음성조작에 의한 인식기의 성능평가

* 정 성 윤* 정 현 열* 김 경 태**
* 영남대학교 전자공학과 ** 한남대학교 정보통신공학과

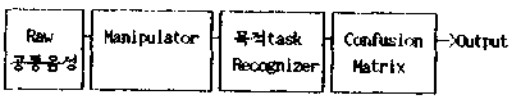
Performance Assessment of Recognizer by means of Speech Manipulation

* Sung-Yun Jung*, Hyun-Yeol Chung*, Kyung-Tae Kim**
* Dept. of Electronic Eng. Yeungnam University,
** Dept. of Information & Communication Eng. Hannam University

I. 서론

음성조작에 의한 인식기를 평가하기위해서는 인식기의 성능에 많은 영향을 미치는 요인들을 모아서 평가항목으로 구성한 뒤, 항목별로 음성데이타를 조작해서 이들에 대한 인식결과를 구하고, 인식결과와 조건들간의 상관관계를 분석해야한다. 그리고 평가항목 뿐만이 아니라 평가할 인식기의 범위도 정해야하는데, 상용화된 인식기와 컴퓨터에 인식을 시뮬레이션하는 경우를 나누어서 생각할 수 있다.

상용인식기의 경우 사용목적에 따라 입력되는 음성의 종류가 다르기 때문에 목적 타스크에 무관하게 강제적으로 공통 음성으로 평가하는 경우와 목적 타스크의 음성에 맞추어서 평가하는 방법을 생각할 수 있다. 아래 그림은 상용인식기를 평가하는 전체 블럭도이다. 목적 타스크에 무관하게 평가하는 경우는 숫자나 단모음 등의 가장 보편화된 음성을 공통음성으로 선정하고 이러한 음성을 평가항목에 따라 적절히 조작하는 조작기가 있으면 평가가 가능하다. 그러나 목적타스크의 음성에 맞추는 경우는 인식기를 개발할 때 사용한 음성을 이용해서 조작하는 정도를 동일하게 해 주면 평가가 가능하다. 그리고 인식결과를 상세히 분석하기위해서는 Confusion Matrix형태로 결과를 나타내어야 하기 때문에 마지막 출력단 전에 Confusion Matrix로 결과를 나타낼 수 있는 처리과정이 추가 되어야한다. 물론 상용인식기의 경우에는 인식성능뿐만이 아니라 처리속도, 사용의 편리함등등을 평가항목에 추가시켜야할 것이다.

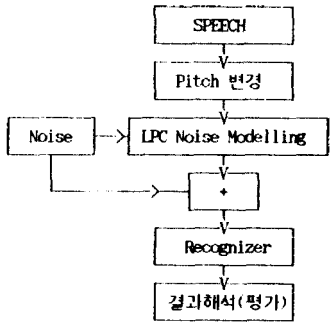


그러나 국내의 경우 상용인식기는 거의 전무한 상태이다. 음성인식의 대부분의 연구가 컴퓨터에 각자가 개발한 인식알고리즘을 시뮬레이션해서 인식결과를 확인하는 정도이다. 따라서 본 연구에서도 이러한 인식기에 국한해서 성능평가를 행한다. 이러한 경우의 인식기도 상용인식기의 경우와 마찬가지로 생각할 수 있다. 아직까지 공통의 음성데이타베이스가 구축되어 있지 않기때문에 평가할 인식기의 목적타스크의 음성을 조작해서 인식결과를 분석한다.

