ISME

ارائه یک الگوریتم جدید برای تشخیص احساس گوینده برای تعامل انسان و ربات

علی نیک‌نژاد[1]
دانشجوی کارشناسی ارشد

علی غفاری[2]
استاد

علیرضا خدایاری[3]
دانشیار

آسان‌ترین ارتباط بین انسان و ماشین از طریق گفتار است و از ملزومات این ارتباط، درک احساس انسان توسط ماشین است. در این مقاله الگوریتم جدیدی برای تشخیص احساس ارائه شده است. در الگوریتم پیشنهادی با الهام گرفتن از شنوایی انسان، با هدف افزایش سرعت و دقت در تشخیص، از ویژگی‌های سیگنال صدا ضرایب کپسترال فرکانسی مل را استخراج کرده و ویژگی‌هایی بهینه انتخاب می‌شوند. سپس با استفاده از ترکیب طبقه‌بندهای ماشین بردار پشتیبان و مدل مخلوط گاوسی، تشخیص احساس انجام میشود. نتایج حاصل از پیاده‌سازی این الگوریتم برای دو زبان فارسی و آلمانی ارائه و با نتایج الگوریتم‌های دیگر برای همان پایگاه‌داده مقایسه شده است. نتایج برای پایگاه‌داده استاندارد آلمانی ۸۹٪ و برای پایگاه داده غیر استاندارد فارسی ۶۸٪ بدست آمده، این نتایج عملکرد مناسب الگوریتم را در طراحی سیستم-های کنترل و هدایت ربات‌ها نشان می‌دهد.

واژه های راهنما : تشخیص احساس، احساس گفتاری، ویژگی های صدا، ضرایب کپسترال فرکانسی مل.

۱- مقدمه

از آنجایی که گفتار می‌تواند سریع‌ترین و کاراترین روش تعامل بین انسان و کامپیوتر باشد، درسال‌های اخیر پژوهش‌های زیادی برای طراحی و ایجاد سیستم های تعاملی بر پایه گفتار انجام شده است. از اینرو پیشرفت های بسیاری در زمینه پردازش گفتار و تبدیل گفتار به رشته‌ای از کلمات و تحلیل محتوای گفتار انجام شده است. اما هنوز فاصله زیادی با داشتن یک تعامل طبیعی بین انسان و ماشین وجود دارد. یکی از ملزومات اصلی تعامل انسان با انسان، درک احساس است که در پژوهش‌ها، طراحی و ایجاد سیستم‌های تعاملی مورد توجه قرار نگرفته است. بنابراین ماشین قادر به درک احساسات گوینده نیست. این موضوع باعث شد که در سال‌های اخیر یک زمینه تحقیقاتی با نام تشخیص احساس از روی گفتار معرفی شود که حالت احساسی گوینده را از روی گفتار مشخص کند. توجه به این نکته ضروری است که تشخیص احساس از روی گفتار می‌تواند برای استخراج معانی مفید از روی گفتار استفاده شود و از این جهت کارایی سیستم‌های تشخیص گفتار را بهبود

---

[1]دانشجوی کارشناسی ارشد مکاترونیک، واحد تهران جنوب، دانشگاه آزاداسلامی، تهران، ایران
[2]استاد، دانشکده فنی و مهندسی واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران
[3]نویسنده مسئول، دانشیار، گروه مهندسی مکانیک واحد پردیس، دانشگاه آزاد اسلامی، تهران، ایران arkhodayari@yahoo.com
تاریخ دریافت: ۹۵/۰۸/۰۵، تاریخ پذیرش: ۹۵/۱۱/۰۵

بخشد[۱]. گفتار انسان به صورت یک موج تولید شده توسط تارهای صوتی است که پارامترهای آن برای انتقال اطلاعات مدوله شده‌اند. ویژگی‌های فیزیکی و وضعیت روانی چگونگی پارامترهای این موج گوینده را تعیین می‌کند و در نتیجه انتقال اطلاعات گفتار در شرایط خود خواسته و یا ناخواسته، تحت تاثیر قرار می‌گیرند. این مساله بیانگر وجود الگوهایی در گفتار هنگام انتقال احساسات است.

