# چگونگی اعمال pre-trained Wav2Vec2.0 در تشخیص خودکار گفتار در یک استاندارد ارتباطات کنترل ترافیک هوایی

دوره کارآموزی عصر گویش پرداز

زهرا رحیمی

۱۴ مرداد ۱۴۰۱

# فهرست مطالب

## ۱ مقدمه

تشخیص خودکار گفتار (ASR) می تواند گفتار را به متون قابل خواندن توسط کامپیوتر ترجمه کند. در کنترل ترافیک هوایی(ATC)، راه اصلی ارتباط بین کنترلرهای ترافیک هوایی(ATCo) و خلبانان گفتار رادیویی است. تشخیص خودکار گفتار برای ترجمه گفتار کنترلرهای ترافیک هوایی و خلبان به سیستم ترافیک هوایی معرفی شده است که می تواند برای کاهش بار کاری و اطمینان از ایمنی پرواز استفاده شود. کارهای اخیر بر روی پیش‌آموزشی تحت نظارت خود(self-supervised pre-training) بر روی داده‌های گفتاری بدون برچسب در مقیاس بزرگ برای ساخت مدل‌های صوتی (AM) قوی (E2E) متمرکز شده است که بعداً می‌توانند در کارهای downstream مانند تشخیص خودکار گفتار تنظیم شوند. ما سناریویی را با تجزیه و تحلیل استحکام مدل‌های Wav2Vec2.0 و XLS-R در downstream ASR برای حوزه ارتباطات کنترل ترافیک هوایی هدف قرار می‌دهیم.

ATC با هدایت هواپیما در هوا و روی زمین از طریق ارتباطات صوتی بین کنترلرهای ترافیک هوایی (ATCOs) و خلبانان سر و کار دارد. اینها توسط یک دستور زبان و واژگان کاملاً تعریف شده کنترل می شوند که باید برای فراهم کردن یک پرواز امن و قابل اعتماد از ترافیک هوایی و در عین حال پایین نگه داشتن هزینه های عملیاتی تا حد امکان رعایت می شوند. علی رغم علاقه به یک سیستم تشخیص گفتار خودکار برای ارتباطات کنترل ترافیک هوایی، یک سیستم کاملاً کاربردی در بازار وجود ندارد که از جمله دلایل آن می توان به این دو اشاره کرد:

۱. کارایی های پایین سیستم های معرفی شده (به جای تأخیر در انجام وظایف آنها، بهره وری هاATCO را افزایش می دهد) در نتیجه اهمیت تشخیص درست دستورات و کاهش خطای تشخیص دستورات خلبان (CA-WER (call-sign word-error-rate)

۲. فقدان داده‌های گفتاری مشروح در مقیاس بزرگ (کمتر از ۵۰ ساعت داده‌های گفتاری منبع باز) و هزینه تولید بالای آن، آن را تقریبا غیرعملی می‌ساز

## ۲ انگیزه انجام پژوهش

پژوهش های ما سناریوی عدم تطابق دامنه را با پاسخ به سه سوال زیر پوشش می دهد.:

عملکرد مدل های **E2E** از پیش آموزش دیده در حوزه های جدیدی مانند **ATC** چقدر بی نقص و قدرتمند هستند؟

نتایج ما (جدول ۳) تأیید می کند که مدل‌های انتها به انتها یا E2E که توسط self-supervised-learning یا SSL پیش آموزش شده اند (pre-trained) (مانند Wav2Vec2) یک نمایش قوی از گفتار را یاد می‌گیرند. تنظیم دقیق (fine-tune) در یک کار downstream (مانند ASR) از نظر محاسباتی ارزان‌تر از آموزش از ابتدا است، و برای دستیابی به نتایج قابل مقایسه با ASR مبتنی بر روش هیبریدی، به داده‌های درون دامنه(in-domain) کمتری نیاز دارد. ما همچنین این فرضیه را مطرح می‌کنیم که مدل‌های چند زبانه E2E مانند XLS-R در داده‌های گفتاری ATC

که حاوی انگلیسی لهجه دار (یعنی مجموعه‌های LiveATC-Test و ATCO2-Test) هستند، بهتر هستند. به دلیل بازنمایی کلی گفتار آموخته‌شده در طول SSL.

