
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

رساله دکتری مهندسی کامپیوتر- هوش مصنوعی

تشخیص احساس از گفتار بااستفاده از نگاشت مجموعه ویژگی‌ها به فضای چندبُعدی احساس

نگارنده: شادی لنگری

استاد راهنما

دکتر حسین مروی

استاد مشاور

دکتر مرتضی زاهدی

مهر ۱۳۹۹

شماره: ۷۱۶/ف ک
تاریخ: ۹۹/۹/۳۶
ویرایش:

باسمه تعالی

مدیریت تحصیلات تکمیلی

فرم شماره ۱۱: صورت جلسه نهایی دفاع از رساله دکتری (Ph.D)

بدینوسیله گواهی می شود آقای/خانم **شادی لنگری** دانشجوی دکتری رشته **مهندسی کامپیوتر-هوش مصنوعی** به شماره دانشجویی **۹۱۲۵۴۶۵** ورودی **۹۱** در تاریخ **۱۳۹۹/۰۷/۲۹** از رساله **نظری**☐ / **عملی**☐ خود با عنوان : **تشخیص احساس از گفتار با استفاده از نگاشت مجموعه ویژگی‌ها به فضای چندبُعدی احساس**

دفاع و با اخذ نمره ۱۹/۰۱ به درجه : عالی نائل گردید.

الف) درجه عالی: نمره ۲۰-۱۹ ☑	ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۷ ☐
ج) درجه خوب: نمره۱۶/۹۹– ۱۵ ☐	د) مردود: کمتر از ۱۵ ☐

ردیف	هیئت داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱	دکتر حسین مروی	استاد/ اساتید راهنما	دانشیار	
۲	دکتر مرتضی زاهدی	مشاور/ مشاورین	استادیار	
۳	دکتر حسن فرسی	استاد مدعو داخلی / خارجی	استاد	
۴	دکتر علی سلیمانی	استاد مدعو داخلی / خارجی	دانشیار	
۵	دکتر منصور فاتح	استاد مدعو داخلی / خارجی	استادیار	
۶	دکتر علی‌رضا تجری	سرپرست (نماینده) تحصیلات تکمیلی دانشکده	استادیار	

مدیر محترم تحصیلات تکمیلی دانشگاه:

ضمن تأیید مراتب فوق مقرر فرمائید اقدامات لازم در خصوص انجام مراحل دانش آموختگی آقای/خانم **شادی لنگری** بعمل آید.

نام و نام خانوادگی رئیس دانشکده :

تاریخ و امضاء و مهر دانشکده:

تقدیم به پدر و مادر عزیزم به پاس حمایت بی دریغ شان

تشکر و قدردانی

از استاد صبور، حامی و شایسته؛ جناب آقای دکتر مروی که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این رساله را بر عهده گرفتند؛ و از استاد گرامی، جناب آقای دکتر زاهدی، که زحمت مشاوره این رساله را متقبل شدند، کمال تشکر و قدردانی را دارم.

تعهدنامه

اینجانب **شادی لنگری** دانشجوی دوره کارشناسی ارشد (دکتری) رشته **مهندسی کامپیوتر- هوش مصنوعی** دانشکده مهندسی کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **تشخیص احساس از گفتار بااستفاده از نگاشت مجموعه ویژگی‌ها به فضای چندبُعدی احساس** تحت راهنمائی **آقای دکتر حسین مروی** متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ ۱۳۹۹/۷/۲۹

امضای دانشجو

مالکیت نتایج و حق نشر

چکیده

یکی از موضوعات مهم در تعامل میان انسان و کامپیوتر، ایجاد سیستمی است که بتواند مانند یک انسان بشنود و عکس‌العمل مناسب نشان دهد. سیگنال گفتار علاوه بر متن گفته شده، حاوی اطلاعات و ویژگی‌های مهمی راجع به گوینده از جمله سن، جنسیت، لهجه، گویش، میزان سلامتی و احساسات و استرس گوینده نیز است. یکی از زمینه‌هایی که می‌تواند به نزدیک شدن به این هدف کمک نماید، برقراری ارتباط کلامی بین انسان و ماشین و همچنین درک احساسات انسانی از سوی ماشین و ارائه واکنش مناسب به آن است. این امر منجر به طراحی سیستم تشخیص خودکار احساس از گفتار شده است. اکثر تحقیقات این حوزه براساس طبقه‌بندی نمونه‌ها در کلاس‌های گسسته و یا فضای پیوسته احساسی طراحی و پیاده‌سازی شده‌اند. یکی از چالش‌های این مساله انتخاب مدل نمایش احساسات کاراتر و انتخاب یک مجموعه ویژگی موثر در طبقه‌بندی احساسات در هر دو فضا است. از این‌رو در این رساله یک نگاشت بین این دو فضا با استفاده از مجموعه ویژگی‌های مناسب پیشنهاد شده است. هدف از رویکرد پیشنهادی دستیابی به یک بازنمایی از فضای ویژگی است که ماهیت ابعاد فضای احساسی را بهتر به تصویر می‌کشد. سیستم تشخیص احساس از گفتار پیشنهادی این رساله در دو فاز طراحی و پیاده‌سازی می‌شود. در فاز اول یک سیستم تشخیص احساس بهبودیافته با استفاده از یک روش استخراج ویژگی نوین انطباقی مبتنی‌بر ضرایب فرکانس- زمان و یک روش انتخاب ویژگی ترکیبی تکاملی پیشنهاد شده است. در فاز دوم یک نگاشت بین مجموعه ویژگی پیشنهادی فاز اول و فضای سه بعدی احساس ایجاد می‌شود. مدل پیشنهادی با استفاده از پایگاه داده گفتار احساسی برلین EMO-DB، مجموعه دادگان شنیداری صوتی و تصویری SAVEE، مجموعه داده‌های احساسی رادیو درام فارسی PDREC و پایگاه داده گفتار احساسی پیوسته آلمانی VAM ارزیابی و اعتبارسنجی شده است. نتایج تجربی نشان می دهند که مدل پیشنهادی به طور موثری کلاس‌های مختلف احساسی را در پایگاه داده EMO-DB با دقت ۹۷,۵۷%، در مجموعه دادگان SAVEE با دقت ۸۰٪ و در PDREC با دقت ۹۱,۶۴٪ شناسایی می‌کند.

کلمات کلیدی: تشخیص احساس، پردازش گفتار، استخراج ویژگی، انتخاب ویژگی، تعامل کامپیوتر- انسان.

لیست مقالات مستخرج از پایان نامه

1. Shadi Langari, Hossein Marvi, and Morteza Zahedi, “Efficient speech emotion recognition using modified feature extraction,” Informatics Med. Unlocked, vol. 20, p. 100424, 2020.
Elsevier, DOI: 10.1016/j.imu.2020.100424

2. Shadi Langari, Hossein Marvi, and Morteza Zahedi, “Improving of feature selection in speech emotion recognition based-on hybrid evolutionary algorithms,” Int. J. Nonlinear Anal. Appl., vol. 11, no. 1, pp. 81–92, 2020. DOI: 10.22075/ijnaa.2020.4227

۳. شادی لنگری، حسین مروی، مرتضی زاهدی، "بهبود کارایی تشخیص احساس از روی گفتار با انتخاب ویژگی‌های بهینه توسط الگوریتم ژنتیک بهبودیافته با جستجوی فاخته"، هفتمین کنگره مشترک سیستم‌های فازی و هوشمند ایران، دانشگاه بجنورد، بجنورد، ایران، ۱۳۹۷

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل۱ : مقدمه

۱-۱ مقدمه

گفتار سریع‌ترین، طبیعی‌ترین و متداول‌ترین روش ارتباطی بین انسان‌ها است. سیگنال گفتار یک سیگنال پیچیده محسوب می‌شود که علاوه‌بر پیام درحال انتقال، شامل مشخصات گوینده از جمله جنسیت، سن، زبان، گویش، احساسات و ... نیز هست. با پیشرفت تکنولوژی، تعامل بین انسان و ماشین نیز افزایش یافته و یکی از روش‌های بهبود و تسریع این تعاملات، ارتباط گفتاری بین ماشین و انسان است. لذا در دهه‌های اخیر محققان روی روش‌های مختلفی جهت افزایش کارایی این ارتباط گفتاری در سیستم‌هایی مانند شناسایی گفتار و شناسایی گوینده مطالعه نموده‌اند. یکی از اهداف مهم در سیستم‌های تشخیص گفتار، ایجاد سیستم‌هایی است که بتوانند مانند انسان بشنوند و عکس‌العمل مناسب نشان دهند. این امر منجر به بررسی یکی از تحقیقات مهم و چالشی در سال‌های اخیر به‌نام تشخیص اتوماتیک احساس از گفتار شده‌است [1]. تشخیص احساس از گفتار می‌تواند در ارتباطات روزمره انسان با ماشین نقش مؤثری داشته و باعث افزایش دقت، سرعت و صمیمیت در این تعامل گردد. ارزش عملی و کاربردی تشخیص احساس از گفتار را می‌توان در صنعت خوردوسازی، گوشی‌های هوشمند، مراکز تماس تلفنی آنلاین و ... مشاهده نمود. به‌عنوان مثال، می‌توان از آن برای قضاوت در مورد صحت و فوریت یک تماس اضطراری استفاده کرد. همچنین می‌تواند برای مسیریابی خدمات تماس اضطراری برای تماس‌های اضطراری با اولویت بالا مورد استفاده قرار گیرد. در کابین‌های هواپیما، تشخیص گفتار استرس بین کنترل ترافیک هوا و خلبانان می‌تواند ایمنی حمل و نقل هوایی را بهبود بخشد. همچنین در تجزیه و تحلیل گفتار پزشکی قانونی، شناسایی وضعیت عاطفی می‌تواند به اجرای قانون کمک کند تا وضعیت تماس گیرندگان تلفنی را ارزیابی کند یا به آن‌ها در مصاحبه‌های مشکوک کمک کند [1]. در این فصل ابتدا به بیان مساله تحقیق می‌پردازیم. سپس انواع پیکربندی‌ها و چالش‌های موجود در مسئله تشخیص احساس از گفتار بررسی می‌شوند. و در ادامه هدف، نوآوری‌ها و دستاوردها، و ساختار رساله را شرح خواهیم داد.

۱-۲ بیان مساله

امروزه سیستم‌های بازشناسی احساس از سیگنال‌های گفتار به دلیل افزایش تراکنش میان انسان و ماشین نقش مهمی در زندگی روزمره پیدا کرده و نیاز به این سیستم‌ها افزایش روز افزون داشته است. اما مشکلاتی نیز در این سیستم‌ها وجود دارد. اولین مساله در تحلیل احساسات، دشوار بودن تفکیک ویژگی‌های احساس می‌باشد، چرا که ویژگی‌های مختلف در احساسات، به افراد مختلف و حالت فعلی گوینده در زمان بیان جملات احساسی مانند حوصله، وضعیت درونی، نگرش و خصوصیات شخصیتی فرد به شدت وابسته است. دومین دلیل پیچیدگی احساسات این است که اغلب اوقات در هنگام برقراری ارتباط بین اشخاص، احساسات کامل، خالص و پایه بروز نمی‌کنند، بلکه معمولاً ترکیبی از احساسات مختلف در یک لحظه ممکن است بروز کنند. بنابراین احساس پدیده‌ای مبهم، پیچیده و مرکب است که برای جداسازی، تشریح و تشخیص بسیار دشوار می‌باشد.

تشخیص احساسات گوینده، موضوعی است که طرفداران زیادی در بین محققان پیدا کرده و در زبان‌های مختلف دنیا تحقیقات زیادی در این زمینه انجام گرفته است. بیشتر تحقیقات انجام گرفته در مورد چهار احساس اصلی غم، شادی، عصبانیت و ترس به همراه حالت طبیعی است. دکتر کریستین جونز، رئیس هیات مدیره شرکت افکتیو مدا در این باره می‌گوید: وقتی فرد افسرده یا ناراحت است، تن صدایش پایین می‌آید و حرف زدنش آرام می‌شود، اما هنگامی که عصبانی می‌شود، تن صدا بالا می‌رود و شدت حرف زدنش بیشتر می‌شود. یعنی زمانی که صحبت می‌کنیم، احساسات خود را از ده‌ها راه نشان می‌دهیم؛ مکانیزم این سیستم تنها بر چگونگی صحبت کردن استوار است و به‌اینکه چه واژگانی به‌کار می‌رود توجهی ندارد. این سیستم می‌تواند چگونگی احساسات فرد را در همان لحظه‌ای که صدای او را می‌شنود و بدون فاصله زمانی، بیان کند [۲]. در تشخیص احساس از گفتار، با استخراج و انتخاب مجموعه ویژگی‌های مناسب از داده‌های گفتار احساسی و مدل کردن طبقه‌بند با این مجموعه ویژگی‌ها، عمل تشخیص انجام می‌شود. به‌طورکلی در اکثر تحقیقات مربوط به تشخیص احساس از گفتار چهار

بخش کلیدی بررسی شده‌اند: ۱- مجموعه دادگان گفتار احساسی[1] ۲- استخراج ویژگی[2]، ۳- انتخاب (کاهش) ویژگی[3] و ۴- طبقه‌بندی[4]. مروری بر تحقیقات گذشته نشان می‌دهد که توسعه این سیستم‌ها با بهبود و توسعه این بخش‌های کلیدی ادامه یافته است.

۱-۳ انواع پیکربندی‌های مسئله تشخیص احساس از گفتار

هدف از تشخیص گفتار احساسی، تشخیص خودکار حالت احساسی یک انسان (مرد/ زن) از روی گفتار بیان شده توسط آن فرد است. اولین سوال مطرح در تعریف مساله تشخیص احساس از گفتار این موضوع است که: احساس چیست؟ احساس یک حالت ذهنی است که خود به خود از طریق تلاش آگاهانه مطرح می‌شود و اغلب با تغییرات فیزیولوژیکی همراه است. می‌توان گفت احساس یک تجربه‌ی کوتاه‌مدت فیزیولوژی-روانی است که درنتیجه‌ی عوامل بیوشیمیایی (درونی) و محیطی (بیرونی) است [۲]. از آنجا که احساسات نتیجه‌ی تجربیات ذهنی هستند، نمی‌توان قوانین یکسان یا مدل‌های جامع برای نمایش آن‌ها پیدا کرد، اما دو گرایش در تاریخ روانشناسی، بسته به اینکه احساسات را گسسته در نظر بگیریم یا پیوسته، وجود دارد. رویکرد گسسته عبارت است از تعیین کلاس‌های پایه عاطفی. در این روش، گروه‌های احساسی مجزا به‌عنوان احساسات اصلی تعریف می‌شوند. اِکمان[5] هفت گروه اساسی و اصلی احساس انسانی را به شکل زیر تعریف کرده است: شادی، غم، عصبانیت، نگرانی، خستگی، تنفر و خنثی بودن. سایر احساسات را می‌توان به صورت ترکیبی از احساسات اصلی بیان کرد [۲]. از طرف دیگر، رویکرد پیوسته شامل تعریف N بُعد برای بیان یک احساس است. مشهورترین آن‌ها یک فضای

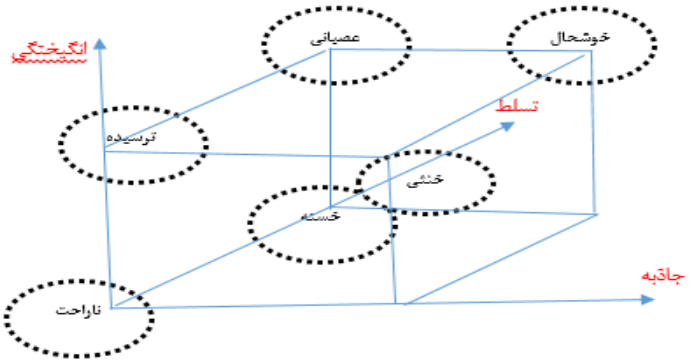
[1] Emotional Speech Corpus
[2] Feature Extraction
[3] Feature Selection (Reduction)
[4] Classification
[5] Ekman

احساسی سه‌بُعدی شامل: برانگیختگی[1] تسلط[2] و جاذبه (ظرفیت)[3] است. هر احساس می‌تواند به‌صورت ترکیب خطی از این سه پارامتر بیان شود. جاذبه مثبت یا منفی بودن یک احساس را بیان می‌کند. درجه‌ی هیجان یا درگیری فرد در حالت احساسی توسط برانگیختگی اندازه‌گیری می‌شود، و قدرت، نفوذ و توانایی احساسات را بیان می‌کند[۳]. به‌عنوان مثال، حس شادی یک احساس با جاذبه مثبت، انگیختگی بالا و قدرت زیاد است. مقادیری که این سه پارامتر خواهند داشت به صورت جدول ۱-۱ مشخص می‌شوند [۳].

جدول ۱-۱ مقادیر پارامترهای فضای سه‌بُعدی احساس

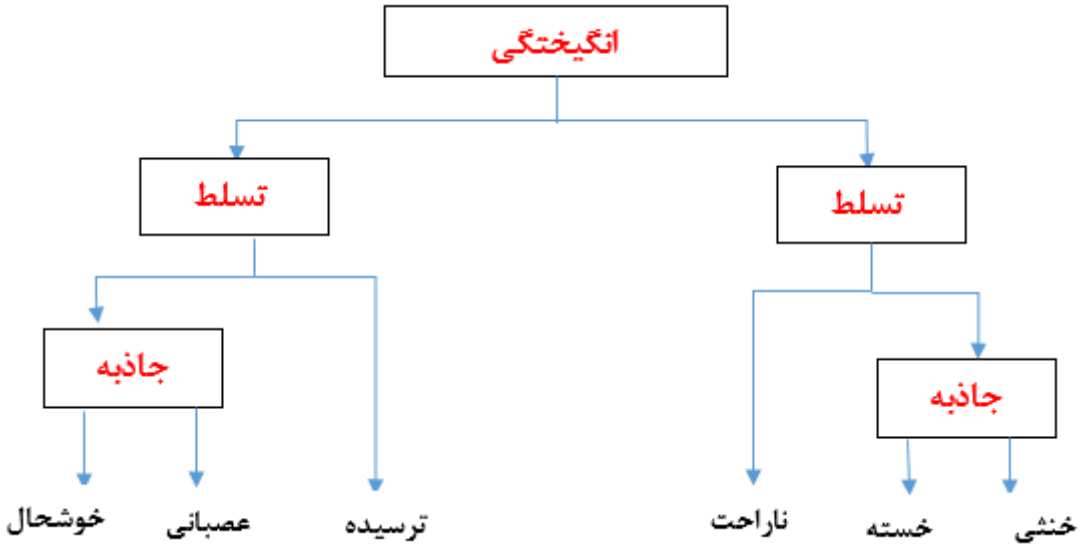
پارامتر	دامنه مقادیر
برانگیختگی	ساکت[4] تا برانگیخته[5]
جاذبه	منفی تا مثبت
قدرت	ضعیف تا قوی

هر دو رویکرد بیان شده را می‌توان با قراردادن احساسات گسسته در فضای احساسات پیوسته ترکیب کرد. شکل ۱-۱ نشان دهنده‌ی موقعیت شش احساس پایه در فضای سه‌بعدی است. همچنین شکل ۱-۲ دسته‌بندی این احساسات را براساس مقادیر بیان شده در فضای سه‌بُعدی نشان می‌دهد.

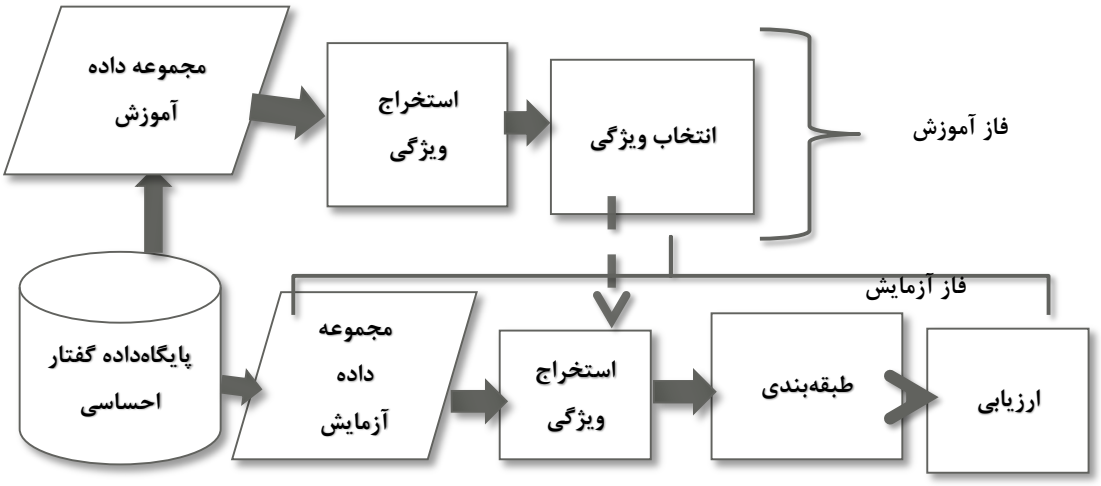


شکل ۱-۱ مدل سه بُعدی احساس با شش احساس پایه [3]

[1] Activation
[2] Potency (Dominance)
[3] Valence
[4] Calm
[5] Excited



شکل ۱-۲ دسته‌بندی احساسات پایه براساس مدل سه‌بُعدی احساس [3]



شکل ۱-۳ معماری متداول یک سیستم تشخیص احساس از گفتار

پروژه‌های مختلف تشخیص احساس معمولاً روش‌های متفاوتی را اتخاذ می‌کنند. با این حال، بیشتر کارهای موجود در زمینه تشخیص احساس را می‌توان به روش کلی شکل ۱-۳ خلاصه کرد. در ادامه برای هر بخش از بلوک دیاگرام شکل ۱-۳، توضیح مختصری ارائه خواهد شد، و در فصل آینده به بررسی آن‌ها در سایر تحقیقات می‌پردازیم.

- **استخراج ویژگی**

استخراج ویژگی معمولا یک فرآیند سه مرحله‌ای است. اولین مرحله، مرحله‌ی پیش پردازش است که سیگنال گفتار در آن باید نرمالیزه و نویزگیری شود. مراحل دوم و سوم مربوط به استخراج ویژگی‌های محلی و عمومی از سیگنال است. یکی از بخش‌های مورد توجه در این تحقیق، روی این بخش و بخش انتخاب ویژگی است.

- **انتخاب ویژگی**

در این بخش با استفاده از الگوریتم‌های مطرح در انتخاب ویژگی‌ها، یک مجموعه مناسب از میان خصیصه‌های استخراج شده از بخش قبلی انتخاب می‌شود.

- **طبقه‌بندی**

در این مرحله ابتدا یک طبقه‌بند مناسب انتخاب شده و سپس با استفاده از مجموعه ویژگی‌های استخراج و انتخاب شده از مجموعه دادگان سیگنال‌های گفتار، طبقه‌بند آموزش داده شده و آزمایش می‌شود.

- **ارزیابی**

در این مرحله برای ارزیابی نتایج طبقه‌بند می‌توان از روش‌های مختلفی از جمله محاسبه معیار دقت[۱] از روی ماتریس سردرگمی[۲]، روش‌های یادآوری و دقت[۳] و یا رسم منحنی ROC[۴] برای مقادیر مختلف آستانه و محاسبه مساحت زیر این منحنی به عنوان نشانه ی کیفیت طبقه‌بند استفاده کرد. همچنین می‌توان از محاسبه همبستگی بین نتایج طبقه‌بندی توسط انسان با نتایج حاصل از طبقه‌بندهای سیستم نیز استفاده کرد.

---

۱ Accuracy

۲ Confusion Matrix

۳ Recall and Precision

۴ Receiver Operating characteristics

۱-۴ چالش‌های سیستم‌های تشخیص احساس از گفتار

اگرچه پیشرفت‌های بسیاری در سیستم‌های تشخیص احساس از گفتار رخ داده است، اما هنوز موانع و چالش‌های مختلفی برای رسیدن به تشخیص با دقت بالاتر و موفق‌تر وجود دارد که باید برطرف شوند. بعضی از این چالش‌ها [۴] [۵] [۶] [۷] عبارتند از:

- ✓ مدل نمایش احساسات : کدام مدل نمایش احساسات بهتر و کارآتر است؟ مدل دسته‌بندی شده براساس احساسات پایه یا مدل نمایش در فضای چند بعدی؟
- ✓ نبودن پایگاه‌های داده با داده‌های گفتار واقعی‌تر و طبیعی‌تر
- ✓ یکسان نبودن پایگاه داده‌ها در تحقیقات و عدم قابلیت مقایسه کارآیی آن‌ها
- ✓ عدم وجود روش‌های استخراج ویژگی و روش‌های طبقه‌بندی مناسب و مشترک برای تمام پایگاه داده‌ها و مدل‌ها: در هر تحقیق نتایج مجموعه ویژگی و طبقه‌بند روی مجموعه داده خاصی ارائه شده و نمی‌توان نتیجه گرفت که آیا این مجموعه ویژگی‌ها و یا طبقه‌بندها روی پایگاه داده‌های دیگر نیز به خوبی عمل می‌کنند؟
- ✓ وابستگی به حالت، فرهنگ، زبان و شخصیت گوینده: بیان احساسات توسط گفتار در هر گوینده با توجه به سن و فرهنگ و جنسیت و شخصیت روانی و اجتماعی‌اش می‌تواند متفاوت باشد.
- ✓ نرخ تشخیص پایین برای بعضی از حالات احساسی: مثلا در اکثر تحقیقات نرخ تشخیص شادی و عصبانیت پایین است.
- ✓ قابلیت کاربردی تحقیقات موجود: در اکثر تحقیقات بررسی شده هیچگونه کاربرد عملی و خاصی مطرح نشده است و سیستم‌ها و روش‌های ارائه شده به‌صورت کاربردی به کار گرفته نشده‌اند.

۱-۵ هدف رساله

هدف اصلی این رساله، توسعه یک سیستم تشخیص احساس از گفتار با هدف بهبود دقت تشخیص است. سیستم پیشنهادی در این رساله در دو فاز مدل شده است. در فاز اول یک سیستم تشخیص کلاس‌های گسسته احساسی از گفتار پیشنهاد شده است که روی دو بخش کلیدی مساله تشخیص احساس از گفتار، یعنی استخراج و انتخاب ویژگی‌های موثر در طبقه‌بندی دقیق احساسات، متمرکز است. در این فاز، علاوه‌بر استفاده از ویژگی‌های متداول سایر تحقیقات، روشی برای استخراج یک ویژگی جدید تطبیق یافته در حوزه زمان-فرکانس پیشنهاد شده است. همچنین یک الگوریتم ترکیبی تکاملی نیز برای انتخاب بهترین مجموعه ویژگی قبل از مرحله طبقه‌بندی ارائه شده است. در فاز دوم نیز یک نگاشت بین مجموعه ویژگی پیشنهادی و فضای سه بعدی احساس مدل و ارزیابی می‌شود.

۱-۶ نوآوری‌ها و دستاوردهای رساله

به‌طور خلاصه دستاوردها و نوآوری‌های این رساله عبارتند از :

- در این رساله ابتدا مروری بر روش‌های انتخاب ویژگی در تشخیص احساس از گفتار انجام شد. در این راستا یک الگوریتم انتخاب ویژگی ترکیبی، برای انتخاب ویژگی‌های مربوطه با در نظر گرفتن فضای جستجوی محلی و عمومی پیشنهاد و نتیجه مطالعات اولیه در مقاله ۳ و نتایج مطالعات و آزمایش‌های بیشتر در مقاله ۲ گزارش شده است.
- در فاز بعدی یک روش استخراج تطبیقی با استفاده از مولفه‌های حوزه زمان-فرکانس معرفی و مدل شده است. نتایج این مطالعات در مقاله ۱ گزارش شده است.
- دستیابی به بالاترین عملکرد SER در سه مجموعه داده با سه زبان مختلف
- ارائه یک نگاشت بین فضای ویژگی‌های آکوستیک گفتار و فضای سه بعدی احساس.

۱-۷ ساختار رساله

فصل‌بندی و ساختار این رساله به‌صورت زیر تدوین شده است. در ۰ پیشینه پژوهش و کارهای مرتبط مورد بررسی قرار خواهند گرفت. در فصل۳ رویکرد پیشنهادی رساله تشریح می‌گردد. در ۰ رویکرد پیشنهادی مورد ارزیابی و تجزیه و تحلیل قرار گرفته و نتایج آزمایش‌های انجام شده و مقایسه‌های صورت گرفته گزارش می‌شوند. درنهایت در ۰ کارهای انجام شده جمع‌بندی و پیشنهاداتی برای کارهای آتی ارائه خواهد شد.

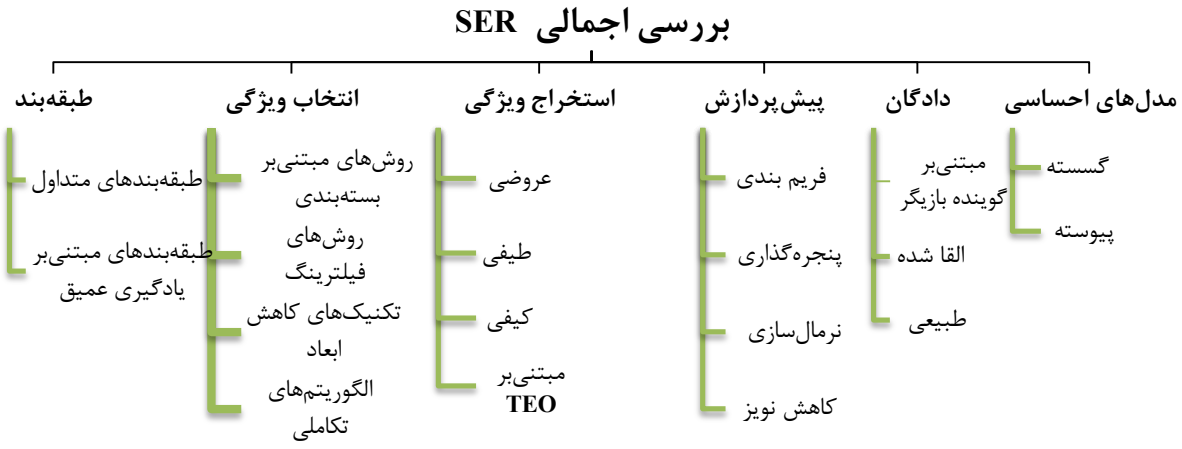
فصل ۲ مروری بر ادبیات موضوع و پیشینه پژوهش

۲-۱ مقدمه

به طور کلی تشخیص احساس از گفتار از چهار بخش دادگان گفتار احساسی، استخراج ویژگی، انتخاب ویژگی (کاهش) و طبقه‌بندی تشکیل شده است. در این فصل ابتدا مروری بر مفاهیم پایه و مبانی نظری موضوع خواهیم داشت و سپس جدیدترین تحقیقات انجام‌شده در حوزه تشخیص احساس از گفتار را در چهار بخش مورد بررسی و مطالعه قرار می‌دهیم. این مطالعات براساس نوآوری‌های انجام‌شده در هر بخش از سیستم تشخیص احساس از گفتار دسته‌بندی شده‌اند.

۲-۲ مبانی نظری و مفاهیم پایه

برای درک بهتر سیستم تشخیص احساس از گفتار (SER) مطابق شکل ۲-۱ آن را به چند بخش مجزا تفکیک کرده و مفاهیم پایه و نظری در هر بخش را بررسی می‌نماییم. همانطور که در شکل مشخص است رویکردهای مختلفی برای مدل‌سازی احساسات وجود دارد که هنوز هم یک مساله‌ی حل نشده است، با این‌حال، در اکثر تحقیقات معمولا دو مدل گسسته و چند بعدی استفاده می‌شود. یک سیستم SER به یک طبقه‌بند، یعنی یک ساختار یادگیری با ناظر، نیاز دارد تا برای تشخیص احساسات در سیگنال‌های گفتاری جدید آموزش یابد. چنین سیستم تحت نظارتی، ضرورت داده‌های برچسب‌دار را که احساسات در آن‌ها تعبیه شده است، به ارمغان می‌آورد. همچنین قبل از استخراج ویژگی‌ها، داده‌ها به پیش پردازش نیاز دارند. ویژگی‌ها برای یک فرآیند طبقه‌بندی ضروری هستند، چرا که داده‌های اصلی را به مهمترین مشخصه‌های آن تقلیل می‌دهند. در فرآیند استخراج ویژگی، مشخصه‌های مختلفی از سیگنال استخراج می‌شود که برای ترکیب یا کاهش ابعاد و تعداد ویژگی‌ها می‌توان از روش‌ها و الگوریتم‌های انتخاب ویژگی استفاده کرد. در ادامه تمام حوزه‌های مطرح شده در شکل ۲-۱ (به‌ترتیب از راست به چپ) بررسی و تشریح می‌شوند.



شکل ۲-۱ مروری بر سیستم‌های تشخیص احساس از گفتار (SER)

۲-۲-۱ انواع مدل‌های احساس

برای پیاده‌سازی موفق سیستم تشخیص احساس از گفتار، احساسات باید با دقت تعریف و مدل شوند. در مورد تعریف احساسات اتفاق نظر وجود ندارد و رویکردهای مختلفی برای مدل‌سازی احساسات وجود دارد که هنوز هم یک مساله‌ی حل نشده است، و به گفته Plutchik، بیش از نود تعریف از احساس در قرن بیستم ارائه شده است [۸]. احساسات حالت‌های روانشناختی پیچیده‌ای هستند که از چندین مؤلفه مانند تجربه شخصی، واکنش‌های فیزیولوژیکی، رفتاری و ارتباطی تشکیل شده‌اند. با این‌حال، در اکثر تحقیقات معمولا دو مدل گسسته و چند بعدی استفاده می‌شود. همانطور که اِکمان و همکاران [۹] توصیف کرده‌اند نظریه احساسات گسسته بر اساس شش دسته احساسات پایه: غم، شادی، ترس، عصبانیت، انزجار و تعجب، است. این احساسات کاملا ذاتی و مستقل از نظر فرهنگی هستند و برای مدت کوتاهی تجربه می‌شوند، اما سایر احساسات با ترکیب این احساسات پایه به‌دست می‌آیند [۹]. از آن‌جا که در زندگی روزمره مردم از این مدل برای تعریف احساسات مشاهده شده خود استفاده می‌کنند، طرح برچسب‌گذاری براساس دسته‌های احساس پایه کاملا طبیعی است، بنابراین بیشتر سیستم‌های SER

موجود بر این دسته‌بندی‌های احساسی پایه تمرکز دارند. با این وجود، این مدل برچسب‌گذاری گسسته احساسات قادر به تعریف برخی از حالات عاطفی پیچیده‌ی مشاهده شده در ارتباطات روزانه نیست.

مدل احساسی چندبعدی یک مدل جایگزین است که از تعداد کمی از ابعاد نهفته برای توصیف احساساتی مانند برانگیختگی[۱] (تحریک)[۲]، تسلط[۳] (قدرت)[۴] و جاذبه (ظرفیت)[۵] و کنترل استفاده می‌کند [۱۰]. در این مدل احساسات مستقل از یکدیگر نیستند و این ابعاد جنبه‌های قطعی و عاطفی احساسات‌اند. یکی از مدل‌های پرکاربرد چندبعدی، یک مدل دو بعدی است که از برانگیختگی در یک بعد، در مقابل جاذبه در بعد دیگر استفاده می‌کند. بعد جاذبه مثبت یا منفی بودن یک احساس را توصیف می‌کند و بازه مقادیر آن بین ناخوشایند[۶] و دلپذیر[۷] است. بعد برانگیختگی، قدرت احساس درک شده را تعریف می‌کند، که ممکن است هیجان‌زده یا بی‌احساس باشد و بازه مقادیرش از بی‌حوصلگی تا هیجانات دیوانه‌وار متغیر است. در مدل سه بعدی یک بعد تسلط یا قدرت هم وجود دارد که به قدرت ظاهری فرد با بازه مقادیری بین ضعیف و قوی اشاره دارد. به‌عنوان مثال، بعد سوم به‌ترتیب با درنظرگرفتن قدرت یا ضعف فرد، خشم و ترس را از هم متمایز می‌کند.

مدل نمایش چند بعدی معایب زیادی دارد. این مدل به اندازه کافی بصری نیست و برای برچسب زدن به هر احساس ممکن است آموزش خاصی لازم باشد. علاوه‌بر این، برخی از احساسات مانند ترس و عصبانیت یکسان تشخیص داده می‌شوند و برخی از احساسات مانند تعجب را در این مدل نمی‌توان طبقه‌بندی کرد، زیرا ممکن است بسته به زمینه احساسات غافل‌گیرانه دارای جاذبه مثبت یا منفی باشد.

---

[۱] Activation
[۲] Arousal
[۳] Dominance (Potency)
[۴] Power
[۵] Valence
[۶] Unpleasant
[۷] Pleasant

۲-۲-۲ انواع دادگان گفتار احساسی

یک موضوع مهم در ارزیابی سیستم‌های تشخیص احساس از گفتار، کیفیت دادگان مورداستفاده برای توسعه و تعیین کارایی سیستم‌ها است [۱۱]. در تهیه پایگاه داده گفتار احساسی، انواع روش‌ها و قواعد جمع‌آوری داده وجود دارد که منجر به توسعه مجموعه دادگان متنوعی شده است. مسائلی از قبیل ویژگی‌های گویندگان، طول و تعداد کلمات جملات، زبان گفتار، فضا و روش ضبط سخنرانی‌ها، و ... در تهیه دادگان نقش اساسی دارند. چالش مهمی که باید در ارزیابی تشخیص گفتار احساسی در نظر گرفته شود، درجه طبیعی بودن پایگاه داده استفاده شده برای ارزیابی عملکرد آن است. اگر از یک پایگاه داده با کیفیت پایین استفاده شود ممکن است نتیجه‌گیری نادرستی به‌دست آید. بنابراین، طراحی یک پایگاه داده مناسب و استاندارد بخش مهمی برای طبقه‌بندی خواهد بود. همچنین نرخ طبقه‌بندی، توسط تعداد و انواع کلاس‌های احساسی موجود در پایگاه داده نیز مورد مقایسه قرار می‌گیرد [۱۱].

کارایی بیشتر پایگاه‌های داده گفتار احساسی به شناسایی عملکرد سیستم تشخیص دهنده احساسات محدود شده‌اند. برخی از محدودیت‌های پایگاه داده احساسات گوینده در زیر خلاصه شده‌اند [۱۲]:

۱) در تعدادی از پایگاه‌های داده احساسات گوینده به اندازه کافی به‌صورت واقعی و واضح شبیه‌سازی نشده است.

۲) در برخی از پایگاه‌های داده، کیفیت ضبط گفتار خوب نیست. بنابراین فرکانس نمونه‌برداری تا حدی پایین است.

۳) رونوشت‌های آواشناختی برای برخی از پایگاه‌های داده آماده نشده است. بنابراین، استخراج محتویات زبان‌شناختی از سخنان این پایگاه‌های داده مشکل است.

۲-۲-۲-۱ جمع‌آوری داده

در تهیه داده‌های صوتی قواعد خاصی وجود دارد، از جمله اینکه پایگاه‌داده باید شامل کلمات و عبارات مناسبی باشد که معرف مجموعه گفتاری زبان مورد نظر است. از طرفی گویندگان نیز باید با انواع مختلف

جنسیت، سن و سطح سواد جهت پوشش دادن مشخصه‌های گفتاری مختلف، انتخاب شوند. همچنین تهیه یک رکورد اطلاعات توصیفی از داده‌های جمع‌آوری شده، بدون شک برای پژوهشگران علاقه‌مند به تحلیل احساسات گوینده مفید است. برای هر داده جمع‌آوری شده اطلاعات توصیفی از قبیل زبان گفتار، تعداد و موضوعات حرفه‌ای، سیگنال‌های روانشناسی دیگر که احتمال دارد همزمان با گفتار ضبط شود، هدف از داده‌های جمع‌آوری شده، ضبط حالت‌های گفتار و انواع احساسات (طبیعی، تقلیدی، استخراج شده) هم می‌توانند درنظر گرفته شوند [۱۳].

