Multi-Modal Speech Emotion Recognition Using Speech Embeddings and Audio Features

** 들어가기 전에…

- 음성데이터 학습은 이렇게 하는구나
- speech2vec이 이렇게 생겼구나
- multi-modal (task) 모델이 이렇게 생겼구나

Review

0. Abstract

1. 모델 프로포절
    a. 2 modalities : audio & text(→speech embeddings)
    b. dense MLP projection layer
2. [성능 비교#1] ENC2에 word2vec VS **speech2vec**
3. [성능 비교#2] 기존 멀티모달 감정인식모델
4. 스피치 임베딩 dimension [50] - 작은거

1. Introduction

- SER는 발화에서 감정을 식별하는 태스크를 말한다.
- DNN(deep neural network)이 speech 관련 태스크에서 state-of-art 결과를 보인다.
    - speech recognition
    - speaker identification
    - language identification

- 최근에 등장한 인공신경망을 활용한 SER 모델은 HMM, SVM, 결정트리 기반 모델 등 전통적 머신러닝 모델보다 성능이 좋다.
- pretrained Alex-net 모델로 감정 식별 태스크를 위한 전이학습에 사용될 수 있다.
- speech feature를 모델에 넣을 때 text data를 guide signal로 함께 넣으면 모델 퍼포먼스가 크게 향상될 것이다.
    - face emotion features를 guide signal로

- **ENC1 : acoustic encoder (audio)**
    - Bi-Directional LSTM
    - input : sequence of speech features
        - MFCC, chroma-gram, zero-crossing rate, …
    - 마지막 time step에서 *hidden representation*을 생성한다.
- **ENC2 : speech embedding encoder**
    - Bi-Directional LSTM
    - input : sequence of word level speech embeddings (차원고정)
    - 마지막 time step에서 *hidden representation*을 생성한다.

- 동시에
    - ENC2 → 오디오 신호에서 의미 정보(semantic information) encoding
    - ENC1 → 오디오 신호에서 감정관련 특성(emotion-related features) encoding

- 이 때
    - speech embeddings는 인코더-디코더 프레임워크로 훈련된다.
    - 중심단어들은 MFCC 특성으로 전환되어 LSTM 인코더에 들어간다
    - 디코더 네트워크는 문맥단어를 위해 MFCC를 생성하고자 한다(mse loss 훈련)

- 분류 다중 퍼셉트론(MLP for classification)

2. Proposed Model

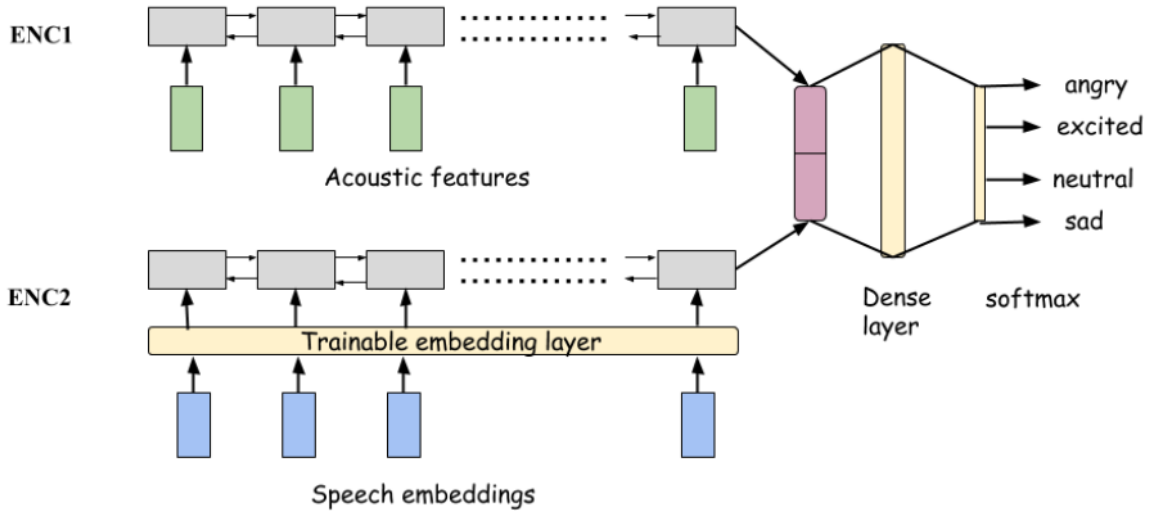

Figure 1: *Proposed multi-modal emotion recognition model*

1. **Acoustic feature extraction**
    a. 음성 주파수 & 음성 에너지 기반 특성
    b. 총 34 가지 (13 MFCC, 13 chromagram, 8 spectral)
        i. spectral features 종류 : zero crossing rate, short-term energy, short-term entropy of energy, spectral centroid and spread, spectral entropy, spectral flux, spectral roll-off.

