《主 題》

음성입출력장치의 성능평가

이 용 주* · 김 경 태** · 정 현 열*** · 조 철 우****
(*원광대 컴퓨터공학과, **한남대 정보통신공학과,
***영남대 전자공학과, ****창원대 제어계측공학과)

■ 차 례 ■

요 약

음성인식 및 합성 시스템으로 대표되는 음성입출력장치의 개발 및 실용화에 따라 이들의 성능을 객관적으로 평가할 수 있는 방법이 중요하게 되었다. 그러나 국내에서의 이분야에 대한 체계적인 연구는 많지 않다. 본고에서는 이와 관련한 지금까지의 국내외 연구예들을 검토하고 우리 실정에 맞는 평가법 작성을 위한 방법을 모색하고자 한다. 음성입출력장치의 객관적인 평가법이 확립되면 연구 개발자는 여러방식의 우열을 객관적으로 파악할 수 있고 신뢰성있는 시스템을 효율적으로 개발할 수 있으며 관련제품의 사용자 입장에서는 제품간의 성능비교가 가능하게 될 것이다.

Ⅰ. 서 론

음성을 맨머쉰인터페이스의 수단으로 사용하기위한 연구들이 각국에서 활발히 진행되고 있다. 이에 따라 음성분석, 부호화, 합성, 인식 등의 응용시스템 개발이 활발하고 또한 일부는 이미 상품화 되고 있다. 이러한 시스템의 연구 개발시, 애매한 성능평가기준은 전체 개발의 효율을 떨어뜨리고 최악의 경우에는 기간내 개발이 실패할 수도 있다. 지금까지 주로 사용되고 있는 평가방법으로는 일정한 척도(예를 들어 인식율, 명료도)를 주로 사용하는데 이를 제대로 정의하고 사용하는 것 또한 쉽지않은 문제이다. 음성합성 시스템의 경우에는 어떠한 텍스트가 입력되더라도 이해도와 자연성이 확보된 음성을 만들어내는 것이 목표이다. 따라서 그 과정에서 텍스트해석, 읽기형태의 변환, 발성스타일의 선택, 리듬, 음색, 억양 등의 구현이 종합적으로 잘 이루어져야 한다. 아울러, 단어 이해도 및 음운명료도와 함께 자연성도 확보되어야 한다. 뿐만 아니라 실제 응용에 적합하여야 하며 어떤 사용자에게도 잘 맞아야할 것이다. 따라서 평가항목의 설정 또한 단순하지가 않다. 즉 단순히 한두가지의 척도만으로 성능을 대표하기는 어렵다. 간단한 단어 인식시스템의 경우도 인식 결과중 대치, 탈락, 삽입, 거부 등을 어떻게 볼것인가, 대상이 아닌 단어나 그밖의 잡음에 반응하는 경우는 어떻게 할것인가, 등에 따라 평가 방법은 단순하지않다. 따라서 인식율을 정의한다 하더라도 미리 부수적인 요인들을 함께 정의 해야 한다. 이러한 문제들을 조직적이고 체계적으로 검토하여 객관적인 평가법이 확립된다면 연구자의 입장에서는 각종 방식의 우열을 객관화 할 수 있을 것

이고, 제품개발자의 입장에서는 더 효율적인 개발단계를 거칠 수 있을 것이며, 사용자의 입장에서도 상품의 객관적인 비교, 선택이 가능한 것이다. 이와 같은 관점에서 본고에서는 이 분야의 연구 동향을 소개한다.

Ⅱ. 음성인식(음성입력) 시스템의 평가

다양한 인식방법들에 기초한 인식기들은 개발자 개개인에 의해 독자적인 방법으로 인식실험이 행해지고 성능이 평가되기 때문에, 서로 다른 인식기들간의 객관적인 비교, 평가가 불가능하게 되어 성능이 더 향상된 인식기의 개발을 어렵게 하는 문제점이 있다. 따라서 개발된 인식기들의 성능을 객관적으로 비교하여 상세한 성능의 정보를 진단하고, 나아가 인식기의 최종 성능한계를 예측할 수 있는 평가방법을 개발해야할 필요가 있다. 그리고 평가한 결과를 개발자에게 통고함으로써 개발자 스스로가 더 향상된 성능의 인식기를 개발할 수 있도록 유도할 수 있다.

인식기를 가장 객관적으로 평가하는 방법은 모든 장소에서 모든 계층의 사람들이 모든 상황의 상황(물리적 상황, 정신적 상황 등)에 대해 직접 테스트해서 인식결과를 서로 비교하면 된다. 그러나 이러한 방법은 시간과 경제적인 측면에서 실현이 어렵다. 따라서 이러한 어려움을 극복하기위해 여러계층의 사람들이 여러 환경에서 발성한 음성을 녹음한 공통의 평가용 데이타베이스를 이용한다. 그러나 이것을 사용하여 평가하는데에도 약간의 문제점이 있다. 왜냐하면, 상세한 성능진단을 위해서는 다양한 조건에서 녹음한 음성데이타가 필요한데, 각 조건에 대한 인식기의 성능평가를 위해 그 상황에 알맞는 음성을 일일이 녹음해서 평가하는 것은 매우 힘드는 일이다. 따라서, 대표적인 음성데이타가 수집되어 있을 때, 수집된 음성데이타로써 각 조건에 해당하는 음성을 조작하는 기술이 필요하다. 음성데이타를 조작하게되면, 조작하는 정도에 따라 음성데이타의 성질을 정량화할 수 있고, 인식결과를 평가할 때에 평가척도로 사용할 수 있게 된다.

음성조작에 의해 인식기를 평가하기 위해서는 인식기의 성능에 많은 영향을 미치는 요인들을 모아서 평가항목으로 구성한 뒤, 항목별로 음성데이타를 조작해서 이들에 대한 인식결과를 구하고, 인식결과와 조건들간의 상관관계를 분석해야 한다. 여기서는 먼저 인식시스템의 성능평가관련 국내외 연구현황과 함께 인식시스템의 구성요소, 인식성능에 영향을 미치는 요소, 성능측정기술 및 평가방법에 대해 논한다.

2.1 국내외의 현황

인식시스템의 객관적인 성능평가를 위해서는 interactive한 성능평가 방법, 도구, 그리고 음성데이타베이스의 개발 등에 관한 연구가 필요하다. 이러한 연구는 외국의 경우 연구소, 대학 기업 등이 개별적 혹은 공동으로 방법론을 연구하고 객관적인 평가기준을 마련하여 활용하고 있다. 국내에서도 이 분야의 필요성은 느끼면서도 본격적이고 지속적인 연구는 없었고 단발적이고 산발적인 연구만 조금씩 이루어지고 있다.

• 미 국

미국에서는 음성처리기술의 성능평가법에 대한 필요성을 일찍부터 인식하고 AVIOS(American Voice Input Output Society)나 Speech Tech 등을 통해 연구결과들을 발표하여 왔다. 특히 인식시스템의 성능평가법에 대해서는 1982년에 당시의 NBS(National Bureau of Standard)에 의한 "음성입출력기술 표준화에 관한 워크샵" 및 NRC(National Research Council)에 의한 "가혹한 환경하에서의 음성자동인식에 관한 조사회", 그리고 "음성인식에 있어서의 robust에 관한 회의" 등을 시작으로 미국 음향학회, IEEE의 ICASSP 등에서 관심을 가지고 다루어져 왔다. 특히 NBS(나중에는 NIST, National Institute of Standard Technology)에서는 Dr. Pallet를 중심으로 DARPA 프로젝트의 지원을 받아 음성DB를 포함한 활발한 많은 활동을 전개하고 있다.

• 유 럽

지역적으로 다양한 언어를 가진 여러나라들이 인접해 있는 유럽은 EC통합에 따라 언어간에 원활한 소통이 필요하여 국가간의 대규모 연구가 이루어지고 있고 그중에서 특히 ESPRIT의 SAM 프로젝트는 성능평가법을 집중적으로 연구하고 있다. 또한 ESCA (European Speech Communication Association)는 2년마다 주최하는 Eurospeech를 전후하여 "Speech I/O assessment and speech DB" 워크샵을 개최해 오고 있다.

