韓國 컴퓨터情報學會 論文誌
第10卷 第6號, 2005. 12.

2005-10-6-1-6

주변 잡음 환경에 강한 화자인식 알고리즘 연구

정 종 순*

A study on the robust speaker recognition algorithm in noise surroundings

Jong-Soon Jung*

요 약

대부분의 화자인식 시스템은 음성 분석을 통해 화자의 특징을 음향 파라미터 형태로 추출하여 화자의 표준패턴을 만든 후, 입력된 미지의 음성패턴과의 차이를 계산하여 허용 여부를 최종적으로 판단한다. 화자인식에 사용하는 파라미터는 화자의 특징을 충분히 표현함과 더불어 발성 시마다 변동이 작은 것이 바람직하다. 따라서 본 논문에서도 이를 위해서 다음과 같이 제안하였다. 벡터 양자화모델에 비잡음 환경에 강한 스펙트럼 특징과 잡음 환경에 강한 운율정보를 화자인식 시스템에 이용할 것을 제안하였다. 훈련과정에서 코드북 형성시 실제 데이터를 스펙트럼 특징과 운율 특징을 조합하여 원하는 모델 수만큼 만들었다. 인식과정에서는 입력된 테스트패턴을 각 모델간에 거리 측도로 비교하여 가부를 결정하였다. 실험결과 스펙트럼 특징과 운율 특징을 각각 이용할 경우 보다 좋은 인식율을 얻었으며, 특히 잡음 환경에서 안정된 인식율을 확보하므로 상용화의 가능성을 한층 높혔다.

Abstract

In the most of speaker recognition system, speaker's characteristics is extracted from acoustic parameter by speech analysis and we make speaker's reference pattern. Parameters used in speaker recognition system are desirable expressing speaker's characteristics fully and being a few difference whenever it is spoken. Therefore we suggest following to solve this problem. This paper is proposed to use strong spectrum characteristic in non-noise circumstance and prosodic information in noise circumstance. In a stage of making code book, we make the number of data we need to combine spectrum characteristic and prosodic information. We decide acceptance or rejection comparing test pattern and each model distance. As a result, we obtained more improved recognition rate than we use spectrum and prosodic information especially we obtained stational recognition rate in noise circumstance.

▸ Keyword : 음율정도(prosodic information), 코드북(code book), 피치율(pitch ratio)

• 제1저자 : 정종순
• 접수일 : 2005.10.13, 심사완료일 : 2005.11.07
* 상지영서대학 인터넷영상미디어 부교수

Ⅰ. 서 론

최근 화자인식은 많은 발전과 연구가 이루어지고 있지만 아직 풀어야할 많은 문제를 가지고 있다. 이들 문제 중 대부분은 화자 내 변화, 채널 변화 그리고 녹음 환경의 변화로부터 발생된다. 배경 잡음이 있을 경우 언어구사 스타일이나 운율특징은 변화하지 않으나 스펙트럼 특징은 변화한다. 스펙트럼은 비잡음 상태에서는 운율특징보다 좋은 성능을 보였으나, 잡음환경에서는 많은 성능저하 현상이 나타났다[1].

본 논문에서는 환경에 강한 화자인식 시스템을 위해 두 가지 측면 - 특징 파라미터와 모델 구성을 고려하였다. 우선, 발성자가 가지는 70 - 80 %의 운율 정보를 나타내기 위해 운율 특징을 사용할 것을 제안하였다. 운율정보와 발성습관은 일반적인 스펙트럼 특징과 대조적으로 배경 환경 변화에 민감하지 않았다. 스펙트럼 정보는 잡음 환경에서 인식율이 급격히 저하되었으나 운율 정보는 인식율에 급격한 저하 없이 안정되게 나타났다. 따라서 본 논문은 운율정보를 이용하는 방법에 초점을 맞추었다. 그러므로 스펙트럼 특징과 운율 정보를 조합한 모델을 제안하였다. 그리고 두 번째로 운율 정보를 벡터 양자화 모델에 이용한 화자인식 시스템을 제안하였다. 이들 모델은 기존의 벡터 양자화 모델보다 잡음 환경에 강인함을 보였다.

Ⅱ. 기존 인식 알고리즘

본 절에서는 화자인식에 사용되는 기법들 즉, VQ, HMM, GMM(Gaussian Mixture Model)에 대해 각각의 기법에 대한 간단한 설명과 비교를 하고자 한다.

