특집

한국어 음성인식

정 차 균
포스데이타(POSDATA) 기술연구소

요 약

한국어 음성인식기술은 1980년대를 전후로 비약적인 발전을 하여왔다. 본 논문에서는 한국어 음성인식에 적용된 기술을 검토하고, 문제점을 고찰해서 향후 한국어 음성인식의 나아갈 방향을 조명해 본다.

I. 서 론

사람이 사용하는 음성은 여러 의사소통 수단 중 가장 자연스러운 통신수단이며, 인간과 인간의 의사소통에서 음성보다 효율적인 매체는 존재하지 않는다. 사람이 사용하는 음성을 컴퓨터로 하여금 자동으로 인식하게 하는 기술을 통칭해서 음성인식이라 한다. 기술적으로는 인간의 발성기관에 의해 발생된 물리적인 음성신호 파형에 내재해 있은 의미 있는 정보(화자의 발성의도)를 컴퓨터로 하여금 자동으로 인식하게 하는 기술로 음성신호처리의 한 분야이다.

음성인식시스템의 성능은 화자종속/독립, 인식단어 수, 고립단어/연속음성인식에 의해서 크게 구분하고 있다[1]. 화자종속 시스템은 훈련된 화자의 특색에 맞게 시스템이 설계되어, 화자독립시스템보다 정확도 면에서는 우월하나, 다수의 화자를 인식하여 서비스하는 곳에서는 편리성 및 기능 면에서 화자독립인식시스템이 반드시 필요하다. 한편 양자의 특색을 반영한 화자적응 방법은 화자독립시스템을 특정한 사용화자에 대해서 실시간에 적응하게하여 인식률을 높이는 시스템이다. 인식단어 수는 보통 1000 단어 이상을 인식하는 시스템을 대단위 인식시스템이라 한다. 단어수가 증가함에 따라, 인식난이도가 로가리듬(logarithm)하게 증가하는데, 이는 유사어의 증가 및 복잡도(perplexity)의 증가에 기인한다. 또한 최근의 인식단어 문제에서 인식시스템에 모델링된 단어 이외의 outlier 해결 방법에 대한 많은 연구가 진행되

고 있다.

고립단어와 연속음성인식은 통상 단어와 단어 사이가 200 msec의 휴지 공간이 있을 때를 기준으로 구분한다. 음성인식 알고리즘은 실시간에 인식 결과를 출력해야 하는데, 200 msec의 휴지 공간은 패턴매칭의 탐색영역(search space)을 축소해 줌으로써 알고리즘의 설계를 간단히 할수있게 한다. 또한 연속음성신호에서는 단어와 단어 사이에서 발생하는 조음현상(coarticulation)으로 인해 음성신호가 크게 변화한다. 조음현상을 보다 정밀히 모델링하기 위해서 단어자체에 대한 모델과 더불어 단어와 단어의 천이과정을 별도로 모델링하는 문맥모델(context)이 주로 사용된다.

단어인식과정에서는 단어와 단어가 결합할 수 있는 구문규칙(syntax knowledge)을 이용하여 엔트로피를 줄여 인식하는 디코딩기법이 주로 사용된다[3]. 현재는 규격화된 문장인식이 아닌, 자연스런 대화체(spontaneous speech)를 인식하려는 시도가 활발히 이루어 지고 있으며, 영어에 적용된 기술을 한국어의 특색에 맞게 변형하여 적용하고 있다.

한국어 음성인식 연구에서 영어에 적용된 인식기법을 한국어에 적용하기 어려운 부분은 언어영역이 포함되는 언어모델링 단계부터이다. 언어 모델링이 포함되지 않는 시스템은 언어모델을 사용하는 시스템에 비해 엔트로피를 줄이지 못해, 문장단위의 자연스러운 연속음성에 있어 인식시간 및 인식률에 있어 많이 저하된다. 또한 특수한 영역의 사전지식을 효과적으로 이용한 특수한 용도의 음성인식시스템에 대한 실용화 연구도 활발히 진행되고 있으며, 자동차 운전 중의 음성다이얼링 및 음성명령시스템과 같은 특수한 응용 분야에서는 음성인식의 효율이 점차 입증되어 실용화되고 있다.

