مقاله

تشخیص حس وابسته به گویندهٔ گفتار فارسی با استفاده از ویژگی‌های آکوستیکی

منصوره کرمی*
کارشناس‌ارشد هوش مصنوعی
دانشگاه صنعتی شریف
karami@ce.sharif.ir

پریا جمشیدلو
کارشناس‌ارشد زبان‌شناسی رایانشی
دانشگاه صنعتی شریف
jamshilou@mehr.sharif.ir

حسین صامتی
دانشیار دانشکدهٔ کامپیوتر
دانشگاه صنعتی شریف
sameti@sharif.edu

تاریخ دریافت: ۱۳۹۲/۱۲/۱۴ تاریخ پذیرش: ۱۳۹۲/۱۲/۲۰

چکیده

بیان احساس در ارتباطات روزمره از جایگاه ویژه‌ای برخوردار است. از جمله بسترهای نمود احساس، گفتار است. از این‌رو، یکی از جنبه‌های مهم در طبیعی‌سازی ارتباط میان انسان و ماشین، تشخیص حس گفتار و تولید بازخورد متناسب با احساس درک‌شده است. باوجود پیشرفت‌های گسترده در حوزهٔ پردازش گفتار، استخراج و درک احساس پنهان در گفتار انسان، همچون خشم، شادی و جز این‌ها، از یک‌سو و تولید گفتار احساسی مناسب از سوی دیگر، همچنان یکی از چالش‌های مهم برای ساخت ماشین‌های هوشمند محسوب می‌شود. در این مقاله، یک سیستم وابسته به گوینده برای تشخیص حس گفتار فارسی ارائه شده است. مراد از تشخیص حس وابسته به گویندهٔ گفتار، شناسایی خودکار حالت احساسی یک یا چند گویندهٔ خاص با استفاده از نمونه‌های گفتاری آنهاست. در طراحی سیستم معرفی‌شده، از روش‌های آماری استفاده شده است و معماری آن شامل دو بخش اصلی، استخراج ویژگی و آموزش مدل دسته‌بند می‌باشد. در مرحلهٔ استخراج ویژگی، ۲۸ ویژگی آکوستیکی شامل اطلاعات مربوط به فرکانس گام، ساخت سه فرمنت اول و دامنه از نمونه‌های گفتار احساسی دو گوینده (یک مرد و یک زن) به‌طور مجزا و به ازای شش حس متفاوت خشم، تنفر، ترس، شادی، غم و خنثی استخراج شده است. پس از تشکیل بردار ویژگی، سه روش دسته‌بند، شامل ماشین بردار پشتیبان[۱]، k نزدیکترین همسایه[۲] و شبکهٔ عصبی[۳]، آموزش داده شده‌اند. در نهایت، سه روش پیاده‌سازی‌شده با استفاده از پاره‌گفتارهای احساسی دادهٔ آزمون ارزیابی شده و دقت و صحت و بازخوانی آنها مشخص شده است. با مقایسهٔ عملکرد سه روش دسته‌بند مشخص شد که بیشترین دقت برای گویندهٔ مرد و زن به‌ترتیب مربوط به ماشین بردار پشتیبان (۹۷ درصد) و شبکهٔ عصبی (۹۳ درصد) بوده است. این در حالی است که در آزمون انسانی صورت‌گرفته، میانگین دقت انسان در تشخیص حس پاره‌گفتارهای احساسی دادهٔ آزمون ۷۸ درصد و کمتر از دقت روش‌های دسته‌بند گزارش‌شده در سیستم پیاده‌سازی شده است.

واژگان کلیدی: تشخیص حس، وابسته به گوینده، ویژگی آکوستیکی، گفتار فارسی

۱. مقدمه

سیگنال گفتار سریع‌ترین و طبیعی‌ترین روش برای ارتباط متقابل میان انسان‌هاست. این واقعیت محققان را برآن داشته است تا از گفتار به‌عنوان روشی سریع در ارتباط میان ماشین و انسان استفاده کنند. برای نیل به این هدف، ماشین نیازمند آن است که از هوشمندی تشخیص گفتار انسان برخوردار باشد. طی پنج دههٔ اخیر، تحقیقات فراوانی در حوزهٔ گفتار انجام شده است. اما با وجود این پیشرفت‌ها هنوز نتوانسته‌ایم به ارتباط طبیعی بین انسان و ماشین برسیم؛ زیرا ماشین‌ها قادر به تشخیص حس گوینده در گفتار نیستند. همین موضوع، انگیزهٔ ایجاد دانشی به نام تشخیص حس گفتار[۴] شده است که هدف آن استخراج احساس گوینده از لابلای گفتار اوست. به‌علاوه، تشخیص حس گوینده به درک بهتر معنا کمک می‌کند و کارایی سیستم‌های پردازش گفتار را افزایش می‌دهد [۱].

در اموری که ارتباط متقابل میان انسان و ماشین وجود دارد، تشخیص حس گفتار امری ضروری به‌نظر می‌رسد. از جملهٔ این کاربردها ایجاد موتور جستجوی احساسی و دسته‌بندی نامه‌های الکترونیک و پیام‌های صوتی براساس حس گوینده است. از دیگر کاربردهای آن نیز می‌توان به سیستم ترجمهٔ خودکار اشاره کرد که در آن حالت احساسی فرد نقش مهمی در ارتباط طرفین و انتقال مفاهیم و معانی دارد. همچنین در سیستم‌های تشخیص گفتار کابین خلبان، مشخص شده است که داده‌های آموزشی استفاده‌شده با گفتار تأکیدی، کارآیی بهتری نسبت به گفتار معمولی دارد [۲]. تشخیص حس گفتار در مراکز مخابراتی نیز مفید است؛ هدف اصلی در این کاربردها تشخیص احساس ناامیدی یا مردم‌آزاری از گفتار گوینده است. به‌علاوه، پزشکان با استفاده از سیستم تشخیص حس گفتار می‌توانند برخی از بیماری‌های خاص همچون درخودماندگی یا اوتیسم[۵]، پارکینسون[۶]، بیماری‌های قلبی و سرطان را تشخیص دهند. در این نوع بیماری‌ها، گفتار بیمار تحت تأثیر بیماری قرار می‌گیرد و عاری از احساس تولید می‌شود.

