다양한 음성을 이용한 자동화자식별 시스템 성능 확인에 관한 연구

홍수기(국립과학수사연구소)

<차 례>

1. 서론
2. 음성자료 및 실험방법
2.1. 음성자료
2.2. 실험방법
3. 실험결과 및 고찰
4. 결론

<Abstract>

Variation of the Verification Error Rate of Automatic Speaker Recognition System with Voice Conditions

Soo Ki Hong

High reliability of automatic speaker recognition regardless of voice conditions is necessary for forensic application. Audio recordings in real cases are not consistent in voice conditions, such as duration, time interval of recording, given text or conversational speech, transmission channel, etc. In this study the variation of verification error rate of ASR system with the voice conditions was investigated. As a result in order to decrease both false rejection rate and false acception rate, the various voices should be used for training and the duration of train voices should be longer than the test voices.

* 주제어: ASR system, FAR, FRR

1. 서 론

음성을 이용하여 개인을 식별하는 기술은 통신수단을 이용하여 신원을 확인하는 경우나 보안 시스템에서 매우 유용하게 활용될 뿐만 아니라 유괴 사건, 공공건

물 폭파 협박 사건, 독극물 투입 협박 사건 등과 같은 범죄 사건에서 범인을 확인하기 위한 필수적인 기술이다. 음성에 의한 개인식별(이하 "화자식별"이라 함) 방법으로는 음성학자들에 의한 청각적인 방법, 성문(Voice Spectrogram)을 이용하는 시각적인 방법 및 음성의 특징 추출 및 식별판단이 컴퓨터에서 이루어지는 기계적인 방법으로 구분된다. 현재 법과학 분야에서 활용되고 있는 청각적인 방법과 시각적인 방법은 비교 대상자가 많은 경우 시간이 많이 소요되어 긴급한 사건인 경우 효율적으로 대처하기 곤란하고, 식별결과가 전문가의 숙련도에 의존되어 주관이 개입되지 않을 수 없는 단점이 있다. 이런 화자식별의 단점을 보완 할 수 있는 화자식별 방법이 기계적인 방법, 즉 자동화자식별(Automatic Speaker Recognition; 이하 ASR이라 함.) 방법이다. 이 방법에 관한 연구(Rosenberg, 1975; Bunge, 1977; Furui, 1980; Ney, 1981)는 1970년대부터 세계적으로 활발히 이루어져 왔으며 전송선로 영향을 보전하는 방법이나 강인 기법들(Naik, 1994; Reynolds, 1995)이 제시되어 왔다. 법과학 분야에서 비교 대상 음성들은 다양한 전송선로를 통하게 되고, 표준화되지 않은 녹음 시스템을 이용하여 녹음된다. 뿐만 아니라 화자들은 대부분이 비협조적이다. 범죄 사건에서 활용 가능한 ASR 시스템은 음성길이나 녹음조건에 관계없이 최소한 95% 이상의 신뢰도가 요구되어 아직까지 상용화되어 있는 시스템이 구축되어 있지 않다.