• 일 본

일본음향학회와 전자정보통신학회가 공동으로 운영하는 음성연구회를 중심으로 관련연구 및 발표가

있어 왔고 1985년에는 음향학회 주최로 "시험용 음성의 표준화"라는 워크샵을 통해 음성합성 및 인식도 포함한 각 분야에서의 실험용음성에 대한 문제점 및 공통화 가능성을 토의하였다. 또, 합성음성의 실용화에 따라 "합성음성 품질평가법"에 관한 워크샵도 있었다. 또한 NTT, KDD 등에서는 오래전부터 음성선송품질을 중심으로한 평가법 및 실용화 연구가 추진되어 CCITT 등에서의 규격 권고작성에 기여하고 있다. 그리고 일본이 주관하여 창설한 ICSLP(International Conference of Spoken Language Processing)에서도 음성DB 및 평가법에 대한 워크샵을 개최오고 있고 최근에는 이를 바탕으로 COCOSDA(Cordinating Committee for Speech Database and assessment)가 창립되어 각국이 연구협력을 강화하고 있다.

• ESPRIT PROJECT 2589 (SAM)

유럽의 ESPRIT project는 음성입출력기술의 성능평가를 1989년부터 본격적으로 과제화하여 종합적으로 수행하고 있다. 이 연구에는 8개국의 28개 연구소, EC내의 6개 연구소, EFTA(유럽자유무역협회)의 2개 연구소가 참여하여 작업환경(SESAM)을 통일화하여 공동연구가를 하고 있다. SAM의 작업환경(SESAM)은 다양한 연구소에서 연구한 결과를 서로 공유해서 공동의 연구를 할 수 있도록 한 것이다.

ESPRIT PROJECT 2589로 수행된 이 세부과제의 정식명칭은 Multi-lingual Speech Input/Output Assessment, Methodology and Standardisation으로써, 음성 입출력 시스템의 성능평가를 interactive하게 지원하기 위한 방법, 도구, 그리고 데이타베이스의 개발을 위한 연구이다. 이 프로젝트의 주요목적은 사용자와 시스템 응용에 있어서 요구사항들이 적절히 반영된 성능평가를 수행하기 위한 강력한 프로토콜의 확립에 있으며, 따라서 잡음효과, 음성환경 및 전송왜곡들이 중요 요소로서 작용되어 실제적인 응용분야에서의 성능평가의 방법들과 밀접한 관계를 갖고 있다.

이 프로젝트는 크게 3단계의 work party(WP)로 나뉘어 진다.

- WP1 : 음성 입, 출력 및 화자인식 모듈의 평가
- WP2 : Interactive 대화 시스템의 평가
- WP3 : 세부구조 및 표준화

WP1에서는 모든 평가방법들을 개개의 음성기술 응용분야에 적용한다. 음성 입력모듈의 평가는 효과적인 성능평가를 하기위한 프로토콜의 설계와 도구을 요구하는 보다 복잡한 시스템의 개발이다. 이들은 중, 대규모의 어휘인식기, 화자나 환경에 적응적으로 대처하는 인식기를 포함한다. 여기서는 화자의 유사성과 다변성, 그리고 이에 따른 실험적 데이타베이스의 분석을 통해 화자인식 시스템의 평가에 대해서도 연구한다.

WP2는 통합시스템과 그 응용분야의 평가이다. 경험과 공식적 두가지 기준을 사용하여 대응적 시스템 내의 대화 모듈에 대한 평가가 이루어진다. 기존 연속적인 음성대응 시스템에 기준하여 전체 시스템 성능평가의 방법과 측정에 대해 연구된다. 음성입력에 관한 환경적 응용요소가 고려되며, 성능의 특성을 규정하기 위해서 해석적 방법과 응용관련 데이타 얻기 위해서 디지틀 모의 실험방법이 사용된다. 마지막으로 관련되는 인간적 요소에 대한 연구가 수행해지며 기존 시스템에 준하여 시험된다.

WP3는 WP1, WP2를 위한 보조수단을 제공한다. 이는 소프트웨어 개발 명세, 워크스테이션과 관련된 도구 명세, 데이타베이스 조합과 수집, 그리고 프로젝트 방법과 도구 보급을 위한 망구성들을 포함한다. 또한 국내적으로나 국제적으로 평가분야에 있어서 표준화를 위한 기초작업을 한다.

연구의 내용 및 방법은 ① 음성 입력 모듈 평가, ② 음성 출력 모듈 평가, ③ 화자 확인 시스템 평가, ④ 대화 모듈 평가, ⑤ 대응 시스템 평가, ⑥ 성능평가에 있어서 환경 및 응용 요소, ⑦ 성능평가에 있어서 인지 요소, ⑧ 도구의 통합구현, ⑨ 워크스테이션 명세 및 개발, ⑩ 데이타베이스 조합, ⑪ 분배, ⑫ 세부구조 및 표준화가 있다.

2.2 인식 성능에 영향을 미치는 요소

음성인식시스템의 성능은 인식 알고리즘과 같은 중심기술 이외에 인식률에 영향을 미치는 여러가지 요인이 있고, 이들을 포함한 평가는 상당히 어렵다. 따라서 인식시스템을 정확하게 평가하기 위해서는 인식률에 영향을 미치는 모든 요인을 조사해서 항목별로 분류한 뒤, 각 항목을 평가해서 전체 인식률과의 상관관계를 조사하면 된다. Lea의 "What causes speech recognizers to make mistakes?" 라는 논문에서 80가지 이상에 달하는 많은 요인들을 제시한 바 있다. 인식시스템의 인식 정확도에 영향을 주는 원인은 다음의 세가지 변동성으로 설명할 수 있다.

- 내부 화자 변동성(intra-speaker variability)
- 상호 화자 변동성(inter-speaker variability)
- Context variability

이러한 변동들은 인간요인(Human factors), 언어요인(Language factors), 채널과 환경요인(Channel and environmental factors), Task요인, 알고리즘요인(Algorithmic factors), 성능과 응답요인(Performance and response factors) 등과 관련이 있다.

위의 요인들 외에 많은 요인들이 있다. 이러한 요인들을 모두 평가대상으로 선정해서 각각에 대한 영향을 조사하여 전체적인 인식기 평가를 해야함이 당연하나 이렇게 하려면 많은 연구인원과 시간, 자본이 투자되어야 한다. 따라서 전체를 모두 평가항목으로 성할 것이 아니라 인식률에 가장 영향을 많이 미치는 요인들을 선정해서 이들 항목을 우선적으로 평가해야한다.

Lea가 제시한 여러 요인들 중에서 인식률에 특히 영향을 많이 미치는 것들은 크게 세 분야로 생각할 수 있다. 이것들은 인간적 요인, 언어요인, 그리고 환경요인이다. 따라서 세가지 각 요인에 대한 세부적인 평가항목을 선정해야 한다. 인간적 요인에는 발성속도, 성별, 방언, 성도특기, 발음습관, 화자의 정문스펙트럼, 건강상태, 심리적 상태와 스트레스 등이 있고, 언어요인으로는 어휘수, 단어간의 음향적 유사성, 조음정도, 억양, 강세, 운율의 패턴 등이 있다. 그리고 대표적인 환경요인으로서는 노이즈 레벨(SNR), 노이즈의 형태(White, pink, tonal, impulse) 등이 있다.

2.3 성능 측정의 척도 및 평가 방법

가. 퍼센트 인식률(또는 에러율)

인식능력의 척도로 가장 많이 사용하고 있는 기술로서 인식정확도(Recognition Accuracy)와 대체 에러율(Substitutionary Error Rate)로 성능을 표현한다. 이러한 종류의 측정들의 단점은 주어진 테스트에서 나온 에러수를 제외한 다른 성능에 관한 정보는 없다. 즉 에러분포를 계산할 수 없고 다른 에러보다 중요한 에러를 판단하는 규정이 없다. 또한 에러율은 어휘크기에 영향을 받지만 어휘 크기나 어휘 어려움(difficulty)에 대한 성능 비교가 불가능하다.

