What is the state of the art in speech emotion recognition

Letian Li

M.Sc. in Artificial Intelligence and Machine Learning
Research Intern & Software Development Engineer

- **SER Introduction**
- **State of the Art Models**
- **Popular Datasets**
- **My Implementation**

SER Introduction

What is the state of the art in speech emotion recognition?

Speech Emotion Recognition

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech.

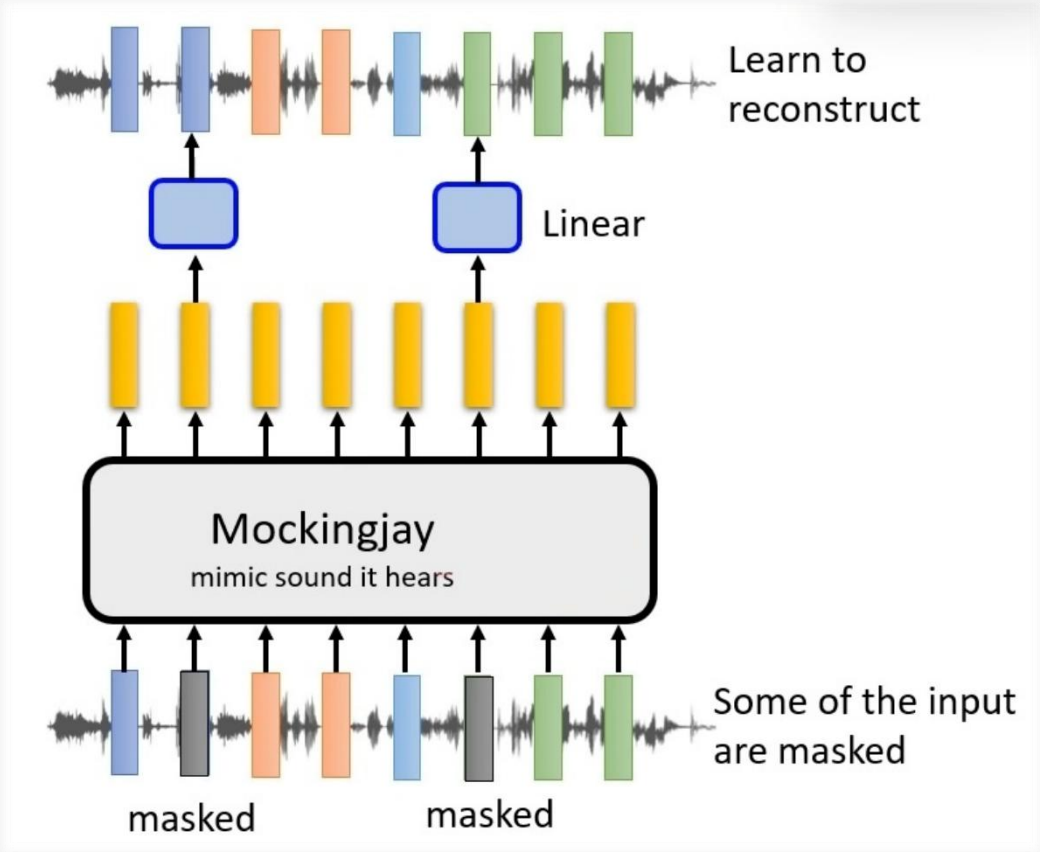
Nowadays Solution

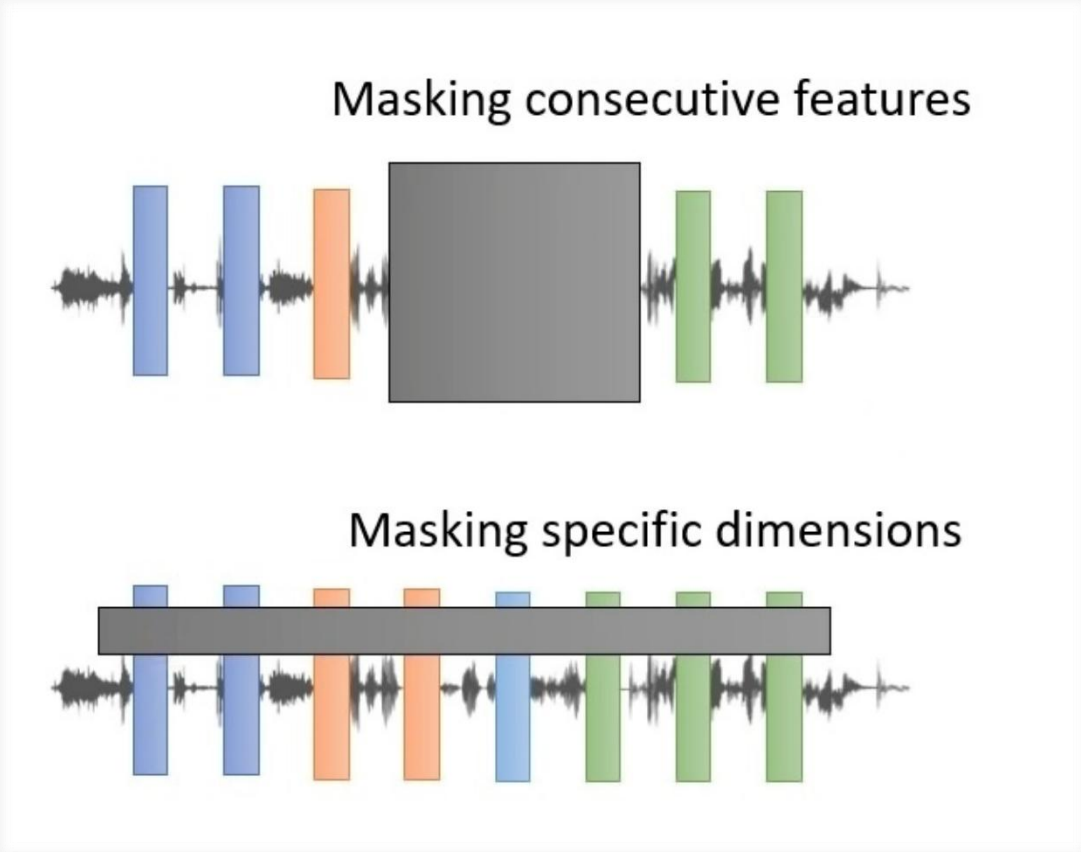
- **Neural Network**
- **End to end model**
- **Self-supervised learning**
  - Pre-trained on large amounts of unlabeled audio data.
  - Effectively capture knowledge of audio before implementing the model into downstream tasks.
  - Self-supervised approaches can usually outperform traditional speech recognition systems.

State of the Art Models

What is the state of the art in speech emotion recognition?

The main idea of self-supervised learning from unlabeled audio data





State of the Art Models

What is the state of the art in speech emotion recognition?

Some examples of self-supervised learning speech models

Categories of self-supervised learning	Speech Models
Generative	Mockingjay[1], APC[2]
Predictive	HuBERT[3]
Contrastive	CPC[4], Wav2vec series[5][6]
Bootstrapping	Data2vec[7]
Regularization	DeLoRes[8]

[1] Liu, Andy T., et al. "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020. https://ieeexplore.ieee.org/abstract/document/9054458.
[2] Chung, Yu-An, and James Glass. "Generative pre-training for speech with autoregressive predictive coding." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020. https://ieeexplore.ieee.org/abstract/document/9054438.
[3] Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 3451-3460. https://ieeexplore.ieee.org/abstract/document/9585401.
[4] Van den Oord, Aaron, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv e-prints (2018): arXiv-1807. https://ui.adsabs.harvard.edu/abs/2018arXiv180703748V/abstract.
