مقاله پژوهشی

بازشناسی احساسات از روی گفتار بر پایه بهره‌گیری از شبکه‌های عصبی پیچشی و روش افزایش دادگان

معصومه شفیعیان*
استادیار، دانشکده فنی و مهندسی رسانه، دانشگاه صدا و سیما
shafieian@iribu.ac.ir

وحید احمدیان
کارشناسی ارشد مهندسی صدا، دانشکده فنی و مهندسی رسانه، دانشگاه صدا و سیما
vahidahmadianbonab@gmail.com

مجید بهداد
استادیار، دانشکده فنی و مهندسی رسانه، دانشگاه صدا و سیما
behdad@iribu.ac.ir

تاریخ دریافت: ۱۴۰۱/۰۱/۱۱ تاریخ پذیرش: ۱۴۰۱/۰۵/۳۱

چکیده

هدف از سیستم‌های بازشناسی احساس از روی گفتار ایجاد ارتباط عاطفی بین انسان و ماشین است. چراکه بازشناسی احساس و اهداف انسان از روی گفتار، به بهبود تعاملات بین انسان و ماشین کمک می‌کند. بازشناسی احساس از روی گفتار برای محققان در دهه گذشته یک مسأله چالش‌برانگیز بوده است. اما با پیشرفت در حوزه هوش مصنوعی این چالش‌ها کم‌رنگ‌تر شدند. هدف از این پژوهش، استفاده از روش‌های یادگیری عمیق در جهت بهتر کردن کارایی این سیستم‌ها است. کار انجام شده از چندین مرحله تشکیل شده است. در مرحله اول از شبکه‌های عصبی پیچشی سه بعدی برای یادگیری ویژگی‌های طیفی زمانی گفتار استفاده شده است. در مرحله دوم برای قدرتمند کردن مدل پیشنهادی از ساختار هرمی جدید شبکه‌های عصبی پیچشی سه بعدی اتصال داده شده؛ که یک معماری چند مقیاسه از شبکه‌های عصبی پیچشی سه بعدی روی ابعاد ورودی است، بهره گرفته شد. در نهایت برای یادگیری ویژگی‌های طیفی زمانی استخراج شده از ساختار جدید (ساختار جدید هرمی شبکه‌های عصبی پیچشی سه بعدی) با درنظر گرفتن رابطه مکانی و زمانی اطلاعات به‌صورت کامل، از شبکه کپسول زمانی استفاده شد. در نهایت بر ساختار پیشنهادی که یک ساختار قدرتمند برای ویژگی‌های طیفی زمانی است نام MSID 3DCNN + Temporal Capsule نهاده شد. پژوهش انجام شده و مدل نهایی بر روی ترکیب دو پایگاه داده گفتار معمولی و گفتار آوازی از پایگاه داده راودیس که یک پایگاه داده چند حالته است انجام شد. نتایجی که با استفاده از مدل پیشنهادی به‌دست آمد؛ نسبت به مدل‌های مرسوم، قابل توجه است. در این پژوهش برای شش کلاس احساسی به تفکیک جنسیت، دقت ۸۱/۷۷ درصد به‌دست آمد.

واژگان کلیدی: بازشناسی احساس از روی گفتار، شبکه‌های عصبی پیچشی سه بعدی چند مقیاسه، شبکه کپسول زمانی، پایگاه داده راودیس.

۱. مقدمه

با وجود پژوهش‌های گسترده، چالش‌های فراوانی در سیستم‌های بازشناسی احساس از روی گفتار وجود دارد. با توجه به اینکه احساس انسان پدیده‌ای پیچیده، مبهم و مرکب است؛ در اغلب مواقع در هنگام برقراری ارتباط بین افراد، احساس‌های کامل، پایه و خالص بروز نمی‌کنند؛ بلکه در یک لحظه ممکن است ترکیبی از چند احساس بروز داده شود [۵]. ازاین‌رو گاهی جداسازی، تشخیص و تشریح محتوای احساسی گفتار، حتی توسط عوامل انسانی بسیار مشکل است. علاوه‌بر آن نحوه‌ی بروز احساس‌ها در گفتار به فرهنگ، زبان، سن، جنسیت گوینده و بسیاری عوامل دیگر وابسته است [۶]. تمامی موارد گفته شده مسأله بازشناسی احساس از روی گفتار را پیچیده‌تر می‌کند.

اگر در یک دید کلی به سیستم‌های بازشناسی احساس از روی گفتار نگریسته شود؛ می‌توان دریافت که این سیستم‌ها با سه چالش اصلی پایگاه داده، نوع ویژگی استخراج شده و دسته‌بند استفاده شده، روبه‌رو هستند. نحوه حل این چالش‌ها در بهبود میزان صحت نتایج این سیستم‌ها بسیار مؤثر است. در ادامه به توضیح این چالش‌ها پرداخته می‌شود:

• چالش انتخاب پایگاه داده

پایگاه داده، بخش مهمی از سیستم‌های شناخت احساس از روی گفتار است. زیرا فرایندهای دسته‌بندی به داده‌های برچسب‌دار وابسته است. از طرف دیگر کیفیت داده‌ها با دقت این سیستم‌ها رابطه مستقیم دارد؛ چراکه از داده‌های ناقص یا کم‌کیفیت به پیش‌بینی‌های نادرست می‌رسیم. بنابراین اولین مسأله برای رسیدن به نتیجه بهتر در این سیستم‌ها، بهره‌گیری از پایگاه داده مناسب است [۱-۳].

یکی دیگر از چالش‌های مربوط به دادگان، چالش چند زبانی است. پایگاه‌های داده امروزی با توجه به وقت‌گیر بودن جمع‌آوری داده، بیشتر به یک زبان خاص جمع‌آوری شده‌اند. لذا بیشتر سیستم‌های موجود برای یک زبان خاص به‌وجود آمده‌اند. برای غلبه بر این چالش باید پایگاه‌های داده بزرگی که شامل چندین زبان باشد را جمع‌آوری کنند.

• چالش استخراج ویژگی از سیگنال گفتار

ویژگی‌ها، جنبه مهمی از اساس کار شناخت احساس از روی گفتار را بیان می‌کنند. تاکنون ویژگی‌های بسیاری در سیستم‌های بازشناسی احساس از روی گفتار استفاده شده است. با این حال هیچ ویژگی یا گروهی از ویژگی‌ها برای دسته‌بندی احساس وجود ندارد که سیستم‌ها بهترین دقت را در آن ویژگی یا ویژگی‌ها داشته باشند و این مطالعات تاکنون صرفاً تجربی بوده است. از طرف دیگر، گفتار یک سیگنال پیوسته با طول متغیر است؛ که هم دارای اطلاعات زبانی و معادل متنی و هم دارای احساس است. ویژگی مورد نظر برای بازشناسی احساس نیز متناسب با رویکردی که وجود دارد، انتخاب می‌شود. لذا انتخاب اینکه از چه ویژگی یا ویژگی‌هایی استفاده شود تا دقت سیستم‌های بازشناسی احساس از روی گفتار بهبود یابد. یکی دیگر از ضرورت‌های حل مسأله است [۱-۳].