퍼센트인식률이나 에러율은 고립단어 인식과 연속음성 인식의 경우를 나누어서 계산해야 한다. 고립단어 인식의 경우에 에러율 계산은 틀린단어에 의해 정해단어(Correct word)가 대체된 대체(substitution)에러수만 구하면 된다. 그러나 연속음성의 경우에는 세 종류의 에러-대체(substitution), 탈락(delection), 삽입(insertion)-가 발생한다. 대체와 탈락은 분명히 에러임을 알 수 있지만 삽입을 에러로 계산해야 하는지의 여부는 분명하지 않다. 초기에는 이러한 삽입을 에러로 간주하지 않았지만 최근에는 에러로 계산하는 추세이다. 기대하지 않았던 단어가 삽입된 경우에는 분명히 에러로 계산해야 되기 때문이다. 따라서 삽입에러의 의심스러운 특징들을 분명히 하고 모든 시스템에 대한 비교를 가능하게 하기 위해서 두 종류의 에러율(인식률)을 생각할 수 있다. 즉, 삽입을 에러로 간주하지 않는 경우와 에러로 간주하는 경우 모두에 대해서 인식률(에러율)을 구하면 된다. 전자의 경우를 Percent Correct라 하고, 후자의 경우를 Word Accuracy라 한다.

인식의 정확도(Recognition accuracy)를 결정하기 위해 맞는 단어열에 대한 인식단어열을 정렬(align)시키고, 그 다음에는 정답, 대체, 탈락, 삽입 단어수를 계산한다. 이러한 정렬(alignment)은 NBS(National Bureau of Standards)에서 제공한 Dynamic programming을 사용해서 얻을 수 있다.

나. 혼동 매트릭스

이것은 퍼센트인식률의 한 단점을 보완할 수 있는 기법인데 에러가 어휘 전체에 퍼져있는지, 단지 몇개의 항목에만 집중되어 있는지를 알 수 있기 때문에 유용한 진단적인 정보를 제공할 수 있다. 혼동 매트릭스의 각 성분 M_{ij}는 단어(혹은 음소 등)i의 100개의 테스트에 대해서 단어 i가 j로 인식된 횟수의 평균치 추정이다.

다. 상대 정보 손실도(Relative Information Loss : RIL)

혼동 매트릭스로부터 유도된 정보이론적인 측정법으로서 성능측정을 정의하기 위해 엔트로피를 사용한다. 이것은 1982년 Woodald와 Nelson이 제안한 것으로서 H(X|Y)/H(X)를 RIL(상대정보손실)로 정의하였고, 애매도(equivocation)에 대한 함수로써 기울기가 1/H(X)이고, 원점을 지나면서 0에서 1까지의 값을 갖는 RIL함수이다. 장점은 에러의 분포를 고려할 수 있고, 속도 왜곡 모델(Rate distortion model)과 관련되어 사용할 때에는 음성입력 시스템(Voice entry system)에 개개의 에러값을 반영할 수 있다. 그러나 이러한 측정의 문제점은 인식기들이 다른 어휘에 테스트될 때 비교평가를 할 수 없다는 것이다.

라. Human Equivalent Noise Ratio(HENR)

1977년 Moore는 어휘의 난이도(difficulty)를 고려한 측정법을 기술했다. 이것은 인간의 단어 인식 처리

모델에 기초를 두고 있는데 임의의 노이즈 레벨에서 임의의 입력단어 집합에 대한 혼동 매트릭스를 예측할 수 있다. 즉, 같은 단어집합에 대해 같은 인식률을 얻기까지 모델의 노이즈레벨을 변화시키므로서 인식기의 성능을 dB. H로 표현하는데 이러한 측정을 HENR이라한다. 즉, 정해진 인식률에 대해 SNR레벨이 낮을수록 성능이 좋은 인식기로 판정할 수 있다.

마. EVC(Effective Vocabulary Capability)

1980년 Martin Taylor가 제안한 방법으로서, 허용할 수 있는 에러율(tolerable error rate)에 대해서 인식기가 조절할 수 있는 최대 어휘크기를 측정하는 기법이다. 이러한 측정을 계산하기 위한 기술이 아직 연구중이다.

바. Perplexity

어휘의 난이도에 대한 척도로서 주로 문법에 의해 강요된 제한의 측정이나 주어진 문법의 불확실성 레벨의 측정을 복잡도(Perplexity)로 구한다. Perplexity를 정의하기 전에 먼저, 인식하는 동안 문법이 어떻게 불확실성(uncertainty)을 감소시키는지를 고려해야한다. 문법이 없으면, 전체 어휘는 각 결정점에서 고려되어야하지만, 만약 문법을 사용하는 경우에는 고려대상으로부터 많은 후보를 제거하거나 다른 후보들보다 더 확률이 높은 몇몇 후보를 지정할 수 있다. 결정점(j)에서의 이러한 제한은 엔트로피(H)나 최적엔코딩기법(Optimal encoding scheme)을 사용해서 다음단어를 명시하는데 필요한 비트(bits)수로 측정할 수 있다.

사. Recognizer Sensitivity Analysis(RSA)

RSA는 음성의 변동효과를 해석하기위해 개발된 방법이다.

RSA방법은 인식성능에 에러를 유발하는 음성의 변동성이 측정가능한 소수의 파라메타들에 의해 특성화될 수 있다는 전제에서 시작된다. 파라메타들과 인식성능과의 관계를 결정하므로써 간단한 스코어링 방법으로서는 불가능한 인식기의 기본동작에 대한 고찰이 가능하다.

아. Recognition Assessment by Manipulation Of Speech(RAMOS)

RAMOS는 음성인식기의 성능을 음성발성 파라메타와 음성전송 파라메타의 변동함수로서 측정해서, 내부화자, 상호화자, 스트레스의 영향, 노이즈(SNR, 노이즈형태) 등의 요인에 관한 성능을 평가하는 시스템이다. RAMOS는 최소차이(minimal-difference) 단어집합으로 된 CVC형태의 소수의 단어 데이타베이스를 사용한다. 평가방법은 자연음성(natual speech)에서 관찰된 물리적 파라메타의 변동이나 다양한 환경조건하의 화자들의 변동에 해당하도록 단어들을 Analysis-Resynthesis 방법으로 조작하는 것이다. 먼저 관련 파라메타의 변동을 정의하기 위해 대표적인 음성토큰들을 해석한다. 다양한 조건에 대한 음성발성 파라메타의 변동을 통계적으로 정량화하기 위해 대표적인 음성토큰들을 녹음한다. 특히 데이타 수집시 고려할 사항은 내부화자변동, 상호화자변동, 남/여 음성의 변동, 물리적인 스트레스의 영향 등이다.

음성발성 파라메타 변동과 음성전송 파라메타의 변동에 관한 조작이 끝나면, 조작된 음성을 인식기에 테스트함으로서 인식결과를 얻는다. 이때 인간의 인식성능과 비교를 위해 청취를 위한 전문인을 두어 인식실험을 병행할 수도 있다. 인식된 결과는 MDS(Multi-Dimensional Scaling)분석을 위해 혼동 매트릭스 형태로 나타낸다. MDS분석은 데이타 중에 숨겨진 구조를 찾아내어 그 구조를 소수 차원의 공간에서 기하학적으로 표현한다. 즉 데이타에 포함된 정보를 추출하기위해 MDS를 탐색수단으로 사용하여 다양한 입력조건에 대한 인식기의 동작을 연구하기 위한 방법이다.

