컴퓨터에 의한 言語의 合成과 認識

李 明 鎬*

목 차

1. 서 론

최근의 컴퓨터 기술은 인간과 기계사이의 음성통신(voice communication)의 새로운 가능성을 제시하고 있다. 과거 수년간에 걸친 연구의 결과로 인간과 기계사이의 음성통신 기술은 인간의 능력을 확장하고 사회적인 요구를 만족시키고 생산성을 증가시키는등 새로운 통신업무의 영역을 형성하게 되었다. 지금까지 알려진 컴퓨터는 터미날을 통하여 인쇄된 형태로 교환되는 assembler와 compiler를 사용하고 있다. 그런데 만일 컴퓨터가 인간에게 일어나는 음성통신과 같은 능력을 갖게 된다면 컴퓨터의 사용가치와 편리성은 더욱 커질 것이다. 이렇게 되면 도처에서 흔히 볼수 있는 전화는 컴퓨터의 터미날로 이용될 수 있을 것이다.

일반적으로 인간과 기계사이의 통신유형은 저장정보(stored information)에 대한 컴퓨터의 음성판독(computer voice readout stored information), 음성신호에 대한 호출자의 동일인 증명(automatic verification of a caller) 및 음성명령의 자동인식(automatic recognition of spoken commands)등의 세 가지로 분류할 수 있다. 이들의 응용범위를 좀더 확장하면 음성지시로 동작되는 전화시설의 설치(voice-directed installation of telephone equipment), 음성에 의한 신용고객의 식별(authentication by voice of a credit customer) 개인의 요청에 의한 특수정보의 판독, 전화번호의 저장이나 여행예약의 자동접수등과 같은 음성제어에 의한 업무수행에 이르기까지 매우 다양하다. 그러나 이와같이 사람과 말 할 수 있고, 사람의 말을 들을 수 있는 기계의 실용성은 언어합성(speech synthesis)과 언어인식(speech recognition)을 얼마나 경제적인 방법으로 실현시키느냐에 달려 있다.

따라서 본 연구에서는 디지탈 컴퓨터 기술에 의한 언어합성과 인식에 대한 기초적인 이론을 설명하고 음성통신의 응용에 관하여 기술하여 보고자 한다.

2. 통신의 디지탈 기술응용

최근에 개발되는 대부분의 전자기기들은 다량의 정보를 저장하고 복잡한 계산을 수행할 수 있는 능력을 보유하고 있다. 디지탈 이론에 대한 지식의 확장과 디지탈 처리를 위한 경제적인 전자기기의 등장은 통신과 정보처리(information processing)업무의 다양화를 위한 새로운 가능성을 제시하였다. 마이크로 전자공학의 획기적인 발전과 디지탈 기술의 응용이 확대됨에 따라 전화사용자들은 계산, 통신 및 제어를 위하여 좀더 개선된 시스템의 혜택을 받게 되었다. 컴퓨터의 엄청난 계산능력은 가입자 개인으로 부터 수요 중계소 및 최종제어소에 이르는 전화회로망의 모든 기술수준을 발전시킬 수 있을 것으로 기대된다. 그러나 이러한 계산능력은 사용자에게 편리하고 용이해

* 正会員 : 延世大 工大 電気工学科 助教授 · 工博

야 한다. 디지탈 컴퓨터는 대단히 빠른 속도로 산술연산을 수행하며 다량의 정보를 확실하게 저장시키고 호출(access)할 수는 있지마는 사람과 통화하기는 그렇게 쉬운 일이 아니다.

한편 사람들은 자연스럽게 발생되는 언어가 통신을 위한 가장 효과적인 수단임을 알게 되었다. 따라서 가장 큰 관심사는 컴퓨터를 이용하여 인간의 언어생활에 쓰이는 자연스러운 언어로 의사를 교환할 수 있도록 하는데 있다. 간단히 말해서 컴퓨터에게 그 자신의 소리로 사람에게 말할 수 있는 입(mouth)과 인간의 요구사항을 들을 수 있는 귀(ear)의 기능을 갖게 하자는 것이다.

컴퓨터에게 그 자신의 기계소리로 말할 수 있는 능력을 갖게 한다는 것은 곧 언어합성(speech synthesis) 기술에 달려있다. 한편 컴퓨터에게 들을 수있는 능력을 갖도록 한다는 것은 발생된 명령의 자동인식(automated recognition)을 의미한다. 언어인식(speech recognition)의 대부분의 예에서 볼 수 있듯이 기계는 '누가(who) 말했는가'에 구애되지 않고 '무엇을(what) 말했는가'에 더 관련된다. 그러나 특수한 정보가 요구되는 경우에는 talker의 identity를 증명하는 것 또한 중요하다. 말할 수 있고 들을 수 있고 이해할 수 있는 능력을 가진 기계라면 사람과 상호 대화할 수 있어야 한다. 전화통신에서는 호출자가 전화번호의 자리수를 말하거나 혹은 호출된 사람의 이름을 말하는 것으로 원하는 숫자를 돌릴 수 있을 것이다. 특히 말하고 들을 수 있는 기계의 용도는 신체장애자(handicapped persons)들의 통신욕망에 직접적인 영향을 주게 되었다. 시각출력표시방식(visual output display)의 자동언어인식(automatic speech recognition)은 청각장애자(hearing handicapped persons)를 위한 보조수단으로 사용될 수 있다. 이동장애자들(motion handicapped persons)은 가정이나 직장에서 기계를 자동제어방식으로 조절할 수 있는 방법을 찾아야 할 것이다.