در پاره‌ای از موارد نیز بیمار قادر به تشخیص گفتار غیرجدی از گفتار جدی نیست. تشخیص حس گفتار در حوزهٔ سرگرمی نیز برای تولید اسباب‌بازی‌های هوشمند و بازی‌های رایانه‌ای کاربرد فراوانی دارد. هدف نهایی در حوزهٔ تشخیص حس گفتار، ساخت سیستمی است که بتواند احساس گویندگان را به ازای گفتار چندزبانه و به‌صورت بی‌درنگ[۷] استخراج کند. برای نیل به چنین هدفی، پایگاه دادگان بسیار بزرگی مورد نیاز است که گویندگان و زبان‌های گوناگونی را پوشش دهد. اما چون هنوز چنین پایگاه دادگانی در دسترس نیست، نخستین گام تشخیص حس به‌صورت وابسته به گوینده[۸] و برای یک زبان خاص است. این مقاله به تشخیص حس دو گویندهٔ فارسی‌زبان با استفاده از ماشین بردار پشتیبان، شبکهٔ عصبی و k نزدیکترین همسایه می‌پردازد. در بخش دوم، پیشینهٔ تحقیق بررسی می‌شود. سپس، روش کار، ارزیابی سیستم و نتایج به‌دست‌آمده بررسی و در نهایت، به کارها و فعالیت‌های آتی اشاره می‌شود.

۲. پیشینهٔ تحقیق

نخستین موضوعی که در حوزهٔ تشخیص حس گفتار وجود دارد این است که باید مجموعه‌ای از احساسات مهم مشخص شود که بتوان از آن در سیستم تشخیص حس گفتار به‌صورت خودکار استفاده کرد. زبان‌شناسان حالت‌های احساسی متفاوت را - که در زندگی ما جریان دارند - تعیین کرده‌اند. شابیگر، اکانر و آرنولد[۹] در مجموعه‌ای ۳۰۰ حالت احساسی گوناگون را بیان کرده‌اند؛ اما دسته‌بندی این تعداد احساس متفاوت قدری مشکل است [۲]. بیشتر محققان موافق با نظریهٔ پالت[۱۰] هستند. این نظریه بیان می‌دارد که هر حس را می‌توان به احساسات اصلی تجزیه کرد؛ درست مثل اینکه هر رنگ را می‌توان با ترکیب رنگ‌های اصلی تولید کرد. احساسات اصلی شامل عصبانیت، ترس، خوشحالی، ناراحتی، تعجب و تنفر می‌باشد [۳]. به این

موارد، احساسات نمونهٔ اولیه گفته می‌شود. انواع گوناگون روش‌های دسته‌بند شامل مدل مخفی مارکوف[11]، مدل مخلوط گوسی[12]، ماشین بردار پشتیبان و شبکهٔ عصبی برای کاربرد تشخیص حس گفتار مورد استفاده قرار گرفته‌اند. در واقع، هیچ‌گونه توافقی درخصوص مناسب‌ترین روش دسته‌بند برای کاربرد تشخیص حس گفتار وجود ندارد. همچنین، هر روش دسته‌بندی، مزایا و محدودیت‌های خاص خود را دارد. به‌منظور ترکیب مزایای روش‌های دسته‌بند، ترکیبی از آنها نیز مورد استفاده قرار گرفته است. البته، مدل مخلوط گوسی و مدل مخفی مارکوف بیشترین کاربرد را در حوزهٔ تشخیص حس گفتار داشته‌اند. در ادامه به بررسی کارهای انجام‌شده در این حوزه می‌پردازیم.

از جمله روش‌های دسته‌بند رایج، که در بسیاری از کاربردهای تشخیص الگو استفاده می‌شود، شبکهٔ عصبی مصنوعی[13] است. این شبکه در مقایسه با مدل مخلوط گوسی و مدل مخفی مارکوف مزایایی دارد. شبکهٔ عصبی مصنوعی به مدل‌سازی کارآمدتر نگاشت غیرخطی معروف است. معمولاً، عملکرد آن زمانی‌که تعداد نمونه‌ها نسبتاً کم است، بهتر از مدل مخفی مارکوف و مدل مخلوط گوسی است. تقریباً می‌توان تمام شبکه‌های عصبی مصنوعی را به سه نوع اصلی طبقه‌بندی کرد:

۱. ام. ال. پی.[14]
۲. شبکهٔ عصبی بازگشتی[15]
۳. شبکهٔ مبتنی بر تابع رادیال[16]

البته مورد آخر به‌ندرت در تشخیص حس گفتار استفاده می‌شود. معمولاً از شبکهٔ عصبی ام. ال. پی. در تشخیص حس گفتار استفاده می‌کنند. دلیل آن نیز پیاده‌سازی آسان‌تر و الگوریتم آموزش خوش‌تعریف‌تر آن در صورت مشخص‌کردن کامل ساختار شبکهٔ عصبی مصنوعی است. البته، دسته‌بند شبکهٔ عصبی مصنوعی پارامترهای طراحی بسیاری دارد؛ پارامترهایی چون فرم تابع فعال‌سازی نورون، تعداد لایه‌های مخفی و تعداد نورون‌ها در هر لایه که معمولاً به شیوهٔ تک‌کاره تنظیم می‌شوند. در حقیقت، عملکرد شبکهٔ عصبی مصنوعی به‌شدت به این پارامترها بستگی دارد. بنابراین در برخی سیستم‌های تشخیص حس گفتار از بیش از یک شبکهٔ عصبی مصنوعی استفاده می‌شود. در مرجع [۱]، هدف اصلی دسته‌بندی هشت احساس شادی، رنجش، ترس، غم، چندش، خشم، تعجب و خنثی بیان شده است. دسته‌بندی پایه، یک شبکهٔ عصبی One-Class-in-One است که شامل هشت زیرشبکهٔ عصبی ام. ال. پی. و یک کنترل‌کنندهٔ منطقی تصمیم است. هر کدام از زیرشبکه‌های عصبی شامل دو لایهٔ مخفی به‌علاوهٔ لایه‌های ورودی و خروجی هستند. لایهٔ خروجی تنها شامل یک نورون می‌شود که خروجی‌اش یک مقدار آنالوگ بین صفر تا ۱ است. هر زیرشبکهٔ عصبی برای تشخیص یکی از هشت احساس بالا آموزش داده شده است. در مرحلهٔ تست، خروجی هر شبکهٔ عصبی مصنوعی احتمال تولید یک ورودی بردار گفتار حاوی یک حس خاص را تعیین می‌کند. کنترل منطقی تصمیم یک فرضیهٔ مبتنی بر خروجی هشت زیرشبکهٔ عصبی است. این بخش بر دادگان گفتاری ضبط‌شده حاوی یکصد گوینده اعمال می‌شود. هر گوینده صد کلمه را هشت بار به ازای هر هشت حس خوانده است. بهترین دقت دسته‌بندی ۵۲/۸۷ درصد مربوط به آموزش پاره‌گفتار مرتبط با ۳۰ گوینده و تست روی مابقی پاره‌گفتارها می‌شود. به‌عبارت دیگر دسته‌بندی مستقل از گوینده است. دقت مشابهی در مرجع [۴] به‌دست آمده است که برای تمام کلاس‌ها یک شبکهٔ عصبی در نظر گرفته شده است. در این مطالعه چهار توپولوژی مورد بررسی قرار گرفته است. در تمامی آنها شبکهٔ عصبی تنها یک لایهٔ مخفی داشته است که شامل ۲۶ نورون بوده. لایهٔ ورودی ۷ یا ۸ نورون و لایهٔ خروجی ۱۴ یا ۲۶ نورون داشته است. بهترین دقت دسته‌بندی ۵۱/۱۹ درصد بوده است. البته مدل‌های دسته‌بندی وابسته به گوینده بودند. نتیجهٔ بهتر در مرجع [۵] گزارش شده است. در این مطالعه سه نوع شبکهٔ عصبی مصنوعی اعمال

