Detecting Escalation Level from Speech with Transfer Learning and Acoustic-Lexical Information Fusion

Ziang Zhou Yanze Xu Ming Li

Duke Kunshan University

ziang.zhou372@dukekunshan.edu.cn

Natural Language Processing Workshop
April 22, 2022

Content

1. Introduction

Escalation Detection

From Speech

Escalation refers to the conflict elevating process in the middle of human-to-human conversations, which can be in the form of speech and text. Escalation Detection Task is a paralinguistic challenge that aims to respond to such scenarios and pre-alert the administrators to take precautions.

Traditional escalation detection tasks heavily relied on the overlap detection of human conversations. Statistical analysis and Machine Learning methods have been applied to the overlapping parts of conversation only, leaving out the rest of the conversation.

Escalation Detection

From Speech

Speech with no overlap can also contain valuable information, including

- Semantic information
- Acoustic patterns that may indicate escalation

Escalation Detection

From Text

Textual escalation detection has been widely applied to the customer service systems of e-commerce companies' to alert and prevent potential conflicts in advance.

Once an increasing escalation level of the customers is detected, special agents will take over and settle the dissatisfied customers. This mechanism can forestall potential conflict from worsening and protects the feelings of their customer service employees.

Escalation Detection

Datasets

Both present unscripted interactions between actors, where friction appears as they spontaneously react to each other based on short scenario descriptions. The transcriptions are manually annotated afterward.

- **TR** (*Lefter et al. 2013*): Dataset of Aggression in Trains. Consists of 21 scenarios of unwanted behaviours in trains and train stations (e. g., harassment, theft, travelling without a ticket) played by 13 subjects.
- **SD** (*Lefter ey al. 2014*): Stress at Service Desk Dataset. Contains scenarios of problematic interactions situated at a service desk (e. g., a slow and incompetent employee while the customer has an urgent request) from 8 subjects.

Challenge: Total duration less that 30 minutes.

Transfer Learning

Motivation: Gideon et al. demonstrate that emotion recognition tasks can benefit from advanced representations learned from paralinguistic tasks (*Gideon et al. 2017*). This implies that emotional representation and paralinguistic features are interconnected to some degree.

Assumption: Emotional recognition tasks may as well benefit the escalation detection task.

There are many well-annotated speech corpora in emotion recognition; thus, we expect to raise the performance of the escalation task by applying transfer learning on speech emotion datasets.

Textual Embeddings

Motivations

- Datasets contain transcribed conversation scripts in Dutch. The semantic meaning of these scripts can be indicators of potential escalation.
- Escalation can be inferred from textual modality as well.

Challenges

Recordings are very short, mostly ranging from 3-7 seconds, thus scripts are often incomplete.

Contributions

1. The first work demonstrates that paralinguistic tasks, such as escalation detection can benefit from advanced emotional representations learned from speech emotion datasets.
2. Proposed a pipeline for escalation level detection under extremely low resource restrictions.

2. Methodology

Pipeline Overview

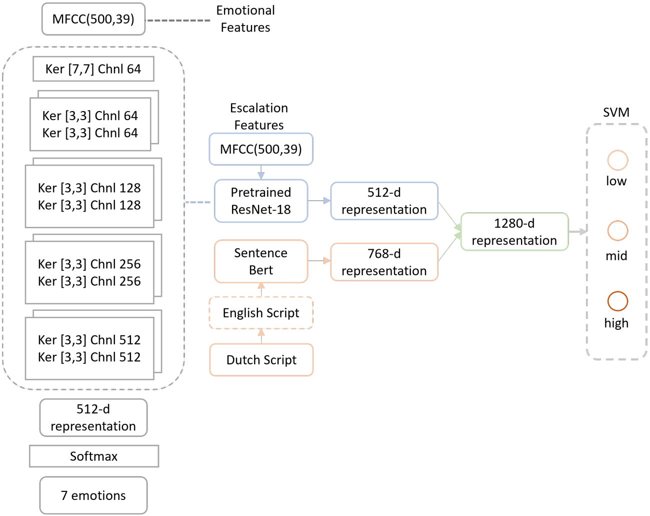

Figure: Pipeline of Escalation Level Detection System

Pretrain Speech Emotion Classifier

Datasets

We selected four well-annotated speech emotional datasets for joint sentimental analysis.

- **RAVDESS** (*Livingstone et al. 2018*): A gender balanced multimodal dataset with 7356 pieces of data.
- **CREMA-D** (*Cao et al. 2014*): A high quality visual-vocal dataset, containing 7442 recordings from 91 professional actors.
- **SAVEE** (*Fayek et al. 2015*): A male-only audio dataset.
- **TESS** (*Dupuis et al. 2010*): A female-only audio dataset.

Eventually, we gathered 2167 samples for `Angry`, `Happy` and `Sad` emotions each; 1795 samples for `Neutral`; 2047 samples for `Fearful`; 1863 samples for `Disgusted` and 593 samples for `Surprised` emotion.

Pretrain Speech Emotion Classifier

Features

Acoustic Features

Mel-frequency cepstral coefficient (MFCC) is one of the most common acoustic features. We vectorize the emotional audios by extracting their MFCC. The signal is first pre-emphasized with a coefficient of 0.97. The *winlen* of each frame is set to 0.025, the *winstep* parameter is set to 0.01; the window function is *hamming* function; the *nfilt* is set to 256. The frequency range is set from 50Hz to 8000Hz.

Finetune on Escalation Datasets

Voice Activity Detection (VAD)

WebRTC-VAD[1]: Filter out the unvoiced segments in the audios from temporal domain.

