□ 기술해설 □

음성정보처리 기술을 이용한 정보검색

한국과학기술원 오 영 환*

목 차

1. 서 론

현대 정보화 사회에서는 많은 정보가 매일 쏟아져 나오고 있다. 정보고속도로의 실현과 통신위성의 발전에 따라, 언제, 어느 곳에서나 원하는 정보를 손끝에서(Information at your fingertips) 얻을 수 있는 시대가 도래하리라 기대되고 있으며, 사무실, 가정, 사회복지, 엔터테인먼트, 학교, 예술 등 모든 분야에서 정보화 사회로의 이행이 가속화되고 있다. 이러한 지식과 정보의 홍수 속에서는 각각의 개별적 사실에 대한 지식보다는 필요한 정보가 위치하는 장소를 알고, 이에 접근할 수 있는 정보 검색 능력이 보다 중요시되며, 정보화 사회의 혜택을 모든 사람이 누리기 위해서는 자연스럽고 편리한 정보 입출력 방법의 개발이 요구된다.

정보화 사회의 핵심 기술인 컴퓨터가 출현한 지는 근 반세기가 지났으며, 눈부신 기술 발전이 이루어졌으나, 아직도 사람들이 쓰기에 편하고 친근감을 느끼기에는 여러 가지 부족한 면이 있다. 가까운 예로, 초창기부터 현재에 이르기까지 컴퓨터와 인간의 의사소통의 장치로 쓰여온 자판과 영상화면단말기(visual display terminal: VDT)는 이동이 불편하고, 장시간 사용할 시에 눈, 어깨, 손목 등의 신체적인 장애와 정신적, 생리적 장애 등을 야기시키며, 일정 수준의 교육을 받지 않고서는 사용에 제한을 받는 단점이 있다. 이러한 단점을 극복하고자 휴대형 개인정보단말기(PDA)에 적용되는 펜입력장치 등이 등장했으나, 사람들끼리 의사를 소통할 때와 마찬가지로 말, 표정, 몸짓 등 사람에게 익숙한 수단을 그대로 사용할 수 있는 보다 편리한 입출력 장치의 출현에 많은 관심이 모아지고 있는 실정이다.

음성은 가장 자연스럽고 널리 쓰이는 사람들 사이의 의사소통 수단으로서, 이를 이용한 기계와의 의사소통에 많은 연구가 수행되고 있다. 음성을 이용한 정보 입출력을 위해서 인간이 발성한 음성 속에 내재되어 있는 언어정보를 자동으로 추출하여 인식하는 음성인식 기술과 주어진 문서를 인간의 말로 변환하는 음성합성 기술을 기계에 부여하여야 한다. 음성을 사용한 정보입출력은 다음과 같은 장점을 가진다[1].

첫째, 음성은 인간의 기본적이고 자연스런 의

*종신회원

사소통 수단이므로 특별한 훈련이 필요없이 기계의 사용이 가능하므로, 보다 편리하고 자연스러운 정보 입출력이 가능하다.

둘째, 음성을 통한 정보 전달과정 중에도 동시에 손, 발, 눈의 사용이 가능하므로 병렬적인 작업이 가능하다.

셋째, 음성은 글로 쓰는 것보다는 8~10배, 키보드에 비해 3~4배 정도 빠르므로 신속한 입력이 가능하다.

넷째, 전화선을 이용한 원격지 정보입력과 출력이 가능하므로, 시간과 장소에 따른 제약을 완화시킨다.

이러한 장점으로 인해 여러 선진국에서는 자국어의 음성인식과 음성합성에 대한 연구가 지난 수십 년간 진행되어 이제 실험실 수준에서의 단계를 지나 다양한 형태의 상품으로 나타나고 있다. 본 고에서는 음성을 이용한 정보검색의 기반기술인 음성인식과 음성합성의 연구 현황과 이를 정보 검색에 적용한 예를 살펴보고자 한다. 2장과 3장에서 음성인식과 음성합성의 기본 과정과 이에 관련된 기술 동향을 알아보고, 4장에서 음성정보처리를 정보검색에 이용한 실례를 살펴본 후, 5장에서 결론을 맺고자 한다.

2. 음성 인식의 기본과정 및 기술 동향

인간이 발성한 음성 속에 내재되어 있는 언어정보를 자동으로 추출하여 인식하는 음성인식 기술에는 여러 가지 기법이 사용되고 있다. 현재에 주로 사용되는 음성인식 기술은 통계적인 패턴정합 접근방법으로서 전형적인 시스템의 개요도는 그림 1과 같다.

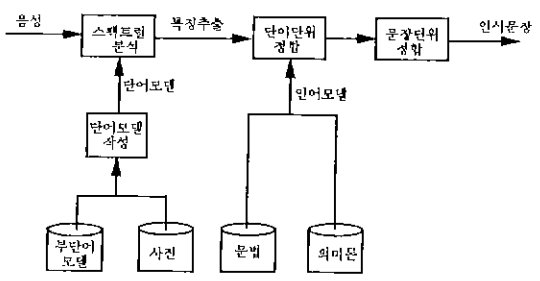

그림 1 연속음석인식기의 개요도

그림 1의 시스템은 스펙트럴 분석 단계에서는 입력된 음성의 특징을 나타내는 스펙트럴 특징을 추출하고, 단어 단위의 정합에서 사전에 따라 부단어(subword) 모델을 연결한 단어모델을 작성한 후, 이를 입력 특징벡터와 정합하여 가장 유사한 단어를 찾는다. 문장 단계의 정합에서는 언어 모델을 이용하여 가장 높은 확률로 발생할 단어 열을 구한다. 구문론적인 규칙과 의미론적인 규칙은 사람이 직접 작성하거나, 과제의 제한(task constraint)이나 통계적인 언어모델에 따라 구성된다. 탐색과 인식 단어열의 결정은 모든 단어 열을 고려한 후, 가장 가능성이 높은 단어 열로 선택된다. 다음절에서는 음성인식의 각단계별 설명과 기술동향을 살펴보기로 하겠다.

2.1 특징 추출

음성으로부터 유효한 특징의 추출을 위해 지난 50여년간 많은 연구가 진행되어 왔지만, 최적의 특징 추출 방법에 대해서는 여러가지 의견이 있다. 유효한 특징 추출 방법이 가져야 할 조건으로는 유사한 음성을 구별해낼 수 있는 판별능력이 있어야 하고, 발성자와 시간에 따른 변이가 적어야 하며, 인지 및 발성 모델로도 설명이 가능하여야 한다. 대부분의 시스템은 DFT(discrete Fourier transformation)나 LPC(linear predictive coding)에 기반한 특징추출 방법을 사용한다. 가장 널리 사용되는 특징은 켑스트럼과 음성의 동적인 특성을 반영하기 위해서 켑스트럼의 미분 값들을 사용하는 것이다[2,3,4]. 또한 사람이 주파수 정보를 인지할 때의 특성에 따라 각 주파수 대역마다의 가중을 하는 mel 또는 bark 간격의 주파수 분석법을 사용한다[5,6].

위의 특징추출 방법 이외에도 아래와 같은 방법들이 연구되고 있다. 고정된 길이의 음성에 대한 동일한 주파수 해상도를 갖는 단점을 극복하고자 사용되는 시간-주파수 분석방법(time-frequency analysis), 인간의 청각특성을 이용한 분석 방법, 음성의 조음 과정에 기반한 특징의 추출방법, 운율이나 지속시간등의 정보를 이용하는 방법, 판별분석을 이용하여 인식에 유효한 특징만을 추출하는 방법 등이 연구되고 있다.

