음성인식 시스템의 입 모양 인식개선을 위한 관심영역 추출 방법

한재혁 [1], 김미혜 [2]
[1] 충북대학교 컴퓨터공학과 연구원
[2] 충북대학교 컴퓨터공학과 교수

haraisi22@gmail.com, mhkim@cbnu.ac.kr

RoI Detection Method for Improving Lipreading Reading in Speech Recognition Systems

Jae-Hyeok Han[1], Mi-Hye Kim[2]
[1,2]Dept. of Computer Engineering, Chungbuk National University

입 모양 인식은 음성인식의 중요한 부분 중 하나로 이를 개선하기위한 다양한 연구가 진행되어 왔다. 기존의 연구에서는 주로 입술주변 영역을 관찰하고 인식하는데 초점을 두었으나, 본 논문은 음성인식 시스템에서 기존의 입술영역과 함께 입술, 턱, 뺨 등 다른 관심 영역을 고려하여 음성인식 시스템의 입모양 인식 성능을 비교하였다. 입 모양 인식의 관심 영역을 자동으로 검출하기 위해 객체 탐지 인공신경망을 사용하며, 이를 통해 다양한 관심영역을 실험하였다. 실험 결과 입술영역만 포함하는 ROI 에 대한 결과가 기존의 93.92%의 평균 인식률보다 높은 97.36%로 가장 높은 성능을 나타내었다.

1. 서론

최근에는 음성 인식 기술이 상용화되어 모바일 및 웨어러블 기기, 스마트 홈 시스템, 차량용 인터페이스 등 여러 분야에서 다양하게 사용되고 있다. 이렇게 사용되는 음성 인식 시스템에서의 인식률을 높이기 위해 특징 추출, 모델 보정, 분류 알고리즘과 같은 원리 분야에서 여러 방법이 연구되어 왔다[1]. 또한, 소음이 심한 환경처럼 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 향상시킬 수 있다는 연구 결과가 나타났다[2].

입 모양 인식을 실험하는 Meier, Stiefelhagen, Yang, Waibel(2000)의 연구[3], 김용기(2019) 연구[4] 등의 경우 모두 관심 영역을 입술 주변 영역 한 가지로 설정하고 입 모양을 인식하는 실험을 수행하였다.

이와 다르게 사람의 독화의 경우 기본적으로 입의 모양을 가지고 상대방의 발화를 인지하지만, 입술만이 아닌 표정, 턱, 혀의 움직임에도 의존하며(한민경 1996)[5], 뺨의 들썩거림은 m-p-b 발음을 구별할 때 사용할 수 있는 시각적 단서이다(Stork and Hennecke 1996:525-531)[6]. 이처럼 사람의 입 모양 인식에 있어서 입술 주변 영역이 아닌 다른 영역에서의 변화들도 중요한 정보이다.

따라서 본 논문에서는 음성 인식 시스템의 입 모양

인식에서 기존의 입술 영역의 성능을 사람의 독화에서 사용되는 변인들이 포함된 다른 관심 영역과 비교하여 어떤 관심 영역이 음성 인식 시스템의 입 모양 인식에서 가장 높은 성능의 영역인지 검증하고자 한다. 사람이 하는 독화에서의 변인들을 참고해 기존의 관심 영역인 입술 영역과 함께, 입술과 턱이 포함된 영역, 입술과 턱 그리고 뺨이 포함된 얼굴의 하관 영역, 이렇게 3 가지 영역을 관심 영역으로 제안한다. 그리고 여러 개의 관심 영역을 검출하기 위해 객체 탐지 인공신경망을 사용해 관심 영역을 자동으로 검출하는 방법을 제안한다. 본 논문에서 제안된 관심 영역의 인식 실험을 수행하는 과정은 다음과 같다. 첫째, 객체 탐지 인공신경망을 본 논문에서 제안된 3 가지 관심 영역에 대해 학습시키고, 학습된 인공신경망을 사용해 원본 데이터에서 관심 영역을 탐지한다. 그리고 탐지된 관심 영역을 검출하고 이중 선형 보간을 사용한 크기 정규화와 그레이레벨화, 히스토그램 평활화 과정을 수행한다. 둘째, 특징들을 생성하고, 같은 관심 영역의 특징끼리 결합한 뒤 결합된 특징들을 차원 축소한다. 셋째, 차원 축소된 각 관심 영역별 특징을 평가하여 화자 독립 입 모양 인식에서의 가장 높은 성능을 가진 관심 영역을 검출한다.