• چالش دسته‌بند

اما اینکه استفاده از کدام دسته‌بند می‌تواند در سیستم‌های بازشناسی احساس از روی گفتار تأثیر مثبتی داشته باشد و اینکه کار به نتیجه مطلوبی برسد یکی دیگر از چالش‌های این سیستم‌ها است. در ادامه توضیح داده خواهد شد که با توسعه پژوهش‌ها در حوزه هوش مصنوعی، مخصوصاً یادگیری عمیق، چطور می‌توان بر این چالش غلبه کرد و سیستم‌های مقاوم‌تر و بهتری را ارائه داد. در این مقاله یک سیستم بازشناسی احساس از روی سیگنال گفتار معرفی می‌شود که در آن برای بهبود نتیجه‌ی بازشناسی احساس، از مدل‌های مبتنی بر شبکه‌های عصبی پیچشی سه بعدی با معماری جدید (هرمی)، همراه با شبکه‌های کپسول زمانی استفاده شده است.

در سال‌های اخیر عملکرد الگوریتم‌های یادگیری عمیق از الگوریتم‌های یادگیری ماشین بالاتر رفته است. از الگوریتم‌های یادگیری عمیق که در حوزه تشخیص احساس از روی گفتار به‌صورت وسیع استفاده می‌شوند، می‌توان به شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی اشاره کرد [۱-۴]. در ادامه ایده‌هایی که اخیرا در این حوزه ارائه شده است مرور خواهد شد.

ایده ترکیبی در دادگان: در مرجع [۸] که توسط کارول در سال ۲۰۲۰ انجام شد است، از ترکیب چهار پایگاه داده به زبان‌های لیتوانی و انگلیسی، آلمانی و اسپانیایی استفاده شد که نتایج سیستم چند زبانه در مدل مبتنی بر CNN[۱] نزدیک به حالت‌های تک زبانه بود. این ایده را می‌توان به این صورت تعریف کرد که هدف از این ایده، قدرتمند ساختن سیستم‌های تشخیص احساس از روی گفتار نسبت به چالش چند زبانی است. چراکه با توجه به نبود دادگان لازم، سیستم‌های تشخیص احساس از روی گفتار امروزی بیشتر برای یک زبان خاص هستند. لذا ترکیب کردن چند دادگان و قدرتمند ساختن این سیستم‌ها، یکی از ایده‌های اصلی محققان حوزه سیستم‌های تشخیص احساس از روی گفتار است.

ایده هم‌آمیزی در دادگان: با توجه به مرجع [۹] که توسط ناین و همکاران در سال ۲۰۱۸ انجام شده است؛ این ایده را می‌توان به این صورت تعریف کرد که هدف از این ایده، رسیدن به دقت بهتر در سیستم‌های تشخیص احساس و عملی کرد آنها است. چراکه در عمل فقط با صدای فرد موردنظر مواجه نیستیم. عواملی مانند چهره، عملکرد فرد و موارد دیگر در تشخیص احساس وی تأثیرگذار است. لذا پژوهش فوق با معرفی یک سیستم هم‌آمیزی دادگان، با استفاده از ویدیو و صدا با استفاده از CNN+DNN[2]، سعی در بهتر کردن سیستم‌های تشخیص احساس داشت.

ایده ترکیبی در ویژگی‌ها: با توجه به مرجع [۱۰] که توسط لین و همکاران در سال ۲۰۱۹ انجام شده است در این پژوهش از ترکیب ویژگی MFCC[3] و انرژی و دسته‌بندهای SVM[4] و GB[۵] استفاده کرده‌اند که نتیجه SVM بهتر بود. در این ایده می‌توان گفت هرقدر تعداد ویژگی‌های به‌کار رفته در سیستم‌های تشخیص احساس از روی گفتار بیشتر باشد دقت این سیستم‌ها افزایش می‌یابد. لذا ایده ترکیب کردن چندین ویژگی، می‌تواند تأثیر مثبتی در این سیستم‌ها داشته باشد.

ایده چند پنجره در ویژگی: با توجه به مرجع [۱۱] که توسط چاپنری و همکاران در سال ۲۰۱۵ انجام شده است؛ این ایده را می‌توان به این صورت تعریف کرد؛ زمانی‌که ما برای استخراج ویژگی از یک پنجره استفاده می‌کنیم، باعث کاهش بایاس سیستم می‌شود. اما به دلیل واریانس بالا، نتایج ضعیفی حاصل می‌شود. زمانی‌که از چند پنجره استفاده می‌کنیم واریانس را نیز کاهش می‌دهیم و دقت بهتری را به‌دست می‌آوریم. لذا استفاده کردن از سیستم‌های چند پنجره را پیشنهاد می‌دهند.

در مرجع [۱۲] که توسط بدشاه در سال ۲۰۱۷ انجام شده است. از ویژگی STFT[۶] و دسته‌بند مبتنی‌بر CNN و پایگاه داده برلین استفاده کرده است. نتایج به‌دست آمده بهتر از روش‌های مرسوم یادگیری ماشین بود. در مرجع [۱۳] که توسط کامبهار و همکاران در سال ۲۰۱۹ انجام شده است، از ویژگی MFCC و دسته‌بند مبتنی بر LSTM[۷] و دادگان RAVDESS بهره گرفته‌اند. نتیجه این شد که برای دسته‌بندی داده‌های پیچیده از LSTM استفاده کنند.

با توجه به مرجع [۱۴] که توسط اتینه و همکاران انجام شده است نیز می‌توان گفت ترکیب کردن ایده در یادگیری عمیق می‌تواند به سیستم‌های تشخیص احساس از روی گفتار کمک مضاعفی کند. برای مثال در مرجع [۱۴] از ترکیب کردن CNN با LSTM سیستمی با دقت بهتر ایجاد کردند. چراکه از CNNها برای استخراج ویژگی‌های سطح بالا بهره گرفتند اما یکی از مشکلات مهم CNNها از دست رفتن اطلاعات زمانی بود. برای درنظر گرفتن

اطلاعات زمانی از LSTMها بهره بردند. به این صورت وابستگی‌های طولانی مدت و وابستگی‌های کوتاه مدت زمانی در ویژگی‌های استخراج شده را نیز مدل کردند و به نتایج بهتری رسیدند.

ایده مدرن در دسته‌بند: بیشترین سهم در ایده‌های مدرن، مربوط به یادگیری عمیق است. در ادامه به چند مورد آن اشاره می‌کنیم. در مرجع [۱۵] که توسط گیوزو در سال ۲۰۲۰ انجام شده به روش چند مقایسه اشاره شده است. در این روش باید ابتدا گفت روش‌های موجود در سیستم‌های تشخیص احساس یک واقعیت را درنظر نمی‌گیرند. آنکه سیگنال‌های گفتاری غالبا در مقیاس‌ها و فرکانس‌های زمانی متفاوت، نسبت به شکل خام، ویژگی‌های متفاوتی از خود نشان می‌دهند. لذا پیشنهاد می‌شود یادگیری در مقیاس‌های مختلف انجام شود. در مقاله فوق نیز سیستم‌های چند مقیاسه زمانی پیشنهاد شده است. در مرجع [۱۶] که توسط لین و همکاران در سال ۲۰۱۹ انجام شده، به مکانیزم توجه اشاره شده است. در این روش، باید گفت که، سیگنال گفتار، فقط دارای گفتار نیست. از قسمت‌هایی مانند؛ سکوت، گفتار و نوفه تشکیل شده است. برای بهتر کردن این سیستم‌ها بعضی اوقات از فیلترهای حذف نوفه یا از تکنیک حذف سکوت استفاده می‌کنند. اما یکی دیگر از روش‌ها، مکانیزم توجه است. در مکانیزم توجه، قسمت‌های مهم سیگنال، که در آنها گفتار وجود دارد پر رنگ‌تر از قسمت‌های کم اهمیت می‌شود (وزن بیشتری داد می‌شود). بدین‌گونه کارایی سیستم بهتر می‌شود. در مرجع [۱۷] که توسط استولار و همکاران در سال ۲۰۱۷ انجام شده، استفاده از روش انتقال یادگیری را پیشنهاد می‌دهند. همان‌طورکه می‌دانید پایگاه‌های داده موجود، از لحاظ تعداد داده کم هستند. برای غلبه بر این موضوع، یکی از روش‌ها، تکنیک افزایش دادگان است. اما روش تا حدودی نتیجه‌بخش بود. لذا روش دیگری به نام انتقال یادگیری به‌وجود می‌آید. در انتقال یادگیری از مدل‌هایی که قبلا روی پایگاه‌های داده بزرگ آموزش داده شده‌اند استفاده می‌کنند. این روش، تأثیر مثبتی روی گرفتن نتیجه مطلوب از داده‌هایی با تعداد کم، مخصوصا داده‌های تصویری دارد.