짧은 시간동안의 학습데이터를 갖는 화자의 특징 파라미터는 그 화자에 필요한 특성을 직접적으로 타나낼 수 있다. 그러나 이러한 표현 방법은 그 학습데이터가 클 때는 비효율적인 수도 있다. 왜냐하면 그 저장 능력이나 요구되는 계산량이 학습데이터의 크기에 비례하여 매우 높아지기 때문이다. 따라서 학습 데이터을 압축하는 효율적인 방법이 제시되었는데 이렇게 원래 신호를 압축하여 그 신호를 대변할 수 있는 벡터로 표시하는 방법을 VQ라 한다[3][4]. VQ 코드북은 특정화자의 특징을 나타내는 효율적인 방법으로써 원래 음성을 나타내는 특징벡터(파라미터)의 작은 집합으로 구성되어 있다. 이러한 VQ 코드북을 얻는 방법으로는 평균 오차를 최소화함으로서 얻어지는 Lloyd 알고리즘이 잘 알려져 있다. 또한 이러한 방식 외에도 최대 기준을 이용한 이른바 커버링 알고리즘, 거리 오차를 계산하는 데 nearst neighbor 규칙 대신 K-nearst neighbor 규칙을 사용하는 등 여러 가지 방식으로 변형된 방법이 사용되기도 한다. VQ를 이용한 화자인식 시스템은 1987년 Soong등에 의해 고립 숫 자음에 대해 실험되었다.

HMM기법은 많은 음성인식에 성공적으로 사용되어지고 있는 기법이다. 이 방법은 음성인식뿐 아니라 화자인식의 경우에도 많은 장점을 가진다. 더욱이 강인 음성 모델을 요할 때는 적은 학습데이터를 가지고도 놀랄만한 성능을 보이고 있다. HMM에서의 음성은 음성을 구성하는 어떤 상태에서 다른 상태로의 천이로 구성된 하나의 연결된 고리로 생각될 수 있다. HMM에서의 상태는 결국 가리워진 즉, 보이지 않는 상태이지만 간접적으로는 음성의 스펙트럼 정보에 의해서 구해질 수 있다. HMM에서의 파라미터는 한 상태에서 다른 상태로의 천이 확률과 각각의 상태에서 발견될 특징 벡터의 확률이다.

Poritz는 1982년에 문장 독립 화자인식을 위한 ergodic HMM을 제안했는데 이 ergodic HMM은 모든 상태의 천이를 허용한 방법이다. 이 방법에서는 하나의 발성음은 음향학적 특징 공간에서 5개의 상태를 가지는 HMM을 통해 각 상태열로 특징지어진다. 5개의 상태는 개의 다른 범위 즉, 유성음, 묵음, 비음, 유음, 탁음, 파찰음으로 나누어지는 음성의 연결인 셈이다. 이러한 Poritz의 연구는 화자인식의 기초가 되는 실험이고 1991년 Tishby는 Poritz의 이론을 확장하여 각각의 상태에 2개에서 8개의 최대 성분을 갖는 연속 확률 분포를 이용한 8개의 상태를 갖는 ergodic AR HMM을 제시하였다[6][7]. 즉 이 방법은 Poritz의 방법보다 분해도가 높아진 것이다. 이밖에도 Naik등은 1989년 장거리 전화선을 통한 화자의 음성을 이용한 HMM방식을 제시하여 이를 화자인식에 적용하였다. 이 방법은 단어 중심의 HMM으로서 문정 종속을 바탕으로 한 실험이었다. 이 방법에서는 모든 공분산(covariance)을 이용하여 최대 4.6%의 오차를 보였는데 이는 6.2%의 오차를 보인 DTW

방식의 성능을 훨씬 앞서는 것이다. 1990년 Rosenberg 등은 PLUs(Phone Like Units)와 ASUs(Acoustic Segment Units)을 이용한 화자인식을 위한 HMM을 제시하였다. PLUs는 발성음의 음성 표기를 근거로 하며 ASUs는 발성의 음성적 성질을 직접적으로 나타낸 것이다. 각각의 부단어 단위는 2개 내지는 3개의 상태를 가지는 left-to-right HMM에 의해 표시되며 각 상태에서 발견될 확률은 가우시안 최대 분포로 나타내어지는 연속확률 분포로써 구할 수 있다. 이러한 HMM 파라미터는 또한 1986년 Rabiner등에 의해 K-평균 알고리즘을 가지고 추정할 수도 있다.