چه مقدار داده برچسب‌دار **ATC** (اعم از صوتی و متنی) در مرحله تنظیم دقیق مورد نیاز است تا نتایج قابل مقایسه با مدل‌های مبتنی بر هیبریدی باشد؟

ما یک مطالعه مقایسه ای از ۵ دقیقه (یادگیری چند شات) تا حدود ۱۵ ساعت گفتار برچسب دار (یعنی از ۱۰۰ تا ۱۵ هزار گفته) انجام می دهیم. علاوه بر این، ما افزایش عملکرد به‌دست‌آمده از رمزگشایی با beam search را با استفاده از یک مدل زبان درون دامنه (LM) به‌جای رمزگشایی بر اساس مدل حریصانه یا greedy بررسی می کنیم.

حتی اگر مدل‌های **Wav2Vec2** و **XLS-R** از نظر طراحی قابلیت **streaming** ندارند، آیا چنین مدل‌های **E2E** می‌توانند در برنامه‌های بلادرنگ مانند **ATC** استفاده شوند؟

بسیاری از برنامه های کاربردی (به عنوان مثال، ATC) به موتورهای ASR streaming نیاز دارند. ما تاخیر عبور و رمزگشایی هر دو مدل E2E، یعنی Wav2Vec2 و XLS-R را در طول استنتاج ارزیابی می کنیم.

## ۳ دادگان

ما روی دو مجموعه آموزشی و چهار مجموعه تست به زبان انگلیسی با لهجه‌های مختلف آزمایش می کنیم (جدول ۱). به این نکته توجه داریم که جمع آوری داده های کنترل ترافیک هوایی به دلیل شرایط نویز، حریم خصوصی داده ها، میزان گفتار و لهجه زبان، چالش برانگیز و پرهزینه است.

NATS و ISAVIA: داده های گفتاری توسط ارائه دهندگان خدمات ناوبری هوایی (ANSP) برای پروژه HAAWAII جمع آوری و حاشیه نویسی می شود. دو مجموعه داده عبارتند از، الف) رویکرد لندن (NATS) و (ب) ایسلندی (ISAVIA). در مجموع، ۳۲ ساعت داده های رونویسی دستی برای آموزش و ۲ ساعت برای آزمایش وجود دارد. هر دو مجموعه داده به‌عنوان گفتار با کیفیت خوب و با فرکانس ۸ کیلوهرتز فهرست‌بندی می‌شوند. برای دیدن جزئیات بیشتر به جدول ۱ مراجعه کنید.

ATCO2-Test: مجموعه توسعه و ارزیابی موجود به عنوان منبع باز و ارائه شده در Interspeech در سال ۲۰۲۱. داده ها از ارتباطات کنترل ترافیک هوایی از فرودگاه های مختلف واقع در استرالیا، جمهوری چک، اسلواکی و سوئیس تشکیل شده است. این دیتاست حاوی ترکیبی از ضبط‌های پر سر و صدا و با لهجه انگلیسی است. این اولین مطالعه‌ای است که تشخیص خودکار گفتار انتها به انتها(E2E ASR) را برای ATCO2-Test ارزیابی می کند، به‌عنوان مثال، نرخ خطای کلمه یا WERهای فهرست‌شده در اینجا می‌توانند به‌عنوان خطوط پایه برای تحقیقات آینده مورد استفاده قرار گیرند.

