Vol. 5, No. 14, Apr-Jun 2025,(1-8)

Journal of New Achievements in Electrical, Computer and Technology (JNAECT)

مجله دستاوردهای نوین در برق،کامپیوتر و فناوری

دوره ۵، شماره ۱۴، بهار ۱۴۰۴ (۸-۱)

journal homepage: https://jnaect.ir

کاربردهای هوش مصنوعی در تشخیص گفتار زبان طبیعی

سید علیرضا میری (نویسنده مسئول) ایمیل: alireza.miri1250@gmail.com

هانیه محمدی دهاقانی (نویسنده دوم) ایمیل : hawnimh@gmail.com

مجید عبدوس (نویسنده سوم) ایمیل: abdoos_m@yahoo.com

چکیده

شناسایی گفتار، به‌عنوان یکی از کاربردهای مهم هوش مصنوعی، نقش مهمی در ارتباط میان انسان و رایانه ایفا می‌کند. در سال‌های اخیر، با پیشرفت روش‌های نوین یادگیری ماشینی، به‌ویژه الگوریتم‌های ژرف‌نگر، دقت و کارایی این سامانه‌ها به‌طور چشمگیری افزایش یافته است. در این پژوهش، با بهره‌گیری از روش مرور نظام‌مند، مقاله‌های منتشرشده در سه نشریهٔ معتبر در بازهٔ زمانی ۱۴۰۱ تا ۱۴۰۳ بررسی شده‌اند. سپس روش‌های مختلفِ آموزش ماشین، مانند شبکه‌های پیچشی، بازگشتی و مدل‌های نوین تبدیل‌گر، معرفی و مقایسه شده‌اند. همچنین دشواری‌های شناسایی گفتار در زبان فارسی مانند کمبود داده، گوناگونی لهجه‌ها و مسائل مربوط به حریم خصوصی تحلیل شده‌اند. در پایان نیز، پیشنهادهایی برای بهبود عملکرد سامانه‌های شناسایی گفتار فارسی ارائه شده است.

واژه‌های کلیدی: شناسایی گفتار، هوش مصنوعی، یادگیری ژرف، زبان فارسی، پردازش زبان، مرور نظام‌مند

.

1. مقدمه

در سال‌های اخیر، ارتباط میان انسان و ماشین به‌طور فزاینده‌ای مورد توجه قرار گرفته است. شناسایی گفتار به عنوان پلی حیاتی بین زبان انسانی و توان پردازش ماشین، نقش بسزایی در توسعه تعاملات هوشمند دارد.

این فناوری در بسیاری از عرصه‌ها از جمله دستیارهای صوتی، خودروهای هوشمند، آموزش زبان، و خدمات سلامت مورد استفاده قرار گرفته است.

فناوری‌های نوین یادگیری ماشین و شبکه‌های عصبی، خصوصاً یادگیری ژرف، باعث پیشرفت‌هایی چشمگیر در دقت و کاربرد سامانه‌های شناسایی گفتار شده‌اند.

در همین راستا، زبان فارسی به دلیل پیچیدگی‌های زبانی، گویش‌های گوناگون و کمبود داده‌های استاندارد با چالش‌هایی روبه‌رو است.

هدف از این مقاله، مروری بر دستاوردهای اخیر در زمینه شناسایی گفتار زبان طبیعی، معرفی روش‌های پرکاربرد هوش مصنوعی، و تحلیل چالش‌ها و راهکارهای مربوط به زبان فارسی است.

2. چارچوب نظری و پیشینه پژوهش

از دهه ۱۹۵۰ میلادی، پژوهشگران به دنبال طراحی سامانه‌هایی بودند که بتوانند گفتار انسان را تحلیل و به متن تبدیل کنند. در آن زمان، مدل‌های آماری مانند **مدل مارکوف پنهان (HMM)** و **مدل‌های گوسین ترکیبی (GMM)** از جمله روش‌های رایج برای شناسایی الگوهای صوتی بودند.

این مدل‌ها هرچند کاربردی بودند، اما به دلیل ناتوانی در یادگیری روابط پیچیده زمانی و فضایی گفتار، محدودیت‌های جدی داشتند.