본 연구의 목적은 음성의 길이 및 녹음 환경에 따른 ASR 시스템에 의한 화자확인율 변화를 확인하여 이 시스템에 의한 화자식별의 신뢰도를 증가시킬 수 있는 방법을 제시하는 것이다. 본 연구에서 이용하는 ASR 시스템은 인식방법으로 문장종속뿐만 아니라 문장독립 경우에도 높은 화자인식률을 보이는 것으로 보고된 [Reynolds, 1995] GMM(Gaussian Mixture Model)을 이용하며, 음성 길이 변화에 따라 mixture 개수를 조절할 수 있고 음성 비교에 사용되는 특징들로는 MFCC, △-MFCC, 및 Power Spectrum을 선택적으로 이용할 수 있도록 개발된 시스템이다. 또한 인식 문턱치도 조절이 가능하도록 되어 있다. 이 시스템을 이용한 과거의 연구(박우식, 2000; 홍수기, 2001)에서 가장 효율적인 음성의 길이는 30-40초간의 음성이고, 전화음성인 경우 확인 문턱치를 약 15% 낮춘 경우 다른 사람을 동일인으로 확인하는 에러율(False Acception Rate; FAR)을 5%이하로 유지하며 인식률은 증가시킬 수 있다고 보고하였다. 또한 훈련 음성이 다양하고 음성 길이가 테스트 음성보다 긴 경우 높은 인식률을 얻을 수 있다는 것이 확인되었다. 그러나 실제 범죄 사건에서는 대부분의 경우 다양한 음성 자료를 얻는 것이 용이하지 않고 실제 비교 가능 음성길이가 30초보다 훨씬 짧다. 그러므로 본 연구에서는 좀 더 짧고 다양한 길이(약 2초에서부터 약 20초까지)의 음성들을 이용하여 음성 길이 변화에 따른 화자확인율을 조사하고, 녹음 시간 간격, 주어진 문장을 읽는 경우와 자연스런 대화를 하는 경우에 따른 화자확인율을 비교한다. 또한 높은 확인율을 나타내는 음성들과 에러율이 가장 높은 음성들의 스펙트로그램을 비교 분석하여 화자확

인 에러에 가장 영향을 미치는 원인을 파악하고자 한다.

2. 음성자료 및 실험방법

2.1. 음성자료

첫 번째 그룹인 남성 화자 10명이 발음한 음성들은 전화기를 통한 한국어 표준 발음 진단용 이야기[이현복, 1984] 중 일부 (191개 음절)를 읽은 음성과 실험자와 자연스런 대화를 한 음성이다. 이 첫 번째 그룹은 하루에 두 번 씩 같은 음성을 발음하였고 일주일에 한 번씩 3주간 녹음하였다. 두 번째 그룹인 7명의 남자 화자들이 발음한 음성은 전화기를 통한 일반적인 협박 내용을 읽은 음성과 실험자와 자연스런 대화를 한 음성이다. 두 번째 그룹 중 두 명의 화자는 첫 번째 그룹에도 속해져 있었다. 두 번째 그룹의 음성 녹음 간격은 일정하지 않으며 녹음 시간 간격은 최고 7년이었다. 전화 음성을 녹음하기 위해서 카플러를 이용하였다. 첫 번째 그룹의 음성들은 디지털 녹음기 [SONY TCD-D10]로 녹음하였고, 두 번째 그룹의 음성들은 일반 카세트 녹음기 [SONY TC-D5M]로 녹음하였다.

2.2. 실험방법

ASR 시스템에서 화자확인에 이용하기 위해서 테이프에 녹음된 음성들은 CSL (KAY, 4300B)를 통하여 wave 파일로 만들었고, Sampling rate는 11 kHz로 하였다.

훈련음성들은 음성 길이 및 녹음조건에 따라 4가지(이하 각각 Train 1, Train 2, Train 3, 및 Train 4 라고 함)로 구분하고, 테스트 음성은 두 가지(이하 각각 Test 1, 및 Test 2 라고 함)로 구분하여 각각 화자확인을 행하였다. 훈련음성들은 10-14명의 음성에서 편집한 음성들로 구성되었다. 두 가지의 테스트 음성들은 각각 200개와 45개로 15명의 화자들 음성에서 편집한 음성들이다.

Train 1 음성들은 첫 번째 그룹 10명의 화자들의 음성으로 둘째 날과 셋째 날에 녹음된 음성 중 주어진 문장을 읽은 음성에서 2-3 초간의 음성들로 편집한 5개 문장으로 10명의 화자 모두 같은 문장들을 발음한 것이다. Train 2 음성들은 첫 번째 그룹 중 8명 및 두 번째 그룹 중 3명의 음성으로 첫 번째 그룹 음성들은 Train 1에서 이용한 둘 째날 녹음된 3개의 문장과 세 번째 날 녹음된 음성 중 자연스런 대화 중에서 3-6초간으로 편집한 음성으로 구성되어 있고, 두 번 째 그룹 음성들은 같은 날 협박내용을 읽는 음성 중에서 3-4초간으로 편집한 음성과 자연스런 대화 음성 중 4-5초간 음성으로 편집한 음성들로 구성되었다. Train 3과 Train 4 음성들은 두 번째 그룹에서 한 사람을 제외한 14명의 화자 음성으로 구성되었다. Train

