Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

화자인식을 위한 관측신뢰도 기반 변형된 HMM 디코더

Modified HMM Decoder based on Observation Confidence for Speaker Identification

Md. Tariquzzaman[1], 민소희[1], 김진영[2], 나승유[2]

[1] 광주광역시 북구 용봉동 전남대학교 전자정보통신공학과
E-mail: tareq_ict_ju@yahoo.com, minsh@chonnam.ac.kr
[2] 광주광역시 북구 용봉동 전남대학교 전자컴퓨터공학부
E-mail: beyondi@chonnam.ac.kr, syna@chonnam.ac.kr

요 약

음성신호는 잡음 또는 전송 채널의 특성에 의하여 왜곡되고, 왜곡된 음성은 음성인식 및 화자인식의 성능을 크게 저하시킨다. 이러한 문제점을 극복하기 위해 본 논문에서는 Gaussian mixture model (GMM)에 적용된 신호대잡음비 (SNR)기반 신뢰도 가중 기법[1][2]을 Hidden Markov model(HMM) 디코더에 변형하여 적용하였다. HMM 디코더 변형은 HMM 상태별 관측확률을 논문 [1]에서 제시된 신뢰도로 가중함으로써 이루어졌다.

제안한 방법의 성능을 확인하기 위해 ETRI에서 만든 한국어 화자인식용 휴대폰 음성 DB를 사용하여 문맥종속 화자식별 실험을 하였다. 실험결과 기존 방법에 비해 제안한 방법의 화자인식률이 크게 향상됨을 확인 할 수 있었다.

Key Words : 화자인식, HMM 디코더, 부정확한 관측, 관측신뢰도

1. 서 론

음성은 잡음 환경 하에서 듣는 사람으로 하여 화자가 누구인지를 쉽게 알 수 있도록 개개인의 특성을 잘 전달해주는 가장 중요한 통신 매체중 하나이다. 그러므로 화자의 정보를 추출하여 화자를 인식하는 시스템관련 기기 개발에 관심이 모아져왔다[2][3][4].

화자인식기술은 국가안전, 통신시스템보호, 컴퓨터 네트워크 보호, 사이버거래등과 같은 다양한 분야에 활용되고 있으며 앞으로도 계속적인 연구를 통한 그 발전 가능성은 매우 크다. 또한 휴머노이드와 같은 인공지능 로봇의 등장으로 자동화자인식 시스템에 대한 요구가 더욱 절실하다.

그러나 음성정보를 사용한 화자인식은 전송 매체의 채널 왜곡이나 주변 환경의 잡음, 코덱의 왜곡 등에 의해 쉽게 손상되어 실생활에서 인식률이 상당히 저하되고 있다. 이러한 문제점을 극복하기 위한 많은 알고리듬이 연구 되었으며, 크게 두 가지 접근방법으로 분류할 수 있다. 첫째는 잡음에 강인한 파라미터를 추출하는 방법이고 [6][7][8], 둘째는 화자의 모델을 잡음에 맞도록 적응시키는 모델 적응방법이다 [9].

최근 새로운 연구 방안 중 하나인 논문 [1][2]는 관측신뢰도의 개념을 도입하여 부정확한 관측을 가지고 있는 문제를 해결하기 위해 변형된 GMM 학습과 인식방법을 제안한 것이다.

본 논문에서는 GMM에 적용된 SNR 기반 관측신뢰도 가중 기법을 HMM 모델에 적용하여 성능을 검토하고자 한다. HMM 기반 인식기에 적용하기 위하여, HMM 디코더의 관측확률을 논문 [1][2]와 유사하게 관측신뢰도를 이용하여 가중하는 변형된 HMM 디코더를 제시한다. ETRI에서 구축된 문맥종속 화자인식용 휴대폰 인식 DB를 사용하여 제시한 알고리듬의 성능을 평가하고 검증하고자 한다.

