멀티 VQ 코드북을 이용한 화자확인 시스템의 성능개선

이재희*, 이상철*, 정연해**
동서울대학*, 한국전력기술인협회**

The Improvement Performance of Speaker Verification System Through the Multi-Vector Quantization Codebook Structure

Lee Jae-Hee*, Lee Sang-Cheol*, Jung Yeon -Hai**
Dong Seoul College*, KEEA**

Abstract - In this paper, we propose the new method that separate the existing common VQ code book into two parts, one is the common VQ code book which is the half of existing common VQ code book, another is the personal speaker VQ code book which accommodate the personal speaker characteristic, variation to improve the performance of the text-dependent speaker verification system using discrete HMM. We apply the propose method in this paper to the text-dependent speaker verification system using discrete HMM and have the improvement performance of about 0.24% compared to existing method

1. 서 론

최근에 연구된 생체인식기술 중에서 가장 각 개인의 고유한 특성을 잘 나타내는 것은 홍채, 지문, 음성 순으로 알려져 있다. 그러나 홍채나 지문 등의 생체인식기술은 기본적인 컴퓨터 환경 이외에 부가적인 장비(홍채 스캐너, 지문 스캐너)를 필요로 한다. 현재로는 그 장비가 표준화 되어있지 않아 보편적으로 많이 사용되기에는 여러 가지 문제점들이 있다. 음성정보를 사용하는 화자인식의 경우는 멀티미디어 컴퓨터의 보급으로 인해 음성 디바이스 인터페이스(Device Interface)가 표준화 되어있다. 또한 휴대용 통신단말기중 휴대폰, PDA등에는 마이크 및 음성 디바이스가 기본 장비로 장착되어 있어 별도 부가장비의 휴대 및 장착 없이 음성을 이용한 사용자확인에 사용될 수 있는 간편한 생체인식 기법으로 최근에 선호되고 있다.

2. 화자확인 시스템의 개요

2.1 화자확인시스템의 개념

화자확인시스템의 전반적인 구성은 그림1과 같다

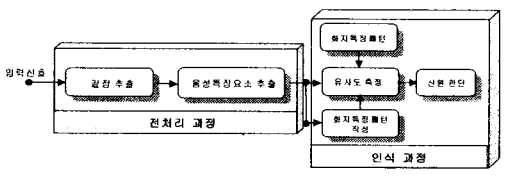

그림 1 화자확인시스템의 구성
Fig. 1 The structure of speaker verification system

화자확인 시스템의 구성은 크게 전처리 과정과 인식과정으로 나눌 수 있다. 전처리 과정에서는 음성구간과 묵음구간을 분리하는 끝점(End-Point) 추출 후 각 개인의 특성을 잘 나타내는 음성 특징요소(Feature Parameter)를 추출한다. 인식과정에서는 추출된 음성특징요소와 이전에 학습되어 등록된 화자특징패턴간의 유사도를 측정하고 그 값을 문턱치와 비교한 후 확인여부를 수행한다. 입력된 음성신호에는 화자의 정보와 더블어 잡음이 섞여 있다. 이 가운데는 발음한 화자의 신원에 관한 정보외의 정보도 있다. 화자 인식을 위한 관점에서 바라볼 때, 화자의 신원에 관한 정보 이외에 다른 모든 정보는 일종의 잡음이다. 이러한 잡음은 화자 인식률 저하의 원인이 된다. 그러므로 화자신원에 관한 정보만 포함하고 그 이외의 모든 정보는 억압한 음성 특징요소를 전처리 과정에서 추출해야 한다. 잡음에 강인한 음성특징요소 추출에 관한 연구도 지속적으로 진행되고 있다[1] 화자인식시스템의 인식과정에 수행되는 인식방법으로는 크게 벡터 양자화(Vector Quantization), 동적정합법(Dynamic Time Warping : DTW)와 같은 패턴정합을 이용하는 방법과 HMM(Hidden Markov Model)을 이용한 확률적인 방법 그리고 인간의 두뇌를 모델링한 신경회로망(Neural Network)을 이용한 방법등이 있다. 이중에서 문장종속 화자확인시스템에서 사용되는 대표적인 인식방법으로는 주로 HMM 기법을 사용하고 HMM기법 중에서도 실제 환경에서 화자확인스시템 구현하기 위해서는 이산 HMM(Discrete HMM)을 사용하고 있다[2].

