입 모양 인식을 통한 구어 텍스트화 인터페이스

휴먼지능정보공학과 201710784 **신은화**
휴먼지능정보공학과 201710876 **박희지**
휴먼지능정보공학과 201710805 **최영윤**
휴먼지능정보공학과 201710809 **최희수**
휴먼지능정보공학과 201710812 **한우정**

목차

1. 프로젝트 소개

프로젝트 소개 | 주제 선정 배경

World Health Organization | Health Topics | Countries | Newsroom | Emergencies | Data | About Us

Launch of the World Report on Hearing

3 March 2021 11:00 – 12:00 CET | Virtual

BACKGROUND

Globally over 430 million people experience disabling hearing loss, a number that could grow to nearly 700 million by 2050. When unaddressed, hearing loss poses a significant challenge for all age groups, hindering language development, communication, cognition,

Related

World Report on Hearing

World Hearing Day 2021

2020년 주요 지표별 등록장애인 현황

(단위 : 천 명, %)

합계: 2,633(100)

성별 구분	인원	비율
남성	1,521	57.8
여성	1,112	42.2

연령별 구분	인원	비율
0~9세	32	1.2
10~19세	59	2.2
20~29세	98	3.7
30~39세	122	4.6
40~49세	243	9.2
50~59세	452	17.2
60~64세	314	11.9
65~69세	289	11.0
70~79세	585	22.2
80세~	440	16.7

장애유형별 구분	인원	비율
지체	1,207	45.8
시각	252	9.6
청각	396	15.0
언어	22	0.8
지적	217	8.2
뇌병변	250	9.5
자폐성	31	1.2
정신	104	3.9
신장	98	3.7
심장	5	0.2
호흡기	12	0.5
간	14	0.5
안면	3	0.1
장루요루	15	0.6
뇌전증	7	0.3

장애정도별 구분	인원	비율
심한 장애	985	37.4
심하지 않은 장애	1,648	62.6

시도별 구분	인원	비율
서울	394	15.0
부산	176	6.7
대구	126	4.8
인천	146	5.5
광주	70	2.7
대전	73	2.8
울산	51	1.9
세종	12	0.5
경기	570	21.6
강원	102	3.9
충북	98	3.7
충남	134	5.1
전북	132	5.0
전남	141	5.4
경북	181	6.9
경남	189	7.2
제주	37	1.4

HOME > 통계뉴스 > 통계뉴

김세진 기자

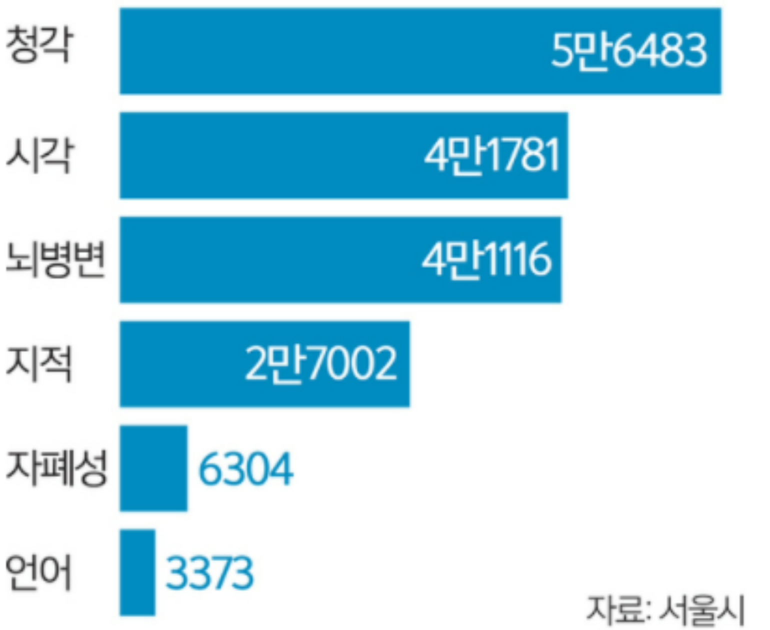
장애대학생 온라인 줌 강의 '산 넘어 산'

수강신청 난관…편의 부족 시험·과제 '끙끙'