본 연구에서는 잡음의 종류(백색잡음, 저주파잡음, 고주파잡음), 신호대 잡음비에 따른 인식기의 성능을 분석한다. 기존의 잡음 첨가는 음성에 백색잡음을 더하는 방법이었다. 이러한 방법은 실제의 잡음환경을 모사하기에는 부족한 점이 많다. 따라서 잡음환경에 대한 좀 더 현실성 있는 음성을 시뮬레이션하기 위해 룸바드 영향을 고려하였다. 본 연구에서는 룸바드 영향을 고려하기위해 룸바드 음성을 녹음해서 그 특징을 분석한 후, 룸바드 음성을 시뮬레이션 하였다. 룸바드 음성의 시뮬레이션은 먼저, 룸바드 음성을 분석한 결과를 기초로해서 피치를 변화시키고, LPC NOISE MODELLING 방법을 사용하여 음성의 스펙트럼에너지를 변화시켰다. 이렇게 조작된 룸바드 음성은 인식기의 입력 테스트 음성으로 사용되었다. 그리고 룸바드 영향을 고려하지 않은 경우와 고려한 경우에 대한 인식실험의 결과를 비교분석하였고, 실험에 사용한 인식기의 환경요인(잡음의 종류, SNR변동)에 대한 성능을 해석하였다.

II. 평가 방법

Lea의 논문에 의하면, 인식시스템의 인식 정확도에 영향을 주는 원인은 한마디로 변동성(variability)이라는 것이다. 이것은 7가지 요인과 밀접한 관계가 있고, 이 중에서 특히 인식률에 영향을 많이 미치는 요인은 인간적 요인, 언어요인, 환경요인등의 세가지이다. 본 연구에서는 환경요인에 대한 인식기의 성능을 평가하는 것이 주 목적이다. 따라서 세가지 종류의 잡음(백색잡음, 저주파 및 고주파 잡음)과 다양한 신호 대 잡음비(35dB-5dB)에 따른 인식률을 조사하여 인식기의 성능지수를 신호대 잡음비로 표현한다. 평가실험을 위한 전체 블럭도는 그림 1에 나타내었다.



III. 환경요인에 대한 음성의 조작

음성에 잡음이 더해지는 잡음환경에 대한 인식기 평가를 위해서는, 먼저 음성데이타를 잡음환경에 가깝도록 조작해야한다. 가장 많이 사용하는 방법은 깨끗한 음성에 잡음을 더하므로서 잡음이 있는 음성신호를 만들어내는 것이다. 그러나 Noisy speech을 좀 더 현실성 있게 시뮬레이션하기 위해서는 롬바드 영향을 고려해야한다. 왜냐하면, Fig.1처럼 배경잡음이 존재하는 상황에서, 화자는 자신의 발성을 변경하기 때문이다. 따라서 단순히 잡음을 더한 음성과는 차이가 있게된다.

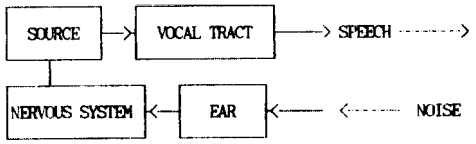


Fig.1 Noise Interaction on the Vocal System

III.1. Lombard Effect 성질

전술한대로, 주위에 잡음이 존재할 때에 화자가 자신의 발성을 변경해서 발성하는 것을 롬바드영향이라고 했다. 배경잡음의 강도에 따라 발성이 어느정도 변하는지에 대한 정량적인 연구는 행해지지 않았으나, 몇몇 연구가들에 의해 롬바드 영향의 정성적인 연구가 이루어졌다. 이들의 연구결과를 보면, Lombard Effect는 음성인식기의 성능저하에 중요한 영향을 미치는 것으로 알려져 있다. 이러한 연구들의 결과들을 종합하면 다음과 같다.

첫째, 잡음환경에서, 화자들은 Vocal effort를 증가시키고 좀 더 강력한 음성신호를 산출한다. 둘째, 잡음에서 발성한 음성의 길이(duration)는 증가하는 경향이 있다. 세째, 평균기본주파수(F0)는 중대한 증가가 있다. 네째, 잡음이 있을 때에는, 최상위포만트들(upper formants)이 더 강렬(intense)하고, 스펙트럴기울기(spectreal slope)가 향상되어진다. 다섯째, Lombard Effect의 가장 큰 결과는 잡음만을 더한 음성과 비교할 때 명료도(intelligibility)의 향상이다.