2. 관련 연구

독화(Lipreading)란 입술이 움직이는 모양을 보고 상대편이 하는 말을 알아내는 방법으로 청각 장애인이나 오디오 비주얼 음성 인식 시스템과 같이 자동 독화를 포함하는 시스템에서 사용하는 방법이다.

청각 장애인이 사용하는 독화의 경우 오래전부터 제한된 범위에서 입의 모양만으로 발화를 이해할 수 있는 교육을 실시하고 있다[7]. 그리고 말의 청각적, 시각적 인지를 위해서 독화를 하는 많은 청각 장애인들은 화자의 얼굴 표정뿐만이 아니라 입술, 턱, 그리고 혀의 움직임에 의존한다[5].

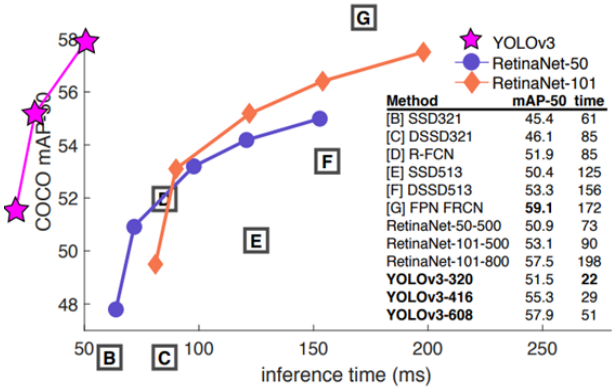
음성 인식 시스템의 자동 독화의 경우, 기존의 독화에서 입술이 움직이는 모양을 보고 말을 알아내는 것과 같이 독화에 사용할 관심 영역을 주로 입술 영역만으로 지정하고 자동 독화를 수행했었다.

김용기(2019)[4]는 입술 영역을 너비, 높이 130 * 100의 크기로 사람이 직접 검출하고 검출된 입술 영역에서 여러 특징을 추출하고 평가 후 최적화된 특징을 선별해 결합한 뒤 결합된 특징들을 차원 축소해 동적 정합의 템플릿으로 차원 축소된 특징들을 평가하고 은닉 마르코프 모델을 사용해 자동 독화를 수행했다.

사람이 하는 독화의 관련 연구에 따르면 사람이 하는 독화에서는 입 모양뿐만 아니라 독화 과정에 포함될 수 있는 변인 중 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 중요한 것이라는 점[8], 뺨의 들썩거림은 m-p-b 발음을 구별할 때 사용할 수 있는 시각적 단서임 점[6], 사람이 발화할 때 음소마다 특별한 안면부 및 구강부의 특징을 갖고 있다는 점[9], 사람이 독화를 할 때 관심 영역에 뺨과 턱까지 포함되는 것이 좋다는[6] 여러 이전 연구 결과에서 확인할 수 있듯이 일반적인 독화에서 입술 영역 외의 다른 변인들이 추가적인 단서가 되고, 인식률에 영향을 미치는 것을 알 수 있다. 그리고 음성 인식 시스템의 입 모양 인식에서도 영상의 해상도 확장에 따른 관심 영역 범위가 증가했을 때 인식률의 증가를 확인할 수 있었다[10].

오디오 비주얼 음성 인식 시스템에서 독화의 성능을 개선하기 위해 관심 영역을 설정해 영상 데이터에서 관심 영역을 검출하고 특징을 추출해 사용한다.

