실시간 문맥독립 화자인식 시스템의 성능향상을 위한 수정된 가중모델순위 결정방법

김민정*, 오세진**, 석수영*, 정호열*, 정현열*
*영남대학교 정보통신공학과
**대구과학대학 디지털 정보통신계열

Modified Weighting Model Rank Method for Improving the Performance of Real-Time Text-Independent Speaker Recognition System

Min-Joung Kim*, Se-Jin Oh, Su-Young Suk*, Ho-Youl Chung*, Hyun-Yeol Chung***
*Department of Information and Communication Eng., Yeungnam University
**Digital Information & Communication Div., Taegu Science College
{manjuk, lera}@orgio.net, osj@mail.taegu-c.ac.kr, {hoyoul, hychung}@yu.ac.kr

요 약

현재까지 개발된 화자식별 시스템 중 가중모델순위 (Weighting Model Rank; WMR)방법을 이용한 화자인식 시스템이 비교적 높은 인식성능을 나타내고 있다. WMR 방법은 각 화자에 대한 프레임 유사도의 순위에 따라 지수함수 가중치로 대치시키는 방법을 사용하고 있으나, 이 방법은 유사도 본래의 변별력이 전체 계산에서 고려되지 않는 문제가 있었다. 이를 해결하기 위해 본 논문에서는 각 화자의 프레임 유사도와 지수함수를 이용한 가중치를 곱한 값을 이용하여 전체 스코어를 계산하도록 하는 수정된 가중모델 순위방법(Modified Weighting Model Rank; MWMR)을 제안한다. 제안한 방법의 유효성을 확인하기 위하여 316명의 화자를 대상으로 하여 인식실험을 실시한 결과, 학습 프레임이 10,000일 경우, MWMR 방법에서 98.1%의 화자 인식률을 얻어 WMR 방법에 비해 약 2.0%의 향상된 인식결과를 보여 제안한 방법의 유효성을 확인할 수 있었다.

1. 서 론

화자식별이란 여러 명의 등록화자 중 발성화자를 식별하는 것을 말한다. 이러한 화자식별 기술은 개인의 음성 특징이 유일하다는 사실을 근거로 하고 있으며 최근의 인터넷 기술의 발전과 더불어 편리하고 완전한 보안을 위한 인증 방법으로 각광을 받고 있으며 특히, 문맥독립 화자식별의 경우 보안성이 높아 이에 관해 많은 연구가 이루어지고 있다[1]. 하지만, 화자식별 및 검증을 위한 화자인식은 고정도의 인식률을 요구하는 특성상 실제 사용 가능한 시스템구현에는 어려움이 많다. 이에, 본 논문에서는 현재까지 개발된 화자인식 시스템 중 화자식별 성능이 높다고 알려진 가중모델순위 (Weighting Model Rank; WMR)방법[2]을 구현하여 기존의 화자식별 시스템[3][4]에 적용하였다. WMR 방법은 식별화자를 결정하는 전체 스코어를 계산할 때, 프레임 단위로 계산된 유사도 대신에 유사도의 상대적 위치에 따라 결정된 지수함수 가중치를 사용함으로서, 프레임 단위에서 화자의 변별력을 높이는 방법이다. 이 방법은 최대 유사도 (Maximum Likelihood; ML) 방법보다는 프레임 단위에서 화자들간의 변별력을 키울 수 있는 장점이 있지만, 유사도 대신에 지수함수 가중치를 사용하기 때문에 유사도 본래의 변별력은 무시되는 문제점이 있다. 따라서, 이 문제점을 해결하고 더욱 향상된 화자식별 성능을 얻기 위해 본 논문에서는 전체 스코어를 계산할 때, 프레임 유사도에 지수함수 가중치를 곱하여 유사도 본래의 변별력을 더욱 크게 하는 수정된 가중모델 순위방법 (Modified Weighting Model Rank; MWMR)을 제안하고 제안된 방법의 유효성을 실험을 통하여 확인하고자 한다.

