음성인식/합성을 위한 기본 개념과 표기법의 정립

정 국
한국외국어대학교

Basic Concepts and Transcription for Speech Recognition and Synthesis

Kook Chung
Hankuk University of Foreign Studies

I. 서 론

1.0 연구의 배경

현재 음성인식/합성을 위한 여러가지 개념들 특히 음성인식/합성이라는 개념 자체까지도 이를 연구하는 여의 전문분야에 따라 서로 다르며 또한 음성표기의 방법도 사람에 따라 서로 다르다. 특히 이것은 음성공학자와 언어학자간에 두드러진다. 이와같이 연구자나 연구분야에 따라 개념이나 표기법이 다르면 이들간의 공동연구에 장애가 될 수 밖에 없다. 그러나 자세히 살펴보면, 이것은 단순히 서로간에 개념이나 표기법이 다른 것에 그치는 것이 아니라, 근본적으로 기본개념이나 표기법이 정립되어 있지 않은 것같이 보인다. 이것이 사실이라면 이것은 단순한 공동연구에만이 아니라, 음성인식/합성의 연구 자체에 근본적인 장애가 될 수 있다.

1.1 연구의 목적

본 연구의 목적은 음성인식/합성을 위한 개념과 표기법 정립의 문제와 필요성을 밝히고, 이를 위해 기본적인 개념과 표기법의 안을 제시함으로서, 앞으로의 완전한 정립을 위한 기반을 마련하여, 궁극적으로 음성인식/합성의 연구에 이바지하자는 데 있다.

음성인식/합성 방면에서는 언어학(특히 음성·음운학)적 개념이 잘 알려져 있지 않음으로서 생기는 문제가 대부분이므로 개념과 표기법을 주로 이러한 측면에서 제안하기로 하되, 언어이론의 설명은 최대한 지양하고 음성인식/합성에 실질적으로 도움될만한 것만 제시하기로 한다.

1.2 개념과 표기법의 관련성

여기서 개념과 표기법의 정립을 함께 논하는 이유는 이들이 상호관련이 되기 때문이다. 표기의 대상은 음성언어인 바, 음성언어에 대한 올바른 개념이 없으면 제대로 표기가 될 수가 없을 뿐 아니라, 표기법이 개념에 맞게 정립되지 않으면 개념이 살아날 수 없다. 표기법은 바로 개념의 표상이라 할 수 있으며, 개념은 표기법의 내용이라 할 수 있기 때문이다.

1.3 주요내용과 순서

이 글에서는 음성인식/합성이 문자그대로 '음성'의 인식/합성이 아니라 '음성언어'의 인식과 합성이라야 한다는 생각을 바탕으로 음성언어의 개념과 음성언어에 관련된 음성 및 음운 등의 기본 개념들을 단계적으로 확인해 보겠다.

그리고 이러한 개념을 바탕으로 음성언어의 기본적인 표기법의 예를 보이겠다.

개념 및 표기법을 논함에 있어 각각 먼저 문제점을 밝히겠다.

II. 기본개념

2.0 기본개념의 문제

음성인식/합성에 있어서의 개념상 오도될 가능성이 있는 것은 다음과 같은 기본적인 개념들이다.

음성인식/합성이라는 개념 자체
음성과 음성언어의 개념
음소 및 음성이라는 개념
음운규칙 개념

음성인식/합성이라는 개념 자체가 문제라는 것은 '음성'인식, 합성이라고 함으로서 인식·합성의 대상이 '음성'이라고 생각하기 쉽다는 것이다. 그러나 인식·합성의 대상은 '음성'이 아니다. 음성이란 소리(sound, 또는 phone)로서 자음/모음 따위의 개개 음을 말하며, 그 개개음도 실제 물리적으로 구현된

음성학적 차원의 음을 말한다. 물리적인 음의 인식은 음성인식에서 궁극적으로 추구하는 바가 아니다. 가령 '있고'라는 말은 '음성'적으로 보면 [it˥ko]가 될 수 있는데, 음성인식은 이와같은 물리적인 [it˥ko]의 인식을 거쳐, 이것이 '있고'라는 말임을 이해하여야 그 목적이 달성된다. 다시 말하면 자음모음 따위 소리자체의 인식이나 합성이 목적이 아니라 '있고'라는 '말'의 이해가 목적이라는 것이다. 이 '말'이 곧 '음성언어'이다.

