연속 음성인식에 있어서의 음운론의 역할을 재고함*

김 기 호
(고려대학교)

1. 소 개

기계 번역(Machine Translation)의 궁극적 목표는 단순한 문서 번역을 넘어 아마도 자연스런 대화체의 두 다른 언어를 실시간으로 정확히 통역해 주는 다언어 자동 통역 장치를 만드는 데 있다고도 볼 수 있다. 실제로 외국을 여행할 때나 국제회의에 참석할 때, 국제 통화를 할 때, 그리고 외국 영화를 감상할 때나 유선 방송을 통해 외국 소식이나 스포츠 중계등을 들을 때 등등 다언어 자동통역 장치의 실용적인 효용성은 무한한 것이다. 이러한 구어체의 자동통역을 향한 첫번째 과정이 바로 연속 음성 인식의 지식을 기계 번역에 도입하는 일이므로, 대화체의 문장을 효과적으로 인식하는 연속 음성 인식의 중요성은 더욱 강조된다고 할 수 있다.

그러나 한국어 음성인식의 경우 소규모 어휘의 화자의존적 고립 단어 또는 연결 단어에 대한 실험 결과들이 최근 몇몇 보고되고 있을 뿐, 대규모 어휘의 화자독립적 한국어 연속음성인식에 대해서는 아직도 초보 단계를 벗어나지 못하고 있다 (김순협 1991, 진용옥 1992 참조). 따라서 본 고의 목적은 한국어 연속음성인식에 있어서의 음운부의 역할을 재조명하므로써 장기적인 안목에서 보다 효율적인 음성인식 알고리듬을 구현하는데 도움을 주고자 하는데 있다. 이를 위해 먼저 제 1항에서 음성인식의 간략한 역사적 배경과 음성인식의 몇가지 기본 방식들을 소개하며, 제 2항에서는 음성인식의 몇가지 기존 모델을 소개 한 후, 제 3항에서 이들 모델에서 공통적으로 해결해야 할 연속 음성인식에 있어서의 음운부의 역할의 문제를 고려하고자 한다. 그리고 제 4항에서는 오류 분석의 탐색과정에서 어떻게 자질 접근 방식이 분절음 접근 방식보다 우월한지를, 그리고 제 5항에서는 유표성 이론이 음성인식에 있어서 어떻게 효과적으로 이용될 수 있는지를 보이도록 하겠다. 끝으로 제 6항에서는 음운론의 관점에서 대화체의 연속 음성인식을 위한 음성인식의 모델을 제시하도록 하겠다. 즉 음운부의 역할이 입력된 음성파형으로부터 단순히 음소의 연쇄만을 추출해 내는 것이라는 기존의 입장에서 더 나아가 연속음성인식에 있어서의 그 역할이 어떻게 형태부과 통사부 및 의미(상황)부와 연결되어 확대될 수 있는지를 보이도록 하겠다.

1.1 음성인식의 역사적 개관

음성인식에 대한 관심은 1930년대 Vocoder의 개발로 거슬러 올라 가지만 실질적인 연구는 1940년대의 음성 스펙트로그램의 개발 이후 Potter, Kopp & Green의 스펙트로그램 판독실험 등으로 더욱 구체화되기 시작하였다고 볼 수 있다. 그 후, 1950년대와 1960년대에는 주로 화자의존적 소규모 어휘의 고립 단어의 음성인식에 많은 연구가 있었고, 그 후 1970년대 미국방성의 후원을 받은 ARPA계획 이후에야 비로서 연속 음성인식에 대한 본격적인 연구가 시작되었다고 할 수 있다. 1975년 ARPA계획의 결과로 보고된 HEARSAY-II system과 WHIM system, 그리고 HARPY system과 같은 1,000 단어 어휘의 음성인식 장치가 본격적인 대규모 어휘의 연속음성인식 실험에 있어서 비교적 성공적인 것으로 보고되었고(Lea 1980 참조), 같은 해 일본 NTT의 Itakura는 효율적인 단어인식을 위해 DTW(dynamic time warping)이라는 새로운 음성인식 방식을 소개하였다.

연속음성 인식에 대한 연구는 1980년대에도 계속되었는데, 1982년 벨연구소의

Wilpon et al.은 화자독립의 고립단어 1129개에 대한 음성인식 실험에서 91%의 인식율을 보였다고 보고하였으며, 1983년 카네기-멜론 대학의 FEATURE System은 문법의 도움없이 자질에 기초한 방식만을 사용하여 90% 이상의 정확도로 영어철자를 인식한 것으로 보고되었다. 그리고 1985년 통계적 처리를 중요시하는 IBM 음성인식팀에서는 Tangora System을 개발하였는데, 비록 화자종속적이지만 5,000단어 규모의 문장에 대해 97%의 높은 인식율을 보여주고 있다. 문맥의존적 음소인식 모델을 사용한 Bolt, Beranek & Newman의 BYBLOS System 역시 997 단어의 연속음성인식에 93%의 인식율을 보인 것으로 보고되었고, 1988년 HMM(Hidden Markov Model) 모델을 이용한 카네기-멜론 대학의 Sphinx System은 ARPA계획의 기존 997개 단어의 문장을 화자독립적으로 997, 60, 20의 문법 난이에 따라 각기 71%, 94%, 96%의 인식율을 보인 것으로 보고하였다. 이와같이 일본, 유럽, 미국의 음성인식연구팀들의 대규모 어휘의 화자독립 연속음성인식에 대한 연구는 1990년대에도 계속 진행되고 있으며 현재 상당한 수준에 이르고 있는 것으로 보고되고 있다 (김기호 1991b, Nirenburg et al. 1992 참조).

한국에서도 비록 기초수준에 불과하지만 소규모 단어의 인식에 대해서는 여러실험이 진행 보고되고 있으며 (김순엽 1991, 진용옥 1992 참조), 장기계획의 일환으로 한국어의 연속음성인식을 위한 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 요소기술개발 보고서」(1991) 등.

1.2. 음성인식의 기본 방식

주어진 음성파형으로부터 음성 특징들을 추출하여 적절한 음소와 단어로 연결시키려는 방식들 중 현재 주로 사용되고 있는 방식들로는 Dynamic Time Warping(DTW)을 이용한 패턴 매칭 방법과 Hidden Markov Models(HMM)를 이용한 음성인식 방식, Connectionist Network(CN)을 이용한 신경망 음성인식 방식, 그리고 음성 음운 지식을 이용하는 전문가 음성인식 방식(Knowledge-based approach) 등을 들 수 있다.

