http://dx.doi.org/10.14400/JDC.2016.14.8.233

발화구간 검출을 위해 학습된 CNN 기반 입 모양 인식 방법

김용기*, 임종관**, 김미혜*
충북대학교 컴퓨터공학과*, KAIST**

Lip Reading Method Using CNN for Utterance Period Detection

Yong-Ki Kim*, Jong Gwan Lim, Mi-Hye Kim***
Dept. of Computer Engineering, Chungbuk National University*
Dept. of Mechanical Engineering, KAIST**

요 약 소음환경에서의 음성인식 문제점으로 인해 1990년대 중반부터 음성정보와 영상정보를 결합한 AVSR(Audio Visual Speech Recognition) 시스템이 제안되었고, Lip Reading은 AVSR 시스템에서 시각적 특징으로 사용되었다. 본 연구는 효율적인 AVSR 시스템을 구축하기 위해 입 모양만을 이용한 발화 단어 인식률을 극대화하는데 목적이 있다. 본 연구에서는 입 모양 인식을 위해 실험단어를 발화한 입력 영상으로부터 영상의 전처리 과정을 수행하고 입술 영역을 검출 한다. 이후 DNN(Deep Neural Network)의 일종인 CNN(Convolution Neural Network)을 이용하여 발화구간을 검출하고, 동일한 네트워크를 사용하여 입 모양 특징 벡터를 추출하여 HMM(Hidden Markov Model)으로 인식 실험을 진행하였다. 그 결과 발화구간 검출 결과는 91%의 인식률을 보임으로써 Threshold를 이용한 방법에 비해 높은 성능을 나타냈다. 또한 입모양 인식 실험에서 화자종속 실험은 88.5%, 화자 독립 실험은 80.2%로 이전 연구들에 비해 높은 결과를 보였다.

주제어 : 이미지 프로세싱, 시청각음성인식, 입모양 인식, 발화구간검출, 심화신경망

Abstract Due to speech recognition problems in noisy environment, Audio Visual Speech Recognition (AVSR) system, which combines speech information and visual information, has been proposed since the mid-1990s,. and lip reading have played significant role in the AVSR System. This study aims to enhance recognition rate of utterance word using only lip shape detection for efficient AVSR system. After preprocessing for lip region detection, Convolution Neural Network (CNN) techniques are applied for utterance period detection and lip shape feature vector extraction, and Hidden Markov Models (HMMs) are then used for the recognition. As a result, the utterance period detection results show 91% of success rates, which are higher performance than general threshold methods. In the lip reading recognition, while user-dependent experiment records 88.5%, user-independent experiment shows 80.2% of recognition rates, which are improved results compared to the previous studies.

Key Words : Image Processing, AVSR, Lip Reading, Motion Segmentation, DNN

Received 20 June 2016, Revised 1 August 2016
Accepted 20 August 2016, Published 28 August 2016
Corresponding Author: Mii-Hye Kim
(Dept. of Computer Engineering, Chungbuk National University)
Email: mhkim@cbnu.ac.kr

ISSN: 1738-1916

1. 서론

인간과 컴퓨터의 상호작용 수단 중 하나인 음성은 인간과 기계 간 정보교환의 효율적인 방법 중 하나이다. 음성 인터페이스는 기존의 마우스, 키보드를 사용하는 사용자 인터페이스보다 더 직관적이고 다양한 형태의 상호작용이 가능하다는 이점이 있다[1].

최근 음성인식기술은 이미 상용화되어 다양한 제품에 적용하여 출시되고 있다. 그러나 음성인식은 소음이 존재하는 실제 환경에서 인식률이 하락하는 문제점이 있다. 이를 극복하기 위해 음성 신호만으로 여러 가지 방법을 적용시켜 소음 환경에서 음성 인식률의 하락을 막는 방법들이 연구되어 왔다[2]. 그 중 한 가지 방법으로 소음환경에서의 음성인식 문제점으로 인해 1990년대 중반부터 음성정보와 영상정보를 결합한 AVSR(Audio Visual Speech Recognition) 시스템이 제안되었고, Lip Reading은 AVSR 시스템에서 시각적 특징으로 사용되었다[1,2].

본 연구는 효율적인 AVSR 시스템을 구축하기 위해 입 모양만을 이용한 발화 단어 인식률을 극대화하는 동시에 하나의 네트워크로 발화구간 검출과 입 모양 인식이 가능한지 확인하는데 목적이 있다. 이 두 가지 목적을 달성하기 위해 발화구간 자동 검출로 학습시킨 CNN (Convolution Neural Network)을 이용하여 발화구간 검출 실험을 하였고, 동일한 네트워크로 입 모양 특징을 추출하였다. 추출된 입 모양 특징을 차원 축소하여 HMM(Hidden Markov Model)으로 인식 실험을 진행하였고, 그 결과 평균 인식률이 80.2%로 나타났다.

2. 이전 연구들의 분석

Lip Reading은 <Table 1>과 같이 AVSR 시스템이 제안되면서부터 사용하는 특징점, 특징점 차원의 축소 방법, 인식단의 설계에서 연구가 차별화 되어 왔다.

입 모양 인식 연구는 크게 입술 영역 내의 여러 정보들을 특징으로 생성하는 방법과 입 모양을 근사화한 좌표를 기반으로 입술 움직임을 나타내는 벡터들의 집합을 특징으로 생성하는 방법, 이 두 가지를 결합하는 방법 등으로 나눌 수 있다.

입술 영역 내의 정보를 기반으로 한 특징을 이용한 연구에서는 입술 영역의 영상의 화소값을 특징으로 삼아 시간상에서 영상값의 변화에 중점을 둔 연구[3,4]와 조음기관 전체의 영상 정보를 사용하되 Sobel 연산자를 이용한 edge 정보, 입술 영역 움직임에 특화시킨 Optical Flow를 특징으로 사용하는 연구[5,6]가 이루어졌다.