GMM[5]는 다음 식(2.1)과 같이 가중 밀도 합으로 표시된다.

$$c_m = \sum_{R=1}^{M-1} \left(\frac{p}{m}\right) c_k a_{m-R} \quad \cdots\cdots (2.1)$$

여기서 p(wi)는 p(x | wi)의 가중 또한 사전 확률이다. 또한 p(wi)의 모든 합은 적절한 확률 밀도 함수의 특성대로 1이 된다. 혼합 모델에 흔히 사용되는 밀도함수는 평균과 분산 그리고 가중치만으로 충분히 그 분포의 특성을 나타낼 수 있는 가우시안 함수를 많이 사용하고 있다. 혼합 모델의 학습은 EM(Estimate Maximize)알고리즘[6]을 통해 이루어진다. 혼합의 개수는 미리 정해 놓고 처음에는 모든 데이터를 초기 개수로 랜덤하게 또는 임의의 클러스터링 기법을 이용하여 파라미터를 통해 구하며 각각 클러스터링의 특징 벡터의 부분은 가중치를 두게 되는데 여기서 클러스터내의 특징벡터는 가우시안 분포인 경우 평균, 분산이 되게 된다. 이 과정을 M단계라 한다. 이러한 학습 알고리즘으로써의 EM 알고리즘은 국부 최대로의 수렴을 보장한다. 혼합 모델은 VQ 알고리즘과 어떤 성분이나 클러스터링을 이용한다는 점에서 비슷하다[9][10].

III. 제안한 운율정보를 VQ에 이용한알고리즘

3.1 데이터베이스

256개(16*16)의 발성음으로 구성된 데이터베이스-256은 16명(여자: 8명, 남자: 8명)으로부터 추출했다. 각 화자는 16개의 문장을 발성했으며 발성 문장은 〈표 1〉에 보였다. 발성음의 평균길이는 약 2.36sec이고, 최소길이는 1.63sec이며 최대길이는 3.19sec이다. 이들 발성음은 1-16 일련번호를 주었고, 10개(1-10)의 문장은 각 화자의 학습 데이타로 사용했다. 잡음이 없는 경우, 6개(11-16) 문장은 테스트용으로 사용했다. 잡음 환경에서는 4종류의 화이트 가우시안 노이즈를 음성에 섞어 사용했다. 따라서 잡음인 경우, 문장의 전체 수는 1024개(256*4)이다.

표 1. 데이터베이스에 사용한 16개 문장
Table 1. The 16th sentence using Database

일련번호	발성 문장
1	나는 정종순 입니다.
2	나는 김정규 입니다.
3	나는 조동욱 입니다.
4	나는 김산 입니다.
5	나는 김현 입니다.
6	나는 남궁욱분 입니다
7	안녕하세요.
8	이용해 주셔서 감사합니다.
9	음성통신 연구실입니다.
10	다녀오겠습니다.
11	열려라 참께
12	나는 정남자 입니다
13	나는 박성환 입니다.
14	독서의 계절은 따로 없다
15	학교 갑니다
16	이겨라 한국축구

3.2 코드북(codebook) 개선

이상적으로 화자음성의 특징을 나타내기 위해서는 학습 데이터가 화자음성의 특징을 모두 포함하고 있어야 한다[8]. 그러므로 본 논문에서 특정화자의 특성을 잘 나타낼 수 있는 요소 중 하나인 운율정보를 언어학습데이터로 추가하여 사용하였다.

VQ에서, 각 소스 벡터는 코드북이라는 미리 저장된 코드워드 셋 중의 하나로 코드화된다[8] . 코드북은 코드벡터와 학습 벡터 사이의 평균 양자화 왜곡을 최소화하는 것으로 설계한다. 음성에 대한 하나의 VQ 코드북은 특정 음성을 포함하는 학습벡터로부터 설계한다. 따라서 화자인식에서 사용하는 VQ source 코드는 반복되는 특정음성으로 구성되는 학습 벡터로부터 설계된 코드북에 의해서 나타낸다. 화자확인에 대한 하나의 VQ 코드북 $\bar{C}$ 은 학습 연속식 t로부터 얻은 평균 왜곡을 최소화하는 것으로 설계된다. 식 (3.1)은 이를 위한 것이다.

$$\sum_{p=1}^{P} d\,(\bar{t}_p, \bar{c}_B) \quad \cdots\cdots (3.1)$$

$\bar{C}_B$ 는 부호화 음성 세그먼트로 $\bar{t}_p$ 부터 얻은 코드워드이다. 그리고 벡터 왜곡거리 측도 d는 식 (3.2)으로 정의된다.

