소 특 집

음성인식 기술 개요 및 향후 과제

김 회 린

한국정보통신대학원대학교 공학부

I. 서 론

지난 20세기의 통신기술의 발전은 서로 멀리 떨어져 있는 사람들간의 대화를 가능케 하는 유선 전화기의 발명으로부터 시작하여 그야말로 비약적인 발전을 거듭하여 현재는 언제, 어느 곳에서나 누구와도 서로 대화할 수 있는 이동통신 전화기가 보편화되어 있다. 이러한 통신 기술은 기본적으로 사람과 사람 사이의 통신의 자유도를 향상시키는 방향으로 발전되어 왔는데, 최근에는 이를 뛰어 넘어 사람과 기계 사이에 음성을 이용한 대화를 가능케 하는 기술에 대한 연구 및 개발에 큰 관심이 집중되고 있다. 이와 같은 voice-activated man-machine interface 기술에 대한 요구는 원래 음성을 전송할 때 전송 효율을 최대한 향상시키기 위해서는 음성을 신호 자체가 아니라 그 음성을 문자로 표현한 symbol들로 변환하여 전송할 때 가장 효율적으로 전송할 수 있다는 발상에서 출발하였다. 그러나, 만일 사람이 발성한 음성을 기계, 즉 컴퓨터가 문자로 변환하고 이를 이해할 수 있다면, 이는 우리가 공상과학 소설이나 영화에서 무수히 접해 왔던 대로 다양한 분야에서 그 파급효과가 대단히 클 것으로 쉽게 예상할 수 있다.

그러면, voice-activated man-machine interface 기술의 구성 요소는 무엇일까? 이는 사람과 사람 사이의 대화 과정을 보면 쉽게 이해할 수 있다. 즉, 사람의 두뇌에서 생성된 개념이 일정한 규칙을 가지는 특정 언어 형태로 완성되어 입의 조음기관에서 공기의 진동으로 전파되어 나오는 과정이 음성언어의 발생 과정이며, 이를 우리는 "음성합성" 과정이라고 한다. 다음으로 공기 중에서 전파되는 과정을 바로 옆의 사람이 아니라 멀리 떨어진 사람에게 전달되도록 할 때 이를 우리는 음성통신, 특히 "음성부호화" 과정이라고 한다. 한편, 전달되어 온 음성 신호를 청각기관인 귀를 통해 받아들여 모종의 신호변환 과정을 거쳐 두뇌로 전달하여 음성신호 내에 포함되어 있는 음성언어를 인지하는 과정을 "음성인식" 과정이라고 한다. 이렇게 음성언어의 발생/전달/인지 과정을 음성합성/음성부호화/음성인식의 3단계로 나누어 볼 때, 음성합성 및 음성부호화 기술은 아직 미흡한 측면이 많지만 그런대로 상용화되어 비교적 널리 이용되고 있는 반면, 음성인식 기술은 아직 크게 대중화되지 못하고 있는 실정이다. 이는 근본적으로 음성합성 및 음성부호화 시스템은 최종 수신측이 매우 지능적인 인간이기 때문에 비록 자연스럽지는 못해도 청취한 음성이 무슨 내용인지를 사람이 판단할 수 있는 수준까지 기술이 발전한 때문이라고 볼 수 있지만, 음성인식의 경우는 수신측이 훨씬 덜 지능적인 컴퓨터이므로 잘못 인식된 결과에 대한 지능적 재처리가 매우 힘들어서 아직 널리 이용되고 있지 못한 실정이다.

본 고에서는 사람의 말, 즉 음성언어를 컴퓨터가 알아 듣는 음성인식 기술에 대하여 II장에서 기본적인 시스템 구현 방법 및 상용화를 어렵게 만드는 요인들에 대하여 살펴보고, III장에서는 최근의 기술개발 동향 및 향후 극복해야 할 기술적 과제를 간략히 살펴본 후, IV장에서 결론을 맺기로 한다.

