롬바드 음성을 이용한 음성인식기의 성능 평가

Performance Assessment of Speech Recogniger using Lombard Speech

정 성 윤*, 정 현 열*, 김 경 태**

(Sung Yun Jung*, Hyun Yeol Chung*, Kyung Tae Kim**)

요 약

한국어 음성인식기의 성능평가를 위한 기초 연구로서 인식기의 성능에 영향을 끼치는 여러 요인 중 잡음환경 하에서의 롬바드 영향을 입은 음성을 인식하는 경우 인식기의 성능평가와 분석에 관해 논하였다.

성능평가에 있어서는 표준 음성데이타를 잡음환경에서 발성한 것에 가깝게 조작해서 롬바드 영향을 고려한 경우와 그렇지 않은 경우에 대해 평가항목(잡음의 종류, 신호대 잡음비)에 따라 인식실험을 행한 결과, 잡음의 종류는 인식성능에 영향을 미치지 않음을 알 수 있었고, 인식률 90%를 한계치로 했을 경우 롬바드 영향을 고려하지 않았을 때는 신호대 잡음비가 10dB 정도에서, 롬바드 영향을 고려한 경우에는 30dB정도에서 동일한 인식률을 나타내어 롬바드 영향을 고려한 경우가 20dB 정도의 인식률 저하를 가져와 실제 평가시 롬바드 영향을 고려해야 함을 알 수 있었다.

분산분석의 결과로부터는 여러 종류의 인식기를 다양한 평가항목에 대해 평가할 때, 각 평가 항목이 인식성능에 미치는 영향을 정량화할 수 있음을 알 수 있었다.

ABSTRACT

This paper describes the performance assessment test and the analysis of test results on a Korean speech recognizer which recognizes Lombard effect received speech in noisy environment, as a basic performance assessment research.

In the assessment test, standard speech data were first manipulated close to speech uttered in a noisy environment, and then performance assessment tests were carried out along with the assessment items(the type of noise, SNR)in two ways-one with Lombard effect received speech(LES), the other with not received(NLES).

As a result, when 90% of recognition rate is set to be a recognition limit, it was achieved at 10dB SNR point with LES, while at 30dB with NLES. This 20dB of SNR difference indicates Lombard effect should be considered in real world assessment test. The type of noises didn't affect performance of recognizers in our tests.

ANOVA analysis, in evaluating several kinds of recognizers, showed every assessment item affecting the recognition performance could be quantified.

*영남대학교 전자공학과
Dept. of Electronic Eng. Yeungnam University
**한남대학교 정보통신공학과
Dept. of Information & Communication Eng. Hannam University
접수일자 : 1994년 4월 29일

Ⅰ. 서 론

음성은 인식하는데 있어 인간과 비슷한 능력을 가진 음성인식시스템은 아직 개발되지 않았지만 음성인식 기술은 1960년대 초부터 연구가 진행되어 많은 발전을 거듭하여 왔다. 특히, 최근 10여년간의 급속한 디지털 신호처리 기술의 발전에 힘입어 유럽, 미국, 일본 등지에서는 상용 단어인식시스템이 개발되어 있는 상태이다. 이와같은 상용제품이 개발되기까지는 음성인식 알고리즘 개발에 있어서도 지속적인 진전이 있어 많은 새로운 방법들도 제안되었다.[1-5]

그러나 다양한 인식방법들에 기초한 인식기들은 개발자 개개인의 독자적인 방법에 의해 인식실험이 행해지고 성능이 평가되기 때문에, 서로 다른 인식기들 간의 객관적인 비교 평가가 불가능하게 되어 각 인식기의 성능에 대한 신뢰성이 결여되고 있는 한편 성능이 더 향상된 인식기의 개발을 어렵게 하는 문제점이 있다. 따라서 개발된 인식기들의 성능을 객관적으로 비교 평가함은 물론 인식기의 최종 성능한계까지도 예측할 수 있는 평가방법을 개발해야 할 필요가 있다. 그리고 평가한 결과를 각 인식기의 개발자에게 통고함으로써 더 향상된 성능의 인식기의 개발을 유도할 수도 있다.

인식기의 성능평가에 관한 연구는 1989년에 본격적으로 시작되어 유럽에서 가장 체계적으로 이루어져 왔다. 유럽에서는 EC통합에 따라 각국 언어간 자유로운 통신을 위해 ESPRIT(european strategic program for research and developement in Information technology)계획의 일환으로 다국 언어 인식 시스템의 개발을 추진하고 있는데, 그 중에 성능평가만을 전담하는 SAM(speech assessment methodologies)프로젝트에는 8개 나라의 28개 연구소, EC내의 6개 연구소, EFTA(유럽자유무역협회)의 2개 연구소가 참여하여 작업환경(SESAM)을 통일화하여 공동연구를 하고 있다.[6,7] 특히, 미국은 음성인식, 합성기의 성능평가의 필요성을 일찍부터 인식하고, NIST를 중심으로 유럽등과 국제적인 연구협력활동으로 확대해가고 있는 실정이다.[8]

그러나, 한국어의 경우 음성합성기에 대한 성능평가에 대해서 규칙합성 연구가 활발해지면서 개발자 나름대로의 검토가 행해지고 있는 단계이지만 음성인식기의 성능평가에 대한 체계적인 연구는 아직 착수되고 있지 않은 상태이다.

따라서, 본 연구는 한국어 음성인식기의 성능평가를 위한 기초연구로서 인식기의 성능에 영향을 끼치는 여러가지 요인중 가장 중요하다고 생각되는 환경요인에 의한 인식기의 성능평가에 중점을 두어 특히 잡음환경 하에서의 롬바드 영향을 입은 음성을 인식하는 경우 인식기의 성능을 평가하는 데 연구의 범위를 한정하여 성능 평가시 롬바드 영향을 고려해야함을 보이고, 분산분석을 통하여 인식성능에 영향을 미치는 정도를 정량화하기로 한다.