۲. روش پیشنهادی

چگونگی بهره‌گیری از روش‌های یادگیری عمیق با توجه به ضعفی که در سیستم‌های بازشناسی احساس از روی گفتار شناسایی شد، اهمیت دارد. به‌عنوان مثال ضعفی که شبکه‌های عصبی پیچشی دو بعدی دارند از دست دادن اطلاعات زمانی است. در روش پیشنهادی این ضعف توسط شبکه‌های عصبی پیچشی سه بعدی برطرف می‌شود. از ضعف‌های دیگر پژوهش‌های انجام شده، بحث کار کردن در چند مقیاس است. چراکه سیستم‌ها در مقیاس‌های متفاوت ویژگی‌های متفاوتی از خود نشان می‌دهند. برای کم‌رنگ کردن این موضوع و قوی‌تر کردن مدل پیشنهادی، در این پژوهش از ساختار هرمی جدیدی از ۳D CNNهای (ساختار چند مقیاسه) بهره برده می‌شود. در نهایت برای اینکه خروجی حاصل از این مدل با حفظ موقعیت زمانی و مکانی یاد گرفته شود، از شبکه کپسول زمانی استفاده می‌شود. شبکه کپسول زمانی، همان شبکه کپسولی است با این تفاوت که لایه کانوولوشنی ابتدایی در این شبکه تغییر داده شده و از لایه ConvLSTM2D[۸] استفاده شده است، تا بتوان خروجی سه بعدی شبکه چند مقیاسه را با یادگیری فریم‌های بلند مدت مفید و مهم، به خروجی دو بعدی برای ورود به شبکه کپسول، تبدیل کرد.

یکی از قسمت‌های مهم در سیستم‌های بازشناسی احساس از روی گفتار، انتخاب دادگان مناسب است. در این پژوهش برای آموزش و آزمودن مدل از پایگاه گفتاری داده‌ای که محصول سال ۲۰۱۸ بوده و توسط لیوینگ استون و همکارش به‌وجود آمده، استفاده شده است که یک پایگاه داده‌ای جدید معتبر و چند حالته از گفتار و آواز احساسی از نوع صدای بازیگری محسوب می‌شود و به همین دلیل جدید

بودن، پژوهش‌های کمتری از این پایگاه بهره برده‌اند. این پایگاه داده با به‌کارگیری ۲۴ بازیگر حرفه‌ای با تفکیک جنسیت زن و مرد به‌وجود آمده است. حالت‌های احساسی موجود در آن از ۸ حالت شادی، ناراحتی، عصبانیت، ترس، شگفت‌زدگی، انزجار، آرام بودن، طبیعی بودن ایجاد شده است. هر بازیگر دو جمله زیر را با احساس‌های متفاوتی گفته است:

Kids are talking by the door.

Dogs are sitting by the door.

عبارات فوق با دو شدت نرمال و قوی برای هر احساس گفته شده و ضبط شده است به جز حالت طبیعی که فقط در شدت نرمال گفته و ضبط شده است [۱۸].

در این پژوهش، از هر دو دادگان گفتاری و آوازی استفاده شده دادگان گفتار دارای ۱۴۴۰ داده و پایگاه داده آوازی دارای ۱۰۱۲ داده است. لذا با توجه به مطلبی که درمورد انواع ایده‌های موجود در سیستم‌های تشخیص احساس بیان شد. یکی از ویژگی‌های سیستم پیشنهادی، ترکیب داده‌های گفتاری و آوازی است.

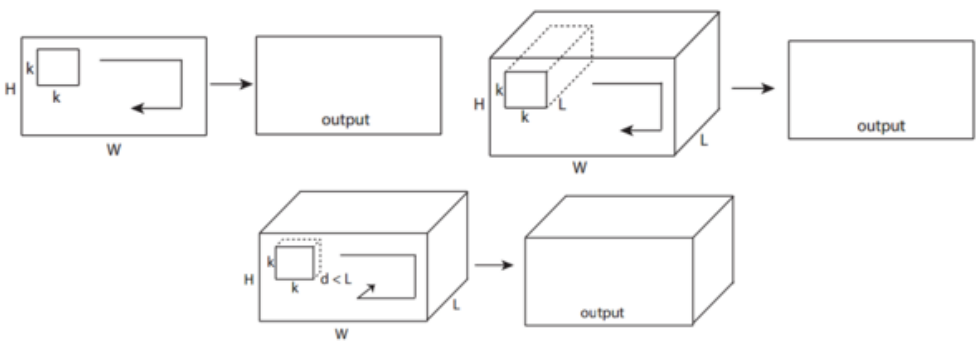
از نظر ماهیت نیز این دادگان، یک دادگان تقریبا کامل است زیرا وابسته به جنیست خاصی نیست و از طیف گسترده‌ای از احساس و در سطح شدت‌های مختلف ساخته شده و برای هرکلاس از تعداد مجموعه داده برابر استفاده شده است.

همچنین در آزمایش‌هایی که توسط شنونده انسان انجام شده، احساس‌های آرام و طبیعی خیلی سخت از هم تشخیص داده می‌شوند. ازاین‌رو این دو کلاس، یک کلاس درنظر گرفته شدند [۱۹].

در ادامه مراحل مختلف انجام روش پیشنهادی به تفصیل بیان می شود.

۲-۱. ویژگی‌های طیفی زمانی و D CNN۳

در این مرحله، فایل صوتی دریافت می‌شود اما یک سری تفاوت‌هایی در اینجا نسبت به کارهای قبلی وجود دارد. داده صوتی ورودی، باید شکل سه بعدی به خود بگیرد. به‌عبارتی از داده خام ویژگی طیفی زمانی استخراج شود. چراکه با استفاده از مرجع [۲۰] این نتیجه به‌دست آمد که در استخراج ویژگی‌های سطح بالا توسط D CNN۲ ایرادی وجود دارد. این شبکه‌ها نمی‌توانستند رابطه زمانی بین ویژگی‌ها را یاد بگیرند. برای این کار نیاز به ترکیب شدن با LSTMها دارند. در این پژوهش نیز در جهت رفع این مشکل و به‌دست آوردن ویژگی‌های سطح بالا با حفظ ترتیب زمانی و با الهام گرفتن از مرجع [۲۰] از D CNN۳ها استفاده می‌شود. به‌عبارتی دیگر ویژگی‌های سطح بالای طیفی-زمانی توسط این شبکه‌ها به‌دست می‌آید.