شواهد نشان می‌دهد که پژوهش‌ها بر روی تشخیص احساسات گوینده به برخی از احساسات محدود شده است. اکثر داده‌های جمع‌آوری شده احساسات گوینده شامل پنج یا شش احساس است، باوجود اینکه دسته‌بندی احساسات در دنیای واقعی بیشتر از این است. در دهه ۱۹۷۰، تئوری پالت پیشنهاد شد که در تلاش بود تا همه احساسات را به‌عنوان ترکیبی از بعضی از احساسات اصلی (مانند چیزی که در ترکیب رنگ‌ها رخ می‌دهد)، توصیف کند. قابل ذکر است که این نظریه پذیرفته نشده است و لغت " احساسات پایه‌ای" هم‌اکنون به‌طور گسترده‌ای بدون اشاره به اینکه احساسات می‌توانند با همدیگر ترکیب شوند، مورد استفاده قرار می‌گیرد. به‌طورکلی پذیرفتن احساسات پایه‌ای، عمومی‌تر است. اِکمان احساسات پایه‌ای زیر را پیشنهاد داده است: عصبانیت، ترس، غم، احساس خوشی، فریب‌خوردگی، خشنودی، خرسندی، تحریک، نفرت، تحقیر، تکبر، شرمساری، گناه، خجالت و راحتی. احساساتی که پایه‌ای نیستند، احساسات سطح بالا نامیده می‌شوند و این نوع احساسات در داده‌های جمع‌آوری شده کمتر دیده می‌شوند [۱۳].

همچنین برای تحلیل پارامترها، سیگنال‌های گفتار تحریف نشده و بدون نویز پس زمینه مورد نیاز است. به منظور بررسی احساسات گوینده، گفتارهای بی‌طرف برای ضبط سخنان یکسان در موقعیت‌های احساسی مختلف نیز موردتوجه است. در نتیجه ضبط باید با روش معین در شرایط آزمایشگاهی انجام شود. در برخی از آزمایشات روانشناختی، سعی می‌شود احساسات ویژه‌ای را به شخص در حال آزمون القاء کنند. اما بنا به دلایل اخلاقی وارد کردن احساسات منفی به شخص در حال آزمون ناخوشایند است.

احساسات شبیه‌سازی شده توسط افراد، تخمین خوبی برای احساسات گوینده واقعی است [۱۲]. اما ساده‌ترین راه برای جمع‌آوری سخنرانی‌ها با احساس مشخص این است که بازیگران آن را شبیه‌سازی کنند. بطور یقین یک بازیگر خوب می‌تواند گفتارهایی را تولید کند که به خوبی طبقه‌بندی شوند. در این راستا در تحقیقات سیگنال گفتار احساس واقعی یک فرد با سیگنال گفتار یک بازیگر که حالت احساسی را تقلید می‌کند مقایسه و تفاوت‌هایی بین این سیگنال‌ها مشاهد شده، اما بطور کلی نحوه صحبت کردن، محدوده فرکانس اصلی و ناپایداری هر دو سیگنال یکسان است. با این حال استفاده از بازیگران در طراحی پایگاه‌داده توصیه نمی‌شود، چرا که آن‌ها در ساختن محتوای احساسی مبالغه می-کنند و واضح است که این سیگنال‌ها طبیعی نخواهند بود. بنابراین یک روش دیگر انتخاب تصادفی از افراد در دسترس است، به این‌صورت که در جلسه انتخاب مقدماتی، افرادی که انتخاب می‌شوند یک سخن را با هدف احساسی مورد نظر می‌گویند و آن سخنان بطور مستقیم با یک میکروفون در هارد دیسک ثبت خواهد شد [۱۲].

به‌طورکلی می‌توان پایگاه داده‌ها را در سه دسته کلی زیر تقسیم کرد:

- پایگاه داده‌های گفتار احساسی مبتنی‌بر گوینده بازیگر[۱] (بازی شده)
- پایگاه داده‌های گفتار احساسی القا شده[۲]
- پایگاه داده گفتار احساسی طبیعی[۳]

دسته اول از مجموعه سخنرانی‌های بیان شده توسط بازیگران باتجربه، آموزش‌دیده و حرفه‌ای تئاتر یا رادیو جمع‌آوری شده است که اکثر پایگاه‌های داده گفتار احساسی از این روش جمع‌آوری شده‌اند. در مقایسه با سایر روش‌ها ایجاد چنین پایگاه داده‌ای نسبتاً آسان‌تر است. با این‌حال، به گفته محققان این نوع گفتار نمی‌تواند احساسات زندگی واقعی را به اندازه کافی درست منتقل کند، و حتی شاید اغراق‌آمیز

[۱] Acted (Simulated) speech emotion databases

[۲] Elicited (Induced) speech emotion databases

[۳] Natural speech emotion databases

باشد. دسته دوم گفتار القا شده، گفتاری است که احساسات یک گوینده معمولی در آن تحریک شده است. این نوع گفتار می‌تواند طبیعی یا شبیه‌سازی شده باشد و از گفتار ادا شده توسط بازیگران حرفه‌ای برای تشخیص احساسات گوینده، قابل اطمینان‌تر و به احساسات واقعی نزدیک‌تر است، زیرا بازیگران حرفه‌ای ممکن است در بیان احساسات اغراق کنند. و دسته آخر پایگاه‌های داده‌ای است که از گفتارهای طبیعی تشکیل شده، که هر نمونه یک گفتار خودانگیخته با احساسات واقعی است. پایگاه داده ای که با سخنرانی خودجوش داده‌های واقعی ایجاد شده است. این داده‌ها شامل اطلاعات ثبت شده از مکالمات مرکز تماس، ضبط مکالمات کابین خلبان در شرایط غیرعادی، مکالمه بین بیمار و پزشک، مکالمه همراه با احساسات در مکان‌های عمومی و تعاملات مشابه است. جمع‌آوری این داده‌ها دشوارتر است زیرا ممکن است هنگام پردازش و توزیع آن‌ها مشکلات اخلاقی و حقوقی به‌وجود آید.

۲-۲-۲-۲ مسائل مطرح در طراحی پایگاه‌های داده

در این بخش چند مساله مورد توجه در طراحی پایگاه‌های داده در سیستم‌های تشخیص احساس از گفتار را بررسی می‌نماییم:

- ✓ **انتخاب ویژگی‌ها:** برای استخراج ویژگی‌های گوینده، باید چند عامل را در نظر داشت، از جمله این ویژگی‌ها نباید به نویز محیط انتقال مانند کانال‌های مخابراتی حساس باشند، همچنین این ویژگی‌ها نباید به حالت روانی و فشارهای محیطی که گوینده در آن در حال صحبت کردن است وابسته باشند.
- ✓ **مدل‌های گویندگان:** در سیستم‌های بازشناسی گوینده، مدل هر گوینده شامل خصوصیات آماری او است. معمولا در مدل کردن گویندگان از یکی از دو مدل پارامتری (مانند مدل گوسی) و غیرپارامتری (مانند مدل مراکز خوشه‌ها یا حالات چندگانه) استفاده می‌شود. همچنین سایر موضوعات طراحی مانند سن و جنس نیز مورد توجه قرار می‌گیرند. اکثر بانک‌های اطلاعاتی

دارای گویندگان بزرگ‌سال هستند، اما مجموعه دادگان با گویندگان در رده سنی کودکان و سال‌خوردگان نیز وجود دارد.

- ✓ **طول گفتار آموزشی** : با افزایش این زمان، نتایج مطلوب‌تری حاصل می‌گردد.
- ✓ **انتخاب گفتار مناسب** : پیشنهاد می‌گردد سکوت و بخش‌های غیرگفتاری حذف گردد. همچنین گفتارهای کم انرژی و بی‌واک در تمایز بین گوینده‌ها کم اثر بوده و پیشنهاد می‌گردد از اطلاعات مدل حذف گردند.
- ✓ **محیط نویزی:** نویز پشت زمینه یک مشکل عمومی است و بهتر است برای تخمین آن از نویز پشت زمینه در زمان سکوت استفاده و با اعمال آن به مدل، تخمینی از مدل بدون نویز داشته باشیم.
- ✓ **تنوع گوینده :** سیگنال گفتار هر شخص در حالت‌های خوشحالی، ناراحتی و خستگی متفاوت است و به این ترتیب کارایی سیستم را کاهش می‌دهد (خصوصیات آماری یک فرد ثابت نیست). ملاحظات دیگر شامل تکرار سخنرانی‌ها با بازیگران مختلف، احساسات متفاوت و جنسیت متفاوت است.

فهرستی از مجموعه دادگان برجسته در مساله تشخیص احساس از گفتار در جدول ۲-۱ خلاصه شده است. در این رساله جهت ارزیابی روش پیشنهادی آزمایشات متعددی با چهار مجموعه داده انجام شده است که شامل: پایگاه داده احساسی برلین (Emo-DB)، دادگان احساسی SAVEE، پایگاه داده گفتار احساسی فارسی درام (PDREC) و مجموعه داده مدل احساس پیوسته VAM می‌باشند.

جدول ۲-۱ فهرستی از مجموعه دادگان برجسته در مساله تشخیص احساس از گفتار

پایگاه‌داده	احساسات	زبان	نوع	گویندگان	اندازه	دسترسی
(Emo-DB[۱]) [14]	شادی، غم، عصبانیت، ترس، تنفر،خستگی، خنثی	آلمانی	مبتنی‌بر گوینده بازیگر	۱۰ (۵ زن، ۵ مرد)	۷۰۰ فایل	عمومی و رایگان
(SAVEE[۲]) [15]	شادی، غم، عصبانیت، ترس، تنفر، تعجب، خنثی	انگلیسی	مبتنی‌بر گوینده بازیگر	۴ (مرد)	۴۸۰ فایل	رایگان
PDREC[۳] [۱۶]	عصبانیت، خستگی، نفرت، ترس، ناراحتی، تعجب، خوشحالی، خنثی	فارسی	طبیعی (رادیو نمایش)	۳۳ (۱۵ زن، ۱۸ مرد)	۷۴۸ فایل	خصوصی
Vera-Am-Mitting (VAM) [17]	جاذبه، برانگیختگی، تسلط	آلمانی	طبیعی(برنامه تلویزیونی)	۴۷ گوینده	۹۴۷ فایل	عمومی و با پرداخت مجوز
(DES[۴]) [18]	عصبانیت، شادی، غم، تعجب، خنثی	دانمارکی	مبتنی‌بر گوینده بازیگر	۴ (۲ زن، ۲ مرد)	۲۶۰ فایل	عمومی و با پرداخت مجوز
(CASIA[۱]) [19]	عصبانیت، ترس، خنثی، ناراحتی، تعجب، خوشحالی	ماندارین(چینی)	مبتنی‌بر گوینده بازیگر	۴ (۲ زن، ۲ مرد)	۱۲۰۰۰ فایل	عمومی و با پرداخت مجوز
(IEMOCAP[۲]) [20]	شادی، غم، عصبانیت، ترس، ناامیدی، خنثی	انگلیسی	مبتنی‌بر گوینده بازیگر	۱۰ (۵ زن، ۵ مرد)	۱۱۵۰ فایل	عمومی و با پرداخت مجوز
FAU Aibo Emotion [21]	شادی، غم، عصبانیت، ترس، تنفر، خستگی، خنثی	آلمانی	طبیعی (مدرسه برلین)	۵۱ دانش‌آموز ۱۰-۱۳	۹ ساعت	عمومی و با پرداخت مجوز
Persian ESD [۳] [۲۲]	عصبانیت، شادی، غم، ترس، چندش، خنثی	فارسی	مبتنی‌بر گوینده بازیگر	۲ (۱ زن، ۱ مرد)	۴۶۸ فایل	خصوصی
دادگان گفتار احساسی سهند[۲۳]	خنثی، تعجب، شادی، غم، عصبانیت	فارسی	مبتنی‌بر گوینده بازیگر	۱۰ (۵ زن، ۵ مرد)	۵۰ دقیقه	خصوصی
(SUSAS[۴]) [24]	چهار حالت تحت استرس: خنثی، عصبانی، گوش‌خراش، اثر لومبارد	انگلیسی	طبیعی و شبیه‌سازی	۳۲ (۱۳ زن، ۱۹ مرد)	۱۶۰۰۰ فایل	عمومی و با پرداخت مجوز
(BHUDES[۵]) [25]	عصبانیت، نفرت، ناراحتی، تعجب، خوشحالی	ماندارین(چینی)	گویندگان غیرحرفه‌ای	۷	۷۰۰ فایل	خصوصی
IITKGP-SESC [26]	عصبانیت، نفرت، دلسوزی، ترس، خوشحالی، تعجب، خنثی، طعنه‌آمیز	تلوگو (هندی)	مبتنی‌بر گوینده بازیگر	۱۰ (۵ زن، ۵ مرد)	۱۲۰۰۰ فایل	خصوصی

[۱] Berlin Emotional Database

[۲] Surrey Audio-Visual Expressed Emotion

[۳] پایگاه‌داده گفتار احساسی فارسی درام

[۴] Danish Emotional Speech Database

۲-۲-۳ پیش‌پردازش

اولین قدم پس از جمع‌آوری داده‌های مورداستفاده برای آموزش طبقه‌بند در سیستم SER، پیش‌پردازش است. برخی از تکنیک‌های پیش‌پردازش برای استخراج ویژگی‌ها استفاده می‌شوند، درحالی‌که برخی دیگر برای نرمال‌سازی ویژگی‌ها به‌کار می‌روند تا تغییرپذیری گویندگان و ضبط‌ها بر روند شناسایی تأثیر نگذارد. فرایند پیش‌پردازش شامل فریم‌بندی[۶]، پنجره‌گذاری[۷]، نرمال‌سازی[۸] و کاهش نویز[۹] است.

- **فریم‌بندی:** فریم‌بندی سیگنال که به آن قطعه‌بندی گفتار نیز می‌گویند، فرآیند تقسیم سیگنال‌های گفتاری پیوسته به بخش‌هایی با طول ثابت است. از آنجا که سیگنال‌های گفتار غیر ایستا[۱۰] هستند، احساسات می‌توانند در طول گفتار تغییر کنند. اگر چه گفتار برای مدت کوتاهی مانند ۲۰ تا ۳۰ میلی‌ثانیه ثابت است، با فریم‌بندی سیگنال گفتار می‌توان این حالت نیمه ثابت را تقریب زد و ویژگی‌های محلی را به‌دست آورد. به‌علاوه، ارتباط و اطلاعات بین قاب‌ها را می‌توان با هم‌پوشانی ۳۰٪ تا ۵۰٪ این قطعه‌ها حفظ کرد. در نتیجه به علت حفظ شدن اطلاعات مربوط به احساس در گفتار، فریم‌های با اندازه ثابت در سیگنال‌های گفتاری پیوسته برای استخراج ویژگی و طبقه‌بندی، مناسب هستند [۷].
- **پنجره‌گذاری:** مرحله بعدی بعد از فریم‌بندی سیگنال گفتار، معمولاً اعمال یک تابع پنجره روی فریم‌ها است. از تابع پنجره‌گذاری برای کاهش اثرات نشتی در هنگام تبدیل فوریه سریع

---

۱ Chinese Emotional Speech Corpus

۲ The Interactive Emotional Dyadic Motion Capture Database

۳ پایگاه‌داده گفتار احساسی فارسی

۴ Speech Under Simulated and Actual Stress Database

۵ Beihang University Database of Emotional Speech

۶ Framing

۷ Windowing

۸ Normalization

۹ Noise Reduction

۱۰ Non-Stationary

که به‌علت ناپیوستگی در لبه سیگنال‌ها ایجاد شده استفاده می‌شود. از متداول‌ترین توابع پنجره می‌توان به پنجره مستطیلی، پنجره هنینگ و پنجره همینگ اشاره کرد.

- **نرمال‌سازی:** یک مرحله مهم در پیش‌پردازش نرمال‌سازی ویژگی‌ها است که برای کاهش اثر تنوع گوینده، تنوع ناشی از محتوای واژگانی زمینه‌ای و تنوع ضبط بدون از دست دادن قدرت جداسازی ویژگی‌ها استفاده می‌شود. با استفاده از نرمال کردن ویژگی‌ها، قابلیت تعمیم ویژگی‌ها افزایش می‌یابد. نرمال‌سازی را می توان در سطوح مختلف مانند سطح ویژگی، سطح داده و یا سطح مدل انجام داد [۲۷].
- **کاهش نویز:** در زندگی واقعی، نویز موجود در محیط همراه با سیگنال گفتار ضبط می‌شود که این مساله بر میزان تشخیص تأثیر می‌گذارد. بنابراین برخی از تکنیک‌های کاهش نویز مانند حداقل میانگین مربعات خطا[۱] (MMSE) و دامنه لگاریتمیک حداقل میانگین مربعات خطا برای حذف یا کاهش این اطلاعات نویزی استفاده می‌شوند.

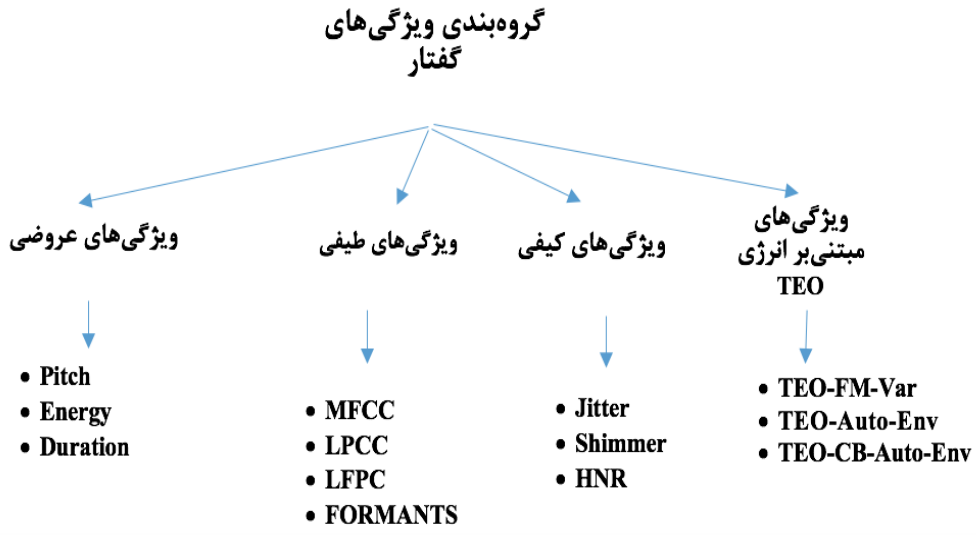
۲-۳ انواع روش‌های استخراج ویژگی در تشخیص احساس از گفتار

از آن‌جا که انتخاب مناسب ویژگی‌های گفتار بر عملکرد طبقه‌بندی تاثیر می‌گذارد، یکی از مراحل اصلی در روند تشخیص احساس از گفتار، مرحله استخراج ویژگی و تولید پارامترهایی از سیگنال گفتار جهت پردازش است. سیگنال گفتار ویژگی‌های زیادی دارد که عموماً به طیف لحظه‌ای سیگنال گفتار یا شکل مجرای گفتار مربوطند. پردازش این تعداد ویژگی برای کاربردی خاصی، همانند تشخیص احساس، کاری منطقی و عملی نخواهد بود، بدین منظور تبدیل‌هایی روی سیگنال گفتار انجام می‌شود تا بتوان ویژگی های مفید را استخراج نمود. استخراج ویژگی به دو دلیل انجام می‌شود: اولا سبب تمرکز روی اطلاعات

[۱] Minimum Mean Square Error

موجود در سیگنال می‌شود که این امر منجر به بهبود میزان شباهت و عدم شباهت میان کلاس‌های مختلف می‌شود. ثانیا داده‌ها را به نحو قابل ملاحظه‌ای کاهش داده، و محاسبات نیز به میزان زیادی کم می‌شود [۲۲].

در تحقیقات مختلف یک توافق عمومی درباره یک مجموعه ویژگی مناسب برای سیستم تشخیص احساس از گفتار وجود ندارد، اما در اکثر تحقیقات ویژگی‌های مختلفی مانند ویژگی‌های محلی ، ویژگی های سراسری ، ویژگی‌های پیوسته ، ویژگی‌های کیفی ، ویژگی‌های عروضی ، ویژگی‌های طیفی (ویژگی‌های مربوط به اطلاعات دستگاه صوتی)، ویژگی‌های مبتنی بر انرژی TEO ، و ویژگی‌های منبع تحریک بررسی شده‌اند [۲۳]. شکل ۲-۲ یک گروه‌بندی از انواع ویژگی‌ها را نشان می‌دهد.



شکل ۲-۲ یک گروه‌بندی از انواع ویژگی‌های گفتار

در ادامه، روش‌های استخراج هریک از این ویژگی‌ها را بررسی خواهیم کرد و سپس در پایان این بخش مروری خواهیم داشت بر تحقیقات مختلفی که در روش‌های استخراج ویژگی نوآوری داشته‌اند.

۲-۳-۱ ویژگی‌های عروضی

اطلاعات عروضی یا اطلاعات زبرزنجیری، جریان گفتار را ساختار می‌بخشد و شامل دیرش[۱]، شدت[۲]، لحن، زیر و بمی صدا[۳] و واحدهای صوتی است [۱۱]. انسان‌ها در حین تولید گفتار، الگوهای دیرش (مدت زمان)، لحن و شدت را به توالی واحدهای صوتی تحمیل می‌کنند. تلفیق این محدودیت‌های عروضی (مدت زمان، لحن و شدت) گفتار انسان را طبیعی‌تر می‌کند. در اکثر زبان‌ها ویژگی‌های عروضی (نوایی) نقش مهمی در انتقال اطلاعات معنایی به شنوندگان دارند و بخش اساسی رفتار زبانی را تشکیل می‌دهند. در واقع عروض یا نوا را می‌توان به‌عنوان ویژگی‌های گفتاری مرتبط با واحدهای بزرگتر مانند هجاها، کلمات، عبارات و جملات در نظر گرفت [۲۸]. در سنتز احساسات، تمرکز زیادی بر ویژگی‌های عروضی، یعنی توصیف لحن، شدت و ریتم گفتار در کنار ویژگی‌های کیفیت صدا وجود دارد [۲۹]. از این‌رو می‌توان از اطلاعات به‌دست آمده از این ویژگی‌ها برای تشخیص احساسات گوینده استفاده کرد. در بسیاری از تحقیقات هم‌بستگی‌های آکوستیک این ویژگی‌ها مانند انرژی، دیرش، و گام[۴] و مشتقات آن‌ها به‌عنوان ویژگی‌های موثر در تشخیص استفاده شده‌اند [۳۰] [۳۱] [۵]. همچنین در تحقیقات مختلف، این ویژگی‌ها بیشتر به‌صورت سراسری و ایستا در سطح کل سیگنال استخراج شده‌اند و تحقیقات کمتری روی ویژگی‌های محلی کار کرده‌اند [۳۲].

عروض را می توان در چهار سطح اصلی دسته‌بندی کرد: (الف) سطح زبان‌شناختی[۵]، (ب) سطح بیان[۶]، (ج) سطح ادراک صوتی[۷]، و (د) سطح ادراکی[۸] هستند [۳۳]. در سطح زبان‌شناختی، عروض به ارتباط عناصر مختلف زبانی یک جمله برای رسیدن به یک سطح موردنیاز از طبیعی بودن اشاره دارد.

---

۱ Duration
۲ Intensity
۳ Intonation
۴ Pitch
۵ Linguistic Level
۶ Articulatory Level
۷ Acoustic Realization Level
۸ Perceptual Level

به‌عنوان مثال، تفاوت‌های زبانی که می‌تواند از طریق تمایز بین پرسش و پاسخ یا تأکید معنایی بر یک عنصر برقرار شود. در سطح بیان، عروض از نظر فیزیکی به صورت مجموعه‌ای از حرکات مفصلی (حرکات شمرده شمرده) آشکار می‌شود. بنابراین، ابراز عروضی گفتار معمولاً شامل تغییرات دامنه حرکات مفصلی و همچنین تغییرات فشار هوا است، و فعالیت عضلات در سیستم تنفسی و همچنین در امتداد دستگاه صوتی منجر به انعکاس امواج صوتی می‌شود [۲۸]. ادراک صوتی عروض را می‌توان با استفاده از تجزیه و تحلیل پارامترهای آکوستیک مانند فرکانس پایه[۱] (F0)، شدت و مدت زمان (دیرش) مشاهده کرد. به‌عنوان مثال، هجاهای دارای تنش و استرس دارای فرکانس پایه بالاتر، دامنه بزرگتر و طول مدت بیشتری نسبت به هجاهای بدون استرس هستند. در آخرین سطح یعنی سطح ادراک، امواج صوتی گفتاری به گوش شنونده‌ای وارد می‌شود که به‌وسیله پردازش ادراکی، اطلاعات زبانی و زبان‌شناسی را از طریق نوا (عروض) استخراج می‌کند. در حین ادراک، عروض را می‌توان با توجه به تجربه ذهنی شنونده مانند مکث، طول، ملودی و بلندی صدای گفتار درک‌شده بیان کرد [۲۸]. پردازش یا تحلیل عروض از طریق تولید گفتار یا مکانیسم‌های ادراک دشوار است. از این‌رو از خصوصیات آکوستیک گفتار برای تجزیه و تحلیل عروض استفاده می‌شود.

در شکل ۲-۳ سه تصویر برای نمایش ویژگی‌های عروضی یک جمله که با پنج احساس مختلف بیان شده، نشان داده شده است [۳۴]، شامل: الف) مدت زمان (دیرش) دنباله هجاها، (ب) کانتورهای انرژی و (ج) کانتورهای گام. از تصویر اول می‌توان مشاهده کرد که الگوهای مدت زمان برای همه احساسات تقریبا یکسان و مشابه است همچنین شکل نشان می‌دهد که برای برخی از احساسات مانند ترس و خوشحالی، مدت زمان هجاهای ابتدایی بیشتر است، و به نظر می‌رسد برای احساسات خوشحالی و خنثی هجاهای میانی طولانی‌تر هستند، اما برای احساسات ترس و عصبانیت هجاهای آخر جمله طولانی‌تر(شکل ۲-۳-الف). از تصاویر مربوط به انرژی مشاهده می‌شود که گفتار با احساس خشم بالاترین انرژی را در طول کل سیگنال دارد. در کنار احساس خشم، ترس و خوشحالی نیز تا حدودی انرژی بیشتری نسبت

[۱] Fundamental Frequency

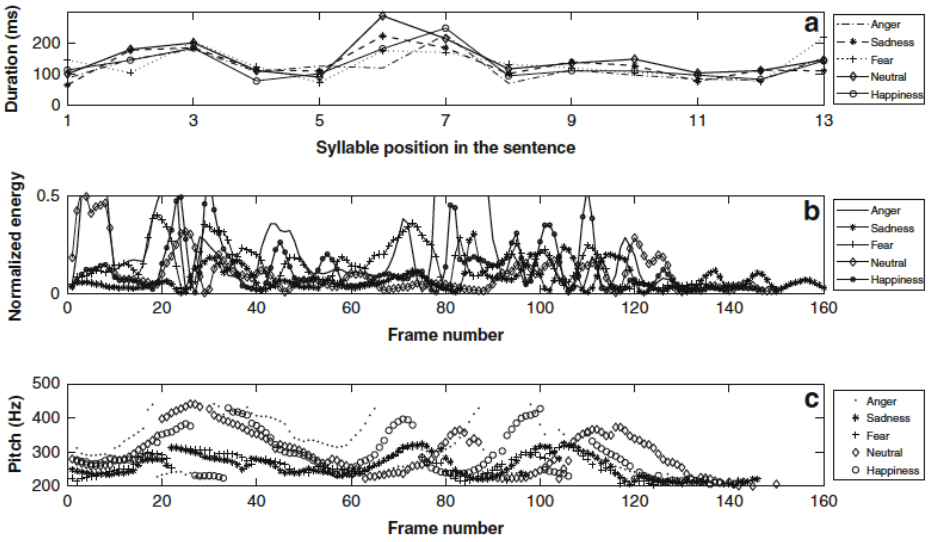
به دو احساس دیگر نشان می‌دهند. از پویایی خطوط انرژی می‌توان برای تشخیص ترس و خوشحالی استفاده کرد (شکل ۲-۳-ب). از شکل(۲-۳-ج) می‌توان مشاهده کرد که احساسات عصبانیت، شادی و خنثی دارای مقادیر گام کمی بالاتر از دو احساس دیگر هستند. با استفاده از پویایی[۱] (تغییر مقادیر عروضی برحسب زمان) کانتورهای گام، تشخیص آسانی بین احساسات خشم، شادی و خنثی امکان‌پذیر است، حتی اگر آن‌ها دارای مقادیر متوسط مشابهی باشند. بنابراین، با توجه به شکل ۲-۳ می‌توان انگیزه اساسی تحقیقات برای استفاده از پویایی ویژگی‌های عروضی در تشخیص احساس از گفتار را درک کرد. با مشاهده کانتورهای گام از شکل (۲-۳-ج)، مشاهده می‌شود که قسمت‌های اولیه نمودارها (دنباله ۲۰ مقدار گام) اطلاعات تشخیص‌دهنده یکسانی را ندارند. ویژگی‌های استاتیک برای احساسات خوشحالی و خنثی تقریباً یکسان هستند. با این‌حال، ممکن است از ویژگی‌های استاتیک برای تشخیص احساسات خشم، ناراحتی، و ترس استفاده شود زیرا مقادیر گام استاتیک آن‌ها بین ۲۵۰ تا ۳۰۰ هرتز گسترش یافته است. به طور مشابه ، ویژگی های پویا برای همه احساسات، به جز ترس، تقریباً یکسان است. می توان روند اولیه کاهش و افزایش تدریجی کانتورهای گام را برای احساسات عصبانیت، شادی، خنثی و غم مشاهده کرد، در حالی که برای احساس ترس کانتور گام با روند صعودی شروع می شود. خصیصه های تشخیص‌دهنده‌ی محلی یکسانی ممکن است در قسمت‌های اولیه، میانی و پایانی سیگنال گفتار در پارامترهای انرژی و مدت زمان نیز مشاهده شود. این پدیده نشان می‌دهد که طبقه‌بندی احساسات براساس ویژگی‌های عروضی سراسری یا محلی مستخرج از کل گفتار ممکن است گاهی دشوار باشد.

در جدول ۲-۲ خلاصه‌ای از تحقیقات انجام‌شده در تشخیص احساس از گفتار با استفاده از ویژگی‌های عروضی فهرست شده است. از جمله تحقیقاتی که از ویژگی‌های عروضی برای تشخیص احساس از گفتار استفاده کرده‌اند، تحقیق Lee و همکارانش [۳۵] است که ویژگی‌هایی مانند نرخ عبور از صفر[۲]، متوسط

[۱] Dynamic

[۲] Zero Crossing Rate

مربع ریشه انرژی[۱]، نسبت هارمونیک به نویز[۲] و ۱۲ ضریب کپسترال MFCC و مشتقات آن را از سیگنال‌های گفتار استخراج کرده و سپس با استفاده از روش درخت تصمیم دودویی سلسله مراتبی و طبقه‌بند ماشین بردار پشتیبان[۳] (SVM)، مدلی برای طبقه‌بندی احساسات عصبانیت، خوشحالی، غم و طبیعی پیشنهاد نموده‌اند.



شکل ۲-۳ نمایش ویژگی‌های عروضی یک جمله با پنج احساس: الف) مدت زمان (دیرش) دنباله هجاها، (ب) کانتورهای انرژی و (ج) کانتورهای گام

یکی دیگر از تحقیقات از مجموعه دادگان زبان چینی ماندارین استفاده کرده است که شامل ۵۴۰۰ گفتار در شش دسته احساس: غم، عصبانیت، ترس، تعجب، شادی و تنفر است [۲]. در این تحقیق یک مدل سه-سطحی برای دسته‌بندی پیشنهاد شده که در هر سطح با روش انتخاب ویژگی فیشر و ماشین بردار پشتیبان SVM ویژگی‌های موردنظر انتخاب شده و سپس با تحلیل مؤلفه اصلی[۴] (PCA) کاهش ویژگی صورت گرفته است و در ادامه از شبکه عصبی مصنوعی برای طبقه‌بندی استفاده کرده است.

---

[۱] Root Mean Square Energy

[۲] Harmonics to Noise Ratio

[۳] Support Vector Machine

[۴] Principal Component Analysis

متوسط نرخ تشخیص درست برای هر سطح به‌ترتیب ۸۶٫۵٪، ۶۸٫۵٪ و ۵۰٫۲٪ گزارش شده‌است.

جدول ۲-۲ خلاصه‌ای از تحقیقات انجام‌شده در تشخیص احساس از گفتار با استفاده از ویژگی‌های عروضی

منبع تحقیق	ویژگی‌های مورد استفاده	هدف و رویکرد تحقیق
[۳۶]	استخراج ویژگی‌های مبتنی‌بر گام، انرژی و توان از سطوح فریم، هجا و کلمات	• تشخیص ۴ احساس به زبان ماندارین • ۹۰٪ عملکرد تشخیص احساس با ترکیب ویژگی‌های سطوح فریم، هجا و کلمه
[۳۷]	ویژگی‌های مبتنی‌بر مدت زمان (دیرش)، انرژی و گام	• تشخیص احساسات در زبان ماندارین • انتخاب بهترین ویژگی‌ها با الگوریتم انتخاب رو به جلوی ترتیبی (SFS) • طبقه‌بندی: شبکه‌های عصبی مدولار
[۳۸]	کانتورهای F0 و آماره‌هایی مانند: میانگین، حداکثر، حداقل مقادیر و دامنه F0	• استخراج ویژگی‌ها از سطح جمله و سطح بخش صدادار از ۴دادگان پایگاه‌داده (۳ مجموعه دادگان انگلیسی، یک مجموعه دادگان آلمانی) • موثر بودن ویژگی‌های سطح جمله نسبت به ویژگی‌های سطح بخش صدادار • ۷۷٪ عملکرد تشخیص احساس
[۳۵]	ویژگی‌هایی مانند نرخ عبور از صفر، متوسط مربع ریشه انرژی، نسبت هارمونیک به نویز، ۱۲ ضریب کپسترال MFCC و مشتقات آن	• پایگاه‌داده AIBO و پایگاه‌داده USC IEMOCAP • روش درخت تصمیم دودویی سلسله‌مراتبی و طبقه‌بند ماشین بردار پشتیبان • متوسط نرخ تشخیص ۵۸٫۴۶٪ و ۴۱٫۵۷٪ به‌ترتیب روی پایگاه‌داده‌های USC IEMOCAP و AIBO
[۲]	انرژی، نرخ عبور از صفر، انرژی برابر با نرخ عبور از صفر، گام، فرمنت‌های اول تا سوم، طیف مرکزی، فرکانس برش طیف، چگالی همبستگی، بعد فرکتال، انرژی پنج باند فرکانس Mel	• یک مدل سه-سطحی برای دسته‌بندی با روش انتخاب ویژگی فیشر و ماشین بردار پشتیبان SVM در هر سطح • تحلیل مؤلفه اصلی و شبکه عصبی مصنوعی برای طبقه‌بندی نهایی • متوسط نرخ تشخیص درست برای هر سطح به‌ترتیب ۸۶٫۵٪، ۶۸٫۵٪ و ۵۰٫۲٪

۲-۳-۲ ویژگی‌های طیفی (ویژگی‌های مربوط به اطلاعات دستگاه صوتی)

ویژگی‌های مستخرج از دستگاه صوتی اصولا به‌عنوان ویژگی‌های طیفی شناخته می شوند. مشخصه‌های دستگاه صوتی در حوزه فرکانس نشان داده می‌شوند، به همین علت ویژگی‌های طیفی با تبدیل سیگنال حوزه زمان به سیگنال حوزه فرکانس با استفاده از تبدیل فوریه به‌دست می‌آیند. به‌طورکلی برای استخراج ویژگی‌های دستگاه صوتی یک فریم گفتار به طول ۲۰-۳۰ میلی‌ثانیه مورد نیاز است. برای تجزیه و تحلیل ویژگی‌های مختلف از طیف گفتار مانند فرمنت‌ها[۱]، پهنای باند آن‌ها، انرژی طیفی و شیب سیگنال یافتن تبدیل فوریه یک فریم سیگنال ضروری است. ضرایب کپسترال یک فریم گفتار با گرفتن تبدیل فوریه روی طیف اندازه لگاریتم به دست می آید. در تحقیقات، ویژگی‌های طیفی برای مدل کردن الگوی زیر و بمی و فرکانس گام گفتار یک گوینده استفاده می‌شوند [۳۹]. تغییرات در فرکانس گام همبستگی مهمی با احساسات در گفتار دارد و همچنین در تحقیقات اثبات شده که این تغییرات منجر به تغییرات در سایر پارامترهای عروضی مانند طول و انرژی می‌شود. در واقع ویژگی‌های طیفی نشان‌دهنده اطلاعات آوایی هستند که مستقیما از طیف سیگنال گفتار مشتق شده‌اند. این ویژگی‌های مشتق‌شده از طیف، با استفاده از بانک‌های فیلتر، به سهم برابر و یکسان تمام مولفه‌های فرکانسی در پردازش یک سیگنال گفتار تاکید دارند. در اکثر تحقیقات مربوط به تشخیص احساسات از گفتار، ویژگی‌های طیفی مانند ضرایب کپسترال پیش‌بینی خطی[۲] (LPCC)، ضرایب کپسترال فرکانس مل[۳] (MFCC)، ضرایب پیش‌بینی خطی ادراکی[۴] (PLPC) و فرمنت‌ها به‌عنوان اطلاعات مستخرج از دستگاه صوتی استفاده شده‌اند. این ویژگی‌های دستگاه صوتی به‌عنوان ویژگی‌های سگمنتال[۵]، طیفی یا سیستمی نیز شناخته می‌شوند [۱۳]. شکل ۲-۴ مشخصه‌های طیفی را برای هشت احساس از مجموعه دادگان IITKGP-

---

۱ Formants

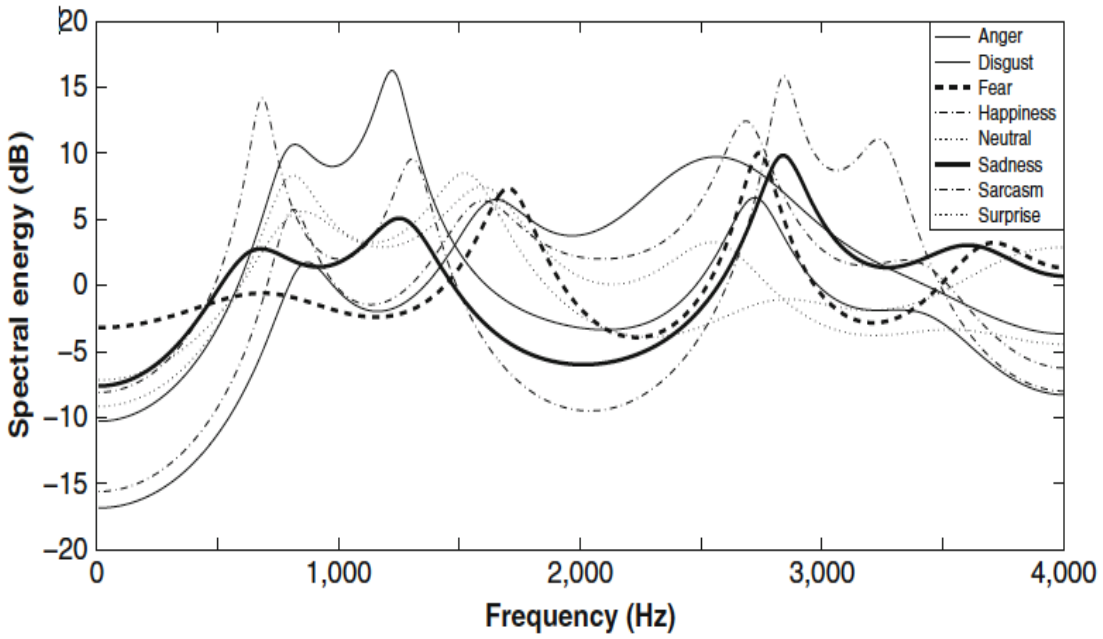
۲ Linear Prediction Cepstral Coefficients

۳ Mel Frequency Cepstral Coefficients

۴ Perceptual Linear Prediction Coefficients

۵ Segmental

SESC نشان می‌دهد. طیف‌های ترسیم شده در شکل ۲-۴ نشان‌دهنده یک ناحیه ثابت واکه / A / از یک جمله که در هشت احساس مختلف بیان شده است [۳۴]. همانطور که در شکل مشخص است تیزی قله‌های فرمنت‌ها، مکان فرمنت‌ها، پهنای باند فرمنت‌ها و شیب طیفی دارای خصوصیات متمایزی برای احساسات مختلف هستند. بسیاری از تحقیقات مربوط به تشخیص احساسات از گفتار با استفاده از ویژگی‌های طیفی مشتق‌شده از کل سیگنال گفتار از طریق یک رویکرد متداول پردازش فریم انجام می‌شود. در ادامه ابتدا به بررسی انواع این ویژگی‌ها می‌پردازیم و سپس مروری بر تحقیقات انجام‌شده روی این دسته ویژگی‌ها خواهیم داشت.