3 음성은 같은 날 녹음된 자연스런 대화 음성 중 3개 문장으로 구성되었고 전체 음성 길이는 약 10초였다. Train 4 음성은 같은 날 녹음된 음성 중에서 전체 음성 길이는 약 20초이고 주어진 문장을 읽는 음성이나 자연스런 대화 음성에서 한 개 나 두 개로 편집하였다. Test 1 음성들은 5초 이하의 음성들이고 Test 2의 음성들은 약 10초간의 음성들로 구성되었다.

ASR 시스템의 환경 조건 중 mixture 개수 및 특징파라미터들(MFCC, △-MFCC, 및 Power Spectrum)을 변화시키면서 Train 1, 2, 3, 및 4 각각의 음성들을 Test 1, 및 2 각각의 음성들과 화자확인 실험을 행하여 화자들 및 각각의 훈련 음성들에 따른 확인율과 에러율을 구하여 주어진 문장을 읽는 경우와 자연스런 대화간, 녹음된 시간간격, 및 음성의 길이에 따른 확인율 변화를 분석한다.

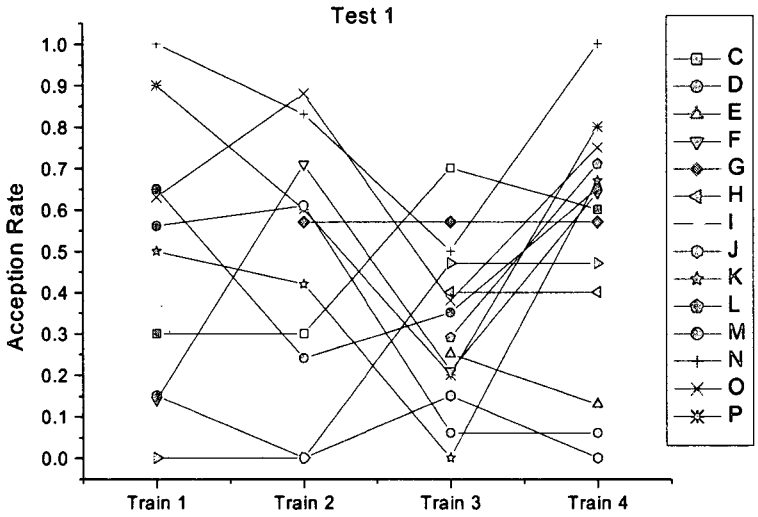
3. 실험결과 및 고찰

<표 1>은 mixture 개수를 8로 설정하고 특징 파라미터를 MFCC 만을 이용한 경우 각 Train 음성들에 따른 전체적인 확인율과 에러율을 보여주고 있고, 그림 1, 2, 3, 및 4는 화자들 및 Train 음성 각각에 따른 화자확인 결과이다.

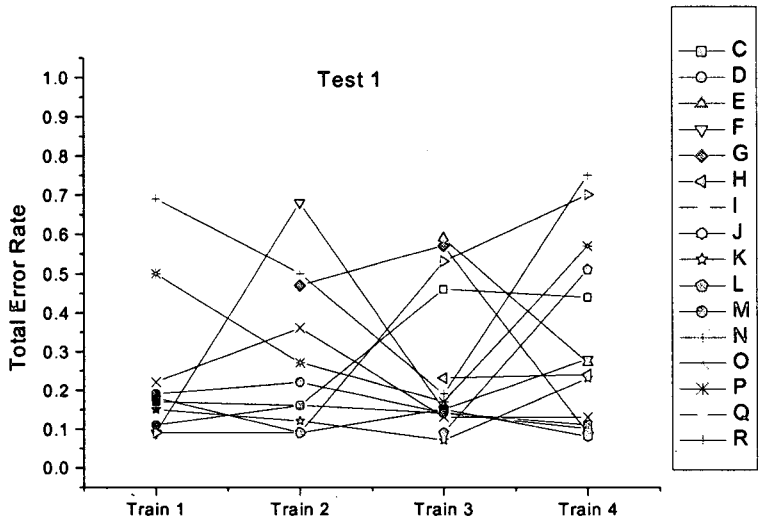
<표 1> Train 음성들에 따른 전체적인 확인율 및 에러율.

