نوع مقاله: پژوهشی

ارائه روشی مبتنی بر یادگیری عمیق جهت واژه‌یابی در گفتار فارسی

حسین سهلانی[۱]، استادیار

۱- گروه فتا، دانشکده اطلاعات و آگاهی، دانشگاه جامع علوم انتظامی امین –sahlani_h@yahoo.com

چکیده: با توسعه فناوری، تحلیل بر روی گفتار انسان توسط ماشین‌های هوشمند، رشد زیادی پیداکرده است، یکی از پیشنیازها، واژه‌یابی در گفتار اســـت. ســـیســـتم های واژه‌یاب وجود و یا عدم وجود کلمات کلیدی را در گفتار بررســـی می کند ولی به دلیل ناکافی بودن مجموعه داده، منابع پرداز شی، وجود نویزها و ... توسعه یک سیستم واژه‌یاب ایده آل با دقت بالا د شوار است. در این پژوهش با ضبط ۴۲ ساعت از گفتار افراد مختلف به‌صورت اختصاصی در کنار سایر دادگان موجود در این حوزه و با اتکا بر روی الگوریتم‌های شبکه عصبی عمیق طراحی، آموزش و ارزیابی گردید که می‌تواند دقت واژه‌یابی و تولید یک متن متناظر با گفتار را افزایش دهد، معماری و روش پیشنهادی در دو قسمت تشکیل شده قسمت اول، طوری طراحی شده که هدفش تولید واج متناظر با گفتار، بدون دا شتن اطلاعات مدل زبانی ا ست ولی قسمت دوم با لایه‌های متعدد و دا شتن مدل زبانی سعی در تولید یک متن صحیح با توجه به دامنه زبان فارسی دارد. روش پیشنهادی ترکیبی ارائه‌شده با دقت ۸۸،۰۱٪ می‌تواند کلیدواژه‌های موجود در گفتار را تشخیص و با دقت ۹۹،۸۰٪ عدم وجود واژه‌ها را مشخص نماید و دقت بالاتری نسبت به کارهای مشابه داشته باشد.

واژه‌های کلیدی: تبدیل گفتار به متن، واژه‌یابی، دادگان گفتار فارسی، پردازش گفتار، شبکه‌های عصبی عمیق.

A deep learning method for voice word spotting in Persian language

Sahlani, Hossein[1], assistant professor

1. FATA Department, Faculty of Information and Awareness, Amin University of Law Enforcement Sciences, Tehran, Iran

Abstract: With the development of technologies related to audio data recording and transmission, as well as the advancement of artificial intelligence science, the analysis of human speech by intelligent machines has grown greatly. One of the most important technologies in speech processing in the last decade has been word search in audio. By receiving keywords from the user, the word search system can check the presence or absence of that word in the audio file and report the result to the user. Due to the insufficient data set, it is difficult to develop an ideal software that can find all the user's words exist in the audio. In this research, by collecting 42 hours of Persian audio data along with other data available in this field and relying on deep neural network algorithms were designed, trained and evaluated. These two architectures are complementary and can increase the accuracy of word search. First architecture is designed in such a way that its goal is to produce the phoneme corresponding to the sound, without having the information of the language model, but the second architecture with multiple layers and having a language model tries to produce a correct text according to the domain of the Persian language. The presented combined proposed method can detect the keywords in the audio with 88.01% accuracy and determine the absence of words with 99.80% accuracy and has higher accuracy than similar methods.

Keywords: Audio to text conversion, audio word search, Persian audio corpus, speech processing, deep neural networks.

تاریخ ارسال مقاله: ۱۴۰۱/۱۲/۰۴

تاریخ اصلاح مقاله: ۱۴۰۲/۰۶/۱۹

تاریخ پذیرش مقاله: ۱۴۰۲/۰۷/۱۰

نام نویسنده مسئول: حسین سهلانی

نشانی نویسنده مسئول: ایران - تهران – گروه فتا، دانشکده اطلاعات و آگاهی، دانشگاه جامع علوم انتظامی امین

Digital Object Identifier 10.22034/JASP.

۱- مقدمه

تشخیص گفتار به‌سرعت در حال تبدیل‌شدن به یک فناوری پرکاربرد و مؤثر است. فناوری‌های تشخیص گفتار در حال تغییر نحوه تعامل افراد با دستگاه‌ها، خانه‌ها، اتومبیل‌ها و مشاغل خود هستند و کاربرد آن در صنایعی مانند تجارت، بانکداری، بازاریابی و مراقبت‌های بهداشتی به‌سرعت آشکار می‌شود. آینده تعامل انسان و سیستم، کار با سامانه‌های گفتار است و نه با صفحه‌کلید و ... . با پیشرفت فناوری، بحث کمک گرفتن از هوش مصنوعی برای تبدیل گفتار به نوشتار و واژه‌یابی به عنوان یکی از شاخه‌های مهم هوش مصنوعی مطرح شده است.

در این پژوهش با بهره‌گیری از هوش مصنوعی و یادگیری عمیق می بایست متن متناظر با گفتار تشخیص داده‌شده و سپس با جستجو در آن متن، حضور یا عدم حضور واژه‌ی مدنظر کاربر مشخص شود. با پیشرفت شبکه‌های عصبی، تغییرات زیادی در الگوریتم‌های هوش مصنوعی ایجاد شده به‌طوری‌که حوزه شناسایی گفتار درگذشته بیشتر با روش‌های مبتنی بر مدل مخفی مارکوف و با ابزارهایی مانند HTK صورت می‌گرفت [1]، که برای انجام این مهم نیازمند دادگان مبتنی بر برچسب برای هر یک از واج‌ها و استخراج ویژگی‌های متداول از آنها مانند MFCC از فایل‌های گفتاری بود که این خود زمان و نیروی انسانی مختص خود را می‌طلبد؛ اما در سال‌های اخیر این کار با استفاده از الگوریتم‌های یادگیری عمیق صورت می‌گیرد. این الگوریتم‌ها برای تبدیل گفتار به متن ابتدا ویژگی‌های مشخصی را از گفتار استخراج می‌کنند، سپس با کمک مدل آموزش داده‌شده این ویژگی‌ها را به دنباله‌ای از واج‌های احتمالی تبدیل می‌کنند و سپس به کمک مدل زبانی این واج‌ها را به رشته‌ای متنی تبدیل می‌کنند. این رویکرد که با عنوان انتها به انتها مطرح است، سیگنال خام گفتار و متن متناظر آن را لازم دارد و نیاز به دادگان با دنباله برچسب واجی نیست. به عبارتی شبکه‌هایی مانند RNN, CNN کار استخراج ویژگی را بر عهده دارند و ویژگی‌های استخراج‌شده را به یک طبقه بند می‌دهند تا عمل یادگیری را انجام دهد[2] .

پیاده‌سازی سامانه‌ای که بتواند با دقت بالایی شناسایی گفتار را انجام دهد، همواره یکی از چالش‌های این حوزه بوده است. هرچند با استفاده از روش‌های یادگیری عمیق در این بخش، دقت‌های قابل‌قبولی گزارش‌شده است [3]، اما به دلیل این‌که استفاده از روش‌های یادگیری عمیق نیازمند دادگان حجیم است، برای زبان‌های با داده کم این روش‌ها کارا نیستند. برای زبان فارسی دادگان فارس‌دات وجود دارد که شامل دو نسخه کوچک و بزرگ است. اکثر مقالات این حوزه برای زبان فارسی از دادگان فارس‌دات استفاده کرده‌اند [5] [4]. در سال‌های اخیر شرکت موزیلا پروژه‌ای را راه‌اندازی کرد تا بتواند با آن مشکل کمبود داده برای سایر زبان‌ها را تا حد زیادی حل کند. این پروژه که با کمک کاربران هر زبان شروع به جمع‌آوری داده‌ها کرده تا به امروز توانسته حجم زیادی داده گفتاری را تهیه نمایند.

هدف مقاله پیش رو، ارائه سامانه واژه‌یاب گفتار فارسی مبتنی بر یادگیری عمیق با دقت بالا می‌با شد که حین کار در کنار دادگان رایگان موزیلا، مجموعه داده گفتار فار سی منا سب جمع‌آوری و پالایش و ایجاد گردید که بعد از پیش‌پردازش‌های صورت گرفته روی ا صوات موجود در دادگان تهیه شده (ا ستخراج ویژگی، حذف نویز و ...) بهینه سازی و بومی سازی یک معماری شبکه عصبی عمیق (تعیین تعداد پالایه‌ها، تعداد لایه‌ها و نورون‌های هر لایه و...) آموزش و مدل‌سازی گردید و نهایتاً کارایی روش پیشنهادی با مدل‌های مختلف ارزیابی گردید که نتایج نشان دهنده بهبود عملکرد روش پیشنهادی با دقت واژه‌یابی ۸۸,۰۱ درصد برای تشخیص وجود کلیدواژه‌ها در گفتار و دقت ۹۹,۸۰ درصد برای تشخیص عدم وجود واژه‌ها در گفتار می‌باشد. و از دستاوردهای جانبی این پژوهش می‌توان به ساخت یک سامانه تبدیل گفتار به متن اشاره کرد که توانسته متن حاصل از گفتار را اصلاح و خروجی قابل‌قبولی ارائه دهد همچنین یکی از محصولات مهم این پژوهش که در آینده بیشتر قابل استفاده خواهد بود ساخت پایگاه داده اختصاصی می‌باشد که در بخش‌های بعدی به‌صورت کامل تشریح شده‌اند.

۲- پیشینه تحقیق

باز شنا سی گفتار، گ ستره و سیعی از کاربردها از قبیل فهم ارتباط میان ان سان و رایانه، شنود مکالمات تلفنی با اهداف امنیتی و ... را شامل می شود که در آن نیازی به بازتولید تمام عبارات و کلمات ادا شده توسط گوینده نمی‌با شد و تنها باز شنا سی دقیق تعدادی از کلمات مهم کفایت می‌کند که به این آ شکار سازی کلمات کلیدی در گفتار پیو سته ورودی، واژه‌یابی گفتار[1] اطلاق می‌شود [6].

رویکردهایی که برای آموزش سامانه‌های واژه‌یاب گفتار مورد استفاده قرار می‌گیرند را می‌توان به دو دسته تقسیم‌بندی کرد: رویکردهای مبتنی بر مدل مخفی مارکف و رویکردهای مبتنی بر طبقه‌بندی[7]

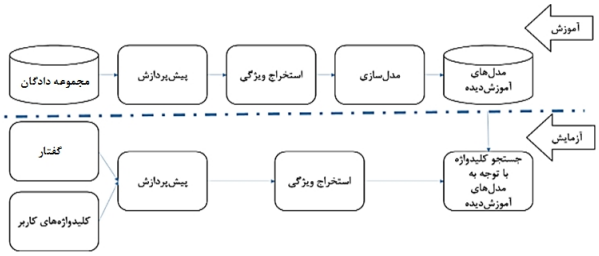
از مزایای روش‌های واژه‌یابی گفتار مبتنی بر مدل مخفی مارکف سرعت مناسب، استفاده از اطلاعات وابسته به محتوا می‌باشد، در مقابل، به دلیل در اختیار نبودن تخمین‌های دقیق در آموزش مدل، الگوریتم آموزش آن در بهینه‌های محلی گرفتار شده و دقت واژه‌یابی گفتار کاهش می‌یابد. رویکردهای واژه‌یابی گفتار مبتنی بر مدل مخفی مارکف به دو دسته تق سیم می شوند: واژه‌یابی گفتار مبتنی بر باز شنا سی گفتار پیو سته با دادگان و واژه‌یابی گفتار مبتنی بر جستجوی واجی. در رویکرد اول، ابتدا با ا ستفاده از یک باز شناس گفتار مبتنی بر دادگان، آر شیوهای گفتاری بزرگ به شبکه‌های واجی یا کلمه‌ای تبدیل می‌شوند. سپس با به‌کارگیری روش‌های جستجو در شبکه‌های واجی یا کلمه‌ای، کلمات ورودی مورد جستجو قرار می‌گیرند که نیاز به وجود حجم زیاد داده آموزشی برچسب زده‌شده و پیچیدگی محاسباتی بالا از چالش‌های اساسی آن محسوب می‌شود. در رویکرد دوم (جستجوی واجی)، زیر مدل‌های مستقل از محتوا یا وابسته به محتوا ساخته می‌شوند و از اتصال

[1] Keyword spotting

زیر مدل‌ها، دنباله واجی کلمه کلیدی حاصل می‌شود. ضعف اساسی این بالا بودن خطاهای درج، حذف و جایگزینی می‌باشد[8].