این الگوها می‌توانند پایه‌های سیستم‌های تشخیص خودکار احساس انسان با استفاده از گفتار را تشکیل دهند. اهمیت چنین سیستم‌هایی با نیاز به بهبود سطح طبیعی بودن و بهره‌وری گفتار مبتنی بر واسط‌های ماشین-انسان افزایش یافته است[۵,۴,۳,۲]. زبان شناسان در تعریف احساس در گفتار دو نظریه ابعادی و نظریه گسسته را تعریف کرده اند. در نظریه ابعادی احساس به صورت دوبعدی [۶] و سه بعدی [۷] تعریف می شود. در نظریه گسسته احساسات به صورت کلاس‌های جداگانه در نظر گرفته می‌شوند.

زبان شناسان احساساتی که انسان درزندگی با آن‌ها مواجه می شود را تعریف کرده‌اند[۸] و مجموعه‌هایی توسط محققان که شامل ۳۰۰ حالت احساسی، ارائه شده‌اند[۹]. روش کلاس‌بندی شده از آنجایی که احساسات را به تعدادی کلاس محدود تقسیم می‌نماید[۱۰] در زمینه‌ی روش‌های مبتنی بر شناخت الگو مفید است. تعداد زیادی از کلاس‌های احساسی در تحقیقات مختلف معرفی شده‌اند [۱۱] و لیست جامعی از برچسب‌های احساسی مورد استفاده در تحقیقات مختلف بر مبنای مطالعات[۱۲] ارائه شده است.

کارهای زیادی در راستای ترکیب احساس با هوش مصنوعی [۱۳،۱۴] به منظور تسهیل در تعامل ماشین با انسان انجام شده است که نه تنها موجب بالا بردن قابلیت تشخیص صحیح احساس گوینده، بلکه باعث بهبود تکنولوژی بیان موثر احساس از طریق تغییرات صوتی و اشارات صورت و بدن نیز گشته است. در مرجع [۱۵] ربات طراحی شده می‌تواند احساسات بازیکنان را تشخیص و در حین بازی شطرنج با کودکان شکلک‌هایی را نشان دهد. سطح تعامل با کاربران و علاقه آنان را به طور چشمگیری افزایش می‌دهد.

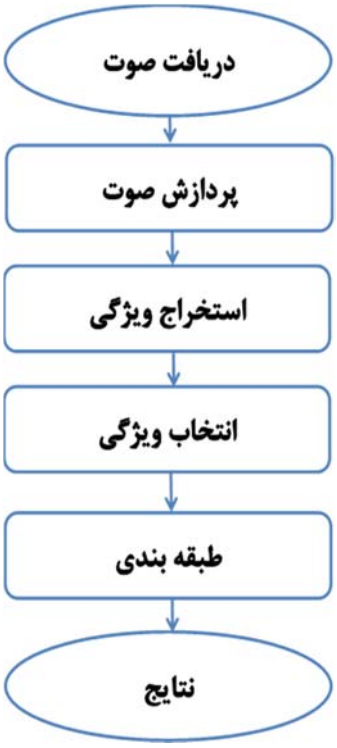
در مرجع [۱۶] با در نظر گرفتن پنج کلاس احساسی (طبیعی، خشم، اضطراب، ناآرامی و شادی) و چهار ترکیب مختلف، ویژگی‌های استخراج شده از اپراتور انرژی[1] و دو طبقه بند شبکه عصبی احتمالی[2] و مدل مخلوط گاوسی امتحان و مقایسه شده اند. در همه موارد دقت طبقه بندی برای زنان (۳۸٪ - ۶۲٪) و مردان (۳۲٪ - ۵۳٪) می‌باشد. به طور کلی در الگوریتم‌های پیشنهادی برای تشخیص احساس، سیگنال صدای مورد نظر از متن صدا جدا می شود و سپس از این سیگنال ویژگی‌های مورد نظر از صدا استخراج می شود و مورد تحلیل قرار می‌گیرد. شماتیک این الگوریتم در شکل (۱) نشان داده شده است. در الگوریتم تشخیص احساس، به طور کلی دو سازوکار برای انتخاب از بین ویژگی‌های موجود وجود دارد که عبارتند از روش‌های فیلتری و روش‌های لفاف[۱۷]. تفاوت اصلی بین این دو رویکرد در این است که انتخاب ویژگی مبتنی بر فیلتر، نتایج طبقه‌بندی را حساب نمیکند و انتخاب بهینه براساس افزایش همزمان اندازه تفکیک بین کلاس‌ها و کاهش اندازه تفکیک داخل از کلاس‌ها است. از طرف دیگر در روش لفاف، انتخاب بهینه براساس حداکثر سازی نرخ طبقه‌بندی صحیح است. در مرجع[۱۸]، برای طبقه‌بندی گفتارهای ثبت شده در یک مرکز تماس برای طبقه‌بندی خشم، نفرت، ترس، شادی، غم و تعجب از چهار روش مختلف طبقه بندی استفاده شده است.

