سامانه‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Summer 2023, 4 (2), 11-23
DOR:

Received: 5 Apr 2023
Accepted: 17 May 2023

مقاله پژوهشی

ParsAirCall: Automated Conversational IVR in Airport Call Center Using Deep Transfer Learning

Soheil Tehranipour[1], Mohammad Manthouri[2*], Samaneh Yazdani[3]

1. MS Student, Department of Computer Engineering, North Tehran Branch, Islamic Azad University, Tehran, Iran. **s.tehranipour@iau-tnb.ac.ir**
2. Assistant Professor, Department of Electrical and Electronic Engineering, Shahed University, Tehran, Iran. **Corresponding Author*, **mmanthouri@shahed.ac.ir**
3. Assistant Professor, Department of Computer Engineering, North Tehran Branch, Islamic Azad University, Tehran, Iran. **s_yazdani@iau-tnb.ac.edu**

Abstract

Introduction: In this paper, we introduce the ParsAirCall toolkit, which is a tool for automatic recognition of Persian numbers in airport systems. It leverages deep transfer learning to improve performance in real and operational scenarios of voice-controlled smart telephone systems at airports across the country. In today's world, with the advancements in artificial intelligence, traditional systems for interacting with callers in telephone calls are not efficient, and this efficiency will be enhanced through automation and the automation of repetitive tasks.

Method: ParsAirCall distinguishes itself by surpassing competing models in the Persian language, achieving heightened accuracy with fewer parameters and optimized computing resources. Addressing the challenge posed by limited data for Persian speech recognition, we meticulously curated a 30-hour telephony dataset, serving as the cornerstone for training the final ParsAirCall model. Embracing the innovative QuartzNet architecture, our deep transfer learning strategy empowers ParsAirCall to capture nuanced features in Persian speech, ensuring superior performance in number recognition tasks associated with airport telephone calls.

Results: Experiments were conducted on both our collected telephony dataset and the Common Voice project, demonstrating ParsAirCall's efficiency in achieving a 2.7% WER (Word Error Rate) in number recognition in airport telephone calls.

Discussion: ParsAirCall emerges as a versatile tool, poised for seamless integration as a service into any Persian-language airport telephone system. Its practical application extends to number recognition in airport call centers, exemplifying the transformative impact of advanced technologies in streamlining communication processes within critical operational environments. ParsAirCall can be easily integrated as a service into any Persian-language airport telephone system, making it a practical tool for number recognition in airport call centers and telephone systems.

Keywords: Call center, automatic speech recognition, Deep transfer learning, Airport smart systems.

سامانه‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران
واحد نجف‌آباد

ParsAirCall: تلفن گویای محاوره‌ای خودکار در مرکز تماس فرودگاهی با استفاده از یادگیری انتقالی عمیق

تاریخ دریافت: ۱۴۰۲/۰۱/۱۶
تاریخ پذیرش:۱۴۰۲/۰۲/۲۷

دورهٔ چهارم، تابستان ۱۴۰۲
شمارهٔ دوم، صص: ۱۱-۲۰

سهیل تهرانپور[۱]، محمد منثوری[۲*]، سمانه یزدانی[۳]

۱- کارشناسی‌ارشد، هوش مصنوعی و رباتیک، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران. s.tehranipour@iau-tnb.ac.ir
۲- استادیار، گروه کنترل، دانشگاه شاهد، تهران، ایران. (نویسنده مسئول) mmanthouri@shahed.ac.ir
۳- استادیار، گروه کامپیوتر، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران. s_yazdani@iau-tnb.ac.edu

چکیده: در دنیای امروز با عنایت به پیشرفت‌های هوش مصنوعی، سامانه‌های سنتی موجود در تعامل با مخاطبین در تماس‌های تلفنی کارآمد نخواهندبود و این ناکارآمدی با اتوماسیون و خودکارسازی فعالیت‌های تکراری بهبود خواهدیافت. در این مقاله، ابزاری با نام ParsAirCall معرفی شده‌است که ابزاری برای باز شناسی خودکار گفتار فارسی اعداد در سامانه‌های فرودگاهی است که از یادگیری انتقالی عمیق برای بهبود عملکرد در محیط‌های واقعی و عملیاتی سامانه‌های هوشمند تلفنی گویا در فرودگاه‌های کشور می‌باشد. ParsAirCall با تعداد کمتر پارامترها و منابع محاسباتی، بهینه‌تر و نیز دقت بالاتر از مدل‌های رقیب عمل‌می‌کند. به‌دلیل محدودیت داده‌های موجود برای تشخیص گفتار در زبان فارسی، یک مجموعه دادگان ۳۰ ساعتهٔ تلفنی جمع‌آوری و برچسب‌گذاری شده و برای آموزش مدل نهایی مورد استفاده قرارگرفته‌است. آزمایش‌ها روی هر دو مجموعه داده تلفنی جمع‌آوری‌شده توسط ما و پروژهٔ Common Voice انجام‌شدند، که کارایی ParsAirCall را در دستیابی به ۲٫۷٪ میزان WER در تشخیص اعداد در تماس‌های تلفنی فرودگاهی نشان‌می‌دهد. ParsAirCall می‌تواند به‌راحتی به‌عنوان یک سرویس در سیستم‌های تلفنی فرودگاهی فارسی زبان ادغام‌شود، که آن را به یک ابزار قابل‌اجرا برای تشخیص اعداد در مراکز تماس و سیستم‌های تلفنی فرودگاهی تبدیل‌کند.

واژه‌های کلیدی: مرکز تماس، بازشناسی خودکار گفتار، یادگیری عمیق انتقالی، سامانه‌های هوشمند فرودگاهی.

۱. مقدمه

امروزه خدمات پس از فروش و پشتیبانی مشتریان یک بخش اساسی در ادارهٔ کسب و کارهای مدرن می‌باشد که به‌ویژه در سال‌های اخیر به دلیل مدل‌های جدید شغلی، استفاده گسترده از هوش مصنوعی، گسترش روند استفاده از تکنولوژی‌های نوظهور و بالاتر رفتن سطح انتظارات برای تجربه کاربری دگرگون شده‌است.[۱] ایده‌ها، پروژه‌ها و فناوری‌های جدید زیادی در حال کمک به رشد بازار هستند که همگی سعی در بهبود وضعیت موجود و رفع مانع می‌باشند. راه‌حل‌های سنتی مراکز تماس در عمل قادر به عملیات و پاسخگویی به مشتریان نخواهندبود؛ به‌عنوان نمونه می‌توان به افزایش فزایندهٔ حجم تماس‌های تلفنی به مراکز تماس برای دریافت خدمات پشتیبانی اشاره‌نمود که مشخصاً پس از شیوع کووید-۱۹ و متعاقباً افزایش بی‌سابقهٔ خدمات از راه دور، این صنعت را دچار آسیب جدی ساخت.