Ⅱ. 음성인식기의 평가 방법

음성인식기를 가장 객관적으로 평가하는 방법은 모든 장소에서 모든 계층의 사람들이 모든 경우의 상황(물리적상황, 정신적 상황 등)에 대해 직접 테스트를 해서 인식결과를 서로 비교하면 되지만, 이러한 방법은 시간과 경제적인 측면에서 실현이 어렵다. 이러한 어려움을 극복하기 위해서 일반적으로는 여러 계층의 사람들이 여러 환경에서 발성한 음성을 녹음한 공통의 평가용 데이타베이스를 이용한다. 그러나, 이 경우 상세한 성능진단을 위해서는 다양한 조건에서 녹음한 음성데이타가 필요한데, 각 조건에 대한 인식기의 성능평가를 위해 그 상황에 알맞는 음성을 일일이 녹음해서 평가하는 것은 매우 힘들다. 따라서, 표준적인 음성데이타가 수집되어 있을 때 수집된 음성데이타로서 각 조건에 해당하는 음성을 조작하는 기술이 필요하다.[9] 음성데이타를 조작하게되면, 조작하는 정도에 따른 각 음성데이타가 갖고 있는 특징을 정량화할 수 있고, 인식결과를 평가할 때에 평가척도로 사용할 수 있게 된다. 그림 1은 일반적인 음성조작에 의한 인식기의 평가 블럭도이다. 그림에서처럼 성능평가를 위해서는 공통음성이나 연구실에서 개발할 때 사용한 음성을 토대로 조작기와 결과 해석기를 거쳐서 성능을 평가한다.

인식기를 평가하기 위해서는 먼저 인식기의 성능에 많은 영향을 미치는 요인들을 모아서 평가항목을 구성해야 한다. Lea[10]는 인식성능에 영향을 미치는 80가지 이상의 많은 요인들 중에서 특히 인간적 요인, 언어요인, 환경요인의 세 가지 요인이 상대적으로 영향력이 크며 이중에서도 특히 잡음레벨(SNR), 잡음형태등의 환경요인이 가장 크게 인식성능을 저하시키는 요인으로 보고하고 있다.

이를 참고로 하여 본 연구에서는 환경요인에 의한

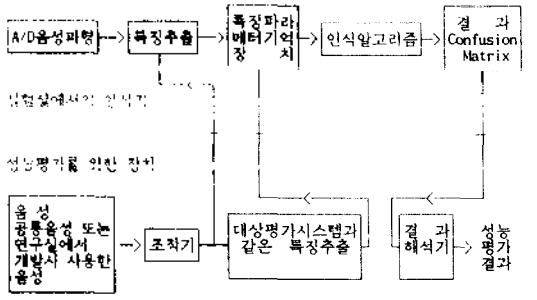


그림 1. 음성인식기의 성능평가 블럭도
Fig 1. Block diagram of performance assessment for speech recognizer

인식기의 성능평가에 중점을 두고 3종류의 잡음(백색, 저주파, 고주파잡음)과 4종류의 신호대잡음비(35, 25, 15, 5dB)를 평가항목으로 설정한 후, 음성을 잡음환경에 가깝게 조작해서 각 평가항목에 따른 인식기의 성능을 분석한다. 이 때 실제 잡음환경 하에서 발생된 음성을 합성하기 위해 롬바드 영향을 고려하기로 한다.

Ⅲ. 롬바드 영향을 고려한 음성의 합성

잡음환경 하에서 발생된 음성에 대한 인식기의 성능을 평가하기 위해서는 먼저 음성데이타를 잡음환경에서 발성한 것에 가깝도록 조사해야 한다. 이를 위한 한 방법으로 일반적으로는 깨끗한 음성에 잡음을 더하므로써 잡음이 있는 음성신호를 만들어내는 방법을 많이 사용하고 있다. 그러나, 배경 잡음이 존재하는 상황에서 화자는 자신의 발성을 변경해서 발성하는 경향(이를 롬바드 영향이라고 한다)이 있으므로 보다 정확한 평가를 위해서는 이를 고려하지 않으면 안된다. 그림 2에 잡음이 있는 환경에서 롬바드 영향을 받아 발성이 변화되는 과정을 보인다.

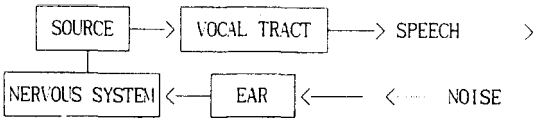


그림 2. 잡음환경 하에서의 발성변화 과정
Fig 2. Process of utterance variation in noisy environment

따라서, 깨끗한 음성에 백색잡음을 첨가한 경우와 롬바드 영향을 고려한 음성에 잡음을 첨가한 경우에 대하여 실제 인식결과에 어느 정도의 영향을 미치는가를 파악하기 위하여 롬바드 영향을 고려하지 않고 단순히 백색잡음만을 혼입한 경우와 롬바드 영향을 고려한 경우로 나누어 고찰하기로 한다.

3.1 음성에 백색잡음이 혼입된 경우

깨끗한 음성에 백색잡음이 혼입된 경우를 묘사하기 위해서는 먼저 백색잡음을 발생시켜야 한다. 백색잡음은 평균치가 0이고, 전력스펙트럼밀도가 전 주파수에 걸쳐 일정한 정상적 확률과정을 갖는 것으로 하여 C언어의 rand 함수를 사용하여 −0.5에서 0.5사이의 난수를 발생시킨다. 이를 원 음성에 부가할 때에는 잡음발생시에 임의의 실수를 곱하므로써 잡음의 전력을 변경시킬 수 있도록하여 원하는 SNR을 얻도록 하였다.

