بازشناسی مقاوم اعداد گفتار فارسی با شبکه عصبی عمیق

علی نصر اصفهانی[۱]، مهدی بکرانی[۲]، روزبه رجبی[۳]

[۱] دانشگاه صنعتی قم، دانشکده مهندسی برق و کامپیوتر، گروه مخابرات و الکترونیک
Alinasresf8@gmail.com

[۲] دانشگاه صنعتی قم، دانشکده مهندسی برق و کامپیوتر، گروه مخابرات و الکترونیک
Bekrani@qut.ac.ir

[۳] دانشگاه صنعتی قم، دانشکده مهندسی برق و کامپیوتر، گروه مخابرات و الکترونیک
Rajabi@qut.ac.ir

چکیده

از چالشهای مهم در بازشناسی اعداد در گفتار وجود نویز در صدای دریافتی دستگاه‌های دیجیتال و تشابهات وجهی اعداد می‌باشد. برای مقابله با این چالش‌ها در این پژوهش، علاوه بر درنظرگرفتن واحد کلمه به‌جای واحد واج، انجام عملیات داده‌افزایی به‌منظور بهبود عملکرد سیستم، یک ساختار ترکیبی از دو شبکه عصبی کانولوشنال باقیمانده و شبکه عصبی واحد بازگشتی گیتی دوطرفه برای بازشناسی اعداد فارسی گسسته صفر تا نه از گفتار ارائه شده است. نتایج حاصل نشان می‌دهند که دقت بازشناسی گفتار روش پیشنهادی برای داده‌های آموزش و اعتبارسنجی به ترتیب ۹۸/۵۳% و ۹۶/۱۰% است. این نتایج نسبت به روش‌های مبتنی بر شبکه عصبی LSTM دارای عملکرد بهتری است.

کلمات کلیدی: بازشناسی ارقام مجزا، داده‌افزایی[۱]، شبکه عصبی کانولوشنال باقیمانده[۲]، شبکه عصبی واحد بازگشتی گیتی دوطرفه[۳]

۱. مقدمه

امروزه با پیشرفت تکنولوژی، استفاده از هوش مصنوعی در کاربردهای مختلف پردازش صوت افزایش چشمگیری یافته است. یکی از این کاربردها در پردازش اعداد و ارقام است که در گفتار روزمره انسان‌ها کاربردهای فراوانی در خریدوفروش و قیمت‌ها، شماره‌های تلفن، سیستم‌های رزرواسیون و رمز کارت‌های اعتباری و غیره دارند.

پژوهش‌ها و کارهای تحقیقاتی متعددی در زمینه بازشناسی اعداد انجام‌گرفته است. در سال ۱۹۹۹ بازشناسی ارقام گسسته از طریق تلفن توسط همایون‌پور و همکاران [۱] انجام شده است. در این روش ترکیب الگوریتم آموزش شبکه عصبی پرسپترون[۴] چندلایه و الگوریتم برنامه‌ریزی پویا[۵] برای بهبود کیفیت بازشناسی بکار گرفته شده است. تعداد ۴۰ گوینده برای آموزش و ۱۰ گوینده برای اعتبارسنجی روی پایگاه‌داده تلفنی متشکل از ارقام صفر تا نه مورداستفاده قرار گرفته است. نتایج بازشناسی برای داده‌های آموزشی ۹۶% و برای داده‌های اعتبارسنجی ۸۱ % بوده است. در سال ۲۰۰۱ سامانه‌ای متشکل از شبکه عصبی مدل مخفی مارکوف[6] (HMM) برای اعداد گسسته و پیوسته توسط اکبری و همکاران [۲] پیاده‌سازی شده است. در این روش پایگاه‌داده اعداد دورقمی فارسی بکار گرفته شده است و هجا به‌عنوان واحد پایه در نظر گرفته شده است. در حالت استفاده از ضرایب MFCC[7] کیفیت بازشناسی برای اعداد پیوسته ۹۸/۷ % و برای اعداد گسسته ۸۹/۷ % بوده است. در حالت استفاده از ضرایب LFCC[8] کیفیت بازشناسی برای اعداد پیوسته و گسسته به ترتیب ۹۲/۱ % و ۸۰ % است. در سال ۲۰۰۳ بررسی روش‌های مبتنی بر شبکه عصبی HMM، شبکه عصبی MLP و ترکیب آنها به‌منظور بازشناسی اعداد فارسی برای اعداد پیوسته و گسسته توسط آقای همایون‌پور و همکاران [۳] انجام شده است. در این روش ابتدا سیگنال گفتار از سکوت باهدف تعیین محدوده اعداد گسسته و تشخیص نویز زمینه تفکیک می‌شود. سپس نویز با استفاده از روش تفاضل طیفی از سیگنال صحبت جدا می‌شود. در نهایت ویژگی‌های سیگنال بدون نویز استخراج و به یک سامانه دسته‌بندی وارد می‌شود.

