特輯

음성 입출력 기술의 성능평가법 및 데이타베이스

金 敬 泰
漢南大學校 情報通信工學科

Ⅰ. 서론

음성을 맨머쉰 인터페이스의 수단으로 사용하기 위한 연구들이 각국에서 활발히 진행되고 있다. 이에 따라 음성분석, 부호화, 합성, 인식등의 응용시스템 개발이 활발하고 또한 일부는 이미 상품화되고 있다. 이러한 시스템의 연구 개발시, 애매한 성능 평가기준은 전체 개발의 효율을 떨어뜨리고 최악의 경우에는 기간 내 개발이 실패할 수도 있다. 지금까지 주로 사용되고 있는 평가방법으로는 일정한 척도(예를 들어 인식률, 명료도)를 주로 사용하는데 이를 제대로 정의하고 사용하는 것 또한 쉽지 않은 문제이다. 음성합성시스템의 경우에는 임의의 텍스트가 입력되어도 이해도와 자연성이 확보된 음성을 만들어내는 것이 목표이다. 따라서 그 과정에서 텍스트의 해석, 읽기 형태의 변환, 발성 스타일의 선택, 리듬, 음색, 억양 등의 구현이 종합적으로 잘 이루어져야 한다. 아울러, 단어이해도 및 음운 명료도와 함께 자연성도 확보되어야 한다. 뿐만 아니라 실제 응용에 적합하여야 하며 어떤 사용자에게도 잘 맞아야 할 것이다. 따라서 평가항목의 설정 또한 단순하지가 않다. 즉 단순히 한 두가지의 척도만으로 성능을 대표하기는 어렵다. 간단한 단어 인식시스템의 경우도 인식 결과 중 대치, 탈락, 삽입, 거부 등을 어떻게 볼 것인가, 대상이 아닌 단어나 그밖의 잡음에 반응하는 경우는 어떻게 할 것인가 등에 따라 평가 방법은 복잡하다. 따라서 인식률을 정의한다 하더라도 미리 부수적인 요인들을 함께 정의해야 한다. 이러한 문제들을 조직적이고 체계적으로 검토하여 객관적인 평가법이 확립된다면 연구자의 입장에서는 각종 방식의 우열을 객관화할 수 있게 되고, 제품 개발자의 입장에서는 더욱 효율적인 개발을 할 수 있을 것이며, 사용자의 입장에서도 상품의 객관적인 비교, 선택이 가능할 것이다. 이와 같은 관점에서 본고에서는 음성 입출력 기술의 성능평가법과 성능평가를 위한 음성 데이타베이스에 대하여 기술한다.

Ⅱ. 음성입력(인식) 기술의 성능평가

1. 개요

인식시스템을 가장 객관적으로 평가하는 방법은 모든 장소에서 모든 계층의 사람들이 모든 경우의 상황(스트레스, 물리적 상황, 정신적 상황 등)에서 직접 테스트해서 인식결과를 서로 비교하면 된다. 그러나 이러한 방법은 시간과 경제적인 여러 문제로 실현이 어렵다. 따라서 실제의 여러 상황에 직접 테스트를 하는 것이 아니라, 실험실에서 간편하게 테스트를 하면서도 객관적이고, 진단적이며, 예측적인 평가결과를 도출할 수 있는 평가법을 마련해야 한다. 이렇게 하려면 가능한한 실제의 상황을 그대로 시뮬레이션한 평가 절차와 측정된 인식결과로서 진단적이고 예측적인 평가를 할 수 있는 기술들이 개발되어야한다. 실제의 상황과 비슷하게 평가하기 위한 가장 좋은 방법은 여러 계층의 사람들이 여러 환경에서 발성한 음성을 녹음한 공통의 평가용 데이타베이스를 이용하는 것이다. 만약, 공통의 평가용 데이타베이스가 구축되어있다면, 적절한 평가항목을 선정해서 항목별 인식결과를 구하고, 인식결과와 평가항목간의 상관관계를 해석함으로서 전반적인 인식시스템의 성능을 진단하고 예측할 수 있다.

인식시스템을 평가하기 위한 절차로서 유럽의 평가 프로젝트인 SAM에서 구현한 인식기평가 구성도가 그림 1이다. SAM에서는 그림 1의 각 모듈을 공통 환경인 SESAM 내에서 참여한 나라별로 분담해서 연구를 추진하고 있다.

