مقاله پژوهشی

ارائه دو روش داده‌افزایی برای بازشناسی گفتار با دادگان محدود: پوشاندن تدریجی و پوشاندن آگاه از فراوانی کلمات

مریم اسدالله‌زاده کرمانشاهی، احمد اکبری ازیرانی و بابک ناصرشریف

چکیده: کمبود داده، چالش اصلی بازشناسی گفتار مبتنی بر شبکه‌های عصبی عمیق است و داده‌افزایی یک راه‌حل مؤثر برای این مسئله می‌باشد. این مقاله ضمن ارائه طبقه‌بندی جامع روش‌های داده‌افزایی در بازشناسی گفتار، به بررسی اثربخشی مهم‌ترین روش‌های این حوزه یعنی روش‌های مبتنی بر پوشاندن در شرایط محدودیت دادگان می‌پردازد. روش‌های مورد بررسی دو روش قدرتمند SpecAugment و پوشاندن کلمه هستند. این روش‌ها علی‌رغم کارایی اثبات‌شده در شرایط دادگان فراوان، در شرایط دادگان محدود، کمتر مطالعه شده‌اند. در تحقیق حاضر، پس از تحلیل معایب روش پوشاندن کلمه در شرایط دادگان محدود، دو روش نوآورانه برای رفع این ایرادات ارائه می‌دهیم: (۱) پوشاندن تدریجی که آموزش را با پوشاندن در سطح فریم آغاز و سپس به پوشاندن کلمه تغییر می‌دهد؛ (۲) پوشاندن آگاه از فراوانی کلمات که ابتدا کلمات پرتکرار و سپس کلمات کم‌تکرار پوشانده می‌شوند. آزمایشات روی مجموعه ۱۰۰ ساعتی پیکره LibriSpeech نشان می‌دهد روش پیشنهادی اول به WER ۶٫۸٪ در مجموعه تمیز و ۱۸٫۲٪ در مجموعه چالش‌برانگیز رسیده که به‌ترتیب ۶٫۸٪ و ۴٫۲٪ بهبود نسبت به روش رقابتی SpecAugment حاصل کرده است. روش پیشنهادی دوم نیز به WER ۶٫۶٪ در مجموعه تمیز و ۱۷٫۳٪ در مجموعه چالش‌برانگیز رسیده که به‌ترتیب ۹٫۶٪ و ۸٫۹٪ بهبود نسبت به SpecAugment کسب کرده است.

***کلیدواژه:* بازشناسی گفتار، پوشاندن کلمات، داده‌افزایی، دادگان محدود.**

۱- مقدمه

تحقیق در زمینه بازشناسی خودکار گفتار[1] (ASR) در سال‌های اخیر رشد قابل توجهی داشته است و بسیاری از شرکت‌های تحقیقاتی مطرح دنیا در این زمینه سرمایه‌گذاری کرده‌اند. این فناوری پایه‌گذار بسیاری از کاربردها در تعامل انسان و کامپیوتر است [۱]. تاریخچه ASR به مدل‌های مخفی مارکف[2] (HMM) [۲] و مدل‌های ترکیبی مبتنی بر مدل مخفی مارکف-شبکه عصبی عمیق[3] (HMM-DNN) [۳] تا [۵] برمی‌گردد و به‌تازگی به مدل‌های انتها به انتها[4] (E2E) مانند مدل‌های ASR مبتنی بر ترنسفورمر[5] [۶] و [۷] و مدل‌های بنیادین گفتار[6] [۸] و [۹] پیشرفت کرده است.

دلیل عمده‌ی پیشرفت سیستم‌های بازشناسی گفتار استفاده از شبکه‌های عصبی عمیق (DNN) با تعداد پارامترهای زیاد و حجم بالای داده‌های آموزشی است. برای آموزش مدل آکوستیکی[7] ساعت‌ها داده‌ی گفتاری همراه با متن معادل آن نیاز است. جمع‌آوری داده گفتار-متن از نظر هزینه زمانی، مالی و نیروی انسانی کار سختی می‌باشد. علاوه بر این، داده‌های گفتاری باید از نظر گوینده، سن، جنسیت، لهجه و شرایط محیطی تنوع کافی داشته باشند.

در حال حاضر سیستم‌های مرز دانش[8] برای بازشناسی گفتار، سیستم‌های انتها به انتها هستند که برای دستیابی به دقت بالا به حجم زیادی داده آموزشی نیاز دارند [۱]. این مدل‌ها در شرایط کمبود دادگان، به‌ویژه در زبان‌های کم-منبع[9]، مستعد بیش‌برازش[10] هستند. بنابراین، استفاده از حجم مناسب و کافی داده آموزشی برای بهره‌گیری مؤثر از شبکه‌های عصبی عمیق یک ضرورت به نظر می‌رسد.

تأثیر میزان داده در دقت بازشناسی گفتار: برخلاف سیستم‌های سنتی HMM و HMM-DNN که پس از یک آستانه مشخص، افزایش داده آموزشی تأثیر چندانی بر عملکردشان ندارد، مدل‌های انتها به انتها (E2E) با افزایش حجم داده، همچنان بهبود قابل‌توجهی در دقت نشان می‌دهند [۴]. شکل ۱ این تفاوت اساسی را در سیستم بازشناسی گفتار به‌خوبی نمایش می‌دهد، به‌طوری که با افزایش داده آموزشی، نرخ خطا در مدل E2E، به‌طور مداوم کاهش می‌یابد [۱۰]. بنابراین مسئله‌ی کمبود دادگان یک مسئله واقعی و جدی در زمینه بازشناسی گفتار و یکی از موضوعات مهم پژوهش است.

حل مسئله کمبود دادگان در بازشناسی گفتار: بسیاری از تحقیقات اخیر در تلاش هستند که با وجود حجم کم داده آموزشی با به‌کارگیری روش‌ها و الگوریتم‌های خاصی به دقت بالایی برسند. از جمله این روش‌ها می‌توان به روش‌های داده‌افزایی[11] [۱۱] تا [۱۴]، یادگیری نیمه‌نظارتی[12]

این مقاله در تاریخ ۱۳ خرداد ماه ۱۴۰۳ دریافت و در تاریخ ۲ آبان ماه ۱۴۰۴ بازنگری شد.

مریم اسدالله‌زاده کرمانشاهی، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران، (email: m_asadolahzade@comp.iust.ac.ir).

احمد اکبری ازیرانی (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران، (email: akbari@iust.ac.ir).

بابک ناصرشریف، دانشکده مهندسی کامپیوتر، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران، (email: bnasersharif@kntu.ac.ir).

1. Automatic Speech Recognition
2. Hidden Markov Model
3. Deep Neural Network
4. End to End
5. Transformer
6. Speech Foundation Models
7. Acoustic Model
8. State of the Art
9. Low-Resource
10. Overfit
11. Data Augmentation
12. Semi-supervised Learning

[۱۵] و [۱۶]، یادگیری خودنظارتی[1] [۷] و [۱۷]، یادگیری انتقالی[2] [۱۸] تا [۲۰] و یادگیری چندزبانی[3] [۲۱] و [۲۲] اشاره کرد. این روش‌ها به داده داده آموزشی برچسب‌خورده کمتری نیاز دارند و نه تنها در بازشناسی گفتار برای زبان‌های کم- منبع بلکه برای زبان‌های غنی مانند انگلیسی نیز مفید هستند.

از میان این روش‌ها، داده‌افزایی به‌دلیل سادگی پیاده‌سازی، کارایی محاسباتی و انعطاف‌پذیری بالا از اهمیت ویژه‌ای برخوردار است. این روش با اعمال تغییرات بر روی داده‌های موجود، نمونه‌های جدیدی از داده‌ها تولید می‌کند و باعث افزایش حجم داده‌ها و ایجاد تنوعات جدید در داده آموزشی می‌شود که به افزایش مقاومت و تعمیم‌پذیری مدل کمک می‌کند. داده‌افزایی به‌عنوان یک تنظیم‌کننده[4] عمل کرده و امکان وقوع بیش‌برازش را کاهش داده و عملکرد مدل را در زمان استنتاج و آموزش بهبود می‌بخشد [۱۱] و [۲۳].

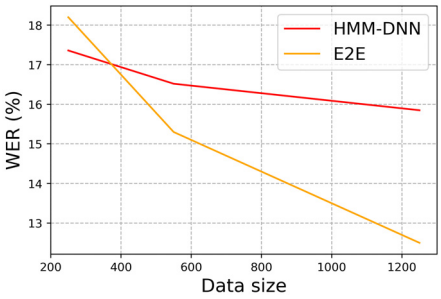
یکی از دستاوردهای مهم داده‌افزایی در ASR، معرفی روش SpecAugment [۱۱] است که با اعمال سه تغییر مختلف بر روی طیف‌نگار[5] شامل پوشاندن زمانی[6]، پوشاندن فرکانسی[7] و پیچش زمانی[8] به زمانی[8] به بهبود مدل کمک می‌کند. روش داده‌افزایی SpecAugment چندین سال است که یک مؤلفه و جزء ثابت در آموزش مدل‌ها برای وظایف مختلف حوزه گفتار باقی مانده است و همچنان به‌عنوان معیار پایه و مرجع مهمی برای مقایسه در تحقیقات جدید محسوب می‌شود. این روش از جمله روش‌های مبتنی بر پوشاندن و در سطح فریم است. در این دسته، روش دیگری [۲۴] نیز وجود دارد که عمل پوشاندن را در سطح واحد کلمه انجام می‌دهد و توانسته است در شرایط دادگان زیاد و در مقیاس ۹۶۰ ساعت داده آموزشی دقت بالایی بدست آورد.

با این حال، در تحقیقات فعلی، کارایی روش‌های مطرح داده‌افزایی در شرایط کمبود دادگان به‌طور جامع مورد ارزیابی قرار نگرفته است. این مسئله بسیار حائز اهمیت است که روش‌های داده‌افزایی بتوانند حتی با میزان محدود داده آموزشی نیز عملکرد مطلوبی داشته باشند. در این پژوهش، به بررسی دقیق این موضوع پرداخته‌ایم.

در طی آزمایش‌ها دریافتیم که روش SpecAugment [۱۱] در هر دو شرایط دادگان زیاد و دادگان کم عملکرد مطلوبی دارد، اما روش پوشاندن مبتنی بر واحد کلمه [۲۴] وضعیت متفاوتی را نشان می‌دهد. این روش علی‌رغم برتری نسبت به پوشاندن مبتنی بر فریم (SpecAugment) در شرایط دادگان زیاد، عملکرد به‌مراتب ضعیف‌تری در شرایط دادگان کم از خود نشان می‌دهد. این تناقض، نقطه آغاز پژوهش حاضر بوده است.

در این مقاله، پس از تحلیل علت این موضوع، دو روش نوآورانه برای مقابله با این چالش و بهبود عملکرد ASR در شرایط کمبود دادگان ارائه می‌کنیم که عملکرد بهتری نسبت به روش‌های موجود نشان داده‌اند:

الف) روش پوشاندن تدریجی[9]: روش پیشنهادی اول ما بر پایه یادگیری مبتنی بر برنامه آموزشی[10] (CL) [۲۵] بنا شده است که آموزش مدل ابتدا با داده‌های آسان انجام شود و سپس با داده‌های دشوار ادامه



شکل ۱: تأثیر افزایش میزان داده آموزشی (برحسب ساعت) بر نرخ خطای بازشناسی در مدل‌های HMM-DNN و E2E [۱۰].

یابد. برخلاف رویکردهای فعلی، فرآیند آموزش ابتدا با اعمال داده‌افزایی در سطح فریم (وظیفه ساده‌تر) آغاز شده و سپس به داده‌افزایی در سطح کلمه (وظیفه پیچیده‌تر) اعمال می‌شود. این روش به مدل امکان می‌دهد تا ابتدا با وظیفه ساده‌تر آموزش داده شده و برای مواجهه با پیچیدگی‌های بیشتر آمادگی پیدا کند.