LiveATC-Test: مجموعه آزمایشی از داده‌های LiveATC4 ضبط‌شده از کانال‌های رادیویی VHF در دسترس عموم، به عنوان بخشی از پروژه ATCO2 جمع‌آوری می‌شود و شامل ضبط‌های

آزمایشی و کنترلر ترافیک هوایی با انگلیسی لهجه‌دار از فرودگاه‌های واقع در ایالات متحده، جمهوری چک، است. ایرلند، هلند و سوئیس. ما LiveATC-Test را به عنوان مجموعه داده‌های گفتاری با کیفیت پایین در نظر می گیریم، یعنی نسبت سیگنال به نویز (SNR) از ۵ تا ۱۵ دسی‌بل می‌رود. (SNR شاخصی که میزان کیفیت سیگنال را نشان می دهد و SNR بالای ۱۰ الی ۱۵ نشان دهنده کیفیت مقبول است)

| | Characteristics | | |
|---|---|---|---|
| **Dataset** | Train / Test | SNR [dB] | WER [%]† |
| **NATS** | 18h / 0.9h | ≥20 | 7.7 |
| **ISAVIA** | 14h / 1h | 15-20 | 12.5 |
| **ATCO2-Test** | - / 1h | 10-15 | 24.7 |
| **LiveATC-Test** | - / 1.8h | 5-15 | 35.8 |

جدول ۱: ویژگی های دادگان آموزشی و آزمایشی

# ۴ تشخیص خودکار گفتار

راه اندازی ما بر اساس دو مجموعه داده تنظیم دقیق (fine-tune) شده است. اول، ما از ۳۲ ساعت داده های مشروح از NATS و ISAVIA استفاده می کنیم و ویژگی های آن را در جدول ۱ فهرست می کنیم. در حال حاضر از این مجموعه های تنظیم دقیق به عنوان مجموعه های تنظیم دقیق ۳۲ ساعت و ۱۳۲ ساعت یاد می کنیم.

## ۱.۴ تشخیص خودکار گفتار مبتنی بر روش هیبریدی

همه آزمایش‌ها با جعبه ابزار Kaldi انجام می‌شوند. مدل‌های پایه از شش لایه کانولوشن و ۱۵ شبکه عصبی تاخیر زمانی فاکتوریزه شده (حدود ۳۱ میلیون پارامتر قابل آموزش) تشکیل شده‌اند. ما دستورالعمل استاندارد آموزش زنجیره ای Kaldi's LF-MMI را دنبال می کنیم. ویژگی های ورودی MFCC با وضوح بالا با میانگین نرمال کپسترال آنلاین (CMN) هستند. ویژگی ها با i-vectors گسترش یافته اند. در هنگام رمزگشایی از مدل زبانی APRA سه تایی استفاده می کنیم. این مدل برای ۵ دوره در ۱۳۲ ساعت گفتار کنترل ترافیک هوایی(شامل NATS و ISAVIA) آموزش داده شده است. WER در آخرین ستون جدول ۱ فهرست شده است.

## ۲.۴ تشخیص خودکار گفتار انتها به انتها

ما نتایج چهار پیکربندی مدل‌های Wav2Vec2/XLS-R را گزارش می‌کنیم که از فرم پلات HuggingFace واکشی شده‌اند. از این پس، ما این مدل‌ها را به‌عنوان زیر برچسب‌گذاری می‌کنیم:

| | |
|---|---|
| W2v2-B BASE model | • 95M parameters<br>• Pre-trained on train-set 960h LibriSearch |
| W2v2-L: Large-960h model | • 317M parameters<br>• Pre-trained then fine-tuned with LibriSearch 960h train-set |
| W2v2-L-60K: Large-960h-LV60K model | • same as w2v2-L but uses LibriSpeech + 60k h from Libri-Light during the pre-training phase |
| W2v2-XLS-R: XLS-R model | • 300M parameters pre-trained on 436k h of publicly available data in 128 languages |