"학습권 보장" 비대면 교육운영 매뉴얼 필요

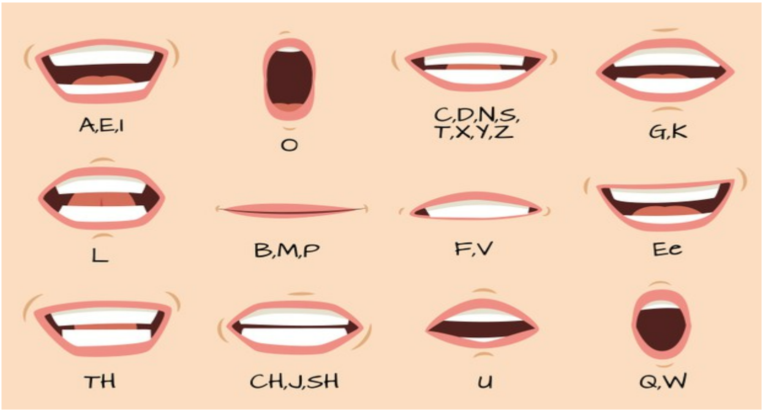
의사소통에 어려움을 겪는 유형의 장애인 (단위: 명)

*2020년 7월 서울시 등록 기준



자료: 서울시

프로젝트 소개 | 주제 선정 배경



A. 음성인식 기반

소음문제에 취약

부정확한 발음은 제대로 인식되기 어려움

문제점 보완 가능

B. 입술 모양 인식 기반

소음문제 없음

부정확한 발음에도 강함

2. 영어 *Data*를 이용한
입 모양 인식

영어 Data

1) 사용한 DataSet

Achraf Ben-Hamadou

Search this site

Navigation

::Homepage

::Publications

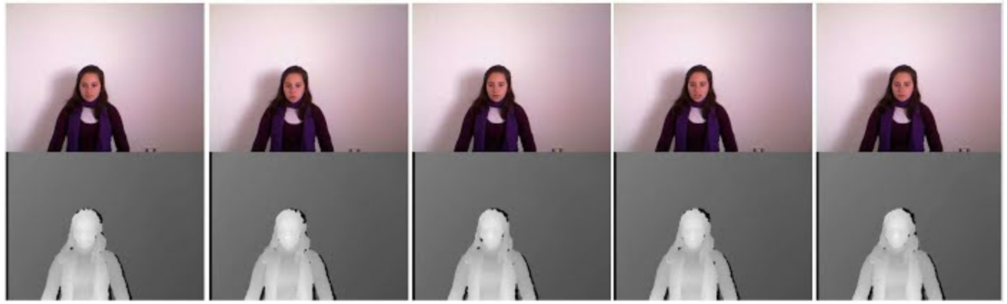
::Datasets

MIRACL-VC1

::Calendar

::Datasets >

MIRACL-VC1



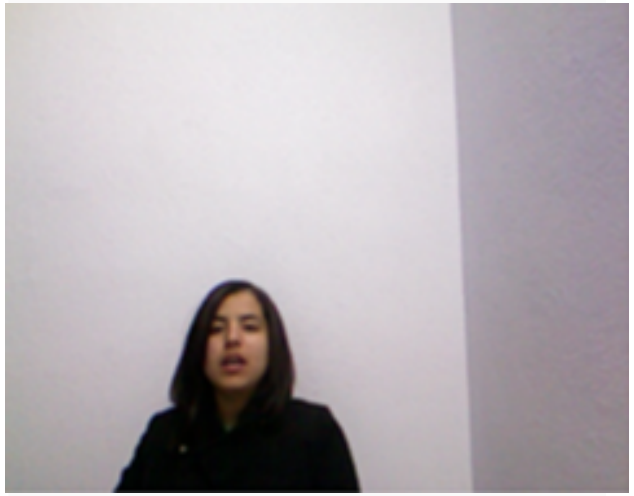
MIRACL-VC1 is a lip-reading dataset including both depth and color images. **It can be used for diverse research fields like visual speech recognition, face detection, and biometrics.** Fifteen speakers (five men and ten women) positioned in the frustum of an MS Kinect sensor and utter ten times a set of ten words and ten phrases (see the table below). Each instance of the dataset consists of a synchronized sequence of color and depth images (both of 640x480 pixels). The MIRACL-VC1 dataset contains a total number of 3000 instances.

