호텔예약을 위한 자동통역 시스템

구 명완, 김 응인, 김 재인, 도 삼주, 강 용범, 박 상규,
손 일현, 김 우성, 장 두성, 이 종락*, 김 진영**

한국통신 연구개발원 소프트웨어연구소 음성언어연구팀

*한국통신 연구개발원 통신시스템개발센터, **전남대학교 전자공학과

An Experimental Speech Translation System for Hotel Reservation

Myoung-Wan Koo, Eung-In Kim, Jae-In Kim, Sam-Joo Doh,
Yong-Bum Kang, Sang-Gyu Park, Il-Hyun Sohn, Woo-Sung Kim,
Du-Seong Chang, Jong-Rak Lee* and Jin-Young Kim**

Spoken Language Research Team, Software Research Laboratory,
Korea Telecom Research Laboratories
*Systems Development Center, Korea Telecom Research Laboratories
**Dept. of Elec. Eng., Chonnam National University

< 요 약 >

이 논문에서는 한국에 있는 손님이 한국어 만을 사용하여 일본 호텔을 예약할 수 있도록 해 주는 한일간 자동통역 시연 시스템에 관해 기술하였다. 이 시스템은 한국어 음성인식부, 한일 기계번역부, 한국어 음성합성부로 구성되어 있다. 한국어 음성인식부는 기본적으로 HMM(Hidden Markov Model)을 이용하는 화자독립, 약 300 단어급 연속음성인식 시스템으로서 전향 언어 모델로 바이그램 언어 모델, 후향 언어 모델로는 의존 문법을 사용하여 N-Best 문장을 생성해 낸다. 실험 결과 단어 인식률은 Top1 문장에 대해 약 94.5%, Top5 문장에 대해 약 98.5%이었으며, 문장 인식률은 Top1 문장에 대해 약 81.2%, Top5 문장에 대해 약 94.7%의 인식률을 얻었다. 인식 시간은 길이가 다른 여러 문장들에 대해 약 0.1 ~ 3초가 걸렸다. 기계번역부에서는 음성인식에서 의존 문법을 사용하여 분석된 파싱 결과를 이용, 직접 번역 방식을 채택하여 일본어를 생성한다. 음성 합성부는 반음소를 합성의 기본단위로 하고, 합성방식으로는 주기 파형 분해 및 재배치 방식으로 하였다. 실험 환경은 2 CPU를 장착한 SPARC 20 workstation이었으며 실시간 특징 추출을 위해 TMS320C30 DSP 보드 1개를 이용하였다.

1. 서 론

각종 정보통신 기술의 발달로 인해 사회가 다변화되고 국제화됨에 따라 다른 언어권과의 의사소통의 필요성이 증대되고 있으며, 자동통역에 대한 관심이 고조되고 있다. 자동통역 기술은 음성인식, 음성합성, 기계번역, 이 세가지 기술이 결합된 첨단기술이다. 한국통신에서는 호텔예약이라는 제한된 범위내에서 자동통역을 제공해 주는 자동통역 시연 시스템을 개발하였다. 이 논문에서는 이 시스템에 대해 기술하였고, 이를 통해 자동통역 시스템의 가능성을 제시하고자 한다.

본 시스템에서 다루고 있는 문제영역은 호텔예약으로 한정하였다. 이는 자동통역이라는 기술 자체가 아직까지 실용화가 덜었기 때문에 그 영역을 한정하지 않고는 개발하기가 어렵기 때문이며, 또한 서비스의 관점에서도 호텔예약이라는 영역이 매우 유용할 것이라 판단되었기 때문이다. 이 한정된 영역 내에서 우리는 호텔예약을 위해 실제로 사용되고 있는 자료들을 수집, 분석하여 이를 바탕으로 이 시스템에서 처리할 수 있는 말들에 대한 문법을 작성하였다. 이 문법의 예를 그림 1에 보였으며 이 문법에서 사용되고 있는 전체 단어수는 약 300단어이다. 또 이 문법으로 생성가능한 전체 문장수는 100만 문장 이상이다.

(여보세요) {(방) 예약(을)|방을 예약)}{하고
{싶습니다|싶은데요|하려고 {하는데요|합니다}} .

{성글|더블|트윈}((으)로) (하나|둘) {부탁합니다|(해)
{주세요|주십시요}}.