ب) روش پوشاندن آگاه از فراوانی کلمات[11]: در روش پیشنهادی دوم، برخلاف پژوهش‌های پیشین که بر انتخاب تصادفی کلمات برای پوشاندن تکیه داشته‌اند، یک الگوریتم انتخاب هدفمند مبتنی بر فراوانی کلمات طراحی و پیاده‌سازی کرده‌ایم. در این رویکرد، ابتدا کلمات پرتکرار جمله پوشانده می‌شوند و با پیشرفت آموزش، کلمات کم‌تکرارتر برای پوشاندن انتخاب می‌شوند. این استراتژی با شناسایی و اولویت‌بندی کلمات برای پوشاندن، به بهینه‌سازی فرآیند یادگیری در شرایط محدودیت دادگان کمک می‌کند.

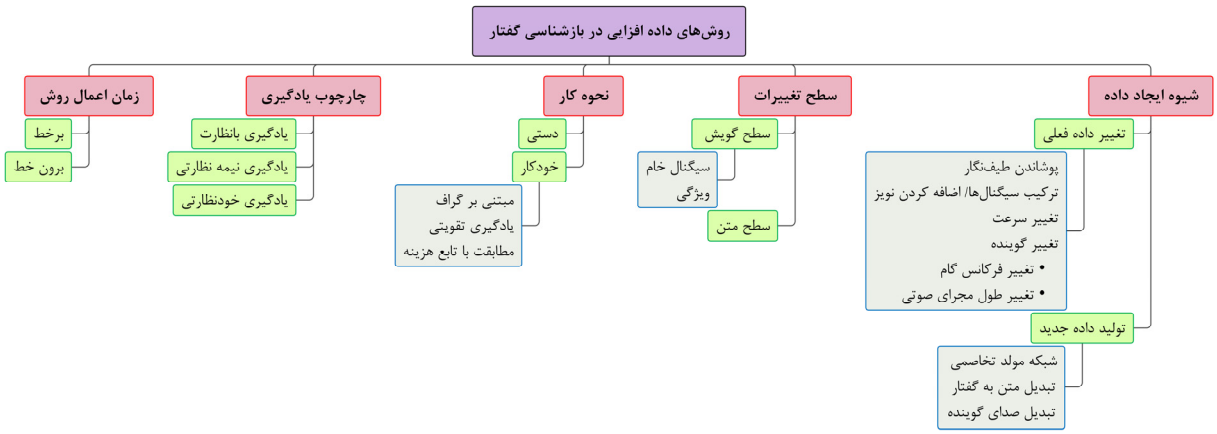
نتایج آزمایش‌های ما نشان می‌دهد که هر دو روش پیشنهادی، نسبت به روش‌های SpecAugment و پوشاندن کلمه در شرایط محدودیت دادگان بهبود قابل‌توجهی در دقت بازشناسی بدست آورده‌اند.

دستاوردهای اصلی این پژوهش به شرح زیر است:

- ارائه یک طبقه‌بندی جامع از روش‌های داده‌افزایی در حوزه بازشناسی گفتار براساس معیارهای کلیدی، همراه با تحلیل مقایسه‌ای از نقاط قوت و ضعف هر روش.
- بررسی و مقایسه دقیق عملکرد روش‌های پیشرفته داده‌افزایی مبتنی بر پوشاندن در شرایط محدودیت دادگان آموزشی (دادگان ۱۰۰ ساعتی).
- معرفی روش «پوشاندن تدریجی» که کارایی سیستم بازشناسی گفتار را در شرایط کمبود دادگان آموزشی بهبود می‌بخشد.
- معرفی روش «پوشاندن آگاه از فراوانی کلمات» که با انتخاب هدفمند کلمات جهت پوشاندن و جایگزینی روش‌های تصادفی متداول، دقت بازشناسی را به میزان قابل توجهی افزایش داده است.

این مقاله در ۶ بخش تنظیم شده است. پس از بخش اول که مقدمه است، در بخش ۲، مروری بر کارهای گذشته در زمینه روش‌های داده‌افزایی در بازشناسی گفتار خواهیم داشت و طبقه‌بندی و مقایسه‌ای از این روش‌ها ارائه خواهیم کرد. در بخش ۳ روش‌های داده‌افزایی را در شرایط دادگان محدود بررسی خواهیم کرد و روش‌های پیشنهادی را شرح خواهیم داد. در بخش ۴ تنظیمات آزمایش و پارامترها، شیوه ارزیابی روش‌ها و پیکره معرفی خواهند شد. در بخش ۵ نتایج حاصل از روش‌های پیشنهادی ارائه و مورد تحلیل قرار می‌گیرد. بخش ۶ نیز به جمع‌بندی و نتیجه‌گیری اختصاص خواهد داشت.

1. Self-supervised Learning
2. Transfer Learning
3. Multi-lingual
4. Regularizer
5. Spectrogram
6. Time masking
7. Frequency Masking
8. Time Warping
9. Gradual Masking
10. Curriculum Learning
11. Word Frequency-Aware Masking



شکل ۲: طبقه‌بندی روش‌های داده‌افزایی در بازشناسی گفتار.

۲- مرور و طبقه‌بندی روش‌های داده‌افزایی در بازشناسی گفتار

در این بخش روش‌های داده‌افزایی را از ۵ منظر مختلف طبقه‌بندی و مرور کرده‌ایم که در شکل ۲ قابل مشاهده است که معیارهای طبقه‌بندی با مستطیل قرمز نشان داده شده‌اند و در ادامه نیز توضیح داده می‌شوند.

۲-۱ شیوه ایجاد داده: داده‌افزایی مبتنی بر تغییر داده فعلی یا تولید داده جدید

بیشتر روش‌های داده‌افزایی از نوع «تغییر داده فعلی» هستند و روی همان داده گفتاری موجود تغییراتی اعمال می‌کنند. روش‌های پوشاندن طیف‌نگار [۱۱]، ترکیب با سیگنال دیگر [۲۶]، اضافه کردن نویز، تغییر سرعت [۱۱] و [۱۲]، تغییر گوینده با تغییر فرکانس گام و طول مجرای صوتی [۲۷] و [۲۸] در این دسته قرار می‌گیرند. دسته‌ی دیگر با کمک مدل‌های شبکه مولد تخاصمی[1] (GAN)، تبدیل متن به گفتار[2] (TTS) و یا روش‌های تبدیل صدای گوینده، یک داده گفتاری جدید تولید و سنتز می‌کنند [۲۹] و [۳۰]. روش‌های دسته اول از نظر پیاده‌سازی و اجرا ساده هستند و روش‌های دسته دوم هر چند می‌توانند تنوع بیشتری ایجاد کنند اما مشکلاتی مانند پیچیدگی آموزش، نیاز به منابع سخت‌افزاری و امکان تولید گفتار غیرطبیعی و مصنوعی دارند.

۲-۲ سطح تغییرات: داده‌افزایی در سطح گویش و متن

تغییرات داده‌افزایی در بازشناسی گفتار می‌تواند در دو سطح گویش[3] یا متن مربوط به آن اعمال شود.

داده‌افزایی در سطح گویش: بسیاری از روش‌های ارائه شده در این حوزه، تغییرات را در سطح گویش اعمال می‌کنند و برچسب متنی را تغییر نمی‌دهند. داده‌افزایی در سطح گویش نیز خود می‌تواند در دو سطح سیگنال خام یا سطح ویژگی اعمال شود:

- داده‌افزایی در سطح سیگنال خام: روش‌هایی که تغییرات را روی سیگنال خام اعمال می‌کنند مانند تغییر سرعت سیگنال [۱۲] و اضافه کردن نویز به سیگنال.
- داده‌افزایی در سطح ویژگی: در این شیوه، تغییرات روی ویژگی‌های استخراج شده از سیگنال مانند طیف‌نگار اعمال می‌شود. بسیاری از روش‌ها مانند [۱۱]، [۲۴] و [۲۶] در این دسته قرار می‌گیرند. مزیت داده‌افزایی در سطح ویژگی این است که برخلاف روش‌های مبتنی بر سیگنال خام، نیازی به استخراج مجدد ویژگی نیست که همین امر باعث افزایش سرعت اجرای آموزش می‌شود و همچنین امکان اعمال روش داده‌افزایی را به‌صورت برخط فراهم می‌کند.

داده‌افزایی در سطح متن: این روش‌ها جملات متنی جدیدی را ایجاد می‌کنند از جمله این روش‌ها می‌توان به تولید داده با کمک قواعد گرامری [۳۱] و جایگزینی کلمات جمله از روی یک لغت‌نامه صوتی[4] [۳۲] اشاره کرد.

هدف داده‌افزایی در سطح گویش ایجاد تنوعات گفتاری و هدف داده‌افزایی در سطح متن ایجاد تنوعات زبانی با ایجاد ساختارهای گرامری و لغوی جدید است.

۲-۳ نحوه کار: داده‌افزایی خودکار و دستی

اکثر روش‌های موجود در داده‌افزایی گفتار از نوع دستی هستند بدین معنا که پارامترها و سیاست‌های هر روش داده‌افزایی توسط انسان تصمیم‌گیری و ارزیابی می‌شوند ولی در روش‌های خودکار[5] تلاش می‌شود می‌شود این موارد به‌صورت بهینه توسط یک الگوریتم یا مدل دیگر تصمیم‌گیری و تعیین شود.

برای روش‌های داده‌افزایی خودکار از تکنیک‌هایی همچون روش‌های مبتنی بر گراف [۳۳]، یادگیری تقویتی [۳۴] و مطابقت با تابع هزینه [۳۵] و [۳۶] استفاده شده است تا پارامترها و زیرمجموعه بهینه از روش‌های داده‌افزایی تعیین شوند.

هرچند روش‌های خودکار می‌توانند در اکثر موارد پاسخ بهینه‌ای پیدا کنند ولی نیاز به فرآیند آموزش مجزا و پیاده‌سازی پیچیده برای الگوریتم‌ها دارند و نسبت به روش معمولی و دستی هزینه‌بر هستند.

۲-۴ چارچوب یادگیری: داده‌افزایی در شیوه‌های مختلف آموزش مدل

داده‌افزایی گفتار یک روش منعطف است که در همه شیوه‌های آموزش مورد استفاده قرار می‌گیرد.

داده‌افزایی می‌تواند علاوه بر شیوه آموزش بانظارت، در آموزش نیمه‌نظارتی نیز مؤثر باشد و روی گفتار با و بدون برچسب متنی اعمال شود و همچنین برای تولید نسخه نویزی شده از داده اصلی نیز استفاده شود [۱۵].

1. Generative Adversarial Network
2. Text to Speech
3. Utterance-Level
4. Audio Dictionary
5. Automatic

انجام داده‌افزایی در آموزش خودنظارتی در مرحله پیش‌آموزش شبکه با گفتار بدون متن [۳۷]، در مرحله یادگیری متقابل[1] [۳۸] و یا در مرحله تنظیم دقیق[2] نیز باعث بهبود دقت می‌شود.

۲-۵ زمان اعمال روش: روش‌های برخط و برون‌خط در داده‌افزایی

روش‌های برخط[3]: در این روش تغییرات حین آموزش مدل اعمال می‌شوند. یک مثال معروف روش SpecAugment است که عمل پوشاندن را روی طیف‌نگار اعمال می‌کند. مزیت این دسته روش‌ها نیاز به فضای ذخیره‌سازی کمتر و امکان ایجاد تنوع بیشتر در هر دوره آموزشی[4] است، اما اگر هزینه محاسباتی روش، بالا باشد می‌تواند باعث سربار محاسباتی هنگام آموزش شود.

روش‌های برون‌خط[5]: در این روش، تغییرات روی داده‌ها قبل از آموزش مدل انجام می‌شود مانند روش [۱۲] که عمل تغییر سرعت روی سیگنال خام اعمال می‌شود و پس از این تغییر، ویژگی‌ها باید مجدداً از سیگنال تغییریافته، استخراج شوند. بنابراین باید تغییرات قبل از آموزش مدل اعمال و داده‌های تغییریافته در حافظه دیسک ذخیره شوند و نکته منفی دیگر این روش امکان ایجاد تنوعات محدود در داده است.