در پاسخ به وضعیت پیش‌آمده، بسیاری از کسب و کارها از فناوری‌های جدیدتر مانند VoIP و هوش مصنوعی برای اجرای عملیات خدمات پشتیبانی و ارائه سرویس به مشتریان خود استفاده می‌کنند. یکی از مهم‌ترین خدمات جدید در مراکز تماس مدرن، استفاده از تشخیص خودکار گفتار (ASR) است که به مدیران مراکز تماس کمک‌می‌کند تا انبوهی از داده‌های تلفنی از منابع مختلف مانند واکنش صوتی تعاملی (IVR) و امکان ضبط تماس‌ها را داشته‌باشند. به این ترتیب سامانهٔ هوشمند نصب‌شده در مراکز تماس، به تماس‌گیرنده اجازه‌می‌دهد تا فقط از طریق گفتار خود با مرکز تماس ارتباط برقرارکند؛ این در حالی است که در گذشته و راهکارهای سنتی، از صفحه کلید برای واردکردن داده‌های موردنظر استفاده‌می‌شد. در مراکز تماس نوین، مسیردهی و تکمیل تماس مخاطب، با تبدیل گفتار به متن و سپس پردازش آن متن برای ارجاع تماس به مقصد تماس‌گیرنده انجام‌می‌شود. علاوه بر این، در ترکیب با دیگر برنامه‌ها، فناوری‌های موجود به تماس‌گیرندگان اجازه‌می‌دهد تا خدماتی مانند بررسی موجودی حساب و احراز هویت خود را آنی و بدون نیاز به صحبت با اپراتور انسانی و یا ورود اطلاعات به‌صورت دستی انجام‌دهند که در نهایت منجر به تجربهٔ بهتر مشتری و پشتیبانی ۷/۲۴ در مراکز تماس می‌شود. اقدامات مذکور نمونه‌ای از مواردی است که منجر به صرفه‌جویی در هزینه و افزایش سطح رضایت در خدمات مشتری خواهدشد.[۲]

از سوی دیگر، در چند سال اخیر، چندین روش تبدیل گفتار به متن، توسط مدل‌های End-to-End (E2E) با استفاده از شبکه‌های عصبی و تکنولوژی‌های جدید نوظهور ساخته‌شده‌اند. برای آموزش مدل‌های گفتار به متن از ابتدا، نیاز به مقدار قابل‌توجهی دادهٔ برچسب‌خورده و توان محاسباتی فوق‌العاده داریم؛ با وجود این که بیش از ۶۰۰۰ زبان در سراسر جهان صحبت‌می‌شوند، تنها معدودی از زبان‌ها هستند که دارای مجموعه داده‌های غنی برای آموزش مدل‌های با کیفیت برای گفتار به متن می‌باشند. یادگیری انتقالی یک رویکرد معتبر برای توسعهٔ مدل‌های گفتار به متن برای زبان‌های با منابع کم مانند فارسی است که از مدل آموزش داده‌شده برای یک زبان دیگر استفاده می‌کند.

هدف ما از این رویکرد، جبران کمبود داده مورد نیاز به جهت آموزش مدل‌های گفتار به متن فارسی برای استفاده در مراکز تماس و مراکز تلفن می‌باشد. طبق تحقیقات صورت‌گرفته، گفتار به متن فارسی تا به امروز در مراکز تماس هنوز عملیاتی نشده‌است. با توجه به افزایش محبوبیت مراکز تماس در میان فارسی‌زبان، ما تأثیر یادگیری انتقالی بر سامانه گفتار به متن فارسی به‌خصوص چگونگی اتوماسیون کامل فرآیند تماس در سیستم پاسخ صوتی تعاملی در صورت برقراری ارتباط با تماس‌گیرنده برای ارتباط با مخاطبین را بررسی‌نموده‌ایم. در نهایت در این مقاله، نتایج نهایی یادگیری انتقالی در زبان فارسی را با درنظرگرفتن مدل‌های مختلف روز دنیا ارائه‌می‌دهیم. همچنین، ما کشف‌کردیم که اگر مجموعه داده نهایی کوچک‌تر، هم‌اندازه یا حتی به‌طور قابل‌ملاحظه‌ای بزرگتر از مجموعه داده آموزشی باشد، به‌طرز عجیبی، استفاده از یک مدل پیش‌آموخته همچنان نسبت به آموزش کلیه داده‌ها از ابتدا، نتایج بهتری خواهدداد. نوآوری‌های مقاله حاضر به شرح زیر خواهدبود:

- ارائه یک سیستم تشخیص گفتار فارسی تک‌زبانه برای تشخیص اعداد فارسی با دقت بالا با استفاده از یادگیری انتقالی.
- عملکرد بهتر ParsAirCall در مقایسه با سایر معماری‌های توسعه داده‌شده برای زبان فارسی.
- با توجه به هدف ما جهت استفاده عملیاتی در مراکز تماس، ParsAirCall از نظر مصرف منابع سخت‌افزاری و بار پردازشی بهینه شده است.
- در راهکار نهایی، ParsAirCall جهت استفاده در مراکز تماس و هوش مصنوعی گفتگومحور به زبان فارسی پیاده‌سازی و بهینه شده‌است.

در بخش ۲ به بررسی مطالعات اخیر در حوزه بازشناسی گفتار خواهیم‌پرداخت و در بخش ۳ مجموعه دیتاستِ این پژوهش ارائه خواهدشد. بخش ۴ روش پیاده‌سازی و بخش ۵ تحلیل گزارش نهایی خواهدبود.

۲. پیشینه پژوهش

در طول دههٔ گذشته، پردازش تصویر، بازشناسی خودکار گفتار، پردازش زبان طبیعی و دیگر حوزه‌های هوش مصنوعی، پیشرفت عظیمی داشته‌است. این پیشرفت عمدتاً به‌دلیل حضور بلامنازع یادگیری عمیق و انواع مختلف آن می‌باشد. یادگیری عمیق، این امکان را می‌دهد که در صورت

ساخت مدل با داده‌های آموزشی متنوع و کیفیت بالا، پیش‌بینی‌ها بسیار دقیق صورت‌گیرد و همان‌طور که در پیشتر اشاره‌شد، با استفاده از راهکارهای ارائه‌شده، اغلب یکی از معماری‌های استاندارد و معروف موجود استفاده‌می‌شود. از سوی دیگر، اگر مجموعه داده بزرگی برای مسئلهٔ موردنظر در اختیار نباشد، چارهٔ کار در استفاده از یادگیری انتقالی است؛ در این حالت، با استفاده از یک شبکهٔ از پیش آموزش‌دیده بر روی حجم عظیمی از داده‌ها و تنظیم برخی از لایه‌های انتهایی آن، مسئله موردنظر قابلیت اجرا خواهدداشت.

اغلب تحقیقات اولیه با محوریت بازشناسی خودکار گفتار، در دهه‌های ۵۰ و ۶۰ میلادی بر روی کلمات به‌صورت ایزوله و وابسته به صحبت کاربر انجام‌شده‌است. تلاش‌هایی نیز در سطح واج صورت‌گرفت، اما با توجه به دشواری بالقوهٔ این پروژه، موفقیت‌های چشمگیری حاصل نشد. به عنوان یکی از ابتدایی‌ترین موارد پیاده‌سازی شده، در اوایل سال ۱۹۵۲ درآزمایشگاه‌های موسوم به آزمایشگاه بل، سیستم‌های کلمات جداشده ارائه‌گردید که قابلیت تشخیص ارقام خوانده‌شده توسط یک مرد بالغ را داشت؛ نکته قابل‌توجه، نرخ دقت ٪۹۹ در تشخیص ارقام مجزای این روش می‌باشد. این سیستم برای تشخیص کل کلمات، تنها قادر به تشخیص در فرکانس‌های موجود بوده و هیچ الگویی از هجاها یا صامت‌ها یا مصوت‌ها یا هر نوع واحد زیرکلمه‌ای در این سیستم‌ها وجودندارد. کلمه به‌عنوان یک واحد درنظرگرفته‌می‌شود و در طول طبقه‌بندی، تمام کلمات برای یافتن بهترین تطابق با یکدیگر مقایسه‌می‌شوند. [۳]