YOLO(You Only Look Once)는 Redmon 등이 발표한 객체 탐지를 위한 인공신경망 모델이다[11]. YOLO 는 기존의 객체 탐지 모델에 근접한 정확도를 가지면서 상대적으로 고속으로 객체를 탐지하고자 개발된 모델로 객체 탐지 시 이미지 전체를 단 한 번만 모델에 통과시키고, 단 하나의 통합된 인공신경망 모델을 사용하며, 실시간으로 객체 탐지를 하는 특징을 가지고 있다. 2018 년에는 이전 모델에서 5 가지 변경 점을 통해 성능을 개선한 v3 가 발표되었다[12]. (그림 1)은 COCO 데이터 셋을 사용했을 때 YOLO v3 와 다른 신경망들의 성능을 비교한 결과이다.

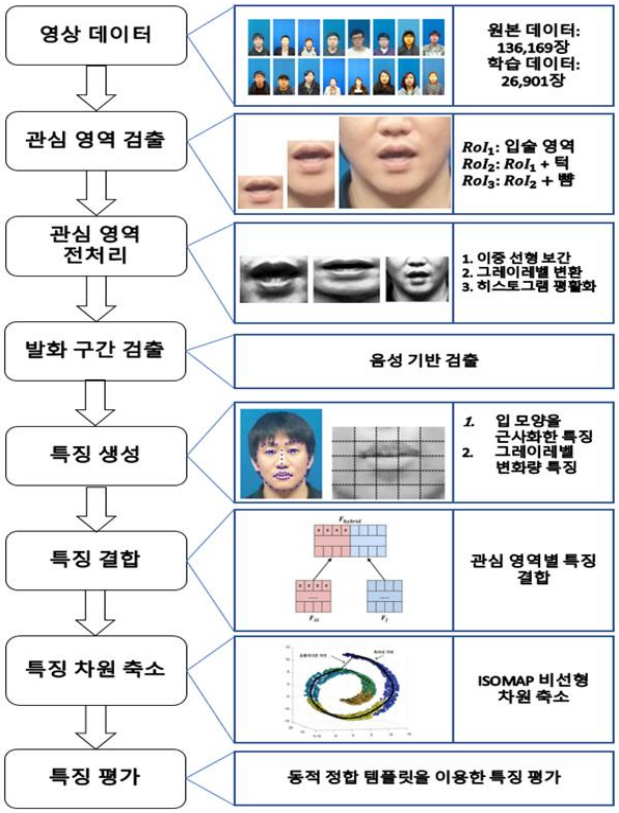


(그림 1) COCO 데이터 셋에서의 YOLO v3 객체 탐지 성능.

3. 실험 방법

본 연구에서는 기존의 입술 영역과 함께 발화할 때 변화하는 턱과 뺨 같은 다른 영역들을 관심 영역으로 포함하는 것이, 사람이 하는 독화에서 입술이 아닌 다른 변인들이 독화의 인식률에 영향을 미쳤던 것처럼 오디오 비주얼 음성 인식 시스템의 입 모양 인식에서도 입술이 아닌 변인들이 인식률에 어떤 영향을 미치는지, 기존의 입술 영역과 비교해 검증하고자 한다. 이 검증을 위해 본 연구에서 발화할 때 변화하는 부위들인, 턱과 뺨을 관심 영역에 포함해 3 가지 관심 영역인 입술 영역, 입술과 턱 영역, 입술과 턱, 뺨을 포함한 영역을 설정한 뒤 객체 탐지 인공신경망을 사용해 검출하고, 객체 탐지 인공신경망 모델을 사용해 검출한 관심 영역들과 사람이 직접 검출한 입술 영역 관심 영역의 성능을 비교하고자 한다.