2.2 인식단위

그림 1에서 보는바와 같이 대부분의 시스템은 부단어를 인식단위로 하여 단어사전에 따라 부단어를 연결하여 입력 음성과의 정합도를 구한다. 따라서 적절한 기본 인식단위의 선정은 매우 중요하다. 인식단위를 선택하는 데는 다음 세 가지를 고려하여야 한다.

첫째는 음소가 환경에 따라 변화하는 조음결합 현상을 흡수할 수 있고(sensitivity), 둘째는 적은 자료로 충분히 학습시킬 수 있고(trainability), 셋째는 인식 단위가 발성시마다 일관된 성질(consistency)을 가지고 있어야 한다.

인식단위로는 단어, 음소, triphone 등이 사용된다. 단어는 조음결합 현상을 포함하여 음향학적 변화를 흡수하므로, 단어를 단위로 하는 인식 시스템은 가장 좋은 성능을 낸다. 그렇지만, 많은 어휘를 필요로 하는 응용분야에서는 각단어가 충분히 학습할 만큼 자주 발생하지 않으므로 학습성(trainability)의 문제가 있고, 새로운 단어의 추가시 그 단어에 대한 많은 자료가 필요하므로 대용량의 음성인식에서는 사용되지 않는다. 음소를 인식단위로 할 경우, 음소는 그 수가 적으므로 적은 자료로도 학습이 가능한 장점이 있으나 문맥에 따른 조음결합 현상을 모델링하지 못하는 민감도(sensitivity)의 문제점을 안고 있다. 이러한 문제 때문에 최근의 시스템들은 triphone 등의 부단어를 인식단위로 사용한다. Triphone은 문맥 종속적인 음소로서, 같은 음소라도 앞뒤 문맥에 따라 서로 다른 단위가 되므로 문맥에 따른 조음결합 현상을 모델링 할 수 있고 단어에 비해 학습성(trainabiliy)면에서 유리하다. 그러나 triphone도 수가 많아 각 단위를 충분히 학습할 수 없으므로 서로 유사한 triphone들을 묶어 하나로 취급하는 generalized triphone을 사용하고 있다.

2.3 인식 알고리즘

음성 인식에 사용되는 방법은 동적정합법(dynamic matching) HMM(hidden Markov model), 신경 회로망, 전문가의 지식을 이용한 지식 공학적 방법 등 다양하다.

동적정합법은 음성인식에 동적 프로그래밍을 적용한 방법이다[7]. 같은 단어를 발성할 경우라도 화자, 감정, 주변환경에 따라 각기 다른 지속시간을 갖는다. 동적정합법은 기준패턴과 입력되는 음성과의 길이가 다른 경우 두 패턴 사이의 거리를 측정하기 위해서 기준음성 패턴의 각 프레임과 그에 대응하는 입력음성 패턴의 프레임간을 동적 프로그래밍의 기법을 사용하여 거리를 구한다. 동적정합법은 인식대상어휘가 작은 고립단어 인식에 주로 이용되며, VLSI의 발전에 따라 칩으로 구현되어 상용화되어 있다. 또한 기준 패턴을 쉽게 만들 수 있기 때문에 사용자의 요구에 따라 음성인식 시스템의 업무 내용을 용이하게 변경할 수 있다.

HMM은 높은 인식 율과 편리한 학습성으로 음성인식에 가장 널리 쓰이는 방법으로서 음성의 시간적 변화를 모델링하는 천이확률과 스펙트럴 변화를 모델링하는 출력확률로 구성된다[8]. HMM의 모델파라미터의 추정을 위해서 여러가지 방법이 제안되고 있다. ML(maximum likelihood) 추정을 이용하고 자동적인 학습방법인 Segmental ML을 이용하는 방법이 가장 널리 쓰인다[9]. ML방법은 안정적인 모델파라미터의 추정을 위해서 많은 자료를 필요로 하므로 deleted interpolation 이나 Bayesian smoothing 등의 평활화 방법이 사용되기도 한다[10]. MMI(maximum mutual information)추정 방법은 ML방법이 모델에 해당되는 자료만 이용하는 반면에, 전체 자료의 정보를 이용하여 상호정보를 최대화하는 파라미터를 얻는다[11]. 이밖에도 새로운 화자와 환경에 적응하기 위해 사후확률(maximum a posteriori)추정법과 전체적인 확률을 최대화하지 않고, 인식 오류를 최소화하는 최소분류에러(minimum classification error)방법이 사용되고 있다[12].

신경회로망은 인간의 뇌세포를 간단히 모델링하고 이들을 연결시켜줌으로써 인간의 뇌가 하는 역할을 수행시켜주는 알고리즘이다. 음성인식에서는 음성의 정적인 특성뿐만 아니라 시간적인 변화를 모델링해야 하므로 다층퍼셉트(multilayer perceptron)를 변형시킨 시간지연신경망

(time delay neural network)과[13], 동적정합법과 다층퍼셉트를 이용하여 패턴을 비선형 예측하는 신경예측모델(neural prediction model) 등이 이용되고 있다[14]. 이밖에 순환(Recurrent) 신경회로망은 회로망의 입력과 출력단 사이에 궤환(feed back)연결이 존재하여 시간에 따라 그 특성이 변하는 구조로서, 궤환 연결선은 현재 입력에 과거의 상태를 반영하여 분류될 수 있는 효과를 지닌다[15]. 최근에는 신경회로망만을 음성인식에 적용하기보다는 HMM이나 기존의 알고리즘에 신경회로만을 결합하는 방식이 연구되고 있다.

지식 공학적 방법은 전문가의 경험적 지식을 추론 규칙의 형태로 표현하여 지식베이스화하여 시스템을 구성한다[16,17]. 주로 주파수 스펙트럼으로부터 각 음소의 고유한 음성학상의 특성을 기술하고, 이것을 기초로하여 인식 규칙(rule)을 작성한다. 지식 공학적인 방법은 종래의 음성연구의 흐름에 충실하여, 음성연구에 의한 지식의 축적에 크게 의존하는 전통적인 시스템이라 할 수 있으나, 자료에 기반한 방법들인 HMM이나 신경회로망에 비해 인식율은 저조한 편이다.

2.4 단어 사전 구성

그림 1의 음성인식 개요도에서 단어모델은 사전을 참조하여 얻은 부단어 열을 연결하여 구성된다. 사전의 구성은 표준적인 사전으로부터 하나의 발음만을 이용하여 구성하는 방법이 있다. 이는 문자나 음성자료의 특성을 반영하지 않는 자료-독립적인 방법이다. 자료-종속적인 사전구성은 자료의 특성이 사전의 구성에 반영되게 하는 방법으로서 다음과 같다.

첫째, 연속음성의 경우 단어의 경계 등에서 단어의 발음은 크게 변화한다. 따라서 복수의 발음 또는 발음 망(pronunciation network)으로 사전을 구성하는 방법[18], 둘째, 언어적인 표기에 의한 사전은 음성의 모든 음향학적인 특성을 표현할수 없으므로 음성언어 사전을 음성자료로부터 직접 학습시키는 방법[18,19], 셋째, 학습 자료의 음향학적인 특성은 모든 단어의 특성을 포함할 수 없으므로 음운론의 규칙으로부터 통계적 또는 결정론적인 방법에 의해 복수의 발음 또는 발음 망을 구성하는 방법이 사용되고 있다[20,21].