جدول۲: پیکربندی مدل های Wav2Vec2/XLS-R

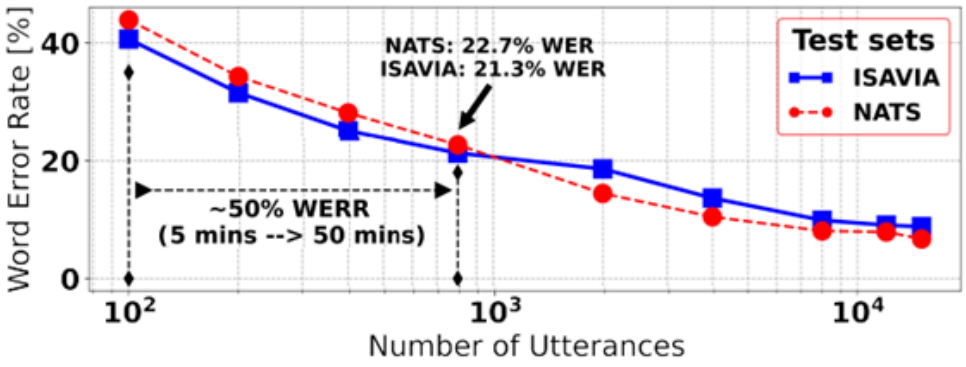
همه آزمایش‌ها از یک مجموعه فراپارامتر (hyperparameter) استفاده می‌کنند. رمزگذار ویژگی (feature encoder) در کل مرحله تنظیم دقیق به روز نمی شود (روش رایج در سناریوهای با منابع کم). ما هر مدل را برای ۱۰ هزار گام با یک مرحله گرم کردن(Warm-up) ۵۰۰ گامی تنظیم می کنیم (تقریباً ۵ درصد از کل به روز رسانی ها). سرعت یادگیری به صورت خطی تا 1e-4 در طول گرم کردن افزایش می یابد، سپس به صورت خطی تحلیل می رود. ما هر مدل را روی NVIDIA GeForce RTX 3090 با اندازه دسته (batch) ۷۲ (اندازه دسته ۲۴، انباشتگی گرادیان ۳) تنظیم دقیق می کنیم. ما از واژگان مبتنی بر کاراکتر با ابعاد ۳۲ استفاده می کنیم.

### ۱.۲.۴ تقویت داده (**data augmentation**)

توالی ورودی را با احتمال $p = 0.075$ و $M = 12$ فریم متوالی ماسک می کنیم. ما همچنین از شاخص فعال سازی ۰۵.۰ استفاده می کنیم. این فراپارامترها از پیاده سازی اصلی Wav2Vec2 پیروی می کنند.

### ۲.۲.۴ مدل زبانی (**language model**)

ما همه رونوشت‌های متنی را به هم متصل می‌کنیم و مدله ای زبانی ARPA ۲، ۳ یا ۴تایی را آموزش می‌دهیم. مدل های زبانی با همجوشی کم عمق با رمزگشای CTC مبتنی بر پایتون ادغام می شوند. مدل زبانی ۴ تایی به طور سیستماتیک در مقایسه با ۲ تایی ها نتایج بهتری در همه مجموعه های آزمایشی انجام دادند (حدود ۲ درصد WERR نسبی). ما نتایج را فقط با مدل زبانی ۴تایی گزارش می کنیم. همچنین ما $\alpha = 0.5$ و $\beta = 1.5$ را تنظیم کردیم، که مربوط به طول و وزن نرمال سازی امتیاز مدل زبانی است. اندازه beam را روی ۱۰۰ قرار می دهیم.




شکل ۱: شکل ۱- تاثیر اندازه داده تنظیم دقیق روی نرخ خطای کلمه

### ۳.۲.۴ آموزش افزایشی

با موفقیت اخیر مدل‌های انتها به انتها که با آموزش تحت نظارت خود از پیش آموزش داده شده‌اند، تعیین مقدار داده‌ای که یک مدل واقعاً برای انجام مؤثر در یک کار downstream نیاز دارد، از اهمیت ویژه‌ای برخوردار است. این امر به ویژه برای کارهای با منابع اندک مانند کنترل ترافیک هوایی که چند ده ساعت داده برچسب‌گذاری شده برای آموزش یا تنظیم دقیق در دسترس است، بسیار مهم است. در اکثر این نمونه ها، داده های یک فرودگاه به فرودگاه های دیگر به خوبی تعمیم نمی یابد و این به دلیل تغییر دامنه قابل توجه AM (لهجه، نرخ بلندگو)، و همچنین تغییر دامنه LM (تسلط هواپیماهای مختلف و دستورات مختلف بسته به فرودگاه) می باشد. ما عملکرد مدل را در مقابل اندازه‌های مختلف داده‌های تنظیم دقیق تحلیل می کنیم. ما با چهار سناریو یادگیری چند شات با کمتر از یک ساعت (نزدیک ۱۰۰۰ عبارت) داده های تنظیم دقیق آزمایش کردیم. در مجموع، ۹ مدل فقط بر روی NATS (خط چین قرمز) یا فقط بر روی داده های ISAVIA (خط مستقیم آبی) تنظیم شده اند که در شکل ۱ نشان داده شده است (محور x به تعداد عبارت های استفاده شده در هنگام تنظیم دقیق در مقیاس لگاریتم اشاره دارد).

