한국음향학회지 제23권 제8호 pp. 583-589 (2004)

연속음성 인식기를 위한 벡터양자화기 기반의 화자정규화

Vector Quantizer Based Speaker Normalization for Continuos Speech Recognition

신 옥 근*
(Ok-keun Shin*)

*한국해양대학교 IT공학부

(접수일자: 2004년 7월 5일; 수정일자: 2004년 11월 2일; 채택일자: 2004년 11월 30일)

포만트 등의 음향학적인 정보를 이용하지 않는 연속음성인식 (CSR)을 위한 벡터 양자화기 기반의 화자 정규화 방법을 제안한다. 이 방법은 앞서 제안한 간단한 숫자음 인식기를 위한 화자정규화 방법을 개선한 것으로, 코드북의 크기를 증가시켜 가면서 벡터양자화기를 반복적으로 학습시킴으로써 정규화된 코드북을 구한 다음, 이를 이용하여 시험용화자의 워핑계수를 추정한다. 코드북 생성과 워핑계수 추정을 위해 모음 음소의 집합과 자음과 모음을 포함한 모든 음소의 집합 등 두 가지 음소집합을 이용하여 실험하였으며, 추정한 워핑계수에 상응하는 구간선형 워핑함수를 이용하여 인식기의 학습과 시험에 사용될 특징벡터를 워핑하였다. TIMIT 코퍼스와 HTK toolkit을 이용한 음소인식 실험을 수행하여 제안하는 방법의 성능을 조사한 결과, 포만트를 이용한 워핑 방법과 비슷한 성능을 가짐을 확인하였다.

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

Proposed is a speaker normalization method based on vector quantizer for continuous speech recognition (CSR) system in which no acoustic information is made use of. The proposed method, which is an improvement of the previously reported speaker normalization scheme for a simple digit recognizer, builds up a canonical codebook by iteratively training the codebook while the size of codebook is increased after each iteration from a relatively small initial size. Once the codebook established, the warp factors of speakers are estimated by comparing exhaustively the warped versions of each speaker's utterance with the codebook. Two sets of phones are used to estimate the warp factors: one, a set of vowels only, and the other, a set composed of all the phonemes. A piecewise linear warping function which corresponds to the estimated warp factor is adopted to warp the power spectrum of the utterance. Then the warped feature vectors are extracted to be used to train and to test the speech recognizer. The effectiveness of the proposed method is investigated by a set of recognition experiments using the TIMIT corpus and HTK speech recognition tool kit. The experimental results showed comparable recognition rate improvement with the formant based warping method.

Keywords: Speech recognition, Speaker normalization, Warping, Vector Quantizer, CSR
ASK subject classification: Speech processing (2.4)

I. 서 론

화자정규화는 발화자의 성도의 차이에 기인하는 음향학적인 변이를 최소화함으로써 화자독립 음성인식기의 성능을 향상시키는 방법이다. 이 방법은 크게 발화의 음향학적인 특성, 즉 포만트를 직접 추출한 다음 이를 이용하여 워핑계수를 추정하는 방법과 주어진 발화에 모든 워핑계수를 적용하여 일련의 특징벡터들을 추출한 다음 이들을 인식 모델 혹은 통계적인 모델과 비교하여 워핑

책임저자: 신 옥 근 (okshin@bada.hhu.ac.kr)
606-791 부산시 영도구 동삼동 1번지
한국해양대학교 IT공학부
(전화: 051-410-4572; 팩스: 051-404-3986)

계수를 추정하는 라인서치(line search) 방법으로 나눌 수 있다. 전자의 방법은 계산량이 적은 대신 context에 따라 쉽게 변하는 포만트의 특성 때문에 일관성 있게 찾아내기 어려워[1] 실시스템에 적용한 예는 보기 드물다. 반면에 후자의 방법은 워핑계수를 추정하는데 필요한 계산량이 과다하다는 단점이 있어 이를 개선하기 위한 연구가 활발하다. 후자의 방법 중 대표적인 것으로 Lee 등이 제안한 방법[2]을 들 수 있는데 이들은 먼저 인식기의 반복적인 학습을 통해 학습용 특징벡터의 워핑계수를 구하고 통계적인 모델을 만든 다음, 시험용화자의 특징벡터를 이 모델과 비교하여 워핑계수를 구하였다. 이 방법은 좋은 성능을 가지는 것으로 알려져 있지만 반복적인 인식기의 학습에 따르는 과다한 계산량 뿐 아니라, 학습용 데이터의 워핑계수는 인식기의 반복적 학습을 통해 구하는 반면, 시험용 데이터의 워핑계수는 통계적인 방법으로 구함으로써 발생하는 비대칭성의 문제가 있을 수 있다.

