□ 기술해설 □

음성합성 기술의 현재와 전망

LG종합기술원 이준우·김세린·이종석

1. 서 론

현대 정보화 사회에서 인간은 많은 정보들을 다양한 매체를 통하여 받아들인다. 이와 같은 정보 수집을 보다 편리하고 신속하게 하기 위하여 다양한 매체의 통합인 멀티미디어 환경하에서의 인간과 기계의 정보교환을 편리하게 하기 위한 Man-Machine Interface기술이 중요하게 대두되고 있다. 음성은 인간의 자연스러운 의사소통의 도구이므로 음성을 통한 기계와의 정보교환 기술은 매우 중요하다.

인간의 말은 성대의 울림에 의한 공기의 진동이 성도를 통해 입 밖으로 나옴으로써 생성된다. 말은 정보를 가지고 있으며 정보는 말의 변별적 특성에 의해 표현된다. 이 변별적 특성은 입과 혀의 모양 및 위치에 따라 성도의 모양이 달라짐으로써 성대의 울림에 의해 발생한 공기 진동이 특정 주파수에서 공진을 일으킴으로써 발생하거나 또는 공기의 흐름을 막거나 틔움으로써 발생한다. 음성합성이란 이와 같은 인간의 발성 기관을 인공적으로 만듦으로써 인위적 음성을 만들어 내는 것이다.

음성합성에 대한 연구는 다년간 제반 음성 관련 기술에 비해 선행되어 왔다. 초창기의 노력은 주로 전기-음향학적 이론에 근거를 둔 기본적 조음 모델을 이용하여 인간의 음성생성 메카니즘을 시뮬레이션 하는 것에 모아졌다. 이러한 모델링이 여전히 음성합성 연구에 있어서 중요한 목표중의 하나임에는 틀림이 없으나 컴퓨터 분야의 발전으로 음성합성의 연구분야는 단순히 음성의 생성뿐만 아니라 문장의 처리를 포함하는 즉, 문장-음성 변환(Text-to-Speech conversion)이라는 영역으로 넓어지게 되었다[1]. 이러한 기술은 음성학적인 이론과 음향학적인 분석에 의해 얻어지는 일단의 규칙들로 이루어지므로 흔히 규칙 합성이라 불리어진다.

본 고의 구성은 TTS 시스템의 구조와 각 부분을 형성하고 있는 기술들을 설명하고 음성합성 기술의 동향 및 응용 현황을 소개하는 것으로 이루어져 있다.

2. 문장-음성 변환장치(TTS system)

그림 1은 전형적인 TTS 시스템의 구성을 보여주고 있다. 이는 언어 처리부, 운율 처리부, 그리고 합성부로 이루어져 있는데 언어 처리부에서는 입력된 문장들이 전처리기와 구문사전을 이용하여 음성학적인 표현으로 변환된다. 이러한 언어 처리부의 결과를 이용하여 운율 처리부에서는 음성의 기본 주파수, 지속시간,

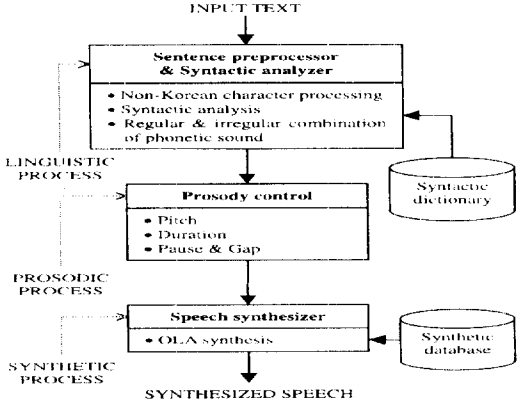

그림 1 문장-음성 변환 장치의 구조

휴지구간 등과 같은 운율정보를 결정한다. 합성부에서는 전단에서 추출한 운율정보를 이용하여 데이타베이스의 합성 단위들을 결합시킴으로써 연속적인 합성음이 얻어진다.

3. 관련 기술들

3.1 합성단위 선정

음성합성은 기본 합성 단위들을 운율정보에 맞게 결합함으로써 구현된다. 그러므로 합성단위는 결합시 발생할 수 있는 문제점들을 최소화할 수 있어야 한다. 합성단위가 가져야 하는 조건들은 다음과 같다[2].

1. 운율정보의 제어가 용이해야 한다.
2. 결합부에서 스펙트럼 불연속성을 최소화 시킬수 있어야 한다.
3. 자연스러운 에너지 궤적이 유지되어야 한다.
4. 적당량의 데이타베이스 구축이 가능해야 한다.

합성단위로는 구절(phrase), 단어(word), 음절(syllable), 음소(phoneme) 등의 단위와 이들을 음성합성에 알맞은 형태로 변형시킨 반음절(demi-syllable), diphone, triphone, CV (Consonant-Vowel), VC, VCV, CVC 등 여러 가지가 사용된다[3, 4]. 합성단위는 클수록 합성음의 자연성은 증가하나 조합하여 만들수 있는 대상은 제한되며 합성단위가 작은 경우는 그 반대이다. 주로 제한 어휘 합성시에는 구절 및 단어를 합성단위로 많이 이용하며 무제한 어휘 합성시에는 음절 이하의 단위가 많이 이용된다.