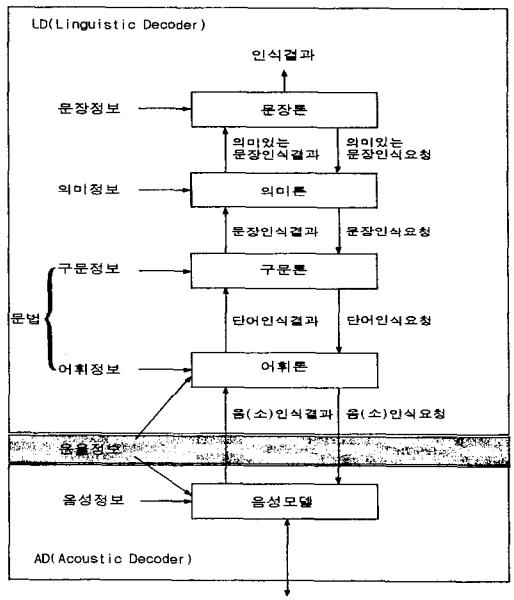
현재의 한국어 음성인식은 몇 명의 화자종속/독립 단어인식 시스템이 주를 이루고 있으며, 실험실 레벨에서 화자독립/연속 시스템에 대한 연구가 진행되고 있다. 인식하는 단어 수는 10여년전에 비해 현격하게 늘었지만, 아직 실환경에서의 화자독립과 연속인식의 문제에 대한 해결은 요원한 상태이다. 본 논문에서는 현재의 음성인식시스템에 사용되는 요소 기술의 개념을 살펴보고, 향후 한국어 음성인식의 연구방향에 대해 알아본다.

II. 음성인식 시스템

본 장에서는 음성인식 시스템에 대한 개괄적인 설명을 한다. 음성인식시스템이 어떻게 구성되는가와 음성모델링 부분에 대한 기본적인 이론에 대해 설명을 한다.

1. 음성인식시스템의 구조

그림 1은 일반적인 음성인식시스템의 구조[1]를 나타낸 것이다.



〈그림 1〉 음성인식시스템 구조

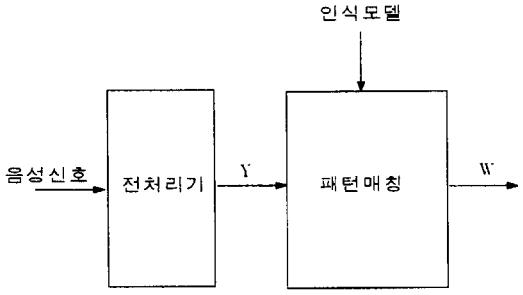
음성인식은 음성신호로부터 음소/음절 혹은 단어를 인식하는 AD(Acoustic decoder)와 AD의 인식결과와 언어학적인 정보를 종합해서 문장을 인식하는 LD(Linguistic decoder)로 구성되어 있다. 일반적으로 AD분야를 연구하는 학문을 음성인식(speech recognition)이라하며, LD분야를 다

루는 분야를 NLP(natural language processing)라 한다. 두 분야는 독자적으로 연구되는 경향이 있으며, 두 분야의 인터페이스에 대한 많은 연구가 진행 중이다.

음성인식방법에는 top-down 방식과 bottom-up 방식이 있다. Top-down 방식은 AD에서 출력한 인식 값을 LD가 언어학적 정보를 종합해서 문장을 인식하는 방법이며, bottom-up 방식에서는 LD가 문법에 맞는 여러 문장을 생성하고, 이를 AD가 문장과 음성신호의 유사도를 계산하여 문장을 인식하는 시스템이다. 실시간에 인식하기 위해서는 문장단위의 문맥정보(문장론), 단어의 의미, 문법규칙을 망라한 언어모델에 대한 시스템이 필요하다. 하지만 현재의 음성인식시스템에서는 문법규칙을 이용하는 정도이며, 단어의 의미를 사용하는 음성이해(speech understanding)시스템에 대한 연구는 미진한 편이다.

