음성에 의한 Man-Machine Communication 기술의 현황

殷 鍾 官
(正 會 員)
韓國科學技術院 電氣·電子工學科 教授

요 약

본 논문에서는 음성에 의한 man-machine communication의 핵심기술인 음성인식 및 합성의 전반적인 기술에 관하여 그 현황을 알아본다. 먼저 음성인식에서 해결되어야 할 문제점들을 고찰하고 격리단어 인식, 연결단어 인식, 그리고 연속언어 인식의 기술 현황을 기술한다. 격리단어 인식에서는 pattern matching 방법에서 사용되는 입력어휘의 특징 추출, reference와의 유사도 측정, 유사도 측정 결과에 의한 인식결정에 관해서 논한다. 연결단어 및 연속언어 인식에서는 현재 연구가 되고 있는 "bottom-up approach"와 "top-down approach"에 관해서 설명하고 이들 방법의 어려운 점들을 고찰한다. 다음 음성 합성에서는 기존의 여러가지 합성 방식을 검토하고 이들의 장단점을 기술한다. 마지막으로 한 예로서 한국어 text-to-speech 변환 시스템에 관하여 기술한다.

I. 서 론

음성은 인간의 가장 기본적이며, 친밀할 뿐 아니라 가장 오래된 정보전달의 수단으로써, 말의 내용뿐 아니라 말하는 사람의 나이, 성별, 습관 등 많은 정보를 포함하고 있다. 이와 같은 음성을 인간과 기계와의 정보 전달의 수단으로 사용한다면, 많은 사람들이 computer 등 현대 문명의 이기를 쉽게 이용할 수 있을 뿐 아니라, 인간의 생활에도 많은 편리함을 제공 받을 수 있을 것이다. 이러한 이유 때문에 많은 전자공학, computer 공학자와 언어 학자들은 음성을 인간과 기계사이의 정보전달의 수단으로 만들기 위하여 음성 인식과 합성 기술에 관한 광범위한 연구를 현재 활발히 하고 있다.

음성에 의한 man-machine communication기술을 크게 분류하면, 먼저 computer가 인간이 발음한 음성을 정확히 알아듣기 위한 음성인식(speech recoginition) 기술과 입력된 text를 유창하며, 명확하고, 자연스러운 합성음으로 만들어 주는 음성합성 (speech synthesis) 기술로 분류할 수 있다.

먼저 음성인식 분야를 살펴보면 지난 10여 년간 격리단어와 연결단어의 인식에 있어서는 많은 발전이 이루어져서 현재는 미국, 일본 등지에서 몇 종의 상업용 제품이 나와 있다. 이들 인식시스템들은 대부분 격리단어 인식 시스템으로 특정 화자의 발음에 대하여 잡음이 섞인 환경 아래서에도 95%이상의 인식율을 갖는다. 단어 인식시스템의 성능이 향상됨에 따라서 그 응용분야도 점점 더 복잡해지고 다양해지고 있다. 예를 들면 각종 자료의 수정 및 관리, 항공안내 및 예약, 구술하는 문장의 자동인식, 음성으로 타자기를 구동 시키거나 전화를 거는 것등으로 그 응용 예는 광범위하다.

음성 인식은 세가지 분야로 나뉘어져 연구되고 있다. 첫째 분야는 격리단어의 인식으로서 여기서는 명료하게 구분하여 발음하는 단어를 인식한다. 지금까지 여러 종류의 시스템들이 실용화 되었으나, 격리단어 인식을 위해서는 각 단어 사이를 구분하여 발음해야 하므로 실제 사용에는 어려운 점이 있다. 두번째 분야는 연결단어 인식으로서 이 인식 시스템은 좀더 자연스럽게 발음하는 연결단어를 다루므로 격리단어 인식시스템의 단점을 어느 정도 보완할 수 있다. 세번째 분야는 연속음성 인식으로서 이 경우는 자연스럽게 발음하는 문장 및 구절을 인식하는 것이다. 음성 인식분야 중 연속음성 인식은 가장 기술적으로 어려우므로 실용화 되기까지는 오랜 시

간이 걸릴 것이 예측된다.

한편 음성합성은 제한된 수의 단어를 조합해서 일정한 문장을 자연음으로 만들어 내는 비교적 간단한 합성기술부터 시작해서 음성의 기본이 되는 음소, 음절 또는 반음절 등을 가지고 어휘의 수에 제한 없이 어떠한 단어 또는 문장도 합성할 수 있는 고도의 기술을 요하는 무제한 음성합성기술등 그 연구 분야가 음성 인식과 같이 방대하다. 음성을 합성하는 방법은 시간영역에서의 합성방법, 음성발생 model을 사용한 합성방법, 인간의 발성기관의 일부인 성도(vocal tract)를 simulation 하여 합성하는 방법등을 들 수 있다.

음성합성 응용분야는 매우 광범위한 데 그중 몇가지 예를 들면 다음과 같다.

- 말하는 각종 computer 또는 전자 장비
- 시각 장애자를 위한 reading machine
- 성대 장애자를 위한 talking machine
- 음성에 의한 경고 system
- Teaching machine 및 training machine
- 장난감 등

이러한 응용분야 이외에도 한국어 음성합성 system은 일기예보, 금융잔고, 증권동향 등 정보가 빠르게 변화하는 곳에서 이를 이용함으로써 전화선을 통하여 음성으로 사용자들에게 여러가지 정보를 service 할 수 있으며, 자동차를 운전 중인 경우와 같이 음성으로 정보를 얻어야 하는 경우에 유용하게 사용된다.

본 논고에서는 음성에 의한 man-machine communication의 기본 기술인 음성 인식과 합성기술의 연구동향과 현황, 그리고 문제점들을 검토하고자 한다. 서론에 이어 제Ⅱ장에서는 격리단어, 연결단어, 연속음성의 인식기술에 관하여 기술한다. 제Ⅲ장에서는 음성합성의 전반적인 기술과 한국어 음성합성에서 필요한 여러가지 중요한 점들을 검토한다. 마지막으로 제Ⅳ장에서 결론을 맺는다.

Ⅱ. 음성 인식 기술

1. 격리단어 인식

격리단어 인식은 1960년대 이래로 여러 방향에 걸쳐서 연구하여 왔으나 1970년에 Velichko와 Zagoruyko가 dynamic programming의 기법을 음성 인식에 도입하기 전까지는 큰 진전이 없었다. 그 후에는 발음속도의 변화를 비선형적 방법으로 보정하기 위하여 dynamic programming 방법을 사용하는 연구가 많이 수행되었다.[1]

격리단어 인식에서는 feature vector들이 미지의 입력 음성으로부터 매 10-30ms 간격으로 얻어지면, training data로부터 구하여진 이미 저장되어 있는 reference sequence들과 feature vector들의 sequence를 비교하여 가장 근사하게 일치하는 것을 찾아낸다. 이 비교 과정에서 가장 중요한 단계는 각각의 reference sequence와 입력 sequence의 시간축을 조정하는 것이다. 이 alignment 방법중 가장 간단한 것은 입력 sequence의 양 끝점과 reference sequence의 양 끝점을 일치시키고 사이 사이에 입력 데이타를 탈락시키거나 채워 넣는 방법이다. 그러나, 발음속도의 변화는 비선형적이므로 좀더 복잡한 방법을 사용하여야 한다. 즉, 입력 sequence와 reference sequence가 가장 근사하게 일치할 수 있도록 dynamic programming 방법으로 선택한 비선형적 warping 함수에 의해서 입력 데이타의 시간축을 변환시키는 방법이다. 이 방법을 dynamic time warping(DTW) 기법이라 하며 격리 및 연결단어 인식에서 좋은 효과를 보고 있다.