مدل‌های مارکوف پنهان(HMM) در دهه ۷۰ میلادی، مدل‌های مخلوط گاوسی (GMM) نیز در دهه ۸۰ میلادی معرفی‌شدند، اما تا دهه ۹۰ میلادی زمان لازم بود تا هر سه الگوریتم در یک راهکار ادغام‌شوند؛ که آن را HMM Toolkit و به‌اختصار HTK نامیدند. HTK، اولین سامانه جهت ترکیب تمام اجزای اصلی شناسایی گفتار مبتنی برGMM-HMM مدرن بود که در سال۱۹۹۲ به‌عنوان الگوریتم متن‌باز در اختیار عموم قرارگرفت.[۴] استفاده از روش مرسوم تشخیص گفتار، یعنی HMM به همراه GMM، به حدی رسید که بر روی پروژه‌های تشخیص گفتار پیوسته مستقل از گوینده با واژگان بسیار گسترده کارایی بالایی پیداکرد و عمدهٔ پروژه‌های تشخیص خودکار گفتار، از این الگوریتم استفاده‌نمی‌کرد. اما در این‌گونه روش‌ها فرض‌های مختلفی درنظرگرفته‌می‌شود که برخی غیرواقعی بوده و در عمل چالش بزرگی را پیش‌روی کاربردهای تجاری و صنعتی می‌گذارد. به‌عنوان مثال فرض‌می‌شود که ویژگی‌های استخراجی در یک بخش الفبایی با یکدیگر همبسته نیستند که این مورد نادرست است. از طرف دیگر، به‌منظور پیاده‌سازی عملیاتی پیچیدگی‌های بسیار زیادی را برای مدل نهایی به همراه خواهندداشت. روش‌های مبتنی‌بر شبکه‌های عصبی راهی برای کاهش وابستگی‌های سیستم به فرضیات غیرواقعی درباره گفتار معرفی‌نمودند. زیرا شبکه‌های عصبی، توانایی یادگیری همبستگی بین ویژگی‌ها را دارا می‌باشد؛ بنابراین گامی رو به جلو برداشته‌شد و کارایی بیشتر روش HMM موجب تولید سیستم‌های تشخیص گفتار بر پایه آن شد.

اما بازشناسی گفتار، پیوسته یک موضوع حل‌نشده باقی‌مانده‌است که همچنان چالش‌های بی‌شماری در تولید یک سامانهٔ کامل با دقت بالا را پیش‌روی خود می‌بینند. مشکلات موجود در دنیای واقعی عبارتند از: اندازهٔ بزرگ دیکشنری، تشخیص‌دهنده‌های مستقل از گوینده (در مقابل وابسته به گوینده که ساده‌تر است)، پدیده گفتار ناخواسته (کلمات خارج از دیکشنری) لهجه و گویش‌های ناآشنا، حالت‌های مختلف صحبت (لحن)، نویز پس‌زمینه، صداهای انعکاسی اتاق، صحبت‌های عامیانه که در دایره لغات آموزشی نیستند، همگی بخشی از این چالش‌ها به‌شمارمی‌روند.[۵]

شبکه‌های عصبی مصنوعی (ANN) را می‌توان برای طبقه‌بندی واحدهای گفتاری مانند کلمات یا واج‌ها به‌کاربرد اما نمی‌توان از آن‌ها برای تشخیص گفتار استفاده‌کرد، زیرا تعداد دنباله‌های کلمات ممکن در یک گفتار نامحدود است و شبکه عصبی دارای محدودیت اندازه است؛ حال آنکه افزایش اندازهٔ ورودی در یک شبکه عصبی موجب افزایش تعداد نورون‌های لایه مخفی و پارامترها می‌شود که چالش‌ها و سختی‌های خود را دارد. ایده اصلی این‌گونه است که از شبکه‌های عصبی برای طبقه‌بندی‌های محلی و از HMM برای مدل‌کردن ساختار دنباله‌ای داده‌ها استفاده‌شود. HMM توانایی بخش‌بندی خودکار دنباله‌های ورودی در طول آموزش را دارد و حتی می‌تواند یک روش اصولی برای تبدیل کلاس‌بندی‌های شبکه به دنباله‌ای از برچسب‌ها تهیه‌کند. بنابراین برخلاف شبکه‌های عصبی، ترکیب‌های ANN-HMM را می‌توان به‌صورت مستقیم در کارهای طبقه‌بندی زمانی مانند بازشناسی خودکار گفتار به‌کار برد.[۶]

در ادامه روش‌های پیشین، انتشار مقاله هینتون و همکاران در سال ۲۰۱۲ با عنوان «شبکه‌های عصبی عمیق برای شناسایی گفتار»، قدرت شبکه‌های عصبی به‌ویژه یادگیری عمیق را در تشخیص گفتار تثبیت‌نمود.[۷] نخستین رویکردها برای رسیدن به این هدف، سعی در درک معنای گفتار با استفاده از تبدیل آن به متن و سپس بررسی قواعد نوشته‌شده توسط انسان بود. با این‌حال، چنین سیستم‌های قاعده‌محوری، محدود به حوزه‌های خاصی بودند که برای آن‌ها طراحی شده‌بود و عموماً قادر به مقابله با ورودی‌های غیرمنتظره یا ورودی‌های جدید نبودند و در نهایت ثابت‌شد که برای به‌دست آوردن پیچیدگی زبان طبیعی بسیار محدودند. به کل بخش استخراج ویژگی و محاسبهٔ احتمال برچسب‌ها پردازش صوتی و بخش محاسبه احتمال (انتخاب برچسب بر حسب ویژگی‌های استخراجی)، مدل صوتی گفته‌می‌شود؛ که در [۷] از آن استفاده‌شده‌است.

از طرف دیگر، طی ۲۰ سال گذشته، رویکرد آماری بازشناسی گفتار خودکار نیز متداول‌شده‌است که از مدل‌های ریاضی برای یادگیری خودکار قوانین استفاده‌می‌کند. در نتیجه، به‌جای نوشتن قوانین سخت و پیچیده که انعطاف زیادی هم ندارند، فرآیند پیشرفت بازشناسی گفتار خودکار، به سمت ایجاد ویژگی‌هایی که بیانگر روابط بین داده‌هاست، هدایت‌گردید. با این همه، تهیه و استخراج ویژگی‌های مهندسی زمان‌بر

است، زیرا این ویژگی‌ها معمولاً وظیفه‌محور بوده و به تخصص در حوزه موردنظر نیاز دارند.

