pISSN 2005-8063
eISSN 2586-5854
2018. 03. 31.
Vol.10 No.1
pp. 33-38

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2018.10.1.033

잔향 환경 음성인식을 위한 다중 해상도 DenseNet 기반 음향 모델

Multi-resolution DenseNet based acoustic models for reverberant speech recognition

박 순 찬 · 정 용 원 · 김 형 순*

Park, Sunchan · Jeong, Yongwon · Kim, Hyung Soon

Abstract

Although deep neural network-based acoustic models have greatly improved the performance of automatic speech recognition (ASR), reverberation still degrades the performance of distant speech recognition in indoor environments. In this paper, we adopt the DenseNet, which has shown great performance results in image classification tasks, to improve the performance of reverberant speech recognition. The DenseNet enables the deep convolutional neural network (CNN) to be effectively trained by concatenating feature maps in each convolutional layer. In addition, we extend the concept of multi-resolution CNN to multi-resolution DenseNet for robust speech recognition in reverberant environments. We evaluate the performance of reverberant speech recognition on the single-channel ASR task in reverberant voice enhancement and recognition benchmark (REVERB) challenge 2014. According to the experimental results, the DenseNet-based acoustic models show better performance than do the conventional CNN-based ones, and the multi-resolution DenseNet provides additional performance improvement.

Keywords: convolutional neural network, DenseNet, multi-resolution, speech recognition

1. 서론

최근 음성인식 성능은 심층신경망(deep neural network: DNN) 기반 음향모델의 도입으로 크게 향상되었으나, 잔향 환경에서의 원거리 음성에 대해서는 여전히 개선의 여지가 많이 남아있다. 잔향이란 실내 환경에서 소리가 벽이나 천장 등에 의해 반사되어 시간차를 두고 강도가 약해진 음성이 함께 들어오는 현상을 말한다. 잔향 효과는 실내 공간이 넓을수록, 화자와 마이크 사이의 거리가 멀어질수록 그 영향이 커지게 된다. 최근 음성인식 기술의 적용 범위가 스마트폰을 이용한 근거리 음성인식에서 스마트 스피커, 로봇 등을 이용한 원거리 음성 인식으로 확장됨에 따라 잔향 환경에서의 왜곡에 강인한 음성인식 기술의 필요성이 대두되고 있다.

잔향 환경 음성인식 성능을 개선하기 위한 다양한 방법들이 연구되고 있는데, 합성곱 신경망(CNN, convolutional neural network) 기반의 음향 모델이 한 가지 방법이다. CNN은 영상인식 분야에서 먼저 그 효과가 확인되었고[1], 이후 연구에서 음향 모델에 적용되어 성능 향상 효과를 보여주었다[2]. CNN 기반의 음향 모델은 기존의 전결합 신경망(fully-connected neural network) 구조에 비해 스펙트로그램의 시간과 주파수 축의 정보

* 부산대학교, kimhs@pusan.ac.kr, 교신저자

Received 8 February 2018; Revised 12 March 2018; Accepted 12 March 2018

를 활용할 수 있다는 장점이 있다. 컴퓨터 비전 분야의 연구에서는 다양한 방법을 통해 보다 많은 수의 합성곱층(convolutional layer)으로 CNN 모델을 구성하였고, 이러한 구조가 뛰어난 성능 향상 효과가 있음을 보여주었다. 그러나 모델 구조가 더욱 깊어지면서 멀리 떨어진 합성곱층까지 정보가 전달되지 않는 문제점이 나타나게 되었고, 이러한 문제를 해결하기 위해 ResNet[3], Highway Network[4]와 같이 서로 떨어져 있는 합성곱층 사이를 이음으로써 정보를 효과적으로 전달하는 구조들이 제안되었다. 특히, DenseNet은 크기가 동일한 특징 맵(feature map) 사이의 연결(concatenation)을 통해 많은 수의 합성곱층들이 서로 이어지도록 하였으며, 최근 영상인식 분야에서 뛰어난 성능을 보여주었다[5].