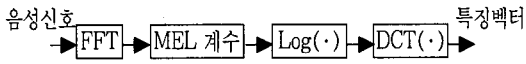
2. AD의 구조

음성인식은 발음된 음성패턴이 주어졌을 때, 인식모델과 패턴매칭을 하여, 가장 근접한 모델의 계수(W)로 인식하는 과정이다



〈그림 2〉 음성인식시스템의 구조. Y는 특징벡터열을 나타내며, W는 인식시스템에서 인식하는 인식단위의 결과를 나타낸다.

음성신호는 ADC(sampling/quantization)을 거쳐, 음성신호전처리기로 입력된다. 음성신호전처리기는 시간도메인(time-domain)의 음성신호를 주파수도메인(frequency domain)으로 변환하여, 음성에 내재하는 정보가 다음 단의 인식기에서 보다 효과적으로 인식하도록 변환하는 처리기이다. 시간도메인의 음성신호는 화자의 발성환경 차이에 따라 변화가 심하여, 음성신호에 내재하는 정보를 처리하는데 주파수 영역의 신호가 사용된다. 주로 사용되는 전처리기는 FFT, LPC, cepstrum이며, 이들의 시간적 변화를 입력을 사용하기도 한다. [4]에서는 여러 전처리 방법에 대한 비교가 되어 있으며, 가장 많이 사용되는 MEL 스케일 기반의 특징벡터를 계산하는 방법을 그림 3에 도시하였다.



〈그림 3〉 음성신호 전처리기(특징벡터 생성)

여기서, MEL 계수는 $m(f)=1000\cdot\log(1+\frac{f}{1000})$에 의해서 변환된 MEL 주파수 영역에서, 균등하게 분배된 영역으로부터 특징벡터를 획득하게 된다. DCT(discrete cosine transform)는 cosine 변환을 말하며, KLT 변환의 근사적 변환으로 보통 이미지 압축에 많이 사용되는 변환 방식이다.

III. ADs(Acoustic Decoders)

음성인식에 사용되는 음성인식알고리즘은 DTW, HMM, NN 등이 있으며, 기본적인 이론은 패턴매칭에 기인한다. 현재는 HMM과 NN의 장점을 결합한 NN-HMM의 복잡형(hybrid)이 많이 연구되고 있다.[5, 6]

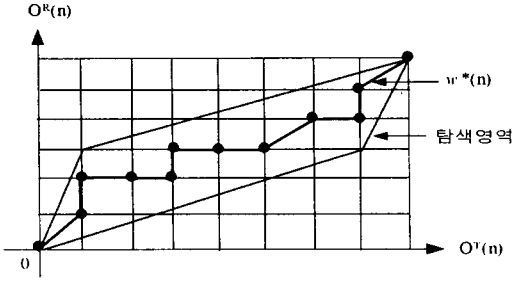
1. DTW(Dynamic Time Warping)

초창기의 음성인식알고리즘은 주로 DTW[1, 3]에 의해서 연구되어 왔고, 근래도 소규모의 인식시스템에 적용되고 있다. 같은 화자가 같은 환경에서 같은 단어를 발음하더라도 단어의 길이가 심하게 변화한다는 현상을 해결하기 위해서 도입되었다. 이론은 최적화 이론(principle of optimality)에 기

인하며, 다음과 같은 수식으로 정리된다.

$$R^*=\arg\min_{(R,\,w(n))}\|O^R(w(n))-O^T(n)\| \quad (1)$$

위 수식에서 R은 인식단어 참조 모델의 인덱스이고, O^R은 참조모델의 특징벡터이다. 그리고 O^T는 입력되는 음성의 특징벡터 열이고, $w(n)$는 O^R과 O^T의 최적의 매칭을 결정하는 굴곡함수(warping function)이다.