یک فرض اساسی در یادگیری آماری باناظر این است که داده‌های آموزشی برچسب‌دار دارای توزیع مستقل و یکنواخت هستند که در نهایت مدل را بر روی نمونه‌های آزمایشی، اجرامی‌کنند. اما در صورت پایین‌بودن خطای آموزش، نمی‌توان عملکرد مناسب در مواجهه با داده‌های آزمون را تضمین‌نمود. به‌ویژه زمانی‌که با داده‌های صوتی و متنی سروکار داریم، این فرض معمولاً به این معنی است که داده‌های آموزشی و آزمایشی باید از یک دامنه باشند. مثلاً، اگر کسی بخواهد یک اپلیکیشن با کاربرد تشخیص موجودیت‌های اسمی (NER) برای وبلاگ‌های شخصی ایجادکند، در حالت عادی باید به‌جای آموزش مدل بر روی مقالات خبری مالی (که اغلب برچسب‌دار هستند)، آن را با استفاده از موجودیت‌های اسمی آموزش‌دهد زیرا توزیع داده‌ها در حوزه خبری و حوزه‌های متفاوت است.[۸] یادگیری عمیق باناظر، متکی بر مقدار کافی از داده برچسب‌گذاری‌شده برای آموزش است؛ به همین دلیل، زبان‌شناسان محاسباتی تلاش زیادی کرده‌اند تا پیکره‌های متنی برچسب‌گذاری‌شده مختلفی برای مقاصد گوناگون تهیه‌کنند.[۹] از سوی دیگر، پیکره‌های ثبت‌شده موجود، حتی در صورت کار بر روی دامنه‌ای دیگر نیز، مفید خواهندبود. بنابراین، استفاده از روش‌هایی که بتوانند به‌عنوان میان‌بر، تولید نرم‌افزارهای نهایی را سرعت‌دهند، می‌تواند ارزش افزودهٔ بالایی را به ارمغان آورد. بنابراین، برای رسیدگی به مشکل گلوگاهی با فقدان وجود تعداد کافی از داده‌های برچسب‌دار، یک سوال تحقیقاتی مهم و جالب این است که چگونه از داده‌های برچسب‌دار موجود از دامنه‌های مرتبط برای آموزش طبقه‌بندها در دامنه‌ای که ما به آن علاقه داریم، استفاده‌کنیم. به‌طور واضح‌تر، دامنه‌ای که داده‌های برچسب‌دار کافی ندارد ولی مورد علاقه ماست را هدف و دامنه‌ای که داده‌های برچسب‌دار کافی دارد، دامنهٔ منبع تعیین می‌کنیم. جهت حلاولین روش‌های شناسایی گفتار پیوسته، از تابع هدف طبقه‌بندی زمانی متصل یا CTC استفاده‌کردند[۱۱]. مدل توالی به توالی با رویکرد مکانیزم توجه (Attention Mechanism) خیلی زود تبدیل به مدل محبوب محققین شد. یک نسخهٔ قابل‌توجه از مدل توالی به توالی با توجه، LAS است که در ادامه پس از استفاده از در مدل‌های بازگشتی، روش Wav2Letter معرفی و بر روی لایه‌های پیچشی و اصلاح تابع اتلاف CTC ساخته‌شد. [۱۲] در طول پنج سال گذشته، مدل‌های پیوس ته کارآمدتر شده‌اند، اما معمولاً در مجموعه داده‌های عظیم آموزش می‌بینند که تنها در دسترس شرکت‌های بزرگ هستند. لذا، پژوهش حاضر با هدف چگونگی تعمیم این مدل‌های پیوسته، برای مجموعه داده‌های کوچک‌تر، انجام‌شده‌است. در اواخر سال ۲۰۱۴، روش DeepSpeech تحت‌نظر اندرو اِنگ، پیشنهادشد که از مدل شبکه عصبی بازگشتی (RNN) استفاده‌نموده‌است.[۱۳]

مشکل مذکور، راهکار موسوم به تطبیق دامنه پیشنهادشده‌است که از داده‌های برچسب‌دار دامنه منبع برای کمک به آموزش طبقه‌بند در دامنه هدف استفاده‌می‌کند. [۱۰] این موضوع یکی از زیر مجموعه‌های اصلی یادگیری انتقالی است. طبق مقالات و تحقیقات منتشرشده در حوزهٔ یادگیری ماشین به‌ویژه در سال‌های اخیر، شاهد رشد روزافزون توجه‌ها به یادگیری انتقالی و تطبیق دامنه هستیم. در محدوده پنج سال گذشته، شبکه‌های عصبی عمیق، به‌عنوان یک دسته خاص از مدل‌های یادگیری ماشین، با توجه به کارآیی فوق‌العاده‌ای که از خود نشان‌دادند به مدل محبوب و منتخب محققین و متخصصین علوم داده و هوش‌مصنوعی تبدیل‌شدند که باتوجه به نوع عملکردشان، نیاز به مهندسی ویژگی و تعیین ویژگی‌ها به‌صورت دستی و توسط نیروی انسانی خبره را کاهش‌می‌دهد. در نتیجه انرژی نیروی انسانی بر تعیین مناسب‌ترین معماری و تنظیمات آموزشی برای هر کار متمرکز خواهدشد. نسل فعلی مدل‌های بازشناسی گفتار خودکار مبتنی بر شبکه عصبی این برتری را دارند که می‌توانند بر روی مقادیر بسیار زیاد داده‌های برچسب‌دار آموزش ببینند. در بازشناسی گفتار خودکار و متعاقب آن، پردازش زبان طبیعی و دیگر حوزه‌های یادگیری ماشین، روش استاندارد برای آموزش یک مدل به این ترتیب خواهدبود که ابتدا باید تعداد زیادی از نمونه‌های جمع‌آوری‌شده برچسب‌گذاری‌شود. سپس این داده‌ها به‌عنوان ورودی به یکی از مدل‌های موجود ارائه می‌شود و این مدل طی فرآیند آموزش، یادمی‌گیرد که از ورودی به خروجی نگاشت‌شود؛ این مورد با عنوان یادگیری باناظر شناخته‌می‌شود. تعداد نمونه‌هایی که باید برای هر موضوع جدید برچسب زده‌شوند، با توجه به فراوانی زبان‌ها، وظایف و حوزه‌ها در دنیای واقعی و تنظیمات مدل پیشنهادی، متفاوت بوده و به‌عنوان چالش‌های دنیای واقعی تلقی‌می‌شوند. در سال‌های اخیر، بازشناسی خودکار گفتار پیوسته بر روش‌های سنتی، ترجیح داده‌شده‌است زیرا برای تخمین پارامترها تنها یک تابع جهت محاسبه Loss وجود دارد. این مقدار Loss، به ما اجازه‌می‌دهد تا مستقیماً مسئله مورد نظرمان را بدین ترتیب، مدل کنیم و بهترین رونوشت ممکن برای یک فایل صوتی را به‌دست بیاوریم. با این حال، با حذف HMM ها از رویکرد ترکیبی، مدل‌های پیوسته، به اطلاعات هم‌ترازی که معمولاً در آموزش مدل‌های سنتی مورد استفاده قرارمی‌گیرند، دسترسی ندارند. به این ترتیب، مدل تشخیص گفتار سرتاسری، باید هم‌ترازی متن و صدا را در طول آموزش بیابد؛ که قطعاً کار ساده‌ای نیست.

در ادامه روش مذکور، مدل توسعه‌یافته آن با نام DeepSpeech2 در سال ۲۰۱۶ ارائه‌گردید که نسبت به روش پیشین تا ۴۳٪ دقیق‌تر و ۷ برابر سریعتر می‌باشد. این مدل بر روی ۱۱,۹۴۰ ساعت داده زبان انگلیسی و ۹,۴۰۰ ساعت داده زبان چینی آموزش داده‌شده‌است. یکی دیگر از ابداعات این روش استفاده از تابع بهینه‌ساز جدیدی به نام SortaGrad است که از استراتژی آموزش برنامه‌درسی بهره‌می‌گیرد. معماری اصلی این مدل، در مجموع شامل ۱۱ لایه است

که در ابتدا سه لایهٔ پیچشی، سپس هفت لایه بازگشتی دوطرفه و در نهایت یک لایه تمام متصل را شامل‌می‌شود. خروجی نهایی همانند مدل DeepSpeech به الگوریتم CTC تحویل‌داده‌می‌شود.[۱۴]

هیسائو و همکاران در سال ۲۰۲۰، در ادامه روش LAS، نسخه آنلاین آن را ارائه‌نمودند تا نگاه‌ها به نسخه‌های بعدی این روش یعنی DeepSpeech3 دوخته‌شود که طبق گفته شرکت بزرگ Baidu می‌تواند گام بزرگ دیگری در بازشناسی خودکار گفتار در تمامی زبان‌ها باشد.[۱۵]

