음성인식과 음성합성에 있어서의 음성학과 음운론의 역할

김 기 호(고려대 영어영문학과)

1.0 소 개

음성인식과 음성합성의 활용은 자판없는 타자기에서부터 말하는 전자비서, 자동 호텔 예약장치, 자동 언어 통역 장치, 인간의 말을 알아듣고 말을 하는 인공지능을 가진 로보트에 이르기까지 실로 다양하다. 이러한 음성인식과 음성합성에 대한 관심은 이미 1930년대의 Vocoder와 1940년대의 스펙트로그램의 개발과 함께 시작되었지만 최근 컴퓨터의 급속한 발전에 힘입어 소규모 고립 단어 인식에서 이제는 대규모 어휘의 구어체 연속음성인식으로 그 관심 영역이 확대되고, 음성합성에 있어서도 보다 자연스런 음의 합성으로까지 확대되었다.

본 논문의 목적은 바로 이러한 연속음성인식과 자연스러운 음성합성을 위하여 음성학과 음운론의 지식이 어떻게 활용될 수 있는지를 보이는데 있다. 본 논문은 다음과 같이 구성되어 있다. 먼저 제 2장에서는 음성인식에 대한 간략한 역사적 개괄과 함께 음성인식의 몇가지 기본적 방식과 기존의 몇가지 모델들을 소개하고, 제 3장에서는 음성 합성의 기본적인 방식들을 소개하고자 한다. 그 후 제 4장에서는 스펙트로그램상에 나타나는 변별자질들과 운율자질들의 음향음성적 특성들을 살펴보고 이들 자질들이 어떻게 음성 인식과 음성 합성에 응용될 수 있는지를 보이고자 한다. 끝으로 제 5장에서는 음운부의 역할이 기존의 모델에서보다 더 확대된 연속음성인식의 모델을 제시하고 아울러 바람직한 음성인식과 음성합성을 위해서는 조음음성학과 음향음성학적 지식은 물론 보다 심도 깊은 청각음성학적 지식과 심리음향인지(psycho-acoustic) 실험이 필요함도 지적하고자 한다.

2.0 음성인식

2.1 음성인식의 역사적 개관

음성인식에 대한 관심은 1930년대 Vocoder(Duley 1936)의 개발로 거슬러 올라 가지만 실질적인 연구는 1940년대의 음성스펙트로그램의 개발 이후 Potter, Kopp & Green(1947)의 스펙트로그램 판독실험 등으로 더욱 구체화되기 시작하였다고 볼 수 있다. 그 후, 1950년대와 1960년대에는 주로 화자의존적인 소규모 어휘의 고립 단어의 음성인식에 많은 연구가 있어 왔고, 본격적인 연속음성인식에 대한 연구는 1970년대 미국방성의 후원을 받은 ARPA계획(1971-76) 이후에 시작되었다고 볼 수 있다. ARPA계획 결과 1975년에 HEARSAY-II system, WHIM system, HARPY system 등과 같은 1,000 단어 어휘 수준의 비교적 성공적인 연속음성인식 장치들이 소개되었다(Lea 1980 참조). 한편 일본 NTT의 Itakura(1975)는 같은 해에 DTW(dynamic time warping)라는 보다 효율적인 패턴 비교의 음성인식 방식을 소개하였다.

연속음성 인식에 대한 연구는 1980년대에도 계속되었는데, 1982년 벨연구소의 Wilpon et al.은 화자독립의 고립단어 1129개에 대한 음성인식 실험에서 91%의 인식율을 보인 것

으로 보고하였고, 1983년 카네기-멜론 대학의 FEATURE System(Cole et al. 1983)은 문법의 도움없이 자질에 기초한 방식만을 사용하여 90% 이상의 정확도로 영어철자를 인식한 것으로 보고되었다. 그리고 통계적 처리를 중요시하는 IBM 음성인식팀에서는 1985년에 비록 화자종속적이지만 5,000단어 규모의 문장에 대해 97%의 높은 인식율을 보여준 Tangora System을 개발하였다. 그리고 문맥의존적 음소인식 모델을 사용한 Bolt, Beranek & Newman(1987)의 BYBLOS System 역시 997 단어의 연속음성인식에 93%의 인식율을 보인 것으로 보고되었고, 1988년 HMM(Hidden Markov Model) 모델을 이용한 카네기-멜론 대학의 Sphinx System(Lee 1988)은 ARPA계획의 기존 997개 단어의 문장을 화자독립적으로 997, 60, 20의 문법 난이에 따라 각기 71%, 94%, 96%의 인식율을 보인 것으로 보고하였다. 대규모 어휘의 화자독립 연속음성인식에 대한 연구는 1990년대에 들어서도 일본, 유럽, 미국의 음성인식연구팀들에 의해 계속 진행되고 있으며 현재 상당한 수준에 이르고 있는 것으로 보고되고 있다.

한국에서도 비록 기초수준에 불과하지만, 소규모 단어의 인식에 대한 여러 실험들이 진행되고 있으며 (김순엽 1991 참조), 한국어의 연속음성인식을 위한 장기계획 보고서들도 제출되고 있다: 한국과학기술원의 「한국어 음성인식 시스템 개발 연구보고서」(1989)와 「한국어특질 및 대화체 기계번역에 관한 연구 보고서」(1991), 그리고 한국전자통신연구소의 「연속음성인식 기술개발에 관한 연구 보고서」(1988)와 「자동통역전화를 위한 요소기술 개발 보고서」(1991), 한국통신의 「한국어 특질에 관한 연구: 자동통역 전화시스템 구현을 위한 음운 및 문법구조 연구」(1993) 등.

2.2 음성인식의 기본 방식

주어진 음성파형으로부터 음성 특징들을 추출하여 적절한 음소와 단어로 연결시키려는 음성인식 방식들 중에서 현재 주로 사용되고 있는 방식 몇가지를 들면 다음과 같다.

가) Dynamic Time Warping(DTW)을 이용한 패턴 매칭 방법
나) Hidden Markov Models(HMM)를 이용한 음성인식 방식
다) Connectionist Network(CN)을 이용한 신경망 음성인식 방식
라) 음성 음운 지식을 이용하는 전문가 음성인식 방식

패턴 매칭 방식(Template-based approach)은 입력된 음성 자료의 스펙트럼을 분석하여 미리 입력 저장된 데이타 베이스의 표준 패턴과 대조시켜 가장 유사한 것을 찾아내는 음성인식 방식이다. 특히 일본의 Itakura에 의해 제시된 DTW 방식은 입력된 표준 패턴과의 비교시에 나타날 수 있는 시간적 차이의 영향을 최소화 하기 위하여 주어진 음성입력에 시간적 변화까지를 보완한 패턴 비교방식으로 소규모의 제한된 단어 인식에 비교적 좋은 결과를 보여주기 때문에 간단한 응용분야에 실제로 많이 이용되고 있는 방식이다. 그러나 이 방식은 화자 독립의 대규모 어휘의 구어체 연속음성인식에서는 인식율이 낮아지는 문제가 있다.

HMM 방식은 러시아 수학자 Markov에 의해 1900년대 초에 소개된 Hidden Markov Models에 기초하여 통계적 모델을 이용하여 음성을 인식하는 방식이다. 이 방식은 음성분할(segmentation)과 음성인식을 동시에 통계적으로 해결할 수 있기 때문에 1970년대 중반 이후 주로 대용량의 연속음성인식에 많이 이용되고 있으며, 현재까지 알려진 대용량 시스템으로서는 가장 성공적인 것으로 간주되고 있다.

CN의 신경망 회로 방식은 인간 두뇌의 생물학적 신경 계통을 모방한 인공신경망을 이용하여 여러 연결마디들에 음성 특질들을 분산 분포시켜 음성인식을 실현하고자 하는 방식으로 최근 새로이 부상되고 있는 연속음성인식 방식이다. 그리고 전문가 음성인식 방식(Knowledge-based approach)은 음향음성적 지식과 음운 지식을 바탕으로 마련된 화자독립의 연속음성인식 시스템으로 일본의 SPREX (A Speech Recognition Expert)나 V. Zue를 중심으로 한 MIT 대학의 음성인식팀에서 주로 이용하고 있다. 한편 최근에는 이들 방식들을 몇가지 혼합하여 각각의 장점만을 이용할 수 있는 혼용(hybrid) 방식이 사용되기도 한다.

2.3 음성인식의 기존 기본 모델들

고립단어 또는 연결 단어등 주로 단어 인식에 많이 사용되고 있는 음성인식 방식으로 VQ (Vector Quantization)을 이용한 음성인식 방식이 있는데, 이를 도식으로 나타내면 다음 그림 1과 같이 나타낼 수 있다.

그림 1. VQ를 이용한 음성인식 시스템

그림 1에서 보는 바와 같이 먼저 입력된 아날로그 음성신호를 저역 여과(Lowpass Filtering)시켜 표본화하는 전처리과정을 거친 후, LPC, PITCH, LPC-CEPTRUM 등을 이용하여 음성요소의 특징을 추출한다. 그 후 VQ를 이용한 음성인식은 미리 저장해 둔 특징벡터중에서 가장 잘 매칭되는 하나의 벡터와 매칭시켜 단어를 인식하게 한다. 이때 DTW와 같은 방식을 사용하여 시간적 차이를 보정한 후 가장 유사한 패턴과 매칭되도록 한다. 그러나 그림 1의 모델은 주로 단어인식에 사용되는 방식이기 때문에, 연속음성인식을 위한 모델에서는 어떠한 특징들을 양자화하여 추출할 것인지 그리고 대화체의 연속된 문장 인식에 있어서의 음성학과 음운론의 역할에 대해 좀 더 명확하게 규명해야 할 필요가 있다.

