심층신경망에 기초한 음성인식음향모형화에서 여러가지 특징들의 성능분석

김광림, 리정철

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《첨단돌파전은 현대과학기술의 명맥을 확고히 틀어쥐고 과학기술의 모든 분야에서 세계를 앞서나가기 위한 사상전, 두뇌전입니다.》

심층신경망(DNN: Deep Neural Network)에 기초한 음성인식음향모형화에서 여러가지 특징추출방법들의 영향을 조사하는 단계는 다음과 같다.

① 음성인식체계의 특징추출부분을 분리한다.

② MFCC(Mel－Frequency Cepstral Coefficient)특징, FFT(Fast Frequency Transform)크기스펙트르, 생음성신호(처리되지 않은 음성파형자료)우에서 훈련된 음향모형들을 평가한다.

론문에서는 입력특징으로 선행한 특징추출방법과는 다른 생음성신호를 리용하는 방법에 중점을 두고 음성인식음향모형화를 실현하기 위한 한가지 방법을 제안하였다.

1. 문 제 설 정

선행연구[1]에서는 여러개의 은폐층을 가진 신경망의 일반근사화특성으로부터 심층신경망으로 자료기지에서 필요한(비선형) 특징추출단계들을 학습할수 있는 방법을 제안하였다.

선행연구[2]에서는 심층신경망의 훈련자료량이 생음성신호에 기초한 특징발견에 어떤 영향을 주는가를 조사하는 방법을 제안하였다.

선행연구에서 제안한 방법들은 입력으로 생음성신호를 리용하면 인식정확도가 낮아지고 훈련자료량이 증가하면 생음성신호와 스펙트르에 기초한 망호상간 성능이 떨어지는 결함을 가지고있다.

론문에서는 심층신경망으로 발견해야 할 특징들의 범위와 어느 정도의 특징들을 발견할수 있는가에 대하여 실험적으로 조사하였다.

2. 생음성신호를 리용한 음향모형화방법

음향모형훈련은 조선어음성자료기지의 10h분음성우에서 프레임별교차엔트로피기준을 리용하여 진행한다.

개발모임과 평가모임은 각각 1h, 40min길이에 해당한 음성으로 이루어진다. 개발모임은 학습률과 같은 일부 초파라메터들을 결정하는데 리용된다.

대규모훈련자료우에서 여러가지 방법들의 성능을 평가하기 위하여 같은 자료기지로

부터 취한 500h자료모임에 대한 실험들을 진행하였다.

입력특징으로서 오른쪽 문맥과 중심, 왼쪽 문맥의 프레임들로 구성된 17개의 프레임 묶음을 리용하였다.

훈련단계에서 묶음의 크기는 512개로서 우연적으로 뒤섞은 훈련모임으로부터 취해진다. 무게들은 예비훈련을 통하여 초기화된다.

표 1에 조선어음소인식실험에 리용된 특징추출방법을 보여주었다.

표 1. 조선어음소인식실험에 리용된 특징추출방법

특징	프레임당 입력벡토르차원	문맥크기/프레임수	표본창문/ms	기타
시간신호	320	12	20	표본화주파수(16kHz) 량자화비트수(16bit)
크기스펙트르－|FFT|	256	10	25	FFT점수(512개)
CRBE	20, 40	10	25	3각형MEL려파기를 리용
MFCC	13, 16, 20, 40	10	25	GMM에서 LDA변환을 적용
PLP	13	10	25	비트척도, 제형려파기를 리용

인식단계에서는 음소2그람언어모형(LM: Language Model)을 리용한다. 그리고 정확도 기준으로 음소오유률(PER: Phone Error Rate)을 리용한다.

모든 실험들에서는 매 층에 2 000개의 은페유니트(Hidden Unit)들을 가진 6개의 은페층을 리용한다. 4 500개 마디를 가진 출력층은 380개의 문맥질문으로 구축된 음소결정나무의 잎(즉 공유상태)들에 대응된다.

심층신경망의 구조를 그림 1에 보여주었다.