---

[1] Teager
[2] PNN

در مرجع[۱۹]، از سیستم تشخیص احساس برای مقایسه چندین طبقه‌بند در گفتار تمیز و نویزی استفاده شده است. یک روش متداول در تشخیص احساس، استفاده از ویژگی‌ها و ابزارهای طبقه‌بندی می‌باشد. با این حال، یکی از موانع عمده در مقابل این رویکرد عدم مشخص بودن جهت گیری تحقیقات است. بسیاری از تحقیقات رایج در زمینه تشخیص احساسِ گفتار بر روی نمایش ارتباط بین دامنه مساله و تکنیک‌های کلاس-بندی متمرکز شده‌اند.

از طرفی دیگر کلاس‌بندهای رایج تقریبا در تمام سیستم‌های تشخیص احساس استفاده شده‌اند. که تاکنون هیچ کلاس‌بندی به عنوان کلاس‌بند برتر در این زمینه در بین محققان مورد توافق قرار نگرفته است. ولی ترکیب کلاس‌بندهای مختلف برای کلاس‌بندی احساسات کمتر مورد توجه قرار گرفته است. هدف اصلی ما در این مقاله ترکیب کلاس‌بندها به بهترین نحو ممکن برای کلاس‌بندی هر چه دقیق‌تر احساسات می‌باشد. در این مقاله الگوریتم جدیدی برای تشخیص احساس گوینده با استفاده از ویژگی‌هایی مبتنی بر ضرایب کپسترال فرکانسی مل[1] و روشی برای انتخاب ویژگی‌ها و چگونگی ترکیب قوی‌ترین کلاس‌بندهای حوزه تشخیص احساس معرفی شده است. مشخصه‌ی مهم در سیستم‌های تشخیص خودکار احساسات، مستقل از گوینده بودن آنها است. سیستم پیشنهادی مستقل از اطلاعات متنی سیگنال گفتار است.



شکل۱– روند کلی تشخیص احساس در گفتار

[1] Mel-frequency cepstral coefficients

۲- الگوریتم تشخیص احساس گوینده با تحلیل گفتار

در الگوریتم پیشنهادی در ابتدا صدای گوینده دریافت می‌شود و سپس عملیات پیش‌پردازش به منظور آماده کردن سیگنال صوت بر روی سیگنال ورودی انجام می‌شود و در مرحله بعدی، ویژگی‌های مورد نظر از سیگنال صوت استخراج شده و ویژگی‌های مناسب انتخاب شده و توسط طبقه‌بند و کلیشه تصمیم، نتیجه نهایی که احساس گوینده است، حاصل می‌شود. در ادامه عملیات و مراحل الگوریتم، توضیح داده می‌شود.

۲-۱- پیش پردازش

در ابتدا مرحله پیش پردازش انجام می شود. در این مرحله سیگنال پیش تاکید[1] می‌شود، بدین مفهوم که توسط یک فیلتر بالاگذر فرکانس‌های پایین آن حذف شده و فقط فرکانس‌های بالای آن باقی می‌ماند. این فیلتر، فیلتر پیش تاکید نام دارد. نتیجه این فیلتر حذف نویزهای مورد انتظار است. بطور کلی هر منبع صوتی که مزاحم صدای مورد نظر ما باشد، به آن صدای نویز می‌گویند. معمولا نویزی که توسط دستگاه الکترونیکی تولید و در نتیجه ذخیره می‌شود دارای گستره پهن و دارای اشکال تصادفی و فرکانس پایینی هستند. نویزهای فرکانس بالا نویزهایی ناشی از عدم سکوت در محیط هستند، که مورد انتظار نیستند.