شکل ۲-۴ طیف یک ناحیه ثابت واکه / A/ از یک جمله که در هشت احساس مختلف [34]

- **ضرایب کپسترال پیش‌بینی خطی (LPCC)**

یکی از ویژگی‌های مستخرج مربوط به اطلاعات دستگاه صوتی، ضرایب کپسترال مشتق‌شده از تحلیل پیش‌بینی خطی یعنی LPCC است. تشخیص احساسات توسط انسان به دو عامل بستگی دارد: بیان آن توسط گوینده و همچنین درک آن توسط شنونده. هدف از استفاده از ضرایب LPCC در

نظر گرفتن مشخصه‌های دستگاه صوتی گوینده در حین انجام تشخیص خودکار احساسات است. از این جهت، ایده اصلی تحلیل پیش‌بینی خطی این نکته است که می‌توان نمونه‌ی nام یک سیگنال گفتار را با استفاده از یک ترکیب خطی از p نمونه قبلی مطابق معادله ۲-۱ به‌دست آورد.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots a_p s(n-p)$$ معادله ۲-۱

در رابطه فوق ضرایب $a_1 . a_2 . a_3 . \cdots$ به‌عنوان ضرایب پیش‌بینی خطی در هر فریم محاسبه می‌شوند. برای محاسبه مجموعه مقادیر یکتای این ضرایب، باید مجموع مربع تفاوت بین نمونه‌های گفتار پیش‌بینی‌شده و نمونه‌های گفتار واقعی، یعنی خطای پیش‌بینی، حداقل شود. برای این کار ابتدا از خطای پیش‌بینی برحسب هر یک از ضرایب a_k مشتق گرفته و سپس مساوی صفر قرار داده تا ضرایب به‌دست آیند. ضرایب کپسترال C_m به‌دست آمده از a_k به‌صورت بازگشتی از معادله ۲-۲ به‌دست می‌آیند.

$$C_0 = \log_e p$$
$$C_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} C_k a_{m-k} \quad \text{معادله ۲-۲}$$
$$for\ 1 < m < p$$

- **ضرایب کپسترال فرکانس مل (MFCC)**

دسته دوم ویژگی‌های مستخرج از اطلاعات طیفی، ضرایب کپستروم فرکانس مل (MFCC) است که نمایانگر طیف توان زمان کوتاه یک فریم گفتار با استفاده از تبدیل خطی کسینوسی لگاریتم طیف توان روی یک مقیاس فرکانس مل غیرخطی است. همان‌طور که در تحقیقات به‌دست آمده، سیستم شنوایی انسان سیگنال گفتار را در یک روش غیرخطی پردازش می‌کند. همچنین اثبات شده که مولفه‌های فرکانسی پایین‌تر یک سیگنال حاوی اطلاعات مشخص‌تر واج هستند، از این‌رو یک بانک فیلتر مقیاس مل غیرخطی برای پیش‌تاکید مولفه‌های فرکانسی پایین‌تر سیگنال در میان فرکانس‌های بالاتر استفاده می‌شود. برای به‌دست آوردن ضرایب MFCC بعد از پیش‌تاکید سیگنال گفتار و تقسیم آن به فریم‌هایی با طول ۲۰ میلی‌ثانیه و هم‌پوشانی ۱۰ میلی‌ثانیه و اعمال پنجره همینگ روی هر فریم، اندازه طیف برای

هر فریم با استفاده از تبدیل فوریه گسسته به‌دست می‌آید. سپس سیگنال ضرایب فوریه از میان یک بانک فیلتر مل عبور داده می‌شود تا طیف مل بدست آید. در پایان روی لگاریتم ضرایب فرکانسی مل به‌دست آمده یک تبدیل کسینوسی گسسته اعمال می‌شود تا ضرایب MFFC موردنظر استخراج شوند. از آنجا که این ضرایب شامل اطلاعات هر فریم مشخص هستند معمولا به‌عنوان ویژگی‌های ایستا لحاظ می‌شوند. برای به‌دست آوردن اطلاعات بیشتر درباره داینامیک‌های زمانی سیگنال می‌توان مشتقات اول و دوم این ضرایب را نیز به‌عنوان ویژگی در کنار ضرایب اصلی استفاده کرد [۴۰] [۴۱]. روال استخراج ضرایب MFCC و مشتقات آن در معادله ۲-۳ تا معادله ۲-۵ آمده است.

معادله ۲-۳

$$X(k) = \sum_{n=۰}^{N-۱} x(n) e^{\frac{-\pi n k ۲ j}{N}} \; ; ۰ \le k \le N-۱$$

معادله ۲-۴

$$s(m) = \sum_{k=۰}^{N-۱} \left[\left| X(k) \right|^{۲} H_m(k) \right] ; ۰ \le m \le M-۱$$

معادله ۲-۵

$$C(n) = \sum_{m=۰}^{M-۱} \log_{۱۰} \left(s(m) \right) \cos\left(\frac{\pi n (m - ۰.۵)}{M} \right)$$

$$n = ۰, ۱, ۲, \cdots, C-۱$$

در روابط فوق ضرایب فوریه گسسته، طیف مل اندازه فوریه و ضرایب MFCC به‌ترتیب با روابط $X(k)$، $s(m)$ و $C(n)$ مشخص شده‌اند. که N تعداد نقاط مورد استفاده در محاسبه تبدیل فوریه گسسته، M تعداد فیلترهای بانک فیلتر مل و C نیز تعداد ضرایب MFCC است. همچنین سیگنال اولیه $x(n)$ و $H_m(k)$ نیز مربوط به فیلتر بانک مل است.

- **ضرایب پیش‌بینی خطی ادراکی (PLPC)**

ایده اصلی روش PLP این است که قدرت تفکیک فرکانسی و حساسیت نسبت به تغییر فرکانس در گوش انسان در فرکانس‌های مختلف، یکسان نیست و به‌علاوه حساسیت گوش نسبت به شدت و انرژی صوت در فرکانس‌های مختلف متفاوت است. بر این اساس نشان‌داده شده است که در گوش انسان میزان

احساس بلندی با ریشه سوم انرژی آن متناسب است. این روش برای به‌دست آوردن تخمینی از طیف شنوایی از سه مفهوم روان‌شناختی شنوایی استفاده می‌کند: (۱) وضوح طیفی باند بحرانی (رزولوشن فرکانسی گوش انسان را شبیه‌سازی می‌کند) ، (۲)منحنی بلندی صدا (احساس یکسان نبودن بلندی صدا در فرکانس‌های مختلف را جبران می‌کند)، و (۳) قانون قدرت شدت صدا (رابطه‌ای غیرخطی بین شدت صدا و بلندی صدای دریافت‌شده شبیه‌سازی می‌کند.). بنابراین در مقایسه با تحلیل پیش‌بینی خطی معمولی (LPCC)، تجزیه و تحلیل PLP با شنوایی انسان سازگارتر است.

- **فرمنت‌ها**

همان‌طور که در تحقیقات بیان شده، نواحی با دامنه بالاتر در یک طیف سیگنال گفتار، مانند فرمنت‌ها، نسبتا کمتر تحت تاثیر نویز قرار می‌گیرند [۴۲]. با این دیدگاه، می‌توان از پارامترهای فرمنت به‌عنوان ویژگی‌های مکمل ویژگی‌های کپسترال استفاده کرد. همچنین این نکته نیز باید موردتوجه قرار گیرد که ویژگی‌های کپسترال فقط اطلاعات دامنه (انرژی) طیف توان گفتار را به‌کار می‌برند، درحالی‌که فرمنت‌ها از اطلاعات فرکانس هم استفاده می‌کنند. به‌طورکلی فرمنت‌ها توالی‌هایی از اشکال دستگاه صوتی را نشان می‌دهند که می‌توان با تحلیل آن‌ها با استفاده از قدرت، موقعیت و پهنای باندشان اطلاعات خاص احساسی دستگاه صوتی را استخراج کرد.

در ادامه در جدول ۲-۳ خلاصه‌ای از تحقیقات انجام‌شده در تشخیص احساس از گفتار با استفاده از ویژگی‌های طیفی و مستخرج از دستگاه صوتی فهرست شده است.

۲-۳-۳ ویژگی‌های کیفی

کیفیت صدا توسط خصیصه‌های فیزیکی دستگاه صوتی تعیین می‌شود. تغییرات غیرارادی ممکن است یک سیگنال گفتاری ایجاد کند که احساسات را با استفاده از ویژگی‌هایی مانند جیتر[۱]، شیمر[۲] و نسبت هارمونیک به نویز (HNR) متمایز کند. یک همبستگی قوی بین کیفیت صدا و محتوای احساسی گفتار

[۱] Jitter

[۲] Shimmer

وجود دارد [۴۳]. Jitter تغییرات فرکانس پایه بین چرخه‌های ارتعاشی متوالی است، در حالی که shimmer تغییرات دامنه است. Jitter معیار بی‌ثباتی فرکانس است، درحالی‌که shimmer ناپایداری دامنه است. نسبت هارمونیک به نویز، اندازه سطح نسبی نویز در طیف فرکانس واکه‌ها است. این نسبت بین مؤلفه‌های تناوبی به غیرتناوبی در سیگنال‌های گفتار صدادار است. این تغییرات به عنوان تغییر در کیفیت صدا مشاهده می‌شوند. در جدول ۲-۳ خلاصه‌ای از تحقیقات انجام‌شده در تشخیص احساس از گفتار با استفاده از ویژگی‌های کیفی فهرست شده است.

جدول ۲-۳ خلاصه‌ای از تحقیقات انجام‌شده در تشخیص احساس از گفتار با استفاده از ویژگی‌های طیفی و کیفی

تحقیق	ویژگی‌ها	هدف و رویکرد تحقیق
[۴۴]	MFCC	• پایگاه داده از کنسرسیوم داده های زبانی • برچسب گذاری هر فریم با استفاده از خوشه بندی MFCC چند الگویی • دقت تشخیص ۶۶,۴٪ با طبقه بند HMM
[۴۵]	ویژگی‌های طیفی: MFCC و مشتقات- ویژگی‌های عروضی- ویژگی‌های کیفی: جیتر و شیمر- ویژگی‌های مبتنی‌بر Teo	• پایگاه داده صوتی و تصویری ۱۳۹ نوجوان (۹۳ زن و ۴۶ مرد) • طبقه بند GMM و SVM • ترکیب خصوصیات گلوت با ویژگیهای عروضی و طیفی- دقت طبقه بندی (بین ۶۷٪ - ۶۹٪ برای مردان و ۷۰٪ - ۷۵٪ برای زنان) • دقت طبقه بندی با ویژگی‌های مبتنی‌بر Teo (بین ۸۱٪ - ۸۷٪ برای مردان و ۷۲٪ - ۷۹٪ برای زنان)
[۴۶]	۱۲ تابع توصیف کننده سطح پایین (LLD) و ضرایب رگرسیون مشتق مرتبه اول آن ها: ZCR, RMS Energy, F0, HNR, MFCC	• پایگاه داده: ABC , SUSAS • طبقه بند Denoising autoencoders و SVM • ۶۴,۱۸٪ میانگین میزان تشخیص در ABC، و ۶۲,۷۴٪ میانگین دقت در SUSAS
[۴۷]	ترکیبی از انرژی ، ویژگی‌های عروضی گام، و ویژگی‌های طیفی MFCC	• پایگاه داده: گفتار احساسی برلین و اسپانیا • یک روش همجوشی ویژگی با استفاده از LDA, RDA, SVM, KNN • ۲۰٪ بهبود عملکرد در مقایسه با عملکرد هر طبقه بندی با ویژگی های جداگانه
[۴۸]	F0، احتمال صدا، ZCR، MFCC با لگاریتم انرژی و مشتقات مرتبه اول	• یک مدل شبکه عصبی RNN و یک روش یادگیری قدرتمند با یک مدل حافظه کوتاه مدت بلند مدت BLSTM • بهبود دقت تشخیص تا ۱۲٪ نسبت به شبکه عصبی DNN
[۴۹]	• توصیف کننده های LLD مستخرج متشکل از F0، انرژی فریم، ZCR و MFCC	• پایگاه داده: IEMOCAP • آموزش خودکار توسط Deep RNN و توصیف کننده های LLD مستخرج متشکل از F0، انرژی فریم، ZCR و MFCC • دقت سیستم پیشنهادی: با ویژگیهای طیفی خام ۶۱,۸٪ - با ویژگی های LLD دارای ۶۳,۵٪

۲-۳-۴ ویژگی‌های منبع تحریک

ویژگی‌های گفتاری مشتق‌شده از سیگنال منبع تحریک به‌عنوان ویژگی‌های منبع شناخته می‌شوند. سیگنال منبع تحریک پس از مانع شدن بر خصیصه‌های دستگاه صوتی (VT) از گفتار بدست می‌آید. این مهم بدین صورت است که ابتدا اطلاعات VT با استفاده از ضرایب فیلتر (LPC) از سیگنال گفتاری پیش‌بینی شده و سپس با فرمول معکوس فیلتر جدا می‌شوند [۲۸]. سیگنال حاصل به‌عنوان باقی‌مانده پیش‌بینی خطی[۱] شناخته می‌شود و بیشتر شامل اطلاعات مربوط به منبع تحریک است. به دلایل مطرح‌شده در زیر [۲۸]، تلاش‌های بسیار کمی در تحقیقات برای استخراج اطلاعات منبع تحریک در توسعه سیستم‌های تشخیص احساس از گفتار انجام شده است:

۱. محبوبیت ویژگی‌های طیفی.

۲. سیگنال تحریک (باقی‌مانده LP) به‌دست آمده از تجزیه و تحلیل LP به دلیل مولفه‌های غیرقابل پیش‌بینی سیگنال گفتار، بیشتر به‌عنوان یک سیگنال خطا مشاهده می‌شود.

۳. باقی‌مانده LP اساساً شامل روابط مرتبه بالاتر است و پیاده‌سازی و حل این روابط مرتبه بالاتر نامشخص و دشوار است.

۲-۳-۵ سایر روش‌های استخراج ویژگی

روند اخیر تحقیقات در زمینه تشخیص احساس از گفتار، بر استفاده از ترکیبی از ویژگی‌های مختلف برای دستیابی به بهبود عملکرد تشخیص تأکید دارند. ویژگی‌های کیفی، منبع تحریک، دستگاه صوتی و عروضی که در بخش‌های قبلی مورد بحث قرار گرفتند، عمدتاً حاوی اطلاعات منحصر به فرد سیگنال گفتار، و از نظر ذاتی مکمل یکدیگر هستند. درنتیجه انتظار می‌رود یک ترکیب هوشمند از ویژگی‌های مکمل باعث بهبود عملکرد مورد نظر سیستم شود.

[۱] Linear Prediction Residual

در بسیاری از تحقیقات، ویژگی‌های استخراج شده از گفتار به دو دسته تقسیم شده‌اند: توصیف‌گرهای سطح پایین[۱] (LLD) و ویژگی‌های عملکردی[۲] [۵۰]. LLD ها شامل ویژگی‌های عروضی و ویژگی‌های طیفی و مشتقات آن‌ها مانند گام (F0)، انرژی، فرمنت‌ها، ضرایب کپسترال فرکانس مل (MFCC) و ضرایب کپسترال پیش‌بینی خطی (LPCC) هستند. ویژگی‌های عملکردی مشتقات آماری مانند میانگین، حداکثر، حداقل و نرخ عبور از صفر را شامل می‌شوند. در ۱۰ سال گذشته، علی‌رغم محبوبیت LLDها، اشکال جدیدی از ویژگی‌ها مانند ضرایب کپسترال بسته موجک کویفلت[۳] (CWPCC) [۵۱] و کیسه کلمه‌های صوتی[۴] (BoAW) [۵۲][۵۳] نیز ارائه شده‌اند . با این‌حال، کاربرد این ویژگی‌ها در SER محدود بوده و مقایسه کمی در مورد عملکرد آن‌ها با LLDها انجام‌شده است. به‌طوری‌که می‌توان گفت به احتمال زیاد تمرکز اکثر تحقیقات بر روی LLDها است، و هر ویژگی جدیدی که توسط محققان ارائه می‌شود ، با LLDها محک زده می‌شود.

در یکی از تحقیقات، آلونسو و همکارانش یک سیستم تشخیص احساس از گفتار با استفاده از دو نوع ویژگی شامل دو ویژگی عروضی و چهار ویژگی زبان‌شناسی مربوط به گام و تعادل انرژی طیفی، و طبقه‌بند SVM برای طبقه‌بندی احساساتی مانند خوشحالی، خستگی، خنثی، عصبانیت و غم در دادگان آلمانی (EMO-DB)، انگلیسی (LDC) و لهستانی پیشنهاد داده‌اند [۵۴]. نتایج این تحقیق نشان می‌دهد که دقت طبقه‌بندی مختلف برای سه پایگاه داده در مقایسه با روش‌های دیگر در برنامه های زمان واقعی پیچیدگی پایین‌تری دارد (دقت‌های مختلفی مانند ۹۴,۹٪ برای EMO-DB، ۸۸,۳۲٪ برای LDC، و ۹۰٪ برای پایگاه‌داده لهستانی). یکی دیگر از ویژگی‌های مورداستفاده در SER ویژگی‌های مستخرج از ضرایب فوریه است. در تحلیل فوریه، یک سیگنال به ارتعاشات سینوسی تشکیل‌دهنده‌اش تجزیه می‌شود و اگر پریودیک باشد می‌تواند برحسب سری‌هایی از موج‌های سینوسی و کسینوسی

[۱] Low-Level Descriptors (LLDs)
[۲] Functional Features
[۳] Coiflet Wavelet Packet Cepstral Coefficients
[۴] Bag of Audio Word

هارمونیکال مربوطه مانند ضرایب صحیح یک فرکانس پایه توصیف شود [۴۱]. به‌عبارت دیگر، یک سیگنال گفتار می‌تواند به‌صورت خروجی عبور یک شکل موج تحریک چاک‌نای از میان یک فیلتر خطی زمان-متغیر نمایش داده شود، که خصیصه‌های تشدیدشده مجرای صوتی را مدل می‌کند. در تحقیقی دیگر [۴۱] ویژگی‌های مربوط به پارامترهای فوریه و مشتقات مرتبه اول و دوم آن‌ها برای تشخیص احساس از گفتار استخراج شده‌اند. در این تحقیق سیگنال گفتار $x(m)$ به L فریم تقسیم‌شده به‌طوری‌که می‌توان آن را به‌صورت ترکیبی از پارامترهای فوریه مطابق با روابط (۶) و (۷) نمایش داد:

معادله ۲-۶

$$x(m) = \sum_{k=1}^{M} H_k^l(m) \left(cos\left(\pi 2 \frac{f_k^l}{F_s} m \right) + \varphi_k^l \right)$$

معادله ۲-۷

$$\mathrm{H(k)} = \sum_{m=0}^{N-1} x(m) e^{-j\frac{\pi 2}{N}mk} \quad k = 0.1.2.\ldots.N-1$$

بخش هارمونیک مدل گفتار یک نمایش سری فوریه از مؤلفه‌های پریودیک سیگنال گفتار است، که این هارمونیک‌ها شامل فرکانس، دامنه و فاز هستند. همانطور که در رابطه ۶ نمایش داده شد است H_k^l، پارامتر فوریه فریم Lم است و H_i دامنه پارامتر فوریه iم است (مقادیر متوسط H_i). بنابراین یک بردار ویژگی گفتار H_k برای همه فریم های سیگنال گفتار از فریم ۱ تا فریم L تخمین زده می شود.

قابل ذکر است که از سال ۲۰۱۲ به بعد کارهای زیادی روی تشخیص احساس با استفاده از ویژگی‌های مستخرج از دستگاه صوتی و اطلاعات عروضی انجام شده است، از جمله نتایج منتشرشده در یک تحقیق [۵۵] نشان می‌دهد که مجموعه ویژگی‌های مستخرج از دستگاه صوتی در کنار اطلاعات عروضی و منبع تحریک می‌توانند نرخ تشخیص را تا ۷۹,۱۴٪ بهبود بخشند. چن و همکارانش [۲] روی دادگان گفتار احساسی زبان ماندارین(BHUDES) علاوه‌بر ویژگی‌هایی مانند گام، فرمنت‌ها، انرژی و ضرایب MFCC، بُعد فرکتال و چگالی همبستگی نیز از هر فریم استخراج شده و با روش‌های تحلیل مجزاساز خطی فیشر و تحلیل مؤلفه اصلی کاهش ویژگی انجام شده است. سپس یک مدل تشخیص احساس سه-سطحی برای طبقه‌بندی دوبه‌دوی کلاس‌ها تشکیل داده که در سه سطح به‌ترتیب با متوسط نرخ تشخیص

۸۶,۵٪، ۶۸,۵٪ و ۵۰,۲٪ برای هر سطح طبقه‌بندی انجام می‌دهد. در تحقیقی دیگر [۵۶] که روی دادگان SAVEE انجام شده، از پارامتر کپسترال مبتنی‌بر زیر-باند (SBC)[۱] و ضرایب MFCC برای طبقه‌بندی شش کلاس احساسی استفاده کرده است که بهترین نرخ تشخیص ۷۹٪ گزارش شده‌است. مائو و همکارانش [۵۷] از معماری شبکه عصبی کانولوشنال برای یک مدل دومرحله‌ای تشخیص احساس از گفتار استفاده کرده‌اند که مرحله اول شامل یادگیری ویژگی‌های ثابت محلی با استفاده از یک رمزگذار خودکار اسپارس برای طیف‌نگار گفتار است، و در مرحله بعد با استفاده از تحلیل مؤلفه اصلی (PCA) ویژگی‌های مجزاساز جهت تشخیص استخراج می‌شوند. نتایج به‌دست آمده روی چهار مجموعه دادگان در این تحقیق نشان می‌دهد که متوسط نرخ تشخیص درست ۷۹% است.

۲-۴ انواع روش‌های انتخاب ویژگی و کاهش ابعاد در تشخیص احساس از گفتار

همان‌طورکه در بخش قبلی بیان شد، در تشخیص احساس از سیگنال گفتار ویژگی‌های مختلفی استفاده می‌شوند که این ویژگی‌ها گاهی از کل سیگنال گفتار و گاهی از یک فریم از آن محاسبه و استخراج می‌شوند. باتوجه به پیشرفت سریع کامپیوتر و تکنولوژی در زمینه پایگاه‌داده، سرعت جمع‌آوری داده‌ها بسیار بیشتر از سرعت پردازش اطلاعات است و به دلیل افزایش متغیرهای مربوط به یک نمونه، نیاز به پیش‌پردازش‌هایی قبل از پردازش اصلی داده‌ها حائز اهمیت است. انتخاب ویژگی‌ها از مهم‌ترین روش‌ها و تکنیک‌ها جهت پیش‌پردازش داده برای کاوش داده است. انتخاب ویژگی فرآیندی است که زیرمجموعه‌ای از ویژگی‌ها را براساس یک معیار بهینه‌سازی انتخاب می‌کند. در مساله انتخاب ویژگی، هدف کاهش اندازه داده‌های گفتاری، با انتخاب بخش‌های مهم و مرتبط از ویژگی‌ها و حذف موارد نامناسب، یا تولید چندین ویژگی جدید است که شامل اطلاعات بیشتری است. انتخاب ویژگی در مرحله

[۱] Sub-Band based Cepstral Parameter (SBC)

قبل از پردازش، در کاهش ابعاد، حذف داده‌های نامناسب، افزایش دقت یادگیری و بهبود تشخیص، به طبقه‌بند کمک می‌کند. کاهش ابعاد، فرآیند کاهش تعدادی از ویژگی‌هایی است که در طبقه‌بندی کمتر در نظر گرفته می‌شوند. در SER برای ذخیره‌سازی و محاسبات مورد نیاز طبقه‌بندی و داشتن یک درون‌یابی در تفکیک ویژگی‌ها، کاهش ابعاد انجام می‌شود و می‌توان کاهش ابعاد را به ویژگی‌های انتخابی و ویژگی‌های مستخرج تقسیم کرد [۱۱] [۵۸]. انتخاب ویژگی و کاهش ابعاد گام‌های مهمی در تشخیص احساسات هستند. ضرورت استفاده از الگوریتم‌های انتخاب ویژگی در این سیستم‌ها به این دلیل است که در بین ویژگی‌های زیادی که موجود است مجموعه مشخص و خاصی از ویژگی‌ها برای مدل‌سازی احساسات وجود ندارد. همان‌طورکه در تحقیقات بیان‌شده، تعداد زیاد ویژگی‌ها باعث می‌شود طبقه‌بندها با مزاحمت ابعاد[۱]، افزایش زمان آموزش و بیش‌برازش[۲] مواجه شوند و میزان پیش‌بینی تحت تأثیر قرارگیرد. بنابراین انتخاب ویژگی فرآیند انتخاب یک زیرمجموعه مرتبط و مفید از مجموعه داده‌ها، و شناسایی و حذف ویژگی‌های غیرضروری، زاید یا بی‌ربط در راستای افزایش دقت مدل پیش‌بینی است [۵۹]. اگرچه بسیاری از روش‌های انتخاب ویژگی در مطالعات مختلف SER استفاده شده‌اند، اما برخی از این تکنیک‌ها پس از کاهش ابعاد موفقیت و دقت SER را کاهش می‌دهند [۶۰].

در تحقیقات عموما در فاز انتخاب ویژگی روش‌های فیلترینگ، مبتنی‌بر بسته‌بندی[۳]، الگوریتم‌های تکاملی و روش‌های کاهش ابعاد استفاده می‌شوند. در روش‌های فیلترینگ از یک تکنیک رتبه‌بندی ویژگی‌ها به‌همراه معیاری برای دسته‌بندی آن‌ها استفاده می‌شود، مانند روش‌های رتبه‌بندی اطلاعات متقابل[۴] و معیار همبستگی[۵]. در این روش‌ها برای هر ویژگی یک رتبه یا امتیاز (با در نظر گرفتن یک حد آستانه) محاسبه می‌شود و سپس براساس این رتبه‌بندی ویژگی‌ها دسته‌بندی شده و ویژگی‌های با امتیاز کمتر حذف می‌شوند. در روش‌های مبتنی‌بر بسته‌بندی از یک تابع دسته‌بندی برای ارزیابی

---

۱ Curse of Dimensionality
۲ Overfitting
۳ Wrapper-based
۴ Mutual Information (MI)
۵ Correlation criteria

شایستگی زیرمجموعه‌های مختلف ویژگی استفاده شده و سپس زیرمجموعه ویژگی‌های مناسب توسط یک الگوریتم جستجو انتخاب می‌شوند. الگوریتم‌های تکاملی نیز روش‌های اکتشافی هستند که از فرآیندهای تکامل بیولوژیکی الهام گرفته و برای طیف وسیعی از مشکلات بهینه‌سازی مفید هستند. برخی از روش‌های انتخاب ویژگی و کاهش ابعاد مورداستفاده در SER عبارتند از: تجزیه و تحلیل مؤلفه‌های اصلی[۱] (PCA)، تجزیه و تحلیل خطی[۲] (LDA)، رویکرد بسته‌بندی با انتخاب رو به جلو، انتخاب ویژگی رو به جلو[۳] (FFS) و انتخاب ویژگی رو به عقب[۴] (BFS) و انتخاب رو به جلو شناور متوالی[۵] (SFFS) [۶۱][۶۲][۶۳][۶۴]. به‌عنوان مثال در مطالعه‌ای [۶۵] با استفاده از درختان تصمیم‌گیری و ارائه یک الگوریتم دومرحله‌ای، از بین ویژگی‌های استخراج‌شده عروضی و کپسترال، یک مجموعه ویژگی انتخاب شده است. در این تحقیق در مرحله اول با یک درخت تصمیم‌گیری c4.5 و یک الگوریتم جنگل تصادفی، ویژگی ها در دو دسته قرار می‌گیرند و سپس با یک روش رای‌گیری اکثریت، یک مجموعه ویژگی انتخاب می شود. نتایج با استفاده از طبقه‌بند KNN و مجموعه دادگان چینی ٪۷۲,۲۵ دقت تشخیص را نشان می‌دهند. در یکی از تحقیقاتی که در سال۲۰۱۶ انجام شده [۶۶] یک روش بهینه‌سازی تکاملی برای جستجوی یک بانک فیلتر پیشنهاد شده است که حداکثر دقت طبقه‌بندی را در تشخیص گفتار با استرس و احساسی داشته باشد. نتایج این تحقیق روی دو مجموعه دادگان هندی و آلمانی به‌ترتیب ۹۱,۳% و ۴۲,۵% تشخیص درست گزارش شده است. در جدول ۲-۴ خلاصه‌ای از تحقیقات انجام‌شده مربوط به روش‌های انتخاب ویژگی در تشخیص احساس از گفتار با استفاده از فهرست شده است.

[۱] Principal Component Analysis
[۲] Linear Discriminate Analysis
[۳] Forward Feature Selection
[۴] Backward Feature Selection
[۵] Sequential Floating Forward Selection

جدول ۲-۴ خلاصه‌ای از تحقیقات انجام‌شده تشخیص احساس از گفتار با بررسی روش‌های انتخاب ویژگی

تحقیق	ویژگی‌ها	هدف و رویکرد تحقیق
[۶۷]	استخراج ویژگیهای جدیدی تحت عنوان "ویژگی های هارمونی جدید" براساس ادراک هارمونیک سایکواکوستیک از نظریه موسیقی- ویژگی‌های طیفی - ویژگی‌های عروضی- ویژگی‌های کیفی	• پایگاه‌داده گفتار احساسی برلین • انتخاب ویژگی‌های منتخب با روش SFFS • دقت ۷۳,۵٪ با طبقه بند GMM
[۶۸]	Intensity Spectrogram Pitch, MFCC, HNR Long-Term Average Spectrum	• پایگاه‌داده گفتار احساسی لهستانی • الگوریتم های انتخاب ویژگی های مبتنی بر همبستگی • بهبود دقت طبقه بند شبکه عصبی بیزین از ۷۴٪ به ۹۶٪- طبقه بند SVM از ۶۴٪ تا ۹۶٪- شبکه عصبی RBF از ۶۷٪ به ۸۸٪
[۶۹]	Energy, MFCC, Zero-crossing rate of time signal, Voicing Probability, F0	• پایگاه‌داده گفتار احساسی برلین • انتخاب ویژگی با روش سریع مبتنی بر همبستگی (FCBF) و الگوریتم نرخ فیشر • تشخیص احساس از گفتار با استفاده از تکنیک رای اکثریت بر روی تکنیک های یادگیری شبکه عصبی، درخت تصمیم گیری، SVM و KNN
[۷۰]	۳ مجموعه ویژگی از جمله MFCC ، پارامترهای فوریه و ویژگی های استخراج شده با جعبه ابزار باز گفتار و موسیقی مونیخ توسط ابزار استخراج فضای بزرگ (openSMILE)	• دادگان گفتار احساسی برلین و گفتار احساسات سالمندان چینی • استفاده از روش جستجوی هارمونی[۱] برای انتخاب و کاهش ویژگی ها • استفاده از اعتبار سنجی متقابل ۱۰ برابر در LIBSVM برای ارزیابی تغییر دقت بین زیر مجموعه های انتخاب شده و مجموعه های اصلی

۲-۵ انواع روش‌های طبقه‌بندی در تشخیص احساس از گفتار

در سیستم تشخیص احساسات گوینده بعد از محاسبه ویژگی‌ها، بهترین ویژگی‌ها برای طبقه‌بندی انتخاب می‌شوند. انواع مختلفی از طبقه‌بندها برای این سیستم‌ها پیشنهاد شده‌است، از جمله

[۱] Harmony Search

طبقه‌بندی‌کننده‌های متداول و الگوریتم‌های یادگیری عمیق. در حقیقت هیچ روش مورد توافقی برای طبقه‌بندی احساسات وجود ندارد و مطالعات فعلی بیشتر تجربی است. بنابراین می‌توان گفت در SER هر طبقه‌بند مزایا و محدودیت‌های خاص خودش را دارد، به‌طوری‌که بعضی از تحقیقات با ترکیب طبقه‌بندها سعی نموده‌اند این محدودیت‌ها را حل نمایند. عملکرد طبقه‌بندهای گروهی اغلب بالاتر از طبقه‌بندهای منفرد است. انواع مختلف معماری در طبقه‌بندهای گروهی موجود است. یکی از روش‌ها تغذیه داده‌های مشابه برای هر طبقه‌بندی و مقایسه و رتبه‌بندی نتایج به‌دست‌آمده برای تصمیم نهایی است. روش دیگر استفاده از طبقه‌بندی سلسله مراتبی است. در این روش، داده‌های ورودی به یک الگوریتم تغذیه می‌شوند، سپس نتیجه در یک روش سلسله مراتبی به نوع دیگری از طبقه‌بند داده می‌شود، سپس تصمیم نهایی گرفته می‌شود [۷].

۲-۵-۱ روش‌های متداول طبقه‌بندی

سیستم‌های SER معمولاً از الگوریتم‌های طبقه‌بندی متداول و کلاسیک استفاده می‌کنند. این طبقه‌بندها به ورودی X، خروجی Y و یک تابع نیاز دارد تا X را به Y نگاشت کند. الگوریتم یادگیری این تابع نگاشت را تقریب می‌زند، تا به پیش‌بینی کلاس ورودی جدید کمک کند. الگوریتم یادگیری به داده‌های دارای برچسب نیاز دارد که نمونه‌ها و کلاس آن‌ها را مشخص کند. پس از اتمام آموزش، از داده‌هایی که در طول آموزش استفاده نشده است، برای تست و آزمایش عملکرد طبقه‌بندی استفاده می‌شود. از جمله الگوریتم‌های این دسته می‌توان به مدل مخلوط گوسی[۱]، k-نزدیکترین همسایه[۲]، مدل مخفی مارکوف[۳]، ماشین بردار پشتیبان[۴]، شبکه های عصبی مصنوعی و درخت تصمیم‌گیری اشاره کرد. علاوه‌بر استفاده از طبقه‌بندهای منفرد، از روش‌های گروهی و ترکیبی نیز برای SER استفاده می‌شود که چندین طبقه‌بند

[۱] Gaussian Mixtures Model (GMM)
[۲] K-nearest neighbors (KNN)
[۳] Hidden Markov Model (HMM)
[۴] Support Vector Machine (SVM)

را با هم ترکیب می‌کنند تا نتایج بهتری به‌دست آورند. به‌عنوان مثال Nwe و همکاران نشان دادند که LFPC با استفاده از طبقه‌بند HMM عملکرد بهتری نسبت به MFCC و LPCC دارد. میانگین و بهترین میزان تشخیص در این تحقیق به ترتیب ۷۷,۱٪ و ۸۹٪ درصد گزارش شده است [۷۱]. لین و همکاران از HMM و SVM برای طبقه‌بندی پنج احساس عصبانیت، شادی، غم، تعجب و حالت خنثی استفاده کرده‌اند. در این تحقیق ۳۹ ویژگی برای طبقه‌بند HMM استخراج شده، و سپس الگوریتم SFS برای انتخاب ویژگی اعمال می‌شود. همچنین از تفاضل بین انرژی‌های زیرباندهای مقیاس فرکانس مل، یک بردار جدید ساخته شده و عملکرد K-نزدیکترین همسایه با استفاده از این بردار آزمایش شده است. نتایج تحقیق در حالت وابسته به گوینده نشان می‌دهد که با HMM دقت ۹۹,۵٪ و با طبقه‌بند SVM دقت ۸۸,۹ است [۷۲]. از جمله تحقیقاتی که با استفاده از طبقه‌بندهای ترکیبی انجام شده است می‌توان به لی و همکارانش [۳۵] اشاره کرد که روش پیشنهادی‌شان سیگنال گفتار ورودی را از طریق لایه‌های پی‌درپی طبقه‌بندهای باینری طبقه‌بندی می‌کند. آن‌ها از رگرسیون لجستیک بیزین به‌عنوان طبقه‌بند باینری استفاده کرده و نشان داده‌اند که نتایج نسبت به طبقه‌بند SVM به‌اندازه 7.44٪ بهبود یافته است.

۲-۵-۲ روش‌های طبقه‌بندی مبتنی‌بر یادگیری عمیق

از آن‌جا که بیشتر الگوریتم‌های یادگیری عمیق مبتنی‌بر شبکه‌های عصبی مصنوعی هستند معمولاً به شبکه‌های عصبی عمیق اشاره می‌کنند. اصطلاح "عمیق" از تعداد لایه‌های پنهان ناشی می‌شود که می‌تواند به صدها لایه برسد، درحالی‌که یک شبکه عصبی سنتی شامل دو یا سه لایه پنهان است. در سال‌های اخیر، عملکرد الگوریتم‌های یادگیری عمیق از الگوریتم‌های یادگیری ماشین سنتی پیشی‌گرفته است. مزیت برخی از این الگوریتم‌ها عدم نیاز به مراحل استخراج ویژگی و انتخاب ویژگی است. همه ویژگی‌ها به‌طور خودکار با الگوریتم‌های یادگیری عمیق استخراج و انتخاب می‌شوند. الگوریتم‌های

یادگیری عمیقی که به‌طور گسترده‌تر در دامنه SER استفاده می‌شوند، شبکه‌های عصبی کانولوشن[۱] (CNN) و شبکه‌های عصبی بازگشتی[۲] (RNN) هستند.