정확한 발음사전의 구성을 위해서는 음성의 음향학적 실현과 언어적인 표기가 일치하여야 한다. 이를 위해서 사전의 변이를 직접적으로 나타내는 확률적인 단어 모델링이 연구되고 있다.

2.5 언어 모델

문장을 인식할 경우에는 문법을 사용하여 인식된 문장이 적합한지를 결정하여야 한다. 구문론 적인 규칙과 의미론적인 규칙은 인식대상과제에 따라 결정되어 지며, 이러한 언어정보를 이용해서 신호처리 단계에서 얻어지는 불완전한 인식 결과를 보정한다. 언어처리 알고리즘은 문법을 단어인식기와 결합하는 방법에 따라 통계적 모델과 구문규칙 모델로 나눌 수 있다. 통계적 모델은 단어와 단어사이의 연관관계를 확률로 표시하는 bigram과 trigram이 있고, 이는 HMM 모델과 쉽게 결합되어 대용량 어휘의 인식과제에 널리 이용되고 있다[22]. 구문 규칙 방식은 언어학에서 연구된 구문론에 의한 규칙에 의해 매 단어 다음에 연결될 수 있는 단어의 종류를 제한함으로서 문장을 인식하는 방법이다[23]. 대용량의 음성인식을 위해서는 보다 발전된 언어모델이 필요하다. 특히 많은 자료로부터 얻어진 언어모델이라도 다른 인식과제로 대상이 바뀌면 잘 적용되지 않으므로 새로운 과제로 기존의 언어모델을 적응시키는 방법에 대한 연구가 필요하다.

2.6 탐색 및 인식 문장 결정 방법

인식 문장을 찾기 위한 방법은 통합적(integrated)과 모듈적(modular) 접근으로 나누어진다. 통합적 방법은 음향분석, 사전, 구문, 의미적인 지식 원을 하나의 유한 상태망으로 묶어서 사용하는 방법으로 현재의 음성인식에 사용되고 있다. 그러나 보다 전체적인 정보인 운율과 trigram 확률을 유한상태망에 포함시키기 어렵고, 대용

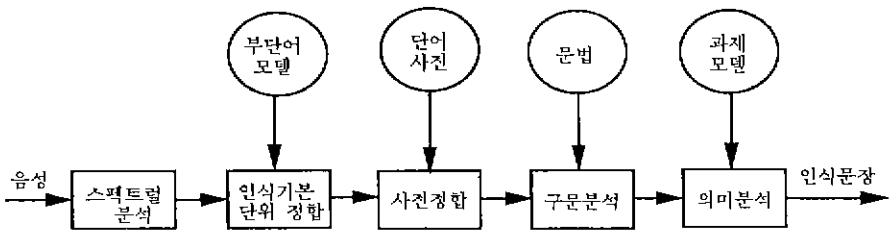

그림 2 모듈적 연속음성인식기의 개요도

량의 음성인식의 경우 모든 지식원을 통합한 유한상태망은 너무 방대하여 탐색이 불가능하다. 모듈적 접근 방법은 그림 2에 보는 바와 같이 기본 단위의 정합, 사전 정합, 구문 및 의미 분석이 독립적으로 행해지므로, 각 분야가 독립적으로 설계되고 평가될 수 있으며, 여러 분야의 전문가가 협력하여 시스템을 구성할 수 있다.

모듈 구조의 탐색방법은 계산량이 적은 반면에 다른 지식원을 이용할 수 없다는 약점이 있다.

탐색방법의 예를 들면 다음과 같다. 프레임 동기 빔(Frame-Synchronous Beam) 탐색은 모든 지식원을 이용하여 유한상태망을 구성한다. 계산량을 감소시키기 위해 목 구조 사전(tree lexicon)과 phone look-ahead기술을 사용한다[24]. 음성신호에 내재된 언어적인 정보는 국부적인 방법(localized manner)으로 나타나므로, 모든 언어적 사건을 매시간마다 고려할 필요는 없다. 이러한 특성을 이용한 best-first 탐색전략을 사용하는 방법이 Stack decoding과 A* 탐색이다[25]. Multi-Pass 결정전략을 사용하는 tree-trellis와 전향 후향(forward-backward) 탐색은 하나의 경로에 따른 탐색을 하는 기존의 방법에 비해, 처음에는 여러 경로에 따른 탐색을 수행한 후 마지막에 세밀한 언어모델과 정합방법을 이용하여 최종 인식문자를 결정하는 방법이다[26].

3. 음성합성의 기본과정 및 기술 동향

음성합성 기술은 입력된 문장에 대해 음운변환 및 운율제어 규칙 등을 적용하여 음성신호로 변환하는 장치로, 음성정보 서비스의 활성화로 주목받고 있는 기술의 하나이다. 음성합성 기술의

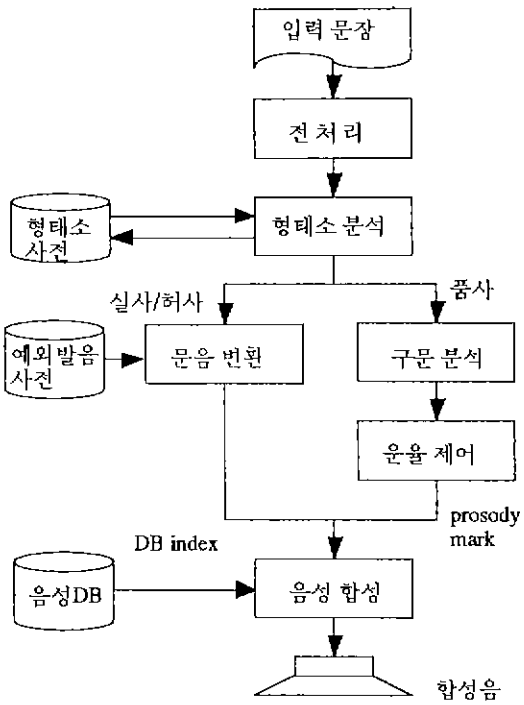

그림 3 문서-음성 변환 시스템의 구성

발전에 힘입어 오디오텍스(Audio-tex), 자동음성응답 시스템(ARS), 음성우편 시스템(VMS) 등이 개발 실용화되고 있어 음성정보서비스의 대중화가 점차 이루어지고 있다[30]. 음성합성의 기본과정은 크게 언어처리부, 운율제어부, 음성생성부의 3부분으로 구성된다[그림 3].

언어처리부에서는 입력된 문장에 포함된 기호나 약어 등을 발음에 알맞은 구술적인 표현으로 변환시킨 후 구문론과 관련된 사전 정보에 의해 문장의 구조를 분석하고 의미를 파악한다. 운율제어부에서는 언어 처리부에서 추출해낸 정보에

의해 끊어 읽기, 강세, 억양 등을 지정하는 운율 기호를 결정한다. 마지막으로 음성생성부에서는 전단에서 입력된 음운기호열과 운율기호열을 조합하여 음성출력을 내보낸다. 문서-음성 변환 시스템은 위와 같은 3가지 요소들로 구성되며 합성을 위한 기본단위, 코딩방식, 운율정보 추출 과정 등에 따라 여러가지 형태의 시스템으로 구현될 수 있다.

