مقاله پژوهشی

بازشناسی مقاوم به نویز ارقام مشابه فارسی مبتنی بر شبکه LSTM و ویژگی‌های طیفی گفتار

شیما طبیبیان

چکیده: یکی از چالش‌های بازشناسی ارقام مجزای فارسی، مشابهت تلفظ برخی از ارقام مانند "صفر و سه"، "نه و دو" و "پنج، هفت و هشت" می‌باشد. این چالش منجر به بازشناسی یک رقم به جای رقم مشابه شده و دقت بازشناسی را کاهش می‌دهد. در این مقاله، یک راهکار ترکیبی مبتنی بر حافظه کوتاه‌مدت ماندگار (LSTM) و مدل مخفی مارکف (HMM) برای رفع چالش مذکور ارائه شده که نرخ بازشناسی ارقام فارسی مبتنی بر HMM را به طور متوسط ۲٪ و در بهترین حالت ۸٪ بهبود داده است. با توجه به تشدید چالش بازشناسی ارقام مشابه فارسی در شرایط نویزی، در ادامه کار مقاوم‌سازی بازشناسی ارقام مشابه فارسی مورد توجه قرار گرفت. به منظور افزایش مقاومت بازشناس مبتنی بر LSTM، از ویژگی‌های مقاوم به نویز مستخرج از طیف گفتار مانند آنتروپی طیفی، درجه از هم پاشی، فرکانس نیمساز، همواری طیفی، فرمانت اول و نرخ گذار از صفر مبتنی بر تابع همبستگی استفاده گردید. استفاده از این ویژگی‌ها، ضمن کاهش تعداد ویژگی‌ها برای بازشناسی ارقام مشابه فارسی از ۳۹ ضریب به حداکثر ٤ و حداقل ۱ ضریب، به طور متوسط به ترتیب بهبود ۱۰، ۱۳، ۱۵ و ۱۳ درصدی مقاومت بازشناس ارقام مشابه را در شرایط متنوع نویزی (۳۰ حالت مختلف حاصل از پنج نوع نویز سفید، صورتی، همهمه، کارخانه و ماشین و شش نسبت سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل) در مقایسه با بازشناس‌های مبتنی بر HMM، LSTM، شبکه باور عمیق با ویژگی‌های مل کپستروم و شبکه عصبی کانولوشنی با ویژگی‌های مل اسپکتوگرام به همراه دارد.

***کلیدواژه:* بازشناسی ارقام مجزا، زبان فارسی، مشابهت تلفظ ارقام، مدل مخفی مارکف، حافظه کوتاه‌مدت ماندگار، مقاوم‌سازی.**

۱- مقدمه

امروزه با پیشرفت تکنولوژی، شاهد افزایش چشم‌گیر استفاده از گفتار در عرصه‌های مختلفی از زندگی انسان‌ها هستیم. از سوی دیگر، ارقام نقش مهمی در زندگی انسان‌ها بازی می‌کنند. رمز کارت‌های اعتباری، شماره‌های تلفن، انجام عملیات بانکی در تلفن‌بانک‌ها، پرداخت برخط قبوض و سیستم‌های رزرواسیون نمونه‌هایی از حضور ارقام در زندگی روزانه انسان‌ها هستند. در تمام کاربردهای مذکور، ارقام از طریق لمس یا فشردن دکمه مورد نظر ارسال می‌شوند. استفاده از گفتار به جای لمس یا فشردن دکمه، علاوه بر محبوبیت و سادگی بیشتر، این امکان را فراهم می‌کند که کاربران بدون از دست دادن تمرکز، هم‌زمان به کار دیگری هم مشغول باشند. این قبیل ارتباطات مبتنی بر گفتار، نیازمند استفاده از سیستم‌های کارای بازشناسی و سنتز گفتار می‌باشد.

در زمینه بازشناسی اعداد فارسی، کارهای تحقیقاتی متعددی انجام شده‌اند. در سال ۱۳۷۱ فکری و همکاران [۱] با استفاده از مدل مخفی مارکف[1] (HMM) یک سیستم بازشناس ارقام گسسته فارسی را در محیط بدون نویز و میکروفنی با دقت بازشناسی ۹۶٫۰۹ درصد برای دادگان آزمون پیاده‌سازی کردند. در سال ۱۳۷۸ بابابیک [۲] با استفاده از تلفیق HMM و شبکه عصبی به دقت ۹۸٫۵ درصد در بازشناسی ارقام گسسته میکروفنی رسید. در تحقیق دیگری در سال ۱۳۷۸، بابایی‌زاده و همکاران [۳] یک مدل ترکیبی شبکه عصبی (MLP)[2] و HMM گسسته پیاده‌سازی کردند که بر روی ارقام صفر تا نه و کلمات بله و خیر در محیط میکروفنی، دارای دقت بازشناسی ۹۷٫۹ درصد بود. در سال ۱۳۷۷ رستم‌زاده و همکاران [۴] با استفاده از HMM پیوسته در محیط میکروفنی به دقت ۹۹٫۷۵ درصد دست یافتند. در سال ۱۳۷۸، نجاری و همکاران [۵] با استفاده از مدل پیشگوی شبکه عصبی بر روی دادگان تلفنی شامل اعداد صفر تا نه به دقت ۸۱ درصد رسیدند. در سال ۱۳۷۹، صیادان و همکاران [۶] یک HMM تک‌حالته با هشت مخلوط گوسی را با یک HMM پیوسته با پنج حالت و ۱۶ مخلوط گوسی در هر حالت مقایسه کردند که دارای نرخ بازشناسی ۱۰۰ درصد در مقایسه با نرخ بازشناسی ۹۴٫۱۷ درصد برای مدل مخفی پیوسته بود. در سال ۱۳۷۹ سیستمی توسط اکبری و همکاران [۷] پیاده‌سازی شد که دقت بازشناسی سیستم مذکور با استفاده از ضرایب مل کپستروم، انرژی و مشتق این ضرایب، بر روی پایگاهی متشکل از اعداد دورقمی فارسی در حالت گسسته، ۸۹٫۷ درصد و برای حالت پیوسته، ۹۸٫۷ درصد بود. در سال ۱۳۸۲ همایون‌پور و همکاران [۸] به بررسی روش‌های مبتنی بر HMM، شبکه عصبی MLP و ترکیب آنها برای بازشناسی ارقام فارسی، که به صورت گسسته، متصل و پیوسته از طریق تلفن بیان شده‌اند، پرداختند. در نهایت بهترین نرخ دقت برای بازشناسی ارقام گسسته و متصل با مدل‌های مبتنی بر کلمه بر روی پایگاه دادگان FARSDIGITS۱ [۹] به ترتیب برابر با ۹۹٫۱ درصد و ۸۳٫۷ درصد برای سیستم مبتنی بر HMM بوده است.

در سال‌های اخیر، بازشناسی گفتار مبتنی بر شبکه‌های عصبی عمیق از اهمیت بسزایی برخوردار شده است. شبکه‌های عصبی که برای بازشناسی گفتار مورد استفاده قرار می‌گیرند، باید قادر به مدل‌سازی طبیعت تغییرپذیر با زمان گفتار باشند. از جمله می‌توان به شبکه‌های عصبی بازگشتی

این مقاله در تاریخ ۸ فروردین ماه ۱۳۹۹ دریافت و در تاریخ ۸ بهمن ماه ۱۳۹۹ بازنگری شد.

شیما طبیبیان (نویسنده مسئول)، پژوهشکده فضای مجازی، دانشگاه شهید بهشتی، تهران، ایران، (email: sh_tabibian@sbu.ac.ir).

1. Hidden Markov Model
2. Multi Layer Perceptron

(RNN)[1] و مدل‌های مبتنی بر آنها مانند شبکه مبتنی بر حافظه کوتاه‌مدت ماندگار (LSTM)[2] و شبکه‌های عصبی کانولوشنی اشاره کرد [۱۰] تا [۱۶]. در [۱۷] یک روش بازشناسی ارقام مجزا مبتنی بر شبکه LSTM ارائه شده است. دقت این روش بر روی دادگان اعداد تک‌رقمی انگلیسی تهیه‌شده در محیط آزمایشگاهی ۹۹٫۹۶ درصد بوده است. در حوزه روش‌های یادگیری عمیق، شبکه‌های عصبی عمیق (DNN)[3] و شبکه‌های باور عمیق (DBN)[4] نقش مهمی در بهبود روش‌های بازشناسی گفتار داشته‌اند. در [۱۸] یک سیستم بازشناس ارقام مجزا مبتنی بر DNN ارائه شده است. در این روش از DBN برای مقداردهی اولیه DNN استفاده شده که دقت بازشناسی بر روی دادگان انگلیسی TIDIGIT ۸۶٫۰۶ درصد بوده است. در یکی از پژوهش‌های اخیر [۱۶] به منظور بازشناسی ارقام مجزای زبان پشتو از شبکه‌های کانولوشنی (CNN)[5] عمیق بهره گرفته شده است. معماری شبکه مذکور از حداکثر چهار لایه کانولوشنی عمیق و به دنبال آن لایه‌های ReLU و تجمیع بیشینه[6] بهره گرفته است. میانگین دقت حاصل بر روی دادگان حاوی ارقام پشتو برابر با ۸۴٫۱۷ درصد بوده که در مقایسه با سیستم بازشناس ارقام مجزا مبتنی بر DNN ۷٫۳۲ درصد عملکرد بهتری داشته است.

چالش اصلی بازشناسی ارقام مجزای فارسی، مشابهت تلفظ برخی از ارقام به یکدیگر و افت دقت بازشناسی به دلیل رخداد خطای جایگزینی بالا در این موارد می‌باشد. مشابهت تلفظ ارقامی مانند "سه و صفر"، "نه و دو" و "پنج، هفت و هشت" بیشترین تأثیر را در افت دقت بازشناسی مجزای ارقام فارسی دارد. با توجه به آن که دقت سیستم‌های بازشناس گفتار در شرایط نویزی تنزل پیدا می‌کند، چالش مذکور در شرایط نویزی پراهمیت‌تر می‌شود.

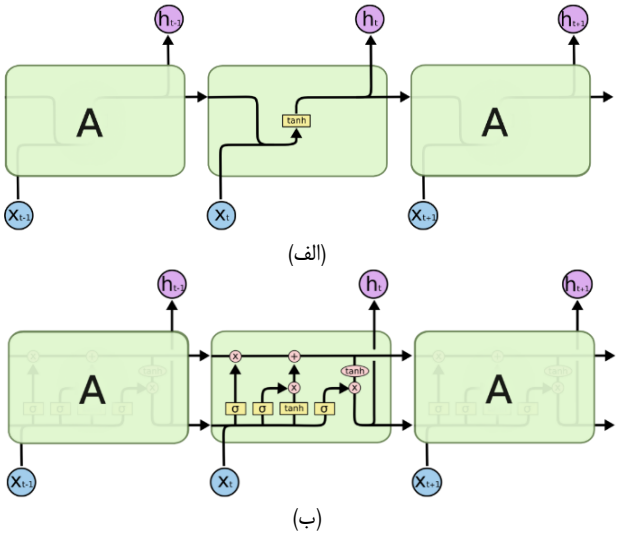
روش‌های مقاوم‌سازی بازشناسی گفتار به سه رویکرد مختلف دسته‌بندی می‌شوند. دسته اول از رویکردها، با استفاده از تکنیک‌های بهسازی گفتار در مرحله پیش‌پردازش سعی در جبران اثرات مخرب نویز و بهبود کیفیت سیگنال گفتار دارند [۱۹]. دسته دوم رویکردها، اقدام به مقاوم‌سازی بازشناسی گفتار در سطح مدل می‌کنند. به این ترتیب که یا از ابتدای کار، سیستم بازشناس گفتار بر روی دادگان نویزی آموزش داده می‌شود و یا مدل‌های آموزش داده شده در شرایط تمیز، با شرایط واقعی آزمون (که نویز هم می‌تواند بخشی از این شرایط باشد) تطبیق داده می‌شوند [۲۰]. طبیعی است که در حالت دوم، نیازی به حجم زیادی از دادگان نویزی نبوده و بخش کوچکی از آن برای تطبیق مدل کافی است. دسته سوم رویکردها مقاوم‌سازی بازشناسی گفتار را در سطح ویژگی انجام می‌دهند. روش‌های مقاوم‌سازی در سطح ویژگی به دو دسته تقسیم می‌شوند: در دسته اول، مانند روش‌های هنجارسازی میانگین و ضرایب کپستروم [۲۱]، هدف، بهبود ویژگی‌ها و حذف اعوجاج و آثار مخرب نویز از ویژگی‌ها می‌باشد. در دسته دوم، هدف استخراج ویژگی‌هایی (ویژگی‌های مقاوم به نویز) است که ذاتاً کمتر تحت تأثیر اعوجاج‌های محیطی قرار بگیرند [۲۲]. از جدیدترین روش‌ها برای استخراج ویژگی‌های مقاوم به نویز، استفاده از شبکه‌های عصبی عمیق مانند شبکه‌های باور عمیق و شبکه‌های خودرمزگذار عمیق، شبکه‌های عصبی کانولوشنی و ... می‌باشد [۲۳] و [۲۴]. همچنین استخراج ویژگی‌های ساده‌ای مانند فرمانت‌ها، انرژی و آنتروپی طیفی از طیف سیگنال گفتار در کنار سادگی پیاده‌سازی، عملکرد بازشناسی گفتار را در شرایط نویزی تا حدود زیادی بهبود می‌دهد [۲۵]. علاوه بر رویکردهای مطرح برای مقاوم‌سازی بازشناسی گفتار که فارغ از نوع روش بازشناسی گفتار به کار گرفته می‌شوند، برخی از روش‌های بازشناسی گفتار، به دلیل ماهیتشان، از مقاومت بیشتری در شرایط نویزی برخوردار هستند. از این میان، به عنوان نمونه در حوزه بازشناسی ارقام مجزا می‌توان به رویکردهای مبتنی بر LSTM و ماشین بردار پشتیبان (SVM)[7] اشاره نمود [۲۶] و [۲۷].

در سال ۱۳۸۷، حجازی و همکاران [۲۶] به طور خاص بر روی حل مشکل شناسایی ارقام مجزای فارسی با تلفظ مشابه (مانند "سه و صفر"، "دو و نه" و "هفت و هشت") متمرکز شدند. آنها از یک سیستم ترکیبی متشکل از HMM و SVM استفاده کردند. نتایج ارزیابی‌ها بر روی دادگان ضبط‌شده آزمایشگاهی حاکی از بهبود دقت بازشناسی ارقام مجزا در مقایسه با رویکرد مبتنی بر HMM به میزان ۲۱٪ در شرایط تمیز و ۲۴٪ در شرایط نویزی می‌باشد.

یکی از چالش‌های استفاده از ماشین بردار پشتیبان و برخی از رویکردهای مبتنی بر شبکه‌های عصبی در بازشناسی گفتار در مقایسه با مدل‌های مخفی مارکف، در نحوه ارائه خروجی می‌باشد. هر دو گروه برای پردازش و استخراج ویژگی از گفتار نیازمند تقطیع سیگنال گفتار هستند. خروجی حاصل از رویکرد مبتنی بر HMM حاوی واحدهای آوایی شناسایی‌شده و مکان رخداد آنها و خروجی حاصل از رویکردهای مبتنی بر شبکه عصبی یا ماشین بردار پشتیبان حاوی واحدهای آوایی شناسایی‌شده در همان سطح قاب می‌باشد. تبدیل خروجی در سطح قاب به خروجی در سطح واحد آوایی نیازمند تشخیص مرز شروع و پایان واحدهای آوایی است. این کار در رویکرد مبتنی بر HMM به دلیل ساختار و معماری خاص آن به راحتی انجام می‌گیرد، در حالی که در دو رویکرد دیگر تبدیل به یک چالش می‌شود که باید به طریق مناسب حل شود. به عنوان مثال در [۲۸] به منظور بهبود بازشناسی ارقام فارسی در دسته‌بندهای مبتنی بر قاب (SVM و DBN)، استخراج نتایج در سطح کلمه از نتایجِ در سطح قاب و اصلاح مرزبندی نواحی سکوت و گفتار، الگوریتمی ارائه شده که با بهره‌گیری از قوانین ساده، خطای در سطح قاب را به طور قابل توجهی کاهش می‌دهد.

از میان روش‌های مبتنی بر شبکه‌های عصبی، LSTM با توجه به ساختار و طبیعتش در به خاطر سپاری وابستگی‌های بلندمدت، اگرچه ویژگی‌ها را در سطح قاب دریافت می‌کند، اما این قابلیت را دارد که خروجی را در سطح واحد آوایی ارائه دهد. در [۲۹]، یک راهکار ترکیبی مبتنی بر HMM و LSTM برای بهبود بازشناسی ارقام مشابه فارسی (به طور خاص "صفر و سه" و "دو و نه") ارائه شده است. بازشناس ارقام مجزای مبتنی بر HMM عبارت گفتار ورودی را دریافت کرده و در خروجی، ارقام و سکوت‌های بازشناسی شده به همراه مکان زمانی رخداد آنها در عبارت گفتار ورودی را تحویل می‌دهد. چنانچه رقم بازشناسی شده "سه"، "صفر"، "دو" و "نه" بود، برای بازشناسی دقیق‌تر، آن بخش از عبارت گفتار ورودی که حاوی این رقم می‌باشد، به دسته‌بند دودویی مبتنی بر LSTM متناظرش داده شده و نتیجه بازشناسی دسته‌بند دودویی به عنوان تشخیص نهایی لحاظ می‌شود. رویکرد ارائه‌شده، نرخ بازشناسی ارقام مشابه فارسی مبتنی بر HMM را به طور متوسط حدود ۳٫۷۵ درصد

1. Recurrent Neural Network
2. Long Short Term Memory
3. Deep Neural Network
4. Deep Belief Network
5. Convolutional Neural Network
6. Max Pooling
7. Support Vector Machine



شکل ۱: ساختار بازگشتی، (الف) شبکه RNN و (ب) شبکه LSTM [۳۱].