راهکار دیگر استفاده از مدل‌های مبتنی بر طبقه‌بندی می‌باشد که در این بین شبکه‌های عصبی، ویژه ماشین بردار پشتیبان، شبکه‌های عصبی عمیق از اقبال بیشتری برخوردار بوده‌اند [9]. پیچیدگی‌های محاسباتی و زمانی بالا، عدم امکان استفاده از اطلاعات وابسته به محتوا و عدم امکان استفاده از مدل‌های زبانی از نقاط ضعف مدل‌های مبتنی بر طبقه‌بندی بوده که اقدامات مختلفی برای آن‌ها انجام شده است. چراکه در این مدل واژه‌یاب گفتار به عنوان یک دسته‌بند دودویی در نظر گرفته شده که کلاس جملات حاوی کلمات کلیدی را از کلاس جملات فاقد کلمات کلیدی تفکیک می‌کند. رویکرد واژه‌یابی گفتار مبتنی بر طبقه‌بندی که در ادامه این مقاله هم تنها به آن اشاره شده از دو بخش اصلی تشکیل شده است: استخراج ویژگی و الگوریتم یادگیری [10].

چارچوب کلی اکثر سامانه‌های واژه‌یاب در تحقیقات پیشین در شکل (۱) آورده شده است، این ساختار به دو بخش آموزش[1] و آزمایش[2] تقسیم می شود مرحله‌ی آموزش شامل فراهم آوری مجموعه داده گفتار، پیش‌پردازش صدا، استخراج ویژگی و درنهایت مدل‌سازی می‌باشد. مرحله ی آزمایش به‌مانند مرحله‌ی آموزش است با این تفاوت که به‌جای بخش مدل‌سازی، به جستجوی کلمات و یا واج‌ها بر روی مدل آموزش دیده پرداخته خواهد شد تا وجود و یا عدم وجود یک کلیدواژه در گفتار مشخص شود. در ادامه هر یک از مراحل تشریح شده است.



شکل ۱: چارچوب کلی سامانه‌های واژه‌یاب

۲-۱- پیش‌پردازش[3]

در محیط‌های واقعی گفتار انسان با عوامل تخریب‌کننده صدا ترکیب‌شده و سرانجام یک گفتار ناخالص تولید می‌شود. عوامل تخریبی صدا باعث کاهش کارایی نهایی سیستم تشخیص گفتار می‌شوند، ازاین‌رو باید به نحوی مؤثر، تأثیر این عوامل را در صدا کاهش داد. عوامل تخریبی صدا به دودسته‌ی ثابت و غیرثابت تقسیم می‌شوند. دسته‌ی ثابت نویزهایی هستند که در طول زمان در یک‌روند متناوب تکرار می‌شوند مثل نویز سفید و دسته‌ی غیرثابت نویزهایی هستند که در طول زمان می‌توانند مقادیر مختلفی بگیرند و اثر متفاوتی بر روی سیگنال صدا بگذارند مانند صدای موسیقی یا صحبت فرد دیگری در پس‌زمینه. به‌صورت کلی کاهش تأثیر نویزهای غیرثابت بسیار سخت‌تر از نویزهای ثابت می‌باشد، زیرا در نویزهای ثابت یک الگو وجود دارد که در طول زمان تکرار می‌شود و با شناسایی آن الگو می‌توان آن را شناسایی کرده و تأثیر آن را کاهش داد، منتهی در نویزهای غیرثابت الگوی مشخصی وجود ندارد و باید با استفاده از روش‌های پیچیده‌ی غیرخطی به‌مانند یادگیری عمیق، تأثیر این نویز را کاهش داد. یکی از مهم‌ترین شبکه‌های عصبی مربوط به کاهش نویز، شبکه‌های خودرمزکننده هستند [11].

در بخش پیش‌پردازش، بیشترین تحقیقات بر روی بهبود کیفیت صدا و افزایش نرخ نمونه‌برداری در سامانه‌های تشخیص گفتار طراحی می‌شوند. روال‌های مرسوم شامل روش‌های تبدیل سیگنال آنالوگ به دیجیتال، پنجره گذاری[4]، تبدیل فوریه سریع، محاسبه‌ی توان اسپکترال[5] و تولید ضرایب کپسترال فرکانس مِل MFCC[6] یا موارد دیگر به‌مانند پیش‌بینی خطی، پیش‌بینی خطی ادراکی[7] و غیره می‌باشند. به‌صورت کلی دو روش برای افزایش نرخ نمونه‌برداری در تحقیقات پیشین استفاده‌شده است. روش‌های سنتی درون‌یابی مانند Spline که با متوسط گیری و یا برازش توابع چندجمله‌ای بر روی‌داده‌ها، سعی بر درون‌یابی فرکانس‌های ازدست‌رفته دارند و دومین روش، روش‌های مبتنی بر یادگیری می‌باشند. روش‌های یادگیری، با مدل‌سازی از روی‌دادگانی با فرکانس پایین، دادگانی با نرخ وضوح‌بالا تولید می‌کنند. محققین در [12]یک مدل یادگیری عمیق خودرمزکننده مستقل برای افزایش وضوح سیگنال صدا در دامنه‌ی زمان پیشنهاد دادند که از خصوصیات مهم این طرح، افزایش ۴ برابری نرخ نمونه‌برداری فرکانس می‌باشد.

۲-۲- استخراج ویژگی

بعد از اعمال پیش‌پردازش‌های مختلف بر روی سیگنال گفتار سیگنال وارد قسمت استخراج ویژگی می‌شود. در این بخش، استفاده از فیلتر بانک‌ها رونق زیادی دارد و فیلتر بانک‌های مِل و بارک از مرسوم‌ترین فیلتر بانک‌های استخراج ویژگی به‌حساب می‌آمدند. سیگنال صدا با کانوالو شدن در فیلترهای مختلف، وزن‌هایی از آن فیلتر دریافت می‌کند که با کنار هم قرار دادن آن‌ها می‌توان به یک ویژگی منحصربه‌فرد از سیگنال صدا رسید [13]. در ادامه این ویژگی‌ها برای این‌که مشخص شود هر واج به چه احتمالی در گفتار وجود دارد به الگوریتم‌های برآوردگر احتمال واج[8] منتقل می‌شود. برآوردگر احتمال واج به همراه مدل HMM و مدل زبانی n-gr توسط رمزگذار[9] برای رمزگذاری گفتار استفاده می‌شود. کلمات

[1] Train
[2] Test
[3] Preprocessing
[4] Windowing
[5] Computing power spectrum
[6] Mel frequency cepstral coefficients
[7] Perceptual linear prediction
[8] Phone likelihood estimator
[9] Decoder

خروجی سپس به رمزگشا[1] فرستاده می‌شوند تا آن را به شکل قابل‌خواندن برای انسان تبدیل کند[14] .

۲-۳- مدل‌سازی و فناوری یادگیری عمیق

جهت مدل‌سازی واج‌ها یا کلمات، محققان درگذشته از روش‌های **آوایی و زبانی** استفاده می‌کردند. **روش‌های آوایی** تنها بر روی تشخیص آوا و واج‌ها به‌صورت مستقل کاربرد دارد در این رویکرد گفتار به کمک تعدادی واحد آوایی (مانند کلمه، هجا، سه واجی یا واج) مدل می‌شود و برای بازشناسی نیز از تشخیص این واحدها و کنار هم قرار دادن آن‌ها، متن متناسب با گفتار تشخیص داده می‌شود. **روش‌های زبانی** به گرامر زبان، هم‌خوانی واج‌ها و کلمات با یکدیگر و یا احتمال وجود کلمه‌ها در آن زبان تمرکز دارد. مدل‌های زبانی معمول مورداستفاده در سامانه‌های تشخیص گفتار امروزی شامل روش‌های **گرامری و آماری** هستند. درروش **گرامری** سعی می‌شود که به جملات خروجی، ساختار گرامری آن زبان اعمال شود و درروش **آماری** احتمال پشت سرهم آمدن کلمات (مثل مونوگرام یا احتمال وقوع کلمات در زبان، بایگرام یا آمار وقوع دو کلمه پشت سر هم در زبان و ...) مورداستفاده قرار می‌گیرند. خروجی شامل لیست واژگانی است که توسط سیستم بازشناسی می‌گردند، در سامانه‌های بازشناسی گفتار پیوسته با تعداد واژگان زیاد، علاوه بر لیست خود کلمات، اطلاعات مختلفی در مورد هر کلمه مانند احتمال وقوع آن در زبان، احتمال وقوع آن بعد از سایر کلمات، نقش (های) گرامری در جمله و ... را نیز شامل می‌شود.

مدل‌های آوایی نیز در تحقیقات گذشته بیشتر به دو صورت ۱- مدل‌های بدون نظارت همچون مدل مخفی مارکوف یا مدل مخلوط گاوسی و ۲- مدل‌های با نظارت مانند ماشین بردار پشتیبان[2] (SVM) بودند که با مرزبندی بین واج‌ها و همچنین اختصاص وزن‌های مختلف به داده‌های گفتار مختلف، می‌توانستند با دقت مناسبی واج‌ها را شناسایی و تشخیص دهند. در مدل‌سازی بدون نظارت با اینکه دقت مناسبی در داده‌های مختلف مشاهده می‌شد، اما هزینه‌ی آموزش و محاسبات بالا بود. در سال‌های گذشته روش‌های بانظارت شبکه عصبی بازگشتی توسعه پیدا کردند که ازنظر دقت، نسبت به روش‌های بدون نظارت برتری محسوسی داشتند [15]. محققان توانستند با ارائه‌ی شبکه‌ی عصبی عمیق[3](DNN) تمام متصل از نظر دقت و سرعت از روش‌های بدون نظارت سبقت بگیرند. با گذشت زمان شبکه‌ی عصبی کانولوشنی (CNN) [4] با استفاده از اسپکتوگرام صدا و تکیه بر شبکه‌های عصبی کانولوشنی توانستند دقت بیشتری نسبت به DNN ها کسب کنند. باوجود کارایی مناسب CNN ها، در نظر نگرفتن وابستگی بلندمدت داده‌ها، اشکال اصلی این روش بود. ازاین‌رو با تولید روش‌های ترکیبی از شبکه‌ی عصبی بازگشتی RNN [5]و CNN مدلی‌هایی به نام شبکه‌های عصبی کانولوشنی بازگشتی طراحی شد که در محیط‌های نویزی پایداری بیشتری نسبت به روش CNN از خود نشان می‌دادند. مدل‌های بازگشتی واحد بازگشتی دروازه GRU [6] و حافظه ی کوتاه‌مدت بلند LSTM[7] دو مورد از مدل‌های متداول و پرکاربرد از شبکه‌های بازگشتی هستند [16] .

