최신 기계학습 기반 음성인식 기술 동향


박 현 신
KAIST


김 성 웅
Qualcomm


진 민 호
Qualcomm


유 창 동
KAIST

Ⅰ. 서 론

현재 음성인식기술은 전화상 안내서비스 및 다양한 기기에서 사용되고 있다. 자동차의 내비게이션에서 음성으로 목적지를 설정하거나, 스마트TV에서 채널을 변경하거나, 스마트폰에서 연락처를 찾고 일정을 관리하며 인터넷 검색을 할 때에 음성인식 기술을 사용할 수 있다. 이렇게 음성인식 기술은 우리 일상생활에 서서히 스며들면서 대중들에게 큰 관심을 받고 있다.

지난 반세기 동안 음성인식기술은 많은 발전을 이루었다. 1952년 미국 통신업체 에이티엔티(AT&T) 벨연구소(Bell Laboratories)에서 숫자인식기술을 개발하여 음성인식 연구가 시작되었으며, 1963년 IBM이 16개의 영어단어인식기를 개발하였고, 1970년대는 미국 국방부 산하 국방첨단연구사업국(DARPA)에서 큰 규모의 음성이해연구(Speech understanding research) 프로젝트를 진행하여 1,000단어 연속음성인식기를 개발하였다. 1980년대는 IBM이 은닉마르코프모델(Hidden Markov Model, HMM)을 활용한 대규모 음성시스템을 개발하면서 인식할 수 있는 단어가 1만 단어로 늘어났으며, 전성기가 시작되었다. 1990~2000년대에는 HMM기반 음성인식 시스템이 주를 이루었으며, 음성인식 오류를 최소화 하기위한 변별학습(discriminative learning), 잡음이나 반향 등에 강인한 음성인식 기술 등이 개발되었다. 최근에는 HMM의 한계를 극복하여 더 좋은 성능을 보이는 기계학습에 기반한 기술들이 소개되고 있다. KAIST에서는 large margin semi-Markov model(LMSMM)과 Gaussian process dynamic system(GPDS)에 기반한 음향모델들이 소개 되었고 토론토 대학의 Hinton 교수 연구실에서는 깊은신경망(deep neural network)에 기반한 음향모델이 개발되어 주목을 받고 있다.

〈그림 1〉 음성인식 기술의 원리

〈그림 2〉 음성인식을 위해 가장 널리 사용되고 있는 MFCC 특징벡터 추출 과정

더욱 자세한 내용은 [1]을 참고하길 바란다.

본문은 2장에서 음성인식의 기초 기술에 대해서 간략히 설명하고 3장에서 음성인식의 심화 기술에 대해서 설명하고 4장에서 결론을 맺도록 한다.

Ⅱ. 음성인식의 기초 기술

일반적으로 음성인식이란 입력된 음성을 기계가 문자열로 전환하는 것을 말하고 음성이해는 전환된 문자열의 의미를 출력하는 것을 의미한다. 〈그림 1〉은 음성인식 기술을 간단히 나타내는 순서도다. 음성인식기술은 발성의 형태에 따라 고립단어인식과 연속어인식으로 분류되며, 화자에 따라 화자독립인식과 화자종속인식으로 나뉘며, 음성신호의 획득방법(마이크의 위치)에 따라 근거리인식과 원거리인식으로도 구별된다. 이번 장에서는 고립단어인식을 위한 특징추출, 음향모델의 정의 및 학습, 인식알고리즘에 대해서 간단히 알아보기로 한다.

음성인식이란 입력된 음성을 기계가 문자열로 전환하는 것을 말하고 음성이해는 전환된 문자열의 의미를 출력하는 것을 의미한다.

1. 전처리 및 특징 추출

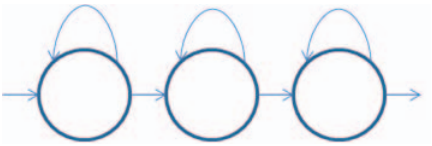
음성신호에는 언어적 의미 뿐만 아니라 잡음, 잔향, 개별화자의 특징 등 다양한 정보가 포함되어 있다. 이러한 음성신호에서 언어적 의미만 추출하기 위한 방법으로, 다양한 전처리 기술들이 이용되고 있다. 예를 들면 음성의 방향추정 (direction of arrival estimation) 기술, 빔포밍 (beamforming) 등을 이용한 음성강화 (speech enhancement) 기술, 여러 음성 분리 (blind source separation, BSS) 기술 등이 전처리 과정에서 이용되고 있다.

