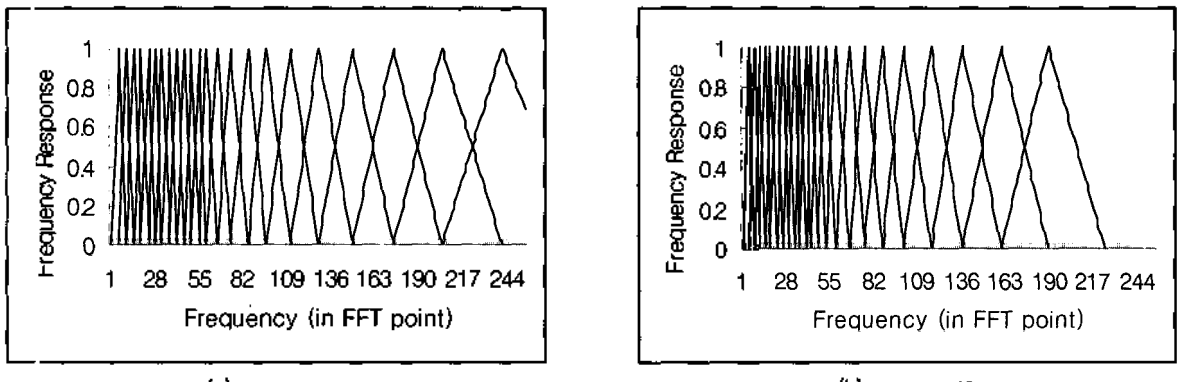
모든 음절은 330 샘플단위로 나누어 30 msec 크기의 프레임들로 나누었으며 프레임 이동구간을 10 msec로 하여 이웃하는 프레임과는 전후 각각 20 msec가 오버랩되게 하였다. 프레임 단위의 음성 데이터에 $1-0.95z^{-1}$ 프리 엠퍼시스 필터와, Hamming 윈도우를 적용하여 처리한 다음, 512 포인트 FFT (fast Fourier transform)를 통해 파워 스펙트럼을 구하였다. 이 파워 스펙트럼에 멜 스케일 필터를 적용하여 24차의 MFCC (Mel frequency cepstrum coefficient)를 특징벡터로 구하고 이를 실험에 이용하였는데, 이 때 화자의 워핑계수에 따라 멜 스케일 필터를 워핑하였다. 다음의 그림 3에 워핑된 멜 스케일 필터의 예를 보인다.

본 연구에서 사용한 기본적인 벡터 양자화기는 LBG 알고리즘[13]을 이용하여 구현하였으며 코드북의 크기는 512이다. 이 벡터 양자화기는 워핑계수를 추출하기 위해 이용되는 동시에 이산 HMM의 학습과 인식을 위한 코드를 추출하기 위해서도 이용된다. 다만 워핑계수 추출 실험에서 벡터 양자화기의 코드 북 크기의 영향을 관찰하기 위해 256크기의 코드북을 이용할 때는 예외이다.

4.3. 실험결과 및 고찰

아래의 표 1에 본 연구에서 수행한 인식 실험결과를 보인다. 이 표의 'Baseline'은 워핑을 고려하지 않은 종래의 이산 HMM 인식기의 인식성능이며 본고에서 제안하는 워핑 방법의 성능을 비교하기 위한 기준이 된다. 'Baseline'과 'CMS'를 제외한 모든 실험에서 인식모델들은 각각 최적으로 워핑된 특징벡터들로 학습되었으며, 인식 시험시에도 각각의 양자화기로 워핑한 시험용 데이터로 시험하였다. 'Warped (256 VQ)'은 워핑계수를 추출할 때 벡터 양자화기의 코드 북 크기를 256으로 설정하였을 때의 인식결과이다. 이 때 워핑계수 추출에 사용하는 코드 북의 크기에 따른 영향을 관찰하기 위해 인식기를 위한 양자화기의 코드 북 크기는 다른 실험에서와 같이 512로 하였다. 'Warped (512 VQ)'는 워핑계수 추출 및 인식기를 위한 코드 북의 크기를 모두 512로 하였을 때의 인식 결과이며, 'CMS (Cepstral Mean Subtraction)'는 워핑하지 않은 MFCC에서 화자별로 MFCC의 평균을 뺀 것 (CMS-MFCC)을 특징벡터로 했을 때의 인식 결과이다. 마지막으로 'Warped CMS'는 CMS-MFCC를 특징벡터로 하여 본고에서 제안하는 워핑을 적용한 다음, 최적의 코드 북을 생성하고 인식기를 학습, 시험하였을 때의 결과이다.