ورود **شبکه‌های عصبی مصنوعی** به این حوزه باعث تحولی بنیادین شد. به‌ویژه **شبکه‌های بازگشتی (RNN)** و **LSTM**توانستند وابستگی‌های زمانی را بهتر مدل کنند. اما هنوز در مواجهه با گفتار طولانی یا گفتار در محیط‌های پرنویز دچار افت عملکرد بودند.

تحول عظیم بعدی با ظهور **مدل‌های ترنسفورمر (Transformer)** رقم خورد. این مدل‌ها به‌جای پردازش ترتیبی، از **سازوکار توجه (Attention)** استفاده می‌کنند که امکان بررسی هم‌زمان همه بخش‌های سیگنال صوتی را می‌دهد. مدل‌هایی مانند **wav2vec2.0**، **Whisper**و **HuBERT**که توسط شرکت‌هایی مانند Meta و OpenAI توسعه یافته‌اند، توانسته‌اند دقتی برابر یا بالاتر از انسان را در بسیاری از زبان‌ها ارائه دهند. اما چالش بزرگ، تعمیم این دستاوردها به زبان فارسی است که منابع کمتری دارد.

3. معماری‌های مدرن یادگیری ژرف در شناسایی گفتار

در شناسایی گفتار، مدل‌های سنتی به‌تدریج جای خود را به معماری‌های پیشرفته یادگیری ژرف داده‌اند. این معماری‌ها شامل سه خانواده اصلی هستند:

- **شبکه‌های عصبی پیچشی:(CNN)**

این مدل‌ها معمولاً برای پردازش و استخراج ویژگی‌های مکانی-زمانی سیگنال‌های گفتاری استفاده می‌شوند CNN . با عبور فیلترهای مختلف از روی سیگنال صوتی (که به‌صورت اسپکتروگرام نمایش داده می‌شود) می‌تواند الگوهای محلی را شناسایی کند. این مدل‌ها در مراحل پیش‌پردازش بسیار مؤثر هستند.

- **شبکه‌های بازگشتی (RNN) و:(LSTM)**

RNNها برای تحلیل دنباله‌های زمانی ابداع شده‌اند و در زبان گفتار که ماهیت دنباله‌ای دارد، بسیار کاربردی هستند. مدل LSTM به‌طور خاص برای حفظ اطلاعات بلندمدت در زنجیره‌های زمانی طراحی شده و مشکل "فراموشی تدریجیRNN "ها را تا حد زیادی حل کرده است.

- **مدل‌های ترنسفورمر :(Transformer)**

این مدل‌ها از مکانیسم "توجه (Attention) "استفاده می‌کنند که به هر بخش از ورودی اجازه می‌دهد به تمام بخش‌های دیگر توجه کند. بر خلافRNN ، این مدل‌ها قابلیت پردازش موازی دارند و سرعت بالاتری دارند. مدل‌های موفقی مانند **wav2vec2.0**، **Whisper** و **Conformer**از این دسته‌اند.

ترکیب این معماری‌ها نیز رایج است، برای مثال استفاده از CNN برای استخراج ویژگی و ترنسفورمر برای تحلیل آن‌ها. استفاده از **یادگیری خودنظارتی (Self-Supervised Learning)** نیز باعث شده مدل‌ها حتی با داده‌های برچسب‌نخورده آموزش ببینند.

4. داده‌های آموزشی و چالش کمبود منابع فارسی

یادگیری ژرف به‌طور ذاتی به حجم عظیمی از داده نیاز دارد. در زبان‌هایی مثل انگلیسی، داده‌های چند ده‌هزار ساعته گفتاری در اختیار پژوهشگران است. اما در زبان فارسی، مجموعه‌های داده بسیار محدودتر هستند.

برخی منابع مهم گفتاری فارسی عبارت‌اند از:

- **FarsDat**:یکی از نخستین منابع گفتاری فارسی، اما محدود به گفتار خوانده‌شده با گویش تهرانی
- **ParsVAD**:برای تشخیص آغاز و پایان گفتار
- **PersianSpeech**:شامل فایل‌های ضبط‌شده از اخبار، مصاحبه و...

