12th International Young Scientists Conference on Computational Science

Multimodal prediction of profanity based on speech analysis

Ivan Smirnov, Anastasia Laushkina
ITMO University

Introduction

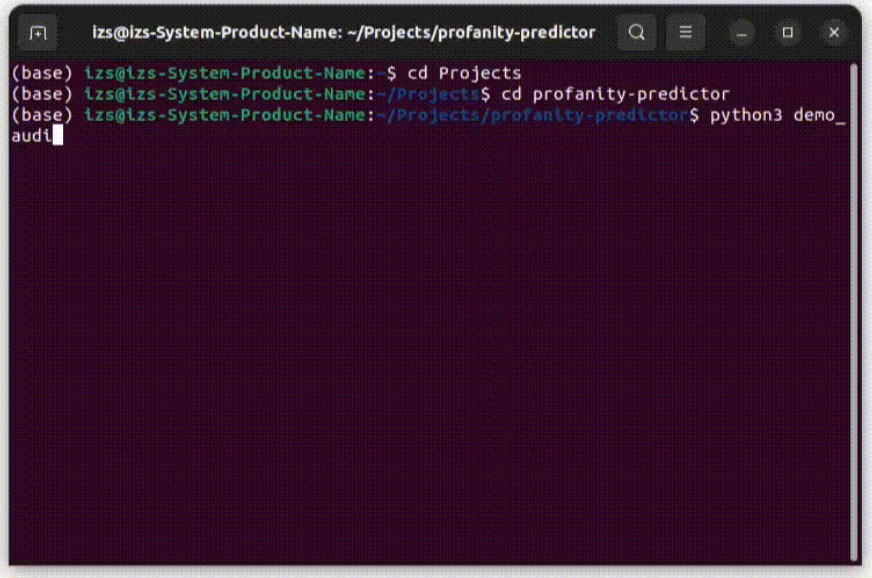
Many profane words have strong connotations and can be offensive; they may be used to provoke confrontations and even violence.

- Profanity affect mental health
- The information space is becoming insecure
- The content of speech broadcasts may be unexpected

The ability to predict profanity allows us to prevent it.

Relevance

Speech prediction task

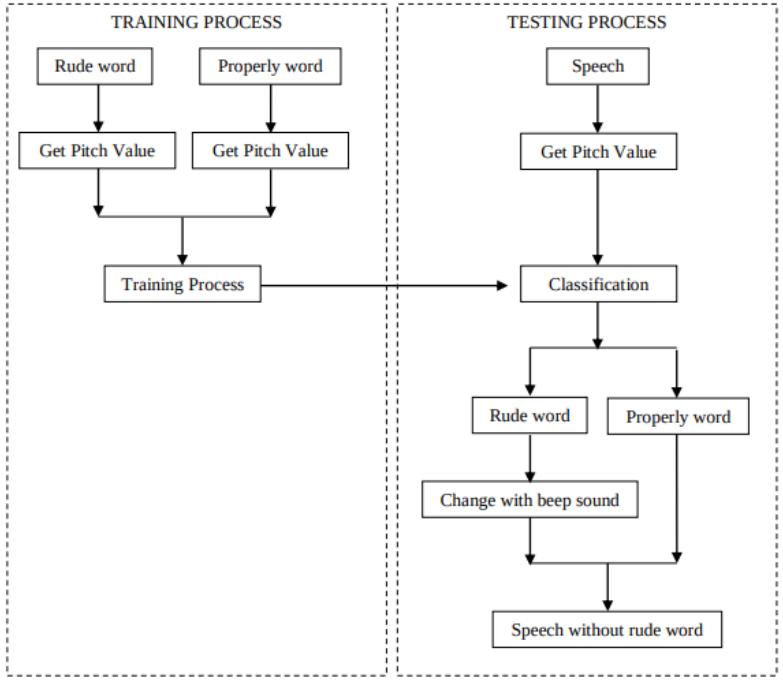


Realms of application

- Automatic censoring during broadcasts
- Moderation of audio messages
- Call center moderation