표 1에서 볼 수 있는 바와 같이 코드 북의 크기가 512일



(a) $\alpha = 0.88$ (b) $\alpha = 1.12$

그림 3. 워핑계수 α가 0.88일 때와 1.12일 때의 멜 스케일 필터뱅크

Fig. 3. Warped Mel scale filter bank when $\alpha = 0.88$ and $\alpha = 1.12$.

표 1. 인식 실험결과
Table 1. Results of recognition experiments.

Baseline	85.44	14.56	-
Warped (256 VQ)	86.67	13.33	8.4
Warped (512 VQ)	89.64	10.36	28.8
CMS	87.90	12.10	-
Warped CMS	90.77	9.23	23.7

때, CMS를 적용하지 않은 경우 오인식률이 약 29%의 감소하였다. 또 CMS를 적용한 다음 워핑한 경우 워핑하지 않은 CMS와 비교하면 약 24%의 오인식률이 감소되어 제안하는 벡터 양자화기를 이용한 워핑계수 추출 방법이 화자 정규화에 유용하게 이용될 수 있음을 확인할 수 있다.

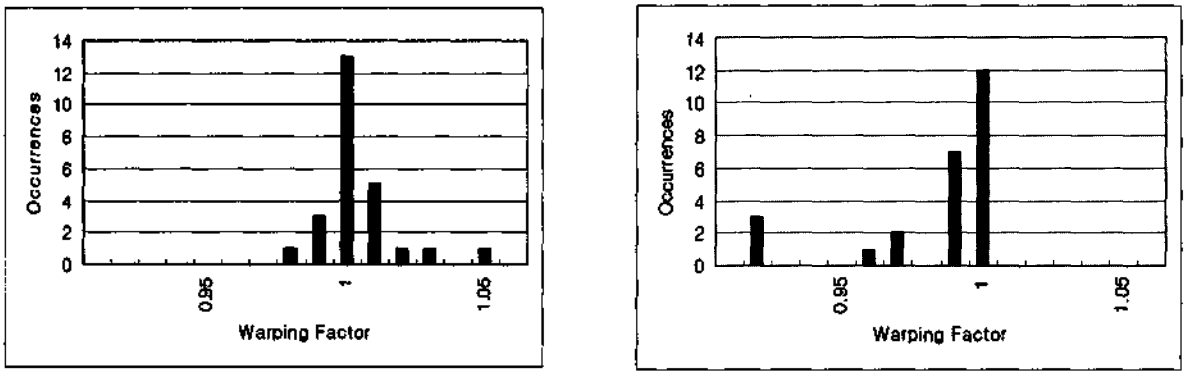
CMS는 인식성능을 향상시키기 위해 일반적으로 많이 쓰이고 있는 방법인데[14] 본 실험에서도 Baseline에 비교하여 17%정도의 오인식률이 감소함을 확인할 수 있었다. 또 Lee[6] 등은 실험을 통하여 그들이 제안한 워핑 방법이 CMS를 이용한 방법보다 더 우수한 성능을 갖는다는 것을 보였는데, 본 연구에서도 제안하는 방법으로 워핑했을 때 통상의 CMS를 이용한 방법보다 더 우수한 성능을 갖는다는 것을 확인할 수 있었다. 뿐만 아니라, 제안하는 워핑 방법은 CMS 방법과 병행할 경우 더 좋은 성능을 가져 CMS와 병행할 수 있음을 알 수 있다.

한편 코드 북의 크기가 256일 때와 512일 때의 경우를 비교하면 워핑계수를 추출하기 위한 최적의 코드 북의 크기에 따라 성능의 차이가 현저하게 달라지는 것을 관찰할 수 있다. 이것은 이산 HMM 인식기를 위한 코드 북의 크기가 인식성능에 영향을 미치는 것과 마찬가지로 워핑계수 추출을 위한 코드 북의 크기도 적절해야 함을 의미하는 것으로 해석될 수 있을 것이다.