이 방법의 단점을 보완하는 한 가지 방법으로 본연구의 선행연구[3]에서 벡터양자화기를 이용하여 학습용 벡터의 워핑계수를 추정하는 한편, 그 결과로 얻어지는 양자화기를 이용하여 시험용 벡터의 워핑계수를 추정하는 방법을 제안한 바 있다. 이 선행연구에서는 우리말의 단음절 숫자음 (영~십, 백, 천)을 대상으로 화자 정규화 실험을 하였으며, 간단한 반복적인 벡터양자화기의 학습으로 4~5%의 WER (word error rate)를 줄일 수 있었다. 그러나 이 방법은 다양한 음소와 변이음, 많은 수의 화자를 포함하는 연속음성 인식 (CSR: continuous speech recognition)의 문제에서는 효과적이지 못하였다. 분석결과, 선행연구에서 제안한 양자화기를 이용하는 방법을 CSR에 그대로 적용할 경우, 최종적인 양자화기의 질은 초기 양자화기의 질에 의해 결정됨을 알 수 있었다. 본고에서는 선행연구의 양자화기를 반복적으로 이용하되 양자화기의 크기를 점진적으로 증가시킴으로써 적절한 상태의 초기 양자화기를 제공하는 방법으로 화자 정규화가 이루어질 수 있음을 보인다. 본 연구에서는 HTK 음성인식 소프트웨어[4]와 TIMIT 음성 코퍼스[5]를 이용한 음소단위의 인식실험을 수행하였으며 제안한 방법으로 2.1%의 절대인식률향상을 얻을 수 있었다.

본 논문의 구성은 다음과 같다. 먼저 II장에서는 본 연구에서 이용한 구간 선형 워핑함수에 대하여 설명한 다음, 정규화 벡터양자화기가 주어졌을 때 시험용 발화의 워핑계수를 추정하는 방법을 설명한다. III장에서는 벡터양자화기를 학습시켜 정규화 벡터양자기를 구하는 방법을 두 개의 프로세서로 나누어 기술한다. IV장에서는 제안하는 방법의 성능을 알아보기 위한 음소 인식 실험에 관해 기술한 다음, V장의 결론으로 끝맺는다.

II. 벡터양자화기와 구간선형 워핑함수를 이용한 화자정규화

이 장에서는 정규화 양자화기가 주어졌을 때 이를 이용하여 워핑계수를 추정하고 적용하는 방법을 설명한다.

추정한 워핑계수 또는 함수를 적용하여 인식기의 학습과 인식을 위해 사용될 워핑된 특징벡터를 구하는 방법은 두 가지를 생각할 수 있다. 한 가지는 추정한 워핑계수, 또는 함수를 이용하여 프레임별 파워 스펙트럼을 직접적으로 decimation/interpolation함으로써 워핑하는 방법[6]이며, 또 다른 한 가지 방법은 Lee[2]등이 사용한 방법과 같이 워핑계수에 따라 멜 스케일 필터의 대역폭을 늘이거나 줄이는 방법이다. 본 연구에서는 전자의 방법을 이용한다.

워핑계수 추정에 사용되는 특징벡터는 모든 화자들의 평균적인 F4 위치[1]까지의 스펙트럼 성분만을 이용하여 추출한 12차원의 MFCC벡터이다. 따라서 워핑계수가 적용된 특징벡터는 주어진 계수에 따라 스펙트럼 신호를 워핑한 다음 F4까지의 성분을 취하여 만든 MFCC이다. 본 연구에서 이용한 워핑계수는 0.02간격을 갖는 0.78~1.22사이의 선형계수이다. 위에서 설명한 특징벡터를 이용해 구한 화자별 계수는 구간선형 워핑함수 (2.1절)로 확장되어 인식기의 학습 및 시험 벡터를 워핑하기 위해 사용된다.

