음성인식 기술의 최근 동향과 국내 연구개발 현황

殷鍾官, 李惠洙, 丘明完, 具俊謨, 金會麟
韓國科學技術院 電氣 및 電子工學科, 音聲情報研究센타

Ⅰ. 서 론

근래 컴퓨터, 반도체, 신호처리 기술등이 빠르게 발전함에 따라 인간과 기계가 말로써 대화를 할 수 있는 가능성이 이제 현실화 되게 되었다. 말은 통신의 가장 간편하면서도 빠르고 정확한 수단으로서 모든 기계를 말로 작동 시킬 수 있다면 그 기술적인 파급효과는 물론, 경제적, 사회적인 면에서도 그 영향이 참으로 지대할 것이다. 이 기술을 음성인식 기술이라고 하는데 음성에 의한 man-machine interface 기술의 핵심이 된다.

음성 인식 기술은 1970년초 부터 연구가 진행되어 많은 업적이 이루어졌지만 현재까지 인간과 비슷한 능력을 지닐 수 있는 인식 기술은 아직 개발되어 있지 않다. 그러나 최근 10년간의 기술 개발로 미국, 일본 등지에서는 격리 단어를 인식할 수 있는 상업용 제품이 나와 있으며 연속 단어를 인식할 수 있는 시스템들도 연구실에서 개발이 되고 있다. 특히 대용량 단어로 이루어진 연속단어 인식 시스템은 음성학, 언어학, 컴퓨터 기술학, 음성 신호 처리 기술등이 집합이 되어야만 개발될 수 있는 첨단 기술 분야로서 실용화 될 수 있다면 그 응용 범위는 무궁무진하다고 할 수 있겠다.

본 논고에서는 이러한 음성 인식 기술 변화의 추이를 살펴보고 현재 우리나라에서 개발되고 있는 음성 인식 기술을 재 조명시켜 음성 인식 기술 발전의 토대로 삼겠다. 먼저 1장의 서론에 이어 2장에서는 최근 10년간에 이루어진 음성인식 기술, 특히 대용량 단어로 이루어진 연속 음성 인식 시스템의 핵심 기술을 분야별로 나누어서 서술하였다. 이들 분야로서는 음성의 특징을 추출하는 음성분석 기술, 분석된 특징으로부터 음성을 인식하는 인식기술, 기존의 음성인식 시스템이 새로운 화자에게도 사용할 수 있게 하는 화자 적응 기술 및 인식 대상이 되는 언어를 분석 처리하여 연속 음성 인식을 가능하게 하는 언어처리 기술등이 있다. 또한 이러한 분야를 통합한 인식시스템 개발 상황등을 기술하였으며 인식 시스템 성능 평가에 사용되는 데이타베이스(DB) 구성 현황에 대하여도 고찰하였다. 3장에서는 국내의 음성인식 기술 개발 현황을 학계와 연구소로 나누어서 개발사례, 연구 계획등을 재 조명 하였으며 4장에서는 결론을 맺겠다.

Ⅱ. 음성 인식 기술

1. 음성의 분석

음성 인식을 위한 음성의 특징추출 연구는 지난 수십년에 걸쳐 이루어져 왔으며 그 결과 현재 여러가지 음성특징 파라메터가 개발되어 음성 인식기에 적용되고 있다. 음성 신호로부터 적절한 음성 특징을 추출해 내기 위한 연구는 크게 세가지 분야에서 이루어져 왔는데 이는 음성의 발성과정을 모델링하는 방법, 음성의 인지 과정을 모델링하는 방법, 그리고 음성 신호 자체를 주파수 영역에서 해석하는 방법이다.

첫번째로 발성과정을 모델링하는 것은 사람이 음성을 발생시킬 때 각각의 발음이 성대로부터 성도를 거쳐 입술에까지 이르는 과정을 모델링하고 이로부터 특징 파라메터를 구하는 방법이다. 이러한 방법 중 가장 대표적인 방법이 all-pole 모델로 모델링하는 것으로써 이로부터 얻어지는 특징 파라메터가 잘

알려진 linear predictive coding(LPC)계수이다.[1] 이 LPC 계수는 모음에 대해서는 비교적 정확히 모델링 하지만 자음이나 비음등에 대해서는 성능이 저하되는 단점이 있다. 이러한 단점을 보완하기 위해 pole-zero 모델링을 적용하기도 하는데 이는 모델에 zero를 추가함으로써 비음과 같은 경우를 보다 정확히 표현할 수 있는 반면 계산량이 증가하게 된다. LPC 계수에 의한 특징 벡터들 사이의 distance는 잘 알려진 측정 방법으로 Itakura-Saito measure를 들 수 있는데 이는 계산량이 많고 잡음에 민감하다. 이와 다른 방법으로 LPC 계수로부터 cepstral 계수를 구하고 이에 적절한 weighting을 가하여 이를 특징 벡터로 사용하면 Euclidean distance로 특징벡터 사이의 거리를 측정할 수 있고 또한 잡음에도 robust한 음성 특징을 구할 수 있음이 밝혀졌다.[2] 그외에도 LSF (line spectral-pair frequency)로 음성 특징을 표현하고 이를 음성 인식에 이용하는 경우도 발표되었다.[3]

두번째로 인지 과정을 모델링하는 것은 발음된 음성에 대하여 사람의 귀가 어떻게 그 특징을 구별해 내어 이를 두뇌(brain)로 전달시키는 지를 연구하고 이를 토대로 음성인식에 유용한 특징을 추출하는 것이다. 이러한 방법중 가장 간단한 방법이 filter-bank 출력을 이용하는 것이다. 이때 이 filter-bank의 각각의 주파수 대역폭은 사람의 귀의 주파수 인지 특성을 실험하여 결정하며 실험 결과 비선형의 주파수 대역으로 분할하는 것이 인식 시스템의 성능향상에 도움이 되는 것으로 밝혀졌다.[4] 또한 청각기관의 인지 과정으로 부터 얻은 지식을 이용하여 더욱 정밀한 음성 특징을 추출하는 방법이 있는데 이를 auditory model에근거한 특징 추출 방법이라 한다.[5] 이 방식은 계산량이 증가하는 단점에 비해 인식율의 향상이 크지 않은데 이것은 아직 청각기관의 인지 과정에 대한 정보를 정확히 이해하지 못하기 때문이다.

세번째로 음성신호 자체를 주파수 영역으로 변환하고 이의 시간적인 변화를 관찰하여 이로부터 유용한 음성특징을 추출해 내는 방법이 있다. 이 방식은 수 많은 음성 샘플의 spectrogram을 관찰하고 이를 실제 발음학상 음운 기호와 비교하여 적절한 음성특징을 결정하게 된다. 이러한 음성특징으로는 에너지 정보, formant 정보, time duration 정보 등 여러가지가 있는데 상당기간 연구되어 왔음에도 불구하고 좋은 인식 성능을 보여주지 못하고 있다. 이는 앞에서와 마찬가지로 이해의 부족과 이러한 음성 특징을 효율적으로 제어 하기에 어려움이 있기 때문이다. 그러나 이러한 음성 특징이 앞서의 모델에 근거한 음성 특징이 지니지 못한 특성을 보완적으로 표현해 주기 때문에 이를 적절히 활용하면 보다 성능이 우수한 인식 시스템을 구현할 수 있음이 입증되었다.

지금까지 설명한 세가지 방향에서의 음성 특징들은 각기 독립적으로는 음성신호에 포함되어 있는 특성을 충분히 표현해 주지 못하기 때문에 실제로 음성 인식기에 적용될 때에는 이들을 상호 보완적으로 결합시키는 것이 바람직하다. 이러한 음성 특징들은 음성 신호의 어느 한 순간에서의 특성을 표현해 주므로 이를 instantaneous feature라고도 말한다. 그러나 사람이 음성을 인지할 때는 어떤 순간에서의 특징뿐만 아니라 음성 특징의 시간에 따른 변화 과정에도 큰 영향을 받기 때문에 이 정보를 표현해 주기 위하여 dynamic feature를 함께 사용하는 것이 바람직할 것이다. Dynamic feature를 사용했을 경우의 성능 향상에 관한 연구 보고가 그동안 많이 발표되어서 이를 많은 인식시스템이 활용하고 있다.[6]

음성 특징 파라메터를 인식 시스템에 적용시킬 때에 continuous한 값을 그대로 사용할 수 있지만 이를 discrete한 symbol로 변환하여 처리할 수도 있다. 이와 같이 discrete한 symbol을 사용하면 음성 신호의 정보량을 감축시키게 되고 결국 전체 계산량을 크게 줄일 수 있는 이점이 있다. 이러한 변환 알고리즘중 대표적인 방법이 vector quantization(VQ) 방식으로서 음성 신호의 부호화나 영상처리 분야에서 널리 이용되고 있다. 그동안 꾸준히 연구가 이루어져 total distance를 최소화 시킬 수 있는 알고리즘들이 개발되었는데 이중 대표적인 K-means 알고리즘이다.[7] 최근의 연구 결과 음성 인식에 있어서 total distance를 최소화 시키는 것이 가장 우수한 인식 결과를 보여주지는 않고 대신 classification rate를 높일 수 있는 VQ 알고리즘이 더 나은 인식 성능을 보여 준다는 연구보고가 있었다.[8] 이 방식은 인간의 두뇌 속에서의 정보처리 과정을 인공적으로 모델링한 neural network 이론을 적용한 것으로서, Kohonen의 self-organizing feature map 알고리즘을 Bayesian optimal classification rule에 접근하도록 수정한 LVQ(learning VQ)알고리즘이다.