비슷한 경우로 언어합성(speech synthesis)은 보통의 키 보오드에 인쇄한 영문으로 부터 직접적으로 발성에 결함이 있는 사람(voice-impaired individual)에게 도움을 줄 수 있을 것이다. 광인식기(optical recognizer)를 갖춘 문장-언어합성기(text-to-speech synthesizer)는 맹인을 위한 reading machine 역할을 하게될 것이다. 이와같은 대화능력은 디지탈 기계의 능력을 인간의 요구에 응용하는 획기적인 발전을 의미한다. 만인 컴퓨터의 기능이 사람의 능력에 더욱 가까와지면 통신기능은 무엇을 시뮬레이션하기 위하여 쓰일 것인가? 그림 1은 이에 대한 예를 나타낸다. 인간에 의한 언어발생(speech generation)은 메세지의 대뇌형성(cerebral formulation) 정보의 언어코드화 과정을 거쳐서 예정된 수령인(intened recipient)에게 이해시킨다. 또한 신경 근육의 제어에 의하여 sound timber 시이퀀스를 발생하는 음성발생시스템은 주어진 언어의 음운(phonemes)으로 해석된다. 한편 다량의 유한정보를 규정하는 이산기호로 나타낸 언어코드의 시이퀀스 성분은 음성

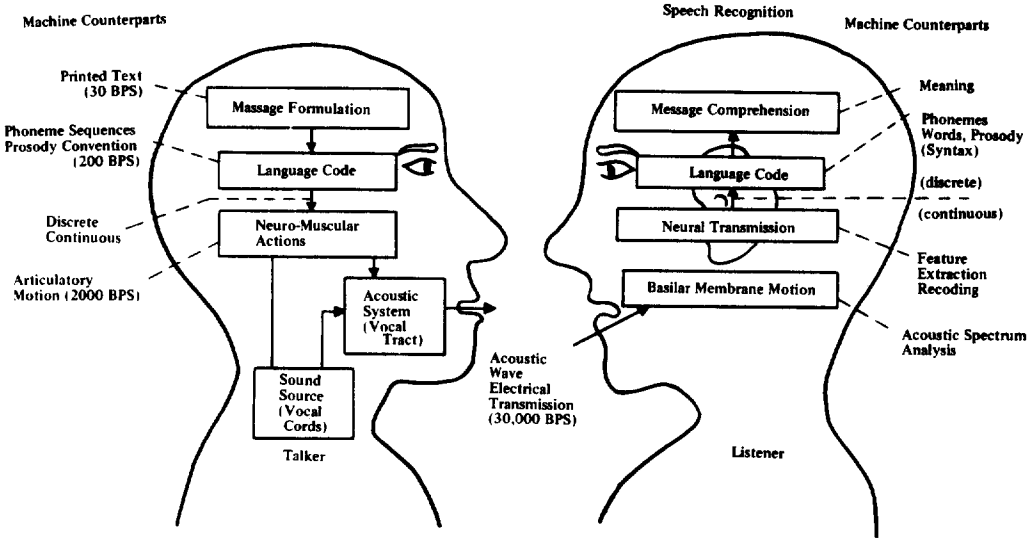

그림 1. 인간의 언어발생 및 인식모델

출력(acoustic output)을 발생시키는 변환기로 명령을 지령하게 된다. 이러한 명령은 신경에 의하여 제어되는 신경-근육운동이라고 볼 수 있다. 예를 들면 성대(vocal cord)가 특정한 주파수와 강도로 진동하게 하는 명령 혹은 "ee"모음, "ah"모음을 적당하게 발음할때의 혀, 턱 및 입을 움직이게 하는 명령이다. 한편으로 언어에 대한 사람의 인식은 청각에 의하여 수신된 음성파(acoustic wave)의 주파수 해석을 통하여 이루어진다. 이러한 주파수 해석의 결과는 상호 언어규약(language convention)에 따라 해석되고 이해되는 전기적인 신경신호(electrical neural signal)로 변환된다.

3. 컴퓨터의 언어합성(컴퓨터의 입)

talking machine을 제작하기 위하여 그림 1에 나타낸 인간의 언어발생 요소를 몇가지 의미에서 시뮬레이션해 볼 필요가 있다. human talker를 시뮬레이션하는 방법이 그림 2에 나타나 있다. 이 경우에 컴퓨터는 인간이나 기계에 의하여 메세지를 말하기 적당한 형태로 개념화시켜야 한다. 기계 메세지(machine message)는 대개는 인쇄된 영문(English text)으로 되어 있어야 하며 또한 이 문장은 talker의 음성과 똑같은 형태로 전화선을 따라서 전송될 수 있는 알기 쉬운 기계음성신호로 변환되어야 한다. 물론 문제가 되는 것은 어떠한 기술이 기계로 하여금 소리를 발생할 수 있도록 하느냐 하는 것이다. 가장 직접적인 방법은 사람이 말한 단어(words)와 귀절(phrase)에 대하여 디지탈적으로 저장된 파형을 활용하는 것이다.

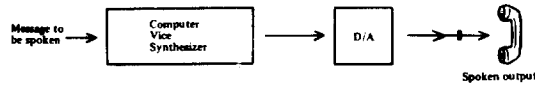

그림 2. 언어발생기계의 요소

3.1 저장된 파형에 의한 컴퓨터의 음성응답

컴퓨터의 음성응답의 가장 간단한 형태는 인간이 말한 단어(words)와 귀절(phrases) 즉 그의 음성파형(acoustic waveforms)이 간단히 디지트화 되어 컴퓨터의 기억장치에 기억되어 있는 어휘(vocabulary)를 사용하는 것이다. 보통의 데이타 요약(data compression)의 형태는 디지탈형으로 부호화 하는데 이용된다. 이 목적에 적합한 encoding technique은 adoptive differential pulse-code modulation(ADPCM)이다.

