



آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی

Language Processing and Digital Humanities

Text Localization in Audio

Final Project - NLP Course - Dr. Asgari

Mahdi Abootorabi
Arshan Dalili
Aryan Ahadiniya

Parsa Haghighi
Saeed Foroutan
Sina Rashidi

March 2023

Outline

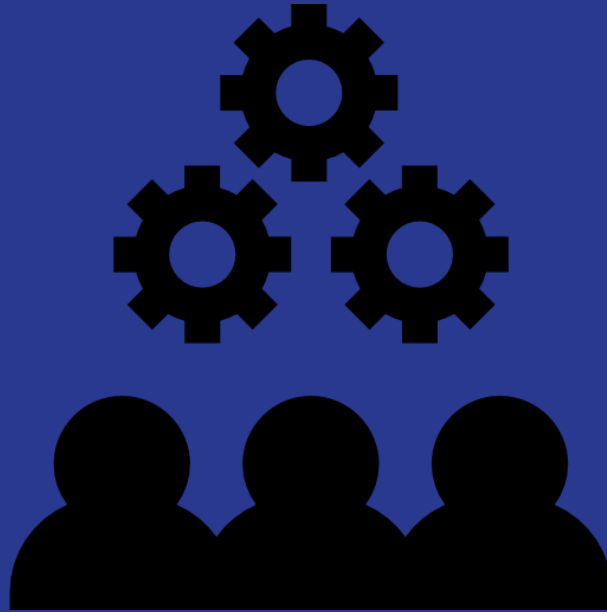
- Introduction
- Data Processing Pipeline
 - Collecting audio files
 - Audio segmentation
 - Proposed ASR model
 - Keyword extraction model
- Datasets
 - English
 - Persian
 -
- Model
 - Proposed Model
 - Baseline
- Results

Introduction

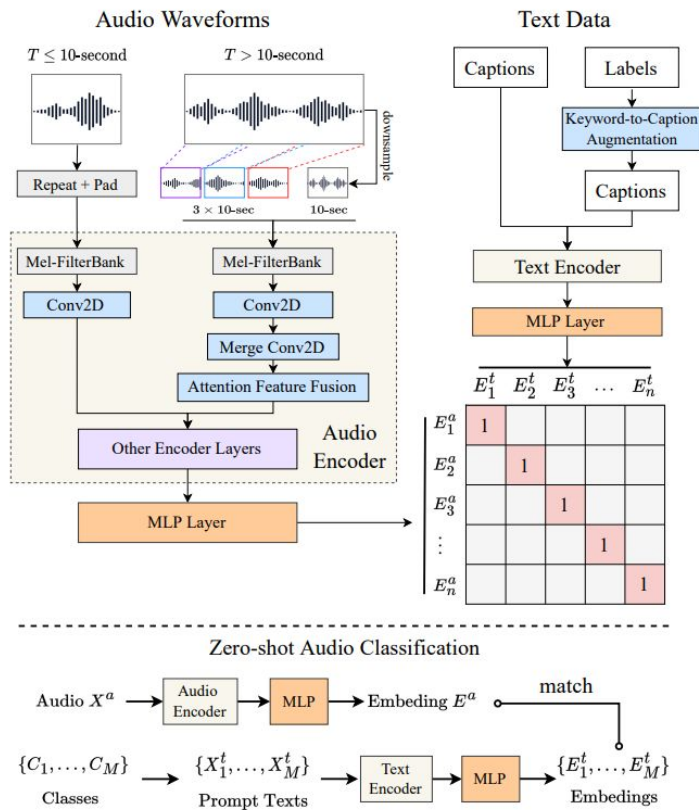
Text localization in audio involves the identification and localization of relevant text segments within an audio stream. This task is crucial in efficiently identifying speech segments that correspond to the words in a query text, thereby enhancing the search process. Text localization finds application in several domains, including retrieving old voice messages stored on social platforms and searching for content in audio such as tutorials or music.