이러한 연구결과를 바탕으로 롬바드 영향을 고려한 음성을 시뮬레이션하기위해서, 실제의 롬바드 영향을 입은 음성을 녹음하여 시간영역,주파수 영역에서 통계적인 분석을 행한 후, 분석결과를 바탕으로 롬바드 음성을 시뮬레이션한다.

III.2 롬바드 음성의 녹음 및 분석

III.2.1 롬바드음성의 녹음

잡음을 포함하고 있는 주파수 대역에 따라 화자가 음성을 어떻게 발성하는지를 고찰하기 위해 백색잡음을 발생시켜서 무잡음, 50, 60, 70, 80, 90 dB등의 6가지 잡음레벨 대해 녹음을 행한다. 발성내용은 10개의 숫자음과 7개의 단모음으로 구성한다. 발성자는 표준어를 사용하는 남성화자 2명과 경상도 방언을 쓰는 남성화자 1명을 선정해서 3회 발성을 시켰다. 전체적인 녹음절차는 Fig. 2에 나타나 있다.

녹음된 롬바드 음성중, 5모음(ㅏ, ㅔ, ㅣ,ㅗ,ㅜ)에 대해 디지탈 신호처리 기술을 사용하여 Fig. 3의 절차에 따라 분석을 행한다. 디지탈화 된 음성은 수작업 처리를 거쳐 파일 이름과 기록 환경, 그리고 음성파형의 정보 및 레이블링 정보를 저장한다. Fig. 4는 수작업처리의 한 예를 나타낸 것이다.

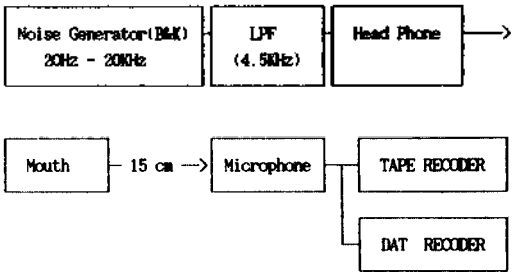
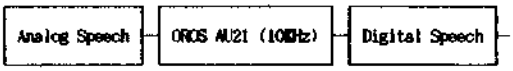


Fig. 2 다양한 잡음레벨(SPL)에 따른 롬바드 음성의 녹음절차



——> 수작업 처리 ——> 파일명,음성파형,레이블링 정보 저장

Fig. 3 롬바드 음성의 디지탈 분석 절차

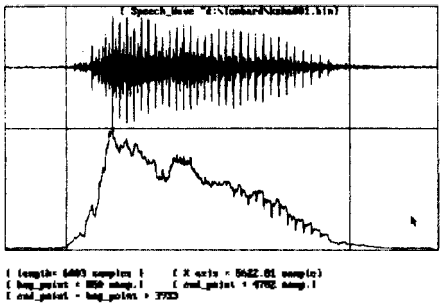


Fig. 4 수작업 처리과정의 한 예 (Mouse로 Click)
/ ㅏ/에 대한 음성파형과 에너지

III.2.2 롬바드음성의 특징분석

III.2.2.1. 피치의 변동량 해석

피치분석에는 고정밀도의 캡스트럼 방법이나 LPC모델의 예측잔차 방법등이 있으나 본 연구에서는 단모음을 분석대상으로 하기 때문에 정밀도는 약간 떨어지지만 계산이 쉬운 자기상관함수(autocorrelation) 방법을 사용한다.

본 연구에서는 한 프레임을 32msec로하고 16msec씩 천이해서 분석을 행하였다. Fig. 5은 한 프레임에 대해 화자 jdy의 /우/에 대한 음성파형과 자기상관 수열을 나타낸것이다. 자기상관관계가 가장 높은 부분이 이 음성의 피치주기가 된다.