한편 time sequence의 정보가 일반적으로 생각하는 것 보다는 크게 영향을 미치지 않는다는 연구결과를 토대로 vector quantization(VQ)에 의한 음성 인식방법이 최근에 연구되었다.[2] 이 방법을 적용하려면 어휘중에 있는 각각의 단어에 대응하는 single codebook이나 multiple codebook을 training sequence로부터 설계해야 한다. 각 codebook은 적절한 크기를 갖는 대표적인 feature vector들로 구성함으로써 입력 feature vector와 가장 유사한 codeword를 찾는데 사용된다. 만약 입력 sequence의 feature vector들이 특정한 codebook에 대하여 가장 적은 accumulated distance를 갖는 것으로 나타내면 이 codebook에 해당하는 단어가 이 시스템이 인식한 단어가 된다. 이러한 방식에는 time sequence에 의한 정보는 전적으로 무시할 수 있다. 그러나, 이 방법은 양끝이 유사한 단어들에 대해서는 잘못된 결과를 보일 수 있으므로, 이러한 문제점을 보완하기 위해서 multi-section VQ나 matrix quantization (MQ)등의 개선된 algorithm이 제안되었다.[3]

이상에서 서술한 dynamic programming 기법이나 VQ방법과는 전혀 다른 통계적 접근 방식이 격리단어 인식에 적용되기도 한다. 한가지 예는 hidden Markov model(HMM) 방법인데 여기서는 모든 단어는 Markov chain의 불연속 확률 함수로 modeling된

다.[4] Markov chain은 state transition network라고도 하며 이 network은 state transition의 확률, 출력 발생의 확률, 그리고 state의 수 등 많은 parameter가 있어야 하는데 이러한 parameter들은 많은 양의 training data로부터 추정한다. 이 방식의 음성 인식 시스템에서는 입력 feature vector가 출력 feature vector를 발생시키기 위해서 VQ 방식으로 입력된다. 그러면 feature vector의 sequence로부터 시스템은 Viterbi algorithm을 사용하여 어떤 reference 단어가 가장 유사한 지 결정한다. 따라서, 이 방식에 의한 인식은 DTW에 의한 것보다 더욱 효과적으로 이루어질 수 있는데 최근의 연구결과에 의하면 HMM 성능이 DTW나 VQ 방식의 성능과 비슷하다는 것을 보여주고 있다.

위에서 기술한 격리단어 인식의 방법중 현재 가장 많이 쓰고 있는 DTW를 이용한 pattern matching 격리단어 인식 과정이 그림 1에 도시되어 있다. Pattern matching에 의한 격리단어 인식은 입력 음성의 특징 추출, 저장된 reference pattern과 test pattern과의 유사도 측정, 그리고 유사도 측정 결과에 의한 결정의 세단계를 거치게 된다. 이에 관하여 아래에 설명한다.

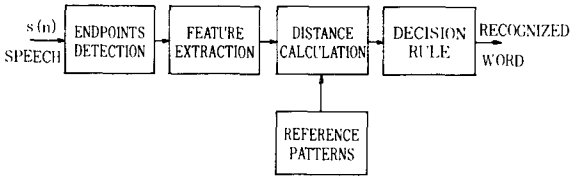

그림 1. 격리단어 인식 시스템의 블럭도

1) 특징 추출

음성 인식에서의 특징 추출은 기본적으로 많은 양의 음성 정보의 데이타를 감소하는 것으로서 이점은 음성 전송에서의 data reduction과 같다. 특징 추출은 memory size, 계산시간, 그리고 hardware 구현의 용이성을 고려하여야 하는데, 그 방법으로서는 filter bank에 의한 short-time spectrum 측정, homomorphic processing 방법, 선형 예측 방법(LPC) 등을 들 수 있다. Filter bank 특징 추출 방법은 입력 음성을 10개 내지 30개의 bandpass filter를 통과 시킨 후 rectification, low-pass filtering을 하여 그 출력을 양자화 한다. 사용되는 filter bank의 간격은 입력의 음성의 대역폭(일반적으로 100Hz 부터 3000 내지 6000Hz) 과 사용되는 filter bank의 수에 따라 다르나 보통 1000Hz 이하는 선형적으로, 그 이상에서는 logarithmic scale로 간격을 갖는다. 이 filter bank 특징 추출 방법은 구현이 용이할 뿐만 아니라 computation수가 비교적 적어 real time processing이 가능하기 때문에 현재 많이 사용되고 있다.

Homomorphic processing 방법은 입력 음성을 discrete Fourier transform(DFT) 과 logarithmic nonlinear operation으로 성도의 parameter (즉 spectral envelope) 와 여기신호(excitation signal)를 분리시키는 방법이다.[5] 분리된 spectral envelope는 전술한 filter bank 방식처럼 band를 10개 내지 30개로 나누어 양자화가 되는데 band를 나누는 방법은 선형 scale로 나누는 방법, 청각을 기초로 한 critical band scale로 나누는 방법등이 있다.

LPC방법은 원래 저전송 속도의 vocoding에 사용되었으나 LPC parameter set자체가 성도의 특성을 나타내기 때문에 음성 인식에도 널리 사용되고 있다.

2) 류사도 측정

입력 음성의 특징이 추출되면 test되는 단어의 특징과 기 저장된 reference 단어의 특징의 류사도를 측정하여야 한다. 류사도를 측정하기 위해서는 distance measure를 사용하여야 하는데 그 예로는 euclidean distance, spectral distance, LPC log likelihood measure등을 들 수 있다.

류사도를 측정하는 데에 있어서 한가지 중요한 문제점은 사람이 말하는 속도는 각 화자에 따라 다르기 때문에 test pattern과 reference pattern의 distance 측정으로만은 단어인식을 정확하게 할 수 없다는 점이다. 따라서 인식률을 높이기 위해서는 test pattern의 끝점 검출(end point detection) 이 정확하게 되어야 하고 두 pattern의 time alignment를 하여야 한다. Time alignment의 가장 효과적인 방법은 앞서 기술한 dynamic time warping(DTW) 방법 또는 보다 개선된 level-building DTW 방법으로 알려져 있는데 이들은 dynamic programming 기법(예 : Viterbi algorithm) 에 근본을 두고 있다. DTW 방식의 기본적인 문제는 end point와 alignment의 path가 주어진 조건하에서 test pattern과 reference pattern의 거리를 최소화하는 최적 warping path를 구하는 것으로서 지금까지 몇가지 algorithm이 제안

되었다. DTW방식에서 test pattern을 reference pattern에 warping시키는 것을 그림2에 도시하였다. DTW는 보통 distance 계산과 동시에 수행되는데 이 방법의 사용은 음성인식에서 가장 획기적인 발전으로 간주되고 있다.

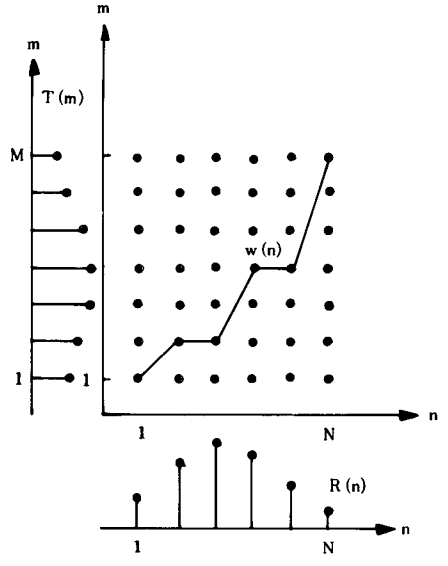

그림 2. DTW에서 path m=w(n) 을 통해 T(m)을 R(n)에 warping 시키는 예

3) 류사도 측정 결과에 의한 인식 결정

Pattern matching에 의한 인식에서의 마지막 단계는 입력되는 미지의 test pattern을 저장된 여러 개의 reference pattern 중에서 가장 가까운 것과 matching 시키는 일이다. 이를 위해서 여러가지 방법이 연구되었으나 가장 잘 알려진 방법은 nearest neighbor rule(NN rule)과 K-nearest neighbor rule (KNN rule)이다. NN rule은 test pattern과 reference pattern 사이의 평균 distance가 가장 적은 것을 인식된 word로 선택하는 것이다. 한편 KNN rule은 각 reference 단어가 둘 또는 그 이상의 reference pattern으로 주어질 때 사용된다. V개의 reference 단어 각각에 대해 P개의 reference pattern이 있다하고 i번째 reference pattern이 j번째 나오는 것은 $R^{i,j}$, $1<i<v$, $1<j<p$라고 표시하자. 이때 DTW의 distance를 $D^{i,j}$ ($D^{i,[1]}<D^{i,[2]}<\cdots<D^{i,[P]}$)로 표시하면 KNN rule에서는 average distance를

$$r^i=\frac{1}{K}\sum_{K=1}^{K} D^{i,[K]}$$

로 표시되고 계산된 r^i가 가장 적은 것을 인식된 단어로 간주하게 된다. KNN rule은 화자독립 시스템에서와 같이 단어당 template수가 많을 때는 NN rule(K=1)보다 K가 2 또는 3일 경우 인식률이 좋은 것으로 알려졌다.