3.2 롬바드 영향(Lombard effect)을 고려한 경우

진술한 바와같이 주위에 잡음이 존재할 때에 화자가 자신의 발성을 변경해서 발성하는 것을 롬바드 영향이라고 한다. 배경잡음의 강도에 따라 발성이 어느 정도 변하는가에 대한 정량적인 연구는 아직 행해지지 않고 있으나, 음성발성시에 존재하는 잡음의 심리적인 영향은 단순히 잡음만을 더한 경우보다 인식결과에 더 큰 영향을 끼친다는 사실이 Rajasekaran[11]에 의해 밝혀진 이래 배경잡음이 존재하는 상황에서 음성신호의 변화를 명백히 규명하기 위해 몇가지 연구들이 수행되어 왔다.[12,13]

이러한 연구들에 의해 밝혀진 결과들을 간략하면 다음과 같다.

1. 잡음환경 하에서 화자들은 Vocal effort를 증가시키므로써 더 강력한 음성을 발성한다.
2. 잡음환경 하에서 발성한 음성의 지속시간(duration)은 증가하는 경향이 있다. 이 증가는 비선형적이고 음소들에 종속적이다.
3. 평균 기본주파수(Fo)는 매우 큰 증가를 보인다.
4. 상위 포만트들(upper formants)이 더 강렬(intense)하고, 스펙트럼의 기울기(spectreal slope)가 커진다.
5. 잡음만을 부가한 음성과 비교할 때 명료도(intelligibility)가 향상 된다.
6. 롬바드 영향은 인식성능에 상당한 영향을 미친다.

의 인식률과를 인식기의 성능평가에 이용하기 위해서는 주성분에 대한 정량적인 평가용 데이타가 필요하나 반동에 대한 정확한 정량적인 데이타를 제시하지 못하고 있다. 따라서, 여기서는 실제 롬바드 영향을 입은 음성을 녹음하여 시간 영역, 주파수 영역에서의 정량적 분석을 통해 인식기의 성능평가시 롬바드 영향을 입은 음성을 사용해야 함을 밝히고, 이를 실제 인식기의 성능평가시 사용키로 한다.

3.3 롬바드 음성의 녹음 및 분석

3.3.1 롬바드 음성의 녹음

잡음환경 하에서의 롬바드 영향을 입은 음성을 녹음하기 위해 잡음의 레벨은 무잡음, 50, 60, 70, 80, 90dB의 6가지로 하고 이 각각의 잡음레벨에 대해 5개의 단모음(ㅏ, ㅔ, ㅣ, ㅗ, ㅜ)을 발성한 것을 녹음한다. 이 때 화자는 표준어를 사용하는 남성화자 2명과 경상도 방언을 쓰는 남성화자 1명으로 하고 각 화자가 3회 발성한 것을 분석용 음성데이타로 한다. 롬바드 음성을 녹음할 때 특히 주의할 사항은 각 잡음레벨의 잡음을 화자의 음성과 함께 녹음해서는 안된다.

전체적인 녹음과정을 그림 3에 보였다.

그림 3의 롬바드 음성의 녹음과정은 다음과 같다.

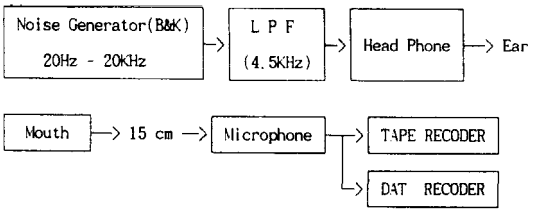


그림 3. 롬바드 음성의 녹음과정

Fig 3. Recording process for Lombard effect received speech

먼저, 발성자는 방음실에 들어가서 헤드폰을 쓰고 편안하게 앉는다. 발성자가 방음실의 환경에 익숙하고 난 뒤, 발성자에게 발성할 리스트가 적혀있는 명령안내서를 주고 15분 정도 조용한 상태에서 발성연습을 시킨다. 명령안내서에는 발성할 때에 주의할 사항과 발성할 내용이 잡음레벨에 따라 랜덤하게 적혀져 있다. 발성시 각 화자는 한 단어를 발성한 후, 반드시 1-2초 정도 쉬어야 한다.

발성연습이 끝나면 잡음발생기를 이용하여 20Hz에서 20KHz 주파수대역의 백색잡음을 발생시켜 4.5kHz LPF를 거친 후 헤드폰의 양쪽에 들려질 수 있도록 한다. 이때 잡음레벨은 6종류로하고 각 잡음레벨에 대해 3회 발성을 한다.

발성된 음성은 고질의 콘덴서 마이크로폰을 사용하여 tape recoder와 DAT에 음성을 기록한다. 마이크로폰에서 입까지의 거리는 15cm정도로 한다.

3.3.2 롬바드 음성의 분석

롬바드 영향을 입은 음성의 변동 특징을 조사하기 위하여 각 잡음레벨에서 발성한 음성의 평균에너지와 지속시간, 피치를 음향적 파라메터로 하여 그림 5와 같은 절차에 따라 분석을 행한다. 즉, 먼저 3명이 3회 발성한 모음 270개(3명×3회×5모음×6조건)는 4.5KHz LPF를 통과시킨 후, 10KHz sampling, 12bit 양자화 한 다음, 수작업 처리를 거쳐 파일 이름과 기록 환경, 음성파형의 정보 및 레이블링 정보와 함께 저장한다.

이와 같이 디지탈화된 롬바드 음성에 대해 3.2절에서 확인된 롬바드 영향의 특징을 조사하기 위해 기본주파수(피치), 에너지, 지속시간의 세 가지 파라메터에 대한 변동량을 분석하기로 한다.