본 논문에서는 DenseNet 구조를 음성인식에서의 음향 모델에 적용하고, 이전 연구에서 잔향 환경 음성에 대한 성능 개선 효과를 보여준 다중 해상도(multi-resolution) CNN을 확장한 다중 해상도 DenseNet 구조를 제안한다. 그리고 REVERB (REverberant Voice Enhancement and Recognition Benchmark) challenge 2014데이터를 통해 제안한 구조의 음향 모델과 기존 CNN 기반 음향 모델의 잔향 환경 음성인식 성능을 비교한다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 CNN 기반 음향 모델의 특징에 대해 간단히 설명하고, 3장에서는 DenseNet에 대한 설명과 이를 음향 모델에 적용하기 위한 방법에 대해 살펴본다. 4장에서는 제안한 다중 해상도 DenseNet 구조에 대해 설명하고, 5장에서 실험 환경과 결과에 대해 기술한 후 6장에서 결론을 맺는다.

2. CNN

일반적으로 DNN 기반의 음향 모델에서는 입력 특징으로 LMFE (log mel filterbank energy)가 주로 사용된다. LMFE는 스펙트로그램에 대한 멜 필터뱅크(mel filterbank)의 대역통과 필터(band pass filter)들의 출력을 구하고, 그 결과에 로그를 취함으로써 얻어진다. 음향 모델은 프레임 단위로 음소 정보를 추정하지만, 입력 특징에는 대상 프레임과 전후의 일부 프레임을 함께 사용하는 문맥 윈도우(context window)를 적용하여 시간에 따른 변화에 대한 정보를 제공한다. 따라서 DNN의 입력은 시간과 주파수 축을 갖는 2차원 행렬 형태로 나타낼 수 있다. 그런데 전결합 신경망에서는 입력과 출력의 모든 값들이 이어지기 때문에, 입력 특징이 시간과 주파수 축에 대한 정보를 잃어버리고, 단일 벡터와 같이 취급된다. 반면, CNN은 입력 특징을 시간과 주파수 축에 대한 정보를 보존한 2차원 행렬 형태로 처리할 수 있다는 장점이 있다.

CNN은 크게 합성곱층과 풀링층(pooling layer)으로 구성되며, 각층의 입력과 출력은 2차원 행렬 형태의 특징 맵들의 집합으로 이루어져 있다. 합성곱층은 입력 특징 맵과 훈련 가능한 파라미터로 이루어진 2차원 행렬 형태의 필터의 합성곱을 통해 출력 특징 맵을 얻는 역할을 수행한다. 합성곱층은 복수의 특징 맵을 출력할 수 있는데, 한 개의 출력 특징 맵을 구하기 위해서 모든 입력 특징 맵을 서로 다른 필터와 합성곱하고, 그 결과를 모두 더한다. 여기에 훈련 가능한 파라미터인 바이어스(bias)를 더한 후 활성함수를 통과하여 한 개의 출력 특징 맵이 얻어지는데, 모든 출력 특징 맵에 대해 동일한 과정을 수행한다. 따라서 한 합성곱층에는 입력과 출력의 특징 맵 수를 곱한 수만큼의 필터가 필요하다. <그림 1>의 왼쪽에 합성곱의 예시를 나타내었다.

풀링층은 입력 특집 맵을 일정한 크기의 2차원 영역들로 나누고, 각 영역을 한 개의 값으로 변환함으로써 특징 맵의 차원을 줄이는 역할을 수행한다. 풀링은 특정 영역에서 어떤 값을 선택하는지에 따라 여러 종류로 나뉘는데, 대표적으로 최댓값을 선택하는 최댓값 풀링(max pooling)과 평균값을 선택하는 평균값 풀링(average pooling)이 대표적이다. <그림 1>의 오른쪽에 최댓값 풀링의 예시를 나타내었다.

일반적으로 CNN 기반 음향 모델은 문맥 윈도우가 적용된 LMFE 입력에 대해 합성곱층과 풀링층이 교차로 반복되며, 시간과 주파수 영역에서 특징을 추출한 후, 분류를 위해 전결합층(fully-connected layer)으로 이어진다.