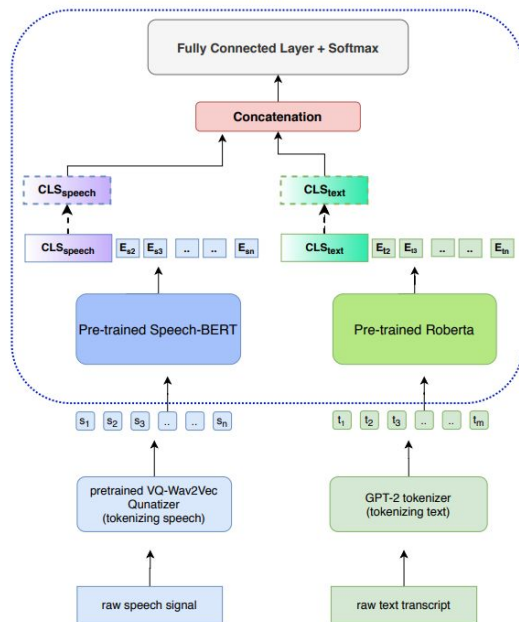
Related Works



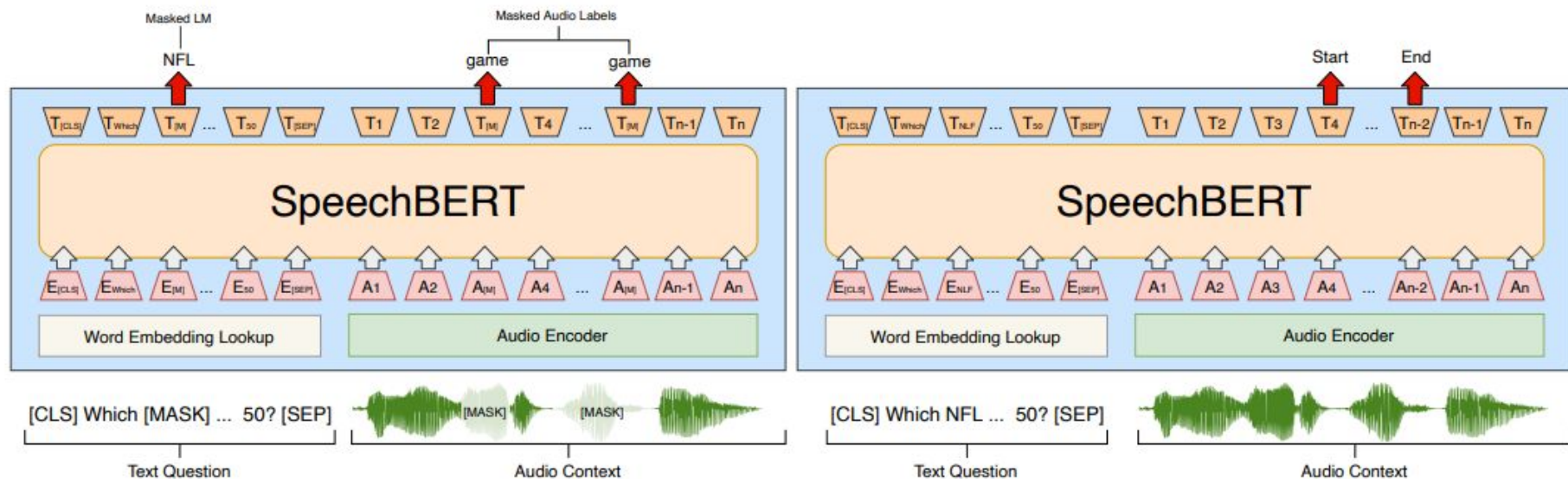
CLAP



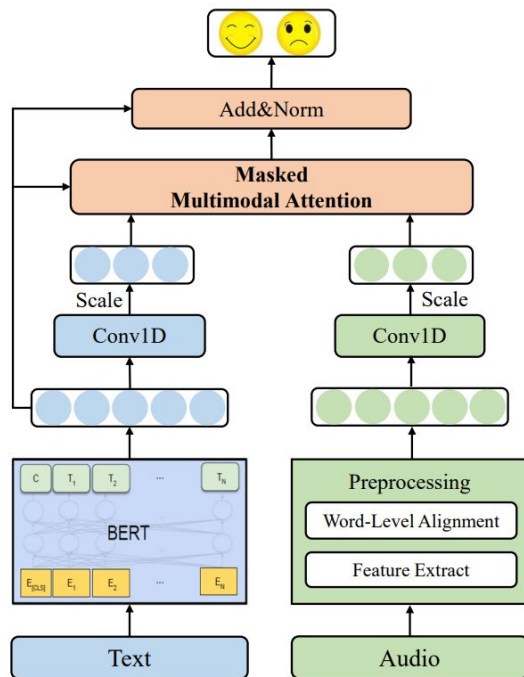
Jointly Fine-Tuning “BERT-like” Self Supervised Models to Improve Multimodal Speech Emotion Recognition



SpeechBERT: An Audio-and-text Jointly Learned Language Model



CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis



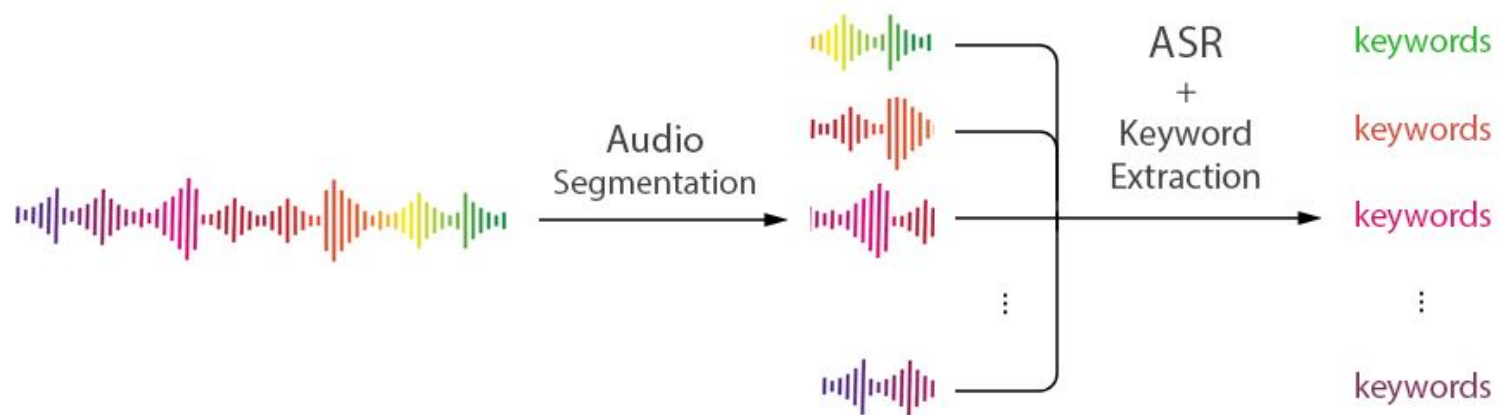
Our Approach

- Create datasets for Persian and English languages
- Create baseline model used cascade ASR and keyword extractor models for solving this problem (for this we may need to create some models for different tasks)
- Create a model which uses contrastive learning and without ASR to solve this problem and for building joint space between keywords and voices



Data Processing Pipeline

Pipeline



Collecting audio files

English Language Dataset:

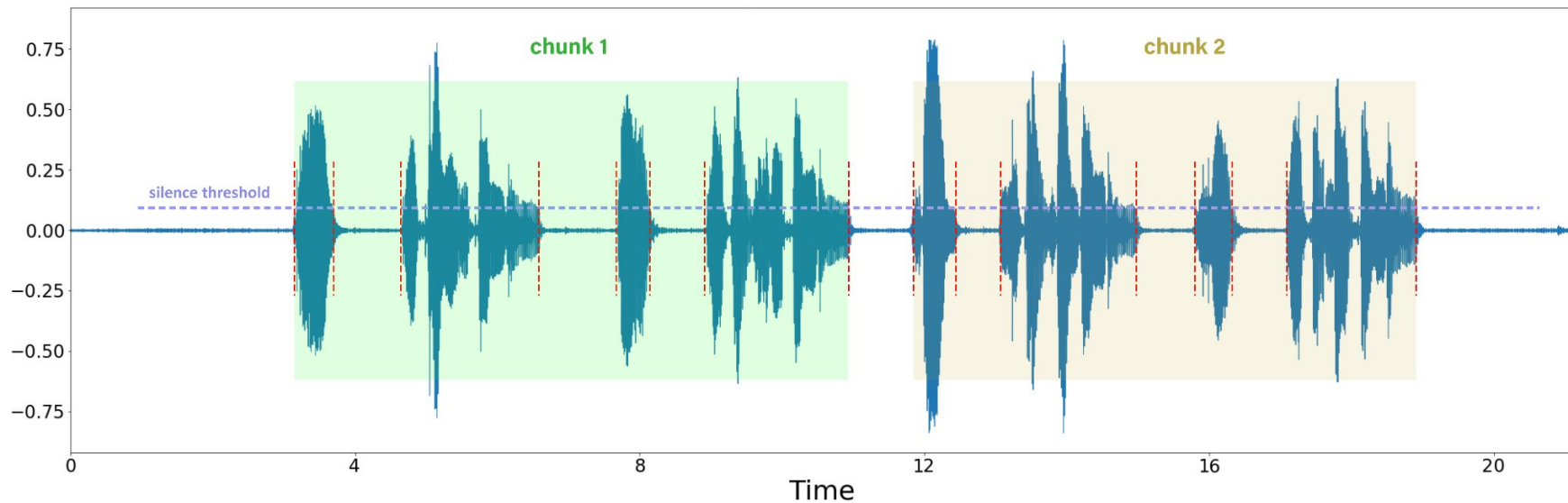
- Used a portion of the LibriSpeech dataset
- Audio chunks have our desirable feature
- Transcript available for each chunk
- Keyword extraction model used to create desired dataset

Persian Language Dataset:

- Existing datasets comprised of very short chunks
- New dataset creation necessary
- Farsi podcast selected
- Podcast in the form of an interview with multiple speakers
- Total duration of 70 hours



Audio segmentation



Persian and English ASR

- Wav2vec2 pretrained Models
- Conformer
- U2++_conformer
- Custom Model



Persian Keyword extraction

- PKE and Perke and Perkey packages
- Bert based Language Model
- YAKE algorithm
- Multi-RAKE algorithm
- Used Our fine-tuned Persian Summarizer



English Keyword extraction

- RAKE algorithm
- YAKE algorithm
- Bert based algorithm
- Maximal Marginal Relevance



Datasets



English Dataset

- Based on LibriSpeech: Small Dataset
- 3K Relevant Audios and Texts
- Create Keywords For Each Speech
- Create Sampled WaveForms For Them
- Train Test Validation Split
- Create Test Dataset With Negative Samples
- Save Bert Embedding and Wav2vec2 Embedding For Each Pairs