ID	Words	ID	Phrases
1	*Begin*	1	*Stop navigation.*
2	*Choose*	2	*Excuse me.*
3	*Connection*	3	*I am sorry.*
4	*Navigation*	4	*Thank you.*
5	*Next*	5	*Good bye.*
6	*Previous*	6	*I love this game.*
7	*Start*	7	*Nice to meet you.*
8	*Stop*	8	*You are welcome.*
9	*Hello*	9	*How are you?*
10	*Web*	10	*Have a good time.*

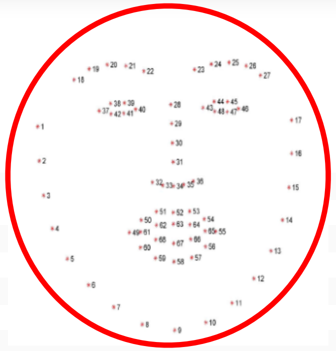
10개의 단어와 10개의 문장

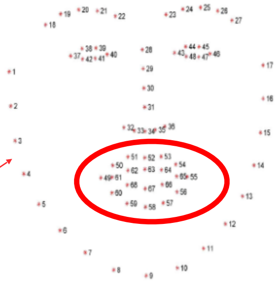
영어 Data

2) DataSet 전처리



원본





얼굴DataSet

+augmentation

입술DataSet

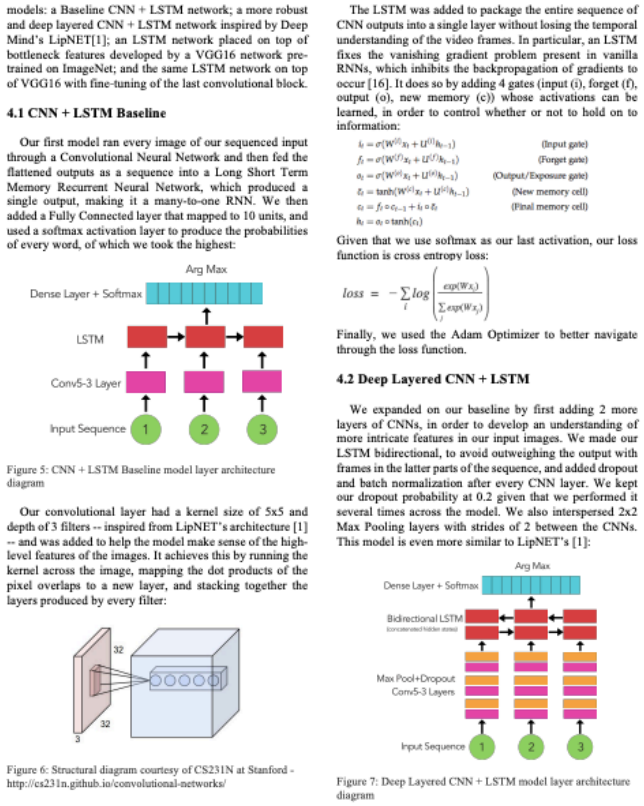
+augmentation

얼굴영역 검출

입술영역 검출

영어 Data

3) 모델 설명



얼굴 데이터셋
CNN, VGG 모델

참고논문
A. Gutierrez and Z.-A. Robert, Lip Reading Word Classification, ed, 2017.

영어 Data

3) 모델 설명

Data. Lip

model	train_accmodel	train_loss	val_acc	val_loss
CNN + LSTM	1.0	0.003	0.66	1.53
✓VGG-16 + LSTM	0.93	0.19	0.68	1.29
Xception + LSTM	0.97	0.08	0.54	1.76
✓MobileNet + LSTM	0.97	0.09	0.73	1.09
EfficientNet + LSTM	0.91	0.25	0.66	1.09

Data. Face

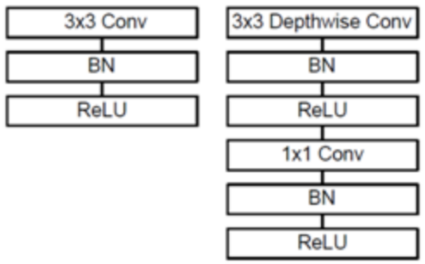
model	train_acc	train_loss	val_acc	val_loss
CNN + LSTM	0.99	0.04	0.49	2.28
VGG-16 + LSTM	0.98	0.08	0.54	1.59
Xception + LSTM	0.96	0.12	0.51	1.99
MobileNet + LSTM	0.73	0.87	0.30	2.28
EfficientNet + LSTM	0.97	0.09	0.57	1.54