그림 1. 문법의 예

본 시스템은 음성인식부, 기계번역부, 음성합성부로 이루어져 있다. 음성인식부는 한국 손님이 일본호텔을 예약하고자 할때 사용하는 말들을 인식하는 부분으로서 화자독립 연속음성인식을 수행하고 있다. 기계번역부는 인식된 한국어 결과를 받아서 이를 일본어로 번역시켜 주는 부분이다. 음성합성부는 일본에서 전송한 한국어 텍스트를 받아서 음성합성을 통해 한국손님에게 알려주는 부분이다. 이 시스템에서 수행하는 작업은 한국 손님이 일본의 호텔을 예약하고자 할 경우에 필요한 부분, 즉 한국어 음성인식, 한->일 기계번역, 한국어 음성합성이다. 또 이에 대응되는 일본 호텔측의 부분은 일본 KDD에서 개발하였으며 이 두 시스템들은 다이얼업(dial-up) 모뎀을 통하여 서로 데이타를 교환할 수 있다. 단 데이타 전송의 안정성을 보장하기 위해 일본 KDD와의 국제 대모에는 전용회선을 사용하여 데이타를 전송하였다. 전체 시스템은 TMS320C30 DSP 보드를 장착하고, 2개의 CPU를 갖는 SPARC 20 workstation 상에서 구현되었다.

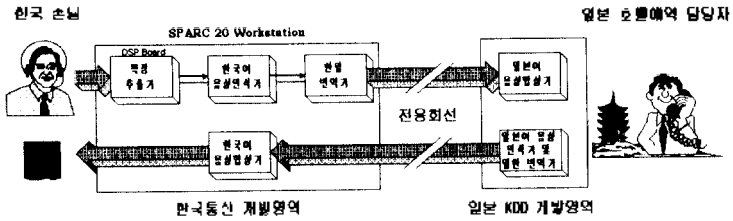

그림 2. 시스템 구성도

2. 음성인식부

본 시스템에서는 음성인식부에 문법과 조음화현상 등 연속음성의 특징을 고려한 N개의 최적문장을 찾을 수 있는 한국어 연속음성인식 알고리즘을 사용하였다[1]. 음성이 입력되면 음성의 특징이 추출되고 추출된 특징을 이용하여 단어 단위의 비교가 이루어진다. 비교결과는 후보단어의 열이된다. 문장인식은 언어모델을 이용하여 수행되며 결과는 문장이 된다.

2.1 특징추출

음성신호는 8kHz, μ-law 8비트로 샘플링되고 $1-0.95z^{-1}$의 전달함수를 갖는 필터를 사용하여 pre-emphasize된다. 이 음성은 프레임 단위로 분할되어 처리되는데 각 프레임은 20msec의 길이를 가지며 10msec씩 중첩된다. 매 프레임은 LPC 분석이 수행되고 이 LPC 계수를 이용하여 cepstral 계수가 구해진다. 각 프레임에서 구한 LPC 계수는 아래의 가중치 윈도우(weight window) W_c에 의해 가중치가 계산된다.

$$W_c(m) = 1+\frac{\theta}{2}\sin\left(\frac{\pi m}{\theta}\right) \qquad 1\le m\le Q$$

여기서 Q는 LPC 차수이다. 음성인식에는 weighted LPC cepstral 계수와 그들의 델타(difference), 이차델타(second order difference), 로그파워의 일차, 이차델타 값을 벡터 양자화하여 사용한다.

2.2 음소 모델

HMM에 근거한 음성인식 시스템은 음성인식의 기본단위가 필요한데 본 시스템은 음소와 유사한 단위(phoneme-like unit)를 사용하였다. 기본 유니트 갯수는 56개를 사용하였으며 조음화현상을 고려하여 문맥종속 음소를 구하였다. 음소 모델은 7개의 상태와 12개의 전이를 가지며 그 전이들은 세개의 그룹으로 묶을 수 있으며 같은 그룹의 전이는 같은 출력확률을 갖게 된다[2].

본 시스템은 아래와 같이 세 종류의 조음화현상을 고려하여 300개의 문맥종속 음소를 구하였다.