이 방법은 매우 간단하게 32~24 K bits/s의 속도로 디지탈화 할 수 있다. 메세지를 발생시키기 위하여 인쇄된 문장입력은 이에 해당되는 단어(words)와 귀절(phrases)에 대한 파형을 순서대로 호출시킨다. 분리된 파형성분은 컴퓨터 프로그램에 의하여 서로 연결되거나 또는 성분파형사이의 silent interval의 주기를 조정한다. 연결된 파형은 보통의 D/A변환기에 의하여 아나로그 형태로 변환된다. 이러한 경우에 합성 혹은 더 정확하게는 조립절차가 간단하고 알맞은 크기의 단일 컴퓨터는 동시에 공통어휘(common vocabulary)로부터 수개의 서로 다른 메세지를 발생시킬 수 있다. 이에 관한 기술이 그림 3에 나타나 있다. 이 과정은 한 가지 중요한 이점이 있는 반면에

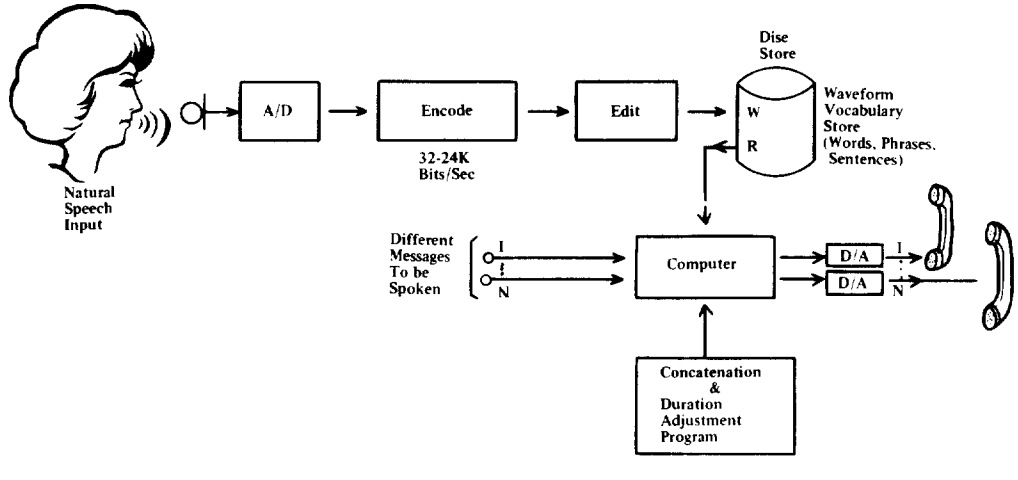

그림 3. 컴퓨터 음성의 응답시스템

몇가지 단점을 가지고 있다. 기계음성의 quality 는 정상인의 음질에 접근한다. 그러나 상대적으로 매우 높은 digital rate 때문에 파형성분은 큰 기억용량을 필요로 하며 또한 파형성분은 발생된 메세지의 음율적요인(prosodic factor)을 제어하는데 유연성을 갖지 못한다. 단어와 귀절은 그들 사이의 silent interval의 조정을 제외하고는 기록된대로 정확하게 사용되어져야 한다. 단어와 귀절은 절편간의 간격을 smooth하게 한다거나 또는 문맥에 꼭 맞도록 소리의 억양을 다양하게 변화시킬 수는 없다. 이와같은 이유는 컴퓨터에 의하여 메세지를 구성할 수 있는 어휘성분의 수가 기계의 기억용량에 의하여 제한을 받기 때문이다. 그러므로 컴퓨터의 다양한 기능과 메세지의 처리범위는 제한을 받게 된다. 따라서 이러한 기술은 고도의 정확성과 특성을 요구하는 비문맥 메세지(noncontextual message) 즉 예를 들면 전화번호의 음성판독(voice readout)에 아주 적절하다. 기계음성의 폭넓은 다양성을 응용하기 위하여 그림 3에 나타난 기술이 이용된다. 예를 들면 전화요금징수국에 수령된 전화사용료의 자동음성고시와 그 분야의 전문가에 의하여 수집된 컴퓨터의 음성지시등을 들 수 있다.

그러나 실제로 인간처럼 대화를 하기 위해서는 컴퓨터가 적절한 메세지를 다양하게 창조해 내거나 혹은 종합할 수 있어야 한다. 그리고 이것은 최소의 처리부담과 디지탈 기억량을 사용하여 우수한 quality 의 언어를 발생시킬 수 있어야 한다. 언어발생의 이러한 다양성과 경제성을 고려하여 기계는 인간언어발생의 아주 상세한 것 까지도 유사화시켜야 한다. 이렇게 하여 발생된 언어가 합성언어(synthetic speech)이다.

3.2 언어의 변수적 합성

인간의 언어처리에 대한 보다 상세한 시뮬레이션 과정이 그림 4에 나타나 있다. 그림 4에서 볼 수 있듯이 말로 표현될 메세지의 인쇄된 문장은 먼저 선택된 언어의 습관에 따라 분리성 어음(speech sound)을 나타내는 연속적인 기호 즉 발성을 위한 운율법(phonemes)으로 변환된다. 이러한 현상은 문자-음(letter-to-sound)변환을 위한 프로그램 알고리즘이나 저장된 발음사전에서 각 문장의 단어를 찾음으로써 수행된다. 이러한 기술혼합의 대표적인 것으로서 컴퓨터 1개의 실행문이 1,000개의 알고리즘법칙과