بهبود داده است. لیکن، ایده ارائه‌شده در [۲۹] تمام حالت‌های رخداد مشابهت در تلفظ ارقام فارسی را لحاظ نکرده است. همچنین از آنجا که مقاوم‌سازی ایده‌های ارائه‌شده در شرایط نویزی هدف کار مذکور نبوده است، دسته‌بندهای مبتنی بر LSTM از ضرایب مل کپستروم در مرحله استخراج ویژگی بهره می‌گیرند که ویژگی مقاومی در شرایط نویزی محسوب نمی‌شود. به علاوه ارزیابی‌های این کار بدون بهره‌گیری از اعتبارسنجی متقابل k–لایه[1] انجام شده است.

با توجه به عملکرد مطلوب رویکرد مبتنی بر ترکیب HMM و LSTM این ایده به ذهن رسید که از ابتدا LSTM را اعمال نموده و با حذف HMM، سرعت بازشناسی را بالاتر برد. این ایده به همراه ارزیابی مقاومت بازشناس ارقام مجزای فارسی مبتنی بر LSTM در نسبت‌های سیگنال به نویز ورودی پایین (۵– تا ۱۰ دسی‌بل) در [۳۰] ارائه شده است. در شرایط تمیز چنانچه انتظار می‌رفت، بهبود دقت بازشناس ارقام مجزای مبتنی بر LSTM نسبت به HMM حاصل شد. لیکن، این میزان بهبود به اندازه زمانی که در رویکرد ترکیبی ارائه‌شده در [۲۹] با استفاده از دسته‌بندهای مبتنی بر LSTM نتایج بازشناسی HMM اصلاح می‌شد، نبود. زیرا در حالت دوم، تمرکز دسته‌بند دودویی مبتنی بر LSTM، بر تفکیک دو رقم مشابه از هم و نه دسته‌بندی تمام ارقام مجزای فارسی بود. در شرایط نویزی، اگرچه نتایج ارزیابی‌ها حاکی از عملکرد بهتر بازشناس ارقام مجزای مبتنی بر LSTM در مقایسه با HMM بود، کماکان مشابهت تلفظ ارقام فارسی بیشترین دلیل افت دقت بازشناس مبتنی بر LSTM بود. به علاوه، اگرچه بازشناس مبتنی بر LSTM در شرایط نویزی عملکرد بهتری نسبت به HMM داشت، به دلیل استفاده از ضرایب مل کپستروم در مرحله استخراج ویژگی و عملکرد ضعیف این ضرایب در شرایط نویزی، از دقت بازشناسی مطلوبی (حدود ۵۷٪) در شرایط نویزی برخوردار نبود. بنابراین به منظور تفکیک مقاوم به نویز ارقام مشابه نیاز به استفاده از ویژگی‌هایی بود که بتوانند ضمن ایجاد بیشترین تمایز بین ارقام مشابه، در شرایط نویزی نیز دستخوش تغییرات زیادی نشوند. طبیعی است که اگرچه ممکن است این ویژگی‌ها برای دسته‌بندی ارقام مشابه از هم، عملکرد بسیار قابل قبولی داشته باشند، ویژگی‌های مناسبی برای تفکیک تمام ارقام مجزای فارسی از هم محسوب نمی‌شوند. بنابراین لازم است از یک رویکرد ترکیبی استفاده شود که در مرحله اول با بهره‌گیری از ویژگی‌های رایج حوزه بازشناسی گفتار دسته‌بندی ارقام مجزای فارسی را انجام دهد. سپس در مرحله دوم، ارقام مشکوک به خطای جایگزینی برای دسته‌بندی دقیق‌تر به دسته‌بند دیگری داده می‌شوند که از ویژگی‌های خاص‌تری برای تمایز میان ارقام مشابه استفاده می‌کند. دسته‌بند اول می‌تواند مبتنی بر LSTM، HMM یا هر رویکرد دیگری آموزش یابد. لیکن به دلیل سرعت بهتر روش‌های مبتنی بر HMM، مدل‌سازی مناسب ماهیت تغییرپذیر با زمان گفتار توسط آنها و دقت مطلوبشان در بازشناسی کلمات مجزا با اندازه دادگان کوچک، در کار حاضر، مشابه با کار ارائه‌شده در [۲۹]، دسته‌بند اول مبتنی بر HMM آموزش یافته است. به منظور تعمیم و تکمیل کار ارائه‌شده در [۲۹]، در این مقاله، یک راهکار ترکیبی مبتنی بر HMM و LSTM ارائه شده که علاوه بر جبران نقایص مذکور در ارزیابی‌های انجام‌شده در [۲۹]، دارای نوآوری‌های زیر می‌باشد:

۱) بهبود دقت بازشناسی ارقام فارسی در تمام حالت‌های رایج مشابهت تلفظ ارقام فارسی شامل "صفر و سه"، "دو و نه" و "پنج، هفت و هشت"

۲) مقاوم‌سازی بازشناسی ارقام مشابه فارسی با استفاده از ویژگی‌های مقاوم به نویز مستخرج از طیف سیگنال گفتار به جای ویژگی‌های رایج مل کپستروم

۳) کاهش ابعاد بردار ویژگی برای بازشناسی ارقام مشابه فارسی از ۳۹ ضریب مل کپستروم به چهار، سه و یک ویژگی مستخرج از طیف گفتار به ترتیب برای تمایز "صفر از سه"، "پنج، هفت و هشت" و "دو از نه"

این مقاله به شکل ذیل ساختاردهی شده است. در بخش دوم به معرفی شبکه بازگشتی LSTM خواهیم پرداخت. رویکرد پایه مبتنی بر HMM و رویکرد ترکیبی پیشنهادی در بخش‌های سوم و چهارم ارائه خواهند شد. سپس در بخش پنجم به معرفی و تحلیل ویژگی‌های مقاوم به نویز مستخرج از طیف گفتار خواهیم پرداخت. دادگان مورد استفاده برای آموزش و آزمون مدل‌ها در بخش ششم معرفی می‌شود. در بخش هفتم به ارائه و تفسیر نتایج آزمایش‌ها خواهیم پرداخت و در انتها، مقاله در بخش هشتم جمع‌بندی می‌شود.

۲- شبکه بازگشتی مبتنی بر LSTM

شبکه مبتنی بر حافظه کوتاه‌مدت ماندگار یا LSTM یکی از انواع شبکه‌های عصبی بازگشتی (RNN) می‌باشد که توانایی یادگیری وابستگی‌های بلندمدت را دارد. شبکه LSTM برای اولین بار توسط هاکریتر و اشمیدبر در سال ۱۹۹۷ معرفی شد [۳۱]. در حقیقت هدف از طراحی شبکه LSTM، حل مشکل به یادسپاری وابستگی‌های بلندمدت در RNNها بود. RNNها، ساختاری بسیار مشابه با شبکه‌های چندلایه پرسپترون (MLP) دارند، با این تفاوت که نورون‌های لایه مخفی علاوه بر یال‌های رو به جلو، یک یال هم به صورت بازگشتی و با احتساب یک زمان تأخیر، از خودشان به خودشان دارند. چنین ساختاری به یادسپاری وابستگی‌های کوتاه‌مدت[2] را تضمین می‌کند، لیکن امکان یادگیری وابستگی‌های مربوط به گذشته‌های دور[3] را ندارد. برای رفع این مشکل نورون‌های مخفی با یک بلاک حافظه با ساختار پیچیده‌تری، جایگزین شد و منجر به ظهور شبکه‌های LSTM گردید. شکل ۱ تفاوت ساختاری RNN و LSTM را نشان می‌دهد.

1. K-Fold Cross Validation

2. Short Term

3. Long Term

چنانچه شکل ۱ نشان می‌دهد، در RNN (شکل ۱- الف)، ماژول‌های تکرارشونده ساختار بسیار ساده‌ای تنها شامل یک لایه تانژانت هایپربولیک دارند. ورودی هر شبکه، شامل خروجی ماژول قبلی و ورودی جدید می‌باشد. خروجی هر ماژول حاصل اعمال تابع تانژانت هایپربولیک بر ترکیب وزن‌دار دو ورودی مذکور می‌باشد. در شبکه بازگشتی LSTM (شکل ۱- ب)، ساختار ماژول تکرارشونده به منظور به خاطرسپاری وابستگی‌های طولانی‌مدت، پیچیدگی بیشتری دارد. ورودی هر ماژول شامل ورودی جدید و دو خروجی از ماژول قبلی است. همچنین هر ماژول دارای دو خروجی می‌باشد. برای تولید هر یک از این دو خروجی از اعمال توابع فعالیت سیگموئید و تانژانت هایپربولیک بر خروجی‌های ماژول قبلی و ورودی جدید و ترکیب‌های وزن‌دار آنها استفاده شده است. جزئیات مربوط به این بلاک حافظه، اجزای تشکیل‌دهنده و روابط مربوطش در [۳۱] و [۳۲] ارائه شده‌اند.

۳- بازشناسی ارقام مجزای فارسی مبتنی بر HMM

برای آموزش سیستم بازشناس مبتنی بر HMM از جعبه ابزار HMMToolkit (HTK) [۳۳] و دادگان (CPHPD)[1] [۳۴] استفاده گردیده و بازشناس مذکور از سه بخش استخراج ویژگی، آموزش و آزمون مدل تشکیل شده است.

در بخش استخراج ویژگی، ابتدا گفتار ورودی با کمک پنجره همینگ به قاب‌های ۲۵ میلی‌ثانیه با همپوشانی ۴۰ درصد تقطیع می‌شود. سپس از هر قاب گفتار، ویژگی‌های مل کپستروم استخراج شده و با استفاده از روش نرمال‌سازی (CMVN)[2]، نرمال‌سازی می‌شوند [۲۱]. تعداد مدل‌های مخفی مارکف، به ازای ارقام صفر تا نه و سکوت، ۱۱ مدل می‌باشد. هر مدل، دارای شش وضعیت اصلی و دو وضعیت ورودی و خروجی است. بر اساس نتایج ارائه‌شده در [۳۴] بهترین انتخاب برای تعداد مخلوط‌های گوسی در هر وضعیت، از میان اعداد ۴، ۸، ۱۶ و ۳۲، چهار و بهترین انتخاب برای ویژگی‌های مل کپستروم از میان چهار دسته ۱۲ ضریب مل کپستروم و یک ضریب انرژی (در مجموع ۱۳ ضریب)، ۱۲ ضریب مل کپستروم و یک ضریب انرژی به همراه مشتقات اولشان (در مجموع ۲۶ ضریب)، ۱۲ ضریب مل کپستروم و یک ضریب انرژی به همراه مشتقات اول و دومشان (در مجموع ۳۹ ضریب) و ۱۲ ضریب مل کپستروم و یک ضریب انرژی به همراه مشتقات اول، دوم و سومشان (در مجموع ۵۲ ضریب)، دسته سوم می‌باشد.

خروجی HMM به ازای هر سیگنال گفتار ورودی شامل سکوت ابتدا و انتهای فایل و برچسب مربوط به رقم اداشده می‌باشد. در صورتی که رقم بازشناسی‌شده یکی از ارقام دارای تلفظ مشابه در زبان فارسی باشد، آن قسمت از سیگنال گفتار ورودی به بررسی دقیق‌تری نیاز دارد. برای انجام این ایده، در بخش بعد رویکردی ترکیبی مبتنی بر HMM و LSTM ارائه شده است.

٤- رویکرد ترکیبی پیشنهادی مبتنی بر LSTM و HMM برای رفع چالش مشابهت تلفظ ارقام فارسی

با توجه به این که بیشترین خطاهای جایگزینی مربوط به بازشناسی ارقام "سه و صفر"، "دو و نه" و "پنج، هفت و هشت" می‌باشد، دو دسته‌بند دودویی برای بازشناسی "سه و صفر" و "دو و نه" و یک دسته‌بند سه‌کلاسه برای بازشناسی "پنج، هفت و هشت" مبتنی بر LSTM آموزش داده می‌شوند. بازشناس ارقام مجزای مبتنی بر HMM عبارت گفتار ورودی را دریافت کرده و در خروجی، ارقام و سکوت‌های بازشناسی‌شده به همراه مکان رخداد آنها در عبارت گفتار ورودی را تحویل می‌دهد. چنانچه رقم بازشناسی‌شده "سه"، "صفر"، "دو"، "نه"، "پنج"، "هفت" و "هشت" بود، آن بخش از عبارت گفتار ورودی حاوی این رقم به دسته‌بند دودویی یا سه‌کلاسه مبتنی بر LSTM متناظرش داده شده و نتیجه بازشناسی دسته‌بند به عنوان تشخیص نهایی لحاظ می‌شود. دیاگرام مربوط به روش ترکیبی پیشنهادی در شکل ۲ ارائه شده است [۳۴].

۵- استخراج ویژگی‌های مقاوم به نویز به منظور مقاوم‌سازی بازشناسی ارقام مشابه فارسی

در بخش قبل، یک رویکرد ترکیبی مبتنی بر LSTM و HMM برای رفع چالش مشابهت ارقام مشابه فارسی ارائه شد. در شرایط نویزی، رخداد خطاهای جایگزینی در بازشناسی ارقام مشابه فارسی افزایش می‌یابد. استفاده از ویژگی‌های مقاوم به نویز بهبود دقت بازشناسی ارقام مشابه فارسی را در شرایط نویزی به همراه دارد. به منظور تصمیم‌گیری در خصوص این ویژگی‌ها، طیف ارقام مشابه فارسی مورد بررسی و تحلیل قرار گرفت. نتایج این بررسی‌ها در سه زیربخش مجزا برای ارقام مشابه "صفر و سه"، "دو و نه" و "پنج، هفت و هشت" ارائه شده‌اند.

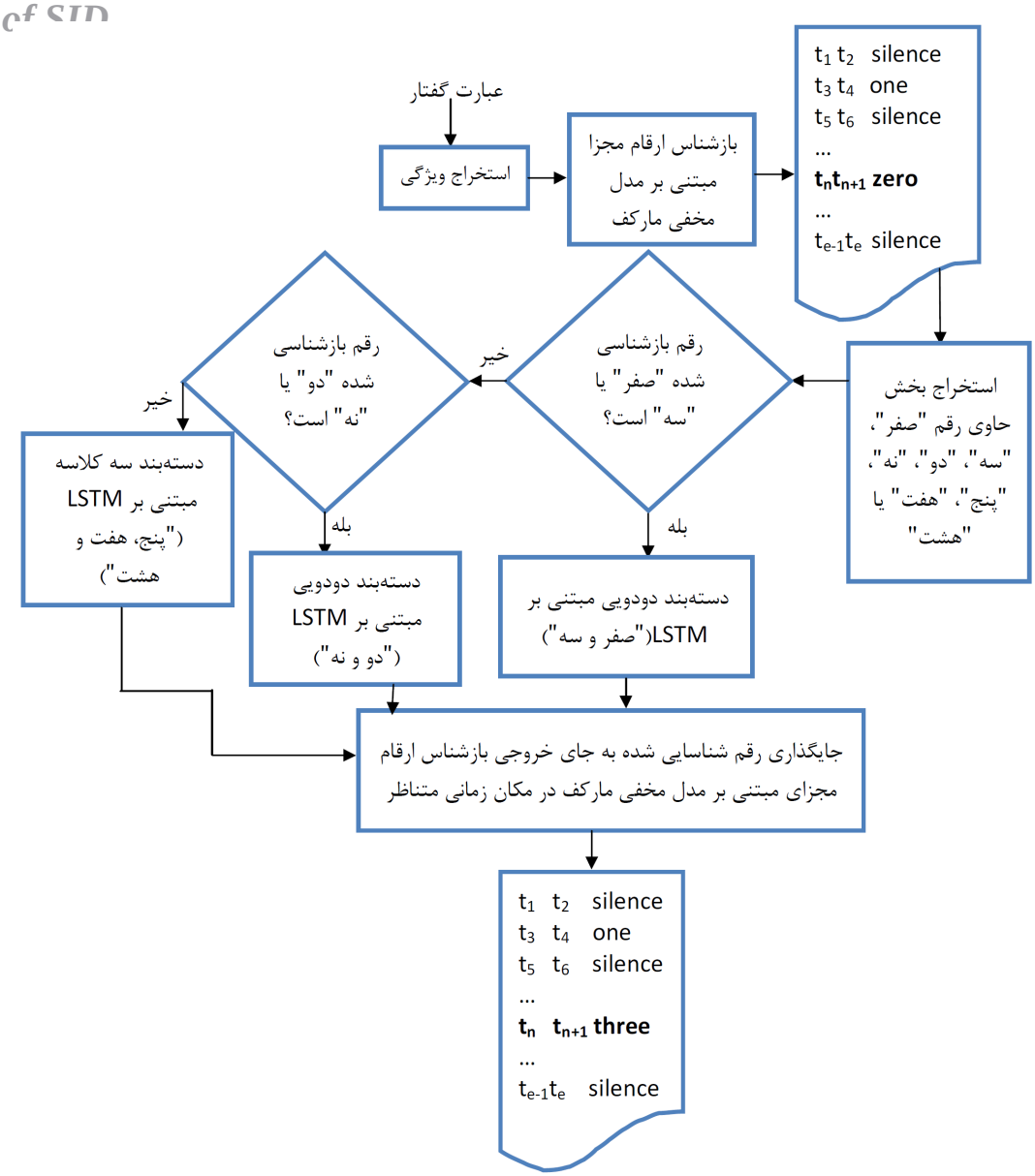
۵-۱ تحلیل طیف و سیگنال ارقام "صفر" و "سه"