음성인식을 위해서는 전처리를 거친 음성 신호로부터 음향 특징벡터의 시퀀스를 추출해야 한다. 〈그림 2〉는 음향 특징벡터를 추출과정을 나타내는 순서도다. 일반적으로 음성 신호를 10 ms 마다 25 ms 구간으로 STFT (short-time Fourier transform)를 수행한 뒤, 인각의 청각 모델을 모방한 mel-scale filterbank를 통해서 각 대역의 에너지들을 얻는다. 이 에너지의 log 값에 DCT (discrete cosine transform)을 수행하여, 최종적으로 MFCCs (mel frequency cepstral coefficients)를 얻는다. MFCCs 로 이루어진 특징벡터는 주로 13차의 기본 계수와 그 계수들의 1차 미분, 2차 미분 값을 추가하여 얻은 39차의 특정벡터를 일반적으로 사용한다. 다른 음성특징으로는 perceptual linear predictive (PLP) 분석을 통한 특징과 linear predictive cepstral coefficients (LPC) 특징 등이 존재한다. 학습을 통한 특징 추출 방법도 존재한다. 대표적인 것으로는 특징 공간에서의 변별학습 (discriminative training)인 fMPE (feature-space minimum phone error)와 신경망 (neural network, NN)을 이용한 특징추출 방법 등이 있다.

특징벡터의 후처리 방법으로는 화자기반 cepstral mean and variance normalization (CMVN)과 문장단위 CMS (cepstral mean substraction)가 있다. 또한 추출한 특징벡터들에 대하여, PCA (principal component analysis)나 LDA (linear discriminant analysis)를 수행하여 잡음에 강인한 특징을 추출하기도 한다.

2. 음향모델

은닉마르코프모델(이하 HMM)[2]은 음성, 이미지, 비디오, 음악, 금융 데이터와 같은 시간적 또는 공간적 계



〈그림 3〉 left-to-right 3 state hidden Markov model

열 데이터를 표현하는데 자주 사용되는 모델이다. 〈그림 3〉은 음성인식에서 주로 사용되는 HMM의 구조이다. 3개의 은닉상태로 이루어지며 각 은닉상태는 오른쪽으로의 전이와 자기 자신으로의 전이만 허락된다. 각 은닉상태에서 관측벡터가 발생하는데, 일반적으로 정규혼합모델(Gaussian mixture model, GMM)을 사용하여 관측벡터를 표현한다. 시간 t에서 특징 벡터 x_t가 HMM의 i번째 은닉상태에서 발생할 확률을 K개의 정규분포로 이루어진 GMM으로 표현하면 다음과 같다.

$$p(x_t|\Lambda_i) = \sum_{k=1}^{K} w_{ik}\mathcal{N}(x_t;\mu_{ik},\Sigma_{ik})$$

여기서 GMM의 파라미터 집합 $\Lambda_i = \{w_{ik},\mu_{ik},\Sigma_{ik}\}_{k=1}^{K}$은 혼합 가중치 ω_{ik}, 평균 벡터 μ_{ik}, 그리고 공분산 행렬 Σ_{ik}로 구성된다. 위의 GMM 파라미터에 HMM의 은닉상태간의 전이확률 $\{p(s_i|s_j)\}_{i,j=1}^{S}$과 초기은닉상태 확률 $\{\pi_i\}_{i=1}^{S}$을 추가하여 HMM-GMM 모델이 구성된다. 음성의 특징벡터 계열 $X = \{x_t\}_{t=1}^{T}$이 주어졌을 때, 이에 대한 HMM의 우도(likelihood)는 다음 식과 같다.

$$p(X|\Lambda) = \sum_{s\in\mathcal{S}}[\pi_{s_1}p(x_1|\Lambda_{s_1})\prod_{t=2}^{T}p(s_t|s_{t-1})p(x_t|\Lambda_{s_t})]$$

3. 음향모델의 학습

음성인식을 위한 음향모델인 HMM의 파라미터들은 학습데이터로부터 추정되어야 한다. HMM의 파라미터 추정을 위한 기본적인 방법으로는 학습데이터에 대한 최대우도추정(maximum likelihood estimation)이 있다. 하지만 학습데이터에는 HMM-GMM의 우도함수에 포함되어 있는 몇 가지의 정보가 결여되어 있기 때문에, 이 문제를 해결하기 위하여 주로 EM 알고리즘을 사용한다. EM알고리즘은 E-step과 M-step을 번갈아 가면서 수행하는 알고리즘으로서, E-step에서는 모르는 변수에 대한 사후확률을 추정하여 목적함수의 기대치를 정의하고 M-step에서는 목적함수의 기대치를 최대화하는 파라미터를 추정하는 과정으로 구성된다. EM알고리즘은 극대점을 찾는 알고리즘으로 구한 값이 최대점이 아닐 수 있기 때문에 적절한 초기값 설정도 중요한 이슈중 하나다.

학습데이터에 대한 HMM의 은닉상태 계열이 주어졌을 때, 특정 은닉 상태에서의 GMM 파라미터 추정을 위한 EM 알고리즘은 다음 두 스텝을 번갈아 가면서 수행한다.

