The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.4, pp.189~198
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.4.189

Region of Interest Extraction and Bilinear Interpolation Application for Preprocessing of Lipreading Systems

Jae Hyeok Han† · Yong Ki Kim†† · Mi Hye Kim†††

ABSTRACT

Lipreading is one of the important parts of speech recognition, and several studies have been conducted to improve the performance of lipreading in lipreading systems for speech recognition. Recent studies have used method to modify the model architecture of lipreading system to improve recognition performance. Unlike previous research that improve recognition performance by modifying model architecture, we aim to improve recognition performance without any change in model architecture. In order to improve the recognition performance without modifying the model architecture, we refer to the cues used in human lipreading and set other regions such as chin and cheeks as regions of interest along with the lip region, which is the existing region of interest of lipreading systems, and compare the recognition rate of each region of interest to propose the highest performing region of interest In addition, assuming that the difference in normalization results caused by the difference in interpolation method during the process of normalizing the size of the region of interest affects the recognition performance, we interpolate the same region of interest using nearest neighbor interpolation, bilinear interpolation, and bicubic interpolation, and compare the recognition rate of each interpolation method to propose the best performing interpolation method. Each region of interest was detected by training an object detection neural network, and dynamic time warping templates were generated by normalizing each region of interest, extracting and combining features, and mapping the dimensionality reduction of the combined features into a low-dimensional space. The recognition rate was evaluated by comparing the distance between the generated dynamic time warping templates and the data mapped to the low-dimensional space. In the comparison of regions of interest, the result of the region of interest containing only the lip region showed an average recognition rate of 97.36%, which is 3.44% higher than the average recognition rate of 93.92% in the previous study, and in the comparison of interpolation methods, the bilinear interpolation method performed 97.36%, which is 14.65% higher than the nearest neighbor interpolation method and 5.55% higher than the bicubic interpolation method. The code used in this study can be found a https://github.com/haraisi2/Lipreading-Systems.

Keywords : Artificial Neural Network, Region of Interest, Interpolation, Object Detection, Lipreading

입 모양 인식 시스템 전처리를 위한 관심 영역 추출과 이중 선형 보간법 적용

한 재 혁† · 김 용 기†† · 김 미 혜†††

요 약

입 모양 인식은 음성 인식의 중요 부분 중 하나로 음성 인식을 위한 입 모양 인식 시스템에서 입 모양 인식 성능을 개선하기 위한 여러 연구가 진행됐다. 최근의 연구에서는 인식 성능을 개선하기 위해 입 모양 인식 시스템의 모델 구조를 수정하는 방법이 사용됐다. 본 연구에서는 모델 구조를 수정하는 것으로 인식 성능을 개선하는 기존의 연구와 달리 모델 구조의 변화 없이 인식 성능을 개선하는 것을 목표로 한다. 모델 구조의 수정 없이 인식 성능을 개선하기 위해, 사람이 하는 입 모양 인식에서 사용되는 단서를 참고해 입 모양 인식 시스템의 기존 관심 영역인 입술 영역과 함께 턱, 뺨과 같은 다른 영역을 관심 영역으로 설정하고 각 관심 영역의 인식률을 비교해 가장 높은 성능의 관심 영역을 제안한다. 또한, 관심 영역 크기를 정규화하는 과정에서 보간법의 차이로 인해 발생하는 정규화 결과의 차이가 인식 성능에 영향을 준다고 가정하고 최근접 이웃 보간법, 이중 선형 보간법, 이중 삼차 보간법을 사용해 동일한 관심 영역을 보간하고 각 보간법에 따른 입 모양 인식률을 비교해 가장 높은 성능의 보간법을 제안한다. 각 관심 영역은 객체 탐지 인공신경망을 학습시켜 검출하고, 각 관심 영역을 정규화하고 특징을 추출하고 결합한 뒤, 결합된 특징들을 차원 축소한 결과를 저차원 공간으로 매핑하는 것으로 동적 정합 템플릿을 생성했다. 생성된 동적 정합 템플릿들과 저차원 공간으로 매핑된 데이터의 거리를 비교하는 것으로 인식률을 평가했다. 실험 결과 관심 영역의 비교에서는 입술 영역만을 포함하는 관심 영역의 결과가 이전 연구의 93.92%의 평균 인식률보다 3.44% 높은 97.36%의 평균 인식률을 보였으며, 보간법의 비교에서는 이중 선형 보간법이 97.36%로 최근접 이웃 보간법에 비해 14.65%, 이중 삼차 보간법에 비해 5.55% 높은 성능을 나타내었다. 본 연구에 사용된 코드는 https://github.com/haraisi2/Lipreading-Systems에서 확인할 수 있다.

키워드 : 인공신경망, 관심 영역, 보간법, 객체 탐지, 입 모양 인식

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업임(IITP-2024-2020-0-01462).
※ 이 논문은 2023년 ACK 2023의 우수논문으로 "음성인식 시스템의 입 모양 인식개선을 위한 관심영역 추출 방법"의 제목으로 발표된 논문을 확장한 것임.
† 정 회 원 : 충북대학교 컴퓨터공학과 박사과정
†† 정 회 원 : 충북대학교 연구원
††† 정 회 원 : 충북대학교 컴퓨터공학과 교수
Manuscript Received : February 5, 2024
First Revision : March 7, 2024
Accepted : March 8, 2024
* Corresponding Author : Mi Hye Kim(mhkim@cbnu.ac.kr)

1. 서 론

최근, 모바일, 웨어러블 기기, 스마트 홈 시스템, 차량용 인터페이스 등 여러 분야에서 음성 인식 기술이 상용화되어 다양하게 사용되고 있다. 이렇게 사용되는 음성 인식 시스템에서의 인식률을 높이기 위해서 특징 추출, 모델 보정 및 분류 알고리즘과 같이 원리 분야에서 여러 방법이 연구됐다[1]. 또한, 소음이 심한 환경과 같이 음성만으로 음성 인식이 힘든 상황을 극복하기 위해 음성 인식과 입 모양 인식을 결합하는 오디오 비주얼 음성 인식(Audio-Visual Speech Recognition) 시스템이 제안되었고, 소음 환경에서의 음성 인식 시스템의 성능을 향상할 수 있다는 연구결과가 나타났다[2].

입 모양 인식을 실험하는 Meier et al.의 연구[3], Kim의 연구[4] 등에서는 모두 실험에서의 관심 영역을 입술 주변 영역 한 가지로만 설정하고 실험을 수행하였다.

사람의 입 모양 인식의 경우 입의 모양을 가지고 상대방의 발화를 인지하는 점은 음성 인식 시스템의 입 모양 인식과 같지만, 입술만이 아닌 표정, 턱, 혀의 움직임에도 의존하며[5], 뺨의 들썩거림은 m-p-b 발음을 구별할 때 사용할 수 있는 시각적 단서이기도 하다[6]. 이처럼 사람의 입 모양 인식에 있어서 입술 주변 영역이 아닌 다른 영역에서의 변화들도 중요한 정보로 사용된다.

음성 인식 시스템에서 입 모양을 인식하기 위해서 관심 영역을 검출한 뒤, 검출된 관심 영역에 여러 전처리를 한 뒤 전처리된 데이터를 사용한다. Kim의 연구[4]에서는 음성 인식 시스템의 입 모양 인식에서 격자를 적용한 입술 영역 안의 영상 정보 특징을 검출하기 위해 관심 영역을 검출한 뒤 검출된 관심 영역의 크기를 정규화한 뒤 회색조 이미지로 변경, 히스토그램 평활화를 수행하여 실험을 수행하였다.