[1] Data augmentation
[2] Residual Convolutional neural network
[3] BiGRU
[4] MLP
[5] Dynamic programming
[6] Hidden Markov Model
[7] Mel Frequency Cepstrum Coefficients
[8] Linear Frequency Cepstral Coefficient

بهترین کیفیت بازشناسی بین شبکه‌های عصبی HMM، شبکه عصبی MLP و ترکیب آنها متعلق به شبکه عصبی HMM با مقادیر ۹۹/۱ % و ۸۳/۷ % به ترتیب برای اعداد گسسته و پیوسته در پایگاه‌داده FARSDIGIT1 [4] است .

در سال‌های اخیر بازشناسی گفتار مبتنی بر شبکه عصبی عمیق بسیار موردتوجه قرار گرفته است. در سال ۲۰۱۶ یک سامانه بازشناسی اعداد گسسته مبتنی بر شبکه عصبی عمیق DNN[1] توسط داناشری و همکاران[۵] ارائه شده است. در این روش از شبکه عصبی باور عمیق DBN[2] برای مقداردهی اولیه شبکه عصبی عمیق DNN استفاده می‌شود. کیفیت بازشناسی اعداد حاصل از این روش بر روی داده‌های انگلیسی TIDIGIT[6] ۸۶/۰۶ % است. در سال ۲۰۲۰ رحیم اله و همکاران [۷] برای بازشناسی اعداد گسسته پشتو از شبکه عصبی کانولوشنی (CNN) بهره گرفته‌اند. برای تولید بردار ویژگی از داده صوتی روش MFCC مورداستفاده قرار گرفته است و معماری شبکه عصبی بکار گرفته شده از چهار لایه کانولوشنی عمیق با تابع فعال‌سازی ReLU و لایه تجمیع بیشینه[3] تشکیل شده است. در این روش میانگین کیفیت بازشناسی اعداد پشتو ۸۴/۱۷ % است. در سال ۲۰۲۲ ویریری و همکاران [۸] روشی با تلفیق دو شبکه عصبی RNN و LSTM ارائه کردند و در ادامه روش خود را بر روی دادگان اعداد انگلیسی آزمایش کردند. آنها توانستند به‌دقت بازشناسی گفتار ۹۹% دست پیدا کنند. همچنین سوتیسنا و همکاران در سال ۲۰۲۲ [۹] از شبکه‌های یادگیری انتقالی[4] همچون AlexNet, GoogleNet به‌منظور بازشناسی گفتار استفاده نمودند. پس از مقایسه بین این دو روش، آنها به‌دقت بازشناسی گفتار ۷۲ % برای AlexNet و ۶۶ % برای GoogleNet دست پیدا کردند.