DTW 방식은 주어진 음성입력에 시간적 변화까지를 보완하여 미리 입력된 표준 패턴과 비교하여 가장 유사한 것을 채택하는 패턴 비교방식(Template-based approach)으로써 소규모의 제한된 단어 인식에는 비교적 좋은 결과를 보여주고 있으며, 실제로 간단한 응용분야에 많이 이용되고 있는 방식이지만 대규모 어휘의 연속음성인식에서는 한계가 있는 방식이라고 볼 수 있다. HMM 방식은 통계적 모델을 이용한 음성인식 방식으로 음성분할(segmentation)과 음성인식을 동시에 통계적으로 해결하고자 하는 방식이다. 따라서 주로 대용량의 연속음성인식에 많이 이용되고 있으며, 실제로 현재까지 알려진 대용량 시스템으로서는 가장 성공적인 것으로 간주되고 있다. 그리고 CN의 신경망 회로 방식은 인간 두뇌의 생물학적 신경 계통을 모방한 인공신경망을 이용하여 많은 간단한 연결마디들에 음성 특질들을 분산 분포시켜 음성인식을 실현하고자 하는 방식으로 최근 새로이 부상되고 있는 연속음성인식 방식이다. 한편, 전문가 음성인식 방식은 음향음성적 지식과 음운 지식을 바탕으로 마련된 화자독립의 연속음성인식 시스템으로 일본의 SPREX (A Speech Recognition Expert)나 V. Zue를 중심으로 한 MIT 음성인식팀에서 주로 이용하고 있다.

이들 방식들의 구체적인 장단점과 알고리듬 등은 관련된 논문들을 참조하기로 하고 본 고에서는 한국어 연속음성인식에서 이들 음성인식방식들에게서 공통적으로 직면하고 있는 음운부의 역할이 무엇인지, 특히 구어체의 문장인식에 있어서의 음운부의 역할을 주로 논의하고자 한다. 이에 앞서 다음 2항에서는 음성인식의 기존 몇몇 모델들을 간략히 소개하겠다.

2. 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

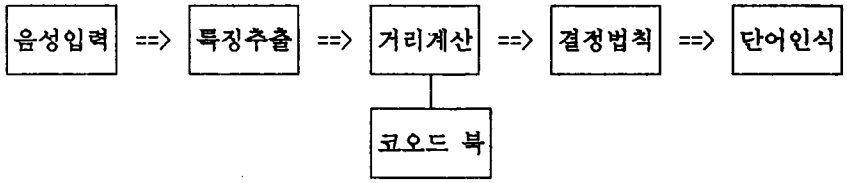


그림 1에서 보는 바와 같이 먼저 입력된 아날로그 음성신호를 저역 여과(Lowpass Filtering)시켜 표본화하는 전처리과정을 거친 후, LPC, PITCH, LPC-CEPTRUM 등을 이용하여 음성요소의 특징을 추출한다. 그 후 VQ를 이용한 음성인식은 미리 저장해 둔 특징벡터중에서 가장 잘 매칭되는 하나의 벡터와 매칭시켜 단어를 인식하게 한다. 이때 DTW와 같은 방식을 사용하여 시간적 차이를 보정한 후 가장 유사한 패턴과 매칭되도록 한다. 그러나 그림 1의 모델은 주로 단어인식에 사용되는 방식이기 때문에, 연속음성인식을 위한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

한편 대화체의 연속음성인식을 위한 기본 모델에서는 단어보다 하위단위인 음절, 겹친이음(diphone), 또는 음소를 기본 인식단위로 삼고 있는데, 일반적으로 다음 그림 2에서와 같이 음소를 기본으로 하고 있다.

그림 2. 연속음성인식의 주요 부문들

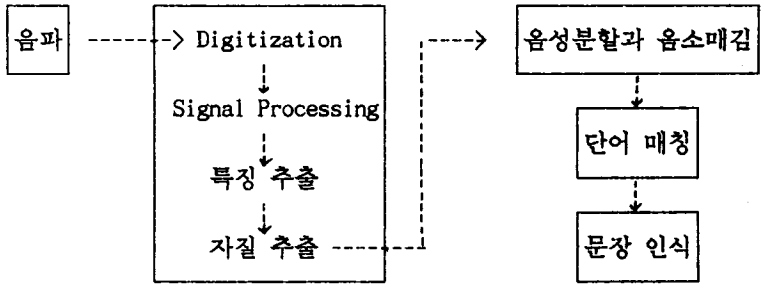


그림 2에서 보는 바와 같이 주어진 음파를 먼저 아날로그 신호에서 디지탈 신호로 바꾼다. 그 후 디지탈 신호는 필터를 통과하여 시간(가로축)과 주파수(세로축)를 함수로 음향 에너지(세기)를 나타내 주는 스펙트로그램 모양으로 바뀌며, 여기에서 영교차율(zero-crossoing rate), 여러 주파수대의 에너지, 포먼트 흐름(formant track) 등의 패러미터를 추출한다. 이러한 패러미터로부터 음성자질들을 추출해 내며, 이러한 자질로부터 음소단위가 도출된다. 그리고 이러한 음소연쇄로부터 순차적으로 단어를

매칭하고 문장을 인식하게 된다(Church 1987 참조). 물론 여기서도 하위 단위 과정에서 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)

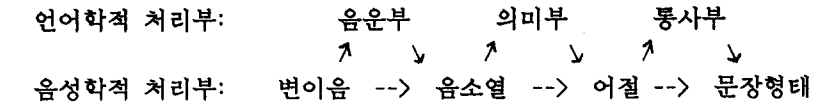


그림 3의 음성인식 모델에 의하면 음성처리부로부터 음성분할(segmentation)이후 인식된 변이음(phoneme-like unit)을 입력 단위로 하고 있는데, 음운부의 역할이 음소 유사단위로부터 일련의 음소열을 찾아내는 일에 국한되어 있다. 물론 음운부에서는 구개음화, 유성음화, 비음화와 같은 한국어 음운규칙들과 허용가능한 음소연쇄, 그리고 *#CC, *CCC, *CC#, *-r#,, *rC 등과 같이 허용되지 않는 음소연쇄를 규정해 주는 한국어 음소배열제약(phonotactic constraints) 등을 이용하여 각 해당 음소 마다 허용하는 인접음소와 허용하지 않는 인접음소를 설정하여 주므로 가능한 주위 음소열을 예측하는 일을 한다. 그러나 음성처리부에서 주어진 음성 파형으로부터 어떠한 음성 특질들을 어떻게 추출할 것인지에 대해 좀 더 구체적인 언급이 필요하다. 더우기 그림 3의 모델에서는 음운론의 역할이 단지 변이음으로부터 음소열을 찾아내는 역할만을 하는 것으로 기술되고 있으나, 다음 제 3항에서 필자는 음운부의 역할이 연속음성인식의 경우 이보다 더 확장되어 어떻게 형태부, 통사부, 및 의미부와 연결되어야 하는지를 보이도록 하겠다.