II. 음성인식 기술 개요

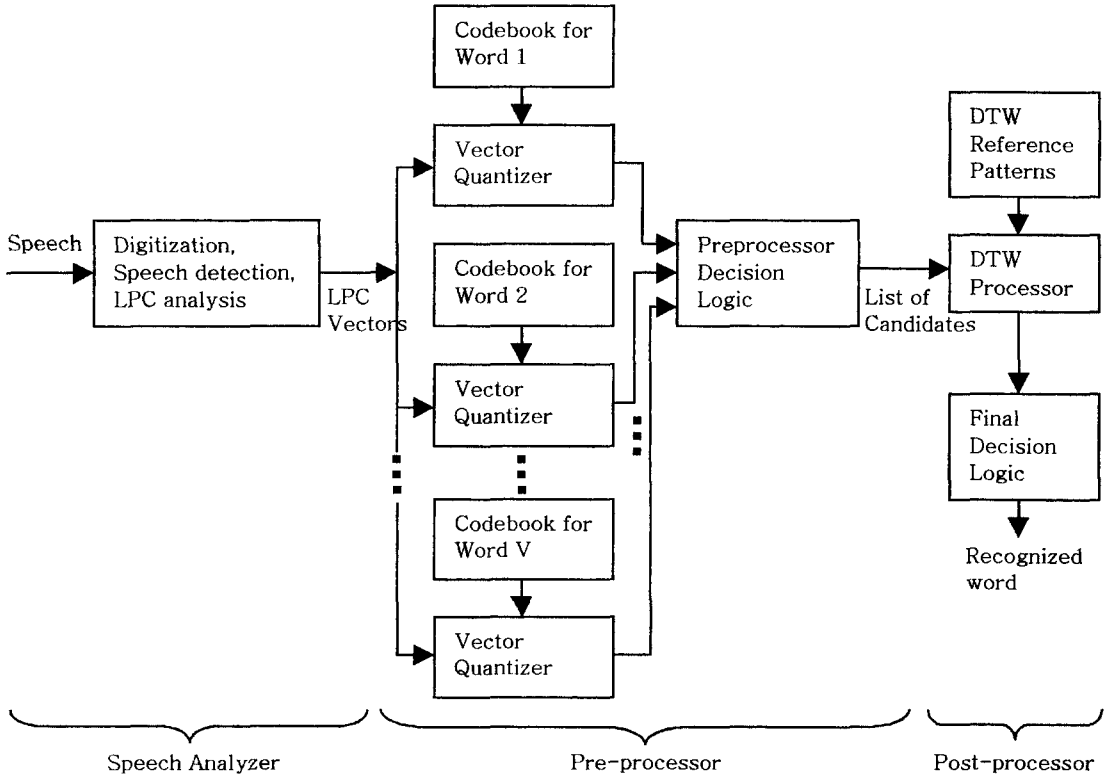
앞서 서론에서 기술하였듯이, 음성인식은 공기 중에서 전달되어 온 음성신호를 마이크를 통하여 받아 들인 후 이를 처리하여 그 신호 내에 포함된 음성언어를 문자의 형태로 변환하는 과정까지를 의미한다. 물론 사람은 그 말 속에 내포된 의미를 파악하는 과정을 포함하는 넓은 의미의 음성이해 과정을 수행하지만, 여기에서는 단지 좁은 의미의 음성인식 과정에 대해서만 살펴보기로 한다.

1. 간단한 음성인식시스템 구현 방법

컴퓨터가 사람의 말을 인식하는 과정은 일종의 패턴인식 과정으로 볼 수 있다. 즉, 사람이 발성하는 특정 단어의 신호를 패턴화하여 컴퓨터 메모리 안에 기억시킨 다음 새로운 음성이 입력되면 입력된 음성이 저장되어 있는 패턴들 중 어느 것과 가장 유사한 지를 판단해 내면 되는 것이다. 이러한 과정은 사람이 태어나서 말을 배울 때 여러가지 단어나 문법을 수많은 반복훈련 과정을 통해 두뇌 속에 패턴화하는 훈련과정과 훈련된 패턴을 새로 입력된 음성과 비교하여 입력 음성이 무엇인지를 알아내는 인식과정으로 크게 나눌 수 있다. 따라서 훈련과정과 인식과정은 서로 유기적인 관련을 가지고 있어서 그 방법상의 일관성이 유지되어야 하는 것은 당연하다.