마지막으로 음성특징 추출에 있어서 고려할 사항은 발음 환경에 관한 문제이다. 이를 위해서 실제로 개발한 인식 시스템을 어떤 주변환경 속에서 사용할 것이냐에 따라 그 환경에 영향을 받지 않는 특징 파라메터를 추출해 내어야 할 것이다. 이러한 잡음에

는 white noise와 어떤 특정한 colored noise가 있을 수 있는데, 이러한 잡음의 영향을 받지 않기 위해 전처리 과정에서 미리 filtering 시키는 방법과 이러한 잡음에 robust한 음성 특징을 찾아내는 방법이 있을 수 있다. 또한 인식될 음성이 전화선과 같은 특정 channel을 통할 경우에는 그 channel에서의 주파수 왜곡이나 기타 잡음이 음성 특징에 영향을 끼칠 수 있으므로 이를 보상해 주기 위한 연구가 수행되어 왔다.

2. 음성의 인식

앞절에서 살펴본 것처럼 음성을 분석하기 위한 기술이 발전함과 더불어 기계가 인간의 음성을 인식하는데 필수적인 음성 인식 방법도 매우 큰 진전을 이루었다. 즉, 적은 수의 어휘가 격리 단어의 형태로 한사람의 화자에 의해서 발음되는 경우에만 효과적인 음성 인식을 할 수 있었던 과거와는 달리 많은 수의 어휘가 연속 음성의 형태로 화자의 구분없이 자연스럽게 발음되는 경우에도 음성인식이 가능하도록 하는 경우가 진행되고 있다. 이를 위하여 많은 기술들이 도입되어 음성 인식에 효과적으로 사용되고 있다. 예를 들면, 많은 수의 어휘를 인식하기 위해서 벡터양자화 기술, 새로운 인식 단위의 선정 방법, 시간감축 방법, 사전구성 방법등이 연구되었다. 자연스러운 연속 음성을 인식하기 위해서 적은 수의 인식 단위를 이용하여 문장 단위의 음성을 모델링하는 방법, 사용하는 언어의 문법이나 구문등을 이용하는 방법, 자연스러운 음성에서 나타나는 다양한 음가의 변화를 수용할 수 있는 인식 단위의 선정 방법등이 연구되었다. 다양한 화자의 음성을 인식하기 위해서는 다수의 표준패턴을 선택하는 방법과 새로운 화자에 대하여 인식 시스템을 적응시키는 화자적응 방법, 화자독립적인 음성 특징의 추출 방법등을 연구하였다. 이상에서 언급한 것 이외에도 여러가지 방법이 음성인식에 이용되었고 음성을 인식하기 위한 기본적인 접근 방식도 많은 발전을 이룩하였으며 새로운 방법들이 제안되었다.[2]

그동안에 널리 사용된 음성인식 방법은 동적 프로그래밍 기법을 이용하여 기준이 되는 음성과 입력 음성을 비교하여 최소거리 법칙에 의하여 입력 음성을 인식하는 dynamic time warping(DTW) 방법이다.[10] 이 방법은 한 사람 혹은 여러 사람으로 하여금 인식 대상 어휘를 여러번 발음하도록 한 후 clustering 알고리즘[11]에 의하여 기준 패턴을 선택하는 training 과정과 이 기준 패턴과 입력 음성을 비교하는 음성인식 과정으로 나뉘어 진다. 음성은 동일한 사람이 동일한 단어를 발음하여도 발음 속도가 매번 다르므로 패턴 사이의 유사도를 측정하려면 이 시간축상의 변화를 고려하여야 한다. 이 시간축상의 변화를 보상해 주는 효과적인 방법인 DTW에 관하여 살펴보면 다음과 같다.

길이가 M frame인 입력 음성을 $T=\{T(1), T(2), \cdots T(M)\}$, 길이가 N frame인 기준 패턴을 $R=\{R(1), R(2), \cdots R(N)\}$라 하고 $T(m)$ 및 $R(n)$은 일차원 함수라고 가정하면 DTW 알고리즘은 총 distance 함수

$$D=\sum_{n=1}^{N} d(R(n), T(w(n))) \qquad (1)$$

를 최소화 시키도록 (m, n) 평면상의 최적 경로 $m=w(n)$을 결정하는 방법이다. 여기서 $d(R(n), T(w(n)))$은 **R**의 n번째 frame과 **T**의 $w(n)$번째, 즉 m번째 frame사이의 distance를 의미한다. 이상의 과정에 대한 예가 그림 1에 그려져 있다. DTW 알고리즘에 위하여 최적 path를 구할 경우, (m, n) 평면상에서 이 path가 지날 수 있는 범위에 대하여 몇가지 제약 조건이 주어져 시간축을 지나치게 신장 또는 수축시키

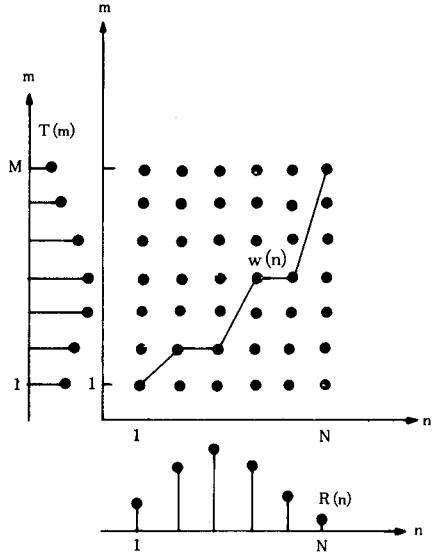


그림 1. Path $m=w(n)$을 통해 $T(m)$을 $R(n)$에 warping 시키는 예

지 않도록 하는데 그 중에서 Itakura와 Sakoe 및 Chiba가 제안한 제약 조건이 널리 쓰인다.[12,13]

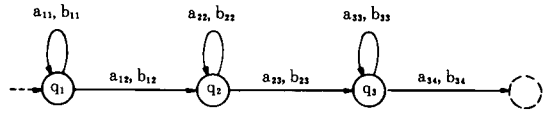
DTW 알고리즘을 이용하여 격리단어 뿐만 아니라 연결 단어까지도 인식하기 위하여 one-stage dynamic programming(OSDP)[14] 방법이나 level-building dynamic time warping(LBDTW)[15] 방법 등이 제안되어 연결 숫자의 인식이나 200 단어 정도로 구성되는 문장의 인식등에 이용되었다. DTW를 이용하는 음성인식 방법은 높은 인식율을 얻을 수 있다는 장점이 있으나 많은 수의 기준패턴을 구성하기 위해서는 많은 시간과 노력이 필요하며 인식 과정에 걸리는 시간이 길기 때문에 100여 단어 정도의 어휘를 화자 종속으로 인식하는 경우에 적합하나 대용량 단어의 인식등에는 적합하지 못하다는 단점이 있다.

이러한 단점을 극복할 수 있는 새로운 방법으로 hidden Markov modeling(HMM)을 이용하는 음성인식 방법이 제안되어 많은 연구가 이루어져 왔다.[16,17] 이 방법은 음성을 상태 천이 확률 및 각 상태에서의 출력 symbol의 관찰 확률을 갖는 Markov process로 가정한 후에 trainining data를 통하여 상태 천이 확률 및 출력 symbol 관찰 확률을 추정하는 training 과정과 추정된 모델에서 입력 음성이 발생할 확률을 계산하는 인식 과정으로 나누어 진다. 이상의 과정을 살펴보기 위하여 아래와 같이 정의하자.

T = 출력 symbol sequence의 길이
N = state의 수
M = 출력 symbol의 갯수
$Q = \{q_1, q_2, \cdots q_N\}$ state들의 집합
$V = \{v_1, v_2, \cdots v_M\}$ symbol들의 집합
$\mathbf{A} = \{a_{ij}\}$, $a_{ij} = \Pr(q_j \text{ at } t+1 | q_i \text{ at } t)$ 상태천이 확률
$\mathbf{B} = \{b_{ij}(k)\}$, $b_{ij}(k) = \Pr(v_k \text{ at } t | q_i \text{ at } t, q_j \text{ at } t+1)$ 출력 symbol 관찰 확률
$\Pi = \{\pi_i\}$, $\pi_i = \Pr(q_i \text{ at } t=1)$ 초기 확률

이상의 정의를 이용하면 HMM은 $\lambda = (\mathbf{A}, \mathbf{B}, \Pi)$로 표시되며 그 간단한 예가 그림 2에 그려져 있다.