(그림 2)는 본 실험의 흐름도이다. 본 실험에서는 16 명의 화자가 각각 10 개의 단어를 최소 10 회에서 최대 16 회까지 발화하는 영상 데이터를 사용한다. 데이터 중 일부를 학습 데이터로 사용해 YOLO v3 신경망 모델을 입술 영역, 입술과 턱 영역, 얼굴 하관 영역에 대해 학습시킨 뒤 YOLO v3 신경망 모델에 영상 데이터를 입력 데이터로 입력해 각 영상의 프레임에 대해 3 가지 관심 영역의 좌표를 검출한다. 그리고 검출된 관심 영역의 좌표를 사용해 파이썬 프로그램을 통해 검출된 영역을 원본 영상 데이터에서 잘라서 편집한다. 각 편집된 관심 영역들을 크기를 정규화하기 위해 너비와 높이를 130 * 100 화소 크기로 조정한다. 조정된 관심 영역을 그레이레벨로 변환한 뒤 히스토그램 평활화를 수행해 전처리된 관심 영역을 생성한다. 발화 구간은 영상의 음성 데이터를 기준으로 검출한다. 원본 영상 데이터에서 입 모양을 근사화한 특징을 생성하고, 각 관심 영역에 대해서 그레이레벨 변화량 특징을 생성한다. 각 관심 영역별로 두 특징 집합을 결합하고 ISOMAP 을 사용하여 결합된 3 가지 관심 영역의 각 특징 집합 차원을 비선형 차원 축소한다. 최종적으로 동적 정합 템플릿 결정 및 인식 방법으로 최종적으로 생성된 각 3 가지 관심 영역의 결합 및 차원 축소된 특징과 이전 연구에서의 결과를 평가, 비교한다.



(그림 2) 관심 영역별 입 모양 인식 실험의 전체 흐름도.

4. 실험 및 결과

본 연구에서는 입 모양 인식 실험을 위해(김용기 2019)의 논문에서 사용했던 입 모양 인식을 위한 데이터 셋을 사용했다. 데이터 셋은 스마트폰의 Wake Up 기능을 위해 “하이”라는 단어를 붙인 10 개의 실험 단어로 구성되어 있다. 실험 단어는 <표 1>과 같다.

<표 1> 실험 단어 데이터 셋

번호	실험 단어	전체 발화수

ω_1	하이갤럭시	200
ω_2	하이알라딘	202
ω_3	하이스마트폰	203
ω_4	하이카메라	202
ω_5	하이메시지	212
ω_6	하이카카오톡	206
ω_7	하이전화걸기	219
ω_8	하이네비게이션	219
ω_9	하이이메일	204
ω_{10}	하이시스트란	214

검출된 관심 영역의 범위는 (그림 3)에 나타나 있다.

제안한 RoI	평균 인식률	이전 연구	평균 인식률
RoI_1	97.36	Y^3_{ISOMAP}	93.92
RoI_2	93.82		
RoI_3	92.49		

(A) 원본 데이터 (B) 입술 (C) 입술, 턱 (D) 하관

(그림 3) 검출된 관심 영역.

추출된 각 관심 영역 집합을 이중 선형 보간법을 사용해 너비와 높이를 130 * 100 화소 크기로 조정했다. 검출된 관심 영역들에서 MATLAB 을 사용해 그레이 레벨 변화량 특징을 생성하고, 원본 데이터에서 입술을 근사화한 특징을 생성했다. 생성된 특징을 MATLAB 을 사용해 특징 결합 및 결합된 특징에 대

관심 영역	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
RoI_1	97.00	99.50	100	99.50	100	98.06	98.63	92.69	98.53	89.72
RoI_2	84.50	97.52	100	99.50	98.11	87.86	93.15	99.09	99.02	79.44
RoI_3	95.00	85.64	100	98.51	92.45	88.83	89.04	87.21	97.06	91.12

한 ISOMAP 차원 축소를 수행했다. 최종적으로 생성된 관심 영역별 특징에서 각 단어 데이터 군집에서 동적 정합으로 가장 중심에 있는 데이터를 템플릿으로 결정하고, 결정된 템플릿을 기준으로 하여 단어별 인식률을 평가했다.

각 관심 영역의 단어별 인식률은 <표 2>에 나타나 있다.

<표 2> 각 관심 영역의 단어별 인식률

그 결과 입술 영역만을 포함하는 관심 영역인 RoI_1 의 10 개 단어의 평균 인식률은 97.36%, 입술과 턱을 포함하는 관심 영역인 RoI_2 의 10 개 단어의 평균 인식률은 93.82%, 얼굴 하관을 전부 포함하는 관심 영역인 RoI_3 의 10 개 단어의 평균 인식률은 92.49%로 모든 관심 영역의 인식률은 90%가 넘고, 이전 김용기(2019)의 연구에서 사람이 검출한 입술 영역으로 실험한 Y^3_{ISOMAP} 의 10 개 단어의 평균 인식률인 93.92%와 비교했을 때 RoI_1 은 더 높은 인식률을, 나머지 영역은 더 낮은 인식률을 나타냈다. 각 관심 영역의 10 개 단어에 대한 평균 인식률은 <표 3>에 나타나 있다.