이제 음성인식 과정의 이해를 돕기 위하여 우선 미리 결정된 소규모 어휘에 대해서만 고립단어의 형태로 컴퓨터에 입력할 때 이를 인식해 내는 비교적 간단한 고립단어 인식기를 예로 들어 음성인식 과정을 설명해 보기로 하자. 이러한 인식기의 구조의 예제를 〈그림 1〉에 도시하였다. 이 그림에서 볼 수 있듯이 인식기 전체 과정을 세 단계로 구분하였는데, 이것은 음성분석기, 음성인식 전처리기, 음성인식 후처리기 이다. 우선 음성분석기에서는 마이크를 통해 입력된 analog 음성신호를 digital 신호로 변환하는 digitization 과정과 입력된 신호 내에 음성이 존재하는 구간을 검출해 내는 speech detection 과정을 수행한다. 이 음성 검출과정은 다음단계에서의 패턴비교 과정에서의 연산량을 감축하고 더불어 패턴비교의 정확도를 향상시키는데 도움을 주기 위하여 사용한다. 다음으로 매우 중요한 과정이 음성신호가 가지고 있는 특성을 계수화하는 특징벡터 추출 과정이다. 이 과정은 보통 20~30 msec 단위로 음성신호를 블록화하여 각 frame별로 특징벡터를 계산해 내는데, 여기서 중요한 것은 이 frame별 특징벡터가 등록되어 있는 어휘를 구별하는데 도움이 되는 특성을 충분히 포함하고 있어야 한다는 것이다. 최근 널리 이용되고 있는 특징벡터 추출 방법으로는 LPC(Linear Predictive Coding)에 기반을 둔 cepstral 계수나 사람의 청각특성을 주파수 영역에서 고려한 MFCC (Mel-Frequency Cepstral Coefficients) 등이 있다.

이제 이렇게 얻어진 특징벡터 열(sequence)을 가지고 각 어휘의 표준패턴을 미리 컴퓨터에 저장해 두는 훈련과정을 수행한다. 이 과정은 그림에 별도로 도시되어 있지 않으나, 그림에서와 같은 인식기 구조를 갖기 위해서는 다음과 같은 훈련과정을 거쳐야 한다. 먼저 다양한 화자로 하여금 인식대상 어휘를 수회 반복 발성케 하여 훈련용 음성 DB를 구축하고 이로부터 각 고립단어별로 특징벡터 열을 추출해 놓는다. 다음으로 각 어휘별로 모아진 수십개 혹은 수백개의 음성 특징벡터를 벡터양자화하여 각 어휘 별로 codebook을 만들어 이를 인식기 전처리기의 표준패턴으로 저장한다. 특기할 점은 이 표준패턴에는 특징벡터의 시간축 상에서의 변화정보를 가지고 있지 않아서 정확한 패턴비교에는 사용할 수 없고, 다만 수많은 인식대상 어휘들 중에서 가능성이 높은 몇개 혹은 몇십개 정도의 어휘를 추출해 내는데 이용한다는 것이다. 다음으로 보다 정밀한 패턴비교에 사용될 각 어휘별 표준패턴을 결정하기 위하여 시간축 상에서의 frame 수의 불일치를 극복하고 및 시간축 상에서의 변화정보를 효과적으로 이용하기 위하여 동적 프로그램의 일종인 DTW(Dynamic Time Warping) 기



〈그림 1〉 소어휘 고립단어 인식기 구현 예제

법을 각 어휘별 표준패턴 훈련과정에 적용한다. 이 방법도 기본적으로는 벡터양자화 과정의 일종으로 볼 수 있으며, 단지 전처리 과정과의 차이는 패턴간의 유사도(similarity)를 계산할 때 시간축 상에서의 효과적인 mapping을 위하여 DTW 기법을 추가로 사용한다는 점이다.