2.1 구간선형 워핑함수

워핑함수는 선형, 구간선형, 비선형 등 다양한 함수들이 연구되고 있으며, Molau 등[7]은 비교적 간단한 구간선형 (piecewise linear) 워핑함수가 Edie[8]등이 제안한 비선형 워핑 방법과 비슷한 성능을 가짐을 WSJ 코퍼스를 이용한 실험을 통하여 보였다. 이들이 이용한 구간

1) F3의 평균 주파수에 해당하는 2.5KHz보다 3.6KHz까지의 주파수 성분을 포함시키는 것이 가장 좋은 결과를 얻을 수 있었으며 3.6KHz는 평균적인 F4의 주파수[9]에 해당한다. 벡터양자화기를 이용한 본 연구에서는 모음뿐만 아니라 자음을 포함한 발화 전체를 이용하여 워핑계수를 추출한 경우에도 F3보다는 F4를 이용함으로써 가장 좋은 결과를 얻을 수 있었다.

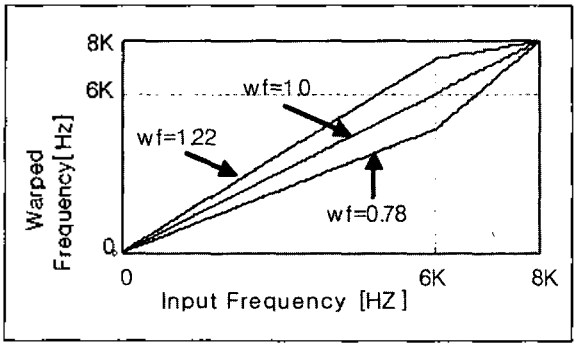

그림 1. 본 연구에서 이용한 구간선형 워핑함수
Fig. 1. Piecewise Linear Warping Function.

선형 워핑함수는 샘플링 주파수가 16KHz인 스펙트럼에서 0~7KHz 사이는 추정한 선형 계수를 적용하여 워핑하고 나머지 7KHz에서 Nyquist 주파수인 8KHz까지는 끝점이 Nyquist 주파수에서 만나도록 하였다. 그러나 본 연구에서 택한 워핑계수의 범위는 0.78~1.22이어서 Molau 등의 워핑함수를 그대로 이용할 수 없으며, 그림 1에 보이는 것처럼 0~6KHz, 6~8KHz의 두 구간으로 나눈 구간선형 워핑함수를 이용한다.

2.2 정규화 벡타양자화기를 이용한 워핑계수 추출

정규화 양자화기 V 가 주어졌을 때, 미지의 화자 u 의 워핑계수 $\hat{\alpha}_u$ 는 다음 식 (1)과 같이 이 화자의 발화의 프레임별 양자화 오차의 합을 최소화하는 워핑계수 α 를 찾음으로써 구한다.

$$\hat{\alpha}_u = \arg\min_{\alpha \in A} \sum_{x_u} d(x_u^{\alpha}, V(x_u^{\alpha})) \tag{1}$$

여기서 $d(,)$ 는 두 인수 사이의 유클리드 거리, A 는 모든 워핑계수의 집합, x_u^{α} 는 화자 u 의 특징벡터 x_u 를 계수 α 로 워핑한 것, 그리고 $V(x_u^{\alpha})$ 는 x_u^{α} 를 양자화기 V 로 양자화 했을 때의 reproduction vector이다. 식 (1)은 벡터양자화기 V 가 정규화된 학습용 벡터들을 이용하여 학습되었다고 가정하였으므로, 최적으로 워핑된 시험용 벡터가 최소의 양자화 오차를 갖는다는 것을 의미한다.

III. 정규화 벡타양자화기의 생성

2.2절의 식 (1)을 이용하여 시험용 화자의 워핑계수를 추정하기 위해서는 정규화된 학습용 특징벡터로 학습된 양자화기 V 를 필요로 한다. 그러나 학습용 화자의 워핑계수는 알려져 있지 않으므로 본 연구에서는 양자화기를 반복적으로 학습시킴으로써 정규화 양자화기를 생성하는 방법을 제안한다. 이 방법은 다음과 같이 두 개의 프로세스로 구성된다. 먼저 LMP (Local Minimization Process)라 부르는 하위 프로세스는 주어진 크기의 코드북을 반복적으로 학습시킴으로써 국부적으로 최적의 코드북을 생성한다. 상위 프로세스는 PRP (Progressive Refinement Process)라 부르며, 초기의 작은 크기에서 시작하여 코드북의 크기를 점차 늘려가며 LMP를 구동함으로써 최종적으로 원하는 크기의 코드북을 생성하게 한다. 그림 2에 이 두 프로세스의 개념을 도식적으로 나타내었다.