입술 · 얼굴 data에 대한 5개의 모델 성능 실험

전체적으로 Lip data에서 더 좋은 성능

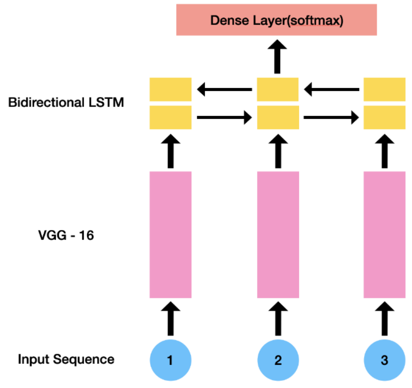
MobileNet+LSTM & VGG-16+biLSTM 좋은 성능

1



MobileNet + LSTM

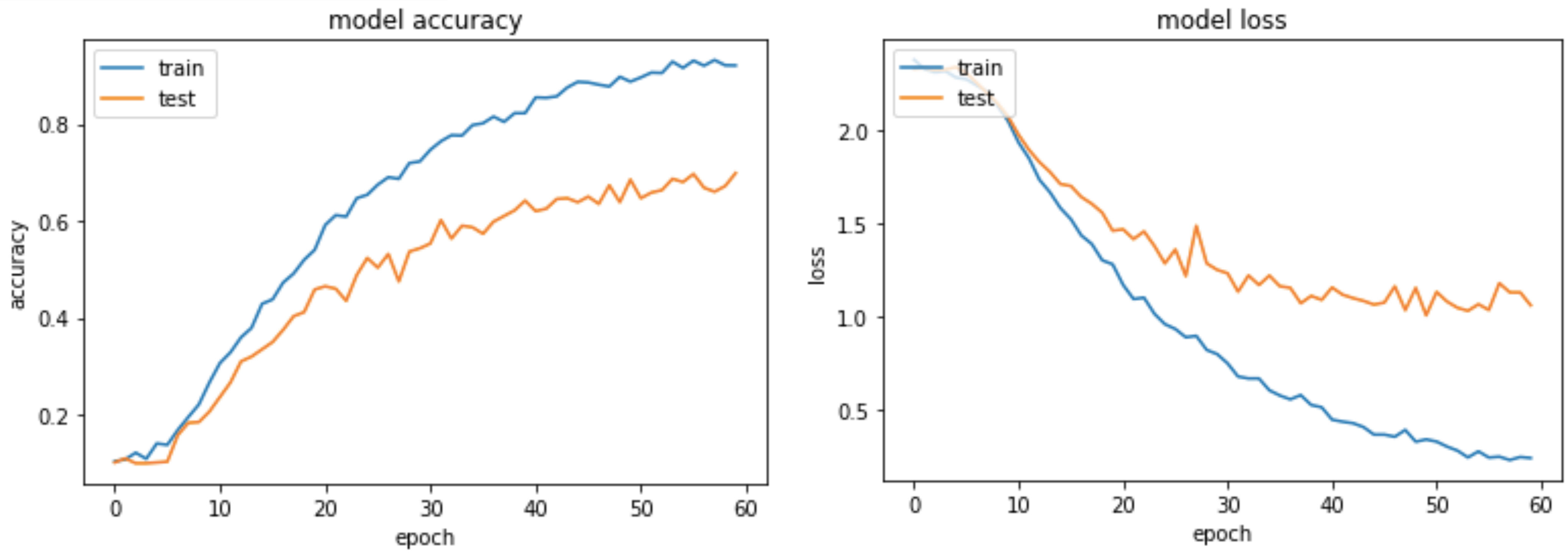
2



VGG-16 + LSTM

4) 실험 결과

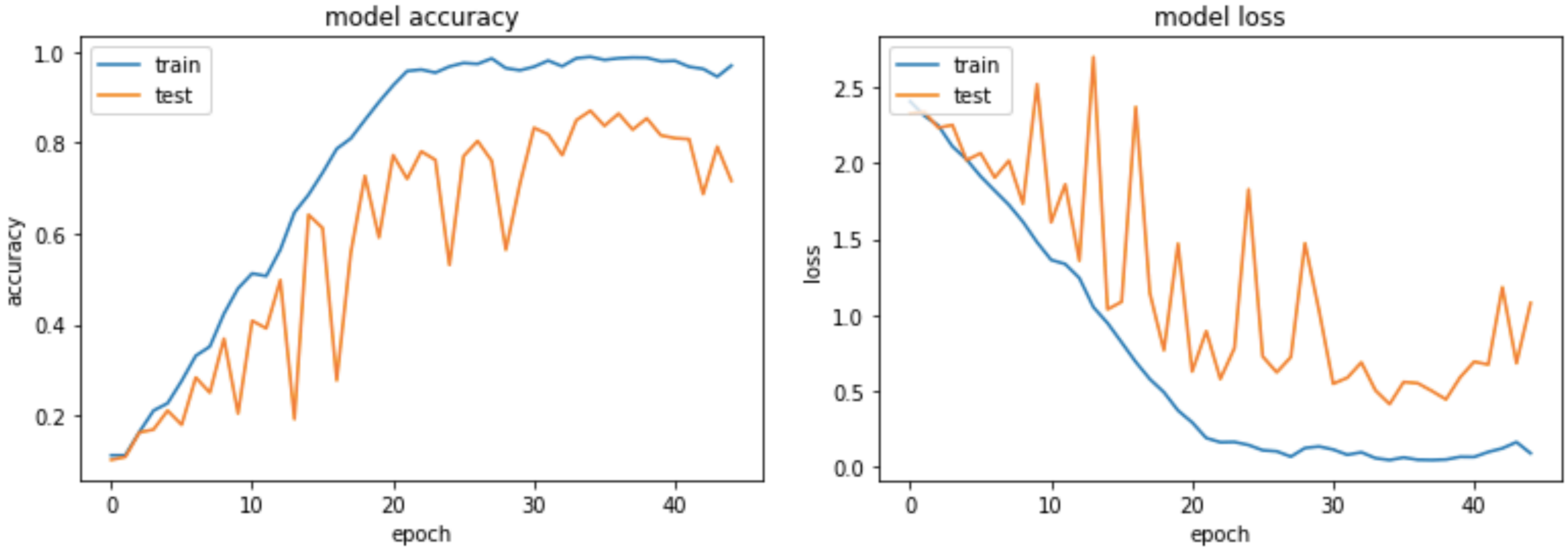
VGG-16 + LSTM 모델



영어 Data

4) 실험 결과