• E-step

주어진 음성특징벡터 계열 중 t번째 음성특징벡터가 GMM k번째 정규분포에 대해 어느 정도 영향을 받는지를 나타내는 척도를 현재까지 추정된 파라미터를 사용해서 다음과 같이 계산한다.

$$r(z_{tk}) = \frac{w_k\mathcal{N}(x_t;\mu_k,\Sigma_k)}{\sum_{j=1}^{K}w_j\mathcal{N}(x_t;\mu_j,\Sigma_j)}$$

• M-step

위에서 구한 척도를 이용해 새로운 파라미터를 다음과 같이 구한다.

$$\mu_k = \frac{1}{T_k}\sum_{t=1}^{T}r(z_{tk})x_t$$

$$\Sigma_k = \frac{1}{T_k}\sum_{t=1}^{T}r(z_{tk})(x_t-\mu_k)(x_t-\mu_k)^T$$

$$w_k = \frac{T_k}{T},\quad T_k = \sum_{t=1}^{T}r(z_{tk})$$

4. 인식 알고리즘

지금까지 음성신호로부터 특징을 추출하고 음향모델을 학습하는 방법에 대해서 알아봤다. 이제 학습된 음향모델이 주어지고, 미지의 음성신호가 들어왔을 때, 이를 인식하는 알고리즘에 대해서 알아본다.

관측된 신호 X가 주어졌을 때, HMM을 사용하여 인식을 하기 위해서는, 각 HMM의 은닉상태 계열을 추정하고, 우도를 계산해야 한다. 이를 위해서 일반적으로 Viterbi 알고리즘이 사용된다. 우선 t번째 관측신호의 i번째 상태에 대한 Viterbi score를 다음과 같이 정의한다.

$$\gamma_t(s_i) = \max_{s_j} p(x_t|s_i)p(s_i|s_j)\gamma_{t-1}(s_j)$$

이 스코어의 의미는 모든 j에 대해서 t-1번째 관측신호의 j번째 상태에 대한 Viterbi score와 s_i에서 s_j로 전의하는 확률을 곱한 것 중 최대값에 i번째 상태에서 t번째 관측신호가 발생할 확률을 곱한 값이다. 이 Viterbi score 모든 상태에 대해서 t=1부터 순차적으로 계산한 뒤, t=T에서 Viterbi score가 최대가 되는 상태를 찾고, 그 상태에 다다르게 된 과거의 상태를 역으로 찾아가면 관측신호에 대한 HMM의 은닉상태계열과 그의 우도를 계산할 수 있다. 모든 HMM에 대해서 동일한 작업을 수행한 뒤 우도가 최대가 되는 HMM를 찾으면 관측신호에 대한 인식결과를 얻을 수 있다.

Ⅲ. 음성인식 심화 기술

1. 변별 학습 (discriminative learning)

기계학습에 있어서 변별학습이란 서로 다른 모델간의 거리를 최대화 하는 기법을 일컫는다. 앞서 살펴본 최대우도추정은 자기 자신의 데이터를 충실히 생성하기 위한 모델을 학습하기 위한 기법인 반면, 변별학습은 자기와 다른 모델간의 거리를 최대화 하는, 즉 식별 성능에 대한 최적성을 보장하는 기법이다. 일반적으로 음성인식 시스템에 있어서 변별 학습(discriminative learning)이 최대우도추정보다 좋은 인식성능을 보인다. 결국 음성인식에 있어서 변별 학습이란, 단어 오인식 률 (word error rate, WER)을 최소화하도록 음향모델을 학습하는 것이다. 하지만 직접적으로 WER을 최소화 하도록 음향모델을 학습하는 것은 어렵기 때문에, WER를 근사화한 식별에러율(classification error rate, MCE)[3]를 최소화하도록 학습한다. MCE 추정은 Bayes' decision rule에서 나온 방법으로 최대우도추정보다 좋은 성능을 보인다. 다른 방법으로는 maximum mutual information (MMI)[4] 기준으로 학습하는 방법이 있다. 이는 음성 데이터와 레퍼런스인 단어 계열간의 상호정보량을 최대화 하는 방법이다. 이는 위의 MCE 추정에 있어 negative MCE와 밀접한 관련이 있다. MMI 학습은 또한 conditional maximum likelihood(CML) 추정과도 깊은 관련이 있다. 또 다른 방법으로 minimum phone error(MPE)[5] 기준에 의한 방법이 있다. MCE는 단어 단위의 오인식률을 최소화 했던 것과 비교해, MPE는 음소 단위의 오인식률을 최소화 하는 것을 목적으로 음향모델을 학습한다.

Semi-Markov model(SMM)은 세그먼트 기반의 마르코프 구조를 사용하여 입력으로 들어오는 순차 발화 데이터의 음소분할과 라벨예측을 동시에 수행하는 모델로서, 음소 세그먼트 내에서의 모든 관측치들 사이의 통계적 상관성을 고려한다.