관심 영역의 크기를 정규화할 때 사용되는 보간법에 따라 정규화된 관심 영역의 화소값에 차이가 발생한다. 입 모양 인식에서 관심 영역의 화소값을 특징으로 사용했을 때, 다른 보간법으로 전처리한 관심 영역의 차이가 결과에 영향을 미칠 가능성이 존재한다.

2020년 Martinez et al.의 연구[7], 2021년 Ma et al의 연구[8], 2023년 Vatadande et al의 연구[9] 등의 경우 입 모양 인식 시스템의 인식률을 높이기 위해서 인식에 사용되는 신경망의 구조를 변경하는 방법을 사용했다.

입 모양 인식 시스템의 구조적 변화 없이 인식률을 높일 수 있다면 기존에 존재하는 인식 시스템들의 인식률을 추가적으로 높일 수 있을 것이다. 따라서 본 논문에서는 입 모양 인식 시스템에서 시스템의 구조적 변화 없이 인식률을 개선하기 위해 기존의 입술 영역을 사람의 입 모양 인식에서 사용되는 변인들이 포함된 다른 관심 영역과 비교하여 어떤 관심 영역이 입 모양 인식 시스템에서 가장 높은 성능의 영역인지 제안하고자 한다. 사람이 하는 입 모양 인식의 변인들을 참고하여 기존의 관심 영역으로 사용되는 입술 영역과 함께, 입술과 턱이 포함된 영역, 입술과 턱 그리고 뺨이 포함된 얼굴의 하관 영역, 이렇게 3가지 영역을 관심 영역으로 사용했다. 그리고 여러 개의 관심 영역을 검출하기 위해 객체 탐지 인공신경망을 사용해 관심 영역들을 학습시켜서 관심 영역을 자동으로 검출하는 방법을 제안한다.

또한, 관심 영역의 크기를 정규화하는 전처리 과정에서 사용되는 보간법이 인식 실험 결과에 영향을 미친다고 가정하고 어떠한 보간법이 가장 높은 성능의 보간법인지 제안하고자 한다. 2차원 값의 보간에 사용되는 3가지 보간법인 최근접 이웃 보간법(nearest neighbor interpolation), 이중 선형 보간법(bilinear interpolation), 이중 삼차 보간법(bicubic interpolation)을 실험에 사용할 보간법으로 선정했다.

본 논문에서 제안된 관심 영역의 인식 실험을 수행하는 과정은 다음과 같다. 첫째, 객체 탐지 인공신경망을 앞에서 제안된 3가지 관심 영역에 대해 학습시키고, 학습된 인공신경망을 사용해 원본 데이터에서 3가지 관심 영역을 탐지한다. 둘째, 탐지된 관심 영역들을 검출하고 회색조 이미지로 변경한 뒤 이중 선형 보간법을 사용해 크기를 정규화한 3가지 집합을 생성, 각 집합에 대해 히스토그램 평활화 과정을 수행한다. 추가로 보간법 간 비교를 위해 첫 번째 관심 영역인 입술 영역에 최근접 이웃 보간법과 이중 삼차 보간법으로 크기를 정규화한 집합을 각각 생성한 뒤 앞과 동일하게 각 집합에 대해 히스토그램 평활화 과정을 수행한다. 셋째, 생성된 각 전처리된 관심 영역 집합으로 격자 기반 그레이레벨 변화량 특징을 생성하고 원본 영상에서 입술 모양을 근사화한 특징을 생성한 뒤, 각 집합마다 그레이레벨 변화량 특징과 입술 모양을 근사화한 특징을 결합하고 결합한 특징들을 차원 축소한다. 넷째, 차원 축소된 각 집합의 특징을 동적 정합 템플릿을 생성해 평가하여 화자 독립 입 모양 인식에서 3가지 관심 영역 중 가장 높은 성능을 가진 관심 영역을 검출하고, 3가지 보간법 중 가장 높은 성능을 가진 보관법을 검출한다.

2. 관련 연구

2.1 입 모양 인식

입 모양 인식(독순술, lipreading)은 입술, 얼굴, 혀의 움직임을 보고 대화 내용을 파악하는 기술로 청각장애인이 대화 중인 상대방이 말하고 있는 내용을 파악할 때 사용한다. 또한, 오디오 비주얼 음성 인식 시스템(Audio-Visual Speech Recognition System)과 같이 사람이 발화한 음성을 입 모양으로 인식하는 시스템에서 사용하는 방법이다.

청각장애인이 사용하는 입 모양 인식의 경우 오래전부터 제한된 범위에서 입의 모양만으로 발화를 이해할 수 있는 교육을 시행하고 있다[10].

입 모양 인식 시스템 연구들의 경우, 기존의 사람이 하는 입 모양 인식에서 입술이 움직이는 모양을 보고 대화 내용을 알아내는 것처럼 입 모양 인식에 사용하는 관심 영역을 주로

입술 영역 혹은 턱과 같은 입술 주변을 포함하는 범위로 지정하고 자동 입 모양 인식을 수행했다. 입 모양 인식 시스템의 전통적인 특징 추출 및 인식 방법은 입술 감지 및 추출, 특징 추출, 특징 변환, 분류(인식)의 단계로 진행된다[11]. 2019년 Kim[4]은 입술 영역에서 여러 특징을 추출하고 평가한 뒤 최적화된 특징을 선별해 결합한 뒤 결합한 특징들을 ISOMAP을 사용해 차원 축소한 뒤 동적 정합(Dynamic Time Warping)으로 템플릿을 생성하여 그 템플릿으로 차원 축소된 특징들을 평가하고 은닉 마르코프 모델을 사용해 자동 입 모양 인식을 수행하는 것으로 높은 인식률을 가지는 특징을 찾고 입 모양을 인식하는 방법을 제안했다. 2020년 Martinez et al[7]은 시간적 컨볼루션 네트워크(temporal convolutional nework)를 사용해 최근 연구에서 사용되는 양방향 게이트 순환 유닛(bidirectional gated recurrent unit) 레이어를 대체하는 것으로 입 모양 인식에 대한 학습을 단순화하고, 가변 길이 증강으로 데이터 세트의 과적합을 회피하는 것으로 인식률을 높이는 방법을 제안했다. 2021년 Ma et al[8]은 깊이 방식 분리 가능한 컨볼루션(depthwise separable convolution)으로 기존의 신경망 모델의 컨볼루션 구조와 교체하는 것으로 입 모양 인식률을 높이는 방법을 제안했다. 2023년 Vayadande et al[9]은 삼차원 컨볼루션 뉴럴 네트워크(3 dimention convolution neural network)와 양방향 장단기 메모리(bidirectional long short term memory)를 사용한 모델을 사용입 모양 인식에 사용하는 것으로 인식률을 높이는 방법을 제안했다.

사람의 입 모양 인식과 관련된 연구들에 따르면 사람이 하는 입 모양 인식에서는 입 모양 외에도 입 모양 인식 과정에 포함될 수 있는 여러 정보를 포함한다. 사람의 기본적인 조음 동작에는 턱 높이, 입술 모양, 혀 위치에 대한 동작이 있다[12]. 화자의 조음 운동에 의해 나타나는 시각 정보, 전언에 수반되는 화자의 안면 표정, 손짓과 몸짓, 제시 상황 그리고 전언에 포함된 통사구조와 활용어 맥락 등은 효과적인 독화 지도 방법의 모색에 중요하다[13]. 뺨의 들썩거림은 m-p-b 발음을 구별할 때 사용할 수 있는 시각적 단서가 될 수 있고, 사람이 입 모양을 인식할 때 관심 영역에 뺨과 턱까지 포함되는 것이 좋다는 연구 결과가 있다[6]. 또한 사람이 발화할 때 음소마다 특별한 안면부 및 구강부의 특징을 갖고 있다는 점[14] 등 여러 이전 연구 결과에서 일반적인 입 모양 인식에서 입술 영역 외의 다른 변인들이 추가적인 단서가 되고, 인식률에 영향을 미치는 것을 알 수 있다. 그리고 음성 인식 시스템의 입 모양 인식에서 영상의 해상도 확장에 따른 관심 영역 범위가 증가했을 때, 인식률의 증가를 확인할 수 있었다[15].