Training 과정에서는 training data로 부터 모델 λ의 여러가지 parameter를 추정하게 된다. Parameter를 추정하는 방법으로서는 training data와 모델사이의 likelihood를 최대로 하는 maximum likelihood(ML) training 방법,[18] 모델과 training data 사이의 mutual information을 최대화 하는 maximum mutual information(MMI) training,[19] 방법과 training data에 대하여



$A = \{a_{ij}\}$, $1 \le i, j \le 4$

$B = \{b_{ij}\}$, $b_{ij} = \{b_{ij}(k)\}$ $1 \le i, j \le 4, 1 \le k \le M$

그림 2. 3개의 state와 M개의 출력 symbol을 갖는 간단한 left-to-right HMM의 예

인식 과정을 수행하여 오인식이 발생하는 경우에 parameter를 수정하여 주는 corrective training(CT)[20] 등이 제안되었다. 이 가운데에 가장 널리 쓰이는 ML training 방법은 forward 확률 $\alpha_i(t) = \Pr(q_t = i | v_1, v_2, \cdots v_t)$와 backward 확률 $\beta_i(t) = \Pr(q_t = i | v_{t+1}, \cdots, v_T)$를 정의하면 아래와 같은 방법으로 구하여 진다.

$$\alpha_i(t) = \begin{cases} 0 & t=0, i \neq \text{initial state} \\ 1 & t=0, i = \text{initial state} \\ \sum_j \alpha_j(t-1)\, a_{ji}\, b_{ji}(v_t) & t > 0 \end{cases} \quad (2)$$

$$\beta_i(t) = \begin{cases} 0 & i \neq \text{final state}, t=T \\ 1 & i = \text{final state}, t=T \\ \sum_j a_{ij}\, b_{ij}(v_{t+1})\, \beta_j(t+1) & 0 \le t < T \end{cases} \quad (3)$$

$$\gamma_{ij}(t) = \Pr(q_t = i, q_{t+1} = j | v_1^T) = \alpha_i(t-1)\, a_{ij}\, b_{ij}(v_t)\, \beta_j(t) / \alpha_N(T) \quad (4)$$

$$a_{ij} = \sum_{t=1}^{T} \gamma_{ij}(t) \Big/ \sum_{t=1}^{T} \sum_k \gamma_{ik}(t) \quad (5)$$

$$b_{ij}(k) = \sum_{t: y_t = k} \gamma_{ij}(t) \Big/ \sum_{t=1}^{T} \gamma_{ij}(t) \quad (6)$$

위와 같은 과정을 거쳐서 추정된 HMM parameter들은 충분하지 못한 training data 등에 의하여 발행되는 인식율의 저하를 막기 위하여 smoothing 과정이나 interpolation 과정등을 거쳐서 최종적인 model parameter로 결정된다.[21]

인식과정은 추정된 각 모델에 대하여 입력음성 $O = \{o_1, o_2 \cdots o_T\}$이 발생할 확률 $\Pr(O|\lambda)$가 계산되며 가장 높은 확률을 갖는 모델이 대표하는 어휘나 문장으로 인식되어 진다. 이 확률 $\Pr(O|\lambda)$는 다음과 같은 Viterbi scoring방법에 의하여 결정된다.

$$V_i(t)=\begin{cases}0 & t\neq 0, i=\text{initial state}\\ 1 & t=0, i=\text{initial state}\\ \underset{j}{\text{Max}}\ V_j(t-1)\,a_{ji}b_{ji}(o_t) & t>0\end{cases} \quad (7)$$

$$Pr(O|\lambda)=V_N(T) \quad (8)$$

이상에서 설명한 기본적인 알고리즘외에 인식 시스템의 성능을 향상시키기 위하여 음성에 관한 여러가지 정보(음소의 지속시간, 음운의 탈락, 변화, 첨가등)와 HMM 과정에 사용되는 VQ 과정의 왜곡을 없애거나 줄이기 위한 방법(continuous 또는 semi-continuous HMM)등 여러가지 방법이 제안되고 연구되었다.[22] 특히, HMM이 갖는 장점중의 하나인 단어 이하의 음성인식 단위, 예를 들면 음소, diphone 등으로 부터 단어나 문장등의 모델을 쉽게 구성할 수 있다는 점을 이용하면서 음성 인식율을 향상시킬 수 있도록 음성 인식 단위를 선정하는 방법에 관한 연구가 수행되었다.[23] HMM을 이용한 음성 인식 시스템은 인식 소요시간이 짧고 적은 수의 음성인식 단위로 부터 쉽게 어휘들을 모델링 할 수 있다는 장점을 갖고 있어서 DTW 보다 앞선 방법으로 평가받고 있으며 따라서 이에 관한 많은 연구가 필요하다.

최근에는 neural network을 이용하여 음성을 인식하는 방법이 제안되어 주목을 받고 있다. Neural network은 인간의 신경 조직을 모방한 것으로 입력을 받아들여 입력의 합에 nonlinearity를 가한 후 이를 출력으로 내보내는 node들을 상호 연결함으로써 구성된다. 이 node들의 연결 방식에 따라서 여러가지의 형태의 neural network이 구성되며 그 동작은 node간의 연결의 세기를 결정하여 주는 weight 값에 큰 영향을 받는다.[24] Minsky와 Papert에 의하여 perceptron이 제안된 후에 다층 구조를 갖는 multi-layer perceptron(MLP)의 weight를 training 시킬 수 있는 back propagation training 알고리즘이 발표된 후에 급격히 연구가 진행되고 있다.[25]

그림 3에 각 layer의 갯수가 M, N인 two-layer perceptron의 예를 도시하였다. 여기서 θ는 각 node가 갖는 threshold 값이며 w_{ij}는 i번째 node에서 j번째 node를 연결하는 weight를 뜻한다. $f(\cdot)$는 임의의 nonlinear function이나 대개의 경우 sigmoid 함수의 형태를 갖는다. 하위 layer는 입력을 받아들여 입력의 가중치의 합에 대하여 nonlinearity를 가하여 이를 상위 layer의 입력으로 전달하게 되며 최종적으로 최상위 layer에 있는 node의 출력이 전체 neural network의 출력이 된다.

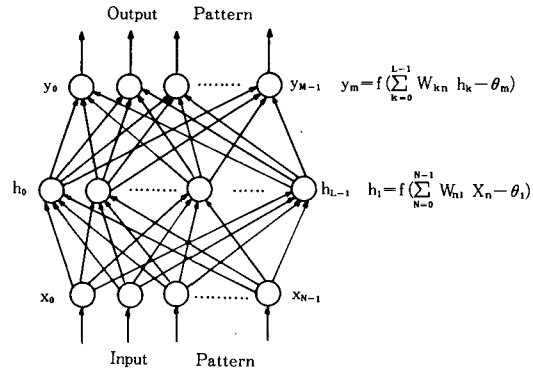


그림 3. Two-layer pereceptron의 예

Neural network이 갖는 장점은 parallel processing이 가능하므로 계산 속도를 높일 수 있다는 점과 neural network의 일부가 파손되어도 전체적인 동작에 큰 영향을 주지 않는다는 점외에 neural network의 성능을 입력 data가 주어짐에 따라서 점진적으로 향상시킬 수 있도록 weight 값을 learning 할 수 있다는 것이다.[26] 이러한 이유로 neural network을 이용하여 음성을 인식하는 많은 연구가 진행되어서, supervised learning을 수행하는 MLP, LVQ2, TDNN 등을 이용하거나 unsupervised learning을 수행하는 feature map등을 이용하여 음소등을 인식하는 연구가 수행되었다.[27-29] 그러나 neural network은 static pattern의 인식에는 우수한 성능을 보이지만 시간에 따라 변화하는 dynamic pattern의 인식에 취약한 점이 있다. 이를 해결하기 위하여 최근에는 여러가지 recurrent neural network[30]을 이용하거나 DTW나 HMM 등과 결합하여 음성인식을 수행하고자 하는 연구가 진행되고 있다.[31] 또한 음성 인식에 적합한 neural network의 구조나 training시간의 감소에 관한 연구도 수행되고 있다.

마지막으로 음성에 관한 지식을 이용하여 음성을 인식하는 knowledge-based 음성 인식 방법에 관하여 살펴보겠다. 이 방법은 음성에 관한 지식을 구축하고 이를 여러가지 rule에 의하여 구성되는 inference engine을 이용하여 입력 음성을 인식하도록 하는 방법이다. 이러한 방법중 가장 대표적인 것은 spectrogram reading 과정에서 얻은 지식을 바탕으로 expert system을 구성하여 음성을 인식하는 Zue의 연구이다.[32] 이 방법은 전문적인 spectrogram reader는

spectrogram을 보고 음성을 인식할 수 있다는 점에 착안하여 연구 되었으나, 인간이 시각정보를 처리하는 과정을 expert system으로 효과적으로 구현할 수 없기 때문에 다른 방법에 비하여 음소의 분할이나 labeling 등의 비교적 제한된 영역에서 사용되고 있다.[33]

이상으로 음성인식 방법중 널리 이용되고 있는 네 가지 접근 방법과 연구 동향에 관하여 살펴보았다. 다음 절에서는 화자적응 방법에 관하여 살펴보도록 하겠다.

3. 화자 적응

음성 인식 시스템은 인식 대상에 따라 특정 화자의 음성만을 인식하는 화자 종속 시스템과 화자에 관계없이 비슷한 인식율을 얻을 수 있는 화자 독립 시스템으로 나눌 수 있는데 현재의 기술로는 화자 종속 시스템의 인식율과 비슷한 화자 독립 시스템은 개발되어 있지 않다. 그러나 화자 종속 시스템은 화자가 바뀔 때마다 특정 화자에 맞게끔 새롭게 화자 종속 시스템을 구성하여야 하기 때문에 많은 데이타와 시간이 필요하게 되고, 그러한 이유로 다양한 화자의 음성을 인식하여야 하는 분야에서는 별로 효용이 없다.

화자 적응이란 화자 종속 시스템을 화자가 바뀔때마다 특정 화자에 대하여 새롭게 화자 종속 시스템을 재구성하는 것이 아니고 기존의 시스템 정보를 충분히 이용하고 새로운 화자의 음성 정보를 약간만 사용하여서 화자 종속 시스템의 성능과 비슷하게 하는 방식이다. 실제로 화자 적응이란 완벽한 화자 독립 시스템을 만들 수 없으므로 화자 종속 시스템을 화자에 독립적으로 사용할 수 있도록 하는 방식이지만, 동일한 화자라도 화자의 건강 상태, 기분, 또는 발음할 때의 주변 환경에 따라 음성이 다양하게 변하기 때문에 특정 화자에 맞게 구성된 음성인식 시스템을 이러한 조건에 따라 시스템 내부의 정보를 바꾸어 줄 때에도 이용될 수 있다. 이러한 생각은 인간이 새로운 환경, 혹은 새로운 화자의 음성에 적응하면서 음성을 인식한다는 사실과 유사하다.