3.3.2.1 기본주파수의 변화량 분석

기본주파수는 음성의 높낮이를 나타내게 되므로 피치라고도 하는데 기본 주파수의 시간축에 대한 변화패턴은, 음성에 포함되는 악센트, 억양, 강세등의 운율적 특징을 반영하기 때문에 음성분석 파라메터로 많이 이용된다. 피치분석에는 켑스트럼 방법이나 LPC모델의 예측잔차 방법등이 있으나 본 연구에서는 단모음을 분석 대상으로 하기 때문에 정밀도는 약간 떨어지지만 계산이 쉬운 자기상관함수(autocorrelation) 방법을 사용하여 기본 주파수를 추출하였다. 이때 각 모음의 기본 주파수는 정상부분 총 10개의 프레임에 대한 평균치이다.

이렇게 하여 얻은 세 명의 화자에 대한 각 잡음레벨에 따른 다섯 모음의 평균 기본주파수의 특성은 그림 4와 같다. 그림 4로부터 잡음의 레벨이 증가함에 따라 기본주파수가 크게 증가함을 알 수 있고 역으로 피치주기는 짧아지는 것을 알 수 있다. 시간영역에서 피치주기를 조작하기 위해 각 잡음레벨에 따른 피치주기의 평균변화량을 조사하여 백분율로 나타내면 표 1과 같게 된다. 이 때 피치주기의 평균변화량은 식

표 1. 피치주기의 평균변화량

	50dBSPL	60dBSPL	70dBSPL	80dBSPL	90dBSPL
평균변화량	91 %	81 %	72 %	65 %	57 %

(1)로부터 구한다.

$$\text{피치주기의 변화량} = \frac{1}{M}\sum_{i=0}^{M-1} \frac{\text{각 잡음레벨에서의 피치주기}}{\text{무잡음일 때의 피치주기}} \times 100 \qquad (1)$$

여기서, M은 각 잡음레벨에서의 모음의 갯수이다.

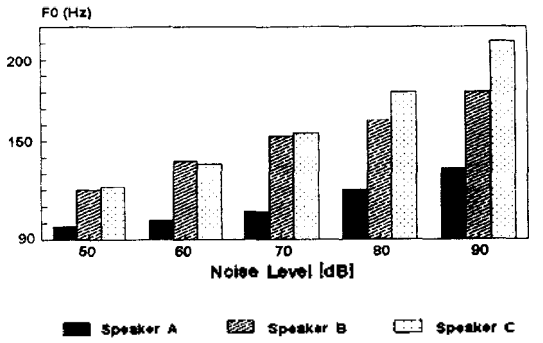


그림 4. 세 화자의 평균 기본주파수 특성.
Fig 4. Characteristics of average F0 of 3 speakers.

3.3.2.2 평균에너지와 지속시간의 변화량 분석

평균 에너지는 시찰에 의해 음성의 끝점을 검출한 후 두 끝점사이의 평균 음성에너지에 대해 대수로그를 취한 식(2)를 사용하여 구하였다.

$$E = 10\log((1/N)\sum_{i=0}^{N-1}(s[i]\cdot s[i])) \qquad (2)$$

여기서 E는 끝점 검출된 음성의 평균에너지이고 s[i]는 i번째 음성샘플을 나타내며, N은 두 끝점사이의 음성 샘플 수를 나타낸다.

그림 5는 3명의 화자에 대한 다섯 모음의 평균 에너지의 분포이다. 잡음 레벨이 증가함에 따라 지속시간이 에너지가 선형적으로 증가함을 나타낸다. 3명의 화자 모두 비슷한 기울기의 증가를 나타내기 때문에 선형함수로 근사화할 수 있음을 알 수 있다.

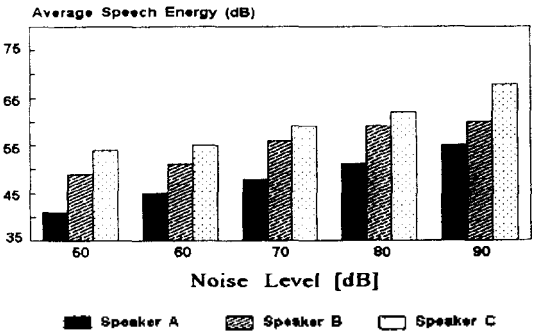


그림 5. 세 화자에 의해 발성된 5모음의 평균 에너지.
Fig 5. Average Energy distribution of 5 Vowels uttered by 3 speakes.

신호대 잡음비에 따른 롬바드 음성의 특징 파라메터들을 분석하기 위해서는 실험에 사용한 잡음레벨 단위인 SPL(sound presure level)을 식 (2)에서의 단위와 동일하게 하여야 잡음의 에너지와 음성신호의 에너지를 같은 단위 내에서 계산할 수 있다. 이를 위해 각 레벨(SPL)에서 녹음한 잡음을 A/D변환하여 식 (2)를 사용하여 잡음의 에너지를 구한 후, 신호대 잡음비에 따른 음성에너지의 특성을 식 (3)과 같이 일차의 선형함수로 나타내어 LPC 잡음 모델링에 의한 롬바드 음성의 합성에 이용한다. 표 2에 잡음의 에너지를 나타낸다.

$$\text{잡음에너지[dB]} = -1.18 \times \text{SNR} + \text{원음성에너지} \times 1.44 \qquad (3)$$

표 2. 잡음에너지(SPL dB와 dB)

SPL dB	50 dB	60 dB	70 dB	80 dB	90 dB
dB	31 dB	37 dB	43 dB	49 dB	55 dB

이상의 결과를 종합하면, 잡음의 레벨에 따라 피치, 평균에너지가 변하기 때문에 인식기의 평가용 음성데이타는 롬바드 영향을 고려한 것이라야 함을 알 수 있다. 따라서, 본 연구에서는 롬바드 영향을 고려한 음성을 합성해서 인식기의 성능을 평가하도록 한다.