روش دیگری نیز توسط دپارتمان هوش‌مصنوعی شرکت فیس‌بوک، ارائه و با نام تجاری Wav2letter معرفی‌شد. کولوبرت و همکاران در اواخر سال ۲۰۱۶ آن را ارائه‌کردند [۱۶] که با تغییر نگرش در معماری و نحوه استفاده از بازنمایش داده‌های صوتی، توانست دریچه‌ای برای شروع مطالعات بعدی بگشاید. این روش دو سال بعد یعنی سال ۲۰۱۸ در مقاله‌ای با نام Wav2letter++ ادامه‌یافت که خود را با عنوان سریعترین روش متن‌باز معرفی‌نمود. طبق اعلام تیم فنی فیس‌بوک، روش Wav2Vec2.0 تنها با داشتن ۱۰ دقیقه داده صوتی برچسب‌دار می‌تواند به دقت بالایی در بازشناسی خودکار گفتار برسد. [۱۷] در این مدل‌ها که عموماً در یادگیری انتقالی، به‌ویژه برای زبان‌های با داده برچسب‌دار کم استفاده‌خواهندشد، امکان بهبود عملکرد سامانه‌های بازشناسی در داده‌های تلفنی نیز خواهدداشت [۱۸]

اخیراً نیز الگوریتم Whisper توسعه‌یافته توسط OpenAI نتایج چشم‌گیری را در تشخیص گفتار با استفاده از یادگیری خودنظارتی به همراه یک تکنیک خوشه‌بندی نوآورانه نشان‌داده‌است. هنگام مقیاس‌بندی به ۶۸۰،۰۰۰ ساعت از نظارت چندزبانه و چندوظیفه‌ای، مدل‌های حاصل تعمیم خوبی به بنچمارک‌های استاندارد دارند و در مقایسه با انسان‌ها، مدل‌ها عملکرد دقیقی داشته‌اند. این پیشرفت‌ها در زمینه یادگیری خودنظارتی قابلیت توسعه بسیار بزرگی را برای پروژه‌های علمی که از یادگیری ماشین استفاده‌می‌کنند، دارند زیرا اجازه‌می‌دهند مدل‌های دقیق‌تری با داده‌های کمتر برچسب‌خورده آموزش‌داده‌شوند.[۱۹]

با وجود مطالعات اخیر در زمینه تشخیص گفتار فارسی، چالش‌ها همچنان در دستیابی به تشخیص به‌صورت زمان واقعی باقی‌مانده‌اند، به‌خصوص در مراکز تماس که داده‌های تلفنی ممکن است چالش‌های بزرگی داشته‌باشند. به عبارت دیگر، نیاز به تحقیقات بیشتری برای توسعه سیستم‌های تشخیص گفتار اتوماتیک واقعی و تعاملی در محیط‌های عملیاتی مراکز تماس و همچنین مدیریت پیچیدگی‌های داده‌های تلفنی وجوددارد. با حل این چالش‌ها، سیستم‌های تشخیص گفتار خودکار، پتانسیل تحولی در مراکز تماس به زبان فارسی را دارند و می‌توانند ارتباط مشتریان و عوامل مرکز تماس را بهبودبخشند.

۳. مجموعه دادگان

داده مورد نیاز جهت بررسی و پیاده‌سازی و آموزش یک سیستم بازشناسی گفتار دارای ساختار مشخصی است به شکل تعداد مشخصی از فایل‌های صوتی شامل گفتار مرجع و متن متناظر آن گفتار. با توجه به نیاز الگوریتم‌های یادگیری عمیق به دادگان آموزشی زیاد و رشد قدرت محاسبات توسط کارت‌های گرافیکی در سال‌های اخیر، جهت به‌دست‌آوردن مدلی مطلوب، حجم دادگان آموزشی بسیار مهم است. مجموعه دادگان گفتاری جامعی تهیه و جمع‌آوری شده‌است که شامل ۳۰ ساعت دادگان گفتاری برچسب‌دار در زبان فارسی می‌باشد. همان‌طور که پیشتر تبیین‌گردید، یکی از چالش‌های موجود، نبود داده‌های کافی در زبان فارسی جهت آموزش و ساخت مدل برای اجرا در این زبان است؛ بنابراین ارتباط تنگاتنگ با صنعت می‌تواند بخش زیادی از این مسئله را هموارسازد. با عنایت به توافقات صورت‌گرفته با شرکت‌های دانش‌بنیان فعال در حوزه راهکارهای نسل جدید مخابراتی و مراکز تماس، امکان تأمین بخشی از این داده‌ها جهت ساخت یک مدل جدید در زبان فارسی حاصل‌شد. طبق جدول ۱، ۳۰ ساعت فایل صوتی ضبط‌شده در مکالمات تلفنی مراکز تماس فارسی زبان، شامل ۱۰هزار فایل صوتی (هر فایل صوتی با مدت زمان تقریباً ۱۰ ثانیه) گردآوری و برچسب‌گذاری شده‌است. این مجموعه دادگان شامل فایل‌های صوتی تلفنی گسترده‌ای از اعداد ارقام به‌صورت مجزا، شماره تلفن، کدملی، شماره شبا و ترکیب‌های مختلف عددی در زبان فارسی است.

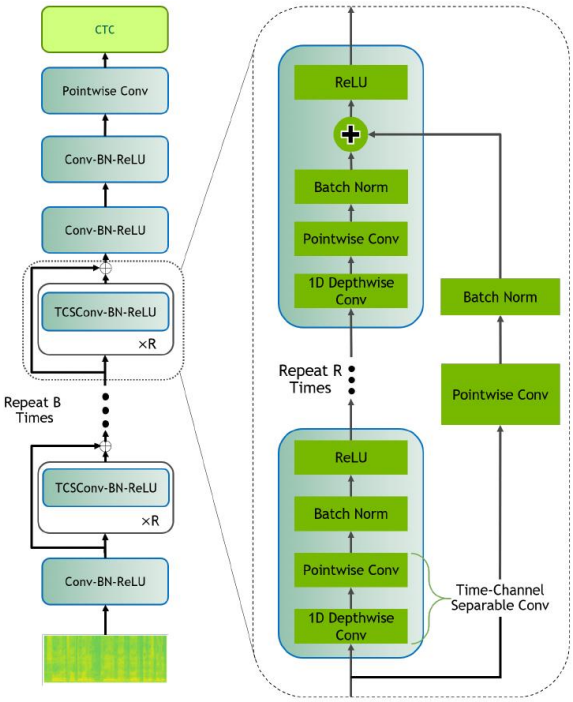
جدول ۱: مجموعه دادگان

مدت زمان فایل صوتی (ساعت)	تعداد فایل صوتی	دیتاست مورد استفاده
۳۵۵	۳۷۶۵	Common Voice
۳۰	**۱۰۰۰۰**	مرکز تماس، برچسب‌گذاری‌شده
۳۸۵	**۱۳۶۷۵**	**مجموع**

یکی دیگر از منابع دادگان ما در این تحقیق، پروژهٔ متن‌باز Common Voice است که توسط شرکت Mozilla برای ایجاد یک پایگاه داده رایگان جهت تشخیص گفتار در سال ۲۰۱۷ آغاز شده‌است.[۲۰] این پروژه توسط داوطلبانی پشتیبانی‌می‌شود که نمونهٔ جملات را با میکروفون ضبط‌می‌کنند و فایل‌های ضبط‌شده سایر کاربران را بررسی و تأیید می‌کنند. جملات رونویسی‌شده در یک پایگاه داده صوتی که تحت مجوز مالکیت عمومی CC0 موجود است جمع‌آوری خواهدشد. این مجوز اطمینان‌می‌دهد که توسعه‌دهندگان می‌توانند از پایگاه داده برای برنامه‌های صوتی به متن بدون محدودیت و هزینه استفاده‌کنند. تمامی ۹،۲۸۳ ساعت ثبت‌شده در این مجموعه داده نیز شامل فرادادهٔ جمعیتی مانند سن، جنس و لهجه است که می‌تواند به آموزش دقت موتورهای تشخیص گفتار کمک‌کند. این مجموعه داده در حال حاضر، شامل ۷،۳۳۵ ساعت تأییدشده به ۶۰ زبان است، اما همچنان در حال گسترش و افزودن داده‌های صوتی می‌باشد.