검출된 입 모양 근사화에 기반한 특징을 이용한 연구에서는 발화시 주요한 특징을 입술의 윤곽으로 국한시켜 Snake Shape Model, Active Shape Model, Active Appearance Model, Convex Hull 등으로 입술 윤곽을 근사화 하여 입술의 움직임을 특징으로 하는 연구 등이 진행되어 왔다[2,7,8,9,10].

그리고 입술 영역 내의 정보를 기반으로 한 특징과 입

<Table 1> Previous Approaches

Approach	Feature Generation Method	Recognizer	Experiment Word	Recognition Rate
[1]	points are designated for lip shape approximation, Grid based Features (Optical Flow, Sobel, Gray)	DTW	6 Isolated Word	60.56%
[2]	ASM	HMM	One, Two, Three, Four	90.6%
[3]	Binary Image of Lip and Teeth	NN	/a/, /i/, /u/, /e/, /o/	58.6%
[4]	Gray Feature of Lip Region	HMM	Il, I, Sam, Sa, O, Yuk, Chil, Pal, Gu, Gong, Young	66.3%
[5]	Optical Flow	SVM	14 English Phonemes	95%
[6]	Optical Flow	SVM	14 English Phonemes	97%
[7]	AAM	HMM	ah, eh, f, ao, t, uh, w, k, p, iy, aa, ch, oo	75.26%
[8]	ASM	SVM	a, e, i, o, u	74.73%
[9]	points are designated for lip shape approximation	HMM	One, Two, Three, Four	68.3%
[10]	Convex Hull	HMM	Isolated Word	68%
[11]	Snake Model, Inner Mouth Area	NN	ba, bi, bou	72.73%
[12]	ASM, Binary Image of Inner Mouth	HMM	Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine	84%

모양 근사화에 기반한 특징을 결합한 연구도 진행되었다[11,12].

또한 위 특징을 검출 한 후, 불필요한 정보를 제거함으로써 정보량을 압축하려는 시도로 PCA(Principal Component Analysis, DCT) 등이 적용되었다[13].

인식기는 HMM, SVM(Support Vector Machine), NN(Neural Network) 등이 공통적으로 사용되었다.

기존의 방법들을 분석 평가한 본 연구의 이전 연구에서는 입술 영역의 화소 단위 변화량보다 입술 특정 부위의 위치 변화를 검출하는 것이 더 정확한 입술의 움직임을 추정할 수 있으며, 단어 인식률이 개선됨을 보여주었다[1]. 상대적으로 낮은 인식률을 보이는 고립 단어에 대해 우리는 다양한 특징점의 성능을 DTW(Dynamic Time Wrapper)를 사용해 점검하였다. 입술 영역을 화소 단위 혹은 화소를 지역별 격자 단위로 묶어 명암, Oprical Flow, Sobel 연산자를 사용하는 경우, 명암 잡음으로 인하여 검출된 입술 영역에 대해 영상 흔들림 현상이 관찰되기 때문에 인식률 저하 현상이 발생하였다. 입술 움직임 정보를 입술 윤곽을 근사화한 좌표를 이용해 추출하는 경우가 더 주효한 특징이 되는데, 특히 입술의 가로, 세로 비율이 발화 단어를 인식하는 가장 중요한 정보임을 재확인하였다. 모음에 주도적으로 영향을 받는 음성인식과는 달리 특정 자음 발화시 발생하는 입술의 닫힘 현상이 단어인식에 주요한 특징이 될 수 있음을 확인하였다.

그러나 입술의 가로, 세로 비율의 변화를 통해 입술의 움직임을 검출하는 방법은 다음과 같은 문제점을 안고 있다. 첫 번째, 입 모양 인식 시스템에서 입술 윤곽을 검출하는 방법 및 입술 영역의 움직임, edge 정보, 영상 정보 등을 추출하는 단계는 정확성과 안정성에 있어 매우 중요한 문제이다[14]. 그러나 영상 처리 시 입술 영역의 검출에 대한 많은 시도들이 이루어져 있으나, 정확한 입술 영역 검출은 여전히 어려운 문제이다[13,14,15]. 기존 우리의 연구에서 검출된 입술 영역의 흔들림 문제가 인식률 저하로 나타나는 문제가 발견되었다.

두 번째, Lip Reading의 자동화에 필수적인 입술 움직임 구간 검출 역시 입술의 윤곽 검출이나 입술 영역 영상값 등에 종속적[1,16,17]인데 영상 잡음으로 인한 낮은 검출 성공률이 움직임 구간 검출의 성공률마저 저해하는 것으로 나타났다.

위에서 제기된 입술 영역 혹은 특정 입술 부위 검출의 어려움과 움직임 구간 검출 저해 문제를 동시에 해결하기 위해 우리는 CNN을 이용한 발화구간 검출 방법과 그와 동일한 네트워크를 이용한 입술 움직임 특징 추출을 제안한다. 본 논문에서는 동일한 네트워크를 사용할 경우, 학습시간 및 실행시간을 단축시킬 수 있으며, 발화 구간 검출과 입 모양 인식에 대한 별도의 특징 처리가 불필요하다는 연구결과를 도출하였다.

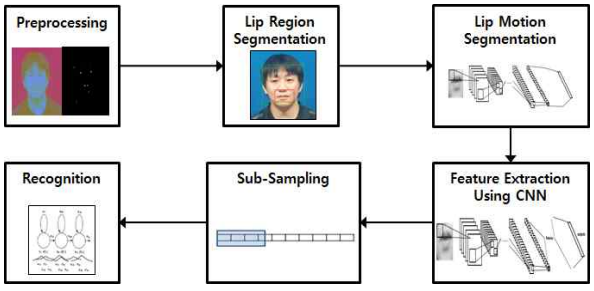
하드웨어 및 인터넷 환경이 급속도로 발전함에 따라 DNN(Deep Neural Network)은 기존의 영상 처리 분야 연구인 물체 인식 및 검출, 얼굴 인식, 장소 인식, 사람의 자세 인식 및 검출, 인간의 골격 인식, 깊이 추정 등의 문제에서 기존의 영상 처리 알고리즘의 성능을 크게 넘어섰다[18,19]. 또한 인간이 설계한 특징 생성 방법을 DNN 특징들로 대체하는 연구들이 등장하고 있다.