شده است. اولی یک دسته‌بند ام. ال. پی. دولایه است. پایگاه دادهٔ مورد استفاده در این بررسی حاوی ۷۰۰ پاره‌گفتار برای احساساتی چون شادی، خشم، غم، ترس و خنثی است. زیرمجموعه‌ای از داده حاوی ۳۶۹ پاره‌گفتار براساس تصمیمات انسانی انتخاب و ۷۰ درصد آن به‌طور تصادفی به‌عنوان دادهٔ آموزش و ۳۰ درصد آن به‌عنوان دادهٔ آزمون جدا شده است. میانگین دقت دسته‌بندی حدود ۶۵ درصد گزارش شده است. متوسط دقت دسته‌بندی برای شکل دوم که از روش ادغام قائم به‌ذات[۱۷] استفاده شده است ۷۰ درصد بوده است. در نهایت، میانگین دقت دسته‌بندی در شکل سوم ۶۳ درصد گزارش شده است که توضیحات آن مشابه توضیحات سیستم قبلی است. بهتربودن عملکرد دسته‌بندی در این مطالعه نسبت به دو مطالعهٔ دیگر به‌علت استفاده از پیکره‌های احساسی متفاوت در هر مطالعه است.

یک مثال برجسته برای دسته‌بندهای تمایزی عمومی، ماشین بردار پشتیبان است. ماشین بردار پشتیبان عموماً مبتنی بر توابع کرنل[۱۸] است که برای نگاشت غیرخطی ویژگی‌های اصلی به فضایی با ابعاد بالاتر مورد استفاده قرار می‌گیرد؛ جایی‌که می‌توان داده را با استفاده از یک دسته‌بند خطی به‌خوبی دسته‌بندی کرد. ماشین بردار پشتیبان در بسیاری از کاربردهای تشخیص الگو مورد استفاده قرار می‌گیرد و نسبت به سایر دسته‌بندهای مشهور عملکرد بهتری دارد. مزایای آن نسبت به مدل مخفی مارکوف، مدل مخلوط گوسی، بهینگی سراسری الگوریتم آموزش و وجود کران‌های تعمیم به دادهٔ آن وابسته است. البته در برخورد با موارد جداناپذیر تاحدی مکاشفه‌ای[۱۹] عمل می‌کند. در حقیقت، هیچ شیوهٔ منظمی برای انتخاب توابع کرنل وجود ندارد؛ بنابراین جداسازی ویژگی‌های انتقالی را تضمین نمی‌کند. در بسیاری از کاربردهای تشخیص الگو شامل تشخیص حس گفتار توصیه می‌شود که جداسازی کاملی از دادهٔ آموزش نداشته باشیم تا از مسئلهٔ بیش‌برازش[۲۰] جلوگیری به‌عمل بیاید [۲]. ماشین بردار پشتیبان در اصل یک جداکنندهٔ دودویی است. یک تشخیص الگوی چندکلاسی می‌تواند به‌وسیلهٔ ترکیب ماشین‌های بردار پشتیبان دو کلاسی حاصل شود. به‌طور معمول دو دید برای این هدف وجود دارد: یکی از آنها رویکرد "یک در مقابل همه" برای دسته‌بندی هر جفت کلاس و کلاس‌های باقی‌مانده است. دیگری، رویکرد "یک در مقابل یک" برای دسته‌بندی هر جفت است. برای مسائل چندکلاسی، رهیافت کلی کاهش مسئلهٔ چندکلاسی به چندین مسئلهٔ دودویی است. هریک از مسائل با یک جداکننده دودویی حل می‌شوند. سپس خروجی جداکننده‌های دودویی ماشین بردار پشتیبان با هم ترکیب و به این ترتیب مسئلهٔ چندکلاس حل می‌شود. دسته‌بند ماشین بردار پشتیبان در بسیاری از مطالعات برای مسئلهٔ تشخیص حس گفتار مورد استفاده قرار گرفته است. تقریباً عملکرد تمامی آنها مشابه هم است؛ بنابراین تنها به توضیح یک مورد از آنها اکتفا می‌کنیم.