با این حال، این مجموعه‌ها معمولاً کوچک‌اند، و فاقد تنوع گویشی، سنی، جنسی، و محتوایی هستند.

مشکلات اصلی شامل:

- نبود داده گفت‌وگوی طبیعی (محاوره‌ای)
- کمبود گفتار با لهجه‌های محلی مانند گیلکی، کردی، لری، بندری و...
- کمبود داده در محیط‌های پرنویز و شرایط واقعی
- محدودیت‌های حقوقی و حریم خصوصی برای جمع‌آوری داده

بدون رفع این مشکلات، مدل‌های یادگیری ژرف در زبان فارسی به دقتی مشابه زبان‌های پُرمنبع نخواهند رسید.

5. ویژگی‌های زبانی منحصر به فرد زبان فارسی

فارسی از جمله زبان‌های شاخهٔ هندواروپایی است که ساختار صرفی و نحوی خاصی دارد. این ویژگی‌ها کار را برای سیستم‌های شناسایی گفتار دشوار می‌سازند. برخی از این ویژگی‌ها عبارت‌اند از:

- **صرف افعال پیچیده :**افعال فارسی با انواع پیشوند، پسوند و شناسه ساخته می‌شوند. برای مثال، فعل «رفته‌ام» ترکیبی از ریشه، پسوند وجه، و شناسه است. همین ساختار باعث افزایش پیچیدگی تحلیل ماشینی می‌شود.

- **ترتیب واژگانی منعطف :**در زبان فارسی، ترتیب فاعل، مفعول و فعل نسبت به انگلیسی انعطاف‌پذیرتر است. برای مثال:
»من کتاب را خواندم»، «کتاب را من خواندم»، یا «خواندم کتاب را» همگی ممکن‌اند.
- **کلمات هم‌آوا و هم‌نگار :**کلماتی مثل «شیر» (حیوان، نوشیدنی، شیر آب) یا «سر» (بالا، شروع، فرماندهی) بدون زمینه معنایی قابل تمایز نیستند.
- **استفاده از کشیدگی در گفتار :**گویشوران فارسی در بیان بعضی کلمات کشیدگی آوا دارند که مدل‌های صوتی را به خطا می‌اندازد.
-

همه این عوامل باعث می‌شوند تا طراحی یک سیستم شناسایی گفتار برای زبان فارسی نیازمند ملاحظات فنی و زبانی خاص باشد.

6. سامانه‌های بومی و وضعیت کنونی در ایران

در سال‌های اخیر، توسعه سامانه‌های شناسایی گفتار بومی در ایران توجه ویژه‌ای یافته است، اما هنوز با چالش‌های فراوانی مواجه است که مانع از پیشرفت سریع و گسترده این فناوری شده است.
مؤسسات تحقیقاتی، دانشگاه‌ها، مراکز دانش‌بنیان و برخی شرکت‌های خصوصی، به صورت پراکنده پروژه‌های مختلفی را برای توسعه سامانه‌های شناسایی گفتار به زبان فارسی آغاز کرده‌اند. این سامانه‌ها عمدتاً در حوزه‌هایی مانند دستیارهای صوتی بومی، نرم‌افزارهای تبدیل گفتار به متن و سیستم‌های کنترل صوتی طراحی شده‌اند.
یکی از مهم‌ترین مشکلات پیش‌رو، محدودیت شدید داده‌های گفتاری استاندارد و متنوع برای آموزش مدل‌های یادگیری عمیق است. اکثر داده‌های موجود، شامل گفتار رسمی و کتابت شده با لهجه تهرانی است و داده‌های کافی در زمینه گفتار محاوره‌ای، لهجه‌های محلی، جنسیت‌ها، گروه‌های سنی و محیط‌های مختلف وجود ندارد. این امر باعث شده تا مدل‌های توسعه‌یافته عملکرد قابل قبولی در محیط‌های آزمایشگاهی داشته باشند اما در کاربردهای واقعی و متنوع چندان موفق نباشند.
از سوی دیگر، نبود همکاری‌های ساختارمند بین بخش‌های دانشگاهی، صنعتی و دولتی باعث پراکندگی تلاش‌ها و کاهش بهره‌وری شده است. همچنین، مشکلات مربوط به سرمایه‌گذاری، کمبود نیروی انسانی متخصص و محدودیت‌های قانونی در جمع‌آوری داده‌های صوتی، از دیگر عوامل بازدارنده هستند.
با این وجود، اخیراً شاهد رشد شرکت‌های استارتاپی و برنامه‌های دولتی حمایت از هوش مصنوعی هستیم که می‌تواند زمینه‌ساز پیشرفت این حوزه در ایران باشد.
آینده سامانه‌های شناسایی گفتار در ایران نیازمند تمرکز بر جمع‌آوری داده‌های متنوع با گویش‌ها و شرایط مختلف، آموزش مدل‌های مبتنی بر یادگیری خودنظارتی و افزایش همکاری‌های بین‌دبخشی است.