طیف و سیگنال دو رقم "صفر" و "سه" در شکل ۳ ارائه شده است. چنانچه شکل ۳ نشان می‌دهد، سیگنال گفتار مربوط به دو رقم "سه" (شکل ۳- الف، بالا) و "صفر" (شکل ۳- الف، پایین) تفاوت محسوسی با هم دارند. این تفاوت در طیف مربوط به هر دو رقم واضح‌تر شده است. چنانچه شکل ۳- ب نشان می‌دهد، طیف گفتار مربوط به رقم "صفر" (شکل پایین) از چهار بخش مجزا (متناسب با چهار واج تشکیل‌دهنده‌اش ("س"، "اِ"، "ف" و "ر") تشکیل شده است. این در حالی است که طیف گفتار مربوط به رقم "سه" (شکل بالا)، تنها از دو بخش مجزا تشکیل شده است ("س" و "اِ"). اگر بتوانیم ویژگی‌هایی را از طیف گفتار استخراج کنیم که به نوعی تمایز بین طیف‌های دو رقم را منعکس کنند، دقت بازشناسی دو رقم مشابه به طور محسوسی افزایش می‌یابد. با توجه به مشابهت رقم "سه" به بخش اول رقم "صفر"، قاعدتاً ویژگی‌های مستخرج از طیف باید نماینده بخش دوم رقم "صفر" باشند تا بیشترین تمایز بین دو رقم را ایجاد نمایند. بخش دوم رقم "صفر" حاوی دو واج "ف" و "ر" می‌باشد که در دسته‌بندی واج‌ها به ترتیب در دو دسته سایشی‌های بی‌واک و شبه‌واکه‌ها قرار می‌گیرند. بر اساس نتایج حاصل در [۲۵]، چهار ویژگی آنتروپی طیفی [۳۵]، هموار طیفی [۳۶]، درجه از هم پاشی و فرکانس نیمساز [۳۷] در مقایسه با ضرایب مل کپستروم و مشتقاتشان از مقاومت محسوسی در شرایط نویزی برخوردار هستند. همچنین به خوبی می‌توانند منعکس‌کننده ویژگی‌های طیفی بخش دوم رقم "صفر" باشند. بنابراین می‌توانند به جای ضرایب مل کپستروم برای تفکیک دو رقم "سه" و "صفر" استفاده شوند.

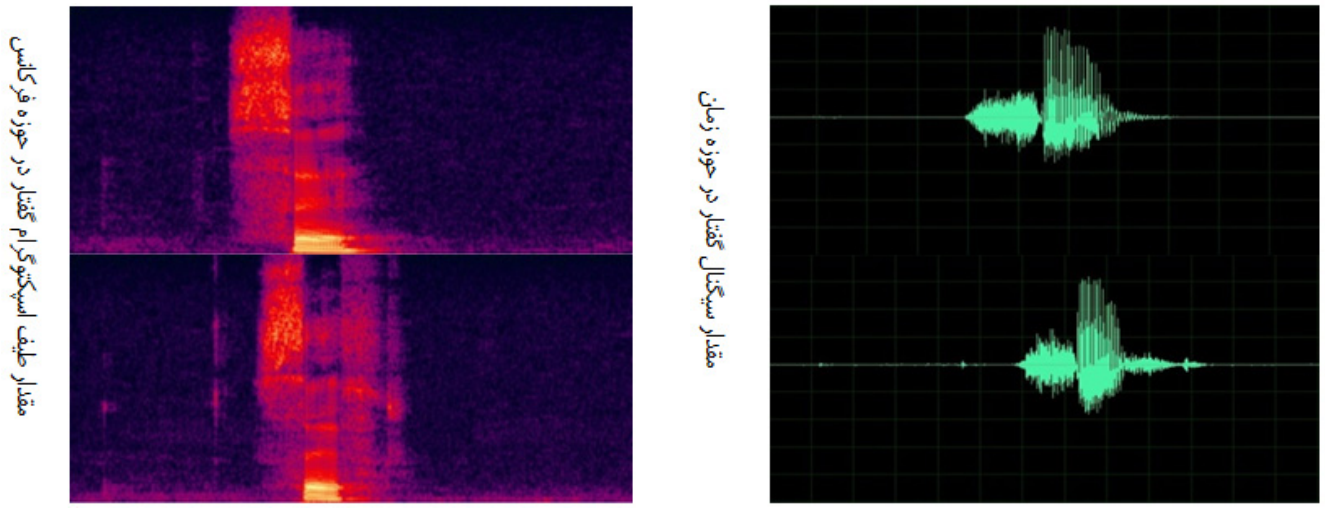
۵-۲ تحلیل طیف و سیگنال ارقام نه و دو

طیف و سیگنال گفتار مربوط به دو رقم "دو" و "نه" در شکل ۴ ارائه شده‌اند. چنانچه بخش الف از شکل ۴ نشان می‌دهند، سیگنال زمانی دو رقم "دو" (شکل بالا) و "نه" (شکل پایین) به لحاظ مقدار انرژی تفاوت

1. Cellphone-Based Persian Digits
2. Cepstral Mean and Variance Normalization



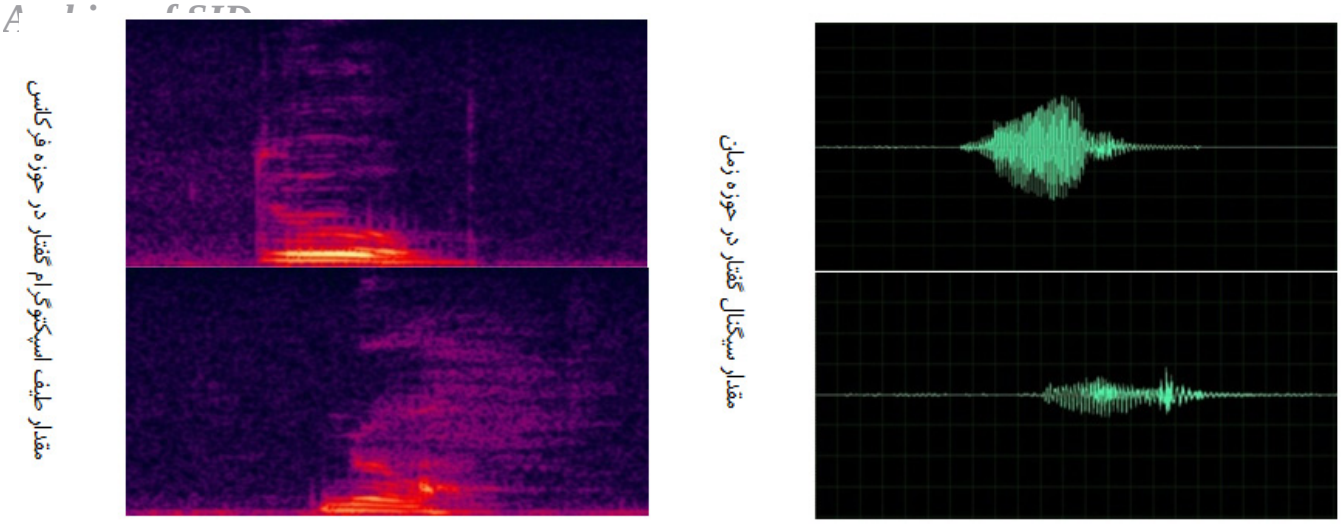
شکل ۲: دیاگرام رویکرد ترکیبی پیشنهادی برای اصلاح خطای بازشناسی ارقام فارسی با تلفظ مشابه.



نمونه‌ها در حوزه زمان
(الف)

نمونه‌ها در حوزه فرکانس
(ب)

شکل ۳: (الف) سیگنال گفتار ارقام "سه" (شکل بالا) و "صفر" (شکل پایین) در حوزه زمان و (ب) طیف اسپکتوگرام گفتار ارقام "سه" (شکل بالا) و"صفر" (شکل پایین).



نمونه‌ها در حوزه زمان
(الف)

نمونه‌ها در حوزه فرکانس
(ب)

شکل ۴: (الف) سیگنال گفتار ارقام "دو" (شکل بالا) و "نه" (شکل پایین) و (ب) طیف گفتار ارقام "دو" (شکل بالا) و "نه" (شکل پایین).

محسوسی با یکدیگر دارند. انرژی صحبت یک گویشور با گویشور دیگر متفاوت است. ممکن است یک گویشور رقم "نه" را با همان انرژی که گویشور دیگری رقم "دو" را ادا می‌کند، بیان کند. اگرچه در مورد یک گویشور دو رقم "نه" و "دو" در اغلب موارد، تفاوت انرژی محسوسی نسبت به یکدیگر دارند، اما وقتی طیف گویشوران مختلف را در یک کاربرد مستقل از گوینده در نظر می‌گیریم، این ویژگی نمی‌تواند چندان متمایزساز محسوب شود. به علاوه، انرژی طیف سیگنال تحت تأثیر نویز تغییر می‌کند. به عبارت دیگر انرژی سیگنال نویزی به اندازه انرژی نویز با انرژی سیگنال تمیز تفاوت داشته و تفکیک دو رقم "دو و نه" در شرایط نویزی با کیفیت شرایط تمیز نخواهد بود.

خطوط موازی در طیف گفتار یک سیگنال مکان قرارگرفتن فرمانت‌ها را نشان می‌دهد. نگاهی به طیف گفتار (شکل ۴- ب) رقم "دو" (شکل بالا) و "نه" (شکل پایین) تفاوت محسوسی را در شدت فرمانت اول این دو رقم به خصوص در بخش ابتدای دو رقم نشان می‌دهد. بخش اول رقم "نه" دربرگیرنده واج "ن" است که خیشومی و واکدار است. بخش اول رقم "دو" دربرگیرنده واج "د" می‌باشد که انفجاری و واکدار است. بخش دوم رقم "دو" واج "اُ" بوده و یکسان تلفظ می‌شوند. بخش آخر رقم "نه" که در رقم "دو" وجود ندارد، واج "ه" است که با توجه به مکان وقوعش در پایان کلمه و قرارگرفتنش در دسته نجوایی‌ها با انرژی کمی ادا شده و نمی‌تواند نقش متمایزسازی داشته باشد. بنابراین تنها بخش متمایزکننده دو رقم "نه" و "دو"، بخش ابتدای این "دو" رقم است که در هر دو مورد واکدار می‌باشد. از این رو، ویژگی‌های طیفی متمایزکننده نواحی واکدار و بی‌واک چنانچه پیش‌بینی می‌شود، برای تمایز ارقام "سه و صفر" کارساز باشند، در مورد تفکیک دو رقم "نه و دو" از یکدیگر کارایی نخواهند داشت. بنابراین تنها ویژگی متمایزساز ارقام "نه و دو" فرمانت اول می‌باشد. فرمانت اول با توجه به این که ویژگی خاص سیگنال‌های واکدار گفتار است، چندان تحت تأثیر نویز قرار نگرفته و ویژگی مقاومی در شرایط نویزی محسوب می‌شود.

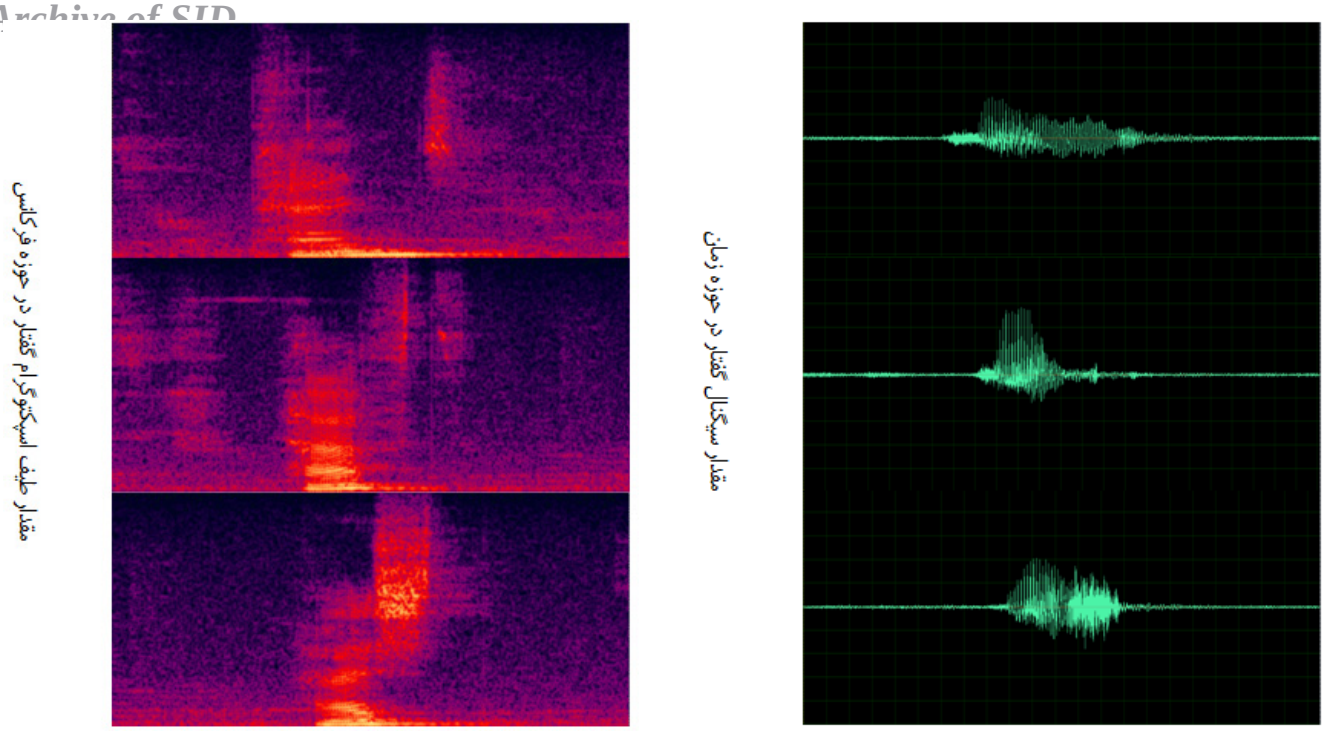
۳-۵ طیف و سیگنال ارقام پنج، هفت و هشت

طیف و سیگنال گفتار مربوط به سه رقم "پنج، هفت و هشت" در شکل ۵ ارائه شده‌اند. چنانچه شکل ۵ نشان می‌دهد، سیگنال گفتار مربوط به رقم "پنج" (شکل ۵- الف، بالا) و دو رقم "هفت" و "هشت" (شکل ۵- الف، وسط و پایین) تفاوت محسوسی با هم دارند. به ویژه این تفاوت در خصوص دو رقم "پنج" و "هشت" محسوس‌تر است. این تفاوت در طیف مربوط به ارقام واضح‌تر شده است. چنانچه شکل ۵- ب نشان می‌دهد، طیف گفتار مربوط به هر سه رقم از چهار بخش مجزا (متناسب با چهار واج تشکیل‌دهنده‌شان ("پ"، "اَ"، "ن" و "ج")، ("ه"، "اَ"، "ف" و "ت") و ("ه"، "اَ"، "ش" و "ت") تشکیل شده است. چنانچه مشخص است، دنباله آواهای ارقام "هفت" و "هشت" تنها در واج سوم با هم تفاوت دارند. هر دو واج "ف" و "ش" در دسته واج‌های سایشی غیر واکدار قرار می‌گیرند، لیکن "ش" در قیاس با "ف" نویزگونه‌تر می‌باشد. چنانچه طیف مربوط به ارقام و دنباله آواهای تشکیل‌دهنده‌شان نشان می‌دهد، تنها آوای متفاوت در هر سه رقم، آوای سوم می‌باشد. آوای سوم رقم "پنج" از دسته واج‌های خیشومی واکدار و آوای سوم دو رقم "هفت" و "هشت" از دسته واج‌های سایشی غیر واکدار است. همچنین آوای آخر رقم "پنج" از دسته واج‌های انفجاری- سایشی واکدار می‌باشد. این در حالی است که آوای آخر ارقام "هفت" و "هشت" از دسته واج‌های انفجاری غیر واکدار محسوب می‌شود. یکی از ویژگی‌های طیفی که بر اساس مقادیر آن بتوان به صورت قابل قبولی بخش‌های واکدار و غیر واکدار را از هم تفکیک نمود، نرخ گذار از صفر می‌باشد. این ویژگی تحت تأثیر نویز محیط تغییر می‌کند. در [۳۸] رابطه‌ای برای محاسبه نرخ گذار از صفر پیشنهاد شده که در مواجهه با نویز مقاوم می‌باشد. این رابطه به صورت زیر است

$$SNR(n)=\frac{1}{2K}\sum_{k=0}^{K-1}\left|sign(\varphi_n(k))-sign(\varphi_n(k-1))\right| \qquad (1)$$

که φ_n معرف تابع همبستگی[1] قاب مشخص‌شده با اندیس n از سیگنال گفتار می‌باشد. همچنین $sign$ تابع علامت است که برای آرگومان مثبت عدد یک و برای آرگومان منفی عدد منفی یک برمی‌گرداند. K نیز معرف تعداد کل نمونه‌ها در یک قاب می‌باشد. بنابراین برای تمایز سه رقم "پنج" از ارقام "هفت" و "هشت" یکی از ویژگی‌های متمایزکننده مقدار نرخ گذار از صفر مربوط به واج‌های سوم و چهارم می‌باشد. به منظور تمایز دقیق‌تر رقم "هفت" از رقم "هشت" و همچنین رقم "پنج" از دو رقم

1. Autocorrelation



نمونه‌ها در حوزه زمان
(الف)

نمونه‌ها در حوزه فرکانس
(ب)

شکل ۵: (الف) سیگنال گفتار ارقام "پنج" (شکل بالا)، "هفت" (شکل وسط) و "هشت" (شکل پایین) و (ب) طیف گفتار ارقام "پنج" (شکل بالا)، "هفت" (شکل وسط) و "هشت" (شکل پایین).