2. HMM 디코더의 변형

2.1 기본 HMM 디코딩 원리

HMM 모델은 연속 HMM과 이산 HMM으로 나뉘는데, 연속 HMM이 음성인식, 화자인증 또는 인식 등에서 우수한 성능을 보인다. 연속 HMM은 연속 관측 확률밀도함수를 사용하는데 아래의 식 (1)로 표현된다.

$$b_i(x_t) = \sum_{m=1}^{N_m} g_m \Phi_{im}(x_t; \mu_{im}, \Sigma_{im}) \quad (1)$$

$$1 \le i \le N_s$$

$$c_{im} \le 0,\ 1 \le i \le N_s,\ 1 \le m \le N_s \quad (2)$$

$$\sum_{m=1}^{N_m} c_{im} = 1,\ 1 \le i \le N_s$$

위 식에서 x_t는 관측벡터이며 Φ_{im}는 커널함수로 i번째 상태의 가중치가 c_{im}, 평균벡터 μ_{im} 공분산 Σ_{im}과 N_m은 총 개수이다.

커널함수 Φ_{im}는 가우시안분포이며, HMM 모델은 위 관측모델을 포함하여 다음과 같이 표기된다.

$$\lambda = (A, B, \pi)$$

이때 A는 상태천이행렬, B는 방사행렬 이고, π는 초기 확률벡터이다.

관측패턴 $\{X_t\}(t = 1,2,....,T\ )$와 HMM이 $\lambda = (A, B, \pi)$로 이루어진 확률 $P(\boldsymbol{X}|\lambda)$를 구하기 위해 우리는 순방향 과정을 사용하였는데, 다음과 같다.

순방향 과정

확률 $P(\boldsymbol{X}|\lambda)$를 구하기 위해 일반적인 상태열을 $Q = [q_1, q_2, ...q_t, ...q_T]$로 나타내었다. 순방향 변수 $\alpha_t(i)$는 식 (3)으로 표현된다.

$$\alpha_t(i) = P(x_1, x_2, ..., x_t,\ q_t = \boldsymbol{S_i}|\lambda) \quad (3)$$

위 식은 모델 λ이 주어지고 시간 t에서 상태변수가 S_i일때 $t = 1$에서 t까지 관측패턴의 확률이다. 다음 식은 $\alpha_t(i)$를 구하기 위한 반복적인 과정이다.

i) $\alpha_1(i) = \pi_i b_i(x_1),\ 1 \le i \le N_S$ (4)

ii) $\alpha_{t+1}(j) = \left(\sum_{i=1}^{N_S} \alpha_t(i) a_{ij}\right) b_j(x_{t+1})$ (5)

$$1 \le t \le T-1,\ 1 \le j \le N_s$$

iii) $\alpha_T(i) = P(x_1, x_2, ... x_t, ..., x_T,\ q_T = \boldsymbol{S_i}|\lambda)$ (6)

또한 $P(\boldsymbol{X}|\lambda)$은 $\alpha_t(i)$의 전체 합이므로

$$P(X|\lambda) = \sum_{i=1}^{N_s} \alpha_T(i) \quad (7)$$

과 같다.

HMM 분류기는 출력값을 로그함수의 형태를 취하므로 $ll(\boldsymbol{X}|\lambda_n)$으로 간략화 하며 화자인식에 대한 결정규칙은 $\arg Max_i ll(X|\lambda_i)$로 나타낸다.

2.2 관측신뢰도 기반 변형된 HMM 디코딩

모든 측정은 주변의 잡음 또는 관측 오차가 존재하므로, 정확한 측정을 불가능하게 만든다. 논문 [1][2]에서 관측신뢰도라는 개념을 도입하여 부정확한 관측을 가지고 있는 문제를 해결하기 위해 변형된 GMM 학습과 인식방법이 성공적으로 적용되었다.

그러므로 HMM 디코딩 과정에 관측신뢰도를 사용하는 것은 충분히 타당한 접근방법이다. 각각의 관측된 벡터에 대한 적당한 신뢰도 값을 가지고 있을 때, 이를 ρ_t라고 표현하자. ρ_t는 t 번째 관측벡터의 신뢰도를 나타내는 멤버쉽 (membership) 값으로 0과 1사이의 값을 갖는다. 따라서 본 논문에서 관측 패턴은 $\{x_t, \rho_t\}$, $t = 1,2,....., T$와 같으며, 모델은 HMM $\lambda = (A, B, \pi)$와 같다.

제안된 HMM 디코딩 방법은 기본적으로 2.1절의 식 (4)~(7)의 과정을 반복한다. 단, 식 (5)에 사용되는 관측확률 $b_j(x_{t+1})$를 관측신뢰도 ρ_{t+1}에 의하여 가중함으로써 변형되게 된다. 따라서 본 논문에서 제안하는 변형된 디코딩(순방향과정)은 식 (5)를 식 (8)과 같이 변형함으로써 이루어진다.