합성단위가 작으면 합성에 필요한 전체 데이타량이 작아져 각각의 단위를 제어하기는 쉬우나, 합성단위의 연결시 발생하는 상호 조음현상을 정확히 구현하기 어려우며 결합부의 불연속성 및 이에 따른 음질 저하 현상이 생긴다. 따라서 이와 같은 문제점을 최소화하기 위해서 모음의 안정된 구간에서 음절을 절단한 반음절 및, CV, VC, VCV 또는 수정된 형태의 음절이 합성단위로 쓰이고 있다[5]. 또한 합성단위를 한 가지로 제한하지 않고 음소 환경에 따라 합성단위가 자동적으로 생성되는 COC(Context Oriented Clustering)방법이 제안되었다[6].

3.2 텍스트처리 기술

TTS 시스템에 입력 가능한 텍스트는 간단한 초등학교 교과서부터 학술 논문까지, 문어체에서 구어체까지 다양하다. 텍스트에는 한글뿐만 아니라 숫자, 약어, 기호, 전문 용어 등이 들어 있다. 텍스트 정규화 모듈은 입력 텍스트를 분석한 뒤, 한글 이외의 문자들을 한글로 바꿔주는 모든 일을 한다. 정규화된 한글 텍스트는 형태소 분석기에서 형태소별 분리, 품사등의 정보를 얻게 된다. 이러한 정보들을 이용하여 구문분석기는 자연스러운 운율정보를 생성하기 위하여 문장의 구문정보를 추출해낸다. 운율생성을 위한 정보는 각 구의 위치, 문장성분, 상호결합 관계 등으로 구성된다.

표기되는 문자와 발음되는 음소는 서로 정확하게 일치하지는 않는다. 따라서 음성합성을 위하여 문자열을 음소열로 변환하는 과정이 필요하게 된다. 이 과정은 음운 변동 규칙에 의하여 대부분 자동적으로 이루어지는데, 대표적인 규칙으로는 대표음화, 연음, 경음화, 격음화, 자음접변, ㄴ첨가, 구개음화 등이 있다. 그러나 경음화, ㄴ첨가 등은 규칙으로 처리하지 못하는 부분이 존재하므로 이 경우는 불규칙 변환 사전을 이용하여 처리하여야 한다.

3.3 운율정보 생성 및 제어기술

운율이란 발성시 나타나는 억양, 강세, 리듬 등의 특성을 말하는데 이는 기본 주파수, 음소의 지속시간, 음량, 휴지구간 길이 등에 의해 결정된다. 사람이 한번 숨을 쉬어 발성하는 말의 단위를 발화 단위라 하는데 발화 단위 내에서는 기본 주파수가 점차 낮아지는 경향을 갖는다. 이를 억양의 기본 기울기라 한다. 억양의 기본 기울기에 단어, 음절의 강세 및 문장 구조에 따른 억양 패턴이 더해져서 전체 억양 패턴이 구성된다[7].

음소의 지속시간 및 휴지구간 길이는 억양과 함께 합성음의 자연도를 결정하는 중요한 요소이다. 음소의 지속시간은 음소 자체의 성질뿐만 아니라 주변의 음소 환경, 한 단어내의 음

소 갯수, 단어 내에서의 음소의 위치, 강세여부 등 다양한 요소에 의해 영향을 받는다.

휴지구간 길이도 음소의 지속시간과 마찬가지로 전후의 음소환경에 의해 영향을 받는데, 그 이외에 발화단위 사이에서 긴 휴지구간이 존재한다. 그런데 발화단위는 하나의 발화단위 내의 어절 수, 음절 수뿐만 아니라 문장의 구조 및 의미에 의해 결정되므로 문장 구조의 분석을 위해서는 형태소 분석, 구문 분석, 의미 분석 등이 필요하며 운율 처리부에서는 이러한 정보를 이용하여 운율 정보를 추출한다.

3.4 음성합성 기술

음성을 출력하기 위한 마지막 단계는 앞단에서 얻어진 운율정보와 합성단위 데이타를 이용하여 파형을 합성하는 것이다. 합성음 생성부즉, 합성기의 종류는 조음 합성기, 포만트 합성기, 그리고 연결형 합성기로 크게 나뉘어진다.

조음 파라미터에 의한 음향적 표현은 인간이 음성을 발성한다는 입장에서 성도의 음향 특성을 전기적으로 흉내낸 것이다[8]. 조음 파라미터로는 성도의 단면적 및 턱, 입술, 혀 등 조음기관의 위치를 이용한다. 이 방식은 인간의 발성기관의 움직임에 대응하기 때문에 파열음과 같은 Non-stationary한 음의 생성도 쉽게 이루어진다. 그러나 정밀한 합성필터를 제어하기 위한 성도의 형상 데이타를 획득하는 것이 대단히 어려워 직접 조음기관의 움직임을 관측하는 수법보다 선형예측기법을 이용하여 간접적으로 조음 파라미터를 추정하는 방법이 시도되어 왔다. 최근에, 각종 계측 기술이 발달하여 이에 관한 직접적인 실험 데이타들이 쌓여 가고 있으므로 앞으로 그 성과를 기대할 수 있게 한다.

조음 파라미터에 의한 음향 표현은 인간이 음성을 발성한다는 입장에서 모델화한 것인데 비해 청취의 관점에서 음성의 주파수 스펙트럼에만 주목하여 전기적으로 흉내낸 것이 포만트 혹은 선형예측 파라미터를 이용한 합성방식이다. 이는 인간의 청취가 음성의 스펙트럼에 근거하고 있다는 점을 이용한 것으로 음성 스펙트럼을 포만트라고 부르는 성도의 3~4개의 공진주파수와 그 대역폭으로 표현하고 포만트에 의한 공진회로를 복수개 접합하여 성도와 등가인 전달특성을 구현한다[9].