شکل ۱. تفاوت شبکه‌های سه‌بعدی با دو بعدی ؛ بالا سمت چپ: کانولوشن دو یعدی، بالا سمت راست: کانولوشن دوبعدی چندین فریم، پایین: کانولوشن سه بعدی

همان‌طورکه در شکل ۱ مشاهده می‌شود از سمت چپ، شکل اول از شبکه پیچشی دو بعدی برای ایجاد نقشه ویژگی استفاده شده است. در شکل دوم از داده‌های ویدئویی استفاده شده است و بعد سوم تعداد کانال درنظر گرفته شده است (فریم‌ها روی هم افتاده‌اند) و لذا از شبکه دو بعدی استفاده شده و در نهایت نقشه ویژگی دو بعدی ایجاد شده است. در شکل سوم، فریم‌ها به‌صورت حجم پشت سر هم افتاده‌اند و تعداد کانال، یک درنظر گرفته شده است. اما از شبکه سه بعدی استفاده شده است.

حال که دلیل استفاده از 3D CNN گفته شد نحوه آماده‌سازی داده برای ورود به این شبکه‌ها توضیح داده می‌شود.

مرحله اول: داده گفتار وارد مرحله پردازش و استخراج ویژگی می‌شود تا آماده ورود به مدل شود. در این مرحله سیگنال پیوسته گفتار با نرخ نمونه‌برداری ۴۴۱۰۰ هرتز تبدیل به سیگنال دیجیتال می‌شود.

مرحله دوم: سیگنال دیجیتال فوق توسط یک فیلتر زمانی به چندین فریم بلند مدت ۳۰۰ میلی‌ثانیه با هم‌پوشانی ۵۰ درصد تقسیم می‌شود. در این قسمت هدف برقرار کردن ارتباط بلند مدت زمانی ویژگی‌هاست.

مرحله سوم: بعد از هرکدام از این گفتارهای فریم شده (۳۰۰ میلی‌ثانیه) ویژگی طیفی Log-Mel به‌دست می‌آید. در هر کدام از ویژگی‌ها، بعد اول طیف و بعد دوم زمان را نشان می‌دهد. در این مرحله، هدف برقرار کردن ارتباط کوتاه مدت زمانی ویژگی‌هاست.

مرحله چهارم: حال این ویژگی‌ها برای رعایت ترتیب اطلاعات زمانی (بعد سوم شامل ترتیب زمانی می‌شود) پشت سر هم قرار می‌گیرند. بدین گونه ویژگی طیفی زمانی لگاریتم مدل استخراج می‌شود.

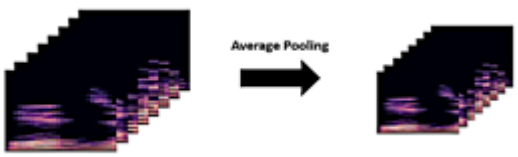
مرحله پنجم: داده آماده ورود به شبکه 3D CNN است.

بعد از ایجاد سیستم ورودی به همراه یک سری ویژگی‌ها، ابتدا در هر تصویر یا به‌عبارتی در هر طیف، ارتباط کوتاه مدت بین ویژگی‌ها در طول زمان را می‌توان دید. سپس بعد سوم ارتباط طولانی مدت بین ویژگی‌ها در طول زمان را نشان می‌دهد.

۲-۲. رویکرد هرمی جدید از 3D CNN[۹]

این پژوهش فقط به نحوه و چینش ویژگی‌های استخراج شده برای ورود به شبکه 3D CNN و استفاده از مدل ساده آن محدود نمی‌شود. به‌منظور اینکه بتوان مدل قدرتمندتری را ایجاد نمود از ایده مرجع [۲۱] که پژوهشی در حوزه تشخیص صداهای محیطی است بهره برده شد.

به این صورت که از ورودی به‌دست آمده قبل از ورود به شبکه 3D CNN، یک Average Pooling گرفته می‌شود. این کار باعث قدرتمند شدن سیستم و چند مقیاسه کردن کار در ابعاد ورودی[۱۰] می‌شود. به‌عبارتی نوآوری دیگری که به پژوهش اضافه شد، استفاده از ایده رویکرد هرمی جدید با دو سطح از شبکه‌های عصبی پیچشی سه بعدی است. این عمل تأثیرات مثبتی روی مدل پیشنهادی گذاشت.



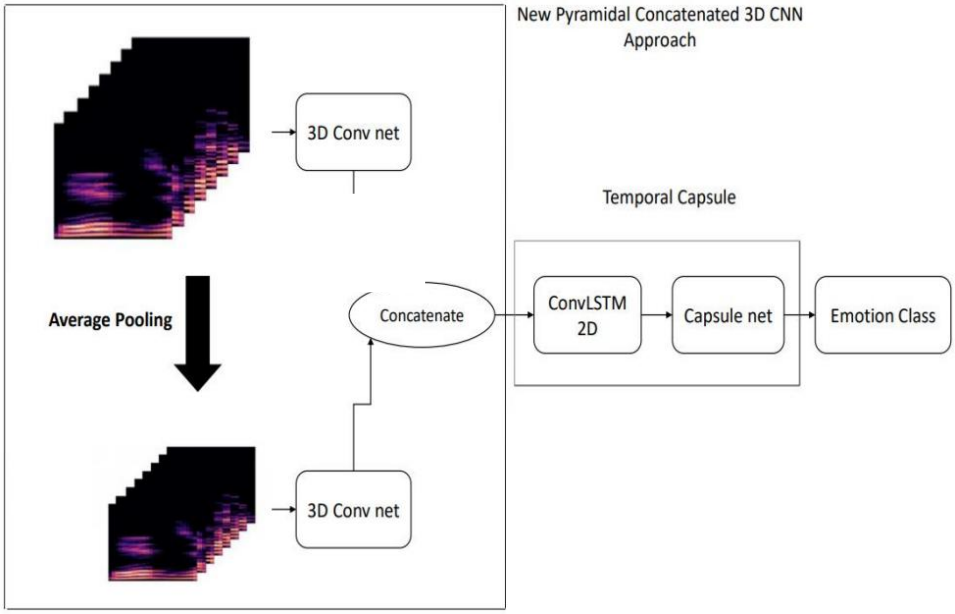
شکل ۲. ساختار هرمی جدید برای 3D CNN

۲-۳. ترکیب شبکه کپسول زمانی

برای ارائه سیستم بهتر سعی کردیم از ایده کپسول در مرجع [۲۲] که مربوط به تشخیص عملکرد انسان از روی ویدئو است، بهره‌مند شدیم. ایده، ایجاد کپسول زمانی از روی ورودی سه بعدی است. به دلیل اینکه مدل پیشنهادی علاوه‌بر ویژگی‌های طیفی، ارتباط ویژگی‌ها در بعد زمانی را نیز دارد به‌منظور جلوگیری از دست دادن اطلاعات مربوط به زمان، لایه ConvLSTM 2D اضافه شد. از طرف دیگر این لایه به‌عنوان لایه ابتدایی شبکه کپسول درنظر گرفته

می‌شود و به‌عبارتی یک شبکه کپسول زمانی[11] ایجاد می‌شود [۲۲].

با توجه به آنچه که گفته شد در جهت بهتر کردن کارایی سیستم‌های بازشناسی احساس از روی گفتار، مدل MSID [12]3DCNN + Temporal Capsule به‌عنوان مدل نهایی انتخاب شد که یک مدل قدرتمند برای سیستم‌های بازشناسی احساس از روی گفتار است. در بخش بعد نتایج مقایسه مدل نهایی با مدل‌های دیگر و عملکرد خوب آن ارائه می‌شود. شکل ۳ مدل نهایی را نشان می دهد.