	Test	Train 1	Train 2	Train 3	Train 4
Acception	Test 1	0.47	0.49	0.36	0.63
	Test 2	0.28	0.39	0.23	0.35
False Rejection	Test 1	0.53	0.51	0.64.	0.37
	Test 2	0.72	0.61	0.77	0.65
False Acception	Test 1	0.21	0.26	0.23	0.31
	Test 2	0.15	0.17	0.14	0.24
Total Error	Test 1	0.24	0.28	0.26	0.32
	Test 2	0.18	0.19	0.18	0.26

(a)

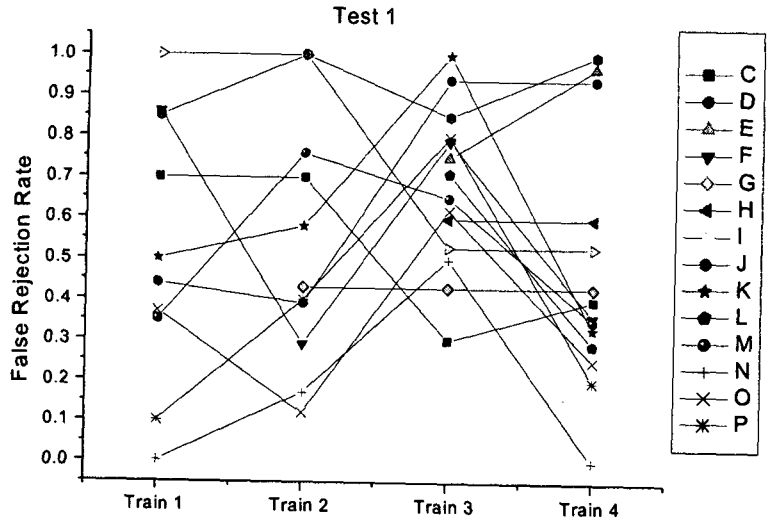


(b)

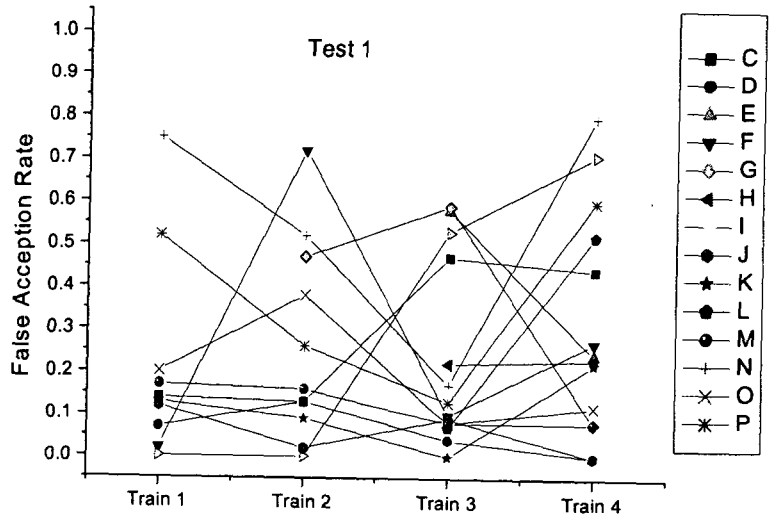


<그림 1> Test 1 음성들의 화자 및 Train 음성들에 따른 확인율 (a) 및 전체 에러율 (b)

(a)