공진특성을 직접 표현하므로 음성 파형과의 대응도 용이하며 합성기 제어의 규칙화도 쉽지만 완벽한 포만트 정보의 자동추출이 어렵다. 한국어 음성합성에 포만트 합성 방식을 적용한 연구가 몇몇 있었으나 자연성에 문제가 있었다.

연결 합성 방식에서는 음성 파형을 일정 단위로 분해하고 생성된 기본 주파수에 따라 분해된 음편을 재배열함으로써 피치 조절이 이루어지며, 지속 시간의 조절은 단순히 음편을 생략 또는 복제함으로써 구현된다[10, 11]. 그림 2는 합성단이 요구하는 피치에 따라 음편을 재배열하고 결합하는 과정을 나타낸다. 이때 처리과정이 전적으로 시간 영역에서만 이루어지므로 실시간 처리가 용이하다.

이 방식은 다시 자연음의 음편과 피치정보를

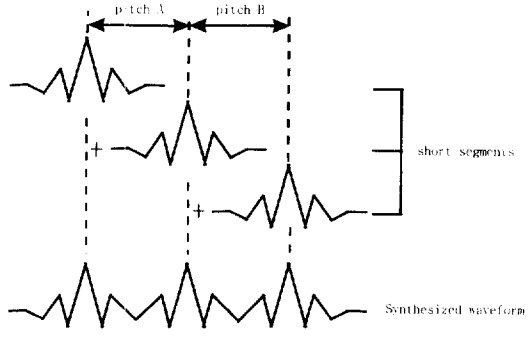

그림 2 단구간 음편의 OLA 합성

이용하는 TD-PSOLA(Time Domain-Pitch Synchronous OverLap and Add)계열과 성도의 임펄스 응답을 이용하는 PSE(Power Spectrum Envelope) 합성방식으로 분류된다. 음성신호를 일정 길이의 음편으로 나누어 이를 주파수 영역으로 변환하면 주파수 포락선 정보와 위상정보가 얻어진다.

이 중 위상정보를 영으로 두고 음성의 피치정보를 제거한 후 시간영역으로 역변환하면 성도 임펄스 응답의 성격을 지니는 좌우 대칭형의 음편이 얻어지는데[12], 이를 중첩하여 연결함으로써 음성을 합성하는 것이 PSE 합성방식이다.

4. 연구 동향 및 응용

4.1 음성 코퍼스 기반의 접근

음성의 특징을 제어할 수 있는 규칙 생성을 위한 부단한 노력이 규칙 기반의 합성음질을 개선시켜 왔다. 그러나 이러한 노력의 과정에서 얻어지는 세세한 규칙과 정선된 제어 파라미터들은 대부분이 특정 시스템 의존적이어서 유사한 시스템 개발자들이 공유하기가 매우 어려운 것 또한 사실이다.

이러한 비합리성을 개선하기 위해 전통적인 규칙기반 접근 방식과는 대조되는 음성 코퍼스 기반의 접근 방식이 대두되고 있다. 코퍼스 기반의 음성 합성방식에서는 음향-음성학적 레이블링과 구문론적 분류정보와 같은 통계적 모델링을 위한 정보들이 주석 처리된 음성 데이타 집합이 구축된다.

이 정보들을 이용하여 음성의 스펙트럼과 운율의 특징 파라미터들이 분석되고 그 결과에 근거하여 제반 규칙 모델들이 생성되고 훈련된다. 모델의 타당성 또는 단점은 정량적으로 나타나고 이를 피드백시켜 보다 개선된 규칙 모델을 생성시킨다.

코퍼스 기반 방식의 이러한 명확한 훈련 과정과 객관성이 보장된 평가 결과는 동일하게 주석 처리된 음성 코퍼스를 보유한 개발자들이라면 누구나 쉽게 공유할 수 있게 되며 현재 음성의 주파수 특성 및 운율제어뿐만 아니라 합성 단위의 결정 방법에도 응용되고 있다[13].

4.2 다국어 지원을 위한 모듈러 아키텍처

그림 1과 같이 TTS 시스템의 구조는 언어처리부, 운율 처리부, 그리고 합성부로 이루어져 있는데 이들은 기능상 독립적으로 동작되는 다수개의 모듈로 분리될 수 있다. 1980년대 이후로 TTS 시스템의 다국어 지원이 강조되고 있는 시점에서 모듈러 아키텍처는 시스템을 언어 독립적 모듈, 언어 의존적 모듈, 그리고 지식원(konwledge source)으로 분리 시킴으로써 특정 언어에 기반한 시스템에 타 언어를 보다 쉽게 이식시킬 수 있게 해준다[14].

이러한 노력들로 인해 언어 규격 테이블만을 제공하여 다국어 지원을 가능케 한다는 다국어 지원의 최종 목표는 현실화되어 가고 있다.

4.3 실용적 측면에서의 합성방식 분류

합성음의 생성방식은 기술적인 측면으로는 3.4에서 언급했듯이 조음 파라미터형, 포만트 등 음성 특징 파라미터형, 연결형으로 나뉘어지지만, 실제 응용면에서는 연결형과 규칙 기반형으로 구분 지을 수 있다.