دیگر از مقدار آنتروپی طیفی واج سوم استفاده می‌شود که بنا بر تعریف، میزان نویزگونه‌بودن سیگنال گفتار را نشان می‌دهد [۳۵].

٦- دادگان

ضبط دادگان صوتی اعداد تک‌رقمی فارسی مبتنی بر تلفن همراه (CPHPD) با کمک نرم‌افزار voice recorder صورت گرفته است. برای هماهنگی بین داده‌های جمع‌آوری شده، صدای افراد با فرمت wav، نرخ بیت ۱۶ کیلوهرتز و به صورت تک‌بانده ضبط شده است. دادگان CPHPD شامل اعداد صفر تا نه می‌باشد که به زبان فارسی و به صورت جداگانه توسط هر نفر بیان شده است. برای انجام این کار از یک گروه ۶۴ نفری که شامل ۳۲ آقا و ۳۲ خانم با رنج سنی بین ۶ تا ۶۸ سال و با میزان تحصیلات مختلف هستند، استفاده کرده‌ایم.

ویرایش دادگان با استفاده از نرم‌افزار cool edit انجام شده است. اگر در بخش‌های سکوت فایل، صداهایی از قبیل کلیک، سرفه، باز و بسته شدن در و ... وجود داشته باشد، توسط نرم‌افزار حذف می‌شوند. سایر نویزهای محیطی با همان شدت باقی می‌مانند. البته با توجه به آن که گویشوران در محیط عاری از نویز اقدام به ضبط صدا کرده‌اند، میزان این نویز قابل توجه نمی‌باشد. برچسب‌گذاری دادگان در سطح کلمه و به صورت دستی با استفاده از نرم‌افزار cool edit انجام شده است. دادگان CPHPD حاوی ۶۴۰ فایل wav (۶۴ فایل به ازای هر رقم) با طول متوسط ۲ ثانیه و ۶۴۰ فایل lab (برچسب متناظر با هر فایل) می‌باشد. سایر جزئیات درباره دادگان در [۳۴] ارائه شده است.

۷- نتایج آزمایش‌ها

در این قسمت نتایج آزمایش‌های انجام‌شده در چهار بخش اصلی ارائه شده است. در بخش اول به ارزیابی بازشناسی ارقام مجزای فارسی مبتنی بر HMM پرداخته شده و سپس کارایی دسته‌بندهای ارقام مشابه فارسی مبتنی بر LSTM مورد تحلیل و ارزیابی قرار گرفته است. در انتهای بخش اول به ارزیابی رویکرد ترکیبی پیشنهادی برای رفع چالش بازشناسی ارقام مشابه فارسی در محیط تمیز خواهیم پرداخت. تمرکز بخش دوم مقاله بر مقاوم‌سازی بازشناسی ارقام مشابه فارسی مبتنی بر LSTM و ویژگی‌های مستخرج از طیف گفتار می‌باشد. در بخش سوم، پیچیدگی زمانی روش پیشنهادی مورد تحلیل قرار خواهد گرفت. در پایان، در بخش چهارم به مقایسه رویکرد ترکیبی پیشنهادی با روش‌های موجود برای رفع چالش مشابهت ارقام فارسی خواهیم پرداخت. به منظور انجام ارزیابی‌ها و افزایش اعتبار صحت نتایج از اعتبارسنجی متقابل چهارلایه استفاده شده و بنابراین دادگان CPHPD [۳۴] به چهار قسمت تقسیم گردیده است. هر قسمت شامل ارقام اداشده توسط ۱۶ گویشور می‌باشد و گویشوران این چهار قسمت کاملاً مستقل از هم هستند. بنابراین برای هر ارزیابی، چهار آزمایش انجام می‌شود. در هر آزمایش، یکی از چهار قسمت به عنوان مجموعه آزمون و سه قسمت دیگر به عنوان مجموعه آموزش در نظر گرفته می‌شود. نتایج این چهار آزمایش برای هر ارزیابی، میانگین‌گیری شده و نتیجه متوسط گزارش می‌شود. ارزیابی‌ها با کمک معیار زیر انجام می‌شوند [۳۹]

$$accuracy = \frac{H - I}{N} \qquad (۲)$$

که H معرف تعداد تشخیص‌های درست، N تعداد کل کلمات قابل شناسایی و I معرف خطای درج است.

همچنین به منظور تعیین تفاوت معنادار آماری میان روش‌های مورد مقایسه از آزمون فرضیه t-test استفاده شده است. آزمون فرضیه t-test تفاوت معنادار میان دو مقدار میانگین را با استفاده از رابطه زیر محاسبه می‌کند [۴۰]

جدول ۱: ارزیابی بازشناسی ارقام مجزای مبتنی بر HMM در سطح رقم بر روی چهار لایه آزمون.

	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	کل
۰	۵۵	۰	۰	۴	۰	۱	۰	۱	۱	۲	۶۴
۱	۰	۵۹	۱	۰	۱	۰	۰	۱	۰	۲	۶۴
۲	۰	۰	۵۶	۰	۱	۰	۰	۲	۰	۵	۶۴
۳	۷	۰	۰	۵۳	۰	۰	۱	۳	۰	۰	۶۴
۴	۰	۰	۰	۰	۵۸	۳	۰	۳	۰	۰	۶۴
۵	۰	۰	۰	۰	۱	۵۹	۰	۴	۰	۰	۶۴
۶	۳	۰	۰	۱	۰	۰	۵۸	۲	۰	۰	۶۴
۷	۰	۰	۰	۰	۱	۲	۰	۵۸	۳	۰	۶۴
۸	۰	۰	۰	۰	۰	۰	۰	۴	۶۰	۰	۶۴
۹	۱	۰	۱	۰	۰	۰	۰	۰	۰	۶۱	۶۴
کل	۶۶	۵۹	۵۸	۵۸	۶۲	۶۵	۵۹	۷۸	۶۴	۷۰	۶۴۰

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (۳)$$

که $\bar{x}_i$ معرف میانگین iام، s_i^2 معرف مقدار انحراف معیار iام و n_i معرف تعداد نمونه‌هایی است که s_i^2 انحراف معیارشان می‌باشد. مقدار t با یک حد آستانه مقایسه می‌شود که این حد آستانه بر اساس یک سطح اطمینان از پیش تعیین شده و جدول T[1] از مقادیر بحرانی t حاصل شده است.

۷-۱ بخش اول: بهبود دقت بازشناسی ارقام مجزای فارسی مبتنی بر LSTM و HMM

در این بخش، ابتدا سیستم بازشناس ارقام مجزای فارسی مبتنی بر HMM مورد ارزیابی قرار می‌گیرد و سپس به تحلیل کارایی دسته‌بندهای ارقام مشابه فارسی مبتنی بر LSTM می‌پردازیم. در انتها نیز رویکرد ترکیبی پیشنهادی برای رفع چالش بازشناسی ارقام مشابه فارسی مورد ارزیابی قرار می‌گیرد.

۷-۱-۱ ارزیابی بازشناسی ارقام مجزا مبتنی بر HMM

با توجه به نتایج حاصل‌شده در [۳۴]، بهترین تنظیمات برای آموزش سیستم بازشناس ارقام مجزای مبتنی بر HMM و تعداد ویژگی‌های استخراج‌شده، چهار مخلوط گوسی در هر وضعیت و ۱۲ ضریب مل کپستروم به همراه مشتقات اول و دومشان و سه ضریب انرژی می‌باشد. نتایج ارزیابی دقت بازشناسی ارقام مجزا مبتنی بر HMM در سطح رقم در جدول ۱ ارائه شده است.

همان طور که جدول ۱ نشان می‌دهد، سیستم بازشناس ارقام مجزا مبتنی بر HMM، بیشترین میزان خطای جایگزینی را در شناسایی ارقام "صفر"، "دو"، "سه"، "پنج"و "هشت" داشته است که به ترتیب در چهار مورد با "سه"، پنج مورد با "نه"، هفت مورد با "صفر"، چهار مورد با "پنج" و چهار مورد با "هشت" اشتباه شده است. رقم "هفت" در سه مورد با "هشت" و دو مورد با "پنج" اشتباه شده است. رقم "چهار" در سه مورد با "هفت" و در سه مورد با "پنج" اشتباه شده است. متوسط دقت بازشناسی ارقام مجزا مبتنی بر HMM برابر با ۹۰/۱۵ درصد بوده است. دقت بازشناسی برای ارقام "صفر"، "دو"، "سه"، "پنج"، "هفت"، "هشت" و "نه" به ترتیب برابر با ۸۶، ۸۷/۵، ۸۳، ۹۲/۱۸، ۹۰/۶، ۹۳/۷۵ و ۹۵/۳ درصد و به طور متوسط برای این هفت رقم برابر با ۸۹/۷ درصد بوده است.

1. T-Table

جدول ۲: ارزیابی دسته‌بندهای دودویی مبتنی بر شبکه LSTM.

عنوان شبکه بهینه	دقت
دسته‌بند دودویی متمایزساز "صفر و سه"	٪۹۶/۹
دسته‌بند دودویی متمایزساز "دو و نه"	٪۹۸/۴
دسته‌بند متمایزساز "پنج، هفت و هشت"	٪۹۱

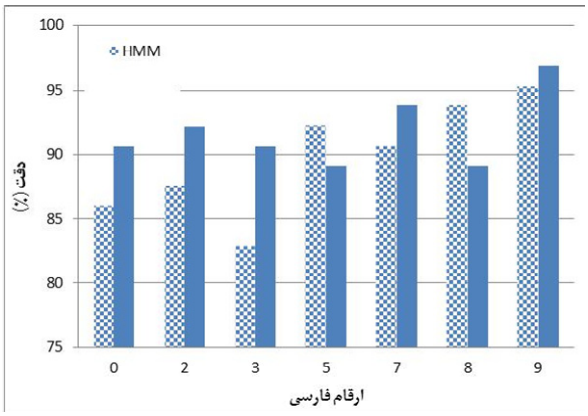
۷-۱-۲ ارزیابی دسته‌بندهای مبتنی بر LSTM برای شناسایی ارقام مشابه فارسی

به منظور آموزش دسته‌بندهای مبتنی بر LSTM، از ارقام "صفر"، "دو"، "سه"، "پنج"، "هفت"، "هشت" و "نه" از دادگان CPHPD استفاده شده است. برای پیاده‌سازی شبکه LSTM از جعبه ابزار یادگیری آموزش عمیق MATLAB ۲۰۱۸ [۴۱] استفاده شده است. این ابزار برای آموزش شبکه LSTM دو حالت "sequence to sequence" و "sequence to label" دارد. در این مقاله به جهت حجم پایین دادگان و ساده‌بودن مسئله بازشناسی ارقام مجزا از حالت دوم و در نتیجه از LSTM به عنوان یک شبکه عصبی ساده و غیر عمیق استفاده شده است. شبکه LSTM مورد استفاده برای هر کدام از دسته‌بندها از یک لایه شامل واحدهای مخفی (بلاک‌های حافظه) با اتصالات کامل و یک لایه ورودی با تعداد ۳۹ واحد و یک لایه خروجی با یک واحد (تنها برچسب خروجی) تشکیل شده است. ویژگی‌های استخراج‌شده مشابه ویژگی‌های مورد استفاده در سیستم بازشناس مبتنی بر HMM (۱۲ ضریب مل کپستروم، یک ضریب انرژی، مشتقات اول و دوم) می‌باشد. بهترین مقادیر مربوط به تعداد واحدهای مخفی در لایه میانی، بیشترین تعداد تکرار الگوریتم آموزش شبکه (max epoch) و کمترین اندازه برای دسته‌بندی دادگان آموزش در هر تکرار از الگوریتم (mini batch size)، از میان ۲۸۸ شبکه با تنظیمات مختلف برای هر یک از این سه پارامتر (تعداد واحدهای مخفی از ۵۰ تا ۴۰۰ با گام ۵۰، mini batch size از ۵ تا ۳۰ با گام ۵ و max epoch از ۵۰ تا ۳۰۰ با گام ۵۰) حاصل شده است. بهترین شبکه از میان ۲۸۸ شبکه با تنظیمات ذکرشده ذخیره گردیده و در مرحله ارزیابی و همچنین در رویکرد ترکیبی پیشنهادی مورد استفاده قرار گرفته است. انتخاب بهترین شبکه بر اساس معیار ارزیابی سیستم (رابطه (۲)) انجام شده و نتایج این ارزیابی‌ها در جدول ۲ ارائه گردیده است. قابل ذکر است که این نتایج، حاصل از متوسط‌گیری از نتایج اعتبارسنجی متقابل چهارلایه با کیفیتی که در بخش دادگان توضیح داده شد، می‌باشد.

۷-۱-۳ اصلاح دقت بازشناسی ارقام مشابه با استفاده از رویکرد ترکیب پیشنهادی مبتنی بر LSTM و HMM

نتایج ارزیابی رویکرد پیشنهادی مبتنی بر LSTM برای رفع چالش بازشناسی ارقام مشابه در شکل ۶ نشان داده شده است.

در شکل ۶ دقت بازشناسی ارقام "صفر"، "دو"، "سه"، "پنج"، "هفت"، "هشت" و "نه" حاصل از اعمال رویکرد ترکیبی پیشنهادی مبتنی بر شبکه LSTM و HMM با رویکرد مبتنی بر HMM مورد مقایسه قرار گرفته است. چنانچه شکل نشان می‌دهد، در شناسایی ارقام "صفر"، "دو"، "سه"، "هفت" و "نه" به ترتیب ۵، ۴/۷، ۷/۸، ۳/۲ و ۱/۶ درصد بهبود حاصل شده است. این مقدار بهبود بر اساس آزمون فرضیه t-test به ترتیب ۹۵، ۹۵، ۹۹/۵، ۹۰ و ۹۵ درصد معنادار می‌باشد. رویکرد



شکل ۶: ارزیابی رویکرد ترکیبی پیشنهادی برای اصلاح چالش مشابهت تلفظ ارقام فارسی.

ترکیبی پیشنهادی برای شناسایی ارقام "پنج" و "هشت" ضعیف‌تر از HMM عمل کرده است. با استفاده از رویکرد ترکیبی پیشنهادی، به طور متوسط دقت بازشناسی برای هفت رقم "صفر"، "دو"، "سه"، "پنج"، "هفت"، "هشت" و "نه" ۲ درصد بهبود داشته است. این مقدار بهبود بر اساس آزمون فرضیه t-test، ۹۵٪ معنادار می‌باشد. دقت بازشناسی ارقام مجزا مبتنی بر رویکرد ترکیبی پیشنهادی در جدول ۳ ارائه شده است.

همان طور که جدول ۳ نشان می‌دهد، سیستم بازشناس ارقام مجزا مبتنی بر رویکرد ترکیبی پیشنهادی، در قیاس با رویکرد مبتنی بر HMM برای ارقام "صفر"، "دو"، "سه"، "هفت" و "نه" خطای جایگزینی کمتری داشته و تنها برای ارقام "پنج" و "هشت" اندکی ضعیف‌تر عمل کرده است. دقت نهایی سیستم بازشناس ارقام مجزا مبتنی بر رویکرد ترکیبی پیشنهادی برابر با ۹۱/۷ درصد بوده که در مقایسه با سیستم بازشناس ارقام مجزا مبتنی بر HMM ۱/۵ درصد بهبود داشته است. این مقدار بهبود بر اساس آزمون فرضیه t-test، ۹۵ درصد معنادار می‌باشد.

۷-۲ بخش دوم: مقاوم‌سازی بازشناسی ارقام مشابه فارسی

برای ارزیابی قدرت تفکیک دسته‌بندها در شرایط نویزی علاوه بر دسته‌بندهای مبتنی بر LSTM بخش قبل، سه دسته‌بند مبتنی بر HMM برای تفکیک "دو از نه"، "سه از صفر" و "پنج، هفت و هشت"، آموزش داده شده است (مجدداً تأکید می‌شود که در تمام ارزیابی‌ها از اعتبارسنجی متقابل چهارلایه استفاده گردیده است). ویژگی‌های استخراج‌شده از طیف سیگنال گفتار برای ارزیابی قدرت تفکیک دسته‌بندها، همان ضرایب مل کپستروم و ضریب صفر انرژی به همراه مشتقات اول و دومشان می‌باشند. در ادامه، به منظور ارزیابی قدرت تفکیک ویژگی‌های مقاوم مستخرج از طیف سیگنال گفتار، دسته‌بند تفکیک‌کننده رقم "سه" از رقم "صفر" مبتنی بر LSTM با استفاده از چهار ویژگی طیفی توضیح داده شده در زیربخش ۵-۱، دسته‌بند دودویی تفکیک‌کننده رقم "نه" از رقم "دو" مبتنی بر LSTM بر مبنای فرمانت اول بخش آغازین دو رقم "دو و نه" و دسته‌بند تفکیک‌کننده سه رقم "پنج، هفت و هشت" از یکدیگر بر اساس نرخ گذار از صفر ارائه شده توسط (۱) واج سوم و چهارم ارقام مذکور و آنتروپی طیفی واج سوم آموزش داده شده‌اند. قابل ذکر است که تمام شش دسته‌بند مذکور بر روی مجموعه آموزش حاوی دادگان عاری از نویز آموزش یافته، ولی در شرایط نویزی مورد ارزیابی قرار گرفته‌اند. برای ایجاد شرایط نویزی، مجموعه آزمون دادگان CPHPD با استفاده از نویزهای سفید، صورتی، ماشین، کارخانه و همهمه از دادگان NOISEX۹۲ و با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل به صورت جمع‌شونده، آمیخته به نویز شده‌اند.