문자그대로의 '음성'의 인식/합성은 진정한 의미의 (음성언어의) 인식/합성의 초보단계에 불과하다. 즉 개별음의 인식/합성을 통해 결국 '(음성)언어' 즉 '말'('글'에 대비되는 '말')의 인식/합성에 도달해야 되는 것이다. 이렇게 볼 때 음성인식/합성은 단순한 '소리'의 인식/합성이 아니라 '말'의 인식/합성이라는 개념을 가져야 할 것이다.

영어에서 음성인식/합성을 sound recognition/synthesis라고 하지 않고 speech recognition/synthesis라고 하는 것은 바로 이 speech가 음성을 가리키는 것이 아니라 음성언어(spoken language) 즉 '말'을 가리키는 말이기 때문이다.[1]

지금까지 말한 음성과 음성언어의 개념상의 문제 외에 또 한가지 중요한 것은 음소의 개념이다. 흔히 공학에서 사용하고 있는 '음소'라는 말은 실은 '음성'을 가리키는 말인 바, 이것은 단순한 용어의 차이에 그치지 않고 음에 있어서의 '음성'과 '음소'라는 두 개의 차원에 대한 개념이 없다는 점에서 중요한 문제가 된다. 왜냐하면 소리 특히 언어음을 단순히 하나의 차원에서만 생각하는 것은 음성인식/합성 자체에 대한 개념과 크게 관계가 있기 때문이다. 이러한 두 개의 차원과 직접 관계된 것이 음성-음운규칙인 바, 이 두 개의 차원과 이를 연결하는 음성-음운규칙의 의의에 대해서는 아래에서 상론하겠다. ('음소' 및 '음성-음운규칙'들의 개념에 대해서는 (1) 참조.)

2.1 (자연)언어, 문자언어와 음성언어

음성언어를 이해하기 위해 자연언어, 문자언어의 개념과 대비시켜 보자.

우선 우리 인간이 사용하고 있는 한국어, 영어 따위의 언어는 컴퓨터언어 등과 대비하기 위하여 자연언어(natural language)라 할 뿐, '자연언어'란 단순히 '(인간)언어'를 가리키는 말에 불과하다. 물론 이 (자연)언어는 인간의 머리속에 존재하며, 이것은 주로 문자나 음성을 통해 표현되나 문자나 음성 외에 몸짓등(예: 手話)으로 표현되기도 한다. 언어를 이와같이 구분할 때, '언어'(language)라는 개념은 좁은 의미로는 머리속에 들어있는 언어를 가리키고, 기록된 것은 문자언어(written language), 발화되어 나오는 언어는 음성언어(spoken language) 또는 '말'(speech)이라고 한다.

흔히 자연언어처리(natural language processing)과 음성언어처리(spoken language processing)가 대비되지만, 실은 전자는 언어처리(language processing)이고 후자는 말 처리(speech processing)로서, 말처리란 음성으로 구현된 언어에 대한 처리이므로, 음성부분만 제외하면 둘 다 본질적으로는 언어처리인데, 마치 이 둘이 하나는 '자연' 또 하나는 '음성'언어의 처리인 양 비치는 것은 단순한 용어의 차이 때문이다. 오히려 자연언어처리에서는 음성언어가 아닌 문자언어를 기반으로 하는 바, 이 문자언어라는 것이 실제 음성언어를 기록한 것이 아니라 머리속에서 만든 이상적인 언어를 기록한 것이므로 '자연스런' 언어의 처리가 아닐 수가 있는 것이다.

2.2 음성언어의 특성

음성언어는 어떠한 특성을 가지고 있는가?

머리속의 '언어'와 발화된 '말'은 후자가 전자에서 나오는 것이므로 본질적으로는 물론 일치하지만, 실질적으로는 언어처리상 중요한 많은 차이가 있다.