نویزهای فرکانس بالا، از لحاظ فرکانس، همانند کلمات گفته شده می‌باشند، بنابراین اثرگذاری زیادی در معنا و تشخیص احساس دارند. بنابریان، وجود نویزهای بالا از محدودیت‌های این روش می‌باشد.

در الگوریتم پیشنهادی پاسخ ضربه این فیلتر از رابطه (۱) به دست می‌آید.

$$H(z) = 1 - \theta z^{-1} \qquad (۱)$$

که مقدار θ در این فیلتر نزدیک به یک (معمولا ۰,۹۵) انتخاب می‌شود.

۲-۲- استخراج ویژگی

در الگوریتم ارائه شده ویژگی‌های مورد نظر که ضرایب کپسترال فرکانسی مل هستند، از سیگنال استخراج می شود. ایده اصلی در استفاده از این ضرایب، الهام گرفتن از خواص شنیداری گوش انسان در دریافت و فهم گفتار است. یک مل[2] واحد اندازه گیری گام درک شده است و به طور خطی به فرکانس گام بستگی ندارد، زیرا عملکرد گوش انسان به گونه ای است که این فرکانس را به همان اندازه فیزیکی آن درک نمی‌کند. فرکانس سیگنال صوت و فرکانس مل مطابق رابطه (۲) با هم مرتبط هستند.

$$F_{mel} = 2595 Log_{10}^{\left(1+\frac{F_{Hz}}{700}\right)} \qquad (۲)$$

برای درک بهتر رابطه فوق، نمودار فرکانس مل بر حسب فرکانس اصلی در شکل (۲) نشان داده شده است. در قسمت استخراج ویژگی‌ها، در ابتدا برای هر تکه ای از هر سیگنال طیف فوریه[3] و دامنه آن با استفاده از روش سریع فوریه[4] محاسبه و استخراج می‌شود. سپس برای هر دامنه با استفاده از رابطه (۲) فرکانس مل محاسبه

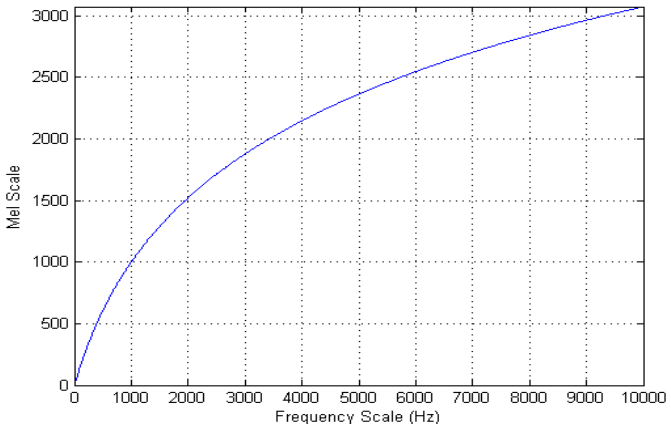
---

[1] Pre-emphasis signal
[2] Mel
[3] Fourier
[4] Fast Fourier transform

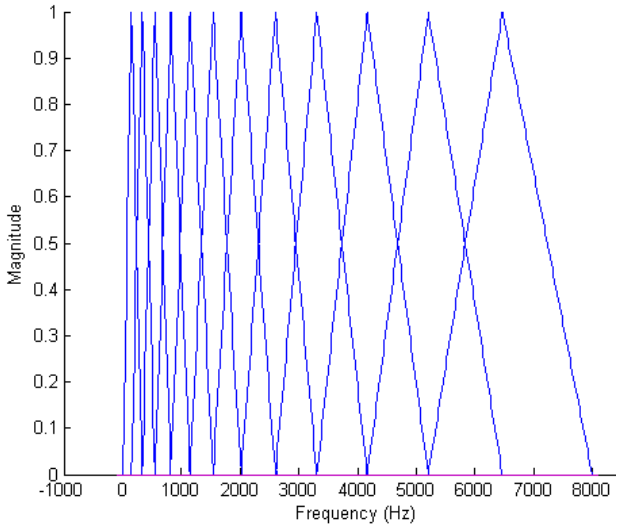
می‌شود. با اعمال این فرکانس‌ها به فیلتر بانک[1] و محاسبه خروجی این فیلتر، می‌توان با استفاده از رابطه (۳) زیر ضرایب کپسترال فرکانسی مل را محاسبه کرد.