인식실험에서는 316명의 남성과 여성이 발성한 단어 음성을 대상으로 제안한 방법과 기존의 ML, WMR 방법에 의한 화자인식률을 비교 검토한다.

2. 기존의 화자식별 방법

2.1 최대 유사도 방법

일반적인 화자식별 방법은 Bayes의 정리[5]에 따라, 식 (1)에서 N명의 화자 중 사후확률 $P(\lambda_i|X)$, $1 \le i \le N$ 를 최대로 하는 모델 λ_{i^*}의 화자 i^*를 찾는 것이다.

$$P(\lambda_i|X) = \frac{p(X|\lambda_i)P(\lambda_i)}{p(X)} \qquad (1)$$

여기에서, 사전정보가 없기 때문에, 화자모델들은 동일하다고 가정한다. 즉, 사전확률 $P(\lambda_i)$는 식 (2)와 같다.

$$P(\lambda_i)=\frac{1}{N},\ 1\le i\le N \tag{2}$$

식 (1)의 분모인 $p(X)$는 발성 X의 빈도에 대한 무조건적인 우도를 나타내며, 모든 화자에 대해 동일한 값을 가진다. 따라서 식 (1)의 분자에서 $p(X|\lambda_i)$의 사후확률이 최대가 될 때의 식별화자는 식 (3)에 의해 결정된다.

$$i^*=\arg\max_i\ p(X|\lambda_i) \tag{3}$$

여기서, i^*는 식별된 화자를 나타낸다

2.2 프레임 단위 최대 유사도 방법

화자검증 시스템에 유사도 정규화 기법을 적용함으로서 시스템의 성능을 향상시킬 수 있다[6,7][8]. 화자검증을 위한 일반적인 방법은 검증을 요구하는 화자 모델 λ_c를 이용하여 발성 $X=x_1,x_2,\cdots,x_T$에 유사도 비 평가(Likelihood ratio test)를 실시하는 것이다[5]. 즉,

$$\ell(X)=\frac{p(\lambda_c|X)}{p(\lambda_{\bar{c}}|X)} \tag{4}$$

여기서, $\lambda_{\bar{c}}$는 모든 다른 가능한 화자의 모델을 나타내며, 사전확률 $P(\lambda_c)$와 $P(\lambda_{\bar{c}})$는 동일하다고 가정한다. 또한 Bayes 정리를 적용하여 로그를 취한 경우의 유사도 비는 식 (5)로 나타낼 수 있다.

$$\Lambda(X)=\log P(X|\lambda_c)-\log P(X|\lambda_{\bar{c}}) \tag{5}$$

유사도 $P(X|\lambda_c)$는 식 (6)에 의해 계산할 수 있다.

$$\log P(X|\lambda_c)=\frac{1}{T}\sum_{t=1}^{T}\log p(x_t|\lambda_c) \tag{6}$$

유사도 $P(X|\lambda_{\bar{c}})$는 백그라운드 화자들의 모델을 사용하여 계산되어지며, B개의 백그라운드 화자모델을 $\{\lambda_1,\cdots,\lambda_B\}$라고 하면, 백그라운드 화자들의 로그 유사도는 다음과 같이 계산할 수 있다.

$$\log P(X|\lambda_{\bar{c}})=\log\left\{\frac{1}{B}\sum_{b=1}^{B}P(X|\lambda_b)\right\} \tag{7}$$

백그라운드 모델에 의한 유사도 정규화는 발성문장의 변화에 따른 변화를 최소화할 수 있기 때문에 시스템의 성능을 향상시킬 수 있다[6].