격리단어 인식에서는 기본적으로 이상의 세단계를 거쳐 인식이 되는데 인식율은 특정 화자를 위한 화자 종속 시스템인 경우에는 현재 96% 이상이 되고 일반 화자 독립 시스템인 경우에는 적어도 85% 이상이 되고 있다. 단어 인식에서 현재 관심의 초점은 인식률을 높이는 일과 인식 어휘의 수를 늘리는 일이다. 인식률은 현재의 수준에서 현저히 높이는 것은 어려운 일이나 단어의 end point detection을 보다 정확히 하고 단어인식의 혼동을 줄일 수 있도록 multi-pass법을 씀으로써 어느 정도 효과를 보고 있다. 어휘수의 증가는 인식률은 물론 processing time과도 직접적으로 관련이 있는데, 이의 해결은 효과적인 인식 algorithm과 사용되는 DSP chip및 processor에 달려있다.

2. 연결단어 인식

소규모의 제한된 어휘량(예를 들면 연결 숫자)을 갖는 연결단어 인식시스템은 연결언어 인식 시스템의 제한된 형태로 볼 수 있으며, 1970년대 초기 이래로 이 연결단어 인식 시스템은 두가지 방식으로 연구되어 왔다. 첫번째 접근방식은 "bottom-up" 방식[6]으로서 단어열을 먼저 특정 단위(예, 음절 또는 음소)에 따라 분할한 다음 이 단위에 의해 분류하는 방식이다. 분할 방법은 각 어휘에 따라 실험적으로 정의되고 최적화된 규칙들에 의해 정해진다. Martin은 연결 숫자 인식 algorithm에 관하여 연구하였는데, 그는 기본 단위를 유사음소로 정하여 여기에 간단한 sequential decision algorithm을 적용하였다.[7] Nakatsu와 Kohda는 음절 단위의 분할에 근거를 두는 연결단어 인식 방법을 제안하였으며[8] 단어 전체를 기본 단위로 사용되는 시스템도 연구하였다. Sambur와 Rabiner는 유성음, 무성음 및 묵음의 구분에 근거를 두고 영어 숫자열에 대한 분할 규칙을 발표하였으며,[9] 최근에는 Zelinski가 통계적 추정에 근거를 둔 분할 방법을 제안했다.[10] 그러나, 이러한 분할 방식에 근거한 인식법은 분할 단

계 이후의 분류 단계에 치명적인 영향을 주는 단어 경계 검출 error를 종종 발생시킬 수 있다.

부정확한 분할에 의하여 발생하는 error를 없애기 위해서 dynamic programming 기법이 격리단어 인식뿐만 아니라 연결단어 인식을 위한 대안으로 사용되어 왔다. 이러한 방법의 장점은 인식 시스템이 미리 갖추어야 할 정보와 training의 양이 적다는 점으로 단지 각각의 단어에 대한 reference 패턴만을 알고 있으면 된다. 또 하나의 장점은 단어의 경계 검출과 비선형적 time alignment, 그리고 인식의 세 가지 동작이 동시에 수행된다는 점이다. 따라서, 단어 경계의 검출이나 비선형적 time alignment에서 비롯되는 인식의 오차는 발생하지 않는다. 이 algorithm은 완전한 단어에 대해서만 일치하려 하므로 단어의 경계는 자동적으로 결정된다.

다음은 어휘량에 관련된 문제점들을 고찰해 보기로 한다. 어휘량이 많으면 많을수록 인식 시스템의 응용분야가 더욱 넓어질 것은 분명하지만 많은 어휘량을 다루면서도 정확하고 빠르게 인식을 하기 위해서는 몇 가지 중요한 과정을 거쳐야만 한다. 첫째로, 단어는 체계적이고 효율적으로 표현될 수 있어야 한다. 많은 양의 어휘를 표현하기 위해서는 음소(phoneme), demisyllable, diphone 또는 음절(syllable) 등의 보다 작은 단위들이 주로 사용된다. 여기서 diphone이란 음성의 두 stationary region의 중심부 사이의 부분을 말한다. 어휘를 이러한 기본 단위로 나타낸 다음에는 미지의 음성을 어떻게 인식할 것인가, 즉 이미 저장되어 있는 단어의 표현양식과 미지의 음성의 음향 정보를 어떻게 비교할 것인가 하는 문제가 제기된다. 이 문제를 해결하기 위해서 많은 사람들이 가설을 세운 후 이를 검증하는 형태의 여러가지 방법을 연구해 왔다. 이는 문제의 일부에 대해서 타당한 가정을 한 후 각 가정의 타당성을 검증하고, 이러한 과정을 되풀이 함으로써 문제를 해결하려는 방법이다. 이러한 방식에서는 단어를 추정하는 부분은 미지의 음성의 음향 정보를 사용하여 우선 비교해 보아야 할 단어들을 선택하며, 단어를 확인하는 부분은 이 음성내의 특정한 부분에서 선택된 단어들과 비교해 보고 어느 단어가 관찰된 음향 신호와 가장 유사한 지를 결정한다. 끝으로 음성 인식을 하는 부분에서는 인식된 연결단어를 출력으로 내어놓게 된다.

또 다른 방식으로 "top-down" 방식이 있다.[11] 이 접근 방식에서는 추정하는 부분이 좀 더 높은 수준의 syntax 및 semantics 그리고 어휘 탐색에 대한 제한조건을 적용하여 단어를 추정한다. 이러한 경우에 단어를 확인하는 부분은 일치하는 단어를 찾아낸 후 시스템으로 하여금 이 단어를 사용하여 또 다른 top-down 방식의 가정을 하도록 한다. 시스템은 이러한 과정을 반복하여 전체 발음한 말과 일치하면서 문법적으로도 가장 합당한 순서의 단어열을 찾아낸다. 이러한 단어의 확인 과정은 계산이 많기 때문에 top-down 방식만의 사용은 거의 하지 않고 있다.

다음은 인식 시스템이 만족스러운 성능을 발휘하기 위한 화자의 수에 대하여 고찰하자. 물론 시스템을 사용할 수 있는 화자가 많을수록 더욱 좋을 것이나 일반적으로 여러 화자들의 음성은 여러 면에서 서로 다르다. 즉, 각 화자의 성도의 크기와 길이, phonetic target의 음향 특성, 발음의 세기와 속도가 다른 조건하에서의 coarticulation의 정도, 그들이 사용하는 방언, 어떤 어휘에 대한 음소의 형태, 그리고 녹음을 한 환경면에서 서로 다르게 된다. 이상에서 나타난 화자에 따른 차이를 해결하는 방법 중 하나는 각 화자에 맞는 template를 갖추어 놓는 방법이다. 이 template 들은 성도의 크기와 길이의 차이, 각 화자들 마다 다른 phonetic target의 음향학적 성질에서 비롯되는 차이, 녹음 환경의 차이, coarticulation 방법이 달라 생기는 차이등을 어느 정도는 보상할 수 있다. 방언과 어휘표현에 있어서의 차이는 어휘측에 다른 어휘들을 첨가하거나 한 가지 일정한 방언을 구사하는 화자에 대해서만 인식 시스템을 운용하므로써 해결할 수 있다. 그러나 많은 수의 화자에 대해서 화자에 특정된 template를 준비하는 것은 간단하지 않으므로, 이런 지루한 작업을 피하기 위해 많은 사람들이 화자독립 인식 algorithm을 개발하고자 노력해 왔다.