Actual methods

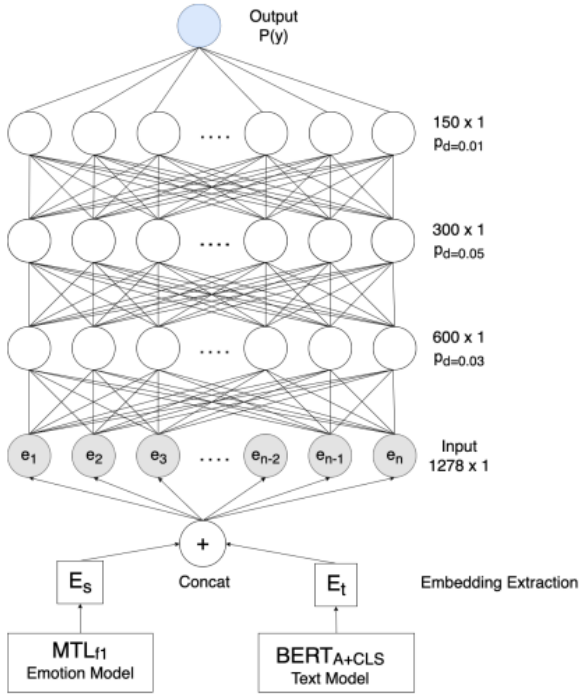
Word frequency matching and classification with Support Vector Machine



Disadvantages:

- Method does not consider sequences in dynamic
- Classification use only one feature statistic information for it

Combining of the text model and the audio emotion classifier

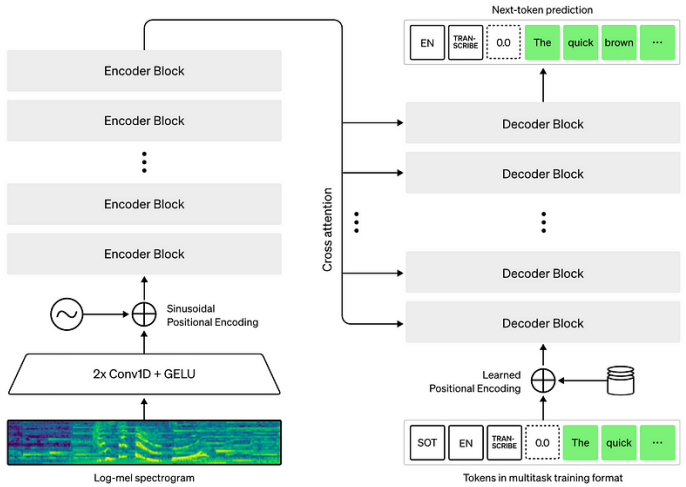


Disadvantages:

- The quality of recognition will vary depending on the quality of the recording
- The proposed models are resource-intensive

Problems

We wanted people to know that ~~how~~ to me where i know and essentially this product is uh what we call scripted changes the way that people are rapid technology.



01 **Speech can be transcribed incorrectly**

02 **ASR models are computationally complex**

Limitations for data

It is necessary to follow the next criteria:

1. **Authenticity:** The audio and video recordings should be authentic and not scripted.
2. **Diversity**: The recordings should come from a diverse set of speakers and contexts
3. **Quality**: The audio and video recordings should be of suitable quality

Dataset

The main problems:

- Unbalanced classes
- Not every “toxic” includes profanity
- The lack of timestamps

Statistics of annotated English datasets after balancing

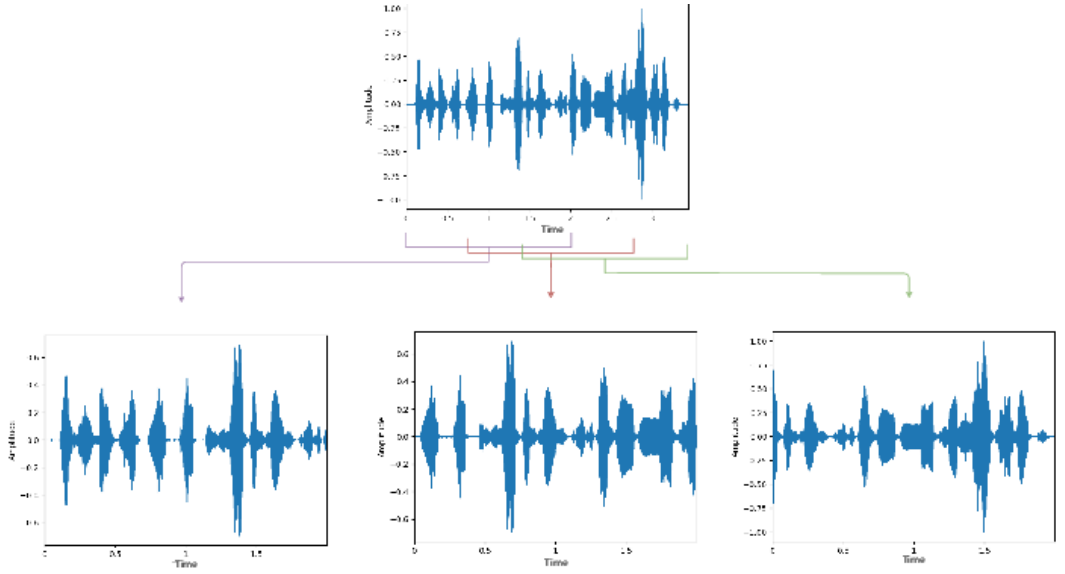
Dataset	Utterances (num)	Toxic (num)	Non-Toxic (num)	Total Duration (hh:mm:ss)
CMU-MOSEI	860	217	643	1:44:25
CMU-MOSI	260	67	193	0:18:03
Common Voice	11,551	2,888	8,663	12:38:17
IEMOCAP	1,090	274	816	1:19:26
LJ Speech	148	40	108	0:14:57
MELD	565	142	423	0:31:05
MSP-Improv	523	129	394	0:36:32
MSP-Podcast	2,772	692	2,080	4:01:57
Social-IQ	479	122	357	0:36:40
Switchboard	1,824	456	1,368	2:28:57
VCTK	199	50	149	0:08:51
Total	**20,271**	**5,077**	**15,194**	**24:39:10**