$$c(i)=\sum_{j=1}^{F} Log\left(X_j\right)\cos\left(\frac{\pi i\left(j-0.55\right)}{F}\right) \qquad (۳)$$

که در آن F تعداد فیلترها و X_j خروجی حاصل از فیلتر j و $c(i)$ ضرایب حاصل است و N تعداد ضرایب را مشخص می‌کند که این مقدار در طراحی‌ها ۱۲ در نظر گرفته می‌شود.



شکل ۲ – نمودار فرکانس مل بر حسب فرکانس اصلی



شکل ۳– پهنای فیلترهای مل بر اساس فرکانس

[1] Filter bank

با توجه به اینکه حساسیت گوش انسان نسبت به تغییر فرکانس در فرکانس‌های بالا، کمتر از حساسیت آن در فرکانس‌های پایین است، در الگوریتم ارائه شده برای فرکانس‌های بالا، فیلترها با پهنای باند بزرگتری طراحی شده است. پهنای فیلترهای مل بر اساس فرکانس در شکل (۳) نمایش داده شده است.

۲-۳- انتخاب ویژگی

با انتخاب مناسب‌ترین زیرمجموعه، از مجموعه ویژگی‌های اصلی، می توان عملکرد طبقه‌بندی کننده را بهبود داده و از سوی دیگر باعث کاهش پیچیدگی محاسباتی شد. بنابراین انتخاب ویژگی‌های موثر در ساختار طبقه‌بندها همواره مورد توجه بسیار بوده است. بنابراین در مرحله انتخاب ویژگی با طراحی و استفاده از الگوریتم های تکاملی ویژگی های بهینه انتخاب می شوند.

با توجه به نوع تابع ارزیابی مورد استفاده روش‌های انتخاب ویژگی به دو گروه اصلی روش‌های فیلتری و روش‌های رپر[1] تقسیم می‌شوند که در روش فیلتری، تابع ارزیابی مستقل از الگوریتم داده کاوی اعمال شده است و در روش رپر الگوریتم داده کاوی تا حد معینی وابسته به تابع ارزیابی است. روش‌های رپر به علت داشتن نتایج دقیق‌تر به عنوان روش بهتر در مسایل یادگیری به رسمیت شناخته شده است اما باید در نظر داشت که پیچیدگی و زمان اجرای این روش در مقابل روشهای فیلتر بیشتر است[۲۱،۲۲].

در الگوریتم ارائه شده در این مرحله بر مبنای روش رپر و با تنظیم و طراحی خصوصیات زمان و دقت، ویژگی های بهینه استخراج می‌شود. بدیهی است در استفاده از روش رپر برای دستیابی به دقت بالا، زمان محاسبات افزایش می‌یابد و همچنین با کاهش زمان محاسبات، دقت کم می‌شود. در این مرحله هدف طراحی حداکثر دقت در حداقل زمان اجرا و محاسبات می‌باشد.

۲-۴- طبقه بندی ترکیبی

در این مرحله ویژگی‌های انتخاب شده با استفاده از طبقه‌بندهای مدل مخلوط گاوسی[2] و ماشین بردار پشتیبان[3] طبقه‌بندی می‌شوند و خروجی این طبقه‌بندها به عنوان ورودی به کلیشه تصمیم داده شده، نتیجه نهایی توسط کلیشه تصمیم ارائه می‌شود[۲۲]. ساختار این سیستم ترکیبی در شکل (۴) نشان داده شده است. برای بدست آوردن پارامترهای مدل مخلوط گاوسی، شامل وزن مخلوط‌های گاوسی و میانگین و کوواریانس توزیع ها از الگوریتم ماکزیمم نمودن امید ریاضی استفاده می‌شود.