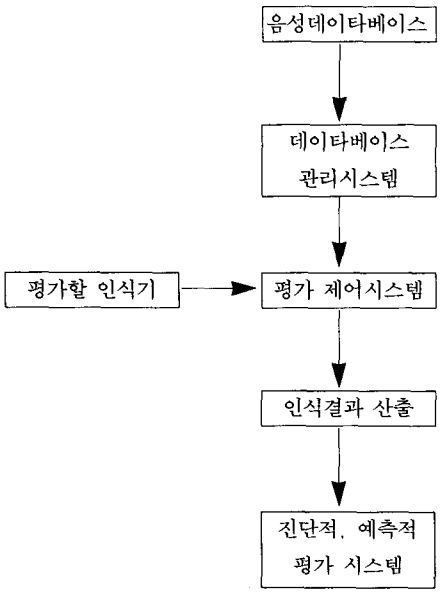

그림 1. 인식기 평가시스템 전체 구성도

2. 평가항목

인식시스템의 성능평가는 궁극적으로 인식률로서 평가되지만, 인식 알고리즘의 중심 기술과 이것 외에 인식률에 영향을 미치는 여러 요인이 있다. 따라서 이들을 포함한 평가는 상당히 어렵다. 그래서 인식시스템을 정확하게 평가하기 위해서는 인식률에 가장 영향을 많이 미치는 모든 요인을 조사해서 항목별로 분류한 뒤, 각 항목을 평가해서 전체 인식률과의 상관관계를 조사하면 된다. 이미 Lea는 1982년에 "What causes speech recognizers to make mistakes?" 라는 제목으로 80 가지 이상에 달하는 많은 요인들을 제시한 바 있다. [1]

이 논문에 의하면, 인식시스템의 인식 정확도에 영향을 주는 원인은 한마디로 변동성이라는 것이다. 이것은 세 가지 변동성으로 설명할 수 있는데, 화자의 내부변동성(intra-speaker variability), 상호화자의 변동성(inter-speaker variability), 콘텍스트 변동성이다. 이러한 변동들은 인간요인, 언어요인, 채널과 환경요인, 타스크 요인, 알고리즘 요인, 성능과 응답요인 등과 관련이 있다. 사실, 이 중에서 인식률에 가장 영향을 많이 미치는 요인은 인간요인, 언어요인, 환경요인 등의 세 가지이다. 따라서 이 세 가지 요인들을 평가항목으로 선정해서 변동성에 대한 인식기의 성능을 평가한다.

인간적 요인에는 발성속도, 성별, 방언, 성도크기, 발성습관, 화자의 성문 스펙트럼, 건강상태, 심리적 상태와 스트레스 등이 있고, 언어요인으로는 어휘수, 단어간의 음향적 유사성, 조음정도, 억양, 강세, 운율의 음운패턴 등이 있다. 그리고, 대표적인 환경요인으로서는 노이즈레벨(SNR), 노이즈의 형태(white,pink,tonal,impulse) 등이 있다.

3. 평가방법

인식시스템을 평가하기 위한 방법들은 대표적인 데이타베이스 사용 방법에서 진단적인 결과를 가지는 인공의 테스트 신호를 사용하는 방법까지로 아래의 다섯 그룹으로 나누어 생각한다. [2]

(1) 일반적인 데이타베이스 사용
(2) 표준시스템(Reference system)에 기초한 방법
(3) 특별한 Calibrated Database 사용
(4) 특별한 어휘에 기초한 진단적인 방법
(5) 인공의 테스트 신호 사용

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

(2)는 인간이나 표준 인식기에 기초한 방법이다. 여기에 해당하는 방법들로서는 Moore의 HENR (Human Equivalent Noise Ratio) 방법, Chollet의 Reference recognition algorithm, Taylor의 EVC(Effective Vocabulary Capability) 등이 있다. HENR은 어휘의 어려움을 고려한 측정법으로서 1977년 Moore가 제안한 방법이다. [3] 이것은 인간의 단어 인식처리 모델에 기초를 두고 있는데 임의의 노이즈레벨에서 임의의 입력단어 집합에 대한 Confusion Matrix를 예측할 수 있다. 그리고 Reference Algorithm은 1982년 Chollet에 의해 구현된 방법으로서 경쟁 알고리즘을 동일한 조건에서 비교하기 위해 모듈 형태로 설계되어 있다. [4]

이것은 데이타베이스 평가는 물론 응용의 복잡성도 평가한다. 1980년 Martin Taylor가 제안한 EVC는, 허용할 수 있는 에러율에 대해서 인식기가 조절할 수 있는 최대 어휘 크기를 측정하는 기법이다. [5]

위의 세 방법들은 고정된 기준, 즉 노이즈, 어휘, 인식알고리즘 등에 기초한 모델을 사용한다. 따라서 어떤 입력변화에 대해 체계적인 에러를 유발할 수 있다. 또한 이러한 방법들은 진단적이지 못하고, 대표적인 어휘를 사용해야 하는 문제점도 있다.

(3)은 Peckham이 제안한 것으로서 RSA (Recogniser Sensitivity Analysis) 방법이다. [6] Calibrated databases의 집합을 사용하는 방법으로서 각 데이타베이스는 인식기의 성능에 영향을 끼치는 환경조건에서 수집된 것이다. 이 방법은 인식성능의 에러에 영향을 주는 음성 변동성이 소수의 측정가능한 파라메터들로 특성화될 수 있다는 전제에 기초를 두고 있다. 이러한 파라메터들과 인식성능 간의 관계를 결정하므로서, 간단한 Scoring 방법으로는 불가능한 인식기의 기본동작 특성에 대해 중요한 고찰을 할 수 있다.

