|  |  |
| --- | --- |
| نام و نام خانوادگی | **محمد پویا افشاری – علیرضا اسمعیل زاده** |
| شماره دانشجویی | **810198351-810198577** |
| تاریخ ارسال گزارش | **1402.09.17** |

|  |  |  |
| --- | --- | --- |
|  | **به نام خدا**  **دانشگاه تهران**  **دانشکده‌ مهندسی برق و کامپیوتر** |  |
| **درس شبکه‌های عصبی و یادگیری عمیق**  **تمرین پنجم** | | |

**فهرست**

[**پاسخ 1** – **SAM** 4](#_Toc152956619)

[۱-۱. **آماده سازی مجموعه داده** 4](#_Toc152956620)

[۱-۱-۱ . افزودن کتاب خانه ها 4](#_Toc152956621)

[۱-۱-۲ . جمع آوری داده 4](#_Toc152956622)

[۱-۱-۳ . resize تصاویر 5](#_Toc152956623)

[۱-۱-۴ . نمونه تصاویر 6](#_Toc152956624)

[۱-۱-۵ . تقسیم داده به دو بخش آموزش و ارزیابی 7](#_Toc152956625)

[**۲-۱**. بارگذاری مدل 7](#_Toc152956626)

[۱-۲-۱ . آماده سازی دیتاست 7](#_Toc152956627)

[۳**-۱.** تقويت داده 10](#_Toc152956628)

[**۴-۱.** بهینه‌ساز، متریک و تابع هزینه 10](#_Toc152956629)

[۵-۱. Fine-Tuneکردن مدل 11](#_Toc152956630)

[۱-۵-۱ . آموزش مدل 11](#_Toc152956631)

[۱-۵-۲ . ذخیره مدل 11](#_Toc152956632)

[**۶-۱.** ازیابی نتایج 12](#_Toc152956633)

[۱-۶-۱ . افزودن کتاب‌خانه ها 12](#_Toc152956634)

[۱-۶-۲ . لود کردن مدل ذخیره شده 12](#_Toc152956635)

[۱-۶-۳ . اجرای داده ازیابی بر روی مدل 13](#_Toc152956636)

[**پاسخ** **۲**  - آشنایی و پیاده سازی مدل Faster RCNN 15](#_Toc152956637)

[۱-۲. **توضیحات مدل ها** 15](#_Toc152956638)

[۲-۲ . پیش پردازش 19](#_Toc152956639)

[۲-۲-۱ . کتابخانه 19](#_Toc152956640)

[۲-۲-۲ . افزودن دیتاست 19](#_Toc152956641)

[۲-۲-۳ . نشان دادن تصاویر نمونه 19](#_Toc152956642)

[۲-۲-۴ . بدست آوردن تعداد کلاس‌ها و تغییر اندازه 20](#_Toc152956643)

[۲-۳ . آموزش شبکه 22](#_Toc152956644)

[۲-۳-۱ . آماده سازی دیتا 22](#_Toc152956645)

[۲-۳-۲ . نمایش تصاویر با Label های مربوط از روی کلاس تشکیل شده‌ی بالا 27](#_Toc152956646)

[۲-۳-۳ . آموزش شبکه و پیاده سازی مدل 28](#_Toc152956647)

[۲-۴ . بررسی داده های تست 31](#_Toc152956648)