이러한 이유로 근래에 세가지 접근 방식이 개발되었는데 첫번째 방식은 수학적 변환에 의해서 성도의 영향을 normalize하는 화자 적응 방법이다. 이 방법은 새로운 화자를 위해서 잠시 동안의 training을 필요로 하는데 이것이 어느 정도의 제약 조건이 된다. 두번째 방법은 statistical training을 채택하는 방식으로서 이 경우에 있어서 어휘는 고정되어 있으나 각 단어에 대해서 몇 가지의 변화된 reference가 발음되는 sample을 기초로 clustering algorithm을 이용하여 통계적으로 구축된다.[12] 그러나 어휘량을 증가시키기 위해서는 새로운 단어에 대하

여 많은 화자들에 의한 training이 필요하고, clustering된 대표값에 잘 맞지 않는 화자들도 가끔 있다.

이러한 문제점을 완화시키기 위해서 수학적 변환이나 통계적인 clustering을 사용하지 않고 임의의 화자에 대해 일정한 특성을 가지는 화자독립 feature를 이용하는 세번째 방법이 연구되어 왔다.[13] 화자 독립적인 feature를 사용하는 시스템은 training이나 어휘량에 대한 제한이 필요 없으므로 큰 관심을 끌고 있다. 그러나, 완전히 화자 독립적인 feature들을 찾는다는 것은 간단하지 않으며 현재까지 얻어진 결과는 통계적 training 방법에 비해 뒤떨어진다.

다음은 계산량의 측면에서 고찰해 보겠다. 지금까지 음성 인식의 핵심적 부분인 dynamic programming(예 : 연결 단어 인식에서의 level-building DTW algorithm) 에서의 계산량을 감소시키려는 연구가 많이 수행되었으며, 만약 dynamic programming 과정 중 기울기에 제한을 두면 계산량을 어느 정도 감소시킬 수 있다. 최근에 궤적분할(trace segmentation) 방법이라 하는 비선형적 시간 압축 기법이 feature frame의 test sequence 중에 있는 frame의 수를 감소시키기 위해서 제안되었다. 이러한 방법들은 rhythm과 articulation rate의 변화가 음성 신호의 균일한 부분(steady region)에 특히 영향을 끼치며, 이들 균일한 부분은 매우 유사한 frame의 연속으로 이루어져 있기 때문에 약간의 frame 만으로 균일한 부분을 대표할 수 있다는 사실들을 기초로 하고 있다. 이러한 방식에서 frame rate는 음성신호에 따라 좌우되는데 균일한 음에서의 frame rate는 낮으며 transition 부분에서는 높게 된다. 이 방식은 다루어야 할 정보의 양을 감소시키며 dynamic pattern matching 전에 시간의 normalization이 이루어지므로 효과적이다.

이상에서 서술한 계산량을 줄이기 위한 세가지 방법은 지금까지는 음성 인식 시스템에서 각각 따로 사용되어 왔다. 그러나, 각각의 장점을 결합한다면 단 한가지만 사용하는 것보다 높은 성능을 가질 것으로 기대된다.[17]

3. 연속 언어 인식

여기에서는 음성 인식 시스템의 기본적인 구성에 대하여 살펴보고 인식의 기본 단위의 종류와 그 특성에 대해 기술한다. 연속 음성 인식 시스템, 또는 음성 이해 시스템(speech understanding system)의 기본적인 block도가 그림 3에 나타나 있다. 연속 음성이 입력되면 먼저 acoustic processor를 거친 다음 문법과 문장론, 구문에 입각하여 음성을 이해하는 방법을 bottom-up approach라 하고 그와 반대로 먼저 전체문장을 추정하고 그 문장에 포함된 단어나 문장들의 가정과 검증을 반복하는 방법을 top-down approach라고 하는데 보통은 두가지 방법을 병행하여 사용하고 있다.[14] 위의 두방법은 앞서 기술한 연결단어 인식에서의 두가지 방법과 같다.

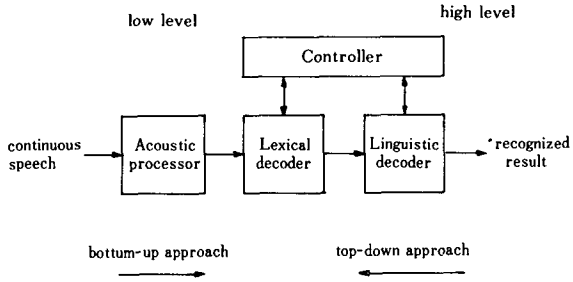

그림 3. 연속 음성 인식 시스템의 블럭도

Acoustic processor는 입력된 연속 음성을 그것으로부터 추출된 특성을 이용하여 인식 단위(recognition unit) 들의 sequence로 바꾸는 과정을 수행하게 되는데 현재는 크게 segmenting acoustic processor와 time-synchronous acoustic processor의 두가지 방법이 사용되고 있다. 그림 4에 acoustic processor의 block diagram이 도시되어 있다. 첫번째 방법은 그림의 실선으로 표시된 부분으로 입력된 연속 음성을 적당한 인식의 단위(recognition unit)로 구분하고 각 segment를 기준 pattern과 비교하여 labeling하는 방법인데 미리 입력 파형이나 spectrogram으로 부터 acoustic knowledge를 알고 있어야 한다. 이 방법은 이용가능한 모든 정보를 이용할 수 있다는 장점이 있는 반면 segmentation error와 labeling시의 error를 감수해야 하는 단점이 있다.

두번째 방법은 그림의 점선으로 나타난 부분인데 연속 음성을 인식 단위로 segmentation하는 것이 아니라 연속 음성을 고정된 길이의 frame으로 나눈

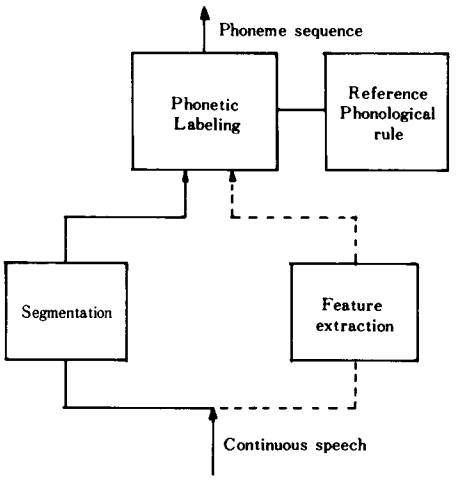

그림 4. Acoustic processor의 블럭도

다음 각 frame을 독립적으로 음소나 음절과 같은 인식 단위로 mapping하는 방법으로 segment 단계에서의 error에 영향을 받지 않는 장점이 있다.

Acoustic process를 거친 신호는 lexical decoder와 linguistic decoder를 거치게 되는데 이들에 대한 세부 block diagram이 그림 5에 나타나 있다. 이 decoder들은 acoustic processor의 결과를 이용하여 단어와 문장을 만들어 내고 문법과 문맥에 관한 knowledge를 이용하여 acoustic processor에서 발생 가능한 error를 교정하는 일을 하게 된다.

Lexical decoder는 word mapper와 phrase mapper로 구성되는데 word mapper는 acoustic processor의 출력인 recognition unit의 sequence를 미리 기억된 인식 단어 사전(lexicon)과 비교하여 가장 정확한 단어를 추정한 다음 그 추정된 단어와 시간 경계점, 그리고 그 단어가 얼마만한 정확도를 가지고 추정되었는가에 관한 정보를 phrase mapper로 전달하고 phrase mapper에서는 음운론적 규칙 (phonological rule)에 따라 가장 발생 가능성이 높은 문장을 만들어 내게 된다.