한편 고립단어가 아닌 대화체의 연속음성인식을 위한 기본 모델에서는 단어보다 하위단위인 음절, 걸친이음(diphone), 또는 음소를 기본인식 단위로 삼고 있는데, 일반적으로 다음 그림 2에서와 같이 음소를 기본으로 하고 있다.

그림 2. 연속음성인식의 주요 부문들

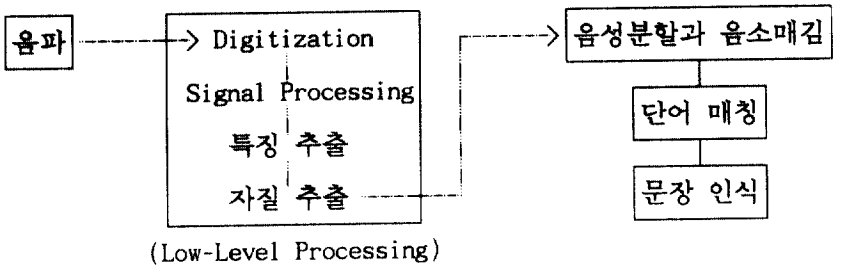


그림 2에서 보는 바와 같이 주어진 음파를 먼저 아날로그 신호에서 디지탈 신호로 바꾼다. 그 후 디지탈 신호는 필터를 통과하여 시간(가로축)과 주파수(세로축)를 함수로 음향에너지(세기)를 나타내 주는 스펙트로그램 모양으로 바뀌며, 여기에서 영교차율(zero-crossoing rate), 여러 주파수대의 에너지, 포먼트 흐름(formant track) 등의 패러미터를 추출한다. 이러한 패러미터로부터 음성자질들을 추출해 내며, 이러한 자질로부터 음소단위가 도출된다. 그리고 이러한 음소연쇄로부터 순차적으로 단어를 매칭하고 문장을 인식하게 된다(Church 1987 참조). 물론 여기서도 하위 단위과정에서 추출하여야 할 특징과 자질이 무엇인지, 그리고 상위단위의 처리과정에서의 음운부의 역할이 무엇인지에 대해 좀 더 구체적인 논의가 있어야 한다.

최근 김종미(1990)는 보다 언어학적 입장에서 한국어 음성인식을 위한 모델로 다음 그림 3의 모델을 제시하고 있다.

그림 3. 한국어 음성인식 모델 (김종미 1990)

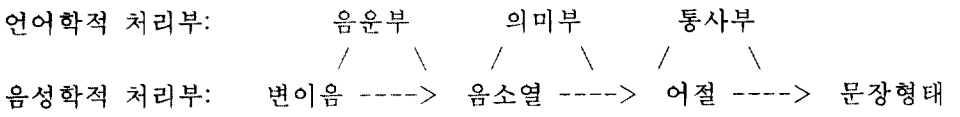


그림 3의 음성인식 모델에 의하면 음성처리부로부터 음성분할(segmentation)이후 인식된 변이음(phoneme-like unit)을 입력 단위로 하고 있는데, 음운부의 역할이 음소 유사단위로부터 일련의 음소열을 찾아내는 일에 국한되어 있다. 물론 여기에는 구개음화, 유성음화, 비음화와 같은 한국어 음운규칙들과 허용가능한 음소연쇄, 그리고 *#CC, *CCC, *CC#, *-r#, *rC 등과 같이 허용되지 않는 음소연쇄를 규정해 주는 한국어 음소배열제약(phonotactic constraints) 등을 이용하여 각 해당 음소 마다 허용하는 인접음소와 허용하지 않는 인접음소를 설정하여 가능한 주위 음소열을 예측하고 있다. 그러나 여기서도 음성처리부에서 주어진 음성 파형으로부터 어떠한 음성 특질들을 어떻게 추출할 것인지가 좀 더 구체적으로 밝혀져야 한다. 더우기 그림 3의 모델에서는 음운부의 역할이 단지 변이음으로부터 음소열을 찾아내는 역할만을 하는 것으로 기술되고 있으나, 앞으로 제 4장에서 논의되겠지만 연속음성인식의 경우 음성학과 음운론의 역할은 이보다 더 확장되어야 한다.

3.0 음성합성

음성합성이란 기계로 하여금 인간의 말소리와 유사한 음을 만들어 내는 것으로 주로 문장을 소리로 변환하는 문장-음성 변환장치를 가리키는데, 이러한 음성합성에는 기본적으로 조음음성학적 접근 방식(Articulatory synthesizer)과 음향음성학적 접근 방식이 있다: 전자는 인간의 성대와 혀 등 발성기관을 모방한 방식인 반면, 후자는 성대 떨림과 구강의 음향적 특성을 나타내기 위해 저주파 소음(buzz) 및 포먼트 생성기를 이용하는 방식이다.

3.1 음성합성의 종류

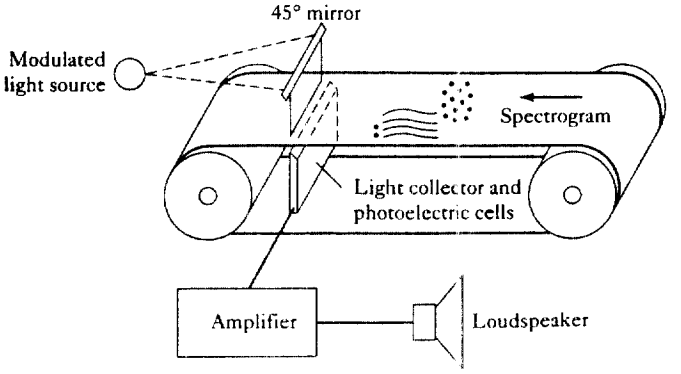
이러한 음성합성의 구체적인 종류는 다음과 같다.

가) 음향합성기(Acoustic Synthesizer): 18세기 후반 독일의 Wolfgan von Kepelen이 고안한 것으로 나무와 가죽으로 인간의 발성 기관을 모방하여 음성과 유사한 소리를 내게 한 장치이다.

나) 전기합성기(Electrical Synthesizer) 1930년대부터 전기 장치를 이용하여 소리를 생성해 내는 장치로 소위 분석/합성(Analysis-synthesis) 장치인 Vocoder (=Voice Coder)가 이에 해당된다. 특히 여러 개의 채널을 이용한 분석/합성 장치를 가리켜 채널 합성기(Channel Vocoders)라고 한다.

다) 스펙트럼 패턴 재생기(Pattern-playback): 1950년대 부터 이용가능한 음성합성기로서 스펙트로그램의 생성과정을 완전히 역으로 이용하여 소리를 합성해 내는 장치이다. 다음 그림 4에서 보는 바와 같이 유사 스펙트로그램을 그린 후 이 모의 스펙트로그램에 빛을 조영하여 통과된 에너지를 다시 소리로 환원하는 장치이다.

그림 4. 스펙트럼 패턴 재생기



라) 포만트 합성기(Formant synthesizer): 성대 떨림을 모사하기 위해서는 저주파 소음 생성장치(buzz generator)를, 마찰음을 내기 위해서는 고주파 소음 생성기(hiss generator)를, 그리고 모음과 공명 자음을 위해서는 포만트 합성기(formant generator)를 이용하는 음성합성 장치이다.

바) 디지탈 컴퓨터 합성기(Synthesis by digital computer): 컴퓨터를 이용하여 디지탈로

소리를 합성하는 것으로 IBM과 호환성인 Speech-Station과 CSL(Computer Speech Lab) 등의 프로그램과 메킨토시용의 Sound Edit와 Signalizer 등의 프로그램에서 이용되고 있다.

사) 조음합성기: 인간의 발성 과정을 그대로 모사하여 소리를 합성하는 장치를 가리킨다. 그러나 실제로 발성을 위해 이용되는 조음기관들과 성대 주위의 발성 근육들의 미세한 움직임을 그대로 모사하는 발성 장치를 개발하기란 쉬운 일이 아니다.

아) LPC 합성기(Linear Predictive Vocoder): 보다 손쉬운 음성 조작을 위해 수학적 방식을 도입한 음성합성 Vocoder를 가리킨다. (Denes & Pinson, 1993, 제 10장 참조.)

이러한 음성합성기에서 생성된 음들이 보다 자연스러운 소리가 되기 위해서는 분절음의 합성과 함께 음장, 피치, 강세 등과 같은 운율적 특성들이 함께 적용되어야만 한다.

3.2 문장/음성 변환기의 처리과정

기계의 음성인식 및 합성에서 반드시 거쳐야 될 단계중에는 발화된 음성을 문장으로 인식하는 과정과 주어진 문장을 음성으로 합성하는 과정이 포함된다. 본 절에서는 글로 쓰여진 문장을 소리로 전환시키는 문자/음성 변환 과정에 대해 간략하게 언급하고자 한다. 문자/음성 변환과정은 일반적으로 언어처리 과정과 음성합성의 두가지 과정을 거치게 된다(구희산 1993 참조). 언어처리 과정은 먼저 다음과 같은 순서로 진행된다.