شکل۳. مدل نهایی

۳. ارزیابی و نتایج

این پژوهش به زبان برنامه‌نویسی پایتون و در محیط ژوپیتر گوگل کولب[13] انجام شده است. در این بخش، پس از بیان معیارهای ارزیابی، آزمایش‌های انجام شده گفته می‌شوند. پژوهش انجام گرفته متشکل از ده آزمایش است که هر کدام از آزمایش‌ها در جهت بهتر کردن سیستم پیشنهادی است.

۳-۱. معیارهای ارزیابی

برای ارزیابی نتایج حاصل از شبیه‌سازی، از سنجه[14]های صحت کلی و ماتریس ابهام[15] استفاده می‌شود. در حوزه یادگیری ماشین و به‌طور خاص دسته‌بندی آماری[16]، ماتریس ابهام که ماتریس خطا نیز نامیده می‌شود جدول خاصی است که نحوه کارکرد یک مدل را نشان می‌دهد. اگر تعداد دسته‌های مسأله را N بگیریم ماتریس ابهام یک ماتریس $N \times N$ خواهد بود. که در آن X_{ii} تعداد نمونه‌های طبقه‌بندی شده در دسته J است به‌طوری‌که کلاس واقعی i است [۷] نحوه محاسبه سنجه صحت برای هر دسته به شکل رابطه ۱ است:

$$AC(i) = \frac{X_{ii}}{X_{i+}} = 1,2,3, \dots , Ni \qquad (۱)$$

که در آن X_{ii} نمونه‌هایی هستند که در دسته مورد نظر درست دسته‌بندی شده‌اند و X_{i+} جمع درایه‌های سطر i ام است، تعداد نمونه‌های دسته i ام را نشان می‌دهد. یک سنجه ارزیابی دیگر برای دسته‌بندی، صحت کلی نسبت به نمونه‌هایی است که درست دسته‌بندی شده‌اند و به تعداد کل نمونه‌ها در فرایند تحقیق را نشان می‌دهد. یک سنجه ارزیابی دیگر برای دسته‌بند، صحت کلی است که نسبت نمونه‌هایی

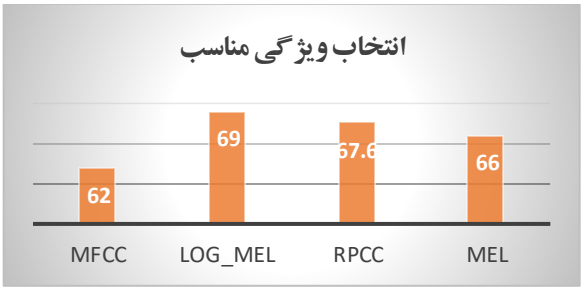
که درست دسته‌بندی شده‌اند را به تعداد کل نمونه‌ها در فرایند تحقیق نشان می‌دهد.

۳-۲. آزمایش‌های انجام شده برای ارائه مدل نهایی

آزمایش‌های زیر در جهت ارائه بهترین مدل با پارامترهای تنظیم شده مناسب است تا بتوان سیستم پیشنهادی قدرتمندی را ارائه داد.

۳-۲-۱. آزمایش انتخاب ویژگی

هدف از این آزمایش با توجه به شکل ۴ به‌دست آوردن بهترین ویژگی طیفی است.

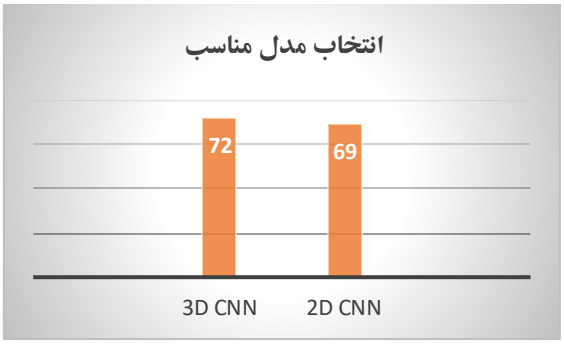


شکل ۴. بررسی ویژگی مناسب برای سیستم پیشنهادی

در این آزمایش با ثابت قرار دادن مدل روی *2D CNN* و استفاده از پایگاه داده *RAVDESS*، ویژگی اسپکتروگرام *Log-Mel* با دقت ۶۹ درصد بهترین دقت را در بین ویژگی‌های آزمایش شده به‌دست آورد؛ از این ویژگی در سیستم پیشنهادی استفاده خواهد شد. همانگونه که گفته شد به‌صورت تجربی، کارایی ویژگی *Log-Mel* در سیستم پیشنهادی مشخص شد.

۳-۲-۲. آزمایش انتخاب مدل

در این آزمایش با توجه به شکل ۵، هدف پیدا کردن مدل مناسب شبکه عصبی پیچشی است.

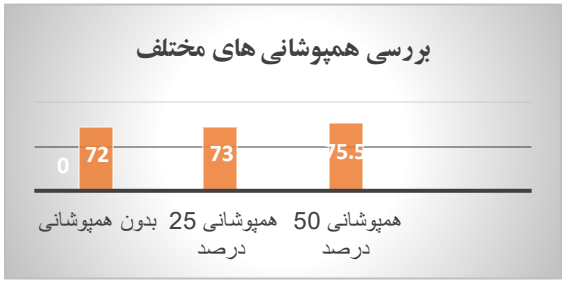


شکل ۵. انتخاب مدل مناسب برای سیستم پیشنهادی

در این آزمایش با قرار دادن *Log-Mel* به‌عنوان ویژگی استخراج شده، پایگاه داده *RAVDESS*، مدل‌های *2D CNN* و *3D CNN* مورد آزمایش قرار گرفتند که مدل *3D CNN* با فریم‌های بلند مدت ۳۰۰ میلی‌ثانیه و به‌دست آوردن دقت ۷۲ درصد، نتیجه بهتری نسبت به مدل *2D CNN* داشت. منطق این مورد این است که شبکه‌های *2D CNN* اطلاعات زمانی را نمی‌توانند یاد بگیرند اما شبکه‌های *3D CNN* این کار را به خوبی انجام می‌دهند.

۳-۲-۳. آزمایش بررسی همپوشانی

در این آزمایش با توجه به شکل ۶، هدف مقایسه همپوشانی در فریم‌های بلند مدت است.



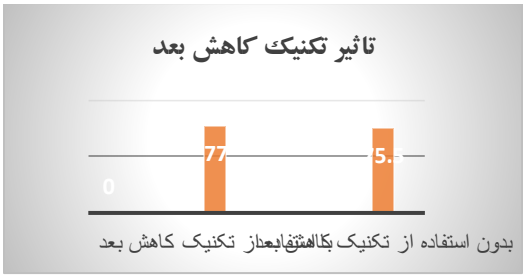
شکل ۶. بررسی همپوشانی‌های مختلف در فریم بلند مدت

در این آزمایش با ثابت قرار دادن ویژگی اسپتروگرام *Log-Mel* و پایگاه داده *RAVDESS* و مدل *3D CNN* و فریم‌های ۳۰۰ میلی‌ثانیه، همپوشانی‌های مختلف مورد بررسی قرار گرفت. که پنجره ۳۰۰ میلی‌ثانیه با همپوشانی ۵۰ درصد دقت ۷۵/۵ درصد را داد که نتیجه بهتری نسبت

به بقیه داشت. برای این مورد می‌توان به یادگیری مغز انسان اشاره کرد. مغز انسان نیز وقتی مطلبی را می‌شنود اگر تکراری از کلمات شنیده شده قبلی را دوباره بشنود یادگیری بهتری خواهد داشت.