이렇게 가정된 문장은 linguistic decoder로 입력되어 인식하고자 하는 특정 언어의 문법(grammar)에 맞는 문장인가를 확인하는 syntactic analysis를 거쳐 문법에는 맞는 문장이라 하더라도 그 의미와 문맥에 어울리는 문장인가를 검증하는 semantic analysis와 pragmatic analysis를 하게 된다. 마지막

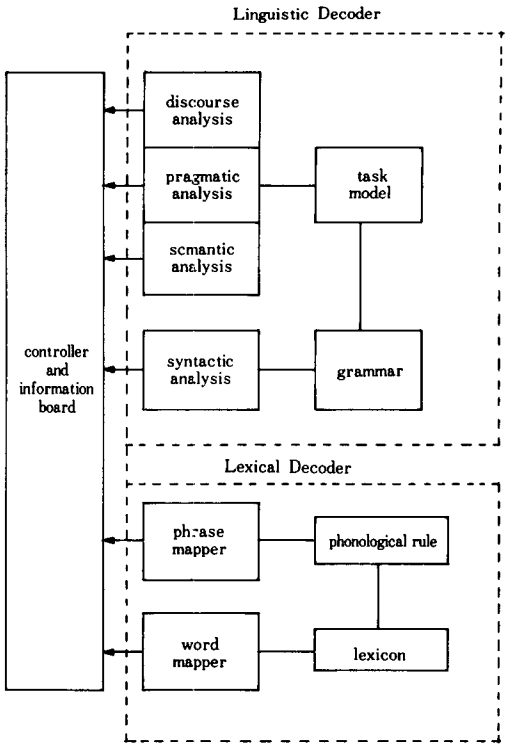

그림 5. Lexical decoder와 linguistic decoder의 블럭도

으로 문장의 stress, accent와 intonation 등으로 부터 얻은 정보를 이용하여 입력된 음성을 이해하게 된다(discourse analysis).

위에 기술한 음성 인식 시스템을 구성하기에 앞서 인식하고자 하는 언어의 음성학 및 음운론에 관한 연구가 선행되어야 함은 물론이다. 특히 연속 음성 인식을 위한 인식의 기본 단위 (recognition unit)를 무엇으로 정할 것인가에 관한 연구들이 필요하다. 지금까지 영어를 대상으로 하는 음성 인식시스템에 대한 연구 결과에 의하면 '단어'를 인식의 기본 단위로 하는 것은 바람직하지 않다.[15] 왜냐하면 연속음성은 모든 단어 경계에 대한 정보를 포함하고 있지 않기 때문에 단어의 경계를 찾아내기가 쉽지 않고 또한 단어 사이의 음운 현상과 조음결합을 고려하여 인식 대상 단어 사전을 구성해야하는 단점이 있는가 하면 인식 대상 단어의 수가 증가함에 따라 계산량의 급격한 증가를 초래하여 비효율적인 시스템이 되기 때문이다. 이런 이유들로 해서 '단어'보다는 subword unit를 인식 단위로 사용하고 있다.

표 1. 각 음성 인식 단위의 장단점

인식 단위	장 점	단 점
allophone	• 일부분은 음향학적으로 쉽게 구별 된다. • 음운론적 규칙의 수가 적다.	• 분류가 정확하지 못하다 • 종류가 많다.
phoneme	• 종류가 적다. • 단어를 쉽게 phoneme으로 표시할 수 있다.	• 음향학적으로 쉽게 분류되지 않는다. • 많은 음운론적 규칙을 필요로 한다.
diphone	• transition 정보를 내포하고 있다.	• 종류가 많다. • 일반적인 음운규칙들을 쉽게 적용할 수 없다.
syllable	• 비교적 쉽게 음절의 위치를 찾아낼 수 있다.	• 정확한 음절은 경계를 찾기 어렵다. • 종류가 많다.

Subword unit에는 phoneme, diphone, demisyllable, syllable 등의 여러가지가 있는데 이것들을 인식의 기본 단위로 사용할 때의 장·단점을 비교해 보면 표 1과 같다. 위의 subword unit 들은 원래 음향학 및 음성학적인 관점에서 정의된 것이 아니고 언어학적인 관점에서 정의된 것이기 때문에 subword unit를 인식 단위로 할때는 음성을 subword unit의 segment로 분리하기가 어렵지만 그 subword unit으로부터 단어를 만들어 내는 lexical decoding이 쉽게 된다. 반면에 subword unit를 음향학적인 관점에서 분류하는 방법이 Wilpon 등에 의해 제안되었는데 이 방법은 분류하는 과정은 수월하지만 음향학적으로 분류된 subword unit과 언어학적으로 분류된 subword unit 사이에 일 대 일 대응관계가 성립하지 않기 때문에 lexical decoding이 어려워지는 단점이 있다.

Ⅲ. 음성합성 기술

1. 음성합성의 기본 방법

과학적 의미를 지니는 최초의 음성 합성기는 1780년경 Kratzenstine에 의하여 설계된 인간의 vocal tract를 모방한 공명기(resonator) 라고 할 수 있다. 그러나 음성합성에 관한 본격적인 연구는 정보화 사회에서 필수적인 man-machine communication을 위한 수단으로써 음성 합성기의 필요성과 computer와 digital 신호처리 기술의 발달에 힘입어 활발히 진행되고 있다.[18~25]

우선 음성 합성기를 분류하여 보면, 작은 phonetic unit를 사용하며 이를 광범위한 언어학적 지식을 바탕으로 한 rule을 적용시켜 임의의 text를 음성으로 합성하는 text-to-speech system과, 미리 coding되어 저장된 문장, 구, 단어를 decoding하여 음성을 합성하는 제한적인 용도의 음성 응답 system으로 분류할 수 있다. Text-to-speech system은 광범위하게 사용될 수 있지만 아직 합성음의 음질이 자연스럽지 못하고 언어에 대한 많은 지식이 필요한 어려운점이 있으며, 음성 응답 system은 합성음의 음질이 text-to-speech system에 비하여 월등히 우수하지만, 그 응용분야가 제한되어 있는 단점이 있다.

음성을 합성하는 방법으로 앞의 서론에서 언급한 바와 같이 3가지 방법으로 구분 할 수 있는데 이에 관해 간략히 설명하면 다음과 같다.[21]

1) Waveform coding 방식에 의한 합성방법

이 방법은 우선 음성 파형을 sampling한 후, pulse code modulation(PCM), adaptive differential PCM (ADPCM), adaptive predictive coder(APC) 등 여러가지 coding 방법에 의하여 음성 파형을 coding하여 computer 내부에 저장한 후, 필요한 것들을 꺼내어 연결시켜 합성음성을 만들어 내는 방법이다.[19] 이 방법은 algorithm이 비교적 간단하며, decoding된 합성음의 음질이 좋은 장점이 있으나, database가 차지하는 양이 매우 많으며, 합성음을 제어하기가 매우 어려워 제한된 수의 단어를 사용하는 음성 응답 시스템에서는 많이 사용되나 일반의 text-to-speech system에서는 거의 사용하지 않는다.

2) 음성발생 model을 사용한 합성 방법

이 방법은 vocal tract의 음향학적 특성을 음성의 입·출력을 관찰함으로써 전달함수의 형태로 modeling하여, 이를 이용하여 음성을 합성하는 방법으로써, 음성 발생 model은 그림 6과 같이 sound source와

vocal tract filter로 구성된다. 이때 vocal tract filter의 coefficient는 LPC, filter bank 방법 또는 homomorphic 신호처리 방법등으로 구할 수 있다. 음성 발생은 그림 6과 같은 model을 이용하여 디지탈 필터를 구성함으로써 실현이 되는데 이 방법은 waveform coding에 의한 합성방법에 비하여 적은 memory로써 음성을 표현할 수 있는 장점이 있다. 또한 vocal tract의 coefficient를 조절함으로써 합성음을 비교적 쉽게 제어할 수 있으므로 text-to-speech system에서 작은 단위의 phonetic unit를 이용하여 문장을 합성하는데 매우 유리하다. 그러나 vocal tract model을 이용한 음성 합성 algorithm은 waveform coding에 의한 합성방법에 비하여 매우 복잡하며, 합성음의 음질도 나쁜 단점이 있다. 이러한 vocal tract model을 이용하여 음성을 합성하는 대표적인 예는 LPC 합성기와 formant 합성기들을 들 수 있다.