가) 전처리 과정: 숫자, 약어, 외래어 등과 같은 특수 문자를 일반 텍스트로 전환 시켜준다.

나) 구문 분석 과정: 입력된 문장을 통사적/형태론적으로 분석한다.

다) 운율정보 추출 과정: 문의 종류와 구절의 위치와 같은 통사 정보와 발화 단위에 내포된 의미 정보에 따라 구절과 억양구의 분절, 구절의 억양 패턴과 길이 조정, 휴지(pause) 등의 적절한 운율 정보를 추출한다.

라) 글자/음운 변환 과정: 구개음화, 경음화 자음동화 등과 같은 음운 규칙들을 적용하여 분석된 각 어절의 글자들을 소리나는대로 변환시켜 준다.

마) 단위 음성 생성: 음소(phoneme), 반음절(half syllalble), 걸친 반음절(demi-syllable), 또는 음절 등의 단위에 따라 합성 단계에서 접속하게 될 음성을 단위별로 미리 저장한다.

이러한 언어처리 과정이 끝나면 음성합성 처리과정에서 추출된 운율 정보와 저장된 단위음성 자료를 근거하여 실제 합성음을 만들게 된다. 앞서 소개한 음성합성 방식들 중에서 포만트 합성과 LSP(Line Spectrum Pair), 켑스트럼(Cepstrum), 파르코(Parco), PSOLA(Pitch Synchronous Overlap and Add) 등의 분석/합성(Vocoder) 방식이 현재 주로 이용되고 있다. 물론 자연성을 높이기 위해서는 반드시 운율적 정보가 음성합성에 포함되어야 할 것이다.

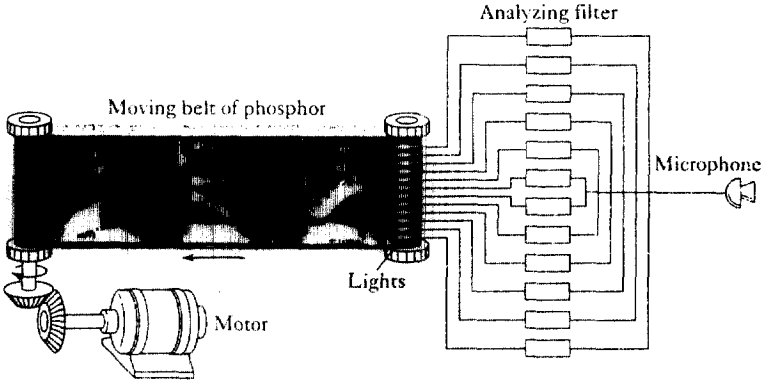
한편 단위음의 합성에 있어서는 각각의 스펙트럼이 가급적 연속적으로 부자연스럽지 않게 자동적으로 결합되도록 하여야 한다. 이때 적절한 분석용 발화 단위의 선택과 함께 정확한 경계 추출과 분할, 그리고 두 단위의 연결 방식 등이 잘 조정되어야만 자연스러운 합성음이 만들어지게 된다 (안성권과 성광모 1992 참조). 이렇게 결합된 음성에 기본주파수(F_0: Fundamental frequency), 강도(intensity)와 지속시간(Duration)과 같은 운율정보를 첨가하여 음의 높낮이와 세기 및 길이가 적절히 조절된 소리가 만들어지게 된다.

4. 스펙트로그램상의 음향음성적 정보

4.1 스펙트로그램과 스펙트로그램 판독 실험

스펙트로그램이란 화자의 음성 신호를 시각적으로 나타낸, 즉 다음 그림 5에서 보는 바와 같이 (가로축의) 시간과 (세로축의) 주파수, 그리고 (밝기의) 음성강도를 나타낸 그림을 일컫는 말이다. 그림 5a는 초기의 스펙트로그램 제작을 나타낸 그림이며, 그림 5b는 'speech anlaysis time'을 *Speech Station*이라는 컴퓨터 음향음성분석프로그램을 통해 만든 것이다.

그림 5. a. 초기의 스펙트로그램



b. 'speech analysis time' (광역스펙트로그램)

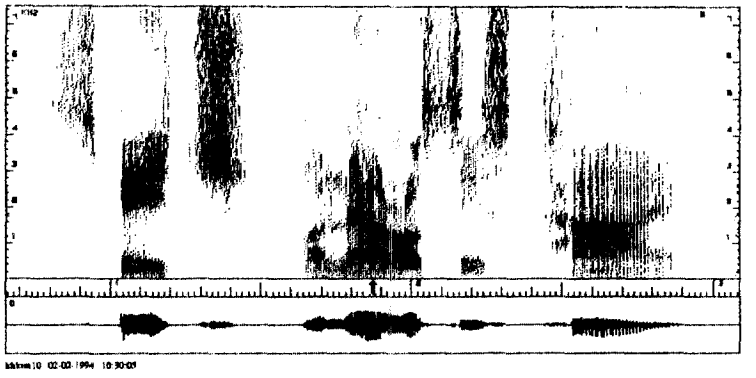


그림 5b에서 볼 수 있듯이 각 음소들은 여러 음향음성적 특성으로 구성되어 있다. 예를들면, 양순 폐쇄음 /p/에는 폐쇄음의 공백과 순음의 음향 특성(후행 모음의 제 2 및 제 3의 초기 공명음대(F_2 & F_3)의 하강 추이(falling transition) 모양)이 나타나 있으며, 파찰음 /tʃ/는 (2 KHz 이상의 소음 특성으로 나타나는) 마찰음 /ʃ/가 폐쇄음 /t/에 후행하는 것으로 나타나 있다. 그러므로 Jakobson, Fant & Halle(1952)의 변별자질이론에서는 이러한 스펙트로그램의 판독에 영향을 받아 자질들이 [consonantal], [vocalic], [grave], [flat], [nasal]에서 보는 바와 같이 음향음성적으로 정의되었다.

그러나 농아들을 위한 언어보조수단으로 실시된 1940년대 말의 Potter, Kopp & Green(1947)의 스펙트로그램 판독실험과 그후의 Svensson(1974)과 Klatt & Stevens(1974)의 판독실험 등은 부분적인 성공에도 불구하고 명백한 방법론이 결여되어 있다는 이유로 인해 긍정적인 평가를 받지 못하였다. 더우기 음운삽입/생략과 같은 음운규칙의 적용으로 인하여 음향음성적 특성들이 음성환경에 따라 스펙트로그램상에 각기 달리 나타나고 있기 때문에 Fant(1960), Liberman et al. (1968) Lindblom & Svensson (1973) 등은 스펙트로그램의 완전한 판독이 불가능한 것으로 간주하였다. 스펙트로그램 판독에 대한 이러한 부정적인 견해로 인해서 Chomsky & Halle(1968)의 자질이론에서는 [strident]라는 자질을 제외하고는 모두 [high], [low], [anterior], [coronal]에서 볼 수 있듯이 자질들이 조음음성학적으로 정의되었다. 한편 스펙트로그램 판독의 부정적 견해는 음성인식방식의 선호에도 영향을 미쳐 1970년대의 ARPA계획에서는 음향음성적 접근 방식인 상향식(Bottom-Up) 음성인식방식보다 통사론과 문법지식을 동원하는 하향식(Top-Down) 음성인식방식이 연속음성인식의 기본 방식으로 채택되었다.

ARPA 계획에서 채택된 음성인식의 과정은 다음과 같다. 먼저 제 1 단계에서는 주어진 입력부의 음성자료를 처리하여 음성기호로 기술하고, 제 2 단계에서 이들 음성기호를 기초로 하여 단어가 될만한 분절음의 연쇄를 찾아 가능한 단어의 연쇄를 만든다. 그러나 음성인식의 하위단계로 구분될 수 있는 제 1 및 제 2 단계에서는 'she prayed'와 'sheep raid' 그리고 'may clean'과 'make lean'이 각기 /ʃipreɪd/와 /meɪklin/의 분절음의 연쇄로 표기되므로써 통사적/의미적 중의성(ambiguity)을 피할 수 없게 된다. 따라서 제 3 단계에서는 음운작용으로 야기된 어휘적 중의성을 포함하여 이와같은 통사적/의미적 중의성의 문제를 해결하기 위하여 상위단계의 정보인 통사론과 의미론의 정보를 이용하여 연속음을 인식하게 된다.

그러나 이와같이 하향식 음성인식방식을 채택한 ARPA 계획은, Klatt(1977)이 지적한 바와 같이, Newell et al.(1971)에서 제시한 바 있는 '인간과 기계와의 의사소통이 가능한 연속음성 인식시스템 개발'의 목표에는 훨씬 못 미치는 실패한 것으로 간주되었다. 그리하여 Cole et al.(1980), Zue(1982), Leung & Zue (1986), Zue & Lamel(1986) 등에서는 의미/통사론적 지식을 지나치게 강조하는 하향식 음성인식방식에 회의를 품고 스펙트로그램에 내포된 음성정보만으로 스펙트로그램상의 문장을 판독할 수 있는지를 실험하였다. 이들은 실험을 통해 스펙트로그램을 직접 눈으로 읽는 방식이 ARPA계획의 음성인식을 포함하여 기존의 여타 인식방법보다도 더 효과적임을 보여주고 있는데, 이러한 사실은 스펙트로그램 판독과정에 동원된 음성 및 음운 지식의 정보들이 연속음성인식 시스템에도 그대로 이용될 수 있음을 보여주는 것이라고 할 수 있다. 이들의 실험에서 주목할 점은 통사/의미론의 정보가 거의 동원되지 않고 음성 및 음운정보만으로도 스펙트로그램이 해독될 수 있음을 보여주었다는 사실이다.