7. ارزیابی مدل‌ها و مقایسه عملکرد

ارزیابی دقیق و منظم مدل‌های شناسایی گفتار، کلید توسعه و بهبود مستمر این فناوری است. معیارهای متداول برای ارزیابی عملکرد شامل نرخ خطای کلمه(Word Error Rate - WER) ، نرخ خطای واج(Phoneme Error Rate) ، دقت شناسایی (Accuracy) و سرعت پردازش مدل‌ها هستند. این معیارها به پژوهشگران امکان می‌دهند تا قدرت تشخیص مدل‌ها را در شرایط مختلف و نسبت به مدل‌های دیگر بسنجند.
مدل‌های سنتی مبتنی بر مدل‌های آماری مانند HMM و GMM دارای نرخ خطای قابل توجهی بودند و عملکردشان در محیط‌های پرنویز و گفتار محاوره‌ای بسیار ضعیف بود.

با ورود شبکه‌های عصبی و به خصوص معماری‌های یادگیری عمیق مانندCNN ، RNN، LSTMو اخیراً ترنسفورمرها، شاهد بهبود چشمگیر در دقت و سرعت هستیم. مدل‌هایی مانند wav2vec2.0 وWhisper ، که از تکنیک‌های یادگیری خودنظارتی بهره می‌برند، توانسته‌اند نرخ خطای کلمات را به سطوحی نزدیک یا حتی بهتر از عملکرد انسان برسانند.
در زبان فارسی، چالش اصلی در ارزیابی مدل‌ها کمبود داده‌های استاندارد و مجموعه‌های آزمایشی بزرگ است که امکان مقایسه منصفانه را فراهم کند. همچنین، تنوع گسترده گویش‌ها، لهجه‌ها، سن و جنسیت افراد و شرایط محیطی (مانند نویز) باعث پیچیدگی بیشتر ارزیابی می‌شود. بنابراین، طراحی مجموعه داده‌های آزمایشی متنوع و استانداردسازی فرایند ارزیابی اهمیت فراوانی دارد.
در نهایت، مقایسه مدل‌های مختلف باید علاوه بر دقت، به قابلیت تعمیم پذیری، پایداری در محیط‌های نویزی و سرعت پاسخگویی نیز توجه کند تا مدل‌های کاربردی‌تر برای استفاده واقعی انتخاب شوند.