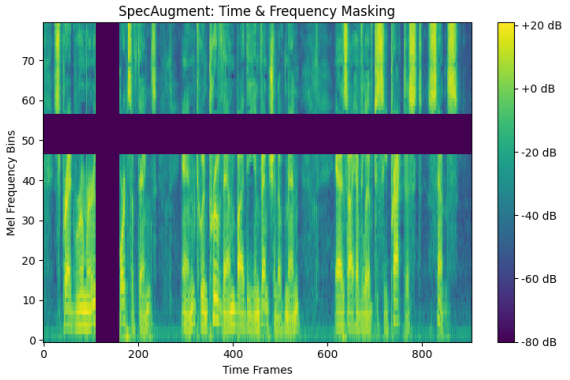
۳- روش‌های داده‌افزایی مبتنی بر پوشاندن و روش‌های پیشنهادی

در این بخش، ابتدا به معرفی و بررسی روش‌های داده‌افزایی مبتنی بر پوشاندن می‌پردازیم. از میان روش‌های موجود، ما بر SpecAugment تمرکز کرده‌ایم زیرا این روش پایه و اساس بسیاری از روش‌های داده‌افزایی در حوزه پردازش گفتار است. روش SpecAugment همچنان در مدل‌های پیشرفته گفتاری مورد استفاده قرار می‌گیرد و معیار مقایسه‌ای برای روش‌های جدید محسوب می‌شود. همچنین مکانیزم پوشاندن مشابه آنچه در SpecAugment می‌بینیم، در مدل‌های زبانی بزرگ[6] [۳۹] نیز به عنوان تکنیک اصلی آموزش مورد استفاده قرار می‌گیرد، که نشان‌دهنده اهمیت و تأثیرگذاری این رویکرد است.

نکته قابل توجه این است که اکثر مطالعات و تحقیقات در زمینه داده‌افزایی برای بازشناسی گفتار عمدتاً روی دادگان با مقیاس بالا (مانند ۹۶۰ ساعت داده آموزشی از پیکره LibriSpeech [۴۰])، متمرکز بوده‌اند. درحالیکه بررسی و تحلیل عملکرد روش‌های داده‌افزایی در شرایط دادگان محدود (به‌طور مثال در مقیاس ۱۰۰ ساعت) موضوع مهمی است که کمتر به آن پرداخته شده است. زیرا اهمیت روش‌های داده‌افزایی در شرایط دادگان محدود بیشتر نمود پیدا می‌کند و این موضوع محور اصلی پژوهش حاضر است.

در این مقاله، ما به بررسی این سؤال می‌پردازیم که آیا روش‌هایی که در مقیاس بالا عملکرد خوبی دارند، در شرایط کمبود دادگان نیز به همان اندازه کارآمد هستند یا خیر. برای این منظور، دو روش مهم از دسته روش‌های مبتنی بر پوشاندن یعنی روش SpecAugment و روش پوشاندن کلمه را مورد آزمایش و تحلیل قرار می‌دهیم. در ادامه، پس از بررسی عملکرد این روش‌ها در شرایط دادگان محدود، روشی را برای حل

1. Contrastive Learning
2. Fine-tune
3. Online
4. Epoch
5. Offline
6. Large Language Models



شکل ۳: نتیجه‌ی اِعمال پوشاندن زمانی و فرکانسی روش داده‌افزایی SpecAugment روی طیف‌نگار.

مشکل عدم کارایی روش پوشاندن واحد کلمه ارائه می‌دهیم و یک روش تکمیلی را نیز در جهت بهبود بیشتر آن معرفی خواهیم کرد.

۳-۱ داده‌افزایی SpecAugment

روش SpecAugment [۱۱] یکی از موفق‌ترین روش‌های داده‌افزایی در زمینه‌ی بازشناسی گفتار تاکنون است. روش SpecAugment، روشی با هزینه‌ی محاسباتی کم و در عین حال ساده از نظر پیاده‌سازی می‌باشد. این روش برخلاف برخی روش‌های گذشته که تغییرات را روی خود سیگنال گفتار اعمال می‌کردند، تغییرات را روی ویژگی‌های استخراجی از آن یعنی لگاریتم طیف‌نگار مل[7] سیگنال ورودی اعمال می‌کند. بنابراین در در این روش نیازی به استخراج مجدد ویژگی نداریم زیرا تغییرات به‌طور مستقیم روی ویژگی‌ها اعمال می‌شود نه سیگنال و درنتیجه این روش به‌صورت برخط و حین آموزش مدل قابل اعمال است. روش SpecAugment شامل سه عملیات اصلی است:

- پوشاندن زمانی: پوشاندن بلوک‌های متوالی در محور زمان
- پوشاندن فرکانسی: پوشاندن بلوک‌های متوالی در محور فرکانس
- پیچش زمانی: کشیدن یا فشرده کردن طیف‌نگار در محور زمان

نمونه‌ی تغییرات پوشاندن زمانی و فرکانسی در شکل ۳ نمایش داده شده است.

عملیات پوشاندن روی طیف‌نگار به‌صورت شکل ۴ تعریف می‌شوند.

در این پژوهش، ما بر روی دو عمل اصلی پوشاندن زمانی و فرکانسی تمرکز می‌کنیم که بیشترین تأثیر را در بهبود مدل‌های ASR دارند. با توجه به اینکه روش پیچش زمانی، روشی کم اثرتر نسبت به دو روش پوشاندن است و پیچیدگی محاسباتی بیشتری دارد [۱۱]، از پرداختن به جزئیات آن خودداری می‌کنیم.

۳-۲ روش پوشاندن کلمه

روش پوشاندن کلمه روشی است که در [۲۴] معرفی شد. برخلاف روش پوشاندن زمانی موجود در الگوریتم SpecAugment که به‌صورت تصادفی تعدادی فریم از طیف‌نگار را می‌پوشاند در روش پوشاندن کلمه، عمل پوشاندن در محور زمان، روی واحدهای کلمه انجام می‌گیرد. در این روش درصدی از کلمات به‌صورت تصادفی انتخاب و قسمت‌های مربوط به آن‌ها در طیف‌نگار پوشانده می‌شوند. در الگوریتم ۲، مراحل روش پوشاندن کلمه آمده است (شکل ۵).

7. Log Mel Spectrogram

ورودی: طیف‌نگار X با ابعاد $\tau \times \upsilon$ (τ تعداد فریم‌های زمانی و υ تعداد باندهای فرکانسی)
خروجی: طیف‌نگار تغییریافته X_{aug}
پارامترها: تعداد پوشاندن در محور زمان m_T، تعداد پوشاندن در محور فرکانس m_F، حداکثر اندازه پوشاندن زمانی T، حداکثر اندازه پوشاندن فرکانسی F

۱. $X_{aug} \leftarrow X$
پوشاندن فرکانسی:
۲. برای i از ۱ تا m_F:
۳. $f \sim \text{Uniform}(0, F)$ // انتخاب اندازه پوشاندن در محور فرکانس
۴. $f_0 \sim \text{Uniform}(0, \upsilon - f)$ // انتخاب نقطه شروع پوشاندن در محور فرکانس
۵. $X_{aug}[:, f_0 : f_0 + f] \leftarrow 0$ // اِعمال پوشاندن فرکانسی
پوشاندن زمانی:
۶. برای j از ۱ تا m_T:
۷. $t \sim \text{Uniform}(0, T)$ // انتخاب اندازه پوشاندن زمانی
۸. $t_0 \sim \text{Uniform}(0, \tau - t)$ // انتخاب نقطه شروع پوشاندن در محور زمان
۹. $X_{aug}[t_0 : t_0 + t, :] \leftarrow 0$ // اِعمال پوشاندن زمانی
۱۰. برگرداندن X_{aug}

شکل ۴: الگوریتم ۱، داده‌افزایی SpecAugment.

۳-۳ تحلیل عملکرد روش پوشاندن مبتنی بر واحد کلمه و واحد فریم در شرایط دادگان محدود

نتایج آزمایش‌ها در بخش ۴ نشان می‌دهد که روش پوشاندن کلمه، علی‌رغم عملکرد خوب در شرایط دادگان زیاد، در شرایط دادگان کم، چندان مؤثر نیست. روش پوشاندن در سطح کلمه عملکرد ضعیف‌تری نسبت به روش پوشاندن در سطح فریم (SpecAugment) دارد.

چرا روش پوشاندن کلمه در شرایط دادگان محدود موثر نیست؟

دلایل اصلی این موضوع را می‌توان به شرح زیر تحلیل کرد:

- **افزایش پیچیدگی وظیفه یادگیری**: پوشاندن واحدهای کلمه وظیفه‌ای دشوارتر برای مدل ایجاد می‌کند، زیرا مدل باید بافت معنایی بزرگتری را درک کند. در شرایط دادگان محدود، این سطح از پیچیدگی، چالش‌برانگیزتر می‌شود زیرا نمونه‌های کافی برای یادگیری این روابط پیچیده وجود ندارد.
- **دشواری بیش از حد در ابتدای آموزش**: یکی از مهم‌ترین عوامل محتمل برای عدم کارایی روش پوشاندن کلمه، سخت‌تر شدن وظیفه یادگیری از همان ابتدای آموزش است.
- **انتخاب تصادفی کلمات برای پوشاندن**: این موضوع به‌خصوص در مراحل اولیه آموزش، زمانی که مدل هنوز الگوهای اولیه را نیز یاد نگرفته است، می‌تواند کار یادگیری را مشکل کند.

این یافته مهم نشان می‌دهد که نمی‌توان استراتژی‌های موفق در شرایط دادگان زیاد را مستقیماً به شرایط دادگان کم تعمیم داد. بنابراین نیاز به رویکردهای جدیدی است که به‌طور خاص برای دادگان کم طراحی شده‌اند. براساس این تحلیل، در ادامه روش‌های جدیدی پیشنهاد می‌دهیم که محدودیت‌های شناسایی‌شده را برطرف می‌کنند و کارایی بهتری برای شرایط دادگان کم ارائه می‌دهند.

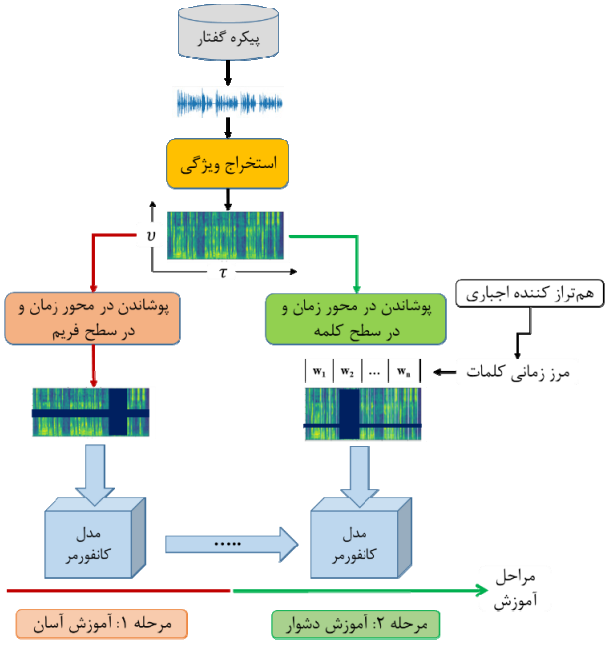
۴-۳ روش پیشنهادی اول: پوشاندن تدریجی

پیشنهاد ما برای حل مشکل مطرح شده و بهبود کارایی، یک روش دو مرحله‌ای برای آموزش مدل است که در شکل ۶ نمایش داده شده است. در این روش، به‌جای استفاده از یک روش داده‌افزایی ثابت در کل فرآیند آموزش، آموزش را به دو مرحله متوالی تقسیم می‌کنیم:

ورودی: طیف‌نگار X و متن متناظر آن $W = [w_1, w_2, \ldots, w_n]$ که شامل n کلمه است، نرخ پوشاندن
خروجی: طیف‌نگار تغییریافته

۱. ابتدا باید نگاشت بین هر کلمه w_i و بخش مربوط به آن را در طیف‌نگار مشخص کنیم:
با استفاده از نتایج هم‌ترازسازی اجباری[1]، زمان شروع t^i_{start} و زمان پایان t^i_{end} هر کلمه w_i را تعیین می‌کنیم.
۲. بخش متناظر با کلمه w_i در طیف‌نگار به‌صورت $X_{w_i} = X\left[t^i_{start} : t^i_{end}\right]$ تعریف می‌شود.
۳. تعیین تعداد کلمات برای پوشاندن: $m = \lfloor n \times \text{نرخ پوشاندن} \rfloor$
۴. تعداد m کلمه از n کلمه به‌صورت تصادفی (بدون تکرار) برای انجام پوشاندن انتخاب می‌شوند.
۵. برای هر کلمه w_i که برای پوشاندن انتخاب شده، بخش متناظر با آن در طیف‌نگار پوشانده می‌شود: $X\left[t^j_{start} : t^j_{end}, :\right] \leftarrow 0$
۶. برگرداندن طیف‌نگار تغییریافته

شکل ۵: الگوریتم ۲، روش داده‌افزایی پوشاندن کلمه.