말은 음성을 수반한다는 점에서 물론 명백한 차이가 있지만 이러한 점 외에 음성언어의 중요한 특성은 아래와 같다. ((14) 참조.)

가. 불완전하고 비문법적인 문장이 많다.

나. 용어가 구체적이 아니고 일반적이다.

다. 단어의 발음변이가 많다.

라. 구문이 단순하다. (복합문, 복합명사 등이 적다.)

마. 내용이 정밀하지 못하고, 비논리적일 수 있다.

바. 더듬거리거나 되풀이하거나 중간에 끊어지는 등 생리적 물리적 방해를 받는다.

사. 군소리나 구어체 용어들이 끼어 간다.

아. 운율적인 요인이 중요한 의미를 갖는다.

아래에 우리말과 영어의 '말'(음성언어)의 실제 예를 들어 본다. (영어에에서 ㅣ는 짧은 쉼을 나타냄.)

뭐 좀 들어보려고요. 여름 그 　해새시라고, 이게 뭐, 수영장 그런 거 뭐 무료로 할 수 있고 그런게 우리 집에 우송돼 왔는데.

A: you know but erm | they used to go out

[1] spoken language와 speech는 비슷한 말로서 각각 우리말의 '음성언어'와 '말'에 해당된다.

in erm August + they used to come + you know the lovely sunsets you get + at that time and
B: oh yes
C: there's a nice new postcard a nice--well I don't know how new it is + it's been a while since I've been here + of a sunset + a new one +
A: oh that's a lovely one isn't it

이들은 위의 특징을 거의 그대로 나타나고 있다.

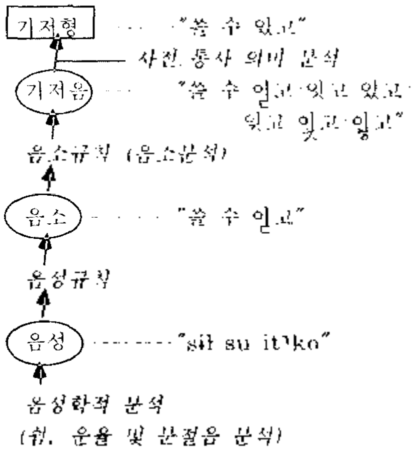
이들 예만 보아도 음성언어가 흔히 우리가 책에서 보는 언어와 얼마나 다른지 알 수 있다. 이러한 음성언어에 대한 연구가 없이는 음성인식/합성은 추상적인 차원에 머물 뿐, 실제 대화의 이해가 되지 못할 것이다.

물론 이러한 음성언어에 대한 연구는 음성·음운 분야에서만이 아니라 통사·의미 분야에서도 행해져야 한다. 현재 국어나 영어등 언어에 대한 연구는 대부분 머리속에 든 추상적인 언어나 문자로 쓰여진 문자언어에 대한 통사·의미 연구일 뿐 실제 발화된 언어에 대한 연구는 거의 없다고 할 수 있다.

더 정확히 말하면 연구 이전에 언어자료수집조차 되어있지 않은 상태이다. 따라서 음성인식/합성은 물론, 언어자체의 이해를 위해서도 음성·음운, 통사·의미 전 분야에 걸쳐 우선 음성언어자료의 수집과 분석이 반드시 있어야 할 것이다.

2.3 음성언어와 음성인식/합성

실제로 음성인식/합성이 어째서 음성언어의 인식·합성인가하는 것을 이해하기 위해서 구체적인 예를 들어 살펴보자. 위의 국어 예에서 '쓸 수 있고'하는 것을 택하여, 인식의 과정을 도표로 그리면 아래와 같이 되겠다.



음성언어의 인식에서 가장 먼저 해야할 것이 물론 음성학적 차원의 인식이다. 이때 이 '음성학적 차원'이라는 것은 단순한 자음/모음따위만을 지칭하는 것이 아니다. 여기에는 음 뿐만 아니라 음의 연결체 즉 음절, 구, 절(문장)도 포함된다. 따라서 음성의 특성을 밝힐 때는 자음/모음 등의 분절음적 특성만이 아니라 음의 연결체 전체 또는 부분의 길이나 고저 등 운율적 특성, 그리고 쉼도 포함되어야 한다. 아마도 쉼을 포함한 운율적 특성에 대한 분석이 분절적 특성에 대한 분석보다 선행하여야 할 경우가 많을 것이다. 왜냐하면 음의 연결체를 작은 단위로 나누는 데에 이들 운율적 특성이 필요하기 때문이다.