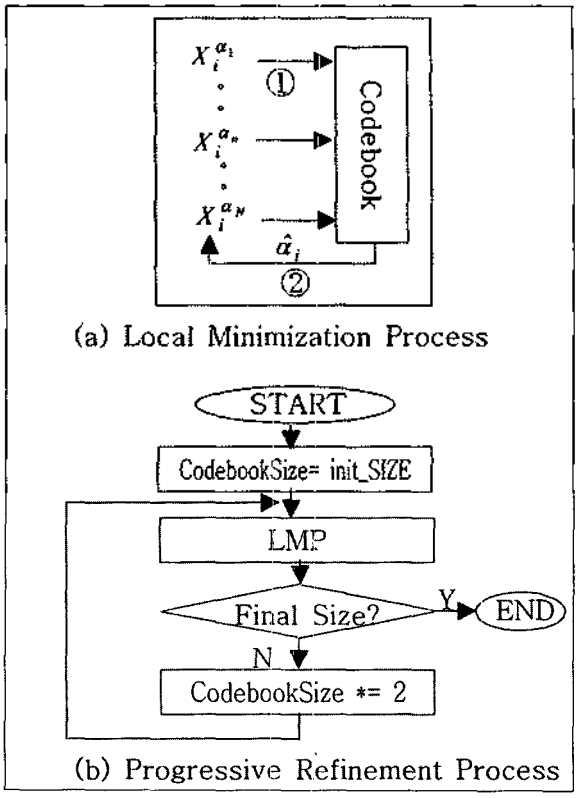

그림 2. 정규화 코드북 생성의 두 가지 프로세스.
Fig. 2. Two Processes of Normalized Codebook Construction.

이들 중 LMP는 본 연구의 선행연구[3]에서 사용된 방법과 같은 것이다. 서론에서 언급한 것처럼 CSR을 위한 화자 정규화에서는 LMP에 제공하는 코드북의 초기 상태가 적절해야하며, 적절한 초기 상태의 코드북을 제공하는 방법 중의 하나는 코드북의 크기를 처음부터 크게 하는 것이 아니라 작은 크기에서 시작하여 점차 늘려 가는 것이다. 다음의 각 절에서 먼저 LMP에 대해 간단히 소

개한 다음, PRP의 메커니즘과 이 프로세서를 이용하여 적절한 초기 상태의 코드북을 제공하는 방법에 대해 기술한다.

3.1. Local Minimization Process

그림 2.(b)에 보인 것처럼 PRP에 의해 코드북의 크기가 정해진 다음 LMP가 구동된다. LMP가 구동되면 먼저 초기 코드북이 만들어지는데 이 코드북은 그 직전의 LMP 구동에 의해 구해진 화자별 워핑계수로 워핑된 특징벡터를 이용하여 만들어지지만, LMP가 최초로 구동될 때에는 워핑되지 않은 특징벡터로 만들어진다. 그림 2.(a)에 보인 것처럼 이렇게 준비된 코드북과 화자별 벡터시퀀스 X_i 를 워핑계수 α_n ($\alpha_n \in A$)으로 워핑한 벡터 $X_i^{\alpha_n}$ 를 비교하여 (그림 2.(a)의 ①), 최소의 양자화 오차를 갖게 하는 계수 $\hat{\alpha}_i$ 를 찾는다. 모든 학습용 화자에 대해서 화자별로 계수를 찾은 다음, 이 계수들로 발화를 워핑하여 새로운 특징벡터의 집합을 구하고 (그림 2.(a)의 ②), 이를 이용하여 다시 코드북을 생성한다. 이상의 과정을 반복하면 모든 화자들의 계수 $\hat{\alpha}_i$ 는 특정한 값에 수렴하여 더 이상 변하지 않으며 이 때 LMP는 종료된다. 이 수렴상태의 코드북은 주어진 초기의 코드북에서 출발하여 구할 수 있는 하나의 최소 양자화 오차를 갖는 상태의 코드북임을 선행연구[3]에서 정성적으로 분석하였다.