در این تحقیق برای زبان فارسی از ورژن ۱۳ پروژهٔ Common Voice شامل ۳۵۵ ساعت داده صوتی تأییدشده و درمجموع ۳۹۲ ساعت داده صوتی، استفاده‌شده که به‌صورت رایگان و متن‌باز در اختیار همگان قرارگرفته‌است. هر ورودی در مجموعه داده شامل یک فایل MP3 و متن مربوط به آن است.

علاوه بر این، پروژهٔ Common Voice در زبان فارسی فاقد یک استاندارد و راهنمای کامل جهت برچسب‌گذاری است و این مسئله موجب مشکلات قابل‌توجهی در فرآیند آموزش مجموعه داده می‌شود. مسئله نرمال‌سازی متون فارسی از بسیاری جهات حائز اهمیت است. متاسفانه به‌دلیل شرایط زبان فارسی که در ساخت کلمات جدید وابستگی زیادی به پیشوندها و پسوندها دارد نحوهٔ نوشتاری کلمات تنوع زیادی دارد. به‌عنوان مثال پسوند جمع "ها" به سه صورت در متون وجوددارد که این تنوع در لغت‌نامه باعث‌می‌شود یک کلمه به سه صورت نوشته‌شود و دقت سیستم را کاهش‌دهد. مثلاً جمع کلمه "درخت" به سه صورت "درختها"، "درخت ها" و "درختها" نوشته‌می‌شود که در نگارش زبان فارسی تنها صورت درست با نیم‌فاصله است. یکی دیگر از مشکلات پیش روی پردازش متن در زبان فارسی، تنوع در یونیکُدهای زبان فارسی ناشی از تفاوت در کیبوردهای مختلف و مشابهت حروف عربی و فارسی در این کیبوردهاست. حروفی مانند "ی" و " ک" بیشترین گوناگونی را از لحاظ یونیکد دارند. این نرمال‌سازی متون فارسی مستقیماً در دقت و نتایج لغت‌نامه آوایی و مدل زبانی تأثیرمی‌گذارد و در نهایت دقت سیستم بازشناسی گفتار را افزایش خواهدداد. بسیاری از نمونه داده‌های صوتی در مجموعه داده مذکور نیاز به نرمال‌سازی متون دارد که شامل حذف ویرگول‌ها و علائم نگارشی، علامت تعجب و کاراکترهای غیرواژه‌نامه‌ای دارد. همچنین تبدیل نمادهای انگلیسی و عربی به فارسی، حذف ایموجی‌ها، جایگزین کردن متن‌های فارسی قابل-فهم با نمادهای ریاضی، تبدیل تاریخ و زمان به متن، جایگزینی اختصارات با تلفظ آن‌ها و حذف برخی از کاراکترهای عربی در مجموعه داده مانند تنوین از اقدامات صورت‌گرفته جهت نهایی‌سازی مجموعه داده فوق است. ماهیت دادهٔ صوتی مناسب آموزش سیستم‌های بازشناسی خودکار گفتار نمی‌باشد.



شکل ۱: معماری مدل QuartzNet [۲۱]

۴. پیاده‌سازی و ساخت مدل

همان‌طور که در تمامی زوایای زندگی همه ما مشهود است، استفاده از سیستم‌های کامپیوتری و نیز وسایل شخصی هوشمند به‌طور فزاینده‌ای رواج‌یافته‌اند. همچنین، به‌دلیل بسیاری از کاربردهای بالقوه این سیستم‌ها، یکی از جنبه‌های اصلی ارتباطی افراد که مکالمه و گفتگو می‌باشد نیز پابه‌پای بقیه کاربردها، باید از علوم روز و صدالبته هوش‌مصنوعی، استفاده‌نماید. هر نرم‌افزاری با محوریت کاربردی مکالمه و گفتگو، از چندین ماژول اساسی دیگر مانند تشخیص خودکار گفتار تشکیل‌شده‌است که مدل‌های استفاده‌شده هر چقدر سبک‌تر و بهینه‌تر باشند، نتیجهٔ نهایی سریع‌تر و جذاب‌تر خواهدبود؛ به‌ویژه در جایی که اکثر دستگاه‌ها کوچک و کم‌حافظه با قدرت پردازش محدود هستند. با این‌حال، بیشتر مدل‌های پیشرفته بسیار بزرگ و سنگین هستند که معمولاً دارای چندصد میلیون پارامترند. با توجه به محدودیت‌های فعلی دستگاه‌ها، عملاً استقرار آن‌ها در مقیاس بزرگ دشوار و حتی در موارد زیادی، غیرممکن خواهدبود. برای مقابله با چالش فوق، از الگوریتم QuartzNet استفاده‌خواهیم‌کرد که یک معماری مبتنی‌بر یادگیری عمیق جهت پروژهٔ بازشناسی خودکار گفتار می‌باشد.[۲۱] این مدل یک مدل End-to-End است که بزرگترین مدل آن، تقریباً ۱۹ میلیون پارامتر دارد و بادقت بسیار قابل‌قبولی، در مجموعه دادگان LibriSpeech، به ۲٫۶٪ خطای کلمه(WER) با استفاده از الگوریتم پردازش زبان طبیعی Transformer-XL می‌رسد. مدل QuartzNet همچنین توانایی یادگیری انتقال خوبی را نشان‌داده‌است و می‌تواند از

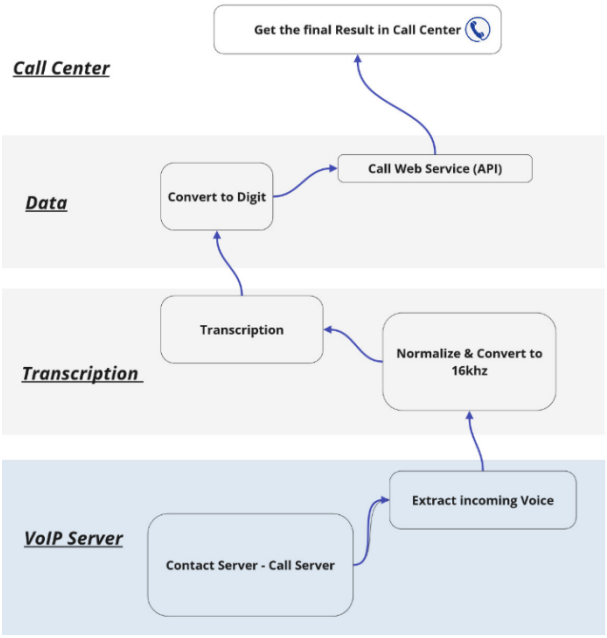
یک نمونه از پیش‌ذخیره‌شده در یک مجموعه دادهٔ کوچک‌تر با دامنه محدود و محتوای تخصصی‌تر تنظیم‌شود تا در نهایت به دقت بسیار بالا در تشخیص خودکار گفتار دست‌یابد. QuartzNet از معماری Jasper مشتق‌شده‌است، هر دو مدل، مبتنی بر پیچش هستند که از بلوک‌های پیچشی، نرمال‌سازی دسته‌ای، ReLU و Dropout تشکیل‌شده‌اند و در انتها یک لایه با مقدار هزینه ناشی از CTC دارند.