2.2이산 HMM(Discrete HMM)

HMM 모델은 음성인식에서 음성 인식에서 음성의 주파수 형태를 통계적인 상태로 잘 표현함으로 그 성능이 좋으며, 가장 많이 사용하는 방법이다. HMM의 확률 파라미터로는 초기 상태를 의미하는 초기 확률 밀도 함수 π와 상태의 천이를 의미하는 천이 확률 밀도 함수 A, 그리고 음성 주파수 영역의 코드워드와 관계되는 관측 확률 밀도 함수 B 등이 있다. 이러한 HMM의 파라미터들은 λ로 쓸 수 있다. 그 표현은 다음 식(2-1)과 같다.

$$\lambda = (A, B, \pi) \qquad (2\text{-}1)$$

이때, 상태천이 확률은 $A = \{ a_{ij} \}$, $a_{ij} = p[q_{t+1}=j \mid q_t=i]$ 로써 표현되며 이것의 의미는 상태 j에서 관측 심볼의 확률 $B=\{ b_j(k) \}$, $b_j(k)=p[o_t=v_k \mid q_t=j]$ 의 천이 정도의 확률을 의미하고 있다. 여기서 초기상태 $\pi = \{ \pi_i \}$, $\pi_i = p[q_1=i]$ 는 각 상태의 초기 확률 값을 의미한다. HMM은 모델에 따라 이산 HMM(Discrete HMM), 반연속 HMM (Semicontinuous HMM), 연속 HMM (Continuous HMM)로 분리된다. 이산 HMM은 임의의 음성 프레임에서 결정된 VQ의 인덱스가 HMM의 관측 확률값를 결정한다. 즉 음성의 관측 열 o_t에 대해 선택된 VQ는 v_k가 된다.

다음 식 (2-2)에서 $\delta(o_t, v_k)$는 임의의 시간에 관측되는 여러 음성 데이타 o_t와 추정 하고자 하는 L개의 코드 워드 중 임의의 k 번째 음성 VQ 데이터 v_k로 이루어진 식이다.

$$\delta(o_t, v_k)=1,\ 0_t=v_k\ \ 1\le k\le L\ \ =0,\ otherwise \qquad (2\text{-}2)$$

이 식에 의하여 각 상태에 대한 관측 확률값을 $B=\{\ b_j(k)\}$를 구한다. 다음 그림3은 6-states 이산 HMM으로 모델링 한 상태 천이가 가능한 음성 (phoneme, word) 모델을 나타낸 것이다.