2.2 관심 영역 검출

YOLO(You Only Look Once)는 Redom 등이 발표한 객체 탐지를 위한 인공신경망 모델이다[16]. YOLO는 기존의 객체 탐지 모델에 근접한 정확도를 가지면서 상대적으로 고속으로 객체를 탐지하고자 개발된 모델로 객체 탐지할 때 이미지 전체를 단 한 번만 모델에 통과시키고, 단 하나의 통합된 인공신경망 모델을 사용하며, 실시간으로 객체를 탐지하는 특징이 있다. 세 번째 모델인 YOLOv3은 2018년에는 이전 모델에서 5가지 변경점을 통해 성능을 개선한 모델로 발표되었다[17]. YOLO는 현재 여덟 번째 모델인 v8이 출시되었다. YOLOv8은 이전 버전보다 성능이 향상되었고, 유연성 및 효율성을 위해 새로운 기능과 개선 사항이 도입되었다. YOLOv8은 실시간 객체 탐지와 이미지 분할, 포즈 추정, 추적 및 분류 등의 분야에 사용되고 있다[18].

2.3 관심 영역의 특징 생성 방법

관심 영역에서 특징을 생성하는 방법은 영상 정보 특징과 변화량 특징, 입 모양을 근사화한 특징으로 나눌 수 있다.

관심 영역에서 입술 영역 안의 영상 정보 특징은 입술과 혀, 치아 등의 조음 기관 전체 움직임을 반영하는 특징이다. 격자를 사용하는 관심 영역의 영상 정보 특징인 그레이레벨 특징은 크기가 같은 격자를 적용한 후 각 격자가 위치한 그레이레벨 화소값의 평균을 이용하여 생성한다. 격자의 평균값을 이용하는 방법은 특징 생성 모델에서 각각 생성된 특징점들의 차원 축소 효과와 더불어 영상 잡음에 의한 영향을 보상할 수 있다[4].

입술 모양을 근사화한 특징은 입술의 가로, 세로 길이로 구성된 특징이다. 여러 연구에서 입술의 좌표를 특징으로 생성하는 방법보다는 가로, 세로 길이인 기하학적 특징들이 더 좋은 성능을 나타내는 것으로 확인되었다[19-22]. 입술 모양을 근사화하는 특징은 입술 영역 내의 영상 정보에 기반한 특징들에 비해 회전, 위치, 크기 변환이 상대적으로 용이하기 때문에 세밀한 입술 영역 영상 추출이 필요 없다는 장점이 있으나 입 내부의 조음 기관인 혀나 치아의 특징이 유실된다는 문제가 있다[4].

2.4 결합된 특징의 차원 축소 방법

차원 축소란 고차원 데이터의 정보 손실을 최소화하면서 데이터의 차원을 저차원으로 바꾸는 것을 말한다. 고차원 데이터는 직관적인 통찰이 어렵고 차원의 저주가 발생하기 때문에 직접 사용하는 대신에 데이터를 차원 축소하는 방법들이 사용되고 있다.

ISOMAP은 데이터의 전체 구조를 반영한 거리 행렬을 이용하여 이 거리 행렬의 값을 최대한 유지하는 저차원 구조를 다차원 척도법으로부터 찾는 방법이다[23].

구조를 반영하는 저차원 구조를 찾는 데 목적이 있다. ISOMAP 알고리즘은 세 단계로 이루어진다. 첫 번째로 어떤 점들이 매니폴드 상에서 서로 가까이 위치하는지 측정해 인접한 이웃 그래프를 구축한다. 이때 이웃 점 간의 관계는 K-Neighborhood 알고리즘을 사용해 정의한다. 두 번째로 그래프상의 모든 가능한 두 점에 대해 두 점 간의 최단 경로 그래프를 계산한다. 세 번째로 모든 점에 대해 이웃 점들과 계산된

거리를 이용하여 거리 행렬을 구하고, 다차원 척도법을 사용하여 저차원 공간으로 사상한 데이터를 구한다.

2.5 동적 정합

동적 정합(Dynamic Time Warping)은 음성신호로부터 추출한 음성 특징 벡터 열을 대상으로 두 특징 벡터 열 사이의 유사도를 측정하는 방법이다. 같은 단어를 발성해도 길이가 다르므로 특징 벡터 열 사이의 비교해야 할 부분을 비선형적으로 결정한다는 의미에서 warping이라고 한다[24].

동적 정합의 알고리즘에서는 두 데이터의 각 성분에 대한 유사도(거리)를 비용으로 설정한다. 두 데이터가 이루는 격자상에서 두 데이터의 시작 시점에서 시작하여, 종료 시점에 이르기까지 비용 표에서 최소 비용으로 이동할 수 있는 최단 경로를 구한다. 최종적으로 비용이 최소인 경로의 비용 합이 동적 정합을 이용한 두 데이터 간 유사도 값이 된다.

Kim의 연구[4]에서는 동적 정합 템플릿 결정 방법의 인식률에 대한 특징 간 높고 낮음의 양상이 딥러닝을 사용하는 은닉 마르코프 모델과 동일한 것을 확인할 수 있다.

3. 입 모양 관심 영역 추출과 이중 선형 보간법 적용

본 연구에서는 입 모양 인식 시스템의 구조 변화 없이 인식률을 개선하기 위해 사람의 입 모양 인식을 참고하여 사람이 발화할 때 입술과 함께 변화하는 턱과 뺨 같은 다른 부위들을 관심 영역으로 포함하는 것이 사람이 하는 입 모양 인식에서 입술이 아닌 다른 변인들이 입 모양 인식의 인식률에 영향을 준 것처럼 입 모양 인식 시스템에도 마찬가지로 영향을 줄 것이라 가정한다. 기존의 입술 영역과 입술이 아닌 다른 변인들이 포함된 영역을 관심 영역으로 설정하고 관심 영역들의 인식률을 비교해 검증한다. 또한, 관심 영역의 전처리 과정에서 관심 영역들을 일정한 크기로 정규화할 때, 각 보간법에 따라 다르게 정규화된 관심 영역의 화소값이 인식률에 영향을 줄 것이라 가정하고 각 보간법으로 전처리된 데이터의 인식률을 비교해 검증한다. 이 검증을 위해 본 연구에서 발화할 때 변화하는 부위들인 입술, 턱, 뺨을 포함하는 3가지 관심 영역인 입술 영역, 입술과 턱을 포함한 영역, 입술과 턱, 뺨을 포함하는 영역을 설정한다. 그리고 객체 탐지 인공신경망 YOLOv3를 사용해 관심 영역들을 검출해 각 관심 영역들의 인식 성능을 비교하고 객체 탐지 인공신경망과 사람이 직접 검출한 관심 영역의 성능을 비교하기 위해 3가지 관심 영역과 사람이 직접 검출한 입술 관심 영역의 성능을 비교한다. 이후 3가지 관심 영역 중 가장 높은 인식률을 보인 관심 영역을 기준으로 영상 보간에 사용되는 3가지 보간법인 최근접 이웃 보간법, 이중 선형 보간법, 이중 삼차 보간법을 사용하여 크기를 정규화하고, 데이터들의 평가 방법으로는 [4]의 평가 방법인 동적 정합 템플릿을 사용한 평가 방법을 선택했다. 동적 정합 템플릿을 생성해 템플릿을 기준으로 데이터를 평가하는 인식 실험을 수행하여 보간법에 따른 인식률을 비교해 검증하고자 한다.