یکی از تحقیقات انجام‌شده در راستای مشکل طبقه‌بندی تشخیص احساس از گفتار مربوط به Fayek و همکارانش [۷۳] است که یک فرمول فرایند مبتنی‌بر فریم برای تشخیص احساس از گفتار پیشنهاد داده‌اند. این تحقیق با استفاده از طیف‌نگاره تبدیل فوریه مبتنی‌بر بانک فیلتر گفتار و یک شبکه عصبی چندلایه عمیق، احتمال وقوع هر کلاس احساسی را برای هر فریم یک گفتار ورودی پیش‌بینی کرده است. نتایج تشخیص درست در این تحقیق روی دادگان IEMOCAP با روش پیشنهادی ۶۴,۷۸% به‌دست آمده است. در مطالعه‌ای دیگر [۷۴] یک شبکه عصبی بازگشتی کانولوشنال مبتنی‌بر توجه[۳] سه بعدی برای یادگیری ویژگی‌های مجزاساز SER ارائه‌شده که از طیف‌نگار مل همراه با مشتقات مرتبه اول و دوم به عنوان ورودی استفاده می‌کند. آزمایش‌های انجام‌شده در این تحقیق با مجموعه IEMOCAP و Emo-DB به‌ترتیب ۶۴,۷۴٪ و ۸۲,۸۲٪ دقت را نشان می‌دهند. جدول ۲-۵ خلاصه‌ای از طبقه‌بندهای استفاده‌شده در تحقیقات مربوط به SER فهرست شده است.

جدول ۲-۵ خلاصه‌ای از طبقه‌بندهای استفاده‌شده در تحقیقات SER

Classifiers	Features	References
Gaussian mixture models (GMM)	Prosodic	[67] [75]
	Spectral	[75] [76]
Support vector machines (SVM)	Prosodic	[77][78][68][65][69][79][2]
	Spectral	[77][78] [79][2][66][80]
Artificial neural networks (ANN)	Prosodic	[68][65][69] [2][81]
	Spectral	[78] [66]
k-Nearest neighbor (KNN)	Prosodic	[69]
	Spectral	[82]
Bayes classifier	Prosodic	[67] [78][68][65]
	Spectral	[78][۸۳]
Linear Discriminant Analysis with Gaussian probability distribution	Prosodic	[65]
	Spectral	[30]
Hidden Markov models (HMM)	Prosodic	[75] [76] [77][5]
	Spectral	[75] [76] [77] [5]

[۱] Convolutional Neural Networks (CNN)

[۲] Recurrent Neural Networks (RNN)

[۳] Attention-based Convolutional Recurrent Neural Network

۲-۵-۳ جمع‌بندی

در این فصل در کنار مروری‌بر ادبیات موضوع و مبانی نظری مربوط به حوزه تشخیص احساس از گفتار به بررسی انواع بخش‌ها در این سیستم و مطالعات انجام‌شده در هر بخش پرداختیم. نتایج نهایی این مطالعات در جدول ۲-۶ به‌صورت خلاصه‌ای از تحقیقات جدید این حوزه فهرست شده است. مطالعات نشان می‌دهند که اگرچه پیشرفت‌های زیادی در سیستم‌های تشخیص احساس از گفتار رخ‌داده است، اما هنوز موانع و چالش‌های متعددی وجود دارد که باید برای تشخیص موفقیت‌آمیز حل شوند. یکی از مهم‌ترین مشکلات جمع‌آوری و تولید مجموعه دادگان است که برای فرایند یادگیری بسیار تاثیرگذار و با اهمیت هستند. اکثر مجموعه‌های داده‌ای که برای SER استفاده می‌شوند یا مبتنی‌بر بازیگران حرفه‌ای و یا القا شده هستند که در شرایط آزمایشگاهی ویژه و به‌دور از نویز و سروصدا ضبط می‌شوند. درحالی‌که، داده‌های زندگی واقعی پر سر و صدا و نویزی هستند و ویژگی‌های آن‌ها بسیار متفاوت است. به‌همین علت یکی از مسائل می‌تواند یافتن مجموعه ویژگی‌هایی باشد که در محیط‌های نویزی نیز تشخیص موفقیت‌امیزی دارند. در این راستا در فصل سوم این رساله یک سیستم تشخیص احساس از گفتار مبتنی‌بر یک روش استخراج ویژگی خاص پیشنهاد شده تا بتوان این مشکلات را حل نمود.

جدول ۲-۶ خلاصه‌ای از تحقیقات جدید در حوزه SER

مرجع	روش پیشنهادی	ویژگی‌های مورد استفاده	نوع طبقه‌بند	دادگان گفتار احساسی مورداستفاده	عملکرد
(Koolagudi2012) [۵۵]	تشخیص و طبقه‌بندی احساسات گفتاری با استفاده از مجموعه ویژگی‌های منبع، دستگاه صوتی و ویژگی‌های عروضی به‌صورت جداگانه و به‌صورت ترکیبی	مجموعه ویژگی‌های مستخرج از: • دستگاه صوتی • اطلاعات عروضی • منبع تحریک	• Autoassociative Neural Network • Gaussian mixture model • Support Vector Machine	• پایگاه داده شبیه سازی تلوگو(IITKGP-SESC) • پایگاه داده گفتار احساسی برلین (Emo-DB)	افزایش نرخ دقت تشخیص تا ٪۷۹,۱۴
(Degaonkar 2013) [۸۴]	تشخیص احساس بر پایه متدی مبتنی‌بر سه روش: روش اول: استفاده از تبدیل موجک (WPT) و مقایسه تعداد ضرایب بزرگتر از آستانه در باندهای مختلف. روش دوم : استفاده از نسبت انرژی باندهای مختلف با استفاده از WPT و مقایسه نسبت انرژی در باندهای مختلف. روش سوم: استفاده از MFCC.	• ضرایب بزرگتر از آستانه در باندهای مختلف WPT • انرژی باندهای مختلف با استفاده از WPT • ضرایب MFCC	• random forest • support vector machine • Bayes • multi-layer perceptron • ensemble classifiers such as AdaBoost,	ایجاد پایگاه داده جدید به زبان Marathi	• نتایج به دست آمده با استفاده از WPT برای حالت عصبانیت، خوشحالی و خنثی به‌ترتیب ٪۸۵ ، ٪۶۵ و ٪۸۰ در مقایسه با نتایج بدست آمده با استفاده از MFCC یعنی به‌ترتیب ٪۷۵ ، ٪۴۵ و ٪۶۰ برای سه احساس است. • براساس ویژگی‌های WPT مدلی برای تبدیل احساسات یعنی خنثی به خشمگین و خنثی به احساس شاد ارائه شده است.
(Ooi2014) [۸۵]	روشی مبتنی‌بر ۲ مسیر اصلی موازی: مسیر ۱: براساس تجزیه و تحلیل فشرده از ویژگی-های مختلف عروضی مسیر ۲: براساس تجزیه و تحلیل ویژگی‌های طیفی: استخراج ویژگی MFCC توسط روش‌های تجزیه و تحلیل مولفه اصلی دو جهته BDPCA و تجزیه و تحلیل خطی مجزاساز LDA و طبقه‌بندی شبکه عصبی RBF- دارای ۳ ساختار زیرمسیر BDPCA + LDA + RBF موازی برای کنترل هر دو احساس در یک زیرمسیر	ویژگی‌های عروضی و طیفی	RBF Neural Network	• پایگاه داده eNTERFACE’05 • پایگاه داده RML	• دقت ٪۷۵,۸۹ در پایگاه داده eNTERFACE’05 • دقت ٪۶۸,۵۷ در پایگاه داده RML
(Deng2014) [۴۶]	یک روش انطباق دامنه بدون نظارت جدید مبتنی‌بر خودرمزگذارهای denoising سازگار برای تجزیه و تحلیل سیگنال گفتاری احساسی با رویکرد کاهش اختلاف بین آموزش و مجموعه آزمون به دلیل شرایط مختلف (مجموعه دادگان مختلف)	۱۲ تابع توصیف‌کننده سطح پایین (LLD) و ضرایب رگرسیون مشتق مرتبه اول آن‌ها: ZCR, RMS Energy, F0, HNR, MFCC	• Denoising Autoencoders • SVM	• پایگاه داده SUSAS • پایگاه داده ABS	• ٪۶۴,۱۸ میانگین میزان تشخیص در ABC • ٪۶۲,۷۴ میانگین دقت در SUSAS

مرجع	روش پیشنهادی	ویژگی‌های مورد استفاده	نوع طبقه‌بند	دادگان گفتار احساسی مورداستفاده	عملکرد
(Mao2014) [۵۷]	استفاده از شبکه عصبی کانولوشن برای یک مدل شناسایی احساسات دو مرحله‌ای: مرحله اول: یادگیری ویژگی‌های ثابت محلی با استفاده از رمزگذار خودکار پراکنده برای طیف-سنجی گفتار مرحله دوم: استخراج ویژگی‌های متمایز با استفاده از PCA برای بهبود تشخیص	• نمایش طیف‌سنجی (ویژگی‌های RAW) • ویژگی‌های TEO • ویژگی‌های آکوستیک: Pitch, Energy-related features, Speech rate, MFCC • ویژگی‌های محلی ثابت LIF	• Convolutional Neural Network • SVM	• پایگاه داده آلمانی EMO-DB • پایگاه داده انگلیسی SAVEE • پایگاه داده گفتار احساسی دانمارکی DES • پایگاه داده ماندرین MES	• میزان شناخت به طور متوسط ۷۹٪ براساس نتایج به دست آمده از چهار مجموعه داده
(Esmaileyan2014) [۸۶]	ارائه ویژگی‌های جدید به دست آمده از طیف-سنجی گفتار با استفاده از تکنیک‌های پردازش تصویر برای تشخیص احساسات (با محاسبه نمودارهای واریوگرام از طیف‌سنجی گفتار).	• ویژگی‌های DCT • ویژگی‌های عروضی • ویژگی‌های طیفی	SVM+FDR	• پایگاه داده آلمانی EMO-DB • پایگاه داده فارسی PDREC	• در EMO-DB: بهبود میزان تشخیص به‌ترتیب از ۸۳,۱۸٪ و ۸۹,۳۶٪ به ۸۶,۸۲٪ و ۹۰,۴۳٪ برای زنان و مردان • در PDREC: بهترین دقت طبقه‌بندی به‌ترتیب برای زنان و مردان ۶۳,۱۸٪ و ۵۷,۳۷٪
(Kuchibhotla2016) [۸۷]	یک روش انتخاب ویژگی دو مرحله‌ای: مرحله اول: انتخاب و ادغام ویژگی‌های مناسب برای تشخیص احساس مرحله دوم: استفاده از تکنیک‌های انتخاب زیرمجموعه ویژگی بهینه (انتخاب متوالی رو به جلو SFS، و انتخاب رو به جلو شناور متوالی SFFS)	• ویژگی‌های عروضی • ویژگی‌های طیفی	• SVM • KNN • LDA • RDA	• پایگاه داده آلمانی EMO-DB • پایگاه داده اسپانیایی SES	• بهبود ۱۵ الی ۲۰ درصدی کارایی طبقه‌بندی با روش انتخاب ویژگی دو مرحله‌ای • ۹۴,۴٪ میانگین دقت در EMO-DB • ۹۰,۱٪ میانگین دقت در SES
(Wang2015) [۴۱]	ارائه یک مدل پارامتر جدید فوریه با استفاده از محتوای ادراکی کیفیت صدا و تفاضل‌های مرتبه اول و دوم برای تشخیص احساس گفتار مستقل از گوینده	• ضرایب MFCC • ویژگی‌های پیشنهادی پارامترفوریه	SVM+PCA	• پایگاه داده آلمانی EMO-DB • پایگاه داده احساسات سالمندان چینی EESDB • پایگاه داده زبان چینی CASIA	• بهبود نرخ های شناسایی به‌ترتیب به ۱۷,۵ ، ۱۰ و ۱۰,۵ درصد در EMO-DB، CASIA و EESDB
(Vignolo2016) [۶۶]	• ارائه یک روش بهینه‌سازی تکاملی برای جستجوی یک بانک فیلتر برای استخراج ضرایب کپسترال اسپلاین تکاملی (ESCC)	• ضرایب کپسترال اسپلاین تکاملی (ESCC) • ضرایب MFCC	SVM	• پایگاه داده آلمانی FAU Aibo • پایگاه داده هندی	• ۴۲,۵٪ میانگین دقت در FAU Aibo • ۹۱,۳٪ میانگین دقت در پایگاه داده هندی
(Fayek2017) [۷۳]	ارائه یک فرمول فرایند مبتنی‌بر فریم برای تشخیص احساس از گفتار بااستفاده از طیف‌نگاره‌تبدیل فوریه مبتنی‌بر بانک فیلتر گفتار و یک شبکه عصبی چندلایه عمیق	طیف‌نگاره‌تبدیل فوریه مبتنی‌بر بانک فیلتر گفتار	• DNN • LSTM-RNN	پایگاه داده IEMOCAP	• پیش‌بینی احتمال وقوع هر کلاس احساسی برای هر فریم یک گفتار ورودی • دقت تشخیص ۶۴,۷۸٪ روی دادگان IEMOCAP
(Yogesh2017) [۸۸]	ارائه روشی برای ایجاد یک سیستم تشخیص احساس/ استرس از گفتار مستقل از موضوع، با استفاده از شناسایی احساس گوینده از صدای آن‌ها، ویژگی‌های جعبه ابزار OpenSmile، ویژگی‌های طیفی مرتبه بالاتر و یک الگوریتم مبتنی بر زیست	• PCM loudness • MFCC • LOG MEL FREQ. BAND 0–7 • LINE SPECTRAL PAIRS FREQ. 0–7 • F0 and F0 ENVELOPE	ELM classifier	• پایگاه داده آلمانی EMO-DB • پایگاه داده انگلیسی SAVEE • پایگاه داده گفتار تحت استرس واقعی و شبیه-سازی‌شده SUSAS	نرخ تشخیص در محدوده • ۹۰,۳۱٪ -۹۹,۴۷٪ در EMO-DB • ۶۲,۵۰٪ -۷۸,۴۴٪ در SAVEE • ۸۵,۸۳٪ - ۹۸,۷۰٪ در SUSAS

مرجع	روش پیشنهادی	ویژگی‌های مورد استفاده	نوع طبقه‌بند	دادگان گفتار احساسی مورداستفاده	عملکرد
	جغرافیایی به‌کمک بهینه‌سازی ازدحام ذرات (PSO) برای انتخاب ویژگی‌ها	• VOICING PROBABILITY • JITTER Local and CONSEC FRAME PAIRS SHIMMER LOCAL			
(Liu2018) [۵۹]	ارائه یک روش تشخیص احساسات گفتاری مبتنی بر مدل یادگیری عاطفی مغز BELبهبود یافته و الهام گرفته شده از مکانیسم پردازش عاطفی سیستم لیمبیک در مغز – استفاده از الگوریتم ژنتیک برای به‌روزرسانی وزن مدل BEL- آزمایش در دو حالت وابسته به گوینده (SD) و مستقل از گوینده (SI)	MFCC و مشتقات مرتبه اول آن LogMelFreqBand	• GA-BEL • SVM • ELM • LDA • PCA • PCA+LDA	• پایگاه داده آلمانی FAU Aibo • پایگاه داده گفتار تحت استرس واقعی و شبیه-سازی‌شده SUSAS • پایگاه داده زبان چینی CASIA	میانگین دقت تشخیص احساس از گفتار • ٪۹۰,۲۸ (SD)، ٪۳۸,۵۵ (SI) در CASIA • ٪۷۶,۴۰ (SD)، ٪۴۴,۱۸ (SI) در SAVEE • ٪۷۱,۰۵ (SD)، ٪۶۴,۶ (SI) در FAU Aibo
(Ozseven2019) [۶۰]	ارائه یک روش انتخاب ویژگی براساس تغییرات مجموعه ویژگی‌ها با توجه به تفاوت‌های احساسات: • میانگین هر ویژگی در مجموعه ویژگی ها بر اساس احساسات • محاسبه تبادل عاطفی ویژگی‌ها (میانگین تفاوت بین ترکیبات دوگانه احساسات) • تنظیم مقدار آستانه برای انتخاب ویژگی‌ها	• ویژگی‌های آکوستیک جعبه ابزار OpenSmile	• SVM • KNN • MLP • PCA • FCBF • SFS	مجموعه دادگان: • SAVEE • EMO-DB • eNTERFACE05 • EMOVO	• کاهش اندازه ویژگی‌ها با استفاده از روش پیشنهادی از ۱۵۸۲ عدد بین ۷۵ تا ۹۸,۶ درصد. • بسته به مقدار آستانه انتخاب شده، کاهش اندازه مجموعه ویژگی‌ها، کاهش حجم کار، و افزایش موفقیت طبقه‌بندی در همه نتایج تجربی
(Shahin2019) [۸۹]	ارائه یک سیستم تشخیص احساسات مستقل از متن و مستقل از گوینده بر اساس یک طبقه‌بند جدید ترکیبی از یک مدل مخلوطی گوسی آبشاری و شبکه عصبی عمیق(GMM-DNN)	• ضرایبMFCC	• طبقه‌بند ترکیبی GMM-DNN	• Emirati speech database (Arabic United Arab Emirates Database)	میانگین دقت تشخیص احساس از گفتار • ٪۸۳,۹۷ دقت تشخیص
(Kerkeni2019) [90]	ارائه یک رویکرد سراسری برای سیستم تشخیص احساس از گفتار با استفاده از تجزیه حالت تجربی (EMD): • تجزیه هر سیگنال با استفاده از EMD به اجزای نوسانی به نام توابع حالت ذاتی(IMF) • استخراج ویژگی‌های جدیدی به نام ویژگی‌های طیفی مدولاسیون (MS) و ویژگی‌های فرکانس مدولاسیون (MFF) براساس مدل مدولاسیون AM-FMو ترکیب آن‌ها با ویژگی‌های کپسترال	• Teager-Kaiser Energy Operator (TKEO) • Modulation Spectral (MS) • Modulation Frequency Features (MFF) • MFCC • SMFCC	• Support Vector Machine (SVM) • Recurrent Neural Networks (RNN)	مجموعه دادگان: • EMO-DB • Spanish corpus	• میزان تشخیص۹۱/۱۶ ٪ در پایگاه داده اسپانیایی با استفاده از طبقه‌بند RNN و ترکیبی از تمام ویژگی های استخراج شده از توابع حالت ذاتی (IMF) • میزان تشخیص ۸۶.۲۲ ٪ در پایگاه داده برلین، با استفاده از طبقه‌بند SVM و ترکیبی از همه ویژگی‌ها
(Issa2020) [91]	• ارائه یک سیستم تشخیص احساس از گفتار با استفاده از یک شبکه عصبی کانولوشنال یک بعدی و ویژگی‌های طیفی و کپسترال. • استفاده از یک روش افزایشی برای اصلاح مدل اولیه برای بهبود دقت طبقه بندی	• ضرایب MFCC • اسپکتروگرام با مقیاس مل • کروموگرام • ویژگی کنتراست طیفی • مدل نمایشTonnet z	• Convolutional Neural Network • SOFTMAX	مجموعه دادگان: • EMO-DB • RAVDESS • IEMOCAP	میانگین دقت تشخیص احساس از گفتار • ٪۷۱.۶۱ برای RAVDESS با ۸ کلاس • ۸۶.۱ ٪ برای EMO-DB با ۵۳۵ نمونه در ۷ کلاس • ٪۹۵.۷۱ برای EMO-DB با ۵۲۰ نمونه در ۷ کلاس • ۶۴.۳ ٪ برای IEMOCAP با ۴ کلاس

مرجع	روش پیشنهادی	ویژگی‌های مورد استفاده	نوع طبقه‌بند	دادگان گفتار احساسی مورداستفاده	عملکرد
(Zheng2020) [92]	تولید ویژگی‌های احساسی سراسری در سیگنال-های گفتاری از زوایای مختلف و استفاده از مدل یادگیری گروهی برای تشخیص احساس در ۳ گام: (۱) طراحی سه نقش تخصصی برای SER : • متخصص ۱ برای استخراج ویژگی های سه بعدی سیگنال های محلی • متخصص ۲ برای استخراج اطلاعات جامع در داده های محلی • متخصص ۳ برای استخراج ویژگی های عمومی (LLDs) (۲) با طراحی یک مدل یادگیری گروهی، هر متخصص می‌تواند به نفع خود بازی کند و احساسات گفتاری را از کانون‌های مختلف ارزیابی کند. (۳) از طریق آزمایشات، عملکرد متخصصان مختلف و مدل‌های یادگیری گروهی در تشخیص احساسات مجموعه دادگان IEMOCAP مقایسه شده و اعتبار مدل پیشنهادی تأیید می‌شود.	• توصیف‌کننده‌های ویژگی آکوستیک سطح پایین(LLDs) • توابع سطح بالای آماری(HSF) • ویژگی‌های محلی و مشتقات زمانی	• Convolutional neural network (CNN) • Bi-directional long short-term memory (BLSTM) • Gated recurrent unit (GRU)	• پایگاه داده IEMOCAP	میانگین دقت تشخیص احساس از گفتار ٪۷۵ در حالت استفاده از طبقه‌بند گروهی
(حریمی۱۳۹۶) [۹۳]	• طبقه‌بندی نمونه‌ها بااستفاده از ویژگی‌های متداول عروضی و طیفی برمبنای سطح برانگیختگی • جداسازی احساسات با سطح برانگیختگی یکسان با استفاده از ویژگی‌های پیشنهادی دینامیکی غیر خطی • استخراج ویژگی‌های دینامیکی غیر خطی از روی مشخصات هندسی فضای فاز بازسازی شده سیگنال گفتار	• ویژگی‌های عروضی • ویژگی‌های طیفی • ویژگی‌های دینامیکی غیر خطی	SVM+FDR	• پایگاه داده آلمانی EMO-DB	• نرخ بازشناسی ٪۹۶/۳۵ و ٪۸۷/۱۸ برای زنان و مردان • متوسط نرخ بازشناسی٪۹۲/۳۴ با توجه به تعداد نمونه‌ها در هر گروه جنسیتی

فصل ۳ روش پیشنهادی

۳-۱ مقدمه

همان‌طور که در فصل ۱ بیان شد، یکی از چالش‌ها در این مساله انتخاب یک مدل کاراتر نمایش احساسات است. سیستم‌های تشخیص احساس از گفتار در دو مدل مبتنی‌بر احساسات گسسته و مبتنی‌بر مدل چندبعدی احساس پیاده‌سازی شده‌اند. در هر دو نوع مدل، استخراج و انتخاب ویژگی‌ها بخش‌های کلیدی و موثر در تشخیص هستند. در این راستا در این رساله دو فاز برای طراحی سیستم تشخیص احساس از گفتار پیشنهادی پیاده‌سازی شده است، به‌طوری‌که در فاز اول با استفاده از یک روش استخراج ویژگی پیشنهادی در فضای زمان-فرکانس ویژگی‌های موثر در طبقه‌بندی احساسات گسسته استخراج شده و سپس در فاز دوم یک نگاشت بین این ویژگی‌های استخراج شده و مدل سه‌بعدی احساس ارائه می‌شود.

هدف از رویکرد پیشنهادی دستیابی به یک بازنمایی از فضای ویژگی است که جوهره ابعاد جاذبه، برانگیختگی و تسلط را بهتر به تصویر می‌کشد. با توجه به اینکه دو بخش کلیدی و چالش‌برانگیز در سیستم‌های تشخیص احساس از گفتار شامل بخش‌های استخراج و انتخاب مجموعه ویژگی‌های موثر جهت طبقه‌بندی کلاس‌های احساسی است، در این راستا ابتدا در فاز اول این رساله یک سیستم تشخیص احساس از گفتار بهبودیافته مبتنی‌بر یک روش استخراج ویژگی نوین و یک الگوریتم ترکیبی انتخاب ویژگی پیشنهاد و آزمایش شده است. در بخش دوم این فصل نمای کلی این سیستم پیشنهادی ارائه شده است. سپس در فاز دوم با استفاده از یک الگوریتم یادگیری یک نگاشت بین ویژگی‌های مستخرج از فاز اول و سه بعد احساس (جاذبه، برانگیختگی تسلط) تخمین زده می‌شود. بخش سوم این فصل مراحل این نگاشت را نشان می‌دهد.

۳-۲ فاز اول: سیستم تشخیص احساس از گفتار پیشنهادی مبتنی‌بر مدل گسسته احساسی

در این رساله یک سیستم بهبودیافته‌ی تشخیص احساس از گفتار مطابق با بلوک‌دیاگرام **شکل ۳-۱** پیشنهاد شده است. این سیستم پیشنهادی مطابق با نمای کلی **شکل ۳-۱** شامل مراحل اصلی زیر است که هر بخش به‌تفضیل در ادامه شرح داده خواهد شد:

۱. پیش‌پردازش

۲. استخراج ویژگی پیشنهادی

۳. انتخاب ویژگی پیشنهادی

۴. طبقه‌بندی و ارزیابی

۳-۲-۱ پیش‌پردازش

پیش‌پردازش شامل بخش‌های پیش‌تاکید، پنجره‌گذاری و فریم‌بندی است، مطابق با مراحل زیر:

- **پیش‌تاکید**: در این مرحله برای مسطح شدن طیفی سیگنال گفتار و به‌دست آوردن یک دامنه قابل مقایسه برای تمام فرمنت‌ها، سیگنال از میان یک فیلتر بالاگذر (FIR) عبور داده می‌شود تا دامنه باندهای فرکانس بالا را افزایش دهد. همانطور که در معادله ۳-۱ ارائه شده است x سیگنال گفتار ورودی و α پارامتر پیش‌تاکید است که مقدار آن معمولا بین ۰,۹ و ۱ است (در حدود ۰,۹۷) [۸۰].

معادله ۳-۱

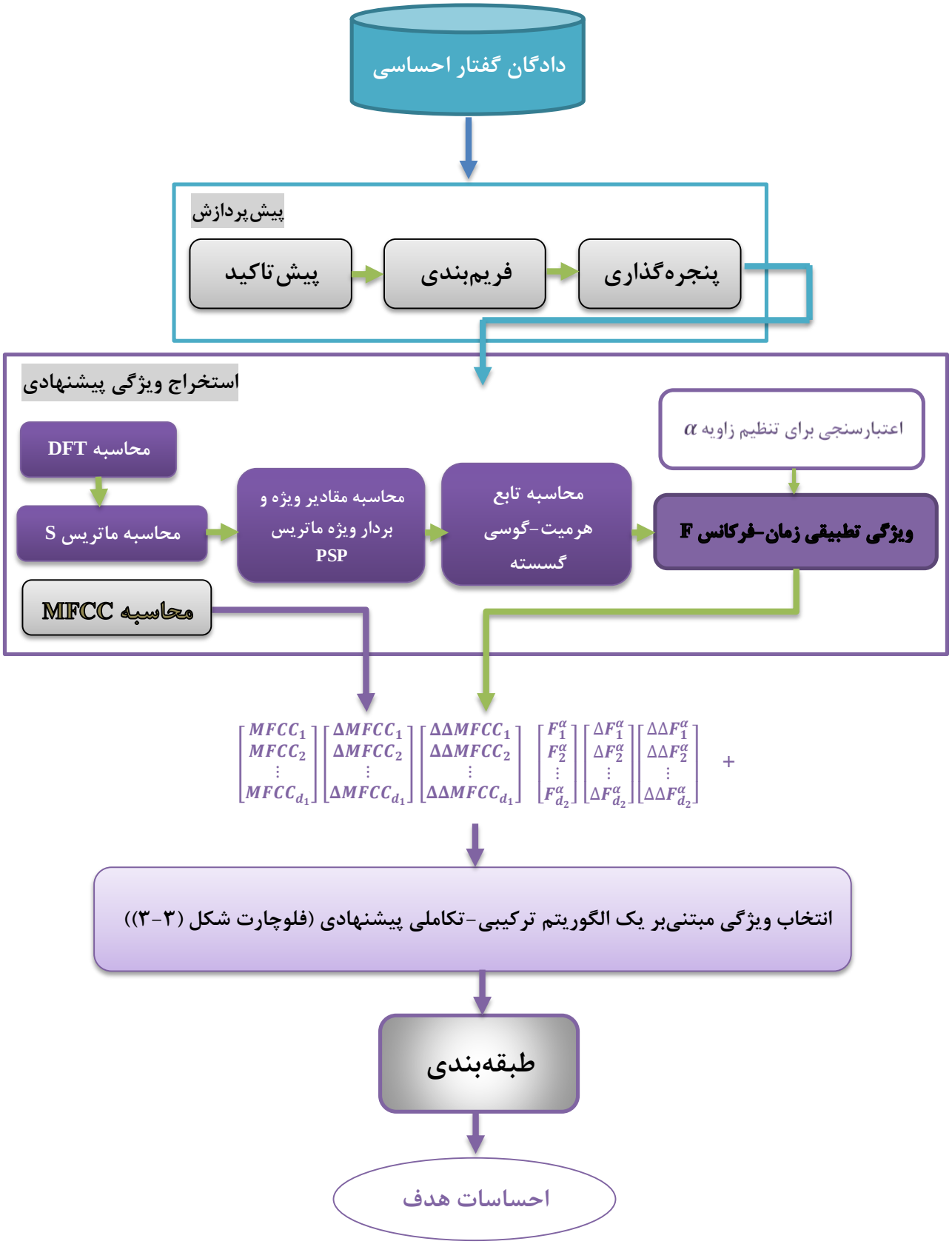
$$x'(n) = x(n) - \alpha x(n-1)$$

❖ **فریم‌بندی:** در این مرحله نمونه‌های گفتار به فریم‌های کوچکی تقسیم می‌شوند. سیگنال ورودی به فریم‌هایی به طول L میلی‌ثانیه با نرخ هم‌پوشانی[۱] $\frac{L}{2}$، تقسیم می‌شود. N نمونه از سیگنال را می‌توان به روش زیر به‌دست آورد:

معادله ۳-۲

$$N = fs * t\_frame$$

[۱] overlap



شکل ۳-۱ سیستم پیشنهادی تشخیص احساس از گفتار مبتنی‌بر مدل گسسته احساسی

❖ **پنجره‌گذاری:** هر فریم گفتاری توسط پنجره استاندارد همینگ[1] (رابطه (۳-۳)) پنجره‌گذاری می‌شود (پنجره در هر فریم ضرب میشود) تا ناپیوستگی‌های سیگنال را در ابتدا و انتهای هر فریم به حداقل برساند.

$$w(n) = 0.54 - 0.64\,cos\left(\frac{2\pi n}{M-1}\right) \qquad 0 \le n \le M-1$$ معادله ۳-۳

۳-۲-۲ روش استخراج ویژگی پیشنهادی

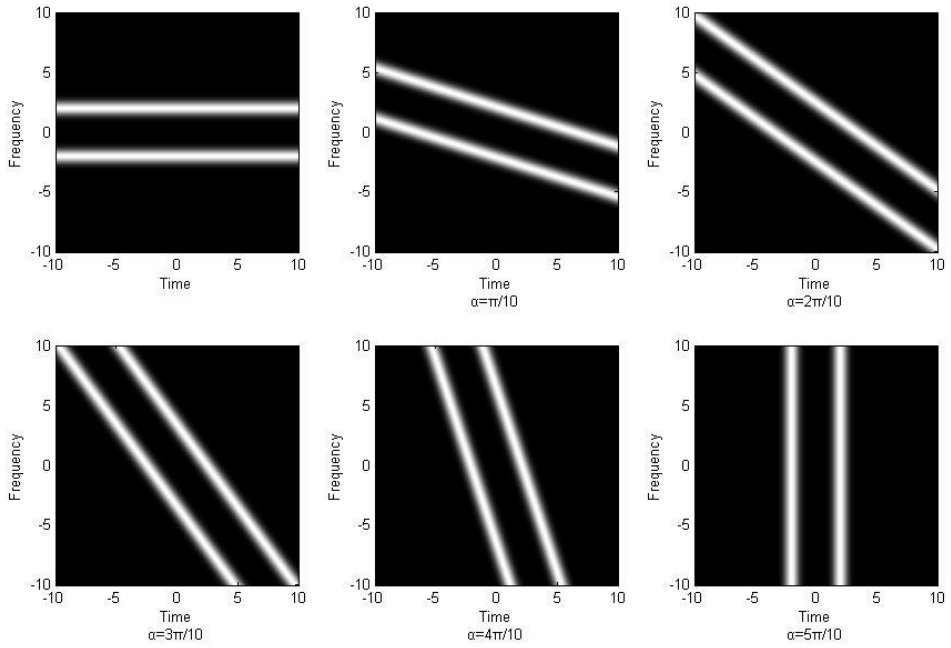
در فصل قبل مشاهده کردیم که روش‌های استخراج ویژگی در SER یا در حوزه زمان هستند یا با استفاده از تبدیل فوریه (که در روش‌های استخراج ویژگی وجود دارد) به خصیصه‌هایی در حوزه فرکانس تبدیل می‌شوند. سیگنال گفتار هم در فرکانس و هم در زمان مدل‌سازی می‌شود و تجزیه و تحلیل فوریه منجر به تولید مؤلفه‌های فرکانسی سیگنال می شود. همان‌طورکه در تحقیقات گزارش شده است اگرچه که تبدیل فوریه برای تجزیه و تحلیل و پردازش سیگنال‌های زمان-ثابت مناسب است، اما نمی‌تواند برای سیگنال‌های زمان-متغیر به عملکرد مطلوبی دست یابد [۹۴]. از آن‌جا که سیگنال گفتار غیرایستا[2] و متغیر در طول زمان است می‌توان از تبدیل فوریه کسری گسسته[3] (DFrFT) استفاده کرد که شامل هر دو نوع مؤلفه‌های زمان و فرکانس سیگنال است. DFrFT خانواده‌ای از تبدیل‌های خطی است که از تبدیل فوریه مشتق می‌شوند و در شاخه تحلیل هارمونیک در ریاضیات مورد استفاده قرار می‌گیرند. تبدیل فوریه کسری را می‌توان بصورت توان nام تبدیل فوریه (که در آن n عدد صحیح است) تعریف کرد؛ بنابراین این تبدیل می‌تواند یک تابع را به هر دامنه‌ای در بین زمان و فرکانس نگاشت کند. همانند تبدیل فوریه، DFrFT نیز دارای ضرایب مختلط است و می‌تواند برای به‌دست آوردن هر دو اطلاعات فاز و اندازه یک سیگنال استفاده شود. همچنین کرنل تبدیل DFrFT، مجموعه‌ای از chirpهای خطی است در

[1] Hamming Window
[2] Non-Stationary
[3] Discrete Fractional Fourier Transform

حالی که کرنل تبدیل فوریه از توابع سینوسی مختلط است. بسته به نوع درجه کسری، سیگنال‌ها را می‌توان در دامنه‌های مختلف نشان داد. این ویژگی به DFrFT درجه آزادی بیشتری در تجزیه و تحلیل سیگنال نسبت به تبدیل فوریه استاندارد می‌دهد. پردازش سیگنال‌های زمان-متغیر در دامنه فوریه کسری به ما این امکان را می‌دهد که سیگنال را با حداقل خطای میانگین مربعات (MSE) تخمین بزنیم [۹۵]. مدل‌سازی سیگنال‌های گفتاری به‌عنوان سیگنال‌های زمان-فرکانس ترکیبی، از دو دیدگاه تولید و ادراک با ویژگی‌های گفتار بهتر تطابق دارد. علاوه‌بر این، کلاس‌های آوایی خاص نمایش و بازنمایی بهتری در حوزه‌های کسری مختلف دارند.

تبدیل فوریه کسری از یک صفحه با دو محور متعامد متناظر با زمان و فرکانس استفاده می‌کند. اگر یک سیگنال $x(t)$ در طول محور زمان نشان داده شود و تبدیل فوریه آن $X(\omega)$ در امتداد محور فرکانس نشان داده شود، آنگاه عملگر تبدیل فوریه F می‌تواند به‌عنوان تغییری در نمایش سیگنال متناظر با یک چرخش محور در خلاف جهت عقربه‌های ساعت توسط یک زاویه $\frac{\pi}{2}$ ترسیم شود. این نکته با بعضی از خصوصیات تبدیل فوریه سازگار است، به‌عنوان مثال دو چرخش پی‌درپی سیگنال با زاویه $\frac{\pi}{2}$ منجر به وارونگی محور زمان خواهد شد. علاوه‌بر این، چهار چرخش پی‌درپی سیگنال را بدون تغییر می‌گذارد زیرا چرخش با زاویه 2π باید سیگنال را تغییر ندهد. یکی از ویژگی‌های بسیار مهم این تبدیل بازیابی سیگنال اصلی از سیگنال نویزی است. بدین‌صورت که این تغییر شکل با توجه به یک پارامتر زاویه آلفا منجر به چرخش سیگنال در فضای فرکانس-زمان می‌شود. تنظیم مناسب پارامتر آلفا در حذف اطلاعات بی‌ربط (نویز) موثر است. شکل (۳-۲) مثالی از نحوه استفاده از پارامتر آلفا در یک تغییر شکل کسری را نشان می دهد. با تنظیم این پارامتر می‌توان یک چرخش مناسب در فضای تبدیل ایجاد کرد تا قسمت نامطلوب سیگنال به‌راحتی حذف شود. همچنین باید به این نکته اشاره کرد که مفهوم نویز در یادگیری ماشین بسته به کاربرد، می‌تواند تعمیم یابد. به‌طور خلاصه هر بخشی از سیگنال که در کاربرد هدف مورد دلخواه نیست نوع تعمیم‌یافته‌ای از نویز است که باید به طریقی تاثیر آن تضعیف یا فیلتر گردد تا

مدل بهتری آموزش یابد. اگر بخواهیم در موضوع تشخیص احساسات این مسئله را شرح دهیم باید اشاره ای به مجموعه دادگان موجود در این کاربرد کنیم. در اغلب این دادگان گویندگان یک جمله مشخص را با احساسات مختلف بیان می‌کنند (بازی می‌کنند). این موضوع نشان می‌دهد که تاثیر کلمات و کلاس‌های آوایی، بر احساس گوینده موردنظر نیست. در نتیجه می‌توان گفت در کاربرد تشخیص احساس ویژگی‌هایی که مرتبط با گفتار و کلمات است بخش نامرتبط به هدف ما بوده و مفهوم کلی از نویز می‌باشند که باید به طریقی تاثیر آن‌ها را کم کرد. برای حل این موضوع در این رساله از قابلیت‌های تبدیل فوریه کسری استفاده کردیم.