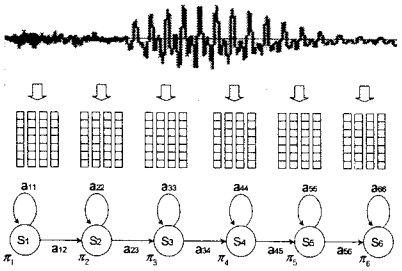

그림 2 상태의 수가 6인 이산 HMM
Fig. 2 Discrete HMM with 6-state

2.3 문턱치(Threshold) 결정

화자확인 시스템은 음성이 입력되면 화자의 음성특징을 추출하여 벡터양자화과정을 수행한 후 공통모델과 개인모델로부터 확률값을 구한 후 그 확률값간의 비례값을 문턱치와 비교하여 화자 확인을 수행한다. 화자확인 시스템을 실제적으로 구현하기 위해서는 사전 문턱치를 결정하는 방법이 필요하다. 일반적인 화자확인 연구에서는 문턱치 결정은 ERR(Equal Error Rate)을 갖도록 사후에 결정된다. 그러나 실제 구현환경에서는 문턱치를 사후에 결정될 수 없으므로 화자 확인 전에 문턱치를 사전에 결정해서 보유하고 있어야 한다. 화자 확인시스템의 에러(Error)는 오류 거부(False Rejection)과 오류 수락(False acceptance)의 두 가지 에러로 나눌 수 있는데, 이 두 가지 에러는 서로 상관관계에 있다. 즉, 오류 수락(FA)에러율을 줄이기 위해 문턱치를 높이면 오류 거부(FR) 에러율이 증가하고 반대로 오류 거부 에러율을 줄이기 위해 문턱치를 너무 낮추면 오류 수락 에러율이 증가하게 된다. 실제 시스템 구현시 이러한 관계를 잘 고려한 적절한 문턱치를 미리 보유하고 있어야 한다.

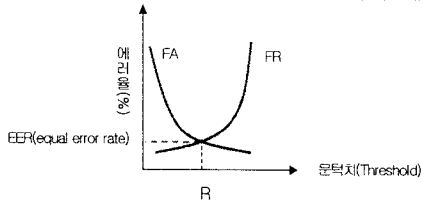

그림 3 문턱치 결정
Fig. 3 The decision of threshold value

화자확인 과정을 V라하고 I 를 확인이 요구되는 신원 그리고 K가 정보라 하고 학습되어 만들어진 참조패턴 T_s와 입력 샘플의 유사도 측정 M과 미리 정해진 문턱치을 θ 로 표현하면 다음 식 (2-3) 같이 나타낼 수 있다.

$$V:(S, I, K)=\begin{pmatrix} 0, M(S,I,T_s)<\theta \\ 1, M(S,I,T_s)\ge\theta \end{pmatrix} \qquad (2\text{-}3)$$

화자인식기법을 이산 HMM을 사용하였을 경우 문턱치 θ 는 참조패턴 T_s와의 비터비 계산(viterbi scoring) 방법에 따라서 달라지므로 정규화 할 필요가 있다. 기존의 우도비(likelihood ratio) 정규화 기법으로는 월드모델(world model)기반의 정규화 기법[3]과 군중 모델(cohort model)기반의 정규화 기법[4]이 가장 많이 사용 되고있다. 월드모델은 문장종속 화자독립모델을 이용한다. 즉, 월드모델과 테스트 샘플간의 비는 화자 S의 틀(Template) P(O|M(S,W))의 우도비를 정규화 하는데 사용된다. M (S_{world}, W)를 문장 W에 대한 월드모델이라 하면 월드모델의 정규화는 식(2-4) 와 같다.

$$R_{world}=\frac{P(O|M(S,W))}{P(O|M(S_{world},W))} \qquad (2\text{-}4)$$

본 연구에서는 월드모델을 이용하여 사전문턱치를 결정하였다. 다음그림4는 기존의 이산 HMM을 이용하여 문장종속 화자확인시스템을 구성하였을 경우의 인식/학습 동작 절차도 이다.

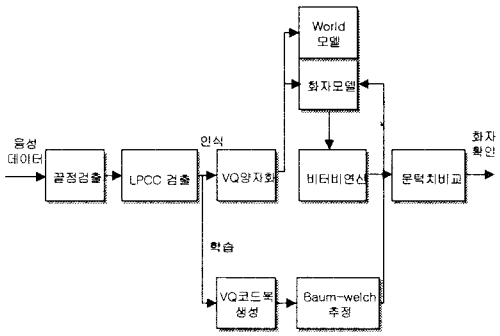

그림 4 이산HMM을 이용한 화자확인시스템 동작도
Fig. 4 The operation diagram of speaker verification system using the discrete HMM