3.2. Progressive Refinement Process

먼저 PRP의 필요성을 설명하기 위해 CSR에서 한번의 LMP 구동만으로 코드북을 학습시킬 경우 발생할 수 있는 문제를 살펴보기로 한다. CSR의 경우, 음소 및 변이음의 종류가 많으므로 코드북의 크기가 커야하고, 이에 따라 코드북의 근접한 centroid 사이의 거리는 가까워진다. 이런 큰 코드북에 어떤 벡터를 가능한 모든 계수로 워핑하여 가장 작은 양자화 오차를 갖는 계수를 찾는다면, 다른 계수를 적용한 같은 발화의 벡터들 사이의 거리가 인접한 centroid 사이의 거리보다 더 커질 수 있게 되고, 결과적으로 잘못된 음소로 분류된 채 코드북을 학습시키게 될 가능성이 높아진다. 이러한 가능성을 방지하기 위해 처음에는 작은 크기의 코드북을 학습시킨 다음 점차 코드북의 크기를 증가시켜 나가는 PRP를 도입한다.

PRP는 그림 2.(b)에 나타낸 것처럼, 최초에 코드북의 크기를 작게 설정하고 (예를 들어, 연속 음성의 경우 16, 혹은 32) 워핑하지 않은 학습용 벡터로 LMP를 구동한다. LMP가 수행되고 나면 화자별로 중간 단계의 워핑계수가 결정되는데, PRP는 이 계수로 워핑한 벡터와 증가시킨 코드북의 크기로 또 다시 LMP를 구동하며, 이 과정을 최종적인 크기의 코드북이 얻어질 때까지 반복한다. 본 연구에서는 LGB 양자화기를 이용하여 LMP를 구동할 때마다 코드북의 크기가 2배로 증가하게 하였다. 아래에 어떻게 PRP를 이용하여 만족스러운 워핑계수를 추정할 수 있는지에 대해 정성적으로 설명한다.

코드북의 크기는 양자화기의 Voronoi 셀의 수와 같으므로, 최초로 LMP를 구동할 때 코드북의 크기가 작다면 LMP는 모든 화자의 벡터들을 이 작은 수의 Voronoi 셀에 할당하게 된다. 첫 LMP의 구동 결과 얻어지는 Voronoi 셀들은 워핑되지 않은 벡터로 만들어진 초기 코드북의 Voronoi 셀들과는 달라지겠지만, 코드북의 크기가 충분히 작다면, 따라서 각 Voronoi 셀을 대표하는 centroid들 사이의 거리가 충분히 크다면, 첫 LMP의 구동에 의해 워핑된 벡터들의 클러스터는 워핑되지 않은 벡터들의 클러스터와 크게 달라지지 않을 것이다. 다시 말해 대부분의 벡터들은 워핑된 후에도 워핑되기 이전과 같은 클러스터에 속하게 될 것이며 이것은 워핑에 의해 생기는 특징벡터의 변화가 cnetroid들 사이의 거리보다 작음을 의미하기도 한다.

LMP는 화자별 발화에 워핑계수의 집합 A 의 모든 계수를 차례로 적용한 다음, 이들을 현재의 코드북과 비교하여 가장 작은 오차를 갖게 하는 계수를 찾아낸다. 따라서 위에서 언급한 것처럼 워핑전후의 벡터들이 같은 클러스터에 속하게 된다면, LMP는 화자별로 워핑된 벡터들을 centroid, 즉 모든 화자들의 평균과 비교하여 가장 가깝게 워핑된 것을 선택하는 것이 된다. 이렇게 코드북의 크기가 작은 상태에서 얻어지는 워핑계수의 값은 정확하지는 않지만 화자별 워핑의 방향, 즉 음성신호의 파워 스펙트럼을 압축해야하는지 아니면 확장해야하는지는 대략적으로 표현할 수 있어 부분적인 워핑을 가능하게 한다.

두 번째와 그 이후의 LMP를 구동할 때 PRP는 코드북의 크기를 증가시킴으로써 이전의 워핑 결과를 더 세분화한다. 이 때의 코드북의 학습 데이터는 근사적이기는 하지만 워핑되어 있어 코드북의 크기가 어느 정도 커져도 잘못 분류됨으로써 틀린 워핑계수를 가지게 될 가능성은 낮아진다.

표 1. 워핑 전후의 인식실험 결과
Tab. 1. Recognition Rates with Warping.