연결형 합성방식은 실제 발성된 음성에서 추출한 음편 또는 성도 임펄스 응답을 이용하여 발성자의 음색을 상당 부분 재생할 수 있어 합성음의 자연성을 높일 수 있다. 그러나 발성자에 지나치게 의존적이어서 다양한 음색이나 발성 스타일을 만들기 위해서는 추가적인 데이타베이스 구축이 요구된다는 단점이 있다. 반면, 규칙 기반형에서는 음향학적 파라미터값을 적절히 바꿈으로써 원하는 음성을 만들 수 있으나 음성 패턴에 관한 광범위한 지식과 이해가 요구된다는 단점이 있다.

이러한 특성들로 인하여 연결형 합성방식은 합성음의 자연성이 부각되는 응용 분야에 적합하며, 규칙 기반형은 메모리 요구량이 문제시되거나 음색변조가 필요한 분야에 주로 적용되고 있다.

4.4 응용 분야

음성 합성기는 PC의 보급과 통신 시장의 확대에 따라 응용분야가 점차 커지고 있다. 음성출력을 갖는 학습기기, 경보음 대신 오류를 음성으로 알리는 상황 진단 시스템, 계산기의 출력 단말, 시각 또는 청각 장애자용 보조기기, 전화번호 안내나 금융정보 안내 서비스와 같은 자동 응답장치 등에 응용될 수 있으며, 향후 음성인식 장치와 연동되어 타 언어의 번역 결과를 자국어로 합성하는 장치로 사용될 수 있다.

특히 최근에는 인터넷, BBS(Bulletin Board System) 등에 합성기를 적용하여 사용자 인터페이스의 편리를 도모하려는 시도가 이루어지고 있으며, 이동통신 단말기 또는 서버에 장착되어 음성메일 등에도 응용이 될 것이다.

5. 결 론

지금까지 본 고에서는 문장-음성 변환 시스템을 중심으로 음성합성 기술 전반에 걸쳐 개략적으로 알아보았다. 음성합성에 대한 연구는 보다 자연스럽고 명료성이 뛰어난 음성을 생성한다는 목표를 위해 지속적으로 이루어질 것이고 향상된 음질은 새로운 응용분야를 창출할 것이 분명하다.

이를 촉진시키기 위해서는 대용량의 데이타베이스의 구축과 더불어 객관성 있는 평가방법이 요구된다. 또한 문장구조의 구문론적인 분석의 정확도를 위해 자연어 처리 분야와 연계하여 문장의 의미론적인 분석이 가능해질 때 비로소 인간과 흡사한 합성음을 기대할 수 있을 것으로 생각된다.

참고문헌

[1] Jonathan Allen, M. Sharon Hunnicutt, and Dennis Klatt. From text to speech-the MITalk system, MIT Press, Cambridge, Massachusetts, 1987.

[2] N. B. Pinto and D. G. Childers, Formant Speech Synthesis : improving production quality, IEEE Trans. ASSP, ASSP-37(12), 1989.

[3] Yunkeun Lee, Seungkwon Ahn, Trend in Speech Synthesis, KITE Review Vol. 20, No. 5, pp. 523~532, 1993.

[4] Byunggoo Kong, Sangryong Kim, and Jeongsu Kim, Speech Enhancement Strategies in Concatenating between Allophones, Preceding of SCAS vol. 10 No. 1, pp. 279~284, 1993.

[5] Joonwoo Lee, Serin Kim, Kew S. Park, Jongseok Lee, Heeyoun Lee, A Korean Text-to-Speech Conversion System, ICSP '97

[6] S. Nakajima and H. Hamada, Automatic generation of synthesis units based on context oriented clustering, ICASSP-88, pp. 659~662, 1988.

[7] H. Fujisaki and H. Kawai, Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese, ICASSP-88, pp. 663~666, 1988.

[8] Parthasarathy and C. H. Coker, Automatic estimation of articulatory parameters,, Computer Speech and Language, 6 : 37~75, 1992.

[9] Dennis H. Klatt, Software for a cascade/parallel formant synthesizer, Journal of the Acoustical Society of America, 67 : 971~995, 1980.

[10] Charpentier, M. Stella, Diphone Synthesis using an Overlap-add Technique for Speech Wave Concatenation, proc. ASSP, pp. 2,015~2,018, Tokyo, 1986.

[11] Moulines, F. Charpentier, Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones, Speech Communication 9, pp. 453~467, 1990.

[12] Imai, Y. Abe, Spectral Envelope Extraction by Improved Cepstral Method, IECE Trans. J62-A, pp. 217~223, 1979.

[13] Yoshinori Sagisaka, Nobuyoshi Kaiki, Naoto Iwahashi, and Katsuhiko Mimura, ATR Talk speech synthesis system, ICSLP, pp. 483~486.

[14] Richard W. Sproat, Joseph P. Olive, A Modular Architecture for Multilingual Text-to-Speech, Progress in Speech Synthesis, Springer-Verlag New York, Inc., pp. 565~573.