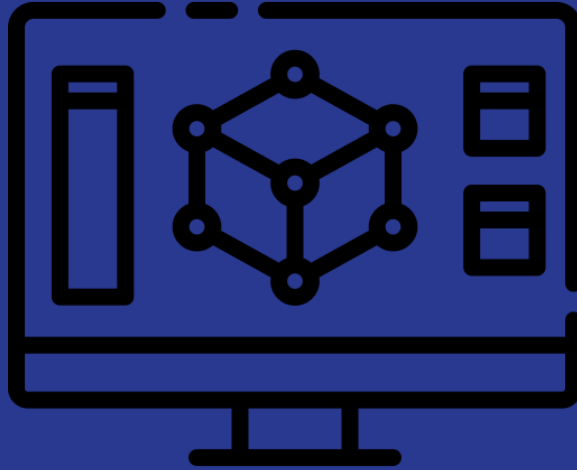


Persian Dataset

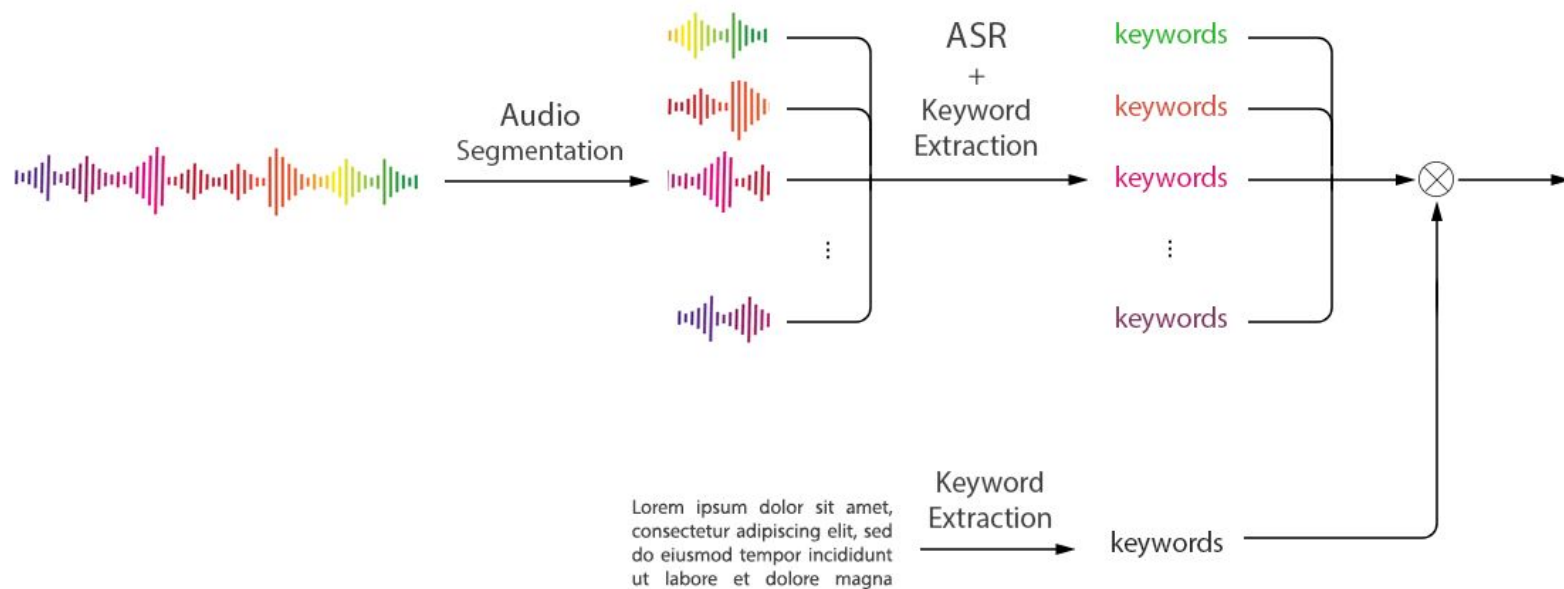
- Based on Radio Marz Podcasts
- Over 70 hours of Speech
- Use Audio Segmentations to Make Each Episodes Into Chunks
- Use Our ASR Models to Find Transcript of Chunks (Future work: Enhanced transcript with language models)
- Create Keywords For Each Chunks