۳-۲-٤. آزمایش تأثیر روش کاهش بعد

در این آزمایش روش کاهش بعد بررسی شد که آیا در روند پژوهش مفید است یا خیر. نتایج در شکل ۷ آورده شده است.

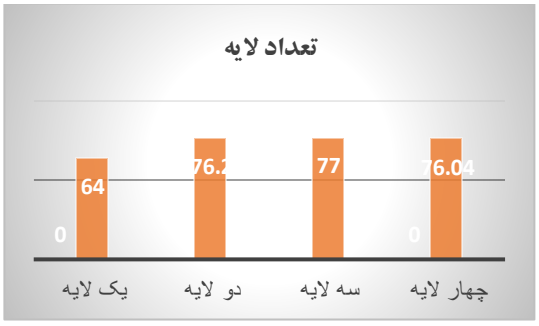


شکل ۷. بررسی تأثیر روش کاهش بعد

در این آزمایش با ثابت نگه داشتن ویژگی اسپکتروگرام *Log-Mel* و پایگاه داده *RAVDESS* و مدل *3D CNN* با پنجره‌های بلند مدت ۳۰۰ میلی‌ثانیه و همپوشانی ۵۰ درصد این پنجره‌ها این نتیجه به‌دست آمد: زمانی‌که از روش کاهش بعد استفاده شود دقت ۷۷ درصد حاصل می‌شود. این عمل با کم کردن محاسبات و پارامترهای آموزش باعث کاهش ابهام مدل و بهتر شدن سیستم پیشنهادی شد.

۳-۲-۵. آزمایش تعداد لایه مناسب

هدف از این آزمایش پیدا کردن تعداد لایه مناسب برای مدل *3D CNN* است.

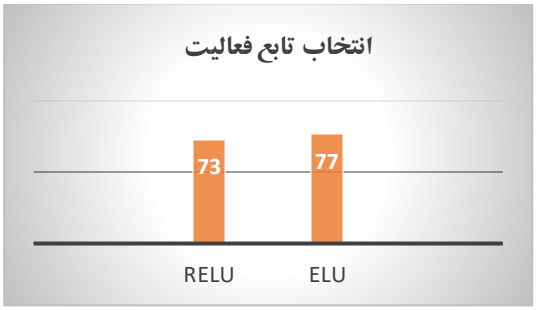


شکل ۸. تعداد لایه مناسب

در این آزمایش با ثابت نگه داشتن شرایط قبلی، مشخص شد که با داشتن سه لایه نتیجه مطلوب‌تری به‌دست می‌آید و نیز مشخص شد که عمیق کردن لایه‌ها تا حدی می‌تواند نتایج را بهتر کند اما بعد از حد معینی نتیجه عکس خواهد شد. برای این مورد می‌توان گفت که تعداد داده‌ها هر چه قدر بیشتر باشد مدل را بیشتر می‌توان عمیق کرد. پس افزایش تعداد داده‌های موجود در روند عمیق کردن مدل می‌تواند نتیجه بهتری داشته باشد.

۳-۲-٦. آزمایش انتخاب تابع فعالیت

در این آزمایش با توجه به شکل ۹، هدف مورد مقایسه قرار دادن دو تابع فعالیت رلو و الو است.

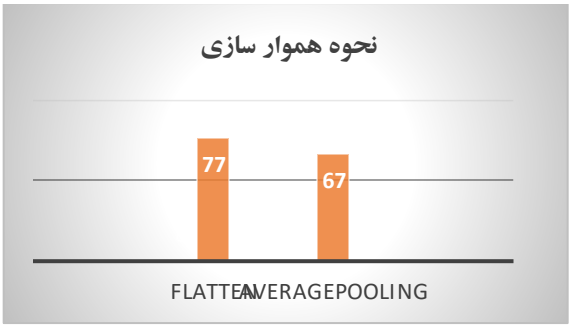


شکل ۹. مقایسه توابع فعالیت

در این آزمایش با ثابت نگه داشتن شرایط قبلی، معلوم شد که استفاده کردن تابع فعالیت الو بهتر از تابع فعالیت رلو است.

۳-۲-۷. آزمایش نحوه هموارسازی

در این آزمایش *Flatten* و *AveragePooling* برای اتصال بین لایه‌های شبکه‌های کانوولوشنی به لایه تمام متصل مورد آزمایش قرار گرفت. نتایج در شکل ۱۰ نشان داده شده است.

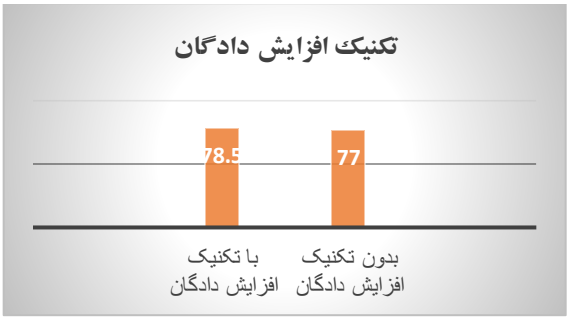


شکل ۱۰. بررسی نحوه ورود به لایه تمام متصل

در این آزمایش با ثابت نگه داشتن شرایط قبلی آزمایش، این نتیجه به‌دست آمد که اگر برای هموارسازی از *Flatten* استفاده شود نتیجه بهتر خواهد بود.

۳-۲-۸. آزمایش تأثیر روش افزایش دادگان

در اینجا هدف بررسی تأثیر روش افزایش دادگان در سیستم پیشنهادی است. نتایج در شکل ۱۱ دیده می‌شود.

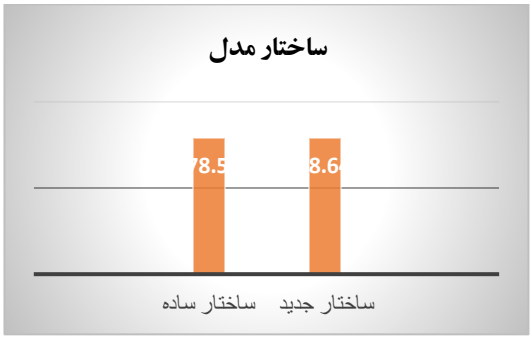


شکل ۱۱. بررسی تأثیر افزایش دادگان

در این آزمایش با ثابت نگه داشتن شرایط قبلی آزمایش و استفاده از لایه *Flatten*، مشخص شد روش افزایش دادگان تأثیر مطلوبی در سیستم دارد. چراکه شبکه‌های یادگیری عمیق نیاز به داده‌های زیادی دارند تا بتوانند اثر مطلوبی روی سیستم‌ها داشته باشند. با استفاده از روش افزایش دادگان، می‌توان در تنظیم بهتر پارامترهای شبکه‌های عصبی پیچشی تأثیر گذاشت.

۳-۲-۹. آزمایش مدل چند مقیاسه

در این مرحله، بررسی تأثیر ساختار جدید شبکه‌های عصبی پیچشی سه بعدی در سیستم پیشنهادی مورد بررسی قرار گرفت. نتایج در شکل ۱۲ نشان داده می‌شود.

