



**گزارش فنی**

**تنظیم دقیق چارچوب Wav2Vec2 در زبان فارسی**

**Fine-Tuned Wav2vec2 In Persian**

**حمیدرضا اکبری**

**خردادماه ۱۴۰۱**



## ۱. هدف

یکی از چارچوب‌های کارآمد در تشخیص خودکار<sup>۱</sup> چارچوب Wav2Vec می‌باشد. در این چارچوب ابتدا با روش یادگیری خودنظارتی<sup>۲</sup> یادگیری اولیه با داده‌های بدون برچسب انجام و در گام دوم با کمک یک روش تحت نظارت<sup>۳</sup> مدل گام قبل تنظیم دقیق‌تری خواهد شد. کارایی مناسب این روش حتی با حجم داده‌های برچسب دار اندک نکته کلیدی در این چارچوب می‌باشد. هدف از این گزارش بررسی عملی این چارچوب بر روی یک مجموعه داده و بررسی میزان کارایی این روش در تشخیص گفتار می‌باشد.

## ۲. مجموعه داده

Common Voice<sup>۴</sup> یک مجموعه‌ای از داده‌های صوتی در دسترس عموم می‌باشد. این مجموعه داده توسط افراد داوطلب در سطح جهانی پشتیبانی می‌شود و تولیدکنندگان برنامه‌های نرم افزار صوتی می‌توانند از این داده‌ها برای یادگیری مدل‌های پیش‌بینی استفاده نمایند. داده‌های مذکور بیش از ۹۰ زبان مهم جهان را پشتیبانی می‌کند. در زیر به برخی از شاخص‌های این مجموعه داده در زبان فارسی اشاره شده است.

اندازه داده	۹ گیگابایت
فرمت فایل صوتی	Mp3
تعداد فایل‌های صوتی	۴,۰۵۸

<sup>1</sup> Automatic Speech Recognition

<sup>2</sup> Self-Supervised Learning

<sup>3</sup> Supervised Learning

<sup>4</sup> <https://commonvoice.mozilla.org/en>



### ۳. پیاده سازی

تنظیم دقیق مدل یادگیری در زبان فارسی بر روی مجموعه داده هدف توسط افراد مختلف انجام گرفته است. برای تسريع در اجرا و ارزیابی نتایج در این تحقیق از یکی از پیاده سازی‌ها<sup>۵</sup> استفاده شده است. پیاده سازی در دو بخش ذیل مدنظر قرار گرفته است.

#### ۳,۱ آموزش و تنظیم دقیق مدل

برای قسمت تنظیم دقیق مدل از محیط کلب<sup>۶</sup> و زبان برنامه نویسی پایتون استفاده شده است. جزئیات پیاده سازی در فایل مورد نظر<sup>۷</sup> وجود دارد که در ادامه مراحل اصلی پیاده‌سازی تشریح می‌شود.

#### ۳,۱,۱ تفکیک داده های آموزش از تست

داده های برچسب دار آموزش و تست به نسبت ۱ به ۱۰ و به صورت تصادفی تعیین می‌شوند. تعداد داده های آموزش ۲۵۵۴ و تعداد داده های تست ۲۸۴ می‌باشد. داده های با فرمت CSV ساخته و ذخیره می‌شوند. برای آگاهی از صحت اطلاعات برخی از داده ها به صورت تصادفی نمایش داده می‌شوند.

#### ۳,۱,۲ پیش پردازش داده ها

عملیات پیش پردازش متن به شرح ذیل بر روی داده های آموزش و تست به صورت جداگانه انجام می‌گیرد. تعداد حروف مجموعه داده‌های آموزش و تست جهت بررسی و مقایسه اطلاعات نمایش داده می‌شود.

- **حذف کاراکترهای خاص:** برخی از کاراکترهای خاص نظیر ؟ از داده های متنی حذف می‌شود. لیست مجموعه این کاراکترها در برنامه مشخص شده است.
- **تبدیل برخی از کاراکترهای خاص:** برخی از حرفهای فارسی نظیر ی و ک به جهت کدینگهای متعدد به یک کد تبدیل می‌شود. لیست این تبدیلهای در برنامه مشخص شده است.
- **عملیات نرمالیز کردن متن:** با کمک تابع کتابخانه hazm متن نرمالیزه می‌شود. به عنوان مثال فاصله ها در برخی از کلمات به نیم فاصله تبدیل می‌شود.

<sup>5</sup> <https://huggingface.co/m3hrdadfi/wav2vec2-large-xlsr-persian>

<sup>6</sup> <https://colab.research.google.com/>

<sup>7</sup> Fine\_Tune\_XLSR\_Wav2Vec2\_on\_Persian\_ShEMO\_ASR\_with\_Transformers\_ipynb.ipynb



### ۳,۱,۳ پیش پردازش فایل های صوتی

از آنجاییکه فایل های صوتی مجموعه داده با نرخ نمونه گیری ۴۸ کیلوهرتز نمونه گیری شده اند در این گام نمونه گیری کلیه فایل های صوتی با نرخ ۱۶ کیلوهرتز انجام می شود.

### ۳,۱,۴ یادگیری

در این مرحله مدل بر اساس اطلاعات برچسب دار آموزش داده می شود. شاخص ارزیابی در طول مدت زمان آموزش میزان خطای کلمات می باشد که تابع `compute_metrics` این محاسبه را انجام می دهد.

### ۳,۲ ارزیابی

هدف از پیاده سازی این بخش تعیین میزان پیش بینی خطا بر اساس مدل آموزش داده شده در مرحله قبل می باشد. این بخش از برنامه با زبان برنامه نویسی پایتون<sup>۸</sup> تحت محیط توسعه نرم افزار پایچارم<sup>۹</sup> بر روی سیستم عامل ویندوز انجام شده است. میزان خطا بر اساس شاخص  $WER^{10}$  تعیین می گردد. بر این اساس خطا بر روی مجموعه داده مورد نظر  $34.87\%$  می باشد که فاصله زیادی با مقاله اصلی<sup>۱۱</sup> دارد. با عنایت به چالش نحوه بارگزاری فایل های mp3 از تابع کتابخانه `audio2numpy` استفاده شده است.

### ۴. سایر مدل ها

افراد متعددی مدل های تنظیم شده مختلفی را طراحی و اجرا نموده اند که از منظر کارایی و بر اساس نتایج اعلامی به شرح جدول ذیل می باشد. میزان تمایز برخی از این مدل ها ضرورت بررسی بیشتر در نحوه طراحی و پیاده سازی آنها را افزایش خواهد داد.

مدل	کارایی (WER)
<code>jonatasgrosman/wav2vec2-large-xlsr-53-persian</code>	۳۰,۱۲
<code>m3hrdadfi/wav2vec2-large-xlsr-persian</code>	۳۴,۸۷
<code>m3hrdadfi/wav2vec2-large-xlsr-persian-v3</code>	۱۰,۳۶

<sup>8</sup> Python

<sup>9</sup> PyCharm

<sup>10</sup> Word Error Rate

<sup>11</sup> <https://arxiv.org/abs/2006.11477>