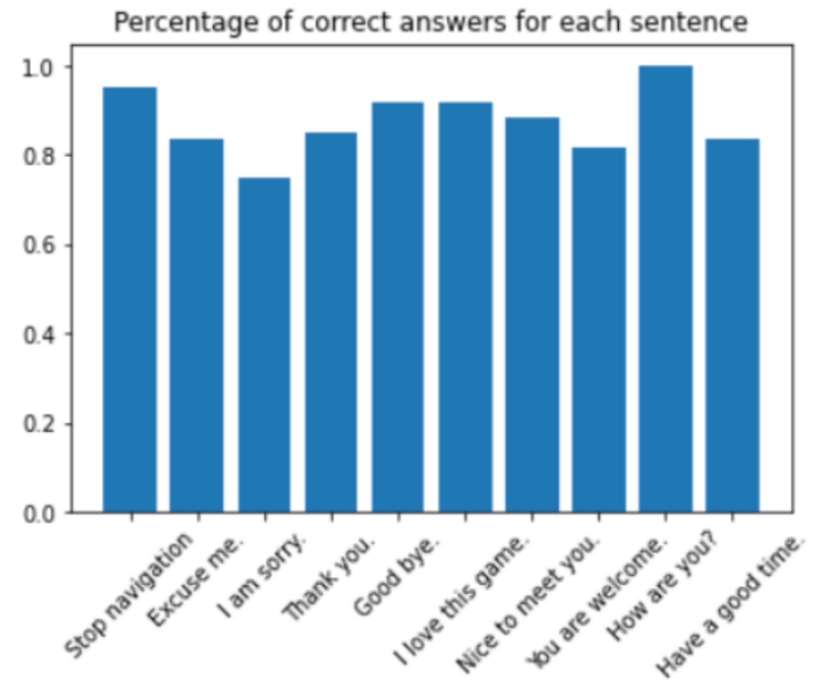
MobileNet + LSTM 모델



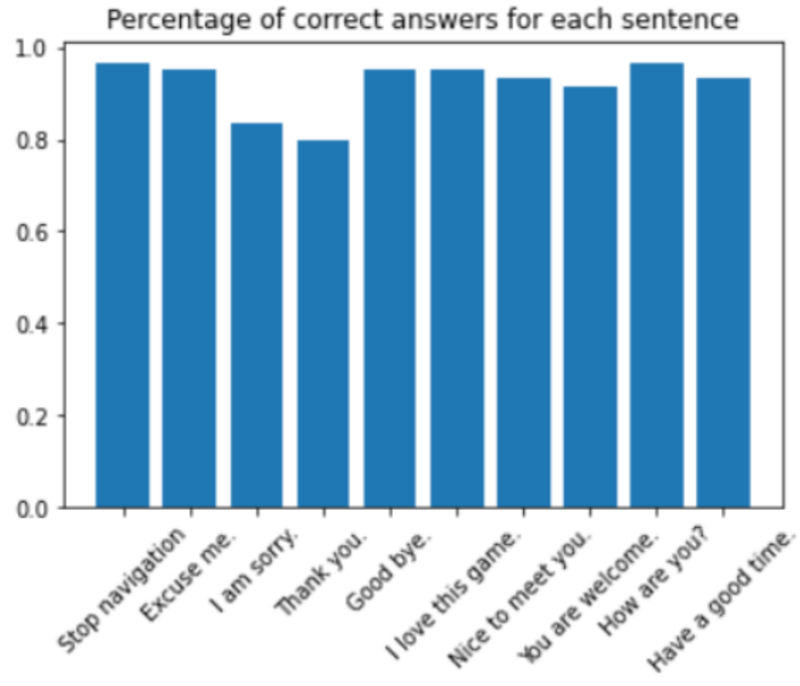
영어 Data

4) 실험 결과

두 모델 예측을 통한 문장 별 정답 비율



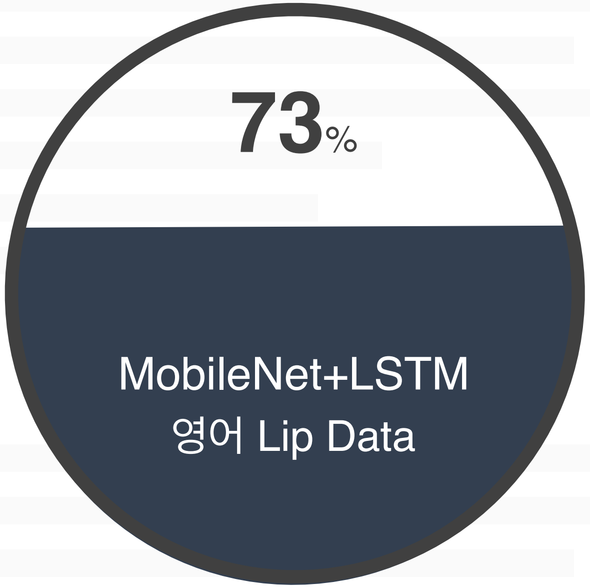
VGG-16 + LSTM



MobileNet + LSTM

영어 Data

4) 실험 결과



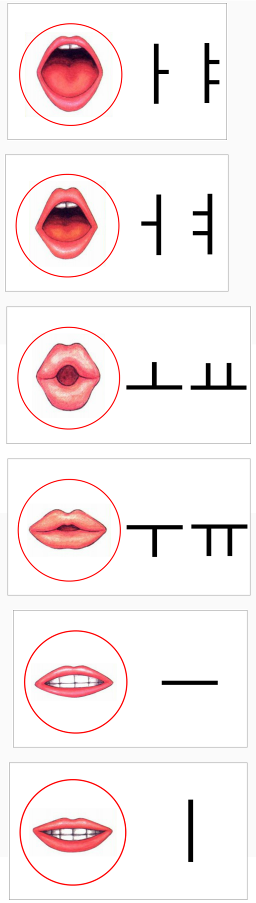
모델 test 결과

이전 논문 대비 20% 가량 높은 성능

3. 한국어 *Data*를 이용한 입 모양 인식

한국어 Data

1) DataSet 수집 방법



한글 모음 별 입모양

문장

- 도와주세요
- 힘내세요
- 누구세요
- 안녕하세요
- 조심히가세요
- 죄송합니다
- 감사합니다
- 좋아요
- 싫어요

입 모양이 좀 더 두드러지게 나타나는
모음 위주의 조합으로 문장을 구성

음절	코드	음절	코드
가	00	요	14
다	01	구	15
사	02	누	16
아	03	주	17
하	04	니	18
감	05	히	19
안	06	싫	20
합	07	심	21
어	08	힘	22
녕	09	게	23
도	10	내	24
조	11	세	25
송	12	와	26
좋	13	죄	27