Data augmentation

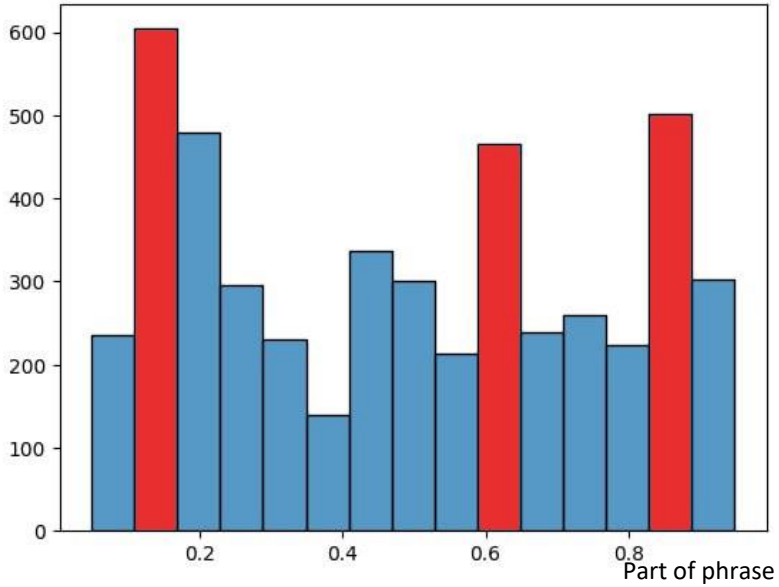
To increase the number of records, the next audio distributions were used with some probabilities:

- Gaussian noise
- Pitch shift
- Time stretch
- High pass filter

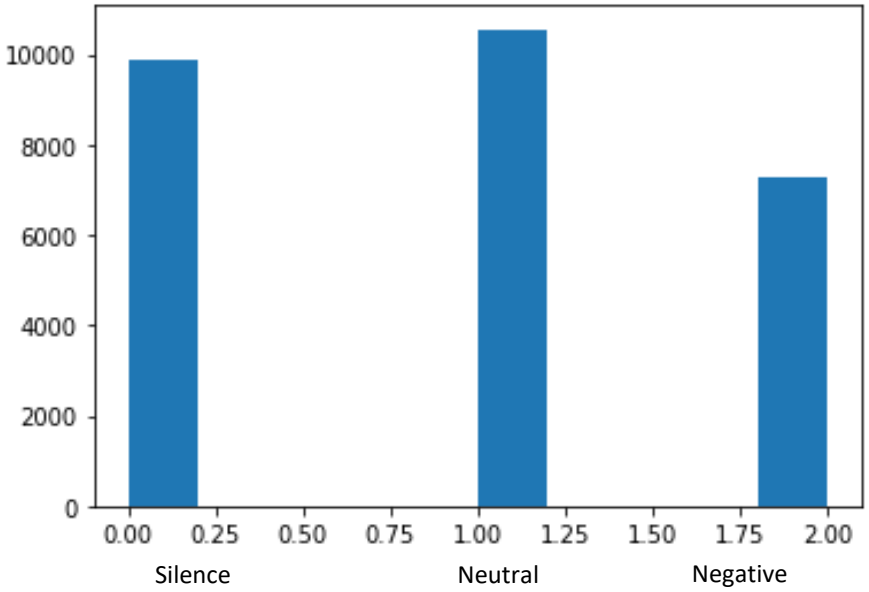
Also, sampling with a sliding window was used.