هر مجموعه آزمایشی (یعنی NATS و ISAVIA) فقط از داده های درون دامنه خود در هنگام تنظیم دقیق و ارزیابی استفاده می کند. گفته های ۱۰۰، ۱،۰۰۰ و ۱۰،۰۰۰ به ترتیب تقریباً ۵ دقیقه (چند شات)، ۱ ساعت و ۱۰ ساعت است.

### ۴.۲.۴ ارزیابی **streaming**

مدل‌های Wav2Vec2 و XLS-R با قابلیت‌های streaming طراحی نشده‌اند، اما می‌توان از قابلیت‌های GPU در طول استنتاج برای ارائه رمزگشایی بلادرنگ استفاده کرد. برای آزمایش این فرضیه، روش زیر را انجام می‌دهیم: صوت را به n تکه با اندازه‌های افزایشی تقریبا به اندازه

۳۰۰ میلی‌ثانیه تقسیم می‌کنیم. سپس هر تکه افزایشی را به مدل منتقل می کنیم و این کار را برای همه تکه‌های صوت انجام می دهیم. در نهایت، میانگین زمان مورد نیاز شبکه برای رمزگشایی تمام تکه‌های یک صوت معین را اندازه‌گیری می‌کنیم. ما این فرآیند را روی ۱۰۰ نمونه تصادفی از مجموعه‌های آزمایشی تکرار می‌کنیم و میانگین زمان تأخیر را گزارش می‌کنیم. ما تأثیری را که در راه‌اندازی streaming بر روی WER ایجاد شده است در نظر نمی‌گیریم.

# ۵ نتایج تجربی

در این مقاله، ما فرض می‌کنیم که مدل‌های انتها به انتها آموزش‌دیده شده تحت نظارت خود، یک نمایش قوی از گفتار را یاد می‌گیرند و در وظایف downstram مانند تشخیص خودکار گفتار تک زبانه یا چند زبانه به خوبی عمل می‌کنند. ما یافته های خود را با پاسخ به سؤالات زیر تقسیم کردیم:

## ۱.۵ شکستن پارادایم، مبتنی بر روش هیبریدی یا انتها به انتها؟

اگرچه مدل‌سازی سیستم تشخیص خودکار مبتنی بر روش هیبریدی برای چندین سال پیش‌فرض بوده است، اما موج جدیدی از معماری‌های انتها به انتها که توسط آموزش تحت نظارت خود برای ترکیب اشتراکی AM و LM آموزش داده شده‌اند جای آن را گرفته‌اند. ما مدل‌های انتها به انتها را با بهترین روش مبتنی بر هیبریدی که با مجموعه تنظیم دقیق ۱۳۲ ساعته در Kaldi آموزش داده شده است مقایسه می‌کنیم (ردیف اول جدول ۳).

برای مدل‌سازی انتها به انتها:

۱. w2v2-L-60k را برای مجموعه‌های آزمایشی NATS و ISAVIA انتخاب می‌کنیم، که فقط در مجموعه ۳۲ ساعته، یعنی داده‌های درون دامنه، به‌خوبی تنظیم شد.

۲. سپس w2v2-XLS-R+برای مجموعه‌های تست ATCO2-Test و LiveATC-Test، که بر روی ۱۳۲ ساعت از داده‌های گفتاری کنترل ترافیک هوایی آموزش داده شد، شامل داده‌های متنوع‌تر و مشابه مدل مبتنی بر هیبریدی

در نهایت w2v2-L-60k ۳۰ درصد کاهش WER نسبی را در NATS و ۴۱ درصد در ISAVIA در مقایسه با روش مبتنی بر هیبریدی به همراه داشت. این بهبود قابل توجه است، حتی اگر مدل پایه بر روی چهار برابر داده های w2v2-L-60k آموزش داده شده باشد (جدول ۳ را ببینید).