3.1 언어 처리부

인간이 문장을 읽을 경우 문장 전체의 의미나 내용을 파악하여야 하는 것처럼 음성합성에 있어서도 높은 품질을 얻기 위해서는 문장의 언어적 이해가 필수적이다. 언어 처리부에서는 입력된 문장에 포함된 기호나 약어 등을 발음에 알맞은 구술적인 표현으로 변환시킨 후 문자의 구조를 분석하고 의미를 파악한다. 문장의 구조를 이해하기 위해서는 형태소 분석, 구문 분석 등이 필요하며 운율 제어부에서는 이러한 정보를 이용하여 운율 파라미터를 추출한다.

3.1.1 형태소 분석

형태소(morpheme)란 일정한 뜻이 있는 언어의 가장 작은 단위, 즉 최소의 유의적 단위로 정의된다. 형태소 분석은 자연언어의 처리에 있어서 가장 기본적 분석과정으로 문장의 구조를 이해하기 위한 구문분석 과정은 물론 음운변환(grapheme-to-phoneme) 과정에서도 형태소 분석 결과를 이용하여 문장을 처리하므로 문서-음성 변환에 있어서 필수적인 단계이다[31,32]. 한국어에 있어서 형태소 분석이란 띄어쓰기의 단위인 어절의 구성요소 즉, 형태소를 밝히는 것으로 정의된다. 문장을 구성하는 최소단위인 어절은 형태소가 모여 이루어진 것으로, 형태소 분석시 그 어절을 구성할 수 있는 모든 가능한 형태소 조합을 찾게 된다. 한국어에는 약 50여개의 형태소 분류가 있으며 분류별로 사전에 저장되는데, 형태소 분석시 주어진 어절에 대하여 가능한 형태소를 찾기 위하여 빈번한 사전 검색이 요구된다. 그러나, 이때 입력된 문장의 모든 어휘가 사전에 등록되어 있다고는 할 수 없으므로 이에 대한 대처가 필요하다. 예를 들어 일반적인 문장에서는 복합어, 새로운 지명 및 인명 등 미지의 어구가 수시로 나타나는데, 보통의 지능을 가진 인간은 이러한 미지어(unknown-word)에 대해서도 문맥 및 과거의 지식에 의해 문법적 역할이나 의미를 추측하여 전체의 대의를 파악할 뿐만 아니라 이를 적절하게 발음할 수 있다. 따라서 문서- 음성 변환에서도 형태소 분석시 미지어에 대한 처리를 적절히 해주어야 할 필요가 있다.

3.1.2 구문 분석

인간이 긴 문장을 적절히 끊어 읽는 것은 문장의 구문구조를 분석해 낼 수 있기 때문이다. 문서-음성 변환 시스템에서 구문 분석기는 자연스러운 운율정보를 생성하기 위하여 문장의 구문정보를 추출해 내는 역할을 한다. 우리말에서의 구의 기능, 상호 결합관계 정보는 구의 마지막에 고정되어 있는 조사, 어미 등의 기능어 문법정보로부터 구할 수 있다. 이와 같은 점에 착안하여 복잡한 처리과정을 생략하고 조사와 어미 등의 형식형태소와 구문요소간의 결합규칙만을 이용하여 구문분석을 한 문서-음성 변환 시스템의 예도 있으나, 임의의 문장을 분석하는데는 미흡한 점이 있다[39].

자연어 처리분야에서 연구되고 있는 구문분석 방법으로는 기존에 많이 쓰여왔던 구-구조문법(phrase structure grammar)에 의한 분석 방법과 최근 한국어 분석에 이용되고 있는 의존문법(dependency grammar)에 의한 분석 방법이 있다. 이 가운데에서 의존문법에 의한 분석이 많은 주목을 받고 있는데, 그 이유는첫째 한국어에 있어서 어순의 자유성에 의한 문제점을 쉽게 해결할 수 있으며, 둘째 구성요소의 불연속성이나 생략 등과 같은 현상에 큰 영향을 받지 않아 매우 견고하다는 점이다.

따라서 구문분석에 의존문법을 이용한다면 한국어의 이해에 매우 효과적일 것으로 생각된다. 그러나, 자연언어 처리에서 사용했던 알고리즘을 그대로 음성합성에 적용하는데는 해결해야할 점들이 있다. 먼저, 자연언어 처리분야에서 구문분석기의 목적은 하나의 문장을 정확히 분석하는데

있다. 따라서 처리과정이 복잡하고 시간이 늦어지고, 어떤 문장은 하나 이상의 모호한 결과가 나오게 된다. 반면, 문서-음성 변환시스템의 파서는 주목적이 운율정보를 추출하는데 있고, 잘못된 문장이라도 실시간에 문서-음성 변환하여야 하므로 분석된 결과는 항상 하나가 나와야 하며, 잘못된 문장이라도 그 자체를 읽어주어야 한다. 따라서 문서-음성 변환을 위해서는 견고하고 실시간 처리가 가능한 구문분석 알고리즘의 개발이 시급하다.

3.2 운율 제어부

운율이란 발성시 나타나는 억양, 강세, 리듬 등의 특성을 말하는데 이는 기본주파수, 음소길이, 음량, 휴지기 길이 등에 의해 결정된다. 운율은 합성음의 이해도와 자연성에 중요한 요소로 작용하며 정보 전달에도 큰 영향을 끼치는데, 운율 제어부에서는 언어 처리부에서 분석된 결과들을 이용하여 운율 파라미터들을 제어한다. 사람이 한번 숨을 쉬어 발성하는 말의 단위를 발화단위라 하는데 발화 단위 내에서는 기본 주파수가 점차 낮아지는 경향을 갖는다. 이를 억양의 '기본 기울기'라 하며, 억양의 기본 기울기에 단어, 음절의 강세 및 문 구조에 따른 억양 패턴이 더해져서 전체 억양 패턴이 구성된다. 음소의 길이 및 휴지길이는 억양과 함께 합성음의 자연도를 결정하는 중요한 요소이다. 음소의 길이는 음소 자체의 성질뿐만 아니라 주변의 음소환경, 한 단어내의 음소 개수, 단어 내에서의 음소의 위치, 강세여부 등 다양한 요소에 의해 영향을 받는다. 휴지기 길이도 음소길이와 마찬가지로 전후의 음소환경에 의해 영향을 받게 되는데, 그 이외에 발화단위 사이에서 긴 휴지기가 존재한다. 발화단위는 하나의 발화단위 내의 어절개수, 음절개수 뿐만 아니라 문장의 구조 및 의미에 의해 결정된다. 일본의 경우 이미 1960년대에 Fujisaki에 의하여 기본 주파수 윤곽을 만들기 위한 모델이 제안되어 널리 쓰이고 있으나, 국내에서는 이러한 획기적인 모델의 제안은 아직 없으며 계속 연구 중에 있다.

3.3 음성 생성부

음성생성부에서는 언어 처리 부에서 변환된 합성 단위음 열과 운율 제어 부에서 생성된 운율기호로부터 실제 음성신호를 합성하게 되는데, 본 절에서는 음성생성과 관련하여 코딩 방법과 합성단위에 대하여 살펴보기로 한다.