이 준 우

1992 경북대학교 전자공학과 졸업(학사)
1995 경북대학교 전자공학과 대학원 졸업(석사)
1995~현재 LG종합기술원 주임연구원
관심분야 : 음성분석, 음성합성

김 세 린

1992 연세대학교 전산과학과 졸업(학사)
1994 연세대학교 전산과학과 대학원 졸업(석사)
1994~현재 LG종합기술원 주임연구원
관심분야 : 음성합성, 자연어처리

이 종 석

1983 서울대학교 제어계측공학과 졸업(학사)
1985 서울대학교 제어계측공학과 대학원 졸업(석사)
1985~현재 LG종합기술원 책임연구원
1995 서울대학교 제어계측공학과 대학원 졸업(박사)
관심분야 : 디지털 신호처리, 음성인식, 음성합성, 음성코딩

● 제25회 임시총회 및 춘계학술발표회 ●

- 일 자 : 1997년 4월 24일(금)~25일(토)
- 장 소 : 충남대학교
- 발표논문 접수마감 : 1998년 3월 3일(화)
- 문의 및 접수처 : 한국정보과학회 사무국
  Tel. 02-588-9246, Fax. 02-521-1352
  서울시 서초구 방배3동 984-1(머리재빌딩) ㉮137-063

□ 기술해설 □

전화망에서의 음성인식기술의 활용현황

한국통신 김재인

1. 서 론

전화는 가장 편리하고 값싼 단말기라고 할 수 있으며, 오늘날에는 전화를 사용하여 많은 유용한 정보를 얻을 수 있다. 그러나 처음에는 음성정보서비스와 같이 기계에서 사용자가 원하는 정보만을 얻기 위해서는 전화기에 달려있는 전화버튼을 이용하여 그 내용을 선택하여야만 해서 불편하였으며 이 역시 전자식 교환기에 연결되거나, 휴대용단말기(휴대폰, PCS)를 가진 경우만 가능하였다. 이러한 불편은 음성인식기능을 가진 전화정보시스템이 개발되어 어느 정도 해결되어가고 있으며, 좀 더 다양한 서비스들이 개발되어 사용자들을 편리하게 하고 있다. 본 논문에서는 전화망을 통해서 음성인식기능을 이용하여 사용할 수 있는 서비스에 관하여 알아보기로 한다.

2. 외국의 서비스 현황

음성인식에 대한 연구는 1952년 미국의 벨 연구소에서 숫자음인식에 대한 논문이 발표된 이후 이제까지 여러 나라에서 활발히 진행되어 오고 있다. 이 기술이 전화망에 사용된 것은 1986년 영국에서 cell-phone에 적용한 이래 1990년에 들어서면서부터 서비스 개발이 활발해져서 현재까지 세계각국에서 시험서비스 또는 상용서비스를 개발하거나 또는 제공하고 있다. 여기서는 유럽과 미주 그리고 일본에 서비스 현황에 대하여 차례로 알아보고 마지막으로 우리 나라에 대하여 설명하겠다.

2.1 유럽

유럽은 유럽통합에 따라 동일한 서비스를 동시에 여러 언어로 제공해야 하기 때문에 하나의 공통된 서비스에 대하여 몇개 나라가 공동으로 개발하면서 각자의 언어에 대한 문제를 해결해주는 연구들이 활발히 진행되고 있다.

2.1.1 프랑스

1992년 이후 몇가지 시험서비스가 개발되었지만, 대부분 CNET의 automatic speech recognition(ASR) PHIL90 의 기술을 사용한 것으로, 이 기술은 화자독립, 단어모델을 사용하고, 안내방송 중에 음성입력이 가능한 barge-in, 인식대상단어가 아닌 경우에도 틀린 인식결과를 내는 것을 방지하는 rejection, 그리고 입력음성 중에서 인식대상단어만 찾아내는 핵심어 추출기능을 지원하며, 화자종속과 allophone model도 지원한다. 현재 여러 분야에 상용서비스가 제공되고 있다[1].

가. FT Public Voice-Mail Service

Yes, no, listen, record, delete, add와 announce 등 7단어를 인식하며 incoming call만을 처리한다. Transvox에 ALCATEL/TITN platform에서 CNET 구현되었으며, 3개 도시(240회선)에 지역주민만 서비스 가능하고, 1997년안에 전국적인 서비스를 할 예정이다.

나. Information Service for People on the Move

유럽의 5개국 교환원(FT, BT, DT, TI, Telefonica)이 협동으로 지원하는 다국어 서비스이며 5개국어의 4단어의 음성명령이 가능하다.

[가]의 경우와 같은 인식기와 H/W를 사용하고 있으나, 45port로 서비스하고 있다. 1995년 7월에 서비스가 개시된 이래 1995년 여름에서 가을까지 매월 25,000 call을 처리했다.

다. France Telecom(FT) Sales Agency Service

FT의 제품 및 서비스들을 근무시간 이외에 제공하는 것으로, 1991년부터 4port정도의 소규모로 시험서비스 중이다. 인식단어는 23개이며 하루에 120통화를 처리한다.

다음은 FT가 아닌 다른 사업자가 제공하고 있는 서비스이다.

라. "Les Baladins" Service

1986년부터 Lannion근교에서 cinema 프로그램들에 대한 정보를 알려주고 있다. 1port로 일주일에 400통화를 처리한다.

마. "MACIF" Service

MACIF는 프랑스에서 가장 큰 자동차보험회사로 고객에게 24시간 내내 상담이나 보험처리결과를 알려주고 있다. 인식단어는 30개("insurance," "vehicle," "car registration number," "robbery," etc)로 PC로 된 server는 동시에 100통화를 처리할 수 있으며, 하루에 수백 통화를 처리하고 있다.

다음은 FT가 현재 field trial 또는 lab testing단계에 있는 서비스들이다.

바. Voice Dialing Service

사용자 한사람당 10~30명의 이름을 등록하여 사용할 수 있다.

사. Automatic Directory Service

CNET사의 Lannion 고용자들의 성명을 인식하며, 인식이 어려운 경우 spelling을 사용한다. 1995년 7월부터 시험서비스 중이다.

아. CINEZOOM

Video on Demand에 대한 예약서비스를 제공한다.

자. Rail Travel Information Service

European RailTel project에 일부로 개발되어 시험서비스를 제공중에 있다.