به طور مشابه، w2v2-XLS-R+ (ردیف آخر، جدول ۳) در هر چهار مجموعه آزمایشی از مدل مبتنی بر ترکیبی پیشی می‌گیرد، اما در ATCO2-Test و LiveATC-Test، دو مورد چالش برانگیز(به دلیل لهجه دار بودن دادگان)، نتایج بسیار قابل توجه است. در مجموع، ۱۹ و ۳۰ درصد WER نسبی در ATCO2-Test و LiveATC-Test به ترتیب به دست آمد (مقایسه w2v2-XLS-R+ نسبت به روش هیبریدی)

| Model (num. params.) | Unlabeled data | NATS Greedy | NATS +LM | ISAVIA Greedy | ISAVIA +LM | ATCO2-Test Greedy | ATCO2-Test +LM | LiveATC-Test Greedy | LiveATC-Test +LM | LS* - | Latency (ms)¶ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Baseline (31M)** | | | | | | | | | | | |
| Hybrid-based † | - | - | 7.7 | - | 12.5 | - | 24.7 | - | 35.8 | - | ~400§ |
| **BASE (95M)** | | | | | | | | | | | |
| w2v2-B | LS | 10.7 | 8.4 | 12.5 | 10.1 | 45.6 | 40.1 | 48.1 | 42.2 | 7.8 | 32/69 |
| **LARGE (371M)** | | | | | | | | | | | |
| w2v2-L | LS | 9.3 | 7.6 | 11.7 | 9.5 | 44.9 | 40.0 | 47.5 | 41.4 | 6.1 | 33/73 |
| w2v2-L-60k | LS+LV | **6.8** | **5.4** | **8.8** | **7.3** | 34.6 | 31.2 | 39.8 | 34.5 | 4.9 | 33/76 |
| w2v2-L-60k+†† | LS+LV | 9.3 | 7.4 | 11.2 | 9.1 | 23.3 | 21.2 | 31.1 | 27.2 | - | -/- |
| **XLS-R (300M)** | | | | | | | | | | | |
| w2v2-XLS-R | ML | 8.4 | 6.5 | 10.5 | 8.2 | 39.1 | 33.8 | 42.9 | 36.7 | 15.4 | 39/76 |
| w2v2-XLS-R+†† | ML | 9.0 | 7.4 | 10.4 | 8.3 | **22.8** | **19.8** | **29.7** | **24.9** | - | -/- |

شکل ۲: WER ثبت شده از چهار مجموعه آزمایشی مطرح شده. WERها به صورت bold و underline به ترتیب مدل‌هایی اشاره می‌کنند که روی داده‌های ۳۲ ساعته و ۱۳۲ ساعته تنظیم شده‌اند.

## ۲.۵ آیا داده های افزوده ی جزئی در دامنه(**partly-in-domain**) ،عملکرد سیستم تشخیص گفتار را افزایش می دهد؟

ما به این سوال با مقایسه مدل‌های تنظیم‌شده دقیق در مجموعه ۱۳۲ ساعت یا ۳۲ ساعت پاسخ می‌دهیم. توجه داشته باشید که NATS و ISAVIA مجموعه‌های گفتاری کنترل ترافیک هوایی درون دامنه‌ای تمیز هستند، به‌عنوان مثال، به‌عنوان درون‌دامنه برای ۳۲ ساعت و در غیر این صورت تا حدی درون دامنه(in-domain) در نظر گرفته می‌شوند (مجموعه ۱۳۲ ساعت). ATCO2-Test و LiveATC-Test را می توان به عنوان مجموعه های پر سر و صدا و تا حدی در دامنه (فرودگاه های مختلف، به عنوان مثال، عدم تطابق صوتی و LM) در نظر گرفت. ما روی w2v2-L-60k و w2v2-L-60k+ که به ترتیب در مجموعه‌های ۳۲ و ۱۳۲ ساعت تنظیم شده‌اند، تمرکز می‌کنیم. توجه داشته باشید که نتایج قابل مقایسه ای بین w2v2-XLS-R و w2v2-XLS-R+ وجود دارد. ما WER را در رمزگشایی حریصانه تجزیه و تحلیل می‌کنیم تا فقط روی AM+LM مشترک تمرکز کنیم. در اینصورت یک تنزل WER برای مجموعه‌های آزمایشی درون دامنه، NATS: 6.8٪ → 9.3٪ WER و ISAVIA: 8.8٪ → 11.2٪ مشاهده کردیم. این عمدتاً برای افزودن داده هایی است که با NATS و ISAVIA مطابقت ندارند. برعکس، کاهش WER قابل‌توجهی در مجموعه‌های دامنه‌های جزئی وجود دارد، ATCO2-Test: 34.6٪ → 23.3٪ و LiveATC-Test 39.8٪ → 31.1٪