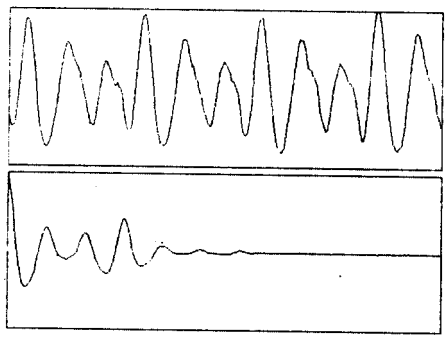


Fig. 5 /우/에 대한 음성파형과 자기상관수열

이렇게 분석한 피치주기는 Fig 6과 같다. Fig 6에서 잡음의 레벨이 증가함에 따라 피치도 크게 증가함을 알 수 있다.

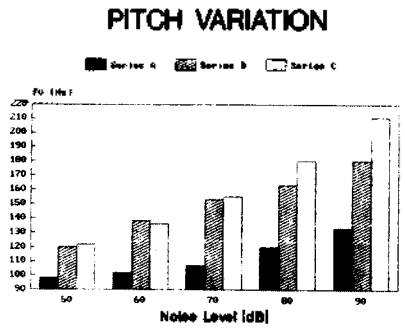


Fig. 6 세 화자의 기본주파수 특성

III.2.2.3. 평균에너지와 지속시간의 변동량 해석

지속시간의 분석은 Fig.4에서 처럼 시간 시간로 음성의 끝점 검출을 해서 구하고, 평균에너지는 끝점 검출된 음성에 대해 대수로 그을 취해서 구했다. Fig. 7은 세 화자에 대한 지속시간의 특성이다. 잡음레벨이 증가함에 따라 지속시간이 항상 증가하지는 않는다. 이것은 녹음대상 음성이 단모음에 국한 되었기 때문이다. 즉, 단어나 문장을 발성할 때는 화자의 발성이 안정되게 계속 지속될 수 있으나, 단모음일 경우에는 지속시간이 비교적 짧아서 잡음레벨에 따른 지속시간의 변동을 정확하게 측정하는 것은 어렵다. 그러나 그림에서 볼 수 있듯이 전반적으로 잡음레벨이 증가함에 따라 지속시간도 증가함을 알 수있다.

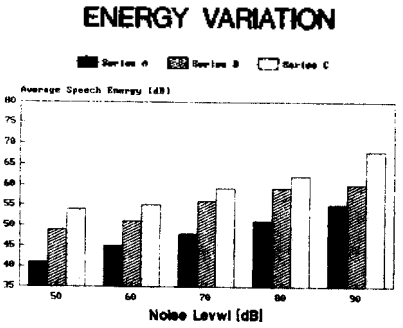


Fig. 7 세 화자의 평균 지속시간 분포

Fig. 8는 세 화자에 대한 평균에너지의 분포이다. 그림에서 알 수 있듯이 잡음레벨이 증가함에 따라 에너지가 선형적으로 증가함을 알 수 있다. 세 화자 모두 비슷한 기울기로 증가하기 때문에 선형함수로 근사화할 수 있다.

III.4 LPC Noise Modelling

이 방법은 Fig.9 에 구현한대로, 잡음에 대한 선형필터 계수를 구해서 음성신호와 필터링을 하므로서 잡음의 영향을 받은 음성신호를 만들어내는 원리이다. 이렇게 만든 신호는 다시 잡음과 더해지므로서 배경잡음에 의해 변경된 음성신호를 시뮬레이션할 수 있다.