일례로 Cole et al.(1980)에서 보고된 Zue의 실험 과정을 간략히 고려해 보자. Zue는 이 실험에 앞서 약 2,000여 시간을 스펙트로그램 판독에 투자한 스펙트로그램판독 전문가라고 할 수 있다. 그가 스펙트로그램 판독 실험과정에서 보여준 제 1 단계는 분절음화(segmentation) 단계로, 스펙트로그램상의 불연속성 (spectral discontinuities), 음장(duration), 공명음대 이동(formant movement) 등을 이용하여 연속적인 음파를 단음(phone)의 단위로 분절하는 과정이다. 'The soldiers knew the battle was won',

'Smoking is bad for your mind and body' 등 정상적인 문장과 'Wake jungles gasoline sudden bright' 등의 비문이 포함된 23개의 영어 문장의 분절음 경계인식에 있어서, Zue는 녹음에 동원된 2명의 남성 화자의 음성특성의 차이와 그리고 스펙트로그램 기록 자체의 결함에도 불구하고 499개의 분절음중 485개의 분절음을 인식함으로써 약 97% 의 높은 정확도를 보여주었다. 그리고 "Say _____ again"이라는 전달구속에 삽입된 'baby', 'elephant' 등 45개의 영어 단어에 대한 분절음 경계인식에 있어서는 3 명의 음성학자가 파악한 201개의 분절음 모두를 정확히 인식하였다.

Zue가 스펙트로그램 판독실험과정에서 보여준 제 2 단계는 이렇게 인식된 분절음에 스펙트로그램상에 나타난 음향음성적 특징과 영어의 음운규칙들에 관한 정보를 이용하여 합당한 분절음을 찾아내는 분절음 분류(labelling)과정이다. 이 과정에서 Zue가 보여준 분절음 분류의 결과는 동원된 3 명의 음성학자들의 음성전사와 비교하여 볼 때 23개의 영어 문장의 경우 약 85% 가량 일치하며 (첫번째 선택에서 약 67% 그리고 두번째 및 세번째 선택에서 각기 약 13% 과 5%의 일치를 보여줌), 전달구 속의 분절음 분류의 경우 약 93%의 일치를 보여주고 있다. 그리고 모음보다는 자음에서 보다 많은 일치를 보여주고 있다. 그런데 여기서 우리가 고려할 점은 연속된 발화의 음성전사란 숙련된 음성학자들 사이에서도 차이가 있다는 사실이다. 실제로 이 실험에 동원된 3명의 음성학자들의 음성전사를 비교하여 보면 이들 사이에도 영어 문장의 경우에 약 85%의 일치만 보여주고 있다는 사실이다. 이러한 사실은 Zue의 분절음 분류가 어느 음성학자의 음성전사와도 견줄 수 있는 거의 완벽한 것이며, 따라서 음성학자가 마치 소리를 듣고 주어진 발화의 음성을 전사할 수 있듯이 전문적인 스펙트로그램 분석가 역시 스펙트로그램만으로 동일한 발화의 분절음을 분류해 낼 수 있음을 보여준다고 할 수 있다. 특히 기존의 하향식 음성인식과 관련하여 우리가 주목해야 할 점은 이 실험에서 Zue는 영어의 통사/의미론적인 상위단위의 정보를 거의 이용하지 않고 단지 영어의 음성적 지식과 분절음 제약을 포함한 음운론적 지식만을 사용하였다는 사실이다.

Zue가 스펙트로그램 판독에 사용한 영어음들의 음향음성적 특성들은 다음 도표 1과 같다(Cole et al. 1980:37).

도표 1. V.Zue가 음성인식을 위해 사용한 영어 음들의 음향음성적 특성

모 음	고/저설 전/후설 /a/ 축약모음	높이와 반비례로 변화됨 F1-F2 사이의 거리와 함께 변화됨 음장: /ɪ/가 /i/가 짧다 F1이 모든 모음중 가장 높다 길이가 짧다. 중립모음의 공명음대 형성
마찰음	유/무성 /s/ /ʃ/	유성 마찰음이 무성 마찰음보다 짧다 소음 > 4 kHz 소음 < 4 kHz
비 음	/m,n,ŋ/	300 Hz 이하의 에너지, 급격한 강도 시작점(amplitude onset) 가짐 모음보다 강도는 약함

		인접 모음의 비음화를 야기시킴
조음위치	순 음 연구개음	모든 공명음대가 아래로 내려옴 F2와 F3가 닫히는 시점에서 겹쳐짐
반전음	/ɚr/ /r/	F3가 2 kHz 아래로 내려감 F3가 F2를 따라감 F3가 F2에 맞다음
설탄음	[ɾ]	길이가 매우 짧다. < 20, 25ms.
폐쇄음	폐쇄기간 파열(burst) 유/무성 공명음대 변이	에너지가 없다. 순 음 : 거의 없다. 치경음 : 고주파수대에 있다. 연구개음: 강한 파열, 또는 이중 파열을 보임 VOT: 무성음이 유서음보다 성대진동시각 더 김 순 음 : 아래로 향함 치경음 : F2 목표위치(locus)가 1800Hz에 있음 연구개음: F2와 F3가 함께 겹쳐짐

4.2 변이음 지표와 음운파싱

연속음성인식에 있어서 파싱과 단어 매칭에 이용되는 음성지표에는 크게 두가지 즉 비교적 문맥에 의존하지 않는 불변지표(invariant cues)와 문맥에 따라 변화를 보이는 변이지표(allophonic or variant cues)가 있다. 그런데 Stevens(1981)의 불변지표이론에서는 영어의 경우 '조음위치', '유성성', '조음자질'과 같은 불변지표들만 단어 검색에 도움이 될 뿐, '기식음화'나 '불파'와 같은 변이지표는 어휘검색에 불필요한 것으로 간주되었다. 그러나 여기서 주목할 점은 동일한 음성적 자질이 언어에 따라 불변(invariant) 자질 또는 변이자질로 간주되고 있다는 사실이다. 예를들어 영어의 순음 파열음 /p/는 'pie'에서와 같이 음절초의 위치에서는 강하게 기식음화되어 발음되지만, 'spy'에서와 같이 /s/음 뒤에서는 기식음화 되지 않으며, 'top'에서와 같이 음절말 위치에서는 불파되어 발음된다. 그러나 영어에서 단순히 기식여부에 따라 구별되는 단어의 쌍은 찾아 볼 수 없다. 다시 말해서 영어의 경우 기식(aspiration)여부는 잉여적, 즉 변별적이지 못한 변이음적(allophonic) 요소인 것이다. 그러나 기식자질은 영어에서는 잉여적이지만 한국어에는 '뿔', '풀', '불'에서 보는 바와 같이 변별적이다. 그런데 여기서 우리의 관심을 끄는 것은 언어에 따라 변이 또는 불변 자질로 간주되는 기식의 음성지표가 그 자체의 독특한 음향음성적 특성에 의해 스펙트로그램상에 손쉽게 구별될 수 있다는 사실이다: 기식성은 파열이후의 긴 소음 기간(long positive voice-onset-time)으로 나타난다. 이러한 사실에도 불구하고 (ARPA계획 기간중인) 1970년대 초에는 이러한 변이음적 자질들이 단어 탐색에 있어서 아무런 도움이 못되는 '소음'에 불과한 것으로 간주되어 보기 (1)과 (2)와 같은 동일 분절음 연쇄의 쌍들은 음운부가 아닌 통사부/의미부에서만 해결되는 것으로 간주되었다.

(1) gray plane vs. grapelane /greɪpleɪn/ [greɪpʰleɪn] vs. [greɪp leɪn]
She payed vs. sheep aid /ʃipeɪd/ [ʃipʰeɪd] vs. [ʃip˭eɪd]
my train vs. might rain /maɪtreɪn/ [maɪtʰrein] vs. [meɪt˭reɪn]
may clean vs. make lean /meɪklin/ [meɪkʰlin] vs. [meɪk˭lin]
I scream vs. ice cream /aɪskrim/ [aɪskrim] vs. [aɪskʰrim]

(2) /uripaksatapaɓekanta/
가) 우리 박사 (모두) 다 방에 간다. 나) 우리 박사(들이) 다방에 간다.
다) 우리(가) 박사다방에 간다. 라) (누군가) 우리박사(라는) 다방에 간다.