بطور خلاصه مجموعه آزمون NATS با هفت درصد کاهش WER نسبی بدتر نسبت به حالت بدون داده های افزونه، تحت تأثیر افزودن داده های جزئی در دامنه قرار گرفت (این درصد برای ISAVIA حدود یک درصد است). با این وجود، مجموعه‌های تست چالش برانگیز به طور چشمگیری بهبود یافتند، یعنی ATCO2-Test، ۴۳ درصد و LiveATC-Test، ۳۳ درصد WER نسبی بهبود داشتند.

### ۳.۵ مدل های از پیش آموزش دیده چند زبانه چه کمکی می کند؟

اگر w2v2-L-60k+ و w2v2-XLS-R را مقایسه کنیم که از تنظیمات دقیق و رمزگشایی beam serach با مدل زبانی استفاده می کنند، نرخ خطای کلمه نسبی در ISAVIA، ATCO2-Test و LiveATC-Test، به ترتیب ۸.۸ درصد، ۶.۶ درصد و ۵.۸ درصد می باشد (بدون پیشرفت در NATS). در عین حال بهبود قابل توجهی در چالش‌برانگیزترین مجموعه‌های تست (SNR: 5- 10 dB) که حاوی گفتار انگلیسی لهجه دار هستند، به عنوان مثال، ATCO2-Test و LiveATC-Test مشاهده می‌شود. بنابراین، مدل‌های پیش‌آموزش‌شده چند زبانه، در مقایسه با مدل‌های پیش‌آموزش‌شده تک‌زبانه، عملکرد کمی را افزایش می‌دهند.

### ۴.۵ برای تنظیم دقیق مدل های **Wav2Vec2** و **XLS-R** به چه مقدار داده نیاز است؟

ما تأثیر مقادیر مختلف از داده های تنظیم دقیق را در طول مرحله تنظیم دقیق روی نرخ خطاهای کلمه، بررسی می کنیم (شکل ۱). همه آزمایش ها بر اساس قوی‌ترین مدل انتها به انتها از جدول ۳ هستند، یعنی w2v2-L-60K WER ها با رمزگشایی حریصانه به دست می آیند، یعنی هیچ مدل زبانی یا اطلاعات متنی صریحا اضافه نمی شود. ما ۱۸ مدل را با تغییر دادن مجموعه داده های آموزشی (یا NATS یا ISAVIA) و مقدار نمونه ها، تنظیم دقیق می کنیم. ما در ابتدا سناریوی یادگیری چند شات (بدترین حالت) را آزمایش کردیم، که در آن تنها ۱۰۰ گفته برچسب دار (۵ دقیقه) برای تنظیم دقیق استفاده شد، و WER ۴۰ درصد برای ISAVIA و ۹.۴۳ درصد برای NATS به دست آورد. علاوه بر این، ۵۰ کاهش WERنسبی با افزایش مقیاس داده های تنظیم دقیق به ۵۰ دقیقه (۸۰۰ گفته) به دست می آید. به طور دقیق، ٪22.7 ← ٪43.9 NATS و WER ٪21.3 ← ٪40.6 ISAVIA. در نهایت، اگر از تمام داده های موجود (حدود ۱۴ ساعت) استفاده شود، به ترتیب به ۸.۸ درصد و ۸.۶ درصد WER برای ISAVIA و NATS می رسیم. این نشان دهنده یک WER نسبی ۸۰ درصد در مقایسه با حالت اولیه یا بدترین حالت (۱۰۰ عبارت) است. از دیگر نتایج اینکه با حدود ۸ ساعت (حدود ۸۰۰۰ عبارت) w2v2-L-60K عملکرد تشخیص خودکار گفتار مبتنی بر هیبریدی را شکست می دهد (که از چهار برابر بیشتر داده های آموزشی استفاده می کند