2. Semi-Markov model (SMM)

기존의 HMM은 오직 이웃하는 관측치(프레임) 사이의 국지적인 통계적 상관성만 있다고 가정하고, 뚜렷한 음소분할 없이 각각의 관측치(프레임)에 대한 음소 라벨을 예측하므로 더 넓은 영역의 상관성을 고려하면서 음소 분할과 라벨인식을 동시에 수행해야 하는 음성인식을 수행하는 데 있어 한계가 있었다. Semi-Markov model(SMM)은 세그먼트 기반의 마르코프 구조를 사용하여 입력으로 들어오는 순차 발화 데이터의 음소분할(만약 마르코프 모델에서 노드에 해당하는 기본 유닛이 음소라고 할 경우)과 라벨예측을 동시에 수행하는 모델로서, 음소 세그먼트 내에서의 모든 관측치들 사이의 통계적 상관성을 고려한다. 또한 HMM은 음소길이를 정확히 모델링하지 못하는데, SMM에서는 음소길이를 직접적으로 모델링 할 수 있다.

SMM의 추론 문제, 즉 입력 음성 신호가 들어왔을 때, 음소 (또는 단어) 레이블 시퀀스를 추정하는 문제의 기준을 MAP으로 하였을 때, HMM에서 사용하던 Viterbi 알고리즘과 비슷하게 다음과 같은 recursion을 이용하여 빠르게 추정할 수 있다.

$$V_t[l] = \max_{d,l'}(V_{t-d}[l'] + \log p(x_{t-d+1}, \ldots, x_t|l) + \log p(l|l'))$$

〈표 1〉 SMM 기반 음소인식 결과
TEST SET PHONE ERROR RATES (%) BY EDIT DISTANCES

	Core test set				Enhanced test set			
	1-mix	2-mix	4-mix	8-mix	1-mix	2-mix	4-mix	8-mix
ML (One-state HMM)	42.8	36.8	34.0	32.2	42.1	36.3	33.4	31.0
One-state LMHMM	31.3	30.7	29.9	28.6	30.2	29.7	29.1	28.0
ML (Three-state HMM)	37.7	33.2	30.1	29.1	37.3	32.5	29.2	28.6
Three-state LMHMM	30.2	28.8	28.0	27.6	29.5	28.2	27.6	27.2
ML (SMM)	35.9	32.1	29.6	28.5	35.1	31.3	28.9	28.1
LMSMM	28.9	28.0	27.3	27.1	28.2	27.5	27.1	26.8

여기서 $V_t[l]$은 처음부터 t까지 마지막 세그먼트의 레이블이 l인 모든 가능한 음소 분할 시퀀스들의 확률값들 중 최대값이다. 이 때 d는 마지막 세그먼트의 길이를 의미한다. 이러한 SMM 추론을 위한 Viterbi 알고리즘의 복잡도는 HMM 추론을 위한 Viterbi 알고리즘에 비해서 $O(|L|^2T)$에서 $O(|L|^2TR)$만큼 늘어난다. 여기서 L은 레이블의 개수, T는 신호의 길이, R은 세그먼트의 탐색구간의 길이를 나타낸다.

참고문헌 [6]에서는 음소 인식을 위한 마진이 큰 차별적 (large margin, LM) SMM을 제안하였다. 이 논문에서 사용하는 LMSMM 프레임워크는 다음과 같이 장기적 상관성을 고려한 특징 맵에 선형적인 비확률적 판별 함수에 기반한다.

$$\mathbf{y}^* = \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}} F(\mathbf{x},\mathbf{y};\mathbf{w}) = \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}} \langle \mathbf{w}, \phi(\mathbf{x},\mathbf{y})\rangle$$

여기서 y는 음소 레이블 시퀀스, x는 음향 특징 벡터, w는 파라미터를 나타내며, Φ(x,y)는 세그먼트 기반 조인트 특징 맵이다. 여기서 세그먼트 기반 조인트 특징 맵은 인접한 음소 레이블 사이의 관계에 의한 천이 특징, 각 세그먼트의 음소 길이 특징, 세그먼트내의 음소내용특징 등을 포함한다. 판별함수의 파라미터 w는 구조적 예측을 위한 마진이 크도록 훈련하는 기법, 즉, structured support vector machine을 이용하여 추정한다. 이러한 파라미터 추정 기법은 곧 많은 제약조건을 가지는 아래와 같은 최적화 문제로 바뀌고, 이 논문에서는 이 최적화 문제를 추계적 경사 추적법을 이용하여 아래와 같이 푼다.

$$\begin{aligned}&\min_{\mathbf{w},\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i \\ &\text{s.t.} \quad \langle \mathbf{w}, \Delta\Phi(\mathbf{X}_i,\mathbf{y})\rangle \geq \Delta(\mathbf{y}_i,\mathbf{y}) - \xi_i \\ &\xi_i \geq 0, \forall i, \forall \mathbf{y} \in \mathcal{Y}\backslash\mathbf{y}_i\end{aligned}$$

여기서 정답 레이블 시퀀스에 의한 판별 함수의 값과 정답이 아닌 레이블 시퀀스에 의한 판별 함수의 값 사이의 차이를 나타내는 마진은 아래와 같다.