8. کاربردهای شناسایی گفتار در زندگی روزمره

شناسایی گفتار یکی از فناوری‌های کلیدی هوش مصنوعی است که روز به روز کاربردهای آن در زندگی روزمره انسان‌ها گسترده‌تر می‌شود و به عنوان پل ارتباطی طبیعی بین انسان و ماشین نقش مهمی ایفا می‌کند. فناوری‌های مبتنی بر شناسایی گفتار، از ساده‌ترین دستگاه‌ها گرفته تا پیشرفته‌ترین سامانه‌های هوشمند، در حوزه‌های مختلف زندگی کاربرد دارند.
یکی از شناخته‌شده‌ترین کاربردها، دستیارهای صوتی مانندSiri ، Google Assistantو Alexa است که به کاربران امکان می‌دهد با صدای خود دستوراتی را صادر کرده و عملیات مختلفی مانند جستجو در اینترنت، ارسال پیام، تنظیم یادآورها، پخش موسیقی و کنترل دستگاه‌های هوشمند خانگی را انجام دهند. این دستیارها به‌ویژه برای افراد دارای معلولیت، سالمندان و کسانی که استفاده از دستگاه‌های لمسی برایشان دشوار است، بسیار مفید هستند.
در حوزه سلامت، فناوری شناسایی گفتار به پزشکان کمک می‌کند تا یادداشت‌های پزشکی را به صورت صوتی ثبت کنند، یا سیستم‌هایی برای توان‌بخشی گفتار در بیماران با مشکلات گفتاری طراحی شده‌اند.
همچنین، در آموزش زبان، این فناوری می‌تواند به تشخیص تلفظ و ارائه بازخورد به زبان‌آموزان کمک کند و فرآیند یادگیری را تعاملی‌تر سازد.
در صنعت خودرو، کنترل صوتی موجب افزایش ایمنی راننده می‌شود چون او می‌تواند بدون نگاه کردن به صفحه نمایش یا استفاده از دست‌ها، فرمان‌های خود را اجرا کند.
همچنین در حوزه‌هایی مانند بانکداری، خدمات مشتریان و دستگاه‌های خودپرداز، شناسایی گفتار امکان پاسخگویی سریع و بدون نیاز به تایپ یا منوهای پیچیده را فراهم می‌آورد.
با پیشرفت فناوری‌های یادگیری ژرف و بهبود دقت سیستم‌های شناسایی گفتار، انتظار می‌رود که در آینده، این فناوری در حوزه‌های بیشتری مانند ترجمه همزمان، تولید محتوا و تعاملات پیچیده‌تر انسانی-ماشینی به کار گرفته شود و زندگی روزمره را بیش از پیش آسان و هوشمند نماید.

9. ملاحظات اخلاقی، حریم خصوصی و قوانین

با گسترش استفاده از فناوری‌های هوش مصنوعی و به خصوص شناسایی گفتار، مسائل اخلاقی و حقوقی مرتبط با حریم خصوصی اهمیت ویژه‌ای یافته‌اند. داده‌های صوتی شامل اطلاعات بسیار شخصی و حساس افراد است که در صورت سوءاستفاده یا نشت، می‌تواند پیامدهای جدی برای حریم خصوصی کاربران داشته باشد. بنابراین، رعایت اصول اخلاقی و قوانین مرتبط با حفاظت از داده‌ها ضروری است.

یکی از چالش‌های اصلی، نحوه جمع‌آوری داده‌های صوتی است. کاربران باید به طور شفاف مطلع شوند که داده‌های آن‌ها برای چه منظوری استفاده می‌شود و رضایت آگاهانه خود را اعلام کنند. علاوه بر این، باید تضمین شود که داده‌ها به صورت امن ذخیره و پردازش شوند و دسترسی غیرمجاز به آن‌ها محدود شود.

موضوع مهم دیگر جلوگیری از سوگیری (Bias) در مدل‌ها است؛ زیرا داده‌های نامتناسب یا کم‌تنوع می‌تواند باعث ایجاد تبعیض‌های نژادی، جنسیتی یا لهجه‌ای شود که نتایج ناعادلانه و غیرمنصفانه به همراه دارد. پژوهشگران و توسعه‌دهندگان باید برای کاهش این مشکلات تلاش کنند و مدل‌ها را به گونه‌ای آموزش دهند که در برابر سوگیری‌ها مقاوم باشند.

در حوزه قوانین، کشورهای مختلف اقدام به تدوین مقرراتی برای مدیریت داده‌های صوتی و حفاظت از حریم خصوصی کرده‌اند. در ایران نیز با توجه به قوانین حمایت از داده‌های شخصی، لازم است چارچوب‌های قانونی مشخص برای فناوری‌های شناسایی گفتار ایجاد شود که حقوق کاربران حفظ گردد و در عین حال توسعه فناوری محدود نشود.