2.4 종합적인 평가 방법

음성인식기의 성능을 종합적으로 평가하기 위해서는 성능에 영향을 미치는 요인들을 모아서 평가항목으로 구성한 뒤, 항목별로 음성데이타를 조작해서 이들에 대한 인식결과를 구하고, 인식결과와 조건들간의 상관관계를 분석해야 한다. 이 때 평가항목 뿐만이 아니라 평가할 인식기의 범위도 정해야하는데 크게 상용화된 인식기와 컴퓨터에 인식을 시뮬레이션하는 경우로 나누어 볼 수 있다.

상용인식기의 경우 사용목적에 따라 입력되는 음성의 종류가 다르기 때문에 목적 타스크에 무관하게 강제적으로 공통음성을 사용하여 평가하는 경우와 목적 타스크의 음성에 맞추어서 평가하는 방법이 있다.

목적 타스크에 무관하게 평가하는 경우는 숫자나 단모음 등의 가장 보편화된 음성을 공통음성으로 선정하고 이러한 음성을 평가항목에 따라 적절히 조작하는 조작기가 있으면 평가가 가능하다. 그러나 목적

테스크의 음성에 맞추는 경우는 인식기를 개발할 때 사용한 음성을 이용해서 조작하는 정도를 동일하게 해 주면 평가가 가능하다. 그리고 최종적으로 인식결과를 상세히 분석하기 위해서는 혼동행렬(Confusion Matrix)형태로 결과를 나타내어야 하기 때문에 마지막 출력단 전에 혼동행렬로 결과를 나타낼 수 있는 처리과정(결과해석기)이 추가 되어야 한다. 물론 상용인식기의 경우에는 인식성능뿐만이 아니라 처리속도, 사용의 편리함 등을 평가항목에 추가시켜야 할 것이다.

인식기의 성능의 평가에는 다음과 같은 방법들이 고려되고 있다.

(1) 일반적인 데이타베이스 사용

(2) 표준시스템(Reference system)에 기초한 방법

(3) 특별한 Calibrated Database 사용

(4) 특별한 어휘에 기초한 진단적인 방법

(5) 인공의 테스트신호 사용

(1)은 가장 흔히 사용하는 방법으로서 대표적인 조건하에서 수집된 음성데이타베이스를 사용한다. 그러나 이것은 특별한 파라메타들의 변동에 대한 제어가 불가능한 단점이 있다.

(2)는 인간이나 표준인식기(Reference Recognizer)에 기초한 방법이다. 여기에 해당하는 방법들로서는 앞에서 소개한 HENR방법, Reference recognition algorithm, EVC 등이 있다.

(3)은 RSA방법으로 Calibrated databases의 집합을 사용하는 방법이며 각 데이타베이스는 인식기의 성능에 영향을 끼치는 환경조건에서 수집된 것이다. 이 방법은 인식성능의 에러에 영향을 주는 음성변동성이 소수의 측정 가능한 파라메타들로 특성화될 수 있다는 전제에 기초를 두고 있다.

(4)는 RAMOS방법이다. 이것은 대표적인 음소들 간의 Confusion 해석으로부터 Reference 조건들의 집합에 관련된 다차원표현을 얻는 방법이다. 이 방법은 좀 더 일반적인 방법, 즉 특별한 음성발성파라메타와 음성전송파라메타의 변동함수로서 인식시스템이나 인식알고리즘의 성능을 명시하는 것이 목적이다.

(5)의 인공(non-speech)테스트신호에 기초한 평가방법은 아직 알려져 있지는 않지만 미래의 주요한 평가방법이 될 것이다. 즉 많은 응용조건들에 대한 공통의 데이타베이스 구축이 상당히 어렵기 때문에 각 조건에 대한 타당성 있는 인공의 테스트 신호를 만들 수 있다면 더욱 더 광범위한 진단적 정보를 얻을 수 있다.

Ⅲ. 합성음성시스템의 평가

음성합성의 최종적인 단계는 규칙음성합성 시스템 또는 문자로부터 음성을 합성할 수 있는 시스템이라고 볼 수 있다. 우리말의 규칙합성에 관한 연구결과가 여러가지 문헌을 통하여 제시되고 있고 상용 시스템까지 등장한 상태이나 이러한 시스템들이 어느 정도로 명료도와 자연성을 갖고 있는지 평가가 된 경우는 거의 없는 실정이며 평가를 수행한 경우도 단지 단어나 음절수준의 인식도를 측정한 경우가 대부분으로 합성기 개발과 성능향상을 위한 진단평가나 합성기의 선택을 위한 비교 척도로서의 평가는 아직 이루어지지 못하고 있으며 지금까지 발표된 합성기에 관한 소수의 평가 마저도 평가과정이나 사용한 데이타베이스가 알려져 있지 않다. 이는 국내의 여건상 지금까지 합성기자체의 개발에 치우친 점이 있어 제대로 된 평가를 할 이유가 없었고 합성음성 평가에 관한 자료를 쉽게 구할 수 없었던 사정도 있다. 그러나 음성합성기의 개발 기술과 합성을 위한 음성분석기술이 어느정도 축적된 지금은 개발자의 입장에서 볼 때 합성음의 명료도와 자연성의 개선을 위하여 보다 손쉽게 행할 수 있는 체계적인 방법이 요구되고 있으며, 차후 합성기의 수요가 다양한 분야로 늘어날 것에 대비하여 구매자의 입장에서도 각 합성기간의 성능을 비교해 볼 수 있는 척도가 필요한 실정이다. 외국의 경우는 DECTalk 등과 같은 뛰어난 성능의 합성기 개발과정에서 여러가지 평가방법이 사용되어 왔고 이를 비교대상으로 현재 개발중인 합성기의 성능평가도 이루어지고 있다. EC의 경우 Esprit계획의 하나로 음성 인식과 합성시스템 평가기술에 대한 대규모의 공동연구가 여러 국가간에 공동으로 이루어지고 있다. 이 기사에서는 지금까지의 외국의 합성음성평가에 관한 연구내용 및 향후 연구방향과 함께 필자 등이 제안한 무의미 낱어에 의한 합성음 평가법을 소개한다.

3.1 합성음 평가의 제요소

합성음 중에서도 규칙합성에 의한 합성음 또는 문자를 음성으로 변환하는 시스템에 의한 합성음의 평가는 평가 대상이 되는 말의 단위에 따라서 혹은 평가자의 관점에 따라서 여러가지 단계로 나누어 수행할 수 있다.

- 평가의 대상 주체에 따라
  - 주관적 평가
  - 객관적 평가

- 평가목적에 따라
  - 시스템 개선과 진단을 목적으로 한 경우
  - 시스템 출력(합성음)의 응용(실용화)를 위한 경우

- 평가단위에 따라
  - 음소단위
  - 음절단위
  - 단어단위
  - 문장단위

- 평가 대상 관점에 따라
  - 이해도
  - 명료도
  - 자연성

- 평가부분에 따라
  - 문장처리부-문자변환부의 평가
  - 분절음-음소, 단어수준의 명료도 평가
  - 초분절음-운율, 억양, 휴지기의 변화 등을 평가

- 기타
  - 전체음질의 평가
  - 언어학적 심리학적 평가
  - 현장시험-실용환경에서의 음질평가, 반응측정

위와 같이 평가기준에 따라서 여러가지 평가 방법이 나올 수 있으며 어떤 기준으로 평가를 하는가에 따라서 필요한 합성음 DB도 달라진다. 합성음 평가에 관한 연구는 크게 필요한 데이타베이스를 정의하는 부분과 평가과정을 정규화하고 평가를 도와주는 도구를 작성하는 부분으로 나눌 수 있다. 이중 데이타베이스를 작성하는 일은 상대적으로 더 중요하고 많은 투자를 필요로 하는 작업이다. 평가작업이나 데이타베이스 작성작업은 모두가 상당한 시간과 비용 노력을 요구하는 작업이므로 가능한 한 일을 효율적으로 수행할 수 있도록 해 주는 도구가 미리 작성되어 있으면 작업의 부담을 덜어줄 수가 있다. 외국의 경우는 DB의 단어군을 자동으로 발생시키고 평가 결과를 분석하여 점수를 매겨주는 시스템을 개발하여 사용하고 있는 경우도 있다.