شکل ۶: مراحل روش پیشنهادی اول: پوشاندن تدریجی.

الف) در مرحله اول مدل را با پوشاندن واحدهای در سطح فریم آموزش می‌دهیم.
ب) در مرحله دوم آموزش را با پوشاندن واحدهای کلمه ادامه می‌دهیم.

این رویکرد دو مرحله‌ای چند مزیت دارد:

- در مراحل اولیه آموزش، مدل با پوشاندن در سطح فریم (که جزئی از روش SpecAugment بوده و برای دادگان کم مناسب است) آموزش می‌بیند و با چالش و وظیفه ساده‌تر مواجه می‌شود.
- پس از کسب دانش پایه و اولیه، مدل با استفاده از پوشاندن کلمه به یادگیری ارتباطات معنایی عمیق‌تر می‌پردازد. در واقع با عمل پوشاندن کلمه، مدل را مقاوم‌تر ساخته و تلاش می‌کنیم مدل بتواند کار یادگیری زبان و مدل‌سازی زبانی را با وظیفه پیش‌بینی کلمه برای بخش‌های پوشانده شده انجام دهد.
- مدل به تدریج با چالش پیچیده‌تر مواجه می‌شود، بدون آنکه از ابتدا با وظیفه دشوار پوشاندن کلمه روبرو شود.

1. Forced Alignment

این رویکرد بر این اصل استوار است که در مراحل اولیه آموزش، وزن‌های شبکه هنوز در حال یادگیری هستند و به مقدار مناسبی همگرا نشده‌اند. با پیشرفت آموزش، وزن‌ها به مقدار بهینه نزدیک می‌شوند و در این مرحله می‌توانیم میزان داده‌افزایی و درنتیجه درجه سختی آموزش را افزایش دهیم.

ارتباط روش پوشاندن تدریجی با دیگر حوزه‌ها: روش پیشنهادی ما به حوزه وسیعی به نام یادگیری مبتنی بر برنامه آموزشی [۲۵] مرتبط است. در این روش یادگیری، در مراحل اولیه آموزش شبکه، نمونه‌های آسان‌تر به شبکه داده می‌شوند زیرا شبکه راحت‌تر این نمونه‌ها را یاد می‌گیرد و سپس نمونه‌های مشکل‌تر از نظر یادگیری به شبکه داده می‌شوند. در این حالت شبکه به دقت بهتری دست پیدا می‌کند.

در مورد کار ما، این نوع یادگیری به این صورت مورد استفاده قرار گرفته است که:

- معیار ما برای سختی یک نمونه گفتاری، میزان پیچیده شدن و سخت شدن داده از نظر یادگیری آن توسط مدل است.
- واحدهای فریم کوچکتر از واحد کلمه هستند و طبیعتاً برای مدل، یادگیری بخش‌های پوشانده شده با واحد فریم آسان‌تر از واحد کلمه است.
- در اینجا ما نمونه‌ها را به دو بخش آسان و سخت تقسیم نمی‌کنیم، بلکه با کنترل میزان داده‌افزایی، خودمان سختی و آسانی آن را تعیین می‌کنیم. بدین‌صورت که در ابتدا شدت داده‌افزایی کم باشد سپس با افزایش مراحل آموزش، شدت داده‌افزایی را زیاد کنیم.

همچنین در این روش، مرحله پوشاندن فریم را تا زمانی ادامه می‌دهیم که مقدار تابع هزینه روی مجموعه اعتبارسنجی نسبت به دوره قبل رو به کاهش باشد. درصورت مشاهده کاهش عملکرد مدل (افزایش تابع هزینه روی مجموعه اعتبارسنجی نسبت به دوره قبل) وارد مرحله آموزش مدل با پوشاندن کلمه می‌شویم.

۳-۵ روش پیشنهادی دوم: پوشاندن آگاه از فراوانی کلمات

ایده اصلی این روش، پوشاندن هدفمند کلمات هر جمله براساس میزان فراوانی آن‌ها در مجموعه داده آموزشی است. به‌جای شیوه معمول تحقیقات گذشته که مبتنی بر پوشاندن تصادفی کلمات است، در روش پیشنهادی کلمات متناسب با تعداد رخدادشان برای عمل پوشاندن اولویت‌بندی می‌شوند و روند یادگیری از شرایط آسان به سخت انجام می‌شود (شکل ۷).

تقسیم‌بندی کلمات برای پوشاندن براساس فراوانی: ابتدا میزان فراوانی کل کلمات پیکره آموزشی را اندازه‌گیری می‌کنیم. سپس کلمات هر فایل گفتاری را برحسب تعداد تکرار آن‌ها مرتب و به دو دسته مساوی تقسیم می‌کنیم:

- **مجموعه پرتکرار** (W_{freq}): نصف پرتکرارترین کلمات جمله در این دسته قرار می‌گیرند.
- **مجموعه کم‌تکرار** (W_{rare}): سایر کلمات جمله که تکرار کمتری دارند در این دسته قرار می‌گیرند.

آموزش دو مرحله‌ای پوشاندن کلمات پرتکرار سپس کم‌تکرار: روش پیشنهادی، آموزش تدریجی با دو مرحله آموزشی زیر است:

- مرحله اول: انتخاب کلمات برای پوشاندن از مجموعه پرتکرار (W_{freq}). در این مرحله، تنها کلمات پرتکرار آن جمله در فرآیند

۱. مرحله پیش‌پردازش:

۱. ۱. تشکیل جدول فراوانی کلمات:

برای هر کلمه w در مجموعه داده آموزشی:

شمارش تعداد رخدادهای w در کل پیکره و قرار دادن در $F(w)$

۱. ۲. تقسیم‌بندی کلمات هر جمله براساس فراوانی به دو مجموعه W_{freq} و W_{rare}:

برای هر جمله گفتاری S_i از مجموعه آموزشی:

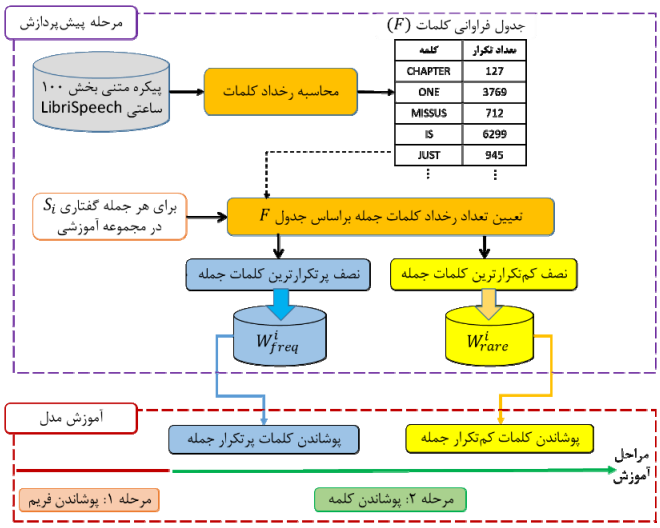
- تعیین تعداد رخداد کلمات جمله S_i از روی جدول فراوانی کلمات F
- قرار دادن نصف پرتکرارترین کلمات جمله در مجموعه کلمات پرتکرار W^i_{freq}
- قرار دادن نصف کم‌تکرارترین کلمات جمله در مجموعه کلمات کم تکرار W^i_{rare}

۲. آموزش مدل:

۲. ۱. آموزش اولیه مدل با پوشاندن کلمات پرتکرار (انتخاب کلمات برای پوشاندن از مجموعه W^i_{freq})

۲. ۲. آموزش نهایی مدل با پوشاندن کلمات کم‌تکرار (انتخاب کلمات برای پوشاندن از مجموعه W^i_{rare})

شکل ۷: الگوریتم ۳، روش پوشاندن آگاه از فراوانی کلمات.



شکل ۸: مراحل روش پیشنهادی دوم، روش پوشاندن آگاه از فراوانی کلمات.

پوشاندن شرکت می‌کنند.

همچنین برای هر دو مرحله آموزش، مدت مساوی دوره آموزشی را درنظرگرفتیم. این تقسیم‌بندی مساوی، زمان کافی برای مدل فراهم می‌کند تا هر دو دسته کلمات کم‌تکرار و پرتکرار به یک اندازه شانس پوشانده شدن و یادگیری توسط مدل را داشته باشند.

- مرحله دوم: انتخاب کلمات برای پوشاندن از مجموعه کم‌تکرار (W_{rare}). در این مرحله، پوشاندن را روی کلمات با رخداد پایین‌تر انجام می‌دهیم.

حذف کلمات خیلی نادر از فرآیند پوشاندن: کلمات با رخداد بسیار پایین از فرآیند پوشاندن حذف می‌شوند، این کلمات تنها یک نمونه در کل پیکره دارند و پوشاندن آن‌ها به معنی حذف تنها نمونه آموزشی از آن کلمه خواهد بود. همچنین این کلمات جزء کلمات خیلی خاص هستند بنابراین آن‌ها را در بافت متن نگه می‌داریم و اصلا پوشانده نمی‌شوند.

به منظور درک بهتر، مراحل روش پیشنهادی دوم در شکل ۸ نمایش داده شده است.

منطق و مزایای روش پیشنهادی پوشاندن آگاه از فراوانی کلمات: رویکرد پیشنهادی بر پایه منطق زیر استوار است و دارای مزایایی به شرح زیر می‌باشد:

- **هم‌راستایی با یادگیری CL:** ابتدا آموزش با پوشاندن کلمات پر

تکرارتر آن جمله آغاز می‌شود. این کلمات (مانند حروف اضافه، حروف ربط، افعال پرکاربرد و افعال کمکی) معمولاً نقش ساختاری در جملات دارند و به دلیل تکرار بیشتر و حضورشان در شرایط و بافت‌های مختلف، راحت‌تر قابل پیش‌بینی هستند. پس از آن، مدل با وظیفه پیچیده‌تر پوشاندن کلمات کم‌تکرار که معمولاً حاوی محتوای معنایی خاص‌تری هستند، آموزش را ادامه می‌دهد. روش پیشنهادی با افزایش سطح دشواری وظیفه (از تشخیص کلمات پرتکرار به سمت کم‌تکرار)، با اصول یادگیری CL هم‌راستا است.

- **توجه به اطلاعات متنی:** برای انتخاب کلمات جهت پوشاندن، به نوع کلمات و تعداد رخداد آن‌ها براساس شرایط پیکره آموزشی با دادگان کم، توجه می‌شود. این درحالی است که دیگر روش‌های داده‌افزایی به محتوای متنی توجه نمی‌کنند و کل فرآیند پوشاندن به‌صورت تصادفی انجام می‌گیرد و برای عمل پوشاندن، همه کلمات یکسان درنظرگرفته می‌شوند.