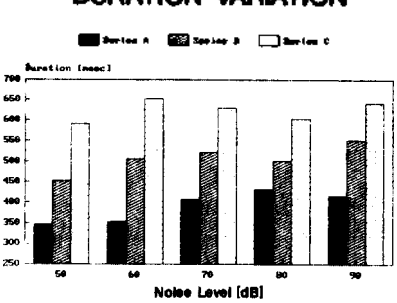


Fig. 8 세 화자의 평균 에너지 분포

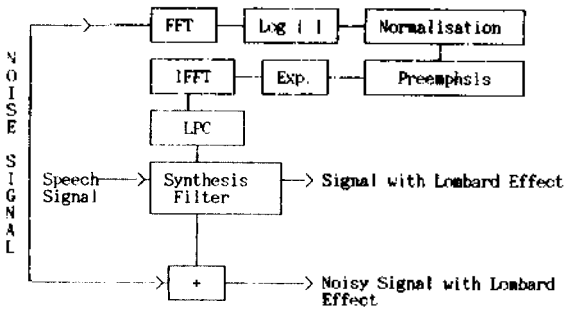


Fig.9 Lombard Simulation by Noise Modelling

앞에서 제시한 LPC Noise Modelling을 사용해서 롬바드 영향을 고려한 음성을 시뮬레이션해서, 원 음성과의 특성을 SNR이 35, 25, 15, 5dB일 때를 Fig. 10에 나타내었다. 그림에서 약간 큰 파형을 나타내고 있는 것이 롬바드 영향을 시뮬레이션한 음성이다. LPC Noise Modelling 실험의 결과, 백색잡음을 포함하고 있는 긴 주파수 대역에서 SNR이 적을수록 신호 에너지가 증가한 사실을 알 수 있다.

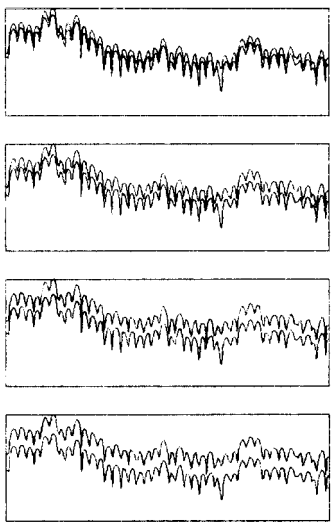


Fig. 10 원 음성과 롬바드 음성의 스펙트럼 특성

(SNR : 35, 25, 15, 5dB)

그러나 이러한 모델링은 스펙트럼이나 음성진폭 등을 변경시킬수 있으나 지속시간이나 화자의 심리적 변동에 따른 효과는 나타낼 수 없다. 따라서 완전한 모델링을 위해서는 LPC모델링 전에 피치를 변화시키는 것이 바람직하다. 피치주기는 앞에서 조사한 피치변동량을 기초로해서 변경시킨다.

피치변경은 피치반분법(pitch halving)과 영삽입방법(zero-insertion)을 사용한다.[10] 이러한 방법으로 원래의 한 프레임 음성에 대해 70% 변경한 음성을 Fig. 13에 나타내었다.

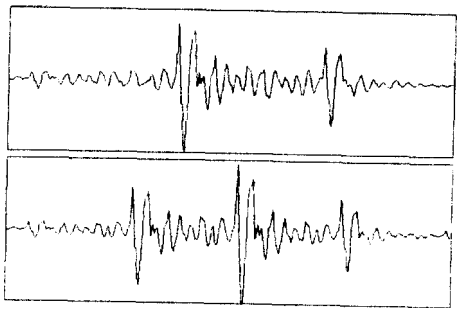


Fig. 11 원래의 음성파형과 70% 피치 변경한 음성파형

Ⅳ. 평가 실험

잡음에 대한 인식실험은 룸바드 영향을 고려한 음성신호와 고려하지 않은 음성신호에 대해 세가지 잡음의 종류와 다양한 신호대 잡음비에 따라 행한다. 그리고 실험에 사용된 인식기는 음소판별필터를 이용한 인식기이다.