문장의 길이와 말의 속도 차이로 발생하는
Frame 관련 문제점을 해결하기 위해 **한 음절 씩 수집**

한국어 Data

1) DataSet 수집 방법

앞말	끝말
도와주	세요
힘내	
누구	
안녕하	
안녕히계	
조심히가	
죄송	합니다
감사	
좋아	요
싫어	

- 문장의 끝 말은 반복
  Ex) ~세요, ~합니다, ~요 등등
  → 앞 말과 끝 말의 분리

- 이후 머신러닝을 통해 앞 말과 끝 말을 합친 후,
  추천 해주는 역할을 하는 모델 구현
  → 단순 classification 보다 성능↑

한국어 Data

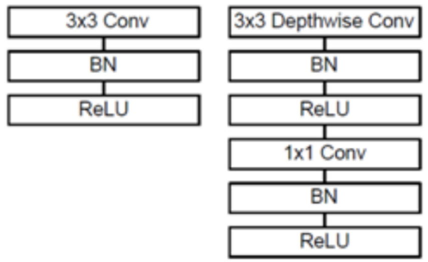
2) DataSet 전처리

입술 영역 검출 후 200X100 사이즈로 cropping

Augmentation으로 데이터 다양화, 증량

한국어 Data

3) 모델 설명



MobileNet + LSTM

영어 Dataset 훈련 과정에서 가장 높은 성능을 보인 MobileNet 사용

음절	코드	음절	코드
가	00	요	14
다	01	구	15
사	02	누	16
아	03	주	17
하	04	니	18
감	05	히	19
안	06	싫	20
합	07	심	21
어	08	힘	22
녕	09	계	23
도	10	내	24
조	11	세	25
송	12	와	26
좋	13	죄	27

수집한 28가지 음절 데이터

모음	단어	코드
ㅏ	하 가 사 아 다	00
ㅏㄴ	안	01
ㅏㅁ,ㅂ	감, 합	02
ㅓㅕ	녕 어	03
ㅗㅛ	도 조 송 좋 요	04
ㅜㅠ	주 누 구	05
ㅣ	히 싫 니	06
ㅣㅁ	힘 심	07
ㅐㅔㅖ	내 계 세	08
ㅘ	와	09
ㅚ	죄	10

유사한 모음과 받침을 묶어 11가지로 분류

3) 모델 설명

```python
def clustering(str,list1, list2):
    score = []
    for i in range(len(str_list)):
        ratio = SequenceMatcher(None, str, list1[i]).ratio()
        score.append(ratio)
    index = score.index(max(score))
    return list2[index]
```

```python
str = '오아'
clustering(str, str_list, answer_list)
```

'좋아'

```python
str = '오아오'
clustering(str, str_list, answer_list)
```

'좋아요'

```python
str = '오아암이'
clustering(str, str_list, answer_list)
```

'좋아합니다'

```python
str = '오아오오'
clustering(str, str_list, answer_list)
```

'좋아요'

```python
str = '암아'
clustering(str, str_list, answer_list)
```

'감사'