이러한 음성분석의 결과로 얻어져야할 것이 음소들이다. 따라서 음성인식은 일단은 음소인식의 차원이 되어야 한다.

음소인식의 차원 이상은 음운론적 차원이다. 즉 음성인식에서 it˺ko라는 것이 인식되었으면 그것은 음성규칙에 의해 음소적으로 '읻고'가 되는데, 음성인식과 음성규칙으로서 할 수 있는 것은 여기까지다. 그러나 이 '읻고'라는 단어는 국어에 존재하지 않는다. 이 '읻고'가 '있고'라는 것을 인식하는 과정은 음운규칙에 의할 수 밖에 없다. [t˺]이 ㄷ이 되는 것은 음성학적으로 설명되는 것이나, ㄷ이 ㅆ으로 되는 것은 음운적인 것으로, 음성학적인 설명이 불가능한 것이며 순수히 언어적인 현상이므로 이를 음운론적 것이라 하는 것이다. 이 때 음운론적으로 분석해서 찾은 음을 기저음이라 부른다. 음운론적으로 이 기저음을 찾는 것이 음성인식의 목적일 것이고, 음성합성은 이와 반대로 이 기저음이 음성학적으로 실현된 모습인 '음성'을 자연스럽게 만들어내는 일일 것이다.

기저음을 찾는 목적은 기저형(Underlying Form)을 찾기 위해서다. 기저형이란 기저음으로 구성된 형태소를 말한다. 이것은 정말 어려운 일일 수 있다. 왜냐하면 위의 예 '읻고'의 기저형은 '입고, 잇고, 있고, 잊고, 잋고, 잎고' 등 얼마든지 있기 때문이다. 이 때에 결정할 수 있는 것은 사전이나 의미론부에서 뿐이다. 우선 '입고, 잋고, 잎고'는 없기 때문에 사전검색에서 제외되며, 다음으로 남은 셋 중 '잇고, 잊고'는 통사론적으로 그리고 의미론적으로 되지 않으므로 남은 것은 '있고' 뿐이 된다.

이러한 기저형 탐색에서 음성언어가 중요하다는 것은 전후 구문이나 문맥으로 기저형을 결정할 때에 위의 예들에서 보인대로 음성언어가 보이는 비문법성, 불완전성 등 때문이다. 이상적인 문자언어의 경우와는 달리 완전한 정보가 없으므로, 음성언어의 특성을 알지 못하면 기저형 결정조차 못하게 되는 것이다.

그렇다면 분절음을 위주로 말한다면 결국 음성인식은 음성인식에서 시작하여 음소인식 내지 기저음 인식을 하는 것이 목표이고 음성합성은 기저음에서 출발하여 음성을 만들어내는 것이 되겠다. 여기에 덧

붙여야할 것이 온율이다. 특히 길이나 고저는 그것이 통사적 또는 의미적 의미가 있는 것을 분절음과 함께 인식/합성하여야 한다.

둘 다 기저음/기저형이 상한선인 것으로 보이지만 실은 기저형의 인식은 단순한 음성인식이 아니라 음성언어의 인식이라 할 것이다. 왜냐하면 기저형의 인식과 사전/통사/의미에 대한 이해는 상호의존적이며, 이 모든 인식-이해는 음성언어에 대한 이해이기 때문이다. 음성언어에 대한 이해가 없는 음성분석이나 사전/통사/의미분석으로는 '음성'의 인식이 결국 불가능하거나 무의미할 것이기 때문이다.

III. 표기법

3.0 표기법의 문제

음성/음소의 표기는 학자마다 다른 상태에서 상호 호환성이 없다는 것이 문제이나, 이 보다 더 큰 문제는 표기의 원칙에 대한 개념이 없다는 것이다. 그래서 여기서는 표기법의 원칙을 위의 개념과 맞추어 설정하고, 통일된 표기법을 정립하기 위한 하나의 안을 제시하고자 한다.