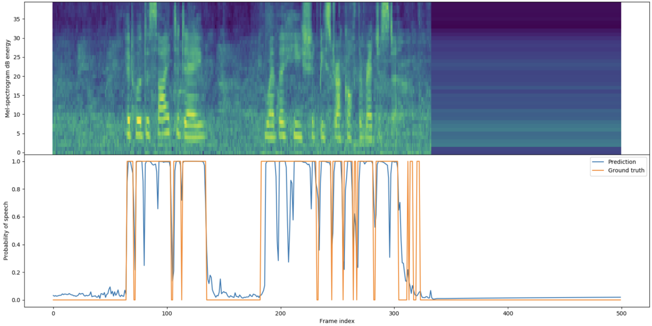

Figure: Voice Activity Detection of WebRTC

[1] https://webrtc.org/

Finetune on Escalation Datasets

Features

We apply the **WebRTC-VAD** toolkit prior to feature extraction for the escalation datasets.
Acoustic Features: MFCC (512-d)
Linguistic Features: Sentence-BERT (768-d)
Concat Features: 1280-d

Multi-lingual Sentence BERT

Reimers et al. 2020

Goal: Various input length, fix-sized output dense vector.

Algorithm Steps:

1. Tokenize input sentence
2. Use transformer like BERT to produce contextualized word embeddings for all input tokens.
3. Apply mean pooling to all word embeddings
4. Obtain fix-sized sentence embeddings.

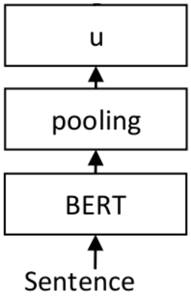

Figure: Architecture

$\text{BERT}_{\text{Base}}$ uses 12 layers of transformers block with a hidden size of 768. Thus the output size of our sentence embedding is 768-d (*Devlin et al. 2018*).

3. Experiments

Model Setup

Frontend Encoder

The pretraining step on speech emotion datasets shares same **ResNet-18** architecture with escalation finetuning.

- **Optimizer**: Stochastic Gradient Descent (SGD), nesterov momentum 0.8
- **Loss Function**: Cross Entropy Loss

$$\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

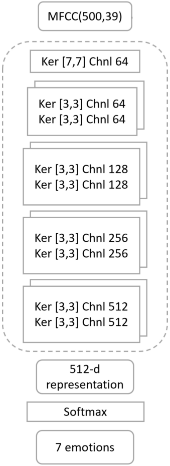

Figure: Architecture

Model Setup

Backend Classifier

Our work did not construct an end-to-end detection system. Instead, we employed Support Vector Machine (SVM) to conduct the backend classification task.

Motivation: Previous work under low resource restrictions has shown that simply replacing fully connected layers with linear SVMs can improve classification performance on multiple image classification tasks (*Tang 2015*).

Evaluation Metric

UAR: Unweighted Average Recall

In multiclass identification tasks, UAR calculates the arithmetic mean of the `recall` scores of each class.

Calculation

$$Recall = \frac{TP}{TP + FN}$$

$$UAR = \frac{\sum_{c=1}^{N} R_c}{N}$$

where R_c stands for the `Recall` score of class c, and N stands for the number of classes.

Experiment Results

VAD

Table: Effects of Voice Activity Detection (VAD) on the devel set. **TE**: Textual Embeddings fused.

Model Name	Precision	UAR	F1-Score
MFCC	0.640	0.675	0.647
MFCC+VAD	0.675	0.710	0.688
MFCC+TE	0.652	0.690	0.664
MFCC+VAD+TE	0.676	0.721	**0.691**
Baseline Fusion	-	**0.722**	-

Experiment Results

pre-trained Models

Table: Effects of fine-tune of pre-trained ResNet-18 on devel set. **PR**: Pre-trained ResNet-18 applied.

Model Name	Precision	UAR	F1-Score
MFCC+VAD	0.675	0.710	0.688
MFCC+VAD+PR	0.807	**0.810**	0.788
MFCC+VAD+PR+TE	0.807	**0.810**	0.788
Baseline Fusion	-	0.722	-

Experiment Results

Extra Attempts

Table: Extra Experiments. **LS**: Label Smoothing.

Model Name	Precision	UAR	F1-Score
Logfbank	0.670	0.743	0.684
Logfbank+VAD	0.711	0.778	0.733
MFCC+VAD+PR+LS	0.781	**0.781**	**0.761**
MFCC+VAD+ResNet-9	0.727	0.749	0.725
Baseline Fusion	-	0.722	-

Experiment Results

Model Fusion

To further leverage the model performance, we attempted three fusion strategies on the best three models, which are *MFCC+VAD+PR*, *MFCC+VAD+PR+LS*, *Logfbank+VAD*. The results are shown as follow.

Table: Model Fusion

Fusion Strategy	Precision	UAR	F1-Score
Concatenate	0.783	0.800	0.779
Mean	0.789	0.805	0.789
Voting	0.810	**0.815**	**0.803**
Baseline Fusion	-	0.722	-

4. Conclusion

Conclusion

1. The multimodal pipeline we proposed for escalation level detection tasks under extremely low resource restrictions is effective.
2. Validated that paralinguistic tasks, such as escalation detection, can benefit from advanced representations in speech emotion recognition tasks.

Acknowledgements

SMIIP Lab

- Ziang Zhou
- Yanze Xu
- Ming Li

Funding

- National Natural Science Foundation of China

The End

Questions? Comments?