	Codebook Size	Accuracy(%)		Improv.
		m3	m4	
Baseline	-	52.67	52.79	-
Vowels	512	54.91	54.53	2.12
All Phones	1024	54.52	54.37	1.73

IV. 실험 및 고찰

이 장에서는 제안한 워핑계수 추정방법의 성능을 시험하기 위해 수행한 실험과정과 결과를 기술한다. 먼저 실험에 이용한 인식기와 코퍼스 및 워핑계수 추정방법에 대해 설명한 다음, 인식 실험결과를 제시하고 같은 코퍼스를 이용한 다른 연구결과와 비교하여 고찰한다.

4.1. 인식기와 코퍼스

본 연구에서는 Cambridge대학에서 개발한 HTK 음성인식 소프트웨어 툴킷을 이용하여 음소단위의 인식 실험을 수행하였다. 이 툴킷을 음소별로 3개의 state를 갖는 diagonal variance, continuous density HMM로 설정하였으며, 각 state는 3개, 또는 4개의 gaussian mixture model (각각 m3, m4로 표기)로 모델링하였다. TIMIT 음성 코퍼스를 이용하여 음소 인식실험을 하였는데, 이 코퍼스의 음성 데이터는 잡음이 차단된 녹음실에서 16KHz, 16bit/sample 의 비율로 샘플링한 것이다. 이 코퍼스의 SA1 및 SA2를 제외한 모든 학습용 데이터로 인식기를 학습시켰으며, 인식시험에서는 모든 시험용 데이터를 이용하였다.

인식 실험에 사용할 특징벡터를 구하기 위해 먼저 음성 데이터를 0.97의 계수를 갖는 1차 프리엠퍼시스 필터로 필터링한 다음, 초당 100개의 비율로 25ms길이의 해밍 윈도우를 적용한 프레임을 추출하였다. 이 프레임으로부터 파워스펙트럼을 구하여 26채널의 멜 필터뱅크를 적용한 다음 12개의 MFCC계수를 구하였다. 최종적인 특징 벡터는 이렇게 구한 12개의 MFCC 계수와 각 프레임의 에너지, 그리고 이들의 1차 및 2차 미분을 포함시킨 39차원의 벡터이다.

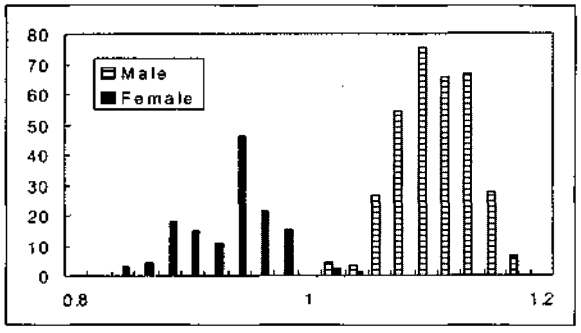

그림 3. 모음을 이용하여 양자화기를 학습시켰을 때의 워핑계수의 분포
Fig. 3. Distribution of warp factors when vowels are used to train vector quantizer.

4.2. 워핑계수 추정 및 정규화

본 연구에서는 워핑계수를 추출하기 위해 두 가지의 음소집합을 이용하여 실험하였다. 하나는 TIMIT 코퍼스의 발화들 중 13개의 모음 (iy, ih, eh, ey, ae, aa, ah, ao, ow, uh, uw, ux, er)을 코퍼스의 label 정보를 이용하여 추출한 모음들의 집합이며, 또 다른 하나는 모음과 자음, 그리고 sp (short pause)까지 포함하는 모든 음소들의 집합이다.

워핑계수를 추정하는데 이용한 특징벡터는 다음과 같이 구하였다. 먼저 주어진 계수에 해당하는 그림 1의 구간선형 워핑함수를 이용하여 발화의 파워스펙트럼을 interpolation/decimation방법으로 워핑하였다. 워핑된 파워스펙트럼에서 F4의 주파수에 해당하는 3600Hz까지의 성분을 추출한 다음 20채널의 멜 필터뱅크를 적용하여 12개의 MFCC계수를 추출하고, 이 것을 이용하여 벡터 양자화기의 학습과 워핑계수 추정에 이용하였다. TIMIT 코퍼스에 포함되어 있는 남성화자 326명, 여자화자 136명의 발화 중, 남여 성비의 균형을 위해 각각 136명씩, 모두 272명의 화자를 임의로 선택하여 양자화기의 학습에 이용하였다. 정규화 양자화기가 만들어진 다음에는 모든 학습용 화자 462명과 시험용 화자 168명의 워핑계수를 이 양자화기와 비교하여 워핑계수를 추정하고, 이 워핑계수에 따라 워핑한 특징벡터로 인식기를 학습