3.1 표기법의 개념과 원칙

표기에서 중요한 것은 위의 3단계 즉 음성단계 음소단계 기저단계의 표기를 구별하는 것이다. 이것도 음운론의 일반 관행을 빌리면 간단하다. 즉 음성단계는 [ ]에 표시하고, 음소단계는 / /에 표시하며, 기저음 단계는 { }에 넣어서 표시하는 것이다.

{있고} ...... 기저

/있고, 잇고, 잊고, 이고, 이고, 이꼬/ 음소

[it ko] 음성

여기서 한글은 음소나 기저형을 표기하기에 가장 좋은 문자이지만 국제적인 기호가 되지 못하고 컴퓨터 활용에 불편하므로 국제적인 로마자 표기로 바꾸어야 한다. 로마자 표기는 그러나 실은 서구언어에 맞게 만들어진 것이어서 문제가 많으므로 가장 우리글에 흡사하도록 맞추어야한다. 그래서 아래와 같은 원칙이 지켜져야 한다. ([2] 참고.)

가. 기저형 및 음소적 표기는 한글 대 로마자가 반드시 1:1로 표기되도록 하여야 한다. 가령 ㄱ을 g나 k로 한다거나 ㅊ을 ch(CH)로 표기해서는 안된다.

나. 음성표기는 (한글과 상관없으므로) 소리나는 대로 로마자로 적되 모든 음성이 동일한 수의 문자로 표기되도록 하여야 한다.

다. 표기시 사용문자는 컴퓨터 자판에 있는 것으로만 하되, 최대한 음성에 가까운 것으로 하고, 다른 언어의 다른 음성을 나타내는 문자(예: f, v)는 국어에 그런 음성이 없는 한 쓰지 않는다.

3.2 표기법의 구체적 실례

위 3.1의 원칙 (가-나-다)에 따른 한 예를 보이면 아래와 같다. ([3] 참고.)

가. 음소는 1음소 1기호로 표기한다. (다만 필요시에는 1기호 H0으로 표시할 수 있다. 예: ㄱ : g 또는 g0. 아래 음성 표기 참조.)

ㄱ : g ㄲ : G ㅋ : k
ㄴ : n
ㄷ : d ㄸ : D ㅌ : t
ㄹ : r
ㅁ : m
ㅂ : b ㅃ : B ㅍ : p
ㅅ : s ㅆ : S
ㅈ : j ㅉ : J ㅊ : c
ㅎ : h
ㅇ : N (받침으로 쓰일 때만)[2]

ㅏ: a ㅓ: e ㅗ: o ㅜ: u
ㅐ: A ㅔ: E ㅚ: O ㅟ: U[3]
ㅡ: I ㅣ: i

반모음 ㅗ, ㅜ: w
반모음 ㅣ: y

ㅑ, ㅕ, ㅛ, ㅠ, ㅒ, ㅖ는 각각 ya, ye, yo, yu, yA, yE로, ㅘ, ㅝ, ㅙ, ㅞ는 각각 wa, we, wA, wE로한다. 이들 음들은 모음 음소에다 반모음 음소 y 또는 w가 앞에 첨가된 것이므로(즉, 두개의 음소이므로) 두 개의 기호로 표기되는 것이 당연하다.

[2] 소리가 없는 초성위치의 ㅇ은 표기하지 않는다. 다만 이 ㅇ은 기저형 표기에서는 필요하게 된다. 예: 바비 [babi] 밥이 {bab i}. 기저형 표기에서 필요할 때에는 영문 에에서 모양대로 '-'를 쓴다.

[3] ㅚ, ㅟ는 각각 wE, wi와 같이 두 개의 음소로 발음될 수 있으며, 이 때는 O, U 대신 wE, wi로 표기한다. ㅢ는 두개의 음소로 간주하여 Ii로 표기한다.

나. 음성은 1음성 2기호로 표기하되, 첫기호는 음소기호와 같게 하고 다음에 부호나 숫자를 붙여서 표기한다.