جدول ۳: ارزیابی بازشناسی ارقام مجزای مبتنی بر رویکرد ترکیبی پیشنهادی در سطح رقم بر روی چهار لایه آزمون.

	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	کل
۰	۵۸	۰	۰	۱	۰	۱	۰	۱	۱	۲	۶۴
۱	۰	۵۹	۱	۰	۱	۰	۰	۱	۰	۲	۶۴
۲	۰	۰	۵۹	۰	۱	۰	۰	۲	۰	۲	۶۴
۳	۲	۰	۰	۵۸	۰	۰	۱	۳	۰	۰	۶۴
۴	۰	۰	۰	۰	۵۸	۳	۰	۳	۰	۰	۶۴
۵	۰	۰	۰	۰	۱	۵۷	۰	۶	۰	۰	۶۴
۶	۳	۰	۰	۱	۰	۰	۵۸	۲	۰	۰	۶۴
۷	۰	۰	۰	۰	۱	۱	۰	۶۰	۲	۰	۶۴
۸	۰	۰	۰	۰	۰	۲	۰	۵	۵۷	۰	۶۴
۹	۱	۰	۰	۰	۰	۰	۰	۰	۰	۶۳	۶۴
کل	۶۴	۵۹	۶۰	۶۰	۶۲	۶۴	۵۹	۸۳	۶۰	۶۹	۶۴۰

در ادامه، ابتدا به ارزیابی قدرت تفکیک دسته‌بندهای مبتنی بر LSTM و HMM در شرایط نویزی پرداخته شده و سپس، میزان مقاوم‌بودن ویژگی‌های مستخرج از طیف توضیح داده شده در بخش پنجم، در شرایط نویزی مورد تحلیل و بررسی قرار خواهند گرفت.

۷-۲-۱ ارزیابی مقاومت ضرایب مل کپستروم در شرایط نویزی

برای آموزش دسته‌بندهای مبتنی بر HMM از همان تنظیمات سیستم بازشناس ارقام مجزایی که در زیربخش ۷-۱-۱ بیان شد، استفاده گردیده است. نتایج ارزیابی دسته‌بندهای مبتنی بر HMM و LSTM مبتنی بر ۳۹ ضریب مل کپستروم در جداول ۴ تا ۶ ارائه شده‌اند.

چنانچه نتایج مندرج در جدول ۴ نشان می‌دهند، دقت بازشناسی دسته‌بند مبتنی بر HMM برای تفکیک دو رقم "سه" و "صفر" در شرایط نویزی با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل نسبت به شرایط تمیز، به طور متوسط به ترتیب ۳۵، ۲۲/۷، ۱۷/۵، ۱۳/۱، ۶/۴ و ۹/۴ درصد افت داشته است. این در حالی است که دسته‌بند مبتنی بر LSTM متناظر در شرایط نویزی مشابه نسبت به شرایط تمیز، به طور متوسط به ترتیب ۳۰/۲، ۲۶/۶، ۲۴/۱، ۱۹/۹، ۱۳ و ۱۳ درصد افت داشته و در مقایسه با دسته‌بند مبتنی بر HMM به جز نسبت سیگنال به نویز ۵- دسی‌بل که عملکرد بسیار بهتری داشته است، در سایر نسبت‌های سیگنال به نویز عملکرد ضعیف‌تری داشته است.

بر اساس نتایج مندرج در جدول ۵، دقت بازشناسی دسته‌بند مبتنی بر HMM برای تفکیک دو رقم "نه" و "دو" در شرایط نویزی با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل نسبت به شرایط تمیز، به طور متوسط به ترتیب ۴۷/۵، ۳۰/۸، ۲۰/۵، ۱۳/۱، ۴/۵ و ۵/۸ درصد افت داشته است. این در حالی است که دسته‌بند مبتنی بر LSTM متناظر در شرایط نویزی مشابه نسبت به شرایط تمیز، به طور متوسط به ترتیب ۲۷/۶، ۲۷، ۲۴/۶، ۲۲/۵، ۱۸/۶ و ۱۶/۷ درصد افت داشته که در مقایسه با دسته‌بند مبتنی بر HMM، در نسبت‌های سیگنال به نویز پایین قدرت تفکیک بهتری داشته است.

بر اساس نتایج مندرج در جدول ۶، دقت بازشناسی دسته‌بند مبتنی بر HMM برای تفکیک سه رقم "پنج"، "هفت" و "هشت" در شرایط نویزی با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل نسبت به شرایط تمیز، به طور متوسط به ترتیب ۴۸/۲، ۳۵/۷، ۲۳/۸، ۱۴/۸،

جدول ۴: ارزیابی قدرت تفکیک دو دسته‌بند دودویی ارقام "سه و صفر" در شرایط نویزی بر حسب دقت بازشناسی (٪) و ۳۹ ضریب مل کپستروم.

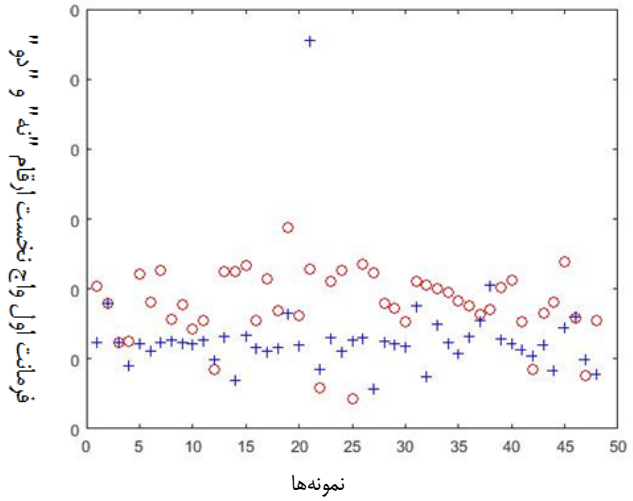
	SNR (db)	سفید	صورتی	کارخانه	ماشین	همهمه	میانگین
HMM	-۵	۶۱٫۷	۶۲٫۵	۵۱٫۵	۴۱٫۴	۴۵٫۳	۵۲٫۵
	۰	۶۷٫۲	۶۸٫۸	۶۱٫۷	۶۴٫۱	۶۲٫۵	۶۴٫۸
	۵	۶۴٫۸	۷۱٫۹	۷۱٫۹	۷۰٫۳	۷۱٫۱	۷۰
	۱۰	۷۴٫۲	۷۴٫۲	۷۳٫۴	۷۵	۷۵	۷۴٫۴
	۱۵	۷۸٫۱	۸۵٫۹	۸۷٫۵	۷۸٫۹	۷۵	۸۱٫۱
	۲۰	۷۸٫۱	۸۲٫۸	۸۰٫۴	۷۵٫۸	۷۳٫۴	۷۸٫۱
	تمیز				۸۷٫۵		
LSTM	-۵	۶۷٫۲	۷۳٫۴	۶۸٫۸	۶۱٫۷	۶۲٫۵	۶۶٫۷
	۰	۶۰٫۲	۷۰٫۳	۷۱٫۹	۷۳٫۴	۷۵٫۸	۷۰٫۳
	۵	۶۱٫۷	۷۱٫۹	۷۶٫۶	۷۴٫۲	۷۹٫۷	۷۲٫۸
	۱۰	۶۲٫۵	۷۷٫۳	۸۳٫۶	۷۸٫۱	۸۳٫۶	۷۷
	۱۵	۶۹٫۵	۸۴٫۴	۸۵٫۲	۷۹٫۷	۸۶٫۷	۸۳٫۹
	۲۰	۷۴٫۲	۸۶٫۷	۸۶٫۷	۸۳٫۶	۸۸٫۳	۸۳٫۹
	تمیز				۹۶٫۹		

جدول ۵: ارزیابی قدرت تفکیک دو دسته‌بند دودویی ارقام "نه و دو" در شرایط نویزی بر حسب دقت بازشناسی (٪) و ۳۹ ضریب مل کپستروم.

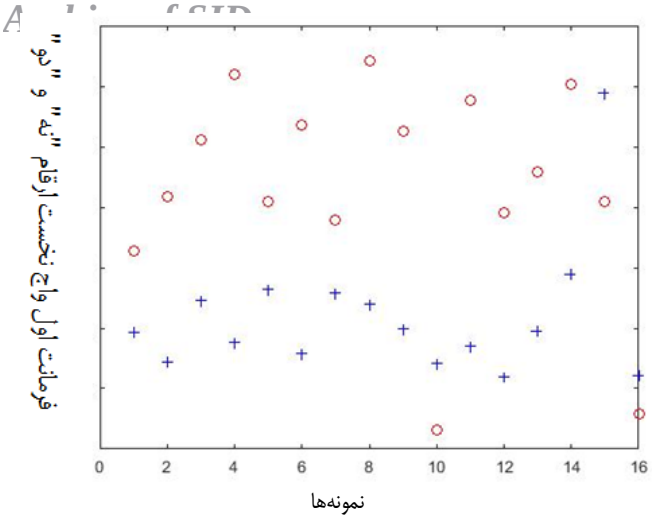
	SNR (db)	سفید	صورتی	کارخانه	ماشین	همهمه	میانگین
HMM	-۵	۶۹٫۵	۵۴٫۷	۳۳٫۶	۲۲٫۷	۱۹٫۵	۴۰
	۰	۷۲٫۶	۶۸	۶۵٫۶	۴۴٫۵	۳۲٫۸	۵۶٫۷
	۵	۷۶٫۵	۷۷٫۳	۷۵	۵۸٫۶	۴۷٫۶	۶۷
	۱۰	۷۸٫۹	۸۲	۷۹٫۷	۶۴	۶۷٫۲	۷۴٫۴
	۱۵	۸۲٫۸	۸۵٫۹	۸۵٫۹	۷۸٫۱	۸۲	۸۳
	۲۰	۸۱٫۳	۸۵٫۲	۸۴٫۴	۶۹٫۵	۷۸٫۱	۷۹٫۷
	تمیز				۸۷٫۵		
LSTM	-۵	۵۴٫۷	۷۵٫۸	۷۹٫۷	۶۹٫۵	۷۴٫۲	۷۰٫۸
	۰	۵۳٫۱	۷۳٫۴	۷۵٫۸	۷۳٫۴	۸۱٫۳	۷۱٫۴
	۵	۵۶٫۳	۷۵٫۸	۷۷٫۳	۷۵	۸۴٫۳	۷۳٫۸
	۱۰	۵۷	۷۷٫۳	۸۰٫۱	۷۷٫۳	۸۷٫۵	۷۵٫۹
	۱۵	۶۱٫۷	۸۱٫۳	۸۵٫۲	۷۹٫۷	۹۱٫۴	۷۹٫۸
	۲۰	۶۵٫۶	۸۴٫۴	۸۵٫۹	۷۹٫۷	۹۳	۸۱٫۷
	تمیز				۹۸٫۴		

جدول ۶: ارزیابی قدرت تفکیک دو دسته‌بند ارقام "پنج، هفت و هشت" در شرایط نویزی بر حسب دقت بازشناسی (٪) و ۳۹ ضریب مل کپستروم.

	SNR (db)	سفید	صورتی	کارخانه	ماشین	همهمه	میانگین
HMM	-۵	۴۰٫۱	۴۶٫۹	۴۹	۳۵٫۹	۴۷٫۹	۴۴
	۰	۴۵٫۳	۶۳٫۵	۶۱٫۴	۴۶٫۹	۶۵٫۱	۵۶٫۵
	۵	۵۷٫۳	۷۳٫۴	۷۲٫۴	۵۹٫۹	۷۹٫۲	۶۸٫۴
	۱۰	۶۶٫۱	۸۱٫۸	۸۳٫۹	۷۰٫۳	۸۴٫۹	۷۷٫۴
	۱۵	۷۲٫۴	۸۴٫۴	۸۳٫۳	۷۸٫۶	۸۶٫۵	۸۱
	۲۰	۷۰٫۸	۸۴٫۴	۸۴٫۹	۷۸٫۶	۸۶٫۵	۸۱
	تمیز				۹۲٫۲		
LSTM	-۵	۴۳٫۷	۵۴٫۷	۵۰	۴۷٫۹	۵۵٫۷	۵۰٫۴
	۰	۴۵٫۲	۶۶٫۷	۶۴٫۶	۵۲٫۶	۶۶٫۱	۵۹
	۵	۴۹٫۵	۷۷٫۶	۷۴	۶۲٫۵	۷۸٫۶	۶۸٫۴
	۱۰	۵۷٫۳	۸۳٫۳	۸۲٫۳	۷۱٫۴	۸۴٫۹	۷۵٫۸
	۱۵	۶۲٫۹	۷۸٫۱	۸۲٫۳	۶۸٫۲	۸۵٫۹	۷۵٫۵
	۲۰	۶۶٫۷	۸۰٫۷	۸۳٫۳	۷۰٫۸	۸۷	۷۷٫۷
	تمیز				۹۱٫۲		



شکل ۷: مقادیر فرمانت اول ارقام "دو" (دایره‌های قرمز) و "نه" (به‌علاوه‌های آبی) برای مجموعه آموزش.



شکل ۸: مقادیر فرمانت اول ارقام "دو" (دایره‌های قرمز) و "نه" (به‌علاوه‌های آبی) برای مجموعه آزمون.

۱۱٫۲ و ۱۱٫۲ درصد افت داشته است. این در حالی است که دسته‌بند مبتنی بر LSTM متناظر در شرایط نویزی مشابه نسبت به شرایط تمیز، به طور متوسط به ترتیب ۴۰٫۸، ۳۲٫۲، ۲۲٫۸، ۱۵٫۴، ۱۵٫۷ و ۱۳٫۵ درصد افت داشته که در مقایسه با دسته‌بند مبتنی بر HMM، در نسبت‌های سیگنال به نویز پایین‌تر قدرت تفکیک بهتری داشته است.

با توجه به نتایج ارزیابی‌ها در این بخش، اگرچه قدرت تفکیک دسته‌بندهای مبتنی بر LSTM برای تمایز ارقام مشابه فارسی در نسبت‌های سیگنال به نویز پایین از قدرت تفکیک دسته‌بند مبتنی بر HMM بهتر است، لیکن به دلیل عدم مقاومت ضرایب مل کپستروم در شرایط نویزی، تنزل دقت هر دو بازشناس در شرایط نویزی قابل توجه می‌باشد. برای افزایش مقاومت بازشناسی ارقام مشابه فارسی مبتنی بر LSTM در شرایط نویزی، در بخش پنجم چند ویژگی طیفی معرفی شد. برای تفکیک "سه از صفر" از چهار ویژگی آنتروپی طیفی، همواری طیفی، درجه از هم پاشی و فرکانس نیمساز، به منظور تفکیک "دو از نه" از فرمانت اول بخش آغازین ارقام و به منظور تفکیک سه رقم "پنج، هفت و هشت" از یکدیگر از نرخ گذار از صفر ارائه‌شده در (۱) واج‌های سوم و چهارم و آنتروپی طیفی واج سوم ارقام استفاده می‌شود. در ادامه نتایج ارزیابی دسته‌بندهای مبتنی بر HMM و LSTM در شرایط نویزی برای ویژگی‌های طیفی مذکور مورد بررسی قرار می‌گیرد.

۷-۲-۲ ارزیابی مقاومت ویژگی‌های طیفی در شرایط نویزی

با توجه به اختلاف محسوس ارقام "دو" و "نه" در بخش آغازینشان، لازم است ابتدا فرایندی برای استخراج نمونه‌های بخش آغازین طیف گفتار هر دو رقم تدوین نموده و سپس فرمانت اول را از آن بخش محاسبه نماییم. برای مرزبندی بخش آغازین هر رقم از نمونه‌های مجموعه آموزش در هر یک از چهار لایه[1] از روش اعتبارسنجی متقابل چهارلایه استفاده شده است. به ازای هر رقم در مجموعه آموزش در هر لایه ۴۸ فایل صوتی و در کل چهار لایه ۶۴ فایل صوتی مستقل از هم موجود می‌باشند. با بررسی تجربی فایل‌های صوتی ضبط‌شده به ازای هر کدام از ارقام "پنج، هفت و هشت" در هر چهار لایه، به این نتیجه رسیدیم که حدود یک‌هفتم ابتدای رقم در برگیرنده واج "ن" در مورد رقم "نه" و واج "د" در مورد رقم "دو" می‌باشد. بر این اساس فرمانت اول از یک‌هفتم ابتدای طیف گفتار این دو رقم استخراج می‌شود. بدیهی است که اگر گویشوری به صورت غیر عادی رقم را تند یا کند ادا کند، یک‌هفتم ابتدای طیف ممکن است دربرگیرنده تمام طول زمانی واج نخست آن رقم نباشد. لیکن با توجه به تنوع سن، لهجه، جنسیت و نرخ صحبت گویشوران لحاظ‌شده در دادگان CPHPD، می‌توان ادعا کرد در صورتی که گویشور با نرخ صحبت معمول و مطابق عرف صحبت کند، یک‌هفتم ابتدای طیف گفتار با دقت قابل قبولی شامل واج نخست ارقام "دو" یا "نه" می‌باشد. شکل‌های ۷ و ۸ به ترتیب نتایج بررسی مقادیر فرمانت اول را برای دو رقم "نه و دو" با استفاده از فایل‌های مجموعه‌های آموزش و آزمون نشان می‌دهند.