DNN의 일종인 CNN은 Yann Lecum이 각 데이터 차원간 기하적 연관성을 가지는 데이터를 효과적으로 인식하기 위해 처음 제안하였으며, ILSVRC(Imagenet Large Scale Visual Recognition Challenge)에서 2012년 압도적으로 높은 성능을 보인 후 DNN을 이용한 방법들이 계속 상위에 위치하고 있다[20]. 학습된 CNN에 가공하지 않은 이미지 데이터를 입력하면 별도의 특징 생성 과정 없이 가공하지 않은 이미지에 대해 Convolution 연산과 Pooling 연산을 여러 차례 번갈아가며 수행함으로써 위치 정보가 제거된 이미지 패턴 정보를 가진 특징들이 생성된다. 또한 CNN을 학습할 때 이미 잡음에 노출된 데이터를 학습 데이터에 포함시킴으로써 성능을 개선하는 연구들이 보고되었다[21]. 따라서 이전 연구들에서 시도되었던 별도의 입술 윤곽 추정 등의 특징생성 과정이 없고, 영상 흔들림으로 인한 입술 영역의 위치 변화에 강인한 특징을 학습과정에서 자동으로 생성할 수 있다. 그리고 Lip Reading의 전처리에 포함되나 Lip Reading을 위한 특징점에 공통적으로 의존하는 기존 입술 발화 구간 검출 방법을 대신하기 위해, Lip Reading 특징점 추출과 함께 CNN을 사용하는 입술발화구간 검출도 제안한다.

3. 발화구간 검출 CNN을 이용한 입 모양 인식 방법

본 연구에서 CNN을 이용한 입 모양 인식을 위해

[Fig. 1]과 같이 전처리, 입술 영역 검출, 발화구간 검출 및 특징 생성을 수행 후, 차원 축소를 거쳐 HMM으로 고립단어 인식 실험을 수행하였다.

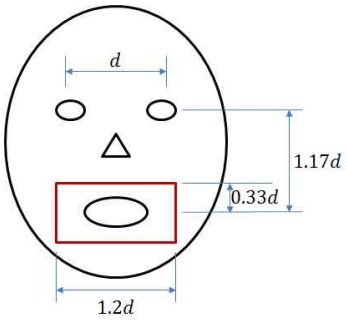


[Fig. 1] Flow Chart of This Work

3.1 전처리 및 입술 영역 검출

모바일로 촬영된 크기 1080x1920으로 입력된 영상 데이터에 대해 크기를 줄이기 위하여 280x500으로 크기를 축소시킨다. 그리고 휘도 보정을 위해 YCbCr 색상공간 변환을 거친 후 히스토그램 평활화를 수행하였다. RGB 색상공간 영상은 R, G, B 각각에 대한 정보를 모두 가지고 있기 때문에 정보량이 많아지고 주변 조명의 변화에 상대적으로 민감하다. 따라서 R, G, B 색상 정보를 밝기 정보인 Y와 색상정보인 Cb, Cr로 변환한다. 이후, 밝기 분포가 특정한 부분으로 치우친 것을 해소하기 위해 넓은 영역에 걸쳐 밝기 분포를 넓히는 Histogram Equalization을 Y 채널에 대해 수행하였다. 영상의 절대적 밝기의 크기보다 대비가 증가할 때 인지도가 증가한다는 점에 근거한다.

전처리 과정 후 입술 영역 검출을 위해 눈 위치에 기반하여 입술 영역을 검출하였다. 먼저 Viola와 Jones가 제안한 Adaboost 알고리즘을 이용하여 얼굴 영역을 검출한다. Adaboost 알고리즘은 Haar-Wavelet 특징을 이용하여 적절한 약 분류기를 선택하고 이에 가중치를 부여하는 알고리즘이다[22]. Historam Equalization을 수행한 Y 채널값으로 구성된 그레이 레벨 영상에서 닫기 연산한 영상과 원 영상의 차를 이진화하여 얼굴영역에서 상대적으로 어두운 눈 영역을 검출할 수 있다[23]. 이후, 눈 위치와 입술 간 거리 비율을 활용하여 입술의 위치를 검출하며 이때 사용하는 얼굴 마스크는 [Fig. 2]와 같이 [23]를 활용하였다.



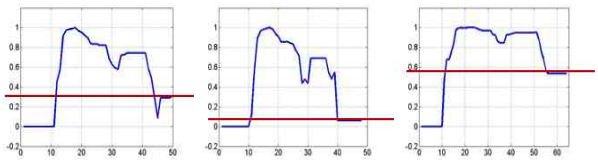
[Fig. 2] Face Mask[23]

이렇게 검출된 입술 영역 영상은 개인 간 눈과 눈 사이의 거리가 모두 다르고, 영상 잡음으로 인하여 동일 화자에서도 검출된 입술 영역의 크기가 모두 다르다. 따라서 동일한 크기로 이미지 크기를 정규화 하였고, 본 연구에서는 이후 CNN에서 사용할 이미지의 크기인 32x32로 크기를 정규화 하였다.

3.2 발화구간 검출

발화구간 검출은 시간에 따라 변화하는 입 모양에 근거하여 입력되는 비디오 스트림을 발화 구간과 비발화구간으로 구분하는 것을 뜻한다.