(4)는 Steeneken이 제안한 RAMOS(Recognizer Assessment by means of Manipulation Of Speech) 방법이다. [2] 이것은 대표적인 음소들 간의 Confusion 해석으로부터 Reference 조건들의 집합에 관련된 다차원 표현을 얻는 방법이다. 이 방법은 좀더 일반적인 방법, 즉 특별한 음성 발성 파라메터와 음성 전송 파라메터의 변동함수로서 인식시스템이나 인식알고리즘의 성능을 명시하는 것이 목적이다.

(5)의 인공(non-speech) 테스트신호에 기초한 평가방법은 아직 알려져 있지 않다. 그러나 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.

4. 음성인식기의 성능측정 기술

인식기의 성능측정은 설정된 평가항목에 대해 적절한 평가방법을 적용해서 얻은 결과를 정량화함을 의미한다. 이것은 단순히 평가된 결과만을 정량화하는 것이 아니라, 진단적이고 예측적인 평가가 가능해야 한다. 즉, 평가한 인식기의 성능한계를 정량화하는 표준기술을 개발해야 한다. 아직까지 성능의 한계를 정량화하는 표준기술은 개발되지 않았지만, 본 절에서는 이제까지 많이 사용된 성능측정 기술에 더해 간략히 기술한다.

1. 퍼센트 인식률(에러율) [7]

퍼센트 인식률은 인식능력의 척도로 가장 많이 사용하고 있는 기술로서 인식정확도와 대체 에러율로 성능을 표현한다. 이러한 종류의 측정들의 단점은 주어진 타스크에서 나온 에러수를 제외한 다른 성능에 관한 정보는 없다. 즉, 에러 분포를 계산할 수 없고 다른 에러보다 더 비중 있는 에러를 판단하는 규정이 없다. 또한 에러율은 어휘크기에 영향을 받고 어휘크기나 어휘 어려움에 대한 성능비교가 불가능하다.

퍼센트 인식률이나 에러율은 고립단어인식과 연속음성인식의 경우를 나누어서 계산해야할 사항이 있다. 고립단어인식의 경우에 에러율 계산은 틀린 단어에 의해 정해단어(Correct word)가 대체된 대체(substitution) 에러수만 구하면 된다. 그러나 연속음성의 경우에는 세 종류의 에러, 대체(substitution), 탈락(deletion), 삽입(insertion)에러 등이 발생한다. 여기에서 탈락에러란 정해단어가 인식된 문장에서 빠지는 경우에 생기는 에러를 의미하고, 삽입에러는 인식된 문장에서 여분의 단어가 첨가된 경우에 생기는 에러를 의미한다. 대체와 탈락은 분명히 에러이지만 삽입을 에러로 계산해야하는지의 여부는 분명하지않다. 초기에는 이러한 삽입을 에러로 간주하지 않았지만 최근에는 에러로 계산하는 추세이다. 사실, 기대하지 않았던 단어가 삽입된 경우에는 분명히 에러로 계산해야되기 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. Percent Correct와 Word Accuracy는 다음과 같이 계산되어진다.

$$\text{Percent Correct} = \frac{\text{Correct}}{\text{Correct Sentence Length}} \times 100$$

$$\text{Error Correct} = \frac{\text{Subs} + \text{Dels} + \text{Ins}}{\text{Correct Sentence Length}} \times 100$$

Word Accuracy = 1 - Error Rate

Correct Sentence Length = Correct + Subs + Dels 이기 때문에, 단어정확도(Word Accuracy)는 아래 식으로 된다.

$$\text{Word Accuracy} = \frac{\text{Correct} - \text{Ins}}{\text{Correct Sentence Length}} \times 100$$

2. Confusion Matrix

이것은 퍼센트 인식률의 한 단점을 보완할 수 있는 기법인데 에러가 어휘 전체에 퍼져있는지, 단지 몇 개의 항목에만 집중되어 있는지를 알 수 있기 때문에 유용한 진단적인 정보를 제공한다. Confusion Matrix의 각 성분 M_{ij}는 단어(혹은 음소 등) i의 100개의 테스트에 대해서 단어 i가 j로 인식된 횟수의 평균치 추정이다.

$$M_{ij} = \frac{\text{단어 i가 j로 인식된 횟수}}{\text{전체단어의 수}} \times 100$$

여기에서 행렬의 대각 성분의 평균이 인식률이 되고 비대각 성분의 평균이 에러율이 된다.

3. 상대정보 손실도(RIL :Relative Information Loss) [8]

Confusion Matrix로부터 유도된 정보이론적인 측정법으로서 성능측정을 정의하기 위해 엔트로피를 사용한다. 이것은 1982년 Woodald와 Nelson이 제안한 것으로서 $H(X\ Y)/H(X)$를 RIL로 정의하였고, 애매도에 대한 함수로서 기울기가 $1/H(X)$이고 원점을 지나면서 0에서 1까지의 값을 갖는 RIL 함수이다. 이것의 장점은 에러의 분포를 고려할 수 있고, 속도왜곡모델과 관련되어 사용할 때에는 음성입력시스템에 개개의 에러값을 반영할 수 있다. 그러나 이러한 측정의 문제점은 인식기들이 다른 어휘에 테스트될 때 비교평가를 할 수 없다는 것이다.

$$\frac{H(X|Y)}{H(X)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} P(Y_j)P(X_i|Y_j)\log_2 P(X_i|Y_j)}{\sum_{i=1}^{n} P(X_i)\log_2 P(X_i)}$$

단, X는 입력심볼, Y는 출력심볼, H는 엔트로피를 나타낸다.