Fig. 1은 본 실험의 순서도이다. 본 실험에서는 16명의 화자가 각각 10개의 단어를 최소 10회에서 최대 16회까지 발화하는 영상 데이터를 사용한다. 데이터 중 일부를 학습 데이터로 사용해 YOLOv3 모델을 본 연구에서 제안한 관심 영역들에 대해 학습시킨 뒤, 학습된 신경망 모델을 사용해 실험에 사용할 영상 데이터의 각 프레임에 관해 관심 영역의 좌표를 검출한다. 그리고 검출된 관심 영역의 좌표를 사용하여 검출된 영역을 원본 영상 데이터에서 추출한다. 각 검출된 관심 영역들을 회색조 영상으로 변환한 뒤 크기를 정규화하기 위해 너비와 높이가 130 * 100화소 크기가 되도록 보간법을 사용해 조정한다. 관심 영역 간의 성능을 비교하기 위해 3가지 관심 영역에 대해서는 이중 선형 보간법을 사용해 관심 영역들의 크기를 정규화한다. 정규화된 관심 영역들에 히스토그램 평활화를 수행한다. 발화 구간은 영상의 음성 데이터를 기준으로 검출한다. 원본 영상 데이터에서 입술 모양을 근사화한 특징을 생성하고, 각 관심 영역에 대해서 격자 기반 그레이레벨 변화량 특징을 생성한다. 각 관심 영역별로 생성된 두 결합된 특징의 시계열 데이터를 ISOMAP을 사용하여 각 관심 영역의 특징 집합 차원을 비선형 차원 축소한다.

이후 차원 축소된 데이터를 평가하기 위해 시계열 데이터 간 유사도를 계산하는 동적 정합(Dynamic Time Warping)을 사용하여 평가를 위한 템플릿을 결정한다. 결정된 템플릿을 기준으로 차원 축소된 특징의 단어별 인식률을 평가한다. 그 중 가장 높은 성능을 보인 관심 영역을 기준으로 하여 최근접 이웃 보간법, 이중 선형 보간법, 이중 삼차 보간법을 사용해 크기를 정규화한다. 3가지 보간법으로 정규화된 각 집합을 앞의 과정으로 다시 반복하여 각 보간법 간의 인식 결과를 평가하고 평가 결과를 비교한다.

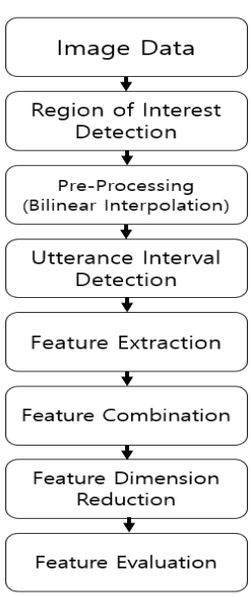


Fig. 1. Experimental Flowchart

입 모양 인식을 위한 관심 영역을 검출하는 방법으로는 영상을 촬영할 때 관심 영역만을 촬영하는 방법[3], 사람이 직접 검출하는 방법[4], 객체 탐지 인공신경망 모델로 검출하는 방법[21] 등이 있다. 사람이 직접 관심 영역을 검출하는 방법은 시간, 노동력, 관심 영역 데이터의 질을 고려했을 때 큰 비용이 든다. 본 연구에서는 앞의 요소를 고려하여 객체 탐지 인공신경망 모델인 YOLOv3를 관심 영역에 대해 학습시킨 뒤 사용하여 원본 영상 데이터에서 3가지 관심 영역을 검출한다. 인공신경망을 학습시키기 위해 원본 영상 데이터의 20%를 무작위로 선택한다. 선택된 영상 데이터에 3가지 관심 영역의 범위를 지정한다. 관심 영역의 범위를 지정한 값과 관심 영역의 범위가 지정된 원본 영상 데이터를 인공신경망의 학습 데이터로 사용한다. 신경망의 학습은 과적합을 막기 위해 학습 중 인식률의 변화가 최소화되는 시점까지 진행한다. 학습이 종료된 YOLOv3 인공신경망 모델을 사용해 전체 원본 영상 데이터에서 관심 영역의 범위의 좌표값을 검출한다. 검출된 좌표값을 사용하여 원본 영상 데이터에서 각 관심 영역을 검출해 .jpg 확장자 파일을 생성했다.

원본 영상 데이터에서 화자가 발화하는 동안 입술과 주변 영역들이 움직이는 것으로 관심 영역의 크기가 변화한다. 크기가 서로 다른 관심 영역의 크기를 보간법을 사용해 130 * 100화소의 크기로 조정한다. 보간법(interpolation)은 알려진 값을 기반으로 원하는 위치에서의 값을 추정하는 방법이다. 음성 인식 시스템의 입 모양 인식에서 화자가 발화하는 동안 입술의 움직임에 따라 검출되는 관심 영역의 크기가 변화하기 때문에 검출된 관심 영역의 크기는 일정하지 않다. 검출된 관심 영역들의 크기를 정규화하기 위해 검출된 관심 영역 영상을 보간법을 사용해 일정한 크기로 보간해야 한다.

이전 Kim의 연구[4]와 Koumparoulis et al의 연구[25]에서 입 모양 인식 실험을 위해 각 관심 영역의 크기를 일정한 크기로 정규화하는 과정이 존재하지만 사용한 보간법을 선택한 이유가 설명되지 않았다. 본 연구에서는 입 모양 인식 실험에서 가장 높은 성능을 가진 보간법을 검증하기 위해 영상 데이터의 보간에 사용되는 최근접 이웃 보간법(nearest neighbor interpolation), 이중 선형 보간법(쌍선형 보간법, bilinear interpolation), 이중 삼차 보간법(쌍삼차 보간법, bicubic interpolation)을 사용해 관심 영역을 보간하고 보간법에 따른 입 모양 인식 실험 결과를 비교한다.

최근접 이웃 보간법은 가장 간단한 보간법으로 영상 크기를 조정할 때, 변환되는 화소의 값을 원본 데이터에서 가장 가까운 위치에 있는 화소의 값을 참조하여 할당하는 방법이다. 단순히 이웃 화소를 복사하여 사용하는 것으로 처리 속도가 빠르다는 장점이 있지만, 원본 영상 내에서만 새로운 화소 값을 찾기 때문에 원래의 영상과 전혀 다른 영상을 출력하는 오류가 발생할 수 있다는 단점이 있다.

이중 선형 보간법은 선형 보간법을 2차원에 적용한 보간법으로 화소당 선형 보간을 세 번 수행하며, 변환되는 화소를 가장 가까운 화소 4개에 가중치를 곱한 값을 합해서 얻는 방법으로 최근접 이웃 보간법보다 더 부드러운 영상을 출력할 수 있다. 선형 보간법은 원본 영상의 두 개의 화소를 지나는 일차 함수를 이용하여 원하는 좌표에서 새로운 화소 값을 계산하는 보간법으로 선형 보간법의 임의의 지점 x에 대한 화소값은 Equation (1)을 이용하여 구한다.

$$f(x) = \frac{d_2}{d_{1+}\ d_2} f(x_1) + \frac{d_1}{d_{1+}\ d_2} f(x_2) \quad (1)$$

Equation (1)에서 x_1, x_2는 참조하는 두 지점이고, d_1, d_2는 x_1, x_2와 x 사이의 거리다.