برای بررسی عملکرد روش‌های موجود، یکی از محبوب‌ترین تابع محاسبه ی خطا روش مدل زمانی پیوسته[8] (CTC) است. در CTC اختلاف متن به‌دست‌آمده حاصل از پیمایش بر روی شبکه‌ی عصبی با متن اصلی محاسبه‌شده و سپس به مقدار خطای تشخیص حاصل‌شده، وزن‌های شبکه‌ی عصبی به هنگام می‌شوند، این روند در پروسه‌ی آموزش و توسعه ی سیستم تا هنگامی‌که وزن‌ها دیگر به هنگام نشوند و یا دقت موردنظر حاصل شود ادامه پیدا خواهد کرد. مدل زمانی پیوسته با در نظر گرفتن حالت‌های مختلفی از خروجی حقیقی برنامه، اختلاف کاراکترهای پیش‌بینی‌شده و کاراکترهای اصلی را به‌صورت میانگین محاسبه می‌کند. مدل‌های زبانی، احتمال رخداد واژه در یک جمله را با توجه به کلمات پیشین و پسین به دست می‌آورند و درصورتی‌که کلمه‌ای به‌اشتباه نمایش داده شود با توجه به دیکشنری از پیش‌ساخته شده آن را اصلاح می‌کند.

در مرحله جستجو در قسمت آزمایش نیز عمدتاً به‌صورت سلسله مراتبی انجام می‌شود. تحقیقات پیشین جهت جستجو در یک گفتار، ابتدا گفتار اصلی را به قاب‌هایی در سطح کلمات تقسیم‌بندی می‌کردند، در ادامه اولین واج از کلیدواژه‌ی موردنظر بر روی قاب‌های تقسیم‌بندی شده به ترتیب جستجو می‌کردند، اگر قاب منتخب دارای این واج بود، واج بعدی در همین قاب آزمایش می‌شد، این روند به همین صورت ادامه داشت تا اینکه با پیمایش بر روی کل گفتار، کلمه‌ی مورد نظر شناسایی شود، در صورت عدم وجود واج‌های پشت سر هم در گفتار، نتیجه گرفته می‌شود کلیدواژه در این بخش حضور ندارد و سیستم جستجو به قاب بعدی می‌رود. جستجوی درختی ساده، جستجو ویتربی، جستجوی مبتنی بر پشته و جستجوی *A، سامانه‌های جستجوی مطرح در این حوزه می‌باشند [15].

۳- روش پیشنهادی

یک نرم‌افزار واژه‌یاب باید تشخیص دهد که آیا کلیدواژه در گفتار وجود دارد یا خیر. لذا ساختار نرم‌افزار واژه‌یاب دارای ۳ مؤلفه‌ی اصلی می‌باشد که مهم‌ترین آن تبدیل گفتار به متن می‌باشد. بعد از آن، متن به‌دست‌آمده باید طبق ساختار مدل زبانی فارسی، غلط‌یابی شده و سپس کلیدواژه‌ها در متن جستجو شوند.

در مرحله اول باید مجموعه داده گفتار فارسی مناسب ساخته و توسعه داده شود سپس از این داده‌ها جهت توسعه مدل یادگیرنده استفاده شود. در این بخش، از روش‌های مبتنی بر یادگیری عمیق استفاده شده به طوریکه در صورت بهره‌گیری از معماری صحیح و وجود دادگان مناسب

[1] Parser
[2] Support vector machine
[3] Deep neural network
[4] Convolutional neural network
[5] Recurrent neural network
[6] Gated recurrent network
[7] Long short term memory
[8] Connectionist temporal classification

می‌توان به استخراجِ ویژگی‌های متمایزکننده و تعمیم‌پذیری بالای این روش ها امید داشت. لذا در این پژوهش از دو معماری جهت ساخت مدل یادگیرنده مبتنی بر شبکه عصبی عمیق استفاده شده است. این دو معماری در کنار یکدیگر مکمل هم بوده و می‌توانند دقت واژه‌یابی و تولید یک متن متناظر با گفتار را افزایش دهند. در ادامه بخش‌های مختلف روش پیشنهادی به‌صورت کامل تشریح شده است.

۳-۱- پایگاه داده مورد استفاده

اولین قدم در توسعه‌ی مجموعه داده گفتار، جمع‌آوری دادگان موجود می‌باشد که در این تحقیق از دو مجموعه داده استفاده شده است پایگاه داده اول مجموعه دادگان فارسی است که می‌توان به عنوان نوآوری این تحقیق در نظر گرفته شود و مجموعه دوم برای شرکت موزیلاست که در ادامه هر یک تشریح می‌شود.

۳-۱-۱- مجموعه دادگان گفتار فارسی

در این مرحله با توجه به وب‌سایت‌های ارائه‌دهنده‌ی کتاب‌های صوتی، کلیپ‌های تصویری و سایر رسانه‌های صوتی به‌مانند پادکست، بیش از ۲۰ گیگابایت فایل گفتار تهیه شد. در تهیه‌ی این اصوات سعی شد از یک گوینده بیش از یک و نیم ساعت فایل وجود داشته باشد تا اطمینان حاصل شود در توسعه‌ی سیستم، صدای گویندگان مختلف در آموزش سیستم نقش دارند. همچنین سعی شد در حالات مختلفی از صحبت، این فایل‌ها جمع‌آوری شوند. به‌طور مثال در این مجموعه داده، حالت‌های محاوره، رسمی، صحبت آرام و تند به زبان فارسی وجود دارد. متنوع بودن مجموعه داده باعث تعمیم‌پذیری مدل نهایی شده و کارایی سیستم را بیشتر خواهد کرد. برای آموزش داده‌ها باید پیش‌پردازش صورت گیرد، زیرا در فایل‌های گفتار بخش‌هایی غیر اطلاعاتی، مانند سکوت زیاد و یا آهنگ در صدا وجود دارند. لذا باید بر روی فایل‌ها پالایشی صورت گیرد تا آماده‌ی تزریق به شبکه‌ی عصبی طراحی شده باشند. ازاین‌رو هر یک از فایل‌های گفتار به فایل‌های زیر ۱۰ ثانیه تقسیم شدند به‌صورتی که اول و آخر هر فایل گفتار سکوت باشد تا داده‌ای تخریب نشود، همچنین سکوت‌های بیش از ۴ ثانیه حذف شدند تا اطلاعات غنی‌تر شوند. برای یکسان‌سازی فایل‌ها جهت تزریق به شبکه‌ی عصبی، تمامی فایل به فرکانس ۱۶۰۰۰ و فرمت wav. تبدیل شدند. لازم به ذکر است با توجه به زمان‌بر بودن برچسب‌گذاری این مجموعه داده، تنها بخشی از این مجموعه داده برچسب‌گذاری و مورد استفاده قرار گرفت.

ارائه‌ی متن فارسی از گفتار نیز باید به دو صورت برای معماری A و B با توجه به دیکشنری تبدیل حروف فارسی به انگلیسی (فینگلیش) در نظر گرفته شود برای معماری A باید به‌صورت خوانش واقعی درآید یعنی تمامی حرکت‌گذاری‌ها برای کلمه باید صورت بگیرد. به‌طور مثال کاربر کلمه‌ی "کتاب" را تایپ می‌کند ولی برای این‌که این کلمه به‌صورت صحیح و دقیق به شبکه‌ی طراحی شده اختصاص یابد، باید به کلمه‌ی "ketab"تبدیل شود تا خوانش واقعی آن لحاظ شود. لذا باید یک لغت‌نامه از صورت فارسی و خوانش واقعی وجود داشته باشد که بعد از اختصاص متن به گفتار، این متن به خوانش واقعی تبدیل شود. همان‌طور که مشخص است، ساختن مجموعه داده بسیار پرهزینه و زمان‌بر هست، ولی برای معماری B نیازی به خوانش واقعی نیست و به همان صورت که نوشته می‌شوند به فینگلیش تبدیل می‌گردند زیرا از مدل زبانی برای اصلاح متن خروجی استفاده می‌شود.

با توجه به موارد عنوان شده ۳۲ ساعت گفتار به همراه معادل آوایی متناظر برای هر فایل تهیه شد. این مجموعه داده با دقت بالا و در حد تمامی کسره‌های آخر و سایر حرکت‌گذاری‌ها هر کلمه مشخص شد، تمامی فایل‌های گفتاری ابتدا به زبان فارسی نوشته شدند و سپس با استفاده از واژه‌نامه‌ی آوایی متناظر به‌صورت معادل آوایی درآمدند. این مجموعه داده شامل ۱۸۳۶۳ عدد کلمه‌ی منحصربه‌فرد است و تعداد کل کلمات ۲۵۱۲۴۱ است. همراه با این مجموعه داده یک واژه‌نامه‌ی آوایی برای کلمات فارسی تولید شد که برای تبدیل کلمه‌های فارسی به کلمه ی آوایی بسیار مهم است.

۳-۱-۲- دادگان موزیلا

در سال‌های اخیر شرکت موزیلا مشکل کمبود داده برای سایر زبان‌ها را تا حد زیادی حل کرده است، این کار با کمک کاربران هر زبان انجام شده و تا به امروز توانسته حجم زیادی از داده‌های گفتاری را تهیه نماید. در این مقاله از مجموعه داده‌ی فارسی رایگان ۱۳۲ ساعته شرکت موزیلا[1] نیز استفاده شده است؛ که با پالایش و منظم سازی فایل‌های این مجموعه داده، درمجموع حدود ۱۶۵ ساعت گفتار به همراه متن گردآوری شد.

در پروژه موزیلا کاربران با ضبط صدای خود و همچنین با تأیید یا عدم‌تأیید صدای سایر کاربران، به تهیه دادگان کمک می‌کنند. اگر حداقل دو کاربر مختلف یک داده را تأیید کنند، آن داده به‌عنوان داده معتبر و اگر حداقل دو کاربر مختلف یک داده را عدم‌تأیید کنند، آن داده به‌عنوان داده نامعتبر شناخته خواهد شد. در جدول (۱) سن گویندگان نشان داده شده است. همچنین ۷ درصد از گویندگان زن، ۷۵ درصد مرد و ۱۸ درصد از گویندگان جنسیت خود را مشخص نکرده‌اند.

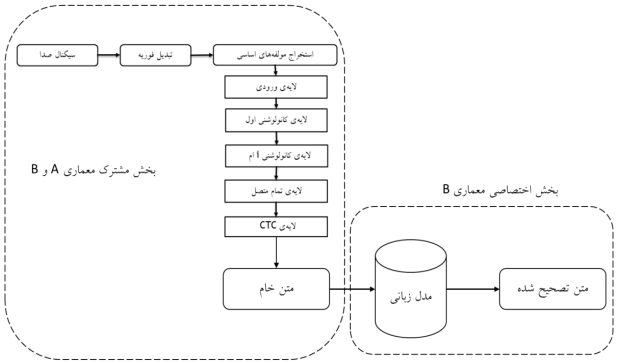
سن گوینده	درصد دادگان
کمتر از ۱۹	۱
بین ۱۹ تا ۲۹	۳۳
بین ۳۰ تا ۳۹	۳۹
بین ۴۰ تا ۴۹	۳
بیشتر از ۵۰	۲
نامشخص	۲۲

جدول ۱: فراوانی سن گویندگان در دادگان موزیلا

[1] commonvoice.mozilla.org/en/datasets, 2020

۳-۲- مدل یادگیری

در این پژوهش از دو معماری جهت تولید متن استفاده شده که مکمل یکدیگر بوده و می‌تواند دقت واژه‌یابی و تولید یک متن متناظر با گفتار را افزایش دهد. به‌صورت کلی ساختار این دو معماری در شکل (۲) نمایش داده شده است. برای ساده‌سازی در ادامه، این دو معماری به‌صورت معماری A و معماری B تعریف می‌شود.