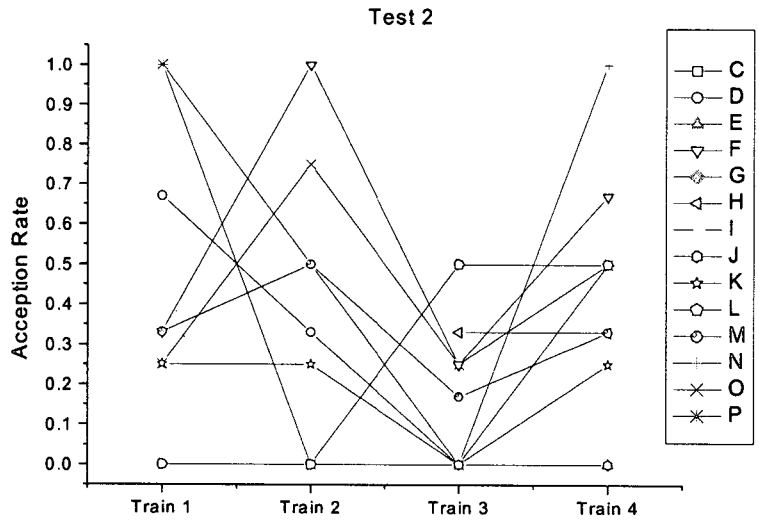


(b)

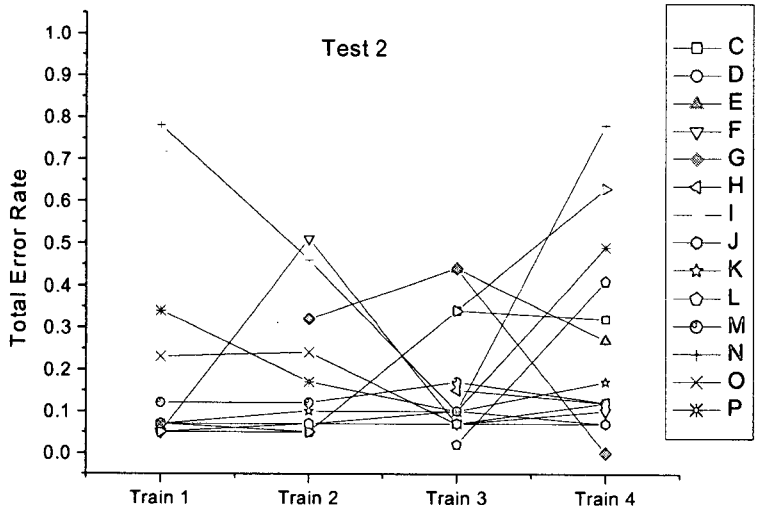


<그림 2> Test 1 음성들의 화자 및 Train 음성들에 따른 FRR (a) 및 FAR

(a)

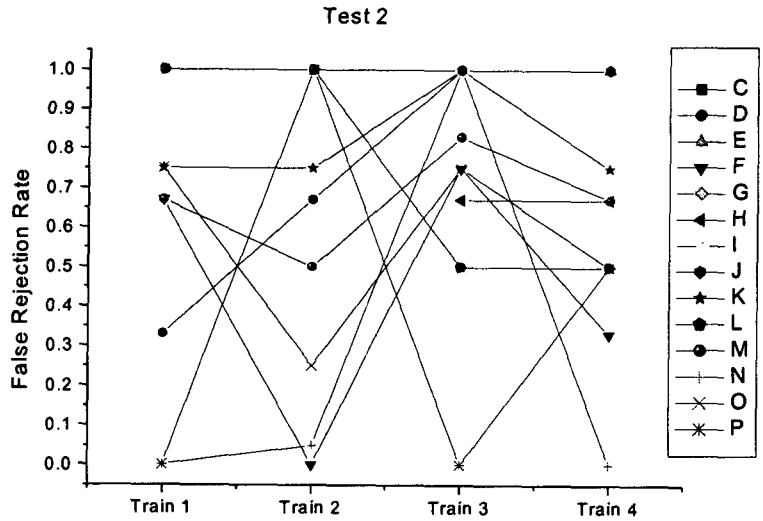


(b)