$$d\,(\bar{t}_p, \bar{c}_B) = \min_i d\,(\bar{t}_p, \bar{c}_i) \quad \cdots\cdots (3.2)$$

코드북은 특정단어를 발성한 한 화자를 표현한다. 코드북 $\bar{C}$ 에서 확인 발성음 v로부터 얻어진 평균 양자화 왜곡 Davg는

$$D_{avg} = \frac{1}{L}\sum_{i=1}^{L} d\,(\bar{v}_i, \bar{c}_B) \quad \cdots\cdots (3.3)$$

본 논문에서는 화자확인 결정을 위해 이 평균 양자화 왜곡을 이용했다. 즉 특징벡터로 구성되는 VQ 코드북은 그 화자의 특성을 특징짓는 효과적인 수단으로 사용되며, 특정화자에 대한 코드북은 각 화자의 학습 특징벡터를 클러스터링하므로 생성된다.

3.3 운율정보를 이용한 코드북

운율 정보를 코드북에 적용하기 위해서 앞 절에서 언급한 피치알고리즘과 피치변경 방법을 선택했다. 운율정보를 코드북에 이용하기 위해서 본 논문에서는 다음과 같은 과정을 거쳤다. 우선, 발성음으로부터 피치 연속식을 추출한다. 그리고 추출된 피치를 필요에 따라 피치변경을 하고 이 피치변경 연속식과 피치 연속식을 이용하여 코드북을 설계한다. 이 코드북 블록도는 (그림 1)에 보였으며 V/UV의 스펙트럼 상에서의 피치 변경법을 이용하여 원래의 음성 스펙트럼을 각각 70%, 80%로 압축하고 110%, 120% 신장시킨 피치 연속식으로 코드북을 구성 하였다. 한 화자당 4개의 코드북을 구성했다. 즉, 발성음으로부터 피치검출 알고리즘에 의해 검출된 피치 연속식과 이것을 피치변경 알고리즘에 의해 스펙트럼 상에서 신장, 압축된 4개의 연속식을 사용하였다.

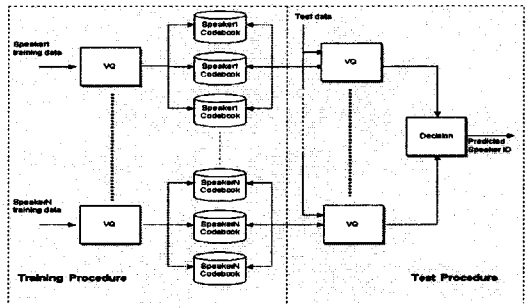


그림 1. 운율정보를 적용한 코드북 블럭도
fig 1. A code book block diagram using prosodic information

테스트 발성음이 입력될 때, 화자의 운율특징 VQ코드북과 테스트 발성음에 피치 연속식과 피치변경 연속식을 비교한다. 이 운율 VQ모델은 (그림 2)에 보였다. 여기서 적당한 크기의 P차를 구하는 것은 중요하다. P값이 너무 크게 되면, 전체 문장과 비교한다. P값이 너무 작으면, 운율 정보가 필요 없다는 것을 의미한다. 이 거리측도는 다음 식으로 정의된다.

$$d\,(t_i, c_B) = \sum_{j=o}^{P} (t_{ij} - c_{Bj})^2$$

$$D_{avg} = \frac{1}{N}\sum_{i=1}^{N} (t_i - c_B) \quad \cdots\cdots (3.4)$$

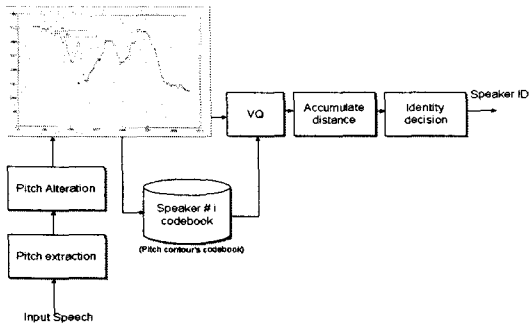


그림 2. 운율정보를 이용한 VQ 모델의 블록도
fig 2. A block diagram of VQ model using prosodic information