در مرجع [۶]، سه رویکرد برای تعمیم دسته‌بند دوتایی ماشین بردار پشتیبان برای حالت چندکلاسه بررسی شده است. در دو رویکرد اول یک دسته‌بند ماشین بردار پشتیبان برای مدلسازی هر حس استفاده شده که در برابر سایر احساسات آموزش داده شده است. در رویکرد اول، تصمیم برای کلاسی اتخاذ می‌شود که بیشترین فاصله را با سایر کلاس‌ها داشته باشد. در رویکرد دوم فاصلهٔ خروجی ماشین بردار پشتیبان به یک دسته‌بند سه‌لایه‌ای ام. ال. پی.، که تصمیم درخصوص خروجی نهایی را اتخاذ می‌کند، داده می‌شود. رویکرد سوم پس از دسته‌بندی سلسله‌مراتبی اعمال می‌شود. سه سیستم با استفاده از پاره‌گفتارهایی که از پیکرهٔ FERMUS III انتخاب شده‌اند ارزیابی شدند. برای دسته‌بندی مستقل از گوینده دقت دسته‌بندی برای رویکرد اول، دوم و سوم به‌ترتیب ۷۶/۱۲، ۷۵/۴۵ و ۸۱/۲۹ درصد بوده است. برای دسته‌بندی وابسته به گوینده دقت دسته‌بندی ۹۵/۹۵، ۸۸/۷ و ۹۰/۹۵ درصد به‌ترتیب برای رویکرد اول، دوم و سوم گزارش شده است. مدل مخفی

مارکوف به‌طور گسترده در کاربردهای مربوط به گفتار مورد استفاده قرار گرفته است. علت آن سازگاری این مدل با سازوکار تولیدی سیگنال گفتار است. در مرجع [۶] سیستمی مبتنی بر مدل مخفی مارکوف برای دسته‌بندی شش احساس گوناگون معرفی شده است. به ازای هر حس و هر گوینده، یک مدل مخفی مارکوف چهارحالته ساخته شده است. مدل‌ها از نوع گسسته‌اند و کُدبوکی با اندازهٔ ۶۴ به ازای دادهٔ هر گوینده ساخته شده است. پایگاه دادهٔ احساسی مورد استفاده در این مقاله حاوی ۷۲۰ پاره‌گفتار است که از ۴۳۲ پاره‌گفتار آن برای آموزش و مابقی آن برای آزمون مدل‌ها استفاده شده است. میانگین دقت دسته‌بند ۷۸/۵ درصد است؛ این در حالی است که میانگین دقت انسانی ۶۵/۸ درصد گزارش شده است. در مرجع [۷] نیز از مدل مخفی مارکوف برای دسته‌بندی شش احساس گوناگون استفاده شده است. پایگاه دادهٔ مورد استفاده حاوی ۶۰ پاره‌گفتار احساسی بوده که گفتار احساسی ۱۲ گوینده را پوشش می‌دهد. مدل‌ها از نوع گسسته بوده، از ویژگی LFPC [۲۱] برای نمایش سیگنال گفتار و آموزش مدل‌ها استفاده شده است. میانگین دقت برای حالت مستقل از گوینده ۷۸ درصد اعلام شده است. مدل مخلوط گوسی یک مدل احتمالاتی است که با استفاده از یک ترکیب محدب از چگالی‌های نرمالِ چندمتغیره، چگالی را تخمین می‌زند. مدل مخلوط گوسی را می‌توان نوعی خاص از مدل مخفی مارکوف پیوسته دانست که تنها حاوی یک حالت است. مزیت مدل مخلوط گوسی نسبت به مدل مخفی مارکوف ملزومات کمتر آن برای مرحلهٔ آموزش و آزمون است. در مرجع [۹] از مدل مخلوط گوسی برای دسته‌بندی پنج حس متفاوت استفاده شده است. پایگاه دادهٔ مورد استفاده حاوی ۷۲۶ پاره‌گفتار احساسی بوده و به‌علت حجم کم آن روش اعتبارسنجی متقابل یکصد لایه اعمال شده است. در مجموع پنج ویژگی مربوط به اطلاعات زیروبمی و انرژی برای آموزش مدل‌ها استخراج شده است. میانگین دقت ۷۸/۷۷ درصد گزارش شده است. در مرجع [۱۰] نیز از مدل مخلوط گوسی برای دسته‌بندی شش حس گوناگون استفاده شده است. پایگاه دادهٔ احساسی مورد استفاده حاوی ۵۲۵۰ پاره‌گفتار احساسی است. روش اعتبارسنجی متقابلِ سه‌لایه اعمال و از مدل مخلوط گوسی ۱۶ مؤلفه‌ای برای مدل‌کردن هر حس استفاده شده است. میانگین دقت برای حالت وابسته به گوینده ۸۹/۱۲ درصد و برای حالت مستقل از گوینده ۷۴/۸۳ درصد گزارش شده است.

روش‌های دسته‌بند بسیار دیگری چون k نزدیکترین همسایه، دسته‌بند فازی و درخت تصمیم در بسیاری از مطالعات دیگر مورد استفاده قرار گرفته‌اند. در کل، مقایسهٔ انواع سیستم‌های تشخیص حس گفتار میسر نیست؛ زیرا در هر یک از مقاله‌های ذکرشده پیکره‌های احساسی متفاوت با پوشش حس‌های گوناگون مورد استفاده قرار گرفته است. همچنین، برخی از این پایگاه داده‌های گفتاری توسط نویسندگان مقاله ضبط شده است و برای سایر محققان قابل دسترسی نیست. بنابراین ادعا درخصوص بهترین دسته‌بند برای کاربرد تشخیص حس گفتار نیازمند مطالعات جامع‌تری است که طی آن انواع روش‌های دسته‌بندی بر روی پیکره‌های قابل دسترس بسیاری اعمال و عملکرد آنها با هم مقایسه شوند.

۳. روش کار

در این مقاله، یک سیستم وابسته به گوینده برای تشخیص خودکار حس گفتار فارسی معرفی شده است. محرک احساسی مورد استفاده در این مطالعه، از دادگان گفتار احساسی فارسی موسوم به Persian ESD [۲۲] گرفته شده است [۱۱]. این دادگان گفتاری شامل مجموعه‌ای از گفتار احساسی ضبط‌شده است که از لحاظ شنیداری تست شده‌اند. این دادگان پنج حس خشم، تنفر، ترس، شادی و غم به‌همراه گفتار خنثی را در بر می‌گیرد. چهارده جملهٔ ثابت توسط دو گویندهٔ خانم و آقا، هر بار به یکی از حس‌های اشاره‌شده، خوانده شده‌اند. این دادگان گفتار احساسی در مجموع حاوی ۱۶۸ پاره‌گفتار و ۱۳ دقیقه و ۱۹

ثانیه گفتار احساسی می‌باشد. مراحل انجام کار شامل استخراج ویژگی از پاره‌گفتارهای احساسی و آموزش دسته‌بند می‌شوند که در ادامه توضیح داده شده‌اند.