3.4 롬바드 음성의 합성방법

3.3절에서 분석한 결론에 따른 평가용 음성데이타를 합성하기 위하여 묵음시 파라메터(피치, 에너지)의 변화량을 조작하므로써 롬바드 음성을 합성하기로 한다.

3.4.1 피치 변경

피치변경에 있어서는 그림 6처럼 내삽(Interpolation)과 솎음(decimation)[11]을 사용한다. 먼저 해밍창(Hamming Window)을 사용하여 단구간의 음성을 취해서 켑스트럼 분석을 한다. 이 때 창의 길이는 320[msec]이고 160msec씩 이동하게 된다. 켑스트럼 분석을 통해 피치주기를 검출한 후 변경할 피치길이(P)만큼의 음성샘플 수를 계산해서 내삽과 솎음을 행한다. 이렇게 행한 각 프레임별 음성신호는 overlap-add 방법을 사용하여 피치가 변경된 음성파형을 합성하게 된다.

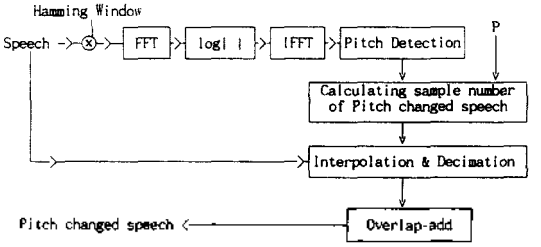


그림 6. 내삽과 솎음을 사용한 피치변경 블럭도
Fig 6. Block diagram of pitch modification by interpolation and decimation

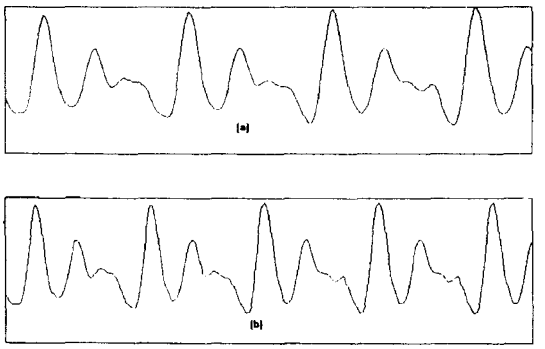


그림 7. 원래의 음성파형(a)와 원래의 피치의 80%로 피치 변경한 음성 파형(b)
Fig 7. Orignal speech wave(a) and modified speech wave with 80% of the original pitch(b)

그림 7에 원음성 (a)과 이러한 방법에 의해 원래의 피치길이의 80% 길이로 변경된 다른 음성 (b)을 나타내었다.

3.4.2 에너지 변경(LPC 잡음모델링)

평균에너지의 변경은 LPC 잡음모델링[16]을 사용하여 식 (3)을 토대로 행한다. LPC 잡음모델링은 그림 8과 같이 잡음에 대한 선형필터 계수를 구한 후 음성신호를 입력하여 필터링을 하므로서 잡음에너지의 영향을 받은 음성신호를 만들어 내는 원리이다. 이렇게 만든 음성신호는 다시 잡음과 더해지므로써 배경잡음에 의해 영향을 받은 음성이 된다.

전체적인 과정은 다음과 같다.

1. 자연로그를 사용하여 잡음신호의 스펙트럼을 계산한다.
2. 이 스펙트럼은 신호대 잡음비에 따른 잡음의 에너지로 정규화한다.
3. 고주파 부분을 강조하기 위해 정규화된 잡음의 스펙트럼에 식 (6)의 프리엠퍼시스(Preemphsis)를 가한다.

$$H(z) = 1 - 0.9375z^{-1} \qquad (6)$$

4. 프리엠퍼시스된 스펙트럼에 지수를 취하고 IFFT하여 자기상관계수를 구한다.
5. 그리고, Levinson-Durbin알고리즘을 사용하여 합성필터의 예측계수를 계산한다.

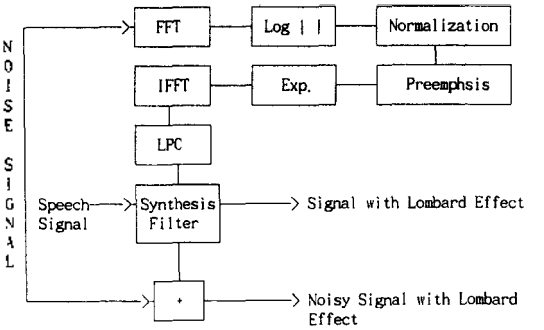
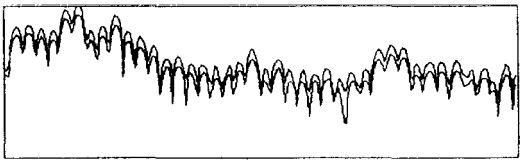


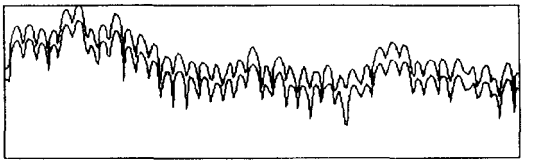
그림 8. LPC 잡음모델링에 의한 롬바드 음성합성
Fig 8. Loambard speech synthesis by LPC noise modeling

그림 9에 백색잡음을 사용하여 신호대 잡음비가 각각 35, 25, 15, 5dB일때의 LPC잡음을 모델링한 롬바

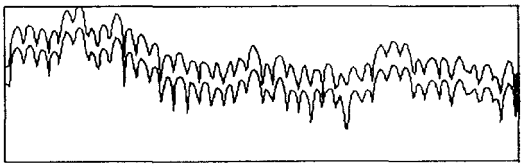
드 음성과 원 음성과의 스펙트럼을 나타내었다. 그림에서 백색잡음을 포함하고 있는 전 주파수 대역에서 신호대 잡음비가 작을수록 스펙트럼 에너지가 증가함을 알 수 있다.