۳-۶ مدل آکوستیکی و شیوه آموزش

برای مدل آکوستیکی از مدل کانفورمر[1] [۶] استفاده شد که یکی از پیشرفته‌ترین معماری‌ها در حوزه بازشناسی گفتار است که با ترکیب هوشمندانه شبکه‌های عصبی پیچشی[2] و ترنسفورمرها طراحی شده است. برای ورودی x به بلوک کانفورمر، پردازش‌ها و خروجی حاصل از این بلوک یعنی y با معادلات زیر محاسبه می‌شود:

$$\tilde{x} = x + \frac{1}{2}\mathrm{FFN}(x) \quad (1)$$

$$x' = \tilde{x} + \mathrm{MHSA}(\tilde{x}) \quad (2)$$

$$x'' = x' + \mathrm{Conv}(x') \quad (3)$$

$$y = \mathrm{Layernorm}\left(x'' + \frac{1}{2}\mathrm{FFN}(x'')\right) \quad (4)$$

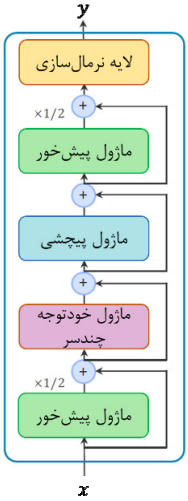
در این معادلات، FFN شبکه پیش‌خور[3]، MHSA ماژول خودتوجه چندسر[4] و Conv ماژول پیچشی است.

مزیت اصلی این ساختار در قابلیت همزمان آن برای مدل‌سازی وابستگی‌های بلندمدت (از طریق مکانیزم توجه) و استخراج ویژگی‌های محلی (از طریق ماژول پیچشی) است که موجب عملکرد برتر آن نسبت به معماری‌های پیشین می‌شود. معماری مربوط به بخش کدگذار[5] مدل کانفورمر در شکل ۹ نمایش داده شده است.

۳-۶-۱ مزایای مدل کانفورمر

در مقایسه با مدل‌های متداول قبلی مانند ترنسفورمرها، کانفورمر دقت بسیار بالاتری بدست می‌آورد که گواهی بر کارآمدی این معماری است [۶]. مدل کانفورمر به دلایل زیر یک روش قدرتمند برای بازشناسی گفتار محسوب می‌شود:

- **ساختار ساندویچی:** این مدل از ساختار ویژه‌ای به نام «ساختار ساندویچی» بهره می‌برد که به چیدمان خاص ماژول‌ها در بلوک کانفورمر اشاره دارد. در این ساختار، شبکه پیش‌خور (FFN) به دو



شکل ۹: معماری مربوط به بخش کدگذار از مدل کانفورمر [۶].

نیمه مساوی تقسیم می‌شود – یک نیمه در ابتدای بلوک و نیمه دیگر در انتهای آن قرار می‌گیرد. ماژول‌های خودتوجه چندسر (MHSA) و پیچشی بین این دو نیمه «ساندویچ» شده‌اند. این طراحی به بهبود جریان گرادیان[6] در طول آموزش کمک می‌کند و عملکرد کلی مدل را افزایش می‌دهد.

- **اتصالات باقیمانده**[7]: وجود اتصالات باقیمانده در هر مرحله باعث می‌شود آموزش شبکه‌های عمیق آسان‌تر شود و مشکل ناپدید شدن گرادیان کاهش یابد.
- **قابلیت پردازش همزمان:** برخلاف مدل‌های شبکه‌های عصبی بازگشتی[8] (RNNها)، بخش‌های توجه و پیچشی قابلیت پردازش موازی دارند که سرعت آموزش و استنتاج را افزایش می‌دهد.

۳-۶-۲ آموزش مدل آکوستیکی با روش طبقه‌بند زمانی پیوندگرا (CTC)

تابع هزینه[9] طبقه‌بند زمانی پیوندگرا[10] CTC [۴۱] متداول‌ترین روش برای آموزش سیستم‌های بازشناسی گفتار انتها به انتها است که مشکل تطبیق بین دنباله ورودی و خروجی با طول‌های متفاوت را حل می‌کند. مسئله‌ی نگاشت دنباله–به–دنباله[11] به‌صورت نگاشت دنباله‌ی ورودی $X=[x_1, x_2, \ldots, x_T]$ به دنباله‌ی هدف $Y=[y_1, y_2, \ldots, y_U]$ تعریف می‌شود. در مورد مسئله‌ی بازشناسی گفتار، X سیگنال صوتی و Y دنباله‌ی برچسب‌ها می‌باشد. روش CTC مدل آکوستیکی را طوری آموزش می‌دهد که احتمال $P(Y \mid X)$ برای هر جفت ورودی–خروجی بیشینه شود.

این روش یک برچسب به نام کاراکتر تهی[12] (با نماد -) نیز تعریف می‌کند که امکان ایجاد ترازبندی‌های[13] مختلف و درنظرگرفتن کاراکترهای تکراری را می‌دهد (مثل وجود دو کاراکتر l در کلمه‌ی "hello"). در نتیجه این توزیع احتمالاتی علاوه بر کاراکترها یا واج‌های موردنظر یک کاراکتر تهی هم دارد. بنابراین شبکه در هر گام زمانی[14]

1. Conformer
2. Convolutional
3. Feed Forward Network
4. Multi-Head Self-Attention
5. Encoder
6. Gradient
7. Residual Connections
8. Recurrent Neural Network
9. Loss Function
10. Connectionist Temporal Classification
11. Sequence-to-sequence
12. Blank Character
13. Alignment
14. Time Step

تصمیم می‌گیرد که یک برچسب واجی یا کاراکتر تهی را به عنوان خروجی تولید کند. با کمک تابع نگاشت چند به یک[1] β ابتدا تمام برچسب‌های غیرتهی تکراری پشت سر هم، ادغام می‌شوند سپس تمام برچسب‌های تهی حذف می‌شوند. مثال زیر، چند ترازبندی مختلف با طول ۷ را نشان می‌دهد که همگی با کمک تابع β منجر به ایجاد یک متن یکسان یعنی "hello" شده‌اند:

$$\left.\begin{array}{l}\beta\left([h,e,l,-,l,o,-]\right)\\ \beta\left([h,-,e,l,-,l,o]\right)\\ \cdots\\ \beta\left([-,h,e,l,-,l,o]\right)\end{array}\right\}=\text{"hello"} \quad (۵)$$

اگر ترازبندی‌های ممکن بین X و Y که بعد از اعمال تابع β منجر به تولید Y می‌شوند را با π نمایش دهیم. احتمال شرطی $P(Y\mid X)$ بدین‌صورت تعریف می‌شود: مجموع احتمالات روی تمام ترازبندی‌های ممکن π بین X و Y که بعد از اعمال تابع β منجر به تولید Y می‌شوند و مقدار آن از طریق (۶) محاسبه می‌گردد

$$P\left(Y\mid X\right)=\sum_{\pi\epsilon\beta^{-1}(Y)}P(\pi\mid X) \quad (۶)$$

یک لایه‌ی بیش‌نرم[2] نیز در خروجی شبکه قرار داده می‌شود که توزیع احتمالاتی روی مجموعه‌ی {برچسب‌ها و تهی} را در هر گام زمانی تعریف می‌کند.

احتمال شرطی دنباله‌ی برچسب $\pi=\left[\pi_1,\pi_2,\ldots,\pi_T\right]$ را با $P(\pi\mid X)$ نمایش می‌دهیم. با درنظرگرفتن فرض استقلال شرطی، حاصل $P(\pi\mid X)$ به‌صورت ضرب خروجی‌های شبکه و از طریق (۷) محاسبه می‌شود

$$P\left(\pi|X\right)\approx\prod_{t=1}^{T}P\left(\pi_t|X\right)=\prod_{t=1}^{T}y_{\pi_t}^{t} \quad (۷)$$

در اینجا $y_{\pi_t}^{t}$ احتمال فعالیت[3] بیش‌نرم مشاهده برچسب π_t در لایه خروجی شبکه در زمان t را نشان می‌دهد.

مقدار تابع CTC به‌صورت منفی لگاریتم درستنمایی[4] تعریف می‌شود و در در واقع تابع هزینه شبکه است و باید کمینه شود

$$loss_{CTC}\left(X,Y\right)=-\ln P\left(Y\mid X\right) \quad (۸)$$

CTC برای محاسبه‌ی بهینه‌ی (۶) از یک الگوریتم برنامه‌سازی پویا[5] که مشابه با الگوریتم پیشرو-پسرو[6] در HMM می‌باشد، استفاده می‌کند. همچنین محاسبه‌ی بهینه گرادیان $\ln P\left(Y|X\right)$ نسبت به خروجی‌های شبکه عصبی با کمک مقادیر میانی[7] این الگوریتم انجام می‌شود.

٤- تنظیمات آزمایش‌ها

در این بخش پیکره، شیوه ارزیابی، تنظیمات مربوط به پارامترهای شبکه، ابرپارامترهای یادگیری و روش‌های داده‌افزایی خواهند آمد.

جدول ۱: مشخصات پیکره‌ی LibriSpeech [۴۰].

مجموعه	زیرمجموعه‌ها	ساعت	زمان (دقیقه) – هر گوینده	تعداد گویندگان زن	تعداد گویندگان مرد
آموزش	train-clean-100	۱۰۰٫۶	۲۵	۱۲۵	۱۲۶
اعتبارسنجی[8]	dev-clean	۵٫۴	۸	۲۰	۲۰
	dev-other	۵٫۳	۱۰	۱۶	۱۷
آزمون	test-clean	۵٫۴	۸	۲۰	۲۰
	test-other	۵٫۱	۱۰	۱۷	۱۶

همچنین جعبه‌ابزار مورد استفاده برای پیاده‌سازی روش‌ها و آموزش مدل، نسخه دوم ESPnet۲٫۰ [۴۲] است که در بستر پایتورچ[9] می‌باشد.

٤-۱ پیکره

پیکره‌ی مورد استفاده LibriSpeech [۴۰] است که در بسیاری تحقیقات اخیر بازشناسی گفتار به عنوان بستر آزمایش[10] مورد استفاده قرار گرفته است. پیکره LibriSpeech شامل ۱۰۰۰ ساعت گفتار خوانده شده از حوزه کتاب‌های صوتی می‌باشد. فایل‌ها با فرکانس ۱۶ کیلوهرتز نمونه‌برداری شده‌اند. مجموعه آموزشی این پیکره حاوی ۹۶۰ ساعت گفتار با ۲۳۳۸ گوینده است. ارزیابی در شرایط دادگان محدود که هدف این تحقیق است با استفاده از بخش ۱۰۰ ساعتی این پیکره یعنی مجموعه train-clean-۱۰۰ انجام می‌گیرد. مشخصات این پیکره در جدول ۱ آورده شده است که از مجموعه آموزشی این پیکره تنها مجموعه ۱۰۰ ساعتی آن مورد استفاده قرار گرفته و در این جدول ارائه شده است.

در این مقاله، نتایج بر روی چهار مجموعه از پیکره LibriSpeech شامل مجموعه اعتبارسنجی (dev) و آزمون (test) و در دو بخش تمیز (clean) و سایر (other) ارائه شده‌اند. منظور از مجموعه سایر، داده‌هایی با گویندگان، کیفیت ضبط یا شرایط گفتاری متنوع‌تر و چالش‌برانگیزتر نسبت به مجموعه تمیز است که به‌طور کلی بازشناسی آن‌ها نسبت به مجموعه تمیز دشوارتر است.