در تمامی روش‌های ارائه شده مبتنی بر شبکه عصبی عمیق تأثیر نویز بررسی نشده است. در سال ۲۰۲۱ یک شبکه مبتنی بر LSTM[5] توسط طبیبیان [۱۰] ارائه شده است که در آن تأثیر نویز بر کیفیت بازشناسی اعداد بررسی شده است. در این روش اعداد گسسته در محیط تلفنی تولید و برچسب‌گذاری شده است. این اعداد در سه دسته کلی اعداد دو و نه، اعداد صفر و سه، اعداد پنج و هفت و هشت به‌عنوان اعداد با تشابهات وجهی قرارگرفته است. شبکه عصبی LSTM توسط هر دسته به‌صورت مستقل آموزش‌داده‌شده است. کیفیت بازشناسی اعداد در شرایط بدون نویز به‌صورت میانگین ۹۱/۷ % است. سپس به داده‌های صوتی نویزهای مختلف اضافه شده است. دقت بازشناسی اعداد در شرایط نویزی به طور میانگین برابر ۶۹/۲۲% است. در سال ۲۰۲۱ ترکیب شبکه عصبی پرسپترون و روش استخراج ویژگی MTDRCC برای بازشناسی اعداد در شرایط نویزی توسط حسینی [۱۱] به‌کارگرفته‌شده است. روش استخراج ویژگی MTDRCC[6] به ترتیب شامل مراحل پیش‌پردازش، تقسیم‌بندی صوت، پنجره همینگ، تبدیل فوریه گسسته، فیلتر گوسین، تبدیل کسینوسی و معکوس تبدیل فوریه است. دقت بازشناسی اعداد فارسی در شرایط بدون نویز ۹۸/۸۵% و در شرایط نویزی ۸۸/۴۹% حاصل شده است .

به دلیل وجود تعداد داده کم برای آموزش شبکه عصبی برای زبان‌های غیر از انگلیسی، روش‌های داده‌افزایی مطرح می‌گردد. در سال ۲۰۲۲ لوناس و همکاران [۱۲] با افزودن نویز سفید، تغییر طول صوت تعداد دادگان خود را افزایش دادند. آنها از ۱۰۰ داده خام به‌عنوان داده ورودی استفاده کرده و پس از اعمال روش‌های داده‌افزایی خود بر روی دادگان از مدل مارکف برای بازشناسی آنها استفاده نمودند.

چالش اصلی در بازشناسی اعداد، تفکیک اعداد با تشابهات وجهی شامل اعداد دو و نه، اعداد سه و صفر، اعداد پنج و هفت و هشت با درنظرگرفتن شرایط نویزی است. در این مقاله روشی مبتنی بر شبکه عصبی کانولوشنی ارائه شده است که در آن از واحد کلمات به‌جای واحد واج استفاده شده است. همچنین به داده‌های صوتی ورودی، نویزهای مختلف شامل بوق یکنواخت، صدای طبیعت، صدای حرکت وسایل نقلیه، صدای همهمه و صدای کارخانه‌ها اضافه شده است.

۲. دادگان

داده‌های مورداستفاده در این پژوهش، فایل‌های صوتی دارای برچسب از ۵۱ گوینده دادگان FARSDIGIT1[4] است. فایل‌های صوتی این پایگاه‌داده شامل اعداد گسسته و پیوسته فارسی صفر تا نُه است. این پایگاه‌داده تلفنی بوده و تقریباً دارای کیفیت سیگنال به نویز ۸/۸ دسی‌بل و نرخ نمونه‌برداری ۱۱۰۲۵ هرتز است. در مجموع از هر عدد تعداد ۵۱۰ داده در دسترس بوده است. به‌منظور افزایش دادگان به جهت جلوگیری از

[1] Deep Neural Network
[2] Deep belief network
[3] Max Pooling
[4] Transfer Learning
[5] Long Short-Term memory
[6] Mel-scale Tow Dimension Root Cepstrum Coefficients

بیش‌برازش شبکه از روش‌های داده‌افزایی زیر استفاده شده است:

- کاهش و یا افزایش سرعت صوت
- اعمال فیلتر reverb
- افزودن نویز زمینه
- شبیه‌سازی محیط سالن

تعداد دادگان ورودی پس از اعمال روش‌های داده‌افزایی بالا ۵ برابر و با احتمال ۱۵ %، ۷/۵ %، ۷۰ % و ۷/۵ % به ترتیب ارائه شده در نظر گرفته شده است. نویزهای انتخابی برای افزودن به داده‌های ورودی نویز بوق، صدای طبیعت، صدای خودرو و موتور، صدای همهمه، صدای مراکز صنعتی با شدت‌های ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل انتخاب شده است. در انتها تعداد کل دادگان این پایگاه، ۲۵۵۰ داده به‌ازای هر رقم و در مجموع ۲۵۵۰۰ عدد داده است.