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N_S} \alpha_t(i) a_{ij}\right) \left\{b_j(x_{t+1})^{\rho_{t+1}}\right\} \quad (8)$$

$$1 \le t \le T-1,\ 1 \le j \le N_s$$

3. 신뢰도기반 화자인식

그림1은 논문[1]에서 제시한 관측신뢰도의 개념을 채용한 HMM 기반 화자식별과정을 보여주고 있다. 그림의 실선은 기존의 화자인식 과정이며 점선은 변형된 화자인식과정을 나타내고 있다.

HMM에 대한 화자인식과정은 다음과 같다.

첫째는 입력음성의 특징파라미터를 추출한다. 본 논문에서는 지금까지 가장 우수한 성능을 보인다고 알려진 멜 캡스트럼(Mel-Cepstrum)을 사용하였다.

둘째는 멜 캡스트럼에 대해 CMS방법을 수행한다. CMS는 채널에 의해 발생하는 채널왜곡을 제거하고 잡음에 의한 파라미터의 오염을 일부 제거하는 성질을 가지고 있다.

셋째는 입력된 CMS 결과 파라미터를 이용

하여 HMM 학습을 수행한다. 학습결과는 화자별로 저장한다. HMM 모델에 대한 발생확률을 계산하여 가장 높은 확률을 갖는 화자로 판단하게 된다.

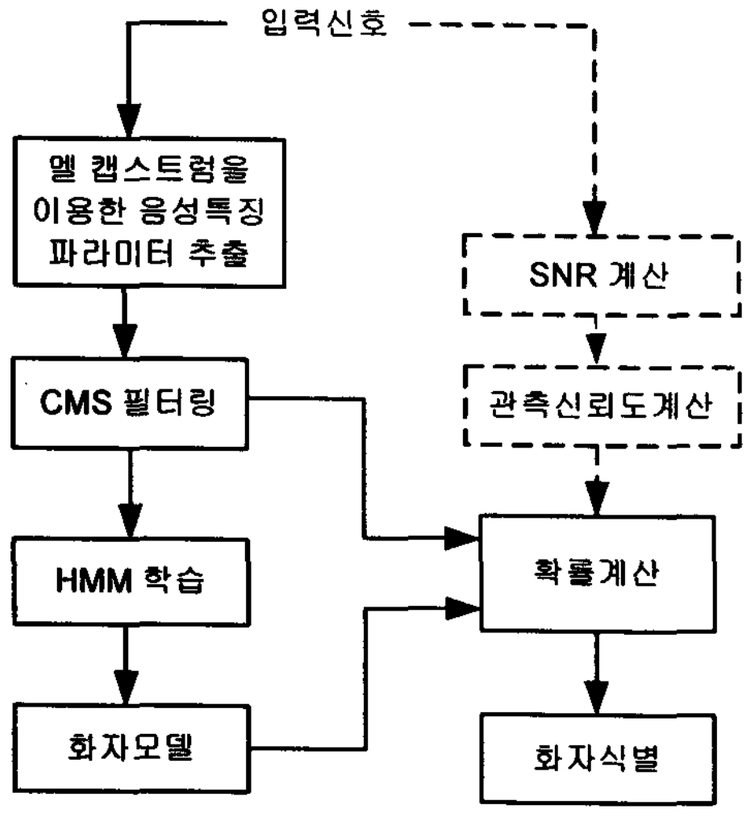

그림1. 관측신뢰도 기반 화자식별 과정

점선으로 표시된 부분은 입력음성으로부터 SNR을 측정한 후 SNR에 따른 관측 신뢰도를 계산하여 입력음성의 확률 계산 때 반영한다. 변형된 HMM 디코딩 알고리듬에 사용된 목적함수는 식 (9)와 같이 표현되는 시그모이드 함수로 잡음신호의 SNR과 관련된다.

$$\rho(SNR) = \frac{1}{1+e^{-a(SNR-b)}} \tag{9}$$

위 식에서 a는 스케일 파라미터이고, b는 이동(shift) 파라미터이다. $a=0.35$, $b=8.5$인 경우 멤버쉽 함수는 그림 2와 같다.