به منظور تعیین مکان واج‌های سوم و چهارم ارقام "پنج"، "هفت" و "هشت" از نمونه‌های مجموعه آموزش در هر یک از چهار لایه از روش اعتبارسنجی متقابل چهارلایه استفاده شده است. به ازای هر رقم در مجموعه آموزش در هر لایه ۴۸ فایل صوتی و در کل چهار لایه ۶۴ فایل صوتی مستقل از هم موجود می‌باشند. با بررسی تجربی فایل‌های صوتی ضبط‌شده به ازای هر کدام از ارقام "پنج، هفت و هشت" در هر چهار لایه، به این نتیجه رسیدیم که کشش زمانی واج‌های تشکیل‌دهنده هر سه رقم تقریباً یکسان بوده و برای دسترسی به واج‌های سوم و چهارم کافی است که طول زمانی رقم به چهار قسمت مساوی تقسیم شده و سومین قسمت به عنوان واج سوم و چهارمین قسمت به عنوان واج چهارم در نظر گرفته شود. بر این اساس، نرخ گذار از صفر مطابق (۱) از سومین و چهارمین قسمت طیف گفتار این سه رقم و آنتروپی طیفی از سومین قسمت این سه رقم استخراج می‌شود. بدیهی است که اگر گویشوری به صورت غیر عادی رقم را تند یا کند ادا کند، سومین یا چهارمین قسمت طیف گفتار ممکن است دربرگیرنده تمام طول زمانی واج سوم یا چهارم آن رقم نباشد. لیکن با توجه به تنوع سن، لهجه، جنسیت و نرخ صحبت گویشوران لحاظ‌شده در دادگان CPHPD، می‌توان ادعا کرد در صورتی که گویشور با نرخ صحبت معمول و مطابق عرف صحبت کند، سومین و چهارمین قسمت طیف گفتار با دقت قابل قبولی شامل واج‌های سوم و چهارم ارقام "پنج"، "هفت" یا "هشت" می‌باشد. دقت بازشناسی ارقام مشابه فارسی مبتنی بر LSTM با استفاده از ویژگی‌های طیفی در شرایط نویزی در جداول ۷ تا ۹ ارائه شده است.

چنانچه نتایج درج‌شده در جدول ۷ نشان می‌دهند، دقت بازشناسی دسته‌بند مبتنی بر LSTM با استفاده از چهار ویژگی طیفی برای تفکیک

1. Fold

جدول ۷: ارزیابی مقاومت سیستم بازشناس ارقام "سه" و "صفر" مبتنی بر LSTM در شرایط نویزی با استفاده از ویژگی‌های طیفی.

SNR (db)	سفید	صورتی	کارخانه	ماشین	همهمه	میانگین
-۵	۷۷٫۳	۷۱٫۸	۷۵	۷۲٫۷	۶۸	۷۳
۰	۷۶٫۶	۸۲٫۸	۸۲	۸۳٫۶	۷۶٫۶	۸۰٫۳
۵	۸۳٫۶	۸۷٫۵	۸۶٫۷	۸۵٫۲	۸۳٫۶	۸۵٫۳
۱۰	۸۸٫۳	۹۰	۸۸٫۲	۸۷٫۵	۸۶٫۷	۸۸٫۱
۱۵	۸۹٫۱	۹۰	۸۷٫۵	۸۷٫۵	۸۸٫۳	۸۸٫۴
۲۰	۹۰٫۱	۹۲٫۲	۸۸٫۳	۸۸٫۳	۸۶٫۷	۸۹٫۲
تمیز			۹۳٫۸			

جدول ۸: ارزیابی مقاومت سیستم بازشناس ارقام "دو" و "نه" مبتنی بر LSTM در شرایط نویزی با استفاده از فرمانت اول.

SNR (db)	سفید	صورتی	کارخانه	ماشین	همهمه	میانگین
-۵	۷۵	۶۶٫۴	۶۸	۵۳٫۹	۷۳٫۴	۶۷٫۳
۰	۸۲٫۸	۷۸٫۱	۸۲	۶۶٫۴	۸۰٫۵	۷۸
۵	۸۶٫۷	۸۸٫۳	۸۳٫۶	۷۲٫۶	۸۳٫۶	۸۳
۱۰	۸۸٫۳	۸۶٫۷	۸۵٫۲	۸۵٫۲	۸۵٫۹	۸۶٫۳
۱۵	۸۵٫۹	۸۶٫۷	۸۵٫۲	۸۶٫۷	۸۶٫۷	۸۶٫۳
۲۰	۸۷٫۵	۸۶٫۷	۸۶٫۷	۸۵٫۹	۸۷٫۵	۸۶٫۹
تمیز			۸۷٫۵			

جدول ۹: ارزیابی مقاومت سیستم بازشناس ارقام "پنج"، "هفت" و "هشت" مبتنی بر LSTM در شرایط نویزی با استفاده از نرخ گذار از صفر.

SNR (db)	سفید	صورتی	کارخانه	ماشین	همهمه	میانگین
-۵	۳۴٫۹	۳۵٫۴	۳۷٫۵	۳۹٫۱	۴۰٫۶	۳۷٫۵
۰	۳۸٫۵	۴۵٫۸	۴۵٫۳	۴۶٫۹	۵۱	۴۵٫۵
۵	۴۷٫۹	۵۴٫۲	۵۷٫۸	۵۴٫۲	۵۵٫۲	۵۳٫۹
۱۰	۵۶٫۳	۶۰	۵۹٫۴	۵۹٫۴	۶۳	۵۹٫۶
۱۵	۵۶٫۳	۶۳	۶۳٫۵	۶۳٫۵	۶۵٫۱	۶۲٫۵
۲۰	۶۰	۶۷٫۷	۶۶٫۱	۶۶٫۱	۶۵٫۶	۶۵٫۱
تمیز			۶۸٫۲			

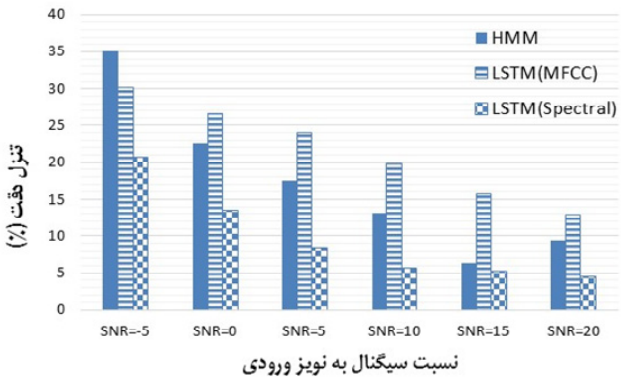
ارقام سه و صفر در شرایط نویزی با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل نسبت به شرایط تمیز، به طور متوسط به ترتیب ۲۰٫۸، ۱۳٫۵، ۸٫۵، ۵٫۷، ۵٫۴ و ۴٫۶ درصد افت داشته است. تأثیر استفاده از ویژگی‌های طیفی در مقایسه با ضرایب مل کپستروم، به وضوح گویای عملکرد بهتر آنها در شرایط نویزی است. نکته قابل توجه دیگر، کاهش تعداد ویژگی‌ها از ۳۹ به ۴ است که کاهش پیچیدگی محاسباتی و پاسخگویی سریع‌تر بازشناس را در پی دارد.

بر اساس نتایج درج‌شده در جدول ۸، دقت بازشناسی دسته‌بند مبتنی بر LSTM با استفاده از فرمانت اول بخش آغازین ارقام "دو" و "نه" در شرایط نویزی با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل نسبت به شرایط تمیز، به طور متوسط به ترتیب ۲۰٫۲، ۹٫۵، ۴٫۵، ۱٫۲، ۱٫۲ و ۰٫۶ درصد افت داشته که در مقایسه با استفاده از ضرایب مل کپستروم، به وضوح عملکرد بسیار بهتری داشته است. نکته قابل توجه دیگر، کاهش تعداد ویژگی‌ها از ۳۹ به تنها یک ویژگی برای بازشناسی تمایز "دو" از "نه" می‌باشد. اگرچه در شرایط تمیز استفاده از فرمانت اول به جای ۳۹ ضریب مل کپستروم، ۱۱ درصد افت دقت را در پی دارد، لیکن عملکرد چشم‌گیر بازشناس مبتنی بر این ویژگی در شرایط نویزی قابل توجه است.

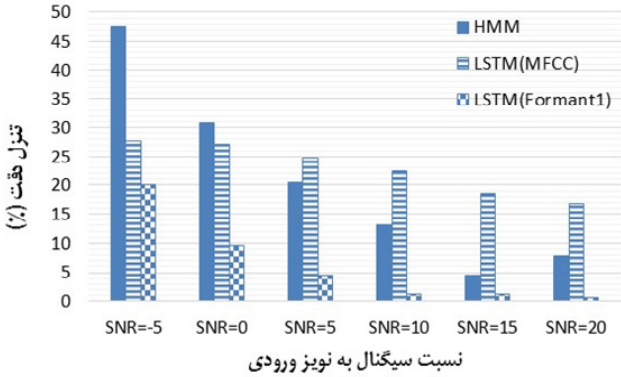
بر اساس نتایج درج‌شده در جدول ۹، دقت بازشناسی دسته‌بند مبتنی بر LSTM با استفاده از نرخ گذار از صفر بخش سوم ارقام "پنج"، "هفت" و "هشت" در شرایط نویزی با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل نسبت به شرایط تمیز، به طور متوسط به ترتیب ۳۰٫۷، ۲۲٫۷، ۱۴٫۳، ۸٫۶، ۵٫۷ و ۲٫۹ درصد افت داشته که در مقایسه با استفاده از ضرایب مل کپستروم، به وضوح عملکرد بسیار بهتری داشته است. نکته قابل توجه دیگر کاهش تعداد ویژگی‌ها از ۳۹ به سه ویژگی برای بازشناسی تمایز این سه رقم از یکدیگر می‌باشد. اگرچه در شرایط تمیز استفاده از سه ویژگی به جای ۳۹ ضریب مل کپستروم، ۲۳ درصد افت دقت را در پی دارد، لیکن عملکرد چشم‌گیر بازشناس مبتنی بر این ویژگی در شرایط نویزی قابل توجه است. مقایسه میان سه بازشناس ارقام مشابه فارسی بعد از میانگین‌گیری روی نتایج مربوط به پنج نویز مختلف در شکل‌های ۹ تا ۱۱ رسم شده است.

نمودارهای رسم‌شده در هر سه شکل معرف میزان افت دقت بازشناس مفروض در شرایط نویزی در قیاس با شرایط تمیز می‌باشد. هرچه این تنزل کمتر باشد، بازشناس مفروض از مقاومت بهتری در شرایط نویزی برخوردار است. چنانچه شکل‌های ۹ تا ۱۱ نشان می‌دهند، ویژگی‌های مقاوم به نویز مستخرج از طیف گفتار در مقایسه با ضرایب مل کپستروم در تفکیک "پنج" از "هفت" و "هشت"، "دو" از "نه" و "سه" از "صفر" به طور کاملاً محسوسی از عملکرد بهتری در شرایط مختلف نویزی برخوردار هستند.

بازشناس ارقام مشابه فارسی "صفر و سه" مبتنی بر LSTM و



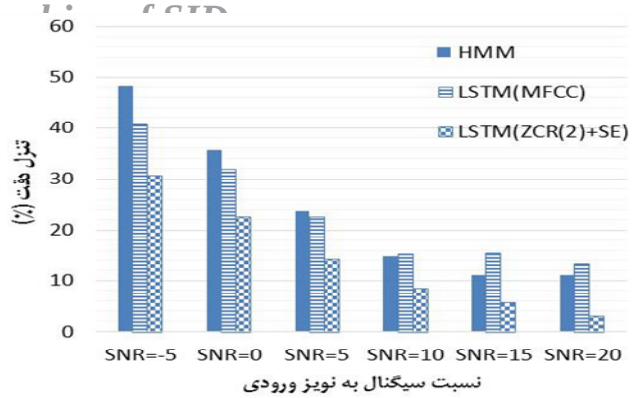
شکل ۹: مقایسه مقاومت بازشناسی "سه" و "صفر" برای سه بازشناس مبتنی بر مدل مخفی مارکف (HMM)، LSTM با ضرایب مل کپستروم ((LSTM (MFCC)) و LSTM با چهار ویژگی طیفی ((LSTM (Spectral)).



شکل ۱۰: مقایسه مقاومت بازشناسی "دو" و "نه" برای سه بازشناس مبتنی بر مدل مخفی مارکف (HMM)، LSTM با ضرایب مل کپستروم ((LSTM (MFCC)) و LSTM با فرمانت اول بخش آغازین دو رقم ((LSTM (Formant۱)).

ویژگی‌های مستخرج از طیف گفتار در شرایط نویزی با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل به ترتیب حدود ۱۵٪، ۹٪، ۹٪، ۸٪، ۱٪ و ۵٪ بهتر از همتای مبتنی بر HMM خود با ضرایب مل کپستروم و ۱۰٪، ۱۳٪، ۱۶٪، ۱۴٪، ۱۰٪ و ۸٪ بهتر از همتای مبتنی بر LSTM خود با ضرایب مل کپستروم عمل کرده است. این مقدار بهبود در تمام نسبت‌های سیگنال به نویز ورودی به جز نسبت سیگنال به نویز ۱۵ دسی‌بل در قیاس با بازشناس ارقام مجزای مبتنی بر HMM بر اساس آزمون فرضیه t-test به ترتیب بیش از ۹۹/۵ درصد معنادار می‌باشد. مقدار بهبود مذکور در نسبت سیگنال به نویز ۱۵ دسی‌بل بیش از ۷۵ درصد معنادار می‌باشد. این مقدار بهبود در تمام نسبت‌های سیگنال به نویز ورودی در قیاس با بازشناس ارقام مجزای مبتنی بر LSTM و ضرایب مل کپستروم و بر اساس آزمون فرضیه t-test بیش از ۹۹/۹۵ درصد معنادار می‌باشد.

بازشناس ارقام مشابه فارسی "دو و نه" مبتنی بر LSTM و ویژگی‌های مستخرج از طیف گفتار در شرایط نویزی با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل به ترتیب حدود ۲۷٪، ۲۱٪، ۱۶٪، ۱۲٪، ۳٪ و ۷٪ بهتر از همتای مبتنی بر HMM خود با ضرایب مل کپستروم و ۷٪، ۱۸٪، ۲۰٪، ۲۱٪، ۱۷٪ و ۱۶٪ بهتر از همتای مبتنی بر LSTM خود با ضرایب مل کپستروم عمل کرده است. این مقدار بهبود در تمام نسبت‌های سیگنال به نویز ورودی در قیاس با بازشناس‌های ارقام مجزای مبتنی بر HMM و LSTM با ضرایب مل کپستروم و بر اساس آزمون فرضیه t-test به ترتیب بیش از ۹۹/۹۵ و ۹۹/۵ درصد معنادار می‌باشد.



شکل ۱۱: مقایسه مقاومت بازشناسی "پنج"، "هفت" و "هشت" برای سه بازشناس مبتنی بر مدل مخفی مارکف (HMM)، LSTM با ضرایب مل کپستروم ((LSTM (MFCC)) و LSTM با نرخ گذار از صفر واج‌های سوم و چهارم و آنتروپی طیفی ((LSTM(ZCR(۲)_SE)).

بازشناس ارقام مشابه فارسی "پنج، هفت و هشت" مبتنی بر LSTM و ویژگی‌های مستخرج از طیف گفتار در شرایط نویزی با نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل به ترتیب حدود ۱۸٪، ۱۳٪، ۹٪، ۶٪، ۶٪ و ۸٪ بهتر از همتای مبتنی بر HMM خود با ضرایب مل کپستروم و ۱۰٪، ۱۰٪، ۸٪، ۷٪، ۱۰٪ و ۱۰٪ بهتر از همتای مبتنی بر LSTM خود با ضرایب مل کپستروم عمل کرده است. این مقدار بهبود در تمام نسبت‌های سیگنال به نویز ورودی در قیاس با هر دو بازشناس ارقام مجزای مبتنی بر HMM و LSTM با ضرایب مل کپستروم و بر اساس آزمون فرضیه t-test بیش از ۹۹/۵ درصد معنادار می‌باشد.

نکته دیگری از بررسی دقیق‌تر نتایج حاصل در شرایط نویزی مختلف حاصل می‌شود؛ زمانی که از ضرایب مل کپستروم به عنوان ویژگی استفاده شده است، رفتار بازشناس‌های ارقام مجزای فارسی در قبال نویزهای مختلف، متفاوت می‌باشد. بازشناس‌های مذکور برای برخی از نویزها عملکرد بهتر و برای برخی عملکردی ضعیف و غیر قابل پیش‌بینی دارند. این در حالی است که با استفاده از ویژگی‌های مستخرج از طیف گفتار، بازشناس ارقام مجزای مبتنی بر LSTM در قبال نویزهای مختلف در نسبت‌های سیگنال به نویز ورودی یکسان عملکرد تقریباً یکسان و قابل پیش‌بینی داشته است. به عبارت دیگر ویژگی‌های مستخرج از طیف پیشنهادی در این مقاله هم نسبت به تغییر نوع نویز و هم نسبت به تغییر نسبت سیگنال به نویز ورودی در قیاس با ضرایب مل کپستروم مقاوم‌تر عمل می‌کنند.

