Speech Emotion Recognition

Tianheng Wang
Yuehan Cai

Syllabus

- Introduction
- Dataset
- Feature Extraction
- Classification
- Results
- Conclusion
- Reference

1 Introduction

Emotion Recognition

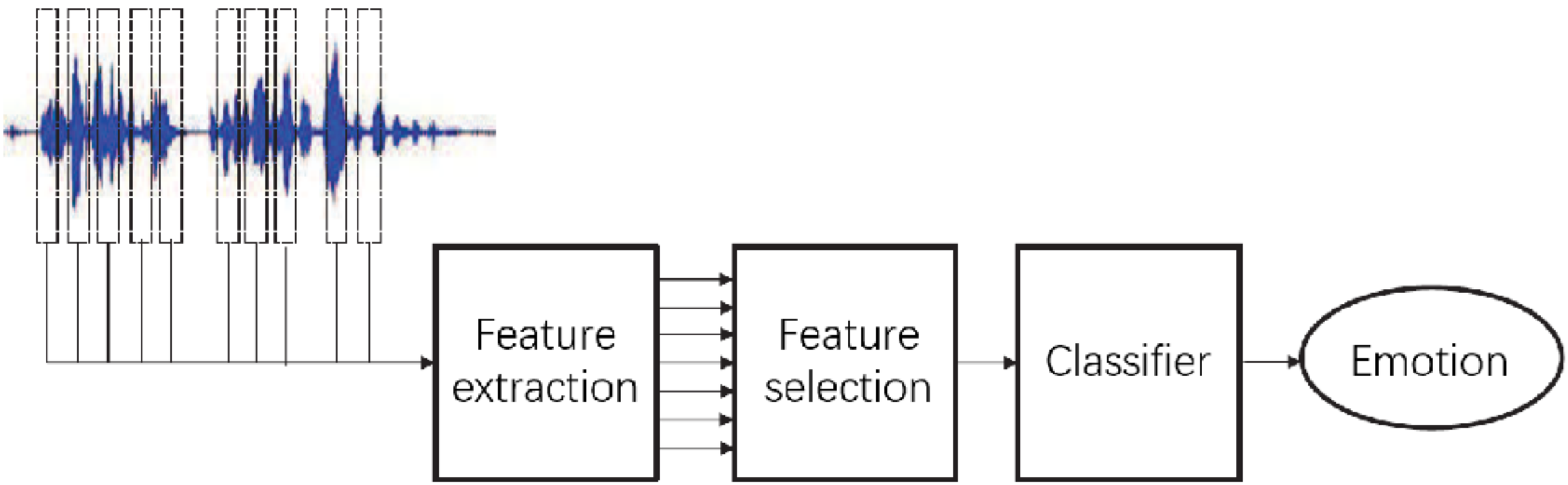
- Classify method: fear, surprised, happy, neutral, sad, angry, disgust.
- Active topic in human-computer interaction

Main Idea

- Extract features from human speech audio
- Preprocessing and select features
- Use machine learning network to recognition

1 Introduction

Framework



2 Dataset

- **RAVDESS**: Ryerson Audio-Visual Database of Emotional Speech and Song
  24 actors (M&F) reading 2 sentences;
  1,248 recordings in total.

- **TESS**: Toronto Emotional Speech Set
  2 actors (F) reading different words;
  1,370 recordings selected.

- 7 emotions: neutral, happy, sad, angry, fear, surprised and disgust.

- Sample rate and number of channels should be unified.

Feature Extraction

- Lots of parameters for emotional characteristics

- Calculated short-term features: pitch, energy, MFCC, ZCR, spectral centroid (and gender)
  Feature Matrix: audioNum * featureNum * windowNum

- Implemented with MATLAB API and LibROSA.
  BUT matlab engine was 100 times slower!

- Trimmed or zero-padded to a fixed length

4 Classification

- SVM
- Pre-trained network (vgg16…)
- Convolutional neural network (CNN)

Pre-processing

- Normalize: Center to the mean and scale to unit variance for each feature
- One hot coding: encode label
  select the class with max probability

SVM

- Support Vector Machine
- non-linear classifier by transforming the input feature vectors into a higher dimensional feature space using a kernel mapping function.

- SVC: based on libsvm, use one versus one method to solve e multi-class problem

VGG16

- Many pre-trained networks in Keras
- Deep learning network used for Image but also can extract features
- Slow, no better performance

CNN

Layer (type)	Output Shape	Param #
batch_normalization_1 (Batch	(None, 56, 57, 1)	4
conv2d_1 (Conv2D)	(None, 54, 55, 32)	320
max_pooling2d_1 (MaxPooling2	(None, 27, 27, 32)	0
dropout_1 (Dropout)	(None, 27, 27, 32)	0
conv2d_2 (Conv2D)	(None, 25, 25, 32)	9248
max_pooling2d_2 (MaxPooling2	(None, 12, 12, 32)	0
dropout_2 (Dropout)	(None, 12, 12, 32)	0
flatten_1 (Flatten)	(None, 4608)	0
dense_1 (Dense)	(None, 256)	1179904
dropout_3 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 7)	1799

Total params: 1,191,275
Trainable params: 1,191,273
Non-trainable params: 2

Feature Selection Method

- Use SVM: fast and high performance
- Important features: MFCC, pitch, gender

5 Results

- SVM: 88.5% (20 MFCCs)
- Pre-trained network: bad performance (~66%)
- CNN: around 84% (50 MFCCs)
- SVM and CNN are suitable for emotion recognition with higher accuracy and faster operation

6 Conclusion

- Use both prosodic and spectral features to realize analysis on emotion with accuracy over 80%
- Dataset small and not random speech
  TESS: 99%, RAVDESS: 64%
- Difficulty in features determine
- More work needed for real-time recognition, try LSTM

7 Reference

- S. R. Livingstone and F. A. Russo, “The Ryerson audio-visual database of emotional speech and song (RAVDESS),” in *Zenodo*, 2018. [Online]. Available: https://zenodo.org/record/1188976. [Accessed May 8, 2018].
- K. Dupuis and M. K. Pichora-Fuller, “Toronto emotional speech set (TESS),” in *TSpace Repository*, 2010. [Online]. Available: http://cie.ed.asu.edu/volume6/number12/. [Accessed May 8, 2018].
- B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, O. Nietok, “librosa: audio and music signal analysis in Python,” in *Proceedings of the 14th Python in Science Conferences (SciPy)*, Jul. 2015, pp. 18–25.
- Y.-L. Lin and G. Wei, “Speech emotion recognition based on HMM and SVM,” in *2005 International Conference on Machine Learning and Cybernetics*, Aug. 2005, pp. 4898–4901.
- T. Seehapoch, S. Wongthanavasu, “Speech emotion recognition using Support Vector Machines,” in *5th International Conference on Knowledge and Smart Technology (KST)*, Feb. 2013, pp. 86–91.
- https://github.com/coreyker/dnn-mgr/blob/master/train_classifier_on_dnn_feats.py
- https://github.com/tuwien-musicir/DL_MIR_Tutorial/blob/master/Music_genre_classification.ipynb
- https://github.com/sdrangan/introml/blob/c2a15e8c2f2e381979d465c1e9f62a6c6967bf84/cnn/vgg16.ipynb

Thank you !