شکل ۳-۲ تاثیر چرخش سیگنال در تبدیل فوریه کسری در حذف نویز

همان‌طور که در فصل‌های گذشته بیان شد، یک فرایند بسیار مهم در SER استخراج اطلاعات اساسی و ضروری سیگنال برای بهبود عملکرد طبقه‌بندی است. در این رساله، با توجه به ویژگی‌ها و خصوصیات مطرح‌شده درباره تبدیل فوریه کسری در بالا، یک روش استخراج ویژگی فرکانس-زمان تطبیقی مبتنی‌بر تبدیل فوریه گسسته کسری (DFrFT) مطابق با بلوک‌دیاگرام پیشنهادی در شکل ۳-۱ پیشنهاد شده است. این ویژگی پیشنهادی طی مراحل مشخص شده در بلوک‌دیاگرام، مطابق با تابع تعریف شده در

معادله ۳-۴ و توسط گام‌های زیر استخراج می‌شود:

- ✓ بسط تبدیل فوریه گسسته (DFT)
- ✓ محاسبه ماتریس S
- ✓ محاسبه بردارهای ویژه و مقادیر ویژه ماتریس *PSP*
- ✓ محاسبه تابع هرمیت گوسی
- ✓ تنظیم پارامتر آلفا با یک مجموعه اعتبارسنجی

معادله ۳-۴

$$F^{a}[m.n] = \sum_{k=0}^{N-1} u_k[m] e^{-j\frac{\pi}{2}ka} u_k[n]$$

در رابطه فوق $u_k[n]$ به kمین تابع هرمیت-گوسی گسسته دلالت دارد، و $\lambda_k = e^{-j\frac{\pi}{2}ka}$ مقدار ویژه توان αم تبدیل فوریه است. همان‌طور که در مطالعه‌ی Candan بیان شده [۹۶]، در حل معادله ۳-۴ دو ابهام وجود دارد که باید حل کرد:

- ❖ ابهام اول در زمان محاسبه توان کسری مقادیر ویژه اتفاق می‌افتد، زیرا عمل محاسبه توان کسری تک-مقداری[1] نیست. این ابهام را با محاسبه مقدار ویژه به‌صورت $\lambda_k^{\alpha} = e^{\frac{-j\pi k\alpha}{2}}$ می‌توان برطرف کرد.
- ❖ ابهام دوم مربوط به ساختار ویژه[2] DFT است. از آن‌جا که ماتریس DFT فقط چهار مقدار ویژه λ_k خاص دارد (یعنی $\lambda_k = exp(-j\pi k / 2) \in \{1. -1. j. -j\}$)، و مقادیر ویژه دیگر به‌طورکلی هدر می‌روند، بنابراین مجموعه بردار ویژه منحصر به‌فرد نخواهد بود. به‌همین دلیل، لازم است یک مجموعه بردار ویژه خاص برای استفاده در معادله ۳-۴ مشخص شود. این ابهام با انتخاب توابع هرمیت-گوسی به‌عنوان توابع ویژه در رابطه حل می‌شود. مطابق با

[1] Single-Valued
[2] Eigen-Structure

تحقیقات [۹۶] اثبات شده است که توابع هرمیت-گوسی که توابع ویژه ماتریس S هستند (معادله ۳-۵)، توابع ویژه ماتریس DFT نیز هستند. همچنین مشخص شده که بردارهای ویژه ماتریس DFT یا S بردارهای زوج یا فرد هستند. بنابراین، ما یک ماتریس P تعریف خواهیم کرد که یک بردار دلخواه را به اجزای زوج یا فرد خود تجزیه می‌کند، به‌عنوان مثال ماتریس ۵-بعدی P در معادله ۳-۶ نشان داده شده است. با در نظر گرفتن تبدیل PSP^{-1}، مشکل یافتن بردارهای ویژه مشترک از S به یافتن بردارهای ویژه ماتریس‌های زوج "Ev" و فرد "Od" کاهش می‌یابد (معادله ۳-۷).

$$S = \frac{-1}{4\pi}\begin{bmatrix} -2 & 1 & 0 & & \\ 1 & 2\cos(\frac{2\pi}{n}) & 1 & \dots & 0 \quad 1 \\ 0 & 1 & 2\cos(\frac{2\pi}{n}2) - 4 & \dots & 0 \quad 0 \\ & & \vdots & & \\ & & \vdots & & \\ 1 & 0 \quad 0 \dots & 1 & 2\cos(\frac{2\pi}{n}(n-1)) - 4 & \end{bmatrix}$$

معادله ۳-۵

$$P = \frac{1}{\sqrt{2}}\begin{bmatrix} \sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

معادله ۳-۶

$$PSP^{-1} = PSP = \begin{pmatrix} Ev & 0 \\ 0 & Od \end{pmatrix}$$

معادله ۳-۷

از آن‌جا که بردارهای ویژه زوج و فرد S به‌ترتیب از بردارهای ویژه ماتریس‌های پوشش صفر Ev و Od استخراج می‌شوند، بنابراین می توان نشان داد که پس از پوشش صفر و تبدیل از طریق P، که چیزی نیست جز یک پسوند زوج و فرد بردار، بردار ویژه Ev با k عبور از صفر، بردار ویژه زوج S با 2k عبور از صفر ($0 \le k \le \lfloor N/2 \rfloor$) را ایجاد می‌کند و بردار ویژه Od با k عبور از صفر، بردار ویژه فرد S با 2k+1 عبور از صفر ($0 \le k \le \lfloor (N-3)/2 \rfloor$) را به دست می‌دهد. بردارهای ویژه زوج و فرد S به‌ترتیب با معادله ۳-۸ و معادله ۳-۹ محاسبه می‌شوند:

$$U_{2k}[n] = P[\hat{e}_k^T|0\cdots 0]^T; \qquad 0 \le k \le \lfloor N/2 \rfloor$$ معادله ۳-۸

$$U_{2k+1}[n] = P[0\cdots 0|\hat{o}_k^T]^T; \qquad 0 \le k \le \lfloor (N-3)/2 \rfloor$$ معادله ۳-۹

در روابط فوق $\hat{e}_k$ بردار ویژه Ev با k عبور از صفر، و $\hat{o}_k$ بردار ویژه Od با k عبور از صفر را نشان می‌دهند.

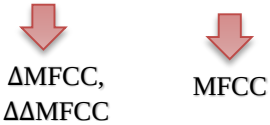
در پایان با توجه به یافتن بردارهای ویژه در معادله ۳-۴، مرحله آخر یافتن بهترین مقدار برای زاویه α در محاسبه مقادیر ویژه $e^{-j\frac{\pi}{2}ka}$ است. از آن‌جا که با تنظیم این پارامتر می‌توان یک چرخش مناسب در فضای تبدیل ایجاد کرد تا قسمت نامطلوب سیگنال به‌راحتی حذف شود، برای تنظیم[۱] این پارامتر آلفا از یک مجموعه اعتبارسنجی[۲] و مطابق با رابطه زیر استفاده کردیم. در فصل آینده چگونگی این اعتبارسنجی و یافتن بهترین مقدار زاویه α را تجزیه و تحلیل خواهیم کرد.

$$Find\ '\alpha'\ that\ \left\{\max_{\alpha} Acc(F_i^{\alpha}[m,n], t_i)\right\}; \qquad 0 < \alpha \le 2\pi \ \ and\ \ \alpha \in \mathcal{R}$$ معادله ۳-۱۰

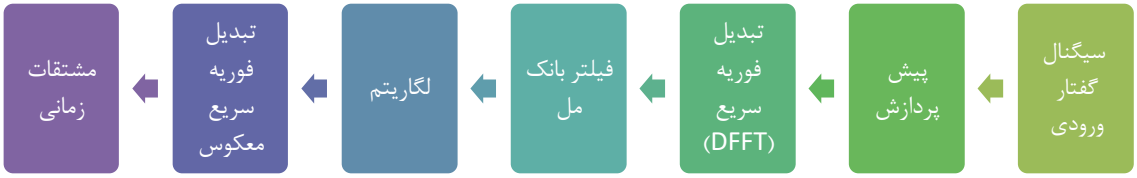
که Acc دقت طبقه‌بند SVM، و $F_i^{\alpha}[m.n]$ و t_i به‌ترتیب بردار ویژگی و برچسب کلاس نمونه iم هستند.

✓ استخراج ضرایب MFCC

برخی از آثار [۹۷] [۹۸] نشان می‌دهند که بدون جایگزینی کل سیستم خطی با رویکردهای غیرخطی جدید، می‌توان با ترکیب ویژگی‌های غیرخطی با MFCC دقت تشخیص بیشتری به‌دست آورد. به‌همین علت در این سیستم علاوه‌بر استخراج ویژگی پیشنهادی، ضرایب MFCC و مشتقات مرتبه اول و دوم آن نیز مطابق با بلوک‌دیاگرام شکل ۳-۳ استخراج می‌شوند.



[۱] Tuning
[۲] Validation Set



شکل ۳-۳ مراحل استخراج ویژگی‌های MFCC

همان‌طور که در شکل ۳-۳ نشان داده شده است، مرحله اول پیش‌پردازش است که در این مرحله نمونه‌های گفتار به فریم‌های کوچکی تقسیم می‌شوند. سیگنال ورودی به فریم‌هایی به طول L میلی‌ثانیه با نرخ هم‌پوشانی $\frac{L}{2}$، تقسیم می‌شود. سپس با استفاده از پنجره همینگ عمل پنجره‌گذاری انجام می‌شود. در مرحله بعدی روی این سیگنال فریم‌بندی و پنجره‌گذاری شده، تبدیل فوریه گسسته سریع اعمال شده و خروجی آن N باند گسسته فرکانس $X(k)$ مطابق معادله ۳-۱۱ خواهد بود.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2(\frac{\pi}{N})kn}$$ معادله ۳-۱۱

مرحله بعدی اعمال فیلتر بانک مل[1] بر روی ضرایب فوریه است. مقیاس مل یک مقیاس اندازه‌گیری گام یا فرکانس شنیده‌شده‌ی یک آوا توسط انسان است. سیگنال گفتار شامل تن‌هایی[2] با فرکانس‌های مختلف f است. مقیاس مل با استفاده از رابطه زیر اندازه‌گیری می‌شود:

$$Mel = 1127.01048 \log_e(1 + (f/700))$$ معادله ۳-۱۲

بعد از اعمال لگاریتم روی سیگنال خروجی از فیلتر بانک مل، و محاسبه معکوس تبدیل فوریه گسسته، دنباله‌ای محدود از نقاط داده یعنی ضرایب MFCC تولید می‌شوند.

ویژگی‌های مستخرج پیشنهادی و ضرایب MFCC باهم ترکیب شده و مشتقات مرتبه اول و دوم

[1] Mel-scaled Filter bank
[2] Tones

آن‌ها (معادله ۳-۱۳) نیز محاسبه می‌شود. سپس بردار ویژگی نهایی به‌عنوان ورودی الگوریتم انتخاب ویژگی ترکیبی پیشنهادی لحاظ می‌گردد.

$$\Delta f_k[i] = f_{k+M}[i] - f_{k-M}[i] \qquad (M = 1 \dots 3\ frames)$$ معادله ۳-۱۳

۳-۲-۳ روش انتخاب ویژگی پیشنهادی

در این مرحله از مدل نیاز به تعریف یک روش انتخاب ویژگی براساس دو معیار است : ۱. به‌حداکثر رساندن دقت طبقه‌بند، ۲. به‌حداقل رساندن تعداد ویژگی‌های استفاده شده برای طبقه‌بندی. با این هدف در این رساله از یک روش ترکیبی متشکل از الگوریتم ژنتیک[۱] و تکنیک جستجوی فاخته[۲] برای انتخاب ویژگی استفاده شده است.

الگوریتم ژنتیک که در اوایل دهه ۱۹۷۰ پیشنهاد شده است، یک روش جستجوی تصادفی سراسری[۳] است که با تولید تصادفی یک جمعیت اولیه از کروموزوم‌ها آغاز شده و با استفاده از عملگرهای مختلف به‌دنبال بهبود این جمعیت براساس یک تابع برازندگی[۴] است. مطالعات زیادی از الگوریتم ژنتیک برای انتخاب ویژگی استفاده کرده‌اند. در GA، یک زیرمجموعه ویژگی توسط یک رشته باینری با طول n، به نام کروموزوم مدل می‌شود، به‌این‌صورت که مقداردهی موقعیت i با عدد یک یا صفر، انتخاب یا عدم انتخاب ویژگی i را نشان می‌دهد. کروموزوم از طریق یک تابع بهینه‌سازی برای رفتن به نسل بعدی مورد ارزیابی قرار می‌گیرد. جمعیت کروموزوم‌ها توسط دو عملگر ادغام[۵] و جهش[۶] توسعه داده می‌شوند.

الگوریتم جستجوی فاخته، یک الگوریتم الهام‌گرفته از طبیعت است که براساس تولیدمثل پرندگانی از نژاد فاخته ایجاد شده است. فاخته معمولا تخم‌های بارورشده خود را در لانه[۷] دیگر فاخته‌ها (میزبان)

[1] Genetic Algorithm (GA)
[2] Cuckoo Search (CS)
[3] Stochastic global search
[4] Fitness Function
[5] Cross-Over
[6] Mutation
[7] Nest

قرار می‌دهد و پرنده میزبان زمانی‌که متوجه می‌شود این تخم‌ها به لانه‌اش تعلق ندارند، دو رفتار احتمالی از خود نشان می‌دهد: یا تخم را از لانه بیرون می‌اندازد یا لانه را ترک می‌کند. اساس این الگوریتم برپایه سه قانون زیر استوار است [۹۹]:

- ✓ هر فاخته یک تخم (راه‌حل) را در لانه‌ای تصادفی قرار می‌دهد.
- ✓ بهترین لانه‌ها از نظر کیفیت تخم (راه‌حل) به نسل بعدی منتقل می‌شوند.
- ✓ تعداد لانه‌های موجود میزبان ثابت است و یک میزبان می‌تواند تخم بیگانه را با احتمال $p_\alpha \in [0.1]$ کشف کند. در این حالت، میزبان می‌تواند تخم را بیرون انداخته و یا لانه را ترک کند و یک لانه جدید در مکانی جدید ایجاد کند.

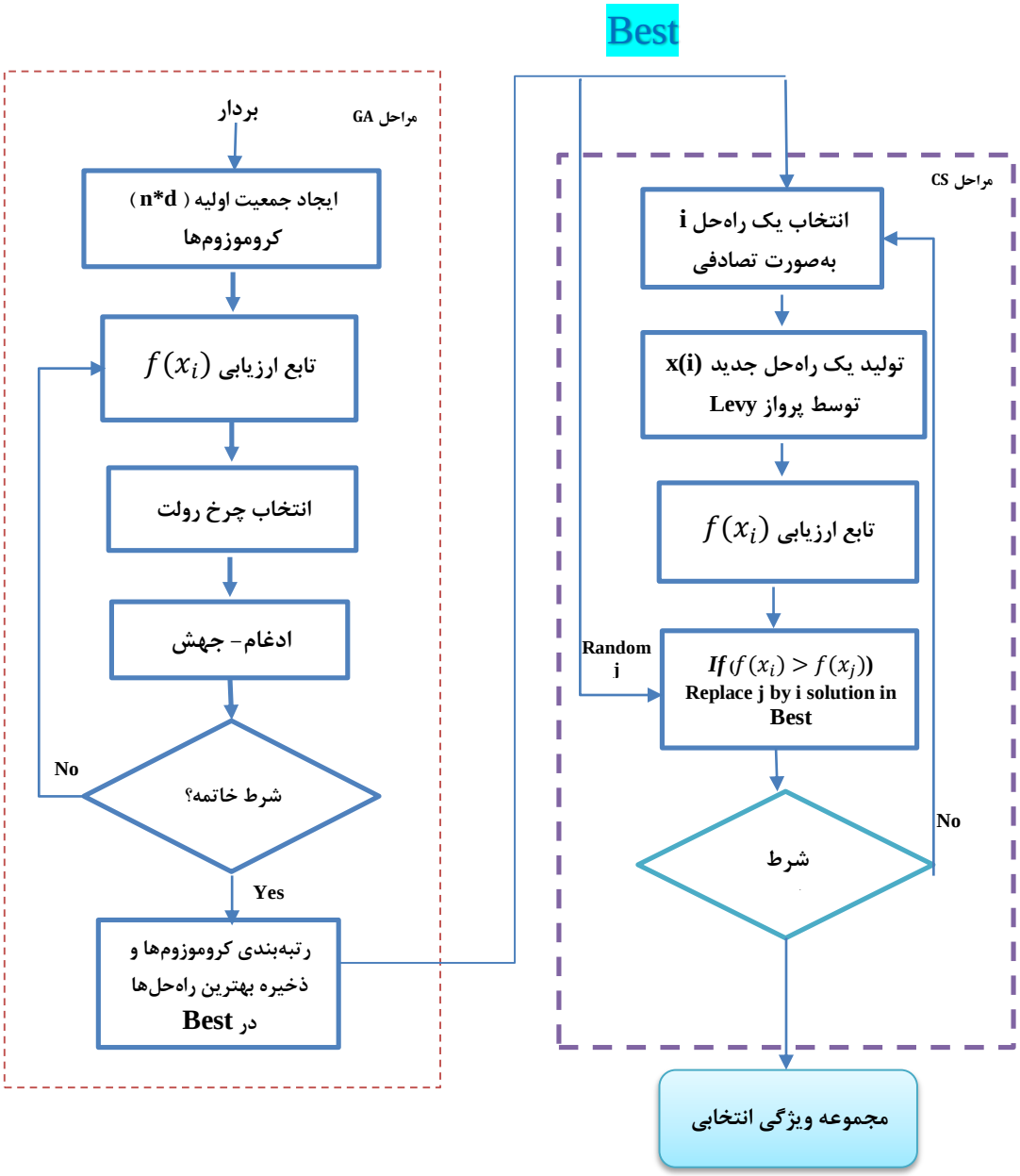
براساس این سه قانون و مفهومی به‌نام پرواز لوی[۱] گام‌های الگوریتم CS پایه تعریف می‌شوند. پروازهای لوی پیاده‌روی‌های تصادفی هستند که جهت آن‌ها نیز تصادفی است و طول گام‌های آن‌ها از توزیع لوی مشتق می‌شود. در طبیعت حیوانات غذای خود را به صورت تصادفی یا شبه‌تصادفی جستجو می‌کنند. به‌طورکلی، مسیر کاوش یک حیوان، گام‌های تصادفی است، زیرا حرکت بعدی براساس مکان یا وضعیت فعلی و احتمال انتقال به مکان بعدی تعیین می‌شود. تحقیقات مختلفی نشان داده‌اند که رفتار پروازی بسیاری از حیوانات و حشرات مشخصه‌های ویژه پرواز لوی را نشان می‌دهند و کاربرد این رفتارها در الگوریتم‌های بهینه‌سازی و روش‌های جستجوی بهینه نتایج امیدوارکننده‌ای دارند [۱۰۰][۱۰۱]. به همین دلیل در پیاده‌سازی الگوریتم جستجوی فاخته برای تولید لانه جدید، از این روش استفاده شده است.

در روش انتخاب ویژگی ترکیبی پیشنهادشده در این رساله، در اولین گام، الگوریتم ژنتیک فضای جستجو را کاوش می‌کند تا محتمل‌ترین و امیدوارکننده‌ترین منطقه در فضای جستجو را جدا کند. در مرحله دوم، جستجوی فاخته برای بهبود جستجوی سراسری و جلوگیری از گرفتار شدن در بهینه‌های

[۱] Lévy Flight

محلی فضای جستجوی محدودشده توسط GA را بسط داده و راه حل های جدید بهتری را می‌یابد. در هر دو مرحله جهت انتخاب بهترین راه‌حل‌ها از طبقه‌بند SVM در تابع برازندگی استفاده شده است. فلوچارت تکنیک انتخاب ویژگی پیشنهادی در شکل ۳-۴ نشان داده شده است.

همانطور که در فلوچارت شکل ۳-۴ نشان داده شده، یک راه‌حل نمایان‌گر یک مجموعه ویژگی است که در الگوریتم ژنتیک با مفاهیم کروموزوم (راه‌حل) و ژن (ویژگی)، و در الگوریتم فاخته به‌ترتیب با مفاهیم لانه و تخم نشان داده می‌شود. برای انتخاب یا عدم انتخاب یک ویژگی، از یک بردار باینری با طولی برابر تعداد کل ویژگی‌ها به‌عنوان راه‌حل استفاده می‌شود به‌طوری‌که عدد ۱ در هر بردار به معنای انتخاب ویژگی مربوطه برای ساخت لانه جدید است و عدد صفر نماد عدم انتخاب آن ویژگی است (شکل ۳-۵).



شکل ۳-۴ فلوچارت الگوریتم ترکیبی انتخاب ویژگی پیشنهادی

۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰	....	d-1	D
۱	۰	۰	۰	۰	۱	۱	۱	۱	۰	.....	۰	۰

شکل ۳-۵ یک نمونه کروموزوم در الگوریتم ژنتیک در روش پیشنهادی

بعد از تولید جمعیت اولیه با تعداد n کروموزوم تشکیل شده از d ژن، ارزیابی جمعیت با استفاده از تابع برازندگی انجام شده و سپس عملگرهای ادغام و جهش و انتخاب بهترین راه‌حل‌ها برای نسل بعدی، روی این کروموزوم‌ها اعمال می‌شوند. یکی دیگر از پیشنهادات این رساله، استفاده از طبقه‌بند SVM در طراحی تابع ارزیابی الگوریتم پیشنهادی است، که این تابع پیشنهادشده از دو معیار دقت طبقه‌بند و تعداد ویژگی‌های انتخاب‌شده استفاده می‌کند. معادله ۳-۱۴ تابع برازندگی پیشنهاد شده را نشان می‌دهد:

$$f(x_i) = w_r * SVM\_Acc(x_i) + 10 * w_n * f\_num(x_i)^{-1}$$ معادله ۳-۱۴

به طوری که $SVM\_Acc(x_i)$ دقت طبقه‌بند SVM با استفاده از مجموعه ویژگی پیشنهادی x_i، و $f\_num(x_i)$ نیز تعداد ویژگی‌های این مجموعه را نشان می‌دهند. همچنین وزنی که براساس میزان تاثیر دقت طبقه‌بند و تعداد ویژگی‌ها در تابع ارزیابی لحاظ شده، به‌ترتیب با وزن‌های w_r و w_n نشان داده شده است. همانطور که در شکل ۳-۴ نشان داده شده، ارزیابی جمعیت و اعمال عملگرهای ژنتیک: انتخاب چرخ رولت، ادغام و جهش تا ارضاء شرط خاتمه به‌طور تکراری ادامه می‌یابند و در پایان بهترین راه‌حل‌ها ذخیره شده و به فاز بعدی الگوریتم (جستجوی فاخته) ارسال می‌شوند.

در فاز دوم الگوریتم پیشنهادی، بهترین راه‌حل‌های تولیدشده توسط GA به‌عنوان لانه‌ها (راه‌حل‌ها) به الگوریتم CS اعمال می‌شوند و در مرحله اول یک راه‌حل تصادفی جدید x_i توسط پرواز لوی تولید می‌شود (معادله ۳-۱۵).

$$x_i^j(t+1) = \begin{cases} 1. & if\ S(x_i^j(t)) > \delta \\ 0 & otherwise \end{cases}$$

معادله ۳-۱۵

$$S(x_i^j(t)) = \frac{1}{1 + e^{-x_i^j(t)}}$$

در رابطه فوق $\delta \sim U(0.1)$ و $x_i^j(t)$ ارزش راه‌حل iام در لانه jام در مرحله t است. پرواز فاخته‌ها

براساس یک تابع سیگموئیدی انجام می‌شود، به‌این‌صورت که ابتدا موقعیت فعلی فاخته‌ها با استفاده از تابع تصادفی δ تبدیل به اعداد بین ۰ و ۱ می‌شود و سپس بهترین پرواز جهت نسل بعدی انتخاب می‌شود. سپس مقدار برازندگی این راه‌حل توسط تابع ارزیابی معادله ۳-۱۴ به‌دست آمده و با مقدار برازندگی راه‌حل تصادفی jم از مجموعه جواب تولیدشده توسط GA مقایسه شده و درصورت بهتر بودن، جایگزین می‌شود. این مراحل تا رسیدن به شرط خاتمه الگوریتم تکرار شده و در پایان یک مجموعه ویژگی بهینه به‌دست می‌آید.

۳-۲-۴ طبقه‌بندی و ارزیابی

۳-۲-۴-۱ طبقه‌بندی

آخرین مرحله در سیستم پیشنهادی تشخیص احساس از گفتار استفاده از طبقه‌بند ماشین بردار پشتیبان (SVM) با هسته تابع پایه شعاعی گوسی[1] برای دسته‌بندی سیگنال‌های گفتار در دسته‌های احساسی و ارزیابی خروجی با معیارهای استاندارد است. در این مرحله ابتدا جهت کاهش تاثیر حاصل از تنوع گوینده در نرخ تشخیص، بردارهای ویژگی قبل از آموزش و آزمایش طبقه‌بندی با روش استانداردسازی-z مطابق با معادله ۳-۱۶ نرمال می‌شوند.

$$\widehat{f^s} = \frac{f^s - \mu_{neu}^{f^s}}{\sigma_{neu}^{f^s}}$$

معادله ۳-۱۶

در رابطه فوق f^s بردار ویژگی مستخرج از سیگنال گفتار s است، و $\mu_{une}^{f^s}$ و $\sigma_{neu}^{f^s}$ به‌ترتیب مقادیر میانگین و انحراف معیار ویژگی‌ها را نشان می‌دهند. سپس بردارهای ویژگی نرمال‌شده جهت آموزش و

[1] Gaussian Radial Basis Function Kernel

آزمایش طبقه‌بند با روش اعتبارسنجی متقابل k-لایه[1] انتخاب می‌شوند. در روش اعتبارسنجی متقابل k-لایه مجموعه داده‌های آموزشی به‌طور تصادفی به k زیرنمونه یا لایه (Fold) با حجم یکسان تفکیک شده، و سپس در هر مرحله از فرایند اعتبارسنجی، تعداد k-1 از این لایه‌ها به‌عنوان مجموعه داده آموزشی و یک عدد به‌عنوان مجموعه داده اعتبارسنجی در نظر گرفته می‌شوند.

پیچیدگی محاسباتی ماشین بردار پشتیبان در مرحله آزمایش وابسته به تعداد بردارهای پشتیبانی است که در مرحله آموزش به‌دست آمده است. معادله ۳-۱۷ رابطه مرز تصمیم‌گیری ماشین بردار پشتیبان را در حالتی که از تابع هسته استفاده شود برای حالت دو کلاسه نشان می‌دهد:

$$\hat{y} = sgn\left(\sum_{i=1}^{n} w_i y_i \mathcal{K}(x_i . x^{'})\right)$$ معادله ۳-۱۷

که در آن $x^{'} \in R^n$ یک نمونه آزمایشی و $x_i \in R^n$ نیز iمین نمونه آموزشی است. $y_i \in \{-1 . +1\}$ خروجی iمین نمونه آموزشی و x_i وزن iمین نمونه آموزشی است. تابع $\mathcal{K}(x_i . x^{'})$ تابع هسته SVM است. انتخاب تابع هسته مناسب در طبقه‌بند SVM از اهمیت خاصی برخوردار است. این انتخاب به جنس داده‌های آموزشی بستگی دارد. به‌عبارت‌دیگر معیار شباهت بین داده‌ها می‌تواند در انتخاب هسته مناسب کمک کند. در اکثر کاربردها معیار شباهت اقلیدسی معیار مناسبی است. در چنین حالتی هسته گوسی که تابعی از فاصله اقلیدسی است، مناسب است. در سیستم تشخیص احساس از گفتار پیشنهادی این رساله از طبقه‌بند SVM با هسته تابع پایه شعاعی گوسی مطابق با معادله ۳-۱۸ استفاده شده است:

$$\mathcal{K}(x_i . x_j) = exp(-\gamma) \left\| x_i - x_j \right\|^2$$ معادله ۳-۱۸

[1] K-fold Cross Validation

۳-۲-۴-۲ ارزیابی عملکرد

گام بعدی، سنجش عملکرد مدل بر اساس ماتریس درهم‌ریختگی[1] و با استفاده از معیارهای ارزیابی عملکرد متداول است. ماتریس درهم‌ریختگی که به آن ماتریس خطا نیز می‌گویند، جدولی است که عملکرد یک مدل یادگیری ماشین تحت نظارت را بر روی داده‌ها توصیف می‌کند. هر ستون این ماتریس نشان‌دهنده نمونه‌های پیش‌بینی‌شده در یک کلاس است، در حالی‌که هر سطر نمونه‌های واقعی آن کلاس را نشان می‌دهد. شکل ۳-۶ نمونه‌ای از این ماتریس را نشان می‌دهد.

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

شکل ۳-۶ یک نمونه ماتریس درهم‌ریختگی

مفهوم هر یک از عناصر ماتریس درهم‌ریختگی به شرح ذیل است:

- ✓ **TN**: تعداد نمونه‌هایی که برچسب (دسته) واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز برچسب آن‌ها را به‌درستی منفی تشخیص داده است.
- ✓ **TP**: تعداد نمونه‌هایی که برچسب واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز برچسب آن‌ها را به‌درستی مثبت تشخیص داده است.
- ✓ **FP**: تعداد نمونه‌هایی که برچسب واقعی آن‌ها منفی بوده اما الگوریتم دسته‌بندی برچسب آن‌ها را به اشتباه مثبت تشخیص داده است.
- ✓ **FN**: تعداد نمونه‌هایی که برچسب واقعی آن‌ها مثبت بوده اما الگوریتم دسته‌بندی برچسب آن‌ها را به اشتباه منفی تشخیص داده است.

[1] Confusion Matrix

محققان براساس این ماتریس و عناصر آن معیارهای ارزیابی کارایی مختلفی را تعریف نموده‌اند که می‌توان برای ارزیابی عملکرد سیستم تشخیص احساس از گفتار پیشنهادی از آن‌ها استفاده نمود. از جمله معیارهای پرکاربرد برای ارزیابی عملکرد SER می‌توان به معیارهای دقت[1]، صحت[2]، حساسیت[3] یا یادآوری[4] و خاصیت[5] اشاره نمود. در ادامه تعریف مختصری از معیارهای ارزیابی به‌کار رفته در این رساله خواهیم داشت.

- **دقت:** این پارامتر نشان‌دهنده این است که چند درصد از الگوهایی که دسته‌بند آن‌ها را مثبت تشخیص داده، در واقعیت هم مثبت هستند. و براساس ماتریس ارائه‌شده در شکل ۳-۶، متوسط دقت برای یک الگوریتم یادگیر مطابق رابطه زیر فرموله و تعریف می‌شود.

$$Average\ Precision = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$$ معادله ۳-۱۹

در تمام روابط محاسبه معیارهای ارزیابی، پارامتر l نمایان‌گر تعداد کلاس‌ها است.

- **صحت:** پارامتر صحت، متداول‌ترین، اساسی‌ترین و ساده‌ترین معیار اندازه‌گیری کیفیت یک طبقه‌بند است و عبارت است از میزان تشخیص صحیح دسته‌بند در مجموع دو دسته. این پارامتر در واقع نشان‌گر میزان الگوهایی است که درست تشخیص داده شده‌اند و براساس ماتریس ارائه‌شده در شکل ۳-۶، متوسط صحت برای یک الگوریتم یادگیر مطابق رابطه زیر فرموله و تعریف می‌شود:

$$Average\ Accuracy = \frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$$ معادله ۳-۲۰

[1] Precision
[2] Accuracy
[3] Sensitivity
[4] Recall
[5] Specificity

- **حساسیت:** حساسیت که همان معیار فراخوان (Recall) است، به معنی نسبتی از نمونه‌های مثبت است که سیستم آن‌ها را به‌درستی به‌عنوان نمونه مثبت تشخیص داده است. در اکثر تحقیقات مربوط به حوزه تشخیص احساس از گفتار معمولا عملکرد سیستم با استفاده از "میانگین بدون وزن صحت[1]" (UA) و "میانگین وزنی صحت[2]" (WA) گزارش می‌شود [۴]. طبق مطالعه سوکولووا و همکارانش، معیارهای UA و WA به‌ترتیب با معیارهای Macro-Average Recall و Micro-Average Recall برابر هستند [۱۰۲]. این معیارها با استفاده از روابط زیر محاسبه می‌شوند:

$$Unweighted\ Average(Macro\ Average\ Recall) = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$$ معادله ۳-۲۱

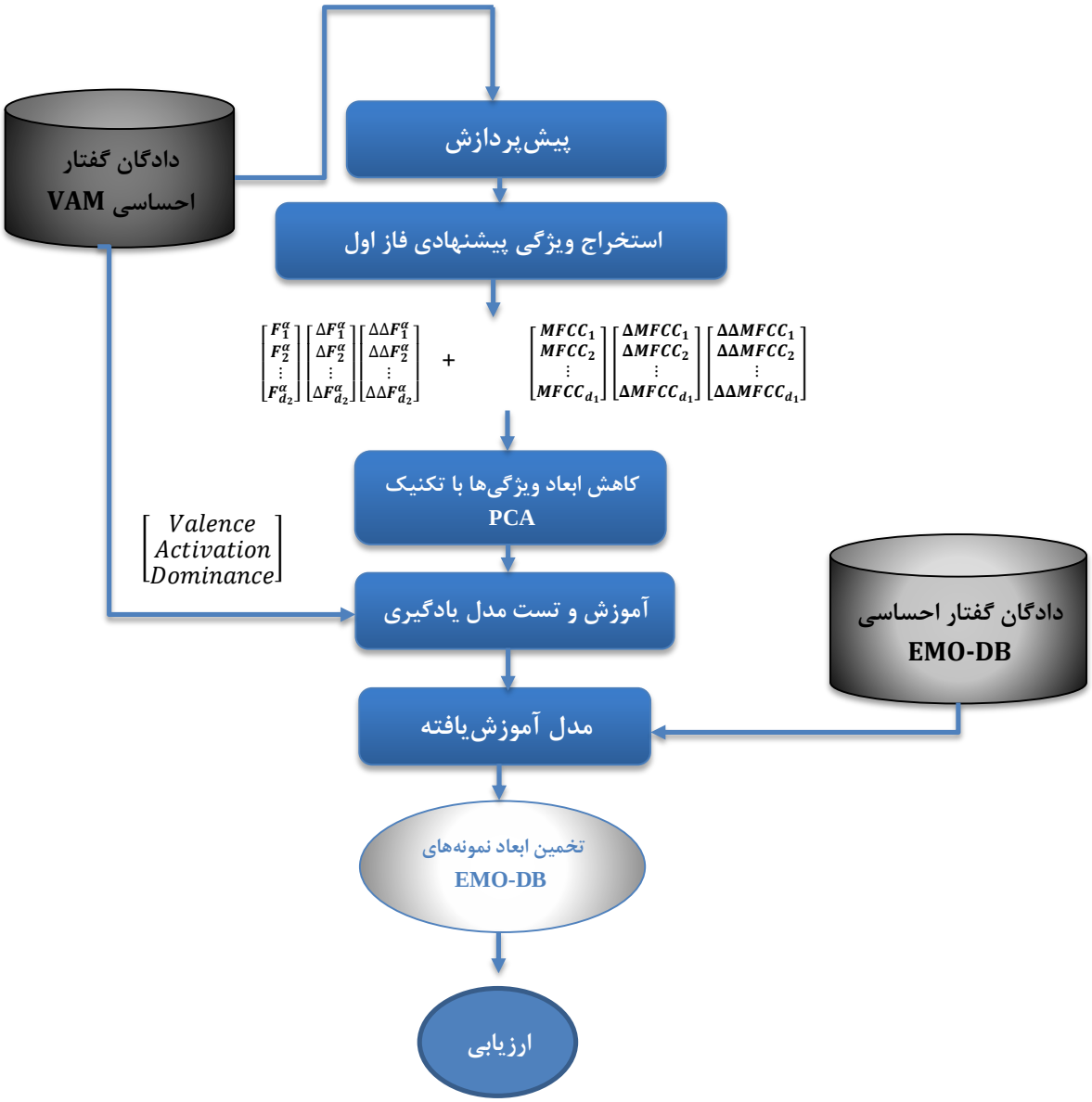
$$Weighted\ Average\ (Micro\ Average\ Recall) = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fn_i)}$$ معادله ۳-۲۲

۳-۳ فاز دوم: سیستم تشخیص احساس از گفتار مبتنی‌بر نگاشت مجموعه ویژگی‌ها به فضای چندبعدی احساس

در فاز دوم مدل پیشنهادی این رساله، با استفاده از ویژگی‌های پیشنهادی فاز اول یک نگاشت بین فضای ویژگی آکوستیک گفتار و فضای سه بعدی احساس تخمین زده شده است. در الگوریتم پیشنهادی این فاز ابتدا با استفاده از روش استخراج ویژگی پیشنهادی فاز اول و نمونه‌های مجموعه دادگان گفتار احساسی فضای پیوسته VAM یک مدل یادگیر آموزش داده می‌شود و سپس ابعاد داده‌های گفتار

[1] Unweighted Average (UA) Recall
[2] Weighted Average (WA) Recall

احساسی گسسته پایگاه داده EMO-DB با این مدل تخمین زده می‌شوند. نمای کلی این فاز در شکل ۳-۷ نشان داده شده است.



شکل ۳-۷ بلوک دیاگرام فاز دوم- نگاشت مجموعه ویژگی‌ها به فضای چندبعدی احساس

همان‌طور که در شکل ۳-۷ مشخص شده است، بعد از فرایند پیش‌پردازش و استخراج ویژگی‌های پیشنهادی از سیگنال‌های گفتار دادگان VAM، قبل از نگاشت به فضای سه بعدی ابعاد فضای ویژگی‌ها بااستفاده از روش تحلیل مؤلفه اصلی PCA کاهش می‌یابد. در ادامه مراحل این فاز شرح داده شده‌اند.

- **کاهش ابعاد ویژگی‌ها با روش تحلیل مؤلفه‌های اصلی PCA**: در این الگوریتم برای کاهش

تعداد ویژگی‌ها از روش نگاشت ویژگی‌ها به یک فضای جدید با ابعاد کوچکتر استفاده می‌شود. هدف این الگوریتم از بین بردن همبستگی بین ویژگی‌های جدید است. به‌طورکلی الگوریتم PCA سعی دارد بهترین نمایش یک داده را با ابعاد کوچکتر ارائه دهد. ورودی این الگوریتم، مجموعه‌ای از ویژگی-های پیشنهادی استخراج شده از دادگان VAM است. برای انجام این تبدیل، ابتدا کواریانس بین این ویژگی‌ها (که دارای میانگین صفر هستند) محاسبه شده و سپس این ماتریس کواریانس با تبدیل A مطابق معادله ۳-۲۳ به یک ماتریس کواریانس قطری تبدیل می‌شود.

$$y = A^T x$$ معادله ۳-۲۳

تبدیل A تبدیلی است که ویژگی‌ها را به فضای جدیدی منتقل می‌کند. ستون‌های این ماتریس، همان بردارهای ویژه ماتریس کواریانس بین ویژگی‌های با میانگین صفر است. یعنی:

$$R_y \equiv E[yy^T] = E[A^T x x^T A] = A^T R_x A$$
$$R_y = A^T R_x A = \Lambda$$ معادله ۳-۲۴

$$R_x \approx \frac{1}{n}\sum_{k=1}^{n} x_k x_k^T$$ معادله ۳-۲۵

بااستفاد از این انتقال، کواریانس بین ویژگی‌های جدید، یک ماتریس قطری خواهد بود:

$$\Sigma_y = A^T \Sigma_x A = \Lambda$$ معادله ۳-۲۶

- **آموزش و تست مدل یادگیری:** در این مرحله یک شبکه عصبی با استفاده از ویژگی‌های به‌دست آمده از مرحله قبل براساس یک تابع خروجی سه مقداره (جاذبه، انگیختگی، تسلط) آموزش می‌یابد. سپس ویژگی‌های مستخرج از پایگاه داده EMO-DB برای تخمین مقادیر سه بعد (جاذبه، انگیختگی، تسلط) به این مدل آموزش‌دیده اعمال می‌شوند.
- **ارزیابی:** در مرحله آخر، مقادیر سه بعد (جاذبه، انگیختگی، تسلط) تخمین زده شده توسط

معیارهای ارزیابی خطای میانگین مربعات[۱] (MSE) و ضریب همبستگی پیرسون[۲] مطابق معادلات زیر ارزیابی می‌شوند.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$ معادله ۳-۲۷

$$CC(Y.\hat{Y}) = \frac{cov(Y.\hat{Y})}{\sigma_Y \sigma_{\hat{Y}}} = \frac{\sum_{i=1}^{N}(Y-\mu_Y)(\hat{Y}-\mu_{\hat{Y}})}{\sqrt{\sum_{i=1}^{N}(Y-\mu_Y)^2}\sqrt{\sum_{i=1}^{N}(\hat{Y}-\mu_{\hat{Y}})^2}}$$ معادله ۳-۲۸

در معادلات فوق Y بردار سه بعدی مقادیر واقعی (جاذبه، انگیختگی، تسلط) و $\hat{Y}$ بردار خروجی تخمین‌زده شده توسط مدل یادگیرنده برای هر نمونه i از N نمونه است. همچنین پارامترهای μ_Y ، $\mu_{\hat{Y}}$، σ_Y و $\sigma_{\hat{Y}}$ به‌ترتیب میانگین و واریانس این مقادیر هستند.