۳-۱. استخراج ویژگی

در این مرحله، ۲۸ ویژگی آکوستیکی از پاره‌گفتارهای احساسی استخراج شدند که حاوی اطلاعاتی در خصوص فرکانس گام[۲۳]، ساختار فرمنتی[۲۴] و دامنه[۲۵] بودند. فهرست ویژگی‌های استخراج‌شده در جدول ۳ ذکر شده است. ویژگی‌های استخراج‌شده در این مطالعه از نوع ویژگی‌های سراسری‌اند. دلیل استفاده از ویژگی‌های سراسری، برتری آنها از لحاظ تعداد پارامترها، دقت و زمان دسته‌بندی نسبت به ویژگی‌های محلی است. ویژگی‌ها به‌صورت جداگانه از پاره‌گفتارهای احساسی مربوط به گویندهٔ زن و مرد استخراج شده‌اند. دلیل این امر، متفاوت‌بودن ویژگی‌ها (متفاوت‌بودن فرکانس گام و ساخت فرمنتی) به ازای صدای زن و مرد و تأثیر آن بر عملکرد سیستم تشخیص حس گفتار است. به‌طور کلی، برای بالابردن دقت دسته‌بندی، کاربرد تشخیص حس گفتار عموماً به‌صورت وابسته به جنسیت انجام می‌گیرد. پس از استخراج ویژگی‌های آکوستیکی، یک بردار ویژگی ۲۸بعدی تشکیل شد. هدف از استخراج جزئیات مربوط به فرکانس گام شامل مقادیر کمینه، بیشینه، میانه، میانگین و انحراف معیار مدل‌کردن دقیق‌تر منحنی زیروبمی است که به‌نظر می‌رسد در تمایز احساساتی با برانگیختگی بالا نظیر حس خشم از سایر احساسات با برانگیختگی پایین بسیار مؤثر است.

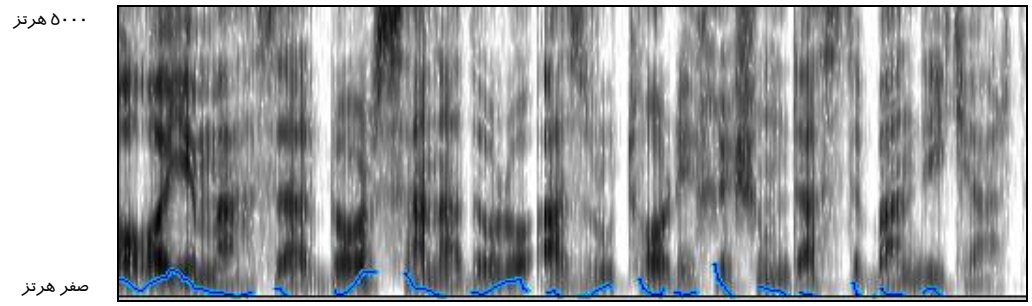
در شکل ۱، طیف مربوط به یک پاره‌گفتار خنثی که توسط گویندهٔ مرد تولیدشده نمایش داده شده است. در این شکل، منحنی آبی‌رنگ نشان‌دهندهٔ منحنی زیروبمی است. شکل ۲، طیف مربوط به همان جمله با حس خشم است که توسط گویندهٔ یکسان تولید شده است. از مقایسهٔ شکل‌های ۱ و ۲ مشخص می‌شود که منحنی زیروبمی در پاره‌گفتار خشمگین نقش به‌سزایی ایفا می‌کند. برای حصول اطمینان از مناسب‌بودن ویژگی‌های آکوستیکی استخراج‌شده، این ویژگی‌ها تست شد. به این ترتیب که تمام ۲۸ ویژگی را به ازای یک حس خاص، مثل خشم، استخراج و روی همان جمله با حس خنثی اعمال و جمله پخش شد. نهایتاً صدای پخش‌شده خشمگین درک شد که این حاکی از مرتبط و مناسب‌بودن ویژگی‌های مورد استفاده بود.

جدول ۱. بردار ویژگی‌های آکوستیکی

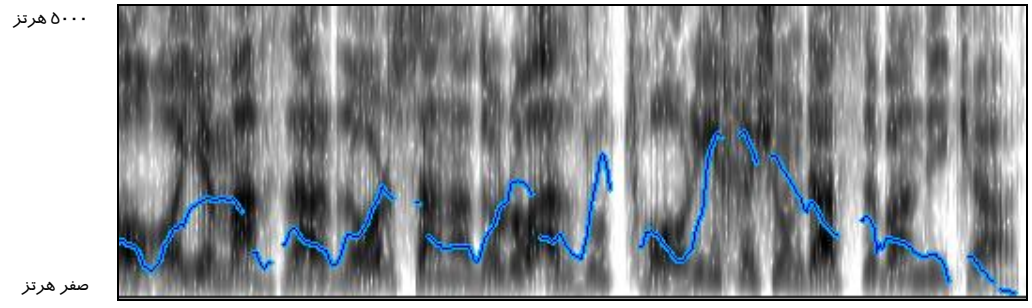
پهنای باند فرمنت اول	بیشینهٔ فرمنت دوم
پهنای باند فرمنت دوم	میانهٔ فرمنت دوم
پهنای باند فرمنت سوم	میانگین فرمنت دوم
کمینهٔ فرکانس گام	انحراف معیار فرمنت دوم
بیشینهٔ فرکانس گام	کمینهٔ فرمنت سوم
میانهٔ فرکانس گام	بیشینهٔ فرمنت سوم
میانگین فرکانس گام	میانهٔ فرمنت سوم
انحراف معیار فرکانس گام	میانگین فرمنت سوم
کمینهٔ فرمنت اول	انحراف معیار فرمنت سوم
بیشینهٔ فرمنت اول	کمینهٔ دامنه
میانهٔ فرمنت اول	بیشینهٔ دامنه
میانگین فرمنت اول	میانهٔ دامنه
انحراف معیار فرمنت اول	میانگین دامنه
کمینهٔ فرمنت دوم	انحراف معیار دامنه