Model



Baseline Model



Web-Based Demo



Audio Localizer

متن جستجو

موج صدا

No file chosen

Choose File

ارسال

© ۱۴۰۲، تمامی حقوق محفوظ است

توسعه داده شده توسط تیم اپوزیسیو فناوری های نوین پزشکی
بخشی از پرداترین زبان های طبیعی دانشکده مهندسی پزشکی

Web-Based Demo



0:00 / 0:30

شروع: 0.0، پایان: 13.25

اساس پژوهشی ستارههای
جرم بزرگترین ستارهها
کنونی امروزه جرم
پژوهشی ستارههای کیهان
خورشید هزار بزرگتر
ستارههای کنونی امروزه
هزار جرم خورشید
جرم خورشید هزار
بزرگتر بزرگترین ستارههای
اساس پژوهشی ستارههای
کنونی امروزه جرم بزرگترین
پژوهشی ستارههای کیهان
خورشید هزار بزرگتر
ستارههای کنونی امروزه
هزار جرم خورشید
جرم خورشید هزار
بزرگتر بزرگترین ستارههای

بر اساس پژوهشی جدید اولین ستارههای کیهان تا بیش از ده هزار برابر جرم خورشید رشد کردند و هزار برابر بزرگتر از بزرگترین ستارههای کنونی بودند امروزه جرم بزرگترین ستارهها

© ۱۴۰۲، تمامی حقوق محفوظ است

توسعه داده شده [لجی، نیا ایوتزایی](#) ریللی [حقیقی](#) فروتن رشیدی
بخشی از [پروژه زبانهای طبیعی دانشگاه صنعتی شریف](#)

Web-Based Demo



0:00 / 0:30

شروع: 0.0، پایان: 13.25

اساس پژوهشی ستارههای
جرم بزرگترین ستارهها
کنونی امروزه جرم
پژوهشی ستارههای کیهان
خورشید هزار بزرگتر
ستارههای کنونی امروزه
هزار جرم خورشید
جرم خورشید هزار
بزرگتر بزرگترین ستارههای
اساس پژوهشی ستارههای
کنونی امروزه جرم بزرگترین
پژوهشی ستارههای کیهان
خورشید هزار بزرگتر
ستارههای کنونی امروزه
هزار جرم خورشید
جرم خورشید هزار
بزرگتر بزرگترین ستارههای

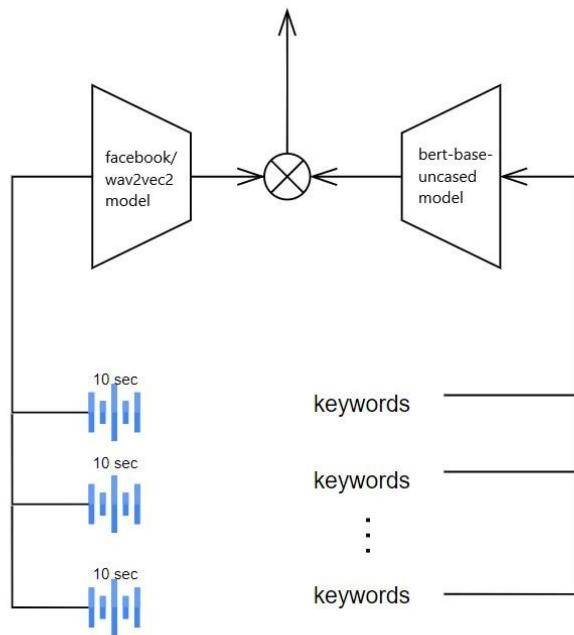
بر اساس پژوهشی جدید اولین ستارههای کیهان تا بیش از ده هزار برابر جرم خورشید رشد کردند و هزار برابر بزرگتر از بزرگترین ستارههای کنونی بودند امروزه جرم بزرگترین ستارهها