그림 4에 워핑계수의 분포를 남여별로 나누어 도시하였다. 이 그림들의 종축은 워핑계수, 횡축은 각 워핑계수에 해당하는 화자의 수이다. 화자별 워핑계수를 추출하기 위해 각 화자별로 평균 19.5개의 음절을 이용하였다. 남성화자 및 여성화자의 워핑계수의 평균은 각각 1.004, 0.98이며 남녀화자의 전체 워핑계수의 평균은 0.99이다.

워핑계수가 1보다 크면 멜 스케일 필터를 확장하고, 1보다 작으면 압축하여 화자 평준화가 이루어지며 워핑계수가 1이면 워핑하지 않은 MFCC를 특징벡터로 이용한다. 일반적으로 남성 성도의 길이가 여성의 그것보다 길고, 포만트 주파수는 성도의 길이에 반비례하므로 남성의 포만트 주파수들은 여성의 포만트 주파수보다 낮다. 따라서 남성의 워핑계수는 1보다 크게 하여 멜 스케일 필터를 확장시킬 필요가 있고, 여성의 경우에는 1보다 작게 하여 압축시킬 필요가 있다. 그림 4에 도시한 본 연구의 실험결과는 이러한 추론에 비교적 상응하는 워핑계수의 분포를 보이고 있다.

그림 5에는 제 3장에 기술한 방법으로 최적의 코드 북을 생성하기 위해 반복적으로 양자화기를 학습시키는 과정에서 각 화자들의 워핑계수가 수렴되어 가는 추세를 보인 것이다. 모두 50명의 화자로 실험하였으나 많은 화자의 추세가 서로 중복되어 서로 다른 추세들만 표시되었다. 초기에 모든 화자들의 워핑계수를 1로 설정한 다음 양자화기를 반복적으로 학습시켜 정상상태 (steady states)에 도달할 때까지 계속하였다. 이 실험에서는 5회의 반복학습 끝에 모든 화자들의 워핑계수가 정상상태에 도달하



(a) Male Speakers (b) Female Speakers

그림 4. 남여화자별 워핑계수의 분포
Fig. 4. Warping factor distribution of male and female speakers.

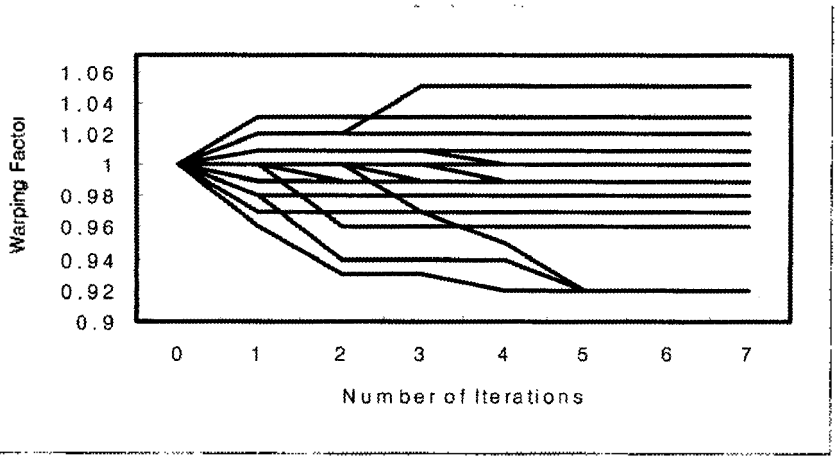


그림 5. 최적의 코드북을 생성하기 위한 코드북 반복 학습과정중 워핑계수의 수렴
Fig. 5. Convergence of warping factors in the course of the training of the optimal codebook.

여 비교적 짧은 시간에 최적의 코드 북을 생성할 수 있었다.