Data analysis

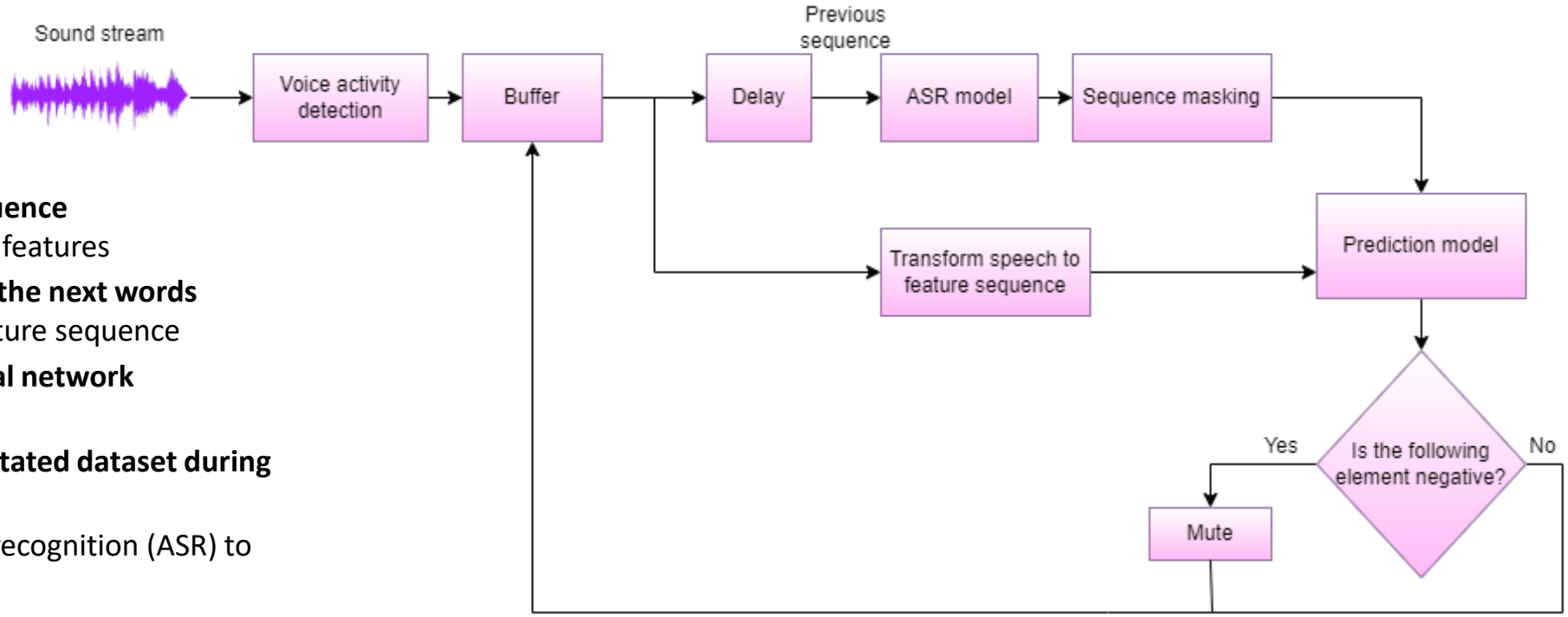


Distribution of the profanity speech in parts of Common Voice dataset phrases



Number of samples for each class after augmentation

Pipeline



- **Describing a word sequence** by a sequence of audio features
- **Predicting the color of the next words** sequence based on feature sequence
- Using a **recurrent neural network for prediction**
- Use **text** from the **annotated dataset during model training**.
- Use automatic speech recognition (ASR) to **correct the prediction of the model**.

Model comparison

Approach	F1 score (The main dataset)	F1 score (MELD dataset)	F1 score (Picked records)	Latency (sec)	Weights (mb)
LSTM (MFCC)	86.6	65.2	73.7	0.347*	51.53
LSTM (MFCC) + Attention	87.1	65.9	76.7	0.348	231.16
LSTM Wav2Vec + Attention	92.0	71.4	86.6	1.171	432.25**
LSTM (MFCC) + ASR + Attention	90.3	67.6	74.6	2.148	231.19

* - the average duration of the element is 0.34 sec
** - it is also necessary to use Wav2Vec feature extractor

Conclusion

1. We propose the method for word's label prediction in real-time fashion.
2. With the proposed approach, it is possible to deal with the latency of speech recognition while using information from it.
3. A trained multimodal prediction model with an F1 score of 74.6%
4. The expansion of textual information can increase performance.

https://github.com/expertspec/profanity-predictor

YSC young scientists conference

Thanks for your attention