이중 선형 보간법으로 임의의 지점 P의 화소값을 구하는 방법은 Equation (2)와 같이 가장 가까운 화소 A, B, C, D를 기준으로 y축 값이 같은 A, B값 사이에서 P와 x의 값이 같은 점 J의 좌표와 화소값을 구한 뒤 I, J간의 선형 보간법을 통해 P의 화소값을 구한다.

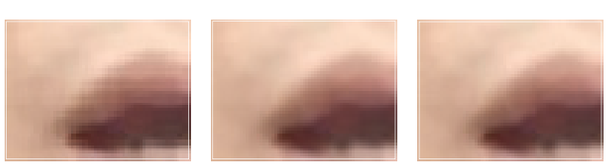
$$\frac{1}{(d_1+d_2)(d_3+d_4)} \begin{Bmatrix} (y_1-y)(x_1-x)f(A)+(y_1-y)(x-x_0)f(B) \\ +\ (y-y_0)(x_1-x)f(C)+(y-y_0)(x-x_0)f(D) \end{Bmatrix} \quad (2)$$

여기서 각 화소의 좌표는 $A(x_0, y_0)$, $B(x_1, y_0)$, $C(x_0, y_1)$, $D(x_1, y_1)$이고, d_1, d_2는 I와 A, B 사이의 거리, J와 C, D 사이의 거리이고, d_3, d_4는 P와 I, J 사이의 거리이다.

이중 삼차 보간법은 삼차 보간법을 2차원에 적용한 보간법으로 원본 영상의 인접한 16개의 화소값과 거리에 따른 삼차 방정식으로 표현된 가중치의 곱을 사용해서 화소값을 결정하는 방법으로 최근접 이웃 보간법의 계단 현상을 해결할 수 있다는 장점이 있지만, 삼차 방정식을 사용하기 때문에 속도가 상대적으로 다른 보간법에 비해 느려진다는 단점이 있다.

Fig. 2는 각 보간법으로 보간된 입술 영역이다.

3가지 관심 영역에 대해서는 이중 선형 보간법을 적용하여 크기를 정규화하고, 가장 높은 인식률을 보인 관심 영역에 대해 나머지 두 보간법인 최근접 이웃 보간법, 이중 삼차 보간법을 적용해 추가로 크기를 정규화한다.



(2A) Nearest neighbor interpolation　(2B) Bilinear interpolation　(2C) Bicubic interpolation

Fig. 2. Region of Lip Interpolated by Each Interpolation

인식기의 인식률을 높이기 위해서는 효과적인 특징을 선택하는 것이 중요하다. 본 연구에서는 Kim의 연구[4]에서 가장 높은 인식률을 보인 특징인 입술의 모양을 근사화한 특징과 그레이레벨 변화량 특징을 결합한 특징을 본 연구에서 사용할 특징으로 선택했다.

입술 모양을 근사화한 특징은 입술의 상하좌우의 좌표를 구한 뒤 바깥 입술의 가로, 세로 길이와 안쪽 입술의 가로, 세로 길이를 계산하여 기하학적인 특징으로 생성한다. 본 연구에서는 Openpose를 사용해 입술의 좌표를 검출하고, 검출된 입술 바깥쪽과 안쪽의 좌표 간 거리를 계산해 입술의 모양을 근사화한 특징을 생성했다.

그레이레벨 변화량 특징은 그레이레벨 특징의 시간 흐름에 따른 변화량을 특징으로 한 특징이다. 각 관심 영역에 크기가 같은 격자를 적용한 뒤, 각 격자 안에 있는 그레이레벨 화소값의 평균을 이용하여 격자 기반 그레이레벨 특징을 생성하고 생성된 특징의 변화량들로 생성한다. 본 연구에서는 각 관심 영역의 데이터를 입력으로 사용했다. 각 관심 영역을 회색조 이미지로 변경한 뒤, 크기 정규화, 히스토그램 평활화를 수행한 뒤 13 * 10개의 격자를 생성해 격자 기반 그레이레벨 특징을 생성했다. 생성된 그레이레벨 특징의 현재 프레임과 이전 프레임의 차이값으로 각 관심 영역별로 그레이레벨 변화량 특징을 생성했다. Table 1은 격자 기반 그레이레벨 특징을 생성하는 과정을 나타낸 의사코드이다.

본 연구에서 선택한 두 가지 특징 중 입술 모양을 근사화한 특징은 입술의 움직임을 나타내는 데 유리하지만, 입술의 가로, 세로 길이만이 사용되기 때문에 입술 내부와 외부의 특징은 유실된다. 격자 기반 특징의 경우 입술 내부와 외부의 특징은 유지되지만, 회전, 위치, 크기의 변환이 불리하다. 입 모양 인식에서 각각의 특징들이 가지고 있는 한계를 극복하기 위해 본 연구에서는 입술의 모양을 근사화한 특징과 각 관심 영역들의 그레이레벨 변화량 특징을 결합한 특징 집합을 생성한다.

Table 1. Grid-based Gray-level Difference Features Generation Pseudocode

```
i_datas = []
datas = []

Input data to i_datas

for each element in i_datas.
        temp = Gray levelize element
        temp = Normalize temp to 130 * 100pixel
        temp = Histogram smooth temp
        for x = 1 to 10
                for y = 1 to 10
                Create grid[y, x] in temp.
                Add grid[y, x] to v
                end for
        end for
        Add v to datas
end for

diff_features = []
prev_data;

for each element in datas.
        if current element is first element.
                Add zero-vector to diff_features.
        else
                Add element-prev_data to diff_features
        end if
        prev_data = element
end for
```

그레이레벨 변화량 특징은 이전 프레임과 비교해 변화량을 나타내는 특징이므로 다른 특징에 비해 1프레임이 부족하므로 특징을 결합하기 전 첫 번째 프레임에 해당하는 데이터를 영벡터로 보간한다. 선택된 특징 데이터들을 서로 결합하고 관심 영역별로 특징 결합을 구성한다.

결합된 특징 집합이 생성되면 차원의 저주 문제를 해결하기 위해 결합된 특징 집합에 대해 비선형 차원 축소 방법인 ISOMAP을 적용해 차원을 축소한다.

본 연구에서는 최종적으로 생성된 시계열 데이터인 각 관심 영역별 결합 및 차원 축소된 특징 집합들을 평가하기 위해 동적 정합 템플릿을 결정하고 결정된 템플릿과 특징집합들을 비교하는 방법을 사용한다.

적절한 템플릿을 결정하기 위해 동적 정합에 의해 도출된 거리 정보를 다차원 척도법(multidimensional scaling)으로 저차원 공간으로 매핑(mapping)한 뒤 각 클래스의 중심을 구한 후 중심과 가장 가까운 데이터를 템플릿으로 결정하는 방법을 사용한다.

동적 정합 템플릿을 결정하는 방법은 시계열 데이터 집합 W의 동적 정합 유사도 행렬을 다차원 척도법을 이용하여 새로운 좌표 공간에 사영된 P를 구하는 것으로 시작한다. P에 대한 i번째 클래스의 중심 C_i을 Equation (3)을 이용하여 구한다.

$$\overline{C_i} = \frac{1}{N_i} \sum_{p_j \in P_i} p_j \tag{3}$$

여기서 P_i는 i번째 클래스에 속하는 데이터의 집합이고, N_i는 i번째 클래스에 속하는 데이터의 수를 나타낸다. 이후 Equation (4)를 이용하여 i번째 클래스의 중심 C_i와 거리를 계산하고, 거리가 가장 가까운 점을 선택한다.

$$d_{it} = \bigwedge_{i=1}^{M} (\| P_{it} - \overline{C_i} \|) \tag{4}$$

여기서 M은 클래스의 개수를 의미하고, P_{it}는 i번째 클래스에서 t회차에서 발화한 시계열 데이터의 저차원 공간 벡터를 의미한다. 그리고 계산된 d_{it}와 대응되는 저차원 벡터 P_{it}를 구하고, 다시 고차원 공간으로 사상되는 시계열 데이터 W_{it}를 구한다. 마지막으로 i번째 클래스의 동적 정합 템플릿은 $V_i^T = W_{it}$로 결정된다[4].