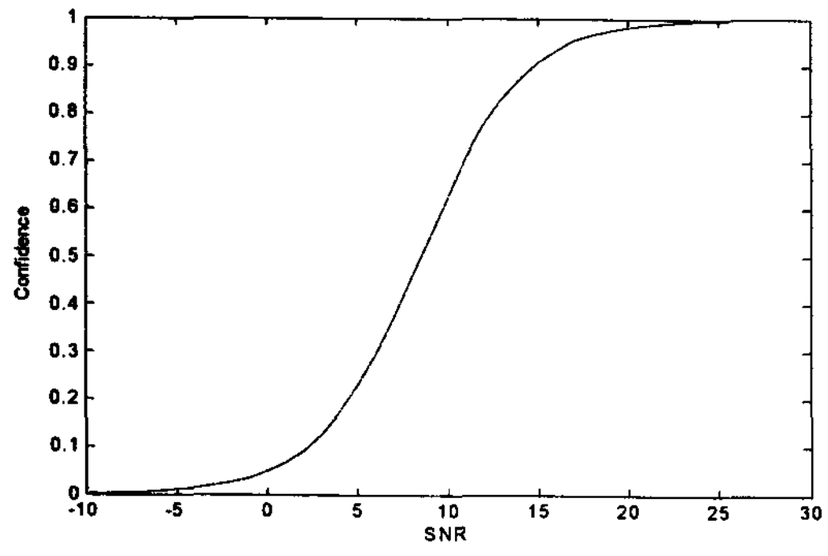

그림2. 관측 멤버쉽 함수의 예 (a=0.35, b=8.5)

잡음 음성은 가우시안 잡음을 더하여 얻는데 식 (10)은 깨끗한 음성으로부터 잡음 음성을 얻는 과정을 수식으로 표현한 것이다.

$$S_n(n) = s(n) + \alpha\eta(n) \tag{10}$$

위 식에서 $s(n)$은 잡음이 섞이지 않은 깨끗한 음성이며, $\eta(n)$은 파워가 1인 가우시안 불규칙 잡음이다. α는 잡음의 양을 결정하기 위한 변수이며, $S_n(n)$은 잡음에 오염된 신호를 의미한다.

4. 실험결과 및 고찰

4.1 데이터베이스

본 논문에서는 제안된 방법의 성능을 확인하기 위하여 ETRI에서 만든 한국어 화자인식용 휴대폰 음성 DB를 사용하여 문맥종속 화자식별 실험을 하였다. 음성데이터의 샘플링 주파수는 8kHz이며, 8 비트 μ-law PCM 방식으로 코딩되어 제공되었고 DB의 전체 화자의 수는 남녀 모두 49명이고, 화자당 음성파일은 모두 20개로 이중 10개씩을 학습용과 실험용으로 나누어 사용하였다. 문맥종속 인식실험에 사용한 음성데이터의 파형으로 발성시간이 약 3초 정도로 화자모델 학습에 사용된 음성데이터는 파일 10개를 합친 평균 약 30초 정도의 분량이다. 실험에서 입력 음성데이터의 한 프레임은 40ms로 하였고, 20ms씩 중첩되어 처리되도록 하고, 음성의 특징벡터는 12차 멜-켑스트럼(mel-cepstrum) 계수와 로그 에너지를 포함하였으며, 채널 왜곡을 보상하기 위하 3절에서 설명한 바와 같이 CMS 방법을 적용하였다.

EM 알고리듬에 의해 HMM 모델의 파라미터를 반복적으로 훈련하여 계산하였다. 이 과정에서 공분산 값은 full covariance를 사용하였으며, 알고리듬의 초기 과정에서는 fuzzy c-means clustering 방법을 사용하였다. 이를 정리하면 표 1과 같다.

표 1. 문장종속 화자식별 실험의 개요

음성 DB	ETRI 휴대폰 화자인식용 음성 DB
샘플링/음성코딩	8kHz/8 bits μ-law PCM
화자 수	49
화자당 학습 음성파일의 개수	10
화자당 테스트 음성파일의 개수	10
프레임길이/중첩	40 ms/20 ms
음성특징벡터	12차 멜 켑스트럼과 로그에너지
채널보상	Cepstral Mean Subtraction
HMM모델	EM 알고리듬, full covariance

4.2 실험결과 및 고찰

제안된 방법을 검증하기 위하여 ETRI 음성 DB를 사용하여 서로 다른 상태를 가진 한 개의 가우시안 분포를 사용한 HMM 모델을 이용하였다. 표 2는 실험결과를 보여주고 있는데, 기존의 방법(baseline)과 변형된 방법(modified)에 대하여 화자인식 실험을 수행한 결과이다. 표에 보인바와 같이 다양한 상태수와 SNR에 대하여 실험이 이루어졌다. 실험을 위하여, 관측 신뢰도 계산시 사용되는 파라미터는 스케일 파리미터 0,35 그리고 이동 파라미터는 8.5를 사용하였는데, 이는 경험적으로 결정된 값이다.