화이트 가우시안 잡음 환경에서 운율 정보 이용한 VQ모델의 인식실험을 수행했다. 실험결과는 (그림 3)에 보였는데, 그림에서 보듯 가우시안 잡음에 민감하지 않음을 알 수 있다. 화이트 가우시안 잡음에서 스펙트럼 특징과 비교한 결과를 (그림 4)에 보였다. 결과에서 보듯, 스펙트럼 특징은 잡음에 상당히 민감함을 알 수 있으며 상대적으로 본 논문에서 제안한 운율 특징은 잡음 환경에서 강함을 나타냈다. 비록 운율특징이 비잡음 상태에서는 스펙트럼 특징처럼 좋은 성능을 보이지는 않았으나, 화이트 가우시안 잡음 환경에서는 스펙트럼 특징보다 많이 강인함을 보였다.

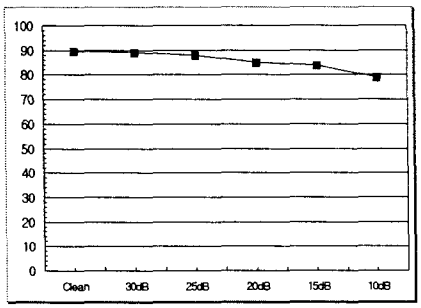


그림 3. 화이트 가우시안 잡음에서 제안한 VQ 모델의 인식률
fig 3. The recognition ratio of VQ model suggested in white gaussian

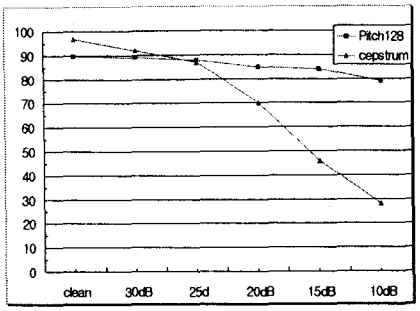


그림 4. 화이트 가우시안 잡음 환경에서의 인식률
fig 4. The recognition ratio in white gaussian noise circumstances

Ⅳ. 조합 모델

4.1 특징 벡터 정규화

조합모델의 신뢰성 있는 구조를 위해서 본 논문에서는 특징벡터의 정규화를 고려했다. 조합모델은 스펙트럼 특징과 운율 특징을 사용하므로, 스펙트럼 특징의 평균과 분산은 운율특징의 평균, 분산과는 다르다[11]. 즉, 이들 동적 영역은 매우 다르다. 그러므로 만약, 본 논문에서 이 특징벡터들 간의 정규화를 실행하지 않는다면 이 모델은 신뢰할 수 없게 될 것이다. 예를 들어, 스펙트럼 벡터의 거리측도 범위는 10,000 - 30,000 사이에 있으나, 운율 특징 벡터는 3,000 - 10,000사이에 있다. 이렇게 동적 영역이 다르므로 정규화를 해야 한다. 본 논문에서는 특징벡터의 분산 정규화를 수행했다. 본 시스템에서는 유클리디언 거리 측도를 이용하였으며, 다음 식과 같이 정의한다.

$$d(x, y) = (x - y)^T (x - y) \quad \cdots\cdots (4.1)$$

x는 N-차 테스트 템플릿을 말하고 y는 N-차 테스트 템플릿을 말한다.

이 조합모델의 마지막 거리측도는 다음 식으로 정의한다.

$$D = \sum_{i=1}^{N} d(x_1, y_1) + \sum_{i=1}^{M} d(x_2, y_2) \quad \cdots\cdots (4.2)$$

x1과 y1은 스펙트럼 특징을 위한 N-차 템플릿이며, x2과 y2는 운율 특징의 M-차 템플릿이다. 본 논문에서는 N=16, M=16을 이용했다. 만약, 정규화를 수행하지 않으면, 거리측도 D는 특히, d(x, y)에 의해서 영향을 받을 것이다. 그러므로 본 논문에서는 정규화를 수행했다. 이 때 사용되는 정규화 거리측도는 다음과 같이 정의된다.

$$d_{Nor}(x, y) = (\frac{x - y}{\sigma})^T (\frac{x - y}{\sigma}) \quad \cdots\cdots (4.3)$$

σ는 XUY 집합에 표준편차이다. 마지막으로 정규화 거리측도 DNor를 구한다.