1) 묵음 모델 : 연속음성을 발음할 경우 단어사이의 묵음은 사람에 따라서 지켜지거나 연이어서 발음을 할 수 있다. 본 시스템에서는 null transition을 만들어서 묵음을 모델링하였다.

2) 단어내 조음화 모델 : 단어내의 조음화 현상을 모델링하기 위하여 트라이폰(triphone)을 사용하였다. 트라이폰 갯수를 결정할 때는 양이 너무 커지지 않게하기 위하여 unit reduction rule을 사용하였다[3].

3) 단어간 조음화 모델 : 단어간 조음화 현상은 매 단어의 앞과 뒤에서 발생한다. 본 시스템은 훈련시에는 단어간 조음화 현상을 하나의 트라이폰으로 모델링하였으며 인식단계에서는 가능한 모든 트라이폰으로 모델링하였다. 그림 3에 단어간 조음화 현상을 고려한 트라이폰 구성을 나타내었다.

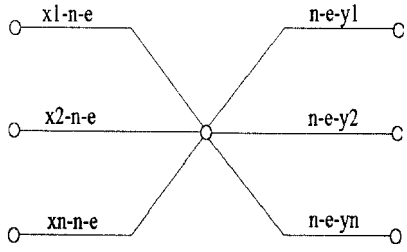

그림 3. 단어간 조음화 모델

2.3 언어처리 알고리즘

연속음성에서 언어처리 알고리즘은 인식시간 및 성능에 중요한 역할을 한다. 본 시스템에서는 구문분석 모델로 바이그램(bigram) 문법을 사용하였다. 바이그램 문법은 단어 w_1 다음에 단어 w_2가 올 확률값 $P(w_1/w_2)$를 훈련시 구하고 인식단계에서는 이 값을 단어 천이 확률값으로 사용한다[4]. 의미모델은 사용하지 않는다.

2.4 탐색 알고리즘

탐색알고리즘은 Viterbi 알고리즘을 사용하였으며 인식시간을 향상시키기 위하여 beam 탐색 알고리즘을 사용하였다. 최적문장의 갯수 N은 5를 사용하였다. N개의 최적문장을 찾는 알고리즘은 word-dependent 알고리즘을 사용하였는데 단어내의 매 state에서 N보다 작은 n개의 가능한 path를 저장한다[5].

2.5 데이타 베이스

음성데이타 베이스는 호텔예약에 많이 쓰이는 약 300개의 단어를 사용하여 20대부터 50대까지의 남성 및 여성으로 구성된 60명이 1인당 약 100문장씩을 발음한 것으로 구성하였다. 화자는 헤드세트(headset) 마이크를 사용하여 발음하고 발음된 음성은 CS4215 코덱이 내장된 A/D 변환장치 SAIB를 통해 컴퓨터에 저장하였다. 데이타 베이스의 구성은 표 1과 같다.

표 1. 데이타 베이스 구성

연 령	사 람 수	
	남	여
20대	15	5
30대	15	5
40대	10	5
50대	5	

3. 기계번역부

이 시스템에 포함된 대화체 기계번역 시스템은 크게 두가지의 부분으로 구별할 수 있다. 그 첫번째 부분인 한국어 분석 과정에서는 의존문법을 이용하여 문장의 후미에서 전향으로 분석해 나가는 방식을 사용하였으며, 구문분석 과정은 음성인식 과정의 후향언어 모델로서의 역할을 동시에 수행함으로서 비문법적인 음성인식 결과를 조기에 제거하는 효과를 거뒀다. 또한 이 시스템은 직접변환 방식을 채택함으로서 그 두번째 부분인 일본어 생성 과정에서 약간의 생성문법만으로 일본어 음성합성기에서 쓰일 수 있는 일본어 문장을 생성할 수 있다. 전체적으로 이 시스템은 실시간 대화체 기계번역이라는 목적으로 전체 시스템의 효율을 극대화할 수 있도록 설계, 구현되었다.

이 시스템은 한국어의 부분자유어순의 특징에 맞게 설계된 의존문법[6]을 구문분석과정에 사용함으로서 기존의 구구조문법을 사용하던 방식에 비해 적은 양의 단위 문법으로도 한국어의 특징에 맞게 구문적 제약을 가할 수 있었다. 또한 한국어를 보다 일반적으로 기술할 수 있기 때문에 통역영역의 확장 및 변경등에 큰 변화없이 적용이 가능하며, 호텔예약과 같은 특정 영역의 자동통역에 적용할 경우에는 어절간의 의존관계에 의미적 제약을 주는 방법으로 의미적 분석까지도 가능하였다.