단어를 발화한 시계열 데이터와 생성된 템플릿들 간의 유

사도를 계산하고 그중 가장 높은 유사도를 보이는 템플릿을 인식 결과로 지정한다. 인식 결과의 템플릿과 실제 데이터가 속하는 단어를 비교하여 특정 단어의 데이터 중 올바른 템플릿을 지정받은 데이터에 해당 단어의 총 발화 횟수를 나눈 값으로 최종 인식률을 구한다.

4. 실험 및 결과

본 연구에서는 이전 입 모양 인식 실험과 비교를 위해 [4]의 연구에서 사용했던 전처리 전 원본 영상 데이터 세트를 원본 영상 데이터 세트로 사용했다. 데이터 세트는 스마트폰의 Wake Up 기능을 위해 “하이”라는 단어를 앞에 붙인 10개의 실험 단어로 구성되어 있다. 각 실험 단어는 200회 이상 발화하였다. 실험 단어의 번호, 실험 단어, 발화 횟수는 Table 2와 같다.

데이터 셋은 1920 * 1080 해상도, 30fps, 조명 및 배경이 일정한 장소에서 촬영되었다.

입력 데이터로 사용될 데이터는 16명의 화자가 10가지 단어에 대해 한 단어 당 10여 회 발화한 동영상 데이터를 프레임 단위로 쪼갠 총 136,169장의 영상 데이터이다.

원본 영상 데이터 136,169장 중 무작위로 선택된 26,901장에 3가지 관심 영역을 사람이 직접 범위를 지정하고, 지정된 좌표값 데이터와 원본 영상 데이터를 YOLOv3의 학습 데이터로 사용해 모델을 학습한다. 학습 데이터 중 80%는 학습용으로 20%는 검증용으로 무작위로 나누어 학습에 사용한다. 모델의 설정으로 앵커값은 원본 데이터 해상도에 맞춘 값을 사용하고, 클래스는 입술 영역 RoI1, 입술과 턱 영역 RoI2, 입술과 턱, 뺨을 포함한 하관 영역 RoI3 3가지로 설정한다. 그리고 그에 따라 3가지 관심 영역을 검출하기 위해 모델의 구조를 조정했다. YOLOv3의 인식률을 최대화하기 위해 학습 모델의 배치와 재분의 값은 각각 64와 16으로 설정했다. 신경망의 학습은 Fig. 3과 같이 학습 중 손실의 변화가 최소화될 때까지 진행했다.

Table 2. Data Description

No.	Word	Number of utterance
ω_1	Hi Galaxy	200
ω_2	Hi Aladin	202
ω_3	Hi Smartphone	203
ω_4	Hi Camera	202
ω_5	Hi Message	212
ω_6	Hi Kakaotalk	206
ω_7	Hi Junhwagerlgi	219
ω_8	Hi Navigation	219
ω_9	Hi Email	204
ω_{10}	Hi Systran	214

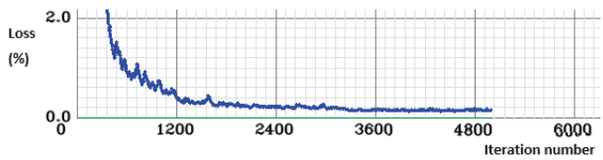


Fig. 3. Loss Rates During Learning

4000번 학습이 반복된 이후 5000번 학습이 반복될 때까지 손실의 변화가 안정되었으므로 평균 손실의 값이 0.1515인 5000번 반복된 시점에서 학습을 종료했다. 관심 영역을 검출하기 위해 사용할 가중치 데이터는 학습이 4000번 반복된 시점의 가중치를 사용했다.

각 관심 영역에 대해 학습시킨 YOLOv3 모델을 사용하여 입술 영역, 입술과 턱 영역, 입술과 턱, 뺨 영역, 3가지 관심 영역을 자동으로 검출하였다.

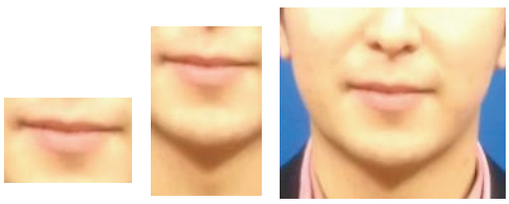
검출 결과로 원본 데이터 136,169장에서 각 3가지 관심 영역과 원본 데이터 중 42장에서 42개의 중복된 관심 영역이 검출되었으나 중복된 관심 영역의 경우 60% 이하의 클래스 예측값을, 올바르게 검출된 관심 영역의 경우 98% 이상의 클래스 예측값을 가졌기 때문에 클래스 예측값을 이용해 올바르게 검출된 관심 영역을 구분할 수 있었다.

Fig. 4는 검출된 각 관심 영역이다. 검출된 관심 영역을 추출한 뒤 추출한 각 관심 영역 집합을 이중 선형 보간법을 사용해 너비와 높이를 130 * 100화소 크기로 조정했다.

각 관심 영역의 그레이레벨 변화량 특징을 생성하고, 원본 데이터에서 입술을 근사화한 특징을 생성했다. 생성된 특징들을 결합한 후 결합된 특징에 대해 ISOMAP 차원 축소를 수행했다. 최종적으로 생성된 차원 축소된 관심 영역 특징에서 각 단어 데이터 군집에서 동적 정합으로 가장 중심에 있는 데이터를 템플릿으로 결정하고, 결정된 템플릿을 기준으로 하여 단어별 인식률을 평가했다.

각 관심 영역의 단어별 인식률을 Table 3과 Fig. 5에 나타나 있다.

각 관심 영역의 단어별 인식률 결과의 경우 입술 영역을 관심 영역으로 한 RoI_1의 평균 인식률은 97.36%, 입술과 턱을 관심 영역으로 한 RoI_2의 평균 인식률은 93.82%, 입술, 턱, 뺨을 관심 영역으로 한 RoI_3의 평균 인식률은 92.49%로 모든 관심 영역의 인식률은 90% 이상의 결과를 보였다. 각 단어별 인식률의 경우 RoI_1이 ω_1부터 ω_7까지 다른 관심 영역에 비해 높거나 동일한 인식률을 나타냈고, ω_8, ω_9에서는 RoI_2이 가장 높은 인식률을, ω_{10}에서는 RoI_3가 가장 높은 인식률을 나타냈다. 단어에 따라서 가장 높은 인식률을 보이는 관심 영역의 차이가 나타났으며, 단어마다 최적화된 관심 영역이 다르다고 판단된다. 평균 인식률이 가장 높은 관심 영역은 입술만을 포함한 RoI_1이였다. 이전 Kim의 연구[4]에서 사람이 검출한 입술 영역으로 본 연구와 동일하게 화자 독립으로 그레이레벨 변화량 특징과 입술을 근사화한 특징을 결합한 특징을 ISOMAP으로 차원 축소한 뒤 동적 정합 템플릿으로 평가한 결과인



