가변 문턱치와 순차결정법을 통한 문맥요구형 화자확인

Text-Prompt Speaker Verification using Variable Threshold and Sequential Decision

안 성 주*, 강 선 미**, 고 한 석*
Sungjoo Ahn, Sunmee Kang, Hanseok Ko

ABSTRACT

This paper concerns an effective text-prompted speaker verification method to increase the performance of speaker verification. While various speaker verification methods have already been developed, their effectiveness has not yet been formally proven in terms of achieving an acceptable performance level. It is also noted that the traditional methods were focused primarily on single, prompted utterance for verification. This paper, instead, proposes sequential decision method using variable threshold focused at handling two utterances for text-prompted speaker verification. Experimental results show that the proposed speaker verification method outperforms that of the speaker verification scheme without using the sequential decision by a factor of up to 3 times. From these results, we show that the proposed method is highly effective and achieves a reliable performance suitable for practical applications.

Keywords: speaker verification, user verification, speaker recognition, biometrics, sequential decision, text-prompt speaker verification

I. 서 론

종전의 유선전화에서 무선 및 개인 휴대폰의 등장과 인터넷의 다양한 서비스가 폭발적으로 보급되면서, 현대를 살고 있는 소비자의 욕구에 적절한 다양한 서비스들이 개발되어 정보화 사회로의 삶을 이끌어가고 있다. 특히 인터넷과 전화망의 만남에서 비롯된 신종 서비스들, 즉 전자상거래, 은행의 현금 지급 서비스, 전화 쇼핑 서비스 등은 이미 널리 이용되고 있으며, 본인 또는 허가를 얻은 사람들만이 접근해서 서비스를 이용하도록 되어있다. 이러한 서비스들은 통상 사용자에게 시간과 공간적인 제약을 뛰어넘는 편리함을 주고 있지만, 얼굴과 얼굴을 맞댄 거래가 아니기 때문에 그 만큼 위험 부담을 안고 있다. 그러므로 많은 사람들은 본인의 정보가 타인에게 노출될 것을 염려해서 많은 이점이 있음에도

* 고려대학교 전자공학과
** 서경대학교 컴퓨터과학과

불구하고 서비스 사용을 망설이고 있다. 따라서 개인의 신분 확인을 위한 보안이 필수적으로 요구된다. 이와 같은 문제점을 해결하기 위해 인간 개개인의 고유한 생체 특성 중에서 인간의 음성을 이용하여 신원을 확인하는 것이 화자확인이다. 이 방법은 화자의 음성 신호에서 개인의 고유한 음성 특징들을 추출하여 이미 학습해 놓은 자신의 모델과 비교해 봄으로써 화자임을 확인하게 된다[1][2][3].

화자확인 방법에는 각 개인의 정해진 단어나 문장을 이용하는 문맥의존형(Text-Dependent) 화자확인과 각 개인이 어떠한 문장을 발음해도 되는 문맥독립형(Text-Independent) 화자확인이 있다. 그러나 문맥의존형 화자확인이 성능은 우수하지만, 비밀단어를 기억해야만 하고 사칭자가 녹음을 해서 사용할 수 있다는 단점이 있다. 그리고 문맥독립형 화자확인은 사용자가 임의의 문장을 발음하므로 편리한 점이 있는 반면에 사용자가 많은 데이터를 입력해야만 그 성능을 유지할 수 있다는 단점이 있다. 따라서 이 두 가지 방법을 보완한 것이 문맥요구형(Text-Prompt) 화자확인이다. 문맥요구형 화자확인은 시스템이 사용자에게 어떠한 단어나 문장을 발음할 것을 요구하기 때문에 비밀단어의 기억이나 녹음 등의 위험은 줄어들고, 성능면에서도 문맥독립형 화자확인 방법보다 우수하다.

기존의 문맥요구형 화자확인 방법들은 사용자에게 음성을 한 번 발음하게 하여 화자확인을 수행한다[4][5]. 그러나 본 논문에서는 단 한번의 발음을 이용하여 사용자를 확인(수락/거절)하는 것이 아니라 사용자에게 두 단어를 발음하게 하고, 각각의 단어에 대해 서로 다른 문턱치(threshold)를 정하고 이때의 우도(likelihood)값을 조합하여 화자확인을 수행한다. 즉, 본 논문은 가변 문턱치(variable threshold)와 순차결정법(sequential decision)을 통한 문맥요구형 화자확인을 제안한다.

본 논문의 구성은 2장에서는 일반적인 화자확인 시스템에 대해 소개하고, 3장에서 새로운 화자확인 방법을 제안한다. 그리고 4장에서 실험결과에 대해 설명하고 마지막으로 5장에서 결론을 맺는다.