그러나 주목해야 할 점은 비록 이들이 동일한 음소열을 갖고 있다 하더라도 변이지표로 간주된 문맥의존적 변이음 특성들이 음향음성적으로는 분명히 구별되어 나타난다는 점이다. 다음 그림 6은 'gray plane [greɪpʰleɪn]'과 'grapelane [greɪp˭leɪn]'을 광역 스펙트로그램으로 나타낸 것이다(Church 1987:49 인용)

그림 (6) "I said 'grape lane' not 'gray plane'" (광역 스펙트로그램)

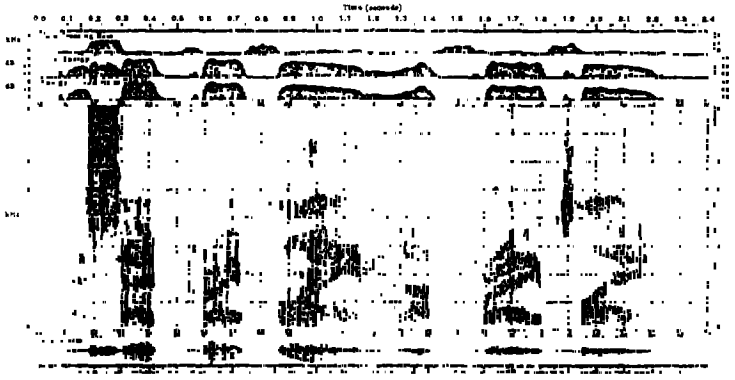


그림 6에서 보는 바와 같이 영어의 경우 /p/, /t/, /k/의 무성 파열음은 음절초에서는 강하게 기식되어 발음되지만 음절말에서는 불파되어 발음된다. 그러므로 Nakatani & Dukes(1978), Nakatani & Schaffer(1978), Church(1987), 김기호(1990) 등에서는 이러한 변이음적 특성들이 더이상 소음이 아니라 오히려 적절하게 단어 경계를 설정하는데 효과적으로 이용될 수 있음을 보여주고 있다. 일례로 통사부나 의미부의 도움없이 어떻게 문장 파싱이 이러한 음성/음운지식에 의해 이루어질 수 있는지를 다음 보기를 통해 살펴보자(Church 1987:180-82).

(3) a. Did you hit it to Tom?
b. [dɪdʒəhɪ ɪʔtʰətʰɑm]
c. [# dɪdʒə # hɪɾ # ɪʔ # tʰə # tʰɑm #]

(4) a. 마찰음 /h/는 "언제나" 음절초의 위치에서
b. 단타음 [ɾ]는 "언제나" 음절말에서
c. 불파음 [ʔ]는 "언제나" 음절말에서
d. 기식음화된 [tʰ]는 "언제나" 음절초에서

(5) 파싱된 음성전사 단어 검색

[dɪdʒə] ==========> dit you
[hɪt] ==========> hit
[ɪʔ] ==========> it
[tʰə] ==========> to
[tʰɑm] ==========> Tom

(3b)의 정밀음성전사를 갖는 (3a)의 영어 문장은 (4)와 같은 영어의 음절 변이음 제약과 음소배열제약(phonotactic constraint)으로 인해 (3c)와 같이 파싱될 수 있다. 그리고 (5)에서 보는 바와 같이 발화된 문장이 일단 음절정도 크기의 단위로 파싱되게 되면 그 후의 단어 검색은 비교적 손쉽게 이루어질 수 있게 된다.

그러므로 1970년대 초기에 소음으로 간주되었던 변이음적 요소는 더이상 소음이 아니라 오히려 음소배열제약과 공명도 위계(sonority hierarchy), 음절화 규칙등의 음운제약과 함께 문장 파싱에 매우 효과적으로 이용될 수 있는 것이다. 한편 변이음적 자질외에도 음장과 피치와 같은 운율적 요소 역시 연속음성인식의 문장 파싱에 효과적으로 이용될 수 있다.

4.3 운율정보와 음운파싱

음성인식에서의 음운부의 역할은 앞서 2.3에서 지적한 바와 같이 일반적으로 주어진 음성파형으로부터 음성 특징들을 추출하고, 그후 음운규칙과 음소배열제약 등을 이용하여 올바른 음소의 연쇄를 도출하는 것으로 한정되어 있다. 그러므로 음운부의 역할은 주어진 발화의 음성파형으로부터 예를들면 보기 (2)와 같은 음소 연쇄 즉 /uripaksatapaɓekanta/를 도출하는 것에 국한되며, 그 후의 적절한 단어와 문장으로의 인식은 형태부와 통사부 및 의미부에서 다루어진다. 그러나 /uripaksatapaɓekanta/의 음소연쇄로부터 바람직한 문장을 도출하기는 쉽지가 않다. 왜냐하면, 주어진 연쇄는 앞의 (2가-라)에서 보는 바와 같이 여러 문장으로 파싱될 수 있기 때문이다. 그러나 실제로 우리가 이러한 음의 연쇄를 일상 대화에서 접하게 될 때, 우리는 별 어려움 없이 그 의미를 파악하게 되는데, 이는 우리가 말하는 음성 파형에는 음소적 특성 뿐만 아니라 형태적 정보와 통사 및 의미적 정보가 충분히 내포되어 있기 때문이다. 일례로 한 단어로 구성된 다음 (6)과 (7)의 문장을 고려해 보자.

(6) a. All right. (That's OK!)
b. All right? (Is that right?)
c. All right. (Are you sure?)

(7) 가) 그래? (정말 그 말이 맞아?)
나) 그래. (그 말이 맞아.)
다) 그래!? (비꼬면서)

(6)과 (7)에서 보는 바와 같이 'All right'과 '그래'라는 단어의 음소 연쇄는 몇가지 뜻을 내포할 수 있다. 그러나 실제로 대화를 하는 청자는 아무런 문제없이 그 의미를 파악할 수 있다. 왜냐하면 'All right'과 '그래'라는 단어의 문장은 그 의미에 따라 다르게 즉, 상승, 하강, 또는 상승하강의 억양으로 발음되며, 바로 이러한 운율적 차이에 의해 달리 해석될 수 있기 때문이다. 다음 그림 7과 8은 피치의 변화를 나타내기 위해 협역

(narrow-band)의 스펙트로그램으로 /ɔlraɪt/과 /kʲræ/를 나타낸 것이다.

그림 7. 스펙트로그램 /ɔlraɪt/ (협역 스펙트로그램)

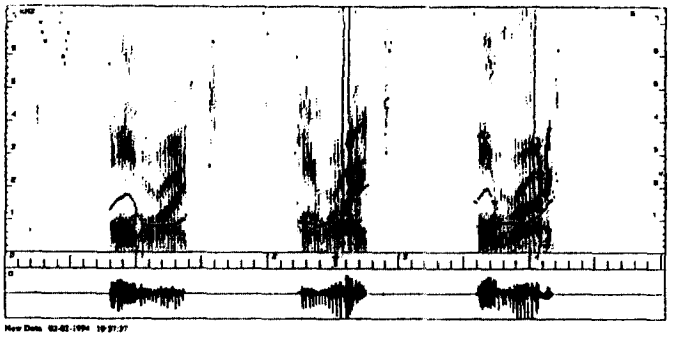


그림 8. 스펙트로그램 /kʲræ/ (협역 스펙트로그램)

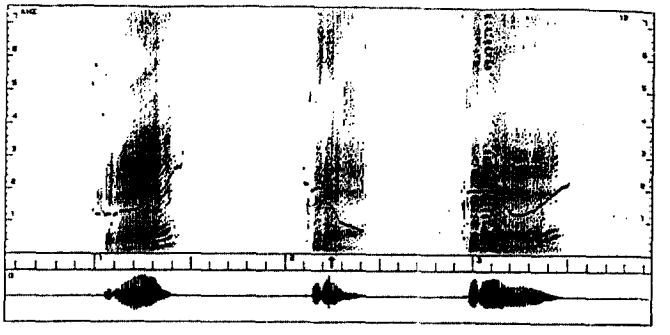


그림 (7)과 (8)에서 보는 바와 같이 /ɔlraɪt/과 /kʲræ/는 내포된 각각의 의미에 따라 각기 다른 억양의 피치 변화를 나타내 주고 있다. 그러므로 운율적 자질인 피치 변화를 이용할 때, 주어진 문장이 평서문인지, 의문문인지, 그리고 비꼬는 투의 문장인지 등을 통사부나 의미(상황)부의 도움없이 파악할 수 있게 된다. 더우기 스펙트로그램 내의 정보는 이와같이 문장 끝 뿐만 아니라 주요단락의 구분에 도움이 되는 많은 정보들을 가지고 있다. '우리박사다방에간다 /uripaksatapaɲekanta/'를광역 스펙트로그램으로 나타낸 그림 (9)와 (10)을 비교해 보자.