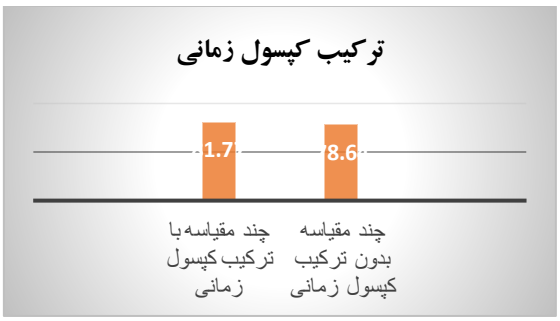


شکل ۱۲. بررسی تأثیر ساختار هرمی جدید برای *3D CNN*

در این آزمایش با ثابت نگه داشتن ویژگی *Log_Mel* و پایگاه داده *RAVDESS* و مدل *3D CNN* سه لایه با تابع فعالیت الو و پنجره‌های بلند مدت ۳۰۰ میلی‌ثانیه و همپوشانی ۵۰ درصد این پنجره‌ها و استفاده از روش کاهش بعد و استفاده از لایه *Flatten* و روش افزایش دادگان این نتیجه حاصل شد که مدل چند مقیاسه با دقت ۷۸/۶۴ اثر به نسبت خوبی در سیستم دارد و باعث قدرتمندتر شدن سیستم پیشنهادی می‌شود.

۳-۲-۱۰. آزمایش ترکیب شبکه کپسول زمانی

در این آزمایش ترکیب کردن شبکه کپسول با ساختار هرمی جدید برای *3D CNN* ، که پیشتر ارائه شده مورد بررسی قرار گرفت. نتایج مدل نهایی در شکل ۱۳ مشهود است.

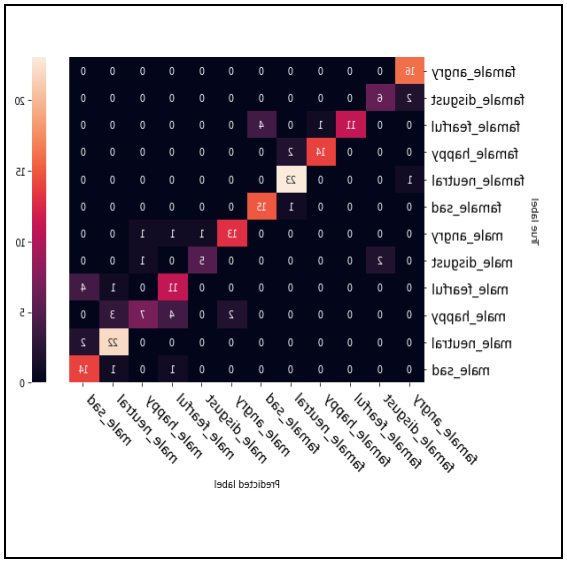


شکل ۱۳. بررسی تأثیر شبکه کپسول

در این قسمت با درنظر گرفتن شرایط ثابت قسمت‌های قبل مشخص شد که مدل نهایی *MSID 3DCNN+ Temporal Capsule* بهترین مدل به‌دست آمده است و به‌عنوان مدل نهایی و پیشنهادی ارائه می‌شود. چراکه خروجی مدل هرمی شبکه‌های عصبی پیچشی سه بعدی، به شکل سه بعدی است، برای اینکه اطلاعات زمانی از دست نرود، از *ConvLSTM 2D* که لایه اولیه شبکه کپسول استفاده شد و خود شبکه کپسول با یادگیری اطلاعات مکانی، در بهتر شدن سیستم پیشنهادی کمک مضاعفی کرد.

۳-۳. ماتریس ابهام مدل پیشنهادی

در این قسمت، مدل نهایی و پیشنهادی که در آزمایش‌های قبلی مورد ارزیابی دقت کلی قرار گرفت برای ارزیابی در ماتریس ابهام نیز مورد بررسی قرار می‌گیرد. در این ماتریس کلاس احساس‌های مختلف در هر سطر مشخص شده‌اند. قطر اصلی این ماتریس مربوط به تعداد احساس‌هایی است که پیش‌بینی درستی از آنها شده است. اگر هر سطر اختصاص داده شده به یک کلاس مورد بررسی قرار گیرد خانه مربوط به قطر اصلی تعداد پیش‌بینی‌های درست آن کلاس و خانه‌های دیگر تعداد پیش‌بینی‌های غلط آن کلاس هستند. این‌گونه می‌توان دقت سیستم را برای هر کلاس به‌دست آورد. ماتریس ابهام برای مدل نهایی (*MSID 3DCNN+ Temporal Capsule*) در شکل ۱۴ نشان داده شده است.



شکل ۱۴. ماتریس ابهام برای مدل پیشنهادی

۴-۳. مقایسه مدل‌های مبتنی‌بر ویژگی اسپکتروگرام *Log_Mel*

در جدول ۱ مدل‌های بررسی شده در این پژوهش و مرجع [۱۹] که از یک مدل *2D CNN* شش لایه استفاده کرده است مورد مقایسه قرار گرفتند. وجه مشترک همه، استفاده از ترکیب پایگاه داده گفتار معمولی و پایگاه داده آوازی و نیز استفاده کردن از ویژگی *Log_Mel* است. در نهایت این نتیجه به‌دست آمد که مدل پیشنهادی، نتایج خوبی را در بر دارد.

جدول ۱. مقایسه مدل های مبتنی بر ویژگی اسپکتروگرام *Log_Mel*

مدل	دادگان	ویژگی	دقت (٪)
مرجع [۱۹]	*RAVDESS*	*Log-Mel*	۷۶
3D CNN با ویژگی های بدست آمده در آزمایش های پژوهش	*RAVDESS*	*Log-Mel*	۷۷
3D CNN با ویژگی های بدست آمده در آزمایش های پژوهش + افزایش دادگان	*RAVDESS*	*Log-Mel*	۷۸.۵
3D CNN چند مقیاسه با ویژگی های بدست آمده در آزمایش های پژوهش (*MSID 3DCNN*)	*RAVDESS*	*Log-Mel*	۷۸.۶۴
MSID 3DCNN + Temporal Capsule	*RAVDESS*	*Log-Mel*	۸۱.۷۷

٤. بحث و نتیجه‌گیری

با توجه به یک سری مشکلات در سیستم‌های بازشناسی احساس از روی گفتار هدف این پژوهش تلاش برای کم رنگ کردن برخی از این مشکلات بود. این مشکلات و راه‌کار ارائه شده و نتیجه به‌دست آمده در این پژوهش را به شرح زیر می‌توان برشمرد:

- *مشکلات مربوط به پایگاه داده:* کم بودن داده‌های آموزش، وابسته بودن به یک سبک گفتار مثلا گفتار معمولی است. که با استفاده از ترکیب داده‌های آوازی و داده‌های گفتار معمولی و نیز افزایش دادگان، تلاش شد نتیجه سیستم‌های *SER*[۱۷] در این چالش بهتر گردد.
- *مشکلات مربوط به حوزه یادگیری عمیق:* درصورت استفاده کردن از شبکه‌های *2D CNN*، یادگیری اطلاعات زمانی ممکن نبود، برای غلبه بر این مشکل از *3D CNN* استفاده شد؛ به‌صورتی که بعد سوم زمان را شامل شود. لذا *3D CNN* برای یادگیری ویژگی‌های طیفی زمانی در حوزه بازشناسی احساس از روی گفتار پیشنهاد می شود.
- *قدرتمند کردن مدل پیشنهادی:* بدین منظور ساختار جدیدی از *3D CNN* ها که یک ساختار چند مقیاسه بر روی ابعاد ورودی است؛ ارائه شد. این مدل جدید ساختار هرمی اتصال داده شده شبکه‌های عصبی پیچشی سه بعدی نامیده شد.
- *مشکلات دسته‌بندهای مرسوم مورد استفاده برای CNN*: برای بهره‌گیری از دسته‌بند، باید خروجی به شکل یک بعدی در آورده شود. این امر باعث از دست رفتن اطلاعات مکانی می‌شود. به همین دلیل ترکیب شبکه کپسول پیشنهاد شد. این ترکیب وضعیت سیستم را بهتر می‌کرد. اما یک مشکل پیش می‌آمد؛ آن هم اینکه ورودی سه بعدی چگونه وارد شبکه کپسول گردد؟
- *ترکیب شبکه کپسول با خروجی ساختار هرمی:* برای این منظور از ایده *ConvLSTM 2D* بهره گرفته شد؛ اینگونه هم اطّلاعات زمانی حفظ شد، هم خروجی سه بعدی به دو بعدی تبدیل شده و وارد شبکه کپسول گردید؛ همچنین با استفاده کردن از لایه *ConvLSTM 2D* به‌عنوان لایه ابتدایی شبکه کپسول، شبکه کپسول زمانی ایجاد شد.

در این پژوهش، همان‌طورکه گفته شد با توجه به ضعفی که در سیستم‌های بازشناسی احساس از روی گفتار شناسایی شد، با استفاده از شبکه‌های عصبی پیچشی سه بعدی بر مشکل از دست دادن اطلاعات زمانی در شبکه‌های عصبی پیچشی دو بعدی غلبه شد. به‌علاوه با پیشنهاد ساختار هرمی اتصال داده شده شبکه‌های عصبی پیچشی سه بعدی، سیستمی قوی‌تر برای کار در مقیاس های مختلف ارائه شد. سپس با ترکیب ساختار بیان شده با شبکه کپسول زمانی، امکان یادگیری خروجی مدل با حفظ موقعیت زمانی و مکانی فراهم شد. بدین ترتیب با نوآوری‌ها و راه‌حل‌های پیشنهاد شده، نتایج مدل *SER* پیشنهادی که بر ترکیب داده‌های گفتاری و آوازی انجام شده است، با به‌دست آوردن دقت ۸۱/۷۷ درصد در مقایسه با پژوهش‌های مشابه عملکرد بسیار بهتری از خود نشان می‌دهد.

٦. مآخذ

[1] Akçay, Mehmet Berkehan, and Kaya Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers", *Speech Communication*, 2020, Vol.116, pp.56-76.

[2] Imani, Maryam, and Gholam Ali Montazer, "A survey of emotion recognition methods with emphasis on E-Learning environments", *Journal of Network and Computer Applications,* 2019, Vol.147, p.102423.

[3] Lugović, S., I. Dunđer, and M. Horvat, "Techniques and applications of emotion recognition in speech. 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2016-Proceedings, 1278–1283", *Google Scholar Google Scholar Cross Ref Cross Ref* (2016).

[4] Swain, Monorama, Aurobinda Routray, and Prithviraj Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review", *International Journal of Speech Technology,* 2018, Vol.21, no.1, pp.93-120.

[5] France, Daniel Joseph, Richard G. Shiavi, Stephen Silverman, Marilyn Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk", *IEEE transactions on Biomedical Engineering,* 2000, Vol.47, no.7, pp.829-837.

[6] Pao, Tsang-Long, Chun-Hsiang Wang, and Yu-Ji Li, "A study on the search of the most discriminative speech features in the speaker dependent speech emotion recognition", In *2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming*, IEEE, 2012, pp.157-162.

[7] Ting, K.M. Confusion Matrix. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. 2011.

[8] Tamulevičius, Gintautas, Gražina Korvel, Anil Bora Yayak, Povilas Treigys, Jolita Bernatavičienė, and Bożena Kostek, "A study of cross-linguistic speech emotion recognition based on 2D feature spaces", *Electronics,* 2020, Vol.9, no.10, p.1725.

[9] Nguyen, Dung, Kien Nguyen, Sridha Sridharan, David Dean, and Clinton Fookes, "Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition", *Computer Vision and Image Understanding,* 2018, Vol.174, p.33-42.

[10] Iqbal, Aseef, and Kakon Barua, "A real-time emotion recognition from speech using gradient boosting", In *2019 international conference on electrical, computer and communication engineering (ECCE),* IEEE, 2019, pp.1-5.

[11] Chapaneri, Santosh V., and Deepak D. Jayaswal, "Multi-taper spectral features for emotion recognition from speech", In *2015 International Conference on Industrial Instrumentation and Control (ICIC),* IEEE, 2015, pp.1044-1049.

[12] Badshah, Abdul Malik, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network", In *2017 international conference on platform technology and service (PlatCon),* IEEE 2017, pp.1-5.

[13] Kumbhar, Harshawardhan S., and Sheetal U. Bhandari, "Speech emotion recognition using MFCC features and LSTM network", In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA),* IEEE, 2019, pp.1-3.

[14] Etienne, Caroline, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch, "Cnn+ lstm architecture for speech emotion recognition with data augmentation", *arXiv preprint arXiv:1802.05630,* 2018.

[15] Guizzo, Eric, Tillman Weyde, and Jack Barnett Leveson, "Multi-time-scale convolution for emotion recognition from speech audio signals", In *ICASSP 2020-2020 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 202, pp.6489-6493.

[16] Li, Chao, Jinlong Jiao, Yiqin Zhao, and Ziping Zhao, "Combining gated convolutional networks and self-attention mechanism for speech emotion recognition", In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2019, pp.105-109.

[17] Stolar, Melissa N., Margaret Lech, Robert S. Bolia, and Michael Skinner, "Real time speech emotion recognition using RGB image classification and transfer learning", In *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*, IEEE, 2017, pp.1-8.

[18] Livingstone, Steven R., and Frank A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English", *PloS one*, 2018, Vol.13, no.5, p.e0196391.

[19] Venkataramanan, Kannan, and Haresh Rengaraj Rajamohan, "Emotion recognition from speech", *arXiv preprint arXiv:1912.10458*, 2019.

[20] Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks", In *Proceedings of the IEEE international conference on computer vision*, 2015, pp.4489-4497.

[21] Demir, Fatih, Muammer Turkoglu, Muzaffer Aslan, and Abdulkadir Sengur, "A new pyramidal concatenated CNN approach for environmental sound classification", *Applied Acoustics*, 2020, Vol.170, p.107520.

[22] Sankisa, Arun, Arjun Punjabi, and Aggelos K. Katsaggelos, "Temporal capsule networks for video motion estimation and error concealment", *Signal, Image and Video Processing*, 2020, Vol.14, no.7, pp.1369-1377.

پی‌نوشت:

1. Convolutional Neural Network
2. Deep Neural Network
3. Mel-frequency cepstral coefficient
4. Support Vector Machine
5. Granular Ball
6. Short Time Fourier Transform
7. Long Short-Term Memory
8. Convolutional LSTM (2 Dimentional)
9. A New Pyramidal Concatenated 3D CNN Approach
10. Multi Scale Input Dimension
11. Temporal Capsule
12. Multy Scale Input Dimension 3 Dimentional CNN
13. Google Colab
14. Measure (معیار)
15. Confusion Accuracy
16. Statistical Classification
17. Speech Emotion Recognition