3. 연속음성인식에 있어서의 음운부의 역할

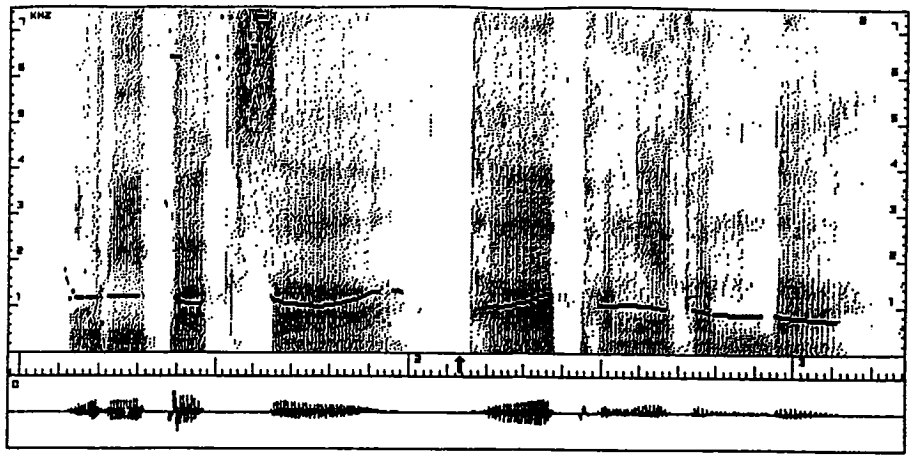
본 항에서는 음운부의 역할을 두가지 부면에서 즉 분절음(segment) 상위단위인 운율적인 정보와 분절음 하위단위인 변별자질의 음운정보로 나누어 설명하고자 한다.

3.1. 분절음 상위단위의 음운 정보

일반적으로 음성인식에서의 음운부의 역할은 앞서 2.2에서 밝힌 바와 같이 주어진 음성파형으로부터 음성 특징들을 추출하고, 그후 음운규칙과 음소배열제약 등을 이용하여 올바른 음소의 연쇄를 도출하는 것에 한정되어 있다. 일례로 그림 4의 스펙트로그램 상에 나타난 음성파형으로부터 음소 연쇄 즉 /uripaksatapaŋekanta/를 도출하는 것에 음운부의 역할이 국한되어 있으며, 그 후 적절한 단어와 문장으로의 인식은 형태부와 통사 및 의미부에서 다루어지고 있다.

그러나 /uripaksatapaŋekanta/의 음소연쇄로부터 바람직한 문장을 도출하기는 쉽지가 않다. 왜냐하면, 주어진 연쇄는 다음 (1)에서 보는 바와 같이 여러 문장으로 파싱될 수 있기 때문이다.

그림 4. /uripaksatapaŋekanta/ '우리박사다방에간다' (광역 스펙트로그램)



(1) /uripaksatapaŋekanta/

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 우리는 일상 대화에서 주어진 음의 연쇄의 의미를 파악하는데 전혀 어려움을 느끼지 않는다. 이러한 이유는 우리가 말하는 음성 파형에는 음소적 특성 뿐만 아니라 형태적 정보와 통사 및 의미적 정보가 충분히 내포되어 있기 때문이다. 일례로 한 단어로 구성된 다음 (2)의 문장을 고려해 보자.

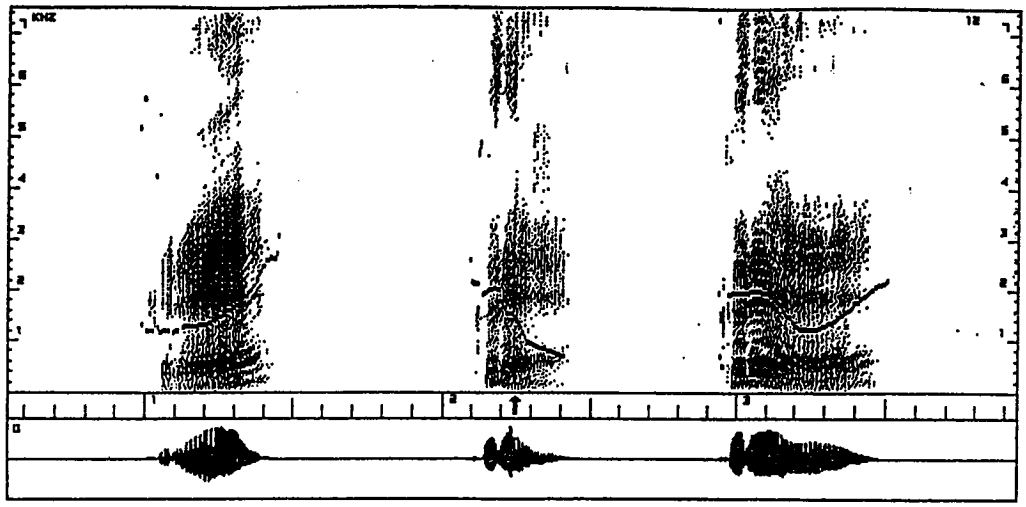
(2) 가) 그래? (정말 그 말이 맞아?)
나) 그래. (그 말이 맞아.)
다) 그래!? (비꼬면서, 그렇다고 치자!)

(2)에서 보는 바와 같이 '그래' /kirɐ/ 라는 단어의 음소 연쇄는 몇가지 뜻을 내포할 수 있다. 그러나 실제로 대화를 하는 청자는 아무런 문제없이 그 의미를 파악할 수 있다. 왜냐하면 '그래'라는 단어의 문장은 그 의미에 따라 달리 즉, 상승, 하강, 또는 상승하강의 억양으로 발음되며, 바로 이러한 운율적 차이에 의해 달리 해석될 수 있기 때문이다. 다음 그림 5는 피치의 변화를 나타내기 위해 협역(narrow-band)의 스펙트로그램으로 '그래' /kirɐ/를 나타낸 것이다.

그림 5에서 보는 바와 같이 /kirɐ/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 이와 같이 문장 끝 뿐만 아니라 주요단락의 구분 분석에 도움이 되는 많은 정보를 가지고 있다. 그림 4와 그림 6의 광역(wide band) 스펙트로그램을 비교해 보자.

그림 4와 그림 6을 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 현

그림 5. '그래' /kirɐ/ (협역 스펙트로그램)



저히 긴 것을 알 수 있다 (/i/ 94ms vs. 278ms ; /a/ 53ms vs. 354ms, 84ms vs. 180ms).

따라서 피치와 음장의 운율적 정보를 이용하면, 주어진 /uripaksatapaŋekanta/ '우리박사다방에간다'의 음소의 연쇄는 각기 다음 (3)과 같이 파싱될 수 있다.