3. DenseNet

컴퓨터 비전 분야를 중심으로 합성곱층의 수를 늘리는 방법을 통해 성능을 향상시키는 방법들이 활발히 연구되었다. 그러나 기존의 CNN 구조에서 단순히 합성곱층의 수를 늘리는 방법으로는 더 이상 성능향상이 이루어지지 않았는데, 이를 해결하기 위해 VGGNet[6]에서는 합성곱층의 필터와 풀링의 크기를 최소화하고, 계층에 따라 필터 수를 적절히 조절하여 합성곱층의 수를 늘리는 방법을 통해 성능 향상 효과를 얻었다. 하지만 이러한 구조 역시 합성곱층의 수를 일정 이상으로 늘리면 성능이 포

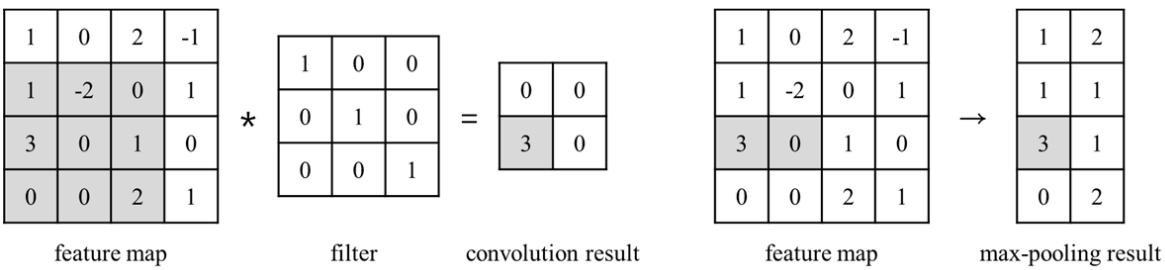

그림 1. CNN에서 합성곱과 최댓값 풀링의 예
Figure 1. Example of convolution and max pooling in CNN, convolutional neural network: CNN

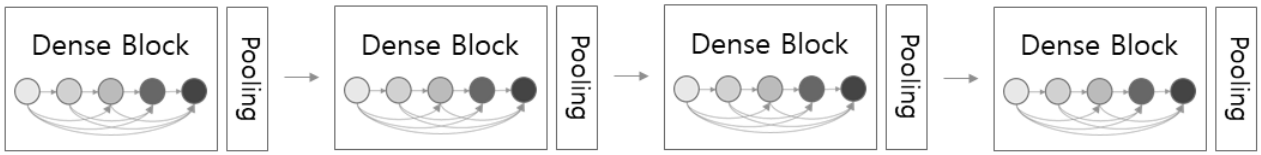

그림 2. DenseNet의 기본 구조
Figure 2. The basic structure of DenseNet

화되거나 오히려 나빠졌는데, 이는 계층 구조가 깊어짐에 따라 멀리 떨어진 합성곱층까지 정보가 전달되지 못하기 때문이었다.

DenseNet은 합성곱층의 입력에서 동일한 크기의 특징 맵들을 모두 연결함으로써 보다 깊은 구조에서도 훈련이 잘 이루어지게 한다[5]. 입력 특징 맵의 가장자리에 적절한 수의 0을 채우게 되면 합성곱층에서 더 이상 특징 맵의 크기가 줄어들지 않고, 풀링 층에서만 특징 맵의 크기가 줄어들게 된다. DenseNet에서는 풀링층을 기준으로 합성곱층들을 나누어, 동일한 크기의 입출력 특징 맵을 갖는 합성곱층들의 집합을 밀집 블록(dense block)이라 한다. 한 개의 밀집 블록 내에서는 합성곱층의 출력이 이후에 오는 모든 합성곱층들의 입력에 연결된다. 따라서 합성곱층의 입력은 이전 합성곱층들의 출력을 모두 연결한 결과가 되고, 출력은 일정한 수의 특징 맵이 되는데, 이때 지정한 합성곱층의 출력 특징 맵 수를 성장률(growth rate)이라 한다. <그림 2>에 DenseNet의 기본 구조를 그림으로 표현하였다.

앞에서 언급한 것처럼 합성곱층의 필터 수는 입력과 출력 특징 맵의 곱으로 결정된다. DenseNet 구조에서 각 합성곱층의 출력 특징 맵의 수는 모두 동일하지만, 입력 특징 맵의 수는 밀집 블록 내의 합성곱층의 수가 늘어남에 따라 함께 증가하게 된다. 따라서 합성곱층의 수가 늘어날수록 전체 파라미터 수가 급격히 증가하여 훈련이 제대로 이루어지지 않을 수 있다. 따라서 DenseNet에서는 전체 모델이 지나치게 방대해지는 것을 막기 위해 성장률을 작은 정수 값으로 제한한다.