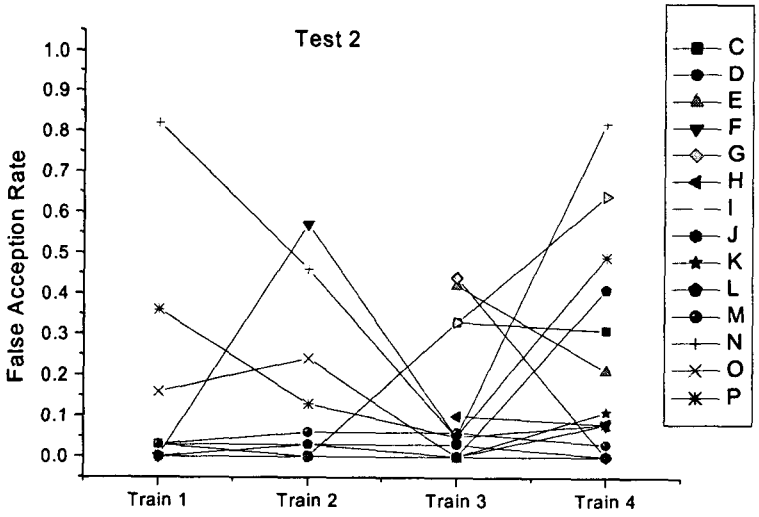


<그림 3> Test 2 음성들의 화자 및 Train 음성들에 따른 확인율 (a) 및 전체 에러율 (b)

(a)



(b)



<그림 4> Test 2 음성들의 화자 및 Train 음성들에 따른 FRR (a) 및 FAR (b)

ASR 시스템에서의 화자확인 결과는 동일한 사람을 정확히 확인하는 확인율이 증가하면 동일한 사람을 다른 사람으로 확인하는 에러율(False Rejection Rate; FRR)은 낮아지게 되고, 다른 사람을 동일한 사람으로 확인하는 에러율(False Acception Rate; FAR)은 증가하였다. <표 1>에 나타난 훈련음성들에 따른 확인율 및 에러율 변화를 보면 네 종류의 훈련음성 모두에서 전체적인 훈련음성과 길이가 유사한 Test 2 음성들이 훈련음성보다 짧은 Test 1 음성들에 비해서 FRR은 커지고 FAR 및 전체 에러율은 낮아지는 경향이 있었고 확인율은 에러율보다 더 많이 낮아졌다. 전체 음성 길이는 길지만 동일한 날 녹음된 음성들로 훈련 음성 수가 가장 적은 Train 4 음성들인 경우 대부분 FRR은 낮아지나 확인율, FAR, 및 전체적인 에러율이 커지는 것으로 나타났다. Train 1 및 2와 전체 음성 길이는 유사하나 동일한 날 녹음된 음성들로 구성되어 있는 Train 3 음성들이 확인율이 가장 낮았다. 주어진 문장을 읽는 음성들로만 구성되어 있는 Train 1 음성들과 자연스런 음성들로만 구성되어 있는 Train 3 음성들을 주어진 문장을 읽는 음성과 자연스런 대화 음성으로 구성되어 있는 Train 2 및 4 음성들과 비교해 보면 Train 1 및 3 음성들이 Train 2 및 4 음성들에 비하여 FRR은 커지고 FAR 및 전체 에러율은 낮아졌으며 확인율은 에러율 보다 더 많이 낮아지는 경향이 있었다.