۳-۴ جمع‌بندی

در این فصل در دو فاز یک سیستم بهبودیافته‌ی تشخیص احساس از گفتار براساس دو مدل گسسته و پیوسته فضای احساسات ارائه شد. در فاز اول با ارائه یک روش استخراج ویژگی نوین و یک روش ترکیبی تکاملی انتخاب ویژگی یک مدل بهینه برای تشخیص کلاس‌های گسسته احساسی پیشنهاد شد. در فاز دوم با استفاده از آموزش یک مدل یادگیرنده توسط ویژگی‌های مستخرج از یک مجموعه دادگان و تخمین مقادیر سه بعد برای ویژگی‌های استخراج شده از یک مجموعه دادگان دیگر، یک نگاشت از فضای ویژگی‌ها به فضای سه بعدی احساس تخمین زده شد. در فصل بعد با استفاده از آزمایش‌های متعدد این سیستم پیشنهادی اعتبارسنجی و عملکرد آن ارزیابی خواهد شد.

---

۱ Mean squared error

۲ Pearson’s Correlation Coefficient

فصل ۴ آزمایش‌ها و تجزیه و تحلیل نتایج

۴-۱ مقدمه

در این رساله در دو فاز یک سیستم تشخیص احساس از گفتار با هدف بهبود دقت تشخیص و براساس دو مدل احساسی پیشنهاد شد. در فاز اول سیستم پیشنهادی روی دو بخش کلیدی مساله تشخیص احساس از گفتار، یعنی استخراج و انتخاب ویژگی‌های موثر در طبقه‌بندی دقیق احساسات، متمرکز بود. در این راستا، علاوه‌بر استفاده از ویژگی‌های متداول سایر تحقیقات، روشی برای استخراج یک ویژگی جدید تطبیق یافته در حوزه زمان-فرکانس پیشنهاد شد. همچنین یک الگوریتم ترکیبی تکاملی نیز برای انتخاب بهترین مجموعه ویژگی قبل از مرحله طبقه‌بندی ارائه شد. در فاز دوم یک نگاشت بین این سیستم پیشنهادی و مدل سه‌بعدی احساس پیشنهاد شد. برای ارزیابی این دو مدل پیشنهادی آزمایش‌های متعددی در هر بخش سیستم انجام شد که در ادامه ارائه خواهد شد. برای ارزیابی عملکرد و نتایج به‌دست آمده هر دو مدل پیشنهادی سیستم تشخیص احساس از گفتار با استفاده از نرم‌افزار MATLAB-R2017 پیاده‌سازی شده است. همچنین برای آموزش و آزمایش سیستم پیشنهادی از چهار مجموعه دادگان استفاده شد که در ادامه به تفصیل شرح داده می‌شوند.

در ادامه این فصل، ابتدا مجموعه دادگان استفاده شده برای ارزیابی روش پیشنهادی معرفی خواهند شد. سپس ارزیابی فاز اول روش پیشنهادی رساله با انجام آزمایش‌ها در سه دسته (ارزیابی روش استخراج ویژگی پیشنهادی- ارزیابی روش انتخاب ویژگی پیشنهادی- تاثیر طبقه‌بند) انجام می‌شود. در بخش بعدی، فاز دوم یعنی نگاشت مجموعه ویژگی‌ها به فضای سه‌بعدی احساس مورد ارزیابی و تجزیه و تحلیل قرار خواهد گرفت.

۴-۲ مجموعه دادگان

در این رساله جهت ارزیابی روش پیشنهادی آزمایش‌های متعددی با چهار مجموعه داده شامل: پایگاه داده احساسی برلین (Emo-DB)، دادگان احساسی SAVEE، پایگاه داده گفتار احساسی فارسی درام

و مجموعه داده مدل احساس پیوسته VAM انجام شده است. در جدول (۲-۱) فصل دوم اطلاعات کلی درباره این چهار مجموعه دادگان ارائه شده است. مجموعه دادگان Emo-DB به زبان آلمانی و شامل ۵۳۵ سیگنال گفتار بیان‌شده توسط ۱۰ بازیگر (پنج زن و پنج مرد) است، که هر بازیگر ۱۰ جمله را در هفت کلاس احساسی: خوشحالی، غم، عصبانیت، ترس، تنفر، خستگی و خنثی بیان کرده‌اند. پایگاه داده SAVEE به زبان انگلیسی است و ۴۲۰ سیگنال گفتار آن توسط چهار گوینده مرد در ۱۵ جمله با هفت دسته احساسی شامل: خوشحالی، غم، عصبانیت، ترس، تنفر، تعجب و خنثی بیان‌شده‌اند. مجموعه دادگان فارسی درام (PDREC) شامل ۷۴۳ سیگنال گفتار احساسی با هشت برچسب: عصبانیت، خستگی، نفرت، ترس، ناراحتی، تعجب، خوشحالی و خنثی است که از رادیو نمایش ضبط شده‌اند. در این پایگاه داده سخنرانی‌های ۳۳ بازیگر (۱۵ زن و ۱۸ مرد) از میان برنامه‌های پخش‌شده از رادیو نمایش ضبط و برچسب‌گذاری شده است. این سه مجموعه دادگان برای ارزیابی نتایج آزمایش‌های فاز اول روش پیشنهادی رساله استفاده شده‌اند. در فاز دوم از مجموعه دادگان گفتار احساس طبیعی با زبان آلمانی VAM استفاده شده که شامل ۹۴۷ سیگنال گفتار است که از یک برنامه جنگ تلویزیونی ضبط شده است. در این پایگاه داده برای هر سیگنال گفتار سه پارامتر: جاذبه، انگیختگی و قدرت (تسلط) لحاظ شده که هر پارامتر دارای دو مقدار میانگین وزنی ارزش احساس[۱] و انحراف معیار برچسب‌گذاری[۲] است. جدول ۴-۱ توزیع مجموعه دادگان گفتار احساسی گسسته استفاده شده در این رساله را نشان می‌دهد.

جدول ۴-۱ توزیع مجموعه دادگان گفتار احساسی استفاده شده در فاز اول- (HP:خوشحالی، SD:غم، AG: عصبانیت، FR:ترس، BD:خستگی، DG:تنفر، NE:خنثی، SR:تعجب، و (ز): زن، (م): مرد.)

دادگان / احساسات	HP	SD	AG	FR	BD	DG	NE	SR	Total
Emo-Db	(ز)۴۴ (م)۲۷	(ز)۳۷ (م)۲۵	(ز)۶۷ (م)۶۰	(ز)۳۲ (م)۳۷	(ز)۴۶ (م)۳۵	(ز)۳۵ (م)۱۱	(ز)۴۰ (م)۳۹	-	(ز)۳۰۱ (م)۲۳۴
SAVEE	(م)۶۰	(م)۶۰	(م)۶۰	(م)۶۰	-	(م)۶۰	(م)۶۰	(م)۶۰	(م)۴۲۰
PDREC	(ز)۳۶ (م)۴۸	(ز)۷۸ (م)۶۷	(ز)۷۳ (م)۱۰۴	(ز)۲۱ (م)۴۲	(ز)۱۵ (م)۱۴	(ز)۱۲ (م)۵	(ز)۸۸ (م)۱۰۶	(ز)۱۶ (م)۲۳	(ز)۳۳۹ (م)۴۰۹

[۱] Average Weighted Emotion Value

[۲] Standard Deviation of Labeling

۴-۳ ارزیابی فاز اول: سیستم تشخیص احساس از گفتار با استفاده از روش پیشنهادی استخراج ویژگی زمان-فرکانس تطبیقی و الگوریتم انتخاب ویژگی ترکیبی-تکاملی

در فاز اول رساله یک سیستم تشخیص احساس از گفتار پیشنهاد شده است که مبتنی بر یک روش پیشنهادی استخراج ویژگی زمان-فرکانس تطبیقی و یک الگوریتم انتخاب ویژگی ترکیبی-تکاملی عمل تشخیص احساسات را انجام می‌دهد. جهت ارزیابی عملکرد این سیستم آزمایش‌های مختلفی را در سه بخش با سه مجموعه دادگان با سه زبان مختلف انجام دادیم. در آزمایش اول نتایج روش استخراج ویژگی پیشنهادی در مقایسه با روش‌های استخراج ویژگی متداول بررسی و مورد ارزیابی و تجزیه و تحلیل قرار خواهد گرفت. در آزمایش دوم نتایج روش انتخاب ویژگی پیشنهادی در مقایسه با روش‌های انتخاب ویژگی متداول ارزیابی می‌شود. و در آخرین آزمایش ارزیابی و تجزیه و تحلیل نتایج سیستم تشخیص احساس از گفتار پیشنهادی با اعمال طبقه‌بندهای مختلف مورد بحث قرار می‌گیرد. قبل از انجام آزمایش‌ها، ابتدا چگونگی تنظیمات مقادیر پارامترهای مختلف موجود در روش پیشنهادی را بیان خواهیم کرد و سپس به بررسی آزمایش‌ها و نتایج آن‌ها می‌پردازیم.

۴-۳-۱ تنظیمات مقادیر پارامترهای روش پیشنهادی

همان‌طور که در فصل سوم بیان شد در تمام بخش‌ها و مراحل سیستم پیشنهادی تشخیص احساس از گفتار پارامترهایی وجود دارند که تنظیم درست و دقیق آن‌ها در نتیجه تشخیص موثر است. در جدول ۴-۲ تمام تنظیمات مربوط به پارامترهای روش پیشنهادی این رساله نشان داده شده است. در پیش‌پردازش پارامترهایی نظیر اندازه فریم و مقدار هم‌پوشانی فریم‌ها براساس مقادیر گزارش‌شده توسط محققان در آزمایش‌های انجام‌شده روی دادگان موردنظر تنظیم شده‌اند.

جدول ۴-۲ تنظیمات مربوط به پارامترهای روش پیشنهادی

بخش موردنظر	پارامتر	مقدار
پیش‌پردازش	فرکانس نمونه‌برداری	$fs = 16kHz$
	طول فریم	$t\_frame = 20\ ms$
	نرخ هم‌پوشانی	$\frac{L}{2} = 10ms$
تعداد ویژگی‌ها	میزان چرخش زاویه	$\alpha = 0.85$
	MFCC	۱۳ ضریب- و مشتقات مرتبه اول و مرتبه دوم
	LPCC	۱۳ ضریب- و مشتقات مرتبه اول و مرتبه دوم
	DFT	۱۲۰ ضریب- و مشتقات مرتبه اول مرتبه دوم
	ویژگی پیشنهادی	۱۹۴ ضریب- و مشتقات مرتبه اول و مرتبه دوم
انتخاب ویژگی پیشنهادی	نرخ ادغام	۰,۸۵
	نرخ جهش	۰,۲۵
	جمعیت اولیه GA	۵۰
	تعداد لانه‌های CS	۱۰۰
طبقه‌بند	پارامترهای کرنل	$c = 10, \gamma = 0.1$
	تعداد لایه‌ها (فولدها)	$k=10$

۴-۳-۲ آزمایش اول: ارزیابی و تجزیه و تحلیل نتایج روش استخراج ویژگی پیشنهادی در مقایسه با روش‌های استخراج ویژگی متداول

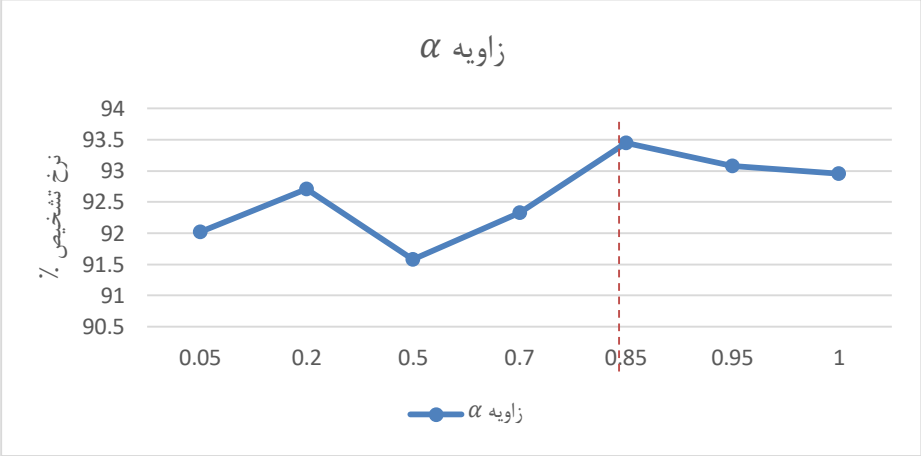
در آزمایش اول نتایج تشخیص احساس از گفتار با استفاده از روش استخراج ویژگی پیشنهادی روی سه مجموعه دادگان EMO-DB، SAVEE و PDREC مورد ارزیابی و مقایسه با سایر روش‌های استخراج ویژگی قرار گرفت. همان‌طورکه در تحقیق Ayadi و همکاران [۱۱] گزارش شده است، ویژگی‌های سراسری به لحاظ نرخ طبقه‌بندی و کارایی محاسباتی نسبت به ویژگی‌های محلی بهتر هستند، بنابراین در این رساله بعد از انجام مرحله پیش‌پردازش و فریم‌بندی شدن سیگنال‌ها، ویژگی‌های موردنظر از هر فریم استخراج می‌شوند. این عمل منجر به ایجاد یک منحنی برای هر مجموعه ویژگی مستخرج از هر سیگنال می‌شود. سپس ۲۰ تابع آماری میانگین، ماکزیمم، مینیمم، میانه، دامنه تغییرات، انحراف معیار،

انحراف متوسط از میانگین، صدک[۱] ۱ام، ۵ام، ۱۰ام، ۲۵ام، ۷۵ام، ۹۰ام، ۹۵ام و ۹۹ام، چولگی[۲]، دامنه بین چارکی[۳]، درجه اوج[۴] و میانگین اصلاح‌شده[۵] ۱۰٪ و ۲۵٪ روی هر منحنی به‌دست‌آمده از هر ویژگی اعمال می‌شوند.

همان‌طور که در سیستم پیشنهادی تشخیص احساس از گفتار بیان شد، آخرین مرحله از فرآیند استخراج ویژگی پیشنهادی، یافتن زاویه بهینه برای چرخش سیگنال در فضای زمان-فرکانس است. در راستای این هدف آزمایش‌های بسیاری با استفاده از روش اعتبارسنجی انجام شد و مقادیر مختلف α مطابق جدول ۴-۳ آزمایش شد. از آزمایش‌ها مشاهد شد که بهترین مقدار برای پارامتر α مطابق شکل ۴-۱ مقدار ۰,۸۵ است.

جدول ۴-۳ بررسی مقادیر مختلف α در تاثیر ویژگی پیشنهادی در کارایی طبقه‌بند

مقادیر α	۱	۰,۹۵	۰,۸۵	۰,۷	۰,۵	۰,۲	۰,۰۵
نرخ تشخیص درست طبقه‌بند(٪)	۹۲,۹۵	۹۳,۰۸۴	۹۳,۴۵	۹۲,۳۳	۹۱,۵۸	۹۲,۷۱	۹۲,۰۲



شکل ۴-۱ مقایسه تاثیر مقادیر مختلف ضریب α در ویژگی پیشنهادی روی دقت طبقه‌بند

[۱] Percentile
[۲] Skewness
[۳] Interquartile Range
[۴] Kurtosis
[۵] Trimmed Mean

بعد از یافتن بهترین مقدار زاویه α، سیستم تشخیص احساس از گفتار پیشنهادی را با استفاده از مجموعه دادگان EMO-DB، SAVEE و PDREC و اعمال تنظیمات جدول ۴-۲، مورد آزمایش قرار دادیم. ماتریس‌های درهم‌ریختگی به‌دست آمده از نتایج آزمایش سیستم پیشنهادی روی هر سه مجموعه دادگان در جدول ۴-۴ الی جدول ۴-۶ ارائه‌شده است.

جدول ۴-۴ ماتریس درهم‌ریختگی SER پیشنهادی با مجموعه دادگان EMO-DB

EMO-DB	غم	خنثی	خوشحالی	ترس	تنفر	خستگی	عصبانیت
غم	۵۹	۱	۰	۱	۱	۰	۰
خنثی	۰	۷۷	۲	۰	۰	۰	۰
خوشحالی	۰	۲	۶۷	۲	۰	۰	۰
ترس	۰	۱	۲	۶۶	۰	۰	۰
تنفر	۰	۱	۰	۰	۴۵	۰	۰
خستگی	۰	۰	۰	۰	۰	۸۱	۰
عصبانیت	۰	۰	۰	۰	۰	۰	۱۲۷

همان‌طور که در ماتریس‌های درهم‌ریختگی مشاهده می‌شود، نتایج روش پیشنهادی روی هر سه مجموعه دادگان دقت بالای ۸۰٪ داشته است. نتایج به‌دست آمده از ارزیابی روش پیشنهادی با استفاده از دادگان EMO-DB نشان می‌دهد که تقریبا کلاس احساسی ۹۷٫۲۱٪ داده‌ها به‌درستی تشخیص داده شده است. سیستم پیشنهادی در مجموعه دادگان SAVEE و PDREC نیز دقت بالایی به‌ترتیب در حدود ۸۰٪ و ۸۷٫۳۱٪ دارد. جدول ۴-۴ نشان می‌دهد که علاوه‌بر اینکه تمام نمونه‌های کلاس‌های احساسی خستگی و عصبانیت در مجموعه دادگان EMO-DB درست تشخیص داده شده‌اند، هیچ نمونه‌ای از کلاس‌های دیگر نیز به غلط در این کلاس‌ها دسته‌بندی نشده‌اند. با توجه به ماتریس‌های درهم‌ریختگی ارائه شده، در مجموعه دادگان SAVEE و PDREC بالاترین دقت تشخیص مربوط به نمونه‌های کلاس احساسی خنثی است.

جدول ۴-۵ ماتریس درهم‌ریختگی SER پیشنهادی با مجموعه دادگان SAVEE

SAVEE	غم	خنثی	خوشحالی	ترس	تنفر	تعجب	عصبانیت
غم	۵۰	۲	۰	۱	۶	۰	۱
خنثی	۳	۵۷	۰	۰	۰	۰	۰
خوشحالی	۱	۰	۴۱	۵	۲	۳	۸
ترس	۱	۰	۵	۴۵	۱	۶	۲
تنفر	۴	۰	۲	۲	۴۹	۱	۲
تعجب	۰	۱	۲	۸	۲	۴۳	۴
عصبانیت	۰	۰	۵	۱	۲	۱	۵۱

جدول ۴-۶ ماتریس درهم‌ریختگی SER پیشنهادی با مجموعه دادگان PDREC

PDREC	عصبانیت	ترس	خوشحالی	غم	خنثی	خستگی	تنفر	تعجب
عصبانیت	۱۶۸	۳	۳	۱	۱	۰	۰	۱
ترس	۶	۵۱	۶	۰	۰	۰	۰	۰
خوشحالی	۶	۰	۶۹	۹	۱	۰	۰	۰
غم	۲	۰	۱	۱۳۳	۹	۰	۰	۰
خنثی	۰	۰	۰	۱	۱۹۳	۰	۰	۰
خستگی	۰	۰	۰	۰	۳	۲۴	۲	۰
تنفر	۰	۰	۰	۰	۰	۳	۱۴	۰
تعجب	۱	۰	۰	۰	۰	۱	۴	۳۴

همچنین عملکرد روش پیشنهادی روی هر سه مجموعه دادگان با استفاده از معیارهای ارزیابی اشاره‌شده در فصل سه در جدول ۴-۷ ارائه شده است. همان‌طور که در جدول مشاهد می‌شود روش پیشنهادی در تشخیص کلاس‌های احساسی مجموعه دادگان EMO-DB بسیار بهتر از دو مجموعه دادگان دیگر عمل کرده است. این امر می‌تواند به‌علت وجود استانداردهای محکم در جمع‌آوری این مجموعه دادگان از جمله: توازن در تعداد گویندگان و جنسیت آن‌ها و تعداد سیگنال‌ها در هر کلاس باشد. همان‌طور که در فصل دوم بیان شد در مجموعه دادگان SAVEE تنها از گویندگان مرد استفاده شده است و تعداد نمونه‌های آن نسبت به دو مجموعه دادگان دیگر کمتر است. این مساله بر میزان پایین بودن دقت تشخیص روش پیشنهادی در این پایگاه داده نسبت به دو مجموعه دیگر تاثیر دارد.

جدول ۴-۷ عملکرد سیستم پیشنهادی تشخیص احساس از گفتار روی سه مجموعه دادگان

دادگان	متوسط صحت[۱]	متوسط دقت[۲]	میانگین وزنی صحت[۳] (WA)	میانگین بدون وزن صحت[۴] (UA)
EMO-DB	٪۹۹,۳۱	٪۹۷,۲۳	٪۹۷,۵۷	٪۹۷,۲۱
SAVEE	٪۹۴,۲۹	٪۸۰,۰۸	٪۸۰	٪۸۰
PDREC	٪۹۷,۸۷	٪۸۹,۹۸	٪۹۱,۴۷	٪۸۷,۹۸

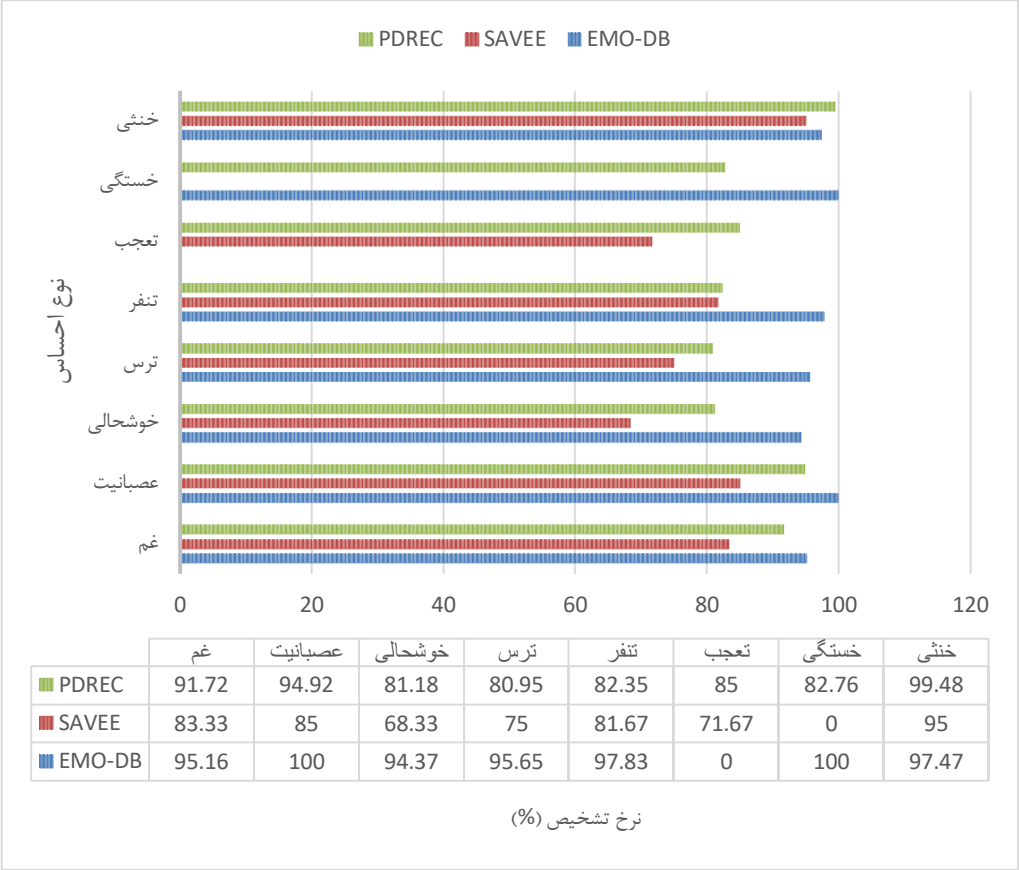
همچنین برای بررسی تاثیر توزیع متوازن کلاس‌های احساسی روی عملکرد سیستم، در شکل ۴-۲ و جدول ۴-۸ نتایج تشخیص هر کلاس توسط سیستم پیشنهادی را در هر پایگاه داده بررسی کردیم. نتایج ارائه شده در شکل ۴-۲ نشان می‌دهد که صد درصد نمونه‌های کلاس‌های احساسی عصبانیت و خستگی در مجموعه دادگان EMO-DB با استفاده از ویژگی پیشنهادی درست تشخیص داده شده‌اند. به‌طورکلی از شکل می‌توان مشاهده نمود که کلاس احساسی عصبانیت در هر سه مجموعه دادگان دقت تشخیص بالایی دارد. کمترین میزان تشخیص نیز مربوط به کلاس احساسی «خوشحالی» در مجموعه دادگان EMO-DB و SAVEE، و کلاس احساسی «ترس» در دادگان PDREC است. در جدول ۴-۸ نیز نرخ تشخیص و تعداد نمونه‌های هر کلاس در سه مجموعه دادگان بررسی شده است.

جدول ۴-۸ نرخ تشخیص و تعداد نمونه‌های هر کلاس در سه مجموعه دادگان (EMO-DB,SAVEE,PDREC) توسط سیستم پیشنهادی

کلاس‌های احساسی / دادگان	EMO-DB	SAVEE	PDREC
غم	٪۹۵,۱۶ (۶۲نمونه)	٪۸۳,۳۳ (۶۰نمونه)	٪۹۱,۷۲ (۱۴۵نمونه)
عصبانیت	٪۱۰۰ (۱۲۷نمونه)	٪۸۵ (۶۰نمونه)	٪۹۴,۹۲ (۱۷۷نمونه)
خوشحالی	٪۹۴,۳۷ (۷۱نمونه)	٪۶۸,۳۳ (۶۰نمونه)	٪۸۱,۱۸ (۸۵نمونه)
ترس	٪۹۵,۶۵ (۶۹نمونه)	٪۷۵ (۶۰نمونه)	٪۸۰,۹۵ (۶۳نمونه)
تنفر	٪۹۷,۸۳ (۴۶نمونه)	٪۸۱,۶۷ (۶۰نمونه)	٪۸۲,۳۵ (۱۷نمونه)
تعجب	—	٪۷۱,۶۷ (۶۰نمونه)	٪۸۵ (۴۰نمونه)
خستگی	٪۱۰۰ (۸۱نمونه)	—	٪۸۲,۷۶ (۲۹نمونه)
خنثی	٪۹۷,۴۷ (۷۹نمونه)	٪۹۵ (۶۰نمونه)	٪۹۹,۴۸ (۱۹۴نمونه)

[۱] Average Accuracy
[۲] Average Precision
[۳] Weighted Average (WA) Recall
[۴] Unweighted Average (UA) Recall

در ادامه برای ارزیابی روش استخراج ویژگی پیشنهادی، سیستم تشخیص احساس از گفتار را با استفاده از پنج مجموعه ویژگی متداول شامل ویژگی‌های عروضی (Pitch, Energy, Duration)، ویژگی‌های طیفی (MFCC, LPCC, Formant)، ویژگی‌های کیفی (Jitter, Shimmer, HNR)، ویژگی‌های مبتنی‌بر انرژی TEO و ویژگی‌های مستخرج از ضرایب فوریه پیاده‌سازی کردیم. جدول ۴-۹ و شکل ۴-۳ عملکرد روش استخراج ویژگی پیشنهادی را در مقایسه با این پنج مجموعه ویژگی نشان می‌دهند. همان‌طور که از نتایج آزمایش‌ها در شکل ۴-۳ مشاهد می‌شود روش استخراج ویژگی پیشنهادی بالاترین دقت تشخیص را در بین سایر روش‌های استخراج ویژگی در هر سه مجموعه دادگان دارد. این جدول نشان می‌دهد که سایر روش‌های استخراج ویژگی نیز روی مجموعه دادگان SAVEE نسبت به دو مجموعه دادگان دیگر کمترین میزان دقت تشخیص را دارند. همان‌طور که بیان شد این مساله از تعداد کم و تنوع پایین گویندگان و نمونه‌های کلاس‌های احساسی در این پایگاه داده ناشی شده است. با توجه به شکل ۴-۳، در مجموعه دادگان EMO-DB بعد از روش پیشنهادی استخراج ویژگی با دقت تشخیص ۹۷,۵۷٪، مجموعه ویژگی‌های طیفی و ضرایب مستخرج از فوریه دارای بالاترین نرخ تشخیص به‌ترتیب در حدود ۸۲,۵٪ و ۷۵,۱۴٪ هستند. از شکل می‌توان مشاهده نمود که در مجموعه دادگان SAVEE و PDREC نیز ویژگی پیشنهادی بالاترین میزان تشخیص به‌ترتیب با مقادیر ۸۰٪ و ۹۱,۴۷٪ را در بین سایر ویژگی‌ها دارد. در جدول ۴-۹ نیز عملکرد روش استخراج ویژگی پیشنهادی در مقایسه با سایر روش‌ها با استفاده از معیارهای ارزیابی کارایی مختلف بررسی شده است. نتایج نشان می‌دهند که چرخش سیگنال در فضای زمان-فرکانس با یک زاویه مناسب α در هر سه مجموعه دادگان منجر به دستیابی به ویژگی‌های موثر در طبقه‌بندی احساسات شده است. در دادگان EMO-DB استخراج این ویژگی پیشنهادی از سیگنال‌های گفتار احساسی استاندارد به‌علت تعداد مناسب و متوازن گویندگان و نمونه‌ها در هر کلاس، بالاترین دقت تشخیص را به‌دست آورده است.



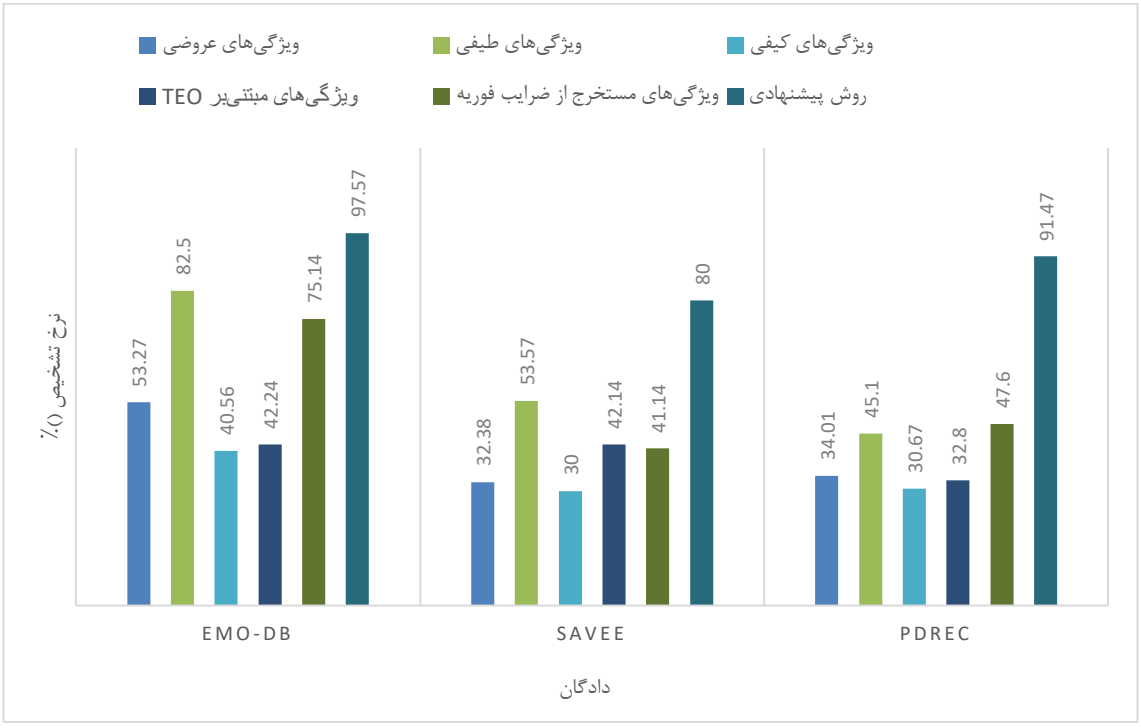
	غم	عصبانیت	خوشحالی	ترس	تنفر	تعجب	خستگی	خنثی
PDREC	91.72	94.92	81.18	80.95	82.35	85	82.76	99.48
SAVEE	83.33	85	68.33	75	81.67	71.67	0	95
EMO-DB	95.16	100	94.37	95.65	97.83	0	100	97.47

شکل ۴-۲ نرخ تشخیص و تعداد نمونه‌های هر کلاس در سه مجموعه دادگان (EMO-DB,SAVEE,PDREC) توسط سیستم پیشنهادی

جدول ٤-٩ مقایسه عملکرد روش پیشنهادی با سایر روش‌های استخراج ویژگی در سه مجموعه دادگان- متوسط صحت (Acc)-متوسط دقت (Prc)-میانگین وزنی صحت (WA)-میانگین بدون وزن صحت (UA)

دادگان/ عملکرد (٪)		عروضی	طیفی	کیفی	مبتنی‌بر TEO	مبتنی‌بر ضرایب فوریه	روش پیشنهادی
EMO-DB	Acc	۸۶,۶۵	۹۴,۹۰	۷۰,۳۰	۷۸,۸۰	۹۲,۹۰	۹۹,۳۱
	Prc	۵۲,۵۶	۸۲,۳۲	۳۰	۴۲	۷۱,۸۴	۹۷,۲۳
	WA	**۵۳,۲۷**	**۸۲,۵۰**	**۴۰,۵۶**	**۴۲,۲۴**	**۷۵,۱۴**	**۹۷,۵۷**
	UA	۴۸,۹۶	۸۱,۶۹	۴۰,۶۰	۴۲,۲۰	۷۱,۸۹	۹۷,۲۱
SAVEE	Acc	۸۰,۶۸	۸۴,۲۷	۶۸	۶۸,۴۰	۸۴,۹۰	۹۴,۲۹
	Prc	۳۱,۹۴	۵۵,۲۰	۲۹,۲۰	۳۰	۴۶,۸۴	۸۰,۰۸
	WA	**۳۲,۳۸**	**۵۳,۵۷**	**۲۹,۰۵**	**۳۰**	**۴۷,۱۴**	**۸۰**
	UA	۳۲,۳۸	۵۳,۵۷	۲۹,۰۵	۳۰	۴۷,۱۴	۸۰
PDREC	Acc	۵۸,۶۰	۸۶,۴۵	۶۳.۲۰	۶۳,۶۰	۸۸,۱۵	۹۷,۸۷
	Prc	۱۳,۵۶	۳۶,۳۳	۲۸	۲۸,۹۰	۳۷,۶۶	۸۹,۹۸
	WA	**۳۴,۰۱**	**۴۵,۱۰**	**۳۲**	**۳۲,۸۰**	**۴۷,۶۰**	**۹۱,۴۷**
	UA	۱۷	۳۳,۶۴	۳۲	۳۲,۸۰	۳۵,۸۹	۸۷,۳۱

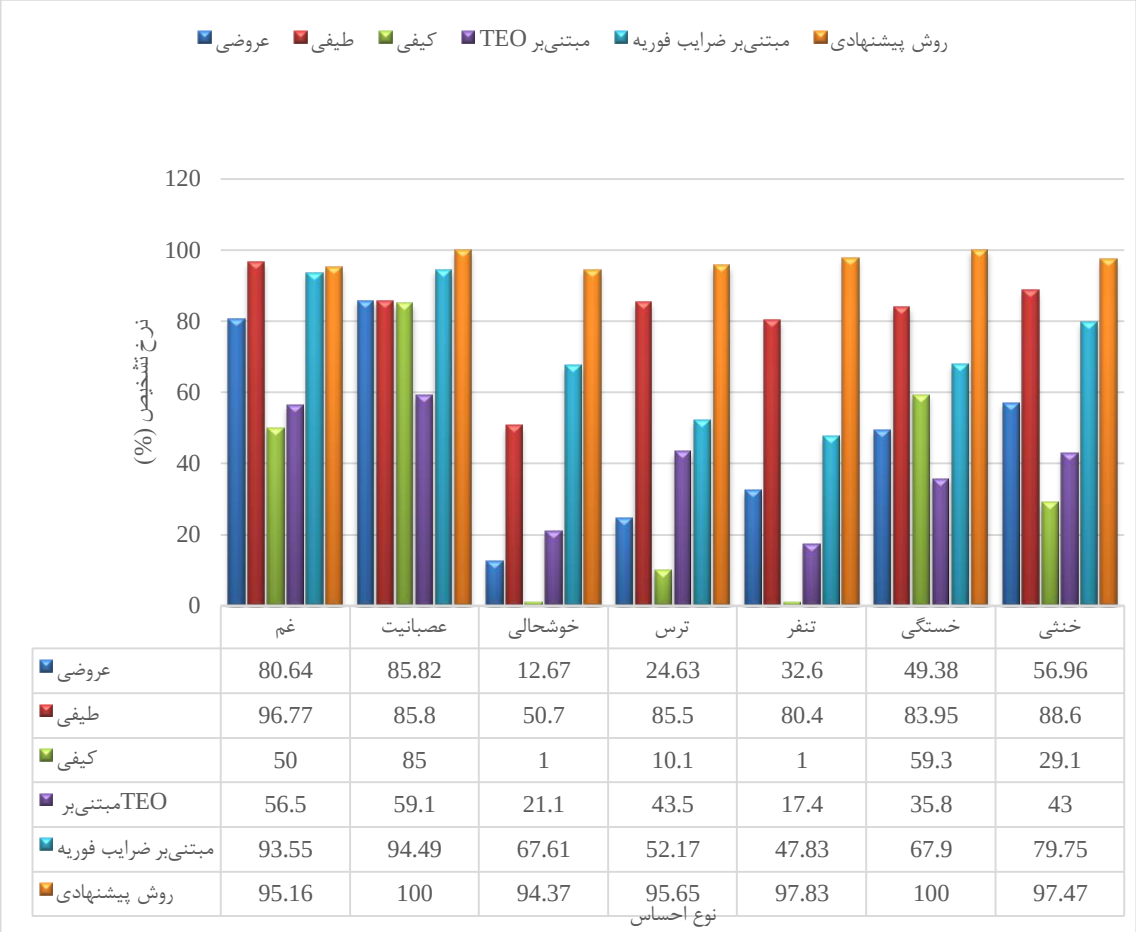
در آخرین مرحله از این آزمایش‌ها به بررسی تاثیر روش‌های استخراج ویژگی در تشخیص هر کلاس در هر مجموعه دادگان می‌پردازیم. در شکل ۴-۴ الی شکل ۴-۶ عملکرد روش استخراج ویژگی پیشنهادی برای تشخیص هر کلاس در سه مجموعه دادگان موردنظر با سایر روش‌های استخراج ویژگی مقایسه شده است. همان‌طور که در شکل ۴-۴ مشخص است در مجموعه دادگان EMO-DB ویژگی پیشنهادی در تشخیص شش احساس: خوشحالی، عصبانیت، ترس، تنفر، خستگی و خنثی نسبت به سایر مجموعه ویژگی‌ها بهتر عمل می‌کند. در حالی‌که دقت تشخیص کلاس احساسی غم با استفاده از مجموعه ویژگی طیفی در حدود ۱٫۵ درصد بهتر از تشخیص احساس غم با استفاده از ویژگی پیشنهادی است.