شکل ۲: معماری کلی شبکه عصبی

همان‌طور مشاهده می‌شود، این دو معماری در کلیات شبیه یکدیگر هستند منتها در چینش لایه‌های شبکه‌ی عصبی و نحوه‌ی محاسبه‌ی تابع خطا و به‌روزرسانی با یکدیگر متفاوت هستند. معماریA ، بسیار ساده طراحی شده و تنها هدفش تولید واج متناظر با گفتار، بدون داشتن اطلاعات مدل زبانی می‌باشد. ولی معماری B کمی پیچیده‌تر بوده و با لایه‌های متعدد و داشتن مدل زبانی هدفش تولید یک متن صحیح با توجه به دامنه‌ی زبان فارسی می‌باشد. در هر دو معماری ابتدا گفتار اصلی را به قاب‌هایی در سطح واج تقسیم‌بندی می‌شوند و از بُعد زمان به بُعد فرکانس انتقال داده می‌شود. سپس تعدادی فرکانسِ مهم که در شنوایی انسان و یا ساختِ سیگنال صحبت انسان از اهمیت بیشتری برقرار است انتخاب می‌شوند. مؤلفه‌های فرکانسی به‌دست‌آمده با عبور از لایه‌های کانولوشنی متعدد و استخراج ویژگی‌های گوناگون در انتزاع‌های مختلف، می‌تواند ابعاد پنهان گفتار انسان را مدل‌سازی کند. لایه یکی مانده به آخر تعداد نورون‌ها به‌اندازه‌ی حداکثر تعداد واج در نظر گرفته‌شده و در لایه‌ی آخر با استفاده از مدل زمانی پیوسته (CTC) اختلاف متن به‌دست‌آمده حاصل از پیمایش بر روی شبکه‌ی عصبی، با متن اصلی محاسبه شده و سپس به مقدار خطای تشخیص حاصل‌شده، وزن‌های شبکه‌ی عصبی به هنگام می شوند، این روند در پروسه‌ی آموزش و توسعه‌ی سیستم تا هنگامی‌که وزن‌ها دیگر به هنگام نشوند و یا دقت موردنظر کسب شود ادامه پیدا خواهد کرد. معماری A به‌گونه‌ای طراحی شده که در انتهای خروجی شبکه‌ی عصبی تنها واج‌های متناظر با گفتار، بدون در نظر گرفتن واج‌های پیشین و یا کلمات پیشین، پیش‌بینی شوند. در این روش از مدل زبانی استفاده نمی‌شود و واج‌ها به‌صورت خام جهت پردازش فرآیند بعدی، یعنی جستجوی کلمه‌ی کلیدی، مورداستفاده قرار می‌گیرند. دلیل استفاده از این معماری سرعت بالای این معماری بود و با استفاده از معماریA ، انتظار می‌رود واج‌هایی که به‌صورت ناقص از کلمات مختلف بیان شود نیز لحاظ شود، بنابراین از این روش در کنار معماری B بهره برده شده است. در ادامه معماری A و B و نحوه‌ی ترکیب آن‌ها بیان شده است.

۳-۲-۱- معماری A

برای مدل‌سازی کاراکترهای فارسی در این مدل تمام‌صورت‌های آوایی در نظر گرفته شده است. به دلیل وجود حروف با صدای یکسان در زبان فارسی، برای بالا بردن دقت مدل، حروف هم‌صدا با یک حرف نشان داده شده و یک حرف نماینده آن دسته از حروف می‌باشد (شکل ۳).

مثال	حروف
آبی==Abi	آ -- A
آسب==asb	أ -- a
احسان==ehsAn	ا -- e
امید==omid	أ -- o
سوله==sule	او -- u
کتابی==ketAbi	ایی -- i
یاور==yAvar	ی -- y
بابک==bAbak	ب -- b
پایان==payAn	پ -- p
طناب==tanab	ت ط -- t
صندلی==sandali	ث س ص -- s
چهار==cAhAr	چ -- c
حوصله==hosele	ح ه -- h
خرداد==xordAd	خ -- x
در==dar	د -- d
زبان==zabAn	ذ ز ض ظ -- z
رادیو==rAdiyo	ر -- r
ژاپن==ZApon	ژ -- Z
شیر==Sir	ش -- S
قالب==qAleb	غ ق -- q
فارغ==fAreq	ف -- f
کار==kAr	ک -- k
گور==gur	گ -- g
لاک==lAk	ل -- l
مردم==mardom	م -- m
نارنج==nArenj	ن -- n
ولی==vali	و -- v

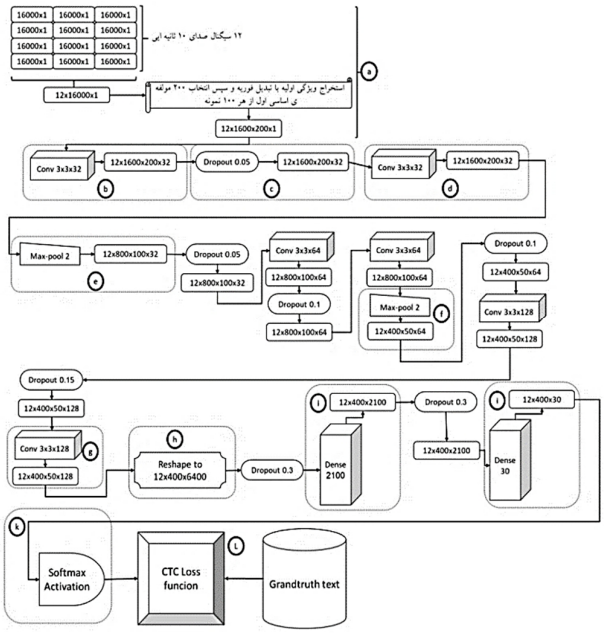
شکل ۳: مدل‌سازی آواهای فارسی در معماری A

در شکل (۴) دیاگرام اصلی معماری A مشخص‌شده است. در فاز آموزش فایل‌های آموزشی وارد شبکه‌ی عصبی شده و بعد از محاسبه خطا و اصلاح وزن‌ها، با برچسب فایل‌ها تا رسیدن به‌دقت مطلوب ادامه می‌یابد.

در تمام لایه‌های میانی این معماری از تابع فعال‌سازی یک‌سوساز خطی[1] و از الگوریتم بهینه‌سازی ADAM جهت به‌روزرسانی وزن‌ها استفاده شده

1 Rectified linear units- RELU

است. میزان یادگیری[1] اولیه در این معماری با توجه به آزمایش‌های متعددِ صورت گرفته برابر 0.001 در نظر گرفته شد.



شکل ۴: شبکه عصبی طراحی شده در معماری A

در این معماری دو نوع استخراج ویژگی صورت گرفته در گام اول به ازای هر قاب از گفتار اصلی با تبدیل فوریه زمان کوتاه ۲۰۰ ضریب ابتدایی آن استخراج می شود که اگر فرکانس نمونه‌برداری ۱۶ کیلوهرتز باشد، ابعاد ماتریس بردار ویژگی 1*200*1600 به ازای هر فایل گفتار می‌شود (شکل (۴) بخشa) بعد از به دست آوردن ۱۲ ماتریس (تعداد فایل‌های گفتار که در هر مرحله برای اصلاح وزن‌ها در نظر گرفته می‌شود) بردار ویژگی 1*200*1600 در مرحله‌ی قبل، در گام دوم این ماتریس ویژگی، جهت کسب ویژگی‌های پنهان از ۳۲ فیلتر کانولوشن دوبعدی ۳*۳ عبور می‌کند تا ویژگی‌های انتزاعی در سطوح مختلف به دست آید که خروجی آن‌یک ماتریس 12×1600×200×32 خواهد بود (شکل (۴) بخش b). در این معماری به ازای هر لایه‌ی کانولوشن یک‌لایه‌ی Dropout (جهت خاموش‌سازی بخشی نورون‌ها) (شکل (۴) بخش c)؛ و به ازای هر دولایه‌ی کانولوشن یک‌لایه‌ی maxpooling (ماکزیمم مقدار هر پنجره) تعریف‌شده است (شکل (۴) بخش e). با حرکت در عمق شبکه، حساسیت به ویژگی های جزئی و نویز بیشتر می‌شود، لذا با اعمال خاموش‌سازی بیشتر، از بیش برازش شبکه جلوگیری می‌شود. در بخش e یک‌لایه maxpooling با ضریب ۲ جهت نصف کردن ابعاد ماتریس اعمال می‌شود تا ابعاد ماتریس به‌صورت12×800×100×32 درآید. ادامه فرآیندها، چرخه تکراری اعمال لایه کانولوشن، Dropoutو maxpooling با ضرایب مختلف است که به ترتیب عبارت‌اند از اعمال لایه‌ی Dropout با ضریب ۰,۰۵، اعمال لایه‌ی کانولوشن ۶۴ فیلتره که درنتیجه ماتریسی به ابعاد12×800×100×64 حاصل می‌شود، اعمال لایه‌ی Dropout با ضریب ۰,۱، اعمال لایه‌ی کانولوشن ۶۴ فیلتره که درنتیجه ماتریسی به ابعاد 12×800×100×64 حاصل می‌شود، اعمال maxpooling با ضریب ۲ که ماتریس ورودی را به ابعاد12×400×50×64 تبدیل می‌کند (شکل (۴) بخشf)، اعمال لایه‌ی Dropout، اعمال لایه‌ی کانولوشن ۱۲۸ فیلتره که درنتیجه ماتریسی به ابعاد 12×400×50×128 حاصل می‌شود، اعمال لایه‌ی Dropout با ضریب ۰,۳، اعمال لایه‌ی کانولوشن ۱۲۸ فیلتره که درنتیجه ماتریسی به ابعاد 12×400×50×128حاصل می‌شود (شکل (۴) بخشg)، در مرحله بعد ماتریس ویژگی، جهت متناسب‌سازی با لایه‌ی خروجی، تغییر شکل پیدا می‌کند و به ماتریسی با ابعاد 12×400×6400 تبدیل می‌شود در حقیقت در این گام دو بعد آخر باهم ترکیب می‌شوند (شکل (۴) بخشh)، در مرحله بعد یک لایه Dropout اعمال می‌شود و سپس به یک‌لایه‌ی تمام متصل[2] 2100 نورونی، فرستاده می‌شود (شکل (۴) بخشi) و به یک ماتریس با ابعاد 12×400×2100 تبدیل می‌شود. سپس یک لایه Dropout اعمال و به یک‌لایه‌ی تمام متصل 30نورونی که همان تعداد واج‌های زبان فارسی (به همراه سکوت) می‌باشد فرستاده می‌شود (شکل (۴) بخشj) درنتیجه در این گام، ماتریس به ابعاد 12×400×30 تبدیل می‌شود. به‌طور دقیق‌تر برای هر فایل گفتار ورودی ماتریسی به‌اندازه‌ی400*30 پیش‌بینی می‌شود که درایه‌های آن برابر احتمال وجود هر کاراکتر می‌باشد، عدد 400به معنای توالی کاراکترها یا همان اندازه‌ی متن می‌باشد که ثابت است، قاعدتاً فایل گفتار زیر 10 ثانیه 400 کاراکتر را در خود ندارند، به همین جهت در پیاده‌سازی محاسبه‌ی خطای تشخیص برنامه یک کاراکتر #=blank در نظر گرفته شده که در بین کلمات قرار می‌گیرد ولی خوانده نمی‌شود به‌طور مثال واژه‌ی ketAb و kk#eee#t###AAA#b و یا #er####hayyy#ote#mmmmو motehayyer برابر خواهند بود. در مرحله بعدی ماتریس انتهایی جهت پیش‌بینی کاراکترهای تعریف‌شده به لایه‌ی فعال‌سازی softmax فرستاده می‌شود (شکل (۴) بخشk). در این معماری، ماکزیمم کاراکتر موجود در فایل 10 ثانیه‌ای ۱۷۸ کاراکتر در نظر گرفته‌شده است (با توجه به نمونه‌های موجود). لذا ماتریس12×400×30 پیش‌بینی‌شده در مرحله قبل، همراه با برچسب(واقعی)، به لایه‌ی آخر برای تعیین خطا ارسال می‌شود (شکل (۴) بخش L). در لایه‌ی آخر با استفاده از مدل زمانی پیوسته یا همان CTC اختلاف متن به‌دست‌آمده حاصل از پیمایش بر روی شبکه‌ی عصبی با متن اصلی محاسبه شده و سپس به مقدار خطای تشخیص حاصل‌شده، وزن‌های شبکه‌ی عصبی به هنگام می‌شوند، این روند در پروسه‌ی آموزش سیستم تا هنگامی‌که وزن‌ها دیگر به هنگام نشوند و یا دقت موردنظر کسب شود ادامه پیدا خواهد کرد. خروجی این معماری در مرحله‌ی آزمایش یک متن خام متناظر با گفتار خواهد بود که واج‌های آن بدون در نظر گرفتن احتمال حضور واج‌های