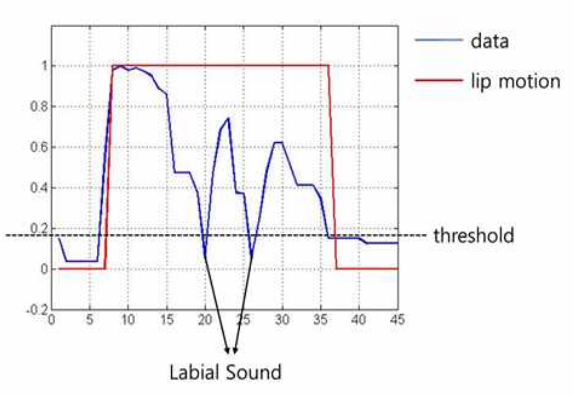
자동음성인식에서는 End Point Detection에 해당하고 Motion Recognition에서는 Motion Segmentation에 해당하는 문제이며 인식률의 향상에 영향을 끼칠 뿐만 아니라 자동화에는 필수불가결한 과정이다[24,25]. 가장 직관적인 발화구간 검출 방법은 입술의 특정 부위의 위치 변화에 임계값을 적용시키는 것인데, 다양한 환경, 입술 모양의 개인차, 영상 흔들림에 대해 보편적인 임계값을 결정하는 것이 까다롭다. 또한 음성은 발화되지 않으나, 입술을 벌린 상태를 지속함으로 인해 발화가 지속되는 것으로 오판단되는 상황이 발생하여, 일종의 Drift Erorr 역할을 한다.



[Fig. 3] Problem for Apply Threshold

[Fig. 3]은 개인별로 달라지는 임계값과 이 Drift Error를 보여준다. 특히 [Fig. 3]-(c)는 55 Sample를 초과하는 구간은 비발화 구간이나 불명확한 위치 정보로 인해 일괄적인 임계값 적용이 불가능하다.

마지막으로, 입술소리인 /ㅁ/, /ㅂ/을 발화시 비발화구간과 동일한 입술의 닫힘현상이 발생하므로 역시 개선이 필요하다. [Fig. 4]은 입술소리와 비발화구간을 Threshold로 구별 할 수 없다는 것을 보여준다.



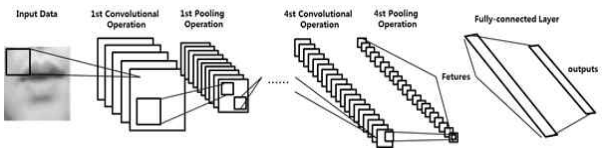
[Fig. 4] Problem for Labial Sound

우리는 발화구간 검출을 위해 Alext Krizhevsky가 CIFAR-10 데이터 셋을 분류하기 위해 제안한 네트워크 모델을 사용하였다[26].

CIFAR-10은 80만장의 32x32 크기 이미지로 구성되어 있고, 이미지들은 10개의 클래스(비행기, 자동차, 새, 고양이, 사슴, 개, 개구리, 집, 양, 트럭)로 분류되어 있는 데이터 셋이다.

Alext Krizhevsky가 제안한 네트워크 모델은 주어진 영상 데이터를 분류하기 위한 최적의 Edge를 자동으로 검출해준다. 이미지에서 Edge정보는 영상 데이터에서 불필요한 정보를 줄여주어 이미지 처리에서 객체 검출 문제에 이용하거나 이미지 인식 문제에서 효과적인 특징이 될 수 있다[6,13,17]. 동시에 해당 객체의 위치 정보를 줄여주기 때문에 입 모양 인식에 있어 이미지 흔들림과 같은 문제를 해결하는데 적합하다.

본 연구에서는 [Fig. 5]와 같이 Convolution 연산과 Pooling 연산을 담당하는 레이어를 4번 중첩 후 여기서 출력된 데이터를 Fully-Connected Layer에 입력시켜 최종적으로 발화구간과 비발화구간으로 출력하였다.



[Fig. 5] This Work`s Network Model

Convolution 연산은 입력된 영상을 스무딩(Smoothing)하여 찾고자 하는 잡음의 영향력을 축소하며, 식 (1)과 같이 2차원 영상 입력에 대해서 $M \times N$ 크기의 커널(Kernel)을 180도 뒤집어 모든 영역에 대해 커널과 화소 사이의 상관도를 계산한다.

$$\begin{aligned} g(x,y) &= h(x,y)*f(x,y) \\ &= \sum_{s=-at=-b}^{a}\sum^{b} h(s,t)f(x-s,y-t) \\ a &= \frac{M-1}{2} \\ b &= \frac{N-1}{2} \end{aligned} \quad (1)$$

여기서 $h(x,y)$는 (x,y) 위치에서의 필터를 의미하며, $f(x,y)$는 (x,y) 위치의 화소값을 의미한다. 본 연구에서는 5×5 크기의 Convolution 커널을 사용하였고, 필터는 가우시안 필터를 사용하였다.

Pooling 연산은 사용하는 커널의 크기가 결정하는 영역 내의 최대값을 해당 영역의 대푯값으로 결정함으로써, 영상의 크기를 $1/(W \times W)$로 축소시킨다. W는 커널 한 변의 크기이다. 이러한 압축 과정을 통해 이미지는 추상화되며 노이즈에 강인해진다[27]. 본 연구에서는 3×3 Pooling 커널을 사용하였다.

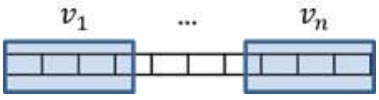
Alext는 이미지 가장자리를 처리하기 위해 모든 레이어에서 동일하게 가장자리를 처리하는 Global Unit을 사용하였으나, 본 연구에서는 사용하지 않았으며 Overfitting을 줄이기 위한 Drop Out 또한 사용하지 않았다.

3.3 CNN을 이용한 특징 생성 및 인식

우리는 [Fig. 5]에 나타난 것과 같이 발화구간 검출에서 학습된 네트워크와 동일한 네트워크를 이용하되 최종 레이어는 생략한다. 위 네트워크가 발화구간을 구분할 정도로 입 모양의 변화를 제대로 반영하므로 동일한 네트워크를 이용함으로써 학습시간 및 실행시간을 단축

시킨다. 크기 32×32의 입력 영상이 Convolution 연산과 Pooling 연산을 4차례 거친 후, 최종 출력이 1024×1 차원 벡터로 변환된다.

추출된 특징 벡터를 [Fig. 6]와 같이 n개의 Sub-Sampling 구역으로 나눈 후 해당 구역의 평균을 계산, 해당 구역의 대푯값으로 삼는다.