DNor 는 다음과 같은 식으로 정의한다.

$$D_{Nor} = \sum_{i=0}^{N} d_{Nor}(x_1, y_1) + \sum_{i=1}^{M} d_{Nor}(x_2, y_2)$$
$$d_{Nor}(x_1, y_1) = (\frac{x_1 - y_1}{\sigma_1})(\frac{x_1 - y_1}{\sigma_1})^T$$
$$d_{Nor}(x_2, y_2) = (\frac{x_2 - y_2}{\sigma_2})(\frac{x_2 - y_2}{\sigma_2})^T \quad \cdots\cdots (4.4)$$

σ1은 스펙트럼 특징의 표준편차이고 σ2는 운율특징의 표준편차이다. 본 논문에서는 전체 화자의 발성음으로 부터 σ1와 σ2를 계산했다.

4.2 조합모델의 거리 계산

앞 절에서 이미 잡음환경에서는 운율특징이 효과적이라는 것을 보였다. 그러나 아직 스펙트럼 특징보다 운율 특징이 더 많이 효과적이라는 것을 알 수 없으므로 이 절에서는 거리측도를 정의한다. 본 논문에서는 거리 계산에 영향을 주는 특정 특징벡터에 가중치를 주어서 구한다. 본 논문에서 사용하는 조합모델의 거리측도는 다음과 같이 정의한다.

$$D_{FIN} = \alpha DSPEC + (1 - \alpha) DPROS \quad \cdots\cdots (4.5)$$

DSPEC은 스펙트럼 거리 측도이고, DPROS는 운율특징의 거리이다. α는 가중치 요소이다. 즉 만약, α가 1이라면, 이것은 스펙트럼 특징만을 이용한다는 것을 의미한다. α가 0이라면, 이것은 운율특징만을 이용한다는 것을 의미한다.

4.3 조합모델의 구조

조합 모델의 구조는 (그림 5)에 보였다. 두 특징벡터들을 이용하기 위해서 분산정규화 과정을 거쳐서 마지막 단계인 거리측도에서는 앞 절 4.2에 언급된 가중치 α를 계산해야 한다. 본 시스템은 화자확인 시스템이므로 마지막에 화자의 확인을 위한 값을 얻게 된다.

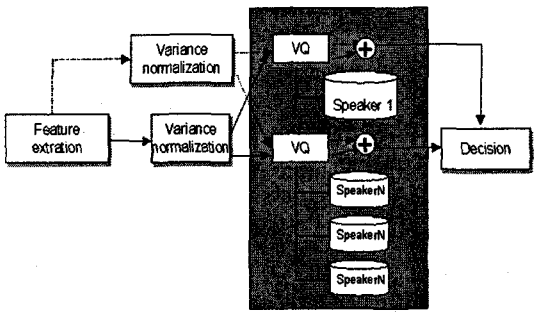


그림 5. 조합모델의 블록도
fig 5. A block diagram of integration model

이 절에서는 최적의 가중치 α값을 얻기 위한 실험이다. 〈표 2〉과 (그림 6)은 α값에 의존하는 인식율을 보였다. 비잡음 환경에서 구한 데이터베이스로 학습데이터를 구했다. 그리고 테스트 데이터는 4가지의 화이트 가우시안 잡음 상태의 음성을 사용했다. 〈표 2〉의 (a)는 α값을 0 ~ 1까지 0.1씩 변화시켰을 때 인식율 표이고, (b)는 α값이 0과 0.05 ~ 0.1사이에서 0.01씩 변화할 때 인식율을 나타냈다. (그림 6)의 (a)는 x축의 α값을 0 ~ 1까지 0.1씩 변화시켰을 때 y축에 인식율 그래프를 보였으며, (b)는 α값을 좀 더 세분해서 0일 경우와 비교했다. 그림에서 볼 수 있듯이, 본 논문에서 제안한 조합모델은 잡음환경에서도 안정된 인식율을 보였다. 그림에서 살펴보면 신호대 잡음비가 30dB일 경우는 α값에 영향을 받지 않고 안정된 인식율을 보였으며, 이와 대조적으로 신호대 잡음비가 10dB일 경우에는 α값에 의존하는 것으로 나타났다. 그러나 가장 좋은 그러나 스펙트럼 특징과 운율특징을 따로 사용할 경우보다 인식율이 안정되어 있음을 알 수 있다. 여기서 가중치 α값이 0.1일 경우가 가장 안정된 인식율을 보였음을 알 수 있다. 좀 더 구체적으로 살펴보면 잡음이 가미될 경우 잡음으로 인해 스펙트럼이 느슨해지거나 약간의 일그러짐이 생겨 잡음 상태에서 인식율 저하가 일어났으나, 조합모델에서 사용한 운율정보 - 정확한 피치를 찾아 피치변경(스펙트럼의 신장과

압축) 가 이 부분을 보안하므로서 인식율 상승효과가 나타나는 것으로 생각된다.