1 Learning rate

2 Fully connected

دیگر تولید شدند قاعدتاً لایه‌ی محاسبه خطا در بخش آزمایش، حذف می شود[1].

۳-۲-۲- معماری B

در شکل (۵) دیاگرام اصلی معماری B نشان داده شده است. این معماری بر اساس روش هانون که شرکت موزیلا آن را با زبان برنامه‌نویسی پایتون و کتابخانه تنسورفلو پیاده‌سازی کرده[2] طراحی شده است[15] . هسته‌ی اصلی این سیستم شبکه‌ی عصبی RNN می‌باشد که با توجه به ویژگی‌های اسپکتوگرام صدا، می‌تواند یک متن متناظر به سیگنال صحبت انسان بدهد. شبکه‌ی RNN جهت مدل‌سازی فرآیندهای مبتنی بر زمان مناسب بوده و می‌تواند از اطلاعات لحظات $t_n,t_{n-1},\dots$ برای تصمیم‌گیری در لحظه ی t_n استفاده کند.

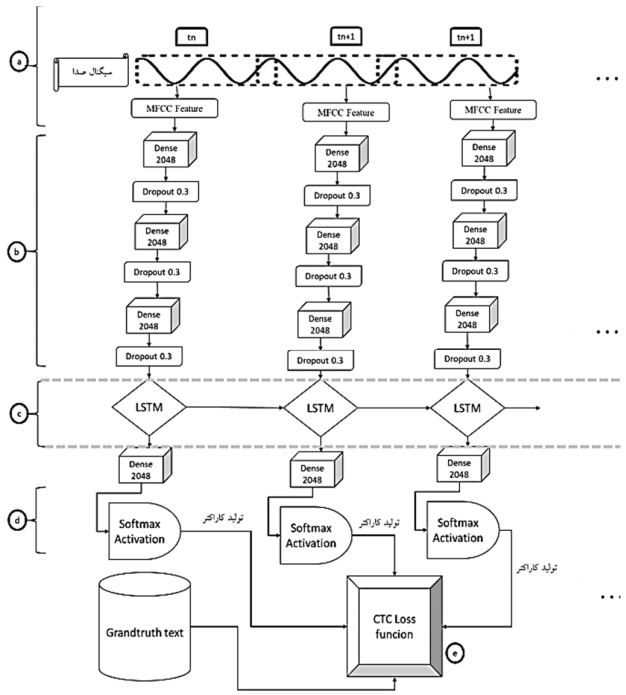
در روش پیشنهادی در این معماری از ویژگی‌های MFCC که پیش‌تر در پیشینه بیان شد جهت استخراج ویژگی استفاده می‌شود. ازاین‌رو با تنظیم گروه داده‌ها برابر ۶۴، (در یک به‌روزرسانی از ۶۴ فایل گفتار، MFCCگرفته‌شده) پیمایش بر روی شبکه‌ی عصبی با استفاده از این ویژگی‌ها شروع می‌شود (شکل (۵) بخشa) این معماری از ۵ لایه‌ی اصلی استفاده می‌کند که ۳ لایه اول غیر بازگشتی، لایه‌ی چهارم بازگشتی و لایه‌ی آخر مجدداً غیر بازگشتی می‌باشد و علاوه بر ۵ لایه‌ی اصلی فوق، از لایه‌ی Dropout نیز با پارامتر 0.3 و الگوریتم بهینه‌سازی ADAM با میزان یادگیری [3]اولیه 0.0001 و مقدار آستانه ۲۰ برای به‌روزرسانی وزن‌ها استفاده شده است.

در گام نخست ویژگی‌های MFCC از ۳ لایه‌ی تمام متصل با تعداد نورون ۲۰۴۸ به همراه لایه‌های Dropoutعبور داده می‌شوند (شکل (۵) بخش b). بعد از گذر از لایه‌های تمام متصل، لایه‌ی چهارم این شبکه‌ی عصبی، شبکه‌ی بازگشتی LSTM می‌باشد. این شبکه درواقع حافظه‌ی بلندمدت شبکه‌ی عصبی بوده و اطلاعات حال و گذشته را با وزن دهی مناسب تنظیم می‌کند، اطلاعات این لایه علاوه بر پیمایش جهت تولید کاراکتر در زمان t_n، به لحظه‌ی t_{n+1}نیز ارسال می‌شود (شکل (۵) بخشc).

لایه‌ی پنجمِ این شبکه یک‌لایه‌ی تمام متصل با ۲۰۴۸ نورون می‌باشد؛ که بعد از اِعمال تابع فعال‌سازیRELU ، جهت تخمین احتمال وجود کاراکترها به لایه‌ایی با تابع فعال‌سازی softmax فرستاده می‌شود (شکل (۵) بخشd). لایه‌ی آخر، لایه‌ی تخمین خطای CTC می‌باشد ((شکل (۵) بخشe)، در این لایه، اختلاف متن به‌دست‌آمده حاصل از پیمایش بر روی شبکه‌ی عصبی با متن اصلی محاسبه شده و سپس به مقدار خطای تشخیص حاصل‌شده، وزن‌های شبکه‌ی عصبی به هنگام می‌شوند، این روند در پروسه‌ی آموزش و توسعه تا هنگامی‌که وزن‌ها دیگر به هنگام نشوند و یا دقت موردنظر کسب شود ادامه پیدا خواهد کرد.

خروجی این معماری در مرحله‌ی اول آزمایش یک متن خام متناظر با گفتار خواهد بود که واج‌های آن (با توجه به اِعمال شبکه LSTM) با احتمال حضور واج‌های دیگر تولید شده‌اند. در مرحله‌ی دوم با استفاده از مدل زبانی، احتمال حضور واج‌ها و کلمات در کنار یکدیگر سنجیده شده و درنهایت متن اصلاح و غلط‌یابی می‌شود. برای ساخت مدل زبانی می بایست یک مجموعه داده از متن‌ها و جملات مختلف در حوزه‌های گوناگون جمع‌آوری‌شده و سپس با یک الگوریتم n-gram، احتمال حضور هر کلمه در حضور کلمات دیگر محاسبه و ذخیره شود (روش n-gram به این صورت است که در یک جمله به ازای هر کلمه، تعداد تکرار n کلمه‌ی قبلی و یا بعدی محاسبه شده و درنهایت، به ازای هر کلمه مشخص می‌شود چه تعداد کلمات دیگری و با چه تکراری حضور دارند). در این پژوهش درمجموع، با تهیه‌ی حدود ۲ میلیون جمله‌ی فارسی یکتا که از 170هزار کلمه‌ی یکتا تشکیل می‌شد و با روش 5-gram مدل زبانی توسعه پیدا کرد (چهار نوع مدل زبانی دو گرام، سه‌گرام، چهار گرام و پنج گرام) لازم به ذکر است در این پژوهش، روش ساختِ مدل زبانی بر اساس kenlm طرح‌ریزی و پیاده‌سازی شد[4]. کد پیاده‌سازی معماری B نیز در لینک زیر آورده شده است.

https://github.com/mahdaviFarAsr/asr.git



شکل ۵: شبکه عصبی طراحی شده در معماری B

[1] لینک پیاده‌سازی معماری A: https://github.com/mahdaviFarAsr/asr.git

2 https://github.com/mozilla/DeepSpeech/

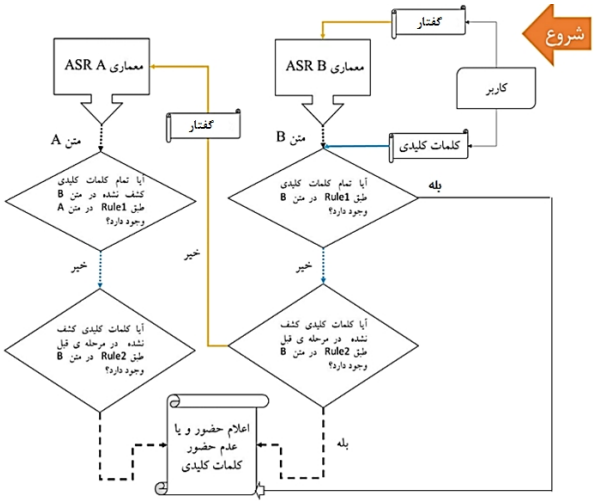
3 Learning rate

4 https://kheafield.com/code/kenlm/

۳-۳- واژه‌یابی

بعد از آموزش شبکه‌ی عصبی در دو معماری A و B و به دست آوردن وزن‌های این دو شبکه، در مرحله‌ی آزمایش دو متن به ازای هر فایل گفتار خواهیم داشت. در متن حاصل از معماریA ، حضور یک واج به حضور واج‌های دیگر وابسته نیست و تنها از مدل آکوستیکی استفاده شده است؛ بنابراین در این معماری، انتظار می‌رود، به ازای هر کلمه‌ای که بیان می‌شود، تعدادی واج متوالی صحیح در آن کلمه پیش‌بینی شود.

در متن حاصل از معماریB ، حضور واج‌ها و کلمات به کلمات پسین و پیشین وابسته می‌باشد، لذا در این معماری کلمات با معنی که در مدل زبانی قبلاً وجود داشتند ارائه خواهد شد. به‌صورت کلی همان‌طور که در قسمت بعد بررسی خواهد شد، دقت معماری B با توجه به فرآیندهای بیان‌شده، بیشتر از معماری A بوده، منتهی در صورت بیان کلماتی خارج از دامنه‌ی زبانی آموزش داده شده در معماریB ، روش معماری A می تواند تا حد مطلوبی تعدادی از واج‌های این کلمات را تشخیص داده و درمجموع مکمل خوبی برای معماری B می‌باشد. در شکل (۶) نحوه‌ی استفاده از این دو معماری جهت واژه‌یابی نشان داده شده است.