© ۱۴۰۲، تمامی حقوق محفوظ است

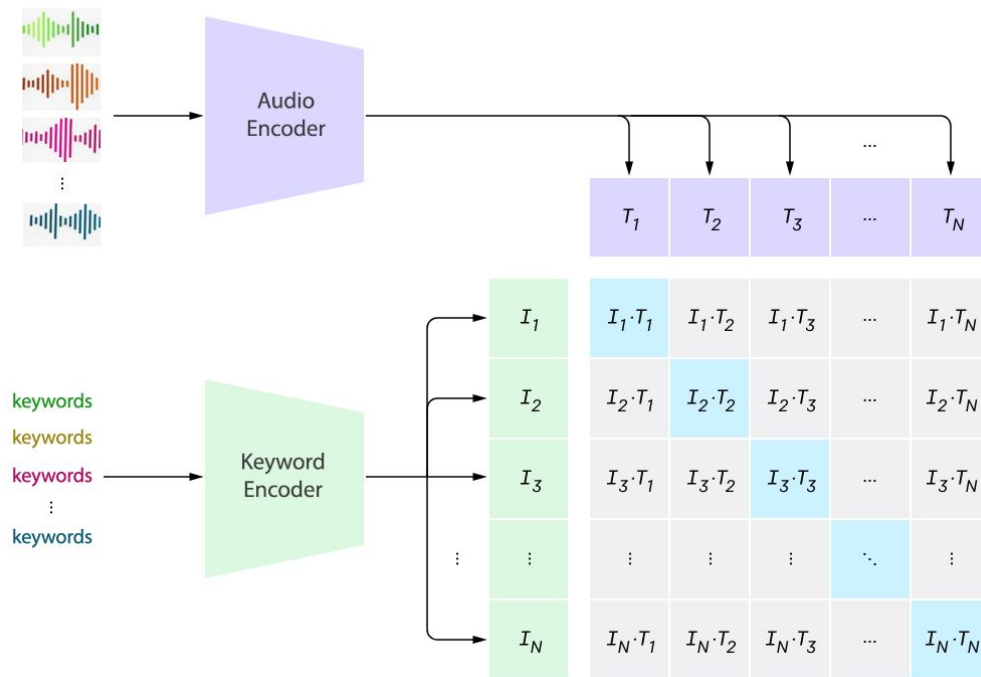
توسعه داده شده [لجی، نیا ایوتزایی](#) ریللی [حقیقی فروتن](#) رشیدی
بخشی از [پروژه زبانهای طبیعی دانشگاه صنعتی شریف](#)