<그림 1, 2, 3 및 4>에 나타난 화자별 확인결과를 보면 화자에 따라 매우 다른 결과를 보여준다. 예로 화자 J인 경우 전체적인 에러율은 매우 낮은 편이나 Test 1 음성들에서는 본인의 음성으로 확인되는 확인율이 매우 낮았고, Test 2 음성에서는 모든 훈련 음성에서 본인의 음성이 전혀 확인되지 않았다. 반면에 화자 N은 Train 3인 경우를 제외하고는 전체적인 에러율이 매우 높고, Test 2에서 Train 2 및 3인 경우를 제외하고는 본인의 음성이 매우 잘 확인됨을 보여주고 있다. Train 4 음성들은 대부분의 화자들이 Test 1에서 높은 확인율을 나타내 주고 있고, Train 1과 3 음성들은 대부분의 화자들이 Test 1 및 2 모두에서 낮은 에러율을 나타내주고 있다. 또한 동일한 화자인 경우 훈련음성에 따라 매우 다른 결과를 보여주기도 하였다. 예로 화자 K는 Test 1에서 Train 4인 경우 높은 확인율과 낮은 에러율을 나타내 주었으나 Train 3인 경우는 낮은 확인율을 나타내주고 있으며, 화자 F인 경우는 Test 2에서 동일한 결과를 보여주고 있다.

4. 결 론

ASR 시스템에 의한 화자확인시 확인율이 증가하게되면 동일한 사람을 다른 사람으로 확인하는 에러율 [FRR]은 감소하게 되므로 화자확인 신뢰도를 향상시키기 위해서는 다른 사람이 동일한 사람으로 확인되는 에러율 [FAR]과 FRR을 동시에 감소시켜야 한다. 실험 결과에 의하면 훈련 음성을 테스트 음성보다 길게 하고 다

양한 음성을 훈련 음성으로 이용해야만 FRR 및 FAR을 동시에 감소시킬 수 있다.

화자들에 따라 확인율 및 에러율 변화가 다양하므로 본인의 음성을 전혀 확인하지 못하는 경우 및 FAR이 매우 큰 화자들에 대한 훈련 음성들의 스펙트로그램을 분석하고 전송선로 특징 및 음성 특징들을 비교하여 에러율을 증가시키는 원인을 파악하여 좀 더 효율적인 음성자료들을 ASR 시스템에서 이용함으로써 신뢰도를 향상시킬 수 있을 것이다.

참 고 문 헌

박우식, 설부경, 홍수기(2002), 자동화자식별시스템의 입력음성에 따른 인식률 비교, 「국과수연보」 32, pp.394～402.

이현복(1984), 「한국어의 표준발음」, 대한음성학회, pp.38～39.

홍수기, 박명철(2001), 실제 사건에서 수집된 음성 데이터에 의한 자동화자식별 시스템의 평가 및 개선 방법에 관한 연구, 「국과수연보」 33, pp.432～440.

Bunge, E., U. Hofker, and H. D. Hohne (1977), The AUROS Project-Automatic Recognition of Speakers by Computer, *FREQUENZ* 31, pp.345～351.

Furui, S., and A. E. Rosenberg (1980), Experimental studies in a new automatic speaker verification system using telephone speech, *Proc. ICASSP* 5, pp.1060～1062.

Ney, H., R. Gierloff and R. Frehse (1981), An automatic system for verification of cooperative speakers via telephone, *Carnahan Conference on Crime Countermeasures,* pp.97～101.

Naik, J. M., and D. M. Lubensky (1994), A hybrid HMM-MLP speaker verification algorithm for telephone speech, *Proc. ICASS* 94.1, pp.153～156.

Reynolds, D. A. (1995), Speaker identification and verification using Gaussian mixture speaker models, *Speech Communication* 17, pp.91～108.

Reynolds, D. A., and R. C. Rose (1995), Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Trans. Speech and Audio Processing* 13, pp.72～83.

Rosenberg, A. E., and M. R. Sambur (1975), New techniques for automatic speaker verification, *IEEE Trans. Acoustics, Speech and Signal Processing ASSP*-23, pp.169～176.

접수일자: 2002년 4월 30일
게재결정: 2002년 5월 24일

▷ 홍수기(Soo Ki Hong)
주소: 경기도 고양시 일산구 장항동 877 호수마을 313-101
소속: 국립과학수사연구소
전화: 02) 2600-4980
E-mail: sooki@nisi.go.kr