(4A) RoI_1 (4B) RoI_2 (4C) RoI_3

Fig. 4. Detected Region of Interest

Table 3. Recognition Rate by Words for Each Region of Interest

Word/RoI	RoI_1	RoI_2	RoI_3
ω_1	97.00%	84.50%	95.00%
ω_2	99.50%	97.52%	85.64%
ω_3	100.00%	100.00%	100.00%
ω_4	99.50%	99.50%	98.51%
ω_5	100.00%	98.11%	92.45%
ω_6	98.06%	87.86%	88.83%
ω_7	98.63%	93.15%	89.04%
ω_8	92.69%	99.09%	87.21%
ω_9	98.53%	99.02%	97.06%
ω_{10}	89.72%	79.44%	91.12%
Average	97.36%	93.82%	92.49%

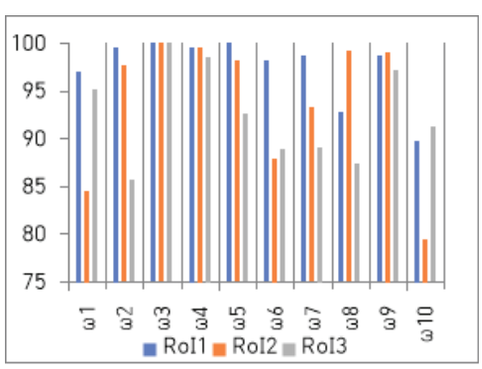


Fig. 5. Recognition Rate by Words for Each Region of Interest

Table 4. Average Recognition Rate by Region of Interest and Previous Experimental Result

RoI	Average recognition rate	Previousexperiment	Average recognition rate
RoI_1	97.36%	Y^3_{ISOMAP}	93.92%
RoI_2	93.82%		
RoI_3	92.49%		

Y^3_{ISOMAP}의 평균 인식률인 93.92%와 비교했을 때 입술 영역인 RoI1은 더 높은 평균 인식률을, 나머지 영역은 더 낮은 평균 인식률을 나타냈다. 각 관심 영역과 이전 연구의 평균 인식률 결과는 Table 4와 같다.

3가지 관심 영역 중 가장 높은 평균 인식률을 보인 입술 영역을 가장 좋은 결과로 보고 입술 영역에 대해 정규화 단계로 돌아가 실험하지 않은 두 가지 보간법인 최근접 이웃 보간법, 이중 삼차 보간법을 사용해 보간한 뒤 각 보간법으로 보간한 관심 영역을 사용해 앞의 실험을 반복하였다.

Table 5. Recognition Rate by Words for Each Interpolation

Word/ Interpolation	Nearest neighbor interpolation	Bilinear interpolation	Bicubic interpolation
ω_1	51.00%	97.00%	91.00%
ω_2	14.85%	99.50%	48.51%
ω_3	100.00%	100.00%	97.54%
ω_4	99.50%	99.50%	98.51%
ω_5	85.38%	100%	95.28%
ω_6	100.00%	98.06%	100.00%
ω_7	96.80%	98.63%	93.15%
ω_8	94.52%	92.69%	94.52%
ω_9	99.50%	98.53%	97.06%
ω_{10}	85.51%	89.72%	92.52%
Average	82.71%	97.36%	90.81%

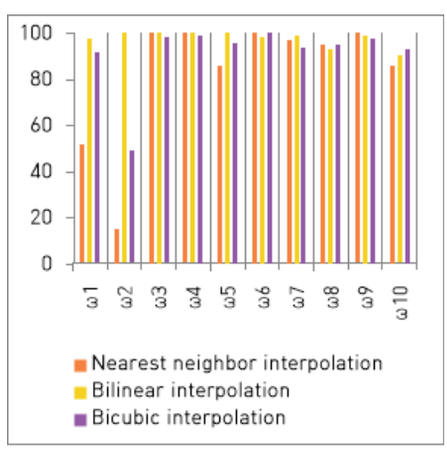


Fig. 6. Recognition Rate by Words for Each Interpolation

실험 결과 최근접 이웃 보간법을 사용한 경우 10개 단어의 평균 인식률은 82.71%, 이중 선형 보간법을 사용한 경우 10개 단어 평균 인식률은 97.36%, 이중 삼차 보간법을 사용한 경우 10개 단어의 평균 인식률은 90.81%이었다.

인식 실험 결과 관심 영역의 범위의 경우 가장 포함하는 범위가 적은 입술 영역만을 포함하는 RoI_1의 평균 인식률이 가장 높았고, 입술 영역과 턱을 포함하는 RoI_2, 입술, 턱, 뺨을 포함하는 RoI_3 순서로 평균 인식률이 낮아지는 결과를 확인했다. 그리고 단어에 따라서 높은 인식률을 보이는 관심 영역의 차이가 있다는 것을 ω_8, ω_9, ω_{10}의 인식률 결과로 확인할 수 있었다.

평균 인식률이 가장 높았던 RoI_1를 사용하여 각 보간법에 대해 인식 실험을 한 결과, 평균 인식률은 이중 선형 보간법이 가장 높았으며, 이중 삼차 보간법, 최근접 이웃 보간법 순서로 인식률이 낮아지는 결과를 확인했다. 본 연구에서 실험한 음성 인식 시스템의 입 모양 인식에서 가장 높은 성능을 보인 보간법은 이중 선형 보간법인 것을 확인할 수 있었다.

같은 특징과 차원 축소 방법으로 실험한 Kim의 연구[4]의 사람이 직접 검출한 입술 영역의 특징별 인식 및 평가 결과와 비교했을 때, 본 연구에서 YOLOv3 모델을 사용하여 자동으로 검출한 입술 영역을 사람이 휴리스틱하게 검출한 입술 영역보다 입 모양 인식 실험에서 높은 성능을 보인 점과 3가지 보간법 중 이중 선형 보간법이 가장 높은 성능을 보임 점을 확인할 수 있었다.

5. 결 론

화자 독립 입 모양 인식 실험의 인식률을 개선하기 위해 많은 방법이 연구됐고, 이를 통해 화자 독립 입 모양 인식의 성능이 개선됐다. 그러나 최근 연구에서 인식 성능을 개선하기 위해서 주로 인식 실험에 사용되는 신경망의 구조를 수정하는 방식이 이루어졌다. 인식 실험의 관심 영역은 주로 입술 영역에 한정되고, 발화 중에 생기는 다른 변인들이 포함된 관심 영역에 대해서는 고려되지 않았었다. 또한, 영상 데이터들을 정규화하는 데 사용하는 보간법에 대한 차이도 고려되지 않았었다. 따라서 본 연구에서 입 모양을 인식하는 방법 구조의 변화 없이 같은 실험 조건에서 관심 영역과 보간법의 차이가 인식률을 개선할 수 있는지 확인하기 위해 포함하는 변인의 수가 다른 3가지 관심 영역에 대해 인식 실험을 수행하고, 3가지 보간법으로 영상 데이터를 정규화하고 인식 실험과 결과를 비교하는 것으로 기존의 관심 영역과 제안한 관심 영역의 성능 차이, 제안한 보간법들의 성능 차이를 검증하는 실험을 제안하였다.

그 결과, 본 연구의 실험 조건과 방법으로 입 모양 인식을 수행했을 때 가장 높은 평균 인식률 성능을 보이는 관심 영역은 입술 영역인 것과 단어에 따라서 최적화된 관심 영역의 범위에 차이가 존재하는 것, 관심 영역 크기 정규화에 사용된 보간법이 이중 선형 보간법일 때 가장 높은 평균 성능을 얻을 수 있다는 것, 정규화에 사용되는 보간법의 차이가 인식률에 영향을 미칠 수 있다는 것을 확인할 수 있었다. 그리고 객체 탐지 인공신경망 모델을 사용해 관심 영역을 검출하는 것이 사람이 직접 검출하는 방법보다 빠르고, 인식 실험에서 더 높은 결과를 얻을 수 있음을 확인했다.