[Fig. 6] Sub-Sampling

따라서 1024차원의 특징점 벡터는 n차원의 벡터로 축소된다. 본 연구에서는 10차원 벡터로 축소하였다.

인식기는 음성과 같은 시계열 데이터 인식 문제에서 가장 뛰어난 성능을 보이는 HMM을 사용하였다. 입력되는 이미지 프레임 별로 검출되는 발화구간 검출과는 달리, 단어 인식에 이미지 프레임의 시계열 정보가 중요하다. 그러나 시계열 데이터는 비선형적인 특성을 갖는 신호에서 얻어지는 데이터들의 불확실성 때문에 확률모델을 주로 사용한다. 본 연구에서는 시계열 데이터에 특화된 HMM을 사용하여 인식 실험을 진행한다[2,4,7,9,10,12,28].

4. 실험 및 결과

4.1 실험 DB

AVSR 시스템을 위한 영어 기반의 학술용 오픈 데이터베이스는 다양한 반면, 한국어 기반 학술용 오픈 AVSR 데이터베이스는 단순한 숫자에 대한 발화 정도로 국한되어 있다. 따라서 본 연구에서 제안한 방법을 검토하기 위한 발화 단어에 대한 데이터베이스를 자체적으로 10개 단어에 대해 10명으로부터 수집한 동영상으로 구축하였다. 본 논문에서 제안한 실험의 주목적은 모바일 환경에서 Wake-Up 기능이 가능한 명령어들에 대해 입모양만을 이용한 발화 단어 인식이다. Wake-Up 기능의 시작을 알리는 “하이”와 다양한 모바일 명령어의 합성어를 한 단어로 구성하여 데이터베이스를 구축하였다.

실험단어는 “하이갤럭시(w1)”, “하이알라딘(w2)”, “하이스마트폰(w3)”, “하이카메라(w4)”, “하이메시지(w5)”, “하이카카오톡(w6)”, “하이전화걸기(w7)”, “하이내비게이션(w8)”, “하이이메일(w9)”, “하이시스트란(w10)”으로 스마트폰 음성명령과 관련된 10개의 단어로 구성하였다.

스마트폰을 가장 많이 사용하는 연령대를 고려하여 피험자는 모두 20대와 30대로 총 10명(남자 5, 여자 5)이며, 10개의 단어에 대해 10회씩 발화하여 총 1000개의 DB를 구축하였다. 실험 DB는 일반적인 스마트폰 동영상 영상 규격인 1080p(1920×1080)를 30fps로 촬영하였다.

4.2 실험 내용

발화구간 검출은 화자독립, 10-fold validation으로 검증하였다. 총 1000개의 발화영상 중에서 9명이 10개의 실험단어를 10회 발화한 900개의 발화 영상을 학습용으로, 1명이 10개의 실험단어를 10회 발화한 100개의 발화영상을 테스트용으로 사용하였다. 이를 화자 10명에 대해 동일한 방법으로 10번 반복하여 획득한 10개의 인식률에 대해 평균을 계산하였다. 또한 발화구간 Labeling은 모든 데이터에 대해 수동으로 검출하였다.

입 모양 인식실험은 화자독립과 화자종속 실험으로 구분하여 동일하게 10-fold validation으로 검증하였다. 화자 독립 실험은 발화구간 검출 실험 방법과 동일하게 진행하였다. 화자종속 실험을 위해 1명이 10개의 실험단어를 9회 발화한 90개의 발화 영상을 학습용으로, 10개의 실험단어를 1회 발화한 10개의 발화 영상을 테스트용으로 사용하였다. 이 과정을 10명에 대해 동일한 방법으로 진행하여 획득한 10개의 인식률에 대해 평균을 계산하였다.

본 연구에서는 CNN을 이용한 실험을 진행하기 위해 CNN의 대표적인 Tool인 CAFFE를 이용하였다. CAFFE는 BVLC에서 공개 소프트웨어로 배포하고 있는 소프트웨어로 핵심 GPU는 c++언어와 CUDA로 작성되어 있으며, 핵심코드를 고칠 필요없이 Protobuf라는 텍스트 설정 파일만을 변경함으로써 CNN의 구조를 바꾸는 것이 가능한 것이 특징이다.

4.3 실험 결과

4.3.1 발화구간 검출 실험 결과

발화구간 검출 인식실험 결과 <Table 2>에 나타난 것과 같이 평균 91%로 나타났다.

<Table 2> Utterance Period Detection Result(%)

	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	mean
Threshold	70.8	70.2	57.8	77	67.7	68.5	62.6	65.9	66.9	67	67.44
CNN	95.4	94.3	95.7	92.1	97.8	74.5	90.3	94.5	97	79.1	91

검출 실험 결과 Threshold를 이용한 검출 결과인 67.44%보다 높은 결과를 나타냈다. 6번 화자와 10번 화자를 제외한 나머지 화자들은 인식결과가 90% 이상으로 나타났으며, 6번화자와 10번화자는 각각 74.45%와 79.14%로 비교적 검출 성능이 낮았다. 실험 결과는 Threshold를 이용한 발화구간 검출 방법의 한계를 CNN을 이용한 발화구간 검출 방법으로 극복할 수 있는 것을 보여준다. <Table 3>은 전체 데이터에 대한 발화구간 검출의 Confusion Matrix이다.

<Table 3> Utterance Period Detection Confusion Matrix

	Non-Utterance Period	Utterance Period
Non-Utterance Period	7568	1420
Utterance Period	6171	42669

발화구간이지만 비발화구간으로 인식된 횟수는 6171로 10.67% 였고, 비발화구간을 발화구간으로 인식한 경우는 1420으로 2.4% 였다. 발화구간을 비발화구간으로 인식한 경우는 입술소리로 인한 것으로 분석되었다.

CNN을 이용하여 발화구간을 검출하는 것은 대체적으로 만족스러운 결과로 나타났으나, 입술소리와 비발화구간을 구분하기 위한 추가적인 연구가 필요하다.