۳. روش پیشنهادی

روش پیشنهادی در این طرح به دو قسمت کلی، شامل استخراج ماتریس ویژگی و معماری شبکه عصبی تقسیم‌بندی می‌شود.

۳-۱ـ استخراج بردار ویژگی با استفاده از تکنیک MFCC [۱۳]

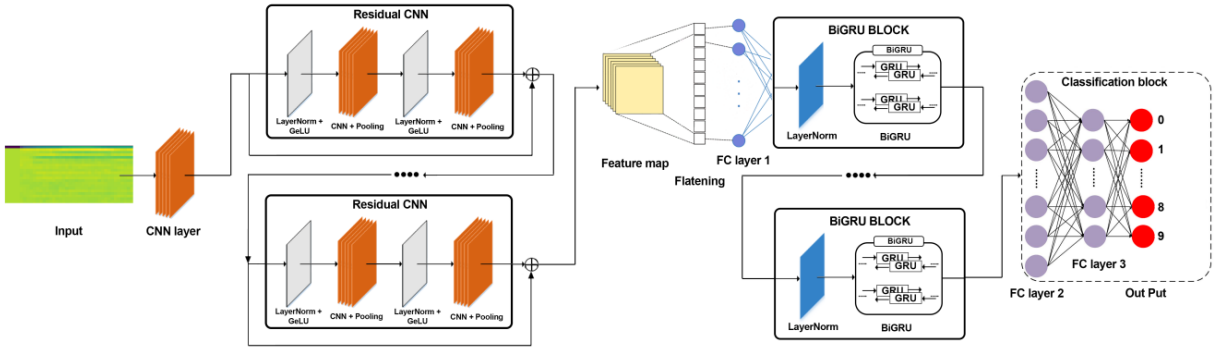
روش MFCC برای استخراج ماتریس ویژگی از سه قسمت ورودی داده، پردازش داده، خروجی ماتریس ویژگی تشکیل شده است. جهت آماده‌سازی دادگان برای روش MFCC اعداد بر اساس کلمات متناظر آنها از یکدیگر جدا و در فایل‌های مجزایی قرار می‌گیرند. پس از دریافت سیگنال دیجیتال به‌منظور یکسان‌سازی طول بردار سیگنال‌ها، ابتدا بیشترین طول بردار به‌عنوان طول مرجع در نظر گرفته می‌شود و سپس با استفاده از روش padding طول دیگر بردارها با بردار مرجع برابر می‌شوند.

در نهایت داده‌ها جهت ورود به روش MFCC آماده می‌شوند. در روش MFCC در مرحله اول یک فیلتر پیش تأکید[1] جهت تقویت فرکانس‌های بالا و متعادل‌سازی طیف فرکانسی مورداستفاده قرار می‌گیرد. این فیلتر باعث بهبود نسبت سیگنال به نویز (SNR) می‌شود. همچنین از خطاهای محاسباتی در فرایند تبدیل فوریه جلوگیری می‌کند.

سپس خروجی فیلتر به بخش‌های کوچکی تقسیم‌بندی می‌شود. برای حفظ مؤلفه‌های فرکانسی حاصل از تبدیل فوریه در مرحله بعدازاین بلوک استفاده می‌شود. در این پژوهش اندازه بخش‌های سیگنال ۲۵ میلی‌ثانیه با ۱۵ میلی‌ثانیه همپوشانی در نظر گرفته شده است. در مرحله بعد یک تابع پنجره همینگ بر روی بخش‌های کوچک سیگنال حاصل از مرحله قبل جهت بهبود طیف و کاهش نشت طیفی اعمال می‌شود. در مرحله بعد تبدیل فوریه بر روی پنجره‌ها محاسبه می‌شود و سپس طیف توان محاسبه می‌شود. در این پژوهش تعداد نمونه‌های تبدیل فوریه برابر ۵۱۲ در نظر گرفته شده است.