각 어휘별로 훈련과정을 통하여 얻은 전처리기에서의 표준패턴인 VQ codebook과 후처리기에서의 표준패턴인 DTW reference pattern들을 가지고 새로 입력된 미지의 음성패턴과 패턴비교 과정을 수행하는 것이 음성인식의 핵심 과정이다. 우선 전처리기에서는 입력된 음성의 시간축 상에서의 변화정보를 무시하고 단지 각 어휘별 codebook과의 총 distortion의 합이 최소가 되는 순서대로 정해진 개수 만큼의 후보 어휘들을 결정한다. 다음으로 후처리기에서는 선택된 후보 어휘들의 DTW reference pattern과 입력음성 패턴을 DTW 방식을 사용하여 정밀한 distortion을 구해서 이들 중 가장 distortion이 작은 값을 가지는 어휘를 최종적으로 인식된 어휘로 결정한다.

여기에서 예로 들은 방법보다 더욱 간단한 인식기를 구현하는 것도 가능한데, 그것은 이 그림에 있는 인식기 전처리기 및 후처리기 중 한가지만을 사용하여 인식 결과를 결정해 버리는 방법이다. 이러한 방법도 인식 대상 어휘 수가 매우 적은 경우에는 비교적 높은 정확도를 가질 수 있다. 이러한 방법들은 주로 1980년대 중반까지 집중적으로 연구되어 매우 다양한 방법들이 개발되었으며, 이러한 방법을 토대로 최근 휴대폰 내부에서의 음성인식 다이얼링 등에 적용되어 상용화된 사례가 있다.

2. 고성능 음성인식시스템 구현 방법

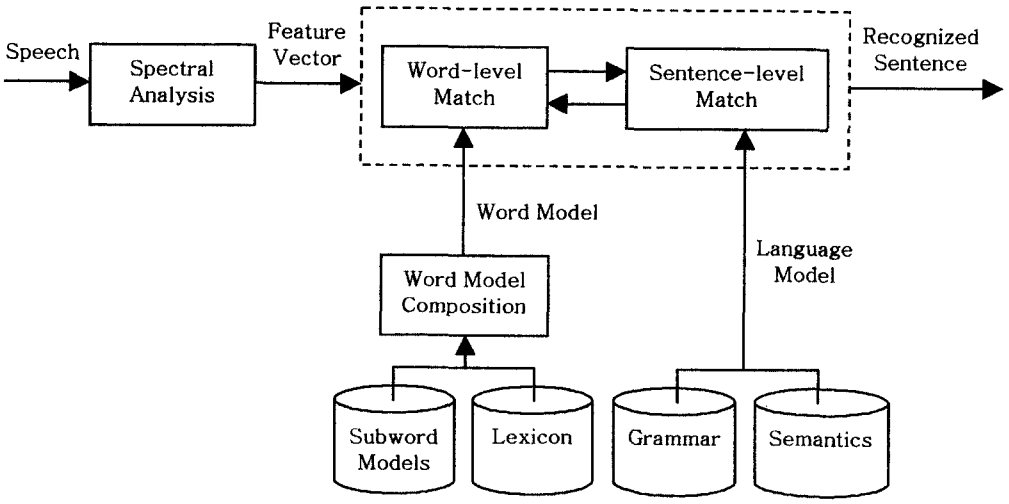
앞에서 예로 들은 소어휘 고립단어 인식기의 구조는 인식대상 어휘 수를 크게 확대하거나 입력음성으로 다양한 문법적 구조를 가질 수 있는 연속음성을 처리하고자 할 때 인식기의 정확도가 크게 저하되는 단점을 가지고 있다. 이를 타개하기 위하여 제안된 인식기 구조가 〈그림 2〉에 그려져 있다. 이 인식기와 앞서 예시된 인식기와의 차이를 정리하면 다음과 같다.