II. 일반적인 화자확인 시스템

일반적인 화자확인 시스템의 기본 구조는 그림 1과 같다. 테스트할 화자의 음성은 먼저 끝점 검출후 특징 파라미터의 추출을 통해 분석된다. 추출된 특징 파라미터 값은 기존에 저장되어 있는 의뢰인의 화자 모델과 비교한다. 이렇게 해서 얻은 우도값을 계산하여 미리 정해진 문턱치와 비교하여 그 의뢰인을 수락할 것인지 거부할 것인지를 판단한다.

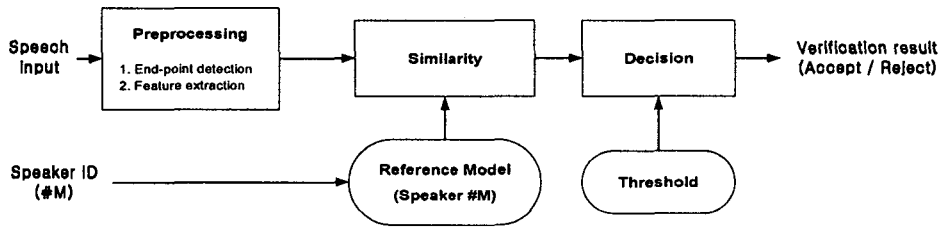

그림 1. 화자 확인 시스템의 블록 다이어그램

화자를 확인할 때 사용하는 방법은 hypothesis 테스트이다. 즉, 요구한 화자가 맞는지 아닌지를 결정하기 위해 입력 음성에 대해 우도비(likelihood ratio)를 적용하는 것이다. 입력음성이 $X=\{x_1,..,x_T\}$이고 모델 λ_c에 해당하는 화자라고 주장했을 경우 우도비는 식(1)과 같이 표현된다.

$$\frac{\Pr(X \text{ is from } \textit{the claimed speaker})}{\Pr(X \text{ is not from } \textit{the claimed speaker})} = \frac{\Pr(\lambda_c \mid X)}{\Pr(\lambda_r \mid X)} \qquad (1)$$

이 우도값을 Bayes's 방법과 균일한 초기(prior) 확률값을 적용하여 로그 도메인에서 보면 식(2)와 같이 표현된다.

$$\Lambda(X) = \log p(X|\lambda_c) - \log p(X|\lambda_r) \qquad (2)$$

이 우도비 $\Lambda(X)$를 문턱치 θ와 비교하여 만약 $\Lambda(X) > \theta$이면 요구한 화자를 수락하고 $\Lambda(X) < \theta$이면 거절한다.

화자확인 시스템에는 두 종류의 에러가 발생한다. 첫 번째는 의뢰인을 거부하는 오인거부율(False Rejection Rate, FRR)이고, 두 번째는 사칭자를 수락하는 오인수락률(False Acception Rate, FAR)이다. 일반적으로 문턱치가 커질수록 FRR은 증가하고 FAR은 감소한다. 반대로 문턱치가 작아지면 FAR은 증가하고 FRR은 감소한다. 그러므로 화자확인 시스템에서는 성능평가의 척도로 FAR과 FRR이 같아지는 동일오류율(Equal Error Rate, EER)값을 사용한다. 즉, EER은 시스템이 얼마나 잘 모델링되었냐를 나타내고, 여러 화자확인 시스템을 비교할 경우, EER이 낮다는 것은 시스템이 더 잘 모델링되었다는 것을 의미한다.

III. 제안한 방법

본 논문은 가변 문턱치와 순차결정법을 통한 문맥요구형 화자확인을 제안한다.

기존의 화자확인 방법들은 2장에서 설명한 것처럼 사용자에게 음성을 한번 발음하게 하여 이 음성을 이용하여 화자확인을 수행한다. 즉, 식 (2)와 문턱치 θ와의 비교를 통해 화자확인을 수행한다. 그러나 본 논문에서는 화자확인을 할 때 사용자에게 먼저 임의의 첫 번째 단어를 발음하게 하여 이때의 우도값을 문턱치와 비교한 결과값을 이용하여 다음 단어의 문턱치를 정하고 사용자에게 두 번째 단어를 발음할 것을 요구하여 새로 정해진 문턱치와 비교를 하여 화자확인을 수행하는 방법을 제안한다.

본 논문에서 제안하는 방법의 자세한 과정은 다음과 같다.

먼저 사용자에게 임의의 첫 번째 단어(X_1)를 발음하게 한다. 그리고 이때의 우도값($\Lambda(X_1)$)을 첫 번째 문턱치(θ_1)와 비교를 하여 두 번째 문턱치(θ_2)를 결정한다. 그 과정은 다음 식과 같다.

$$\begin{aligned} &\text{IF } \Lambda(X_1) > \theta_1+\alpha && \text{Then } \theta_2 = \theta_1 - \beta \\ &\text{IF } \theta_1-\alpha \le \Lambda(X_1) \le \theta_1+\alpha && \text{Then } \theta_2 = \theta_1 \\ &\text{IF } \Lambda(X_1) < \theta_1-\alpha && \text{Then } \theta_2 = \theta_1+\beta \end{aligned} \quad (3)$$

여기서 α와 β는 임의의 상수로서 실험을 통해서 얻을 수 있다.