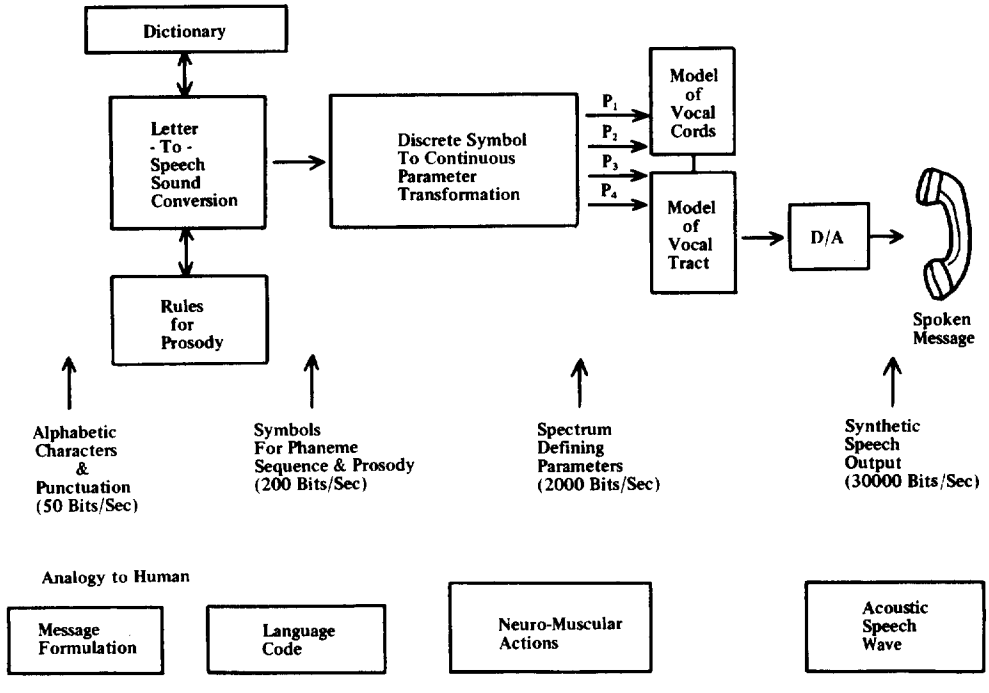

그림 4. 연속변수에 의한 언어합성 과정

1,000개의 단어사전(알고리즘 법칙으로는 처리될 수 없는 발음을 갖는)을 이용할 수 있다. 반면에 다른 한 시스템은 40개의 알고리즘법칙과 3,000개의 단어사전을 이용한다. 이러한 시스템들은 문맥상의 성분을 종합할 수 있는 비교 수행방법을 제시하여 준다. 또한 음운을 발생시키는 과정에서 컴퓨터는 음운의 발음을 포함하는 인쇄문장의 구문(syntax)을 고찰한다. 그런 후 다시 알고리즘법칙과 단어사전에 의하여 메세지의 운율 즉 음소 사이퀀스에 대한 음의 고저 강도 및 주기변화에 대한 기호를 발생시킨다.

이러한 변화는 절대적으로 문맥에 따라 좌우된다. 인간의 lanquage code와 유사한 discrete symbol의 sequence는 계속적으로 변하는 일련의 변수로 변형된다. 이 변수들은 시간에 대하여 비교적 천천히 그리고 smooth하게 변화하고 출력으로 요구되는 음성언어신호(acoustic speech signal)의 주파수 spectrum에 대한 기본적인 표현으로 변화된다. smooth하게 변화하는 변수의 이와같은 특성은 발성근육을 움직이도록 하는 인간의 신경명령 발생과 비슷하다. 이 발성근육은 성대(vocal cord)의 움직임과 발성관(vocal tract)의 모양에 따라 smooth하게 변화한다. 이들 변화는 음성출력으로 연결되는 어음(speech sound)를 발생한다. 이와 비슷한 방법으로 컴퓨터의 spectrum으로 정의되는 변수는 인간의 발성시스템의 음성특성을 모델화한 전기합성기를 동작시키는데 이용된다.

3.3 합성기의 구성요소

언어합성시스템의 최종요소는 인간의 발성관(vocal tract)과 성대(vocal cord)에 관한 개념적모델이 된다. 이 모델은 컴퓨터의 프로그램에서 software로 실행되거나 특별한 디지탈 hardware의 외부요소로 쓰인다. 또한 이 모델은 일련의 언어변수를 입력으로 받아들이며 이것으로 부터 합성음성파형의 디지탈 샘플을 나타내는 binary sequence를 발생시킨다. 디지탈에서 아나로그형으로 변환된 신호는 listner에게 정상적인 아나로그 전송에 적합하다. 합성기 모델은 여러가지 형태가 있으나 널리 알려진 방법으로는 소위 선형예측계수(linear prediction coefficient)방법이 있으며 다른 하나는 모음중에 포함되어 각 모음에 특유한 음색을 주는 입안의 공명음으로써 혀의 위치, 상태와 입술의 위치로 결정되는 "formant"에 기초를 둔 방법이다. hardware의 견지에서 볼 때 합성기는 현재 이용 가능한 microprocessor component로도 편리하게 수행시킬 수가 있다.

3.4 이산기호의 연속변수 변환

이산음소(discrete phoneme) sequence를 언어 합성기를 제어하는데 필요한 연속변수로 변환시키는데는 두가지가 방법이 있다. 그림 5는 이 관계를 나타낸다. 첫째는 현재 널리 사용되고 있는데 인간이 말한 문장, 단어 및 음절로부터 측정되어 기억된 파라