3.3.1 코딩 방법

무제한 어휘 합성을 위한 대표적인 코딩방법으로는 포만트 합성방식과 LPC 계열의 분석합성방식이 있는데 2가지 모두 음원 코딩 법이라는 공통점을 가지고 있다. 1960년 Fant에 의해 음성생성의 디지탈 모델이 발표된 이후 음성합성에 큰 발전을 가져온 음원 코딩 법은 인간의 성도 특성을 모델링하여 특징 파라미터의 시간적 변화 정보에 의해 음성을 합성한다. 파형 코딩 법에 비해 연산량이 많고 음질도 떨어지나, 데이터 압축률이 높고 특히, 특징 파라미터의 변환에 따라 말의 속도, 음높이, 스펙트럼 변환 등이 용이하여 대부분의 무제한 어휘 합성에 적용되어 왔다. 먼저 포만트 합성방식은 순수 규칙합성 방식으로서 성도의 변화/필터특성은 각 음소 그리고 음소간의 포만트 변화를 나타내는 규칙을 사용하여 기술한다. 영어권에서는 이미 오래 전부터 Klatt형 합성기가 상용화되어 쓰이고 있다. 국내에서도 몇몇 대학 및 연구소를 중심으로 포만트 방식의 음성합성에 대한 연구가 이루어져 왔으나, 음성분석에 대한 많은 자료를 필요로 하기 때문에 음질개선을 위해서는 앞으로도 보다 많은 연구가 필요할 것으로 생각된다[33].

LPC 계열의 합성방식은 음성신호의 예측 계수를 이용하여 all-pole 성도모델 필터를 구성하고 음원 신호를 필터링하면 쉽게 음성신호를 합성해낼 수 있다는 장점이 있다. 국내의 경우 초창기에는 대부분의 연구기관에서 비교적 시스템의 구현이 간단하고 쉽다는 장점 때문에 이 방식을 택하였으나, 합성음질에 한계를 느끼고 지금은 별로 사용하지 않고 있다.

최근에는 음성신호를 파라미터화하지 않으면서 피치를 변화시킬 수 있는 PSOLA(Pitch Synchronous Overlap & Add) 방법이 개발되어 많은 연구기관에서 이를 문서-음성 변환에 이용하고 있다. PSOLA 방법은 음성신호의 각 피치 단위의

신호를 피크 값을 중심으로 분석 창을 이용해 음성소편(short term signal)을 만든 후, 합성시 각각의 음성소편을 중첩시켜 연속된 음성신호를 만드는데, 중첩 길이를 변화시킴으로서 피치 주기를 조절할 수 있다[34]. PSOLA 방법은 음성신호를 시간 영역에서 부호화 하므로 음원 코딩법에 의한 합성음에 비하여 음질이 좋고 합성시 연산량이 많지 않아 실시간 처리가 용이하다. 반면 합성시 필요한 데이타량이 많으며 음성소편을 자동적으로 분류해내기 어려우므로 음성소편 사전 구성시 수작업이 많이 필요한 단점이 있다.

3.3.2 음성합성 단위

음성합성 단위는 연속음을 내기 위한 합성의 기본 단위로써, 기존의 음성합성 단위에는 음소, diphone, 반음절, 음절 등이 있으며 이때 알고리즘의 복잡도, 데이터의 개수, 음질 등이 크게 영향을 받는다. 음소를 기본단위로 하여 음성을 합성할 경우 매우 적은 개수의 데이터로 무제한 음성합성을 할 수 있는 장점이 있으나 음소와 음소를 연결시키는 과정에서 일어나는 조음결합 현상을 나타내기 어려우므로 불연속이 생겨 합성음의 명료성이 떨어진다. diphone은 서로 연결할 경우 음소처럼 불연속이 생기지는 않으나 diphone 자체의 수가 상당히 많아 많은 기억용량을 필요로 하게 된다. 음절을 기본 단위로 할 경우 비교적 좋은 결과를 얻을 수 있으나 음절과 음절 사이에 일어나는 조음결합 현상과 천이구간상의 문제가 발생할 수 있고 데이터의 수가 많아지는 단점이 있다.

최근의 합성단위 선택의 동향은 음성의 전후 환경을 고려하여 합성단위를 정의하여 합성음의 음질을 향상시키려는 시도들이 나타나고 있다. 일본 ATR에서 제안한 COC (context-oriented-clustering), 전자통신연구소에서 제안한 CDU (context dependent unit), 삼성종합기술원에서 제안한 modified diphone 등은 모두 이러한 흐름에서 제안된 단위들이다[35]. 이러한 방법들은 최근 하드웨어의 발달로 기억장치의 집적도 및 가격 면에서 유리해진 것을 이용하여 조음결합 법칙의 추출에서 발생하는 노력 및 비용을 절감하고 양질의 합성음을 얻을 수 있다는 장점이 있다.

앞에서 살펴본 바를 정리하면 최근 연구 동향은 연구소 및 기업체 중심의 실용적인 연구가 두드러짐을 알 수 있다. PSOLA 방식과 같은 시간축상의 코딩방법을 문서-음성변환 시스템에 적용하여 합성음질의 명료도 향상을 꾀하였고, 합성단위에 있어서 음소의 환경을 충분히 흡수하고자 많은 수의 합성단위를 음성합성에 이용하는 추세이다[35]. 이러한 합성음의 명료성 확보를 바탕으로 자연언어 처리기법을 적용하여 합성음의 자연성을 향상시키고자 하였으나, 아직 한국어 처리 기술 자체가 완벽한 단계에 이르지 못했고 이를 문서-음성 변환에 이용하려는 시도가 오래되지 않아 이에 대한 기술은 초보적인 수준을 벗어나지 못하고 있다.

4. 음성정보처리 기술을 이용한 정보검색

앞절에서 음성인식과 음성합성의 기본과정에 관련된 기술을 알아보았다. 본 절에서는 이러한 두 기술을 결합하여 정보검색에 적용하는 과정과 각 과정에 관련된 문제점을 알아본 후, 실제 응용예를 살펴보고자 한다.

4.1 음성을 이용한 정보검색 과정 및 문제점

인간의 기본적인 의사소통 수단인 음성을 이용하여 컴퓨터에 저장되어 있는 정보를 검색하기 위해서는, 인간이 발성한 명령어를 컴퓨터가 알아듣게 하는 음성인식 기술과 음성으로 정보를 출력하기 위한 음성합성 기술의 결합이 필요하다. 음성을 인터페이스로 사용하여 정보를 검색하는 기본 과정을 구체적으로 살펴보면 다음과 같다.

첫째, 시스템은 사용자가 시스템에 접속된 것을 감지한다.

전화선을 경유하여 시스템과 사용자가 연결될 경우에는 전화선 인터페이스가 이를 감지하여 시스템을 작동시키며, 사용자가 시스템 앞에서 직접 명령어를 발성하는 경우에는 미리 정해진 특별한 단어를 발성함으로서 시스템을 동작시킨

다. 특히, 전화선을 경유하여 정보검색을 수행하는 시스템은 여러 사용자가 동시에 사용하므로 시스템과 사용자의 효과적인 접속은 중요하다. 전화선을 경유한 정보검색 시스템은 전화기를 컴퓨터의 자판과 터미널의 역할로 사용하므로, 컴퓨터를 보유하지 않고도 컴퓨터를 활용할 수 있는 장점이 있고, 음성인식과 음성합성 기술을 효과적으로 이용할 수 있는 시스템의 구조이므로 음성을 이용한 정보검색시스템에 많이 이용된다. 또한 전화선을 통한 정보검색은 원격지 정보접근을 가능하게 하여 어느 곳에서나 정보검색이 가능하고, 오피레이터의 개입없이 직접 컴퓨터와 연결되므로 무인정보검색이 가능하여, 인력절감의 효과와 함께 24시간 서비스가 가능하여 언제나 서비스를 받을 수 있는 장점이 있다.