٤-۲ ارزیابی

معیار استاندارد برای ارزیابی سیستم بازشناسی گفتار، نرخ خطای کلمه[11] (WER) می‌باشد. در این معیار، دنباله‌ی کلمات مرجع و دنباله‌ی کلمات بازشناسی شده با استفاده از یک الگوریتم برنامه‌سازی پویا به نام فاصله ویرایش[12] با هم مقایسه می‌شوند. اگر تعداد خطاهای جایگزینی، خطاهای حذف و خطاهای درج نسبت به متن مرجع را به ترتیب با S، D و I نمایش دهیم، WER به‌صورت زیر محاسبه می‌گردد:

$$WER=\frac{S+D+I}{N} \quad (۸)$$

که در اینجا N، تعداد کل کلمات در دنباله‌ی مرجع را نشان می‌دهد. معیار دیگری نیز به نام نرخ خطای کاراکتر[13] (CER) وجود دارد که مشابه با معیار WER است با این تفاوت که واحدها به‌جای کلمه، کاراکتر یا همان حروف هستند و خطا در سطح کاراکتر محاسبه می‌شود.

1. Many to One
2. Softmax
3. Activation
4. Liklihood
5. Dynamic Programing
6. Forward-backward
7. Intermediate Values
8. Validation
9. PyTorch
10. Benchmark
11. Word Error Rate
12. Edit Distance
13. Character Error Rate

جدول ۲: پارامترهای معماری کدگذار و کدگشا مربوط به مدل کانفورمر.

	کدگذار	کدگشا
نوع معماری	کانفورمر	ترنسفورمر
تعداد سرهای توجه	۴	۴
تعداد واحدهای خطی	۱۰۲۴	۲۰۴۸
تعداد بلوک‌ها	۱۲	۶
نرخ حذف تصادفی نرون‌ها[1]	۰/۱	۰/۱
اندازه کرنل ماژول پیچشی	۳۱	–

جدول ۳: پارامترها و تنظیمات مربوط به پوشاندن زمانی و پوشاندن فرکانسی روش SpecAugment.

پارام	عملکرد	مقدار
m_T	تعداد عمل پوشاندن در محور زمان	۲
T	حداکثر اندازه پوشاندن زمانی	۴۰ فریم
m_F	تعداد عمل پوشاندن در محور فرکانس	۲
F	حداکثر اندازه پوشاندن فرکانسی	۳۰ باند فرکانسی

جدول ۴: آمار کلمات روی مجموعه ۱۰۰ ساعتی پیکره LibriSpeech.

شاخص	مقدار
تعداد کل فایل‌ها	۲۸۵۳۹
تعداد کل کلمات	۹۹۰۱۰۱
تعداد کلمات منحصربه فرد	۳۳۷۹۸
میانگین تکرار هر کلمه منحصربه فرد	۲۹/۲۹

۴-۳ تنظیمات مربوط به آموزش مدل آکوستیکی

پارامترهای معماری بخش کدگذار و کدگشا[2] از مدل کانفورمر در جدول ۲ آمده است.

تعداد کل پارامترهای مدل استفاده شده برابر با ۳۴/۲۳ میلیون است. همچنین نرخ یادگیری[3] برابر با ۰/۰۰۲ و تعداد گام‌های گرم‌سازی[4] برابر با ۱۵۰۰۰ تنظیم شده‌اند.

۴-۴ تنظیمات مربوط به روش‌های داده‌افزایی

تنظیمات مربوط به پوشاندن زمانی و پوشاندن فرکانسی در روش SpecAugment در جدول ۳ آمده است.

در روش پوشاندن کلمه، به منظور شناسایی محدوده زمانی کلمات در طیف‌نگار از مدل هم‌تراز کننده اجباری مبتنی بر مدل HMM به نام Montreal Forced Aligner [۴۳] استفاده کردیم. همچنین طبق مقاله [۲۴] ۱۵٪ کلمات هر فایل گفتاری برای پوشاندن انتخاب می‌شوند.

آمار کلمات مجموعه داده: برای پیاده‌سازی روش پیشنهادی دوم «پوشاندن آگاه از فراوانی کلمات»، نیاز به آمارگیری از کلمات مجموعه داده آموزشی داشتیم. جدول ۴ آمار استخراج شده از مجموعه ۱۰۰ ساعتی پیکره LibriSpeech را نشان می‌دهد.

1. Dropout
2. Decoder
3. Learning Rate
4. Warmup Steps

جدول ۵: نتایج روش پوشاندن در سطح فریم و کلمه برحسب معیار WER (درصد).

روش		مجموعه اعتبارسنجی		مجموعه آزمون	
		تمیز	سایر	تمیز	سایر
بدون داده‌افزایی		۷/۹	۲۴/۴	۸/۵	۲۴/۹
پوشاندن	واحد: فریم	**۷/٤**	**۲۲/۱**	**۷/۷**	**۲۲/۷**
زمانی	واحد: کلمه	۷/۶	۲۴/۱	۸/۰	۲۴/۵

جدول ۶: نتایج روش پوشاندن در سطح فریم و کلمه برحسب معیار CER (درصد).

روش		مجموعه اعتبارسنجی		مجموعه آزمون	
		تمیز	سایر	تمیز	سایر
بدون داده‌افزایی		۳/۲	۱۲/۳	۳/۳	۱۲/۵
پوشاندن	واحد: فریم	**۳/۰**	**۱۱/۳**	**۳/۰**	**۱۱/٤**
زمانی	واحد: کلمه	۳/۱	۱۲/۴	۳/۲	۱۲/۵

۵- نتایج آزمایش‌ها

در این بخش، نتایج حاصل از پیاده‌سازی روش‌های داده‌افزایی مختلف را بررسی می‌کنیم. ابتدا به مقایسه دو روش پایه داده‌افزایی یعنی پوشاندن در سطح فریم و پوشاندن در سطح کلمه می‌پردازیم و سپس نتایج روش‌های پیشنهادی خود را ارائه می‌دهیم.

۵-۱ مقایسه عملکرد داده‌افزایی مبتنی بر پوشاندن در سطح کلمه و فریم

نتایج روش پوشاندن در سطح کلمه و روش پوشاندن در سطح فریم برحسب معیار WER و CER به ترتیب در دو جدول ۵ و جدول ۶ ارائه شده است.

همانطور که در جداول فوق مشاهده می‌شود، هر دو روش پوشاندن در محور زمان نسبت به حالت بدون داده‌افزایی نتایج بهتری بدست آورده‌اند. با این حال، پوشاندن در سطح کلمه نتوانسته است به نتایج بهتری نسبت به پوشاندن در سطح فریم دست یابد. در بخش ۳-۳ به‌طور مفصل به بررسی ریشه این موضوع پرداختیم که دلیل اصلی این موضوع به ماهیت دشوارتر یادگیری در روش پوشاندن کامل کلمات به‌طور تصادفی برمی‌گردد، به‌ویژه زمانی که داده‌های آموزشی محدود باشند. این مشاهدات، انگیزه اصلی برای ارائه روش‌های پیشنهادی جدید این پژوهش را فراهم کرد که در ادامه به تشریح نتایج آن‌ها می‌پردازیم.

۵-۲ نتایج روش‌های پیشنهادی

جداول ۷ و ۸ نتایج روش‌های پیشنهادی را در مقایسه با روش‌های پوشاندن کلمه و SpecAugment که یک روش مرجع قدرتمند در داده‌افزایی ASR است، نشان می‌دهند.

نتایج نشان می‌دهند که هر دو روش پیشنهادی، عملکرد قابل توجهی نسبت به روش‌های پایه و روش قدرتمند SpecAugment دارند.

تحلیل خطاهای جزئی: برای بررسی دقیق‌تر عملکرد روش‌های پیشنهادی، تحلیل جزئی خطاها به تفکیک نوع خطا (درج، حذف و جایگزینی) در جدول ۹ ارائه شده است.

همانطور که ملاحظه می‌شود، بیشترین نوع خطا از نوع خطای جایگزینی است و خطاهای درج و حذف خیلی کمتر رخ می‌دهند. روش‌های داده‌افزایی پیشنهادی توانسته‌اند به‌طور قابل توجهی خطاهای

جدول ۷: مقایسه نتایج روش‌های پیشنهادی و پایه برحسب معیار WER (درصد).

روش	مجموعه اعتبارسنجی		مجموعه آزمون	
	تمیز	سایر	تمیز	سایر
بدون داده‌افزایی	۷٫۹	۲۴٫۴	۸٫۵	۲۴٫۹
SpecAugment [۱۱]	۷٫۱	۱۸٫۷	۷٫۳	۱۹٫۰
پوشاندن کلمه [۲۴]	۷٫۶	۲۴٫۱	۸٫۰	۲۴٫۵
روش پیشنهادی ۱: پوشاندن تدریجی	**۶٫۶**	**۱۸٫۰**	**۶٫۸**	**۱۸٫۲**
روش پیشنهادی ۲: پوشاندن آگاه از فراوانی کلمات	**۶٫۳**	**۱۷٫۱**	**۶٫۶**	**۱۷٫۳**

جدول ۸: مقایسه نتایج روش‌های پیشنهادی و پایه برحسب معیار CER (درصد).

روش	مجموعه اعتبارسنجی		مجموعه آزمون	
	تمیز	سایر	تمیز	سایر
بدون داده‌افزایی	۳٫۲	۱۲٫۳	۳٫۳	۱۲٫۵
SpecAugment [۱۱]	۳٫۰	۹٫۴	۲٫۹	۹٫۴
پوشاندن کلمه [۲۴]	۳٫۱	۱۲٫۴	۳٫۲	۱۲٫۵
روش پیشنهادی ۱: پوشاندن تدریجی	**۲٫۷**	**۹٫۰**	**۲٫۶**	**۸٫۹**
روش پیشنهادی ۲: پوشاندن آگاه از فراوانی کلمات	**۲٫۵**	**۸٫۷**	**۲٫۶**	**۸٫۵**

جایگزینی را کاهش دهند، که نشان‌دهنده بهبود توانایی مدل در تشخیص صحیح کلمات است.

۵-۳ مقایسه با دیگر روش‌های داده‌افزایی

برای ارائه یک دید بهتر از میزان خطاهای گزارش شده تاکنون، در جدول ۱۰ جایگاه نتایج خود را با روش‌های پیشرفته دیگر که روی مجموعه آموزشی ۱۰۰ ساعتی LibriSpeech آموزش دیده‌اند، مقایسه می‌کنیم. نتایج هر روش به‌طور مستقیم از نتایج همان مقاله گزارش شده‌اند. در این جدول علامت – به‌معنای این است که نتیجه برای آن مجموعه در مقاله مربوطه گزارش نشده است.

هدف این بخش مقایسه دقیق و مورد به مورد بین روش‌ها نیست، زیرا هر مقاله از معماری متفاوتی استفاده کرده است. هدف این جدول کسب اطلاع از آخرین نتایج ارائه شده در زمینه تحقیقات داده‌افزایی روی دادگان محدود ۱۰۰ ساعتی می‌باشد. این جدول می‌تواند به‌عنوان نقطه مرجع مفیدی برای درک جایگاه روش ما و دیگر روش‌ها و محدوده دقت‌های بدست‌آمده در این حوزه باشد.