DenseNet 구조를 음향모델에 효과적으로 적용하기 위해서는 입력 특징을 기존의 CNN 기반 음향모델과 다르게 구성할 필요가 있다. 보다 많은 수의 합성곱층으로 DenseNet을 구성하기 위해서는 밀집 블록의 수 또한 많아질 필요가 있고, 그러기 위해서는 풀링층의 수가 많아져야 한다. 그러나 특징 맵은 풀링층을 지날 때마다 크기가 줄어들기 때문에, 풀링의 크기를 최소화 하더라도 사용 가능한 풀링층의 수는 입력 특징 맵의 크기에 의해 제한된다. 따라서 보다 깊은 구조의 DenseNet을 구성하기 위해서는 입력 특징의 크기를 기존 CNN에 비해 크게 만들어야 한다. 시간 축으로는 문맥 윈도우의 크기를 늘림으로써, 주파수 축으로는 LMFE의 필터 수를 늘리는 방법을 통해 입력 특징의 크기를 확장할 수 있다.

4. 다중 해상도 DenseNet

이전 연구를 통해 다중 해상도 CNN 구조가 기존의 CNN 구조에 비해 잔향 환경 음성인식 성능을 개선함을 확인할 수 있었다[7]. 다중 해상도 CNN 구조는 스펙트로그램으로부터 두 종류의 특징을 추출하고, 서로 분리된 합성곱층과 풀링층의 스트림(stream)을 통과한다. 각 스트림의 출력은 하나로 결합된 후 전결합층을 거쳐 음소 정보를 추정하게 된다. 이때 두 입력 특징은 각각 좁은 문맥 윈도우의 협대역 LMFE와 넓은 문맥 윈도우의 광대역 LMFE로 구성되는데, 이를 통해 잔향 환경 음성의 특징을 보다 잘 반영할 수 있다.

다중 해상도 DenseNet은 기존의 다중 해상도 CNN을 DenseNet으로 확장한 구조이다. 각각의 입력 특징이 별도의 합성곱층과 풀링층을 통해 처리되는 점은 동일하지만, DenseNet에는 전결합층이 포함되지 않기 때문에 두 특징을 통합하는데 다른 방법이 필요하다. 다중 해상도 DenseNet에서는 DenseNet이 특징 맵 사이의 연결을 통해 합성곱층 사이를 잇는다는 점에 착안하여, 다른 특성을 가진 두 특징 맵 사이의 연결을 통해 이를 구현하였다. 전결합층을 통해 두 특징의 정보가 통합되는 다중 해상도 CNN에 비해, 다중 해상도 DenseNet은 시간과 주파수 축에 대한 정보가 보존된 상태로 통합되고, 그 결과가 다시 합성곱층을 통해 처리된다는 특징이 있다.

다중 해상도 DenseNet의 전체 구조를 <그림 3>에 나타내었다. 각 입력 특징이 별도의 밀집 블록들을 거친 후, 그 출력 특징 맵들이 연결되어 단일 밀집 블록의 입력이 된다. 이후 두 특징이 통합된 상태에서 밀집 블록들을 거친 후 선형 연결을 통해 음향모델의 출력을 얻게 된다. 서로 다른 입력 특징에서 얻어진 특징 맵들을 연결하기 위해서는 두 특징 맵의 크기가 동일해야 하는데, 이를 위해 두 입력 특징과 풀링의 크기가 적절히 조절되어야 한다.