شکل ۶: فرآیند واژه‌یابی

همان‌طور که در شکل (۶) مشخص می‌باشد، جهت واژه‌یابی در گفتار، ابتدا کاربر یک سیگنال گفتار به همراه یک مجموعه کلمات کلیدی در سیستم بارگذاری می‌کند. سپس ابتدا کلمات کاربر در متنِ معماری B جستجو شده و در صورت پیدا نکردن کلمات کاربر، این واژه‌ها در متنِ معماری A جستجو خواهند شد. برای تشخیص واژه در گفتار دو قانون به نام Rule1 و Rule2 طبق شکل (۶) طراحی گردیده است که Ruel1 به دنبال شکل صحیح و کامل کلمات در گفتار می‌باشد. در بسیاری از موارد شکل کامل کلمات ممکن است بیان نشود، به‌طور مثال ممکن در متنی به‌جای کلمه "هستیم"، کلمه "هستم" پیش‌بینی شود و یا تعدادی بیشتری از واج یک کلمه به نادرستی بیان شوند. جهت حل این مشکل قانون Rule2طراحی شد. در این قانون به ازای یک بازه‌ی اطمینان از شباهت دو کلمه، وجود و یا عدم وجود یک کلیدواژه در گفتار اعلام می شود. الگوریتم لون اشتیان متداول‌ترین روش جهت تشخیص فاصله یا همان شباهت دو کلمه را می‌باشد. با استفاده از این الگوریتم می‌توان کمترین تعداد عملیات لازم جهت تبدیل یک کلمه به کلمه‌ی دیگر را محاسبه کرد. عملیات تعریف‌شده در فاصله‌ی لون اشتاین در این بخش، درج، جایگذاری و حذف می‌باشد .به‌عنوان‌مثال فاصله لون اشتاین بین "kitten"و "sitting" برابر ۳ است چراکه حداقل سه ویرایش برای تبدیل یک‌رشته به رشته‌ی دیگر وجود دارد، در این پژوهش از بازه‌ی اطمینان ۱ استفاده شده است یعنی کلمات کشف‌شده الزاماً باید در یک واج تفاوت داشته باشند.

۴- پیاده‌سازی و ارزیابی نتایج

پیاده‌سازی روش پیشنهادی تحت زبان برنامه‌نویسی پایتون و با استفاده از کتابخانه‌های Tensorflow و Keras برای پیاده‌سازی معماری‌های A,B و برای کار روی فایل‌های گفتار از کتابخانه‌های wave, scipy, ffmpeg, Pydub, PyAudio استفاده شده است.

جهت سنجش کارایی روش واژه‌یابی می‌بایست دقت پیدا کردن واژه موجود در گفتار و دقت در پیدا نکردن واژه ناموجود در گفتار، به‌صورت مستقل بررسی شود تا معیاری صحیح و همه‌جانبه از کارایی واژه‌یاب طراحی شده به دست بیاید. همچنین در ادامه، به تأثیر پارامترهای شبکه عصبی و مدل زبانی در کارایی روش پیشنهادی با ارائه نمودارها و جداول مختلف پرداخته شده است.

۴-۱- معیارهای سنجش کارایی

هسته اصلی معیار ارزیابی تعریف شده در تبدیل گفتار به متن همان‌طور که پیش‌تر بیان شد بر اساس الگوریتم لون اشتاین می‌باشد. درواقع با استفاده از این الگوریتم می‌توان کمترین تعداد عملیات لازم (درج، جایگذاری و حذف) جهت تبدیل یک کلمه به کلمه دیگر را محاسبه کرد. جهت بیان دقیق‌تر و قابل تفسیر کارایی، فاصله لون اشتاین اگر کمتر از طول رشته حقیقت یا متن اصلی باشد تقسیم بر این طول شده و در عدد ۱۰۰ ضرب می‌شود تا عددی بر مبنای درصد بیان شود. اگر میزان فاصله لون اشتاین بیشتر از طول رشته حقیقت باشد، آنگاه، میزان خطا ۱۰۰ درصد تعریف می‌شود، بنابراین هرچه عدد به‌دست‌آمده کمتر باشد، کارایی بهتر خواهد بود.

جهت آزمایش کارایی دقت واژه‌یابی از ماتریس درهم‌ریختگی استفاده شد. این ماتریس جهت ارزیابی عملکرد الگوریتم‌های هوش مصنوعی در دسته‌بندی داده‌ها استفاده فراوانی داشته و طبق مقادیر آن می‌توان به نتایج قابل تفسیر رسید. نتایج مثبت، یعنی الگوریتم پیشنهادی اعلام می‌کند واژه در گفتار وجود دارد و منفی یعنی واژه در گفتار وجود ندارد؛ بنابراین می‌توان ماتریس درهم‌ریختگی را به‌صورت زیر بیان کرد [18].

معیار اول [1] (TP) می‌باشد که این معیار، صحت وجود واژه‌ها در متن پیش‌بینی‌شده با توجه به تک‌تک واژه‌های موجود در متن برچسب خورده و اصلی، سنجیده می‌شود.

معیار دوم،[2](FP) می‌باشد و نشان می‌دهد چه تعداد واژه که در گفتار وجود نداشتند، به اشتباه در دسته مثبت قرار گرفتند. جهت آزمایش این معیار برای هر فایل گفتار ۲۵ واژه به‌صورت تصادفی استخراج شد تا سنجیده شود آیا این واژه‌هایی که در اصل در گفتار موجود نیستند، سیستم واژه‌یاب آن‌ها را پیدا خواهد کرد یا نه. این ۲۵ واژه به‌هیچ‌عنوان با واژه‌های اصلی فایل گفتار برابر نیستند.

معیار سوم[3] (FN) می‌باشد و نشان می‌دهد چه تعداد واژه که در گفتار وجود داشتند، به اشتباه در دسته منفی قرار گرفتند.

معیار چهارم[4] (TN) می‌باشد و نشان می‌دهد چه تعداد واژه که در گفتار وجود نداشتند، به‌درستی در دسته منفی قرار گرفتند. جهت آزمایش این معیار از همان آزمایش معیار دوم که نرخ تشخیص واژه‌های تصادفی ناموجود در متن اصلی می‌باشد، استفاده شده است.

به‌جای استفاده مستقیم از این ۴ معیار، به‌طور متداول از معیارهای حساسیت[5] (صحت[6]،TPR [7])، تخصیص[8]، نرخ تشخیص نادرست دسته مثبت (FBR)[9] و دقت کل[10] جهت ارزیابی دقیق‌تر استفاده می‌شود که طریقه محاسبه هر یک در جدول (۲) آورده شده است.

جدول ۲: معیارهای ارزیابی

این معیار مشخص می‌کند چه تعداد از واژه‌های موجود در گفتار به‌درستی تشخیص داده شدند.	حساسیت (R) = $\frac{TP}{FN+TP}$
این معیار در مشخص می‌کند چه تعداد از واژه‌های ناموجود در گفتار به‌درستی در دسته منفی قرار گرفتند.	تخصیص= $\frac{TN}{TN+FP}$
این معیار در مشخص می‌کند چه تعداد از واژه‌های ناموجود در گفتار به نادرستی در کلاس مثبت قرار گرفتند.	FBR = $\frac{FP}{TN+FP}$
	دقت کل = $\frac{TP+TN}{TP+FP+TN+FN}$

طبق تعاریف انجام شده در جدول (۲)، کارایی سیستم وقتی مناسب است که مقادیر حساسیت، تخصیص و دقت کل بالا بوده و درعین‌حال معیار FBRکم باشد.

۴-۲- دادگان مورد استفاده

جزئیات دادگان مورد سنجش که Speech168H نام‌گذاری شده در جدول (۳) آورده شده است. به‌صورت کلی در آزمایش‌های صورت گرفته بر روی این مجموعه داده، از 143.25 ساعت داده گفتار فارسی جهت آموزش، 2.77 ساعت داده جهت اعتبارسنجی و 12.38 ساعت جهت آزمایش استفاده شد.

جدول ۳: جزئیات مجموعه داده Speech168H

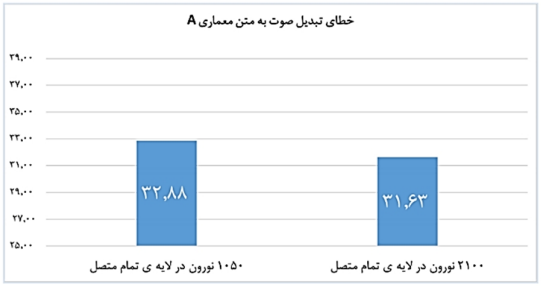
مرحله	تعداد فایل	مجموع ساعت
آموزش	۵۵۲۲۸	۱۴۳,۲۵
اعتبارسنجی	۱۰۰۰	۲,۷۷
آزمایش	۴۴۶۴	۱۲,۳۸

علاوه بر مجموعه داده ذکر شده در جدول (۳)، یک آزمایش دیگر بر روی مجموعه ی داده Mozilla7H جهت سنجش دقیق‌تر کارایی سیستم نهایی صورت گرفته است. در این مجموعه داده سعی شده حتی‌الامکان گویندگان متفاوت باشند و هیچ متن مشابه ای در فاز آموزش و آزمایش نباشد. این مجموعه شامل اصوات محاوره‌ای بوده و درجه سختی آن بسیار بالاتر از مجموعه داده Speech168H می‌باشد. مجموعه داده موزیلای مورد استفاده در این مقاله شامل ۸۴۵۱ فایل گفتار به طول ۷,۶۷ ساعت جهت آزمایش در روش پیشنهادی می‌باشد که از مجموعه داده موزیلای فارسی گردآوری شده است.

۴-۳- نتایج تبدیل گفتار به متن

با توجه به آزمایش‌های صورت گرفته در این تحقیق، سعی شده تأثیر پارامترهای مختلف در عملکرد روش‌های پیشنهادی بررسی شود، لذا در ادامه تأثیر به‌کارگیری تعداد نورون‌های مختلف در لایه‌ها و تأثیر به‌کارگیری مدل زبانی در معماری B و ... نشان داده شود.

یکی از پارامترهای تأثیرگذار در دقت شبکه عصبی، تعداد نورون لایه‌های تمام متصل هستند، به همین دلیل آزمایشی طراحی شد تا با تغییر نورون‌های لایه‌های تمام متصل، کارایی معماری A و B مشخص شود. نتایج آزمایش‌های صورت گرفته بر روی معماریA ، در نمودار شکل (۷) نشان داده شده است.



شکل ۷: کارایی معماری Aبا تغییر تعداد نورون‌ها بر روی Speech168H

همان‌طور که در نمودار شکل (۷) مشخص می‌باشد، خطای تشخیص معماری A با در نظر گرفتن 1050و 2100 نورون در لایه آخر نشان داده

1 True Positive
2 False Positive
3 False Negative
4 True Negative
5 Sensitive
6 Precision
7 True positive rate
8 Specificity
9 False positive rate
10 Total accuracy

شده است. افزایش تعداد نورون‌های لایه آخر توانسته حدود یک درصد خطای تشخیص کاهش داده تا میزان آن به ۳۱,۶۳ برسد. افزایش نورون‌ها تا حدی در کاهش خطا تأثیر داشته، منتهی در صورت افزایش این نورون‌ها پیچیدگی شبکه عصبی نسبت به داده‌های موجود نیز افزایش داشته و سبب بیش برازش شبکه عصبی می‌شود. به همین دلیل افزایش پارامترها بیش از 2100 نورون در لایه انتهایی پیشنهاد نمی‌شود.

نتایج آزمایش‌های صورت گرفته بر روی معماریB ، نیز در نمودار شکل (۸) نشان داده شده است. در این نمودار، خطای تشخیص معماری B با در نظر گرفتن ۱۰۲۴ و ۲۰۴۸ نورون در لایه‌های تمام متصل و همچنین تأثیر اضافه شدن مدل زبانی بر هر یک از این شبکه‌ها نشان داده شده است.