۳-۲. آموزش دسته‌بند

ابتدا پاره‌گفتارهای احساسی به دو دستهٔ آموزش و آزمون تقسیم شدند. که این کار به‌دلیل حجم کم پایگاه‌داده با روش اعتبارسنجی متقابلِ Leave-One-Out صورت گرفت. به‌طوری‌که در هر بار اجرا، به ازای هر حس یکی از چهارده پاره‌گفتار به‌عنوان تست در نظر گرفته و سیزده گفتار دیگر به‌عنوان دادهٔ آموزشی برای آموزش دسته‌بند استفاده شد. در نهایت میانگین نتایج به‌عنوان نتایج نهایی استخراج شد. ویژگی‌ها به ازای گویندهٔ مرد و زن به‌صورت جداگانه استخراج و آموزش داده شدند. همچنین به‌دلیل



شکل ۱. منحنی زیروبمی (آبی‌رنگ) مربوط به پاره‌گفتار خنثی؛ گویندهٔ مرد



شکل۲. منحنی زیروبمی (آبی‌رنگ) مربوط به پاره‌گفتار خشم؛ گویندهٔ مرد

فرمت پایگاه داده، نتایج همگی وابسته به گوینده هستند. برای آموزش سیستم مستقل از گوینده به چندین گوینده نیاز است که در پایگاه داده‌ای که در اختیار داشتیم تنها از یک گویندهٔ زن و یک گویندهٔ مرد استفاده شده بود.

برای دسته‌بندی نیز سه روش پیاده‌سازی و با هم مقایسه شدند. این روش‌ها عبارت‌اند از ماشین بردار پشتیبان، k نزدیک‌ترین همسایه و شبکه‌های عصبی. مزایای ماشین بردار پشتیبان، دقت بالای آن در حالتی که دادهٔ آموزشی کافی وجود ندارد، آموزش نسبتاً ساده و گیرنکردن در بیشینه‌های محلی است. شبکه‌های عصبی نیز در حالتی که حجم دادهٔ آموزش زیاد نیست عملکرد خوبی دارند. مزیت دسته‌بند k نزدیکترین همسایه نیز سادگی پیاده‌سازی آن است. در روش ماشین بردار پشتیبان، از رویکرد "یک در مقابل همه" استفاده شد. بدین ترتیب به تعداد کلاس‌ها مرز ماشین بردار پشتیبان مشخص شد. در دسته‌بند ماشین بردار پشتیبان از تابع کرنل چندجمله‌ای استفاده شد. در روش k نزدیکترین همسایه، با مقداردهی k به یک، برچسب داده‌ها تعیین گردید. در روش شبکه‌های عصبی نیز، از یک شبکهٔ عصبی دولایه استفاده شد که در آن شش خروجی به ازای هر کلاس تعیین گشت. در مرحلهٔ آزمون، پاره‌گفتارهای احساسی با استفاده از مدل‌های آموزش داده‌شده دسته‌بندی شد و به هر یک از آنها، یک برچسب احساس تعلق گرفت.

٤. ارزیابی سیستم

در جدول‌های ۲ تا ۴، نتایج چهارده بار اجرا برای پاره‌گفتارهای احساسی مربوط به گویندهٔ مرد آمده است. همچنین، در جدول‌های ۵ تا ۷، نتایج ۱۴ بار اجرا برای پاره‌گفتارهای احساسی مربوط به گویندهٔ زن ذکر شده است. در جدول‌های ۸ و ۹ نیز میانگین دقت، صحت و بازخوانی به ترتیب برای گویندهٔ مرد و زن آورده شده است. برای ارزیابی سیستم و اینکه مشخص کنیم مقادیر به‌دست آمده در قسمت قبل تا چه‌حد قابل قبول‌اند، یک نمونه آزمون انسانی آماده شد. در این آزمون، چهارده پاره‌گفتار احساسی به‌صورت تصادفی انتخاب و برای شنونده پخش و خواسته

شد تا حس موجود در پاره‌گفتار را تشخیص دهد و گزینهٔ مربوط به آن را در پاسخ‌نامه علامت بزند. در نهایت دقت تشخیص افراد مشخص شد. برای این آزمون، ۵ مرد و ۷ زن، با محدودهٔ سنی ۲۳ تا ۲۷ سال و سطح تحصیلات کارشناسی ارشد، از رشته‌های تحصیلی گوناگون، انتخاب شدند. میانگین دقت تشخیص این افراد ۷۸ درصد محاسبه گردید که پایین‌تر از دقت هر سه دسته‌بند گزارش شد. در بین احساساتی چون عصبانیت، تنفر، ترس، خوشحالی، ناراحتی و خنثی، حس تنفر کمترین دقت و حس خوشحالی و خنثی بالاترین دقت را به‌خود اختصاص دادند.

جدول ۲. مقادیر دقت، صحت و بازخوانی اجراها در روش دسته‌بند ماشین بردار پشتیبان (گویندهٔ مرد)

	۱	۲	۳	۴	۵	۶	۷
دقت	۱	۰/۹۷	۰/۹۷	۰/۹۷	۰/۹۴	۰/۹۱	۱
صحت	۱	۱	۰/۸۳	۰/۸۳	۰/۶۶	۰/۶۶	۱
بازخوانی	۱	۰/۸۵	۱	۱	۱	۰/۸۰	۱
	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴
دقت	۰/۹۷	۱	۱	۰/۹۴	۱	۱	۱
صحت	۱	۱	۱	۰/۸۳	۱	۱	۱
بازخوانی	۰/۸۵	۱	۱	۰/۸۳	۱	۱	۱

جدول ۳. مقادیر دقت، صحت و بازخوانی اجراها در روش دسته‌بند شبکهٔ عصبی مصنوعی (گویندهٔ مرد)

	۱	۲	۳	۴	۵	۶	۷
دقت	۱	۰/۹۴	۱	۰/۹۴	۰/۹۴	۰/۹۴	۱
صحت	۱	۰/۸۳	۱	۰/۸۳	۰/۸۳	۰/۸۳	۱
بازخوانی	۱	۰/۸۳	۱	۰/۸۳	۰/۸۳	۰/۸۳	۱
	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴
دقت	۰/۹۴	۰/۹۷	۱	۱	۱	۱	۱
صحت	۱	۱	۱	۱	۱	۱	۱
بازخوانی	۰/۷۵	۰/۸۵	۱	۱	۱	۱	۱