۵-۴ تحلیل و بررسی نتایج

نتایج روش‌های پیشنهادی چندین نکته مهم را در زمینه داده‌افزایی ASR نشان می‌دهد که در ادامه بررسی می‌شوند:

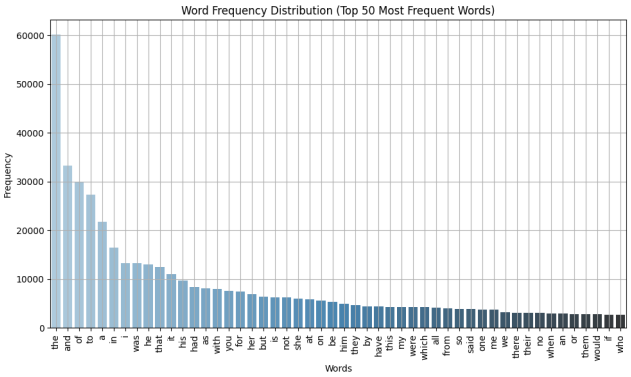
- **تأثیر پوشاندن تدریجی**: پوشاندن تدریجی از واحد فریم به کلمه (روش پیشنهادی اول) به‌طور چشمگیری عملکرد را بهبود می‌بخشد. این نتایج تأیید می‌کند که مدل نیاز دارد ابتدا با چالش‌های ساده‌تر (پوشاندن فریم) مواجه شود و سپس با چالش‌های پیچیده‌تر (پوشاندن کلمه) آموزش ببیند. این رویکرد کاملاً با اصول «یادگیری مبتنی بر برنامه آموزشی» همخوانی دارد که در آن افزایش تدریجی سختی آموزش، منجر به یادگیری بهتر می‌شود.
- **اهمیت گزینش کلمات برای پوشاندن**: راهکار انتخاب کلمات برای پوشاندن (روش پیشنهادی دوم) با در نظر گرفتن فراوانی کلمات، عملکرد را بیش از پیش بهبود می‌بخشد. این نشان می‌دهد که تنظیم سطح دشواری داده‌افزایی براساس آشنایی مدل با کلمات، یک راهکار مؤثر است. کلمات با فراوانی کمتر در پیکره، چالش بیشتری برای مدل ایجاد می‌کنند و بهتر است بعد از کلمات پرتکرار انتخاب شوند تا مدل این نمونه‌های چالش‌برانگیز را بهتر یاد بگیرد. آمار دادگان در جدول ۶ بیانگر تنوع واژگانی نسبتاً بالا در مجموعه داده مورد استفاده است، با میانگین تکرار هر کلمه حدود ۲۹٫۲۹ بار. همچنین توزیع فراوانی کلمات بسیار نامتوازن است که این موضوع را در شکل ۱۰ نمایش داده شده است.

جدول ۹: نتایج روش‌های داده‌افزایی روی مجموعه آزمون بخش تمیز و سایر به تفکیک نوع خطا.

روش	مجموعه آزمون- بخش تمیز			مجموعه آزمون- بخش سایر		
	درج	حذف	جایگزینی	درج	حذف	جایگزینی
بدون داده‌افزایی	۱٫۰	۰٫۷	۶٫۷	۲٫۸	۲٫۵	۱۹٫۶
SpecAugment [۱۱]	۰٫	۰٫۶	۵٫۸	۲٫۲	۱٫۷	۱۵٫۱
پوشاندن کلمه [۲۴]	۱٫۰	۰٫۷	۶٫۴	۲٫۷	۲٫۵	۱۹٫۳
روش پیشنهادی ۱: پوشاندن تدریجی	**۰٫۹**	**۰٫۵**	**۵٫۴**	**۲٫۰**	**۱٫۷**	**۱۴٫۵**
روش پیشنهادی ۲: پوشاندن آگاه از فراوانی کلمات	**۰٫۸**	**۰٫۴**	**۵٫۳**	**۱٫۹**	**۱٫۵**	**۱۳٫۸**

جدول ۱۰: بهترین نتایج داده‌افزایی گزارش شده تاکنون با دادگان آموزشی ۱۰۰ ساعتی LibriSpeech.

روش	نوع روش	مجموعه اعتبارسنجی		مجموعه آزمون	
		تمیز	سایر	تمیز	سایر
ADA-RT [۳۲]	ایجاد جملات متنی جدید	۸٫۵۴	۲۱٫۱۱	۸٫۸۰	۲۱٫۳۲
SegAug [۴۴]	ایجاد جملات متنی جدید	-	-	۶٫۹۹	۲۰٫۱۸
SapAugment [۳۶]	خودکار	-	-	۸٫۳	۲۱٫۵
Policy-SpecAugment [۳۵]	خودکار	۸٫۳	۲۱	۹٫۱	۲۱٫۵
روش پیشنهادی ۱: پوشاندن تدریجی	پوشاندن	۶٫۶	۱۸٫۰	۶٫۸	۱۸٫۲
روش پیشنهادی ۲: پوشاندن آگاه از فراوانی کلمات	پوشاندن	۶٫۳	۱۷٫۱	۶٫۶	۱۷٫۳



شکل ۱۰: فراوانی ۵۰ کلمه از پرتکرارترین کلمات پیکره LibriSpeech.

این نمودار تعداد تکرار برای ۵۰ کلمه از پرتکرارترین کلمات پیکره را نشان می‌دهد. در پیکره مورد بررسی، ۱۲۴۷۵ کلمه تنها یک بار رخ داده‌اند، درحالیکه تعداد محدودی از کلمات با فراوانی بسیار بالا وجود دارند. این مشاهده، پایه اصلی روش پیشنهادی دوم ما را تشکیل می‌دهد که ابتدا کلمات پرتکرار را بپوشانیم و سپس به سراغ کلمات کم‌تکرار برویم.

- **ترکیب مزایای دو روش پوشاندن کلمه و فریم:** روش‌های پیشنهادی ما از مزایای هر دو رویکرد پوشاندن کلمه و فریم بهره می‌برند:
الف) الهام از مدل‌های زبانی قدرتمند: پوشاندن کلمه، واحدهای معنادار و مشخص زبانی را هدف قرار می‌دهد، بنابراین ساختار طبیعی گفتار بهتر حفظ می‌شود. این رویکرد مشابه عملکرد مدل‌های زبانی بزرگ [۲۵] است که در آن بخشی از کلمات ورودی حذف می‌شوند و مدل یاد می‌گیرد با استفاده از اطلاعات بافت[1] اطراف، این کلمات را پیش‌بینی کند.
ب) ترکیب نقاط قوت دو روش: روش‌های پیشنهادی هم از قدرت SpecAugment در تقویت مدل با دادگان محدود بهره می‌برد و هم از توانایی پوشاندن کلمه در یادگیری روابط معنایی عمیق‌تر استفاده می‌کند.
- **پویایی و تطبیق‌پذیری روش‌های پیشنهادی:** یکی از ویژگی‌های کلیدی روش‌های پیشنهادی، تطبیق‌پذیری و پویایی آن‌ها در طول فرآیند آموزش است. در روش پوشاندن تدریجی، واحد پوشاندن در طول آموزش ثابت نیست و در حین آموزش تغییر و از واحد کوچک فریم به واحد بزرگتر کلمه افزایش می‌یابد و با پیشروی آموزش شبکه، درجه سختی داده‌افزایی نیز افزایش می‌یابد. در روش پوشاندن آگاه از فراوانی کلمات نیز کلمات موردنظر برای پوشاندن براساس شرایط آماری‌شان در کل پیکره و مطابق با تعداد رخدادشان نسبت به دیگر کلمات آن جمله انتخاب می‌شوند.
- **اهمیت رویکرد محتوامحور**[2]: برخلاف روش‌های معمول که کلمات را به‌صورت تصادفی می‌پوشانند [۲۴]، روش ما به متن و شرایط آماری کلمات آن حساس است. نتایج آزمایش‌ها، اهمیت رویکرد محتوا محور ما را در مقابل روش‌های داده‌افزایی صرفاً مبتنی بر گفتار مانند SpecAugment [۱۱]، که یک روش مرجع قدرتمند در داده‌افزایی ASRاست و روش پوشاندن کلمه [۲۴] که به‌صورت تصادفی کلمات را می‌پوشاند، تأیید می‌کند.
- **بهبود قابل توجه در شرایط چالش‌برانگیز:** میزان بهبود در مجموعه داده «سایر» (که شامل شرایط صوتی چالش‌برانگیز است) بسیار قابل توجه است. این نشان می‌دهد که روش‌های پیشنهادی در افزایش مقاومت مدل در برابر شرایط دشوار صوتی مؤثر هستند و تنوع داده‌های ایجاد شده توسط روش‌های پیشنهادی، به مدل کمک می‌کند تا در مواجهه با این شرایط، عملکرد بهتری داشته باشد.
- **پیچیدگی محاسباتی روش‌ها:** در این تحقیق بر روش‌های مبتنی بر پوشاندن تمرکز شد که از دسته روش‌های با سرعت مطلوب هستند. زیرا عملیات پوشاندن در تمامی روش‌ها به‌صورت برداری و بهینه پیاده‌سازی شد که بسیار سریع‌تر از پردازش ترتیبی است بنابراین این روش‌ها بار محاسباتی کمی ایجاد می‌کنند. با توجه به یکسان بودن معماری مدل و تنظیمات آموزش شامل دوره آموزشی، روش‌های مورد بررسی از نظر زمان آموزش مشابه عمل کردند. مدت زمان آموزش دو روش پیشنهادی و دو روش مورد مقایسه یعنی پوشاندن کلمه [۲۴] و روش SpecAugment [۱۱] بر روی سیستم GPU RTX 3090 حدود ۱۷ ساعت طول کشید. تنها در روش پیشنهادی دوم، روش پوشاندن آگاه از فراوانی کلمات، نیاز به مراحل پیش‌پردازش مانند محاسبه فراوانی کلمات و مرتب‌سازی کلمات هر جمله براساس فراوانی کلمات بود که این عملیات هزینه بسیاری کمی در حد چند دقیقه داشتند و به‌صورت آفلاین و پیش از آموزش مدل انجام شده و تأثیری بر زمان آموزش نداشته‌اند و تنها یک بار برای کل پیکره اجرا می‌شود. همچنین میزان حافظه مصرف‌شده توسط GPU، ۲۳ گیگابایت بود. زمان اجرای عملیات کدگشایی برای تولید متن در مرحله آزمایش نیز برای کل مجموعه‌های اعتبارسنجی و آزمون به مدت ۵ ساعت به طول انجامید.

۶- نتیجه‌گیری و کارهای آینده

سیستم‌های بازشناسی گفتار با چالش‌های متعددی در زمینه داده‌های آموزشی مواجه هستند. نیاز به داده‌های گسترده، متنوع و با کیفیت بالا، فرآیند توسعه این سیستم‌ها را پیچیده می‌کند. جمع‌آوری این داده‌ها زمان‌بر، پرهزینه و نیازمند منابع انسانی زیادی است. علاوه بر این، داده‌های آموزشی باید از تنوع مناسبی از جهات مختلف مانند سن، جنسیت، لهجه و شرایط محیطی برخوردار باشند تا بتوانند عملکرد مناسبی در شرایط واقعی داشته باشند. این چالش‌ها به‌ویژه برای زبان‌هایی که منابع محدودتری دارند، جدی‌تر است.

چالش اصلی مورد توجه این پژوهش، عملکرد ضعیف برخی روش‌های داده‌افزایی مانند پوشاندن واحد کلمه بود که علی‌رغم کارایی مناسب در شرایط دادگان فراوان، در شرایط دادگان کم عملکرد مطلوبی نداشتند. بدین‌منظور دو روش داده‌افزایی برای بهبود عملکرد ASR در شرایط دادگان محدود معرفی و بررسی شد.

روش اول پیشنهادی ما، پوشاندن تدریجی، از مفهوم یادگیری مبتنی بر برنامه آموزشی بهره گرفت. در این روش، فرآیند آموزش با پوشاندن در سطح فریم (وظیفه ساده‌تر) شروع شده و سپس به پوشاندن در سطح کلمه (وظیفه پیچیده‌تر) گسترش می‌یابد.

روش دوم، روش پوشاندن آگاه از فراوانی کلمات، یک لایه هوشمندی بیشتر به روش اول اضافه می‌کند. در این روش، استراتژی انتخاب کلمات برای پوشاندن براساس فراوانی آن‌ها در پیکره آموزشی انجام می‌شود. این رویکرد براساس این فرضیه طراحی شده است که در ابتدای آموزش، کلمات پرتکرار پوشانده شوند که یک وظیفه آسان‌تر برای مدل تعریف

[1] Context
[2] Content-based

می‌کند زیرا مدل نمونه‌های بیشتر و متنوع‌تری برای تشخیص این کلمات در اختیار دارد.