현재까지 화자 적응을 위하여 개발된 알고리즘은 인식 시스템에 화자 적응만을 위한 모우드의 존재 여부에 따라 정적 적응(static adaptation), 동적적응(dynamic adaptation)으로 나눌 수 있으며 화자 적응시 사용되는 음성 데이타의 내용을 시스템이 미리 인지하고 있는 지에 따라 지도 적응(supervised adaptation), 독자 적응(unsupervised adaptation)으로 구분된다.

정적 지도 적응 알고리즘은 화자 적응만을 위한 특정한 모우드에서 미리 지정된 어휘를 새로운 화자가 발음하면 시스템내에 저장된 정보를 새로운 화자의 특징에 맞게끔 적응시키는 알고리즘으로서 가장 많이 연구되어 왔다. 이러한 이유는 정적 지도 적응 알고리즘이 화자 적응을 위한 알고리즘중 가장 기본적이며 이 알고리즘의 성능 향상이 우선적으로 이루어져야 다른 알고리즘의 성능 향상을 기대할 수 있기 때문이다. 대표적인 정적 지도 적응 알고리즘은 discrete HMM 파라미터 등을 새로운 화자에 맞게 변경시키는 HMM parameter adaptation과 codebook을 새로운 화자에 적응시키는 codebook adaptation 방식으로 구성된다. HMM parameter adaptation은 HMM parameter중 output probability만을 새로운 화자에 맞게끔 주로 변경시키는데 새로운 화자가 적응 모우드에서 발음한 음성 데이타(적응 데이타)와 기존의 시스템에 저장된 동일한 어휘의 음성 데이타로 부터 새로운 화자의 음성 특징을 찾아내어서 output probability를 변경시키는 probabilistic spectral mapping[34]과 fuzzy 개념을 추가한 fuzzy histogram mapping[35] 방식으로 구성된다.

Probabilistic spectral mapping은 BBN 연구소에서 화자 종속 시스템으로 개발한 BYBLOS에 사용되었는데 과정은 다음과 같다. 우선 적응 데이타와 시스템에 내장된 음성 데이타로 부터 DTW를 수행하여 음성 특징의 상관 관계 C(i, j)

$$C(i,j),\ j=\omega(i),\ i=1,2,\cdots I,\ j=1,\cdots J$$

ω : warping function,

I, J : 최대 frame 수 (9)

를 구한 후, 이들을 누적시켜 음성 데이타를 VQ encoding시킨 VQ index 사이의 통계적 횟수 $C(\hat{L}_j, L_i)$를 얻은 다음 probabilistic spectral mapping matrix $P(\hat{L}_j/L_i)$

$$P(\hat{L}_j/L_j)=\frac{C(\hat{L}_j, L_i)}{\sum_{j=1}^{M} C(\hat{L}_j, L_i)} \quad (10)$$

$\hat{L}_j$: 적응 데이타 j번째 frame의 VQ codebook index

L_i : 기존 데이타 i번째 frame의 VQ codebook index

를 구한다. 그러면 새로운 화자의 S_k state($1 \le k \le N$)에서 $\hat{L}_j$ index를 갖는 output probabilily $P(\hat{L}_j/S_k)$는

$$P(\hat{L}_j/S_k) = \sum_{L_i=1}^{M} P(L_i/S_k)\ P(\hat{L}_j/L_i, S_k) \qquad (11)$$

$1 \le L_j \le M$, M : codebook 크기

인 관계식이 성립되며 $P(\hat{L}_j/L_i, S_k)$가 state S_k에 독립적이라고 가정하면 probabilistic spectral mapping matrix와 동일하게 되어 (11)식으로 부터 output probability를 쉽게 구할 수 있다. IBM Japan에서는 통계적 횟수 $C(\hat{L}_j, L_i)$를 구하는데 DTW를 사용하지 않고 기존 시스템의 HMM parameter를 이용하여 적응 데이타를 Viterbi alignment 시켜서 $C(\hat{L}_j, L_i)$를

$$C(\hat{L}_j, L_i) = \sum_{L(\omega,t)=\hat{L}_j} P(L_j/V[L(\omega, t)]) \qquad (12)$$

$L(\omega, t)$: 적응 어휘 $\omega(1 \le \omega \le W)$의 t번째 frame 에서의 VQ index,

$V[L(\omega, t)]$: $L(\omega, t)$의 Viterbi alignment

와 같이 구한후 (10), (11)식을 이용하여 probabilistic spectral mapping 시키기도 한다.[36]

Codebook adaptation은 적응 데이타로 부터 VQ codebook을 구성하여 새로운 화자에 이용하는 방식을 채택하는데 기존의 codebook과의 관계는 적응 데이타와 시스템에 내재해 있는 같은 어휘의 음성 데이타를 DTW를 취하여 codebook mapping table을 구성하여 얻는다. 반면 Grenier는 새로운 화자가 발음한 음성과 기존의 화자가 발음한 음성 사이의 correlation이 최대가 되도록 적응시키는 canonical correlation analysis를 이용하여 codebook adaptation을 수행하였다.[37]

정적 독자 알고리즘은 화자 적응 만을 위한 적응 모우드가 존재하지만 이때 새로운 화자가 화자 적응을 위해 발음하는 적응 어휘가 미리 정해져 있지 않고 인식 대상 어휘중에서 임의의 어휘를 발음하여도 화자 적응이 되는 알고리즘이다. 이 알고리즘은 새로운 화자가 적응 과정 모우드에서 미리 정해진 어휘를 발음할 필요성이 없으므로 편리하지만 적응도가 떨어지는 단점이 있다. 대표적인 알고리즘으로 NTT의 Furui가 개발한 VQ codeword의 상관 관계를 이용한 것이 있는데 VQ codebook을 구성할 때 codeword를 계층 구조의 관계로 바꾼 다음 새로운 화자의 적응 데이타가 입력되면 계층 구조가 변하지 않도록 codeword 값을 새로운 화자에 적합하게 변경시켜 적응시킨다.[38]

동적 독자 적응 알고리즘은 화자 적응만을 위한 적응 과정 모우드가 없이 시스템 내부에서 인식 작업이 일어날 때마다 자동적으로 새로운 화자에 적응되는 알고리즘이다. 그러므로 이 알고리즘을 사용하면 외형적으로 독립 인식 시스템과 같은 역할을 하지만 내부적으로는 화자가 바뀌거나, 혹은 동일한 화자일 경우에라도 인식 작업을 수행할 때 마다 인식 시스템을 사용하는 화자의 발음 환경에 맞게 적응이 된다. 실제로 화자적응 알고리즘의 최종목표는 이와 같은 방식으로 화자 적응이 이루어져야 하지만 현재까지 높게 적응율을 나타내는 알고리즘이 개발되어 있지 않다. CMU의 Stern은 FEATURE시스템에 MAP (maximum a posteriori probabilily)추정 이론을 이용하여 동적 독자 적응 알고리즘을 개발 하였으나 적응 대상 영역의 확장 및 인식율의 개선이 필요하다.[39]

4. 언어 처리

음성 인식기가 처리할 대상이 문장인 경우에는 언어의 문법 및 의미 정보를 이용함으로써 시스템 성능을 향상시킬 수 있다. 일반적으로 문장 음성 인식에 있어서는 음성학적 처리에 의한 인식 오류를 수정하기 위해 상위 레벨 정보인 운율특징(prosodic feature)과 구문론(syntactics)에 의거해 문장 구조를 결정하고 계속해서 의미론(semantics)과 실용론(progmatics)에 따라 최종 인식 결과를 얻게 된다. 음성이 입력되면 먼저 음성학적 해석을 한 후에 문장론과 구문론 등에 입각하여 언어학적 처리가 수행되는 경우를 bottom-up approach라 하고 이와는 반대로 언어학적 처리의 결과를 토대로 필요한 음성학적 해석을 하는 방법을 top-down approach라고 하는데 일반적으로 두가지 방법을 병행해서 사용한다.[40]

그림 4는 음성 인식 시스템중 상위 레벨의 지식을 처리해 주는 부분을 보여 주고있다. 이 그림에서 구문분석은 특정한 순서로 배열된 단어가 문법적으로 맞는 지를 검토하고 문장내에서 특정한 위치에 올 수 있는 단어의 종류도 추정하는데 이용된다. 여기서 문법이란 언어학(linguistics), 자동이론(automata theory) 및 프로그래밍 언어(programming language) 등에서 사용되는 올바른 언어를 표현해 주는 규칙의 집합이다. 이러한 구문 규칙의 구현 방법으로는

phase-structure rule을 사용하는 방식, finite-state 혹은 Markov model을 사용하는 방식, ATN (augmented transition network)을 이용하는 방식 및 production rule을 사용하는 방식등이 있다. 다음으로 의미 분석은 구문론적으로 올바른 문장이 실제로 의미가 있는 지를 결정해 주는 지식을 적용하는 부분으로 화자가 의도하는 의미를 논리적으로 분석한다. 실용분석은 대화를 통하여 문장을 이해할 수 있는 능력을 제공해 주는 부분으로 같은 문장이라도 서로 말하는 사람의 신분과 알고 있는 지식 및 이전에 대화하였던 내용에 따라서 의미가 달라질 수 있기 때문에 언어학적 처리부에서 이러한 정보를 사용하는 것이 바람직 하다. 마지막으로 운율분석이란 강세(stress), 억양(intonation), 정지(pauses) 및 발음속도(timing structure) 등의 분석을 의미하며 이러한 운율 정보를 사용하면 음성 이해력이 향상될 것이라는 가정은 언어처리 연구의 초창기 부터 연구되어 왔지만 현재까지는 운율이 인식 시스템에 성공적으로 구현되지는 못하였다.