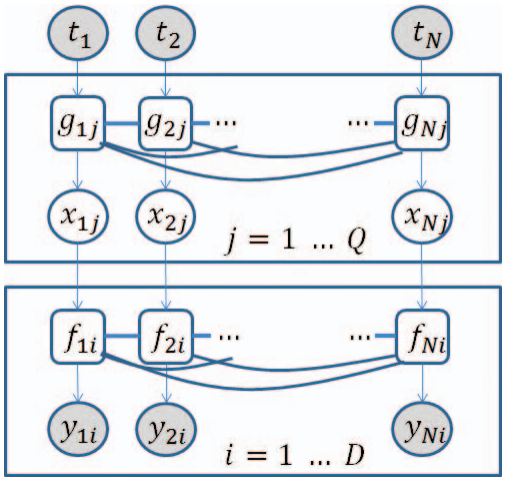
$$\begin{aligned}\langle \mathbf{w}, \Delta\Phi(\mathbf{x}_i,\mathbf{y})\rangle &= F(\mathbf{X}_i,\mathbf{y}_i;\mathbf{w}) - F(\mathbf{X}_i,\mathbf{y};\mathbf{w}) \\ &= \langle \mathbf{w}, \Phi(\mathbf{x}_i,\mathbf{y}_i) - \Phi(\mathbf{x}_i,\mathbf{y})\rangle\end{aligned}$$

제안된 LMSMM은 〈표 1〉에서와 같이 TIMIT 음소 실험에서 기존의 HMM에 비해 더 나은 성능을 보인다.

3. 정규과정동적시스템 (Gaussian process dynamical system)

기존의 HMM은 1) 1차 마르코프 구조로 인해 음성 특징벡터간의 조건부 독립을 필요조건으로 하며 2) 이산 은닉상태와 그의 전이과정이 실제 음성신호를 모델링 하는데 맞지 않다는 점에 착안하여 정규과정동적시스템 (Gaussian process dynamical system, GPDS) 기반 음향모델이 참고문헌 [7]에서 제안되었다.

GPDS기반 음향모델은 음성의 은닉상태의 복잡한 역학적 구조와 은닉상태에서 음성특징에 발현되는 확률분포를 정규과정(Gaussian process, GP)[8]을 이용해서 표현한다. GP는 함수에 대한 확률모델로서 모든 샘플들 간의 상관관계를 커널함수로 표현하는 비모수 베이지안 모델이다. GP에서 임의의 샘플을 취했을 때의 확률분포가 정규분포를 따른다는 특징을 가지고 있으며 자세한 내용은 [8]에서 확인할 수 있다. 〈그림 4〉는



〈그림 4〉 정규과정동적시스템의 확률그래프

GPDS의 확률그래프를 나타내며, 이를 식으로 표현하면 다음과 같다.

$$\begin{aligned} x_{nj} &= g_j(t_n) + \eta_{nj}, \quad \eta_{nj} \sim \mathcal{N}(0, 1/\beta_j^x) \\ y_{ni} &= f_i(\mathbf{x}_n) + \epsilon_{ni}, \quad \epsilon_{ni} \sim \mathcal{N}(0, 1/\beta_i^y) \end{aligned}$$

여기서 함수 $f_i(\mathbf{x}) \sim \mathcal{GP}(\mu_i^f(\mathbf{x}), k_i^f(\mathbf{x}, \mathbf{x}'))$ 는 은닉상태에서 관측신호가 발현되는 과정을 GP로 표현하며, 함수 $g_j(t) \sim \mathcal{GP}(\mu_j^g(t), k_j^g(t, t'))$ 는 은닉상태간의 전이 과정을 GP로 표현하고 있다. 여기서 정규과정의 평균함수는 항상 0의 값을 갖는다고 두고, f_i 와 g_j 의 커널함수에 대해서는 각각 다음과 같은 함수를 사용한다.

$$k^f(\mathbf{x}, \mathbf{x}') = \alpha^f \exp\left(-\sum_{j=1}^{Q} \omega_j^f (x_j - x_j')^2\right),$$

$$k^g(t, t') = \alpha^g \exp\left(-\omega^g (t - t')^2\right) + \lambda t t' + b,$$

k^f 는 RBF 커널함수이며 $\{\alpha_j^f, \omega_j^f\}$의 하이퍼파라미터를 갖는다. k^g는 RBF 커널과 선형 커널을 결합한 커널함수이며 $\{\alpha^g, \omega^g, \lambda, b\}$ 의 하이퍼파라미터를 갖는다.

GPDS 음향모델의 학습은, 위에서 언급한 커널함수의 하이퍼파라미터를 찾는 것으로서, 학습데이터에 대한 GPDS의 우도함수를 변분적 추론(variational inference)을 통해 얻고 이 우도함수를 최대화하는 하이퍼파라미터를 찾기위해 경사법(gradient method)을 사용한다.