جدول ۴. مقادیر دقت، صحت و بازخوانی اجراها و در روش دسته‌بند k نزدیکترین همسایه (گویندهٔ مرد)

	۱	۲	۳	۴	۵	۶	۷
دقت	۰/۸۸	۰/۸۸	۰/۷۷	۰/۸۸	۰/۸۸	۰/۹۴	۰/۸۸
صحت	۰/۶۶	۰/۶۶	۰/۳۳	۰/۶۶	۰/۶۶	۰/۸۳	۰/۶۶
بازخوانی	۰/۶۶	۰/۶۶	۰/۳۳	۰/۶۶	۰/۶۶	۰/۸۳	۰/۶۶
	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴
دقت	۰/۹۴	۰/۸۸	۱	۰/۸۳	۰/۸۸	۰/۸۸	۰/۸۸
صحت	۰/۸۳	۰/۶۶	۱	۰/۵۰	۰/۶۶	۰/۶۶	۰/۶۶
بازخوانی	۰/۸۳	۰/۶۶	۱	۰/۵۰	۰/۶۶	۰/۶۶	۰/۶۶

جدول ۵. مقادیر دقت، صحت و بازخوانی اجراها در روش دسته‌بند ماشین بردار پشتیبان (گویندهٔ زن)

	۱	۲	۳	۴	۵	۶	۷
دقت	۰/۹۴	۰/۹۷	۰/۸۳	۰/۹۴	۰/۸۸	۰/۹۱	۰/۷۷
صحت	۰/۶۶	۰/۸۳	۰/۶۶	۰/۶۶	۰/۶۶	۰/۶۶	۰/۱۶
بازخوانی	۱	۱	۰/۵۰	۱	۰/۶۶	۰/۸۰	۰/۲۵
	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴
دقت	۰/۹۱	۰/۸۸	۰/۹۱	۰/۸۸	۰/۹۷	۰/۹۴	۰/۹۴
صحت	۰/۶۶	۰/۵۰	۰/۸۳	۰/۳۳	۱	۰/۶۶	۰/۶۶
بازخوانی	۰/۸۰	۰/۷۵	۰/۷۱	۱	۰/۸۵	۱	۱

جدول ۶. مقادیر دقت، صحت و بازخوانی اجراها در روش دسته‌بند شبکهٔ عصبی مصنوعی (گویندهٔ زن)

	۱	۲	۳	۴	۵	۶	۷
دقت	۰/۹۷	۰/۹۷	۰/۸۰	۰/۹۴	۰/۹۱	۰/۹۱	۰/۸۸
صحت	۰/۸۳	۰/۸۳	۰/۸۳	۰/۸۳	۰/۸۳	۰/۸۳	۰/۶۶
بازخوانی	۱	۱	۰/۴۵	۰/۸۳	۰/۷۱	۰/۷۱	۰/۶۶
	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴
دقت	۱	۰/۹۷	۰/۹۷	۰/۹۴	۰/۹۷	۰/۹۷	۰/۹۷
صحت	۱	۱	۱	۰/۶۶	۰/۶۶	۱	۱
بازخوانی	۱	۰/۸۵	۰/۸۵	۱	۰/۵۷	۰/۸۵	۰/۸۵

جدول ۷. مقادیر دقت، صحت و بازخوانی اجراها در روش دسته‌بند k نزدیکترین همسایه (گویندهٔ زن)

	۱	۲	۳	۴	۵	۶	۷
دقت	۰/۷۷	۰/۸۸	۰/۷۲	۰/۷۲	۰/۸۳	۰/۹۴	۰/۷۲
صحت	۰/۳۳	۰/۶۶	۰/۱۶	۰/۱۶	۰/۵۰	۰/۸۳	۰/۱۶
بازخوانی	۰/۳۳	۰/۶۶	۰/۱۶	۰/۱۶	۰/۵۰	۰/۸۳	۰/۱۶
	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴
دقت	۰/۸۳	۰/۸۳	۰/۷۷	۰/۷۲	۰/۸۳	۰/۸۳	۰/۷۷
صحت	۰/۵۰	۰/۵۰	۰/۳۳	۰/۱۶	۰/۵۰	۰/۵۰	۰/۳۳
بازخوانی	۰/۵۰	۰/۵۰	۰/۳۳	۰/۱۶	۰/۵۰	۰/۵۰	۰/۳۳

جدول ۸. میانگین مقادیر دقت، صحت و بازخوانی اجراها در سه دسته‌بند (گویندهٔ مرد)

	دسته‌بند ماشین بردار پشتیبان	دسته‌بند شبکهٔ عصبی مصنوعی	دسته‌بند k نزدیکترین همسایه
دقت	۹۷/۸۲ درصد	۹۷/۶۴ درصد	۸۹/۲۹ درصد
صحت	۹۱/۶۷ درصد	۹۵/۲۴ درصد	۶۷/۸۶ درصد
بازخوانی	۹۵/۳۴ درصد	۹۲/۴۳ درصد	۶۷/۸۶ درصد

جدول ۹. میانگین مقادیر دقت، صحت و بازخوانی اجراها در سه دسته‌بند (گویندهٔ زن)

	دسته‌بند ماشین بردار پشتیبان	دسته‌بند شبکهٔ عصبی مصنوعی	دسته‌بند k نزدیکترین همسایه
دقت	۹۱/۰۷ درصد	۹۳/۶۵ درصد	۸۰/۵۶ درصد
صحت	۶۴/۲۹ درصد	۸۵/۷۱ درصد	۴۱/۶۷ درصد
بازخوانی	۸۰/۹۹ درصد	۸۱/۳۱ درصد	۴۱/۶۷ درصد

جدول ۱۰. ماتریس ابهام[۲۶] تست انسانی

	عصبانیت	تنفر	ترس	خوشحالی	خنثی	ناراحتی	خطا (درصد)
عصبانیت	۲۱	۱	۰	۲	۴	۱	۲۷
تنفر	۱	۱۴	۱	۰	۹	۳	۵۰
ترس	۱	۲	۲۰	۰	۱	۱	۲۰
خوشحالی	۰	۰	۰	۲۷	۰	۱	۳
خنثی	۰	۰	۰	۱	۲۸	۰	۳
ناراحتی	۰	۱	۶	۱	۰	۲۱	۲۷
							۲۲

۵. فعالیت‌های آتی

همان‌گونه که قبلاً ذکر شد، هدف نهایی در حوزهٔ تشخیص حس گفتار ایجاد یک سیستم مستقل از گوینده است؛ سیستمی که قادر به استخراج بی‌درنگ حس گفتار باشد. در آینده می‌توان با ایجاد یک پایگاه دادهٔ احساسی بزرگ برای زبان فارسی، گفتار احساسی گویندگان گوناگون را پوشش داد و در گام بعد، به ایجاد سیستمی مستقل از گوینده برای تشخیص حس گفتار فارسی اقدام نمود.