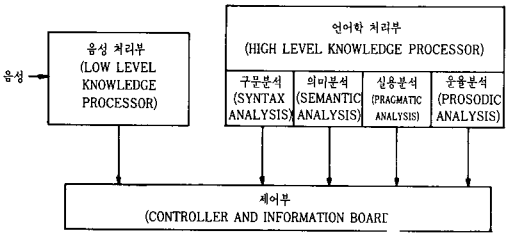


그림 4. 음성 이해 시스템의 구조

언어학적 처리를 수행하는 방향에는 여러가지 방법이 있을 수 있다.[41] 몇가지 대표적인 것을 보면, 문장의 왼쪽에서 오른쪽으로 순차적으로 해석하는 left-to-right method, island라 일컫는 문장의 keyword를 먼저 찾은 후 그 island로 부터 오른쪽 혹은 왼쪽으로 해석하는 island-driven method, 그리고 위의 두가지 방법을 병행해서 사용하는 left hybrid method 등이 있다.

지금까지 기술한 일반화된 언어학적 처리부가 모두 필요한 인식 대상 영역은 natural task이다. 그러나 실제로 구현되어 사용되고 있는 많은 시스템은 언어학적 처리부의 극히 일부만을 사용하는 artificial task를 task domain으로 설정하고 있다. 이 artificial task로서 대표적인 것으로는 new Raleigh language[42] airline information and reservation language[43] 등이 있다. 또한 natural task로서는 business letter의 text[44], patent applications[45] 등이 있는데 여기서는 언어학적 처리를 지식에 토대를 두기 보다는 확률적 모델링을 이용하여 수행하였다. 일반적으로 artificial task는 상대적으로 적은 양의 인식 대상 어휘를 가지는 시스템에 적당하며 미리 정해진 문법에 의해서 언어를 설계하므로 구문분석을 통하여 단어의 오인식을 대폭 줄일 수 있다. 그러나 연구자가 직접 언어를 설계해야 하고 인식 대상 어휘수가 늘어남에 따른 복잡성의 증가로 확장성이 없기 때문에 수천 단어 이상의 어휘에 대해서는 natural task로서 언어학적 처리부를 구성해야 한다.

5. DB 및 시스템

음성 데이타 베이스는 개발된 음성 인식 시스템의 성능을 객관적으로 평가 비교할 수 있으며 음성의 분석과 인식 방법의 개발에 이용될 수 있는 측면과 A/D (analog to digital) 변환기가 없는 연구실에서도 음성 관련 연구를 수행할 수 있게 환경을 조성 시킬 수 있다는 면에서 공동으로 제작 구성되는 경향이 있다.

대표적인 음성 데이타 베이스로는 미 국방성에서 제작한 1000 단어용 DB이다. 이 DB는 DARPA가 지원하는 음성 인식 프로젝트의 성능 평가를 위해 만들었는데 화자 독립, 화자 적응 및 화자 종속 음성 인식 알고리즘에 이용될 수 있다. DB 구성은 BBN, SRI, TI (Texas Instruments), NBS (National Bureau of Standard)의 네 기관이 모여서 각 기관의 상호 협조 아래 제작되었다. BBN은 DB사용 용도, 어휘 선정 및 문장 구성을 담당하였으며 SRI는 화자 선정과 방언, BBN은 제작된 DB의 분배를 분담하였다.[46] 이 DB를 사용한 음성 인식 시스템은 CMU의 ANGEL system, BBN의 BYBLOS system[47] CMU의 SPHINX system[23] 등이 있다.

또다른 DB로는 TI와 MIT가 공동 제작한 TIMIT DB가 있다. TIMIT는 TI와 MIT가 선정한 문장을 630 화자들이 발음한 것으로 구성되어 있는데 화자 독립 음소 인식기의 개발 및 테스트 용으로 만들어 졌다.[23] 최근에는 TIMIT를 시내·외 전화 선로를 통하여 재 녹음을 한 NTIMIT (network TIMIT)가 미국 NYNEX에서 개발 되었다.[48]

일본에서는 자동번역연구소(ATR)에서 음성 데이

타베이스를 구성하고 있으며 영국의 ALVEY 프로그램, 프랑스의 GRECO 프로젝트, 유럽 제국의 ESPRIT 프로젝트 등에서도 음성 DB의 구축에 대한 연구가 활발히 진행되고 있다.

일단 DB가 완성되면 DB의 이용 측면에서도 연구가 진행되어야 한다. 대용량 단어 음성인식 시스템을 구성할 경우 음성의 기본 단위를 단어로 사용하지 않고 음소 혹은 유사 음소를 사용하는데 이것과 화자가 발음한 단어와의 관계를 연구하여야 한다. 이러한 작업을 lexicon table 구성이라고 하는데 CMU의 Lee는 음성학적 지식을 이용하여 lexicon table를 재구성하였으며 이로 인하여 인식율의 향상을 얻었다.[23]

음성 인식 시스템은 앞절에서 기술한 음성의 분석, 인식, 화자 적응, 언어처리 기술 및 DB 구성등이 집대성된 시스템이기 때문에 모든 분야의 연구가 동시에 이루어져야 한다. 음성 인식 시스템을 평가하는데 기준이 되고 있는 관점은 화자 독립 여부, 연속 음성 인식 가능 여부, 인식 대상 단어수, 제한된 문법의 사용 여부에 따라 나눌 수 있는데 현재의 기술 수준으로서는 상기 모든 조건을 완화시키면 인식율이 떨어지게 된다. 이러한 이유로는 음성을 정확히 모델할 수 있는 알고리즘의 미비, 인식 시스템에 음운, 음성 정보의 부적절한 사용, 화자간 혹은 화자내의 변화를 정확히 알 수 없기 때문이다. 그러나 최근에 CMU에서 개발한 SPHINX 시스템은 1000 단어 화자 독립 연속 음성 인식에서 95.8% 인식율을 나타내고 있어 현재까지 개발된 음성인식 시스템중 가장 좋은 성능을 나타내고 있다. 반면 제한된 문법을 사용하고 있기 때문에 인식 대상이 되는 task domain이 미리 선정되어야 한다. 그러므로 현재의 음성 인식 시스템 기술 추세는 미리 선정된 task domain에서 대용량의 연속 단어를 이해하여 요구되는 일을 직접수행 할 수 있는 man-machine interface 용도의 음성 이해 시스템 개발에 역점을 두고 있다. 최근에 개발된 음성 이해 시스템은 다음과 같다.

CMU에서 개발된 SPHINX 시스템은 음소를 기본 단위로 구성한 HMM 알고리즘을 사용하였는데 음소의 조음 현상(coarticulation)을 고려하여 1000 개의 유사 음소를 만들어서 인식율의 향상을 기하였다.

BBN에서 개발한 BYBLOS 시스템은 1000 단어 화자 종속 음성 인식 시스템용으로 개발되어 94.8% 인식율을 나타내고 있으며 화자 적응 과정을 두어 화자 독립적으로 사용할 수 있게 하였다.[47]

IBM에서는 20,000개의 고립 단어를 인식할 수 있는 화자 종속 시스템을 PC에 구성하였으며 200,000 단어로 확장시키고 있다.[49]

일본 ATR에서는 fuzzy VQ와 HMM을 사용하여 1035 단어로 이루어진 화자종속 연속 음성 인식 시스템을 개발하였는데 88.4%의 인식율을 나타냈으며 화자적응을 시킬때는 81.6%의 적응된 화자의 인식율을 나타내었다. 이 시스템은 context-free grammer를 사용하였으며 모음과 자음에 따라 state를 달리한 phone model에 기초를 하였다.[50] 한편, 몇년 전부터 연구되어온 neural network를 이용한 음성인식 시스템은 음소, 혹은 몇개의 고립단어를 인식하고 있는 정도이지만 최근에는 기존의 HMM 방식과 조합하여 시스템을 만들려고 하는 추세이다. 특히 기존의 HMM을 이용한 음성인식 시스템에서 VQ 대신 Kohonen이 개발한 LVQ를 사용하여서 인식율의 향상을 이룩하기도 하였다.

마지막으로 spectrum reading 기술을 이용한 knowledge-based 음성 인식시스템이 미국 MIT에서 개발되었다.[51] 이 시스템은 SUMMIT라고 불리우는데 음성·음향학적 지식을 이용한 화자 독립 연속 단어 음성 인식 시스템으로서 DARPA DB를 이용한 인식 테스트에서 86.4%의 인식율을 나타내었다.

Ⅲ. 국내 연구 현황

1. 학계 연구 활동

국내에서의 음성인식 연구는 그 역사가 매우 짧아 약 7년에 불과해서 선진외국에 비해 10년이상 늦게 연구가 시작되었지만 국내 여러 대학과 연구소에서 꾸준히 연구를 지속한 결과 음성 인식의 기본 기술이 구축되었다고 할 수 있다. 본절에서는 국내에서의 한국어 음성인식 기술 연구중 학계의 연구 활동을 살펴보고자 한다.