3.2 국내외 동향

가. 외국의 동향

표 1은 규칙합성기의 음질평가의 사례의 일부를 보여준다.

음성인식 및 합성기의 평가법의 표준화에 관하여는 CCITT의 스터디그룹 5/12에서 다루고 있다. 아직 합성음의 평가에 관하여는 논의가 진행되고 있는 상태이며 권고안이 나오지 않고 있으나 지금까지 다음과 같은 사항들을 대상으로 논의하고 있다.

① 어떤 파라메타가 측정되어야 하며 어떻게 측정되어야 하는가.
② 이런 파라메타들의 어떤 값을 권고할 것인가.
③ 어떠한 데이타베이스가 표준화 되어야하는가.
④ 언어간의 차이에서 오는 문제점들을 어떻게 다룰 것인가.

CCITT에서 검토된 내용으로는 주관적인 평가법에 있어서의 자료의 준비, 자료의 제시, 실험과정의 계획, 청취조건, 의견 평가용 설문지의 형식 등에 관한 내용을 다루었고 객관적인 평가법에 관하여는 시간-주파수 평면에서 합성음의 자연성에 관한 객관적 평가법을 제시했다.

1993년 3월에 권고안 초안으로 제시된 P.85에서는 음성출력장치의 주관적 성능 품질평가에 관하여 다루고 있는데 내용은 다음과 같이 구성되어 있다.

① 시험의 준비 단계
② 실험의 설계
③ 결과의 통계처리와 보고

이 안에서는 음성의 전송품질을 다룬 권고안 P.80의 내용을 기준으로 하여 특정 응용분야에 따른 전체적인 시스템의 성능평가에 관하여 다루고 있다.

나. 국내현황

국내에서의 합성시스템의 평가에 관한 전문적인 연구는 아직 미미하고 주로 명료도에 관한 몇몇 연구예와 합성시스템 개발자들에 의해 자기가 개발한 시스템에 대한 일부 평가결과가 있을 뿐으로, 이 분야의 체계적인 연구가 필요한 시점이다.

명료도 측정법은 통화품질의 측정을 목적으로 체신부산하 전기통신연구소의 연구를 시작으로 건축분야의 실내음향측정용의 연구예가 있었고 최근 한국전자통신연구소에서 학계와의 공동연구예가 있으나 아직 통일적인 방법으로 두루 받아들여지고 있지는 않다. 최근 필자 등이 무의미 단어에 의한 규칙합성음의 평가에 대한 제안 등을 비롯하여 이분야의 연구가

표 1. 규칙합성기의 음질 평가에 사용된 여러가지 방법들

시험의 종류	비 고
CV 음절	Nusbaum et al. '84
PB-단어표	Schwab et al. '83
DRT 2 가지중 선택	Pratt '86
RDT(rhyme and diskll. test)	Zhou '86
MRT 6 가지중 선택	Schwab et al. '83, Greene et al. '84 Pisoni & H'80, Bernstein & P '80 Pisoni & K'82, Nye & G '73 Graillot et al. '83
MRT 초성/종성 혼합	Greene et al. '84
MRT 개방형응답	Greene et al. '84, Pisoni & H '80 Pisoni & K'8
CVC 무의미 단어	Pols & Olive '83, v. Bez. & P '87
VCV 무의미 단어	v. Bezooijen & P '87
CVVC와 VCCV 무의미 단어	Pols et al. '87
Havard 문장	Schwab et al. '83, Greene et al. '84 Pisoni & H '80, Bernstein & P '80 Nye & G '74
Prose comprehension	Pisoni & H '80, Carlson et al. '76, Bernstein & P '80, Luce '81
Surface properties	Luce '81
Word recall, free	Luce et al. '83
Word recall, serial	Luce et al. '83
+ load	Luce et al. '83
Identify words 4 altern.	Greene & P '82
Repeat digit strings	Greene & P '82
Recognize letters & cons. in words	Laddaga et al. '81
Understand(proper)names	Larreur & S '77, Spiegel '85
Lexical decision	Pisoni '81 ; '82
Word/non-word naming	Pisoni '81 ; '82
Word monitoring	Nusbaum et al. '83
Recall gist of simple story	Jenkins & Franklin '81
Subjective ratings, questionnaires	Nusbaum et al. '84
Noise masking	Pisoni & K '82, Bareri & G '86
Variable speech rate	Luce et al. '83, Carlson et al. '76
Interactive rate variation	Nusbaum et al. '83

기지개를 펴기 시작하고 있다.

합성시스템 평가에 관한 내용에서 보면 대부분의 경우 사용된 대상 단어나 문장을 밝히지 않고 어떤 경우는 피상적인 자기 평가 결과로 그친 경우도 있다. 그리고 기업체의 연구기관이나 상용화된 제품의 경우는 시험결과를 전혀 제시하지 않고 있어서 시스템의 객관적 주관적 평가가 힘든 실정이다. 또한 일부 평가결과를 제시한 시스템도 그후의 개선작업 등에서 추가적인 평가를 수행한 결과가 제시되지 않아 초기의 평가결과가 어떻게 시스템의 개선에 이용되었는지도 알려져 있지 않다. 즉 한국어 합성기가 여러곳에서 연구되고 있고 상용 합성기까지 출현한 상태이나 아직도 합성음의 평가수준은 초보적인 단계에 머물러 있다. 단어단위의 이해도 시험이나 제한된 음소단위의 시험을 수행된 적이 있으나 진단을 위한 시험이나 성능평가를 위한 비교실험을 수행된 적이 없다.

3.3 합성음 평가의 연구방향

합성음의 평가작업은 우선 공통의 평가절차와 데이타베이스의 확립이 선행되어야 하며 이와 관련된 도구들이 개발되어 규칙합성시스템을 연구하는 팀들에게 제공될 수 있어야 한다. 이러한 배경하에서 실제의 평가는 합성기를 개발하는 측에서 직접 수행하는 것이 가장 효율적인 방법이라고 생각된다. 이렇게 함으로써 합성기 개발의 정도를 측정할 수 있게 될 것이다. 출력음성에 대하여는 다른 음성전문가에 의한 공통적인 비교평가도 가능하겠지만 분장처리부와 같

이 외부에 세부적으로 공개되지 않는 부분에 대하여는 일정한 기준하에 자체평가를 수행하는 방법밖에 없다고 생각된다. 그러므로 이러한 부분에 대하여는 공통적으로 사용할 수 있는 DB와 평가 도구를 이용하여 자체적으로 평가를 행하고 결과를 비교해 볼 수 있는 방법이 필요하게 된다.

가. 주관적 평가법 연구

인간의 귀는 가장 좋은 음성인식기이므로 합성음의 평가에 있어서도 가장 좋은 평가도구라고 볼 수 있다. 주관적인 평가법은 합성음성분만 아니라 통신선로 및 통화품질평가 등에서도 사용되어 온 방법이다. 일반적으로 일정한 구조를 갖는 단어나 문장을 들려주고 청취자로 하여금 받아 쓰게하는 방법이 사용되는데 이 경우는 적절한 음성단위별 음성DB의 작성과 청취조건의 확립에 관한 연구가 있어야 한다. 음성DB의 작성은 평가대상에 따라 음절, 문장 등 다른 형태의 DB가 필요하므로 이러한 단어군이나 문장의 선택이 중요하다. 또한 이러한 음성DB군의 작성에는 사용목적에 따라 필요한 음성학적 특성을 가져야 하므로 음성학적 전문가의 참여가 필요하다.