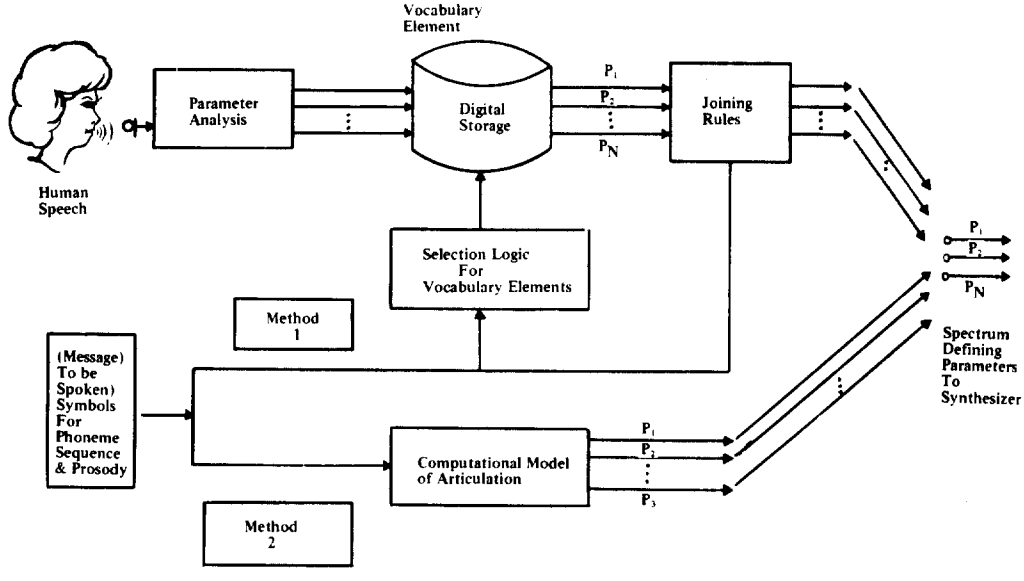

그림 5. speech sound에 대한 이산기호를 변환시키는 두가지 방법

미터 어휘를 이용하는 것이다. 이들 측접 파라미터는 단어(예 : telephene), 음절(예 : phone) 혹은 음절의 일부(예 : pho)와 일치하는 길이의 요소어휘로 이루어진다. 따라서 컴퓨터는 이러한 파라미터를 기억시키기 위하여 500~2,000 bits/s 정도의 기억용량을 필요로 한다. 이것은 본래의 파형을 기억시키는데 필요한 digital rate(보통 24,000~64,000 bits/s) 보다는 훨씬 작은 용량이다. 문자-음(letter-to-sound) 변환에 의하여 발생되는 음소(phoneme) sequence는 어떤 어휘요소가 file로부터 또는 무슨 명령으로 호출되어야 하는가를 지시한다. 일단 어휘요소가 호출되면 이에 대응되는 파라미터들은 smooth하게 연결되어야 한다. 여기서 의미하는 것은 어음(speech sound)이 유성음(voiced), 무성음(unvoiced), 혹은 양쪽 모두 가능한 것인지에 대한 규정을 말한다. 또한 컴퓨터는 수반되는 운율(prosody) 즉 목소리의 고저, 강도 및 문맥에 적절한 음의 지속시간(sound duration)의 변화상태를 결정해야 한다. 예를 들면 언어가 선언적이거나 의문적 혹은 감탄적인 발성으로 감지될지의 여부를 결정하는 것이 바로 운율이다. 음소 sequence의 prosody marker로부터 컴퓨터는 알고리즘법칙에 의하여 음의 고저와 소리의 상도를 제어할 연속 파라미터를 발생하며 이 선택된 어휘요소의 지속시간을 수정하게 된다. 기억된 어휘요소가 word-length 요소로 구성되는 특수한 문맥에서 어음(speech sound)으로의 변환은 매우 간단하고 직접적이다. 그러나 규정된 문맥에 대한 운율의 계산은 여전히 필요하다.

반면에 discrete phoneme sequence를 연속파라미터로 변환시키는 또 다른 방법은 저장된 사람의 음성 혼적이 전혀 없도록 사용하는 것이다. 더구나 이 방법은 알고리즘법칙에 의하여 합성언어 파라미터를 완벽하게 계산할 수 있게 한다. 따라서 예상되는 바와 같이 computational approach는 더욱 어려우며 컴퓨터의 처리 능력이 매우 커야 한다. 가장 필수적인 계산은 인간의 유절음(articulation)의 동특성에 관한 수학적인 설명 즉 입, 턱, 혀, 연구개(velum) 및 후두등의 움직임을 지배하는 물리적 법칙을 정량적으로 설명하는 것이다. 그리고 어떻게 이들 유절음이 어음(speech sound)를 발생시키는데 이용되고 있는가에 대한 수학적인 설명이다. 이에 대한 설명은 현재의 연구과제이며 정량적으로 상세하고 완전해야 한다.

이 두번째의 기술에서 초기에 문자-음(letter-to-sound) 변환에 의하여 발생된 이산음소 sequence는 유절음의 프로그램모델의 운동결과는 컴퓨터가 합성기에 필요한 연속 spectrum 파라미터를 계산하도록 한다.

3.5 언어합성기의 성능

언어합성의 결과는 주로 다음의 세 가지 측정의 표준에 의하여 호출된다. 즉

① 기계음성의 질(quality)
② 메세지의 융통성(versatility)
③ 음성응답을 수행하는데 필요한 processor complexity 등이다.

그림 6은 이들의 관계를 나타낸다.