در ادامه مراحل، برای استخراج باندهای فرکانسی از یک بانک فیلتر با ۴۰ فیلتر مثلثی در مقیاس Mel استفاده می‌شود. استخراج باندهای فرکانسی در مقیاس Mel فرکانس‌های پایین با تراکم بیشتر را بهتر تفکیک می‌کند. برای سهولت در محاسبات از خروجی بانک فیلتر لگاریتم گرفته می‌شود. در مرحله آخر برای مقابله با همبستگی ضرایب حاصل از بانک فیلتر در الگوریتم‌های یادگیری ماشین از تبدیل کسینوسی گسسته استفاده می‌شود. در این مرحله برای حذف جزئیات اضافی دادگان ضریب کپسترال ۱۲ در نظر گرفته شده است.

۳-۲ـ معماری شبکه عصبی



شکل۱: بلوک دیاگرام شبکه عصبی پیشنهادی

[1] Pre-Emphasis

ساختار کلی شبکه عصبی در روش پیشنهادی از مدل DeepSpeech2 [۱۴] الهام گرفته شده است و در شکل (۱) ترسیم شده است. این شبکه عصبی به ترتیب از یک لایه CNN، سه بلوک Residual CNN، یک لایه Fully Connected، پنج بلوک BiGRU و دو لایه Fully Connected تشکیل شده است. لایه CNN برای تبدیل داده صوتی ورودی به بردار ویژگی اولیه و هم چنین تغییر ابعاد ماتریس ورودی بکار گرفته می‌شود. بلوک Residual CNN برای یادگیری ویژگی‌های صوتی استفاده می‌شود. لایه‌های این بلوک‌ها جزئیات بیشتری از دادگان صوتی را نسبت به لایه‌های CNN مورداستفاده قرار می‌دهند و یادگیری بهتری نسبت به آنها دارند [۱۵]. لایه Fully Connected برای تغییر ابعاد ماتریس ویژگی خروجی از Residual CNN مورداستفاده قرار می‌گیرد. خروجی لایه Fully Connected به بلوک BiGRU وارد می‌شود. GRU نوع بهبودیافته شبکه بازگشتی RNN است. شبکه‌های RNN ویژگی‌های صوتی را مرحله‌به‌مرحله پردازش می‌کنند و در پردازش هر فریم از دادگان فریم‌های قبلی استفاده می‌کنند. شبکه بازگشتی BiRNN برای پردازش هر فریم از دادگان فریم‌های قبلی و بعدی استفاده می‌شود و پیش‌بینی دقیق‌تری برای شبکه به دنبال دارد. شبکه GRU در روش پیشنهادی از منابع محاسباتی کمتری نسبت به شبکه LSTM استفاده می‌کند و در برخی موارد عملکرد بهتری دارد.

۳-۲-۱ــ بلوک شبکه عصبی کانولوشنال باقیمانده:

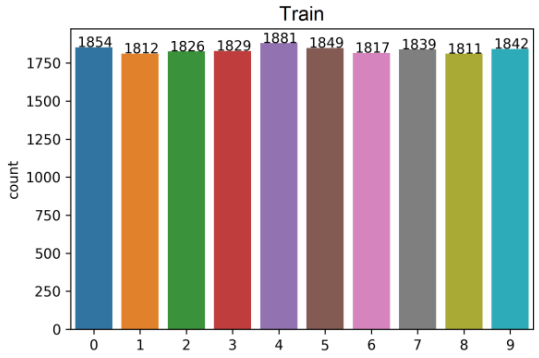
همان گونه که در بلوک دیاگرام شکل (۱) آمده است این بلوک از دو لایه کانولوشنی همراه با لایه تجمیع بیشینه و لایه‌های نرمال کننده تشکیل شده است. دلیل استفاده از لایه نرمال کننده کاهش حجم محاسباتی است. شبکه‌های عصبی کانولوشنی مورداستفاده در این بلوک دارای کرنل سایز ۳ و padding موردنیاز به جهت جلوگیری از تغییر ابعاد و ابعاد ورودی و خروجی شبکه ۳۲ می‌باشد.