4.3.2 입 모양 인식 실험 결과

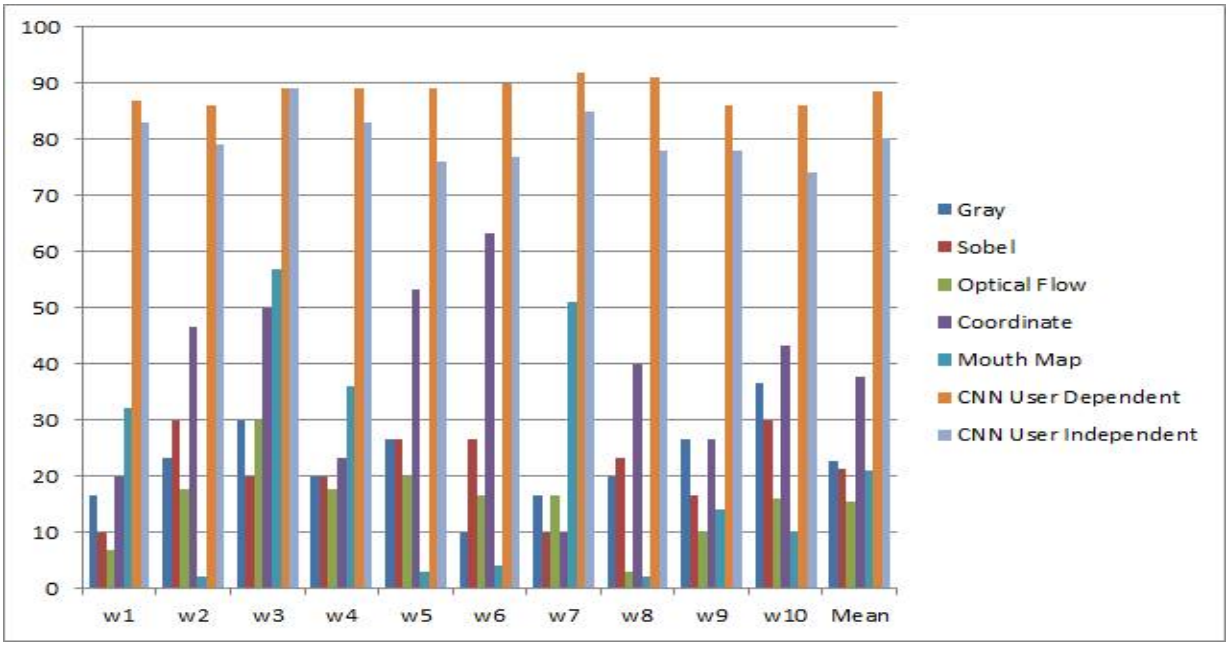
CNN을 이용하여 특징을 추출하고 HMM으로 인식실험을 진행한 실험 결과 <Table 4>, [Fig. 7]와 같이 화자종속 실험에서 평균 88.5%, 화자독립 실험에서 평균 80.2%로 기존 방법들을 분석 평가한 이전 연구[1]에 비해 월등히 높은 인식 성능을 나타냈다.

화자종속실험에서 “하이전화걸기(w7)”가 92%로 가장 높은 인식률을 보였고, “하이알라딘(w2)”, “하이이메일(w9)”, “하이시스트란(w10”이 각각 86%로 인식률이 낮았다. 반면, 화자독립실험에서는 “하이스마트폰(w3)”이 85%로 인식률이 가장 높았고, “하이시스트란(w10)”은 74%로 인식률이 가장 낮았다. [Fig. 7]의 결과에서 나타난 것처럼 화자종속 실험과 화자 독립실험의 실험 결과 인식률 추이가 약한 상관관계로 관찰되며, 화자독립 실험의 인식률 편차가 상대적으로 컸다. 이는 실험단어 간 변별력보다 추출된 특징벡터의 화자 간 변별력이 큰 것으로 판단되며, 각 화자들의 발화할 때의 입술 모양이 모두 다르기 때문으로 사료된다. 특히, “하이카카오톡(w7)”, “하이네비게이션(w8)”, “하이시스트란(w10)”에서 각 화자들이 발화할 때의 입술 모양이 다른 것을 관찰할 수 있었다. [Fig. 8]-(a)에서 나타난 것처럼 “하이카카오톡(w7)”에서 “카”를 발화할 때 치아가 보이는 화자와 그렇지 않은 화자가 관찰되었다. 한편 [Fig. 8]-(b)에서 나타난 것처럼 “하이네비게이션(w8)”과 “하이시스트란(w10)”에서 마지막 음소의 /ㄴ/을 발화할 때 혀가 보이는 화자와 그렇지 않은 화자가 관찰되었다.

<Table 4> Lip Reading Recognition Experiment Result

	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	Mean
Gray	16.67%	23.33%	30%	20%	26.67%	10%	16.67%	20%	26.67%	36.67%	22.67%
Sobel	10%	30%	20%	20%	26.67%	26.67%	10%	23.33%	16.67%	30%	21.33%
Optical Flow	6.7%	17.7%	30%	17.7%	20%	16.7%	16.7%	3%	10%	16%	15.45%
Coordinate	20%	46.67%	50%	23.33%	53.33%	63.33%	10%	40%	26.67%	43.33%	37.67%
CNN User dependent	87%	86%	89%	89%	89%	90%	92%	91%	86%	86%	88.5%
CNN User Independent	83%	79%	89%	83%	76%	77%	85%	78%	78%	74%	80.2%

/haigælʌksi/(w1), /haialladin/(w2), /haismɑ:rtpon/(w3), /haikamera/(w4), /haimesɪdʒ/(w5), /haikakaotok/(w6), /haiʒʌnhwagʌlgi/(w7), /hainævigeiʃn/(w8), /haiimeil/(w9), /haisistlan/(w10)



/haigælʌksi/(w1), /haialladin/(w2), /haismɑ:rtpon/(w3), /haikamera/(w4), /haimesɪdʒ/(w5), /haikakaotok/(w6), /haiʒʌnhwagʌlgi/(w7), /hainævigeiʃn/(w8), /haiimeil/(w9), /haisistlan/(w10)

[Fig. 7] Lip Reading Recognition Experiment Result

(a) (b)

[Fig. 8] Uttered characteristic of speaker

그러나 “하이스마트폰(w3)”의 마지막 음소의 /ㄴ/을 발화할 때에는 모음 /ㅗ/의 특성으로 모든 화자들에게 혀가 관찰되지 않기 때문에 화자 종속 실험과 화자 독립 실험의 결과가 같은 것으로 판단된다.