기존의 일반적인 화자식별 시스템에서는 식별화자를 결정하는데 단일발성으로부터 유사도를 계산하기 때문에 정규화 과정이 필요하지 않지만[6], 본 시스템에서는 프레임 단위 유사도를 사용하기 때문에 정규화 과정이 필요하다. 이러한 유사도 정규화는 식 (8)을 이용하여 수행된다.

$$p_{norm}(x_t|\lambda_i)=\frac{p(x_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B}p(x_t|\lambda_b)} \tag{8}$$

모든 벡터 $x_t(t=1,2,\cdots,T)$에 대해 계산되어진 유사도의 전체 합계를 구하면 각 화자 모델 i에 대한 새로운 스코어가 계산되고, 인식화자는 식 (9)에 이용하여 가장 높은 스코어를 가진 화자로 결정된다.

$$Sc_i(X|\lambda_i)=\frac{1}{T}\sum_{t=1}^{T}\log p_{norm}(x_t|\lambda_i) \tag{9}$$

3. 수정된 가중모델순위 방법

3.1 가중모델순위 방법

WMR 방법은 인식화자를 결정하는 스코어를 계산할 때 테스트 음성과 화자모델들과의 프레임 유사도를 사용하지 않고, 각 프레임에서 계산된 유사도들의 상대적 위치에 따라 가중치를 결정하고, 이 가중치를 인식화자를 결정하는 스코어를 계산하는데 이용한다. 이렇게 함으로써, 프레임 단위에서 높은 유사도 값을 가지는 화자모델은 더 높은 값을, 낮은 유사도 값을 가지는 화자모델은 더 낮은 값을 가지게 되어 화자들간의 변별력을 더 높일 수가 있다.

WMR 방법을 적용하는 순서를 간략히 기술하면 다음과 같다. 첫 번째 단계는 먼저 식(8)을 이용하여, 각 테스트 벡터 $x_t(t=1,2,\cdots,T)$에 대한 프레임 유사도 $p(x_t|\lambda_i), i=1,\cdots,N$를 계산하고, 이를 내림순으로 정렬한다. 즉, 가장 큰 유사도를 가지는 화자모델은 최상위에 위치하게되고, 가장 낮은 유사도를 가지는 화자모델은 최하위에 위치하게 된다. 표 1은 각 프레임에서의 화자모델의 순위와 가중치의 관계를 나타낸 것이다.

표 1. 화자모델의 유사도 순위

순위 r	유사도	가중치 $w(r)$	화자모델
1	p_i^t	$w(1)$	Model λ_i (최대유사도)
2	p_j^t	$w(2)$	Model λ_j
...	...	...	...
m	p_k^t	$w(m)$	model λ_k
...	...	...	...
N	p_p^t	$w(N)$	Model λ_p (최소유사도)

그 다음, 화자모델의 각 순위에 따라 식 (10)에 나타낸 가중치 $w(r)$을 결정한다. 이때 가중치는 지수함수를 이용하면 화자들간의 변별력을 크게 할 수 있다[2].

$$w(r_\lambda)=\exp(A-Br_\lambda),\ r_\lambda=1,\cdots,N \tag{10}$$

여기서, A와 B는 $w(1)\approx N$이 되도록 설정한다. 그림 1은 가중치와 순위의 관계를 나타낸 것이다.

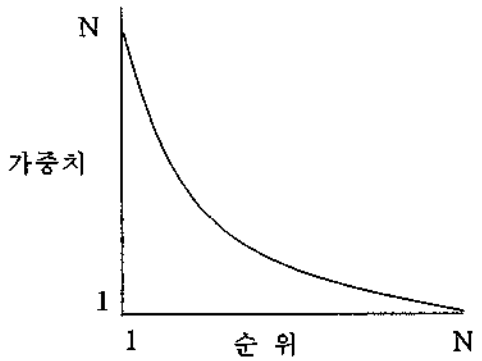

그림 1. 가중치와 유사도 순위

두 번째 단계에서는 각 모델 λ_i의 순위에 해당하는 가중치 $w_t(r_{\lambda_i})$를 유사도 $p(x_t|\lambda_i)$ 대신에 사용하여 전체 스코어 $Sc(X|\lambda_i)$를 계산한다. 전체 스코어 $Sc(X|\lambda_i)$는 $t=1,\cdots,T$에서 모든 가중치를 더하여 얻을 수 있다.

$$\log Sc(X|\lambda_i) = \sum_{t=1}^{T} w_t(r_{\lambda_i}) \qquad (11)$$

여기서, $w_t(r_{\lambda_i})$는 시간 t에서 순위가 r_{λ_i} 인 모델 i의 가중치이다.