나. 객관적 평가법 연구

대개의 합성음 평가는 시험단어음성을 청취자들에게 들려주고 받아쓰게 한 것을 평가하는 주관적인 청취실험에 의한 방법을 사용하고 있다. 이와 같이 인간이 들어서 판단하는 것은 가장 정확한 평가방법이지만 대상자의 선택과 시험조건에 따라서 결과의 편차가 있을 수 있다. 이와 같은 문제점을 해결하는 한 방법으로는 수학적인 방법에 의한 객관적인 평가법을 들 수 있다. 음성의 자연성, 명료도 등을 표시해 주는 파라메타를 수학적인 방법으로 찾아내어 분석하는 것이다. 객관적인 평가법이 개발되면 번거로운 시험평가 절차를 거치지 않고도 음성을 평가할 수 있기 때문에 무척 편리한 방법이 될 것이다.

그러나 어떤 파라메타가 원하는 특성을 나타내는지 알 수가 없기 때문에 아직 행해진 예가 많지 않다. 이와 같은 객관적인 평가는 한편으로는 궁극적인 음성인식의 문제와 연관이 되기 때문에 상당히 어려운 문제이다. 지금까지 수행된 사례에서는 음성의 크기 주파수 및 시간평면의 파라메타의 관계 등을 장시간 통계를 구하여 자연음성의 그것들과 비교하여 자연성을 판정하려는 시도가 있었는데 아직 이러한 방법은 절대적인 방법이 되지 못하고 주관적인 평가법의 보조 수단으로 사용될 수 있는 정도이다. 그러나 객관적 평가와 관련한 파라메타를 찾기 위하여 자연음성과 합성음성의 여러가지 파라메타에 관한 통계적 특성을 파악하려는 노력이 계속되어야 할 것이다.

한국어의 합성음성평가에 있어서는 정규화되고, 제어된 방법에 의한 시험이 실시된 예가 거의 없으므로, 우선 시험종류별 단어표 구성과 문장 DB의 작성 그리고 시험절차를 확립할 필요가 있다. 비록 현재 발표된 합성기들의 음소명료도는 양호하다고 주장하고 있고 실제로 상당히 양호한 성우도 있으나 이를 객관적으로 판단할 수 있는 기준이 없는 상태이므로, 우선 가장 기본적인 명료도 시험에 관한 연구가 수행되어야 하며, 이후 자연성 등에 관한 평가절차를 확립하여 합성음성의 음질개선을 위한 작업을 해야 할 것이다.

그리고 우리나라와 같이 합성에 관련한 음성분석 및 언어처리기술이 개발단계에 있는 경우는 합성음성평가 작업이 음성합성기술의 개선에도 도움을 줄 수 있도록 진단형 평가기술을 확립하여 개발과 음질개선을 해가면서 전체적인 품질평가가 수행되어야 할 필요가 있다.

3.4 무의미단어에 의한 합성음성 평가방법의 예

분절음 단위의 명료도 시험을 위하여는 실제의 단음절 단어를 합성하여 들려주고 6개 또는 2개의 보기 중에서 선택하게 하는 MRT나 DRT방법을 많이 사용하여 왔으나 그러한 방법들이 갖고 있는 몇가지 장점에도 불구하고 다음과 같은 문제점들때문에 무의미 단어를 이용하여 명료도 평가를 하려는 시도가 있다.

무의미 단어에 의한 시험법에 관하여 알아보기 전에 우선 지금까지 여러 곳에서 사용되어온 MRT나 DRT의 방법과 문제점을 살펴보면, MRT법은 여러개의 운이 다른 음절 단어를 보기로 6개정도 제시해 주고 그 중에서 들은 음성을 선택하도록 하는 방법이고, DRT법은 MRT와 같으나 두가지의 보기중에서 선택하도록 하는 방법이다. 지금까지 알려진 MRT와 DRT방법의 단점으로는 다음과 같은 점을 들 수 있다.

1) 피험자가 단어의 뜻을 예측할 수 있으므로 음소만의 명료도를 측정하기가 어려운 점이 있다. 즉, 빨리 학습할 수 있으므로 나름대로 단어를 추측하도록 할 수 있다. 이런 이유로 시험결과에서 개방 시험의 경우 즉, 보기를 주지 않고 했을 경우 오류율이 커진다.

2) 원래 통신채널의 시험방법으로 개발된 것으로써, 대역통과성 잡음 마스킹 등과 같은 아나로그 통신

채널 시험용으로 만들어진 것이므로 합성음의 경우에는 적합하지 않을 수가 있다.

3) 실제 단어군에서 일어날 수 있는 음소간의 혼동현상이 선택된 단어군에 포함되어 있지 않을 경우 그 현상을 검출하지 못한다.

4) 각 언어에 대해 다른 시험법이 개발되어야 한다.

5) 모든 음소결합이 포함되지 않을 수 있다.

반면에 장점으로 훈련되지 않은 청취자에게 적용하기가 쉽고, 점수 내기가 쉽고, 사용 단어수가 적은 사실 등을 들수가 있다.

MRT 방식에 의한 규칙합성시스템의 평가사례를 보면 다음과 같다.

MITalk-79, Prose-2000, DECtalk, Berkely, Infovox Prose 3.0, Votrax Type'n Talk, Echo

CNET-IPG(독일어 : VCV단어 55세트이용)

DRT방식에 의한 규칙합성시스템 평가사례는 다음과 같다.

Pratt(1986, 1987)

DECtalk, Prose2000 v1.2, Call Text, Infovox, TI-Speech, Namal Type & Talk Speech, CNET(불어 : 111개의 단어쌍 이용),

이러한 단점을 고려하여 무의미단어에 의한 명료도 시험이 시행되고 있는데 무의미단어에 의한 명료도 시험의 사례는 다음과 같은 것들이 있다.

①Pols와 Olive(1983), Olive(1977, 1980)

Diphone 규칙합성시스템의 평가에 CVC 무의미단어 사용

PCM부호화된 자연음성, LPC코우딩된 재합성음을 피험자 33인에게 규칙합성음과 같이 들려준다. 결과를 기반으로 음소표현과 규칙합성의 구조향상에 기여한다.

②Pols et al. (1987)

불어 diphone 규칙합성시스템 평가측정, 1250개 diphone의 진단평가용, CVVC, VCCV형태의 무의미단어 사용, LPC코우딩된 재합성음과 자연음성 같이 제시, 4명씩 2집단의 피험자 참여

③Bezooijen과 Pols(1987)

네덜란드어 규칙합성시스템의 프로토타입 평가

④Rolf Carlson, Bjorn Granstrom, Lennart Nord(1992)

KTH(Royal Institute of Technology) 합성기 평가

무의미단어에 의한 평가에 의해 얻을 수 있는 결과로는,

1) 합성음에서 음소의 명료도 판정

2) 분절음 수준에서 합성기의 진단을 위한 자료 획득

3) 필요한 보다 정확한 음소환경을 제시해 줄 수 있다.

4) 실제 사용하는 단어군보다 청취자의 학습효과가 적어 보다 정확한 평가가 가능하다.

무의미단어에 의한 평가에서의 문제점은,

1) 각 언어에 적합한 음절구조 즉 CVC, VCV, CV, VC 또는 음소균형이 있도록 만들어진 무의미 단어표등을 만들어내기가 어렵다.

2) 음소의 조합으로 CVC, VCV, CV, VC 등을 시험할 경우 제시되는 단어의 수가 많아진다.

3) 합성시스템을 구성하는 음성의 기본단위에 따라 평가시 제시한 음성의 단어구조도 바뀌어야 한다.

무의미단어를 이용한 명료도 측정을 위해서, 한국어 발음사전의 빈도조사 통계로부터 음절구조의 빈도분포를 살펴본다. 모든 종류의 음소결합의 경우를 다 포함하여 시험하기는 거의 불가능하므로 빈도수가 많은 음절의 유형을 찾아 평가에 사용한다면 시험결과의 유용성이 클 것이다.