5. 실험 및 결과

5.1. 실험 환경

본 논문에서는 잔향 환경 음성인식 성능 비교를 위해 REVERB challenge 2014 데이터를 사용하였다. REVERB challenge 2014는 음향모델 훈련을 위해 다중 조건 훈련(MCT, multi-condition training) 데이터와 성능 평가를 위한 잔향 환경 음성 데이터를 제공한다. MCT 데이터는 WSJCAM0[8] 데이터의 깨끗한 음성 7,861 문장과 다양한 잔향 환경에서 측정한 RIR(room impulse response)들을 합성곱하고, 배경 잡음을 더하여 생성된다. 제공되는 RIR 데이터는 측정된 방의 크기에 따라 3종류로 구분되며, 각각의 크기에는 2 종류의 방에서 측정한 RIR이 포함된다. 그리고 6개 방에 대해 화자와 마이크 사이의 거리 2종류, 각도 2종류로 총 24개의 RIR로 구성된다. 생성된 MCT 데이터를 통해 다양

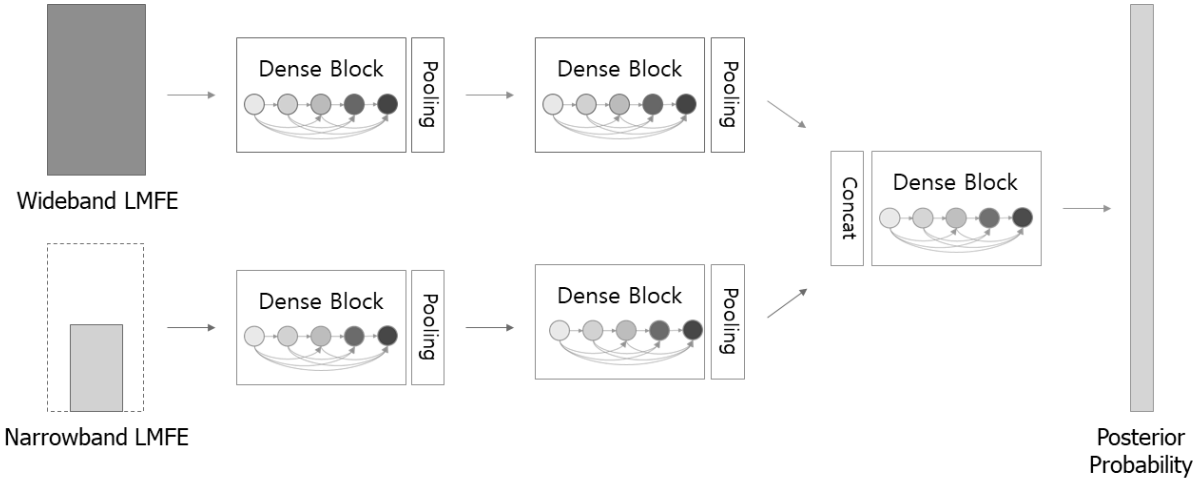

그림 3. 다중 해상도 DenseNet의 구조
Figure 3. The structure of multi-resolution DenseNet

한 잔향 환경에 대응할 수 있는 음향모델을 만들 수 있다.

평가를 위한 데이터는 SimData와 RealData로 구성된다. SimData는 MCT 데이터와 마찬가지로 왜곡이 없는 깨끗한 음성과 측정된 RIR을 합성곱하고, 배경 잡음을 더하여 생성된다. SimData는 3종류의 방과 2종류의 마이크와 화자 사이 거리의 조합을 통해 총 여섯 가지 환경에 대한 데이터로 구분되며, 각 환경별로 360여개의 문장으로 구성된다. 이때 RIR을 측정한 방의 잔향시간, 즉, T_{60} 은 각각 0.25, 0.5, 0.7초이며, 화자와 마이크 사이의 거리는 각각 0.5, 2m이다. SimData는 잔향의 수준이 서로 다른 여러 환경에서의 음성인식 성능을 비교할 수 있게 구성되어있다. RealData는 MC-WSJ-AV[9]의 데이터로, 실제 잔향 환경에서의 발화를 녹음한 음성이다. RealData는 T_{60} 이 0.7초인 방에서 대해 1m와 2.5m 거리에서 녹음된 음성으로 구성되며, 각각 180여개 문장으로 되어 있다. <표 1>에 REVERB challenge 2014 평가 데이터의 환경을 정리하였다.