4. Perplexity [9]

Perplexity는 인식대상의 복잡성의 척도로서 언어모델에서 생성될 수 있는 문의 총 수에 관련된 척도이다. 만약, 언어 L에 있어서 단어 예(혹은 음절, 음소) $W_1^k = W_1, W_2, W_3, \cdots\cdots, W_k$의 출현확률을 $P(W_1^k)$로 하면, 언어 L의 엔트로피는 다음 식으로 정의된다.

$$H_e(L) = \sum_{W_1^k} P(W_1^k)\log_2 P(W_1^k)$$

단어 대신에 음절, 음소의 예를 사용해도 $He(L)$의 값은 변하지 않는다. 또, 한단위당 엔트로피는

$$H(L) = -\sum_{W_1^k} (1/k)P(W_1^k)\log_2 P(W_1^k)$$

로 정의된다. 부호화 정리에 의하면 언어 L의 한 단위당 엔트로피가 $H(L)$이라면, 다음 단위어를 결정하기 위해 평균 $H(L)$회의 Yes/No 질문을 반복하여야 한다. 즉, $2^{H(L)}$회의 같은 출현 확률의 단위어에서 한 단위어를 결정하는 것이 된다. 이것은 정보이론적인 의미에서의 평균 분기수이고, Perplexity라고 부른다. 이것을 $Fp(L)$라 하면 다음 식으로 주어진다.

$Fp(L) = 2^{H(L)}$

5. 성능평가방법의 예 : RAMOS [2]

RAMOS는 음성인식기의 성능을 음성발성 파라메터와 음성전송 파라메터의 변동함수로서 측정해서 화자내부, 상호화자, 스트레스의 영향, 노이즈(SNR, 노이즈 형태) 등의 요인에 관한 성능을 평가하는 평가시스템이다. RAMOS는 최소차이 단어집합으로된 CVC형태의 단어 데이타베이스를 사용하는데, 이것은 초성자음, 종성자음, 모음의 세 그룹으로 구성되어있다. 평가방법은 자연음성에서 관찰된 물리적 파라메터의 변동이나 다양한 환경조건들 하의 화자들의 변동에 해당하도록 단어들을 분석/합성 방법으로 조작하여 관련 파라메터의 변동을 정의하기 위해 대표적인 음성토큰들을 해석한다. 그림 2는 고립단어인식기의 평가에 대한 RAMOS의 전체 구성도이다.

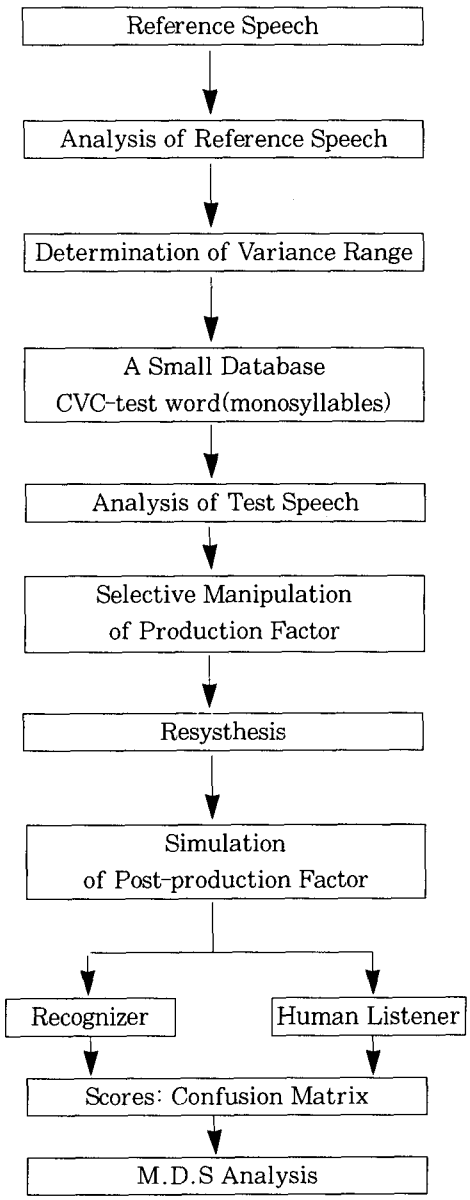

그림 2. RAMOS의 구성도

먼저, 다양한 조건에 대한 음성발성 파라메터의 변동을 통계적으로 정량화하기 위해 대표적인 음성 토큰들을 녹음한다. 특히, 데이타 수집시 고려할 사항은 내부화자변동, 상호화자변동, 남/여 음성의 변동, 물리적인 스트레스의 영향등이다. 녹음은 각 4명의 남,여 발성 화자가 10개의 단어로 된 하나의 문장을 두번씩 발성한다. 이 때 발성은 보통의 방법으로 발성한 것과 물리적인 스트레스를 가한 후 발성한 것으로서 4명 중 2명은 다른 날, 다른 장소에서 녹음한다. 음성은 잡음이 없는 방에서 고질의 콘덴서 마이크로폰을 사용하여 DAT에 녹음한다. 이렇게 녹음된 음성 데이타를 12차 LPC와 BPF 뱅크로 해석하여 특징 파라메터 추출 후, 남/여 음성, 보통/스트레스 음성, 화자의 내부/상호 화자의 변동 조건에 따른 기준 음성의 변동 범위를 정한다. 보통 상호화자의 변동이 화자 내부의 변동보다 변동 폭이 훨씬 더 크다.