그림 9. ##[우리#박사]PP##[다]PP##[방에#간다]PP## (PP=Phonological Phrase)

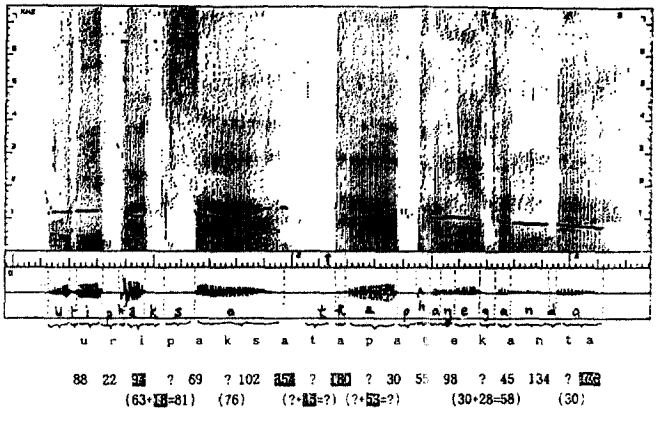


그림 10. ##[우리]PP##[박사다방에#간다]PP##

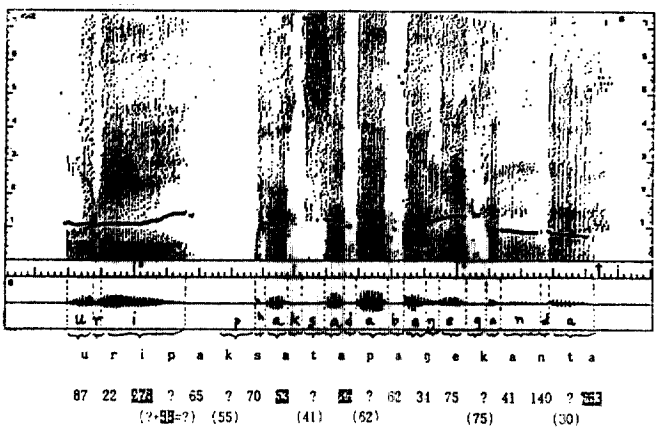


그림 9와 그림 10을 비교해 보면, 매 음운 어절(phonological phrase)이 끝날 때마다 피치가 올라감을 알 수 있으며, 음장(duration)에 있어서도 문장 끝의 모음은 비교적 길며, 또한 어절말의 모음의 길이가 어절 또는 단어 내에서의 모음의 길이에 비해 현저히 긴 것을 알 수 있다 (/i/ 94ms vs. 278ms ; /a/ 53ms vs. 354ms, 84ms vs. 180ms). 따라서 피치와 음장의 운율적 정보를 이용하면, 주어진 /uripaksatapaɓekanta/ '우리박사다방에간다'의 음소의 연쇄는 각기 다음 (8)과 같이 파싱될 수 있다.

(8) 가. #uripaksa# 우리박사 나. #uri# 우리
#ta# 다 #paksatapaɓekanta# 박사다방에간다
#paɓekanta# 방에간다

만일 음운부의 역할이 단순히 주어진 음파로부터 음소의 연쇄만을 도출하는 것이라고 가정한다면, /uripaksatapaɓekanta/의 음소 연쇄로부터 '우', '울', '우리', '우리박', '우리박사다', '우리박사다방', 등등 모든 가능한 단어의 연쇄를 다 검색해야 한다. 이와는 대조적으로 음장과 피치와 같은 운율적 음운 정보를 이용할 경우, (8a)의 경우에는 [uripaksa]PP, [ta]PP, [paɓekanta]PP의 음운 어절에서만 가능한 단어 연쇄를 찾으면 되므로 단어 검색의 시간을 효과적으로 단축시킬 수 있게 된다. 다시말해서 /uripaksa/의 음소 연쇄에서 '우리'나 '울이', '우리박', '우리박사'의 가능한 단어를 거쳐 '우리박사'를, 그리고 /ta/의 음소 연쇄에서는 '다'를, 마지막으로 /paɓekanta/의 음소 연쇄에서는 '바', '방', '방에', '방에 간다'를 거쳐 손쉽게 '우리 박사 다 방에 간다'를 도출할 수 있다.

더우기 4.2에서 지적한 바와 같이 변이음적 정보도 연속음성인식에 도움을 줄 수 있는데, 세 종류의 한국어 폐쇄음 중에서 단지 평음만이 유성음 사이에서 유성음화를 겪는 사실에 근거하여 문장 파싱에 유성의 변이음적 특성을 이용할 수 있다. 그림 9와 그림 10에서 볼 수 있는 바와같이 국어의 평음 'ㄱ', 'ㄷ', 'ㅂ'은 어절앞과 어두에서는 어중에서보다 강하게 기식되어 발음되고 있다. 즉, '우리박사'의 'ㅂ'은 단지 18ms의 기식음 또는 VOT(voice onset time: 성대 진동 개시시간)를 보여주는 반면, '박사다방'의 어절 앞 'ㅂ'은 이보다 긴 35ms의 VOT를 보이고 있다. 그리고 '다'의 어절 앞 'ㄷ'과 '방에'의

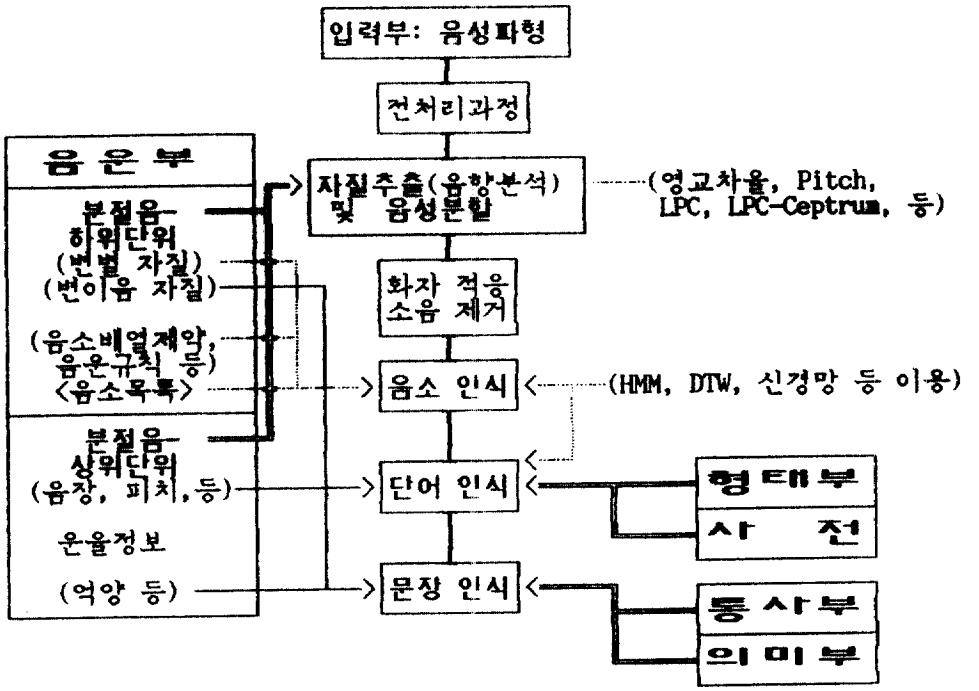
어절 앞 'ㅂ'은 각기 45ms과 53ms의 VOT를 보여주고 있지만, '박사다방'의 경우 단어 내의 'ㄷ'과 'ㅂ'은 유성음화되어 기식음 없이 발음되고 있다. 따라서 앞 절에서 제시한 음장과 피치 등 분절음 상위 단위의 운율 정보 뿐만 아니라 유성음화와 같은 변이음의 분절음 하위 단위의 정보 역시 문장 파싱에 도움이 되고 있다. 그러므로 /uripaksatapaδ ekanta/의 음소 연쇄로부터 각기 상이한 VOT를 이용하여 그림 4의 문장은 '우리박사 다방에 간다'로, 그리고 그림 6의 문장은 '우리 박사다방에 간다'로 파싱하여 두 문장을 쉽게 구별해 낼 수 있게 된다. (이와 유사한 주장의 통계적 보고는 Silva(1991)를 참조하기 바람.)

5.0 결론

5.1 연속음성인식의 모델

스펙트로그램상에는 앞절에서 살펴본 바와 같이 많은 음성/음운정보가 포함되어 있다. 그러므로 보다 효율적인 음성인식을 위해서는 이러한 다양한 음성/음운 정보를 효과적으로 이용할 필요가 있다. 따라서 연속음성인식의 경우 음운부의 역할은 주어진 음파로부터 일련의 음소연쇄만을 도출해 내는 기존의 역할에서 보다 확대되어야만 한다. 특히 분절음 하위단위인 변이음 정보와 분절음 상위단위인 음장, 피치, 억양 등의 운율정보는 단어 검색과 문장 파싱에 매우 효과적으로 이용될 수 있으며 이에 따라 기존의 형태부, 통사부, 및 의미부의 부담을 효과적으로 줄일 수 있게 된다. 따라서 연속음성인식의 모델은 다음 그림 11과 같이 수정되어야 한다.

그림 11. 연속음성인식을 위한 음성인식 모델



먼저 입력된 아날로그 음성신호를 디지탈화한 후, 저역여과시켜 표본화하는 전처리과정을 거친다. 그후, 영교차율, 피치, LPC, 또는 LPC-Cepstrum 등을 이용하여 음성자질과 운율자질 등 음성요소의 특징들을 추출한다. 초분절음적인 운율자질로는 음장, 피치변화, 억양(영어의 경우 강세 첨가) 등을 추출하며, 음성자질로는 모음/비음/마찰음/폐쇄음을 분류해 주는 조음방식자질, 그리고 치음/순음/연구개음을 구별해 주는 조음위치자질 등의 불변(invariant) 자질들 뿐만 아니라 문맥에 의존하는 변이음적(allophonic) 자질들도 모두 추출한다. 왜냐하면 이러한 변이음적 자질들이 운율정보와 함께 초분절음 구문 파싱(suprasegmental phonological parsing)에 이용될 수 있기 때문이다. 이러한 자질들의 조합에 한국어 음소배열제약과 음운규칙 등의 음운정보를 이용하여 음소 목록으로부터 가능한 음소의 연쇄를 도출해 내게 된다. 음소인식을 위해서는 HMM 방식이나 DTW 방식, 또는 신경망 회로 등의 방식을 이용할 수 있으며, 이 방법들은 음소인식 뿐만 아니라 단어 인식에도 이용된다.