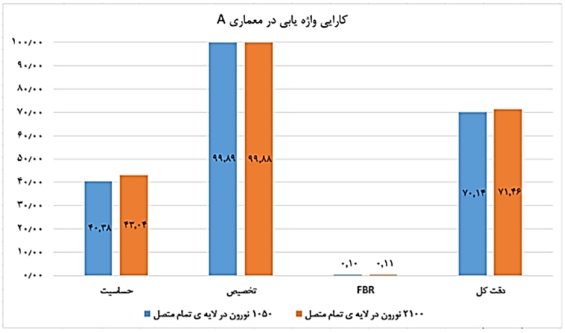
شکل ۸: کارایی معماری B با تغییر تعداد نورون‌ها و مدل زبانی

همان‌طور که در نمودار شکل (۸) مشخص است، افزایش تعداد نورون‌های لایه‌های تمام متصل در معماری B از ۱۰۲۴ نورون به ۲۰۴۸ نورون، موجب کاهش ۲۰ درصدی خطا شده و همچنین اضافه شدن مدل زبانی باعث کاهش خطای تشخیص ۵ درصدی می‌شود. وجود مدل زبانی، در هر دو مدل ۱۰۲۴ و ۲۰۴۸ نورونی، باعث تصحیح املایی کلمات شده و در صورت لزوم می‌تواند کلماتی معنی‌دار را با توجه به کلمات پیشین و پسین در متن پیش‌بینی‌شده، بگنجاند. بهترین عملکرد در معماریB، شبکه عصبی با ۲۰۴۸ نورون در لایه‌های تمام متصل به همراه مدل زبانی می‌باشد که توانسته با خطای تشخیص ۱۰,۲۹ درصد گفتار مورد آزمایش را به متن تبدیل کند.

ساختار معماری B باوجود شبکه عصبی بازگشتیLSTM ، وابستگی بین ورودی‌ها را در نظر می‌گیرد، بنابراین، هر چه تعداد نورون‌های لایه‌های تمام متصل، بیشتر باشد، LSTMدرنهایت بهتر تغذیه شده و می‌تواند ویژگی‌های مناسب‌تری جهت تشخیص واج ارائه دهد. این معماری، مانند معماریA ، در صورت افزایش بیش‌ازحد نورون‌های تمام متصل به سمت بیش برازش حرکت می‌کند.

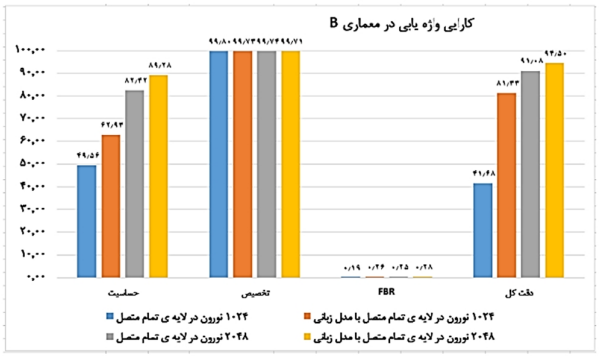
۴-۴- نتایج واژه‌یابی

در این بخش، به بررسی نتایج کارایی واژه‌یابی پرداخته خواهد شد. در نمودار شکل (۹)، معیارهای حساسیت، تخصیص، FBR و دقت کل در معماری A مشخص شده است. همان‌طور که در شکل (۹) مشخص است، معماری A با ۲۱۰۰ نورون در لایه تمام متصل، با دقت ۴۳,۰۴ درصد می‌تواند واژه‌های موجود در گفتار را پیدا کند، همچنین در صورت عدم وجود کلمه در گفتار، با دقت ۹۹ درصد، اعلام می‌کند که کلمه در متن پیدا نشده است و نرخ FBR این معماری در حدود ۱ درصد می‌باشد. درمجموع معماری A می‌تواند با دقت ۷۱,۴۶ درصد واژه‌های کاربر در گفتار را شناسایی کند و یا در صورت عدم وجود واژه در گفتار آن را اعلام کند.



شکل ۹: کارایی معماری Aدر واژه‌یابی بر روی Speech168H

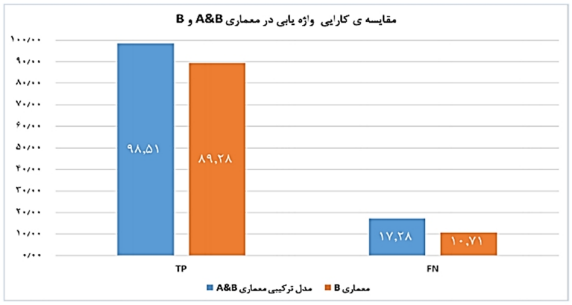
در نمودار شکل (۱۰)، معیارهای حساسیت، تخصیص، FBRو دقت کل در معماری B مشخص شده است.



شکل ۱۰: کارایی معماری Bدر واژه‌یابی بر روی Speech168H

همان‌طور که در نمودار شکل (۱۰) مشخص است، معماری B با ۲۰۴۸ نورون نسبت به ۱۰۲۴ نورون و استفاده از مدل زبانی درمجموع توانسته دقت جستجو را بالا ببرد، در این حالت با حساسیت ۸۹,۲۸ درصد توانسته واژه‌های موجود در گفتار را پیدا کند و همچنین در صورت عدم وجود واژه در گفتار، با دقت ۹۹,۷۱ درصد، پیغام عدم وجود را به کاربر اعلام کند. FBR معماری B مانند معماری A بسیار ناچیز بوده و این بدین معنی است که کلماتی که واقعاً در گفتار نیستند با درصد کمی به کاربر اعلام وجود می‌شوند. معماری B با دقت کل ۹۴,۵۰ درصد می‌تواند واژه‌های موجود در گفتار را تشخیص داده و یا واژه‌های ناموجود را تعیین کند. معماری B نسبت به معماری A حدوداً 23درصد دقت بهتری دارد اما ترکیب این دو معماری همان‌طور که شکل (۱۱) نشان داده شده، حدود ۹۸,۵۱ درصد دقت در TP داشته است درحالی‌که این مقدار برای معماری B۸۹,۲۸ می‌باشد. منتهی مدل ترکیبی معماری A و B در معیار FN دارای ۱۷,۲۸ درصد خطا می‌باشد، درحالی‌که این معیار برای معماری B ۱۰,۷۱ می‌باشد. استفاده از مدل ترکیبی جایی اهمیت پیدا می‌کند که کاربر

می‌خواهد با حساسیت بیشتری نتایج جستجوی کلیدواژه‌ها را مشاهده کند هرچند ممکن است، سیستم واژه‌هایی را نمایش دهد که در گفتار موجود نباشند.



شکل ۱۱: کارایی معماری B&A و B بر روی Speech168H

۴-۵- آزمایش بر روی مجموعه دادهMozilla7H

در جدول ۴ نتایج آزمایش روش پیشنهادی بر روی مجموعه داده Mozilla7H نشان داده شده است.

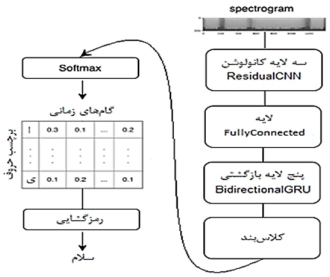
جدول ۴: نتایج آزمایش نرم‌افزار نهایی بر روی دیتاست Mozilla7H

تبدیل گفتار به متن	خطای تشخیص کلمه	۱۴,۳۷
واژه‌یابی	حساسیت	۷۶,۲۱
	تخصیص	۹۹,۸۰
	FBR	۰,۱۹
	دقت کل	۸۸,۰۱

با توجه به آزمایش‌های صورت گرفته روش پیشنهادی تبدیل گفتار به متن بر روی این مجموعه داده ۱۴,۳۷ درصد خطا دارد که با توجه ساعات مجموعه داده آموزشی شبکه عصبی قابل‌قبول می‌باشد. در بخش واژه‌یابی روش پیشنهادی با دقت ۷۶,۲۱ درصد می‌تواند واژه‌های موجود در گفتار را پیدا کند و با دقت ۹۹,۸۰ درصد نیز می‌تواند در صورت عدم وجود واژه آن را به کار اعلام کند. خطای FBR بر روی این مجموعه حدود ۰,۱۹ درصد بوده و این بدین معنی است که کلماتی که واقعاً در گفتار نیستند با درصد کمی به کاربر اعلام وجود می‌شوند. درمجموع روش پیشنهادی می‌تواند با دقت کل ۸۸,۰۱ درصد واژه‌های موجود در گفتار را تشخیص داده و یا واژه‌های ناموجود را تعیین کند.

۴-۶- مقایسه با روش‌های مشابه

در مقاله "سامانه شناسایی گفتار فارسی انتها به انتها با استفاده از پس‌پردازش دادگان پروژه موزیلا" با معماری شکل زیر، با ۱۰۰ ساعت داده موزیلا (۳۲۹۹ گوینده) و همچنین حدود ۵ ساعت داده برای توسعه و ۵ ساعت داده برای آزمون، به درصد خطای تشخیص کلمه ۴۶,۶ درصد رسیده است. [19]



شکل ۱۲: معماری سامانه شناسه گفتار فارسی

در مقاله "تشخیص عبارت‌های گفتاری برای اخبار فارسی صداوسیمای جمهوری اسلامی ایران" با مجموعه داده آموزشی فارس‌دات کوچک و بزرگ و دادگان اخبار فارسی صداوسیمای جمهوری اسلامی به درصد خطای تشخیص کلمه ۷,۹۳ و دقت واژه‌یابی ۸۳ درصد با مدل آموزش دیده با دادگان فارس‌دات کوچک بر روی همین دادگان و درصد خطای تشخیص کلمه ۲,۷۱ دقت واژه‌یابی ۹۲ درصد با مدل آموزش دیده با فارس دات بزرگ بر روی همین دادگان و درصد خطای تشخیص کلمه ۲۸,۲۳ درصد دقت واژه‌یابی ۷۹ درصد با مدل آموزش دیده با فارس دات بزرگ بر روی دادگان اخبار فارسی با استفاده از مدل SGMM رسید. [20]

در جدول (۵) نتایج مقایسه روش پیشنهادی با الگوریتم‌های موجود در زبان فارسی نشان داده شده است.

جدول ۵: مقایسه کارایی سامانه با محصول مشابه

سیستم	دادگان	مقدار داده آموزشی	خطای تبدیل گفتار به متن	دقت کل واژه‌یابی
روش پیشنهادی	Mozilla,Persian corpus و مختص همین پژوهش	۱۵۲	۱۴,۳۷	۸۸
[19]	Mozilla	۱۰۰	۴۶,۶	ندارد
[20]	فارس دات کوچک و بزرگ و دادگان اخبار فارسی	۷۷,۵	۲۸	۷۹

در رابطه با اهمیت پایگاه داده و موارد مشابه در قسمت قبل همان‌طور که در جدول (۵) مشاهده می‌شود روش پیشنهادی توانسته دقت قابل قبولی در مقایسه با کارهای مشابه داشته باشد. لذا علاوه بر مقایسه‌ای که برای اجزای مختلف روش پیشنهادی(تعداد نورون‌ها، نحوه به‌کارگیری معماری ها و ...) در قسمت‌های قبل در این قسمت نیز روش پیشنهادی با روش‌های مشابه مورد مقایسه قرار گرفت و نتایج قابل قبولی از خود نشان داد.