표 1. REVERB challenge 2014 평가 데이터의 환경
Table 1. Environments of REVERB challenge 2014 evaluation set

Condition	SimData						RealData	
	Room1		Room2		Room3			
	Near	Far	Near	Far	Near	Far	Near	Far
Distance (m)	0.5	2.0	0.5	2.0	0.5	2.0	1.0	2.5
T_{60} (s)	0.25	0.25	0.5	0.5	0.7	0.7	0.7	0.7

REVERB, reverberant voice enhancement and recognition benchmark

REVERB challenge 2014의 모든 음성 데이터는 8채널 원형 배열 마이크를 통해 수집되었으나, 본 논문에서는 정면 1개 마이크를 통해 수집된 단일 채널 음성 데이터만을 사용하였다.

5.2. 실험 결과

실험을 위한 음성인식 시스템은 음성인식을 위한 오픈소스 툴킷인 Kaldi[10]와 신경망 훈련을 위한 오픈소스 툴킷인 CNTK[11]를 통해 구현하였다. 음향 모델 훈련을 위한 라벨(label)은 트라이폰(triphone)에 대한 3,268개의 상태 집합(tied state)으로, GMM- HMM(Gaussian Mixture Model – Hidden Markov Model) 모델의 강제 정렬(forced alignment)을 통해 생성하였다. 라벨 생성을 위한 GMM-HMM 모델은 Kaldi의 기본 제공 스크립트를 통해 훈련하였는데, REVERB challenge 2014의 MCT 데이터에 해당하는 WSJCAM0의 깨끗한 음성으로 훈련 및 강제 정렬하여 잔향에 의한 오류를 최소화하였다. 모든 신경망 모델은 확률적 경사 하강법(SGD, stochastic gradient descent)을 통해 상태 집합의 라벨과 출력의 CE(cross entropy)를 최소화하도록 훈련하였으며, 훈련 데이터의 10%를 검증 데이터로 분리하고, 교차 검증(cross validation)을 통해 epoch 단위로 CE의 변화에 맞추어 학습률(learning rate)을 자동 조절하였다.

잔향 환경 음성인식 성능 비교를 위해 CNN, VDCNN(very deep convolutional neural network), DenseNet, MR-DenseNet (multi-resolution DenseNet)의 네 가지 음향 모델을 구성하였다. <표 2>에 각 음향모델의 세부 구조를 정리하였다.

표 2. 실험에 사용된 음향모델의 세부구조
Table 2. Detailed structure of acoustic models for experiments

Model	Input	Convolution		Pooling		FCs
CNN	40×11	-		-		2,048*6
VDCNN	64×17	[3×3]*10				2,048*4
DenseNet	64×17	[3×3]*30 [1×1]*35		[2×1]*2 [2×2]*3		-
MR-DenseNet	64×17	[3×3]*12 [1×1]*15	[3×3]*18 [1×1]*20	[2×1]*2 [2×2]*1	[2×2]*2	-
	64×8	[3×3]*12 [1×1]*15		[2×1]*3		

CNN, convolutional neural network; VDCNN, very deep convolutional neural network; MR-DenseNet, multi-resolution DenseNet

CNN 모델은 비교를 위한 기준으로서 가장 많이 알려진 CNN 기반 음향 모델 구조를 적용하였다. 입력으로는 40차 LMFE를 좌우 5프레임과 함께 총 11프레임을 문장 단위로 평균 및 표준편차에 대해 정규화 하여 사용하였다. 입력은 9×9 크기의 필터

256개로 구성된 첫 번째 합성곱층을 거쳐, 3×1 크기의 최댓값 풀링층으로 연결된다. 풀링층의 출력은 두 번째 합성곱층의 4×3 필터들에 의해 처리되어 256개의 출력 특징 맵을 생성한다. 이들은 각 2,048개의 노드의 전결합층 4개를 거쳐, 선형 연결과 softmax 함수를 통해 상태 집합에 대한 확률로 변환된다.