실험에 사용한 잡음은 White 잡음, 저주파 잡음, 고주파 잡음들이다. 저주파 잡음과 고주파 잡음은 Fig. 11과 Fig. 12에서처럼 백색잡음을 Cutoff 주파수가 1.5KHz, 2KHz인 FIR LPF에 통과시켜서 만든다. 본 연구에서 사용한 LPF는 Kaiser Window를 이용한 필터이다. 그리고 Fig. 12에 저주파 잡음과 고주파 잡음을 사용해서 시뮬레이션한 룸바드 음성의 스펙트럼 특성을 원음성의 스펙트럼 특성과 비교하여 나타내었다.

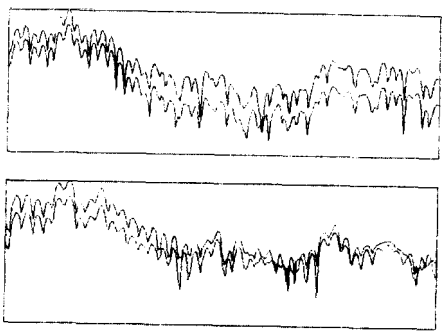


Fig. 12 고주파, 저주파 노이즈에 따른 룸바드 음성의 스펙트럼 특성

Ⅳ.3 실험 결과

실험결과는 아래의 도표, Fig. 13, Fig. 14에 그래프로 나타내었다. 여기에서 다음의 두가지 사실을 알 수 있다.

첫째, 룸바드 영향을 고려하지 않았을 때, 인식률 90%를 인식의 한계치로 가정한다면, 세종류의 잡음에 대한 인식기의 모음의 인식성능은 SNR 10[dB]임을 알 수 있다. 즉, 신호대 잡음비가 10 [dB]이하에서는 모음의 인식이 어렵다는 것을 예측할 수 있는 것이다. 그림에서 A는 백색잡음, B는 저주파 잡음, C는 고주파 잡음일 때의 인식률 곡선이다.

둘째, 룸바드 영향을 고려한 경우는, 대략 30dB에서 25dB 사이에서 인식율이 90%이다. III절에서 언급했듯이 Lombard Effect는 음성인식기의 성능저하에 중요한 영향을 미치는 것이 확인된다. 따라서 이 인식기의 모음에 대한 인식성능이 30dB정도의 신호대 잡음비임을 진단할 수 있고, 그 이하의 비에서는 인식율이 많이 저하됨을 예측할 수 있다.

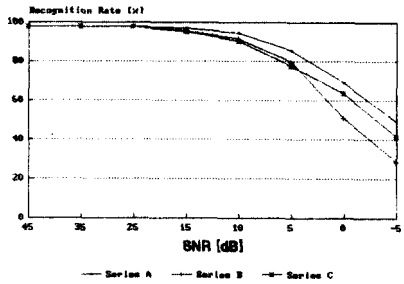


Fig. 13 룸바드 영향을 고려하지 않은 경우의 인식결과
Series A : White noise, Series B : Low freq. noise
Series C : High freq. noise

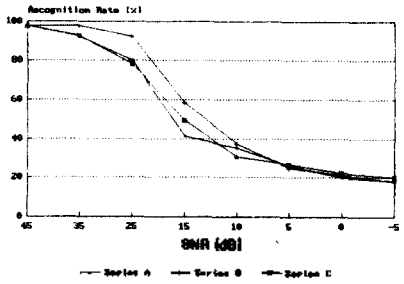


Fig. 14 룸바드 영향을 고려한 경우의 인식결과
Series A : White noise, Series B : Low freq. noise
Series C : High freq. noise

Fig. 14와 Fig. 15를 비교하면, 룸바드 영향을 고려한 경우가 인식율의 저하를 많이 가져옴을 알 수 있다. 이것은 Reference Pattern이 룸바드 영향을 고려한 경우의 음성보다 고려하지않은 경우의 음성에 더 가까운 특성을 나타내기 때문이다.