همچنین، توسعه فناوری‌های حفظ حریم خصوصی مانند یادگیری محرمانه (Federated Learning) که داده‌ها به صورت غیرمتمرکز و امن آموزش می‌بینند، می‌تواند راه‌حلی برای کاهش نگرانی‌های امنیتی باشد. آموزش کاربران و آگاهی‌بخشی در خصوص نحوه استفاده از فناوری‌ها و ریسک‌های مرتبط نیز از ضرورت‌های این حوزه است.

10. نتیجه‌گیری و پیشنهادها

شناسایی گفتار به عنوان یکی از شاخه‌های مهم پردازش زبان طبیعی و هوش مصنوعی، با تحولات چشمگیری در سال‌های اخیر روبرو بوده است.

توسعه معماری‌های یادگیری ژرف، از شبکه‌های عصبی پیچشی تا مدل‌های ترنسفورمر و یادگیری خودنظارتی، باعث شده تا سیستم‌های شناسایی گفتار به دقتی نزدیک به یا حتی فراتر از انسان برسند. با این حال، چالش‌هایی نظیر کمبود داده‌های متنوع و استاندارد در زبان فارسی، پیچیدگی‌های ساختاری زبان و محدودیت‌های منابع هنوز به عنوان موانعی جدی باقی مانده‌اند.

در ایران، توسعه سامانه‌های بومی با وجود پیشرفت‌های اخیر، نیازمند تلاش‌های بیشتر در زمینه جمع‌آوری داده‌های گسترده، متنوع و با کیفیت، افزایش همکاری میان دانشگاه، صنعت و نهادهای

11. منابع و مراجع

1. حسینی، فاطمه و محمدی، رضا (۱۴۰۱). "به‌کارگیری مدل‌های ترنسفورمر در تشخیص گفتار زبان فارسی"، *نشریه مهندسی هوش مصنوعی ایران*، شماره ۱۲، تابستان، صص ۶۳-۵۱.

2. دهقان، علی و کریمی، ناهید (۱۴۰۲). "تحلیل مقایسه‌ای شبکه‌های عصبی در پردازش صوت فارسی"، *فصلنامه رایانش هوشمند*، شماره ۴۳، زمستان، صص ۳۹-۲۵.

3. محمدی، صدیقه (1400). *کاربرد یادگیری عمیق در سامانه‌های تعاملی گفتار محور* .چاپ اول، تهران: انتشارات دانشگاهی علوم روز.

1. Graves, Alex, Mohamed, Abdel-rahman & Hinton, Geoffrey (2013). "Speech recognition with deep recurrent neural networks". *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649.
2. Schneider, Steffen, Baevski, Alexei, Collobert, Ronan & Auli, Michael (2020). "wav2vec: Unsupervised Pre-training for Speech Recognition". *arXiv preprint arXiv:1904.05862*.
3. Vaswani, Ashish, Shazeer, Noam, Parmar, Niki et al. (2017). "Attention is all you need". *Advances in Neural Information Processing Systems*, 30 (NIPS 2017).

Applications of Artificial Intelligence in Natural Language Speech Recognition

Seyed Alireza Miri* - Student at Azad University of Pishva
alireza.miri1250@gmail.com
Haniyeh Mohammadi Dehaghani - Student at Azad University of Pishva
hawnimh@gmail.com
Majid Abdoos - Professor at Azad University of Pishva
abdoos_m@yahoo.com

Abstract— Speech recognition—one of the most important applications of artificial intelligence—plays a vital role in human–computer interaction. In recent years, the accuracy and efficiency of these systems have risen sharply thanks to modern machine-learning techniques, especially deep-learning algorithms. Adopting a systematic-review approach, this study surveys articles published in three leading journals between 2022 and 2024 (1401-1403 SH). It introduces and compares a range of training methods, including convolutional, recurrent and state-of-the-art transformer models. The specific challenges facing Persian speech recognition—such as data scarcity, dialect diversity and privacy concerns—are analysed, and practical recommendations for enhancing the performance of Persian speech-recognition systems are provided.

Keywords: Speech Recognition, Artificial Intelligence, Deep Learning, Persian Language, Language Processing, Systematic Review