분절음 하위 단위의 음성정보중 변이음 음운정보는 음장과 피치 등 분절음 상위단위의 운율 정보와 함께 병렬적으로 음운정보를 처리하면서 주어진 음소 연쇄로부터 음운 어절을 도출해 낸다. 그후 음운 파싱된 음운 어절로부터 형태부의 도움을 받아 가능한 단어열을 사전으로부터 차출 연결시켜 준다. 이러한 가능한 단어 연쇄는 다시 억양과 음장 등의 운율정보와 함께 통사부와 의미부의 도움으로 하나의 완전한 문장으로 파싱되어 인식된다.

요약하면 연속음성인식에서의 음운부의 역할은 주어진 음파로부터 가능한 음소연쇄의 추출이라는 제한된 역할을 벗어나 그림 11에서 보는 바와 같이 전 영역으로, 즉 자질 추출에서부터 음소인식은 물론 단어 인식과 문장인식에까지 확장되어야 한다. 이러한 연속음성인식에 있어서의 음운부의 역할은 역으로 음성합성에 그대로 적용될 수 있다. 자연스러운 음의 합성을 위해서는 음소의 불변지표는 물론 변이지표의 음향음성적 특성과 함께 음장과 피치 등 운율적 특성을 구현해 주어야 할 것이다.

5.2 그밖의 고려사항

5.2.1 청각음성학과 음성인식

소리의 학문인 음성학에는 일반적으로 다음 세 분야를 가르킨다.

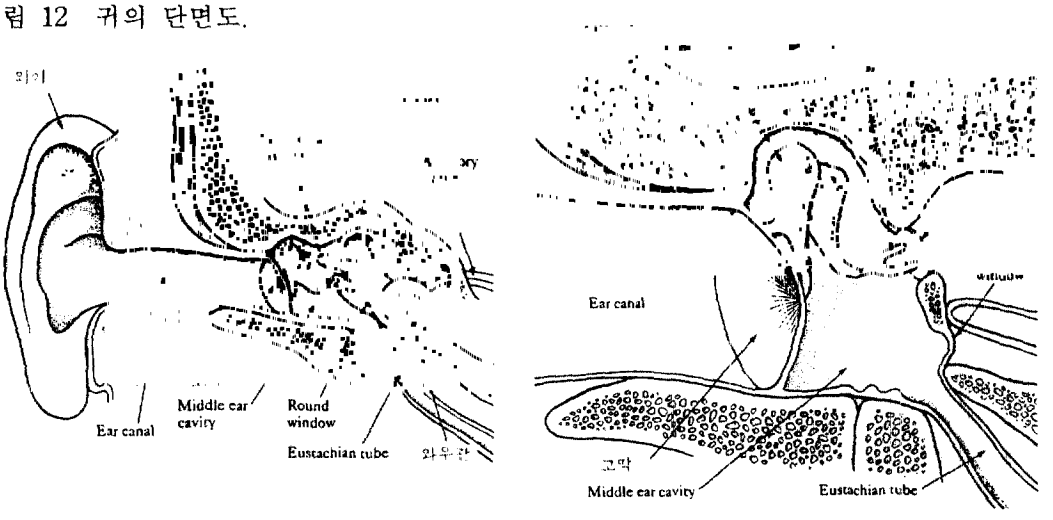
가) 조음 음성학: 말소리의 발성에 대한 연구.

나) 음향 음성학: 말소리의 물리적 특성 연구.

다) 청각 음성학: 말소리의 인식에 대한 연구.

지금까지 논의된 음성인식과 음성합성과 관련된 음성학의 분야는 주로 조음음성학과 음향음성학의 두 분야였다. 이제 본 절에서는 간략하게나마 청각음성학과 음향음성인지(psychoacoustic)실험의 중요성을 지적하고자 한다. 다음 그림 12가는 귀의 단면도를 나타낸 그림이고 그림 12나는 특히 중이 부위를 확대하여 나타낸 그림이다.

그림 12 귀의 단면도.



여기서 주목할 점은 외이의 통로가 하나의 공명체 역할을 하여 소리의 에너지를 수합 증폭 시킨다는 점이다. 실제로 외이의 통로는 약 3,000Hz의 고유 진동수를 가지고 있기 때문에 이 공명 주파수와 유사한 3,000Hz 내외의 소리는 본래의 소리보다 약 2-4배 증폭되어 고막에 이르게 된다. 한편 고막은 소위 가청 주파수인 20Hz에서 20,000Hz까지의 주파수만 감지하며 그 이외의 주파수는 감지하지 못한다.

그리고 중이에 있는 망치뼈(malleus;추골)는 고막이 진동함에 따라 진동하는데 이는 다시 연결된 모루뼈(incus;침골)에 전달되며 이것은 다시 연결된 등자뼈(stapes;등골)로 전달된다. 그런데 여기서 주목할 점은 고막과 타원창(oval window)을 연결시켜주고 있는 중이의 세 뼈마디가 마치 지렛대와 같은 역할을 하여 소리를 몇배 더 증폭시킨다는 사실이다. 뿐만 아니라 고막과 타원창의 크기의 차이(25:1) 로 실제 타원창에 느끼는 공기의 압력은 원래 소리의 압력보다 훨씬 증폭되게 된다. 그리고 흥미로운 사실은 중이의 역할 중 하나는 음향음성학적 수정기능(rectification function)인데, 하나의 단순음이 주어질 때, 이를 홀수 또는 짝수 배 배음하거나 (harmonics)하거나 차음(difference tone) 또는 합음(summation tone)한다는 사실이다.

끝으로 내이의 와우관(cochlear duct)에서는 내이의 타원창에 전달된 음향에너지를 뇌가 분석할 수 있는 전기화학적(electro-chemical) 에너지인 신경 맥박으로 바꾸어주는 일을 하면서 복합음을 분석하는 일을 한다. 그러므로 소리는 귀의 외이, 중이, 내이를 거치면서 실제의 음향적 특성과는 상당히 변화되어 뇌에 전달되고 있음을 알 수 있다. 이에 따라 최근 JASA(Journal of the Acoustical Society of America)의 몇몇 논문에서는 특정 주파수에 대해서 난청인 사람을 피실험자로 하여 음성 인식에 있어서 특정 주파수의 역할에 대하여 보고하고 있다.

그러므로 보다 효율적인 음성인식을 위해서는 단순한 소리의 물리적인 음향음성적 특성 조사에 끝날 것이 아니라 이러한 음향음성적 특성들이 소리를 수집/증폭/분석/전달하는 귀의 외이, 중이, 내이를 통해 어떻게 뇌로 전달되는지에 대한 좀더 깊이있는 연구가 진행되어야 할 것이다.

5.2.2 소리의 범주적 인식과 음성인식

음성인식과 음성합성과 관련하여 좀더 많은 연구가 필요한 분야중 하나가 바로 음향심리실험 분야이다. 먼저 VOT에 의한 자음의 범주적 인식 문제를 고려해 보자. 예를 들어 영어 폐쇄음 [p]와 [b]의 경우, VOT를 조사해 보면 실제로 많은 차이를 보여주고 있는데, 유성음 [b]는 때때로 10ms,또는 20ms 등의 VOT를, 그리고 무성음 [p]는 40ms, 50ms, 또는 60ms등의 VOT를 보여주고 있다. 그런데 흥미로운 사실은 미국인 청자들은 이들을 유무성의 구별이 분명한 범주적인 것으로, 즉 VOT가 25ms 미만인 경우에는 유성자음으로, 그러나 25ms 이상인 경우에는 무성자음으로 인식한다는 사실이다. 다시 말해서 만일 20ms의 차이를 둔 10ms의 VOT와 30ms의 VOT를 가진 두 음이 주어질 경우에는 이를 각기 [b]와 [p]로 인식하게 되지만, 20ms의 동일한 차이를 둔 40ms와 60ms의 VOT를 가진 두음이 주어질 경우에는 아무런 차이를 인식하지 못한다는 사실이다. 그런데 주목할 점은 VOT의 범주적 인식이 언어에 따라 차이가 있다는 사실이다. 즉 한국어와 태국어의 경우에는 3 종류의 폐쇄음이 있기 때문에 영어나 스웨덴어와는 달리 3 가지의 다른 범주로 인식된다는 사실이다(Lisker & Abaramson 1964,1971 참조). 언어음의 범주적 인식에 대한 이러한 심리음향인지실험은 모음의 경우에도 동일하게 적용된다고 볼 수 있다(Liberman 1970, Clark & Clark 1977 참조). 일례로 다음 그림 13과 그림 14를 고려해 보자.