**شکل‌ها**

شکل 1. عنوان تصویر نمونه 4

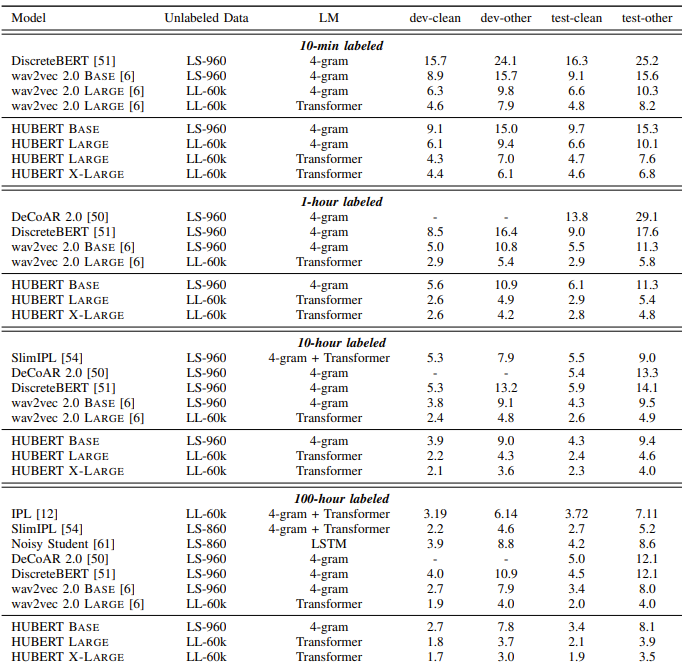
# **پاسخ 1** – **تشخیص احساسات گفتار(SER)**

## ۱-۱. **معرفی مدل HuBERT**

به طور کلی مدل HuBERT شبیه مدل word2vec 2.0 عمل می‌کند با این تفاوت که مقادیر را به جای quantize کردن cluster می‌کند. از مدل تقریبا مشابه BERT استفاده می‌کند و نیز مقدار loss ان ساده تر شده است. جدول زیر تفاوت مدل ها در عمل کرد را نشان می‌دهد:

جدول 1

جدول 2جدول 1. مقادیر مدل در برابر literature review



## ۱-۲. سوالات تشریحی

### ۱-۲-۱. چالش هاي داده هاي صوتی در یادگیري

مسائل یادگیری Self-supervised برای نمایش Speech به سه چالش اصلی تقسیم بندی می‌شوند:

1. به ازای هر گفته ورودی چندین واحد صدا وجود دارد

(There are multiple sound units for each input utterance)

1. مدل بدون اینکه به Transcription متنی دسترسی داشته باشد از روی Audio waveform باید تشخیص بدهد. که این کار را برای تشخیص مدل سخت می‌کند.

(There is no lexicon of input sound units during the pre-training phase)

1. مقادیر Sound unit ها در هر زبان متفاوت و متغیر است. مثلا در هیچ زبانی هر 20ms به عنوان یک Sound unit حساب نمی‌شود و این مقادیر در صوت های گوناگون مختلف است.

(Sounds units have variable lengths with no explicit segmentation)

علاوه بر سه مشکل اصلی ذکر شده در کار کردن بر روی داده‌های مصوت باید در نظر داشته باشیم که گویش هر فرد از یک کلمه با فرد دیگر متفاوت است. در کار بر روی متن ما بیشتر ساختار مند هستیم و همینطور قواعد جمله بندی و دستور زبان را رعایت می‌کنیم که در صوت این امر کمتر دیده می‌شود.

### ۱-۲-۲. رویکرد HuBERT

برای حل مسائل مطرح شده رویکرد‌های مختلفی مطرح شد از جمله آن مدل ‌هایی هستند که کلا بدید Quantization عمل می‌کنند و مدل های با Contrastive loss و Diversity loss. در مقایل HuBERT رویکرد متفاوتی در نظر می‌‍گیرد. رویکر HuBERT بیشتر شبیه BERT است. بنابر این از مدل های Language به عنوان منبع برای Speech representation بهره می‌گیرد.