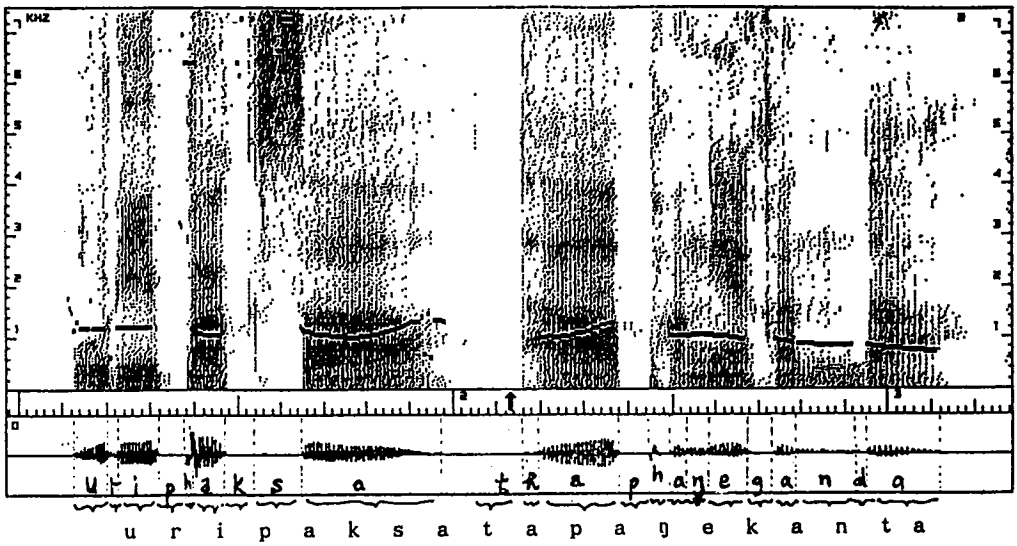
(3) a. #uripaksa# 우리박사 b. #uri# 우리
#ta# 다 #paksatapaŋekanta# 박사다방에간다
#paŋekanta# 방에간다

만일 음운부의 역할이 단순히 주어진 음파로부터 음소의 연쇄만을 도출하는 것이라고 가정한다면, /uripaksatapaŋekanta/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, (3a)의 경우에는 [uripaksa]PP, [ta]PP, [paŋekanta]PP 의 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 /paŋekanta/의 음소 연쇄에서는 '바', '방', '방에', '방에 간다'를 거쳐 손쉽게 '우리 박사 다 방에 간다'를 도출할 수 있다. 따라서 연속음성인식에 있어서의 음운부의 역할은 당연히 기존의 음소의 연쇄 도출에서 통사부와 의미부의 문장 파싱에까지 더 확장되어야만 한다. (음운 파싱(phonological parsing)에 대해서는 Church (1987)와 김기호 (1991a)를 참조하기 바람.)

3.2. 분절음 하위단위의 음운 정보

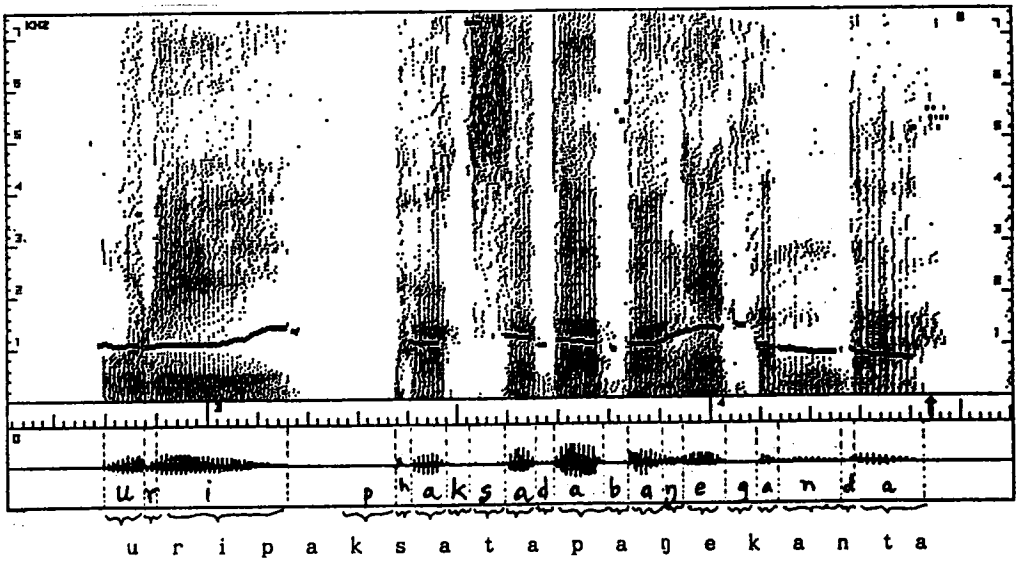
분절음 상위 단위인 음장, 피치 등의 운율적인 음운 정보외에도 분절음 하위 단위인 변별적 자질들의 정보 역시 연속 음성 인식에 매우 중요하다. 최근 김기호(1991a)는 음성인식에 있어서도 자질 접근 방식이 분절음절 접근 방식보다 더 효과적임을 지

그림 4. ##[우리#박사]PP##[다]PP##[방에#간다]PP## (PP=Phonological Phrase)



88 22 [redacted] ? 69 ? 102 [redacted] ? [redacted] ? 30 55 98 ? 45 134 ? [redacted]
(63+[redacted]=81) (76) (?+[redacted]=?) (?+[redacted]=?) (30+28=58) (30)

그림 6. ##[우리]PP##[박사다방에#간다]PP##



87 22 [redacted] ? 65 ? 70 [redacted] ? [redacted] ? 62 34 75 ? 41 140 ? [redacted]
(?+[redacted]=?) (55) (41) (62) (75) (30)

적한 바 있으며, 그럼에도 불구하고 보다 능률적인 자질 접근 방식이 음성인식에 적용되지 못한 이유가 Chomsky & Halle (1968: SPE) 자질 체계의 문제에 부분적으로 기인하였다고 간주하여, 자질기하학 이론의 틀에서 음성 인식을 위해 다음 그림 7과 같이 새로운 계층적 자질모형을 제시한 바 있다.