한국어 음성인식 기술에 대한 본격적인 연구는 한국과학기술원 통신연구실에서 한국전기통신공사의 지원을 받아 1984년 부터 시작되었다.[52-57] 1차년도에는 잡음이 섞인 음성의 개선과 음성 발생 모델로 부터의 성도계수 검출에 대하여 연구하였고, 격리단어 인식을 위해 vector quantization과 matrix quantization을 적용하였다. 2차년도에 격리단어 인식을 위해 시간 정보를 고려한 finite-state VQ 알고리즘을 적용하였고, 연결단어 인식을 위해 단어를 유사음소의 연결로 생각하여 이에 DTW 알고리즘을 적용한 결과를 보고하였다. 또한 DTW 알고리즘에서의 계

산량 축소 방법과 dynamic reference updating 방법에 의한 화자독립성 연구도 수행하였다. 3 차년도에는 한국어 음소인식에 적응 알고리즘인 recursive least squares(RLS) 알고리즘을 적용하여 실험하였고, 격리단어 인식을 위해 finite-state VQ 알고리즘과 HMM 모델링을 결합시킨 방법을 연구하였다. 또한 화자의 인식을 위한 기초연구를 수행하였고, DTW 알고리즘을 이용한 격리단어 인식 기술을 hardware로 구현한 음성인식 전화기를 개발하였다. 4 차년도에는 음소분할 연구를 위해 formant tracking방법을 적용하고, HMM 모델링을 이용한 음소인식 알고리즘에 대해 연구하였다. 또한 대상 어휘수가 200 단어인 연결단어 인식 알고리즘에 대해 연구를 수행하였는데, 이 알고리즘은 level-building DTW 알고리즘에 문법적인 제한을 가한 것으로서 자동 전화번호 안내 시스템으로 구현되었다. 5 차년도에는 어휘수를 대폭 확장하여 1200 단어 인식 시스템을 개발하였는데 이것은 음소의 HMM 모델링을 기본으로 하는 화자종속 문장 인식 시스템이었다. 이 시스템에서는 인식 시간의 단축을 위해 rule-based 알고리즘을 적용하여 후보단어 수를 줄여 주는 처리 과정을 부가하였고, 단어 단위의 인식결과로 부터 문장 전체의 의미적 인식율을 높여주기 위한 언어학적 처리부도 포함되었다. 또한 이와는 별도로 화자 독립성을 부여하기 위하여 음소 모델의 화자 적응 기법에 대한 연구와 음소의 자동 분할을 위한 음소분리 알고리즘도 연구하였다. 마지막으로 6 차년도에는 전년도에 개발된 1200단어 인식 시스템의 음성학적 처리부 성능 향상을 위한 연구와 화자적응 알고리즘의 개선 및 시스템과의 통합, 그리고 인식시간 감축 알고리즘의 개선 연구가 수행되었다.

과학기술원에서 현재 연구중인 대용량 단어 및 문장 인식 시스템의 전체적인 흐름도가 그림 5에 나타나 있다. 먼저 인식할 대상 어휘가 결정되면 이로부터 각 단어의 발음 사전이 구성된다. 발음사전은 음소의 연결로서 표시되며 이때 음운론적인 규칙이 고려되어야 한다. 이와는 별도로 단어의 연결로 이루어진 문장의 문법으로 부터 구문해석기를 설계한다. 각 음소의 기준 모델을 결정하기 위해서는 수십개의 phonetic balanced words로 부터 음소들을 분류해 내고 이를 사용하여 각 음소의 HMM 모델을 훈련한다. 한편 입력 음성과 비교할 후보단어의 수를 줄이면 인식 시간을 감축할 수 있는데, 이를 위해 음소군 분류 및 단어군 분류 알고리즘을 개발하였다. 이

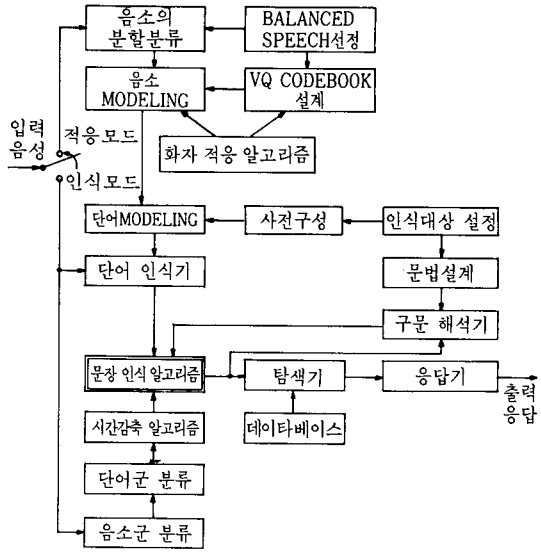


그림 5. KAIST의 대용량 단어 및 문장 인식 시스템

와같이 하여 후보단어가 결정되면 이 단어의 발음사전을 참조하여 음소 모델의 연결로 이루어진 단어 모델을 구성하고 이를 입력음성과 Viterbi scoring을 통해 likelihood를 계산한다. 인식대상 화자가 바뀌면 그 새로운 화자에 대해 음소 모델 파라메터를 바꾸어야 하는데 이를 위해 화자적응 알고리즘을 적용하고 있다.

한편 서울대 전자공학과에서는 음성신호의 분석을 위해 음성신호의 프레임당 평균 진폭의 분포를 이용한 새로운 끝점 검출 알고리즘을 개발하였고, 비강 및 방사 임피던스의 효과를 기존 성도 모델에 포함시킨 일반화된 성도 모델을 결정하고 이 모델로 부터 pole-zero 선형예측 모델을 유도하였다. 음소 분류 인식을 위해서는 간략한 에너지 곡선을 이용한 새로운 비음구간 검출 알고리즘과 종성 내파음 검출 알고리즘을 개발하였다. 연속 숫자음 인식에서는 DTW의 일종인 UELM(unconstrained endpoint local minimum)을 적용하였고, 고립단어 인식 연구에서는 자동으로 초기화되는 K-means clustering 알고리즘, 교정 학습의 HMM, 시간 압축을 사용한 HMM, 그리고 고립단어의 새로운 모델로서 분할 확률모델을 제시하고 인식 실험을 통하여 기존 방법들과 비교하였다. 또한 음소 단위인식을 위해 음성학적 지식의 응용과 신경 회로망의 응용 등을 연구하고 있으며, 소

단위 인식의 기초로서 분할(segmentation) 알고리즘의 연구도 수행되고 있다.[58]

연세대학교의 음향, 음성 및 신호처리 연구실에서는 최근 자동차 소음 환경에서 선형예측 방법에 기초한 네가지 스펙트럼 matching 방법, 즉 log likelihood ratio, LPC cepstrum, spectral slope distance measure, weighted cepstral distance measure 등을 적용하여 숫자음 인식을 수행하였다.[59] 또한 Itakura-Saito distance measure를 이용하여 한국어 음소의 분리 및 인식에 관한 연구도 발표하였다.[60]

광운대학교에서는 146개의 전국 DDD 지역명을 인식하기 위해 12차 LPC cepstrum 계수에 시간 정보를 가지는 multi-section VQ 알고리즘을 적용하여 화자독립 인식 시스템을 구현하였다.[61] 성균관대학교 전자공학과에서는 패턴매칭을 이용한 격리단어 인식에 관한 연구를 수행하였고, 최근에는 연속음성 인식을 위한 음소 단위의 인식에 관심을 집중시키고 있다. 이를 위해 LSP 파라메터를 이용한 음소의 분할 및 분류 연구를 수행하고 이를 단어 인식으로 확장하는 연구를 수행하고 있다.[62] 건국대학교 전자공학과에서는 LSP 방식에 의한 음소 분석과 연결단어 인식에 대해 연구하였고, OSDP 알고리즘을 이용한 한국어 연속 숫자음 인식도 수행하였다.[63] 경희대학교 전자공학과에서는 최근 전화선을 통한 화자의 검증을 위해 전화시스템을 모델링하고 이로부터 원음을 복원하여 검증하는 방법을 연구하였다.[64] 영남대학교 전자과에서는 formant 분석을 통한 한국어 단모음 및 파열음에 대해 연구하였고, 이를 토대로 음소를 인식 단위로 하는 소규모 특정화자 인식 시스템 개발을 수행하고 있다.[65] 그외 한국과학기술원 전산과에서는 spectral-peak-weighted binary spectrum을 이용한 한국어 비음인식에 관한 연구를 수행하였다.[66]

지금까지의 학계 연구동향을 살펴보면, 음소 단위나 단어 단위의 인식 연구가 여러가지 알고리즘으로 구현되어서 그 결과가 발표되고 있지만 그 성능은 아직 높은 수준에 이르지 못하고 있으며 표준적인 음성 데이타베이스가 구축되어 있지 않아서 연구 결과에 대한 비교 평가가 제대로 이루어지지 않고 있다. 앞으로는 공통적인 음성 데이타베이스의 구축과 이를 이용한 보다 깊이있는 한국어 음성 특성에 대한 연구 및 인식 알고리즘 연구, 그리고 언어학적 처리 연구에 학계가 공동으로 연구 노력을 집중해야 할 것이다.