〈그림 4〉 DTW의 기본원리

그림 4는 최적의 $w(n)$를 구하는 과정을 나타낸 것이다. 인식시 DTW는 최적의 $w(n)$에 대해서 최소의 차이(difference)를 갖는 참조모델의 인덱스를 인식 결과로 표시하게 된다.

DTW는 고립단어 기반의 소규모의 단어인식시스템에서는 구조가 간단하고, 규칙화되어서 systolic array와 같은 하드웨어로 1980년대에 구현되어 사용되어 왔다. 하지만, 연속단어인식으로 확장방법에서 level-building 알고리즘, one-pass 알고리즘과 같은 방법이 연구되었지만[1, 3], 복잡한 규칙으로 구현이 어려우며, 이론의 발전에 한계가 있었다. 하지만, 위의 알고리즘은 다음에서 설명되는 HMM을 이용한 연속음성인식 시스템의 기본적 바탕이 되었다.

또한 DTW는 음성인식 시스템에서 가장 필요한 학습은 코드북의 참조모델의 특징벡터를 생성하는 곳에 부여할 수 있다. VQ(vector quantization)에 의해서 다수 화자의 패턴의 중심으로 결정하거나, 각 화자 별로, 각 단어의 중심으로 참조모델을 생성한다. 하지만 이러한 학습능력 외의 능력을 부여하기 어려우며, HMM이나 NN에 비해 인식능력이 떨어져 점차 사라져가고 있다. 특히 단어와 단어사이에서 발생하는 조음현상을 모델링하기 어려워 연속음성인식에 적용되지 못하고 있으며, 소규모의 숫자연결음성인식에 적용되고 있다.

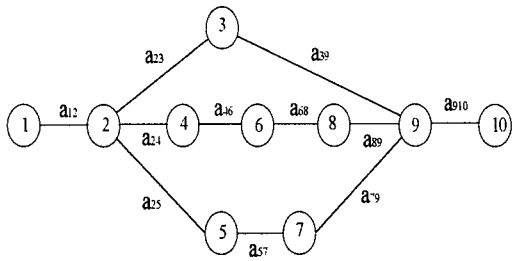
2. HMM(Hidden Markov Model)

음성신호가 Markov 특징을 갖는다는 관측으로부터 연구가 시작된 Hidden Markov Model은 통계적(stochastic) 방법에 의해서 관측특징벡터열과 음성인식모델 간의 유사도를 결정하는 인식알고리즘[7]으로 다음과 같이 간단히 정리할 수 있다.

$$\lambda_w=\arg\min_{\lambda_w}\Pr(Y|\lambda_w)\Pr(\lambda_w),\quad w=1,\cdots,M \quad (2)$$

여기서 $\lambda_w(w=1,\cdots,M)$는 개별 음성(단어)인식모델이고, M은 인식하고자하는 단어수를 나타내며, $Y=\{y_1y_2\cdots,y_N\}$는 관측특징백터열을 나타낸다. 또한 $\Pr(\lambda_w)=\Pr(W)$는 4절에선 설명되는 양으로 사전($a\ priori$)확률값을 나타낸다.

HMM에서는 많은 유사한 시스템이 연구되어 왔지만, 기본적인 인식모델 λ_w는 $\{\pi,A,B\}$로 시작된다. π는 초기상태확률로 초기에 각각의 상태(state)에 존재할 확률을 나타내며, $A=[a_{ij}]$는 상태 i에서 상태 j로 천이하는 확률값을 나타낸다. 음소모델에 많이 사용되는 left-right 모델에서는 $\pi_i=\delta(i)$, $a_{ij}=0(\text{if } i>j)$의 조건을 만족하는 모델이다.