شکل ۴-۳ مقایسه عملکرد روش پیشنهادی با سایر روش‌های استخراج ویژگی در سه مجموعه دادگان (EMO-DB,SAVEE,PDREC)

همچنین همان‌طور که در شکل ۴-۵ نشان داده شده است، در مجموعه دادگان SAVEE ویژگی پیشنهادی برای دسته‌بندی تمام کلاس‌های احساسی به‌طور متوسط ۱٫۵ تا ۲ برابر بهتر از سایر روش‌ها عمل می‌کند. در مجموعه دادگان PDREC نیز با استفاده از روش استخراج ویژگی پیشنهادی، تمام هشت کلاس احساسی با بالاترین دقت دسته‌بندی شده‌اند. در این مجموعه دادگان نیز دقت روش

پیشنهادی بیش از دو برابر دقت سایر روش‌های استخراج ویژگی در تشخیص تمام کلاس‌های احساسی است.

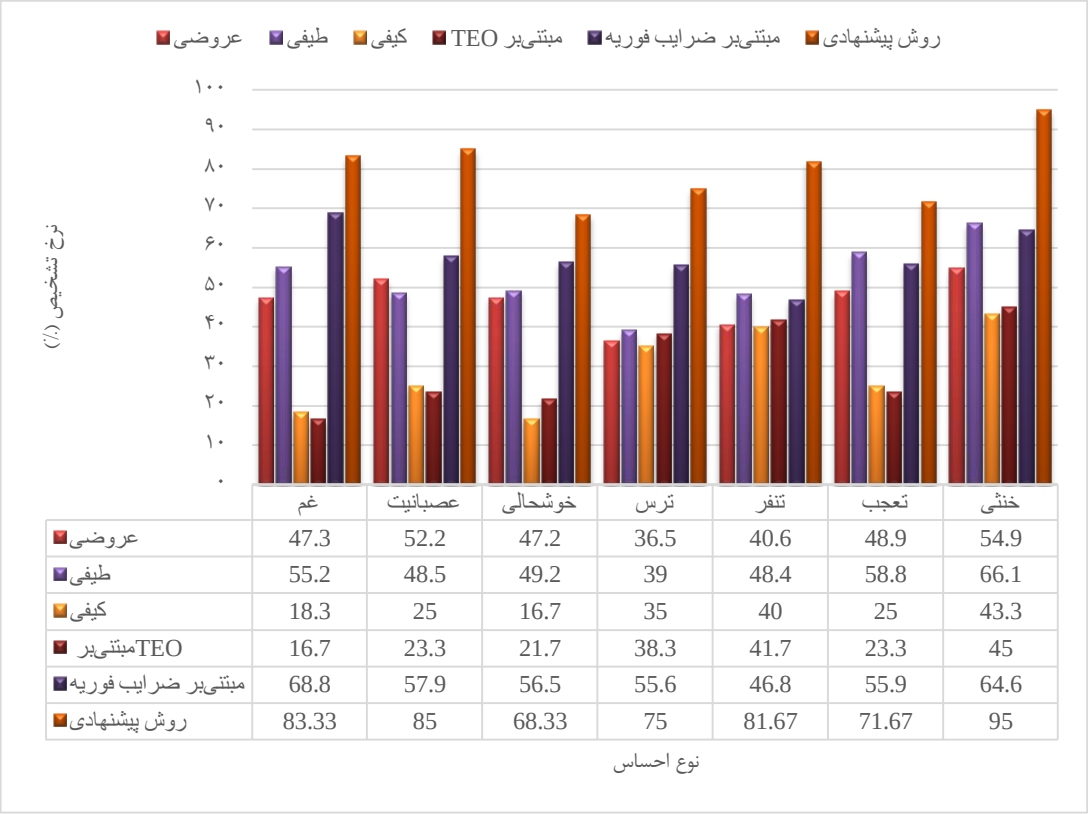


	غم	عصبانیت	خوشحالی	ترس	تنفر	خستگی	خنثی
عروضی	80.64	85.82	12.67	24.63	32.6	49.38	56.96
طیفی	96.77	85.8	50.7	85.5	80.4	83.95	88.6
کیفی	50	85	1	10.1	1	59.3	29.1
مبتنی‌برTEO	56.5	59.1	21.1	43.5	17.4	35.8	43
مبتنی‌بر ضرایب فوریه	93.55	94.49	67.61	52.17	47.83	67.9	79.75
روش پیشنهادی	95.16	100	94.37	95.65	97.83	100	97.47

شکل ۴-۴ مقایسه عملکرد روش پیشنهادی با سایر روش‌های استخراج ویژگی در دادگان EMO-DB

نمودارهای ارائه شده در شکل‌های ۴-۴ الی ۴-۶ نشان می‌دهند که روش‌های استخراج ویژگی طیفی و کپسترال در تشخیص تمام کلاس‌های احساسی بهتر از روش‌های عروضی، کیفی و مبتنی‌بر TEO عمل می‌کنند. اکثر کارهای حوزه تشخیص احساس از گفتار که روی مجموعه دادگان متفاوتی انجام شده‌اند، نیز به تاثیر بالای ویژگی‌های کپسترال و طیفی مانند MFCC و فرمنت‌ها در تشخیص احساس اشاره نموده‌اند. از آن‌جا که ضرایب کپسترال یک فریم گفتار با گرفتن تبدیل فوریه روی طیف اندازه لگاریتم به دست می آید، از ویژگی‌های مشتق‌شده از آن‌ها برای مدل کردن الگوی زیر و بمی و فرکانس گام گفتار یک گوینده استفاده می‌شوند. تغییرات در فرکانس گام همبستگی مهمی با احساسات در گفتار دارد و همچنین در تحقیقات اثبات شده که این تغییرات منجر به تغییرات در سایر پارامترهای عروضی مانند طول و انرژی می‌شود. در واقع ویژگی‌های طیفی نشان‌دهنده اطلاعات آوایی هستند که

مستقیما از طیف سیگنال گفتار مشتق شده‌اند. این ویژگی‌های مشتق‌شده از طیف، با استفاده از بانک‌های فیلتر، به سهم برابر و یکسان تمام مولفه‌های فرکانسی در پردازش یک سیگنال گفتار تاکید دارند.

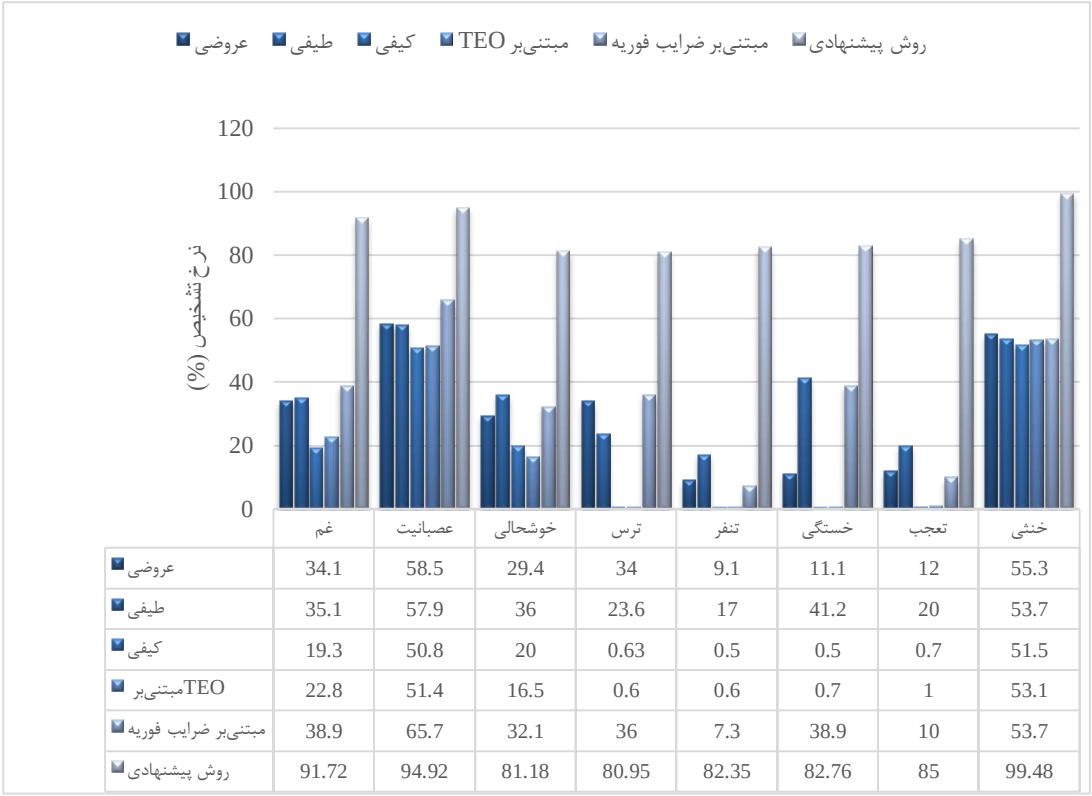


	غم	عصبانیت	خوشحالی	ترس	تنفر	تعجب	خنثی
عروضی	47.3	52.2	47.2	36.5	40.6	48.9	54.9
طیفی	55.2	48.5	49.2	39	48.4	58.8	66.1
کیفی	18.3	25	16.7	35	40	25	43.3
مبتنی‌بر TEO	16.7	23.3	21.7	38.3	41.7	23.3	45
مبتنی‌بر ضرایب فوریه	68.8	57.9	56.5	55.6	46.8	55.9	64.6
روش پیشنهادی	83.33	85	68.33	75	81.67	71.67	95

شکل ۴-۵ مقایسه عملکرد روش پیشنهادی با سایر روش‌های استخراج ویژگی در دادگان SAVEE

نتایج به‌دست آمده در نمودار شکل ۴-۶ نشان می‌دهد که روش‌های استخراج ویژگی مبتنی‌برTEO و ویژگی‌های کیفی در تشخیص نمونه‌های ۵۰٪ کلاس‌های احساسی مجموعه دادگان PDREC یعنی کلاس‌های احساسی تعجب، خستگی، تنفر و ترس کاملا ناتوان هستند. در حالی‌که ویژگی پیشنهادی این رساله به‌طور متوسط در حدود ۸۳٪ نمونه‌های این کلاس‌ها را درست تشخیص داده است. نتایج نشان می‌دهند که با وجود اینکه پایگاه داده PDREC از سخنان بیان‌شده به‌صورت زنده در رادیو نمایش تهیه شده است و نویز محیط و خطا در برچسب‌گذاری انسانی می‌تواند در نتیجه تشخیص تاثیر بگذارد، اما چرخش مناسب این سیگنال‌ها در فضای فرکانس-زمان توسط یک زاویه مناسب منجر به افزایش

دقت تشخیص شده است.



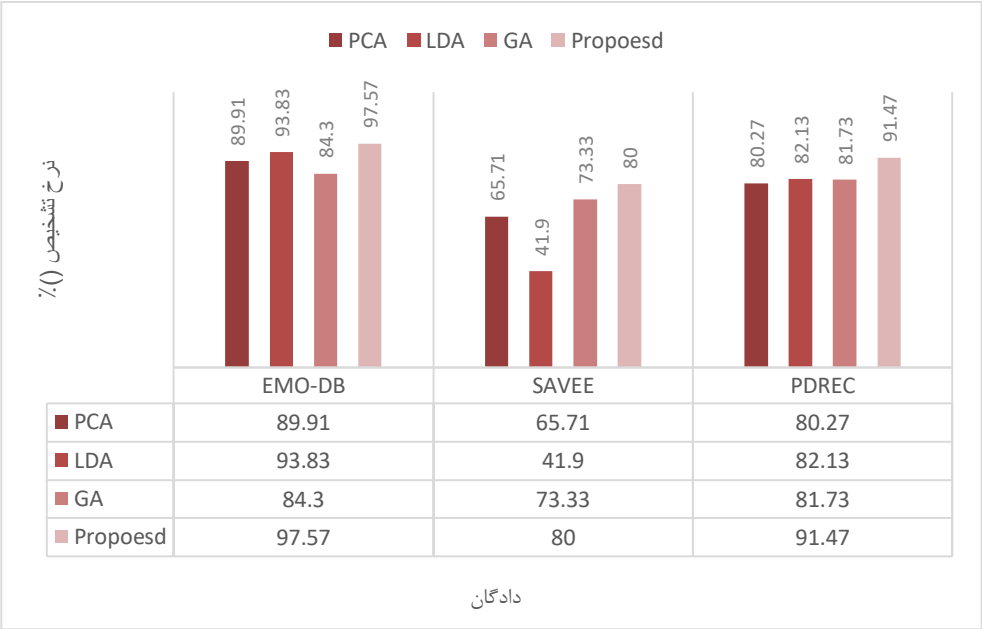
	غم	عصبانیت	خوشحالی	ترس	تنفر	خستگی	تعجب	خنثی
عروضی	34.1	58.5	29.4	34	9.1	11.1	12	55.3
طیفی	35.1	57.9	36	23.6	17	41.2	20	53.7
کیفی	19.3	50.8	20	0.63	0.5	0.5	0.7	51.5
مبتنی‌بر TEO	22.8	51.4	16.5	0.6	0.6	0.7	1	53.1
مبتنی‌بر ضرایب فوریه	38.9	65.7	32.1	36	7.3	38.9	10	53.7
روش پیشنهادی	91.72	94.92	81.18	80.95	82.35	82.76	85	99.48

شکل ۴-۶ مقایسه عملکرد روش پیشنهادی با سایر روش‌های استخراج ویژگی در دادگان PDREC

۴-۳-۳ آزمایش دوم: ارزیابی و تجزیه و تحلیل نتایج روش انتخاب ویژگی پیشنهادی در مقایسه با روش‌های انتخاب ویژگی

در آزمایش دوم برای ارزیابی عملکرد روش انتخاب ویژگی پیشنهادی، بعد از بررسی و آزمایش روش‌های مختلف انتخاب ویژگی، یک سیستم تشخیص احساس با استفاده از مجموعه ویژگی پیشنهادی و سه روش انتخاب ویژگی متداول تحلیل مؤلفه اصلی (PCA)، تحلیل مجزاساز خطی فیشر (LDA) و الگوریتم ژنتیک (GA) پیاده‌سازی و آزمایش شد. نتایج این آزمایش‌ها در شکل ۴-۷ گزارش شده است. همان‌طور که در شکل مشخص است در هر سه مجموعه دادگان، روش انتخاب ویژگی پیشنهادی دقیق‌تر عمل می‌کند. علت این امر ترکیب جستجوی عمومی و محلی در روش پیشنهادی انتخاب ویژگی است. این

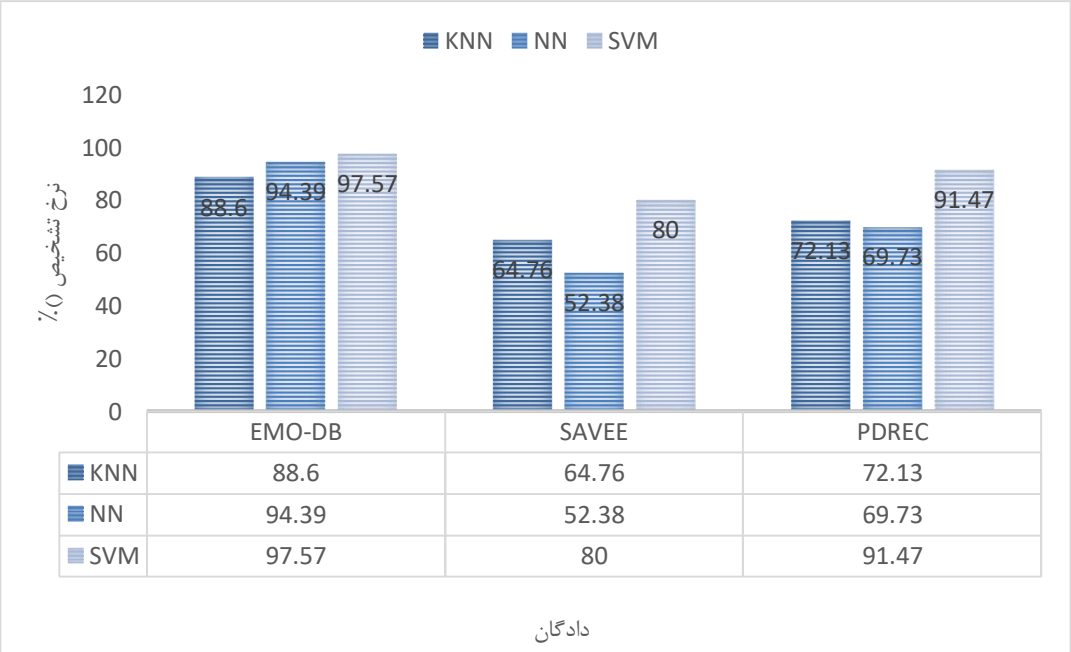
روش با استفاده از الگوریتم ژنتیک فضای جستجو را کاوش می‌کند تا محتمل‌ترین و امیدوارکننده‌ترین منطقه در فضای جستجو را جدا کند. در مرحله دوم، جستجوی فاخته برای بهبود جستجوی سراسری و جلوگیری از گرفتار شدن در بهینه‌های محلی فضای جستجوی محدودشده توسط GA را بسط داده و راه حل های جدید بهتری را می‌یابد. در مجموعه دادگان EMO-DB بعد از روش پیشنهادی با ۹۷,۵۷٪ دقت، روش LDA با دقت ۹۳,۸۳٪ عملکرد خوبی دارد. در مجموعه دادگان PDREC نیز بعد از روش انتخاب ویژگی پیشنهادی با دقت ۹۱,۴۷٪، روش LDA با دقت ۸۲,۱۳٪ بهتر از دو روش دیگر عمل می‌کند. اما با توجه به نتایج شکل ۴-۷، بعد از روش پیشنهادی با ۸۰٪ دقت در دادگان SAVEE، الگوریتم ژنتیک با دقت ۷۳,۳۳٪ بهتر از دو روش دیگر است.



شکل ۴-۷ مقایسه عملکرد روش پیشنهادی انتخاب ویژگی با سایر روش‌های انتخاب ویژگی در سه مجموعه دادگان (EMO-DB, SAVEE, PDREC)

۴-۳-۴ آزمایش سوم: ارزیابی و تجزیه و تحلیل نتایج روش‌های مختلف طبقه‌بندی

در آزمایش سوم برای بهبود عملکرد روش پیشنهادی تشخیص احساس از گفتار، طبقه‌بندهای مختلفی بررسی شد. بعد از بررسی و آزمایش روش‌های مختلف طبقه‌بندی، نتایج سیستم تشخیص احساس پیشنهادی با استفاده از سه طبقه‌بند k-نزدیک‌ترین همسایه (KNN)، شبکه عصبی پرسپترون چندلایه و ماشین بردار پشتیبان (SVM) در شکل ۴-۸ ارائه شده است. همان‌طور که در شکل مشاهده می‌شود طبقه‌بند SVM در هر سه مجموعه دادگان با دقت بیشتری نسبت به دو طبقه بند دیگر عمل دسته‌بندی را انجام می‌دهد. این مساله ناشی از انتخاب تابع هسته مناسب پایه شعاعی گوسی در طبقه‌بند SVM است که به علت جنس داده‌های آموزشی در سیستم تشخیص احساس از گفتار بر افزایش دقت تشخیص موثر است.



شکل ۴-۸ مقایسه عملکرد روش پیشنهادی با استفاده از طبقه‌بندهای مختلف در سه مجموعه دادگان (EMO-DB, SAVEE, PDREC)

۴-۳-۵ ارزیابی کارایی عملکرد روش پیشنهادی در مقایسه با سایر تحقیقات

جهت ارزیابی عملکرد سیستم تشخیص احساس از گفتار پیشنهادی این رساله، در بخش‌های قبلی چندین نوع آزمایش در سه بخش کلیدی SER انجام شد. همان‌طور که نتایج ارائه شده در نمودارها و جداول نشان دادند، روش پیشنهادی بهترین عملکرد را در هر سه مجموعه دادگان EMO-DB، مجموعه SAVEE و پایگاه داده PDREC دارد. در ادامه علاوه‌بر این آزمایش‌ها، برای اعتبارسنجی و امکان‌سنجی معماری پیشنهادی این رساله، عملکرد روش پیشنهادی را با نتایج گزارش شده توسط سایر مطالعات انجام‌شده در حوزه تشخیص احساس از گفتار مقایسه و بررسی نمودیم. در جدول (۴-۱۰) مقایسه عملکرد سیستم تشخیص احساس از گفتار پیشنهادی این رساله با نتایج گزارش شده در چند تحقیق به‌روز با استفاده از معیار دقت ارائه شده است. نتایج جدول نشان می‌دهند که دقت تشخیص با استفاده از روش پیشنهادی در مقایسه با مطالعه اوزسون [۶۰]، که از یک روش انتخاب ویژگی آماری جدید مبتنی‌بر تغییرات احساسات در ویژگی‌های صوتی استفاده می‌کند، برای EMO-DB حدود ۱۳٪ و برای مجموعه دادگان SAVEE در حدود ۷.۵٪ افزایش یافته است. همچنین جدول (۴-۱۰) بهبود دقت تشخیص در PDREC را در مقایسه با سایر مطالعاتی نشان می‌دهد که روی این مجموعه دادگان انجام شده [۱۶] [۸۶] و در آن‌ها به‌ترتیب از ویژگی‌های استخراج شده از ضرایب تبدیل کسینوسی گسسته واریوگرام و ویژگی‌های متداول طیفی و عروضی مورد بحث در بخش‌های قبلی استفاده شده است. همان‌طور که در جدول (۴-۱۰) نشان‌داده شده است، نتایج مطالعه مائو و همکاران [۵۷] دقیق‌تر از نتایج SER پیشنهادی این رساله در مجموعه دادگان SAVEE است، اما در EMO-DB روش پیشنهادی تشخیص احساس از گفتار این رساله نتایج بهتری به‌دست آورده است. این مطالعه از شبکه عصبی کانولوشن برای یک مدل تشخیص احساسات دو مرحله‌ای استفاده کرده است که در مرحله اول، یادگیری ویژگی‌های ثابت محلی با استفاده از یک رمزگذار خودکار پراکنده برای یک طیف‌سنج گفتار انجام شده

است و در مرحله بعدی ویژگی‌های متمایز با استفاده از PCA استخراج شده تا تشخیص را بهبود بخشد. Issa و همکارانش در سال ۲۰۲۰ یک سیستم تشخیص احساس از گفتار با استفاده از یک شبکه عصبی کانولوشنال یک بعدی و ویژگی‌های طیفی و کپسترال پیشنهاد داده‌اند. این سیستم از از یک روش افزایشی برای اصلاح مدل اولیه برای بهبود دقت طبقه‌بندی مجموعه دادگان EMO-DB استفاده کرده است. بدین صورت که در یک مرحله با ۵۳۵ نمونه از این پایگاه داده به دقت ۸۶٫۱٪ دست یافته است، و در آزمایشی دیگر دقت تشخیص را با ۵۲۰ نمونه از این مجموعه ۹۵٫۷۱٪ گزارش کرده است.

یکی دیگر از مطالعاتی که روی مجموعه دادگان EMO-DB و SAVEE توسط یوگش و همکارانش انجام شده [۸۸] از یک الگوریتم بهینه‌سازی ذرات PSO نوین مبتنی‌بر الگوریتم بیوژئوگرافی برای انتخاب ویژگی روی دو دسته ویژگی طیفی استفاده کرده است. مطابق جدول، روش پیشنهادی این رساله در هر دو مجموعه دادگان دقیق‌تر از عملکرد روش ارائه شده در این تحقیق است. همچنین، جدول(۴-۱۰) نشان می‌دهد که نتایج روش پیشنهادی در مجموعه داده‌های EMO-DB دقیق‌تر از نتایج مطالعه انجام شده توسط کوچیبوتلا و همکاران است [۸۷]، که یک روش انتخاب ویژگی بهینه دو مرحله‌ای برای SER را با استفاده از ویژگی‌های آکوستیک (Pitch, Energy, MFCC) معرفی می‌کند. به‌طورکلی، مطابق جدول (۴-۱۰)، روش پیشنهادی تشخیص احساس از گفتار این رساله در مقایسه با سایر مطالعات دارای بالاترین میزان دقت تشخیص است. سیستم پیشنهادی تشخیص احساس از گفتار این رساله با استفاده از چرخش سیگنال در یک فضای زمان-فرکانسی با یک زاویه مناسب و ترکیب ویژگی‌های استخراج شده در این فضا با ویژگی‌های کپسترال MFCC، و انتخاب یک مجموعه ویژگی موثر با تعداد مناسب از این ترکیب به بالاترین میزان دقت تشخیص کلاس‌های مختلف احساسی در مجموعه دادگان با زبان‌های مختلف، تعداد و تنوع گویندگان مختلف دست یافته است.

جدول ۴-۱۰ مقایسه عملکرد سیستم پیشنهادی تشخیص احساس از گفتار با برخی از تحقیقات گذشته

دادگان	تحقیق	نرخ تشخیص (٪)
EMO-DB	Mao et al. (2014) [57]	88.30
	Kuchibhotla et al. (2016) [87]	92.70
	Yogesh et al. (2017) [88]	97.54
	Özseven (2019) [60]	84.62
	Issa (2020) [91]	86.1
	Proposed	97.57
SAVEE	Mao et al. (2014) [57]	86.70
	Yogesh et al. (2017) [88]	78.44
	Liu et al. (2018b) [59]	76.40
	Özseven (2019) [60]	72.39
	Proposed	80
PDREC	Esmaileyan and Marvi (2014a) [16]	51.51
	Esmaileyan and Marvi (2014b) [86]	60.28
	Proposed	91.47

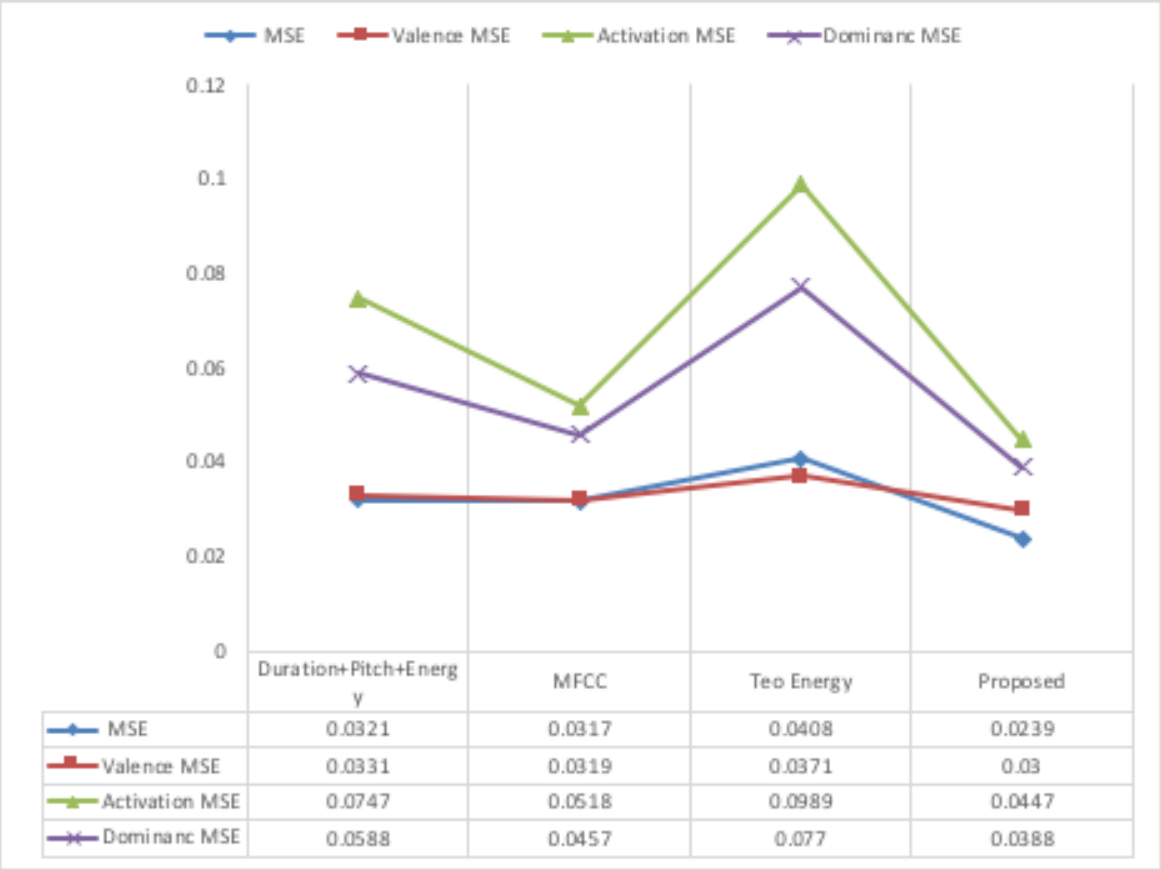
۴-۴ ارزیابی فاز دوم: سیستم تشخیص احساس از گفتار مبتنی‌بر نگاشت مجموعه ویژگی‌ها به فضای چند بعدی احساس

۴-۴-۱ ارزیابی نگاشت

در آخرین مرحله از اعتبارسنجی جهت ارزیابی مدل پیشنهادی نگاشت در فاز دوم، در گام اول بعد از انجام مراحل پیش‌پردازش روی نمونه‌های مجموعه دادگان VAM (مطابق مراحل بیان شده در بخش ۳-۲-۱)، با استفاده از روش استخراج ویژگی پیشنهادشده (بخش ۳-۲-۲) یک مجموعه ویژگی از این نمونه‌ها جهت آموزش مدل یادگیر استخراج شد. برای مدل یادگیر بعد از آزمایش‌های متعدد بهترین نوع مدل یک شبکه عصبی با ۱۵ لایه پنهان درنظر گرفته شد. همچنین قبل از آموزش شبکه با این مجموعه ویژگی، ابتدا با استفاده از روش PCA تعداد ویژگی‌ها به ۵۰ عدد کاهش یافت. در این آزمایش‌ها

علاوه‌بر مجموعه ویژگی پیشنهادی رساله در فاز اول، جهت ارزیابی دقت نگاشت چهار دسته ویژگی دیگر شامل ویژگی‌های عروضی، طیفی، کیفی و ویژگی‌های مبتنی‌بر انرژی TEO نیز از این نمونه‌ها استخراج شد. نتایج حاصل از این آزمایش‌ها در شکل ۴-۹ و شکل ۴-۱۰ نمایش داده شده است. با توجه به شکل ۴-۹ مقدار متوسط خطای تخمین هر بعد با استفاده از ویژگی‌های پیشنهادی کمتر از سایر روش‌ها است. همچنین شکل ۴-۱۰ نشان می‌دهد که مقادیر واقعی ابعاد دارای بیشترین میزان همبستگی با مقادیر تخمین‌زده شده توسط ویژگی‌های پیشنهادی نسبت به سایر روش‌ها است.

در ادامه برای نگاشت بین فضای ویژگی‌های آکوستیک و فضای سه بعدی احساسی، با استفاده از مدل آموزش‌یافته توسط ویژگی‌های پیشنهادی مستخرج از مجموعه دادگان VAM، برای نمونه‌های پایگاه داده EMO-DB مقادیر سه بعد (Valence,Activation,Dominance) تخمین زده می‌شوند. جهت ارزیابی این نگاشت مطابق شکل ۴-۱۱ عملکرد روش پیشنهادی با سایر روش‌های استخراج ویژگی مقایسه شده است. همان‌طور که در شکل مشخص است روش پیشنهادی بالاترین دقت را در مقایسه با سایر روش‌ها دارد.

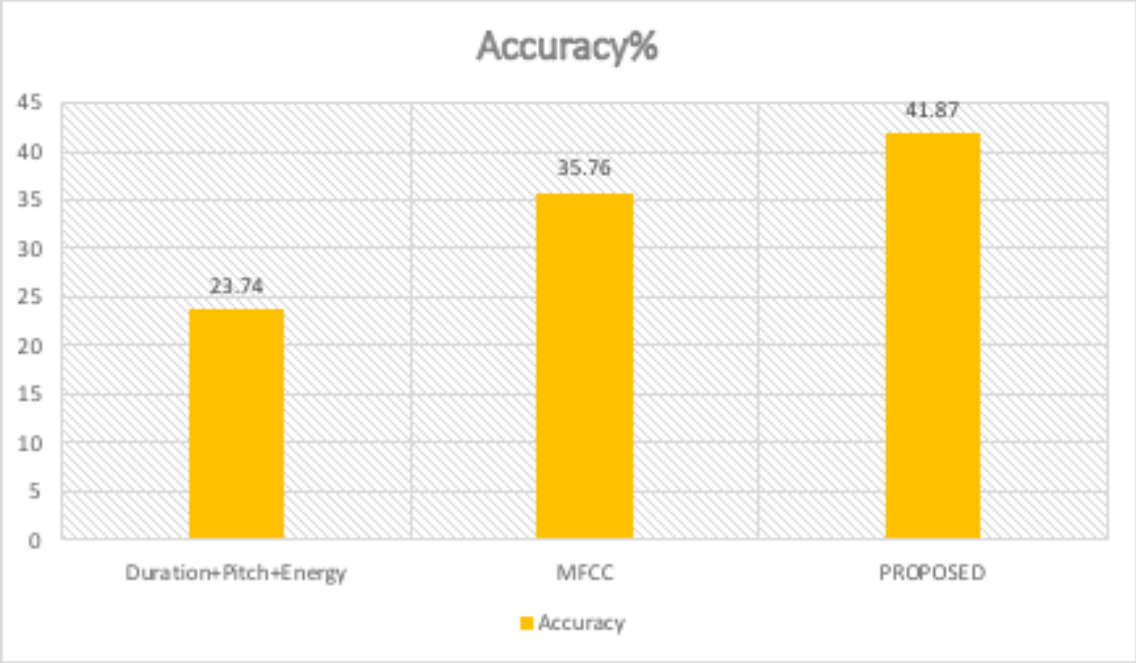


	Duration+Pitch+Energy	MFCC	Teo Energy	Proposed
MSE	0.0321	0.0317	0.0408	0.0239
Valence MSE	0.0331	0.0319	0.0371	0.03
Activation MSE	0.0747	0.0518	0.0989	0.0447
Dominanc MSE	0.0588	0.0457	0.077	0.0388

شکل ۴-۹ ارزیابی میزان خطای مقادیر تخمینی سه بعد بااستفاده از ویژگی‌های مختلف

	Duration+Pitch+Energy	MFCC	Jitter+Shimmer+HNR	Teo Energy	Proposed
Valence CORR	0.3817	0.4365	0.0728	0.1041	0.3981
Activation CORR	0.6174	0.7441	0.04813	0.422	0.7893
Dominanc CORR	0.5867	0.7098	-0.0306	0.3768	0.7568

شکل ۴-۱۰ ارزیابی مقدار همبستگی مقادیر تخمینی سه بعد بااستفاده از ویژگی‌های مختلف

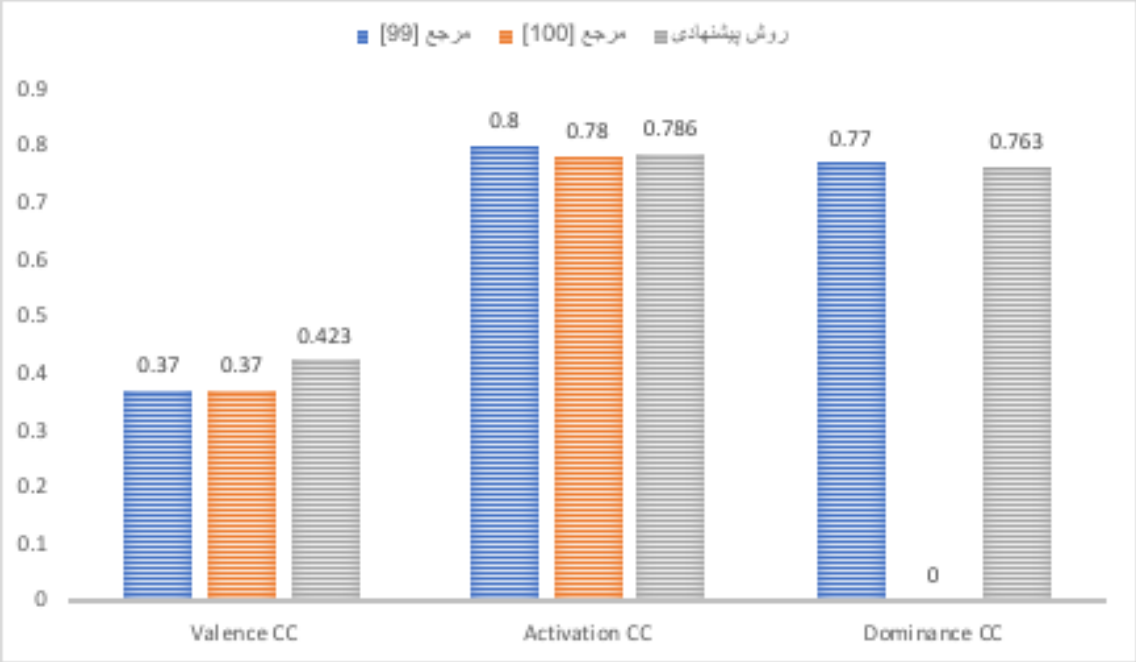


شکل ۴-۱۱ ارزیابی عملکرد روش پیشنهادی در نگاشت به فضای سه بعدی احساس در مقایسه با سایر روش‌های استخراج ویژگی

۴-۴-۲ ارزیابی کارایی عملکرد نگاشت پیشنهادی در مقایسه با سایر تحقیقات

به‌طورکلی تحقیقات کمی در حوزه تشخیص احساس در فضای چند بعدی احساس انجام شده است. شکل ۴-۱۲ نتایج روش پیشنهادی را در مقایسه با نتایج ارائه شده توسط تحقیقاتی در سال‌های ۲۰۰۷ و ۲۰۱۸ برحسب معیار ارزیابی هم‌بستگی ارائه می‌دهد. در تحقیق اول از یک روش رگرسیون بردار پشتیبان برای تخمین مقادیر سه بعد (جاذبه، انگیختگی و تسلط) بااستفاده از ویژگی‌های گام، انرژی و MFCC و روش انتخاب ویژگی SFS استفاده شده است [103]. در تحقیق دوم برای شناخت احساسات در فضای دو بعدی (جاذبه، برانگیختگی) از یک رویکرد کدگذاری اسپارس سلسه مراتبی (SC) استفاده شده است [104]. در سیستم پیشنهادی این تحقیق درکنار ویژگی‌های طیفی و عروضی ویژگی‌های سطح بالای

ادراکی مانند معیارهای مبتنی بر زیروبمی صدا و معیارهای برگرفته از مدل شنوایی انسان نیز استفاده شده است. این مجموعه ویژگی از مجموعه دادگان VAM استخراج شده وسپس بااستفاده از یک مدل دیکشنری کدگذار اسپارس ضرایبی از آن استخراج و جهت طبقه‌بندی با روش رگرسیون بردار پشتیبان (SVR) انتخاب شده‌اند. همان‌طور که از شکل ۴-۱۲ مشخص است مقدار همبستگی بعد جاذبه با استفاده از روش پیشنهادی این رساله به اندازه ۰،۰۵ بهتر از روش مراجع موردبحث است. اما مقدار همبستگی بعد انگیختگی تغییر چندانی نداشته است. به‌علاوه روش پیشنهادی این رساله بعد سوم یعنی بعد تسلط را نیز با مقدار همبستگی نزدیک به مرجع اول تخمین می‌زند.