Proposed Model

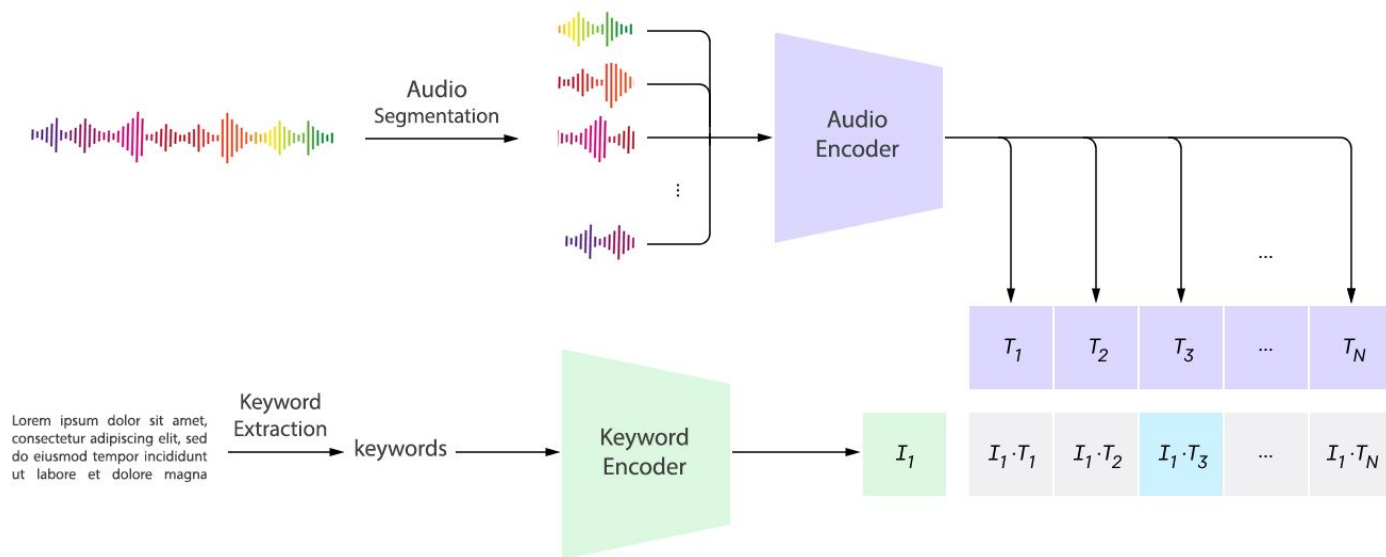
Based on
Contrastive Learning



Training

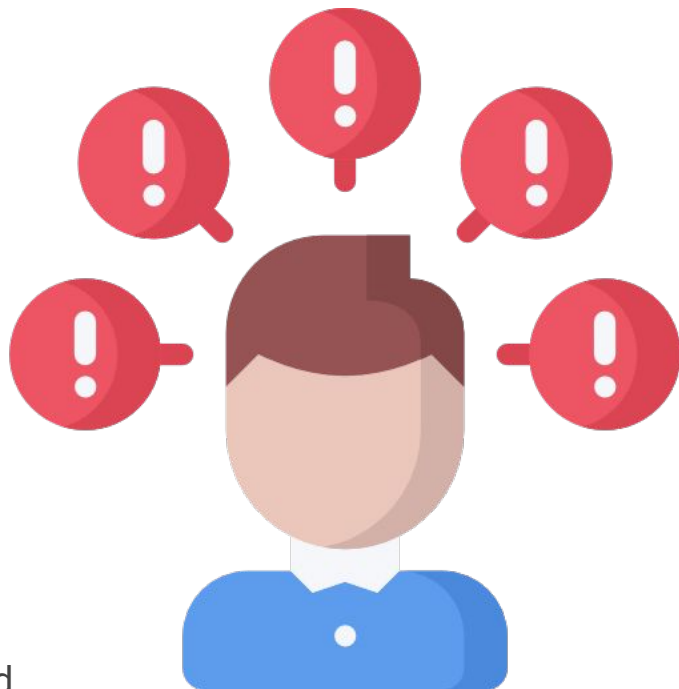


Inference



Proposed Model Problems

- Loss Problem
 - Contrastive Loss
 - SimCLR Loss
 - L1 Loss
 - Custom Loss
- Resource Problems
 - Generating datasets
 - Training Process
 - Cannot Make Architecture More Complicated
 - Storage



Results

Model	Hits@1	MRR	Precision	Recall	F1 Macro	Accuracy
Proposed Model	0.163	0.406	0.5	0.05	0.09	0.1
ASR based Model (Baseline)	0.177	0.418	0.178	0.18	0.176	0.177

Future Works

- Work on Architecture of Models
- Try to Enhance Them and Reach State-of-art Models
- Improve Web Based Demo of Models



References

- [Cross-modal-bert-for-text-audio-sentiment](#)
- [LARGE-SCALE CONTRASTIVE LANGUAGE-AUDIO PRETRAINING WITH FEATURE FUSION AND KEYWORD-TO-CAPTION AUGMENTATION](#)
- [Jointly Fine-Tuning “BERT-like” Self Supervised Models to Improve Multimodal Speech Emotion Recognition](#)
- [SpeechBERT: An Audio-and-text Jointly Learned Language Model for End-to-end Spoken Question Answering](#)



Any Questions?



thank you!