또한 이 시스템은 한국어를 의존문법을 사용하여 구문분석할 때 문장의 후미에서 전향으로 우향우선 구문분석하여 되면 문장이 가져야 할 올바른 구문구조가 문장의 분석초기에 밝혀지기 때문에 문장내에서 구문구조에 어긋난 어절을 빨리 찾아낼 수 있었으며, 이와 같은 특성에 근거하여 기계번역의 구문분석 과정은 음성인식기의 후향언어모델로서의 역할을 같이 수행하고 있다. 이렇게 의존문법을 후향 언어모델로 사용하여 문장을 우향우선 분석하게 되면 오인식된 문장을 찾아내는 계산시간이 적게 소요되므로, 전체적인 음성인식 시간을 줄일 수 있는 장점이 있다. 또한 실제 실험한 결과 기존의 전향탐색만을 하는 연속음성 인식시스템에 비해 가장 높은 가중치를 갖는 인식결과의 경우 문장 오인식률은 6.98%, 단어 오인식률은 10.59%가 감소되었다(5개까지의 인식결과를 고려할 때 문장 오인식율은 17.28%, 단어 오인식률은 8.47% 감소).

그림 4는 실제 대화체 기계번역 시스템의 설명도이다. 여기에서 음성인식의 전향 언어모델을 거쳐 선택된 인식후보에 대하여 우향우선으로 구문분석하며, 그 과정에서 비문법적인 문장을 조기에 걸러내는 역할을 한다. 변환 및 생성 과정에서는 구문분석된 의존트리를 입력으로 일본어를 생성한다. 이와 같은 구성으로 음성인식과 기계번역에서 각자 사용될 수 있는 두번의 구문분석과정을 하나로 통합함으로서 통역시간이 감소시켰으며, 시스템의 구성을 간소화하였다.

4. 음성합성부

음성합성부는 음성학적 전처리부, 운율발생부 그리고 합성부의 중요한 세 부분으로 나누어진다. 음성학적 전처리부에서는 화일이나 키보드입력에 의한 문장을 성분분석하여 음운변동을 수행하고 운율정보발생을 위한 기본적인 구문분석을 수행한다. 다음 운율발생부에서는 전처리의 결과를 받아 한국어에 적합한 억양, 길이, 세기 등의 운율을 발생시킨다. 합성부에서는 합성의 기본단위를 가져와 연결시킴으로써 문장을 만들어내는데, 운율구현 및 음소간의 인터폴레이션(interpolation)을 동시에 수행한다[7].

4.1 합성 단위

본 시스템에서는 새롭게 제안된 반음소(demiphone)를 합성의 기본단위로 사용한다. 반음소는 음소를 그것의 정상상태시점인 중점을 기준으로 해서 다시 양분함으로써 얻어진다. 음소를 양분하여 얻어진 두개의 반음소 중에서 먼저 것을 전반음소(initial demiphone), 나중 것을 후반음소(final demiphone)라고 한다. 반음소의 경계는 음소 및 다이폰의 경계와 일치한다. 따라서 전반음소와 후반음소들을 적당히 결합함에 따라 음소를 만들 수도 있고 다이폰을 만들 수도 있다. 이와 같은 성질 때문에 반음소는 음소와 다이폰의 장점을 동시에 가지게 된다. 다시 말하자면 음소와 마찬가지로 다루기 쉽고 메모리 양을 적게 필요로 하며, 다이폰과 마찬가지로 합성시 얻어지는 합성음성의 음질이 좋게 되는 장점이 있다[8].

4.2 음성학적 전처리

무제한 음성합성 시스템에 있어서 음성학적 전처리 단계는 문자열 정형화부(text preformatting block), 문장구조 추출부(parsing block), 음운변동 처리부(phonetic recoding block)로 세분될 수 있다.