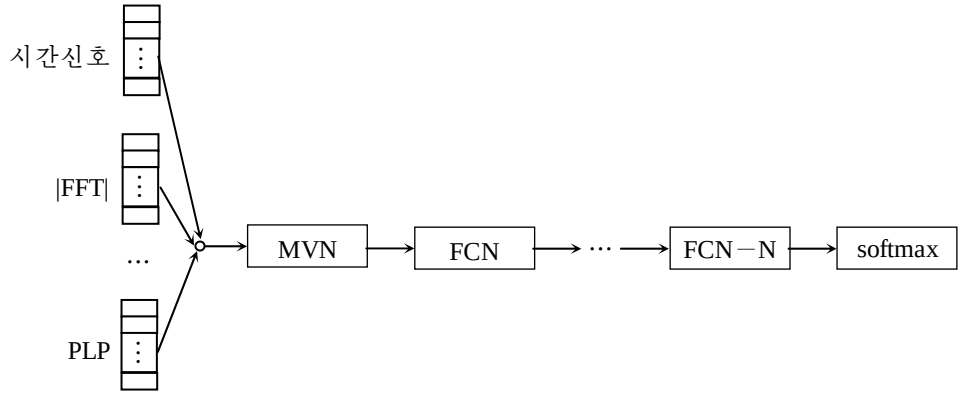

그림 1. 심층신경망의 구조

그림에서 MVN(Mean Varience Normalization)은 입력특징에 대한 평균분산정규화로서 매 발성당 혹은 전체 발성에 대하여 수행될수 있다.

FCN(Forward Connection Network)층은 정결합층으로서 은페유니트는 시그모이드(sigmoid) 혹은 정류선형유니트이다. softmax층은 출력층으로서 공유상태들의 사후확률분포를 모형화한다.

리용된 최종 특징량은 9개의 련속으로 놓인 13차원MFCC프레임들에 선형판별분석(LDA: Linear Discriminant Analysis)을 적용하여 얻어진 40차원특징량이다.

3. 실험결과 및 분석

첫번째 실험에서는 MFCC특징우에서 훈련된 GMM(Gaussian Mixture Model)과 그림 1의 심층신경망음향모형으로 얻어진 기준결과들을 비교하여 진행한다.

음소오유률을 표 2에 보여주었다.

표 2. 음소오유률

특징	모형	개발/%	평가/%
MFCC	GMM	27.4	28.0
MFCC	DNN	23.5	24.1
시간신호	DNN	32.9	34.4

표 2에서는 MFCC특징들을 평균/분산정규화한 결과를 보여주었다. 훈련은 10h음성자료우에서 진행된다. 같은 심층신경망구조를 생음성신호특징을 리용한 훈련에 적용하였다. 표 2에서 보여준것처럼 MFCC에 기초한 심층신경망모형은 GMM을 릉가하지만 생음성신호우에서 훈련된 체계의 음소오유률(PER)은 여전히 더 높다는것을 알수 있다.

두번째 실험에서는 인식정확도가 여러가지 전처리단계들에 얼마나 의존하는가를 알아내야 한다. 이러한 목적밑에 MFCC를 단계별로 갈라내여 그 성능을 측정하였다.

DNN음향모형을 위한 특징전처리와 정규화, 1개 특징벡토르의 차원, 음소오유률을 표 3에 보여주었다.

표 3. DNN음향모형을 위한 특징전처리와 정규화, 1개 특징벡토르의 차원, 음소오유률

특징	차원수	음소오유률	
		개발/%	평가/%
MFCC	16		
MFCC + 대역정규화		24.7	25.2
MFCC + 발성정규화		24.3	24.9
MFCC	20		
MFCC + 발성정규화		23.5	24.1
CRBE + 발성정규화	20	23.9	24.7
	40	23.3	24.1
|FFT|	256		
|FFT| + 대역정규화		24.9	25.4
|FFT| + 발성정규화		24.6	25.0
시간신호	320		
시간신호 + 대역정규화		33.1	34.8
시간신호 + 발성정규화		32.9	34.4

1) 특징결합

표 3의 결과로부터 MFCC가 높은차원FFT와 시간신호특징을 릉가한다는것을 알수 있다. 그러면 낮은차원특징량안에 정보량을 얼마나 늘일수 있는가가 문제로 된다. 앞에서 론의한것처럼 여러가지 단시간특징추출처리공정들은 음성에 대한 특징표현에서 조금씩 차이난다.

