MULTIMODAL SPEECH EMOTION RECOGNITION USING AUDIO AND TEXT

COMP8240 Group J Project Proposal

Jan Arvin Lapuz, Alyssa Lim, Le Van Nguyen, Ramil Zabala

September 1, 2020

» Overview

- 2018 IEEE Spoken Language Technology Workshop (SLT)
- Rank 14 - Computational Linguistics Category
  - Google Scholar
- Deep Dual Recurrent Encoder Model
  - audio signals & text data
  - speech data » signal level » language level
  - emotion categories:
    - angry
    - happy
    - sad
    - neutral
  - accuracy: 68.8% to 71.8% (IEMOCAP dataset)
- CNN Long Short-Term Memory Neural Networks
  - speech recognition
  - natural language processing

» Specifications

Requirements	python 2.7 nltk 3.3 scikit-learn 0.20.0 tensorflow 1.4 Google ASR system OpenSMILE toolkit
Input Features	Mel-Frequency Cepstral Coefficients (MFCCs) Prosodoic Features Word Tokens

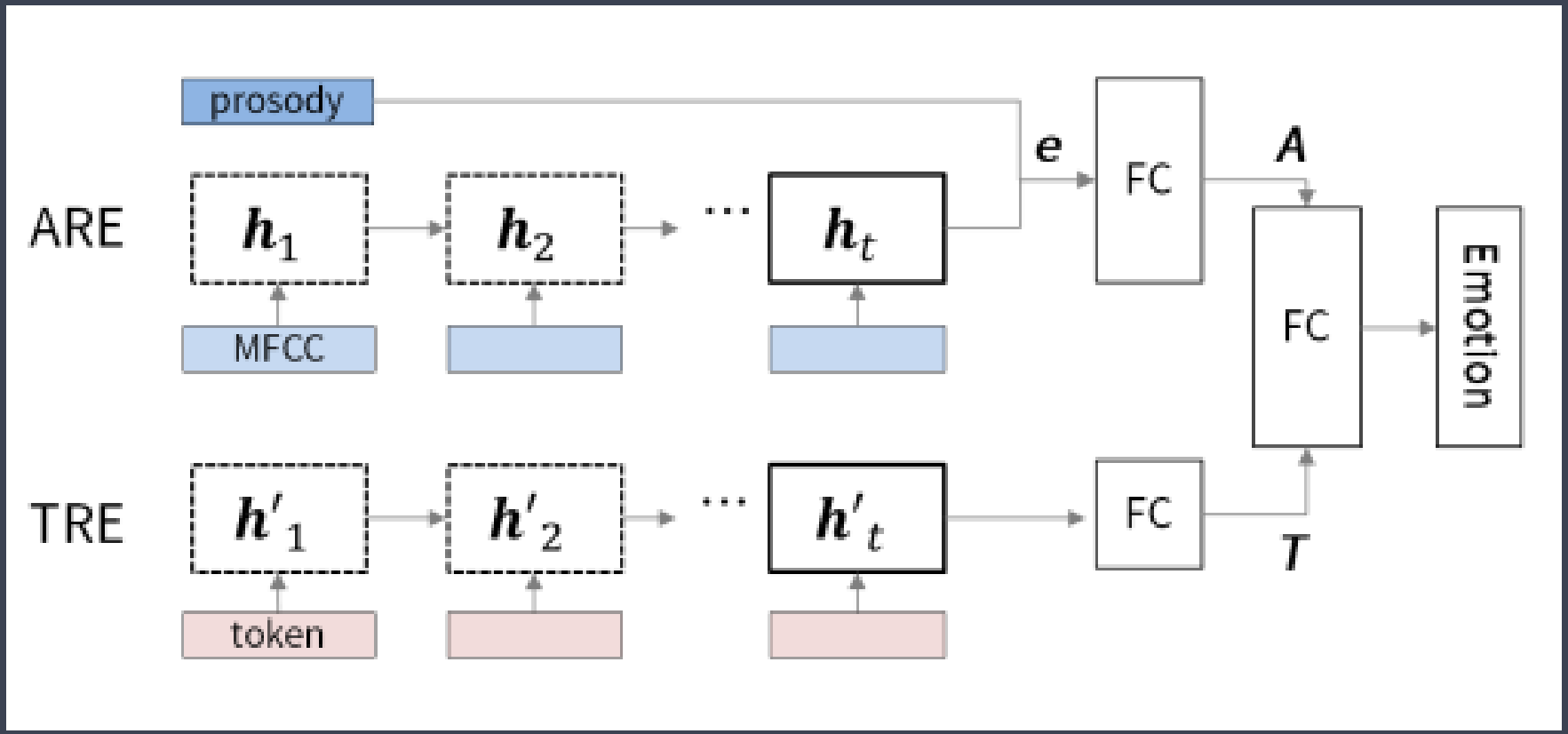
» Specifications

Source Code
: https://github.com/david-yoon/multimodal-speech-emotion

Original Datasets
: IEMOCAP dataset
ASR system processes IEMOCAP audio data

New Dataset
: * 25 new data per member
* existing dataset from other research of the same topic

» Model



Multimodal Dual Recurrent Encoder (MDRE)

» Research Results

Multimodal Speech Emotion Recognition Results

Model	WAP
ARE	0.546 ± 0.009
TRE	0.635 ± 0.018
MDRE	0.718 ± 0.019
MDREA	0.690 ± 0.019
TRE-ASR	0.593 ± 0.022
MDRE-ASR	0.691 ± 0.019
MDREA-ASR	0.677 ± 0.013

Novel Models' Results

Model	WAP
ACNN	0.561
LLD RNN-attn	0.635
RNN(prop.)-ELM	0.628
3CNN-LSTM10H	0.688