각 인식대상 어휘의 표준패턴을 표현하기 위하여 단어보다 작은 단위인 sub-word를 기본 패턴 단위로 사용하고 어휘를 표현하기 위하여 발음사전(lexicon)을 사용하는 것이 첫번째 차이점이다. 이러한 sub-word 단위의 예로는 음소나 음절 등과 같이 발음사전과 유기적으로 연관될 수 있는 단어 이하의 단위이면 어느것이나 가능하다. 따라서 이제는 각 어휘의 표준패턴이 sub-word model과 lexicon으로 대표되며, sub-word model들은 훈련과정을 통하여 결정하고, lexicon은 발음규칙이나 별도의 작업으로 생성시켜 저장해 놓는다.

또 한가지 큰 차이점이 문법적 제약을 가지는 연속음성을 인식해 내기 위하여 채용한 언어모델(language model)이다. 이 언어모델에서는 각 어휘들이 주어진 인식대상 영역에서 어떤 문법 및 의미적 제약을 가져야 하는지를 각종 언어모델링 기법(예 : Finite State Grammar, Context-Free Grammar, Stochastic Grammar 등)을 사용하여 훈련과정 혹은 수작업으로 정의하여 사용하는 것이다.

마지막으로 중요한 차이점이 입력된 미지 음성의 특징벡터와 인식 대상인 연속 어휘 패턴들을 효율적으로 비교하는 검색과정(search) 이다. 이 과정에서는 단어 레벨과 문장 레벨의 패턴 정보가 서로 결합하여 최종적인 인식 문장을 찾아낼 수 있도록 검색공간(search space)을 구성하고 검색한다. 가장 간단한 검색 방법으로 Viterbi beam search 기법이 있으며 이외에도 N-best 결과나 word lattice 결과를 얻기 위한 다양한 방법들이 제안되어 왔다.

특히, sub-word 단위를 모델링 하는데 사용하는 기법으로 최근에는 HMM(Hidden Markov Model)이라는 기법을 주로 사용하는데, 이것은 각 단위 음성을 몇 개의 state sequence로 정의하고 각 state는 각 음성 segment의 특징벡터가 발생하는 확률적 분포로 정의하는 방법이다. 이러한 모델링을 가능케 하는 훈련방법의 대표적인 것으로 maximum likelihood에 기반을 둔



〈그림 2〉 대어휘 연속음성 인식기 기본 구조

forward-backward estimation algorithm과 segmental k-means algorithm이 있다. 이러한 방법들을 적용함으로써 최근에 미국등에서 상품으로 출시된 음성타자기 등이 개발되었고, 현재도 연구개발의 주류를 형성하고 있다.

3. 음성인식시스템의 상용화를 제약하는 요인

최근 10여년간 음성인식 시스템 구현에 필요한 각 모듈별 요소기술이 급속히 발전해 와서 현재는 수십 어휘에서 수만 어휘까지 처리할 수 있는 다양한 음성인식 응용분야에서의 상용 시스템 개발이 크게 진척되었지만, 아직도 이를 현장에서 최종 사용자가 이용하는 데는 극복해야 할 많은 문제점들이 남아 있고, 이것이 상용화 시스템의 확산에 걸림돌이 되고 있다. 이러한 제약 요인들을 요약하면 다음과 같다.

(1) 음성인식의 성능을 저하시키는 요인
- Additive noise
  - 마이크를 통해 입력되는 신호에 부가적으로 포함된 비음성 및 타인의 음성 잡음으로 인한 SNR 저하
  - 원거리에 있는 마이크를 통해 음성을 입력할 때 불가피하게 수반되는 급격한 SNR 저하
  - 전송 채널상에 개입되는 잡음으로 인한 SNR 저하
- Convolutive noise
  - 유선전화에서 4-to-2 hybrid connection의 electrical echo로 인한 음성 왜곡
  - Loud speaker를 통해 출력된 음성이 실내에서의 반향으로 마이크로 다시 입력되어 발생되는 acoustical echo로 인한 음성 왜곡
  - 마이크 및 사운드카드의 상이한 주파수 특성으로 인한 음성 왜곡
  - 전송 채널의 주파수 응답 특성의 차이로 인한 음성 왜곡
- 미리 등록되어 있지 않은 어휘를 신뢰성 있게 제거 시킬 수 있는 미등록어 제거 알고리즘의 불완전성