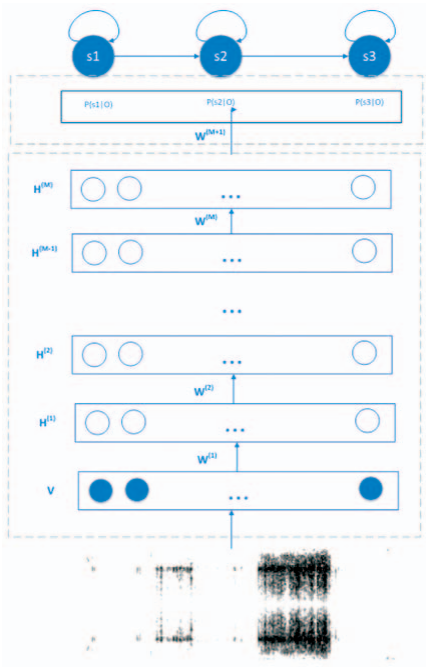
정규과정동적시스템의 음향모델로서의 성능을 검증하기 위해서, 영어음성 데이터베이스인 TIMIT를 사용해서 모델을 학습하고 TIMIT의 core test set에 대해 음소인식실험을 수행한 결과, 비슷한 조건의 HMM모델이 57.8 %의 성능을 보인 것에 비해 GPDS모델이 61.5%의 성능을 보여 GPDS의 음향모델로서의 가능성을 보였다.

현재 GPDS기반 음향모델은 학습시간이 HMM에 비해서 오래 걸린다는 점과, HMM-GMM와 같은 출력분포를 모델링하기위한 혼합모델이 없다는 점이 한계점으로 남아있고 앞으로 이를 해결해야 음성인식시스템에 사용될 수 있을 것으로 보인다.

4. 깊은신경망 (deep neural network)

지난 30년 동안, GMM-HMM 시스템 (Gaussian mixture model 을 이용하여 상태 관측 확률을 계산하는 은닉 마르코프 모델)이 사실 상의 표준(de factor standard) 으로 여겨왔다. 많은 수의 특징 추출 알고리즘, 식별적 학습 알고리즘, 적응 알고리즘들이 GMM-HMM 기반으로 개발되어, 다른 형태의 확률 모델로 GMM-HMM의 성능을 넘기 어려운 것으로 생각되었다. 최근의 깊은 신경망 (Deep neural network)을 이용한 기법들이 몇몇 음성 인식 성능 측정에서 기존의 GMM-HMM 기반 기법에 비해서 향상된 성능이 보고되어, 많은 연구가 이루어지고 있는 추세이다.

가장 흔히 사용되는 깊은 신경망은 〈그림 5〉와 같이 HMM기반의 시스템의 상태 관측 확률 (state observation probability)을 GMM이 아닌 깊은 신경망을 이용하여 모사한 시스템이다. 기존의 GMM-HMM 시스템과의 가장 큰 차이점은 하나의 상태 관측 확률을 계산하기 위해 각각의 음성 프레임이 비선형적으로 기여하게 된다는 점이다. 기존의 GMM-HMM의 경우 한 상태의 관측 확률은 그 상태에 해당되는 프레임들의 우도 값의 선형 합으로 계산된다. SMM의 경우에는 이러한 선형 합을 보다 구조화하여 보다 정확한 추론을 시도하지만, 기본적으로는 우도의 선형합이라는 한계를 벗어나지 않는다. 반면, 최근에 많이 사용되는 〈그림 5〉와 같은 깊은 신경망을 이용한 DBN-DNN (Deep



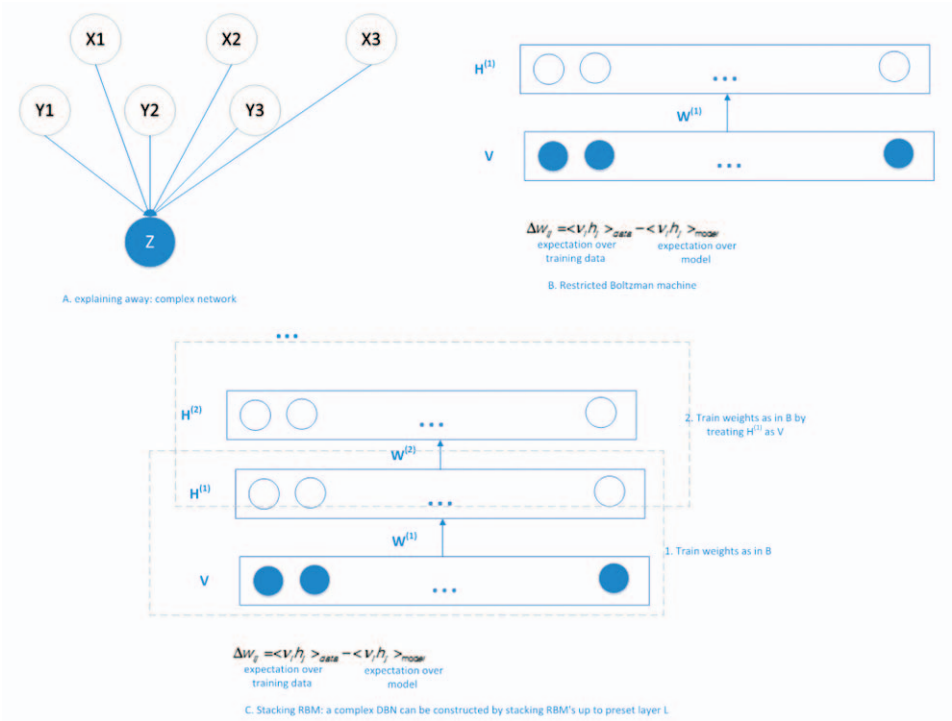
〈그림 5〉 HMM의 상태 관측 확률을 DNN으로 모사한 DBN-DNN 시스템