인식테스트 음성은 CVC 형태의 단음절로서 세 개의 집합, 초성자음, 종성자음, 모음그룹으로 분류한다. 이 그룹들은 요구된 하나의 음소만 다르고 나머지 두 개의 음소는 각 테스트 단음절에서 반복되게 한다. 이렇게 구성된 테스트음성은 위에서 사용한 해석기술을 사용해서 미리 조사된 변동 범위에 따라 인식률에 영향을 미치는 변동 요인별로 특징추출된 음성파라메터를 조작한다. 조작된 음성은 재합성 단계를 거쳐, 환경요인에 관한 성능평가를 위해 노이즈를 첨가한다. 이 때 노이즈는 노이즈의 형태와 SNR에 따라 변화를 주게 된다.

음성발성파라메터 변동과 음성전송 파라메터의 변동에 관한 조작이 끝나면, 조작된 음성을 인식기에 테스트함으로서 인식결과를 얻는다. 이때 인간의 인식성능과 비교를 위해 청취를 위한 전문인을 두어 인식실험을 병행한다. 인식된 결과는 MDS(Multi-Dimensional Scaling) 분석을 위해 Confusion Matrix 형태로 나타낸다. MDS 분석의 목적은 데이타 중에 숨겨진 구조를 찾아내어 그 구조를 소수 차원의 공간에서 기하학적으로 표현하는 것이다. 즉, 데이타에 포함된 정보를 추출하기 위해 MDS를 탐색 수단으로 사용하는 것이다. 여기에서는 다양한 입력 조건에 대한 인식기의 동작을 연구하기 위한 방법이다.

Ⅲ. 음성합성의 성능평가

1. 개요

규칙에 의한 음성 합성기술의 평가는 합성기술의 완성도에 관한 진단적 평가와 합성시스템의 이용목적, 이용자, 이용환경을 포함한 종합적 평가로 나누어 생각할 수 있다. 진단적 평가는 음성합성기술을 개발 연구하는 기술자나 연구자 자신이 개발과정의 피드백 요소로서 각 개발 단계별로 개별적으로 수행

하는 경우가 많지만 최근에는 이를 계통적으로 규범화하려는 움직임을 보이고 있다. [10] [11] [12]

또한, 규칙합성기술이 맨머쉰 인터페이스로서 폭넓게 이용되기 위해서는 이용자인 인간의 음성언어 표출 및 수용과정을 포함한 토탈 시스템의 구성요소로서 합성기술을 바라보면서 종합적으로 평가하는 종합적 평가방법의 확립이 시도되고 있다. 여기서는 우선, 내용을 규칙합성으로 한정하여 성능평가법을 기술한다.

2. 규칙합성의 단계별 성능 평가

규칙합성시스템은 일반적으로 언어처리부, 음운처리부, 음운 및 운율의 음향파라메터 생성부, 음성신호 생성부로 구성된다.

1) 텍스트 전처리 및 음운변환

약어, 구두점 및 숫자 읽기의 정확도를 객관적으로 평가할 필요가 있다. 자동교열시스템(proof reading)과 같이 합성 시스템의 용도에 따라 이 부분의 성능에 민감한 경우는 특히 그러하다. 음운의 읽기변환은 전자화된 발음사전으로 평가 가능하나 우리말의 경우는 특히 같은 휴지구간 내에서의 단어간 결합에 따른 평가법이 필요하고, 영어의 경우는 외래어나 외국이름(인명)의 정확한 발음 평가가 해결되어야 할 문제이다.

2) 분절음 평가

음운 명료도나 단어 또는 문장이해도의 경우 계통적 평가가 필요하여 영어의 경우는 MRT나 DRT 등의 방법이 사용되어 오고 있으나 언어에 종속되므로, SAM(Speech Assessmemt Method) 프로젝트에서는, 무의미 음소열을 중심으로 각 언어에 따른 음운 발생의 제약 조건을 고려하여 랜덤하게 발성 목록을 작성하고, 자극음을 제시한후, 그 결과를 키보드로 받아 스코아를 계산하고, 그 결과를 통계처리하여 알려주는 자동 시험절차를 구성하여 사용하고 있다. 단어나 문장레벨의 이해도 평가를 위해서는 문법적으로는 맞고 의미적으로는 틀린문장을 작성하여 사용하고 있는데 이는 문법적으로나 의미적으로 올바른 문장은 부분적으로 내용을 예측 가능하기 때문이다.