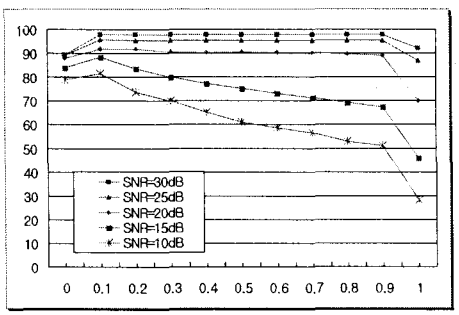
표 2. α값에 따른 조합모델의 인식율
table 2. The recognition ratio of integration model according to α value

(a) α값을 0에서 1까지 0.1씩 변화시켰을 때

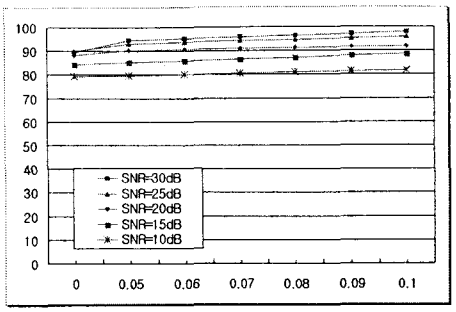
Database	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
30dB	89.50	97.78	97.78	97.78	97.78	97.78	97.78	97.78	97.78	97.78	92.23
25dB	88.91	95.76	95.52	95.46	95.46	95.40	95.40	95.36	95.36	95.36	87.12
20dB	88.14	91.80	91.80	90.63	90.23	90.63	90.23	90.23	89.84	89.06	70.34
15dB	84.25	86.52	83.65	80.16	77.40	75.21	73.16	71.28	69.08	67.26	46.36
10dB	79.11	81.64	73.83	70.14	65.23	60.94	58.59	56.64	52.73	51.17	24.38
Total	85.98	91.10	86.52	86.63	85.22	83.99	83.03	82.26	80.96	80.13	64.08

(b) α값을 α값을 0.05에서 0.1까지 0.01씩 변화시켰을 때

Database	0	0.05	0.06	0.07	0.08	0.09	0.1
30dB	89.50	94.47	95.14	95.80	96.46	97.12	97.78
25dB	88.91	93.10	93.62	94.15	94.67	95.25	95.76
20dB	88.14	90.36	90.65	90.94	91.23	91.52	91.80
15dB	84.25	85.11	85.79	86.47	87.15	87.83	88.52
10dB	79.11	79.57	80.03	80.50	80.95	81.42	81.64
Total	85.98	88.52	89.05	89.57	90.09	90.63	91.10



(a) α값을 0에서 1까지 0.1씩 변화



(b) α값 0과 α값을 0.05에서 0.1까지 0.01씩 변화

그림 6. α값에 따른 조합모델의 인식율 (x축:α값, y축: 인식율)
fig 6. The recognition ration of integration model according to α value

V. 결 론

본 논문에서는 환경에 강한 화자인식 시스템을 위해 두 가지 측면 - 특징 파라미터와 모델 구성 - 을 고려하였다. 우선, 발성자가 가지는 70 - 80%의 운율 정보를 나타내기 위해 운율 특징을 사용할 것을 제안하였다. 운율정보와 발성습관은 일반적인 스펙트럼 특징과 대조적으로 배경 환경 변화에 민감하지 않았다. 스펙트럼 특징을 VQ 모델에 이용할 경우 비 잡음 상태에서는 좋은 인식율을 나타내었으나 잡음 환경에서는 인식율이 50% 이하로 저하되는 단점을 가지고 있었다. 그러나 운율 정보는 인식율에 급격한 저하 없이 안정되게 나타났다. 따라서 본 논문은 운율정보를 이용하는 방법에 초점을 맞추었다.

첫째, 음성의 특징벡터로 운율정보 사용했다. 즉, 입력음성에 동적 변화를 측정할 수 있게 세그먼트로 분할하고 분할된 피치열을 변경하여 인식의 특징패턴으로 사용했다.