جدول ۲: نمونه‌ای از داده‌های نرمال‌سازی‌شده

برچسب نرمال‌سازی‌شده جهت آموزش PARSAIRCALL	برچسب موجود در دیتاست اصلی COMMON VOICE
آنقدر خسته ام که ممکن است بیفتم.	آنقدر خستهام که ممکن است بیفتم.
بسیار مشتاق دیدار ملکه زیبایی بودم	بسیار مشتاق دیدار ملکهی زیبایی بودم!
عمدا آهسته از پله ها پایین رفت	عمداً آهسته از پلهها پایین رفت
صادقانه نظریه خود را اظهار کردیم	صادقانه نظریهٔ خود را اظهار کردیم
مراتب تسلیت صمیمانه مرا بپذیرید	مراتب تسلیت صمیمانهی مرا بپزیرید
اسطوره رستم و سهراب	اسطورهی رستم و سهراب
برنامه جبران عقب افتادگی برای کودکان کوخ نشین	برنامهی جبران عقبافتادگی برای کودکان کوخنشین
چهره چروکیده یک پیرمرد	چهرهی چروکیدهی یک پیرمرد
اختلاف عیقده ای در بین ما نیست	اختلاف عیقدهای، در بین ما نیست
او اهمیت آموزش را مورد تاکید قرار داد.	او اهمیت آموزش را مورد تاءکید قرار داد.
زیرساخت غم انگیز این شعر	زیرساخت غمانگیز این شعر
جلسه کمیته ساعت چهار خاتمه یافت.	جلسهی کمیته ساعت چاهار خاتمه یافت.
به شدت باد تندی می آمد	به شدّت باد تندی میآمد
یکی از جاسوسان کا گه به	یکی از جاسوسان کا گ ب
نهصد و هجده شصت و شش	نهصد و هیجده شصت و شیش

سیگنال خام گفتار که معمولاً با نرخ فرکانس ۱۶کیلوهرتز (میکروفون موبایل یا کامپیوتر) یا ۸کیلوهرتز (سیستم‌های تلفنی) نمونه‌برداری شده‌است، پس از قاب‌بندی و پنجره گذاری که معمولاً با نرخ ۲۵میلی‌ثانیه و میزان همپوشانی ۱۰ تا ۱۵میلی‌ثانیه انجام‌می‌شود را به‌عنوان ورودی به کارمی‌بریم. سپس از قاب ها ویژگی های آن را استخراج‌می‌کنیم. به‌طور معمول از ۱۲ ضریب اول MFCC، مشتق اول و دوم آن به‌عنوان ویژگی استخراجی استفاده‌می‌نماییم و آن را به بخش بعدی انتقال‌می‌دهیم تا احتمال برچسب آن بخش از گفتار را محاسبه‌کند و برچسب مورد نظر را خروجی دهد.

تنظیم نرخ نمونه‌برداری: نرخ نمونه‌برداری باید میان تمامی داده‌ها یکسان و برابر عدد ۱۶ هزار باشد.



شکل ۲: فلوچارت خروجی مدل ParsAirCall از ابتدا تا انتها

تنظیم تعداد کانال: باید تمامی داده‌های ورودی دارای یک کانال صوتی باشند که این مهم به‌ویژه در مکالمات مراکز تماس از اهمیت زیادی برخوردارست.

آموزش مدل شبکه عصبی عمیق: در این بخش پس از پیش‌پردازش داده‌ها و ایجاد یک مدل پایه، مدل‌های شبکه عصبی عمیق پیاده‌سازی شده و درصد خطا بر روی یک داده آزمایشی ثابت بدست آمده است. این مجموعه دادگان آزمایشی شامل ۱۰۰۰ فایل صوتی، به‌صورت کاملاً تصادفی از بین داده‌های جمع‌آوری‌شده تلفنی مراکز تماس انتخاب‌شده‌است. ابزارهای مورد استفاده جهت پیاده‌سازی مدل‌های شبکه عصبی عمیق و آموزش این مدل‌ها در ادامه ارائه‌خواهدشد. به‌منظور اجرای پروسه آموزش مدل‌ها نیز از دو عدد GPU مدل Nivida RTX3070 استفاده‌شده‌است و نحوه آموزش مدل به صورت موازی و Multi-GPU با استفاده از کتابخانه PyTorch Lightning و قابلیت DDP(Distributed Data Parallel) صورت‌گرفته‌است. همچنین ابزارهای دیگری نیز جهت پیش‌پردازش داده و پیاده‌سازی مدل‌های مبتنی بر یادگیری عمیق در این تحقیق به‌کاررفته‌اند که فهرست آن در جدول ۳ آورده‌شده‌است.

استفاده از مدل‌های زبانی جهت تبدیل حروف به ارقام جهت استفاده در مراکز تماس: در فاز انتهایی این تحقیق، به ایجاد زیرساختی جهت ارائه خروجی نهایی به‌صورت ارقام موردنظر جهت

استفاده در سامانه‌های مراکز تماس اطلاع‌رسانی فرودگاهی خواهیم-پرداخت؛ به‌گونه‌ای که مخاطبین پس از تماس تلفنی با مرکز تماس سامانه اطلاع‌رسانی فرودگاه‌های کشور و بیان شماره پرواز مورد نظر، از آخرین وضعیت آن پرواز مطلع خواهندشد. بنابراین با توجه به نیاز مبرم به داشتن ارقام درست جهت ارسال به سامانه‌های ثالث به‌منظور دریافت آخرین وضعیت هر پرواز، این فاز به‌عنوان یکی از اصلی‌ترین بخش‌های سامانه هوشمند ارائه اطلاعات پرواز گویا خواهدبود.

جدول ۳: ابزارهای استفاده‌شده در ParsAirCall

نسخه	توضیحات	نام ابزار
22.04.2 LTS	سیستم عامل	**Linux Ubuntu**
3.10.12	زبان برنامه‌نویسی	**Python**
1.5.3	کار با انواع داده‌ها	**Pandas**
1.2.2	پیش‌پردازش داده‌ها و یادگیری ماشین	**Scikit-learn**
3.7.1	بصری‌سازی	**Matplotlib**
1.4.1	کار با داده صوتی	**Sox**
0.2.0	کار با داده صوتی	**ffmpeg**
0.16.2	نمایش خروجی و لاگ‌ها	**Wandb**
0.9.2	کار با داده‌های صوتی	**Librosa**
2.1.0	یادگیری عمیق	**PyTorch**
2.1.3	انعطاف بالاتر کار با پایتورچ	**PyTorch Lightning**
12.1	ابزار جهت کار با GPU	**CUDA**

۵. ارزیابی روش پیشنهادی و مقایسه با سایر مدل‌ها

نرخ خطای کلمه (WER) یک معیار رایج برای ارزیابی کارایی سیستم‌های شناسایی گفتار خودکار است. این معیار تعداد جایگذاری‌ها، اضافه‌ها، و حذف‌ها را می‌سنجد تا خروجی شناسایی‌شده را با متن مرجع هماهنگ‌کند.

$$WER = \frac{S + D + I}{N} \qquad (1)$$

به این صورت که در مقایسه متن خروجی با داده‌های واقعی، 'S' تعداد کل جایگزینی‌های انجام‌شده و 'D' تعداد حذف‌های انجام‌شده و 'I' تعداد درج‌های انجام‌شده را نشان‌می‌دهند. تعداد تجمعی اصطلاحات در داده‌های واقعی نیز با 'N' مشخص‌می‌شود. ازآنجاکه WER مستقل از زبان است، از آن برای ارزیابی کارایی ASR در هر زبانی، از جمله فارسی، می‌توان‌استفاده‌کرد.