۳-۷ تحلیل پیچیدگی زمانی روش پیشنهادی

سیستمی که تمام آزمایش‌ها بر روی آن انجام شده‌اند، دارای پردازشگر Intel(R) Core™ i۵ – ۳۴۷۰ CPU @۳٫۲۰GHz و RAM چهار گیگابایتی می‌باشد. به منظور تحلیل پیچیدگی زمانی روش پیشنهادی، از معیار RTF[1] استفاده شده که به صورت زیر محاسبه می‌شود

$$RTF = \frac{P}{I} \qquad (۴)$$

که P معادل با طول زمان پردازش بر حسب ثانیه و I معرف طول زمانی فایل مورد پردازش بر حسب ثانیه می‌باشد. روش پیشنهادی از سه قسمت اصلی تشکیل شده است: زمان اعمال بازشناس ارقام مجزای مبتنی بر HMM، زمان لازم برای پردازش خروجی و تشخیص قطعات

1. Real Time Factor

جدول ۱۰: تحلیل پیچیدگی زمانی روش پیشنهادی در مقایسه با روش مبتنی بر HMM.

روش	RTF
روش مبتنی بر HMM	۰/۰۴۶
روش ترکیبی پیشنهادی مبتنی بر LSTM و HMM (ویژگی‌های مل کپستروم)	۰/۰۹۷
روش ترکیبی پیشنهادی مبتنی بر LSTM و HMM (ویژگی‌های مستخرج از طیف)	۰/۰۸۴

جدول ۱۱: مقایسه رویکرد پیشنهادی و موفق‌ترین رویکردهای مورد استفاده در سال‌های اخیر در حوزه بازشناسی ارقام مجزا.

روش	ویژگی‌های مستخرج	متوسط دقت در شرایط تمیز	متوسط دقت در شرایط نویزی
رویکرد مبتنی بر HMM	۳۹ ضریب مل کپستروم	۹۰/۲	۶۹/۵
رویکرد ترکیبی پیشنهادی مبتنی بر LSTM و HMM	۳۹ ضریب مل کپستروم	۹۱/۷	۶۹/۲۲
رویکرد ترکیبی پیشنهادی مبتنی بر LSTM و HMM	ویژگی‌های مستخرج از طیف گفتار	۸۲/۰۳	۷۱/۳
رویکرد مبتنی بر LSTM [۳۰]	۳۹ ضریب مل کپستروم	۹۲	۶۹/۱
رویکرد مبتنی بر DBN [۲۸]	۳۹ ضریب مل کپستروم	۸۷/۰۶	۶۱/۶
رویکرد مبتنی بر CNN	مل اسپکتوگرام	۹۰/۵۲	۶۶/۸

مشکوک به خطای جایگزینی حاصل از مشابهت تلفظ ارقام و زمان اعمال بازشناس ارقام با تلفظ مشابه مبتنی بر LSTM، به منظور محاسبه معیار RTF و تحلیل پیچیدگی زمانی روش پیشنهادی، لازم است زمان‌های پردازش هر یک از سه قسمت مذکور بر حسب ثانیه محاسبه شده و مجموع حاصل بر طول زمانی مجموعه مورد پردازش بر حسب ثانیه تقسیم شود. جدول ۱۰ تحلیل پیچیدگی زمانی روش پیشنهادی را در دو حالت استفاده از ویژگی‌های مل کپستروم و ویژگی‌های مستخرج از طیف گفتار در قیاس با بازشناس ارقام مجزای مبتنی بر HMM نشان می‌دهد.

چنانچه جدول ۱۰ نشان می‌دهد، بازشناس ارقام مجزای مبتنی بر HMM از کمترین میزان RTF برخوردار است که دلیل آن، انجام‌ندادن دو مرحله تشخیص قطعات مشکوک به خطای جایگزینی حاصل از مشابهت تلفظ ارقام و اعمال بازشناس ارقام با تلفظ مشابه مبتنی بر LSTM می‌باشد. اعمال این دو مرحله نیز در حالت استفاده از ویژگی‌های مل کپستروم و ویژگی‌های مستخرج از طیف به ترتیب تنها ۰/۰۵۱ و ۰/۰۴۲ مقدار RTF را افزایش داده که با توجه به کوچک‌تربودن RTF از یک، چندان قابل توجه نمی‌باشد. به علاوه، تشخیص قطعات مشکوک به خطای جایگزینی حاصل از مشابهت تلفظ ارقام در زمان بسیار کوتاهی انجام شده و در محاسبه پیچیدگی محاسباتی روش پیشنهادی تأثیر محسوسی ندارد. همچنین مشخص است که استفاده از ویژگی‌های مستخرج از طیف به جای ویژگی‌های مل کپستروم، به دلیل کاهش ابعاد بردار ویژگی‌ها پیچیدگی محاسباتی کمتری به همراه دارد.

۷-٤ مقایسه روش پیشنهادی با کارهای پیشین

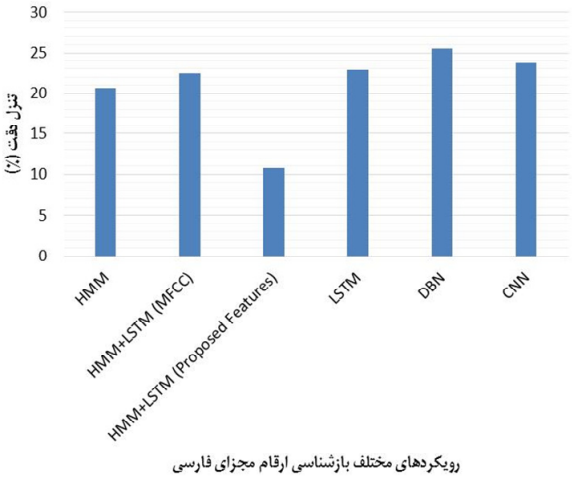
در این قسمت به مقایسه روش پیشنهادی با کارهای پیشین خواهیم پرداخت. این مقایسه در دو زیربخش انجام می‌گیرد. در زیربخش نخست، به مقایسه رویکردهای مطرح بازشناسی ارقام مجزا از هم در شرایط تمیز و نویزی خواهیم پرداخت. قابل توجه است که در حوزه بازشناسی ارقام مجزا، پژوهش‌های اندکی در زمینه مقاوم‌سازی بازشناسی ارقام مجزا مبتنی بر رویکردهای مطرح جدید مانند شبکه‌های عصبی عمیق، شبکه‌های بازگشتی، شبکه‌های کانولوشنی عمیق و ... انجام شده است. با این حال در راستای تکمیل پژوهش حاضر، مقایسه‌ای میان این رویکردها و رویکردهای مورد استفاده در این مقاله انجام شده است. در زیربخش دوم این بخش به مقایسه رویکرد پیشنهادی در این مقاله با رویکردهای موجود برای حل چالش مشابهت ارقام فارسی خواهیم پرداخت.

۷-٤-۱ مقایسه رویکردهای مطرح بازشناسی ارقام مجزا در شرایط تمیز و نویزی

چنانچه در بخش مقدمه نیز مطرح شد، اغلب پژوهش‌های انجام‌شده در حوزه بازشناسی ارقام مجزا بر روی زبان فارسی به دو دهه قبل مربوط می‌شوند. رویکردهای مطرح جدید مانند رویکردهای مبتنی بر شبکه‌های باور عمیق و شبکه‌های عصبی بازگشتی در مقالات محدودی مورد توجه پژوهشگران بوده است. این در حالی است که بر روی زبان‌های غیر فارسی تحقیقات بیشتری در حوزه بازشناسی ارقام مجزا در سال‌های اخیر انجام شده است. بر اساس مطالعات انجام‌شده در [۱۵] تا [۱۷] و [۲۷] موفق‌ترین رویکردهای مورد استفاده در حوزه بازشناسی ارقام مجزا در سال‌های اخیر مبتنی بر شبکه‌های عصبی کانولوشنی (CNN)، شبکه‌های باور عمیق (DBN) و حافظه کوتاه‌مدت ماندگار بوده‌اند. همچنین ویژگی‌های رایج مستخرج از سیگنال‌های گفتار ورودی، در اغلب این رویکردها همان ضرایب مل کپستروم بوده‌اند. نتایج حاصل از مقایسه رویکردهای مذکور و رویکردهای مورد استفاده در پژوهش حاضر، در شرایط تمیز و نویزی (همان شرایط نویزی مذکور در بخش قبلی) در جدول ۱۱ ارائه شده‌اند.

علت اختلاف مقادیر ارائه‌شده در جدول با مقادیر ارائه‌شده در [۲۸] و [۳۰]، اختلاف در شرایط نویزی کار حاضر با کارهای ارائه‌شده در آن مراجع و استفاده از رویکرد اعتبارسنجی متقابل چهارلایه برای اطمینان از صحت اعتبار نتایج در این کار است. به منظور مقایسه بهتر روش‌های ارائه‌شده در جدول ۱۱، میزان تنزل دقت در شرایط نویزی نسبت به شرایط تمیز برای هر روش محاسبه شده و در قالب نمودار شکل ۱۲ نشان داده شده‌اند.

همان طور که شکل ۱۲ نشان می‌دهد، رویکرد ترکیبی مبتنی بر HMM و LSTM که از ویژگی‌های مستخرج از طیف پیشنهادی در این مقاله برای رفع چالش مشابهت ارقام نویزی کمک می‌گیرد، از مقاومت مطلوب‌تری در شرایط نویزی برخوردار است. زیرا سایر رویکردهای مورد مقایسه در این مقاله از ویژگی‌های رایج مل کپستروم استفاده می‌کنند (معماری رویکرد مبتنی بر شبکه عصبی کانولوشنی مبتنی بر معماری ارائه‌شده در [۱۵] بوده و از مل اسپکتوگرام گفتار ورودی به عنوان ویژگی بهره می‌گیرد) که چندان در برابر نویز مقاوم نمی‌باشند. همچنین رویکرد مبتنی بر LSTM نسبت به رویکردهای مبتنی بر CNN و DBN از مقاومت مطلوب‌تری در شرایط نویزی برخوردار است که این نتیجه در



شکل ۱۲: مقایسه مقاومت رویکردهای مطرح بازشناسی ارقام مجزا.

پژوهش‌های نویسندگان دیگر نیز تأیید شده است [۲۷].

۷-٤-۲ مقایسه رویکردهای موجود برای حل چالش مشابهت ارقام فارسی

از میان کارهای به ثبت رسیده در حوزه بازشناسی مجزای ارقام فارسی، تنها تحقیقی که به چالش مشابهت ارقام فارسی پرداخته است، کار آقای حجازی و همکاران [۲۶] است. در این مقاله از ماشین بردار پشتیبان برای بهبود نتایج بازشناسی ارقام فارسی مبتنی بر HMM استفاده شده است. راهکار پیشنهادی این مقاله نیز اصلاح خطای بازشناسی ارقام مشابه مانند "دو و نه"، "سه و یک" و "هفت و هشت" به وسیله ماشین بردار پشتیبان دودویی می‌باشد. مشابهت ارقام بر اساس ساختار آوایی مشابه، تعداد آواهای واکدار مشابه، تعداد و مکان آواهای همخوان مشابه انتخاب شده است. راهکار پیشنهادی در هر دو حالت تمیز و نویزی ارزیابی شده است. دادگان مورد ارزیابی حاوی ۴۰۰ نمونه برای گویشوران آقا و ۵۰ نمونه برای گویشوران خانم در رنج‌های سنی مختلف است که برای ارقام یک تا نه و با نرخ نمونه‌برداری ۱۱۰۲۵ هرتز ضبط شده است. ۳۳۰ نمونه برای آموزش و ۱۲۰ نمونه برای آزمون انتخاب شده‌اند. تعداد ویژگی‌ها ۱۲ ضریب مل کپستروم به همراه یک ضریب انرژی بوده است. بر اساس نتایج حاصل از این کار، دقت روش مبتنی بر HMM بر روی دادگان آزمون در شرایط تمیز برابر با ۷۷٫۲۵ درصد می‌باشد. استفاده از ماشین بردار پشتیبان با بهبود مشکل بازشناسی ارقام مشابه، نرخ بازشناسی ارقام مجزا را در شرایط تمیز به ۹۸٫۵۹ درصد بهبود داده است. در شرایط نویزی اگرچه عملکرد روش ترکیبی بسیار بهتر از روش مبتنی بر HMM است، ولی روش چندان مقاومی در شرایط نویزی محسوب نشده و نسبت به حالت تمیز خودش در بدترین حالت حدود ۶۵ درصد تنزل دقت داشته است. مقایسه‌ای میان شرایط آزمایش و نرخ دقت‌های کسب‌شده از روش ارائه‌شده در [۲۶] و روش پیشنهادی در این مقاله در جدول ۱۲ ارائه شده است. برای عادلانه‌تربودن مقایسه‌ها، نتایج روش پیشنهادی مبتنی بر LSTM تنها برای ویژگی‌های مبتنی بر ضرایب مل کپستروم و در تنها نسبت سیگنال به نویز مشترک یعنی ۱۰ دسی‌بل ارائه شده است.

بر اساس نتایج ثبت‌شده در جدول ۱۲، عملکرد روش ترکیبی پیشنهادی مبتنی بر LSTM و HMM برای رفع چالش مشابهت تلفظ ارقام فارسی به میزان قابل توجهی در شرایط نویزی بهتر از روش مبتنی بر ماشین بردار پشتیبان ارائه‌شده در [۲۶] می‌باشد. با توجه به این که دادگان آزمون و آموزش دو رویکرد یکسان نیست، برای عادلانه‌بودن مقایسات، میزان تنزل دقت هر روش در شرایط آزمایش خودش محاسبه شده است. بر این اساس روش ارائه‌شده در [۲۶] در شرایط نویزی حدود ۶۵ درصد تنزل دقت داشته است. این در حالی است که روش ترکیبی پیشنهادی مبتنی بر LSTM و HMM در شرایط نویزی حدود ۱۵ درصد تنزل یافته است. روش پیشنهادی در این مقاله به میزان ۵۰ درصد نسبت به روش مبتنی بر ماشین بردار پشتیبان بهتر عمل کرده که این مقدار بهبود بر اساس آزمون فرضیه t-test بیش از ۹۹٫۹۵٪ معنادار می‌باشد.

۸- جمع‌بندی

در این مقاله به بررسی چالش مشابهت تلفظ ارقام فارسی در بازشناسی مجزای ارقام فارسی پرداختیم. بازشناس پایه برای بازشناسی ارقام مجزای فارسی مبتنی بر مدل مخفی مارکف و بر روی دادگان صوتی اعداد تک‌رقمی فارسی مبتنی بر تلفن همراه (CPHPD) آموزش داده شد. ارزیابی این بازشناس بر روی دادگان آزمون، حاکی از عملکرد ضعیف آن در بازشناسی ارقام مشابه فارسی مانند "صفر و سه"، "دو و نه" و "پنج، هفت و هشت" بود. به منظور رفع این چالش، رویکردی ترکیبی مبتنی بر حافظه کوتاه‌مدت ماندگار و مدل مخفی مارکف پیشنهاد شد. دقت بازشناس ارقام مجزای مبتنی بر مدل مخفی مارکف پس از اصلاح توسط رویکرد پیشنهادی از ۹۰٫۲ درصد به ۹۱٫۷ درصد بهبود یافته است. همچنین دقت بازشناسی ارقام مشابه از ۸۹٫۷ درصد برای بازشناس مبتنی بر مدل مخفی مارکف به ۹۱٫۸ درصد برای بازشناس مبتنی بر رویکرد ترکیبی پیشنهادی بهبود داشته است.