둘째, 시스템 모델링 측면에서 벡터 양자화모델에 비 잡음 환경에 강한 스펙트럼 특징과 잡음 환경에 강한 운율정보를 화자인식 시스템에 이용했다. 즉, 훈련과정에서는 코드북 형성시 실제 데이터를 스펙트럼 특징과 운율 특징을 조합하여 원하는 모델 수만큼을 만든다. 인식과정에서는 입력된 테스트패턴을 각 모델간에 거리 측도로 비교하여 가부를

결정한다. 실험결과 비잡음 환경(SNR 30dB)일 경우 스펙트럼 특징과 운율 특징을 각각 이용할 경우 보다 더 좋은 97.8% 인식율을 얻었다. 특히 잡은 환경(SNR 10dB)에서는 안정된 81.6%의 인식율 결과를 얻어 같은 환경에서 스펙트럼의 특징을 이용하여 얻은 인식율 28% 보다 좋은 결과를 얻었다. 또한, 음성합성 시스템에서 주로 사용하는 피치변경을 화자인식에 접목시켜 좋은 인식율을 얻었다는 것이다.

본 논문을 계기로 화자인식 시스템에 운율정보를 좀 더 구체적으로 이용할 수 있게 되었으면 한다. 본 논문에서는 화자식별에 문장 의존형 화자확인 시스템이었으나 좀 더 나아가 문장 독립형 화자식별에도 안정된 인식결과를 보이기 위해서는 다음과 같은 실험이 더 있어야 한다.

첫째, 특정화자의 고유한 음성특징을 잘 검출하기 위해서는 운율정보를 이루고 있는 3대요소인 피치정보, 에너지 정보, 지속시간 정보를 충분히 고려해야 할 것이다.

둘째, 잡음이 섞일 경우 잡음으로 인해 스펙트럼이 느슨해지거나 약간의 일그러짐이 생겨 잡음 상태에서 인식율이 저하가 일어났으나, 조합모델에서 사용한 운율정보 - 정확한 피치를 찾아 피치변경(스펙트럼 신장과 압축) - 가 이 부분을 보완하므로서 인식율 상승효과가 있는 것으로 나타났다.

참고문헌

[1] T. Takagi, E. Miyasaka, "A Speech Prosody Conversion System with a high Quality Speech Analysis-Synthesis Method", Proc. EUROSPEECH'93, pp. 995 - pp. 998, September 1993.

[2] B. E. Caspers, B.S. Atal, "Changing Pitch and Duration in LPC Synthesised Speech using Multipulse Excitation," J. Acoust. Soc. Amer., Vol. 73, No.1, pp. 55, 1983.

[3] A. Gersho and R. M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers. 1992.

[4] J. he, Li Liu, and G. Palm "A New codebook Training algorithm for VQ-base speaker recognition", Proc. of ICASSP'97, pp. 1091 - pp. 1094, 1997.

[5] B. S. Atal, "Automatic Speaker Recognition based on Pitch Con-tours," Journal of the Acoustical Society of America, pp. 1687 - pp. 1697.

[6] B. S. Atal, V. Cuoerman and A. Gersho, Advance in Speech Coding, Kluwer Academic Publisher, 1991.

[7] M. J. Carey et al., "Robust Prosodic Features for Speaker Identification", Proc. of ICSLP, 1996.

[8] J. He, Li Liu, and G. Palm, "A New Codebook Training algorithm for VQ-based speaker recognition", Proc. of ICASSP'97, pp. 1091 - pp. 1094, 1997.

[9] J. S. Jung, Y. J. Kyung and H. S. Lee, "A study on the performance improvement of speaker recognition using average pattern and weighted cepstrum", Proc. of the Speech Comm. and Signal Processing workshop, pp. 179 - pp.183, 1995.

[10] 이창윤, "음성인식 시스템에서의 잡음 제거 개선에 관한 연구", 한국컴퓨터정보학회 논문지, pp. 2 - pp. 6, 2002. 6.

[11] 지진구, 윤성일, "음성을 이용한 화자 검증기 설계 및 구현", 한국컴퓨터정보학회지 논문지, pp. 92 - pp. 98, 2000. 9.

저 자 소 개

정 종 순

1990년 2월 서울산업대학교 전자공학과(학사)

1992년 2월 서울시립대학교대학원 전자공학과(공학석사1)

1996년 2월 한국과학기술원 정보 및 통신공학과(공학석사2)

2002년 8월 숭실대학교 전자공학과 (공학박사)

1997년 3월 문경대학 정보처리과 전임

1998년 2월~현재 상지영서대학 인터넷 영상 미디어과 부교수