둘째, 시스템은 사용자에게 서비스 가능한 선택사항을 나열한다.

시스템은 선택사항을 나열하기 위해 음성합성 기술을 사용하여 음성으로 선택사항을 출력한다.

셋째, 사용자가 발성한 선택사항을 인식한다.

사용자의 음성명령을 인식하기 위해서 음성인식 기술을 사용한다. 음성인식 기술은 실험실 환경에서는 실용화가 가능할 정도의 인식론을 가지고 있지만, 잡음이 존재하는 실제 현장에는 상당한 인식률의 저하가 있다. 특히 전화선을 통한 음성인식은 전송선에 따른 채널 잡음과 300~3,300 Hz의 대역제한으로 고주파 대역의 음성정보를 상실하므로 인식률의 저하가 일어난다. 따라서 음성인식을 실제 현장에 적용하기 위해서는 환경변화에 강건한 음성인식기를 사용하여야 하고, 이를 위하여 여러가지 잡음처리 기법이 연구되고 있다. 또한 실제현장에서 음성인식기는 사람의 음성을 항상 정확하게 인식할 수 없으므로 오인식에 따르는 오류를 처리하여야 한다. 사람의 경우에도 대화 도중에 내용을 확실히 이해할 수 없는 경우가 있다. 이러한 경우에 사람은 되묻거나, 알아들은 내용을 다시 확인하는 행동을 한다. 컴퓨터의 경우에도 오인식된 결과를 보정하고 인식된 내용을 확인하기 위해서 사용자와의 상호작용을 하는 대화관리 및 이해 방법이 연구되고 있다.

넷째, 사용자의 선택사항에 따라 원하는 정보를 음성으로 출력한다.

음성으로 정보를 출력하기 위해서는 음성합성 기술을 사용한다. 고정되어 있고 단순한 내용의 음성은 합성하고자 하는 어휘들을 미리 분석하였다가 이들의 조합에 의해 말을 합성하는 제한 어휘 합성 방식을 사용하여 음성을 출력한다. 이 방법은 구조가 간단하고 미리 사람이 발음한 내용을 편집하는 것으로 자연스러운 음질을 갖는 장점이 있으나, 출력하고자 하는 문장에서 저장된 단어의 위치, 억양에 따라 합성가능한 어휘수에 제약이 따르게 되며, 미리 저장된 음성만을 출력할 수 있는 단점이 있다. 따라서 이 방식은 주로 단어 또는 문장 단위의 음편들을 연결한 몇가지의 합성음성만으로도 사용가능한 지하철 안내방송 또는 ARS 등에 이용되고 있다. 시시각각 변화하는 내용을 출력하기 위해서는 임의의 문장을 음성으로 변환하여 주는 무제한 음성합성 기술을 사용한다. 이 방법은 제한 어휘 합성 방식보다 음질이 저하되므로 음질의 향상을 위해 적절한 코딩 방법 및 합성단위의 선택과 운율제어에 대한 연구가 진행되고 있다.

음성을 이용한 자동정보검색시스템으로 이미 상용화되어서 쓰이고 있는 음성응답시스템(ARS: audio response system)을 들 수 있다. 이 시스템은 24시간 무인 서비스를 원격지의 소비자에게 전화선을 통해 제공할 수 있는 장점을 가지고 있으나, 명령어 입력을 위해 푸쉬버튼을 이용하므로 푸쉬버튼이 없는 전화기를 가진 사람은 이를 이용할 수 없고, 숫자음 이외의 명령어는 입력할 수 없다는 단점이 있다. 또한 기능 선택 방법 등에 관한 안내 음성이 계속되므로 사용자에게 거부감을 갖게 하므로, 사용자의 간단한 음성이라도 인식하여 안내과정을 짧게 하여 서비스의 질을 고품질하여야 할 필요성이 있다. 다음 절에서는 음성인식과 음성합성을 결합하여 보다 편리한 정보검색을 가능하게 하는 시스템의 예를 살펴보기로 하자.

4.2 음성을 사용한 정보검색 시스템

음성을 사용한 대표적인 정보검색 시스템으로 일본 NTT사 ANSER(automatic answer network

표 1 ANSER 시스템의 질의 응답의 예

Customer	System
(Call the center)	(Detect the call)
	"Hellow, this is the NTT bank telephone service center. What is your service number? "
"One, one "	
	"You are asking for your account balance What is your branch number?"
"One, two, . "	
	"What is your account number? "
"Three, four, ..."	
	"What is your secret number? "
"Five, six, . "	
	"Your current balance is ¥53,000 yen. If you would like to have your balance repeated, please say 'one more.' If not, say 'OK.'"
"OK "	
	"Thank you very much "

system for electrical request) 시스템을 들 수 있다. 이 시스템은 1981년부터 잔고 조회와 이체확인 등의 금융업계의 정보서비스를 제공하고 있다[36]. 초기에는 푸쉬버튼 전화기로 입력을 받고 음성합성을 이용하여 정보를 출력하였으나, 이후에 음성인식 기능이 부가되었다. 이 시스템은 음성뿐만 아니라 팩시밀리와 퍼스널 컴퓨터 등을 이용해서도 정보의 검색이 가능하다. 무인 자동정보검색에 의한 인력절감의 효과로 인해 70여 개 도시의 은행이 ANSER시스템을 채용하고 있으며 하루에 수십만 번의 검색이 이루어지고 있다. 이 시스템은 유사한 응용분야를 가진 주식 및 증권 시장에도 적용될 수 있으며, 은닉 마르코프 모델과 같은 통계적인 기법을 적용하여 대용량 연속음성인식 기능도 추가할 예정이다.

ANSER 시스템은 전화선상에서 음성인식과 음성합성을 사용하여 정보검색 서비스를 수행한다. 전형적인 질의응답 과정을 표 1에 나타내었다.

ANSER 시스템은 10개의 숫자음과 6개의 명령어를 인식할 수 있으며, 20대에서 60대 사이의 남녀 1,564명의 음성을 학습자료로 사용하였고, 인식률은 96.5%이다. 특징벡터로 LPC(linear predictive coding) 파라미터를 사용하고, 인식 알고리즘은 여러 개의 템플릿을 사용하는 동적 프로그래밍 기법을 사용한다. 음성합성을 위해서 선형예측 계수로부터 유도된 LSP(linespectrum pair) 파라미터를 합성 파라미터로 사용하고, 합성단위로 1000여개의 자음-모음-자음 음절과 400여 개의 자음-모음, 모음-자음, 모음-모음 음절을 사용한다. 음성은 음운학 및 언어적인 규칙에 의해 합성되는 규칙합성방법을 사용한다.

Bell-Northern 연구소에서 시험가동 중인 증권정보 검색시스템은 전화로 회사명을 말하면, 회사명을 인식하여 그 회사의 현재의 주가를 음성으로 출력한다[37]. 인식대상의 어휘는 뉴욕 증권거래소에 등록된 1,991개의 회사명과 12개의 명령어를 합한 2,003개, 토론토 증권거래소에 등록된 회사명 883개이다. 대부분의 음성인식 시스템은 고정된 인식대상어휘를 대상으로 하기 때문에, 학습자료의 수집에서 인식대상어휘를 포함시킬 수 있지만, 증권정보 검색시스템은 인식대상의 회사명이 빈번히 교체되는 특성이 있으므로 인식대상의 회사명을 모두 학습자료로 수집할 수 없다. 이를 처리하기 위해서 43개의 부단어(subword)를 모델링하여, 새로운 회사가 증권거래소에 등록되면, 회사명에 따라 학습된 부단어를 연결하여 회사명의 모델을 구성하는 기법을 사용하였다. 특징벡터로 멜켑스트럼과 LSP 파라미터를 사용하고, 인식 알고리즘은 연속분포 HMM을 사용하였다.