그림 7. 계층적 자질모형 (김기호 1991a)

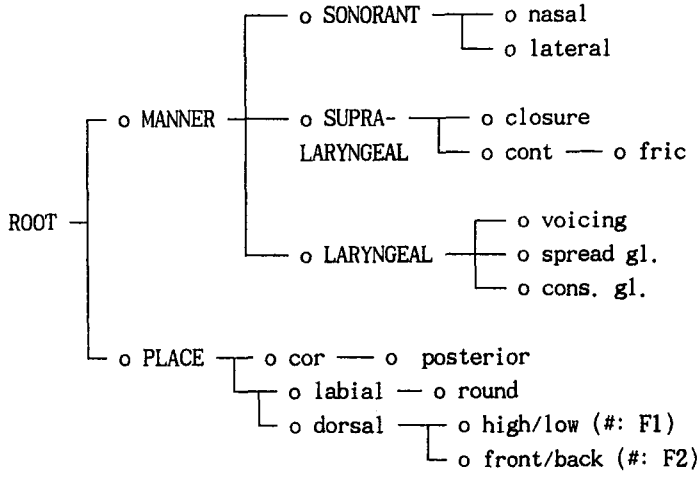


그림 7의 자질체계에서는 SPE의 자질체계와는 달리 자질들이 관련된 부류들로 분류될 뿐 아니라 계층적인 구조를 갖는 것으로 기술되고 있다. 뿐만 아니라 이들은 각기 조음음성학과 음향음성학적 특성들을 갖고 있다. (자세한 점은 김기호 (1987, 1991a, 1991b 참조). 이러한 음향음성적 특성들은 음성인식시 분절음 분리와 음소 표시(segmentation & labelling)에 매우 유용하게 사용될 수 있다. 실제로 스펙트로그램만의 판독으로 음성학자와 같은 정확도로 음성을 인식한 V. Zue의 경우, 주어진 음성파형으로부터 음소를 분류할 때 다음 도표 1의 음향음성적 정보를 이용한 것으로 보고되었다.

이러한 음성적 특성들은 음성인식의 경우 문맥의 영향을 받지 않는 불변 특성(invariant cue)과 문맥의 영향을 받는 변이 특성(variant or allophonic cue)들로 나눌 수 있다. 이중 불변의 변별 자질들은 음성인식에 있어서 음소 분류와 음소 매김에 매우 요긴하게 사용되는 자질들이다(Perkell & Klatt 1986 참조). 이와는 달리 변이자질들은 일반적으로 음성인식에 불필요한 소음에 불과한 것으로 간주되어 왔다. 그러나 최근 Church(1987)는 이러한 변이자질들도 음성인식에 매우 유용하게 이용될 수 있음을 보여주고 있다. 예를들어 기식자질 [aspirated]는 한국어의 경우 '뿔'과 '풀'에서 보는 바와 같이 의미의 차이를 가져오는 변별 자질로써 불변자질로 간주된다. 그러나 영어의 경우에는 'spy'와 'pie'에서 보는 바와 같이 [p]와 [pʰ]는 음소 /p/의 변이음에 불과하므로 변이자질에 불과하다. 그러므로 기식자질은 한국어 음성인식에는 불변 자질로 요긴하게 이용될 수 있지만, 영어의 경우에는 음성인식에 불필요한 소음일 뿐이었다. 그러나 Church는 이렇게 무시되어온 변이자질들이 오히려 음운을 이용한 문장 파싱(phonologcial parsing)에 효과적으로 이용될 수 있음을 보여주었다. 다음 보기 (4)의 예들은 중의성을 갖는 음소의 연쇄들로써 음성인식에 통사론과 의미론의 정보가 필수적이라고 주장된 예들이다.

도표 1. V.Zue가 음성인식을 위해 사용한 음향음성적 특성들 (Cole et al. 1980:37)

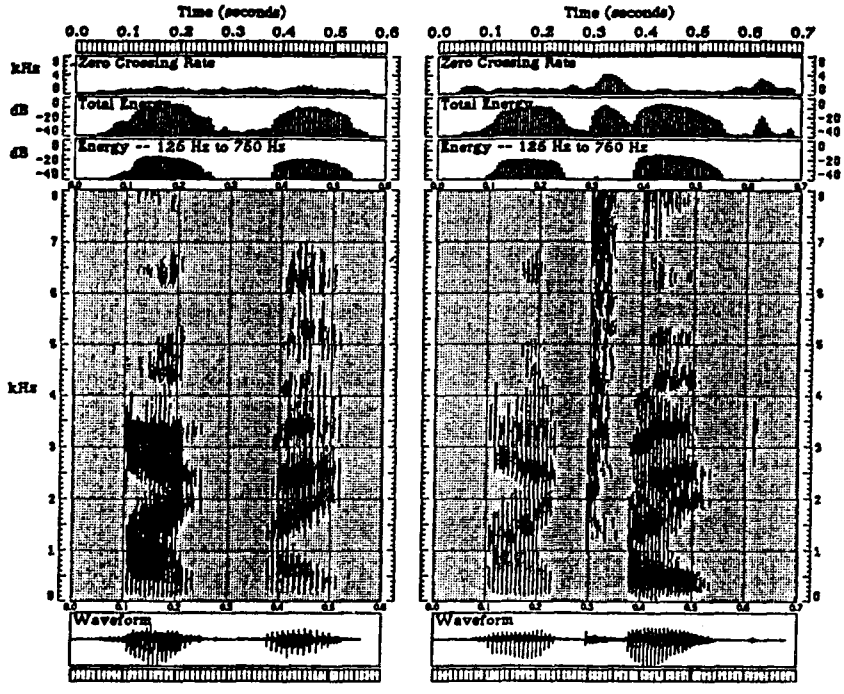
모 음	고/저설 전/후설 /a/ 축약모음	높이와 반비례로 변화됨 F1-F2 사이의 거리와 함께 변화됨 음장: /ɪ/가 /i/가 짧다 F1이 모든 모음중 가장 높다 길이가 짧다. 중립모음의 공명음대 형성
마찰음	유/무성 /s/ /ʃ/	유성 마찰음이 무성 마찰음보다 짧다 소음 > 4 kHz 소음 < 4 kHz
비 음		300 Hz 이하의 에너지, 급격한 강도 시작점(amplitude onset) 가짐 모음보다 강도는 약함 인접 모음의 비음화를 야기시킴
조음위치	순 음 연구개음	모든 공명음대가 아래로 내려옴 F2와 F3가 닫히는 시점에서 겹쳐짐
반전음	 /ɚ / /r/	F3가 2 kHz 아래로 내려감 F3가 F2를 따라감 F3가 F2에 맞다음
설탄음	[ ]	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) a. she prayed vs. sheep raid vs. sheep preyed /ʃipreɪd/
b. night rate vs. nitrate /naɪtreɪt/

만일 음운론의 역할이 주어진 발화로부터 단지 일련의 음소 연쇄만을 찾아내는 일로 끝난다면 보기 (4)의 중의적 해석은 통사론과 의미론에서 처리되어야만 할 것이다. 그러나 다음 그림 8에서 보는 바와 같이 이들의 중의성은 무시되어온 변이자질 [aspirated]에 의해 오히려 더 쉽게 분류될 수 있다.