본 연구의 실험에서 입술 영역 외 영역을 포함하여 관심 영역으로 사용했을 때, 성능이 떨어지는 결과는 사람의 입 모양 인식의 경우 입술의 모양에 더하여 뺨과 턱의 움직임을 참고하는 방식으로 입술이 아닌 변인들이 인식에 도움이 되었지만, 본 연구의 실험에서는 관심 영역의 너비, 높이를 130 * 100화소 크기로 정규화하고, 격자 기반 특징인 그레이레벨 변화량 특징을 사용했다. 동일한 격자에 포함되는 특징이 입술 영역만을 관심 영역으로 할 때는 온전히 입술의 특징만이 포함되지만, 턱, 뺨이 포함되었을 때 같은 격자에 포함되는 특징이 많아져 희석되기 때문에, ω_1부터 ω_7까지의 인식 결과에서 잡음으로 작용한다고 판단한다. 또한, 3가지 보간법을 사용하여 관심 영역을 정규화했을 때, 성능의 차이가 발생하는 결과는 각 보간법으로 영상이 정규화되면서 변화시키는 화소값들의 차이가 결과가 격자 기반 그레이레벨 변화량 특징을 사용하는 실험 조건에서 유의미한 영향을 미칠 수 있는 수준의 차이를 발생시키기 때문이고, 이중 선형 보간법의 부드러운 영상을 출력하는 특징이 각 프레임의 정규화에 의한 변화에 영향을 적게 주었기 때문이라고 판단한다.

본 연구에서 검증한 관심 영역의 범위에 따른 인식률 차이의 결과, 관심 영역의 크기를 조정할 때 사용되는 보간법 차이의 결과 등을 바탕으로 객체 탐지 인공신경망 모델을 사용한 관심 영역의 검출이 입 모양 인식의 성능을 높일 수 있을 것으로 기대한다.

References

[1] Z. Zhang, J. Geiger, J. Pohjalainen, A. E. D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology*, Vol.9, No.49, pp.1-28. 2018.

[2] C. Bregler and Y. Koing, ""Eigenlips" for robust speech recognition," in *Proceedings of the ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing,* Adelaide, Vol.2, pp.669-672, 1994.

[3] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel, "Towards unrestricted lip reading," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol.14, No.5, pp.571-585, 2000.

[4] Y. G. Kim, "Feature selection method for speaker independent lip reading on noisy environments," Ph.D. dissertation, Chungbuk National University, Cheongju, Korea, 2019.

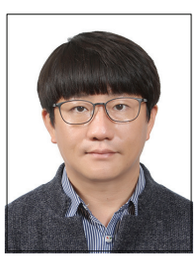
[5] 한민경, "독화에 청각적으로 제공된 기본 주파수(F0) 보완정보," *Communication Sciences & Disorders*, Vol.1, No.1, pp.150-177, 1996.

[6] D. G. Stork and M. E. Hennecke, "Speechreading by humans and machines: models, systems, and applications," *Berlin: Springer Science & Business Media*, pp.525-531, 1996.

[7] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, pp.6319-6323, 2020.

[8] P. Ma, B. Martinez, and M. Pantic, "Towards practical lip-reading with distilled and efficient models," in *Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, pp.7608-7612, 2021.
[9] K. Vayadande, T. Adsare, N. Agrawal, T. Dharmik, A. Patil, and S. Zod, "LipReadNet: A deep learning approach to lip reading," in *Proceedings of the 2023 International Conference on Applied Intelligence and Sustainable Computing*, Dharwad, pp.1-6, 2023.
[10] 최병문, "구화교육," 한국구화학교, 1970.
[11] M. Hao, M. Mamut, N. Yadikar, A.Aysa, and K. Ubul, "A survey of research on lipreading technology," *IEEE Access*, Vol.8, pp.204518-204544, 2020.
[12] 김민정, "임상중심 말소리장애." 1st ed, Seoul: 학지사, 2021.
[13] J. J. O'Neill and H. J. Oyer, "Visual communication for the hard of hearing: History, research, and methods," 2nd ed., New Jersey: Prentice Hall, 1981.
[14] S. H. Cho and C. D. Choi, "Viseme and its teaching strategy for speech-reading and language normalization of people with hearing loss," *Audiology and Speech Research*, Vol.14, No.4, pp.219-226, 2018.
[15] G. Potamianos and C. Neti, "Improved ROI and within frame discriminant features for lipreading," in *Proceedings of the 2001 International Conference on Image Processing*, Thessaloniki, Vol.3, pp.250-253, 2001.
[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, Las Vegas, pp.779-788, 2016.
[17] J. Redmon and A. Farhadi, "Yolov3: An Incremental Improvement," *Computer Vision and Pattern Recognition*, Vol.1804, pp.1-6, 2018.
[18] G. Jocher and A. Chaurasia, "Ultralytics YOLOv8 Docs" [Internet], https://docs.ultralytics.com/ko
[19] J. Luettin and N. A. Thacker, "Speechreading using probabilistic models," *Computer Vision and Image Understanding*, Vol.65, No.2, pp.163-178, 1997.
[20] Y. Lan, B. J. Theobald, R. Harvey, E. J. Ong, and R. Bowden, "Improving visual features for lip-reading," in *Proceedings of the Auditory-visual Speech Processing 2010*, Hakone, paper S7-3, 2010.
[21] B. Sujatha and T. Santhanam, "A novel approach intergrating geometric and Gabor wavelet approaches to improvise visual lip-reading," *International Journal of Soft Computing (IJSC)*, Vol.5, pp.13-18, 2010.
[22] M. Z. Ibrahim and D. J. Mulvaney, "Robust geometrical-based lip-reading using Hidden Markov models," in *Proceedings of the EUROCON 2013*, Zagreb, pp.2011-2016, 2013.
[23] 박혜영, 이관용, "패턴 인식과 기계학습," 1st ed., Gyeonggi-do: 이한출판사. 2011.
[24] 박창순, 이광용, 이형석, 정호영, "생활 속의 임베디드 소프트웨어", 1st ed., Seoul: U-북, 2007.
[25] A. Koumparoulis, G. Potamianos, Y. Mroueh, and S. J. Rennie, "Exploring ROI size in deep learning based lipreading," in *Proceedings of the Auditory-visual Speech Processing 2017*, Stockholm, pp.64-69, 2017.

한 재 혁
https://orcid.org/0009-0000-9397-395X
e-mail : haraisi22@gmail.com
2019년 충북대학교 컴퓨터공학과(학사)
2021년 충북대학교 컴퓨터공학과(석사)
2023년 ~ 현 재 ㈜시아이솔루션 선임
2024년 ~ 현 재 충북대학교 컴퓨터공학과 박사과정
관심분야 : Artificial Neural Network & Digital Therapeutics



김 용 기
https://orcid.org/0000-0002-8646-0758
e-mail : moodeathkyk@gmail.com
2008년 청주대학교 관광경영학과(학사)
2014년 충북대학교 컴퓨터공학과(석사)
2019년 충북대학교 컴퓨터공학과(박사)
2020년 ~ 2023년 우석대학교 컴퓨터공학과 조교수
2024년 ~ 현 재 충북대학교 연구원
관심분야 : Featrue Engineering, AVSR, Lipreading, Machine Learning

김 미 혜
https://orcid.org/0000-0002-3531-0215
e-mail : mhkim@cbnu.ac.kr
1992년 충북대학교 수학과 (이학사)
1994년 충북대학교 수학과 (이학석사)
2001년 충북대학교 수학과 (이학박사)
2004년 ~ 현 재 충북대학교 컴퓨터공학과 교수
관심분야 : Serious Game, Fuzzy Measures & Fuzzy Integrals Digital Therapeutics