이렇게 두 번째 문턱치가 정해지면 사용자에게 두 번째 단어(X_2)를 발음하게 한다. 그리고 이때의 우도값($\Lambda(X_2)$)과 두 번째 문턱치(θ_2)와의 비교를 통해 사용자를 수락할 것인지 거절할 것인지에 관한 화자확인을 수행한다. 즉, (4)와 같다.

$$\begin{aligned} &\text{IF } \Lambda(X_2) \ge \theta_2 \text{ Then Accept} \\ &\text{IF } \Lambda(X_2) < \theta_2 \text{ Then Reject} \end{aligned} \quad (4)$$

본 논문에서는 위에서 설명한 것처럼 단 한 번의 발음을 이용하여 사용자를 확인하는 것이 아니라 사용자에게 두 단어를 발음하게 하고, 각각의 단어에 대해 서로 다른 문턱치를 정하고 이때의 우도값을 이용하여 화자확인을 수행한다.

IV. 실험 결과

4.1 실험 환경

본 논문에서는 문맥 요구형 화자확인 시스템을 고려한 한국어 단어 음성 데이터베이스를 사용하여 성능평가 실험을 수행하였다. 사용된 음성 데이터베이스는 남자 22명이 11단어를 5번씩 발음한 한국어 단어로 구성되었다. 음성은 마이크를 이용하여 컴퓨터로 녹음되었고, 11.025 kHz로 샘플링되었다. 음성 특징으로는 12차원의 Mel Frequency Cepstral Coefficient (MFCC)와 그 1차 미분, 그리고 로그 파워 및 그 1차 미분값, 총 26차의 특징벡터를 사용했다.

그리고 화자모델과 world 모델은 5개의 상태(state)를 가지는 Left-to-Right 연속 HMM (Hidden Markov Model)으로 각 상태는 여러 개의 가우시안(Gaussian) mixture로 이루어졌다. 또한 각 화자의 단어 모델 훈련은 수집된 각 단어의 처음 3개를 사용했다.

4.2 기초 실험

이 실험은 한 번의 발음을 이용하는 일반적인 화자확인 방법으로 실험한 것으로 본 논문에서 제안하는 실험에 대한 성능 비교를 위한 기준 실험으로 수행되었다. 각 화자의 단어 모델을 위해 실험에 사용된 학습 데이터의 개수는 3개이다. 학습은 Segmental K-means 알고리즘을 사용했으며 결정(decision)을 위해 world 모델로 우도값을 정규화(normalization)하였다(식 (2)). 또한 화자모델과 world 모델의 각 상태에 대한 mixture의 수를 1, 2, 4, 6개로 증가시키면서 실험을 하였다.

실험 결과는 표 1과 같다. 실험 결과로부터 mixture의 개수를 2개에서 4개, 6개로 증가시킴에 따라 EER이 증가하는 것을 볼 수 있다. 이 사실은 3개의 적은 학습 데이터를 이용해서 화자의 모델 파라미터를 추정할 때 mixture의 개수가 많아지면, 모델 파라미터들을 정확하게 추정할 수 없기 때문이다. 그러나 mixture의 개수가 1개일 경우는 mixture의 개수가 2개일 때보다 EER이 크다. 그 이유는 1개의 mixture로 모델들을 정확히 잘 표현할 수 없기 때문이다.

표 1. Mixture의 수에 따른 화자확인 EER(%)

Mixture의 수	1	2	4	6
EER(%)	3.37	2.51	6.08	6.60

4.3 제안한 방법에 의한 화자확인 실험

이 실험은 3장에서 설명한 것처럼 본 논문에서 제안한 두 단어를 이용한 가변 문턱치와 순차결정법을 화자확인에 적용했을 경우의 성능 향상을 알아보기 위해서 수행되었다. 화자의 모델은 4.2에서 사용한 것과 같고, mixture의 수는 앞의 실험에서 가장 좋은 성능을 보인 2개를 사용했다. 또한 실험에 사용한 첫 번째 문턱치 θ_1은 4.2절에서 EER을 구할 때 사용한 값을 이용했다. 가능한 모든 두 단어의 조합을 이용했으며, 식 (3)에서의 α와 β의 값을 변화시키면서 실험을 수행하였다. 실험 결과는 표 2와 같다.