1987년도에 수행된 한국어 발음사전에 관한 단어의 발음에서의 빈도조사를 보면 우리말의 단어는 5, 6, 7개 음소의 구성을 갖는 것이 총 단어의 66.87%로서 많은 부분을 차지하고 있으며 다른 1음소-30음소까지의 경우의 개별 종류에 비하여 상대적으로 고빈도를 기록하고 있다. 같은 조사에서 보면 모음의 출현빈도는 아, 어, 이, 오, 우의 순으로 출현빈도를 기록하고 있는데 아, 이, 오 만으로도 전체 모음 빈도의 53.8%를 차지하고 있으며 아,어,이,오,우까지 합하면 75.8%가 된다. 이는 모음들은 자음과의 결합면에서도 다른 모음보다 빈도가 앞서므로 모음을 재합할 경우 빈도수가 많은 모음들을 위주로 택할 수 있다.

음절의 구조면에서는 13가지 음소결합 종류로 나누어 볼 때 CVC가 가장 많고(34. 25%), CCV(또는 VCC) 18.20%, #CV 13.53%의 순으로 빈도를 가지므로(이 세가지가 65.98%를 차지) 이러한 구조의 무의미 음절을 사용할 경우 유리하다고 볼 수 있다. 그러나 이것이 절대적인 단어를 만드는 기준이라고는 볼 수 없다. VC#의 경우는 종성자음을 포함하므로 반드시 포함되어야 할 것이고 합성기의 음성단위와 측정하고자 하는 성질에 따라 적절한 단위가 선택되어야 할 것이다.

음성의 제시과정을 정규화 하는 것은 음성평가에 중요하다. 청취환경이 다를 경우 청취자의 인식율이 크게 달라질 수 있기 때문이다.

앞서진 예에 의해 청취시험자들에서 제어되어야 할 변수로는, 청취자의 수, 대상음성의 종류, 합성기

의 종류, 같이 제시하는 음성의 종류-자연음, LPC부호화된 음성, 잡음의 혼합여부 등이 있다.

필자 등은 지금까지 합성음성의 평가에 많이 사용되어온 MRT, DRT의 단점을 고려하여 무의미단어에 의한 합성음성 평가법에서의 장단점, 문제점을 비교, 무의미단어의 작성 및 제시과정 등을 검토하여 한국어 규칙합성의 명료도를 진단적으로 평가하는 방법을 제안한 바 있다.

이 제안에서는 먼저 음소환경을 고려한 다음절 무의미단어를 컴퓨터 프로그램에 의해 자동작성하는 방법이 제안되었다. 유럽의 SAM계획에서 수행된 것과 같은 기존의 무의미단어 발생기에 관한 연구에서는 주로 단일 음절 무의미단어를 대상으로 명료도를 측정하여 왔다. 독일어, 불어, 이태리어 등에서 사용된 이와 같은 방법들은 통계적으로 구해진 음소나 음소환경의 출현 빈도수를 바탕으로 시험에 사용할 무의미단어를 발생시켜 평가에 사용하는데, 평가에 사용되는 음소환경의 수가 합성기음성에서 발생할 수 있는 여러 음소환경의 경우를 다루기에 충분하지 못하다. 그리고 많은 환경을 포함하기 위해서는 단어의 수가 많아지게 되어 평가시험을 수행하는데 피험자의 협조성 상실 등의 현실적인 어려움을 초래할 수 있다. 이와 같은 어려움을 개선하기 위하여 여기에서는 자연어와 유사한 음소환경의 비를 갖는 다음절 무의미단어를 자동 작성하였다.

다음절 무의미단어를 작성하는 방법은 다음과 같이 두가지로 접근할 수 있다. 먼저 음절수를 제한하고 음소환경의 가지수를 택하는 방법이 있다. 이 방법은 자연단어의 평균음절수가 3-4음절이고 음절수가 많을 경우 효과적인 시험이 어려워질 것이라는 가정하에 택하는 방법이다. 그 다음 음절수를 제한하지 않고 모든 음소환경의 종류수가 전체적으로 일정 비율에 달할 때까지 단어를 발생시키는 방법이다. 후자의 경우에도 시험의 효율성을 위해 50-150개정도로 단어군을 만드는 것이 바람직하다고 생각된다. 본 연구에서는 전자의 방법을 택했다. 그리고 각 음소환경의 수를 모두 포함하더라도 수백개 이상이 되면 실제 실험에의 적용시 효율성이 떨어지므로 단어수를 제한하고 환경수의 비를 자연언어군과 유사하게 하는 방법을 사용하였다. 그러므로 본 단어군은 개개 음소의 정오율을 판별하는 것보다는 오류가 많은 음소환경의 종류를 판별하는데 사용할 목적으로 작성되었다.

단어군 작성은 한국어 발음사전 통계조사로부터 가장 빈도수가 높게 나타난 음소환경인 CVC, VCV와 CC, VV환경을 중심으로 하였다. 단어군 작성의 기준으로는 각 자모음소의 통계적 빈도수를 고려하고, CVC, VCV의 순으로 여러가지 환경이 동일 빈도로 출현하도록 자모음소를 추가하면서 각 음소환경의 출현빈도를 측정하는 방법을 사용하였다.

단어군의 작성은 고정된 음절수에 대하여 출현가능한 음소환경의 종류를 우선 가정하면 개개 종류별 음소환경의 종류의 수를 구할 수 있다. 여기서는 4음절어를 가정하고 고려하는 음소환경을 CVC, VCV, CC, VV로 한정하였기 때문에 가능한 음소환경의 종류수는 10가지로 하고 전체 음소환경의 빈도 수의 비가 자연단어군의 그것과 동일하게 종류수를 조절하였다. 그리고 각 음소환경은 동일한 것이 두번씩 출현하지 않도록 미리 조합된 테이블을 두고 그곳에서 임의로 추출하여 단어군에 추가하였다.

기본적으로는 각 환경이 1회씩 출현하는 것을 목표로 하였고 음소의 빈도수는 고려하지 않았으나 결과에서 보면 거의 출현빈도가 비슷하다. 연구결과는 혼동의 가능성이 적은 6개의 모음과 18개의 초성자음, 종성자음을 결합한 48개로 단어의 갯수를 한정하고 모든 단어가 동일한 음절수를 가지도록 단어군을 작성한다.

작성된 단어군으로부터 합성된 음성을 이용하여 측정대상인 규칙음성합성시스템을 평가하는데 청취실험에는 정상 청력을 가진 평균학력수준의 일반인을 피험자로 한다. 시험은 헤드폰에 의해 합성음성을 청취하여 주어진 답안지에 들은 합성음을 받아쓰게 하는 방법을 사용한다.

몇가지 보완되어야 할 문제점들은 통상적 발음법칙의 반영, 즉 자연언어군에서 어떤 음소뒤에 어떤 음소가 확률적으로 많이 나타나고 어떤 음소가 올 확률이 거의 없다 등의 발음법칙의 적용이 필요하다. 현재의 단어군의 경우는 환경의 종류수만 고려한 나머지 실제로 발음이 어려운 무의미 단어가 만들어지는 경우가 많아 발성이나 인식에 어려움이 예견되는 경우가 있다. 이러한 통계적 데이타가 앞으로 음성학자들에 의해 조사되어야 할 분야이다.

그리고 음소열이 겹칠 때 발생하는 음운변동에서 새로 파생되는 음소환경을 고려하는 것이 필요하다. 현재의 경우는 초성과 종성이 음운환경의 중간에서 변화되는 경우를 모두 고려해 주지 못하였으므로 예상된 빈도수와 실제 단어군에서의 빈도수와 차이를 보였다. 또한 현재 포함되지 않은 소수 환경의 종류를 포함하는 것도 필요하며 각 음소의 빈도수도 유사하

도록 세부적으로 고려되어야 할 필요가 있다. 그러나 음소의 빈도수와 환경의 빈도수를 같이 고려하는 것은 현실적으로 단어수나 음절수의 증가를 초래하며 문제를 더욱 복잡하게 된다. 또 앞으로의 청취실험에서는 동일한 단어에 관한 자연음성을 같이 넣어 자연음성의 인식율과 합성음성의 인식율을 비교하여 실험할 필요도 있다.