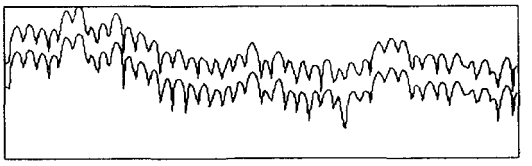
(a) SNR = 35 dB



(b) SNR = 25 dB



(c) SNR = 15 dB



(d) SNR = 5 dB

그림 9. 신호대 잡음비에 따른 원 음성(하측)과 롬바드음성(상측)의 스펙트럼

Fig 9. Orignal(lower line) and Loambard speech(upper line) spectral in different SNRs.

Ⅳ. 인식기의 성능평가 및 분석

여기서는 이와 같은 방법으로 합성한 음성을 이용하여 음성인식기의 성능평가에 대해 기술한다. 이때 인식기는 음소판별필터를 이용한 인식기를 사용한다.[5] 평가 실험의 전 과정을 그림 10에 나타내었다.

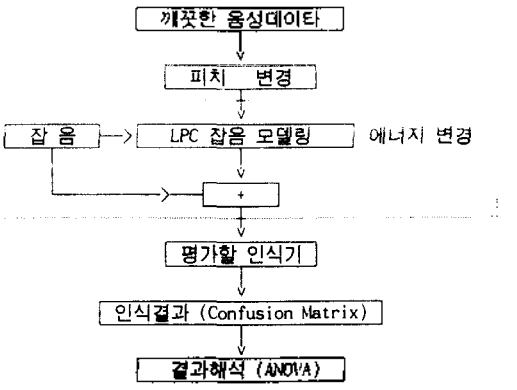


그림 10. 인식기의 성능평가 실험의 전 과정

Fig 10. Flow of performance assessment test for speech recogizer

평가실험은 롬바드 영향을 고려한 음성신호와 고려하지 않은 음성신호에 대해 세 가지 잡음의 종류와 4가지의 신호대 잡음비에 대해 각각 행한다. 여기서 롬바드 영향을 고려한 음성신호는 3.4절의 피치와 에너지를 변경하는 방법으로 롬바드 음성을 합성한 후 잡음을 첨가한 것이고 롬바드 영향을 고려하지 않은 음성신호는 깨끗한 음성에 잡음을 첨가한 것이다.

4.1 음성데이타

인식기를 평가하기 위해 사용하는 깨끗한 음성데이타는 공통의 음성 데이타베이스에서 수집되어야 한다. 그러나, 국내에는 공통의 데이타베이스가 구축되어 있지 않기때문에 공통의 데이타베이스가 구축되어 있다는 가정하에 실험을 행하도록 한다. 따라서 본 연구실이 보유하고 있는 549 단음절 데이타를 공통의 데이타로 가정한다. 이 음성데이타는 성인 남성 화자 3명이 방음실에서 랜덤하게 각 3회씩 자연스럽게 발성한 것으로서 총 549개이다. 이 중에서 본 실험에 사용된 데이타는 화자 1명이 1회 발성한 다섯 모음(/ㅏ/, /ㅔ/, /ㅣ/, /ㅗ/, /ㅜ/)이며, 모음 /ㅐ/는 /ㅔ/에 같이 포함시켜서 사용하였다.

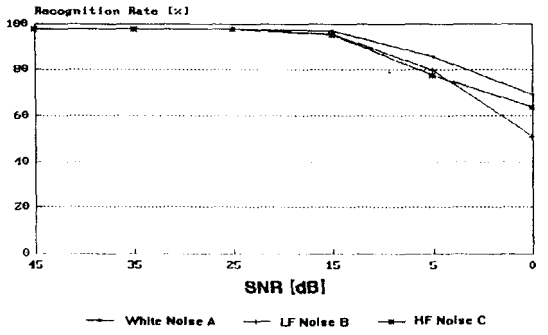
4.2 평가 결과

그림 11에 롬바드 영향의 고려 여부에 따라 평가한 결과를 나타낸다. 여기에서 A는 백색잡음, B는 저주파 잡음, C는 고주파 잡음일 때의 인식율 곡선이다. 실험에 사용한 잡음들은 백색 잡음, 저주파 잡음, 고주파 잡음의 3종으로 저주파 잡음과 고주파 잡음은

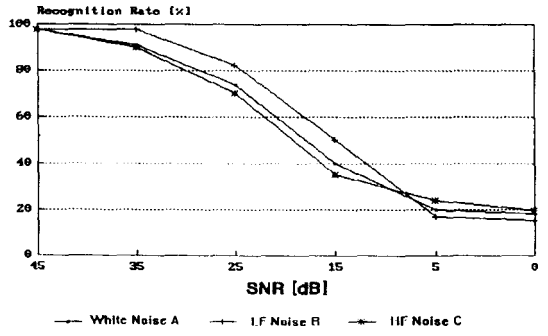
백색 잡음을 Cutoff 주파수가 각각 1.5KHz, 2kHz인 FIR, LPF, HPF에 통과시켜서 만든다. 이때 사용한 윈도우는 Kaiser Window를 이용한 것이다.