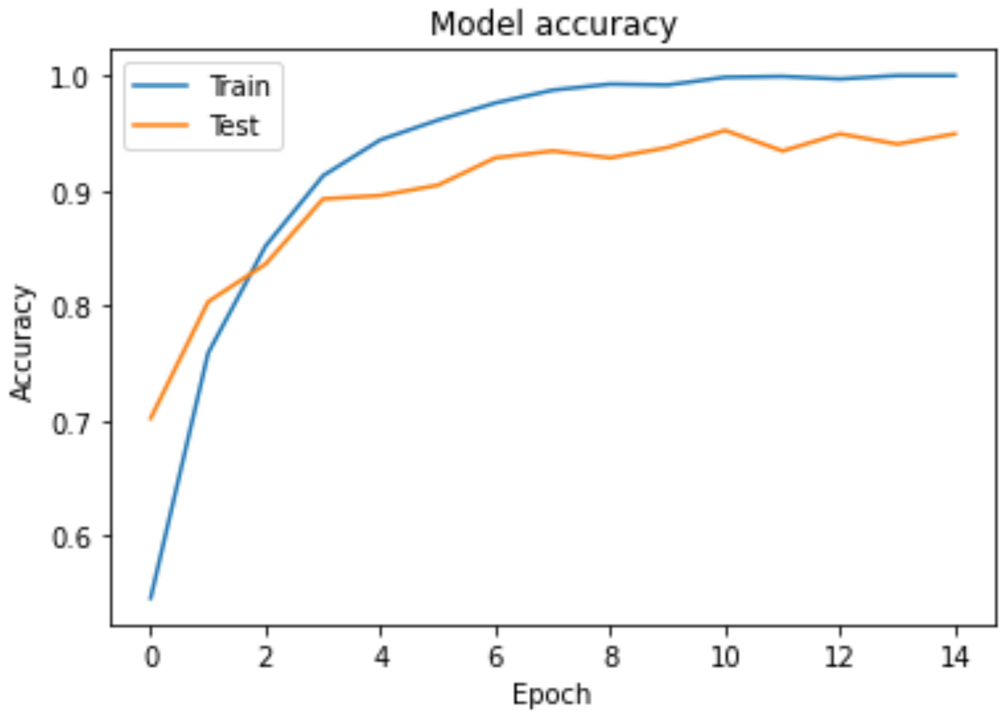
- difflib의 SequenceMatcher 함수 사용

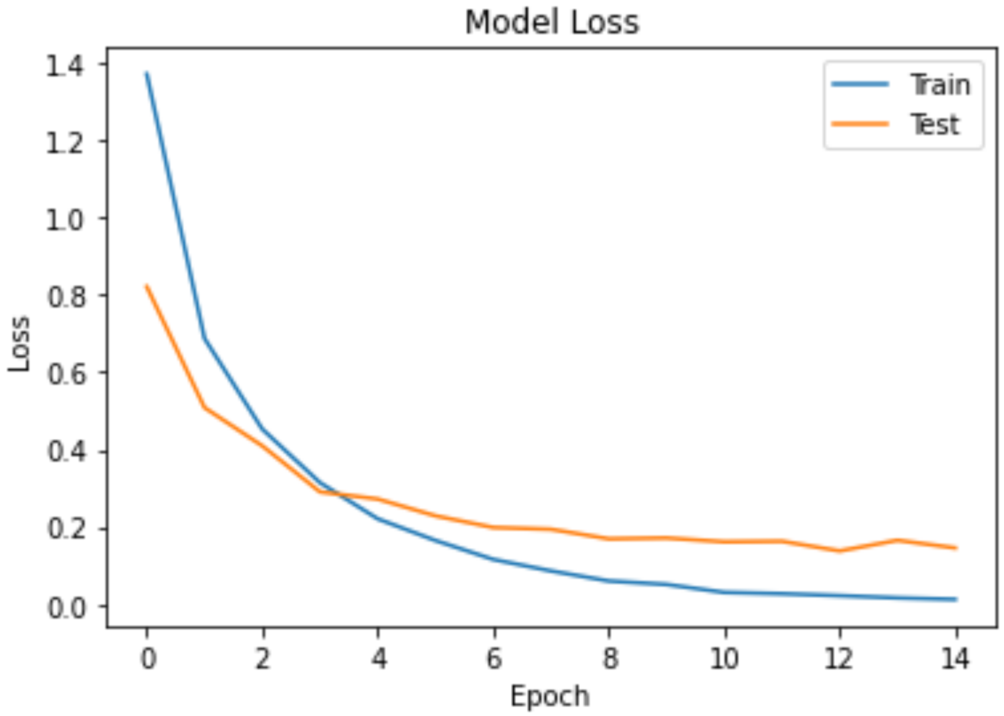
모음이 완전히 일치하지 않아도
가장 근접한 문장(혹은 단어)를 찾아내어 반환
앞말과 끝말의 연결을 통해 좋아, 좋아요, 좋아합니다
등 다양한 표현 가능

한국어 Data

4) 실험 결과

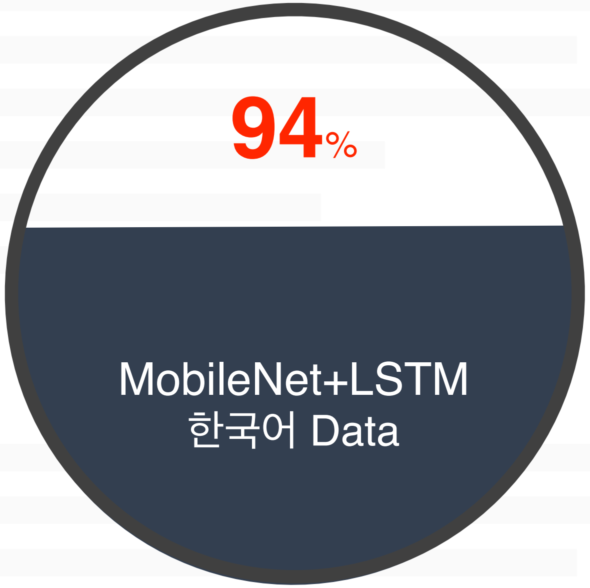
MobileNet + LSTM 모델 성능





한국어 Data

4) 실험 결과



모델 test 결과

기존에 존재하던 영어 Dataset을 이용한 연구보다

직접 수집한 한국어 Dataset을 이용한 연구에서 더 좋은 성능

4. 결론

결론

활용 방안 및 기대효과

Barrier free

비장애인과 장애인간의 원활한 의사소통

수화 통역, 사회복지 분야 활용

동영상 플랫폼에서의 자막 기능

발전 방향성

- 음성 인식 기반 시스템과 결합해 multimodal model로 제작
- KoNLPy를 사용한 형태소 분석, 문장 추천 model제작