2.1.2 영국

영국의 BT는 10년이상 음성처리기술을 전화망에 적용해 왔으며, 그 결과 다양한 서비스를 제공하고 있다. 1986년 세계 최초로 handfree speech dialing cellphone인 "Topaz"를 개발하였으며, 1988년에 "Topaz II"와 1991년 "Azure"로 개량되었다. BT에서는 두 종류의 main platform을 사용하여 모든 서비스를 개발하고 있다. 그중에 대부분은 BT에서 설계하고 현재 Ericsson에서 만들어내고 있는 interactive Speech Application Platform(ISAP)[2]이 있으며, MAP(Minor Application Platform)은 PC based platform으로 main network이외에서 사용될 서비스를 위해 이용될 수 있다.

가. CallMinder

영국 내에서 음성인식기술을 이용한 서비스 중 가장 큰 것으로 사용자가 전화선을 사용하고 있을 때, 대신 응답을 해준다. 음성인식을 이용하여 사용자가 메시지를 검색하거나, 시스템을 사용자가 원하는 대로 변환할 수 있다. 예를 들어 CallMinder가 받기까지의 ring의 횟수를 조정할 수 있다. 1995년 6월부터 서비스되기 시작하여 전국적으로 19개 site에서 50만 이상의 사용자를 확보하고 있다.

나. Automatic Directory Enquiries

1994년 East Suffolk에서 25,000명의 이름에 대해 대어휘인식과 spelt 입력이 가능한 directory enquiry service를 처음 제공한 이래 현재에도 몇개의 시스템이 시험서비스 중에 있다. BT의 한 연구소에서도 4,000명의 근무자의 전화번호를 사람이름으로 찾을 수 있는 서비스를 제공하고 있다.

다. Payment Line

이 서비스는 home shopping이나 고객관리를 위해서 개발되었으며, 개인번호, 이름 주소를 인식하기 위해 대어휘 연속음성인식기술과 확인을 위해 음성합성기술을 사용하였다. 이는 몇 개의 우편주문회사에서 시험 중에 있다.

라. Medical Line

의사들이 원격지에 있는 데이타베이스 환자들의 진료기록을 찾아볼 수 있도록 하는 것으로 음성합성기를 같이 사용하고 있다.

2.1.3 이탈리아

지난 몇 년 사이에 음성인식과 합성기술을

이용한 서비스 개발이 활발히 진행되어 왔다. 그 결과 여러 서비스들 중 RailTel은 Eurospeech '97에서 실시된 음성인식시스템 평가회에서 10여개의 시스템 중에서 1등을 차지하기도 하였다.

가. 1412Reverse Directory Service

전화번호를 입력하면 가입자의 이름과 주소를 알려주는 것으로 1994년 처음에는 pulse dialing이나 DTMF로 번호를 입력하였으나 1996년부터 음성입력과 (yes/no)의 입력이 가능한 시험서비스가 실시되고 있다.

나. Automatic Alternate Billing Services

시외전화나 국제전화의 방법을 선택하는 것으로 연속숫자음과 (yes/no)를 이식할 수 있으며 1996년부터 부분적으로 또는 전부 자동화하였다.

다. Customer Care Service

이 서비스는 12만 Telecom Italia직원에게 생산품의 기술적인 정보를 전달해 준다. 입력은 DTMF나 음성입력이 가능하며, 8자리의 개인번호와 문제의 종류를 5개의 범위중에서 선택하여 입력한다. 1995년 시작하여 내부적으로 운영중에 있다.

1996년 말부터는 연속 숫자음 인식과 보다 자세한 정보를 입력할 수 있는 인식기를 사용하고 있다.

라. NOMINA-Voice Dialing on PABX

이 서비스는 원하는 사람의 이름을 말하면 자동적으로 그 번호를 dialing해준다. 물론 인식결과는 TTS를 사용하여 들려준다. 회사전화번호부에 의해 인식어휘가 자동적으로 생성되며, PC-Dialogic Antares platform을 이용하여 개발되었다. 이는 1994년부터 시작되어 몇 군데서 시험운용중에 있으며, 약 300명에서 1,200명의 이름을 인식할 수 있다.

마. VAD-Voice Activated Dialing

이는 개인적으로 20명까지의 이름과 전화번호를 등록할 수 있으며, 일반전화망에서 이름만으로 dialing을 가능하게 해준다. 화자종속인식기를 사용하였으며 1996년 전반기에 108명의 사용자를 대상으로 시험을 하였다.

바. DEMOS-Automatic Telephone Interview System

이 서비스는 교환원없이 DTMF 또는 ASR을 이용하여 면접을 하는 것이다. 현재 1996년 말부터 시험서비스를 개시하였다.

사. Gas Meter Reading by Phone

전화선을 통해서 고객 코드(10~11자리)에 대한 음성입력과 gas meter를 읽는다. 두 개의 gas회사에서 서비스 중이다.

아. Fax Yellow Pages

원하는 광고주에 대한 새로운 정보를 fax로 전송해주는 서비스로 광고주 전화번호와 fax 전화번호에 대한 숫자입력을 DTMF나 ASR을 이용하여 받는다. 1994년부터 시험을 시작하였다.

자. INFORMACITY Service

Venice에 대한 설명을 해주며, 60회선 규모로 1995년부터 운영해 오고 있다.