۳-۲-۲ــ بلوک شبکه عصبی واحد بازگشتی گیتی دوطرفه:

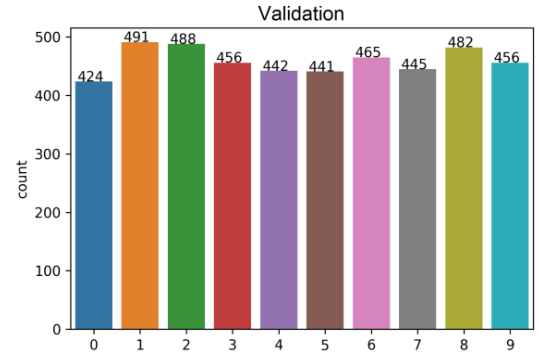
مطابق شکل (۱) این بلوک از شبکه عصبی واحد بازگشتی گیتی و یک لایه نرمال[1] کننده تشکیل شده است. شبکه عصبی مذکور شامل یک لایه ورودی با اندازه ۵ و ده لایه پنهان است.

۴. پیاده‌سازی روش پیشنهادی و تجزیه‌وتحلیل نتایج

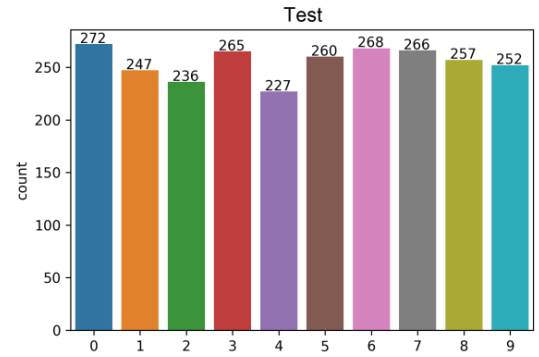
به‌منظور پیاده‌سازی روش پیشنهادی ابتدا دادگان به سه دسته آموزش، اعتبارسنجی، آزمایش تقسیم شدند. شکل‌های (۲ و ۳ و ۴) به ترتیب نمودار فراوانی هر کلاس از اعداد در دسته‌های آموزش[2]، اعتبارسنجی[3] و آزمایش[4] را بیان می‌کند.



شکل ۲: نمودار فراوانی هر کلاس در دادگان آموزش



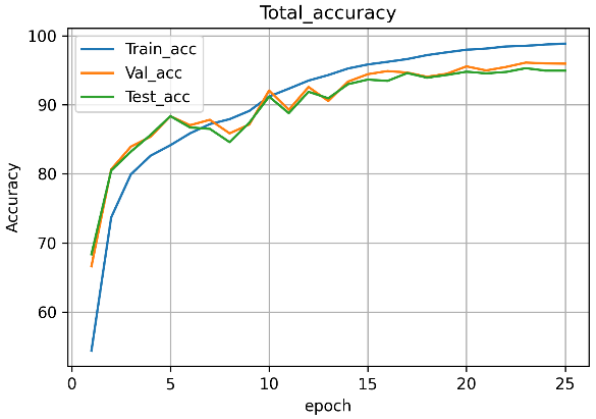
شکل ۳: نمودار فراوانی هر کلاس در دادگان اعتبارسنجی



شکل ۴: نمودار فراوانی هر کلاس در دادگان آزمایش

[1] NormLayer
[2] Train
[3] Validation
[4] Test

این تقسیم‌بندی بعد از ورود همه اعداد همراه با برچسب‌ها به یک متغیر و درهم ریختن آن انجام‌گرفته است. همان‌طور که در بخش قبل بیان شد ماتریس ویژگی‌های دادگان ورودی با استفاده از تکنیک MFCC استخراج شد. برای آموزش شبکه عصبی از سرویس Colab گوگل استفاده شده است. نسخه رایگان این سرویس کارت گرافیک Tesla T4 همراه با ۱۵/۱ گیگابایت رم و همچنین حافظه داخلی برابر با ۷۸/۱۹ گیگابایت جهت ذخیره اطلاعات در محیط کاری و رم به میزان ۱۲/۶۸ گیگابایت جهت انجام پردازش‌ها در اختیار این پژوهش قرار داده است. در ابتدا به بررسی شبکه‌های ساده مثل شبکه عصبی کانولوشنی و یا شبکه عصبی GRU مورد بررسی قرار گرفت که به دلیل کوچک بودن شبکه توانایی یادگیری خوب دادگان نویزی ما را نداشتند و دقت آنها به ترتیب بر روی دادگان اعتبارسنجی برای شبکه عصبی کانولوشنی و GRU برابر ۸۳/۲۲ % و ۷۸/۷۶% شد. در ادامه به بررسی کارکرد شبکه عصبی LSTM بر روی دادگان خود پرداختیم که دقتی برابر ۸۷/۸۶ % به دست آمد که به نسبت دو شبکه قبلی دارای دقت بهتری بود.