3.2 수정된 가중모델순위 방법

위에서 설명한 WMR 방법은 인식화자를 결정하는 전체 스코어를 계산할 때 테스트 음성과 화자모델들과의 프레임 단위 유사도를 사용하지 않고, 계산된 프레임 단위 유사도 대신 이 유사도들의 상대적 위치에 따라 화자모델들을 내림차순으로 정렬한 후, 이 순위에 따라 가중치를 결정한 다음, 결정된 가중치를 인식화자를 결정하는 전체스코어를 계산할 때 사용하였다. 이 방법은 프레임 단위에서 화자모델들 사이의 변별력을 지수함수 가중치를 사용하여 유사도 본래의 변별력보다 크게 할 수는 있지만, 각 프레임에서 유사도값의 크기는 고려되지 않고 유사도의 순위에 따라서만 가중치가 결정되기 때문에 화자의 성도 특성을 잘 표현하지 못한 프레임의 경우에도 유사도의 순위만 높다면 큰 가중치값을 부여함으로서 화자의 변별력을 감소시킨다. 또한, 현재 프레임의 유사도 순위가 이전 프레임의 유사도 순위와 같은 상대적 위치를 차지한다면 이전 프레임과 동일한 가중치가 결정된다는 것이므로 전체 프레임에서 순위의 합계가 동일한 두 화자모델이 존재할 경우 동일한 가중치합계가 나온다는 단점이 있다. 따라서, WMR 방법에서 프레임 유사도의 크기까지 가중치를 결정하는데 고려할 수 있다면, 화자모델들 사이의 변별력을 좀 더 크게 할 수 있을 것으로 기대된다. 따라서 본 논문에서 제안한 수정된 가중모델순위(MWMR) 방법에서는 프레임 단위 유사도 값의 상대적 위치에 따라 결정된 가중치와 프레임단위 유사도 값을 곱한 값을 사용함으로써 프레임 단위 유사도의 변별력에 지수함수 가중치를 주어 화자들 사이의 변별력을 더 크게 할 수 있는 수정된 가중모델순위 방법(Modified Weighting Model Rank, MWMR)을 제안한다. 그림 2는 제안된 MWMR 방법을 나타낸 것으로, 기존의 최대 유사도(ML) 방법에 가중치를 부여하는 과정을 추가한 형태가 된다.

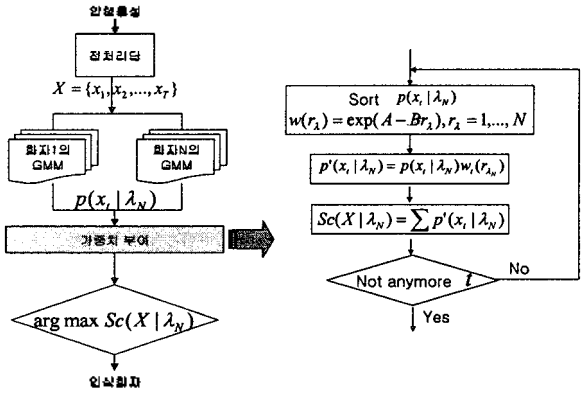

그림 2. 수정된 가중모델순위 방법

4. 인식 실험

4.1 음성 분석 및 데이터

제안한 방법의 화자인식에 대한 유효성을 확인하기 위하여 인식실험을 수행하였다. 인식실험에서GMM (Gaussian Mixture Model)의 혼합수는 인식률 및 계산량을 고려하여 16으로 고정하였으며, 특징 파라미터는 캡스트럼 계수와 회귀계수 값만을 사용하였다. 음성 특징 파라미터의 분석조건을 표 2에 나타내었다.