▼ *MFCC란?*

음성의 고유한 특징을 표현하는 값

음성/음악 등 오디오 신호 처리 분야에서 널리 쓰이는 특징값(Feature) 중 하나인 **MFCC(Mel-Frequency Cepstral Coefficient)**

오디오 신호 → **FFT**(Fast Fourier Transform : 고속 푸리에 변환) 적용 → Spectrum → Mel Filter Bank 적용 → Mel Spectrum → Ceptral 분석 → MFCC

▼ *Chromagram?*

Chroma. pitch class 정보

크로마벡터는(chroma vector) 서양음악의 12 음계로 화성적(harmonic) 특징을 나타낸 것으로, 특정 시간에 존재하는 각 반음(semitone) 들의 에너지 정보를 12차원의 벡터값으로 표현한 것

c. *“compute feature vectors for every 100 ms windows with no overlap” [추가설명]*

d. 특성벡터 전체 수는 200으로 제한 (20초 음성 데이터)

2. **Speech embeddings**

a. 중심단어로 주변단어를 예측하도록 훈련된 skip-gram 모델로 생성된다.

b. Speech2Vec은 인코더-디코더 프레임워크로 만들어졌다. (RNN)

c. 인코더가 MFCC 특성을 high-level representation으로 인코딩하고, 여기서 만들어진 representation을 디코더가 주변단어로 다시 디코딩한다.

d. pretrained model

3. **Acoustic encoder(ENC1)**

a. 매 time step마다 input으로 음성특성값(34-dim acoustic feature)을 받는다.

b. 최종 h 값은 정방향과 역방향 h을 각각 구해서 concat한 값이다.

$$h_t^A = f(x_t^A, h_t^A{}_{-1}, \Theta_A)$$

- f = Bi-LSTM (ENC1)
  - 양방향 정보처리
  - gating mechanism : 어떤 정보가 메모리에 저장될지, 버려질지 조절함
- t = time step
- x = input sequence; acoustic feature
- h = hidden state representation
- Θ = parameter

4. **Speech embeddings encoder(ENC2)**

a. sequence speech embedding vector를 구한다.

b. 발화의 문맥 정보를 하기 위해 설계된 embedding이다.

i. speech embedding은 발화에서 의미정보와 구문정보를 획득하는데 중요한 요소.

ii. pretrained speech embedding layer를 Bi-LSTM 이전 단계에 추가한다.

iii. speech2vec 모델에서 speech embedding 생성 → 각 단어에 대한 fixed-dim embedding을 획득한다.

iv. embedding layer는 (1)학습가능(trainable)하며, (2)역전파 과정에서 Bi-LSTM 모델 파라미터와 함께 업데이트된다.

$$h_n^S = g(x_n^S, h_{n-1}^S, \Theta_S)$$

- g = Bi-LSTM (ENC2)
- x = input speech embedding from sequence
- n = total number of words in sequence
- h = hidden state representation
- Θ = parameter

5. **Speech embeddings and acoustic features for emotion recognition**
   a. ENC1의 마지막 타임스텝 hTA 과 ENC2의 마지막 타임스텝 hNS **fuse** (early fusion)
   b. fused feature의 경로 : dense layer → softmax → emotion label prediction

      ** fused feature에 들어있는 정보들

      - acoustic feature

      - speech embeddings에서 도출한 문맥단어의 의미정보

3. Dataset

- IEMOCAP 데이터셋 사용

4. Experimental Analysis

1. **Acoustic feature** based emotion recognition

   : 성능 별로다.

2. **Speech embedding** based emotion recognition

   : 워드임베딩 기반 모델이 스피치임베딩 기반 모델보다 조금 더 나은 성능을 보인다.

3. (1) + (2)

   : 내 모델 시스템 성능있다.

Table 1: *Uni-Modal emotion recognition models*

Model	Unweighted Accuracy
Acoustic feature only model	55.01%
Speech embedding only model	60.03%
Word embedding only model	60.68%

Table 2: *Performance (%) of single systems and their fusion on IEMOCAP dataset*

System	Unweighted Accuracy
Bi-LSTM	55.03%
E-Vector	57.25%
MCNN + LSTM	64.33%
E-vector + MCNN + LSTM	65.90%
Our system	**68.49**%

▼ 5. Conclusion

- 고안한 multi-modal SER 시스템의 성능 확인(2.59%)

: 오디오 & 스피치 임베딩에서 추출한 정보를 바탕으로 감정 클래스 분류

- 스피치임베딩 모델 성능이 워드임베딩 모델 성능보다 좋다

Terms

- Speech2Vec
  - word2vec의 speech 버전
    - skipgram, cbow에서 임베딩하는 방식을 그대로 채택하여 음성데이터를 임베딩
  - 사전에 녹음된 음성파일로 훈련한다
  - wave2vec
    - ASR(자동스피치인식)이 된다.
    - 자기지도학습을 통한 스피치 인식

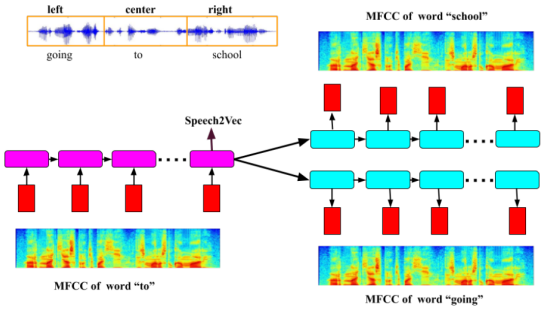

Figure 2: *Speech2Vec model*

        - “AI 훈련을 위한 데이터셋이 구축되어있지 않은 언어에 대한 스피치인식이 가능하다”

- Speech Emotion Recognition(SER)
    - “identifying the emotion of a speech utterance”

- Multi-modal SER
    - Multi-modal
        - 멀티 모달 : 여러가지 형태와 의미로 컴퓨터와 대화하는 환경
        - 사람과 컴퓨터를 연결하여 데이터를 수집 및 분석 → 행동분석, 감정분석이 가능
            - HCI 분야

+) Bi-LSTM : 두 개의 독립적인 LSTM 아키텍처를 함께 사용하는 구조 (예시)

Citations

https://www.isca-speech.org/archive/pdfs/avsp_2019/n19_avsp.pdf