Belief Network – Deep Neural Network) 시스템의 경우, 긴 구간의 음성 특징 (long-term characteristics)이 비선형적으로 한 상태의 관측 확률에 기여한다는 가정하에 모수(parameter) 들이 훈련된다. 반면, 기존의 음성 인식 시스템에서는 각각의 프레임들이 겹침(overlap)이 있음에도 불구하고 각 프레임이 연관이 없으며 (uncorrelated), 서로 다른 차원의 음성 특징 벡터들의 연관은 거의 0에 가깝다고 가정하여 대각 행렬로 분산을 모사하였다. 그렇기 때문에 DBN-DNN의 가정과 같이 비선형적인 결합이 실제 현상을 보다 잘 설명한다면, 보다 적은 수의 모수로 통계 모델을 정확하게 표현할 수 있고, 이러한 특징이 뒤에서 나올 DBN-DNN 시스템의 성능 향상에 기여한다고 볼 수 있다. 많은 경우에 있어서 상태 관측 확률을 제외한 상태 천이 확률 등은 HMM과 같은 구조를 가져감으로써 기존의 HMM 기반의 Viterbi 디코딩 구조를 이용할 수 있도록 한다.

DBN-DNN의 가정과 같이 비선형적인 결합이 실제 현상을 보다 잘 설명한다면, 보다 적은 수의 모수로 통계모델을 정확하게 표현할 수 있고, 이러한 특징이 뒤에서 나올 DBN-DNN 시스템의 성능 향상에 기여한다고 볼 수 있다.

복잡한 신경망 (neural network)을 이용하여 이러한 상태 관측 확률을 모사하는 시도는 예전에도 있어 왔으나 이러한 시스템들은 기존의 GMM-HMM에 비해 괄목할 만한 성능 향상을 얻지는 못했다. 〈그림 6. A〉와 같이 복잡한 신경망을 생각해 보자. Z 절 (node)에 대해 추론하기 위해서는 X1, X2, X3, …, Y1, Y2, Y3 와 같이 많은 절들이 사용된다. 위의 표현상으로는 X1, X2, X3, .., Y1, Y2, Y3가 모두 Z에 대해서 조건부 독립(conditional independence)으로 표현된다. 하지만, Z에 대한 사후 확률(posterior probability)를 설명하기 위해 X1, X2, X3, …, Y1, Y2, Y3, … 가 서로 경쟁하게 되어 결과적으로 한 두 개 (예를 들어 X2, Y2) 만으로 설명할 수 있는 Z를 다른 X1, X3, …, Y1, Y3, … 등이 영향을 주어 부정확하게 확률 관계가 설명되는 소위 explaining away 현상이 발생할 수 있다. 이러한 현상에 대한 대안으로 G. Doddington 등의 연구에서는 〈그림 6. B〉 제한적 볼쯔만 머신(Restricted Boltzman Machine)을 여러 층으로 쌓아 나가는 형태의 깊은 신경망을 제안하였다. 제한적 볼쯔만 머신은 마르코프 랜덤 필터 (Markov Random Field)의 특수한 형태로, 0또는 1의 값을 가지는 visual node (v)와 hidden node (h) 로 구성되어 있으며 hidden node에서 hidden node, visual node에서 visual node로의 연결이 없는 특징을 가진다. 이러한 제한적인 구조를 이용하여 explaining away를 해소하고, 동시에 〈그림 6. C〉에 표현된 바와 같이 RBM을 이용하면 임의의 층수의 DBN을 체계적으로 구축할 수 있기 때문에 다양한 실험이 가능하고, 특히 비전 관련 연구에서 많은 성능 향상을 관찰한 바 있다. (〈그림 6. C〉에 표현된 모수 w 의 추정 방법은 참고문헌 [10]에 자세히 설명되어 있다.)

〈그림 5〉와 같은 DBN-DNN 시스템을 훈련하기 위



〈그림 6〉 A. 복잡한 신경망, B. 제한적 볼쯔만 머신 (RBM), C. Stacking RBM

해서는 선행 훈련 (pretraining)과 차별적 미세 조정 (discriminative fine tuning)을 이용한 형태의 훈련 방법을 이용한다[9,11]. 선행 훈련에서는 DBN으로 표시된 여러 층의 볼쯔만 머신을 음성 프레임의 상태 (state) 정보를 고려하지 않고 〈그림 6. C〉와 같은 방법으로 훈련한다. 이 후, 차별적 미세 조정 과정에서는 먼저 별도의 음성 인식 시스템 (예를 들여 GMM-HMM시스템)을 통해 먼저 상태 순열 정보를 얻어 낸다. 이 후, DBN-DNN을 만들기 위해 깊은 신경망의 마지막 층에 각 상태의 사후 확률을 모사하도록 미리 추정된 상태 순열 정보를 정답으로 하여 일반적인 신경망과 같이 흔히 역전파 (back propagation) 알고리즘을 이용하여 훈련된다.