در ادامه کار، مقاوم‌سازی بازشناس ارقام مشابه فارسی مورد توجه قرار گرفت. در ابتدا، مقاومت دسته‌بند تفکیک‌کننده ارقام مشابه فارسی مبتنی بر ۳۹ ضریب مل کپستروم مورد ارزیابی قرار گرفت. نتایج ارزیابی حاکی از عملکر بهتر دسته‌بند مبنی بر حافظه کوتاه‌مدت ماندگار در مقایسه با دسته‌بند مبتنی بر مدل مخفی مارکف در نسبت‌های سیگنال به نویز پایین‌تر می‌باشد. به منظور افزایش مقاومت دسته‌بند مبتنی بر LSTM و با توجه به عملکرد ضعیف ضرایب مل کپستروم در شرایط نویزی، از طیف گفتار چهار ویژگی طیفی آنتروپی و همواری طیفی، فرکانس نمیساز و درجه از هم پاشی برای جداسازی "سه" از "صفر"، فرمانت اول بخش آغازین برای تفکیک "دو" از "نه" و نرخ گذار از صفر مبتنی بر تابع همبستگی واج‌های سوم و چهارم و آنتروپی طیفی واج سوم برای تمایز "پنج"، "هفت" و "هشت" استخراج شد. دسته‌بند مبتنی بر LSTM و چهار ویژگی طیفی مذکور برای تفکیک "سه" از "صفر" که در دو حالت تمیز و نویزی از کارایی مطلوبی (به طور متوسط ۱۳ درصد بهبود مقاومت در شرایط نویزی در مقایسه با ضرایب مل کپستروم) برخوردار بوده است. دقت بازشناس مبتنی بر فرمانت اول بخش آغازین برای تفکیک دو از نه در شرایط تمیز در مقایسه با ۳۹ ضریب مل کپستروم حدود ۱۱ درصد افت داشت. لیکن از عملکرد قابل توجهی در شرایط نویزی (به طور متوسط ۱۷ درصد بهبود مقاومت در شرایط نویزی در مقایسه با ضرایب مل کپستروم) برخوردار بوده است. دقت بازشناس مبتنی بر نرخ گذار از صفر مبتنی بر تابع همبستگی واج‌های سوم و چهارم و آنتروپی طیفی واج سوم برای تفکیک "پنج" از "هفت" و "هشت" در شرایط تمیز در مقایسه با ۳۹ ضریب مل کپستروم حدود ۲۳ درصد افت داشت. لیکن از عملکرد قابل توجهی در شرایط نویزی (به طور متوسط ۱۰ درصد بهبود مقاومت در شرایط نویزی در مقایسه با ضرایب مل کپستروم) برخوردار بوده است. در هر سه مورد کاهش پیچیدگی محاسباتی به دلیل کاهش تعداد ویژگی‌ها از ۳۹ ضریب به حداکثر چهار و حداقل ۱ ضریب حاصل شده است. تحلیل پیچیدگی محاسباتی روش پیشنهادی نیز مؤید همین مطلب می‌باشد.

جدول ۱۲: مقایسه روش پیشنهادی، روش مبتنی بر HMM و روش ارائه‌شده در [۲۶].

روش	ارقام مورد بازشناسی	متوسط دقت در شرایط تمیز	متوسط دقت در شرایط نویزی
روش ارائه‌شده در [۲۶]	یک تا نه	۹۸٫۵۹	۳۳٫۳
روش مبتنی بر HMM	صفر تا نه	۹۰٫۲	۷۵٫۴
روش ترکیبی پیشنهادی مبتنی بر LSTM و HMM (شکل ۲)	صفر تا نه	۹۱٫۷	۷۶٫۲

در ادامه، روش ترکیبی پیشنهادی برای اصلاح خطای بازشناسی ارقام مشابه در دو بخش با کارهای پیشین مورد قیاس قرار گرفت. در بخش اول رویکرد ترکیبی پیشنهادی مبتنی بر ویژگی‌های مستخرج از طیف با رویکردهای مبتنی بر مدل مخفی مارکف، رویکرد ترکیبی پیشنهادی مبتنی بر ضرایب مل کپستروم، حافظه کوتاه‌مدت ماندگار، شبکه باور عمیق و شبکه عصبی کانولوشنی مورد مقایسه قرار گرفت. نتایج مقایسه عملکرد بهتر (بیش از ۱۰ درصد افت دقت کمتر در شرایط نویزی) رویکرد پیشنهادی را در شرایط نویزی تأیید می‌کند. در بخش دوم، رویکرد پیشنهادی با یک رویکرد مبتنی بر ماشین بردار پشتیبان که برای رفع چالش مشابهت تلفظ ارقام فارسی ارائه شده است، مورد مقایسه قرار گرفت. نتایج مقایسات حاکی از آن است که روش پیشنهادی در شرایط نویزی به میزان ۵۰ درصد نسبت به روش مبتنی بر ماشین بردار پشتیبان بهتر عمل کرده است.

از نقاط قوت رویکرد پیشنهادی، عملکرد قابل قبول آن برای تفکیک ارقام با تلفظ مشابه به ویژه "دو و نه" و "صفر و سه" به دلیل تمرکز بر روی ویژگی‌های متمایزسازتر مستخرج از طیف این ارقام در مقایسه با ویژگی‌های رایج مل کپستروم در شرایط تمیز و نویزی می‌باشد. به ویژه در شرایط نویزی، حتی در نسبت‌های سیگنال به نویز ورودی بسیار پایین مانند ۵- دسی‌بل و صفر دسی‌بل، عملکرد روش پیشنهادی در مقایسه با بازشناس ارقام مجزای مبتنی بر مدل مخفی مارکف به طور قابل توجهی بهتر می‌باشد. این عملکرد بهتر هم در نسبت‌های سیگنال به نویز متفاوت و هم در قبال نویزهای مختلف محسوس است. اگرچه تفکیک "پنج" از "هفت" و "هشت" مبتنی بر ویژگی‌های مستخرج از طیف گفتار در شرایط نویزی با مقاومت مطلوبی صورت می‌گیرد، لیکن در شرایط تمیز به دلیل مشابهت بسیار زیاد آواهای تشکیل‌دهنده ارقام "هفت" و "هشت"، ویژگی‌های مل کپستروم قدرت تفکیک بهتری در مقایسه با ویژگی‌های مستخرج از طیف گفتار دارند. مقاوم‌سازی ویژگی‌های با قدرت تفکیک بالاتر ارقام "پنج"، "هفت و "هشت" در شرایط نویزی در کارهای آتی مورد تمرکز بیشتری قرار می‌گیرد. از سوی دیگر در کار حاضر از حافظه کوتاه‌مدت ماندگار به عنوان یک شبکه عصبی ساده برای تفکیک ارقام مجزای فارسی استفاده شده است. افزایش حجم دادگان ضبط‌شده به منظور آموزش حافظه کوتاه‌مدت ماندگار در قالب معماری‌های پایانه به پایانه و با بهره‌گیری از مکانیزم توجه با هدف بهبود دقت تفکیک ارقام مجزای فارسی از یکدیگر نیز می‌تواند در کارهای آتی مورد توجه بیشتری قرار بگیرد.

مراجع

[۱] ف. فکری، *شناسایی صحبت توسط کامپیوتر*، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شریف، دانشکده مهندسی برق، ۱۳۷۱.

[۲] ح. بابایک، "بازشناسی گفتار با استفاده از تلفیق مدل مخفی مارکف و شبکه عصبی،" *مجموعه مقالات هفتمین کنفرانس مهندسی برق ایران، مقالات مخابرات سیستم*، صص. ۲۰۴-۱۹۹، تهران، ایران، ۲۹-۲۷ اردیبهشت ۱۳۷۸.

[۳] س. بابایی‌زاده، ا. غلام‌پور و ک. نایبی، "بهبود کارایی سیستم‌های بازشناسی گفتار گسسته با ترکیب شبکه‌های عصبی و مدل‌های مارکف پنهان،" *مجموعه مقالات هفتمین کنفرانس مهندسی برق ایران، مقالات مخابرات سیستم*، صص. ۱۹۰-۱۸۳، تهران، ایران، ۲۹-۲۷ اردیبهشت ۱۳۷۸.

[۴] ش. رستم‌زاده، س. م. احدی، ح. شیخ‌زاده، نجار، "بازشناسی گفتار فارسی ناپیوسته، به صورت ناوابسته به گوینده به کمک مدل‌های پنهان مارکف با چگالی پیوسته،" *مجموعه مقالات ششمین کنفرانس مهندسی برق ایران*، صص. ۹۷-۹۳، تهران، ایران، اردیبهشت ۱۳۷۷.

[۵] م. م. همایون‌پور و ا. نجاری، "بازشناسی ارقام ناوابسته به گوینده با استفاده از مدل پیشگوی عصبی،" *مجموعه مقالات هفتمین کنفرانس مهندسی برق ایران*، صص. ۸۱-۷۵، تهران، ایران، ۲۹-۲۷ اردیبهشت ۱۳۷۸.

[۶] ا. صیادیان، ک. بدیع، م. حکاک و م. ر. بیک‌زاده، "ارائه روش آماری FPG-GMM در بازشناسی گفتار،" *مجموعه مقالات هشتمین کنفرانس مهندسی برق ایران*، صص. ۴۰۶-۳۹۸، اصفهان، ایران، ۳۰-۲۸ اردیبهشت ۱۳۷۹.

[۷] ا. اکبری و ب. ناصرشریف، "بازشناسی هجاها در اعداد دورقمی فارسی به وسیله مدل مخفی مارکف،" *مجموعه مقالات ششمین کنفرانس سالانه انجمن کامپیوتر ایران*، صص. ۴۳۷-۴۳۲، اصفهان، ایران، ۴-۲ اسفند ۱۳۷۹.

[۸] م. م. همایون‌پور و ج. کبودیان، "بازشناسی اعداد فارسی بر روی خط تلفن: مقایسه‌ای بین روش‌های آماری، عصبی و هیبرید،" *مجله مهندسی برق*، سال چهاردهم، شماره آ-۵۶، صص. ۱۰۶۵-۱۰۴۵، پاییز ۱۳۸۲.

[۹] دانشگاه صنعتی امیرکبیر، *گزارش نهایی طرح ملی پردازش زبان فارسی*، شورای پژوهش‌های علمی کشور، کمیسیون اطلاع‌رسانی و فناوری اطلاعات، صص. ۶۸-۶۷، ۱۳۸۰.

[10] J. Wu, Y. Chua, M. Zhang, H. Li, and K. C. Tan, "A spiking neural network framework for robust sound classification," *Frontiers in Neuroscience*, vol. 12, Article No.: 836, 17 pp., Nov. 2018.

[11] A. Wazir and J. Chuah, "Spoken Arabic digits recognition using deep learning," in *Proc. IEEE Int. Conf. on Automatic Control and Intelligent Systems*, *I2CACIS'19*, pp. 339-344, Selangor, Malaysia, 29-29 Jun. 2019.

[12] E. Swedia, A. Mutiara, and M. Subali, "Deep learning long-short term memory (LSTM) for indonesian speech digit recognition using LPC and MFCC feature," in *Proc. 3rd Int. Conf. on Informatics and Computing, ICIC'18*, 5 pp., Palembang, Indonesia, 17-18 Oct. 2018.

[13] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, "Bi-directional recurrent end-to-end neural network classifier for spoken Arab digit recognition," in *Proc. 2nd Int. Conf. on Natural Language and Speech Processing, ICNLSP'18*, 6 pp., Algiers, Algeria, 25-26 Apr. 2018.

[14] A. B. Nassif, S. Ismail, A. Imtinan, A. Mohammad, and S. Khaled, "Speech recognition using deep neural networks: a systematic review," *IEEE Access*, vol. 7, pp. 19143-19165, 2019.

[15] R. Sharmin, K. R. Shantanu, and R. H. Mohammad, "Bengali spoken digit classification: a deep learning approach using convolutional neural network," *Procedia Computer Science*, vol. 17, pp. 1381-1388, 2020.

[16] B. Zada and U. Rahim, "Pashto isolated digits recognition using deep convolutional neural network," *Heliyon*, vol. 6, no. 2, Article No.: e03372, 6 pp., Feb. 2020.

[17] A. Graves, D. Eck, and J. Schmidhuber, *LSTM and Timewarping: Spoken Digit Recognition with a Recurrent Neural Network*, Technical Report, No. IDSIA-12-03, pp. 1-9, 2003.

[18] D. Dhanashri and S. B. Dhonde, "Isolated word speech recognition system using deep neural networks," in *Proc. of the Int Conf. on Data Engineering and Communication Technology,* Springer, Singapore, pp. 9-17, Aug. 2017.

[19] S. Vihari, A. S. Murthy, P. Soni, and D. C. Naik, "Comparison of speech enhancement algorithms," *Procedia Computer Science*, vol. 89, pp. 666-676, 2016.

[20] A. Pervaiz, *et al*., "Incorporating noise robustness in speech command recognition by noise augmentation of training data," *Sensors*, vol. 20, no. 8, pp. 2326-2344, 2020.

[21] D. Grozdic, J. Slobodan, S. P. Dragana, G. Jovan, and M. Branko, "Comparison of cepstral normalization techniques in whispered speech recognition," *Advances in Electrical and Computer Engineering*, vol. 17, no. 1, pp. 21-26, Feb. 2017.

[22] V. Mitra, *et al*., "Robust features in deep-learning-based speech recognition," in S.Watanabe, M. Delcroix, F. Metze, and J.Hershey (eds) *New Era for Robust Speech Recognition, Springer, Cham*, pp. 187-217, 2017.

[23] D. Vazhenina and K. Markov, "End-to-end noisy speech recognition using Fourier and Hilbert spectrum features," *Electronics*, vol. 9, no. 7, pp. 1157-1174, 2020.

[24] S. Chang and S. Wegmann, "On the importance of modeling and robustness for deep neural network feature," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'15*, pp. 4530-4534, South Brisbane, Australia, 19-24 Apr. 2015.

[25] S. Tabibian, A. Akbari, and B. Nasersharif, "Keyword spotting using an evolutionary-based classifier and discriminative features," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 7, pp. 1660-1670, Aug. 2013.

[26] S. A. Hejazi, R. Kazemi, and S. Ghaemmaghami, "Isolated persian digit recognition uding a hybrid HMM-SVM," in *Proc. Int. Symp. on Intelligent Signal Processing and Communication Systems,* 4 pp., Bangkok, Thailand, 8-11 Feb. 2009.

[27] L. Ming, Y. Wang, J. Wang, J. Wang, and X. Xie, "Speech enhancement method based on LSTM neural network for speech recognition," in *Proc. 14th IEEE Int. Conf. on Signal Processing, ICSP'18*, pp. 245-249, Beijing, China, 12-16 Aug. 2018.

[۲۸] ش. طبیبیان، "بهبود بازشناسی ارقام مجزای فارسی در تلفن همراه بر مبنای کاهش خطای دسته‌بندی در سطح قاب،" *مجموعه مقالات بیست و چهارمین کنفرانس ملی انجمن کامپیوتر ایران*صص. ۱۳۵–۱۲۸، تهران، ایران، ۲۳۵–۲۲ اسفند ۱۳۹۷.

[۲۹] ش. طبیبیان، "بهبود بازشناسی ارقام مشابه فارسی مبتنی بر شبکه بازگشتی LSTM،" *بیست و چهارمین کنفرانس ملی انجمن کامپیوتر ایران*، صص. ۴۳۸–۴۳۲، تهران، ایران، ۲۳–۲۲ اسفند ۱۳۹۷.

[30] M. M. Naseri and S. Tabibian, "Improving the robustness of persian spoken isolated digit recognition based on LSTM," in *Proc. 6th Int. Conf. of Signal Processing and Intelligent Systems, ICSPIS'20*, 6 pp., Mashhad, Iran, 23-24 Dec. 2020.

[31] S. Hochreiter and J. Schmidhuber, "Long short term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997.

[32] C. Olah, Understanding LSTM Networks, http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 2015.

[33] Hidden Markov Model Toolkit (HTK), Speech vision and robotics group of the Cambridge University engineering department, http://htk.eng.cam.ac.uk/, August 2015.

[۳۴] ش. طبیبیان، "طراحی و جمع‌آوری دادگان صوتی اعداد تک‌رقمی فارسی مبتنی بر تلفن همراه،" *مجموعه مقالات چهارمین کنفرانس پردازش سیگنال و سیستم‌های هوشمند*، ۵ صص.، *تهران، ایران*، ۴–۴ دی ۱۳۹۷.

[35] A. M. Toh, R. Togneri, and S. Nordholm, "Spectral entropy as speech features for speech recognition," in *Proc. of Postgraduate Electrical Engineering and Computing Symp., PEECS'05,* pp. 22-25, Perth, Australia, 2005.

[36] G. Peeters, A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project, Cuidado Project Report, Ircam, pp. 1-25, 2004.

[37] C. Y. Lin, J. S. Rager Jang, and K. T. Chen, "Automatic segmentation and labeling for Mandarin Chinese Speech Corpora for concatenation-based TTS," *Computer Linguistic Chinese Language Processing*, vol. 10, pp. 145-166, 2005.

[38] P. Kathirvel, M. S. Manikandan, S. Senthilkumar, and K. P. Soman, "Noise robust zerocrossing rate computation for audio signal classification," in *Proc. 3rd Int. Conf. on Trendz in Information Sciences & Computing, TISC'11*, pp. 65-69, Chennai, India, 8-9 Dec. 2011.

[39] K. Dietrich and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication*, vol. 38, no. 1-2, pp. 19-28, Sept. 2002.

[40] M. D. Mahony, *Sensory Evaluation of Food: Statistical Methods and Procedures*, CRC Press, 1986.

[41] MathWorks, *Long Short-Term Memory Networks*, https://www.mathworks.com/help/deeplearning/ug/long-short-term-memory-networks.html

شیما طبیبیان تحصیلات خود را در مقاطع کارشناسی رشته مهندسی کامپیوتر- گرایش نرم‌افزار از دانشگاه صنعتی اصفهان در سال ۱۳۸۳ اخذ کرده است. ایشان مقاطع کارشناسی ارشد و دکترای خود را در رشته مهندسی کامپیوتر- گرایش هوش مصنوعی و رباتیک در سال‌های ۱۳۸۶ و ۱۳۹۲ در دانشگاه علم و صنعت ایران گذرانده است. نام‌برده قبل از پیوستنش به دانشگاه شهید بهشتی در سمت استادیار و عضو هیات علمی پژوهشکده فضای مجازی در سال‌های ۱۳۹۳ الی ۱۳۹۶ استادیار پژوهشگاه هوافضا بوده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: بازشناسی گفتار، واژه‌یابی گفتار، بهسازی گفتار، تشخیص فرامین صوتی، طراحی واسط‌های کاربری مبتنی بر گفتار، تشخیص احساس از گفتار، روش‌های یادگیری ماشین، پایش سلامت و روش‌های بهینه‌سازی.