2. 연구소

음성 인식 기술은 상품의 부가 가치를 높일 수 있는 기술이므로 국가의 연구소 뿐만 아니라 기업에서도 연구소를 중심으로 하여 활발한 연구를 진행하고 있으나 아직 상품화 단계에 이른 것은 매우 적은 형편이다. 우선 한국전자통신연구소에서 진행되고 있는 음성 인식에 관한 연구를 살펴보면 그동안 음소나 격리단어의 인식에 관한 연구를 수행하여온 전자통신연구소는 최근에 대어휘 연속음성 인식을 위한 음소인식을 기술을 개발하고 있다. 이 연구의 최종 목적은 phoneme-balanced word 음소단위 DB 구축과 음소 단위에 의한 대어휘 음성인식 기술 개발에 있으며 이를 위하여 단음절 음성 데이타베이스, 단음절 지속시간 분포조사, 어두 파열음의 분할 알고리즘 개발, 음소 특징 추출을 위한 지각 실험 시스템 개발이 이루어 졌다. 또한 음운 balance 단어에 대한 DB가 구성중에 있으며 spectral의 회귀 분석을 이용하거나 모음의 평균 패턴과의 상관을 이용한 단음절 segmentation 방법에 관한 연구가 수행되었다. 그리고 HMM을 이용한 연결 단어 인식 시스템, Hopfield network을 이용하는 단모음 인식 방법과 화자적응 방법에 관한 연구도 수행하는등 활발한 연구 활동을 계속하고 있다.[67]

한국전기통신공사의 연구개발단에서도 음성 인식에 관한 연구를 수행하고 있는데 이곳은 그동안 한국과학기술원에서 수행한 연구과제의 결과를 전수받고 이를 바탕으로 연구를 계속할 예정으로 있다. 전기통신공사는 음성인식 기술을 이용하는 제품의 실용화 보다는 당분간은 음성 인식 기술에 관한 기초 연구를 수행할 것으로 알려져 있다.

기업의 연구소에서도 음성인식에 관한 연구가 진행되고 있으며 그 중 대표적인 연구를 살펴보면 다음과 같다. 삼성종합기술원에서는 숫자와 지명을 포함하는 30여 단어를 화자 종속으로 인식할 수 있는 시스템을 개발하여 상품화를 추진하고 있다. 이 인식 시스템의 특징은 잡음제거 알고리즘을 채택하고 있어서 잡음과 음성이 섞여있는 상황에서도 좋은 인식 성능을 보여준다는 점이다.

금성사 중앙연구소에서는 지난 2년간 격리단어 인식 시스템에 관한 연구를 수행하였다. 그 결과로 200 단어를 인식할 수 있는 화자종속 음성인식 시스템이 디지탈 신호처리 chip을 이용하여 PC plug-in board의 형태로 구현되었다. 현재는 이를 바탕으로 연결 숫자음을 인식할 수 있는 음성 인식 시

스템을 개발하는 것을 목표로 하여 연구가 진행되고 있다.

마지막으로 주식회사 디지콤의 정보통신연구소에서는 화자 종속 또는 독립으로 최대 200 단어를 인식할 수 있는 음성인식 시스템을 개발하였다. 상공부에서 시행한 공업기반 기술 개발 사업의 일환으로 개발된 이 시스템은 특정한 화자가 최대 200 단어까지 임의의 어휘를 선정하여 인식을 수행할 수 있으며 복수의 화자나 임의의 화자도 사용할 수 있도록 지원하는 별도의 software가 갖춰져 있다. 이 시스템은 여러가지 용도로 사용할 수 있으나 특히 컴퓨터를 keyboard의 조작없이 음성으로 동작시킬 수 있는 음성 단말기로 사용될 수 있도록 설계되었다.

Ⅳ. 결 론

음성에 의한 man-machine interface 기술의 핵심이 되는 음성 인식 기술의 최근 동향을 대용량 단어 음성 인식 시스템 기술을 이루고 있는 분야별로 나누어서 기술하였다.

음성의 특징을 추출하는 음성 분석 기술의 최근 동향은 음성의 발생 과정을 단지 모델링하여 음성의 특징을 추출하는 단계에서 벗어나 음성의 인지 과정을 모델링 하고 잡음이 섞인 음성으로부터도 음성 특징을 추출하여 높은 인식율을 얻는 방향으로 바뀌고 있으며, 추출된 특징으로부터 음성을 인식하는 인식 기술은 몇개의 단어를 인식할 때 적합한 DTW 기술 보다 대용량 단어 인식에 적합한 HMM 기술을 사용하는 추세로 바뀌어 가고 있다. 특히 최근에는 neural network를 이용한 음성인식 기술이 개발되었으나 음소, 혹은 몇개의 단어를 인식할 수 있는 정도이며 기존의 HMM 기술과 결합시켜 대용량 단어 인식에 사용할 수 있도록 연구가 진행되고 있다.

화자 적응 기술은 화자 독립인식 시스템의 성능이 화자 종속 인식 시스템에 비하여 떨어지기 때문에 그 대안으로 최근에 연구되는 기술인데 화자 종속 인식 시스템을 기본으로 하여 새로운 화자가 발음한 짧은 음성을 이용하여 기존 시스템의 파라미터로 변경하는 방식을 채택하고 있다. 언어 처리 기술은 음성인식 과정에 언어학적 지식을 이용한다면 인식율의 개선을 이룰 수 있다는 전제하에 연구되어 왔는데 연속 음성 인식에 활용되고 있다. 이러한 기술등을 통합시킨 음성 인식 시스템은 여러 나라에서 개발되었는데 가장 높은 인식율은 나타내고 있는 음성 인식 시스템은 CMU의 SPHINX 시스템이다. 음성 인식을 위한 DB의 기술은 국가의 주도로 표준화되어 개발되는 경향이 있는데 개발된 DB는 인식 시스템의 성능 테스트에 이용된다.

한편 국내의 연구 동향을 살펴보면 학계에서는 KAIST가 대용량 연결 단어 화자 종속 시스템을 개발하였으며 그외는 대부분 단어 혹은 음소 인식 기술에 머무르고 있다. 연구소는 주로 상품화를 위한 연구개발과 기초 연구에 주력하고 있는 실정인데 주로 고립된 단어로 인식하는 수준이다. 그러나 최근 KAIST를 중심으로 음성정보연구센터가 운영되어 관련 기술의 연구가 활성화 되고 있다.

이상에서 살펴본 것처럼 한국에서 한국인에 의해 개발되어야 한다는 특수한 환경아래 세계적인 인식 기술에 비해 많이 뒤떨어진 한국어 음성 인식 기술이 부단히 발전하기 위해서는 관계 기관의 끊임없는 협조와 지원이 뒤따라야 겠고 무엇보다도 관련 기술을 갖고 있는 연구인들의 상호 협조 및 공동 연구 노력이 있어야 겠다.

參 考 文 獻

[1] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, N.J., Prentice Hall, 1978.

[2] J. Junqua and H. Wakita, "A Comparative Study of Cepstral Lifters and Distance Measures for all Pole Models of Speech in Noise," ICASSP, s10a. 3, 1989.

[3] K.K. Paliwal, "A Study of LSF Representation for SD and SI HMM-Based Speech Recognition Systems," ICASSP, s15a. 13, 1989.

[4] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, no. 4, Aug. 1980.

[5] M.J. Hunt and C. Lefebvre, "Speaker Dependent and Independent Speech Recognition Expreiments with an Auditory Model," ICASSP, s5.9, 1988.

[6] M. Nishimura, "HMM-Based Speech Recognition Using Dynamic Spectral Feature," ICASSP, s6.12, 1989.

[7] R.M. Gray, "Vector Quantization," IEEE ASSP Magazine, Apr. 1984.

[8] T. Kohonen, et al., "Statistical pattern

recognition with neural networks: Benchmarking studies," *IEEE , Proc. of ICNN,* vol. 1, pp. 61-68, July 1988.

[9] J. Mariani, "Recent Advances in Speech Processing," Proc. of ICASSP, s9.1, 1989.

[10] C. Myers, L.R. Rabiner and A.E. Rosenberg, "Performance trade-offs in dynamic time warping algorithms for isolated word recognitions," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-28, pp. 623-635, Dec. 1980.

[11] J.G. Wilpon and L.R. Rabiner, "A modified K-means clustering algorithm for use in isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 587-594, June 1985.

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

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

[14] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoust., Speech Signal Processing,* vol. ASSP-32, pp. 263-271, Apr. 1984.

[15] C.S. Myers and L.R. Rabiner, "A level building dynamic time warping algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-29, pp. 284-297, Apr. 1981

[16] L.R. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, Jan. 1986.

[17] L.R. Bahl, F. Jelinek, and R.L. Mercer, "A maximum likelihood approach to contiuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.* vol. PAMI-5, pp. 179-190, 1983.

[18] L.E. Baum et al., "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.,* vol. 41, no. 1, pp. 164-171, 1970.

[19] L.R. Bahl et al., "A Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," Proc of ICASSP, pp. 47-52, 1986.

[20] L.R. Bahl et al., "A New Algorithm for the Estimation of Hidden Markov Model Parameters," Proc. of ICASSP, pp. 493-496, 1988.

[21] K.F. Lee and H.W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-37, pp.1641-1648, 1989.

[22] B.H. Juang and L.R. Rabiner, "Mixture autoregressive HMM for speech signals," *IEEE Trans. Acoust.,Speech, Signal Processing,* vol. ASSP-33, pp. 1404-1413, 1985.

[23] K.F. Lee, *Automatic Speech Recognition,* Kluwer Academic Publishers, 1989.

[24] D.E. Rumelhart, J.L. McCelland and the PDP Research Group, *Parallel Distributed Processing,* The MIT Press, 1986.