جدول ۴: ارزیابی مدل‌های نهایی

نام مدل	تعداد پارامتر (میلیون)	زمان خروجی (ثانیه)	سایز مدل(MB)	WER (%)
ParsAirCall	**۱۹**	**۱**	**۶۷**	**۲,۷**
Whisper - Tiny	۳۹	۲,۲	۷۲	۴,۴
Whisper - Medium	۷۶۹	۲,۷	۱۴۲۰	۶,۱
Wav2vec 2.0	۹۱	۲	۳۶۳	۹,۵
DeepSpeech 2	۵۲	۲	۴۹۵	۱۵,۹

طبق جدول ۴ ، روش پیشنهادی پس از پیاده‌سازی یادگیری انتقالی و Fine-Tuning از سایر روش‌های موجود، بهینه‌تر عمل‌نموده و با درنظر گرفتن مدت زمان لازم جهت ارائه خروجی، برای استفاده در مرکز تماس عملیاتی فرودگاه‌های کشور پیشنهادمی‌گردد. همچنین لازم به ذکر است که میزان منابع مصرفی توسط مدل پیشنهادی، با توجه به سایز کمتر مدل و تعداد پارامترهای به‌مراتب کمتر، قابلیت اجرا در محیط‌های واقعی را داراست به‌گونه‌ای که در هسته تلفنی مرکز تماس تعبیه‌گردد. برای زبان فارسی، حیاتی است که مجموعه داده به‌خوبی مشخص‌شده‌ای وجودداشته‌باشد که این عوامل را درنظر بگیرد. سیستم تشخیص گفتار فارسی تولیدشده، باید برای شناسایی انواع لهجه‌ها، گویش‌ها، و واژگان محاوره‌ای خاص در زبان فارسی آموزش داده‌شود تا WER کمتری به-دست‌آید. ارزیابی باید به‌طور ایده‌آل با استفاده از چندین مجموعه آزمون انجام‌شود که شامل یک محدوده گسترده از گویندگان، گویش‌ها، و شرایط گفتاری است تا درک جامعی از کارایی سیستم به‌دست‌آید.

۶. نتیجه‌گیری و جمع‌بندی

در این مقاله، ابزار ParsAirCall به‌عنوان یک ابزار تشخیص هوشمند اعداد فارسی در سامانه‌های فرودگاهی ارائه‌شده‌است که با استفاده از یادگیری انتقالی عمیق ساخته‌شده‌است. ParsAirCall با توجه به کمترین تعداد پارامترها و نیاز به منابع محاسباتی کمتر و نیز دقت بالاتر از مدل‌های رقیب، از مدل‌های مناسب برای محیط‌های عملیاتی است که می‌تواند به‌راحتی به‌عنوان یک سرویس در سیستم‌های تلفنی فرودگاهی فارسی زبان ادغام‌شود. ParsAirCall به‌دلیل Fine-Tuning بر روی ۳۰ ساعت داده تلفنی جمع‌آوری‌شده و نیز داده‌های پروژهٔ Common Voice پس از نرمال‌سازی آن به‌مراتب به دقت بالاتری بر روی داده‌های واقعی در فاز آزمایشی رسیده‌است. همچنین بهره‌مندی از ایده یادگیری انتقالی و استفاده از مدل Pre-Trained مناسب، دلیل دیگر برای دستیابی به مدلی بهینه است. از ParsAirCall می‌توان به‌عنوان ابزاری برای تشخیص اعداد فارسی در مراکز تماس و سیستم‌های تلفنی فرودگاهی استفاده‌کرد. به‌منظور توسعه پژوهش‌های صورت‌گرفته در حوزه استفاده از هوش مصنوعی و یادگیری

ماشین در راهکارهای ارتباطی به‌ویژه مراکز تماس و مراکز تلفن بومی موارد به شرح ذیل پیشنهادمی‌گردد:

- بررسی تأثیر Data Augmentation داده‌های تلفنی در خروجی نهایی
- پیاده‌سازی مدل های مبتنی بر یادگیری عمیق بهینه‌تر به‌گونه‌ای که قابلیت اجرا در کلاسترهای مختلف داشته‌باشد.
- جمع‌آوری مجموعه دادگان کامل‌تر در فضاهای نوی‌زی و شهری با لهجه‌ها و گویش‌های کامل‌تر

References

1. Deshpande, G., A. Batliner, and B.W. Schuller, *AI-Based human audio processing for COVID-19: A comprehensive overview*. Pattern recognition, 2022. **122**: p. 108289.
2. Agarwal, P., S. Swami, and S.K. Malhotra, *Artificial intelligence adoption in the post COVID-19 new-normal and role of smart technologies in transforming business: a review*. Journal of Science and Technology Policy Management, 2022.
3. Mitreska, M., et al., *Representation Learning for Automatic Speech Recognition: A Review of Speech-to-Text Methods.* 2023.
4. Young, S.J. and S. Young, *The HTK hidden Markov model toolkit: Design and philosophy.* 1993.
5. Veisi, H. and A. Haji Mani, *Persian speech recognition using deep learning.* International Journal of Speech Technology, 2020. **23**: p. 893-905.
6. Shafieian, M., *Hidden Markov model and Persian speech recognition.* International Journal of Nonlinear Analysis and Applications, 2023. **14**(1): p. 3111-3119.
7. Hinton, G., et al., *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.* IEEE Signal processing magazine, 2012. **29**(6): p. 82-97.
8. Farahani, M., et al., *Parsbert: Transformer-based model for persian language understanding.* Neural Processing Letters, 2021. **53**: p. 3831-3847.
9. Gonbadi, L. and N. Ranjbar, *Sentiment Analysis of People's opinion about Iranian National Cars with BERT*.
10. Farhadi, A., et al., *Unsupervised Domain Adaptation for image classification based on Deep Neural Networks.* Intelligent Multimedia Processing and Communication Systems (IMPCS), 2023. **4**(1): p. 27-37.
11. Salazar, J., K. Kirchhoff, and Z. Huang. *Self-attention networks for connectionist temporal classification in speech recognition*. in *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. 2019. IEEE.
12. Perero-Codosero, J.M., et al. *Exploring Open-Source Deep Learning ASR for Speech-to-Text TV program transcription*. in *IberSPEECH*. 2018.
13. Hannun, A., et al., *Deep speech: Scaling up end-to-end speech recognition.* arXiv preprint arXiv:1412.5567, 2014.
14. Amodei, D., et al. *Deep speech 2: End-to-end speech recognition in english and mandarin*. in *International conference on machine learning*. 2016. PMLR.
15. Hsiao, R., et al., *Online Automatic Speech Recognition With Listen, Attend and Spell Model.* IEEE Signal Processing Letters, 2020. **27**: p. 1889-1893.
16. Collobert, R., C. Puhrsch, and G. Synnaeve, *Wav2letter: an end-to-end convnet-based speech recognition system.* arXiv preprint arXiv:1609.03193, 2016.
17. Baevski, A., et al., *wav2vec 2.0: A framework for self-supervised learning of speech representations.* arXiv preprint arXiv:2006.11477, 2020.
18. Peng, C.-J., et al., *Attention-based multi-task learning for speech-enhancement and speaker-identification in multi-speaker dialogue scenario.* arXiv preprint arXiv:2101.02550, 2021.
19. Radford, A., et al. *Robust speech recognition via large-scale weak supervision*. in *International Conference on Machine Learning*. 2023. PMLR.
20. Ardila, R., et al., *Common voice: A massively-multilingual speech corpus.* arXiv preprint arXiv:1912.06670, 2019.
21. Kriman, S., et al. *Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions*. in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020. IEEE.