3) 운율의 평가

음운 명료도나 단어의 이해도가 어느 정도로 확보되었더라도 운율정보의 구현이 불충분하면 자연스럽지 못하다. 정확한 운율정보의 구현이 오직 한 가지만 있는 것이 아니므로 평가의 어려움은 더하다. 따라서 계통적이고 표준적인 운율 평가법은 아직 존재하지 않고 주로 대 비교법(pair comparison)에 의한 성능판단 등이 제한적으로 활용되고 있다.

4) 개괄적인 품질평가

합성음을 문장 또는 문단 단위로 제시하여 일반적 품질, 발음의 명료성, 발음의 정확성, 음색, 강세, 템포, 사실성, 유창성, 자연성 등의 항목을 대상으로 MOS(Mean Opinion Score)를 평가하는 방법이 사용되고 있다.

5) 현장평가

실제로 운영중에 있는 시스템을 대상으로 하는 평가법으로서 예를 들어 스웨덴에서는 시각장애자를 위한 신문낭독 서비스가 제공되고 있는데 이 경우에 제어명령어(부분청취부분 검색, 스킵 등) 사용의 일반패턴, 낭독시간, 낭독하는 텍스트의 양, 낭독속도, 사람이 낭독한 음성과 합성음성의 장단기 기억능력, 경험의 정도 등의 사항을 고려하여 평가가 시도되고 있다.

6) 객관평가

주관적인 평가는 인간이 직접 듣고 평가하므로 많은 시간이 소요된다. 따라서 통신 분야에서는 이해도와 음성품질을 기계적인 방법으로 객관 평가하고자 시도하고 있다. 이 때 주된 영향인자로써 채널특성(전화대역), 환경(잡음,잔향)이 고려되어야 하고 개괄적인 음성특성뿐만아니라 전체적인 S/N비, 평균피치, 휴지의 분포 등도 고려하여야 한다. 이분야는 CCITT에서 특히 중점적으로 다루고 있다.

Ⅳ. 음성 데이타베이스 및 관리시스템

1. 개요

음성 정보처리를 연구하기 위하여 다양한 종류(성별, 연령, 방언, 인원, 발성횟수)의 음성 데이타베이스 구축의 필요성이 날로 더해가고 있다. [13] 국내의 경우 지금까지는 음성 연구를 하고자 하는 각자가 필요에 따라 음성 데이타를 녹음, 이용, 보관하여 왔다. 그러나 음성 연구가 진전됨에 따라 처리하고자 하는 데이타량의 증가가 요구되고, 음성 정보처리 시스템의 연구 개발을 위하여 분석, 합성, 인식의 각종 방법을 비교, 평가할 수 있는 공통 음성 데이타가 요구되고 있다. 이렇게 여러 사람이 이용 가능한 각종의 음성 데이타를 수록, 보관, 공개 하는 일은 매우

중요한 일이며, 연구 개발 과정과 성능평가의 차원에서도 꼭 필요한 일이다.[14] 이와 같은 목적으로 이용되는 음성 데이타를 일반적으로 음성 데이타베이스라 부른다. 음성 데이타베이스를 구축할 경우, 그 이용 분야로서 음성 인식을 비롯하여 음성 합성, 화자 인식 분야로 생각할 수 있다. 각 분야별로 독립적인 음성 데이타베이스가 있으면 좋겠지만 경우에 따라서 어느 방식도 음성 인식용 데이타베이스를 부분적으로 이용할 수 있다.

2. 음성 데이타베이스

공통 이용 가능한 음성 데이타를 수집해서 보관, 공개함은 연구개발 입장에서는 분석 방식과 인식 알고리즘의 개발, 평가에 이용할 수 있고, 사용자 입장에서는 인식장치의 성능평가 비교를 객관적으로 할 수 있다.

또 음성 데이타베이스의 이용가치를 높이기 위해서는 여러 가지 조건(발성내용, 발성자의 범위, 녹음조건 등)에서 수록하는 다양성, 조건 설정이 적절하여 한쪽으로 편중되어 있지 않는 비편중성, 데이타의 다량성, 누구라도 손쉽게 이용할 수 있고 편집이 쉬운 편리성이 있어야 한다.

발성내용으로서, 단음절, 숫자, 지명(도청소재지, 도시명), 역 이름, 성명, 기본 어휘, 틀리기 쉬운 단어 쌍(minimal pair) 등의 단음절, 일기 예보 등의 구 발성, 문장등의 발성등을 생각할 수 있다.

또한, 음성 데이타베이스의 다양성, 비편중성의 요구조건을 만족시키기 위하여 발성자의 연령과 직업, 출신지 등은 가능한 넓은 범위에 걸쳐 있는 것이 바람직하고 대화 방법은 표준어(공통어)를 원칙으로 하지만 궁극적으로는 방언을 첨가하는 것이 필요하다.

데이타 량은 소수화자, 최저 2회 발성으로 하는 것이 바람직한 특정 화자용과 다수 화자의 2회 발성이 바람직한 불특정 화자용으로 나눠서 생각하는 것이 필요하다.