[25] D.E. Rumelhart, G.E. Hinton, and R.J. Willams, "Learning Internal Representations by Error Propagation," in D.E. Rumelhart and J.L. McCelland (Eds.), *Parallel Distributed Processing*, The MIT Press, 1986.

[26] R.P. Lippmann , "An Introduction to Computing with Neural Nets," IEEE ASSP Magazine, April 1987.

[27] M. Miyatake, H. Sawai and K. Shikano, "Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time Delay Neural Networks," Proc. of ICASSP, s8.10, 1990.

[28] E. McDermott and S. Katagiri, "Shift-Invariant, Multi-Category Phoneme Recognition Using Kohonen's LVQ2," Proc. of ICASSP, s.3.1, 1989.

[29] T. Kohonen, "The neural phonetic typewritter," *IEEE Computer,* vol. 21, no. 3, pp. 11 22, March 1988.

[30] H. Bourlard, C.J. Wellekens, "Speech Dynamics and Recurrent Neural Networks," Proc. of ICASSP, s1.9, 1989.

[31] L.T. Niles and H.F. Silverman, "Combining Hidden Markov Model and Neural Network Classifiers," Proc. of ICASSp, s.8.2, 1990.

[32] V.W. Zue, "The use of speech knowledge in

automatic speech recognition," *Proc. IEEE*, vol. 73, no. 11, Nov. 1985.
[33] V.W. Zue et al., "Acoustic Segmentation and Phonetic Classification in the SUMMIT System," Proc. of ICASSP, s8.1, 1989.
[34] R. Schwartz et al., "Rapid Speaker Adaptation Using a Probabilistic Spectral Mapping," ICASSP '87, pp. 633-636, 1987.
[35] S. Nakamura et al., "Speaker Adaptation to HMM and Neural Networks," ICASSP '89, pp. 89-92, 1989.
[36] M. Nishimura et al., "Speaker Adaptation Method for HMM-Based Speech Recognition," ICASSP '88, pp. 207-210, 1988.
[37] Y. Grenier et al., "Spectral Transformations Through Canonical Correlation Analysis for Speaker Adaptation in ASR," ICASSP '86, pp. 2659-2662, 1986.
[38] S. Furui, "Unsupervised Speaker Adaptation Method Based on Hierachical Spectral Clustering," ICASSP '89, pp. 286-289.
[39] R.M. Stern, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. ASSP-35, pp. 751-763, 1987.
[40] W.A. Lea, *Trends in Speech Recognition*, Englewood Cliffs, N.J., Prentice-Hall, 1980.
[41] F. Fallside and W.A. Woods, *Computer Speech Processing*, Englewood Cliffs, N.J., Prentice-Hall, 1975.
[42] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, no. 4, pp. 532-556, April 1976.
[43] S.E. Levinson, "The effects of syntactic analysis on word recognition accuracy," *BSTJ*, vol. 57, no. 5, pp. 1627-1644, May-June 1978.
[44] F. Jelinek, "The development of an experimental discrete dictation recognizer," *Proc. IEEE*, pp. 1616-1624, Nov. 1985.
[45] L.R. Bahl, et al., "Recognition of a Continuously Read Natural Corpus," ICASSP, pp. 422-424, 1978.
[46] P. Price et al., "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," ICASSP '88, pp. 651-654.
[47] Y.L. Chow, et al., "BYBLOS: The BBN Continuous Speech Recognition System," ICASSP '87, pp. 89-92, 1987.
[48] C. Jankavski, et al., "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database," ICASSP '90, pp. 109-112, 1990.
[49] F. Jilinek et al., "Experiments with the Tangora 20,000 Word Speech Recognizer," ICASSP '87, pp. 701-704, 1987.
[50] T. Hanazawa et al., "ATR HMM-LR Continuous Speech Recognition System," ICASSP '90, pp. 53-56, 1990.
[51] V. Zue, et al., "The SUMMIT Speech Recognition System: Phonological Modelling and Lexical Access," ICASSP '90, pp. 49-52, 1990.
[52] 은종관 외, "디지탈 음성처리 기술 연구 개발," 1 차년도 최종보고서, 한국과학기술원, 1984.
[53] 은종관 외, "디지탈 음성처리 기술 연구 개발," 2 차년도 최종보고서, 한국과학기술원, 1986.
[54] 은종관 외, "한국어 음성인식 시스템 개발 연구," 3 차년도 최종보고서, 한국과학기술원, 1987.
[55] 은종관 외, "한국어 음성인식 시스템 개발 연구," 4 차년도 최종보고서, 한국과학기술원, 1988.
[56] 은종관 외, "한국어 음성인식 시스템 개발 연구," 5 차년도 최종보고서, 한국과학기술원, 1989.
[57] 은종관 외, "한국어 음성인식 시스템 개발 연구," 6 차년도 최종보고서, 한국과학기술원, 1990.
[58] 성굉모 외, "서울대학교 전자공학과 음성 신호처리 연구 현황," 음성통신 및 신호처리 Workshop 1989.
[59] 차일환, 윤대희 외, "스펙트럼 매칭 방법에 따른 소음 환경에서의 단독음 인식," 음성통신 및 신호처리 Workshop 1989.
[60] 윤대희, 차일환 외, "음소를 이용한 한국어 음성 신호의 분석과 인식에 관한 연구," 한국음향학회지, vol. 8, no. 5, pp. 70-77, 1989.
[61] 김순협, 조형제 외, "시간 정보와 VQ를 이용한 DDD 지역명 인식에 관한 연구," 한국음향학회지, vol. 8, no. 5, pp. 102-111, 1989.
[62] 박병철 외, "성균관대학교 전자공학과 음성처

리 연구실의 연구현황," 음성통신 및 신호처리 Workshop 1989.
[63] 김형래 외, "건국대학교 전자공학과 음성인식 연구 현황," 음성통신 및 신호처리 Workshop 1989.
[64] 진용옥 외, "전화 음성의 화자 검증을 위한 파형복원," 음성통신 및 신호처리 Workshop1989.
[65] 정현열 외, "음소를 단위로 한 한국어 음성인식 장치 개발을 위한 기초 연구," 음성통신 및 신호처리 Workshop 1989.
[66] K.C. Kim, H.S. Lee and J.W. Cho, "Phonetic Recognition Using Peak Weighted Binary Spectrum," Proc. ICASSP-89, pp. 330-333, 1989.
[67] 김경태 외, "대어휘 연속음성 인식을 위한 음소인식 기술 개발," 최종연구보고서, 9ST5200231230F, 한국전자통신 연구소, 1990.

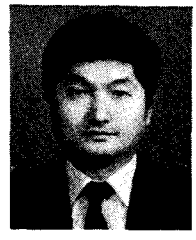
감사의 말씀

본 논고는 지난 5년간 한국전기통신공사의 지원으로 연구된 내용의 일부입니다. 한국어 음성인식 연구를 지원하여 주신 전기통신공사에 심심한 사의를 표합니다. 또한 이 연구를 위해 불철주야 노력을 한 KAIST 통신연구실의 연구원, 박사 및 석사 학생들에게도 아낌없는 치하의 말씀을 드립니다.

筆者紹介

殷鍾官

1940年 8月 25日生
1964年 6月 미국 Univ. of Delaware 전자공학과 졸업 (공학사)
1966年 6月 동대학원 졸업 (공학석사)
1969月 6月 동대학원 졸업 (공학박사)
1969年 9月～1973年 5月 미국 Univ. of Maine 전자공학과 조교수
1973年 5月～1977年 6月 미국 스텐포드연구소 (SRI) 책임연구원
1983年 7月～1989年 6月 한국과학기술연구원 통신공학연구실장
1977年 6月～현재 한국과학기술원 전기 및 전자공학과 교수



李愃洙

1952年 9月 19日生
1975年 2月 서울대학교 공과대학 전기공학과 (학사)
1978年 8月 한국과학기술원 전기 및 전자공학과 (공학석사)
1983年 2月 한국과학기술원 전기 및 전자공학과 (공학박사)
1975年 1月～1975年 10月 현대조선중공업(주) 설계부 사원
1983年 3月～1989年 2月 한국과학기술원 전기 및 전자공학과 조교수
1984年 4月～1985年 5月 미국 Stanford 대학교 Information Systems Lab. Post Doc. 연구원
1989年 3月～현재 한국과학기술원 전기 및 전자공학과 부교수

丘 明 完

1960年 4月 26日生

1982年 2月 연세대학교 전자공학과 졸업(공학사)

1985年 2月 한국과학기술원 전기 및 전자공학과 졸업 (공학석사)

1988年～현재 한국과학기술원 전기 및 전자공학과 (박사과정)

1985年～현재 한국전기통신공사 연구개발단 전임연구원

●

具 俊 謨

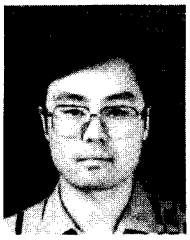
1963年 1月 13日生

1985年 2月 서울대학교 전자공학과 졸업(공학사)

1987年 2月 한국과학기술원 전기 및 전자공학과 졸업 (공학석사)

1987年～현재 한국과학기술원 전기 및 전자공학과 (박사과정)

1987年 7月～현재 (주)디지콤 정보통신연구소 전임연구원



金 會 麟

1961年 3月 9日生

1984年 2月 한양대학교 전자공학과 졸업(공학사)

1987年 2月 한국과학기술원 전기 및 전자공학과 졸업 (공학석사)

1987年～현재 한국과학기술원 전기 및 전자공학과 (박사과정)

1987年 11月 한국전자통신연구소 입소