[5] Schneider, Steffen, et al. "wav2vec: Unsupervised pre-training for speech recognition." arXiv preprint arXiv:1904.05862 (2019). https://arxiv.org/abs/1904.05862.
[6] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in Neural Information Processing Systems 33 (2020): 12449-12460. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.
[7] Baevski, Alexei, et al. "Data2vec: A general framework for self-supervised learning in speech, vision and language." arXiv preprint arXiv:2202.03555 (2022). https://arxiv.org/abs/2202.03555.
[8] Ghosh, Sreyan, Ashish Seth, and S. Umesh. "Delores: Decorrelating latent spaces for low-resource audio representation learning." arXiv preprint arXiv:2203.13628 (2022). https://arxiv.org/abs/2203.13628.

Popular Datasets

What is the state of the art in speech emotion recognition?

IEMOCAP *https://sail.usc.edu/iemocap/*

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an acted, multimodal and multispeaker database, recently collected at SAIL lab at USC. It contains approximately 12 hours of audiovisual data, including video, speech, motion capture of face, text transcriptions.

RAVDESS *https://smartlaboratory.org/ravdess/*

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. The database contains 24 professional actors (12 female, 12 male), vocalizing in 8 different emotions.

SAVEE *http://kahlan.eps.surrey.ac.uk/savee/*

Surrey Audio-Visual Expressed Emotion (SAVEE) database has been recorded as a pre-requisite for the development of an automatic emotion recognition system. The database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total.

My Implementation: *https://github.com/LetianLee/Speech-Emotion-Recognition*

What is the state of the art in speech emotion recognition?

- **Model Selection:**
  - **HuBERT** (The model I used has already been fine-tuned on IEMOCAP dataset.)
- **Deep Learning Framework:**
  - **PyTorch and HuggingFace**
- **Dataset:**
  - **RAVDESS**

Implementation Information:

Training Data: 691

Training Epoch: 3

Training Time: 8 min

Training Accuracy: 58.5% - 72.5% - 75.5%

Test Data: 173

Wrong Predition: 43

Test Accuracy: 75.14%