영어의 경우 음절초에 위치한 폐쇄자음은 강하게 기식되어 발음되지만 음절말에 위치한 폐쇄자음은 불파되어 발음된다. 그러므로 'night rate'과 'nitrate'은 음소적으로는 /naitreit/로 같지만, 음성적으로는 달리 나타난다. 즉 그림 8에서 보는 바와 같이 'night'의 음절말 /t/음은 불파되어 나타나지만, 'nitrate'의 /t/음은 음절초에 위치하므로 강한 기식음을 동반하여 발음된다. 따라서 그림 8에서 보는 바와 같이 음절 환경에 영향을 받는 변이자질인 기식자질에 의해 오히려 (4)의 중위성은 쉽게 해결될 수 있게 되는 것이다.

그림 8. "night rate"과 "nitrate" (광역 스펙트로그램: Church 1987:47 재인용)



한국어의 음성인식의 있어서도 이와같이 변별자질들을 이용하여 문장 파싱에 도움을 줄 수 있다. 한국어에는 다음과 같이 3종류의 폐쇄음이 있다: 가) (약하게 기식하는) 평음 (ㄱ, ㄷ, ㅂ), 나) (강하게 기식하는) 격음 (ㅋ, ㅌ, ㅍ), 다) (기식하지 않는) 경음 (ㄲ, ㄸ, ㅃ). 이들 세 종류의 폐쇄음들 중에서 단지 평음만이 유성음 사이에서 유성음화를 겪는다. (보기: /aka/ [aga] '아가', /kanta/ [kanda] '간다') 그런데 만일 음운부의 역할이 단지 주어진 음파로부터 음소의 연쇄만을 도출해 내는 것만으로 한정한다면, 유성성 [voicing]의 변이음적 특성은 음소 인식에 있어서 아무런 도움도 되지 않는 불필요한 요소에 불과할 것이다. 그러나 앞서 영어의 경우에서와 마찬가지로 한국어의 변이음적 요소인 유성성도 연속음성인식에 있어서 매우 긴요하게 이용될 수 있다. 앞의 그림 4와 그림 6에서 볼 수 있는 바와같이 국어의 평음 'ㄱ', 'ㄷ', 'ㅂ'은 어절 앞과 어두에서는 어중에서보다 강하게 기식되어 발음됨을 알 수 있다. 즉, '우리박사'의 'ㅂ'은 단지 18ms의 기식음 또는 VOT(voice onset time: 성대 진동 개시시간)를 보여 주는데 비하여, '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. 그러므로 앞 절에서 제시된 음장과 피치 등 분절음 상위 단위의 운율 정보 뿐만 아니라 평음의 유성음화와 같은 변이음적 음운 정보 역시 문장의 구문 분석에 효과적으로 이용될 수 있는 것이다. 다시말해서 한국어 평음은 어절 앞에서와 어절 내, 단어 앞에서와 단어 내, 그리고 문장 앞에서 각기 다른 VOT를 가지며, 이에 따라 역으로 평음의 상이한 VOT에 기초하여 단어와 어절 및 문장의 경계를 도출해 낼 수 있게 되는 것이다. 그러므로 /uripaksatapaŋ ekanta/의 음소 연쇄로부터 각기 상이한 VOT를 이용하여 그림 4의 문장은 '우리박사 다 방에 간다'로, 그리고 그림 6의 문장은 '우리 박사다방에 간다'로 음운적 문장 파싱이 가능해지므로 두 문장은 쉽게 구별될 수 있게 된다. (이와 유사한 주장의 통계

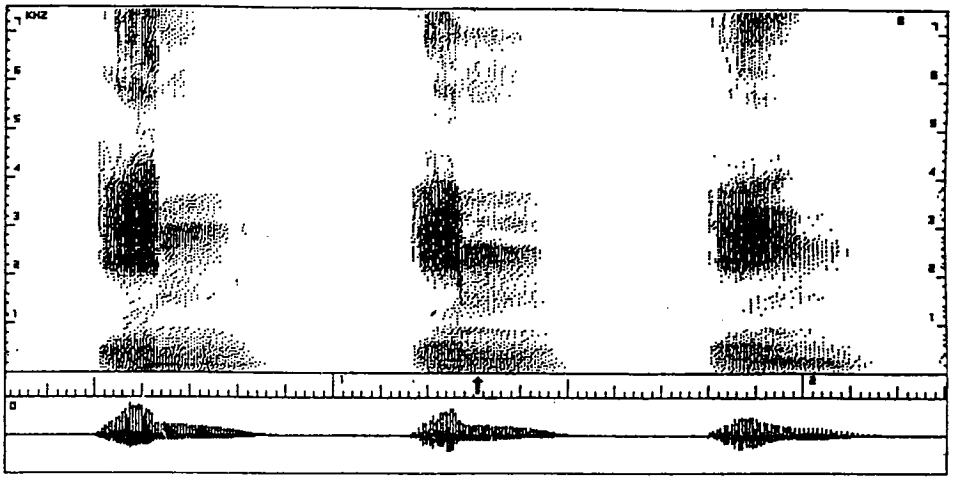
적 보고는 Silva(1991)를 참조하기 바람.)

그러므로 효과적인 연속 음성인식을 위해서는 분절음 상위 단위인 음장, 피치 등의 운율적 음운 정보 뿐만 아니라 분절음 하위 단위의 불변의 변별적 자질들의 특성은 물론 문맥 의존적인 변이자질들의 특성을 효과적으로 이용하는 것이 필수적이라고 할 수 있다.

4. 오류 탐색시에서의 자질 접근 방식의 효율성

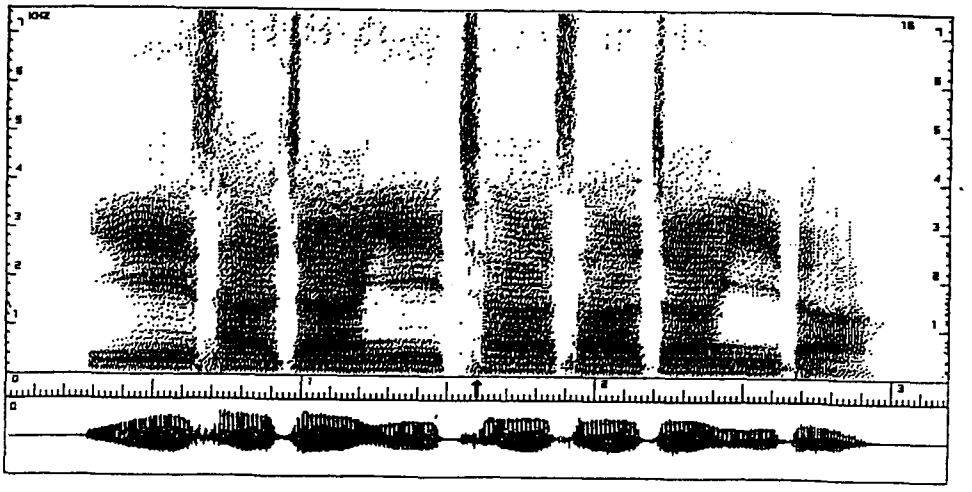
분절음 접근 방식에 비해 자질접근 방식이 갖는 장점중 하나는 음소인식 과정에서 일어나는 오류 발생시 올바른 음소를 찾는데 보다 효과적으로 시간을 절약할 수 있다는 데 있다. 즉 오류 발생시 가능한 모든 음소를 일일이 검색할 필요없이 오류의 가능성이 높은 자질 부류에서만 검색할 경우 시간을 훨씬 효과적으로 줄일 수 있다. 다음 그림 9에서 보는 바와 같이 모음과 비음은 공통적 특성인 공명음대(formants)의 존재로 순수자음과 쉽게 구분되며, 모음과 비음의 차이도 파형과 공명음대의 모양과 강도에 의해 쉽게 구별될 수 있다.