Ⅳ. 결 론

음성인식 및 합성의 평가법에 관한 연구동향을 살펴보았다. 공업기술의 선진화는 측정방법의 표준 측정방법 및 장비의 정밀성 및 객관성에 크게 좌우되는데 이 음성합성 및 인식으로 대표되는 음성입출력 기술의 객관적인 평가법의 확보는 연구단계에서의 고급 알고리즘의 개발에 기초가 되고 최종결과물의 평가에 기준이 된다. 때 늦은 감이 없지 않으나 이러한 평가방법의 확보를 위한 노력이 필요한 시점이다. 다행스러운 것은 비록 부분적이나마 음성DB를 비롯한 개발 및 평가환경의 구축을 위한 공동노력이 각급기관을 중심으로 논의되기 시작하고 있어 다행스럽다. 음성언어에 관한 연구가 보다 고차원이 인간의 언어인지 및 생성과 관련된 문제들의 연구는 인접 학문간의 학제적 공동노력이 필요하며 평가법과 같은 표준화와 관련된 연구는 기관간의 협조체제의 구축과 함께 국제적인 협력에도 관심이 모아져야 할 것이다.

참 고 문 헌

1. 이용주, "음성합성 및 인식의 성능평가와 음성DB," 제9회 음성통신 및 신호처리 워크샵 논문집 1992. 8.
2. 이용주, "음성합성 및 인식의 성능평가와 음성DB" 음성통신 및 신호처리워크샵 (92. 8)
3. 이용주, 김성태, "음성인식 및 합성의 성능평가법에 관하여," 제5회 신호처리 합동학술대회 논문집 제5권 1호, 1992
4. 김성태 외 "음성입출력시스템의 성능평가법 연구" 한국전자통신연구소 위탁연구보고서 1993. 12.
5. 정성윤, 정현열, 김성태, "음성인식기의 환경요인에 대한 성능평가," 제 10회 음성통신 및 신호처리 워크샵, SCAS-10권 1호, pp. 251-255, 1993
6. 한국어 음성인식 및 합성기술 특집, 전자공학회지 20권 5호, 1993
7. 김순협 외, 규칙합성의 이해성 평가를 위한 단어표 구성 및 실험법, 전자공학회지 29권 1호 1992
8. Proceedings of ESCA Tutorial Day and Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, the Netherlands, 20-23 Sep. 1989
9. Preprints of International Workshop on International Coordination and Standardization of Speech database and Assessment techniques for Speech Input/Output, Kobe, Japan, Nov. 23-24, 1990
10. Proceeding of Workshop on International Cooperation and Standardization of Speech databases and Speech I/O Assessment Methods Chiavari, Italy, 26-28 Sep. 1991
11. L. W. Pols, "Evaluating the performance of speech technology systems" IEA Proceedings 15, 1991
12. Nakagawa, "Assessment and database of speech recognition/understanding systems" J. of IEICE, Japan (90. 12)
13. H. Kasuya, "Assessment of synthetic speech produced by rule" Proceeding of spring meeting of ASJ, Japan (90. 3)
14. Multi lingual Speech Input/Output Assessment, Methodology and Standardisation Final Report, SAM-UCL G004, 1992
15. 中川聖一, "音聲認識,理解 SYSTEMの評價와 DATABASE," 日本 電子情報通信學會誌, Vol. 73, No. 12, pp 1304-1310, Decem. 1990
16. Lea W. A., "What causes speech recognizers to make mistakes?," IEEE ICASSP, Vol. 3, pp 2030-2033, 1982
17. J. M. Baker, "Issues and Answers in Evaluation of Automatic Speech Recognizers," JASA, Suppl. 1, Vol. 70, 1981
18. Delogu C. & Paoloni A., "Human factors for automatic speech recognition," FUB-21, 1 March 1992
19. Chollet G. & Gagnoulet C., "On the Evaluation of speech recognizers and data bases using a reference system," IEEE ICASSP, pp 2026-2029, 1982
20. Chollet, "Reference systems for speech recognition research, development and evaluation," NATO-ASI, Bad Windsheim, July 1987
21. Hieronymus J. L. & Majurski W. J., "A reference speech recognition algorithm for benchmarking and speech data-base analysis," IEEE ICASSP, Tampa,

1985

22. K. F. Lee, "Automatic Speech recognition-development of the SPHINX system," Kluwer Academic Publishers, 1989
23. Hunt M. J., "Figures of merit for assessing connected-word recognizers," Speech Communication 9, pp 329-336, 1990
24. Steeneken H. J. M., "Diagnostic information of subjective intelligibility tests," IEEE ICASSP, Dallas, 1986
25. Woodard J. & Nelson J., "An information theretic measure of speech recognition performance," Proc. Workshop on standardization I/O technology, NBS, March 1982
26. Moore R. K., "Evaluating speech recognizers," IEEE Trans. ASSP, Vol ASSP-25, NO. 2, pp 178-183, 1977
27. Taylor M. M., "Issues in the evaluation of speech recognition system," J. Am. Voice I/O Soc., Vol. 3, pp 34-68, 1986
28. W. Lea & J. Woodard, "New procedures for comprehensive assessment of voice entry systems," Proc. Voice Data Entry Syst. Appl. Conf. (American Input-Output Society), Chicago, 1983
29. Winski R. & Kordi K., "Assessment of continuous speech recognizers using RSA," Proc. Eurospeech '91, Genoa, Vol. 2, pp 521-524
30. Steeneken J. M. & Velden J. G., "RAMOS-Recognizer assessment by means of manipulation of speech," Proc. ESCA 1989 Paris, pp 316-319, 1 June 1989
31. 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
32. 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
33. 北脇信彦, "부호화, 합성음성의 음질평가," 일본전자정보통신학회지 70권, 4호, 1987
34. 比企靜雄 외, "음성처리기술의 성능평가법에 관한 요인," 문부성과학연구비 「음성언어에 따른 맨머신 인터페이스의 고도화」 연구보고서, PASL 63-8-1, 1988
35. 前谷英樹, 음성합성기술의 평가, 문부성과학연구비 중점영역연구 「음성언어에 따른 맨머신 인터페이스의 고도화」 연구보고서, PASL 62-8-1, 1987
36. "Experiment in assessing the quality of synthetic speech," Temporary Document No. 70-E, CCITT working party 5/XII, Geneva, 27 Feb. -3 March 1992
37. "Elements for Draft Rec. on synthetic speech assessment," Temporary Document 80-E, CCITT working party 5/XII, Geneva, 27 Feb. -3 March 1992
38. "Report on Question 5/XII, Speech synthesis/recognition systems," Temporary Document No.52(REV. 1)-E, CCITT working party 5/XII, Geneva, 11-19 May 1993
39. "Draft Recommendation P.8S-Subjective Performance Assessment of the Quality of Speech Voice Output Devices," COM 12-6-E, CCITT working party 5/XII, 1993

이 용 주

- 1954. 1. 17. 일생
- 1972. 3 ~ 1976. 2 : 고려대 전자공학과 학사
- 1985. 9 ~ 1987. 8 : 고려대 대학원 전자공학과 석사
- 1988. 3 ~ 1992. 8 : 고려대 대학원 전자공학과 박사
- 1976. 3 ~ 1980. 7 : 공군 통신전자장교
- 1980. 8 ~ 1994. 2 : 한국전자통신연구소 자동통역연구실 (실장, 책임연구원)
- 1994. 3 ~ 현재 : 원광대 컴퓨터공학과 (조교수)
- 관심분야 : 음성정보처리, 휴먼인터페이스, 복지공학기술

김 경 태

- 1949년 5월 9일생
- 1972. 2 : 경북대학교 전자공학과 (공학사)
- 1980. 8 : 인제대학교 전자공학과 (공학석사)
- 1985. 3 : Tohoku Univ., Japan, 전기 및 통신 전공 (공학박사)
- 1991. 2 : 한국전자통신연구소 신호처리연구실 (실장, 책임연구원)
- 현재 : 한남대학교 정보통신공학과, 교수

정 현 열

- 영남대학교 전자공학과 교수

조 철 우

- 창원대학교 제어계측공학과 교수