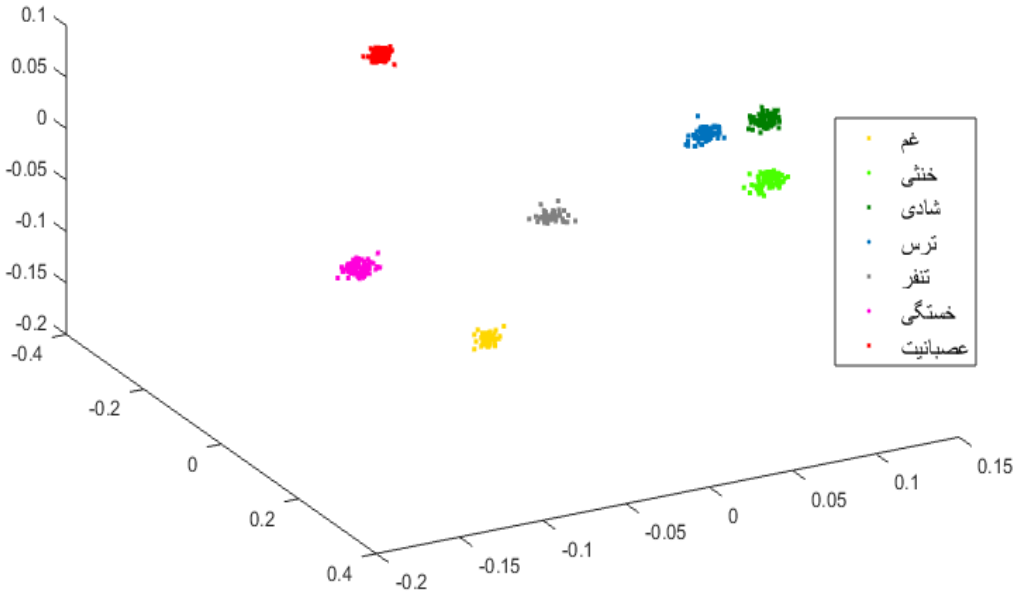


شکل ۴-۱۲ مقایسه روش نگاشت پیشنهادی با سایر تحقیقات

۴-۴-۳ ارزیابی عملکرد نگاشت در تفکیک کلاس‌های احساسی

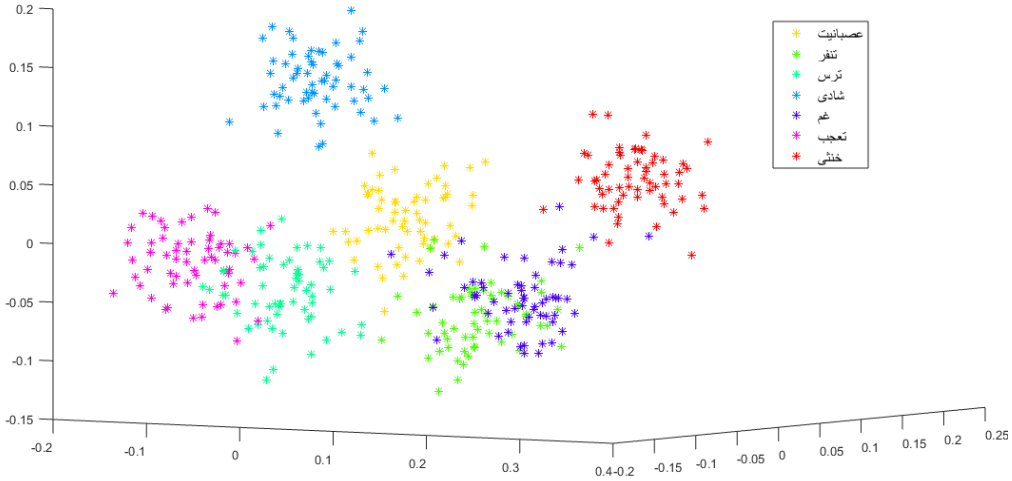
در یکی از آزمایش‌ها برای نمایش تأثیر نگاشت ویژگی‌های استخراج شده در سطح جداسازی هر کلاس در هر مجموعه دادگان، ابتدا فضای ویژگی پیشنهادی با استفاده از روش کاهش ابعاد تحلیل مجزاساز

تعیم‌یافته[1] GDA به یک فضای با ابعاد کوچک نگاشت شد. در این روش، فضای ویژگی اصلی به فضایی با ۳ بعد نگاشت شده و هر یک جداگانه ترسیم شده‌اند (شکل ۴-۱۴ و شکل ۴-۱۵و شکل ۴-۱۵). همان‌طور که در شکل‌ها مشاهده می‌شود، فضای ویژگی پیشنهادی در هر ۳ کلاس قابلیت تفکیک بالایی دارد. مطابق شکل‌ها ویژگی پیشنهادی در پایگاه داده EMO-DB دارای جداسازی بیشتری در کلاس‌ها نسبت به کلاس‌های مجموعه دادگان SAVEE و PDREC است. این تصاویر تا حدودی نشان می‌دهند که داده‌های EMO-DB قابلیت تفکیک بیشتری نسبت به دو مجموعه دادگان دیگر دارند.

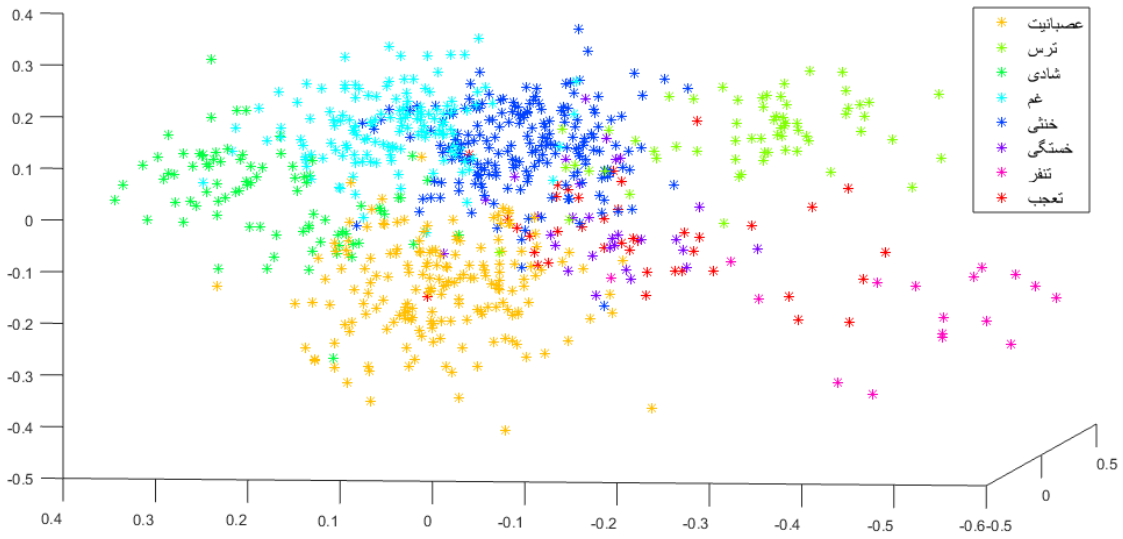


شکل ۴-۱۳ تفکیک‌پذیری کلاس‌های احساسی بااستفاده از ویژگی پیشنهادی در EMO-DB

[1] Generalized Discriminant Analysis



شکل ۴-۱۴ تفکیک‌پذیری کلاس‌های احساسی بااستفاده از ویژگی پیشنهادی در SAVEE



شکل ۴-۱۵ تفکیک‌پذیری کلاس‌های احساسی بااستفاده از ویژگی پیشنهادی در PDREC

فصل ۵ نتیجه گیری و پیشنهادات آتی

۵-۱ بحث و جمع‌بندی

در این رساله در راستای بهبود تعامل انسان و ماشین در کاربردهای روزمره و با توجه به اهمیت برقراری این ارتباط از طریق گفتار، یک سیستم تشخیص احساس از گفتار بهبودیافته پیشنهاد شد. همان‌طور که در مقدمه بیان شد در روان‌شناسی و علوم مرتبط نمایش و تعریف احساسات توسط دو مدل گسسته (برچسب‌گذاری احساسات با برچسب‌ها و دسته‌های مشخص و گسسته احساسی مانند خوشحالی، غم، عصبانیت و ...) و یا مدل پیوسته در فضای چند بعدی احساس (بیان احساسات در یک فضای پیوسته سه بعدی با ابعادی مانند میزان انگیختگی، میزان تسلط و جاذبه) انجام می‌شود. از این‌رو در این رساله مساله تشخیص احساس از گفتار در دو فاز تشخیص کلاس‌های احساسی گسسته بااستفاده از مجموعه ویژگی‌های مستخرج از سیگنال و سپس یافتن یک نگاشت بین این ویژگی‌ها و فضای سه‌بعدی احساس مدل شده است. همان‌طور که در فصل سوم ذکر شد، در فاز اول برای بهبود عملکرد تشخیص کلاس‌های احساسی یک روش استخراج ویژگی انطباقی فرکانس-زمان براساس تبدیل فوریه کسری پیشنهاد شد که مبتنی‌بر چرخش زاویه α در صفحه فرکانس-زمان است. انگیزه استفاده از DFrFT در استخراج ویژگی برای شناخت احساسات این بود که با چرخش زاویه α، سیگنال به فضایی نگاشت می‌شود که در آن سیگنال اصلی از داده‌های نویزی بازیابی می‌شوند که از این خصیصه می‌توان برای تفکیک اطلاعات مفید از اطلاعات نامرتبط سیگنال برای بهبود طبقه‌بندی استفاده کرد. براساس این نگاشت می‌توان دقت تشخیص احساسات را بهبود بخشید. از آن‌جا که درجه چرخش زاویه α بر نتیجه تشخیص تأثیر می‌گذارد، با آموزش و اعتبارسنجی یک مدل یادگیر به‌منظور افزایش دقت تشخیص، بهترین مقدار پارامتر α به‌دست آمد. همچنین از مطالعات و تحقیقات این مساله مشخص شد که از آن‌جا که ویژگی پیشنهادی اطلاعات دامنه فرکانس-زمان را استخراج می‌کند، با ترکیب این ویژگی‌ها با ضرایب MFCC که اطلاعات حوزه کپسترال را ارائه می‌دهند می‌توان مجموعه ویژگی مؤثرتری برای طبقه‌بندی به‌دست آورد. در

ادامه، با توجه به مساله کاهش ابعاد و انتخاب زیرمجموعه ویژگی بهینه در مباحث طبقه‌بندی، بااستفاده از یک الگوریتم ترکیبی انتخاب ویژگی GA-CS، بالاترین میزان تشخیص در هر سه مجموعه دادگان حاصل شد. در راستای ارزیابی عملکرد این سیستم پیشنهادی چندین نوع آزمایش بااستفاده از چهار پایگاه داده استاندارد شامل پایگاه داده گفتار احساسی به زبان آلمانی EMO-DB، مجموعه دادگان گفتار احساسی به زبان انگلیسی SAVEE، پایگاه داده جمع‌آوری شده از رادیو نمایش فارسی PDREC و مجموعه دادگان جمع‌آوری شده از یک برنامه تلویزیونی به زبان آلمانی VAM انجام شد. علاوه‌بر این، عملکرد معماری پیشنهادی با روش‌های استخراج و انتخاب ویژگی متداول نیز مقایسه شد تا اثربخشی SER پیشنهادشده تأیید شود. لازم به ذکر است که روش پیشنهادی در مقایسه با سایر روش‌ها، بالاترین میزان تشخیص را برای سه مجموعه دادگان با سه زبان مختلف (آلمانی، انگلیسی و فارسی) و همچنین تعداد و جنسیت متفاوت سخنرانان مختلف به‌دست آورده است. همچنین، با توجه به نتایج نشان‌داده شده در فصل ۴، SER پیشنهادی دارای بالاترین دقت برای شناسایی اکثر کلاس‌ها در هر سه مجموعه دادگان است. به‌علاوه، برای امکان‌سنجی و اعتبارسنجی سیستم تشخیص احساس از گفتار پیشنهادی، عملکرد آن با سایر مطالعات و تحقیقات انجام‌شده در سه مجموعه داده مقایسه شد که مطابق نتایج ارائه شده در فصل ۴، روش پیشنهادی در مقایسه با سایر مطالعات بالاترین میزان تشخیص را دارد.

در فاز دوم با استفاده از مجموعه ویژگی پیشنهادی مستخرج از مجموعه دادگان VAM سه بعد مدل سه بعدی احساسات یعنی جاذبه، انگیختگی و تسلط تخمین زده شد. برای ارزیابی این مقادیر تخمین زده شده، با استفاده از دادگان EMO-DB اعتبارسنجی و تست انجام شده است. نتایج به‌دست آمده در فصل ۴ نشان دادند که با استفاده از ویژگی‌های پیشنهادی، نگاشت بهتری نسبت به سایر ویژگی‌ها از فضای گسسته به فضای پیوسته خواهیم داشت.

همچنین یکی از نکات بررسی شده در این رساله مساله تفاوت نتایج به‌دست‌آمده در مجموعه دادگان مختلف است. دادگان EMO-DB از نظر تنوع گوینده و استانداردهای ضبط بسیار بهتر از مجموعه

دادگان SAVEE است. در EMO-DB، احساسات توسط ۱۰ بازیگر (۵ زن و ۵ مرد) شبیه‌سازی شده‌اند. این مجموعه داده شامل جملات آلمانی (۵ جمله کوتاه و ۵ جمله طولانی) است که به‌طور گسترده‌ای در ارتباطات روزمره استفاده می‌شوند. اما مجموعه دادگان SAVEE یک مجموعه داده صوتی و تصویری احساسی است که در آن چهار بازیگر انگلیسی در هر احساس ۱۵ جمله را بیان می‌کنند: ۳ جمله مشترک، ۲ جمله خاص برای احساسات و ۱۰ جمله کلی که برای هر احساس متفاوت است. در یکی از آزمایش‌ها برای نمایش تأثیر ویژگی‌های استخراج شده در سطح جداسازی هر کلاس در هر مجموعه دادگان، ابتدا فضای ویژگی اولیه با استفاده از روش کاهش ابعاد تحلیل مجزاساز تعیم‌یافته[۱] GDA به یک فضای با ابعاد کوچک نگاشت شد. . در این روش، فضای ویژگی اصلی به فضایی با ۳ بعد نگاشت شده و هر یک جداگانه ترسیم شده‌اند (شکل ۴-۱۳ و شکل ۴-۱۴). همان‌طور که در شکل‌ها مشاهده می‌شود، فضای ویژگی پیشنهادی در هر ۳ کلاس قابلیت تفکیک بالایی دارد. مطابق شکل‌ها ویژگی پیشنهادی در مجموعه دادگان EMO-DB دارای جداسازی بیشتری در کلاس‌ها نسبت به کلاس‌های SAVEE است. این تصاویر تا حدودی نشان می‌دهند که داده‌های EMO-DB قابلیت تفکیک بیشتری نسبت به دو مجموعه دادگان دیگر دارند.

۵-۲ پیشنهادات آتی

در این بخش براساس نتایج به‌دست آمده از مطالعات، شبیه‌سازی‌ها و آزمایش‌های متعدد انجام شده در این رساله چند پیشنهاد برای ادامه تحقیق مطابق با فهرست زیر ارائه می‌شود.

- ✓ یکی از مهم‌ترین چالش‌ها براس توسعه سیستم‌های تشخیص احساس از گفتار تولید مجموعه داده‌ی مناسب برای بهبود فرآیند یادگیری است. بیشتر مجموعه داده‌هایی که برای SER استفاده می‌شوند توسط گویندگان با احساسات بازی‌شده یا القاشده در اتاق‌های مخصوص و ساکت و

[۱] Generalized Discriminant Analysis

بدون نویز ضبط می‌شوند. درحالی‌که داده‌های واقعی زندگی نویزی هستند و ویژگی‌های متفاوتی نسبت به مجموعه دادگان تولیدشده دارند. اگرچه در این حوزه مجموعه داده‌های طبیعی نیز موجود است، اما تعداد آن‌ها کمتر است، و برای ثبت و استفاده از احساسات طبیعی مشکلات قانونی و اخلاقی وجود دارد. بیشتر سیگنال‌های گفتاری در مجموعه داده‌های طبیعی از برنامه‌های تلویزیونی و رادیویی، ضبط‌های مراکز تماس و موارد مشابه جمع‌آوری شده‌اند که طرفین درگیر از ضبط مطلع هستند. این مجموعه داده ها شامل همه احساسات نیستند و ممکن است احساساتی را که احساس می شود منعکس نکند. علاوه‌بر این ، برچسب‌گذاری این داده‌ها نیز ممکن است بادقت انجام نشود.

- ✓ یکی از مسائل مطرح و به‌روز در یادگیری ماشین کاربرد یادگیری عمیق در استخراج و طبقه-بندی ویژگی‌هاست. از آن‌جا که در روش استخراج ویژگی پیشنهادی این رساله بُعد زمان نیز در دسترس است، در آینده می‌توان از این ویژگی‌ها به‌عنوان ورودی شبکه‌های عمیق مانند LSTM و CNN استفاده کرد تا از قابلیت این شبکه‌ها برای دستیابی به ویژگی‌های سطح بالا بهره برد.
- ✓ یکی دیگر از مسائلی که می‌توان در آینده در مسائل تشخیص احساس از گفتار بررسی کرد، مساله بررسی تاثیر مشخصه‌های گوینده، و اثرات فرهنگی و زبانی روی SER است. در این راستا می‌توان پیاده‌سازی یک سیستم تشخیص احساس از گفتار مستقل از گوینده و زبان و فرهنگ را با تکنیک‌های Cross-Language بررسی نمود. البته درهم‌آمیختگی احساسات بر گفتار در بین زبان‌های مختلف ممکن است به‌عنوان یک چالش در این روش مطرح باشد.
- ✓ چالش دیگر، مساله سیگنال‌های گفتاری آمیخته است. در این حالت سیستم SER مجبور است تصمیم بگیرد که روی کدام سیگنال تمرکز کند. برای حل این چالش می‌توان از الگوریتم‌های تفکیک گفتار در مرحله پیش‌پردازش استفاده کرد.

مراجع

[1] S. Casale, A. Russo, and S. Serrano, "Multistyle classification of speech under stress using feature subset selection based on genetic algorithms," *Speech Commun.*, vol. 49, no. 10–11, pp. 801–810, 2007, doi: 10.1016/j.specom.2007.04.012.

[2] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digit. Signal Process. A Rev. J.*, vol. 22, no. 6, pp. 1154–1160, 2012, doi: 10.1016/j.dsp.2012.05.007.

[3] M. Valstar *et al.*, "AVEC 2014 - 3D dimensional affect and depression recognition challenge," *AVEC 2014 - Proc. 4th Int. Work. Audio/Visual Emot. Challenge, Work. MM 2014*, pp. 3–10, 2014, doi: 10.1145/2661806.2661807.

[4] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009, pp. 312–315.

[5] T. L. Nwe, S. W. Foo, and Liyanage C. De Silva., "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.

[6] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 148–152, 2013.

[7] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, no. October 2019, pp. 56–76, 2020, doi: 10.1016/j.specom.2019.12.001.

[8] A. Ben-Zeev, "The nature of emotions," *Philos. Stud.*, vol. 52, no. 3, pp. 393–409, 1987, doi: 10.1007/BF00354055.

[9] P. EKMAN, W. V. FRIESEN, and P. ELLSWORTH, "Emotion in the human face: Guidelines for research and an integration of findings," *Elsevier*, vol. 11, 2013, doi: 10.1016/b978-0-08-016643-8.50006-9.

[10] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977, doi: 10.1016/0092-6566(77)90037-X.

[11] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011, doi: 10.1016/j.patcog.2010.09.020.

[12] S. T. Sawale, V. P. Kshirsagar, D. Cse, G. M. E. Cse, G. College, and O. Engg, "Emotions and Strategies for Preparation of Emotional Speech Database," vol. 3, no. 1, pp. 42–47, 2010.

[13] C. Kotropoulos and D. Ververidis, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, 2006.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in *9th European Conference on Speech Communication and Technology*, 2005, pp. 1517–1520.

[15] S. Haq and P. J. B. Jackson, “Speaker-Dependent Audio-Visual Emotion Recognition,” in *Int’l Conf. on Auditory-Visual Speech Processing*, 2009, pp. 53–58.

[16] Z. Esmaileyan and H. Marvi, “A database for automatic persian Speech Emotion Recognition: Collection, processing and evaluation,” *Int. J. Eng. Trans. A Basics*, vol. 27, no. 1, pp. 79–90, 2014, doi: 10.5829/idosi.ije.2014.27.01a.11.

[17] M. Grimm, K. Kroschel, and S. Narayanan, “The Vera am Mittag German audio-visual emotional speech database,” *2008 IEEE Int. Conf. Multimed. Expo, ICME 2008 - Proc.*, pp. 865–868, 2008, doi: 10.1109/ICME.2008.4607572.

[18] I. S. Engberg, A. V Hansen, O. Andersen, and P. Dalsgaard, “Design, Recording and Verification of a Danish Emotional Speech Database,” *Proc. Eurospeech 1997*, vol. 4, pp. 1695–1698, 1997.

[19] W. Bao *et al.*, “Building a Chinese Natural Emotional Audio-Visual Database,” *Int. Conf. Signal Process. Proceedings, ICSP*, vol. 2015-Janua, no. October, pp. 583–587, 2014, doi: 10.1109/ICOSP.2014.7015071.

[20] C. Busso *et al.*, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

[21] A. Batliner, S. Steidl, and E. Nöth, “Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus.,” *Proc. Work. Corpora Res. Emot. Affect Lr.*, pp. 28–31, 2008.

[22] N. Keshtiari, M. Kuhlmann, M. Eslami, and G. Klann-Delius, “Erratum to Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD)[Behav Res, 10.3758/s13428-014-0467-x],” *Behav. Res. Methods*, vol. 47, no. 1, p. 295, 2015, doi: 10.3758/s13428-014-0504-9.

[23] M. H. Sedaaghi, “Documentation of the Sahand Emotional Speech Database,” *Electr. Eng.*, pp. 1–55, 2008.

[24] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, “Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database,” *Eurospeech*, pp. 1743–46, 1997.

[25] X. Mao, L. Chen, and B. Zhang, “Mandarin speech emotion recognition based on a hybrid of HMM/ANN,” *Int. J. Comput.*, vol. 1, no. 4, pp. 321–324, 2007.

[26] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, “IITKGP-SESC: Speech database for emotion analysis,” *Commun. Comput. Inf. Sci.*, vol. 40, pp. 485–492, 2009, doi: 10.1007/978-3-642-03547-0_46.

[27] M. Imani and G. A. Montazer, “A survey of emotion recognition methods with emphasis on E-Learning environments,” *J. Netw. Comput. Appl.*, vol. 147, no. July, p. 102423, 2019, doi: 10.1016/j.jnca.2019.102423.

[28] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: A review,” *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012, doi: 10.1007/s10772-011-9125-1.

[29] B. Schuller, “Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends,” *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[30] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, 2005, doi: 10.1109/TSA.2004.838534.

[31] N. Kamaruddin and A. Wahab, “Features extraction for speech emotion,” *J. Comput. Methods Sci. Eng.*, vol. 9, no. 1, pp. 1–12, 2009.

[32] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. I–593.

[33] M. Swain, A. Routray, and P. Kabisatpathy, “Databases, features and classifiers for speech emotion recognition: a review,” *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, 2018, doi: 10.1007/s10772-018-9491-z.

[34] K. S. Rao and S. G. Koolagudi, *Emotion Recognition using Speech Features*. Springer US, 2013.

[35] C. C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” *Speech Commun.*, vol. 53, no. 9–10, pp. 1162–1171, 2011, doi: 10.1016/j.specom.2011.06.004.

[36] Y. Kao and L. Lee, “Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language,” in *Ninth International Conference on Spoken Language Processing.*, 2006, pp. 1814–1817.

[37] C. Yin, R. Sun, and Q. Luo, “Study on speech emotion recognition system in E-learning,” in *International Conference on Human-Computer Interaction*, 2007, pp. 544–552, doi: 10.1109/CHICC.2006.4346832.

[38] C. Busso, S. Lee, and S. Narayanan, “Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 4, pp. 582–596, 2009, doi: 10.1109/TASL.2008.2009578.

[39] D. Neiberg, K. Elenius, and K. Laskowski, “Emotion recognition in spontaneous speech using GMMs,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2006, vol. 2, pp. 809–812.

[40] R. A. Calvo and S. D’Mello, “Affect detection: An interdisciplinary review of models, methods, and their applications,” *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, 2010, doi: 10.1109/T-AFFC.2010.1.

[41] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, “Speech emotion recognition using Fourier parameters,” *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69–75, 2015, doi: 10.1109/TAFFC.2015.2392101.

[42] J. Chen, Y. A. Huang, Q. Li, and K. K. Paliwal, “Recognition of noisy speech using dynamic spectral subband centroids,” *IEEE Signal Process. Lett.*, vol. 11, no. 2 PART II, pp. 258–261, 2004, doi: 10.1109/LSP.2003.821689.

[43] R. Cowie *et al.*, “Emotion recognition in human-computer interaction,” *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[44] N. Sato and Y. Obuchi, “Emotion Recognition using Mel-Frequency Cepstral Coefficients,” *J. Nat. Lang. Process.*, vol. 14, no. 4, pp. 83–96, 2007, doi: 10.5715/jnlp.14.4_83.

[45] L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, “Detection of clinical depression in adolescents’ speech during family interactions,” *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3 PART 1, pp. 574–586, 2011, doi: 10.1109/TBME.2010.2091640.

[46] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, “Autoencoder-based unsupervised domain adaptation for speech emotion recognition,” *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, 2014, doi: 10.1109/LSP.2014.2324759.

[47] S. Kuchibhotla, H. D. Vankayalapati, R. S. Vaddi, and K. R. Anne, “A comparative analysis of classifiers in emotion recognition through acoustic features,” *Int. J. Speech Technol.*, vol. 17, no. 4, pp. 401–408, 2014.

[48] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, vol. 2015-Janua, pp. 1537–1540.

[49] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention Center for Robust Speech Systems,” *IEEE Int. Conf. Acoust. Speech, Signal Process. 2017*, pp. 2227–2231, 2017, doi: 10.1109/ICASSP.2017.7952552.

[50] G. Tamulevičius and T. Liogienė, “Low-Order Multi-Level Features for Speech Emotion Recognition,” *Balt. J. Mod. Comput.*, vol. 3, no. 4, pp. 234–247, 2015.

[51] W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and Recurrent Neural Networks,” 2016, doi: 10.1109/APSIPA.2016.7820699.

[52] M. Schmitt, F. Ringeval, and B. Schuller, “At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016, vol. 08-12-Sept, no. September, pp. 495–499, doi: 10.21437/Interspeech.2016-1124.

[53] K. W. Gamage, V. Sethu, and E. Ambikairajah, “Salience based lexical features for emotion recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5830–5834, doi: 10.1109/ICASSP.2017.7953274.

[54] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, “New approach in quantification of emotional intensity from the speech signal: emotional temperature,” *Expert Syst. Appl.*, vol. 42, no. 24, pp. 9554–9564, 2015, doi: https://doi.org/10.1016/j.eswa.2015.07.062.

[55] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech using source, system, and prosodic features,” *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 265–289, 2012, doi: 10.1007/s10772-012-9139-3.

[56] K. V. Krishna Kishore and P. Krishna Satish, “Emotion recognition in speech using MFCC and wavelet features,” *Proc. 2013 3rd IEEE Int. Adv. Comput. Conf. IACC 2013*, pp. 842–847, 2013, doi: 10.1109/IAdCC.2013.6514336.

[57] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE Trans. Multimed.*, vol. 16, no. 8, pp. 2203–2213, 2014, doi: 10.1109/TMM.2014.2360798.

[58] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Commun.*, vol. 53, no. 9–10, pp. 1062–1087, 2011, doi: 10.1016/j.specom.2011.01.011.

[59] Z. T. Liu, Q. Xie, M. Wu, W. H. Cao, Y. Mei, and J. W. Mao, “Speech emotion recognition based on an improved brain emotion learning model,” *Neurocomputing*, vol. 309, pp. 145–156, 2018, doi: 10.1016/j.neucom.2018.05.005.

[60] T. Özseven, “A novel feature selection method for speech emotion recognition,” *Appl. Acoust.*, vol. 146, pp. 320–326, 2019, doi: 10.1016/j.apacoust.2018.11.028.

[61] P. Jackson, S. Haq, and J. Edge, “Audio-visual feature selection and reduction for emotion classification.,” in *Int’l Conf. on Auditory-Visual Speech Processing*, 2008, pp. 185–90.

[62] J. Sidorova, “Speech emotion recognition with TGI+.2 classifier,” in *EACL 2009 - 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, 2009, no. April, pp. 54–60, doi: 10.3115/1609179.1609186.

[63] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and Y.-H. Chang, “Emotion recognition and evaluation of Mandarin speech using weighted D-KNN classification,” in *the 17th Conference on Computational Linguistics and Speech Processing*, 2005, pp. 203–12.

[64] D. Ververidis and C. Kotropoulos, “Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections,” in *14th European Signal Processing Conference*, 2006, pp. 1–5.

[65] J. Rong, G. Li, and Y. P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” *Inf. Process. Manag.*, vol. 45, no. 3, pp. 315–328, 2009, doi: 10.1016/j.ipm.2008.09.003.

[66] L. D. Vignolo, S. R. M. Prasanna, S. Dandapat, H. L. Rufiner, and D. H. Milone, “Feature optimisation for stress recognition in speech,” *Pattern Recognit. Lett.*, vol. 84, pp. 1–7, 2016, doi: 10.1016/j.patrec.2016.07.017.

[67] B. Yang and M. Lugger, “Emotion recognition from speech signals using new harmony features,” *Signal Processing*, vol. 90, no. 5, pp. 1415–1423, 2010, doi: 10.1016/j.sigpro.2009.09.009.

[68] Ł. Juszkiewicz, “Improving speech emotion recognition system for a social robot with speaker recognition,” *2014 19th Int. Conf. Methods Model. Autom. Robot. MMAR 2014*, pp. 921–925, 2014, doi: 10.1109/MMAR.2014.6957480.

[69] M. K. Sarker, K. M. R. Alam, and M. Arifuzzaman, “Emotion recognition from speech based on relevant feature and majority voting,” *2014 Int. Conf. Informatics, Electron. Vision, ICIEV 2014*, pp. 1–5, 2014, doi: 10.1109/ICIEV.2014.6850685.

[70] Y. Tao, K. Wang, J. Yang, N. An, and L. Li, “Harmony search for feature selection in speech emotion recognition,” *2015 Int. Conf. Affect. Comput. Intell. Interact. ACII 2015*, pp. 362–367, 2015, doi: 10.1109/ACII.2015.7344596.

[71] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Detection of stress and emotion in speech using traditional and FFT based log energy features,” in *The 4th International Conference on Information, Communications and Signal Processing and 4th Pacific-Rim Conference on Multimedia (ICICS-PCM 2003)*, 2003, vol. 3, no. December, pp. 1619–1623, doi: 10.1109/ICICS.2003.1292741.

[72] Y.-L. Lin and G. Wei, “Speech emotion recognition based on HMM and SVM,” in *International Conference on Machine Learning and Cybernetics 2005*, 2005, pp. 4898–4901.

[73] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” *Neural Networks*, vol. 92, pp. 60–68, 2017, doi: 10.1016/j.neunet.2017.02.013.

[74] M. Chen, X. He, J. Yang, and H. Zhang, “3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition,” *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, 2018, doi: 10.1109/LSP.2018.2860246.

[75] T. L. Pao, C. H. Wang, and Y. J. Li, “A study on the search of the most discriminative speech features in the speaker dependent speech emotion recognition,” *Proc. - Int. Symp. Parallel Archit. Algorithms Program. PAAP*, pp. 157–162, 2012, doi: 10.1109/PAAP.2012.31.

[76] E. Bozkurt, E. Erzin, Ç. E. Erdem, and A. T. Erdem, “Formant position based weighted spectral features for emotion recognition,” *Speech Commun.*, vol. 53, no. 9–10, pp. 1186–1197, 2011, doi: 10.1016/j.specom.2011.04.003.

[77] T. Vogt, E. André, and J. Wagner, “Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4868 LNCS, pp. 75–91, 2008, doi: 10.1007/978-3-540-85099-1_7.

[78] Y. Li, L. Chao, Y. Liu, W. Bao, and J. Tao, “From simulated speech to natural speech, what are the robust features for emotion recognition?,” *2015 Int. Conf. Affect. Comput. Intell. Interact. ACII 2015*, pp. 368–373, 2015, doi: 10.1109/ACII.2015.7344597.

[79] D. Bitouk, R. Verma, and A. Nenkova, “Class-level spectral features for emotion recognition,” *Speech Commun.*, vol. 52, no. 7–8, pp. 613–625, 2010, doi:

10.1016/j.specom.2010.02.010.

[80] S. Wu, T. H. Falk, and W. Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011, doi: 10.1016/j.specom.2010.08.013.

[81] “ITERATIVE FEATURE NORMALIZATION FOR EMOTIONAL SPEECH DETECTION Carlos Busso Multimodal Signal Processing (MSP) Department of Electrical Engineering , The University of Texas at Dallas Speech Analysis and Interpretation Laboratory Viterbi School of Engin,” *Signal Processing*, pp. 5692–5695, 2011.

[82] E. Fersini, E. Messina, and F. Archetti, “Emotional states in judicial courtrooms: An experimental investigation,” *Speech Commun.*, vol. 54, no. 1, pp. 11–22, 2012, doi: 10.1016/j.specom.2011.06.001.

[83] D. Ververidis and C. Kotropoulos, “Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition,” *Signal Processing*, vol. 88, no. 12, pp. 2956–2970, 2008, doi: 10.1016/j.sigpro.2008.07.001.

[84] V. N. Degaonkar and S. D. Apte, “Emotion modeling from speech signal based on wavelet packet transform,” *Int. J. Speech Technol.*, vol. 16, no. 1, pp. 1–5, 2013, doi: 10.1007/s10772-012-9142-8.

[85] C. S. Ooi, K. P. Seng, L. M. Ang, and L. W. Chew, “A new approach of audio emotion recognition,” *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5858–5869, 2014, doi: 10.1016/j.eswa.2014.03.026.

[86] Z. Esmaileyan and H. Marvi, “Recognition of emotion in speech using variogram based features,” *Malaysian J. Comput. Sci.*, vol. 27, no. 3, pp. 156–170, 2014.

[87] S. Kuchibhotla, H. D. Vankayalapati, and K. R. Anne, “An optimal two stage feature selection for speech emotion recognition using acoustic features,” *Int. J. Speech Technol.*, vol. 19, no. 4, pp. 657–667, 2016, doi: 10.1007/s10772-016-9358-0.

[88] C. K. Yogesh *et al.*, “A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal,” *Expert Syst. Appl.*, vol. 69, pp. 149–158, 2017, doi: 10.1016/j.eswa.2016.10.035.

[89] I. Shahin, A. B. Nassif, and S. Hamsa, “Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network,” *IEEE Access*, vol. 7, pp. 26777–26787, 2019, doi: 10.1109/ACCESS.2019.2901352.

[90] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub, and C. Cleder, “Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO,” *Speech Commun.*, vol. 114, no. September, pp. 22–35, 2019, doi: 10.1016/j.specom.2019.09.002.

[91] D. Issa, M. Fatih Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” *Biomed. Signal Process. Control*, vol. 59, p. 101894, 2020, doi: 10.1016/j.bspc.2020.101894.

[92] C. Zheng, C. Wang, and N. Jia, “An ensemble model for multi-level speech

emotion recognition," *Appl. Sci.*, vol. 10, no. 1, 2020, doi: 10.3390/app10010205.

[93] خ. یغمایی, "تشخیص احساس از روي گفتار با استفاده از طبقه بند مبتني بر مدل و ویژگیهاي دینامیکي غیر خطي," *مهندسی برق و مهندسی کامپیوتر ایران*, vol. ۱۵, no. ۲, p. ۱۵۲-۱۴۵. and ,علی، حریمی, ع. احمدیفرد, ع. شهزادی

[94] L.B. Almeida, "The Fractional Fourier Transform and Time-Frequency Representations," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp. 3084–3091, 1994, doi: 10.1088/0266-5611/16/2/101.

[95] R. Tao, Y. L. Li, and Y. Wang, "Short-time fractional fourier transform and its applications," *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2568–2580, 2010, doi: 10.1109/TSP.2009.2028095.

[96] C. Candan, M. A. Kutay, and H. M. Ozaktas, "The discrete fractional fourier transform," *IEEE Trans. Signal Process.*, vol. 48, no. 5, pp. 1329–1337, 2000.

[97] V. Pitsikalis and P. Maragos, "Analysis and classification of speech signals by generalized fractal dimension features," *Speech Commun.*, vol. 51, no. 12, pp. 1206–1223, 2009, doi: 10.1016/j.specom.2009.06.005.

[98] A. Ezeiza, K. L. De Ipiña, C. Hernández, and N. Barroso, "Combining mel frequency cepstral coefficients and fractal dimensions for automatic speech recognition," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7015 LNAI, pp. 183–189, 2011, doi: 10.1007/978-3-642-25020-0_24.

[99] M. Mareli and B. Twala, "An adaptive Cuckoo search algorithm for optimisation," *Appl. Comput. Informatics*, vol. 14, no. 2, pp. 107–115, 2018, doi: 10.1016/j.aci.2017.09.001.

[100] C. T. Brown, L. S. Liebovitch, and R. Glendon, "Lévy Flights in Dobe Ju/'hoansi Foraging Patterns," *Hum. Ecol.*, vol. 35, no. 1, pp. 129–138, 2007, doi: 10.1007/s10745-006-9083-4.

[101] I. Pavlyukevich, "Lévy flights, non-local search and simulated annealing," *J. Comput. Phys.*, vol. 226, no. 2, pp. 1830–1844, 2007, doi: https://doi.org/10.1016/j.jcp.2007.06.008.

[102] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[103] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10–11, pp. 787–800, 2007, doi: 10.1016/j.specom.2007.01.010.

[104] D. Torres-Boza, M. C. Oveneke, F. Wang, D. Jiang, W. Verhelst, and H. Sahli, "Hierarchical sparse coding framework for speech emotion recognition," *Speech Commun.*, vol. 99, no. February, pp. 80–89, 2018, doi: 10.1016/j.specom.2018.01.006.

Abstract

One of the most important issues in human-computer interaction is creating a system that can hear and react like a human. In addition to the text, the speech signal also contains important information and characteristics about the speaker, including age, gender, accent, dialect, health and emotions, and stress of the speaker. One of the areas that can help to achieve this goal is the establishment of spoken communication between humans and machines, as well as the understanding of human emotions by the machine and provide an appropriate response to it. This has led to the design of an Automatic Speech Emotion Recognition System. Most research in this area has been designed and implemented based on the classification of samples into discrete classes or emotionally dimensional space. Therefore, in this dissertation, a mapping between these two spaces is proposed using a suitable set of features. The aim of the proposed approach is to obtain a representation of the feature space that would better capture the essence of the emotional dimensions. The proposed speech emotion recognition system in this dissertation is designed and implemented in two phases. In the first phase, an improved emotion recognition system is proposed using a new adaptive feature extraction method based on time-frequency coefficients and an evolutionary hybrid feature selection method. In the second phase, a mapping is created between the proposed feature set of the first phase and the three-dimensional space of emotions. The proposed model has been evaluated and validated using the Berlin EMO-DB Emotional Speech Database, the SAVEE Audio-Video Database, the PDREC Persian Radio-Drama Emotional Database, and the German Continuous Emotional Speech Database VAM. Experimental results show that the proposed model effectively detects different emotional classes in the EMO-DB database with 97.57% of accuracy, in the SAVEE dataset with 80% of accuracy, and in PDREC with 91.64% of accuracy.

Keywords: Emotion Recognition, Speech Processing, Feature Extraction, Feature Selection, Classification, Computer- Human Interaction.

Shahrood University of Technology

Faculty of Computer Engineering

Ph.D. Thesis in. Artificial Intelligence

Speech Emotion Recognition by Mapping Feature Set to Multi-Dimensional Emotional Space

By: Shadi Langari

Supervisor:
Dr. Hossein Marvi

Advisor:
Dr. Morteza Zahedi

September, 2020