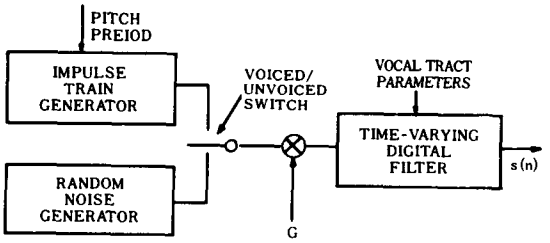

그림 6. Vocal tract filter를 이용한 음성 발생 model

3) Vocal tract을 simulation하여 합성하는 방법

이 방법은 사람이 발음할 때 vocal tract이 변화하는 과정을 X-ray 등을 이용하여 vocal tract을 관찰함으로써 vocal tract을 modeling하며 이를 이용하여 음성을 합성하는 방법으로써, 음성파형을 이용하여 vocal tract을 modeling하는 경우에 비하여 vocal tract을 정확히 modeling할 수 있는 장점이 있다. 그러나 vocal tract을 modeling하기 위한 data가 X-ray 등에 의존함으로 이러한 data를 구하기 어려우며, 관찰된 vocal tract이 3차원이므로 이를 정확히 modeling하는데 기술적인 문제가 있어 아직 합성음의 음질이 나쁘다.

Text-to-speech system을 구성하기 위해서는 그림 7과 같이 해석과 합성 routine을 거쳐야 된다. 여기서 해석 routine은 입력된 text를 발음되는 형태의 code로 바꾸는 역활을 할 뿐 아니라, 구문론적(syntactic)정보와 의미론적(semantic)정보 및 accent와 intonation에 관한 정보를 추출하며, 합성 routine에서는 해석 routine에서 얻은 정보와 음소의 feature parameter, duration, fundamental frequency 등을 사용하여 입력 text에 대응하는 합성 파형을 만든다. 이러한 해석 routine과 합성 routine에 관한 연구는 음성 합성에 필수적인 사항이므로 많은 사람에 의하여 연구 되었으며 지금도 진행 중에 있다.

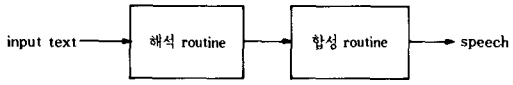

그림 7. 음성 합성을 위한 처리과정

2. 한국어 음성 합성에 관한 고찰

한국어 음성합성 시스템은 한글 text 입력을 이에 대응하는 한국어 음성으로 발생시키는 시스템으로서, 이 시스템의 기본적인 block diagram이 그림 8에 도시되어 있다. 이 시스템을 먼저 개괄적으로 설명하면 음성 표기변환 시스템은 입력된 한글 text를 한국어의 음성학적 특성을 토대로 설정된 발음규칙을 이용하여 발음표기로 변환시켜 준다. 또한 음율(prosody)을 조정하기 위하여 입력된 문장의 형태 (평서문, 명령문, 의문문, 중문 등)를 구별하여 주며 어간과 어미를 분리시키는 역할도 한다. Database에는 합성되는 각 음소 또는 음절등을 LPC, formant 또는 channel bank 방법 등으로 analysis하여 해당되는 feature parameter들이 저장 되어 있다. 제어 시스템에서는 음성 표기 변환 시스템의 출력과 computer 내부에 저장된 database를 이용하여 음절(또는 음소)과 음절을 연결시키는 연결 규칙을 적용시키며, 합성음의 자연스러움과 밀접한 관계가 있는 accent와 intonation을 위하여 음율 조절 규칙을 적용시켜 합성기에 입력 신호를 보낸다. 마지막으로 합성기에서는 feature parameter 추출을 위한 analysis 방법의 역과정을 적용시켜 음성을 합성해서 speaker로 출력시킨다. 이상에 기술한 각 과정의 핵심을 좀 더 자세히 살펴보자.

1) 음성 표기변환

음성 표기변환 algorithm은 한국어 음성합성 system에서 음운학적인 면과 음성학적인 면의 차이를 해결하여 준다. 즉, 합성음의 명료성을 위하여 한국

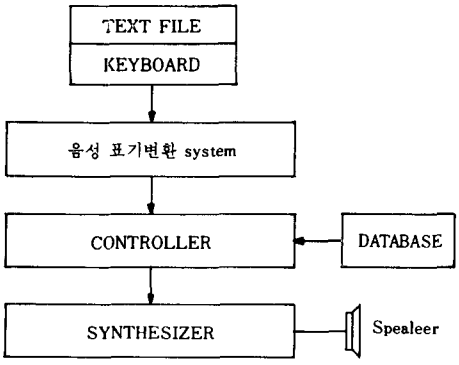

그림 8. 한국어 음성합성 system의 block diagram

어의 음성학적 발음 특성을 토대로 발음 규칙을 설정하며, 이를 이용하여 입력된 한글 text를 정확한 발음표기로 바꾸어 준다. 또한 자연스러운 합성을 만들기 위한 음율에 관한 정보도 제공한다.

따라서 음성 표기변환 algorithm에서는 입력 text가 한국어 정서법에 의한 띄워 쓰기 원칙을 준수하고 있다는 가정 아래, 뚜렷하며, 편안한 발음을 위하여, 그리고 음소가 결합할 때 발생하는 제약조건에 의해 음소가 다른 형태의 음소로 변환되는 음운변동 현상을 처리한다. 이때 음성 표기변환 algorithm에서 사용된 음운변동 현상에는 종성법칙, 연음법칙, 절음법칙, 음운의 첨가 및 탈락, 구개음화, 경음화, 격음화, 자음접변에 관한 법칙을 이용한다. 그리고 음율 조절을 위하여 평서문, 명령문, 의문문, 중문을 구별하며, pattern matching을 이용하여 어간과 어미를 구분한다.

2) Database를 위한 phonetic unit

음성합성 시스템의 database는 어떠한 phonetic unit를 사용하는가에 따라 합성음의 음질 및 database의 memory size 등에 크게 영향을 준다.[28]

한국어의 특성으로 보아 음성합성을 위한 효과적인 phonetic unit로서 음절을 들 수 있다. 음절은 음소의 결합으로 된 가장 기본적인 음소론적 단위로써, 음절의 구성을 공식으로 표현하면 $C^{l}VC^{r}$ (C^{l}=초성 자음, V=음절 핵, C^{r}=종성 자음)로 나타낼 수 있다. 여기서 C^{l}의 위치에는 |ㅇ|을 제외한 18개의 자음과 C^{l}가 없는 경우가 있으며(이 후 empty로 표현), V의 위치에는 10개의 단모음과 11개의 복모음이 온다. 그리고 C^{r}의 위치에는 empty와 7개의 대표음이 온다.