V. 결 론

본고에서는 벡터 양자화기를 이용하여 화자 검증이 가능하다는 점에 주목하여 벡터 양자화기를 이용한 화자를 정규화 방법을 제안하였다. 이 방법에서는 먼저 학습용 발화로부터 최적의 양자화기를 구성 다음, 시험용 발화를 가능한 모든 계수로 워핑하여 차례로 최적의 양자화기와 비교함으로써 화자별 워핑계수를 추출하고, 이 계수를 MFCC 추출과정의 멜 스케일 필터에 적용하여 워핑된 특징벡터를 추출하였다. 제안하는 방법의 효용성을 검증하기 위하여 이산 HMM 기반의 숫자음 인식기를 이용하여 간단한 인식실험을 수행하였다. 모두 13개의 단음절 우리말 숫자음을 인식하는 인식기를 구성하였으며, 워핑하지 않은 Baseline의 경우, 제안하는 방법으로 워핑한 경우, 그리고 CMS (cepstral mean subtraction)을 적용한 경우의 인식율을 각각 조사하고 비교하였다. 실험결과, 워핑을 적용한 경우 Baseline인식기의 성능에 비해 약 29%의 오인식률 감소를 보였으며, 워핑과 CMS와 병행하였을 때는 Baseline에 비해 약 37%의 오인식률을 감소시킬 수 있었다. 이 방법은 최적의 워핑계수를 추출하기 위해 Lee[6] 등의 방법처럼 반복적인 인식 시험을 거치지 않아도 된다는 것과 CMS 기법과 병행하여 적용할 수 있다는 점 등의 장점이 있어 더 많은 연구를 할 가치가 있는 것으로 판단된다. 또 본 연구의 실험에서 사용한 음성 인식기는 이산 HMM (Discrete HMM)이지만 이 연구의 결과는 가우시안 분포를 이용한 연속 HMM (Continuos HMM)에도 적용할 수 있으리라 기대된다.

감사의 글

본 연구는 BK21사업의 부분적인 지원을 받아 수행되었습니다.

참 고 문 헌

1. C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of continuous density HMM parameters," *Proc. of the ICASSP*, 1, 145-148, Toronto, Canada, 1991.
2. C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuos density hidden markov models," *Computer Speech and Language*, 9, 171-185, 1995.
3. J. E. Hamaker, "MLLR : A speaker adaptation technique for LVCSR," Lecture note, http://www.isip.msstate.edu/publications/courses/ece_7000_speech/lectures/current/lecture_10/paper/paper.pdf
4. E. B. Gouvea, "Acoustic-Feature-based Frequency Warping For Speaker Normalization," *Thesis*, Carneigie Mellon University, 1998.
5. P. Zhan and Alex Waibel, *Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition*, Language Technologies Institute Technical Report: CMU-LTI-97-150, Carnegie Melon University, May, 1997
6. L. Lee and R. C. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. on Speech and Audio Processing*, 6 (1), 49-60, Jan. 1998.
7. Y. Ono, H. Wakita and Y. Zhao, "Speaker normalization using constrained spectral shifts in auditory filter domain," *EuroSpeech*, 1, 355-358, 1993.

8. P. G. Pop and E. Lupu, "Speaker verification with vector quantisation," *Proc. Trends and Recent Achievements in Information Technology*, 16-18, May 2002.
9. T. Kinnunen, T. Kilpelainen and P. Franti, "Comparison of clustering algorithms in speaker identification," *SPC*, 222-227, 2000.
10. S. Furui, "Recent advances in speaker recognition," Pattern Recognition Letters, **18**, 859-872, 1997.
11. E. Edie and H. Gish, "A parametric approach to vocal tract length normalization," *ICASSP*, **1**, 346-348, May, 1996.
12. L. Welling, R. Haeb-Umbach, X. Aubert and N Haberland, "A study on speaker normalization using vocal tract normalization and speaker adaptive training," *ICASSP98*, 1998.
13. Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, **COM-28** (1), January 1980, 84-95.
14. C. Avendano, S. Tibrewala and H. Hermansky, "Multiresolution channel normalization for ASR in reverberant environments," *EUROSPEECH*, Rhodes, Greece, 1997.

저자 약력

● 신 옥 근 (Ok-Keun Shin)

1981년: 서강대학교 전자공학과 졸업 (학사)
1983년: 부산대학교 전자공학과 (공학석사)
1989년: 프랑스 Université (공학박사)
1983년~1995년: 한국전자통신연구소 선임연구원
1995년~현재: 한국해양대학교 자동화정보공학부 부교수
※ 주관심분야: 신호처리, 음성신호처리, 음성인식