آیا تشخیص خودکار گفتار real-time در معماری های انتها به انتها، به عنوان مثال، Wav2Vec2 امکان پذیر است؟ ما هر شش مدل از جدول ۳ را در حالت streaming در یک پردازنده گرافیکی NVIDIA GeForce GTX 1080 Ti میان رده آزمایش می کنیم. تأخیرات شامل گذر رو به جلوی مدل، رمزگشایی beam search و رمزگذاری (detokenization) است. نتایج اصلی در جدول ۳ (ستون آخر) گزارش شده است. تأخیر گذر رو به جلو مدل‌های Wav2Vec2 و XLS-R به طور کلی کمتر از ۱۰۰ میلی‌ثانیه است. به عنوان مثال، مدل‌های w2v2-B/L/L-60k و w2v2-XLS-R زمانی که رمزگشایی حریصانه انجام می‌دهند تأخیر کمتر از ۴۰ میلی‌ثانیه دارند. اگر از رمزگشایی beam search با مدل زبانی ۴تایی استفاده شود، تاخیر تقریباً دو برابر می شود. این تحقیق تخریب WER ناشی از استفاده از مدل‌های انتها به انتها در حالت streaming

را پوشش نمی‌دهد.

## ۶ نتیجه‌گیری

این مقاله استحکام مدل‌های Wav2Vec2 از پیش آموزش دیده را در downstream برای کنترل ترافیک هوایی ارزیابی می‌کند. آزمایش‌های ما پیشرفت‌های بزرگی را در تشخیص Wav2Vec2 و XLS-R در مقایسه با تشخیص خودکار گفتار مبتنی بر هیبریدی نشان می‌دهند. از نظر کمی، بین ۲۰ تا ۴۰ درصد کاهش WER نسبی در مجموعه‌های آزمایشی ISAVIA، NATS و از مجموعه‌های چالش‌برانگیز چند لهجه‌ای مانند ATCO2- Test و LiveATC-Test به دست آمد. علاوه بر این، ما نشان دادیم که Wav2Vec2 از پیش آموزش دیده، یک مرحله تنظیم دقیق سریع با مقادیر کمی از داده های سازگار را امکان پذیر می کند، به عنوان مثال، تنظیم دقیق ۵ دقیقه ای، مدلی که WERهای ۴۰ درصد برای ISVAIA و ۹.۴۳ درصد را برای NATS به دست می آورد. علاوه بر این، ما نشان دادیم که حداقل ۴ ساعت از داده‌های درون دامنه، WER قابل قبولی در حدود ۱۰ درصد برای ضبط‌های ISAVIA و NATS ارائه می‌کنند و با استفاده از داده‌های دو برابر بیشتر (یعنی ۸ ساعت) عملکردشان از تشخیص خودکار گفتار مبتنی بر هیبریدی پیشی می‌گیرد. در نهایت، ما اعداد قابل رقابتی در تأخیر برای مدل‌های Wav2Vec2 و XLS-R در یک GPU میان‌رده به‌دست آوردیم، یعنی حدود ۴۰/۸۰ میلی‌ثانیه با مدل زبانی در رمزگشایی جستجوی حریصانه و beam search.

## مراجع

[۱] J. Zuluaga-Gomez et al., “How Does Pre-trained Wav2Vec2.0 Perform on Domain Shifted ASR? An Extensive Benchmark on Air Traffic Control Communications.” arXiv, Mar. 31, 2022. doi: 10.48550/arXiv.2203.16822.