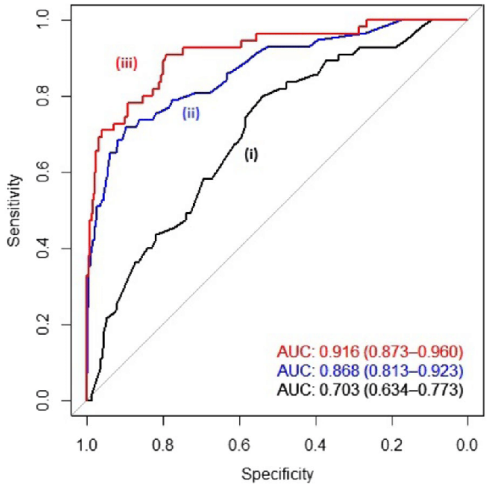
یکی دیگر از معیارهای ارزیابی الگوریتم‌های واژه‌یاب گفتار، نمودار ROC[1] می‌باشد. این معیار در قالب یک منحنی دوبعدی نمایش داده می‌شود که محور افقی آن نماینده تعداد اخطارهای اشتباه[2] بر کلمه کلیدی و محور عمودی آن نماینده تعداد تشخیص‌های درست[3] است.

[1] Receiver Operating Characteristic

[2] Specificity

[3] sensitivity

AUC متوسط سطح زیر منحنی ROC برای نرخ اخطار اشتباه مابین ۰ تا ۱ می‌باشد و هر چه به ۱ نزدیک‌تر باشد بیانگر عملکرد بهتر الگوریتم می‌باشد.



شکل ۱۳: نمودار ROC برای ارزیابی دقت چارچوب‌های تبدیل به متن

همان‌طور که در شکل (۱۳) نشان داده شده (i: روش [۱۹] ، ii: روش[۲۰] ، iii: روش پیشنهادی) روش پیشنهادی در مقایسه با روش های دیگر در نرخ اخطار اشتباه یکسان از دقت تشخیص بالاتری برخوردار بوده و متوسط سطح زیر منحنی ROC در روش پیشنهادی نسبت به روش‌های مورد مقایسه بیشتر می‌باشد.

۵- نتیجه‌گیری و پیشنهادها

به‌صورت خلاصه می‌توان از آزمایش‌های صورت گرفته، نتایج زیر را کسب کرد:

- دقت معماری B، از معماری A بیشتر بوده که دلیل اصلی برتری معماری Bنسبت به معماری A استفاده از شبکه عصبی بازگشتی و اضافه کردن مدل زبانی به عنوان احتمال پیشین به مدل آکوستیکی می‌باشد.
- استفاده از مدل ترکیبی معماری A و B موجب نمایش تعداد بیشتری از نتایج اعلام وجود واژه شده و هنگامی کاربر حساسیت زیادی بر روی پیدا کردن واژه موردنظر خود دارد انتخاب آن مفید می‌باشد.
- پارامترهای شبکه عصبی در صورت افزایش، می‌تواند به‌صورت متناسب افزایش پیدا کند و درنهایت، دقت شبکه عصبی را افزایش دهد، منتها افزایش غیرمتناسب با نمونه‌های آموزشی، موجب بیش برازش می‌شود.
- مدل زبانی می‌تواند متن‌های پیش‌بینی‌شده را غلط‌یابی کرده و متناسب با احتمال وجود جمعی کلمات، به یک متن قابل درک برای کاربر برسد؛ بنابراین اعمال این مدل، موجب افزایش دقت واژه‌یابی و تبدیل گفتار به متن می‌شود.

۵-۱- دستاوردها و نوآوری‌ها:

دستاوردها و نوآوری‌های حاصل از این پژوهش عبارت‌اند از:

۱ - واژه‌یابی با دقت ۸۸،۰۱ درصد:

روش پیشنهادی با دقت ۸۸ درصد می‌تواند کلیدواژه‌های موجود در گفتار را تشخیص دهد و همچنین با دقت ۹۹،۸۰ می‌تواند عدم وجود واژه‌ها را در گفتار پیغام دهد.

۲ - تبدیل گفتار به متن با خطای تشخیص کلمه ۱۴،۳۷ درصد:

یکی از دستاوردهای جانبی این پژوهش ساخت یک سیستم تبدیل گفتار به متن است. این سیستم با داشتن الگوریتم‌های مدل زبانی توانسته متن حاصل از فایل گفتار را اصلاح کرده و خروجی قابل‌قبولی ارائه دهد.

۳ - ساخت یک پایگاه داده گفتار ۱۷۵ ساعته:

یکی از محصولات مهم این پژوهش که در آینده بیشتر قابل استفاده خواهد بود ساخت پایگاه داده اختصاصی گفتار بود. در این پایگاه داده اصواتی با مدت‌زمان زیر 10 ثانیه از گویندگان مختلف وجود دارند که به ازای هرکدام از آن‌ها یک متن نوشته شده است. در این پژوهش به‌صورت اختصاصی ۴۲ ساعت گفتار، برچسب گزاری و همچنین از مجموعه داده موزیلا نیز ۱۲۹ ساعت و از مجموعه داده corpus Persian نیز ۳ ساعت، پالایش و بهینه‌سازی شد تا در مجموع ۱۷۵ ساعت فایل گفتار به همراه متن به وجود آید.

۵-۲- محدودیت‌ها و راه‌کارها

محققین این حوزه با توسعه روزافزون الگوریتم‌ها می‌توانند به پوشش انواع زبان‌ها و لهجه‌ها در زبان مردم ایران‌زمین بپردازند که در این راستا پیشنهادها زیر عنوان می‌گردد.

الف- مهم‌ترین عامل در کارایی سامانه واژه‌یاب و تبدیل گفتار به متن پیشنهادی، افزایش منابع محاسباتی و داده‌ای می‌باشد. این گسترش در سه بُعد باید صورت پذیرد. اولین بُعد، متناسب‌سازی تعداد گفتار محاوره و رسمی می‌باشد. با توجه به اینکه زبان فارسی محاوره بخش اعظمی از مکالمات روزمره را در برمی‌گیرد و نوع گویش و نگارش آن متفاوت از زبان معیار می‌باشد، می‌بایست تعداد اصوات محاوره و رسمی برابر باشد تا در هنگام آزمایش، کارایی بالا رود. دومین بُعد، متناسب‌سازی بین جنسیت مرد و زن است که باید از هر نوع، به‌صورتی یکسان، گسترش مجموعه داده صورت پذیرد. سومین بُعد، متناسب‌سازی بین لهجه مختلف زبانی است تا در هنگام آزمایش، تفاوت لهجه‌ها موجب کاهش دقت نشود.

ب- مسئله مهم دیگر گسترش مجموعه داده جهت تولید مدل زبانی می‌باشد، هر چه قدر متن‌های منحصربه‌فرد در ساخت مدل زبانی بیشتر باشد، دامنه زبانی فراگیرتر شده و موجب افزایش دقت می‌شود.

ج- استفاده از شبکه‌های از پیش آموزش دیده، در زبان‌های دیگر. با توجه به فراوانی مجموعه داده در زبان‌های دیگر و بهینه‌سازی وزن‌های شبکه‌های عصبی حاصل از آموزش این مجموعه داده‌ها، می‌توان از این شبکه‌ها جهت آموزش هر چه بهتر شبکه عصبی در زبان فارسی استفاده نمود. استفاده از این ایده موجب همگرا شدن سریع‌تر شبکه عصبی شده و به نحوی کمبود منابع داده‌ای تا حدی جبران می‌شود.

مراجع

[1] Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, Moore G, Odell J, Ollason D, Povey D, Valtchev V. The HTK book. Cambridge university engineering department. 2002 Dec; 3(175):12.

[2] Han W, Zhang Z, Zhang Y, Yu J, Chiu CC, Qin J, Gulati A, Pang R, Wu Y. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. arXiv preprint arXiv:2005.03191. 2020 May 7.

[3] Zhang Y, Qin J, Park DS, Han W, Chiu CC, Pang R, Le QV, Wu Y. Pushing the limits of semi-supervised learning for automatic speech recognition. arXiv preprint arXiv:2010.10504. 2020 Oct 20.

[4] Sameti H, Veisi H, Bahrani M, Babaali B, Hosseinzadeh K. A large vocabulary continuous speech recognition system for Persian language. EURASIP Journal on Audio, Speech, and Music Processing. 2011 Dec;2011:1-2.

[5] Veisi H, Haji Mani A. Persian speech recognition using deep learning. International Journal of Speech Technology. 2020 Dec;23:893-905.

[6] م. مرادنژاد و ش. طبیبیان، "یک رویکرد پایانه به پایانه برای تشخیص موثر فرامین صوتی فارسی در خودروی هوشمند،" در بیست و هشتمین کنفرانس بین المللی کامپیوتر، انجمن کامپیوتر ایران، تهران، ۱۴۰۱.

[7] Fendji, J. L. K. E., Tala, D. C., Yenke, B. O., & Atemkeng, M. (2022). Automatic speech recognition using limited vocabulary: A survey. Applied Artificial Intelligence, 36(1), 2095039.

[8] Mohanty, P., & Nayak, A. K. (2022). CNN based keyword spotting: an application for context based voiced Odia words. International Journal of Information Technology, 14(7), 3647-3658.

[9] s. mehra and s. susan, "deep fusion framwork for speech command recognation using acoustic and linguistic features," multimedia tools and applications, vol. 23, pp. 1-25, 2023 Mar.

[10] M. BİLGİLİ, "Application of long short-term memory (LSTM) neural network," Turkish Journal of Electrical Engineering and Computer Sciences, 2022.

[11] Weninger F, Erdogan H, Watanabe S, Vincent E, Le Roux J, Hershey JR, Schuller B. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. InLatent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings 12 2015 (pp. 91-99). Springer International Publishing.

[12] Kuleshov V, Enam SZ, Ermon S. Audio super resolution using neural networks. arXiv preprint arXiv:1708.00853. 2017 Aug 2.

[13] Alumäe T, Karakos D, Hartmann W, Hsiao R, Zhang L, Nguyen L, Tsakalidis S, Schwartz R. The 2016 BBN Georgian telephone speech keyword spotting system. In2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017 Mar 5 (pp. 5755-5759). IEEE.

[14] Kaur A, Sachdeva R, Singh A. Latest trends in deep learning for automatic speech recognition system. InArtificial Intelligence and Speech Technology: Third International Conference, AIST 2021, Delhi, India, November 12–13, 2021, Revised Selected Papers 2022 Jan 29 (pp. 62-72). Cham: Springer International Publishing.

[15] Tabibian S, Akbari A, Nasersharif B. Discriminative keyword spotting using triphones information and N-best search. Information Sciences. 2018 Jan 1;423:157-71.

[16] Arik S, Kliegl M, Child R, Hestness J, Gibiansky A, Fougner C, Prenger R, Coates A, inventors; Baidu USA LLC, assignee. Convolutional recurrent neural networks for small-footprint keyword spotting. United States patent US 10,540,961. 2020 Jan 21.

[17] Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, Ng AY. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567. 2014 Dec 17.

[18] V. Yathish, "Loss Functions and Their Use In Neural Networks," 2022. [Online].

[19] س.زارعی، ی.شکفته، "سامانه شناسایی گفتار فارسی انتها به انتها با استفاده از پس پردازش دادگان پروژه موزیلا "،پنجمین کنفرانس ملی فناوری در مهندسی برق و کامپیوتر، دانشگاه گلستان، گلستان، ایران، ۱۴۰۰.

[20] Veisi H, Ghoreishi SA, Bastanfard A. "Spoken term detection for Persian news of Islamic Republic of Iran broadcasting". Signal and Data Processing. 2021 Feb 10;17(4):67-88.