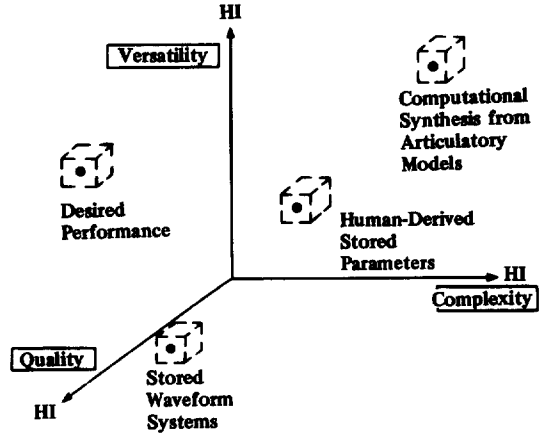

그림 6. 신호의질, 메세지의 다양성 및 복잡성으로 본 언어합성시스템의 평가

performance values에 영향을 주는 주 요인은 저장어휘의 형태와 언어합성기의 실행에 기인한다. 이미 언급한 것처럼 언어파형의 저장기술은 기계음성의 높은 질, 메세지의 융통성 및 문맥상의 어색함이 최소이고 저장파형의 어휘가 크지 않은 경우에 매우 유용하다. 자동광고와 차단 메세지는 이런 방법을 이용하여 효과적으로 만들어진다. 그러나 기계음성이 보다 큰 융통성이나 메세지의 다양성 및 자발성을 갖도록 요구되는 응용에 대해서는 파라미터 기술을 이용하는 합성기술이 매력적인 선택이 될 것이다. 모든 합성기술의 기본은 인쇄된 문장을 말하는 메세지로 변환시키는 능력이다. 현시점에서 기대할 수 있는 가장 좋은 결과는 인간의 언어로부터 측정되어진 spectrum 파라미터의 저장어휘를 사용하여 이룰 수가 있다. vocabulary entry는 전체 절, 단어, 음절 혹은 음절의 단편을 나타낼 수 있다. 컴퓨터는 저장

어휘로부터 연속적으로 vocabulary entry를 선택해야 되고 또한 성분 모두를 smooth하고 자연스럽게 연결해야 한다. 그러나 문맥상의 영향때문에 이 결합절차는 성분이 클 때에만 보다 납득할 만한 결과를 가져오는 경향이 있다. 즉 단어의 길이성분을 나타내는 파라미터들은 음절의 길이성분보다 더 효과적으로 결합할 수 있다. 아주 긴 성분들은 결합값의 빈번한 계산을 요구하지는 않는다. 완전히 자연적인 "within-word" 공통교합(coarticulation) 상태로 남게된다. 그러나 음소의 길이보다는 단어가 더 많기때문에 메세지의 융통성 즉 주어진 디지탈 기억량에 대하여 합성될 수 있는 메세지의 융통성의 궁극적은 목적은 흔적없이 저장되는 인간의 언어를 이용하는 기본원리로부터 언어의 순수한 계산합성에 의하여 약속된다. 처리의 복잡성은 이 방법에서는 더욱 크며 현 단계에서는 소리의 질이 대표적으로 자동화될 경향이 있다. 대표적인 예로 미니컴퓨터나 마이크로컴퓨터는 이와같은 처리에 기여하고 있다. 그러나 이미 장님을 위한 문장판독기가 이 원리로 제작되고 있다. 디지탈 기억장치의 가격이 계속감소됨으로써 단어와 음절의 크기에 관련되는 수천개의 변수어휘를 저장하는 언어합성시스템은 경제성이 더욱 증대될것이다. 컴퓨터가 인간처럼 말할 수 있도록 하려면 "입" 뿐만 아니라 "귀"도 필요하다.

4. 컴퓨터의 언어인식

인간의 언어발생(speech generation)과 언어인식(speech recognition)에 관한 그림 1에 나타난 블럭선도는 내이(inner ear)의 기저세포막에 의하여 이루어지는 peripheral frequency analysis는 spectral information의 신경변환(neural conversion)과 부수적으로 일어나는 부호화 및 언어의 음성패턴의 이해로 이루어진다는 사실을 암시해 준다. 인간이 언어발생에서와 마찬가지로 청력의 peripheral processes 이해는 비교적 좋은편이나 언어인식의 경우는 전혀 알려져 있지 않다. 이러한 상황이 언어인식을 위한 현재와 같은 시스템에 관련된다. 언어인식 실행의 차원(dimension)은 언어의 합성에서와 같이 적어도 3차원이다. 이러한 이유는 시스템이 어떤 talker를 조절할 수 있는지 혹은 그 시스템에 대한 훈련을 받았는지의 여부 혹은 입력언어가 연속적으로 연결되

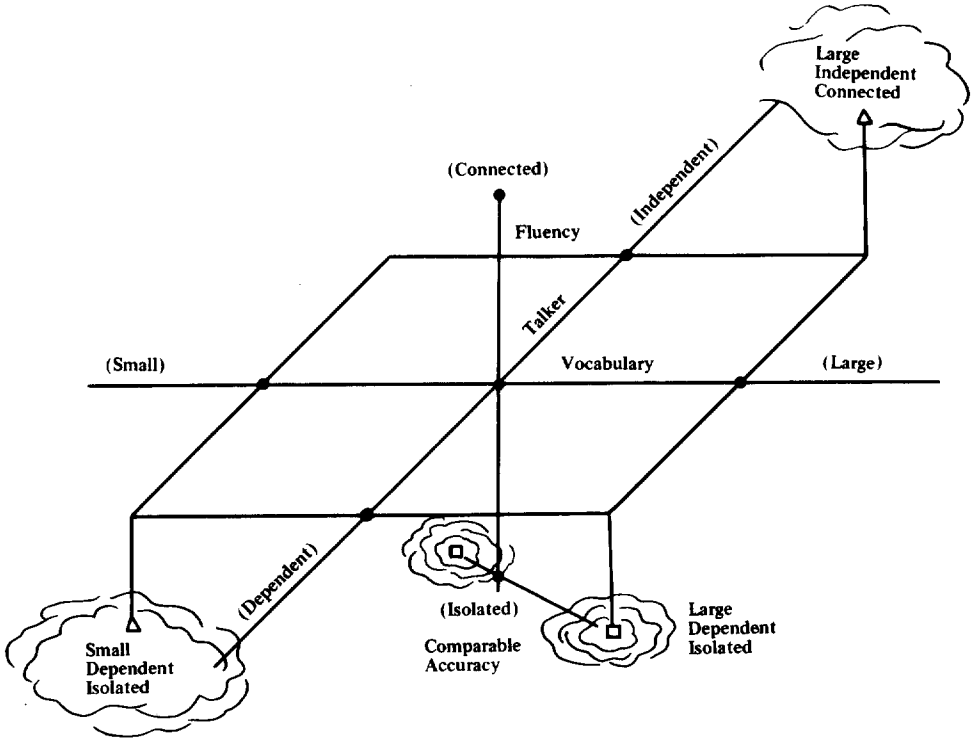

그림 7. 자동언어인식의 3차원

는 발성이 될 수 있는지 그렇지 않으면 개인의 명령이 분리되어야 하는가에 따라 주로 명령어휘의 크기에 맞추어져야 한다. 그림 7은 이때의 관계를 나타낸다.