음성은 여기에 다 수록할 수 없으므로 몇개의 예만 제시하겠다.

가령 국어의 ㄱ (/g/)이 [g], [k], [γ]로 소리난다고 할 때, 즉 음성이 이 세가지라고 할 때, 이들은 각각 [g+], [g-], [g=]으로 표기하거나, [g1], [g2], [g3]으로 표기할 수 있다. 이 때 부호는 물론 +는 유성음, -는 무성음, =는 마찰음을 나타내는 것이다. 숫자로 표기하는 경우에도 부호의 경우나 마찬가지로, 1은 유성음, 2는 무성음, 3음 마찰음 또는 반모음 등을 나타내는 것으로 통일 시킨다. 물론 여기에서 우리는 표기법 이전에 하나의 음소가 가지는 정확한 음성의 종류를 찾아내고 그것을 정확히 정의내릴 수 있어야만 한다는 것은 말 할 필요가 없겠다.

흥미있는 점은, 이와같이 표기법을 따르면 위에 말한 음운분석 중 음성규칙에 의한 음소인식과정이 생략될 수 있다는 점이다. '음성'의 종류를 확실히 파악하고 정확히 분류하기만 하면, 그 음성이 어느 음소에 속하는가하는 것은 두자리 기호 중 첫기호가 말해주므로 별도의 음성규칙에 의한 작업이 필요없게 된다는 것이다. 여기에서 우리는 표기법과 음성분석의 상관성을 볼 수 있다.

IV. 결론

본 연구의 결론은 다음과 같이 요약될 수 있다.

가. 음성인식/합성은 '음성'의 인식/합성이 아니라 '음성언어'의 인식과 합성이다. 이를 위해 '음성'과 '음성언어'의 개념이 다음과 같이 정립되어야 한다.

음성: 인간이 귀로 듣는 언어음(speech sound)에 대한 물리적인 차원의 개념으로서, 음소(phoneme), 기저음(underlying segment)의 언어적인 차원의 개념 과 구별된다.

음성언어: 자연언어가 발화를 통해 구현된 실제언어를 말하며, 이것은 추상적인 머리속의 언어나 문자로 표현된 문자언어와 구별된다. 음성인식/합성은 음성을 통하여 음소 및 기저음을 인식하여 음성언어를 이해/합성하는 것이 그 목표이다.

나. 음성언어의 표기는 위의 개념을 바탕으로 하여야 하며 다음과 같은 원칙을 따름으로서 음성인식/합성에 적극적으로 유용하게 활용되게 만들어질 수 있다.

음소/기저음 표기: 1음소 1기호 주의로 한다.

음성표기: 1음성 2기호주의로 하되, '음소기호+부호'식으로 하여 부호를 떼면 음소표기가 되도록 한다.

이상의 결과가 활용되어 음성언어의 인식/합성 및 나아가서 기계번역에도 도움이 되고, 또 최근 관심의 대상이 되고 있는 휴대용 통역기의 개발에도 기여하게 되기를 바란다. 왜냐하면 특히 휴대용 통역기는 음성언어에 대한 개념이 없이는 발전이 불가능하기 때문이다.

참고문헌

[1] 송민석, 최현탁, 이찬도, 정 곡, 음성언어 인식의 이론적 모델, 한국정보과학회 인간과 컴퓨터 상호작용 연구회 회보 3권 1호, HCI '94 학술대회 발표논문집, pp.1-7, 1994.

[2] 정 곡, 구희산, 이찬도, 김종미, 한선희, 음성인식/합성을 위한 국어의 음성-음운론적 특성연구, 1992년도 교육부 자유공모과제 연구보고서, 1992.

[3] 정 곡, 장석진, 이기용, 김홍규, 이찬도, 김종미, 한선희, 이강혁, 송민석, 한국어 특질에 관한 연구, 자동통역 전화시스템 구현을 위한 음운 및 문법구조 연구, '93 장기기초연구과제 최종 보고서, 1993.

[4] Gillian Brown and George Yule, *Teaching the Spoken Language*, Cambridge Language Teaching Library, Cambridge University Press, 1983