차. RAILTEL-Italian Train Timetable Service[3]

이탈리아 주요 도시간에 철도연결에 대한 정보를 제공한다. 500개의 도시명과 날짜, 숫자 등 고립단어를 인식하며, 4회선규모의 PC-Dialogic Antares platform을 사용하여 개발되었다. 1995년 9월부터 운영하기 시작했다.

타. DIALOGOS-Italian Train Timetable Service

차와는 달리 자연스런 언어를 사용할 수 있게 개발되었다. 1990년 첫 시스템이 개발되었으며 3,500단어를 인식할 수 있다. DEC workstation에서 Dialogic telephone board를 사용하여 한 개의 채널을 수용할 수 있으며 TTS와 연속음성인식, 그리고 담화기술들이 적용되었다. 1995년말 500명의 사용자를 대상으로 서비스되고 있다.

2.1.4 독일

Deutshe Telekom에서는 ASR 기술을 이용하여 다음과 같은 서비스를 제공하고 있다.

가. SCALL-a new paging service

이 서비스는 1995년에 시작되었으며, 사용자는 숫자정보를 특수한 호출기로 보낼 수 있다. 숫자음과 제어를 위한 명령어를 인식할 수 있다.

나. T-Card-an international virtual calling card service

이는 걸려오는 전화를 ASR기능을 이용하여 제어 할 수 있는 음성응답시스템이다. 사용자의 신분확인을 위하여 calling card number와 PIN code를 입력하여야 한다.

다. FAUST-a directory assistance demonstrator

현재 5,000명의 가입자와 25개 도시에 대한 전화번호안내를 해 준다. 이 시스템은 핵심어 추출기능, alphabet 인식과 음성합성기술을 사용하고 있다. 1996년부터 시험중에 있으며, 실제의 전화번호안내서비스에서 40개의 도시이름을 인식하는 반자동 시스템이 사용될 예정이다.

라. Tarifinfo

Deutsche Telekom의 통화요금에 대한 정보를 제공한다. DTMF입력과 정해진 어휘를 인식하며, 교환원과 연결가능하다.

2.1.5 스페인

현재 몇가지 서비스가 시험운용중이며, 1997년에 상업적인 서비스를 개시할 예정이다. 이미 서비스중인 것중에는 "CAMPIN" 시스템이 있는데, 사용자가 Telefonica의 calling card의 PIN을 변경할 수 있게 해 준다. 현재는 숫자음 인식을 하지만, 자연스러운 연속숫자음 인식기능을 추가할 예정이다. 그 외에는 voice dialing, answering machine, automatic directory assistance 등 앞에서 설명한 국가들이 제공하고 있는 서비스들에 대하여 다각적인 시도를 하고 있는 것으로 보인다.

2.2 미국

AT&T는 주로 전화교환원의 일을 자동화하는데 음성처리기술을 적용하고 있다.

가. VRCP-Voice Recognition Call Processing

1992년 다섯 단어{collect, calling card, third number, person, operator}를 인식하는 시스템으로 시작하여 현재 이 다섯 단어에 대한 spotting과 연속단어인식기술이 추가되어 미국 48개 주에서 사용되고 있으며, 일년에 수십억 call을 처리하고 있다. 인식률을 90%이다.

나. Universal Card 24Hr Customer Services(1-800-423-4343)

카드소유자에 대한 신용정보를 조회하는데 연속숫자음 인식기술을 이용하여 카드번호, 비밀번호, 서비스선택 등을 할 수 있다. 현대 한 달에 400만 call이상을 처리하며 97%이상의 인식률을 보이고 있으며, call의 57%를 자동으로 처리하고 있다. 숫자입력방식에서 DTMF보다 음성입력을 더 선호하고 있는 것으로 나타났다(45%@55%).

다. VoiceDialing[5]

NYNEX에서 1992년 12월 처음 개발된 이후 93년 중반부터 실제 서비스에 들어갔으며, 94년 New York와 New England에서만 1만5천 가입자를 확보하였다. 이 서비스는 30, 50, 70명까지 개인의 전화번호 및 인명을 등록하여 사용할 수 있다. On line help기능을 가지고 있으며, DTW를 사용한 화자종속음성인식방법을 사용하고 있다. 1995년부터는 DTW에서 continuous density multi-gaussian mixture HMM방식으로 변환을 시작했으며, 화자독립 명령어와 화자종속이름인식기를 동시에 사용할 수 있도록 개선하고 있다.

2.3 일본

가. ANSER[4]

NTT가 1981년 개발한 시스템으로 은행업무에 대한 정보를 제공한다. 초기에는 voice response기능과 DTMF 수신기능만이 있었고 후에 숫자인식기능이 결합되었다. 1990년까지 일본내 601개의 은행에서 사용하고 있다.

나. Voice Dialing

KDD에서 개발하여 1995년 11월부터 사용하고 있는데 PBX에 환경에서 이름을 말하면 전화를 걸어주며, 5,000명의 이름을 인식한다.

다. Operator Assistance

국제전화에 대한 사용자의 문의사항에 대답하므로 써 KDD의 교환원을 도와주기도 하며, 다른 종류로는 핵심어의 존재여부를 조사하여 외국인의 국제장난전화를 차단하는 역할을 하고 있는 것도 있다.

라. 날씨정보 안내

일본내 각 지역 명이나 휴양소를 말하면 그

곳의 날씨를 알려준다.

마. 새소리 안내

일본 내 야생조류(189종류)의 소리를 들려주며, 1995년 8월부터 서비스를 시작하였다.

바. Home Banking

원하는 은행에 돈을 송금할 수 있게 해주며, 계좌번호, 암호, 이체금액 등을 인식하며 아직 시험중에 있다.