따라서 한국어의 음절구성은 아래와 같이 4가지 형태가 있으며, 이를 근거로 이론적으로 계산되는 음절의 수는 약 3500개 정도 된다. 그러나 실제 사용되는 음절의 수는 음소와 음소의 연결에 제약과 현재 사용되지 않는 음절이 있으므로, 이보다는 훨씬 적은 1096개 정도이다.[27]

$C^{l}=1$, $C^{r}=1$인 경우 : CVC형

$C^{l}=1$, $C^{r}=0$인 경우 : CV 형

$C^{l}=0$, $C^{r}=1$인 경우 : VC 형

음질의 면에서도 음절은 음절을 구성하고 있는 음소들 사이에 존재하는 transition 부분의 정보를 모두 가지고 있으므로, 다른 phonetic unit와 비교하여 볼때 좋은 음질의 합성음을 만들 수 있다. 또한 음절과 음절을 결합시키는 rule이 음소에 비하여 단순하며 음절과 음절사이에서의 coarticulation effect가 음소와 음소사이의 coarticulation effect에 비하여 강하지 못하고 음절과 음절사이의 경계에서 energy level이 골짜기 모양을 하고 있으므로,[26] 음절과 음절을 연결할 때 연결 부위에서 발생하는 spectral envelope상의 discontinuity에 의한 음질의 손상을 감소시켜 좋은 음질의 합성음을 기대할 수 있다. 그러나 database의 수가 비교적 많아 변이음 처리, duration 조절, 각 database와 관련된 parameter의 조절 등 database를 제어하는 일이 쉽지 않은 단점이 있다.

3) 음절의 연결

앞서 기술한 바와 같이 한국어 음절은 초성, 중성, 종성으로 구성되어 있으며, 이들 음절의 음소는 서로의 상호작용으로 인하여 음향학적 특성이 변형된다. 또한 음절로 구성된 단어는 앞뒤로 space, pause에 의하여 다른 단어들과 구별 되어지며, 음절 내부의 음소와 음소 사이에서 처럼 강한 상호작용은 아니지만, 음절과 음절 사이에서 음절의 종성과 다음 음절의 초성이 상호작용을 일으켜, 음절의 음향학적인 특성을 변형시켜 연결된다. 따라서 음절을 phonetic unit로 이용한 한국어 음성합성 system에서는 합성음의 음질 향상을 위하여 음절과 음질이 연결될 때 발생하는 현상을 고려하여 음절을 연결시킬 필요가 있다.

한국어 음절의 구성을 식으로 표현하면 $C^{l}VC^{r}$로 표시 할 수 있으므로, 두 음절이 연결될 때 다음과 같이 표현 할 수 있다.

$$C^{l_1}V^{1}C^{r_1} \cdot C^{l_2}V^{2}C^{r_2}$$

여기서 C는 자음을, V는 모음을 나타내며, 1은 첫 음절, 2는 두번째 음절을 표시한다. 그리고 한글의 특징상 $C^{'1}$에는 empty와 7개의 모음(/k/, /t/, /p/, /η/, /n/, /m/, /l/)이 올 수 있으며, $C^{'2}$에는 /η/를 제외한 모든 자음과 empty가 올 수 있으므로, 두 음절이 연결될 때 발생하는 경우의 수는 다음과 같다.

$$N = 8C^{'1} \cdot 19C^{'2} = 152$$

그러나 국어 음운학에 의하면, 두 음절이 연결될 때 제약 조건이 있으므로, 이 제약 조건을 고려하면 두 음절이 연결될 때 발생하는 경우의 수는 110가지로 줄어든다. 제약조건은 다음과 같다.[26]

- /l/ 초성은 /l/ 이외의 다른 종성과 연결되지 않는다.
- /k/, /p/, /t/ 종성은, 비음 앞에서 /g/, /m/, /n/으로 중화되므로, /m/, /n/ 초성은 /k/, /t/, /p/ 종성에 연결되지 않는다.
- /n/ 초성은 /l/ 종성에도 연결되지 않는다.
- /k/, /p/, /t/ 종성 다음의 /k/, /p/, /t/, /ĉ/는 경음화되며, /h/은 격음화 되므로 /k/, /p/, /t/, /ĉ/, /h/ 초성은 /k/, /t/, /p/ 종성에 연결되지 않는다.

4) 음율의 조절

사람의 음성에는 말하는 사람이 전달하려는 정보뿐 아니라 말하는 사람의 감정, 습관, 나이, 성별등 여러가지 주위 환경에 관한 정보를 포함하고 있다. 따라서 이러한 정보와 밀접한 관계를 가진 음율에 관한 내용을 연구하여 음성합성 system에 적용시키면, 좀더 자연스럽고 명확한 합성음을 기대할 수 있다. 음율에 관한 연구는 주로 음소의 duration, intensity, 그리고 fundamental frequency(이 후 F_0로 표시)와 밀접한 관계가 있는 accent와 intonation에 관한 것으로써, text-to-speech conversion system에서 이를 구현하는 algorithm은 몇가지 있는데 대표적인 algorithm은 schematic algorithm과 naturalistic algorithm은 있다.[29] 여기서 naturalistic algorithm은 미리 저장된 몇개의 표준 fundamental frequency pattern을 이용하여 비교적 정확한 자연 언어의 fundamental frequency pattern을 modeling하려고 한 것이며, schematic algorithm은 미리 저장된 표준 fundamental frequency pattern 없이, target과 target 사이를 interpolation에 의하여 연결함으로써 문장의 전체적인 fundamental frequency pattern을 구현하는 방식이다. 여기에서는 한국어의 accent와 intonation에 관한 특성에 관한 연구와 schematic algorithm을 이용하여 이를 구현하는 방법에 대하여 설명한다.[29,30]

(1) Accent

Accent는 "주위 음절에 대한 특정 음절의 F_0의 상승, duration의 증가, intensity의 증가, 또는 이들의 복합 현상"으로 정의할 수 있다. 한국어에서 F_0, duration, intensity 중에서 accent와 가장 밀접한 관계가 있는 parameter는 F_0이며, 그 다음으로 밀접한 관계가 있는 것이 intensity, 그리고 마지막이 duration으로 나타나 있다. 또한 accent의 위치에 대해서 살펴 보면, 한국어에서는 단어 단위로 accent가 부여되는 경우는 없으며, 모든 accent는 음절 단위로 부여된다. 그리고 accent의 위치는 구(phrase) 또는 문장의 전반부와 중반부에 있는 두 음절로 구성된 단어에서는 두번째 음절에 약간의 accent가 있으며, 문장의 종반부에 있는 단어는 첫 음절에 accent가 있다. 그리고 세 음절로 구성된 단어에서는 두번째 음절에 accent가 위치한다.

(2) Intonation

Intonation은 "문장에서의 pitch의 모양 또는 구(phrase)와 구 사이에서의 accent의 상대적인 높이"로 정의 할 수 있으며, accent와는 달리 말의 뜻을 구별시키는 변별적 기능이 있다. 즉, 어떤 사람이 "밥 먹어"라는 문장을 발음할 때, 문장의 끝 부분에서 F_0가 상승 또는 하강함에 따라 의문 또는 명령의 형태로 말의 뜻이 달라진다. 따라서 이러한 intonation에 관하여 정확한 modeling을 할 필요가 있다.

우선 문장 전체의 F_0 pattern을 살펴보면, 가장 일반적인 형태는 그림 9와 같은 점차 하강하는 pattern이며, 상승-하강의 pattern과 평탄-하강의 pattern도 볼 수 있다. 그리고 pause가 포함된 문장의 경우, F_0의 pattern은 문장의 첫 부분에서 pause가 있는 곳까지 하강하며 pause가 있는 곳에서 F_0가 급격히 상승한 후, 다시 천천히 하강하는 pattern을 가진다. 그리고 F_0의 가장 일반적인 경우인 하강하는 pattern은 거의 모든 언어의 공통된 현상이다. 즉, F_0의 변화폭이 문장의 끝 부분으로 갈수록 좁아지고, top-line은 빠르게 감소하며 base-line은 느리게 감소한다. 그리고 문장의 끝 부분의 F_0의 pattern은 선언문에서는 상승-하강, 예/아니오 의 형태로 대답이 가능한 의문문과 부가의문문에서는 상승, 의문문에서는 상승 또는 상승-하강 그리고

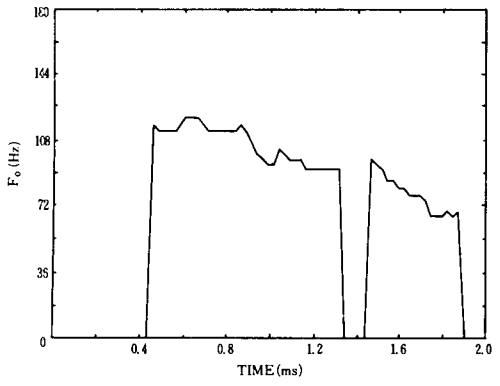

그림 9. 남성의 "형은 등산이 취미다."의 F_0 pattern

감탄문에서는 상승-하강의 형태를 지니고 있다. 따라서 이와 같은 intonation을 정확히 구현하기 위해서는 입력된 한글 문장을 음성합성 system이 이해해야만 가능하다.