V. 결론 및 향후 연구과제

본 연구에서는 환경노인에 대한 인식기의 성능을 조사하기위해, 스펙트럼 이 전 주파수 대역에 걸쳐 거의 일정한 백색잡음과 다양한 신호대 잡음비에 따른 인식기의 인식율을 분석하였다. 그리고 좀 더 현실성에 가까운 잡음환경을 위해 롬바드 영향을 고려한 음성을 시뮬레이션하였고, 잡음을 포함하는 주파수 대역에서의 롬바드 현상을 고려하기위해 저주파 잡음과 고주파 잡음을 실험에 사용하였다.

기본주파수(F0), 평균음성에너지, 지속시간등의 세 가지 음성 분석 파라메터를 사용하여 롬바드 음성을 잡음의 레벨에 따라 분석한 결과, 기본 주파수는 무잡음의 F0를 기준으로 했을 때, 50dB에서 8%, 60dB에서 25%, 70dB에서 40%, 80dB에서 57%, 90dB에서 80%의 변동량을 알 수 있었다. 그리고 음성 에너지는 잡음의 레벨이 증가함에 따라 선형적으로 증가함을 알 수 있었고, 신호대 잡음비에 따른 음성에너지를 조사한 결과 '음성에너지 = -1.1K * SNR + 원래의 음성에너지 * 1.44'라는 선형 관계식을 얻었다. 그리고 지속시간은 일정하게 증가하지는 않았지만 전반적으로 잡음레벨의 증가에 대해 증가하는 성질을 알 수 있었다.

인식실험결과, 롬바드영향을 고려하지 않았을 때는 신호대 잡음비가 10dB정도에서 인식률 90%를 나타내었고, 롬바드 영향을 고려한 경우에는 25~30dB정도에서 동일한 인식률을 나타내었다. 따라서 롬바드 영향을 고려한 경우가 인식기의 인식률을 많이 저하시킴을 알 수 있다. 그리고 실험에 사용한 인식기의 성능은 SNR이 25dB 정도임을 알 수 있었다.

그리고 다양한 조건에 대한 인식결과를 얻었을 때, 이러한 결과들을 각 조건과 상관지어서 전반적인 인식기의 성능뿐만이 아니라 세부적인 인식기의 성능을 정량화하여 도출하는 기술이 개발되어야 한다.

참 고 문 헌

[1] Summers W.Van, Pisoni D.B.,Bernacki R.H.,Pedlow R.I. and Stokes M.A.," Effects of noise on speech production:acoustic and perceptual analysis",JASA,Vol.84,pp917-928,September 1988

[2] Steeneken J.M. & Velden J.G.,"RAMOS-Recognizer assessment by means of manipulation of speech",Proc.ESCA 1989 Paris, June 1989

[3] Lea W.A.,"What causes speech recognizers to make mistakes?",IEEE ICASSP, Vol.3,pp 2030-2033, 1982

[4] A.J.Fourcin,et al.(Ed):"Speech input and output assessment",Ellis Horwood(1989)

[5] "SAM Final Report, Year Three",SAM-UCL-G004, Feb. 1992

[6] Rajasekaran P.K.,Doddington G.R. and Picone J.W.,"Recognition of speech under stress and in noise",ICASSP,N14.10, pp733-736,1986,Tokyo

[7] Pisoni D.B.,Bernacki R.H.,Nusbaum H.C.and Yuchtman M.,"Some acoustic-phonetic correlates of speech produced in noise",ICASSP pp1581-1584,1985

[8] 허성필, 정현열, 김경태, "음소판별필터를 이용한 음성인식에 관한 연구", 영남대학교 석사학위 논문, 1993. 8

[9] 음성 입출력 시스템의 성능평가법 연구, 한국전자통신연구소 중간연구보고서, 1993. 8

[10] 배명진, "고음질 합성을 위한 피치변경법",한국음향학회지 12권 2호, 1993