4.1 분리된 단어인식

인간의 인식에 대한 peripheral recognition의 이해를 정리해 보면 분리된 각 단어의 자동인식 시스템은 그림 8에 보여진바와 spectral pattern comparator 형태를 취하고 있다. 여기서 인간은 기계에 단일

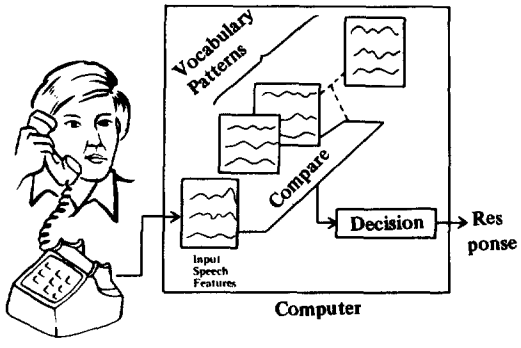

그림 8. 저장된 단어 형판에 의한 언어인식시스템의 기능적요소

발성으로 명령해야 한다. frequency spectrum의 시간적인 특징패턴은 입력되는 인간의 발성으로 측정되며 이들의 특징은 인간이 창안해낸 저장패턴의 어휘와 비교되며 각 단일발성 마다 기계에 받아들여진다. 만일 시스템이 talker-dependent 라면 이들 어휘패턴은 정해진 talker에게 미리 준비되어져야 한다. 독특한 용도의 특징으로서는 앞서 언어의 합성에서 언급한 소위 LPC(linear-prediction-coefficient) 파라미터이다.

4.2 분리된 단어인식에서 복잡성과 talker-independence 중의 trades

한개 이상의 저장된 패턴은 일정한 talker가 한단어를 서로 다른 방법으로 발음하거나 혹은 서로 다른 talker가 한 단어를 여러가지 방법으로 발음하도록하는 일정한 어휘항목으로 이용된다. 어휘항목에 대한 multiple pattern은 기계의 기억장소를 소비하게 되며 추가된 패턴에 대하여 processor가 "closeness of fit"거리측정을 계산할 시간을 필요로하게 된다. 그러나 만일 multiple pattern이 많은 수의 talker를 정확하게 특성화 시킬 수 있다면 자동인식기는 실제로 talker-independent 로 구성할 수 있다. 영어로 말하는 talker가 talker-independent voice dialing에 기본이 되는 디지트를 발생하는 유용한 표현방식을 12개의 서로 다른 패턴까지 나타낼 수 있음을 bell 연구소의 연구를 통하여 증명하였다. 「cost 니 trade는 각 어휘항목에 대한 multiple pattern을 찾거나 저장하기에 필요한 계산속도와 부가되는 기억장치에 좌우된다. speaker-dependent로부터 speaker-independent 인식까지가는 12가지 인자는 디지트라기보다는 분리된 단어 어휘로 표시되어야 한다. 예를들면 사람이름, 도시, 시간, 날짜등이다. 많은 talker에게 특성을 부여하는 multiple pattern은 많은 talker 대하여 행하여진 측정으로부터 결정된다. 따라서 통계상의 집단분석은 이때 많은 인구를 거의 모두특성화 할 수 있는 multiple pattern을 결정한다.

4.3 분리된 단어인식을 위한 equipment complexity와 cost

분리된 단어를 인식하는데 현재 이용되는 처리과정은 그림 9에 상세히 나타나 있다. 그림 9에 나타난 숫자들은 dynamic time warping' LPC 특성해석 그

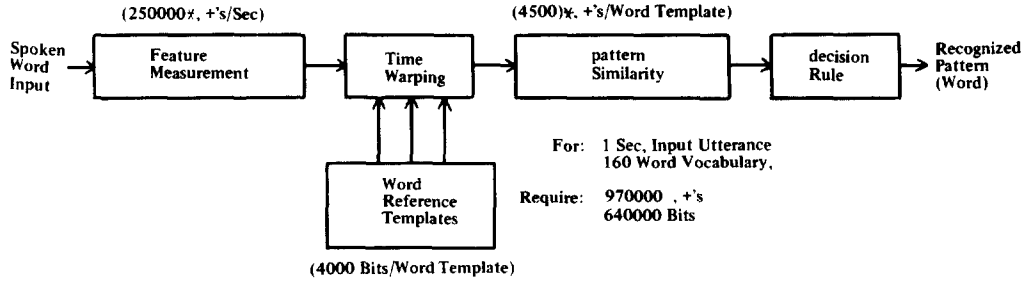

그림 9. 분리된 단어인식시스템의 처리과정

리고 160개의 저장된 단어패턴을 가지는 인지기를 만족시키기에 적당한 정도의 복잡성을 나타낸다. 그림 9에 나타난 것처럼 특징의 계산은 입력음성에 대하여 초당 25만의 정수곱이 필요하다. 패턴의 비교는 형판(template)마다 약 4,500 multiply-adds이 필요하며 각 형판마다 약 4,000bit의 디지탈용량이 필요하다. 160개의 어휘패턴에 대한 디지탈 저장장치는 단 한사람에 의하여 행하여진 많은 단어에 이용되거나 혹은 말하는 사람과 무관한 20개의 단어어휘를 위해 다중형판을 디지탈저장장치를 사용한다. 이와같은 실험시스템이 연구실에서 이루어지고 있다. 이와 같은 실험시스템의 응용으로 비행계획에 대한 정보, 발성된 이름에 의한 저장 다이얼링, 발성된 디지트에 의한 음성호출, 자동전화부를 수정하거나 보조하기 위하여 음성의 상호작용을 이용하는 것 등이다.