3. 화자확인 시스템의 성능개선을 위한 제안방식

3.1 기존 이산HMM을 이용한 화자확인 시스템

기존의 이산 HMM을 이용한 화자인식시스템의 구조는 다음과 그림5와 같다. 화자의 개인특성을 구분하는 척도로 개인 화자모델을 형성하여 월드모델과의 확률값을 비교한후 문턱치보다 크면 본인으로 인식하고 작으면 거절하는 방식이다. 기존 방식의 경우 VQ 코드북은 사전에 화자를 제외한 사람들의 음성 데이터를 가지고 만든다[5]. 이 방식의 경우 개인의 특성이 VQ코드북에는 반영되지 않고 오직 개인 화자 HMM모델에만 반영된다.

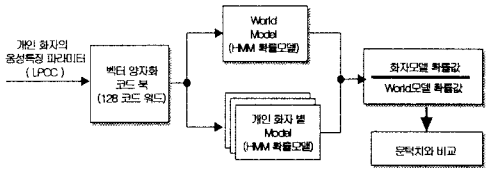

그림 5 기존의 화자확인시스템의 구성도
Fig. 5 The structure diagram of existing speaker verification system

3.2 제안한 화자확인 시스템

본 연구에서는 기존방식의 동일한 VQ코드북 크기를 가지면서도 각 개인의 음성특성 및 음성의 변이성을

보다 잘 반영하기 위해 VQ코드북을 1/2로 나누어 반은 사전에 화자를 제외한 음성데이터를 이용하여 기존의 공통 VQ코드북을 형성하고 반은 개인 음성 데이터로만을 이용하여 개인화자 VQ 코드북을 형성하는 방식을 제안 하였다. 제안하는 방식의 경우 VQ 코드북의 크기는 증가하지 않으면서도 개인화자의 특성이 VQ코드북의 일정영역에 집중적으로 반영되므로 인해 보다 안정적인 화자인식성능을 갖는다. 다음 그림7은 제안한 화자확인시스템의 구성도 이다. 공통 VQ코드북은 화자 본인을 제외한 화자들의 음성데이터를 이용하여 VQ코드북을 형성하고 개인 VQ코드북은 화자 개인의 음성데이터만을 이용하여 형성한다.

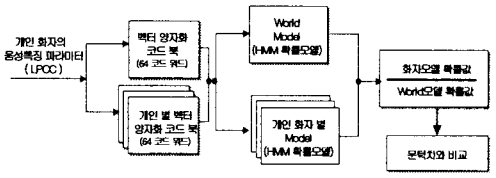

그림 6 제안방식의 화자확인 시스템의 구성도
Fig. 6 The structure diagram of proposing speaker verification system

공통 VQ코드북은 시스템 동작이전에 미리 구성해 놓고 개인 코드북은 화자개인의 음성 데이터를 학습 등록할 때 실시간으로 구성한다. 개인 코드북 구성시 적은 학습 데이터로 안정된 VQ코드북을 구성하기 위해 LBG 알고리듬을 이용한다. 개인 VQ코드북을 두는 구조로 화자확인시스템을 구성하였을 경우 화자인식성능이 개선되는 동시에 개인화자의 음성변화를 VQ코드북 전체에 반영하여 VQ코드북을 갱신하는 것이 아니라 개인 VQ코드북에만 반영함으로써 VQ코드북 갱신시간을 단축 할 수 있고 개인화자의 음성변이성에 적응할 수 있다. 개인 VQ 코드북을 형성하여 화자확인시스템을 구성하는 제안방식은 다음 그림8과 같다.