VDCNN은 VGGNet을 기반으로 기존 연구에서 잡음 및 잔향 환경에서 뛰어난 성능 향상을 보여준 음향모델 구조를 적용하였다[12]. 입력 특징으로는 문장 단위의 평균 및 표준편차에 대한 정규화가 적용된 64차 LMFE를 좌우 8프레임씩, 총 17프레임을 사용한다. 모든 합성곱층에는 가장자리에 적절한 수의 0이 추가되어 특징 맵의 크기가 변하지 않으며, 필터의 크기는 3×3으로 고정된다. 필터의 수는 처음 두 개의 합성곱층에서는 64개로, 다음 네 개의 합성곱층에서는 128개, 나머지 4개의 합성곱층에서는 256개로 구성된다. 두 개의 합성곱층마다 출력에 최댓값 풀링층이 연결되며, 처음 두 개는 2×1, 이후로는 2×2의 크기를 갖는다. 10개의 합성곱층과 5개의 최댓값 풀링층 이후 각 2,048 노드로 이루어진 전결합층 4개로 연결되며, 선형 연결과 softmax 함수를 통해 상태 집합에 대한 확률로 변환된다.

DenseNet은 영상인식 분야의 연구 내용을 바탕으로 실험을 통해 음향 모델에 최적화된 형태로 구성하였다. DenseNet은 VDCNN과 동일한 입력 특징을 사용하며, 5개의 밀집 블록, 65개의 합성곱층, 5개의 풀링층과 상태 집합의 출력을 나타내기 위한 선형 연결과 softmax 함수로 구성된다. 성장률은 32, 밀집 블록에 포함된 합성곱층의 수는 각 4, 8, 12, 24, 12개이고, 나머지 5개의 합성곱층은 풀링층의 입력에 적용된다. 밀집 블록 내의 합성곱층은 2개가 한 쌍으로, 앞의 합성곱층은 성장률의 4배인 128개의 필터로 1×1 합성곱을 수행한다. 뒤의 합성곱층은 3×3 크기의 필터로 성장률 만큼의 특징 맵을 출력한다. 첫 번째 밀집 블록 이후에는 최댓값 풀링이, 나머지 밀집 블록 이후에는 평균값 풀링이 적용된다.

MR-DenseNet은 DenseNet을 다중 해상도 구조로 확장한 형태이며, 총 다섯 단계의 밀집 블록으로 이루어져 있다. 첫 번째부터 세 번째 밀집 블록까지는 각각의 입력 특징이 분리되어 처리되며, 네 번째 밀집 블록의 입력에서 두 스트림의 출력 특징 맵이 연결된다. 두 스트림의 기본적인 구성은 동일하나, 협대역 LMFE를 처리하는 스트림에서는 풀링의 크기를 조정하여 두 스트림의 출력의 크기를 같게 만들어 특징 맵 사이 연결을 가능하게 하였다. 입력 특징으로는 광대역(64~8,000 Hz)의 64차 LMFE 17프레임과 협대역(64~4,000 Hz)의 64차 LMFE 8 프레임을 사용하였다. 광대역 LMFE의 경우, 기준 프레임과 전후 각각 8 프레임씩을 사용하여 총 17 프레임으로, 그리고 협대역 LMFE의 경우, 기준 프레임에 이전 4 프레임과 이후 3 프레임을 합쳐 총 8 프레임의 문맥 윈도우를 구성하였다.

<표 3>에 각 음향모델 구조에 의한 잔향환경 음성인식결과를 정리하였다. VDCNN은 CNN에 비해 상당한 인식오류 감소효과를 보여주었고, DenseNet은 VDCNN에 비해 SimData에서 7.06%, RealData에서 3.74%의 추가적인 오류감소율(error reduction rate)을 보여주었다. 또한 MR-DenseNet는 SimData에서 5.27%, RealData에서 5.19%의 오류감소율을 나타내어, 다중 해상도 구조의 적용이 DenseNet에서 추가적인 성능 향상 효과가 있음을 확인할 수 있었다.

6. 결론

본 논문에서는 최근 영상인식 분야에서 뛰어난 성능을 보여준 DenseNet을 음향 모델에 적합한 형태로 적용하고자 하였다. 또한 이전에 제안한 다중 해상도 CNN 구조를 DenseNet으로 확장한 다중 해상도 DenseNet 구조를 제안하고, REVERB challenge 2014 데이터를 통해 기존의 음향모델과 잔향 환경에서의 음성인식 성능을 비교하였다.