〈그림 5〉 LR-HMM의 상태도(state diagram)

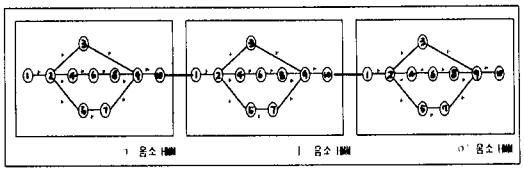
초기의 HMM은 이산분포(discrete probability distribution)를 갖은 $B=[b_{ij}]$로 모델링하였다. 여기서 b_{ij}는 상태 i에서 양자화된 특징벡터 j를 관측할 확률값이다. 근래에 많이 사용하는 연속확률분포를 갖는 CD-HMM은 관측확률분포를 Gaussian mixture로 다음과 같이 모델링한다[7].

$$b_j(y)=\sum_{i=0}^{I} c_i G(y|\mu_i,\ \Lambda_i) \qquad (3)$$

여기서, I와 c_i는 상수이며, $G_i(\cdot)$는 μ_i를 평균으로, Λ_i를 분산으로 하는 Gaussian 함수이다. CD-HMM은 VQ에서 발생하는 손실을 보상하여 인식률을 높이지만, 계산량 및 복잡성이 상대적으로 증가하는 단점이 있다.

HMM에서 훈련패턴이 주어졌을 때, 파라미터를 결정하는 과정이 훈련과정이다. 파라미터 예측은 관측(observation) 벡터가 서로 독립적이라는 가정 하에서 EM(Expectation-Maximization)에 의해서 예측할 수 있다. 기본적인 HMM 모델을 확장하여 각 상태에서 유지하는 상태지속시간(duration)의 크기를 모델링하는 경우에도 적용할 수 있다. 실질적인 문제를 연구하는 분야에서는 훈련시간의 단축 혹은 노이즈에 강인한 효율적인 훈련 방법(discriminate training, parameter tying)이 꾸준히 연구되고 있다.

통상 HMM을 이용한 음성인식시스템은 음(phone)차원에서 모델링으로부터 시작하여, 단어사전의 발음법칙에 의해서 단어모델을 구성한다. 그림 6은 음소HMM을 연결하여 단어 HMM을 생각하는 과정을 도시한 것이다. 계속해서, 단어모델을 문법의 규칙에 맞게 문장을 구성하여, 문장단위의 HMM을 구성할 수 있다.



〈그림 6〉 단어 HMM생성

문장단위의 HMM을 훈련하는 과정은 단어 단위의 HMM을 훈련하는 것과 같은 EM으로 모델 파라미터를 예측할 수 있다. 특히 문장 단위의 훈련으로 모델 파라미터와 함께 음소 및 단어의 끝점에 대한 예측을 자동으로 할 수 있어, 수동적인 음소 라벨링이 필요치 않으며, 문장을 구성하는 음(소) 열에 대한 정보만 필요하게 된다. 하지만 음소단위의 라벨링이 된 음성데이타베이스는 음소 HMM을 훈련하는 데 사용하며, 훈련된 HMM은 문장단위의 훈련과정의 초기치로 사용된다.

3. 신경망(Neural Networks)

신경망을 이용한 음성인식은 인간의 신호처리 과정이 단순한 계산소자(신경 뉴론)들이 복잡한 구조를 이루어 고도의 인지능력을 부여한다는 사실에 기인해서 발전했다. 초기 신경망을 이용한 음성인식 시스템은 SOFM, multilayer perceptron, TDNN, recurrent neural networks 등 단순한 구조에서 점차 발전하였다. 신경망의 특징은 HMM의 관측독립성(observation independence) 가정과 같은 입력벡터에 특별한 가정을 두지 않고 인식 시스템을 설계할 수 있다. 무엇보다 신경망은 학습이론에 의해, 훈련패턴에 내재해있는 음성신호의 특징을 스스로 학습할 수 있는 능력으로 타 인식 시스템에 비해, 부분적인 인식 테스트에서 높은 인식 결과를 보였다. 특히 TDNN[8]과 같은 구조는 짧은 시간(short-time)에 발생하는 시간굴곡현상과 정확한 음소/단어의 끝점에 대한 예측없이 인식을 할 수 있는 특징(time-shift invariance)으로 인해 음소인식 분야에서 많이 사용되어 왔다. 또한 회귀구조 신경망은 긴 시간의 문맥을 학습할 수 있는 특징으로 인해, 음소 인식률에서 HMM에 비해 높은 인식률을 보였다.