۶. تشکر و قدردانی

نویسندگان برخود لازم می‌دانند تا از دکتر محرم اسلامی و خانم نیلوفر کشتیاری برای تسهیل در امر دسترسی به پایگاه داده‌شان تشکر و قدردانی کنند.

۷. مآخذ

[1] Nicholson, J., K. Takahashi, R. Nakatsu, "Emotion recognition in speech using neural networks", *Journal of Neural Comput. Appl.* (2000): 290–296.

[2] Ayadi, M.E., M.S. Kamel, F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases", *Journal of Pattern Recognition 44*, (2011): 572–587.

[3] Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, J. Taylor, "Emotion recognition in human–computer interaction", *IEEE Signal Process. Mag.* 18, (2001): 32–80.

[4] Hozjan, V., Z. Kacic, "Context-independent multilingual emotion recognition from speech signal", *Int. J. Speech Technol.* 6, (2003): 311-320.

[5] Petrushin, V., "Emotion recognition in speech signal: experimental study, development and application", *Proceedings of the ICSLP 2000*, Beijing, China, October 16-20, 2000.

[6] Schuller, B., G. Rigoll, M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", *Proceedings of the ICASSP*, Montreal, Canada, May 17-21, 2004.

[7] New, T., S. Foo, L. De Silva, "Speech emotion recognition using hidden Markov models", *Journal of Speech Commun.* 41, (2003): 603-623.

[8] Schuller, B., G. Rigoll, M. Lang, "Hidden Markov model-based speech emotion recognition", *Proceedings of International Conference on Multimedia and Expo (ICME)*, Maryland, USA, July 6-9, 2003.

[9] Breazeal, C., L. Aryananda, "Recognition of affective communicative intent in robot-directed speech", *Journal of Autonomous Robots 2*, (2002): 83-104.

[10] Schuller, B., "Towards intuitive speech interaction by the integration of emotional aspects", *Proceedings of IEEE International Conference on Systems*, Man and Cybernetics, Hammamet, Tunisia, October 6-9, 2002.

[11] Keshtiari, N., M. Kuhlmann, M. Eslami, G. Klann-Delius, "A database of Persian Emotional Speech", *Paper presented at the 1st Basic and Clinical Neuroscience Congress*, Tehran University of Medical Sciences, 2012.

پی‌نوشت

۶. نخستین‌بار بیماری پارکینسون (Parkinson's Disease) توسط جیمز پارکینسون، دانشمند شهیر انگلیسی، در سال ۱۸۱۷ م توصیف شد و بعدها به افتخار وی، به‌نام بیماری پارکینسون معروف گشت. این بیماری همان لرزش در وضعیت استراحت است و شیوع آن در دوران کهنسالی رایج است، اما گاه در جوانان نیز دیده می‌شود. شیوع این بیماری در جای‌جای دنیا یکسان است؛ یعنی درصد شیوع آن با تغییر در منطقهٔ جغرافیایی تفاوت نمی‌کند. به‌طور کلی این بیماری بر اثر از بین رفتن سلول‌های ترشح‌کنندهٔ ماده‌ای به نام دوپامین رخ می‌دهد. پارکینسون براساس دو علامت یا بیشتر از چهار علامت اصلی بیماری مشخص می‌شود. لرزش دست و پا به‌هنگام استراحت، کندی حرکات، سختی و خشک‌شدن دست و پا و بدن و نداشتن تعادل این چهار علامت اصلی را تشکیل می‌دهند [ویراستار].

7. real-time
8. speaker dependent
9. Schubiger, O'Conner and Arnold
10. Palette
11. hidden Markov model (HMM)
12. Gaussian mixture model (GMM)
13. artificial neural network (ANN)
14. multilayer perceptron (MLP)
15. recurrent neural network

1. support vector machine (SVM)
2. K-nearest neighbor (KNN)
3. neural network (NN)
4. speech emotion recognition

۵. درخودماندگی یا اوتیسم (Autism) نوعی اختلال رشدی (از جنس روابط اجتماعی) است که با رفتارهای ارتباطی - کلامی غیرطبیعی مشخص می‌شود. علائم این اختلال تا پیش از سه‌سالگی بروز می‌کند و علت اصلی بروز آن ناشناخته است. این اختلال در پسران شایع‌تر از دختران است. وضعیت اقتصادی، اجتماعی، سبک زندگی و سطح تحصیلات والدین در بروز این اختلال نقشی ندارد. درخودماندگی بر رشد طبیعی مغز در حیطهٔ تعاملات اجتماعی و مهارت‌های ارتباطی اثر می‌گذارد. کودکان و بزرگسالان مبتلا به اوتیسم، در ارتباطات کلامی و غیرکلامی، تعاملات اجتماعی و فعالیت‌های مربوط به بازی مشکل دارند؛ به‌طوری‌که این اختلال، ارتباط با دیگران و دنیای خارج را برای آنها دشوار می‌کند. در پاره‌ای از موارد رفتارهای خودآزارانه و پرخاشگری نیز دیده می‌شود. در این افراد حرکات تکراری (دست‌زدن، پریدن) پاسخ‌های غیرمعمول به افراد، دلبستگی به اشیاء یا مقاومت در برابر تغییر نیز دیده می‌شود و ممکن است در حواس پنج‌گانه نیز حساسیت‌های غیرمعمول دیده شود. هستهٔ مرکزی اختلال در اوتیسم، اختلال در ارتباط است [ویراستار].

16. radial basis function network
17. bootstrapa
18. Kernel function
19. heuristic
20. over-fitting
21. log frequency power coefficients
22. Persian Emotional Speech Database
23. fundamental frequency
24. Formant structure
25. amplitude
26. confusion matrix