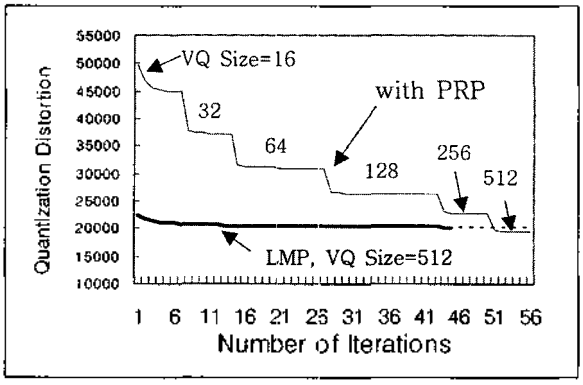

그림 4. PRP를 이용했을 때와 이용하지 않았을 때의 양자화오차의 감소 추세
Fig. 4. Reduction of Quantization Distortions with and without PRP.

및 시험하였다.

워핑계수 추정에 사용한 벡터 양자화기는 LBG 양자화기[10]이며 PRP에서 구동한 양자화기의 크기는 16에서 시작하여 두 배씩 증가 시켰는데, 모음으로 양자화기를 생성하였을 때는 512까지, 모든 음소를 이용하였을 때는 1024까지 증가시켰다.

4.3. 인식실험 결과 및 고찰

4.2절에 기술한 것처럼 양자화기의 학습 및 워핑계수 추출 과정에서는 모음, 또는 모든 음소 전체를 이용한 두 가지로 나누어 실험 하였으나 인식 실험에서는 모든 음소를 이용하여 인식기를 학습시킨 다음 인식시험을 수행하였다. 인식기의 학습에는 SA1 및 SA2 문장은 제외하였으며, 인식기의 시험에는 모든 문장을 다 이용하였다.

아래의 표1에 제안한 방법으로 화자 정규화과정을 거친 다음의 인식결과를 Baseline 인식 결과와 비교하여 보인다. 이 표에서 보는 것처럼 모든 음소를 이용하여 벡터 양자화기를 생성했을 때(All Phones) 보다 모음만을 이용하여 양자화기를 생성했을 때(Vowels) 더 좋은 결과를 얻을 수 있었다. 또 워핑하지 않은 경우에는 state당 4개의 Gaussian mixture를 가지게 했을 때(m4)가 3개의 Gaussian mixture (m3)를 갖게 했을 때보다 인식률이 높았으나 워핑한 다음에는 m3의 인식률이 m4의 그것보다 높아졌다. 이 것은 워핑 결과 화자들 사이의 변이가 줄어들었기 때문이라고 해석된다. 모음을 이용하여 워핑계수를 추정한 결과 Baseline (m4 기준)보다 약 2.1%의 인식률 향상을 얻을 수 있었다.

본 연구와 같이 화자 정규화를 수행한 다음 TIMIT 코퍼스를 이용하여 음소단위 인식 실험을 수행한 연구로 Bacchiani의 것[11]을 들 수 있다. 그는 두 가지 방법으로 정규화 실험을 하였는데, 그 중 하나는 포만트를 이용한 방법으로 모든 화자의 모음 중 F3 주파수의 미디언(Fc)과 미지의 화자의 F3의 미디언(Fs)을 구한 다음, 그 비 (Fs/Fc)를 워핑계수로 추정하였으며, 인식 실험 결과 Baseline 인식률이 43.6%인 HMM인식기에서 2.0%의 절대 인식률 향상을 보였다. 또 다른 한 가지 워핑계수 추정 방법에서 그는 frame 기반 특징벡터가 아닌 segment 기반 특징 벡터를 이용하였는데, 음소별로 하나씩의 Gaussian 분포를 만든 다음, 모든 음소의 분포를 하나의 multi-variate Gaussian mixture로 모델링하고 이 모델을 반복적으로 학습 시켰다. 미지의 화자의 워핑계수를 추정하기 위해 주어진 발화를 모든 계수로 워핑한 다음, 이 모델과 통계적으로 비교하여 워핑계수를 추출하였다. 이 방법으로 그는 2.2%의 인식률 향상을 얻을 수 있었다.