실험을 위한 음성 데이터베이스는 10대에서 50대까지의 남녀 화자가 혼합된 것으로 채집기간과 발성단어가 다른 음성으로 구성되어 있으며, 모델 학습과 평가를 위한 음성 데이터는 무작위로 추출하여 사용하였다. 평가용 화자는 남성 및 여성이 혼합된 102명과 316명의 화자를 대상으로 실험하였다.

표 2. 전처리 분석조건

Sampling Rate	16 kHz
Pre-emphasis coefficient	0.98
Hamming Windows	yes
Frame length	256 points
Frame Shift	120 points
Cepstrum vector dimension	10

4.2. 실험 결과

우선, 102명과 316명의 평가화자에 대해 본 논문에서 제시한 ML, WMR, MWMR 방법에 대해 화자 인식실험을 수행하였다. 이때 화자모델 학습을 위한 음성 데이터의 프레임(frame)은 학습을 위한 최소단어라고 생각되는 4,000 프레임과 학습을 위한 발성시간을 고려한 10,000 프레임을 선택하였고, 평가를 위한 프레임은 350 프레임으로 설정하였다. 각각의 화자인식 방법에 대해 인식실험을 수행한 결과를 표 3과 그림 2, 3에 각각 나타내었다. 표 3과 그림 2, 3에서 본 논문에서 제시한 각각의 화자인식 방법에 대해 4,000 프레임보다는 10,000 프레임을 선택한 경우가 효과적임을 확인할 수 있었다. 전체적으로 ML 방법보다는 WMR 방법이 인식성능을 향상시키는데 보다 효과적임을 알 수 있다. 특히, 본 논문에서 제

안한 MWMR 방법의 경우, 316명의 화자를 대상으로 화자 인식실험을 수행한 결과, 10,000 프레임에서 평균 98.1%의 인식률을 얻어, 다른 방법보다 상대적으로 높은 인식성능을 보여 제안방법의 유효성을 확인할 수 있었다.

표 3. 평균 화자인식률(%)의 비교

화자수	학습 프레임	ML	WMR	MWMR
102명	f4k	85.29	94.12	95.1
	f10k	97.06	99.02	100
316명	f4k	87.66	89.97	90.82
	f10k	93.35	97.78	98.1

f4K : 4,000 프레임, f10k :10,000 프레임

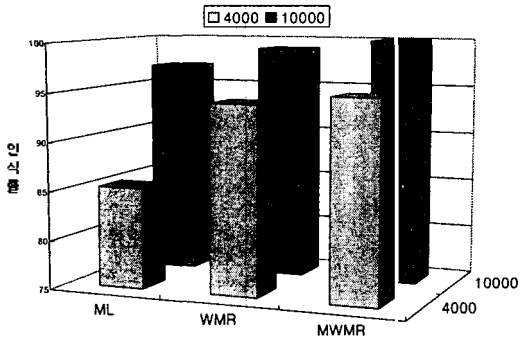

그림 3. 학습프레임 4,000과 10,000일 때 102명의 평균 화자인식률

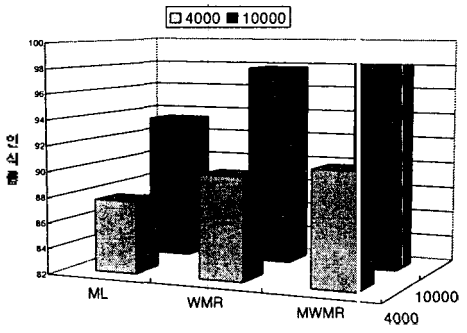

그림 4. 학습 프레임 4,000과 10,000일 때 316명의 평균 화자인식률

두 번째는 학습 데이터의 프레임을 변화시키면서 모델을 학습한 경우, ML, WMR, MWMR 방법에 대해 화자 인식실험을 수행하였다. 표 4에 학습 프레임의 변화에 따른 각 인식방법의 평균 화자인식률을 나타내었다. 표 4에서 알 수 있듯이, 학습 프레임의 변화에 관계없이 본 논문에서 제안한 MWMR 방법이 가장 높은 인식률을 얻어 제안방법의 유효성을 확인할 수 있었다.