신경망은 음소/음절과 같은 인식률 면에서 HMM에 비해 우수하였으나, 수학적 이론이 약해 체계적 발전이 이루어지지 않고 있다. 또한 화자적응 및 대용량의 단어인식 시스템에 홀로 적용하는 데는 한계를 나타내었고, 긴 시간에 발생하는 시간굴곡현상을 모델링하기 어려워, 단독으로 인식 시스템을 구성하는데 제한되었다. 하지만 신경망의

출력값은 사후확률(MAP)과 같다는 증명[9]으로 인해, 신경망을 이용한 시스템은 HMM이나 DTW와의 결합이 이루어져, 기존의 시스템의 인식성능의 향상을 가져오게 되었다.

ML에 의해서 훈련이 이루어지는 HMM은 모델간의 분별력이 저하되어 인식시 오인식의 원인이 되어왔다. 모델간의 분별력을 높이기 위해서 분별훈련(discriminative training) 방법으로 인식 향상을 시도하였다. 신경망은 기본적으로 모델간의 분별력이 높도록 훈련된다. 더욱이 신경망의 정규화된 값은 사후 확률과 같음으로, 신경망을 이용하여, HMM에서 요구하는 ML 값은 다음과 같이 계산할 수 있다.

$$P(Y|\lambda_w)=\frac{P(\lambda_w|Y)P(Y)}{\sum_w P(\lambda_w|Y)P(Y)} \qquad (4)$$

식(4)에서, 필요한 ML $P(Y|\lambda_w)$는 Y에 대한 특별한 가정을 하지 않아도 계산할 수 있으며, EM 알고리즘에 의해서 구한 파라미터에 의한 결과보다 좋은 결과를 보였다. 신경망은 MS-TDNN, NN-HMM과 같이 주로 음성신호 차원에서 신호처리기로 사용되며, HMM은 음소이후의 단어모델/언어모델에 적용되고 있다.

IV. 언어모델

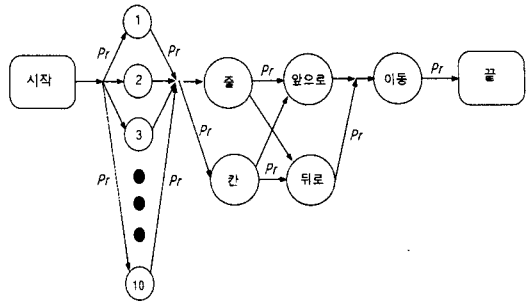
언어모델은 사전확률(*a priori*) $\Pr(W)$를 계산하는 데 사용되는 모델이다. 여기서 $\Pr(W)$는

$$\Pr(W)=P(w_0)\prod_{i=1}^{I}\Pr(w_i|w_{i-1},w_{i-2},\cdots) \qquad (5)$$

과 같으며, 마르코프 가정을 할 경우

$$\Pr(W)=P(w_0)\prod_{i=1}^{I}\Pr(w_i|w_{i-1}) \qquad (6)$$

과 같다. 또한 $\Pr(w_0)$과 $\Pr(w_i|w_{i-1})$는 대량의 텍스트 문장으로부터 사전에 계산해 놓은 값으로부터 얻을 수 있다. 위의 정보를 이용해서 문장을 인식하는 간단한 예를 보자.



〈그림 7〉 간단한 확률 언어모델

그림 7은 워드 프로세서의 간단한 음성 명령을 수행할 수 있는 음성명령인터페이스에서 사용되는 state machine을 도시한 것이다. 위 state machine은 "N 줄/칸 앞으로/뒤로 이동"의 조합으로 구성된 음성명령을 인식할 수 있으며, perplexity는 2.2이다. "줄"이나 "칸"을 인식한 상태에서는 "앞으로"나 "뒤로"의 단어만 연이어 나타날 수 있다. 따라서 인식과정에서 많은 가능성이 배제되어 인식 시스템의 인식률이 향상된다.