(2) 음성 DB 규격화 및 시스템 인터페이스 표준화 미흡
- 음성 DB를 구축할 때 적용될 규격이 공통의 형식을 가지고 있지 못해서 그 동안 축적된 수많은 DB의 효율적 활용 및 재사용이 어려움
- 음성인식 엔진과 이를 이용하는 시스템과의 일관된 인터페이스 규격이 없어서 응용 제품 개발자가 쉽게 음성인식 기능을 채용하기가 어려움. 최근 Microsoft사에서 SAPI나 TAPI 등을 제안하여 산업표준으로 자리잡아 가고 있고, AT&T, Lucent, IBM, Motorola 등을 중심으로 음성인식 기능을 쉽게 웹 페이지 설계에 적용할 수 있는 VXML 표준화 작업이 활발히 진행되고 있으나, 아직 개선해야 할 점이 많음
- 음성인식기 내부의 각 기능 모듈별 방식 및 알고리즘 표준화가 아직 시기상조이어서 이 기술의 급속한 확산을 제약하고 있음

(3) 개발된 음성인식 시스템의 객관적 성능 평가 방법이 정립되지 못하여 제품들간의 공정한 성능 평가가 불가능 하다.

(4) 최종 사용자의 편의성 측면에서의 제약 요인
- 현재의 기술 수준으로는 최종 사용자가 구입한 음성인식기의 성능을 최대한 높이기 위해서는 최종 사용자의 음성으로 인식기를 재훈련해야 하는데, 이것이 사용자로 하여금 불편함을 느끼게 하는 요인이 되고 있음.
- 현재의 인식기는 우리가 알고 있는 모든 어휘를 처리하는 무제한 음성인식이 불가능하므로, 결국 사용자는 이 시스템이 어떤 어휘를 인식할 수 있는지를 항상 기억하고 있어야 한다는 불편함이 있음.
- 또한, 음성인식기의 성능을 최대한 높은 수준에서 유지하기 위해서는 이용자의 발음태도 및 발음속도가 어떤 범위 내에서 매우 협조적이어야 한다는 제약이 있음.

Ⅲ. 최근 동향 및 향후 과제

1. 최근 기술개발 동향

최근의 음성인식 기술 개발은 앞서 기술한 상용화 제약 요인들을 극복하기 위한 연구에 집중되고 있다. 즉, 각종 additive 혹은 convolutive noise가 개입되어 있는 음성신호에서 음성인식기의 성능 향상을 위하여 microphone array를 이용한다든지 적응신호처리 기법을 이용하여 잡음을 제거하는 방법 등 많은 연구가 진행되고 있다. 또한, 미등록어 제거를 정확히 처리해 주는 각종 알고리즘 개발도 최근의 중요한 연구 분야 이다. 한편, 음성 DB 규격화, 시스템 인터페이스 표준화 및 객관적 성능 평가 방법의 개발을 위하여 최근 국내에서도 정보통신부나 산업자원부 등에서 국가적 차원의 활발한 연구 지원이 시작되고 있다.

이와 같은 연구 개발 분야는 현재 개발된 인식 알고리즘에 부가적으로 사용하여 인식기의 상용화를 촉진하는데 초점이 맞추어져 있지만, 실시간 고성능 대어휘 음성인식을 가능케 하기 위한 요소기술, 특히 변별적 능력이 탁월한 음성 특징 벡터 추출 방법과 음향모델 훈련 방법이나 검색 공간을 대폭 축소할 수 있는 언어모델 훈련 방법 등에 대한 연구도 지속적으로 이루어지고 있다.