잔향 환경의 데이터를 통해 음성인식 실험을 진행한 결과, DenseNet 기반의 음향모델이 기존의 VGGNet을 기반으로 한 VDCNN 모델에 비해 좋은 성능을 보여주었다. 그리고 본 논문에서 제안한 다중 해상도 DenseNet 기반 음향모델 실험을 통해 기존의 CNN에서 잔향 환경 음성에 대해 성능 향상 효과를 보여준 다중 해상도 구조가 DenseNet에서도 긍정적인 효과를 나타냄을 확인할 수 있었다. 다중 해상도 DenseNet의 합성곱층 단계에서 두 특징을 통합하는 것이 기존의 다중 해상도 CNN의 전결합층 단계에서 두 특징을 통합하는 것보다 효과적일 것으로 추정되며, 이를 확인하기 위한 추가 실험이 필요하다고 판단된다.

표 3. REVERB challenge 2014 데이터에 대한 실험 결과(word error rate, %)
Table 3. Experimental results on REVERB challenge 2014 evaluation set

Experimental condition	SimData							RealData		
	Room1		Room2		Room3		Average	Room1		Average
Acoustic model	Near	Far	Near	Far	Near	Far		Near	Far	
CNN	5.74	6.74	7.28	12.42	8.16	15.00	9.22	24.66	26.06	25.36
VDCNN	4.88	5.54	5.81	8.80	6.34	10.26	6.94	19.35	20.26	19.81
DenseNet	5.00	5.20	5.29	8.15	6.18	8.90	6.45	18.62	19.51	19.07
MR-DenseNet	4.47	5.12	5.22	7.43	5.71	8.73	6.11	17.66	18.50	18.08

reverberant voice enhancement and recognition benchmark: REVERB, convolutional neural network: CNN; very deep convolutional neural network: VDCNN, multi-resolution DenseNet: MR-DenseNet

감사의 글

본 연구는 산업통상자원부의 산업기술혁신사업으로부터 지원을 받아 수행된 연구임(No.10063424, 실내용 음성대화 로봇을 위한 원거리 음성인식 기술 및 멀티 태스크 대화처리 기술 개발).

참고문헌

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (pp. 1097-1105).

[2] Sainath, T., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A., Dahl, G., & Ramabhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks. *Neural Networks*, 64, 39-48.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

[4] Srivastava, R., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. *Proceedings of the Advances in Neural Information Processing Systems 28* (pp. 2377-2385).

[5] Huang, G., Liu, Z., Maaten, L., & Weinberger, K. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700-4708).

[6] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211-252.

[7] Park, S., Jeong, Y., & Kim, H. (2017). Multiresolution CNN for reverberant speech recognition. *Proceedings of the Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques.*

[8] Robinson, T., Fransen, J., Pye, D., Foote, J., & Renals, S. (1995). WSJCAMO: A British English speech corpus for large vocabulary continuous speech recognition. *1995 International Conference on Acoustics, Speech, and Signal Processing* (pp. 81-84). Detroit, MI. 1995.

[9] Lincoln, M., McCowan, I., Vepa, J., & Maganti, H. (2005). The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments. *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. San Juan* (pp. 357-362).

[10] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2011)* (p. 4). Hawaïi. 11-15 December, 2011.

[11] Yu, D., Yao, K., & Zhang, Y. (2015). The computational network toolkit. *IEEE Signal Processing Magazine*, 32(6), 123-126.

[12] Qian, Y., Bi, M., Tan, T., & Yu, K. (2016). Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12), 2263-2276.

• **박순찬 (Park, Sunchan)**
부산대학교 전기전자컴퓨터공학과
부산시 금정구 부산대학로63번길 2
Tel: 051-510-1704 Fax: 051-515-5190
Email: sunchanpark@pusan.ac.kr
관심분야: 음성인식, 음성신호처리

• **정용원 (Jeong, Yongwon)**
부산대학교 전자공학과
부산시 금정구 부산대학로63번길 2
Tel: 051-510-1704 Fax: 051-515-5190
Email: jeongy@pusan.ac.kr
관심분야: 음성인식, 음성신호처리

• **김형순 (Kim, Hyung Soon)** 교신저자
부산대학교 전자공학과
부산시 금정구 부산대학로63번길 2
Tel: 051-510-2452 Fax: 051-515-5190
Email: kimhs@pusan.ac.kr
관심분야: 음성인식 및 합성, 음성신호처리