Ⅳ. 결 언

지금까지 음성에 의한 man-machine communication에서 가장 핵심 기술인 음성 인식과 합성기술을 알아보고 이 기술의 여러가지 어려운 점들을 토의하였다. 음성은 가장 자연스러운 통신의 수단인 바 인간과 기계간에 음성으로 의사 전달이 가능하게 된다면 여러면에서 편리함은 이루 말할 수 없을 것이다. 이를 위해서 현재 범 세계적으로 많은 연구가 진행되고 있으나 연속언어 인식 기술이 실용화 되기 까지는 앞으로 오랜 시일이 걸릴 것으로 예측된다. 연속인식의 어려운 점은 현재 실용화 되기 시작하고 있는 제한된 어휘의 격리단어 인식 기술을 사용할 수 없다는 점이다. 다시 말해서 격리단어 인식 기술이 실용화 되었다고 해서 연속 음성인식이 곧 실용화 될 것으로 기대하는 일은 하나의 환상에 불과하다.

한편 음성합성 기술은 근래 신호처리와 VLSI 기술이 많이 발전됨에 따라 가까운 장래에 실용화가 될 것이다. 문제는 무제한 어휘의 음성합성에서 얼마만큼의 자연스러운 합성음을 만들수 있고 어떻게 경계적으로 시스템을 구현할 것인가가 관건이다.

음성인식이나 합성 기술이 사용하는 언어에 크게 구애 받는다는 것은 다 잘 아는 사실이다. 따라서 man-machine communication을 위한 한국어 음성인식 및 합성기술은 타 분야와는 달리 외국 기술을 직접 사용할 수 없고 국내에서 자체 개발 되어야만 될 특이성을 갖고 있다. 이를 위해서는 지속적인 연구비 투자와 많은 노력이 요청된다.

謝 辭

본 논문의 내용은 한국과학기술원에서 현재 진행중인 "한국어 음성 인식 기술 개발"과 "무제한 음성합성 시스템 개발" 연구결과의 일부입니다. 음성인식 연구를 가능토록 지난 수 년간 연구비를 지원하여 주신 한국전기통신공사와 음성합성 연구비를 지원하여 주신 과학기술처에 심심한 감사 말씀을 드립니다. 아울러 본 연구에 헌신적인 노력을 다하고 있는 통신공학 연구실의 speech 분야 연구원, 석·박사 학생들의 노고를 치하드립니다.

參 考 文 獻

[1] H. Sakoe, S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43-49, Feb. 1978.

[2] J.E. Shore, D. Burton, J. Buck, "A generalization of isolated word recognition using vector quantization," in Proc. 1983 IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 1021-1024, Apr. 1983.

[3] D.K. Burton, J.E. Shore, J.T. Buck, "Isolatedword speech recognition using multisection vector quantization codebooks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 4, pp. 837-849, Aug. 1985.

[4] L.R. Rabiner, S.E. Levinson, "A speaker-independent, syntax-directed, connected word recognition system based on hidden Markov models and level building," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 3, pp. 561-572, June 1985.

[5] L.R. Rabiner, R.W. Schafter, *Digital Processign of Speech Signal*, Englewood Cliffs, NJ: Prentice-Hall, 1978.

[6] V.R. Lesser, R.D. Fennel, L.D. Erman, D.R. Reddy "Organization of the hearsay II speech understanding system" *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 11-24, Feb., 1975.

[7] T.B. Martin, "Acoustic recognition of a limited vocabulary in continuous speech," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, 1970.
[8] R. Nakatsu, M. Kohda, "Speech recognition of connected words," Trans. IECE Japan, vol. E61, pp. 770-771, Sept. 1978.
[9] M.R. Sambur, L.R. Rabiner, "A statistical decision approach to the recognition of connected digits," *IEEE Trans. Acoust., Speech, Signal Processing,*, vol. ASSP-24, pp. 550-558, Dec., 1965.
[10] R. Zelinski, F. Class, "A segmentation procedure for connected word recognition based on estimation principles," in Proc. 1981 IEEE Int. Conf. Acoust., Speech, Signal Processing, Atlanta, GA, pp. 960-963, Mar.-Apr. 1981.
[11] W. Lea, Ed., *Trends in Speech Recognition*, Englewood Cliffs, N.J., Prentice-Hall, 1980.
[12] L.R. Rabiner and J.G. Wilpon, "Considerations in applying clustering techniques to speaker-independent word recognition," *J. Acous. Soc. Am.*, vol. 66, no. 3, pp. 663-672, 1979.
[13] P. Fonsale, "Connected word recognition system using speaker-independent phonetic features," in Proc. 1983 IEEE Int. Conf. Acoust., Speech, Signal Processing, Boston, pp. 312-315, 1983.
[14] B. Lowerre and R. Reddy, "The Harpy Understanding System," in Trends in Speech Recognition, Englewood Cliffs, N.J., Prentice-Hall, 1980.
[15] Torbjorn Svendsent and Frank K. Soong "On the Automatic Segmentation of Speech Signals," Proc. ICASSP-87, pp. 3.4.1-3.4.4, Dallas, 1987.
[16] J.G. Wilpon, B.H. Juang and L.R. Rabiner, "An Investigation on the Use of Acoustic Sub-Word Units for Automatic Speech Recognition," Proc. ICASSP-87, pp. 20.7.1-20.7.4, Dallas, 1987.
[17] 은종관외, "한국어 음성 인식 시스템 개발 연구," 1986, '87, '88 최종 보고서, 한국과학기술원.
[18] 은종관외, "한국어 무제한 음성 합성 시스템 개발," 최종보고서, 한국과학기술원, 1988.
[19] L.R. Labiner and R.W. Schafer, *Digital Processing of Speech Signals,* Prentice-Hall, Inc., N.J. Englewood Cliffs, 1978.
[20] J.L. Flanagan, "Computers that talk and listen: man-machine communication by voice," Proceedings of the IEEE, vol. 64, no. 4, April 1976.
[21] F. Fallside and W.A. Woods, *Computer Speech Processing,* Prentice Hall International, 1985.
[22] D.H. Klatt, "Review of text-to-speech conversion for English," JASA, vol. 82, no. 3, September 1987.
[23] J.L. Flanagan, L.R. Rabiner, "Speech synthesis," Papers in Acoustics, 1973.
[24] D. Osaughnessy, "Automatic speech *synthesis," IEEE Communication magazine,* 1983.
[25] G. Kaplan, E.J. Lerner, "Realism in synthestic speech," IEEE Spectrum, vol. 21, no. 4, April 1985.
[26] 허 웅, 국어 음운학, 정음사, 1985.
[27] 정 철, 국어 음소 배열의 연구
[28] 이상억, "Papers in Korean Phonetics," 범한서적, 1987.
[29] G. Akers and M. Lenning "Intonation in text-to-speech synthesis: evalution of algorithms," *J. Acoust. Soc. Am.*, vol. 77, Jun. 1985.
[30] A. Culter and D.R. Ladd, "In Prosody: Models and Measurements," Berlin, Heidelbery, New York, Tokyo: Springer-Verlag, 1983.

♣ 用 語 解 說 ♣

Image enhancement : 화상(畵像) 강조

화상 정보를 처리하여 화상내의 측정 정보를 강조하고, 보다 관측하기 쉬운 화상으로 변환하는 조작