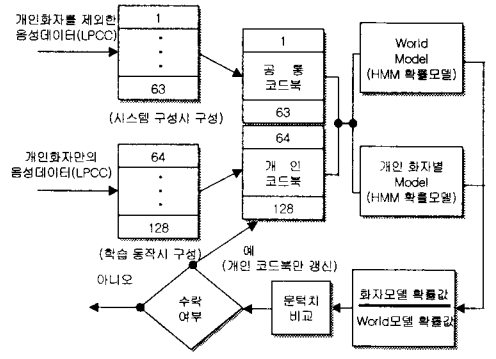

그림 7 제안방식의 화자확인시스템 동작도
Fig. 7 The opeartion diagram of proposing speaker verification system

4.. 제안한 방식의 컴퓨터 모의 실험결과

본 연구에서 제안한 방식과 기존방식간의 성능비교를 컴퓨터 모의실험 하였다. 일반 실험실 환경에서 20대 후반의 남성과 여성을 각각 12명을 선발하여 이중 7명을 회원으로 구성하고 회원으로 등록되지 않은 5명을 사칭자로 하여 오류 거부률(FR)과 오류 수락률(FA)을 각각 평가하여 에러률을 표시하였다. 음성신호의 표본화 주파수는 1102Hz 하였고 양자화 레벨은 16비트로 하였다. 문장 "안녕하세요"와 "하품하지마"를 사용하여 제안한 화자별 개인 코드북을 갖는 방식과 기존의 공통 VQ코드북만을 갖는 방식을 비교 평가하였다. 본 연구에서 제안한 LPC 켑스트럼방식(LPCC)을 적용한 후 기존의 공통 VQ코드북 방식은 VQ코드북 크기를 128개로 구성하였고 제안한 방식은 개인별 VQ코드북 64개, 공통 VQ코드북 크기를 64개로 구성하였다. 학습발음수는 6회로 하였다. 표 1과 표 2는 남성 및 여성화자에 대해서 "안녕하세요"라는 문장을 학습 등록한 후 확인실험을 수행한 인식률 결과이고 표3, 표 4는 "하품하지마"에 대한 남성 및 여성화자에 대한 결과이다.
표 1과 표 2의 결과를 보면 개인별 VQ코드북을 생성하는 제안방식이 기존방식에 비해 오류 수락(FA)회수가 반으로 감소함을 알 수 있다.

표 1 문장 "안녕하세요"의 남성화자 인식률
Table 1 Recognition rate of man speaker's text "An nyeong ha se yo"

		공통 코드북 (기존방식)	개인별 코드북 (제안방식)
M_aaa	FR	1/30	2/30
	FA	1/150	0/150
M_bbb	FR	0/30	0/30
	FA	0/150	0/150
M_ccc	FR	1/30	1/30
	FA	1/150	0/150
M_ddd	FR	0/30	0/30
	FA	0/150	0/150
M_eee	FR	1/30	0/30
	FA	0/150	0/150
M_fff	FR	2/30	2/30
	FA	1/150	1/150
M_ggg	FR	0/30	0/30
	FA	1/150	1/150
총 합계	FR	2.38% (5/210)	2.38% (5/210)
	FA	0.38% (4/1050)	0.19% (2/1050)
평균(FR+FA)/2		**1.38**	**1.28**

표 2 문장 "안녕하세요"의 여성화자 인식률
Table 2 Recognition rate of woman speaker's text "An nyeong ha se yo"

		공통 코드북 (기존방식)	개인별 코드북 (제안방식)
W_aaa	FR	1/30	1/30
	FA	1/150	0/150
W_bbb	FR	1/30	1/30
	FA	1/150	0/150
W_ccc	FR	1/30	1/30
	FA	1/150	0/150
W_ddd	FR	0/30	1/30
	FA	0/150	0/150
W_eee	FR	1/30	0/30
	FA	1/150	1/150
W_fff	FR	1/30	1/30
	FA	1/150	1/150
W_ggg	FR	1/30	0/30
	FA	1/150	1/150
총 합계	FR	2.85% (6/210)	2.38% (5/210)
	FA	0.57% (6/1050)	0.28% (3/1050)
평균(FR+FA)/2		**1.71**	**1.33**

표 3과 표4도 제안방식이 기존방식에 비해 오류 수락(FA)회수가 약 1/2로 감소함을 알 수 있다.