한편 스페인의 Telefonica I+D사는 News Service라는 메뉴 방식의 AUDIOTEX를 개발하였다[38]. 이 정보검색 시스템을 사용자가 발성하는 12개의 단어를 인식하여 이에 해당하는 항목의 뉴스를 제공한다. 이 시스템의 동작은 다음과 같다[그림 4].

사용자의 전화가 걸려오면 전화선 인터페이스가 이를 탐지하여 시스템을 동작시킨다. 시스템은 사용자에게 인사를 하고, 도움말을 들을 것인지를 물은 후, 선택사항을 나열한다. 사용자가 선택사항중 하나를 발성하면 이를 인식하여 원하는 항목의 정보를 제공한다.

특징벡터로 멜켑스트럼을 사용하고, 인식 알고리즘은 연속분포 HMM을 사용하였다. 인식대상이외의 단어를 모델링하여 주변잡음이나 부가적인 단어 속에 내포된 인식대상어휘를 인식하는 방법을 사용하였다.

미국 뉴욕지역의 전화회사인 NYNEX사는 목소리로 전화를 걸 수 있는 서비스를 뉴욕 통아

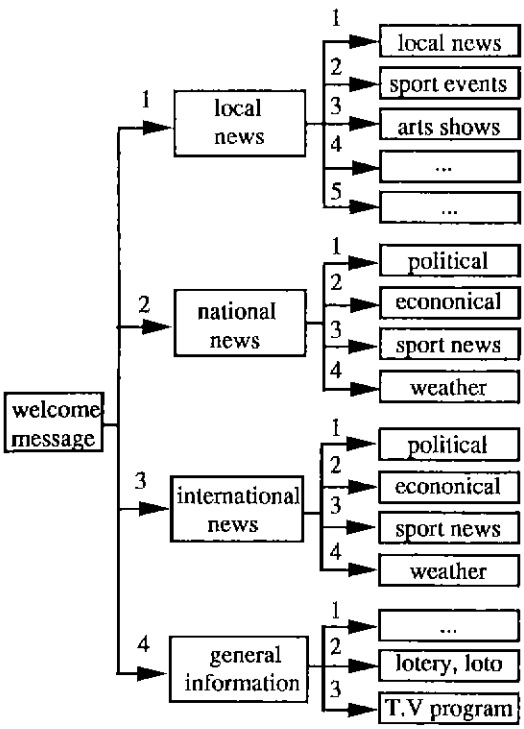

그림 4 News Services

일랜드지역에서 실시하고 있다. 전화 가입자는 서비스에 가입하면 전화를 걸때 다이얼을 누르거나 돌릴 필요없이 단지 수화기를 들어 통화하고자 하는 상대편 이름만 말하면 된다. 그러면 컴퓨터가 이름을 인식하여 등록된 전화번호를 찾아 이를 연결하여 준다. 일반적으로 버튼식 전화의 경우 엉뚱한 번호를 누를 확률이 1백번의 15번 정도라면 이 시스템의 오인식율은 1%로서 보다 정확도가 높고 매우 편리한 서비스를 제공한다.

5. 결 론

본 고에서는 음성정보처리를 이용한 정보검색 시스템의 기반기술인 음성인식과 음성합성에 대하여 알아보았고, 정보검색에 응용한 실제 시스템을 살펴보았다. 미국, 일본, 및 유럽연합 등은 제각기 대규모 국가 사업으로 연구를 지속하여 실험실 단계에서의 음성인식과 음성합성 기술은 이미 상당한 수준에 도달하였고, 제한된 분야에서 이미 상용화되고 있다. 음성인식의 연구는 보다 대용량의 어휘를 가진 음성인식 시스템과 일상 생활에서 나타나는 비문법적인 자연 발화 음성을 인식할 수 있는 연구를 진행중이며, 실험실 환경의 음성인식기는 환경 잡음이나 전송선에 의한 잡음 등에 의해 인식기의 성능이 저하되지 않도록 하는 잡음처리기법이 연구되고 있다. 음성합성의 경우 향상된 명료성을 바탕으로 자연어 처리기법을 적용하여 합성음의 자연성을 향상시키는 연구가 진행중이다.

음성을 이용한 정보검색은 자판과 모니터를 사용한 인터페이스에 비하여 편리하고 자연스러운 장점이 있으므로, 모든 사람에게 정보화 사회의 혜택을 누릴 수 있게 한다. 음성정보처리를 이용한 정보검색은 원격지 정보접근과 무인정보검색이 가능하므로 24시간 서비스가 가능하다. 또한, 정보의 입출력 중에도 손발의 활동에 제한을 가하지 않으므로 차안이나 이동 중에도 정보검색이 가능한 장점이 있으므로 날로 그 중요성이 증대되고 있다. 한국어 인식 및 합성 기술의 발전을 위해서는 대량의 음성자료가 필요하게 되나, 국내에는 구축된 표준적인 자료가 부족하여 연구에 어려움이 따르고 각 연구자들에 의해 만들어진 각 시스템을 검증할 수 없는 문제점이 있다. 또한 인력 및 연구비에서 외국에 비해 부족하여 연구에 어려움이 따르지만, 음성인식과 음성합성 기술은 각 나라 말의 고유한 특성에 좌우되는 경향이 강해, 자국 스스로 관련 기술을 확보하는 것이 바람직하다. 연구개발자들의 협조적인 연구와 관련 업체나 국가 기관으로부터 보다 많은 관심과 투자가 병행되어 지속적인 연구가 필요하다.

참고문헌

[1] 오영환, 패턴 인식론, 정익사, 1991.

[2] Furui, S., "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans. on Acoust., Speech, Signal Processing, Vol. 34, No.1, pp. 52-59, 1986.

[3] Juang, B. H., Rabinar, L. R. and Wilpon, J. G., "On the Use of Bandpass Liftering in Speech

Recognition," IEEE Trans. on Acoust., Speech, Signal Processing, Vol. 35, No.7, pp. 947-954, 1987.

[4] Soong, F. K. and Rosenberg, A. E., "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," IEEE Trans. on Acoust., Speech, Signal Processing, Vol. 36, No.6, pp. 871-879, 1988.

[5] Davis, S. B. and Mermelstein, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continously Spoken Sentences," IEEE Trans. on Acoust., Speech, Signal Processing, Vol. 28, No.4, pp. 357-366, 1980.

[6] Junqua, J. C., Wakita, H. and Hermansky, H., "Evaluation and Optimization of Perceptually-Based ASR Front-End", IEEE Trans. on Acoust., Speech, Signal Processing, Vol. 1, No. 1, pp. 39-48, 1993.

[7] Harvey F. Silverman and David P. Morgan, "The Application of Dynamic Programming to connected Speech Recognition," IEEE ASSP Magazine, pp 6-25, July 1990.

[8] Joseph Picone, "Continous Speech Recognition Using Hidden Markov Models," IEEE ASSP MAGAZINE, april, 1984.