2. 향후 과제

궁극적인 의미에서의 음성인식, 즉 사람의 능력에 버금가는 음성인식 기술을 개발하기 위해서는 결국 사람의 두뇌 속에서 음성언어가 어떻게 생성되고, 이것이 어떻게 발성기관을 동작시키는지를 발견해 내어야 하며, 또한 청각기관이 전달되어 온 음성신호를 어떻게 분석하고 이를 두뇌가 어떤 방법으로 이해하는지를 분석해야 한다. 이와 더불어 사람과 사람 사이의 대화에서는 음성 뿐만 아니라 다른 정보, 즉 동작이나 표정 등과 같은 다른 modality를 함께 사용하므로 이를 이용하는 multi-modal interface에 대한 연구도 수행되어야 한다. 또한, 음성인식기가 어떤 특정 언어에만 잘 동작하지 않고 다국어 음성입력에 대해서도 우수한 성능을 갖도록 하기 위해서는 multi-lingual capability도 가져야 하므로 이에 대한 연구도 향후 지속적으로 수행되어야 한다. 좀더 실질적인 측면으로는, 최근 급속히 확산되고 있는 인터넷 환경에서 음성신호를 처리할 수 있는 음성인식 방법에 대한 연구의 필요성도 급속히 증대되고 있고, 대규모 음성 DB, 예를 들면 방송 음성을 DB화하고 이를 효율적으로 검색할 수 있는 audio indexing & retrieval을 음성으로 가능케 하는 음성인식에 대한 연구 등 향후 연구 과제는 무궁무진하다 하겠다.

Ⅳ. 결 론

본 고에서는 음성인식 기술의 개요 및 향후 과제를 개괄적으로 살펴 보았다. 요약하면, 지난 30여년간 지속적인 기술의 발전이 이루어져 와서 현재는 비록 제한된 분야이기는 하지만 상용화 시스템 및 서비스가 우리 주변에서 급속히 확산되고 있다고 할 수 있다. 하지만, 아직 극복해야 될 수많은 과제를 안고 있어서 상용 시스템 개발과 병행하여 핵심 요소기술에 대한 지속적인 연구가 절실한 형편이다. 또한, 이제 막 출시되고 있는 음성인식 기반 서비스를 이용할 때에도 상기한 여러 가지 제약들을 고려하여 이용하면 나름대로 큰 편리성을 얻을 수 있을 것으로 생각된다.

참 고 문 헌

〔1〕 L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Pretice-Hall, 1993.

〔2〕 Hang-Seop Lee, Hoi-Rin Kim, "Speech Web browser using variable vocabulary word recognition," Proc. of AV-

IOS'97, San Jose, 1997.
〔3〕 김회린, 이항섭, "음성학적 지식 기반 변이음 모델을 이용한 가변 어휘 단어 인식기," 한국음향학회 논문지, 제16권, 제2호, pp. 31-35, 1997.
〔4〕 김회린, 이영직, "Voice Interface 및 인식," 정보처리학회지, 제5권, 제1호, pp. 42-48, 1998.
〔5〕 B. H. Juang, S. Furui, et al. , Special Issue on Spoken Language Processing, in Proceedings of the IEEE, Aug. , 2000.
〔6〕 Kwang-Sik Moon, Yu-Jin Kim, Hoi-Rin Kim, Jae-Ho Chung, "Out-of-vocabulary word rejection algorithm in Korean variable vocabulary word recognition, Proc. of ISCAS2000, pp. V-53-56, Geneva, 2000
〔7〕 S. Furui, et al., Proc. of HSC2001, Kyoto, Apr., 2001.

저 자 소 개

金 會 麟

1961년 3월 9일생, 1984년 2월 한양대학교 전자공학과 졸업, 1987년 2월 한국과학기술원 전기 및 전자공학과 졸업(석사), 1992년 2월 한국과학기술원 전기 및 전자공학과 졸업(박사), 1994년 6월~1995년 5월 : 일본 ATR연구소 방문연구원, 1987년 10월~1999년 12월 : 한국전자통신연구원 음성언어팀 선임연구원, 2000년 1월~현재 : 한국정보통신대학원대학교 공학부 조교수, 〈주관심 분야 : 음성인식, 음성합성, 자동통역, 음성코딩〉