<표 3> 관심 영역별 인식 결과 및 이전연구 인식결과

인식 실험 결과 포함하는 영역이 가장 적은 입술 영역만을 포함하는 RoI_1 의 인식률이 가장 높고 입술 영역과 턱을 포함한 RoI_2, 얼굴 하관 전체를 포함한 RoI_3 의 인식률은 그보다 낮아지는 결과를 보였다. 본 연구에서 제안한 음성 인식 시스템의 입 모양 인식에서 일반적으로 사용하는 입술 영역이 가장 높은 성능을 가진 관심 영역인 것을 확인할 수 있었다.

또한, 같은 특징과 차원 축소 방법으로 실험한 김용기(2019)의 사람이 직접 검출한 입술 영역의 특징별 인식 및 평가 결과의 Y^3_{ISOMAP} 와 비교했을 때, 화자 독립 상황에서의 인식률은 93.92%로 본 연구의 YOLO v3 모델을 사용한 자동으로 검출한 입술 영역

이 사람이 휴리스틱하게 검출한 입술 영역보다 입 모양 인식 실험에서 높은 성능을 보임을 확인할 수 있었다.

5. 결론

본 연구의 실험 조건에서는 관심 영역에서 포함한 변인이 입술 영역에 한정될 때 가장 높은 결과를 얻을 수 있는 것으로 판단되며, 본 연구에서 사용한 방법의 입 모양 인식에서 가장 높은 성능을 보이는 관심 영역은 입술 영역임을 확인했다. 또한, 객체 탐지 인공신경망 모델을 사용해 관심 영역을 검출하는 것이 사람이 직접 검출하는 방법보다 빠르고, 인식 실험에서 더 높은 결과를 얻을 수 있음을 확인했다.

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성(Grand ICT 연구센터) 사업의 연구결과로 수행되었음 (IITP-2023-2020-0-01462)

[1] Zhang, Z., 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*, 9(5), 49. 2018

[2] Bregler, C. and Y. Koing, "Eigenlips for Robust Speech Recognition," *In Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing,* Vol 2. pp. II-669, 1994

[3] Meier, U., R. Stiefelhagen, J. Yang, A. Waibel, "Towards Unrestricted Lip Reading," *International Journal of Pattern Recognition and Artificial Intelligence*, 14(5), 571-785, 2000.

[4] 김용기, 소음 환경에서 화자독립 입 모양 인식을 위한 특징 선택 기법, 충북대학교 대학원 박사학위 논문, 2019.

[5] 한민경, "독화에 청각적으로 제공된 기본 주파수(F0) 보완정보," *Communication Sciences & Disorders*, 1, 150-176, 1996.

[6] Stork. D. G., M.E. Hennecke, "*Speechreading by Humans and Machines*," Springer, 1996.

[7] 최병문, "구화교육," 한국구화학교, 1970.

[8] O'Neill, J. J., H. J. Oyer, "*Visual communication for the hard of hearing: History, Research, and Methods*," Prentice Hall, 1981.

[9] 조소현, 최참도, "청각장애인의 독화/판독과 언어 정상화를 위한 독화소 및 독화교수전략," *Audiology and Speech Research*, 14(4), 219-226, 2018

[10] Potamianos, G., C. Neti, "Improved ROI and within Frame Discriminant Features for Lipreading," *2001 International Conference on Image Processing*, vol.3, 250-253, 2001.

[11] Redmon, J., S. Divvala, R. Girshick, A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Computer Vision and Pattern Recognition*, 779-788, 2016.

[12] Redmon, J., A. Farhadi, "YOLO v3: An Incremental Improvement," *Computer Vision and Pattern Recognition, 2018.*