DBN-DNN 시스템을 훈련하기 위해서는 선행 훈련 (pretraining)과 차별적 미세 조정(discriminative fine tuning)을 이용한 형태의 훈련 방법을 이용한다.

〈표 2〉는 [9]등의 문헌에 있는 여러 시스템들의 성능 비교표이다. 보는 바와 같이 기존의 GMM-HMM 기본 (baseline) 시스템에 비해서 DBN-DNN 시스템이 비교적 낮은 오차를 보여주고 있다. 이러한 DBN-DNN은 특징의 구성 (MFCC/Filterbank/etc) 또는 DBN의 구조 설정 등을 통해 다양한 변이가 가능하기 때문에 아직까지도 많은 추가 연구가 필요하다.

〈표 2〉 GMM-HMM/DBN-DNN 성능 나열[9]

	TIMIT (Phone error rate, PER)	Switchboard Test 1 (Word error rate, WER)	Switchboard Test 2 (Word error rate, WER)
GMM-HMM system	27.3%	18.6%	17.1%
DBN-DNN system	20.0%	18.5%	16.1%

Ⅳ. 결 론

음성인식의 일반적인 특징추출, 음향모델, 음향모델의 학습과 인식기술에 대해서 알아본 뒤 심화기술인 변별학습, SMM, GPDS, 깊은신경망 등에 대해서 알아

보았다. 스마트 기기의 보급으로 인해 대중들이 음성인식기술을 쉽게 접하고 있으나, 아직까지 만족할 만한 성능을 제공하고 있다고 보기 어렵고, 심화 기술들이 실제로 쓰이기 위해서는 대용량 어휘 연속음성인식에 대한 성능평가가 더욱 이루어져야 한다. 앞으로 더욱 더 좋은 음성인식 기술들이 나와 일반 사용자들이 만족할 수 있는 시스템이 개발되기를 기대하는 바이다.

참 고 문 헌

[1] 한국콘텐츠진흥원, "음성인식 기술의 동향과 전망", 2011.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," Proceedings of the IEEE, Vol.64, pp.532-556, 1976.

[3] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error methods for speech recognition," IEEE Trans. Speech Audio Processing, vol. 5, no. 3, pp. 257-265, 1997.

[4] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in Proc. ICASSP, pp. 49–52, 1986.

[5] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in Proc. ICASSP, pp. 105–108, 2002.

[6] Sungwoong Kim, Sungrack Yun, and Chang D. Yoo, "Large Margin Discriminative Semi-Markov Model for Phonetic Recognition", IEEE Trans. Audio, Speech and Language processing, vol.19, no.7, pp. 1999-2012, September 2011.

[7] Hyunsin Park, Sungrack Yun, Jongmin Kim, Sanghyuk Park, and Chang D. Yoo, "Phoneme Classification using Constrained Variational Gaussian Process Dynamical System" in Proc. NIPS, December 2012.

[8] C. E. Rasmussen and C. K. I. Williams, "Gaussian Process for Machine Learning," MIT Press, Cambridge, MA, 2006.

[9] Hinton et. al, "Deep Neural Networks for Acoustic Modeling in Speech Recognition", IEEE Signal Processing Magazine (82) Nov. 2012.

[10] Mohamed, G. Dahl, and G. Hinton, "Deep Belief Networks for phone recognition", in Proc. ICASSP, 2011.

[11] D. Yu, L. Deng, and G. Hinton, "Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition" in Proc. NIPS, 2010.

박 현 신

2007년 3월 Kobe University 학사
2009년 3월 Kobe University 석사
2009년 9월~현재 KAIST 박사과정 재학중

〈관심분야〉
기계학습, 음성인식

김 성 웅

2004년 8월 KAIST 학사
2011년 8월 KAIST 박사
2011년 9월~2012년 3월 KAIST 박사후연구원
2012년 3월~현재 퀄컴연구소

〈관심분야〉
기계학습

진 민 호

2002년 2월 KAIST 학사
2004년 2월 KAIST 석사
2009년 8월 KAIST 박사
2010년 1월~2011년 1월 삼성전기
2011년 1월~현재 퀄컴연구소

〈관심분야〉
음성 인식, 화자 인식, 멀티미디어 식별

유 창 동

1986년 6월
B.S. in Engineering & Applied Science/ California Institute of Technology
1988년 8월
M.S. in Electrical Engineering/Cornell University
1996년 8월
Ph.D in Electrical Engineering/ Massachusetts Institute of Technology
1997년 1월~1999년 3월 한국통신(KT) 연구개발 본부 선임연구원
1999년 3월~현재 한국과학기술원(KAIST) 전기및 전자공학과 교수
2011년 11월~현재 한국과학기술원(KAIST) 국제 협력처 처장

〈관심분야〉
기계학습, 신호처리