[9] Rabiner, L. R., Wilpon, J. G. and Juang, B. H., "A Segmental K-Means Training Procedure for Connected Word Recognition," AT&T Tech. Journal, Vol. 65, pp. 21-31, 1986.

[10] Gauvain, J. L. and Lee, C. H., "Bayesian Learning for Hidden Markov MOdels With Gaussian Mixture State Observations Densities." Speech Communication, Vol. 11, Nos. 2-3, pp. 205-214, 1992.

[11] Bahl, L. R., Brown, P. F., de Souza. P. V. and Mercer, R. L., "Maximum Mutual Information Estimation of Hidden Markov Model Parameter for Speech Recognition" Proc. of ICASSP, Tokyo, pp. 49-52, 1986.

[12] Chou, W., Juang, B. H. and Lee, C. H., "Segmental GPD Training of HMM Based Speech Recognizer," Proc. of ICASSP, San Francisco, pp. 473-476, 1992.

[13] A. Waibel, T. Hanazawa, G.E. Hinton, K.Shikano and K.J.Lang, "Phoneme recognition using time-delay neural networks," IEEE Trans. on Acoust., Speech, Signal Processing, Vol. 37(3), pp. 328-339, March 1989.

[14] Ken-ichi Iso and Takao Watanabe, "Speaker-Independent Speech Recognition Using a Neural Prediction Model," 전자정보통신 학회논문지(일본), Vol. J73-D-II, No. 8, pp. 1316-1321, 1990년 8월.

[15] J.L Elman, "finding structure in time, "CRL Technical Report No. 8801, University of California, San Diego, 1988.

[16] Zue, V, and Lamel, L. "An expert spectrogram reader : A knowledge-based approach to speech recognition," Proc. of ICASSP, pp. 1197-1200, 1986.

[17] Mizoguchi, R., Tsujino, K., and Kakusho, O., "A Continuous Speech Recognition System Based on Knowledge Engineering Techniques," Proc. of ICASSP, pp. 1221-1224, 1986.

[18] Bahl, L. R. et al, "Multonic Markov Word Models for Large Vocabulary Continuous Speech RecognitionA," IEEE Trans. on Acoust., Speech and Audio Processing, Vol. 1, No. 3, pp. 334-344, 1993.

[19] Lee, C. H., Juang, B. H and Soong, F. K., "A Segment Model Based Approach to Speech Recognition," Proc. of ICASSP, New York, pp. 501-504, 1988.

[20] Riley, M. D., "A Statistical Model for Generating Pronunciation Networks," Proc. of ICASSP, Vol. 2, Toronto, pp. 737-740, 1991.

[21] Weintraub, M., et al., "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," Proc. of ICASSP, Glasgow, pp. 699-702, 1989.

[22] Jelinek, F., "The Development of an Experimental Discrete Dictation Recognizer," Proc. of IEEE 73, pp. 1616-1624, 1985.

[23] Ney, H., "Dynamic Programming Parsing for Context-Free Grammar in Continuous Speech Recognition," IEEE Trans. on Signal Processing, Vol. 39, No. 2, pp. 336-340, 1991.

[24] Ney, H., et al., "Improvement in Beam Search for 10,000-Word Continuous Speech Recognition," Proc. of ICASSP, San Francisco, pp. 9-12, 1992.

[25] Bahl, L. R. et al, "A Fast Approximate Acoustic

Match for Large Vocabulary Continuous Speech Recognition," IEEE Trans. on Speech and Audio Processing, Vol. 1, No, 1, pp. 59-67. 1993.

[26] Soong, F. K. and Huang, E. F., "A Tree-Trellis Based Fast Search for Finding the N-Best Sentence Hypothese in Continous Speech Recognition," Proc. of ICASSP, Toronto, pp. 703-706, 1991.

[27] H. Iida, "Prospects for Advanced Spoken Dialogue Processing", IEICE Transactions on Information and Systems, Vol.E76-D, No.1, pp. 2-8, 1993.

[28] Yoichi Takebayashi, Hiroyuki Tsuboi, Hiroshi Kanazawa. "A Real-Time Speech Dialogue System Using Spontaneous Speech Understanding", IEICE Transactions on Information and Systems, Vol.E76-D, No.1, pp. 112-120, 1993.

[29] Y. Yamashita, H. Yoshida, T. Hiramatsu, Y. Nomura, R. Mizoguchi, *"MASCOTS II: A Dialog Manager in General Interface for Speech Input and Output~, IEICE Transactions on Information and Systems, Vol.E76-D, No 1, pp. 74-83, 1993.*

[30] 진용옥, "음성 정보처리 기술 및 음성 정보서비스의 발전과 전망", 음성통신 및 신호처리 워크샵 논문집, pp. 12-26, 1992.

[31] 김상룡, 김정수; "형태소 해석을 이용한 합성 음성의 음운 및 운율 처리", 전자 공학회지 20권 5호, pp. 508-515, 1993

[32] 강승식, "한국어의 형태론적 특성과 형태소 분석 기법", 정보과학회지 12권 8호, pp. 47-59, 1994.

[33] 이정석, 오영환; "한국어 text-to-speech 합성기의 구현", 한국정보과학회 춘계학술발표 논문집 19권 호, pp. 107-110, 1992.

[34] 김상훈, 지민제, 최운천; "한국어 문장/음성 변환에서의 TD-PSOLA 적용", 음성통신 및 신호처리 워크샵 논문집, pp. 291-294, 1993

[35] 공병구, 김상룡, 김정수; "이질음 접속에 의한 음질 저하 및 극복 대책 연구", 음성통신 및 신호처리 워크샵 논문집, pp. 279-284, 1993.

[36] R. Nakatsu, "Anser: an application of speech technology to the Japanese banking industry," IEEE computer, Vol 23. No. 8, pp. 43-48, Aug. 1990.

[37] M. Lennig et al., "Flexible vocabulary recognition of speech." Proc. of ICSLP, pp. 93-96, Oct. 1992.

[38] M. J. Poza et al., "An approach to automatic recognition of keywords in unconstrained speech using parametric models," in Proc. 2nd European Conf. on Speech Comm. and Tech., pp. 471-474, Sep, 1991.

[39] 김연준, 오영환; "한국어 문서-음성 변환 시스템에서의 구문분석에 의한 운율조절에 관한 연구", 음성통신 및 신호처리 워크샵 논문집, pp. 285-290, 1993.

[40] 최환진, 윤성진, 오영환; "음향분절모델과 거리에 기반한 관측확률 평활화를 이용한 한국어 단어 인식에 관한 연구", 한국정보과학회 추계학술발표 논문집 , pp. 629-632, 1994.

[41] 지상문, 윤성진, 오영환; "비모수적 커어널 밀도 추정을 이용한 은닉 마르코프 출력 확률 모델링", 한국정보과학회 추계학술발표 논문집 , pp. 471-474, 1994.

[42] 유하진, 김승주, 오영환; "연속음성 인식을 위한 음소간 퍼지 유사도 관계의 추정", 한국정보과학회 충청지부 추계학술발표 논문집 , pp. 161-171, 1994

오 영 환

1972 서울대학교 공과대학(공학사)
1974 서울대학교 교육대학교(석사)
1980 Tokyo Institute of Technology 정보공학(박사)
1981 충북대학교 공과대학 조교수
1981 서울대학교 공과대학 강사
1983 Univ. of California, Davis연구 교수
1985 한국과학기술원 전산학과 교수
관심 분야 : 음성인식, 음성합성, 패턴인식 등