특징결합과 음소오유률을 표 4에 보여주었다.

표 4. 특징결합과 음소오유률

특징	개발/%	평가/%
MFCC	23.5	24.1
PLP	23.8	24.3
MFCC + PLP	21.0	21.7

표 4에서 보여준것처럼 심층신경망이 매개 특징모임보다 여러개의 특징모임으로부터 더 많은것을 학습할수 있다는것을 알수 있다.

2) 시간신호우에서 훈련된 입력층분석

모든 파라메터들을 분석하지 못한 경우에도 충분히 훈련된 심층신경망의 첫번째 층안에서는 해석가능한 패턴들을 추출할수 있다.

생음성신호우에서 훈련된 첫번째 층무게행렬로부터 취한 4개의 행을 그림 2에 보여주었다.

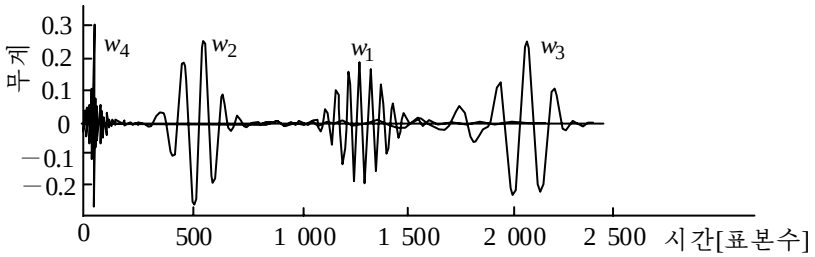

그림 2. 생음성신호우에서 훈련된
첫번째 층무게행렬로부터 취한 4개의 행

그림 2에서 시간은 10ms씩 17개 프레임에 대응한다. 즉 17 × 10ms × 16kHz = 2 720개이다.

따라서 심층신경망을 리용하여 자료로부터 대역통과려파기에 대응하는 한가지 형태의 임풀스응답들과 다른 패턴들을 학습하였다는것을 알수 있다.

발견된 려파기들의 스펙트르특성을 보여주기 위하여 무게행렬안의 매 행을 8 000개 항까지 령채우기하고 크기스펙트르를 계산한 다음 가장 찾기 쉬운 색쪼각무늬의 위치에 따라 행들을 분류하였다.

$$W_i = |FFT\{w_{ij}\}| \in \boldsymbol{R}^{1\times 8\,000},\ 1 \le i \le 2\,000 \tag{1}$$

가우스핵심 g를 가지고 스펙트르를 평활화한 후에 계산하면 다음과 같다.

$$\hat{W}_i = W_i \times g \tag{2}$$

$$f_c^i = \arg\max_{1\le j\le 8\,000} \{\hat{W}_{i,j}\} \tag{3}$$

시간신호우에서 훈련된 첫번째 층무게행렬로부터 취한 재순서화된 행렬의 크기스펙트르를 그림 3에 보여주었다.

매 행이 1개 대역통과임풀스응답으로서 해석될수 있다고 가정하면 색쪼각무늬의 위치는 학습된 전달함수의 중심주파수에 대응한다. 그림 3에서는 얻어진 스펙트르들을 $20\log_{10} W_i$로 표시하였다. 그림 3에서 보여준것처럼 아무런 사전지식도 없이 심층신경망

이 대략적인 청력학적인 분포를 나타내는 대역통과려파기들을 많이 발견하였다는것을 알 수 있다. 또한 낮은 주파수령역의 좁은 대역통과려파기수는 매우 크면서도 중심주파수가 증가함에 따라 려파기의 대역너비가 더 커진다. 즉 중심주파수들의 분포는 비선형이라는 것이다.