문자열 정형화부는 문자열이 입력되면 숫자, 알파벳, 많이 쓰이는 영어단어, 약자 등을 입력한 사전을 참조하여 문자열 속에 있는 모든 익어, 숫자, 특수기호 및 수식, 수사처리를 하여, 발음가능한 문자열과 제어문자 구둑기호(punctuative symbol)들로만 구성된 정형화된 문자열을 생성하며, 1300여 단어의 발음예외사전의 탐색 및 치환과정을 수행한다.

문장구조 추출과정은 조사,어미, 선어말어미 등으로 구성된 형식형태소 사전(약 400단어)과 관형사, 부사, 불완전명사들로 구성된 실질형태소 사전(약 230단어)을 참조하여 정형화된 문자열에 대해 구문해석을 함으로써 수식, 피수식 관계 및 구, 절의 경계점을 검출하고, 구 및 절의 통사적 기능을 결정한다. 음운변동 처리부는 발음예외사전 탐색결과 형태소 분석결과 문장구조 추출결과 및 음운변동 알고리듬을 이용하여 자소 단위의 변환을 수행한다[9][10].

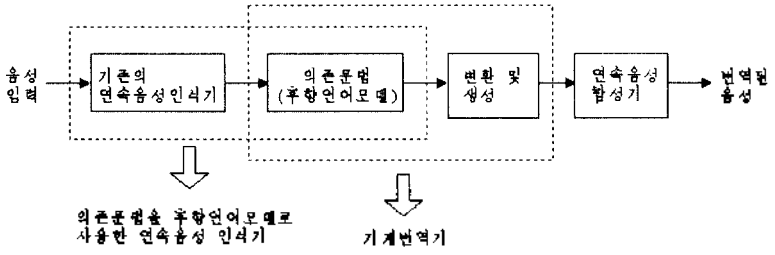

그림 4 음성인식 시스템과 기계번역 시스템의 결합

4.3 운율 생성

운율조절부에서는 운율의 기본 요소인 각 음소의 길이, 억양 그리고 세기를 조절한다. 첫째로, 길이 조절을 위하여서는, 음소의 길이가 음성학적 요인, 구문론적요인 그리고 발음속도에 의하여 변하는 것으로 모델링을 하여 길이조절규칙을 제안하였다. 즉, 임의의 음소의 길이는 전후의 음소의 음성학적 성질에 따라 적절히 늘어드는 것으로 모델링을 하였으며, 문장의 끝이나 억양구의 끝에서는 길이가 늘어나도록 하였다. 둘째로 억양에 있어서는 어절단위, 억양구단위 그리고 절단위의 억양조절규칙을 사용하였는데, 이 조절규칙으로는 기본적으로 어절단위의 피치패턴규칙과 baseline resetting규칙으로 이루어져 있다. 다음 세번째로 각 음소의 세기조절규칙에서는 먼저 음소별 기준세기를 정한 후에 부여된 피치의 크기에 따라서 선형적으로 증감하는 방법으로 사용하였다[11].

4.4 합성 방식

원음성(original speech) 중의 유성음 구간의 신호를 각 성문펄스(glottal pulse)에 의해 만들어지는 한 주기분의 음성파형에 해당하는 단위파형(unit waveform 또는 wavelet)들로 분해하는 주기파형분해 방식과 저장된 단위파형들 중 배치시키고자 하는 위치에 가장 가까운 단위파형을 선택하여 그것들을 서로 중첩시킴으로써 원음성의 음질을 그대로 가지면서도 음성단편의 지속시간(duration)과 피치주파수(pitch)를 임의대로 조절할 수 있게 하는 시간 왜곡식 단위파형 재배치방식을 합성 방식으로 사용한다.

주기적 음성을 그것의 스펙트럼 포락 함수의 시간영역 함수인 임펄스 응답과, 주기적 음성과 주기가 같고 평탄한 스펙트럼 포락을 가진 주기적 피치펄스열 신호로 디콘벌루션(deconvolution)한 다음 이것으로 검출 알고리즘(epoch detection algorithm)[12]과 같은 시간영역에서의 피치검출 알고리즘을 이용하여 주기적 피치펄스열 신호나 시간영역의 음성 파형으로부터 피치펄스들의 위치를 구한 후 피치펄스가 한 주기구간당 하나씩 포함되도록 피치펄스열 신호를 주기적으로 분할하고, 유효지속시간에 따라 파라미터 연장과 영샘플 추가를 한 후, 이 피치펄스신호들을 그 주기구간 동안의 임펄스응답과 다시 콘벌루션시키면 단위파형이 구해진다.