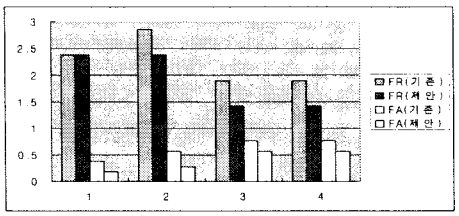

그림 8 각 방식의 성능비교
Fig. 8 The performance comparison of each method

그림 8의 x축의 1과 2는 "안녕하세요" 문장의 남자, 여자의 결과이고 3, 4는 "하품하지마" 문장의 남자, 여자의 결과이다.

표 3 문장 "하품하지마"의 남성화자 인식률
Table 3 Recognition rate of man speaker's text "Ha pum ha gi ma"

		공통 코드북 (기존방식)	개인별 코드북 (제안방식)
M_aaa	FR	2/30	1/30
	FA	1/150	1/150
M_bbb	FR	0/30	1/30
	FA	0/150	0/150
M_ccc	FR	0/30	0/30
	FA	1/150	1/150
M_ddd	FR	0/30	0/30
	FA	1/150	1/150
M_eee	FR	1/30	1/30
	FA	0/150	0/150
M_fff	FR	1/30	0/30
	FA	2 /150	1/150
M_ggg	FR	0/30	0/30
	FA	3/150	2/150
총 합계	FR	1.9% (4/210)	1.42% (3/210)
	FA	0.76% (8/1050)	0.57% (6/1050)
평균(FR+FA)/2		**1.33**	**0.99**

표 4 문장 "하품하지마"의 여성화자 인식률
Table 4 Recognition rate of woman speaker's text "Ha pum ha gi ma"

		공통 코드북 (기존방식)	개인별 코드북 (제안방식)
W_aaa	FR	1/30	1/30
	FA	1/150	0/150
W_bbb	FR	1/30	1/30
	FA	1/150	0/150
W_ccc	FR	1/30	1/30
	FA	1/150	1/150
W_ddd	FR	0/30	1/30
	FA	1/150	0/150
W_eee	FR	1/30	0/30
	FA	1/150	1/150
W_fff	FR	1/30	1/30
	FA	1/150	1/150
W_ggg	FR	0/30	0/30
	FA	1/150	1/150
총 합계	FR	2.38% (5/210)	2.38% (5/210)
	FA	0.66% (7/1050)	0.38% (4/1050)
평균(FR+FA)/2		**1.52**	**1.38**

5. 결 론

본 연구에서는 이산 HMM을 이용하여 문장종속 화자확인 시스템을 구현할 때 화자의 음성특성 과 시간에 따른 음성변이에 적응하기 위해 VQ코드북의 구조를 공통, 개인 VQ코드북, 2개 영역으로 분리하는 방식을 적용하여 화자인식성능을 개선하였다. 제한 방식은 기존 방식에 비해 FR에서는 약간의 인식성능이 개선되었지만 FA에서는 인식성능이 약 2배 정도 향상되었다. 평균적으로 약 2.4%의 인식성능향상을 이루었다.

[참 고 문 헌]

[1] 이재희,"LPC 켑스트럼 가중함수를 이용한 화자확인 시스템의 성능개선", 한국통신학회 논문지,2002 - 12, Vol.27, No. 12T

[2] L. R. Rabiner and B. H. Jung, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[3] C. Fredouille and T. Merlin, "Similarity normaliztion meth ad based on World model and a posteriori probability for speaker verification," Eurospeech, pp. II-983-986, 1999.

[4] Toshihiro Isobe, Jun-ichi Takahashi, "A new cohort nor malization using local acoustic information for speaker ve rification," Proc. ICASSP, Vol. pp. II-841-844, 1999.

[5]Yong Gu, Trevor Thomas, "A hybrid score measurement for HMM - based speaker verification," Proc. ICASSP, V ol. pp.I-317-320, 1999.