دستاوردهای کلیدی این پژوهش عبارتند از:

- **موفقیت رویکرد تدریجی در داده‌افزایی:** نتایج نشان داد که آموزش تدریجی از وظایف ساده به سخت در زمینه داده‌افزایی، در شرایط دادگان محدود، منجر به عملکرد بهتری می‌شود.
- **تأیید اهمیت رویکرد محتوا محور:** روش پیشنهادی دوم که به محتوای متن حساس هستند، عملکرد بهتری نسبت به روش‌های صرفاً مبتنی بر سیگنال نشان داد.
- **بهبود مقاومت مدل در شرایط چالش‌برانگیز:** بهبود قابل توجه در مجموعه داده چالش‌برانگیز نشان می‌دهد که روش‌های پیشنهادی در افزایش مقاومت مدل در برابر شرایط دشوار مؤثر هستند.
- **دستیابی به نتایج رقابتی:** روی مجموعه دادگان محدود ۱۰۰ ساعتی پیکره LibriSpeech، روش‌های پیشنهادی ما بهترین نتیجه تاکنون گزارش شده را به دست آوردند.

به‌طور کلی، این پژوهش نشان داد که با طراحی هوشمندانه روش‌های داده‌افزایی و در نظر گرفتن محتوای متنی، می‌توان عملکرد سیستم‌های ASR را تا حد زیادی بهبود بخشید، به‌ویژه در شرایطی که با کمبود دادگان مواجه هستیم. روش‌های پیشنهادی ما موفق شدند نقاط ضعف روش‌های موجود در شرایط دادگان محدود را برطرف کرده و عملکرد بهتری نسبت به روش‌های داده‌افزایی پوشاندن کلمه و SpecAugment روی پیکره LibriSpeech و هر دو نوع داده تمیز و چالش‌برانگیز ارائه دهند. این یافته‌ها برای توسعه سیستم‌های ASR در زبان‌هایی که با محدودیت منابع مواجه هستند یا در حوزه‌های خاصی که جمع‌آوری داده در آن‌ها دشوار است، بسیار ارزشمند خواهد بود. با استفاده از روش‌های پیشنهادی، می‌توان از دادگان موجود به شکل مؤثرتری استفاده کرد و نیاز به جمع‌آوری داده گسترده که فرآیندی زمان‌بر و پرهزینه است را کاهش داد. همچنین یکی از کارهای پیشنهادی ما برای آینده، توسعه روش‌های پیشنهادی روی زبان فارسی به عنوان یک زبان با منابع محدود است.

مراجع

[1] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 325-351, 2024.

[2] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195-304, Jul. 2008.

[3] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, Nov. 2012.

[4] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Olomouc, Czech Republic, pp. 273-278 Olomouc, Czech Republic, 8-12 Dec. 2013.

[5] M. Asadolahzade Kermanshahi and M. M. Homayounpour, "Improving phoneme sequence recognition using phoneme duration information in DNN-HSMM," *Journal of AI and Data Mining*, vol. 7, no. 1, pp. 137147, Jan.2019.

[6] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, pp. 5036-5040, Shanghai, China, 25-29 Oct. 2020.

[7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, pp. 12449-12460, 2020.

[8] Y. Zhang et al., Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages, arXiv preprint arXiv:2303.01037, 2023.

[9] A. Radford, *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. of the 40th Int. Conf. on Machine Learning*, pp. 28492-28518, Honolulu, HI, USA, 23-29 Jul. 2023.

[10] A. Hussein, S. Watanabe, and A. Ali, "Arabic speech recognition by end-to-end, modular systems and human," *Computer Speech & Language*, vol. 71, Article ID: 101272, Jan. 2022.

[11] D. S. Park *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. 20th Annual Conf. of the Int. Speech Communication Association,* pp. 2613-2617, Graz, Austria, 15-19 Sept. 2019.

[12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. 16th Annual Conf. of the Int. Speech Communication Association*, Dresden, Germany, pp. 3586-3589, Dresden, Germany, 6-10 Sept. 2015.

[13] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. 42nd IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 5220-5224, New Orleans, LA, USA, 5-9 Mar. 2017.

[14] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, and M. Wieling, "Making more of little data: Improving low-resource automatic speech recognition using data augmentation," in *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics,* vol. 1, pp. 715-729, Toronto, Canada, 9-12 Jul. 2023.

[15] D. S. Park *et al.*, "Improved noisy student training for automatic speech recognition," in *Proc. 21st Annual Conf. of the Int. Speech Communication Association*, pp. 2817-2821, Shanghai, China, 25-29 Oct. 2020.

[16] Y. Zhang *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519-1532, 2022.

[17] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *Proc. Int. Conf. on Machine Learning*, pp. 1416-1429, Honolulu, Hawaii, USA, 23-29 Jul. 2023.

[18] M. Asadolahzade Kermanshahi, *Transfer Learning for ASR to Deal with Low-Resource Data Problem*, Technical Report, Tehran, Iran, 2019. https://www.researchgate.net/publication/359159354_Transfer_Learning_for_ASR_to_Deal_with_Low-Resource_Data_Problem.

[19] M. Asadolahzade Kermanshahi, A. Akbari, and B. Nasersharif, "Transfer learning for end-to-end ASR to deal with low-resource problem in Persian language," in *Proc. 26th Int. Computer Conference, Computer Society of Iran,* 5 pp., Tehran, Iran, 3-4 Mar. 2021.

[20] H. Kheddar, Y. Himeur, S. Al-Maadeed, A. Amira, and F. Bensaali, "Deep transfer learning for automatic speech recognition: Towards better generalization," *Knowledge-Based Systems*, vol. 277, Article ID: 110851, Oct. 2023.

[21] A. Babu*et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. 23th Annual Conf. of the Int. Speech Communication Association*, pp. 2278-2282, Incheon, South Korea, 18-20 Sept. 2022.

[22] V. Pratap, *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, Article ID: 97, 52 pp., 2024.

[23] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469-1477, Sept. 2015.

[24] C. Wang, *et al.*, "Semantic mask for transformer based end-to-end speech recognition," in *Proc. 21st Annual Conf. of the Int. Speech Communication Association*, pp. 971-975, Shanghai, China, 25-29 Oct. 2020.

[25] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. of the 26th Annual Int. Conf. on Machine Learning*, pp. 41-48, Montréal, Canada, 14-18 Jun. 2009.

[26] L. Meng, *et al.*, "MixSpeech: Data augmentation for low-resource automatic speech recognition," in *Proc. 46th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 7008-7012, Toronto, Canada, 6-11 Jun. 2021.

[27] D. Fucci, M. Gaido, M. Negri, M. Cettolo, and L. Bentivogli, "No pitch left Behind: Addressing gender unbalance in automatic speech recognition through pitch manipulation," in *Proc. Automatic Speech Recognition and Understanding Workshop*, 8 pp., Taipei, Taiwan, 16-20 Dec. 2023.

[28] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, p. 21, Atlanta, GA, USA, 16-21 Jun. 2013.

Technologies, pp. 4171-4186, Minneapolis, MN, USA, 2-7 Jun. 2019.
[40] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, pp. 5206-5210, South Brisbane, Australia, 2015.
[41] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. of the 23rd Int. Conf. on Machine Learning*, pp. 369-376, 25-29 Jun. 2006.
[42] S. Watanabe, *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. 19th Annual Conf. of the Int. Speech Communication Association*, pp. 2207-2211, Hyderabad, India, 2018.
[43] [43] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: trainable text-speech alignment using Kaldi.," in *Proc.18th Annual Conf. of the Int. Speech Communication Association*, pp. 498-502, 20-24 Aug. 2017.
[44] K. Le, T. V. Ho, D. Tran, and D. T. Chau, "SegAug: CTC-Aligned segmented augmentation for robust RNN-transducer based speech recognition," in *Proc. 50th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 5 pp., Hyderabad, India, 6-11 Apr. 2025.
[29] Y. Qian, H. Hu, and T. Tan, "Data augmentation using generative adversarial networks for robust speech recognition," *Speech Communication*, vol. 114, pp. 1-9, Nov. 2019.
[30] E. Casanova *et al.,* "ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion," in *Proc. 24th Annual Conf. of the Int. Speech Communication Association*, pp. 1244-1248, Dublin, Ireland, 20-24 Aug. 2023.
[31] J. Sun et al., "Semantic data augmentation for end-to-end Mandarin speech recognition," in *Proc. 21st Annual Conf. of the Int. Speech Communication Association*, pp. 1269-1273, Brno, Czech Republic, 30 Aug.-3 Sept. 2021.
[32] T. K. Lam, M. Ohta, S. Schamoni, and S. Riezler, "On-the-Fly Aligned Data Augmentation for Sequence-to-Sequence ASR,"in *Proc. 21st Annual Conf. of the Int. Speech Communication Association*, pp. 1299-1303, Brno, Czech Republic, 30 Aug.-3 Sept. 2021.
[33] G. Wang, *et al.*, "G-Augment: Searching for the meta-structure of data augmentation policies for ASR," in *Proc. IEEE Spoken Language Technology Workshop*, pp. 23-30, Doha, Qatar, 9-12 Jan. 2022.
[34] Z. Jin, *et al*., "Towards automatic data augmentation for disordered speech recognition," in *Proc. 49th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 10626-10630, Seoul, Korea, 14-19 Apr. 2024.
[35] R. Li, G. Ma, D. Zhao, R. Zeng, X. Li, and H. Huang, "A policy-based approach to the SpecAugment method for low resource E2E ASR," in *Proc. Asia Pacific Signal and Information Processing Association*, pp. 630-635, Chiang Mai, Thailand, 7-10 Nov. 2022.
[36] T. -Y. Hu, *et al*., "SapAugment: Learning asample adaptive policy for data augmentation," in *Proc. 46th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 4040-4044, Toronto, Canada, 6-11 Jun. 2021.
[37] A. Sriram, M. Auli, and A. Baevski, "Wav2Vec-Aug: Improved self-supervised training with limited data," in *Proc. 20th Annual Conf. of the Int. Speech Communication Association,* pp. 4950-4954, Incheon, South Korea, 18-20 Sept. 2022.
[38] D. Jiang, W. Li, M. Cao, W. Zou, and X. Li, "Speech SimCLR: Combining contrastive and reconstruction objective for self-supervised speech representation learning," in *Proc. 21st Annual Conf. of the Int. Speech Communication Association*, pp. 1544-1548, Brno, Czech Republic, 30 Aug.-3 Sept. 2021.
[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language*

مریم اسدالله‌زاده کرمانشاهی تحصیلات خود را در مقطع کارشناسی رشته مهندسی کامپیوتر در دانشگاه رازی کرمانشاه در سال ۹۲ به پایان رساند. سپس در سال ۹۴ مدرک کارشناسی‌ارشد خود را در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه صنعتی امیرکبیر اخذ کرد. در حال حاضر وی در حال اتمام مقطع تحصیلی دکتری در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه علم و صنعت ایران می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان بازشناسی گفتار، داده‌افزایی، پردازش زبان طبیعی و خلاصه‌سازی گفتار می‌باشد.

احمد اکبری ازیرانی دانشیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران بوده و دارای بیش از ۲۵ سال سابقه تدریس و پژوهش در زمینه‌های مختلف رشته مهندسی کامپیوتر در این دانشکده می‌باشد. ایشان بیش از ۱۵۰ مقاله در مجلات و کنفرانس‌های معتبر بین‌المللی منتشر نموده است. زمینه‌های پژوهشی مورد علاقه ایشان پردازش داده‌ها، پردازش سیگنال‌ها، شبکه‌های کامپیوتری و امنیت شبکه می‌باشد.

بابک ناصر شریف از سال ۱۳۹۰ عضو هیأت علمی و اکنون دارای مرتبه علمی دانشیار در گروه هوش مصنوعی دانشکده مهندسی کامپیوتر در دانشگاه صنعتی خواجه نصیرالدین طوسی است. وی مدارک کارشناسی ارشد و دکتری خود را در گرایش هوش مصنوعی از دانشگاه علم و صنعت ایران دریافت کرده است. زمینه تحقیقاتی ایشان پردازش گفتار، بازشناسی گفتار، بازشناسی احساس از گفتار، یادگیری عمیق و خودنظارتی برای شاخه‌های مختلف پردازش گفتار و بازشناسی الگو است.