김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제10호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 10 JUCHE105(2016).

음소인식을 위한 심층신경망에서 판별 특징량리용에 대한 연구

리정철, 현성군

지난 30여년동안 대부분 자동음성인식체계들은 HMM을 리용하여 음성신호의 련속적인 구조를 모형화하고있으며 HMM상태들은 음성파형의 스펙트르표현을 모형화하기 위하여 가우스분포혼합모형(GMM: Gaussian Mixture Models)을 리용하였다. HMM에 기초한 체계들은 최대우도기준(ML: Maximum Likelihood)에 기초한 GMM훈련으로 훈련자료를 정합시키고 다시 판별훈련기준으로 ML훈련된 GMM들을 약간 조절하는 방식을 취하고있다.[2] 이외에도 서로 다른 발성자들의 특성을 표준발성자에 맞게 넘기는 발성자적응기술들이 GMM에 적용되여 보충적인 개선을 주고있다.[1] 발성자적응과 판별훈련은 모두 더 좋은 음향특징파라메터를 생성하기 위하여 특징공간령역에 적용될수 있다. 최근에 정결합신경망의 효률적인 훈련방법이 제안되면서 신경망에 대한 관심이 높아지고 GMM대신에 신경망을 리용하여 자연언어처리, 다매체처리분야는 물론 음성인식분야에서도 좋은 성과를 거두고있다.[3] 이러한 체계들은 입력신호로 MFCC특징량이나 려파기출력벡토르를 그대로 리용하고있다.

론문에서는 종전의 체계들에서 리용한 특징량들보다 더 좋은 특징량들을 리용하면 심층신경망의 성능이 더 좋아질수 있다는 가정에 기초하여 망의 입력신호로 발성자적응된 판별특징량을 리용하기 위한 방법을 제안한다.

1. 발성자적응된 판별특징량추출과 심층신경망의 입력층구성

① 발성자적응된 판별특징량추출

론문에서 리용한 발성자적응된 판별특징량을 얻기 위한 처리흐름도를 그림 1에 보여주었다.

그림 1에서 발성자적응된 판별특징량추출과정은 다음과 같다.

우선 발성된 음성은 10ms의 프레임이동을 가진 25ms너비의 하밍창문을 거쳐 39차원 MFCC특징벡토르렬로 변환된다. 여기서 매 특징벡토르는 12차원멜케프스트람결수와 에네르기(13차원MFCC), 그것의 1계 및 2계시간도함수를 포함하여 총 39차원으로 구성된다.

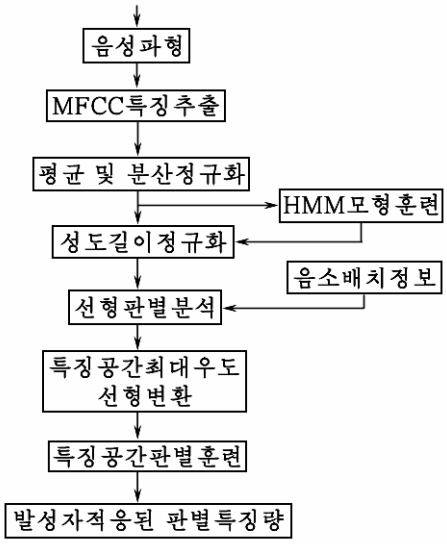

그림 1. 발성자적응된 판별특징량추출과정

얻어진 특징벡토르들은 매 훈련문장들에 대하여 령평균과 단위분산을 가지도록 MVN[1]방법으로 정규화된다.

그리고 성도길이정규화(VTLN[1])기술을 리용하여 발성자의 성도길이를 정규화한다. 이를 위해 매 발성자에 대해서 척도화인자를 학습한 다음 평균성도길이를 가진 표준발성자와 일치하도록 발성음성의 주파수스펙트르를 변형시킨다.

정규화가 진행된 후 문맥정보를 반영하도록 좌우로 총 9개의 특징벡토르들을 련결한 다음 LDA[2]변환을 적용하여 40차원특징벡토르로 축소한다. VTLN을 적용하여 얻어진 특징량들에 대하여 특징공간최대우도선형회귀(fMLLR)를 리용하여 발성자적응을 진행한다.

마지막으로 발성자적응된 판별특징량을 얻기 위하여 BMMI기준[2]에 기초한 대여백 판별훈련을 진행하였다.

론문에서는 특징공간BMMI특징량을 얻기 위하여 ML훈련된 HMM모형을 리용하였다.