(1)롬바드 영향을 고려하지 않은 경우, 인식률 90%를 인식의 한계치로 가정할 때, 신호대 잡음비는 10dB이므로 세 종류의 잡음에 대한 인식기의 성능은 이 점에서 결정되어진다. 따라서 이 이하에서는 인식이 어렵다는 것이 예측되어진다.

(2)롬바드 영향을 고려한 경우, 35dB에서 25dB 사이에서 인식율이 90%이므로 대략 30dB의 신호대 잡음비에서 인식기의 성능이 결정되어진다. 따라서 롬바드 영향을 고려하지 않은 경우의 인식결과와 비교해 볼 때 약 20dB정도의 차를 보이므로 롬바드 영향의 음성인식기의 성능저하에 중요한 영향을 미치는 것을 알 수 있다. 따라서, 환경요인에 관한 성능평가



(a)



(b)

그림 11. 롬바드 영향의 유무에 따른 인식결과
a) 롬바드 영향을 고려하지 않은 경우의 인식결과
b) 롬바드 영향을 고려한 경우의 인식결과

Fig 11. Recognition Results tested
a) with Loambard Effect
b) without Loambard Effect

시 롬바드 영향을 반드시 고려해야 함을 알 수 있다.

(3)앞의 두 경우 모두 잡음의 종류는 인식기의 성능에 거의 영향을 미치지 않는다. 이것은 모음에 대한 인식성능만을 평가했기 때문으로 만약 자음인식 결과를 포함시키면 잡음의 종류는 인식성능에 상당한 영향을 미치리라 생각된다.

4.4 분 석

4.3절의 평가결과에서, 이 인식기의 인식성능은 백색 잡음, 저주파 잡음, 고주파 잡음 등의 잡음의 종류에 대해서는 거의 영향을 받지 않았지만 신호대 잡음비가 30dB 이하에서는 심각한 영향을 받는다는 것을 알았다.

그러나 여러 종류의 인식기의 성능을 다양한 평가항목에 따라 비교 평가할 경우, 인식성능에 가장 큰 영향을 미치는 요인이 무엇인지, 그리고 그영향이 어느 정도인지를 알 필요가 있다. 따라서 여기에서는 성능평가 결과에 대한 분산분석을 행하여 평가항목이 인식성능에 미치는 영향의 정도를 정량화하기로 한다.

분산분석(ANOVA)은 특성값의 분산이나 분포를 분석하는 방법으로써 특성값의 분포를 제곱합으로 나타내고, 이 제곱합을 실험에 관련된 요인별로 분해하여, 순수한 오차에 의한 영향보다도 큰 영향을 주는 요인이 어떤 것인가를 규명하는 방법이다. 본 연구에서는 반복이 없는 이원 배치법을 사용하여 분석을 행하였다.

표 3은 롬바드 영향을 고려하지 않은 경우의 분산분석표이다. 표에서 두인자(factor)는 잡음의 종류와 신호대 잡음비이고 가설의 검정은 유의수준 95%로 하였다.

표 3. "깨끗한 음성+잡음"의 분산 분석표

인 자	제곱합	자유도	제곱평균	F비
잡음의 종류	113570	2	56785	1.2865
SNR	3841152	4	960288	21.7559
잔 차	353111	8	44139	
계	4307833	14		

첫번째 인자인 잡음의 종류가 인식성능에 영향을 미치지 않는다는 가설을 유의수준 95%로 검정하기 위해 가설의 F비를 계산하면 $F(2, 8 : 0.05) = 4.4590$ 이 된다. 잡음의 종류에 대한 F비는 1.2865가 되어 가

설의 F비보다 작다. 따라서 이 가설은 채택되어지며 잡음의 종류는 인식성능에 영향을 미치지 않고, 그 정도가 가설의 F비(4.459)를 100%로 할 때 71%임을 알 수 있다.

신호대 잡음비의 경우에도 위와 같은 가설을 세우고 유의수준 95%로 검정을 하면, 가설의 F비는 F(4, 8 ; 0.05) = 3.8378이 된다. 이 경우에는 신호대 잡음비에 대한 F비가 21.7559가 되어 가설의 F비보다 훨씬 크므로 가설은 기각되어져서, 신호대 잡음비는 인식성능에 영향을 미치는 것임을 알 수 있고, 그 정도가 F(4, 8 ; 0.05)를 100%으로 할 때 467%임을 알 수 있다. 따라서 상당히 많은 영향을 미침을 알 수 있다.

표 4는 롬바드 영향을 고려한 경우의 분산 분석표이다. 표 3의 경우와 동일하게 분석을 행하면, 잡음의 종류에 대해서는 2.03 < F(2, 8 ; 0.05) = 4.4590가 되어 가설은 채택되어지고, 신호대 잡음비에 대해서는 122.17 > F(4, 8 ; 0.05) = 3.8378이 되어 가설은 기각되어진다. 그리고 각각에 대해 가설의 F비를 100%로 할 때, 46%, 3183%로 롬바드 영향을 고려하지 않은 것과 비교할 때 각각 25%, 2671% 정도의 더 많은 영향을 미침을 알 수 있다.

표 4. "롬바드 음성 + 잡음"의 분산 분석표

인 자	제곱합	자유도	제곱평균	F비
잡음의 종류	625605	2	312803	2.03
SNR	75415314	4	18853829	122.17
잔 차	1234643	8	154330	
계	77275562	14		

V. 결 론

본 연구는 한국어 음성인식기의 성능평가를 위한 기초 연구로써 인식기의 성능에 영향을 끼치는 여러 요인 중 잡음환경 하에서의 롬바드 영향을 입은 음성을 인식하는 경우 인식기의 성능평가와 분석에 관한 연구이다.