그림 13

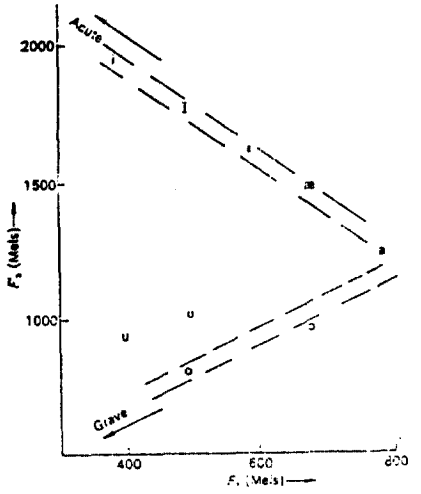


그림 14

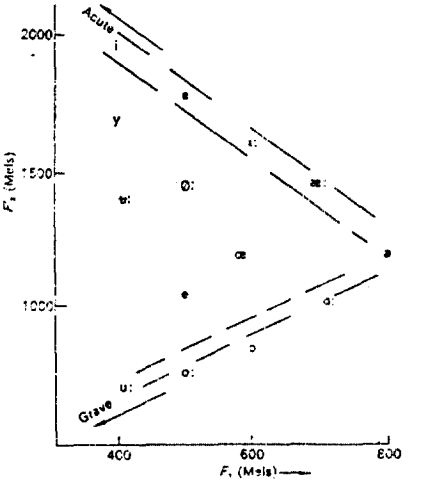


그림 13과 14는 각기 영어와 스웨덴어의 모음을 음향음성적으로, 즉 가로축에 제 1 공명음대의 주파수를 세로축에 제2 공명음대의 주파수를 Mel value로 나타낸 것이다. 그런데 흥미로운 사실은 그림 13과 14에서 보는 바와 같이 이들의 모양이 조음음성학으로 분류한 모음 삼각도의 모양과 매우 흡사할 뿐만 아니라, 이들 두 언어의 모음이 모두 소위 양자역학적 모음(quantal vowel)인 [i], [a], [u] 음들을 가지고 있다는 점이다. 모음 [i],[a],[u]는 여러 언어에서 가장 흔히 나타나는 모음으로 (Trubezkoy 1939, Greenberg 1963, 1966 참조), 의사소통에 가장 효과적인 음이라고 할 수 있으며 유표성(markedness)

이론에선 가장 무표적인, 즉 가장 자연스러운 음으로 간주되고 있다. 그런데 음성인식과 관련하여 우리의 관심을 끄는 것은 영어와 스웨덴어의 모음들이 소위 양자역학적 모음 [i],[a],[u] 음들을 정점으로 비슷한 거리를 두고 배치되어 있다는 점이다. 특히 /i/-/a/축의 모음들을 비교해 보면 비록 다른 모음 기호가 사용되었지만, 영어 모음 /i/,/ɪ/,/ɛ/,/æ/,/a/의 음향음성적 인식의 차이가 (비록 음장의 차이 [ɛ:],[æ:]는 있지만) 스웨덴어 모음 /i/,/e/,/ɛ:/,/æ:/,/a/의 음향음성적 차이와 같이 인식되고 있다는 사실이다. 그러므로 자음 뿐만 아니라 모음의 고저와 전후설성 자질들도 범주적으로 인식되고 있으며, 언어에 따라 다른 상대적 가치를 가지게 된다.

이러한 소리의 범주적 인식과 관련하여 김기호(1991)에서는 5 단계의 구분이 있는 영어와 스웨덴어와는 달리 3 단계의 구분만 있는 한국어의 경우 이들 영어와 스웨덴어 모음을 어떻게 인식할 것인가 하는 문제를 위시하여 김기호(1991)에서 음성인식을 위해 제시한 자질 수형도(feature geometry)의 자질위계구조를 음성인식의 음향심리 실험을 통해 검증해 볼 것을 제안한 바 있다. 예를들어 Miller & Nicely(1955)는 「자음+모음」의 연쇄에 인위적 소음(white noise)을 첨가하여 분절음이 어떻게 인식되어지는지를 실험하였는데, 이들의 실험 결과는 하나의 분절음이 유무성과 비음등 몇가지 통로를 통해 여러 조각으로 인식되며, 이러한 부분적 조각들이 재구성되어 하나의 분절음으로 인식되고 있음을 보여주고 있다. 따라서 이상적인 자질 수형도의 계층구조는 음운 현상을 통해서도 입증되어야 하겠지만 심리음향인지실험을 통해서도 입증되어야 할 것이다.

참 고 문 헌

구희산. 1993. 음성합성의 운율처리를 위한 악센트 연구. 「음성/음운/형태론 연구」 음운론 연구회. 한국문화사. 21-34.

김기호. 1991a. "Revisiting distinctive feature approach in speech recognition,"

김기호. 1991b. 영어 자질이론의 발전과 음성인식과의 관계. 「영어영문학」 37.783-803.

김기호. 1993. 연속음성인식에 있어서의 음운론의 역할을 재고함. 「음성/음운/형태론 연구」 음운론 연구회. 한국문화사. 49-63.

김기호.이용재. (1991) "The role of phonology in speech understanding," *Harvard Studies in Korean Linguistics* 4: 143-156.

김순협. 1991. 국내외 음성인식 기술 동향 및 전망. *Korea-Japan Joint Symposium on Acoustics*, (1991) 183-198.

김종미. 1990. 언어학을 활용한 국어음성인식. 「음성인식 및 신호처리 워크샵」 170-177.

안승권, 성광모. 1992. 포만트 궤적 중첩 방법에 의한 한국어 문자-음성 변환. 「음성통신 및 신호처리 워크샵 논문집」 9: 197-200.

이양희. 1992. 음성합성 기술 현황 및 전망. 「음성 통신 및 신호처리 워크샵 논문집」 9: 88-96.

Blumstein, S.E.and K.N. Stevens. (1979) "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants", *JASA* 66.

Church, K. (1987) *Phonological Parsing in Speech Recognition.* Kluwer AP.

Denes, P. & E. Pinson. (1993) (2nd ed.) *The Speech Chain: The Physics & Biology of Spoken Language.* Freeman.

Holmes, J. (1988) *Speech Synthesis and Recognition.* van Nostrand Reinhold.

Huffman, M. & R. Krakow. (eds.) (1993) *Nasals, Nasalization and the Velum.* Phonetics & Phonology 5. Academic Press.

Iverson, G. & K.-H. Kim. (1987) "Underspecification and hierarchical feature representation in Korean consonantal phonology," *CLS* 23: 182-98.

Kim, K.-H. (1987) *The Phonological Representation of Distinctive Features: Korean Consonantal Phonology.* Ph.D. dissertation, U. of Iowa.

Kim, K.-H. (1990) "Revisiting distinctive feature approach in speech recognition," paper presented at Seoul Internaitonal Conference on Natural Language Processing.

Klatt, D.H. (1977) "Review of the ARPA speech understanding project". *Journal of the Acoustical Society of America*,61.

Klatt, D.H., & K.N. Stevens. (1973) "On the automatic recognition of continuous speech :Implications from a spectrogram-reading experiment", IEEE Transactions on Audio and Electroacoustics, AU-21, 210-217.

Koo, H.-S. 1986. *An Experimental Acoustic Study of the Phonetics of Intonation in Standard Korean.* PhD. Univ. of Texas.

Lea, W. (ed.) (1980) *Trends in Speech Recognition.* Prentice-Hall.

Liberman, A.M, F.S.Cooper, D.P.Shankweiler,& M.Studdert-Kennidy (1968) "Why are speech spectrograms hard to read?" *American Annals of the Deaf*,113.

Lieberman, P.and S.E.Blumstein. (1988) *Speech physiology, speech perception, and acoustic phonetics*, Cambridage University Press.

Lindblom, B.E.F. and S.G.Svensson. (1973) "Interaction between Segmental and Nonsegmental Factors in Speech Recognition," IEEE Transactions on Audio and Electroaoustics, AU-21..

Lisker, L.and A.S.Abramson. (1964) "A cross-language study of voicing in initial stops: acoustical measurements", Word 20,384-422.

Miller, J., R. Kent, & B. Atal. (eds.) (1991) *Papers in Speech Communication: Speech Production, Speech Perception, Speech Processing.* (Three Volumes) The Acoustical Society of America.

Nirenburg, S., Carbonell, J., Tomita, M., and K. Goodman. 1991. *Machine Translation: A Knowledge-Based Approach.* CA: Morgan Kaufmann Publishers.

Perkell, J.S.and D.H.Klatt (1986) *Invariance and variability in speech processes.* Hillsdale, New Jersey, Erlbaum.

Potter, R.,G.Kopp, & H.Green (1947) *Visible speech.* New York: Van Nostrand.

Seneff, S.(1979) "A Spectrogram Reading Experiments", unpublished paper,MIT.

Silva, D. J. (1991) "A prosody-based investigation into the phonetics of Korean stop voicing," *Harvard Studies in Korean Linguistics* 4: 181-195.

Svensson, S.G.(1974) Prosody and grammar in speech perception, Monographs from the Institute of Linguistics, MILOS,2.

Yannakoudakis, E. & P. Hutton. (1987) *Speech Synthesis and Recognition Systems.* Ellis Horwood Limited.