표 2. 기존 방법과 변형된 HMM 디코더 실험결과
(B: baseline, M: modified)

State / SNR	3		5		7	
	B	M	B	M	B	M
8	26.53	42.86	28.57	30.61	28.57	42.86
12	44.90	67.35	48.98	65.31	36.73	59.18
16	46.94	65.31	55.10	73.47	46.94	67.35
20	61.22	73.47	69.39	79.60	61.22	75.51
30	75.51	83.67	83.67	85.71	83.67	87.76
State / SNR	9		11		13	
	B	M	B	M	B	M
8	24.49	38.78	28.57	34.70	30.61	34.70
12	44.90	63.26	44.90	59.18	44.90	63.27
16	51.02	67.35	51.02	73.47	46.94	67.35
20	77.55	85.71	59.18	79.60	65.30	77.55
30	83.67	93.88	87.76	93.88	89.80	93.88

표 2에 의하면 본 논문에서 제안한 변형된 HMM 디코더 방법이 기존의 순방향 디코더에 비하여 우수함을 확인할 수 있다. 모든 상태값에서 대부분의 SNR에 대한 인식률이 10% 이상 대폭 향상되었음을 확인 할 수 있다. 이러한 실험의 결과는 제안한 방법이 잡음에 의한 인식률 저하 문제를 해결할 수 있는 한 방법임을 보여주고 있는 것이다.

5. 결론

본 논문에서는 관측 신뢰도기반 변형된 HMM 디코더를 제안하여 기존 순방향 디코더에 비하여 화자인식 성능을 향상시켰다. ETRI DB를 사용하여 문맥종속 화자식별 실험을 통해 검증하였는데, 잡음 하에서 성능이 크게 향상됨을 확인하였다.

본 논문에서 제안한 방법은 패턴인식, 음성인식, 얼굴인식과 같은 관측신뢰도와 관련한 어떤 분야에도 적용될 수 있다. 향후 논문 [1]과 같이 잡음이 학습시 데이터에도 존재하는 경우에 적용할 수 있는 학습 모델에 대하여 연구하고자 한다.

참 고 문 헌

[1] Jin Young Kim, et. al, "Modified GMM Training for Inexact Observation and Its Application to Speaker Identification," SPEECH SCIENCES Vol.14. No.1, pp.163~175, March 2007.

[2] 민소희, 김진영, 송민규, 나승유,"Particle Swarm 기반 최적화 멤버쉽 함수에 의한 잡음환경에서의 화자인식 성능향상" 음성과학회, Vol.14. No.2, pp.105~114, 2007.6.

[3] J P. Campbell, "Speaker Recognition: A Tutorial," Proceedings of the IEEE, vol. 85, pp. 1437~1462, Sept. 1997.

[4] D. A. Reynolds, "An Overviews of Automatic Speaker Recognition Technology', Proceedings of ICASSP '02 pp vol 4, pp. 4072~4075 May 2000

[5] Zhen Bin, Wu Xihong, Liu Zhimin, C. Huisheng. "An Enhanced RASTA processing for speaker identification." Proc of 2000 ICSLP, pp. 251~254, 2000.

[6] R.J. Mammone, X. Zhang. and R. P. Ramachandran, "Robust Speaker Recognition, A Feature-based Approach." IEEE Signal Processing Magazine, Vol.13, No.5, pp.58~71, 1996.

[7] D. Stephane and R. Christophe, "Robust Feature Extraction and Acoustic Modeling at Multitel : experiments on the Aurora databases." Proc of EuroSpeech -2003, pp.1789~1792 2003.

[8] A. Rosenberg et al., "Cepstral Channel Normalization Techniques for HMM-based Speaker Verification." Proc.ICSLP-94, pp.1835~1838, 1994.

[9] E. Mengusoglu, "Confidence Measure based Model Adaptation for Speaker Verification." Proc. of the 2nd IASTED International Conference on Communications, Internet and Information Technology, 2003.