그림 9. '인' '임' '잉' (광역 스펙트로그램)



한편 비음의 조음위치들도 선행하는 모음의 공명음대의 전이로 구별이 가능하다. 그림 9에서 보는 바와 같이 순비음 'ㅁ'의 경우에는 선행 모음의 F2가 비음 앞에서 급격히 하강하고 있으며, 연구개 비음 'ㅇ'의 경우에는 선행 모음의 F2와 F3가 겹쳐지고 있다. 한편 치경 비음 'ㄴ'의 경우에는 별 변화가 없다. 여기서 주목할 점은 음성인식에서 오류가 발생할 경우에는 주로 조음위치에서 일어나며 조음 방식에서는 별로 오류가 발생하지 않는다는 점이다. 왜냐하면 비음의 조음 방식 자질의 음성적 특성들은 비교적 분명하기 때문이다. 따라서 오류 발생시 모든 분절음과 대조 분석하는 것이 아니라 오류 발생 가능성이 가장 높은 조음위치 자질을 먼저 검색하므로써 오류 검색의 시간을 효과적으로 줄일 수 있게 된다. 오류 검색시의 또 다른 예로 '이사장이 차사장이다'라는 문장의 스펙트로그램인 다음 그림 10을 살펴 보자.

그림 10. '이사장이 차사장이다' (광역 스펙트로그램)



영어에는 여러가지 마찰음과 파찰음이 존재하지만, 한국어의 마찰음과 파찰음에는 단지 치경음과 구개음만 있다. 그런데 이들의 조음방식 자질은 그림 10에서 보는 바와 같이 고주파수 대의 소음(white noise)으로 특징지울 수 있으며, 이러한 특징은 모음이나 전이음, 유음, 비음과 같은 공명자음과는 물론 폐쇄자음과도 명확히 구별된다. 그러므로 오류 발생시 오류 검색은 단지 치경마찰음 또는 경구개 파찰음 내에서만 찾으면 된다.

일반적으로 공명성 자질과 [closure], [continuant] 등의 조음 방식 자질들은 조음위치 자질에 비해 비교적 쉽게 음향음성적으로 구별된다. 그러므로 한국어 음성인식의 경우 그림 7의 자질 모형에서 다음 (5)에서 보는 바와 같이 변별성이 높은 자질 순으로 즉, [공명성] - [조음방식] - [후두자질] 또는 [조음위치]의 순서로 자질들을 찾게 되면 음소판독에 있어 시간을 비교적 효과적으로 절약할 수 있게 된다.

(5) [공명음] : 순수자음 / \ 공명자음 : 모음

[조음방식]: 폐쇄음 마찰음:파찰음 비음 : 유음

[후두자질] & [조음위치]:

즉 먼저 파형과 스펙트로그램상의 공명음대의 유무로써 순수자음과 모음 또는 공명자음을 구분한다. 순수자음인 경우, 조음 방식에 의해 폐쇄음과 마찰 또는 파찰음으로 구분된다. 그후 후두자질과 조음위치로써 평음/경음/격음 또는 순음/치경음/연구개음 등으로 세분된다. 한편, 공명자음인 경우, 먼저 비음과 유음으로 나누인 다음, 비음의 경우에는 조음위치에 의해 순음/치경/연구개 비음으로 세분된다. 그리고 모음의 경우에는 제 1 공명음대(F1)에 의해 고설과 저설 모음으로 구분되며, 제 2 공명음대(F2)에 의해 전설과 후설 모음으로 구분될 수 있다. 한편 오류 발생시에는 음소 탐색의 방향과는 반대로 오류가능성이 가장 높은 자질 부류에서부터 역으로 검색해 나가면 효과적으로 탐색 시간을 줄일 수 있게 된다.

5. 음성인식과 유표성 이론

유표성 이론 역시 단어 탐색과정에 유효하게 이용될 수 있다. 다음 보기 (6)에서 볼 수 있듯이 한국어의 비음동화나 조음위치 동화의 경우 그 방향성은 항상 일정하다. 즉 무표적인 음이 유표적인 음에 동화된다.

(6) 가. 폐쇄음 + 비음(또는 유음) ==> 비음 + 비음
국민 [궁민], 밥물 [밤물], 닫는다 [단는다]
국+만 [궁만], 밥만 [밤만], 닫니 [단니]

나. 치경음 > 순음 또는 연구개음, 순음 > 연구개음
군밤 [군밤] 또는 [굼밤], 군고구마 [군고구마] 또는 [궁고구마], 등등.