4.4 수화자의 증명

정보의 호출이나 조사과정에서 특정한 데이타가 문제가 될 수 있다. 기계는 이러한 것을 감지하여야 하며 또한 이 제한된 데이타를 사용할 권한을 가진사람의 고유성을 어느 정도 입증할 수 있어야 한다. 이 경우 "무엇을 말했는가"를 결정하는 것 뿐만 아니라 "누가 말했는가"가 더욱 중요하다. 증명의 요소와 인식의 요소는 서로 비슷하다. 이 기술은 그림10에 나타나 있다. 그러나 최근에는 이 기계는 identity claim에 따라 개인에게 규정된 cord phrase를 말하도록 요구된다. 이것은 절의 특징을 측정할 수 있으며 그 개인에 대하여 file상의 패턴과 이 결과를 비교할 수 있다. 만일 fit이 알맞으면 이 identity claim은 지속되고 요구된 상호작용이 수행된다. 근래에는 talker-independent word인식에 사용되는 multiple recognition templates에 따른 말하는 사람의

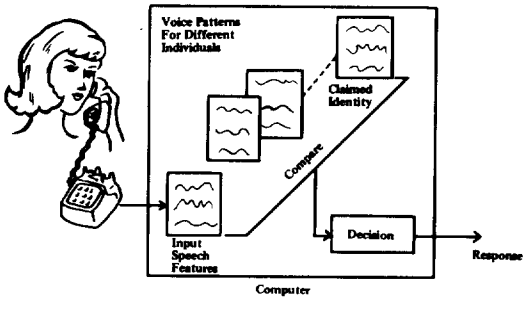

그림 10. Talker's claimed identity의 자동증명에서의 기능적 동작

고유성의 분류가 가능해졌다. 연속되는 단어의 자동인식을 하는데 있어서 개인은 multiple template 어휘를 통해 특수한 패턴을 조합하도록 한다. 이 연속적인 패턴은 말하는 자의 개인적인 특성을 동시에 나타난다. 예를들면 음성에 의하여 credit charge를 만들때 그 사람의 credit card의 수에 대응하는 연속적인 디지트를 말하도록 한다. 이 기계는 말하여진 digit sequence를 인식하여 정확한 계산치를 결정하고 동시에 다중 어휘 template에 대한 거리측정의분포를 감지하여 이 말하는 사람이 account number의 인정된 사용자인가를 확인한다. 근래에는 10진 디지트의 spoken sequence를 사용하여 증명에 있어서 90%이상의 정확도를 얻게되었다.

5. 언어합성과 언어인식의 상호영향

이상의 고찰로부터 합성과 인식의 차원을 서술하는 공통성과 유사성을 확인하였다. 합성과 인식의 통일된 원리는 만일 지식이 합성언어발생을 완벽하게 성취시킬 수 있다면 자동인식의 문제는 동시에 해결된다는 사실을 뒷받침 한다. 이러한 견해는 이상적인 언어합성에 의하여 개인별 talker의 차이, 강도, 방언 및 기호의 이해를 포괄할 수 있다면 거의 정확하다고 볼 수 있다. 보통의 합성에서는 합성기가 한 사람이나 혹은 그 이상의 여러사람의 음을 합성할 수 있는 것만으로도 만족하게 될 것이다. 이와는 대조적으로 인식기는 광범위한 음성의 차이를 극복할 수 있다면 이로 인하여 speaker-independent 인식에 대한 음성차이는 무시될 수 있다. 그러나 이러한 기분으로 독특하고 이산음성심볼에 의하여 합성기를 구동하고 주어진 언어형태를 이상적으로 발생시킬 수 있다면 동일한 합성기는 임의의 언어발성을 적절히 조화시키거나 모방하기 위하여 사용될 수 있을 것이다. 모방으로 얻어지는 이산제어기호의 sequence는 임의의 언어의 복사물과 동일하게 될 것이다. 이러한 방향에 대한 기본적인 연구 노력은 근래에 시작 되었으나 이 야심적인 목표에 대한 진전은 앞으로 기대해 볼 일이다.

참 고 문 헌

[1] L.R. Bahl and F. Jelinek; "Decoding for channels with insertions, deletions and substitutions with applications to speech recognition," IEEE Trans. Inform. Theory, vol. IT-21, pp. 404-411, 1975.

[2] C.H. Coker, N. Umeda, and C.P. Brown; "Automatic synthesis from ordinary English test," IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 293–298, 1973.

[3] J.L. Flanagan; "Computers that talk and listen: Man-machine communication by voice," Proc. IEEE, vol. 64, pp. 405–415, Apr. 1976.

[4] J.L. Flanagan, C.H. Coker, L.R. Rabiner, R.W. Schafer, and N. Umeda," Synthetic voices for computers," IEEE Spectrum, vol. 7, pp. 22–45, Jan. 1970.

[5] J.L. Flanagan, M.R. Schroeder, B.S. Atal, R.E. Crochiere, N.S. Jayant, and J.M. Tribolet; "Speech coding," IEEE Trans. Commun., vol. COM-27, pp. 710–737, Apr. 1979.

[6] F. Itakura; "Minimum prediction residual principle applied to speech recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp. 67–72, 1975.

[7] F. Jelinek; "Continuous speech recognition by statistical methods," Proc. IEEE, vol. 64, pp. 532–556, Apr. 1976.

[8] S. Furui; "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp. 254–272, 1981.