다음 그림 3에 제안한 방법으로 구한 학습용 화자의 워핑계수의 분포를 보인다. 이 분포는 모음 집합으로부터 구한 것이며 양자화기의 크기가 512일 때의 것이다. 이 그림에서 보는 것처럼, 남여 화자의 워핑계수는 확연히 구별되며 남성의 경우 스펙트럼의 확장이, 여성의 경우에는 압축이 필요함을 알 수 있다. 그림 4에는 제안한 방법의 모음을 이용한 실험 (Vowels)에서 양자화 오차가 감소하는 추세, 그리고 PRP를 이용하지 않고 처음부터 양자화기의 크기를 512로 했을 때의 양자화오차를 보인다. PRP를 이용했을 때의 오차가 이용하지 않았을 때의 오차보다 약간 적음을 알 수 있다. 또 PRP를 이용하지 않을 때는 수렴하기까지 많은 LMP iteration (46회)이 필요하였다.

V. 결 론

화자독립 음성인식기를 위한 화자 정규화의 한 가지 방법으로 벡터 양자화기를 이용하는 방법을 제안하였다. 이 방법은 간단한 어휘의 음성인식기를 위해 이미 연구된 화자정규화 방법을 개선하여 연속음성 인식기를 위한 화자 정규화가 가능하게 한 것이다. 이 방법은 같은 코퍼스를 이용한 기존의 연구와 비슷한 성능을 가짐을 음소 인식실험을 통해 확인하였으며, 기존의 포만트를 이용한 방법에서처럼 포만트를 직접적으로 구하지 않고 워핑할 수 있는 장점이 있다. 또 인식기를 반복적으로 학습시킨 다음 통계적인 방법을 이용한 Lee[2]등의 방법과 비교할 때 학습 데이터의 워핑계수를 구하기 위한 인식기의 반복적인 학습을 보다 간단한 양자화기의 반복적인 학습으로 대치한 것으로 볼 수 있다. 그러나 제안한 방법을 이용하여 시험용화자의 워핑계수를 추출하기 위해서는 시험용 발화를 모든 가능한 계수로 워핑해서 특징벡터를 추출한 다음, 양자화기와 비교해야 하는 소위 'exhaustive search'의 단점은 남아 있다. 이를 개선하기 위한 방편으로 통계적인 모델을 이용하는 방법 등을 고려할 수 있으며, 향후 더 많은 연구가 필요하다. 일반적으로, 화자 정규화를 위해 많은 연구를 해 왔음에도 불구하고 대용량 연속음성의 경우 파워스펙트럼의 단순

한 워핑을 통한 화자 정규화로 얻을 수 있는 성능향상은 아직 미미한 수준이다. 주파수 워핑과 더불어 특징벡터의 성분별 크기 정규화, 혹은 LDA를 이용하는 방법 등이 같이 고려되면 좀 더 나은 성능향상이 이루어질 수 있을 것으로 기대된다.

참 고 문 헌

1. P. Zhan and A. 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.
2. 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.
3. 신옥근, "DHMM 음성 인식 시스템을 위한 양자화 기반의 화자 정규화", 한국음향학회지, 22(4), 299-307, 2003.
4. S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, The HTK Book, ver. 3., Microsoft Corp., 2000.
5. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet and N. L. Dahlgren, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus: CDROM, NIST., 1993.
6. S. Umesh, L. Cohen and D. Nelson, "Frequency Warping and the Mel Scale", IEEE Signal Processing Letters, pp.104-107, 9(3), March 2001.
7. S. Molau, S. Kanthak and H. Ney, "Efficient Vocal Tract Normalization in Automatic Speech Recognition", Proc. ESSV, 209-216, Sept. 2000.
8. E. Edie and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization", Proc. ICASSP'96, 346-349, 1996.
9. J. Hogberg, "Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficient", Speech, Music and Hearing Quarterly Progress and Status Report, 33, 41-49, Institutionen for tal, musik och horsel, 1997.
10. Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, 28(1), 84-95, 1980.
11. M.A. Bacchiani, Speech Recognition System Design Based On Automatically Derived Units, Ph. D. Thesis, Boston University, 1999.

저자 이력

• 신 옥 근 (Ok-keun Shin)

1958년 3월 29일생
1981년 서강대학교 전자공학과 졸업 (학사)
1983년 부산대학교 전자공학과 (공학석사)
1989년 프랑스 Universite de Franche-Comte (공학박사)
1983년~1995년 한국전자통신연구소 선임연구원
1995년~현재, 한국해양대학교 IT공학부 부교수
※관심분야: 신호처리, 음성신호처리, 음성인식