화자 간 변별력에 대해 영향을 받지 않는 화자 종속 실험 결과에서 가장 낮은 인식률을 보인 실험단어들의 공통적인 특징은 같은 모음이 연속적으로 나타난다는 것으로 관찰되었다. “하이알라딘(w2)”과 “하이이메일(w9)”은 각각 모음 /ㅏ/와 /ㅣ/가 연속적으로 관찰되고, “하이시스트란(w10)”은 모음 /ㅣ/와 /ㅡ/가 연속적으로 관찰된다. 반면, 인식률이 높은 실험단어 “하이전화걸기(w7)”과 “하이네비게이션(w8)”에서는 입모양의 변화가 상대적으로 매우 동적이다. 이것은 입 모양 특징만을 사용한 인식의 한계로 추후 음성과 결합하였을 때 해결될 수 있을 것으로 판단된다.

또한, 각 특징을 비교한 이전 연구결과에서는 입술소리 /ㅁ/, /ㅂ/, /ㅍ/가 포함된 특정 실험 단어의 인식률이 높게 나타난 반면, 본 연구에서는 입술 소리가 포함된 특정 실험 단어의 인식 성능이 특별히 높은 인식률을 나타내지 않았다. 이는 입술 소리를 비발화구간으로 인식한 것으로 인해 입술 소리의 입 모양 특징이 유실되었기 때문으로 판단된다. 그러나 본 연구에서 제안한 방법을 사용하였을 때 이전 연구에서의 실험 단어 간 인식률 편차가 눈에 띄게 낮아졌다. 특히 이전 연구에서 평균 인식률이 가장 높았던 좌표 기반 특징벡터로 인식 실험 한 실험 단어 중 가장 낮은 성능을 보인 “하이전화걸기(w7)”의 인식 성능이 본 연구에서 제안한 방법을 이용한 인식 실험에서 압도적으로 높아진 것을 확인 할 수 있었다. 따라서 발화구간 검출을 위해 학습된 CNN을 이용하여 추출된 특징벡터가 입 모양의 특징을 잘 반영된 것으로 판단된다.

본 논문에서는 입 모양 인식을 위해서 CNN 기반의 발화구간 검출을 하는 실험 과정을 통하여 하나의 문제를

위해 학습된 CNN을 종속관계에 있는 다른 문제에도 적용시킬 수 있다는 시사점을 발견하게 되었다.

5. 결론

AVSR 시스템에서 Lip Reading System의 성능은 전체 시스템의 성능을 결정짓는 핵심적인 요소이다. 본 연구는 발화구간 검출과 입 모양 인식과 같은 유사하지만 종속적인 문제에서 동일한 CNN 구성을 이용할 수 있는 가능성도 보였다. 그러나 화자종속 실험과 화자 독립실험의 실험 결과의 인식률 추이가 상이한 것으로 관찰되며, 화자독립 실험의 인식률 편차가 상대적으로 컸다. 이는 추출된 특징 벡터의 클래스 간 변별력이 화자 간 변별력보다 높은 것으로 판단된다. 따라서 각 화자의 발화 시 나타나는 입술 모양의 편차를 극복할 수 있는 추가 연구가 필요하다.

추가적으로, Alext Krizhevsky가 제안한 네트워크 구성은 비행기, 자동차, 사람, 동물과 같은 전혀 다른 특징을 지닌 영상을 분류하는 네트워크 구성이다. 그러나 입 모양 인식 문제는 앞서 기술한 것과 같이 입술 모양의 개인간 편차가 심하며, 이를 극복할 수 있는 네트워크 구성이 가능한지에 대해 논의가 필요하다.

마지막으로 이 연구를 통해 소음환경에서의 음성인식 성능 향상에 충분히 도움이 될 수 있을 것이라는 예측을 할 수 있었고, 앞서 기술한 문제를 해결하고, 추가적인 연구가 진행된다면 좀 더 강인한 AVSR 시스템을 구축할 것으로 기대한다.

REFERENCES

[1] Y. K. Kim, J. G. Lim, and M. H. Kim, "Feature Generations Analysis of Lip Image Streams for Isolate Words Recognition." International Journal of Multimedia and Ubiquitous Engineering Vol. 10, No. 10, pp. 337-346, 2015.

[2] Luettin, Juergen, and Neil A. Thacker. "Speechreading using probabilistic models." Computer Vision and Image Understanding, Vol. 65, No. 2, pp. 163-178, 1997.

[3] E. K. Kim, Y. D. Kwon, and J. S. Lee. "Neural Network Vowel-Recongition Jointly Using Voice Features and Mouth Shape Image". Korean Institute of Information Scientists and Engineers Congress 1996, Vol. 23 No. 2A, pp. 693-696, 1996.

[4] J. S., Lee, and C. H. Park, "Automatic Lipreading Using Color Lip Images and Principal Component Analysis" Journal of Information Processing Systems B, Vol.15, No.3 pp. 229-236, 2008.

[5] Shaikh, A. A., Kumar, D. K., Yau, W. C., Azemin, M. C., and Gubbi, J, "Lip reading using optical flow and support vector machines." Image and Signal Processing (CISP), 2010 3rd International Congress on. Vol. 1, 2010.

[6] Shaikh, Ayaz A., Dinesh K. Kumar, and Jayavardhana Gubbi. "Automatic visual speech segmentation and recognition using directional motion history images and Zernike moments." The Visual Computer Vol.29, No.10, pp.969-982, 2010.