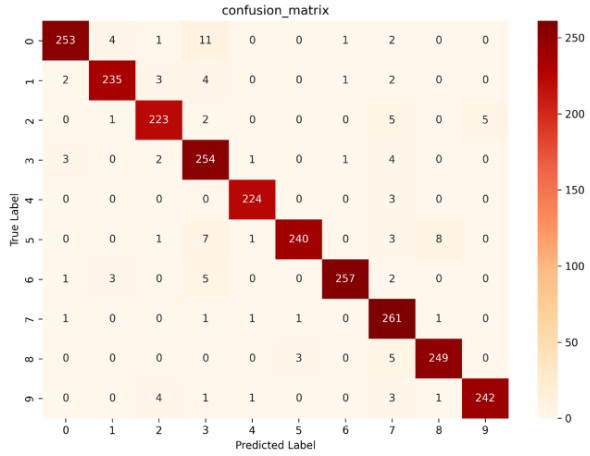
شکل ۵: نمودار دقت بازشناسی بر دادگان آموزش، اعتبارسنجی و آزمایش

مطابق شکل (۵) پس از آزمایش شبکه عصبی پیشنهادی بر روی دادگان تهیه شده با ۲۵ دوره آموزشی به‌دقت ۹۸/۵۳% برای دادگان آموزش و دقت ۹۶/۱۰% برای دادگان اعتبارسنجی دست یافتیم. همچنین دقت شبکه بر روی دادگان آزمایش برابر ۹۵/۹۲ % به دست آمد که مقایسه آن با دیگر شبکه‌های مورد آزمایش در جدول (۱) بیان شده است.

شکل (۶) ماتریس درهم‌ریختگی را نشان می‌دهد. همان‌طور که در شکل مشخص است خطای شبکه بر روی دادگانی که دارای تشابهات وجهی هستند وجود دارد؛ ولی میزان خطای آن به نسبت شبکه‌های مبتنی بر واجی بسیار کمتر است.

جدول ۱: مقایسه شبکه‌های LSTM, CNN, GRU و شبکه پیشنهادی بر روی دادگان تهیه شده

نام شبکه	LSTM	CNN	GRU	شبکه پیشنهادی
دادگان آموزش	۹۱/۱۶	۷۶/۰۵	۸۲/۹۰	۹۸/۵۳
دادگان اعتبارسنجی	۸۷/۴۵	۸۳/۲۲	۸۰/۳۴	۹۶/۱۰
دادگان آزمایش	۸۶/۸۲	۸۳/۴۹	۸۰/۵۹	۹۵/۹۲