음성인식 시스템은 많은 에러를 포함하고 있고, 다분히 확률적인 값을 결과로 나타내는데, NLP에서는 가부(false/true)의 이산된 정보를 기본으로 한다. 보다 탄력적인 언어모델이 음성인식에 적용되기 위해서는 두 분야의 인터페이스에 대한 통합된 연구가 심도 있게 진해되어야 한다.

V. 결론 및 향후 연구방향

음성신호가 속귀(inner ear)와 신경계 및 뇌에서 어떠한 처리를 거쳐 인식는지 알려져 있지 않고 있다. HMM을 사용한 음성인식시스템은 실험실레벨에서는 95% 이상의 인식률을 보이나, 모델

링이 되지 않은 잡음이 존재하는 실환경에서의 인식 성능은 현격한 차이를 보인다. 특히, 정확히 문법을 따르지 않은 대화체(spontaneous dialog) 인식, 제한되지(unrestricted) 않은 연속음성인식 영역은 아직 태동기에 불과하다.

인식단어외의 인식문제(out-lier)는 새로운 단어에 대해서 학습하는 모델을 정립해야하며, 일치된 한국어 단어 및 문장모델을 정립해야 한다. 또한 잡음에 강인한 음성인식방법(robust speech recognition)이 실응용을 위해서는 필히 연구되어야 할 부분이다.

특히, 현재의 음성인식시스템은 선형모델에 기본으로 하고 있으나, 향후 카오스와 같은 비선형수학이론을 응용한 음성인식 시스템[10]이 등장할 것이나, 근시적으로 현재의 기술에 기반한 한국어 언어모델이 지속적으로 사용되어질 것이다. 특히 LD와 AD를 체계적으로 통합한 음성인식시스템에 대한 연구가 활발히 연구되어 질 것이다.

참 고 문 헌

[1] John R. Deller, Jr, John G. Proakis, and John H. L. Hansen, Discrete-Time Processings of Speech Signals, Macmillan Publishing Company, Inc. 1993.

[2] Steve Young, "A Review of large-vocabulary continuous-speech recognition," IEEE Signal Processing, Vol. 13, No. 5, Sept. 1996.

[3] L. Rabiner and G.-H. Juang, Fundamentals of speech recognition, Prentice Hall, 1993.

[4] C. R. Jankowski Jr, Hoang-Doan H. Vo, and R. P. Lippman, "A Comparison of signal processing front ends for automatic word recognition," IEEE Trans. on SAP, ' Vol 3, No. 4, pp. 286--293, 1995.

[5] C. Dugast, L. Devillers, and X. Aubert, "Combining TDNN and HMM in a hybrid systems for improved continuous speech recognition," IEEE Trans. on SAP, vol. 2, pp. 217--223, 1993.

[6] S. Renals, N. Moran, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," IEEE Trans. on Speech and Audio Processing, Vol 1. No.2 Part II, pp. 161--174, 1992.

[7] X.D. Huang, Y. Ariki, and M.A. Jack, Hidden Markov Models for Speech Recognition, Edinbyrgh Univ. Press.

[8] A. Waibel, H. Sawai and K. Shikano, "Modularity and scaling in large phoneme neural networks," IEEE Trans. on ASSP, Vol. 37, pp. 1188--1197, 1987.

[9] J. B. Hampshire and B. A. Pearlmutter, "Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function", Proc. of the 1990 connectionist models summer school, Morgan Kaufmann, 1990.

[10] C.-G. Jeong and Hong Jeong, "Automatic phone segmentation and labeling of continuous speech," Speech communication, Vol 20, pp. 291--311, 1996.

저 자 소 개

鄭 車 均

1966年 6月 18日生
1990年 한국과학기술원 한국과학기술대학 전자전기공학과 학사과정졸업
1992年 포항공과대학 전자전기공학과 석사과정졸업
1996年 포항공과대학 전자전기공학과 박사과정졸업

1996年 2月~현재 : 포스데이타 기술연구소

주관심분야 : 음성인식, 멀티미디어