② 심층신경망의 입력층구성

론문에서는 MFCC특징량과 같은 실수값자료들을 심층신경망의 입력신호로 리용하기 위하여 가우스－베루누이제한볼츠만기계(GRBM[3]: Gaussian-Bernoulli Restrict Boltzmann Machine)를 심층신경망의 입력층에 배치한다.

2. 평 가 실 험

실험을 위하여 표준음성인식자료기지(TIMIT[2])에서 462명의 음성자료가 훈련모임으로 리용되였으며 이와 분리된 50명의 발성자들로 되여있는 개발모임(dev)은 층수와 층크기와 같은 모형파라메터들을 조절하는데 리용되였다.

이때 실험결과들은 개발모임에서 배제된 24명의 검사모임(core test)을 리용하여 얻었다. 맨 밑층의 GRBM을 학습하기 위하여 0.005의 고정된 학습률을 가지고 150번의 갱신단계를, 웃층의 RBM들을 학습하기 위하여 0.08의 학습률로 50번의 갱신단계를 거쳤다. 갱신단계마다 파라메터들을 조절하기 위해 확률그라디엔트하강법을 리용하였다. 훈련과 복호화때에 183개의 클라스기호(61개 음소×3개 상태)를 리용하였으며 목표클라스로 단음소기호와 학습자료에 포함되여있는 2 400개의 3음소기호들을 리용해보았다. 모든 실험들에 리용된 비터비복호기는 정적FSM(Finite State Machine)으로서 언어모형의 그람수에 대한 제한이 없다.

그림 2는 은폐층수를 층당 1 024개로 고정시키고 입력프레임수를 11개로 하였을 때의 서로 다른 입력특징량들의 성능을 보여준다.

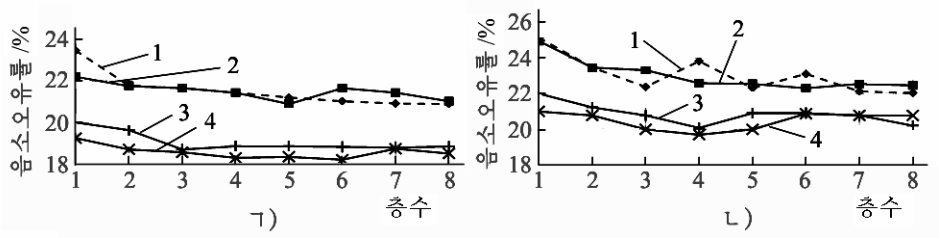

그림 2. 은폐층수에 따르는 서로 다른 특징량들의 성능
ㄱ) 개발모임일 때, ㄴ) 검사모임일 때;
1－MFCC, 2－LDA, 3－LDA+SA, 4－LDA+SA+fBMMI(제안한 방법)

그림 2에서 보는바와 같이 MFCC특징량과 LDA특징량사이에 성능상 의미적인 차이가 없다는것을 알수 있다. 이것은 망이 문맥창문에 포함되여있는 국부적인 판별정보를 리용한다는것을 말해준다.

한편 그림 2를 보면 발성자정보를 리용하여 모든 발성자들을 표준발성자에로 변환하였을 때 정확도가 시종일관하게 2%정도 올라갔다는것을 알수 있다. 또한 fBMMI목적함수를 리용하여 경쟁하는 음소클라스들사이의 여백을 최대화함으로써 입구문맥창문안에서 심층신경망의 국부적인 판별성을 보충하였다는것을 알수 있다. 그리고 특징량형태가 추가됨에 따라 음소오유률이 감소된다는것을 알수 있다.

맺 는 말

발성자적응된 판별특징량을 심층신경망의 입력신호로 리용하기 위한 방법을 제안하였다. 실험결과 발성자적응된 판별특징량을 리용하였을 때 음소오유률이 시종일관 2%정도 개선되였다는것을 알수 있다.

참 고 문 헌

[1] M. J. F. Gales; Computer Speech and Language, 12, 75, 1998.

[2] T. N. Sainath et al.; Proc. ASRU, 359, 2009.

[3] A. Mohamed et al.; IEEE Trans. on Audio, Speech and Language Proccessing, 20, 1, 14, 2012.

주체105(2016)년 6월 5일 원고접수

Study for Using Discriminative Features in Deep Belief Networks for Phonetic Recognition

Ri Jong Chol, Hyon Song Gun

Deep belief networks (DBNs) can be good at modeling windows of coefficients extracted from speech by discovering multiple layers of features that capture the higher-order statistical structure of the data. In this paper, we propose a method for initializing deep belief networks with better features.

Key words: discriminative feature transform, deep neural network