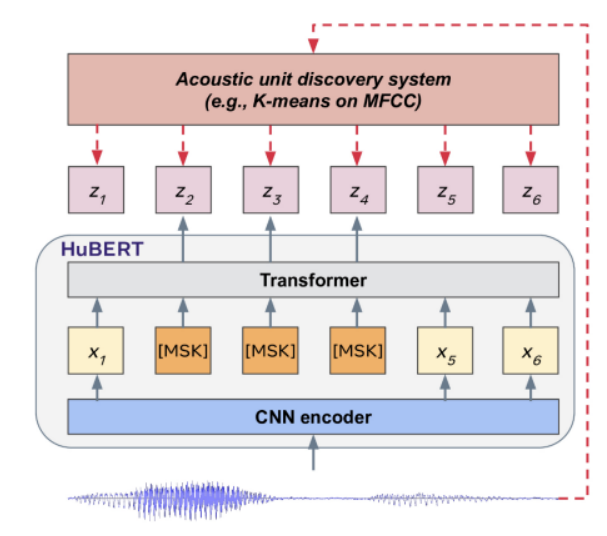


Figure 1شکل 1. شمای کلی فرآیند پیش آموزش HuBERT

برای حل مشکلات مطرح شده در شمای Pre-training مدل نشان داده می‌شود که ابتدا صوت دریافت شده به Encoder می‌رود. در این مرحله Encoder صوت رو به Snippet های 20 میلی ثانیه بخش بخش می‌کند. این حالت Speech representation تولید شده است. سپس نیمی از آنها Mask می‌شوند. از سمت دیگر یک Acoustic unit discovery system می‌بینیم که عمل Clustering را انجام می‌دهد. نتیجه این بخش تولید Representational unit های قابل بازنمایی Transformer است. مدل با انجام Cross entropy میتواند تفاوت بین این Z(k) و [MSK] مقدار MSK را بازنمایی کند.

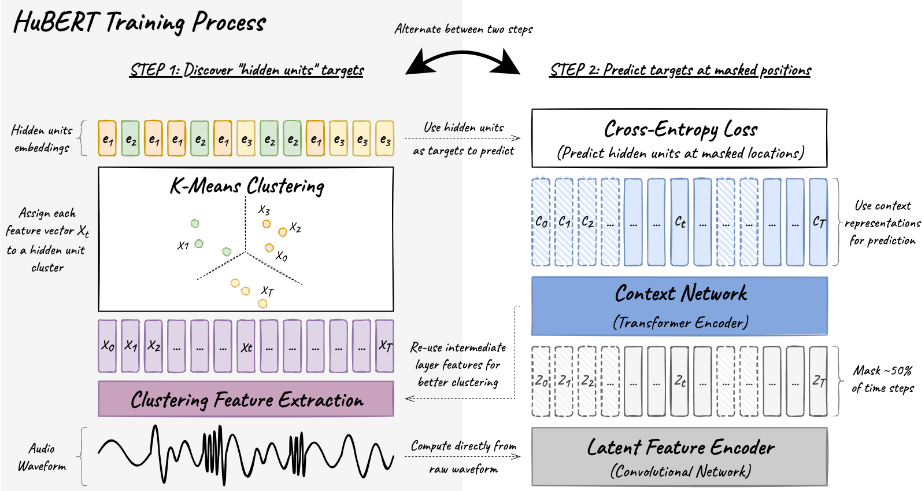


Figure 2شکل2. نمایش شمای train مدل HuBERT

بر اساس مدل یادگیری HuBERT ابتدا از Waveform که به Encoder داده می‌شود ( که در واقع شبکه CNN) است یک سری Hidden unit 20ms استخراج می‌شود. نیمی از این خروجی در مرحله Mask می‌شود. در مرحله Hidden units پیدا کردن به وسیله K-means cluster صورت می‌گیرد. در این مرحله همه ی Audio frame هایی که به یک Cluster assign می‌شود یک Label یکتا می‌گیرند.در نهایت Cluster Embedding های تولید شده ی این مرحله برای مقایسه Transformer Context Network و نهایتا خروجی Prediction با Cross Entropy loss برای Mask location ها می‌رود. درمرحله Prediction از Embedding vector به منظور استفاده می‌شود. در حقیقت مرحله دوم که مرحله Predict noisy target from context هست مشابه BERT صورت می‌پذیرد.

# **پاسخ** **۲** -