표 2. 제안한 방법을 이용한 화자 확인률(%)

alpha	0.3			**0.6**			1.0			2.0		
beta	0.5	1.0	2.0	0.5	1.0	**2.0**	0.5	1.0	2.0	0.5	1.0	2.0
FRR(%)	0.76	0.00	0.00	0.76	0.00	**0.00**	0.76	0.00	0.00	0.76	0.00	0.00
FAR(%)	1.89	1.66	1.75	1.89	1.57	**1.53**	1.98	1.71	1.64	2.24	2.11	1.97
HTER(%)	1.33	0.83	0.87	1.33	0.78	**0.77**	1.37	0.86	0.82	1.50	1.06	0.98

여기서 HTER(Half Total Error Rate)는 FRR과 FAR의 평균값이다. 본 논문에서 제안한 2단어를 이용한 문맥요구형 화자확인을 수행하기 때문에, EER 값은 구할 수가 없고 대신 HTER 값을 이용하여 비교를 하였다.

표 2의 결과를 표 1의 결과와 비교해 보면 전체적으로 본 논문에서 제안하는 방법이 2~3배 이상 성능이 향상된 것을 볼 수 있다. 표 2에서 보면 $\alpha = 0.6$, $\beta = 2.0$일 때 가장 좋은 성능을 보인다. 또한 위의 결과로부터 특정한 α값에서 β의 값이 1.0 이상일 경우 FRR이 0이 되는 것을 볼 수 있는데, 이는 첫 번째 단어에 의해 두 번째 문턱치의 값이 변하고, 따라서 두 번째 단어에 의해 다시 한 번 화자확인이 이루어지기 때문에 그 성능이 향상되었다는 것을 알 수 있다. 이러한 결과로부터 본 논문이 제안하는 가변 문턱치와 순차결정 방법이 화자확인 시스템에 매우 효과적이라는 것을 알 수 있다.

V. 결 론

본 논문은 가변 문턱치와 순차결정법을 통한 문맥요구형 화자확인을 제안했다. α와 β의 값들을 다양하게 변화시키면서 두 단어를 이용하여 화자확인을 수행했으며, 실험 결과 제안한 방법이 한 개의 단어만을 사용하는 경우보다 3배 정도의 성능향상을 보였다. 또한 FAR과 FRR 모두 한 개만의 단어를 사용할 때보다 그 성능이 매우 향상되었다. 이러한 결과로부터 본 논문이 제안하는 화자확인 방법이 시스템의 성능 향상에 매우 효과적이라는 것을 알 수 있다.

앞으로 좀더 성능이 향상된 화자확인을 위해 최적의(optimal) 문턱치를 선택하는 연구가 필요하다. 또한 현재의 단어 단위의 화자모델을 음소나 반음소 단위로 바꿔서 본 논문에서 제안한 문맥요구형 화자확인에 대한 연구를 계속할 계획이다.

감사의 글

본 연구는 한국과학재단 목적기초연구 (1999-2-30200-016-5) 지원으로 수행되었음.

참 고 문 헌

[1] Lee. C. H., Soong. F. K. and Pliwal. K. K., *Automatic Speech and Speaker Recognition Advanced Topics,* Kluwer Academic Publishers, Second Printing, pp. 31-56, 1997.

[2] S. Furui, "An Overview of Speaker Recognition Technology," ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.

[3] Q. Li, B. H. Juang, C. H. Lee, Q. Zhou, and F.K. Soong, Recent Advancements in Automatic Speaker Authentication, IEEE Robotics & Automation Magazine, pp. 24-34, March 1999.

[4] W. D. Zhang, M.W.Mark and M.X.He, "A Two-Stage Scoring method combining world and cohort models for speaker verification," Proc. ICASSP2000, Vol. 2, pp. 1193-1196, 2000.

[5] J.B. Pierrot, J. Lindberg, J. Koolwaaij, H.P.Hutter, D.Genoud, M. Blomberg, F. Bimbot, "A comparison of A priori threshold setting procedures for speaker verification in the CAVE project, Proc. ICASSP98, pp. 125-128, 1998.

접수일자: 2000. 10. 26.
게재결정: 2000. 11. 21.

▲ 안성주
서울시 성북구 안암동 5가 1번지 (우: 136-701)
고려대학교 전자공학과
Tel: +82-2-927-6115 (O), H/P: 019-375-6709
Fax: +82-2-3291-2450
E-mail: sjahn@ispl.korea.ac.kr

▲ 강선미
서울시 성북구 정릉동 16-1 (우: 136-704)
서경대학교 컴퓨터과학과
Tel: +82-2-940-7291 (O), H/P: 011-9760-7144
Fax: +82-2-919-0345
E-mail: smkang@skuniv.ac.kr

▲ 고한석
서울시 성북구 안암동 5가 1번지 (우: 136-701)
고려대학교 전자공학과
Tel: +82-2-3290-3239 (O), H/P: 011-9001-3239
Fax: +82-2-3291-2450
E-mail: hsko@korea.ac.kr