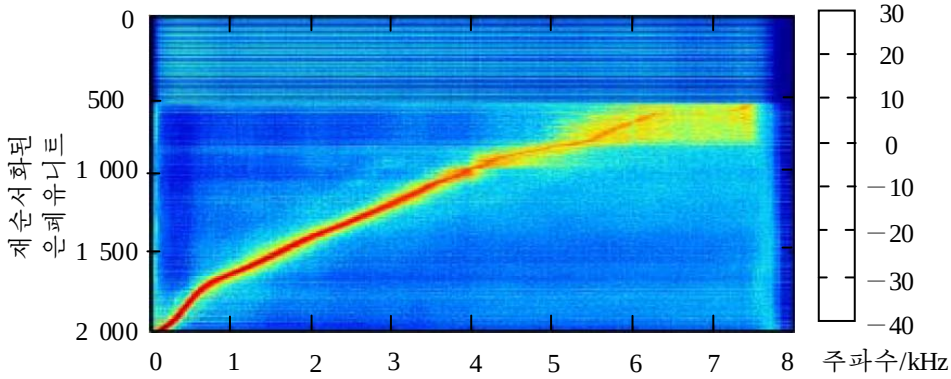

그림 3. 시간신호우에서 훈련된 첫번째 층무게행렬로부터 취한 재순서화된 행렬의 크기스펙트르

특징 및 활성함수비교와 10h자료우에서 훈련, 음소오유률을 표 5에 보여주었다.

표 5. 특징 및 활성함수비교와 10h자료우에서 훈련, 음소오유률

특징	개발/%		평가/%	
	시그모이드	정류선형	시그모이드	정류선형
MFCC	23.5	22.8	24.1	23.6
MFCC + PLP	21.0	20.7	21.7	21.3
|FFT|	24.6	23.3	25.0	23.7
시간신호	32.9	28.1	34.4	29.5

특징 및 활성함수비교와 500h자료우에서 훈련, 음소오유률을 표 6에 보여주었다.

표 6. 특징 및 활성함수비교와 500h자료우에서 훈련, 음소오유률

특징	개발/%		평가/%	
	시그모이드	정류선형	시그모이드	정류선형
MFCC	18.5	18.7	19.1	19.5
MFCC + PLP	18.0	17.6	18.7	18.4
|FFT|	20.0	18.9	21.4	20.1
시간신호	27.3	23.9	28.2	24.5

3) 정류선형세포를 리용한 대규모실험

활성함수를 교체하고 훈련자료량을 증가시킬 때 여러가지 특징들사이의 인식정확도 차가 어느 정도 줄어드는가를 확인하여야 한다.

먼저 시그모이드활성함수를 정류선형세포(ReLU)[3]와 비교한다. ReLU가 정규화에 민감하기때문에 0.000 1값을 가지고 L_2—정규화를 리용한다. 반대로 시그모이드비선형은 정류화를 하지 않을 때 가장 좋은 성능이 나타난다. 표 5에서 보여준 결과로부터 ReLU는

높은 오유률을 가진 체계에 더 적합하다는것을 알수 있다.

다음 500h음성우에서 훈련된 DNN을 가지고 생음성신호를 리용한 실험을 반복한다. 표 6은 얻어진 결과이다. 숨은 층수를 11까지 증가시키면서 실험을 진행하였다. 평가코퍼스에 대하여 21%의 음소오유률을 얻을수 있다. 즉 MFCC와 생음성신호사이의 성능간격을 좁힐수 있다.

맺 는 말

학습된 무게들에 대한 분석은 아무런 사전지식이 없이도 심층신경망이 순수 생음성신호의 시간령역에서 대역통과려파기모임을 학습할수 있다는것을 확인하였다. 결과적으로 다른 정상특징추출처리방법들을 실질적으로 대신할수 있다. 또한 MFCC와 PLP특징결합우에서 심층신경망을 훈련시켜 다른 기타 체계들과 비교함으로써 심층신경망이 특징들사이의 차이로부터 입력자료에 대한 추가적인 지식을 얻는다는것을 확증하였다.

참 고 문 헌

[1] Y. Ephraim et al.; IEEE Signal Processing Letters, 12, 2, 1666, 2005.

[2] C. Plahl et al.; in proc. Interspeech, Aug, 1237, 2011.

주체108(2019)년 11월 5일 원고접수

Performance Analysis of Various Features in Acoustic Modeling for Speech Recognition Using Deep Neural Networks

Kim Kwang Rim, Ri Jong Chol

In this paper we investigate various extractions in acoustic modeling for speech recognition using deep neural network(DNN) and show that using hybrid DNN/HMM acoustic models allows to obtain reasonable recognition results even without any processing of the raw time signal.

Keywords: acoustic modeling, raw signal, neural networks