3. 음성 데이타의 수록과 편집

1) 수록방법

음성 데이타의 수록과 편집을 위하여는 수록방법, 조건, 어디에 기록하느냐 라는 기록메체를 고려하려야 한다. 먼저 수록방법으로서는 두 가지 방법을 생각할 수 있는데 특정한 한 장소에서 수록하는 방법과 여러 장소에서 수록하는 방법이 있다. 전자는 녹음조건이 같은 점은 장점이지만 녹음을 위한 작업이 집중화되어 특정 기관에 부담을 주는 단점이 있고 후자는 녹음을 위한 작업을 분산할 수 있지만 녹음장치의 특성에 분산(오차)이 생긴다. 그러나 녹음장치의 모델과 방식을 통일시키면 특성의 분산(오차)은 최소화할 수 있다.

수록조건으로는 무향실, 방음실, 조용한 방, 잡음환경실, 계산기실 등의 녹음장소와 입력장치(마이크로폰, 전화기), 필터 특성, 표본화 주파수, 양자화 정도를 고려한 A/D 변환의 고려사항이 있다.

기록매체로서는 아날로그 녹음 테이프는 취급이 간단하고 호환성은 좋지만 복사, 테이프 길이의 변화와 더빙에 따른 품질 열화 등이 일어날 수 있다. 디지틀 오디오 테이프(DAT)는 디지틀 더빙이 가능하지만 헤더정보의 기록이 곤란하다. 콤팩트 디스크(CD)는 보존성 면에서는 DAT보다 좋고, 디지틀 자기테이프는 이용 방법은 바람직하지만 보존성과 표본화 주파수가 문제가 된다. 또 계산기 설비와 상당한 작업 시간이 필요하다.

2) 발성내용의 제시방법

일반적으로, 음성을 수집하기 위하여 녹음을 한다고 하면 평상시의 자연스러운 발성이 나오지 않고 긴장하게 되고 조심스럽게 발성이 되어 실제적인 효과를 얻기가 상당히 어렵다. 따라서, 어떻게 하면 보통 때의 발성과 똑같이 발성된 음성을 얻을 수 있을까가 중요한 연구의 대상이 된다. 보통 생각될 수 있는 방법으로 첫번째, 발성단어 혹은 문장의 리스트를 만들어 두고 발성자로 하여금 읽게 하는 방법, 두번째, 발성해야 할 단어를 헤드폰을 통하여 음성으로써 알려주는 방법, 세번째, 발성해야 할 대상 단어, 또는 문장을 개인용 컴퓨터 등의 화면에 디스플레이 시켜 발성자가 읽게 하는 방법, 네번째, 화면 등을 이용하여 원하는 발성이 나오도록 유도 질문하는 방법이다.

대량의 음성 데이타베이스를 구축함을 목적으로 세번째의 발성내용 제시시스템을 구현하여 4연속 숫자음 35개, 숫자음 25개를 각각 96명이 9회씩 발성한 데이타와 계산기 제어용 40개 단어를 60명이 5회씩 발성한 데이타를 수집하고 레이블링하는 데 적용한 결과가 있다.[15]

4. 음성 데이타베이스 작성, 관리[16]

음성을 수록하고 그 음성이 표현하는 내용을 보조정보로 부가하는 데이타베이스를 작성하는 일은 대단

히 많은 수고와 시간을 요하는 작업이다. 따라서 음성 데이타베이스를 작성 지원하는 시스템을 개발하여 자동화하고 작업 효율을 향상시킬 필요가 있다. 또 대량의 음성 데이타의 축적, 보존의 관리 및 이를 유용하게 이용하기 위한 검색을 지원하는 시스템 구축에 대한 연구가 필요하다. 이를 위한 연구 내용으로서는 다음과 같은 것들이 있을 수 있다.

· 연구 목적에 맞는 부음성 데이타베이스를 주 음성 데이타베이스에서 검색하는 방법

· 음성 데이타, 지식 데이타 등의 데이타를 추가하거나 변경을 쉽게 하기위한 관리 시스템 구현

· 음성 레벨의 편집 시스템 (잡음 부분의 제거, 재발성, 정정 발음의 편집, 보조 정보의 데이타 또는 기록 매체상의 데이타 형식의 설정과 신뢰성 향상을 위한 에러 정정 부호의 검토 연구가 있다.

Ⅴ. 결론

본 논문에서는 음성 입출력 기술인 음성인식 시스템과 음성합성 시스템의 성능평가법과 성능평가와 연구를 위한 음성 데이타베이스에 대하여 고찰했다. 다양한 분석기법과 알고리즘, 그리고 서로 다른 음성데이타를 사용하여 구성된 시스템을 객관적으로 평가한다는 것은 매우 어려운 일이다. 음성입력기술의 평가를 위하여는 주로 인식률에 영향을 많이 미치는 요인들에 대한 인식기 성능의 진단적 예측적 평가에 대해 개관하였다. 즉, 알고자 하는 인식기의 특성을 H, 입력 X를 변화시켰을 때 인식률로 표현하는 출력 Y를 구하므로서, 입력 X와 출력 Y의 상관관계에 의해 인식기의 특성 H를 유도하는 것이다.