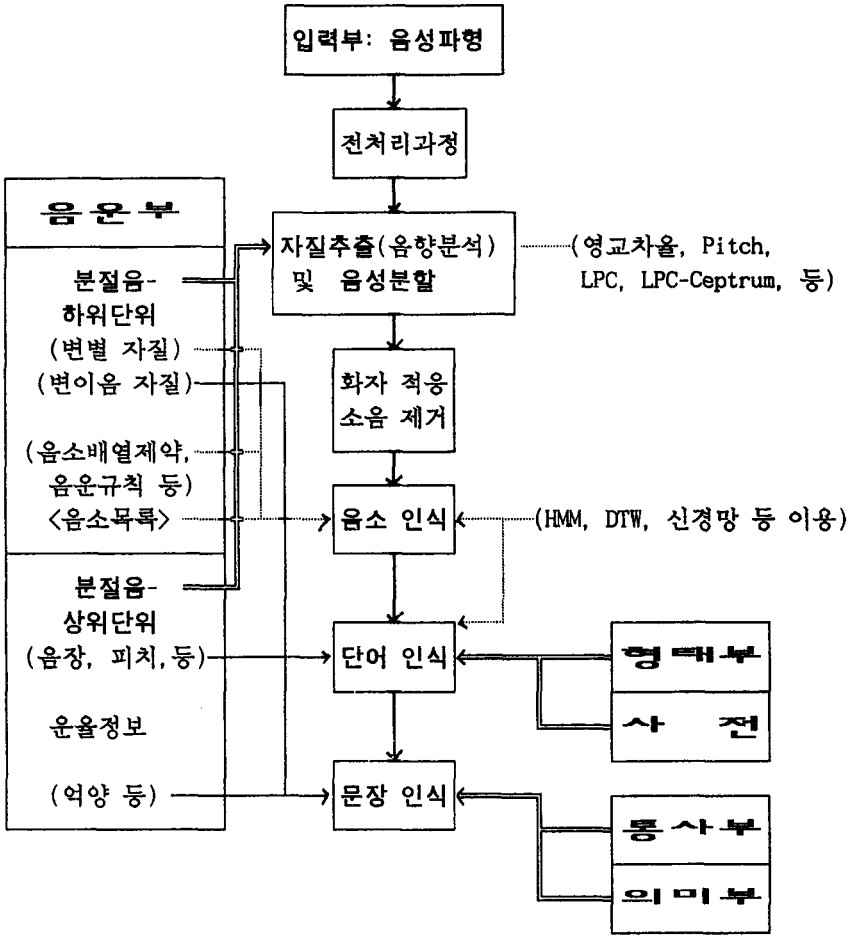
(6)의 보기에서 보는 바와 같이 종성의 폐쇄음은 후행하는 초성의 공명자음 앞에서 동일 조음위치의 비음으로 바뀌며, 치경음은 순음과 연구개음 앞에서 위치 동화를 받는다. 그러나 그 역 방향의 동화는 성립되지 않는다(김기호(1987) 참조). 이때 어휘부 또는 사전의 발음 사전에는 발음대로, 예를들어 '국민'은 [궁민]으로 기재될 수 있다. 그러나 복합어나 조사와의 결합, 또는 연속된 문장속에서의 모든 가능한 변화 상황을 일일히 기술하는 것은 잉여적인 요소의 반복일 뿐이므로 비효율적이며 낭비적이라고 할 수 있다. 오히려 동화의 일정한 방향성을 이용하여 이를 규칙으로 처리하는 편이 훨씬 더 경제적이다. 그러므로 음소인식과 단어 탐색의 과정에서 유표음과 유표음의 연쇄가 나타날 경우, 두가지 가능성, 즉 유표음과 유표음의 연쇄 또는 무표음과 유표음의 연쇄의 경우를 모두 고려하여 단어를 매칭시키도록 하여야 할 것이다. 예를들어 [밤만]의 경우 '밤만'과 '밥만'의 두가지 가능성을 모두 검토하여야 한다.

6. 결론

지금까지 살펴본 음운론의 역할을 고려해 볼 때, 연속 음성인식에 있어서 음운론의 역할은 주어진 음파로부터 일련의 음소 연쇄만을 도출해 내는 기존의 역할에서 훨씬 더 확장되어야 함을 알 수 있다. 즉 음소 분할과 음소 표기, 그리고 효율적인 단어 검색과 문장 파싱을 위해서는 불변의 변별적 자질들의 음향음성적 지식은 물론 변이음적 음운 정보와 분절음 상위단위의 정보인 음장, 피치, 억양 등의 운율정보를 효과적으로 이용하여야만 하며, 이에 따라 음성 인식에 있어서의 기존의 형태부, 통사부, 및 의미부의 부담은 더욱 줄어들게 되는 것이다. 그러므로 연속 음성인식의 모델은 다음 그림 11과 같이 수정되어야 한다.

먼저 입력된 아날로그 음성신호를 디지탈화한 후, 저역여과시켜 표본화하는 전처리 과정을 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 문장 파싱(suprasegmental phonological parsing)에 이용될 수 있기 때문이다. 이러한 자질들의 조합에 한국어 음소배열제약, 유표성 이론, 그리고 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.

그림 11. 연속 음성인식을 위한 음성인식 모델



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.

결론적으로 연속 음성인식에서의 음운론의 역할은 주어진 음파로부터 가능한 음소 연쇄의 추출이라는 제한된 역할을 벗어나 그림 11에서 보는 바와 같이 전 영역으로, 즉 자질 추출에서부터 음소 인식은 물론 단어 인식과 문장 인식에까지 확장되어야 한다.

* 본 논문은 김기호(1992)와 김기호(1993)에서 발표된 것을 일부 발췌하여 부분적으로 수정 보완한 것이다.

참 고 문 헌

김기호. (1987). *Phonological Representation of Distinctive Features: Korean Consonantal Phonology*. Ph.D. dissertation. U. of Iowa.

김기호. (1991a). "Revisting distinctive feature approach in speech recognition," *SICONLP'91 Proceedings*. 21. (서울대학교 어학연구 제 27권 2호. 255-272)

김기호. (1991b). 영어자질이론의 발전과 음성인식과의 관계. 「영어영문학」 제37권 3호, 783-803.

김기호. (1992). 음성인식에 있어서의 자질기하학 이론과 유표성 이론. 제 9회 음성통신 및 신호처리 워크샵 논문집, 35-39.

김기호. (1993). 연속 음성 인식에 있어서의 음운부의 역할. HCI '93 학술대회.

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*, 183-198.

김종미. (1990). 언어학을 활용한 국어음성인식. 음성인식및 신호처리 WORKSHOP, pp. 170-177.

진용옥. (1992). 음성 정보처리 기술 및 음성 정보 서비스의 발전과 전망. 제 9회 음성통신 및 신호처리 워크샵 논문집, 12-26.

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

Cole, R. (ed.) (1980). *Perception and Production of Fluent Speech*. NJ: Lawrence Erlbaum.

Cole, R., A. Rudinsky, V. Zue & D. Reddy (1980). "Speech as pattern on paper," in R. Cole (ed.), *Perception and Production of Fluent Speech*.

Klatt, D. (1980). "Overview of the ARPA speech understanding project (1)," in W. Lea (ed.), *Trends in Speech Recognition*, Prentice-Hall.

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

Nirenburg, S., Carbonell, J., Tomita, M., and K. Goodman. (1992). *Machine Translation: A Knowledge-Based Approach*. Ca: Morgan Kaufmann Publishers.

Perkell, J. & D. Klatt. (1986). *Invariance and Variability in Speech Processes*. NJ: Lawrence Erlbaum.

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

Stevens, K. (1981). "Invariant acoustic correlates of phonetic features," *JASA* 69, suppl. S31.

Waibel, A. & K. Lee. (1990). *Readings in Speech Recognition*. Ca: Morgan Kaufmann.

Zue, V. & L. Lamel. (1986). "An expert spectrogram reader: a knowledgy-based approach to speech recognition," in *ICASSP 86 Proceedings*, 23.1., 1986.

서울시 성북구 안암동. 고려대학교. 영어영문학과 (130-701)

Email: KEEHOKIM@KRKOREA1

Fax: 02-929-1957