2.4 우리나라

국내에서의 음성인식연구는 KT, ETRI, KAIST와 일부 대학연구소에서 지속적으로 연구되고 있다. 전화망에 관련기술을 적용한 경우는 사람이름이나 부서명을 인식하여 전화번호를 안내해 주는 시스템들이 시험적으로 몇 군데서 개발된 적이 있으나, 아직 일반인들을 대상으로 상용서비스를 제공하고 있지 못하고 있다. 다행이 한국통신(KT)에서는 음성인식증권정보서비스, 음성다이얼링서비스 등과 관련 시스템들을 개발하였고, 현재 시험서비스중에 있으며, 또한 여기에 적용된 기술들을 국내시장을 활성화시키기 위하여 업체로 이전하였고, 관련서비스를 개발할 수 있도록 도와준 결과 LG정보통신, 삼성전자, 삼보정보통신 등에서 음성인식전화정보시스템을 개발하여 증권 정보서비스를 구현하였으며, 이를 바탕으로 다양한 서비스를 구상하고 있다. 그리고 컴퓨터 회사에서는 PC용 음성인식기로 개량하여 응용제품을 개발할 예정이다.

가. 증권정보서비스[6]

1995년 11월 9일부터 PC-base로 된 시스템을 2회선 규모로 시험서비스를 시작하였으며, 대용량 회선을 수용할 수 있는 VME-base시스템을 개발하여 이것으로 시험서비스를 준비중에 있다. 이 시스템은 화자독립, 어휘독립기술이 적용되어 인식단어를 추가 및 변경이 자유로우며, barge-in과 echo cancellation 기능이 있다. 약 300개의 CDP를 사용하여 1,000개의 회사명을 실시간으로 인식할 수 있다.

또한 PCS회사 중에서는 한국통신의 음성인식기술을 이용하여 개발된 시스템을 삼성전자로부터 도입하여 증권정보서비스를 1997년 12월 10일부터 제공하고 있으나, 통신망이 불안하여 입력된 말의 일부가 잘리는 경우가 발생하여 인식률이 낮은 형편이며 망이 안정되면 이 문제도 해결되리라 기대하고 있다.

나. Voice Dialing Service

화자종속기술을 적용하여 휴대통신사업자인 SK-telecom과 신세기이동통신등에서 외국의 기술과 장비를 이용하여 1997년 초에 상용서비스중에 개시하였다. 한국통신에서는 화자독립기술을 이용하여 1997년 2월에서 10월말까지 서울지역에서 시험서비스를 실시하였으며 150개의 정해진 단어중에서 선택하여 전화번호를 등록, 사용할 수 있으며, 위치독립과 위치종속서비스가 가능하다. 이 서비스 역시 화자독립 인식기술이 적용되었다.

다. 이름인식 부서안내서비스

1997년말에 개발된 것으로 lab-test를 중에 있다. 1000명정도의 이름을 인식할 수 있지만 이름의 특성상 인식률이 다른 서비스보다 낮은 편이어서 연구소 내에서 시험서비스를 하면서 개선시킬 예정이다. 이 서비스는 사람이름을 말하면 해당사람의 전화번호와 부서를 안내해 주며, 사설교환기에 연결하면 전화도 걸어줄 수 있다.

3. 결 론

본 논문에서는 음성 인식기술을 전화망에 적용한 서비스에 대하여 살펴보았다. 전반적으로 인식단어수는 많지 않으면서 90%이상의 높은 인식률을 가지고 있었으며, 숫자음 인식기술을 많이 적용되었다. 또한 사용자들이 보다 편리하게 사용할 수 있게 barge-in이나 핵심어 추출기능 등에 기술이 적용되고 있었다. 우리나라도 1997년 초에 비록 국내기술을 이용한 것은 아니지만 음성다이얼링서비스가 시작되었으며, 한국통신이 개발한 두 가지 정도의 서비스가 시험서비스중에 있다. 1998년에는 한국통신으로부터 이전 받은 음성인식기술을 바탕으로 관련시장에 서비스개발이 더욱 활성화될 전망이다.

참고문헌

[1] C. Sorin et al. : "Operational and experimental French telecommunication services using CNET speech recognition and text-to-speech synthesis", Speech Communication, Vol 17, pp. 275~286, November 1995.

[2] F.A. Westall, R.D. Johnston, A.V. Lewis, (Editors) : "Speech Technology for Telecommunications", British Telecom Technical Journal, Vol 14, No 1, January 1996.

[3] R. Billi et al. : "Field Trial Evaluation of Two Different Inquiry Systems" ; *Proc.* this IVTTA '96 Workshop.

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

[5] George J. Vysotsky, "VoiceDialing-The first speech recognition based service delivered to custom's home from the telephone network," *Speech Communication,* Vol. 17, pp. 235~247, 1995.

[6] 김재인, 구명완, "음성인식 증권정보시스템의 개발 및 시험운용결과 분석," 음성통신 및 신호처리 워크샵 논문집, 제13회, pp. 185~191, 1996.

[7] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition,* Prentice-Hall, NJ, 1993.

김 재 인

1981 고려대학교 전자공학과 졸업(공학사)
1986 고려대학교 대학원 전자공학과 졸업(공학석사)
1986 금성전기연구소(연구원)
1988~현재 한국통신 멀티미디어 연구소 음성언어팀 선임연구원
1996 고려대학교 대학원 전자공학과 졸업(공학박사)
관심분야 : 음성인식, 화자인식, 음성합성 등 음성처리분야