표 4. 학습 프레임의 변화에 따른 각 인식방법의 평균 화자인식률(%)

화자수	인식방법	4k	5k	6k	7k	8k	9k	10k
102	ML	85.29	89.21	89.21	94.12	96.08	93.14	97.06
	WMR	94.12	96.08	96.08	97.06	98.04	98.04	99.02
	MWMR	95.1	96.08	96.08	97.06	98.04	98.04	100
316	ML	87.66	88.61	89.24	90.82	91.77	90.82	93.35
	WMR	89.87	92.72	93.35	93.35	95.57	95.87	97.78
	MWMR	90.18	93.99	93.99	94.3	96.52	96.2	98.1

5. 결 론

본 논문에서는 기존의 화자식별 시스템의 성능향상을 위하여 WMR 방법을 구현하였으며, 강건한 화자식별을 위하여 MWMR 방법을 제안하였다. WMR 방법은 각 화자에 대한 프레임 유사도의 순위에 따라 지수함수 가중치로 대치시키는 방법을 사용하고 있으나, 이 방법은 유사도 본래의 변별력이 전체 계산에서 고려되지 않는 문제가 있었다. 이를 해결하기 위해 본 논문에서는 각 화자의 프레임 유사도와 지수함수를 이용한 가중치를 곱한 값을 이용하여 전체 스코어를 계산하도록 하는 수정된 가중모델 순위방법(Modified Weighting Model Rank; MWMR)을 제안한다. 제안방법의 유효성을 확인하기 위해 남녀 화자에 대해 화자인식 실험을 수행한 결과, 기존의 WMR 방법의 경우, 기존의 ML 방법보다 약 4.0% 향상된 인식결과를 나타내었으며, MWMR 방법의 경우, WMR 방법보다 약 2.0%의 향상된 인식결과를 보여 제안한 MWMR 방법의 유효성을 확인하였다. 또한, 316명의 화자를 대상으로 하여 학습 프레임이 10,000일 경우, MWMR 방법에서 98.1%의 화자 인식률을 얻어 제안한 방법의 실용화의 가능성을 확인할 수 있었다.

참 고 문 헌

1. 정현열, "음성을 이용한 화자인식 기술의 현황과 전망," 정보과학회지 제19권 제7호, pp. 32-44, 2001.
2. K. Markov and S. Nakagawa, "Text-independent speaker identification on TIMIT database," Proc. of A.S.J., pp. 83-84, 1995. 3.
3. 김민정, 석수영, 정현열, "Gaussian Mixture Model을 이용한 실시간 문맥독립 화자인식에 대한 고찰," 한국음향학회 하계학술발표대회 논문집, 제20권, pp. 123-126, 2001. 7.
4. 김민정, 석수영, 김광수, 정현열, "프레임 레벨 유사도 정규화를 적용한 문맥독립 화자식별 시스템의 구현," 신호처리시스템학회지, 제3권, 1호, pp. 8-14, 2002.
5. K. Fukunaga, "*Introduction to statistical pattern recognition*," Academic Press Inc., 1990.
6. D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, Vol. 17, No.1-2, pp. 91-108, 1995.
7. A. Rosenberg, J. DeLong, C. Lee, B. Juang and F. Soong, "The use of cohort normalized scores for speaker verification," Proc. of ICSLP'92, pp. 599-602, 1992.
8. T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," Speech Communication, Vol. 17, pp. 109-116, 1995. 8.