[7] Lan, Y., Theobald, B. J., Harvey, R., Ong, E. J., and Bowden, R, "Improving visual features for lip-reading.", In AVSP 2010, International Conference on Audio-Visual Speech Processing, pp. 7-3, 2010.

[8] Kim Y. K., Lim J. G., and Kim M. H., "Lip Reading Algorithm Using Bool Matrix and SVM", International Conference on Small & Medium Business, (in Korean), (2015), pp.267 - 268.

[9] Sujatha, B., and T. Santhanam. "A novel approach integrating geometric and Gabor wavelet approaches to improvise visual lip-reading." Int. J. Soft Comput 5, pp.13-18, 2010.

[10] Ibrahim, M. Z., and D. J. Mulvaney. "Robust geometrical-based lip-reading using Hidden Markov models." EUROCON, 2013 IEEE, pp.2011-2016, 2013.

[11] Werda, Salah, Walid Mahdi, and Abdelmajid Ben Hamadou. "Lip localization and viseme classification for visual speech recognition." arXiv preprint arXiv:1301.4558, Vol.5, No. 1, pp. 62-75 2013.

[12] Wang, S. L., Lau, W. H., Leung, S. H. and Yan, H. "A real-time automatic lipreading system." Circuits

and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on. Vol. 2, 2004.

[13] Cetingul, H. E., Yemez, Y., Erzin, E. and Tekalp, A. M, "Discriminative analysis of lip motion features for speaker identification and speech-reading." Image Processing, IEEE Transactions on. Vol.15, No.10, pp. 2879-2891, 2006.

[14] Siatras, S., Nikolaidis, N., Krinidis, M., and Pitas, I., "Visual lip activity detection and speaker detection using mouth region intensities." Circuits and Systems for Video Technology, IEEE Transactions on Vol.19, No.1, pp.133-137, 2009.

[15] Arsic, Aleksandra, Milos Jordanski, and Milan Tuba. "Improved lip detection algorithm based on region segmentation and edge detection." Telecommunications Forum Telfor (TELFOR), 2015 23rd. IEEE, 2015.

[16] G. B. Kim, J. W. Ryu, and N. I. Cho, "Voice Activity Detection using Motion and Variation of Intensity in The Mouth Region", Journal of Broadcast Engineering, Vol.17, No.3, pp.519-528, 2012.

[17] E. K. Kim, "Speech Activity Detection using Lip Movement Image Signals", Journal of the Institute of Signal Processing and Systems, Vol.11, No.4, pp.289-297, 2010.

[18] J. S. Kim, J. G. Nam , and B. T. Zhang, "Deep Learning-based Video Analysis Techniques" Journal of Korean Institute Information Scientists Engineers, Vol.33, No.9, pp.21-31, 2015.

[19] Yun-A Hur, Keun-Ho Lee, "A Study on Countermeasures of Convergence for Big Data and Security Threats to Attack DRDoS in U-Healthcare Device", Journal of the Korea Convergence Society, Vol. 6, No. 4, pp. 243-248, 2015.

[20] G. J. Jang and J. S. Park, "Visual Object Recognition Based on Deep Neural Networks Implemented by CAFFE". Journal of Korean Institute Information Scientists Engineers, Vol.33, No.8, pp.49-54, 2015.

[21] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. pp1097-1105, 2012.

[22] Viola, Paul, and Michael Jones, "Rapid object detection using a boosted cascade of simple features." Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Vol. 1, 2001.

[23] S.I. Chien and Il Choi, "Face and Facial Landmarks Location Based on Log-Polar Mapping," Lecture Notes in Computer Science, pp. 379-386, 2000.

[24] Lim, Jong Gwan, Jaehong Kim, and Dong-Soo Kwon. "Multidimensional evaluation and analysis of motion segmentation for inertial measurement unit applications." Multimedia Tools and Applications, pp.1-28, 2015.

[25] Lim, Jong Gwan, Mi-hye Kim, and Sahngwoon Lee. "Empirical Validation of Objective Functions in Feature Selection Based on Acceleration Motion Segmentation Data." Mathematical Problems in Engineering, 2015.

[26] Krizhevsky, Alex, and G. Hinton. "Convolutional deep belief networks on cifar-10." Unpublished manuscript, 2010.

[27] Maini, Raman, and Himanshu Aggarwal. "Study and comparison of various image edge detection techniques." International journal of image processing (IJIP) Vol.3, No.1, pp.1-11, 2009.

[28] Jun-Yeon Lee, "Forecasting the Time-Series Data Converged on Time PLOT and Moving Average", Journal of the Korea Convergence Society, Vol. 6, No. 4, pp. 161-167, 2015.

김 용 기(Kim, Yong Ki)

- 2008년 2월 : 청주대학교 관광경영학과(경영사)
- 2014년 2월 : 충북대학교 컴퓨터공학과(공학 석사)
- 2015년 3월 ~ 현재 : 충북대학교 컴퓨터공학과 박사 재학
- 관심분야 : Image Processing, AVSR, Lip Reading, Machine Learning, Pattern Recognition
- E-Mail : moodeath.kyk@gmail.com

임 종 관(Lim, Jong Gwan)

・2002년 6월 : 충북대학교 정보통신공학과(공학사)
・2006년 2월 : 한국과학기술원 바이오 시스템 학과(공학 석사)
・2015년 8월 한f국과학기술원 기계공학과 (공학 박사)
・2015년 9월 ~ 현재 : University of Trento 박사후 과정
・관심분야 : Machine Learning, Pattern Recognition, HRI, Sensor Fusion, Signal Processing
・E-Mail : jonggwanlim@gmail.com

김 미 혜(Kim, Mi Hye)

・1992년 2월 : 충북대학교 수학과(이학사)
・1994년 2월 : 충북대학교 수학과(이학석사)
・2001년 2월 : 충북대학교 수학과(이학박사)
・2004년 9월 ~ 현재 : 충북대학교 컴퓨터공학과 교수
・관심분야 : 기능성 게임, 유비쿼터스 게임, 플랫폼, 퍼지측도 및 퍼지적분, 제스츄어 인식
・E-Mail : mhkim@chnu.ac.kr