5. 성능평가

본 자동통역 시스템은 손님측의 발화가 끝나는 시점을 기준으로 통역된 음성이 합성되어 나오는 시간까지 약 3초가 걸리는 실시간에 가까운 시스템이다. 본 시스템에 사용된 음성인식 시스템은 나이에 고르게 분포된 60명의 발성화자의 총 4762발화를 학습자료(training data)로 사용하여 학습자료에 포함되지 않은 4명의 발성화자의 644발화를 선택하여 실험하였을 때 81.21%의 문장인식률을 보였다. 표에서 Top1은 첫번째 후보만을 고려한 인식률이며, Top5는 5번째 후보까지 포함하였을 경우의 인식률이다.

표 2 의존문법을 후향 언어모델로 사용하는 한국어 연속음성 인식시스템의 성능평가표

단어 인식률 (%)		문장 인식률 (%)	
Top1	Top5	Top1	Top5
94.53	98.50	81.21	94.72

6. 결 론

이 논문은 한국 손님이 한국어 만을 사용하여 일본호텔을 예약할 수 있도록 해 주는 호텔예약을 위한 자동통역 시연 시스템에 관한 것이다. 자동통역 기술은 음성인식, 합성, 기계번역이 결합된 첨단기술로서 본 연구에서는 한국어 음성인식, 한->일 기계번역, 한국어 음성합성 시스템이 결합된 자동통역 시연 시스템을 구현하였다. 일본 호텔측의 영역은 일본 KDD에서 개발하였고, 우리가 개발한 시스템과 다이얼업 모뎀 또는 전용선을 통해 통신할 수 있다. 음성인식 시스템은 HMM을 이용하는 300 단어급 화자독립 연속음성인식 시스템으로서 전향언어모델로 바이그램, 후향언어모델로 의존문법을 사용하여 N-best 문장을 생성해 낸다. 기계번역은 직접번역방식을 채택하였고, 음성합성은 반음소를 기본 단위로 주기파형 분해 및 재배치 방식을 사용하였다. 이 시스템을 사용하여 일본 KDD 시연 시스템과의 각개 데모에 성공하였으며 이를 통해 자동통역의 가능성을 제시하였다고 할 수 있다.

참 고 문 헌

[1] 구 명완, "N개의 최적문장을 찾을 수 있는 한국어 연속음성인식 시스템", 음성통신 및 신호처리 워크샵 논문집, pp. 48-51, 1994년 10월.

[2] K. F. Lee, *Automatic speech recognition : the development of the SPHINX system*. Kluwer Academic Publisher, Norwell, Mass., 1989.

[3] C. H. Lee et al., "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language*, vol.4 pp.127-165, 1990.

[4] F. Jelinek, "The development of an experimental discrete dictation recognizer,"*Proc. IEEE*, pp. 1616-1624, 1985.

[5] R. Schwartz and S. Austin, "Efficient, high-performance algorithms for N-best search," *Proc. of the DARPA speech and natural language workshop*, pp. 6-11, 1990.

[6] I. A. Mel'cuk, Dependency Syntax: Theory and Practice, The State Univ. of New York Press, 1988

[7] 김 웅인, 김 재인, "한소리: 무제한 음성 합성 시스템," 음성통신 및 신호처리워크샵 논문집, pp.342-345, 1994년 10월.

[8] 이 종락,"반음소:새로운 음성합성 및 인식단위," 음성통신 및 신호처리워크샵논문집,1993년8월.

[9] 강 용범, 안 치홍, " 무제한 음성합성 시스템을 위한 문장구조 추출에 관한 연구," 음성통신 및 신호처리 워크샵논문집, 1993년 8월.

[10] 강 용범, 김 진영, "무제한 음성합성 시스템을 위한 전처리 과정" 음성통신 및 신호처리워크샵 논문집, pp.334-337, 1994년 10월.

[11] 김 진영, 성 굉모, "한국어의 억양에 관한 연구," *Korean-Japan Joint Symposium on Acoustics*, pp. 292-297, 1991.

[12] C.d'Alessandro, J.S. Lienard, "Decomposition of the Speech Signal into Short-Time Waveform Using Spectral Segmentation," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1988