평가와 분석은 표준 음성데이타를 잡음환경에서 발성한 것에 가깝게 조작해서 롬바드 영향을 고려한 경우와 그렇지 않은 경우에 대해 평가항목(잡음의 종류, 신호대 잡음비)에 따라 인식실험을 행한 후 분산분석을 통하여 각 평가항목이 인식성능에 어느 정도의 영향을 미치는가를 정량화하였다.

그 결과, 잡음의 종류는 인식성능에 영향을 미치지 않음을 알 수 있었고, 인식을 90%를 한계치로 했을 경우 롬바드 영향을 고려하지 않았을 때는 신호대 잡음비가 10dB정도에서, 롬바드 영향을 고려한 경우에는 30dB정도에서 동일한 인식율을 나타내어 롬바드 영향을 고려한 경우가 20dB 정도의 인식율 저하를 가져와 실제 평가시 롬바드 영향을 고려해야 함을 알 수 있었다.

분산분석을 행한 결과, 잡음의 종류에 대해서는 가설의 F비를 100%로 할 때 영향을 미치지 않는 정도가 롬바드 영향을 고려하지 않은 경우에는 71%, 고려한 경우에는 46%로 롬바드 영향을 고려하지 않은 경우에 25% 더 많음을 알 수 있었다. 신호대 잡음비의 경우, 영향을 미치는 정도가 롬바드 영향을 고려하지 않은 경우에는 467%, 고려한 경우에는 3183%로 롬바드 영향을 고려한 경우에 2671% 더 많음을 알 수 있었다. 이로부터 여러 종류의 인식기를 다양한 평가항목에 대해 평가할 때, 분산분석을 통하여 각 평가항목이 인식성능에 미치는 영향을 정량화할 수 있음을 알 수 있다.

참 고 문 헌

1. Sakoe, H. and Chiba, S., "Dynamic Programming algorithm optimization for spoken word recognition," IEEE Trans. ASSP-26, 1, pp.43-49, 1978.
2. Rabiner, L. R., Levinson, S. E., and Sondhi, M. M. "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," Bell Systems Tech. J., 62, 4, pp. 1075-1105, 1983.
3. B. Widrow and M. A. Lehr, "30 Years od Adaptive Neural Networks : Perceptron, Madaline, and Backpropagation," Proc. IEEE, Vol.78, No.9, Special Issue on Neural Networks 1.
4. Victor W. Zue, "The Use of Speech Knowledge in Automatic Speech Recognition," Proc. IEEE, Vol. 73, No.11, pp.1602-1615, 1985.
5. 여상필, 정현열, 김상배, "음소판별필터를 이용한 음성인식에 관한 연구," 영남대학교 석사학위 논문, 1993. 8
6. "SAM Final Report, Year Three," SAM-UCL-G004, Feb. 1992.
7. A. J. Fourcin, et al.(Ed), "Speech input and output assessment," Ellis Horwood(1989)
8. D. Pallet, "Speech input assessment using benchmark tests : Procedures, advantages and limit-

ations," Proc. of ESCA Tutorial Day and Workshop on Speech Input/Outout Assessment and Speech Databases, pp.33-36, 1989.

9. 정성윤, 정현열, 김경태, "음성인식기의 환경요인에 대한 성능평가," 제10회 음성통신 및 신호처리워크샵, SCAS-10권 1호, pp.251-255, 1993.

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

11. Rajasekaran P. K., Doddington G. R. and Picone J. W., "Recognition of speech under stress and in noise," ICASSP, N14, 10, pp.733-736, 1986, Tokyo

12. Pisoni D. B., Bernacki R. H., Nusbaum H. C. and Yuchtman M., "Some acoustic-phonetic correlates of speech produced in noise," ICASSP pp. 1581-1584. 1985.

13. Summers W. Van, Pisoni D. B., Bernacki R. H., Pedlow R. I. and Stokes M. A., "Effects of noise on speech porduction ; acoustic and peceptual analysis," JASA, Vol.84, pp.917-928, September 1988.

14. 정성윤, 정현열, 김경태, "Interpolator와 Decimator를 이용한 샘플링 주파수 변환," 한국음향학회 학술논문발표회 논문집, 제 11권 1호, pp.39-42, 1992.

15. L. R. Rabiner and C. A. McGonegal, "FIR Windowed Filter Design Program-WINDOW," Program for Digital Signal Processing

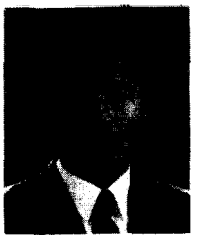
▲정 성 윤(Sung Yun Jung) 1968년 10월 2일생

1991년 2월 : 경북대학교 전자공학과 졸업(공학사)

1994년 2월 : 영남대학교 대학원 전자공학과 졸업(공학석사)

1994년 ~ 현재 : 태일정밀(주) 연구소

▲정 현 열(Hyun Yeol Chung) 1951년 11월 26일생



1975년 2월 : 영남대학교 전자공학과 졸업(공학사)

1981년 2월 : 영남대학교 대학원 전자공학과 졸업(공학석사)

1989년 4월 : 일본東北대학 대학원 정보공학과(공학박사)

1992년 7월 ~ 1993년 7월 : Carnegie-Mellon University Robotics Institute, 방문교수.

1992년 7월 ~ 현재 : 영남대학교 전자공학과 부교수

※주관심분야 : 음성신호처리 및 그 응용

▲김 경 태(Kyung Tae Kim) 1949년 5월 9일생

1972년 2월 : 경북대학교 전자공학과 졸업(공학사)

1980년 8월 : 연세대학교 전자공학과 대학원 졸업(공학석사)

1985년 3월 : 일본東北대학 대학원 전기및 통신전공(공학박사)

1991년 2월 : 한국전자통신연구소 신호처리연구실(실장, 책임연구원)

현재 : 한남대학교 정보통신공학과 교수