شکل ۶: ماتریس درهم‌ریختگی در آخرین دوره دادگان آزمایش

۵. جمع‌بندی

در روش پیشنهادی شبکه‌های عصبی Residual CNN و BiGRU برای بازشناسی اعداد فارسی صفر تا نُه گسسته در شرایط نویزی بکار گرفته شد و برای مقابله با تشابهات وجهی و بهبود کیفیت استخراج ویژگی‌های اعداد، از واحد کلمه به‌جای واحد واج استفاده گردید. شبکه عصبی Residual CNN به دلیل ادغام ماتریس ویژگی حاصل با ماتریس داده ورودی متناظر، جزئیات بیشتری از داده‌های صوتی را استخراج می‌کند. هم چنین شبکه عصبی BiGRU علاوه بر پیش‌بینی دقیق‌تر به‌منظور استفاده از دادگان فریم‌های قبلی و بعدی، دارای بار محاسباتی کمتری نسبت به شبکه عصبی LSTM است. نتایج حاصل از آزمایش‌ها نشان می‌دهند که دقت بازشناسی حاصل از این روش برای دادگان آموزش ۹۸/۵۳%و برای داده‌های اعتبارسنجی۹۶/۱۰ % است. مطابق با این نتایج در شرایط نویزی، روش پیشنهادی به‌صورت متوسط ۲۶/۸۸ % عملکرد بهتری نسبت به روش‌های مبتنی بر واحد واج و LSTM بر روی اعداد فارسی در محیط نویزی و ۷/۶۱% عملکرد بهتری نسبت به روش‌های متفاوت در استخراج ویژگی داشته است.

قدردانی

از شرکت عصر گویش پرداز به جهت در اختیار قراردادن دادگان خود به‌منظور انجام آزمایش‌ها این پژوهش قدردانی می‌گردد.

مراجع

[۱] م. م. همایون‌پور و ا. نجاری، "بازشناسی ارقام نا وابسته به گوینده با استفاده از مدل پیشگوی عصبی، " مجموعه مقالات هفتمین کنفرانس مهندسی برق ایران، صص ۸۱ـ ۷۵، تهران، ایران، ۲۹ـ۲۷ اردیبهشت ۱۳۷۸.

[۲] ا. اکبری و ب. ناصر شریف، "بازشناسی هجاها در اعداد دورقمی فارسی به‌وسیله مدل مخفی مارکف، " مجموعه مقالات ششمین کنفرانس سالانه انجمن کامپیوتر ایران، صص. ۴۳۷ـ۴۳۲، اصفهان، ایران، ۴ ـ۲ اسفند ۱۳۷۹.

[۳] 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 آـ۵۶، صص. ۱۰۶۵ـ۱۰۴۵، پاییز ۱۳۸۲.

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

[۵] D. Dhanashri and S. Dhonde, "Isolated word speech recognition system using deep neural networks," in Proceedings of the international conference on data engineering and communication technology, 2017: Springer, pp. 9-17.

[۶] R. Leonard and G. Doddington, "TIDIGITS dataset," Linguistic Data Consortium, Philadelphia, 1993.

[۷] B. Zada and R. Ullah, "Pashto isolated digits recognition using deep convolutional neural network," Heliyon, vol. 6, no. 2, p. e03372, 2020.

[۸] J. Oruh, S. Viriri, and A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," IEEE Access, vol. 10, pp. 30069-30079, 2022.

[۹] C. Amadeus, I. Syafalni, N. Sutisna, T. Adiono, "Digit-Number Speech-Recognition using Spectrogram-Based Convolutional Neural Network." In 2022 International Symposium on Electronics and Smart Devices (ISESD), 2022: IEEE, pp. 1-6.

[۱۰] ش. طبیبیان، "بازشناسی مقاوم به نویز ارقام مشابه فارسی مبتنی بر شبکه LSTM و ویژگی‌های طیفی گفتار."، مجله مهندسی برق و مهندسی کامپیوتر، سال نوزدهم، شماره ب_۱، صص ۱۷ـ۱، بهار ۱۴۰۰.

[۱۱] S. M. Hoseini, "Recognition of Persian Digits from Zero to Nine using Acoustic Images based on Mel Capstrom Coefficients and Neural Network." in International Journal of Mechatronics, Electrical and Computer Technology (IJMEC), Vol. 11, Oct. 2021, pp. 5059-5064

[۱۲] K. Lounnas, M. Lichouri, and M. Abbas, "Analysis of the Effect of Audio Data Augmentation Techniques on Phone Digit Recognition For Algerian Arabic Dialect," in 2022 International Conference on Advanced Aspects of Software Engineering (ICAASE), 2022: IEEE, pp. 1-5.

[۱۳] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," arXiv preprint arXiv:1003.4083, 2010.

[۱۴] D. Amodei et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in International conference on machine learning, 2016: PMLR, pp. 173-182.

[۱۵] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.