그러나 평가의 최종 목적인 더욱 향상된 인식시스템의 개발이라는 측면에서 볼 때, 각 조건에 대해 얻어지는 정보로서 인식률의 향상을 꾀할 수 있는 방법이 필요하게 된다. 즉, 인식률의 저하를 가져오는 인식시스템의 구성요소를 밝혀내고, 이를 수정해서 인식률의 향상을 꾀한다. 이러한 방법은 인식률에 영향을 미치는 요인들을 평가항목으로 취한 경우와 비교해 볼 때 평가항목을 더 세분화한 것이라 할 수 있고 두 방법 모두 검토할 부분이 아직 많이 있다. 전자의 경우, 경제적인 부담을 줄이고, 실험의 간편성을 위해 실제 상황에서 예측하는 것과 비슷하게 실험실에서 성능을 진단하고 예측한다. 이 경우 각 조건에 대한 실험이 실제 상황에 어느 정도 가까운가 하는 문제가 남게 된다. 따라서 현실성 있는 실험에 대한 깊은 연구가 뒤따라야 한다. 후자의 경우는 아직 시도한 국가나 연구소가 없다. 그러나 이러한 방법은 앞으로 반드시 해결해야 할 과제라고 생각한다. 이 연구가 되어야만 임의의 인식기의 최종 성능 한계를 명시할 수 있다.

합성의 평가법에 관한 연구에서는 주로 규칙합성의 평가법을 중심으로 정리하였다. 음성인식 합성시스템을 올바로 평가하기 위하여 필수적인 음성 데이타베이스에 대하여 기술하였다.

음성언어에 관한 다른 연구도 그러하듯이 인간의 언어인지 및 생성과 관련된 문제들의 연구는 인접 학문 간의 학제적 공동 노력이 필요하며 평가법과 같은 표준화와 관련된 연구는 기관 간의 협조체제의 구축과 함께 나아가서는 국제적인 협력에도 관심이 모아져야 할 것이다.

參 考 文 獻

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

[2] Steeneken J.M. & Velden J.G., "RAMOS-Recognizer assessment by means of manipulation of speech", Proc. ESCA 1989 Paris 3160319, 1 June 1989.

[3] Moore R.K., "Evaluating speech recognizers", *IEEE Trans.* ASSP, vol ASSP-25, NO.2, pp 178-183, 1977.

[4] Chollet G. & Gagnoulet C., "On the Evaluation of speech recognizers and data bases using a reference system", IEEE ICASSP, pp 2026-2029, 1982.

[5] Taylor M.M., "Issues in the evaluation of speech recognition system", *J. Am. Voice I/O Soc.*, vol.3, pp 34-68, 1986.

[6] Peckham J., Thomas T. and Frangoulis E. "Recogniser Sensitivity Analysis : A Method for assessing the Performance

of Speech Recognisers", Speech Communication, pp 317, vol 9, August 1990.
[7] K.F.Lee, "Automatic Speech recognition-development of the SPHINX system", Kluwer Academic Publishers, 1989.
[8] Woodard J. & Nelson J., "An information theretic measure of speech recognition performance", Proc.Workshop on standardization I/O technology, NBS, March 1982.
[9] 中川聖一, "音聲認識,理解 SYSTEMの 評價と DATABASE", 日本 電子情報通信學會誌, vol.73, No.12, pp 1304-1310, Dec. 1990.
[10] 이용주, "음성합성 및 인식의 성능평가와 음성 DB" 음성통신 및 신호처리워크샵 (92.8)
[11] Proceedings of ESCA Tutorial Day and Workshop on Speech Input/Output Assessment and Speech Databases Noordwijikerhout, the Netherlands, 20-23 Sep. 1989.
[12] Proceeding of Workshop on International Cooperation and Standardization of Speech databases and Speech I/O Assessment Methods Chiavari, Italy, 26-28 Sep. 1991.
[13] 김경태, 최준혁, "음성데이타베이스의 동향과 구축", 한국음향학회 음성통신 및 신호처리 워크샵, 1990.8, KAIST.
[14] 板橋秀一, "音聲 Database構想", 日本語學, pp14-23, (平1-3)
[15] 김경태, 이용주, 정유현, "음성 데이타 베이스 수집을 위한 발성내용 제시시스템", 한국음향학회, vol.12, no.1(1993)
[16] J.H.Choi, K.T. Kim, "Construction of a Large Korean speech database and its management system in ETRI", ICSLP 90, Kobe(1990)

筆者紹介

金 敬 泰
1949年 5月 9日生
1972年 2月 경북대학교 전자공학과(공학사)
1980年 8月 연세대학교 전자공학과(공학석사)
1985年 3月 Tohoku Univ., Japan(공학박사)

1978年 1月 ~ 1986年 1月 한국기계연구소
1986年 1月 ~ 1991年 3月 한국전자통신연구소 신호처리연구실
1991年 3月 ~ 현재 한남대학교 정보통신공학과

주관심분야: 음성인식 및 합성기술, 음성처리 기술의 성능평가법, 휴먼인터페이스 기술, 신호처리분야