۲-۴-۱- مدل مخلوط گاوسی

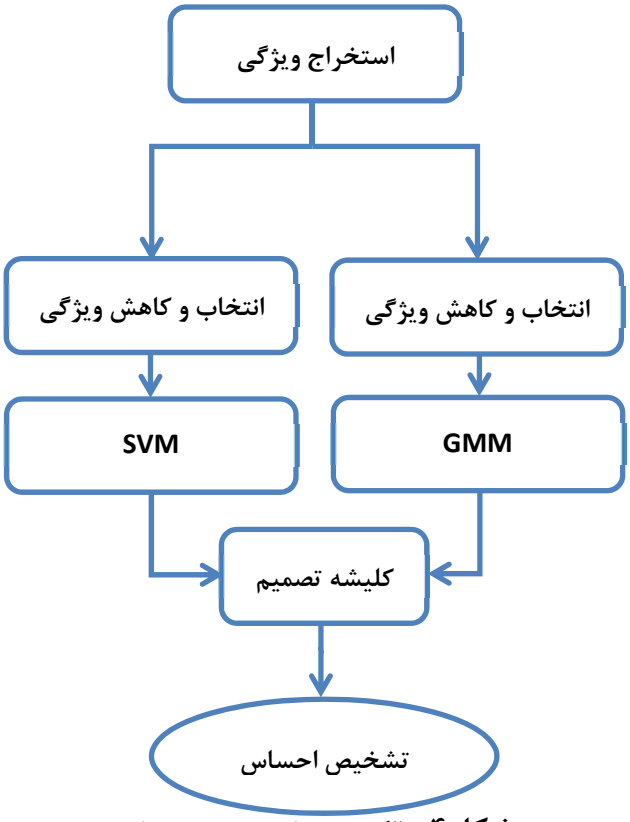
مدل مخلوط گاوسی یکی از مهمترین روش‌های مدل کردن سیگنال گفتار است که در واقع شبیه یک مدل مخفی مارکف یک حالته است که تابع چگالی احتمال آن حالت دارای چندین مخلوط نرمال می‌باشد. احتمال تعلق بردار آزمایشی x به یک مدل مخلوط گاوسی دارای M مخلوط به شکل (۴) بیان می‌شود.

---

[1] wrapper
[2] Bayesian Gaussian mixture
[3] Support vector machine

$$p(x|GMM) = \sum_{i=1}^{M} c_i . N(\mu_i , \Sigma_i)$$

(۴)



شکل ۴– الگوریتم ترکیب بند پیشنهادی

که در آن ci وزن مخلوط، μi و ∑i به ترتیب بردار میانگین و ماتریس کوواریانس توزیع نرمال هستند. ماتریس کواریانس مدل مخروط گوسی معمولا به صورت قطری در نظر گرفته می‌شود، گرچه امکان استفاده از ماتریس کامل نیز وجود دارد. رابطه فوق را می‌توان با استفاده از فرمول تابع چگالی احتمال نرمال به صورت زیر بیان می‌شود.

(۵)

$$p(x|GMM) = \sum_{i=1}^{M} c_i . \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_i - \mu_i)^T \sum_i (x_i - \mu_i)\right)$$

که در آن d بعد فضای ورودی‌ها است. برای بدست آوردن پارامترهای مدل مخروط گوسی، شامل وزن مخلوط-های گاوسی و میانگین و کوواریانس توزیع‌ها از الگوریتم ماکزیمم نمودن امید ریاضی استفاده می‌شود. باید توجه داشت که تعداد مخلوط‌های گاوسی با تعداد نمونه‌های موجود آموزشی رابطه مستقیم دارند و نمی‌توان با مجموعه داده‌ای ناچیز یک مدل مخروط گوسی دارای تعداد بیش از حد از مخلوط‌ها را آموزش داد.

۲-۴-۲- ماشین بردار پشتیبان

ماشین بردار پشتیبان در واقع یک طبقه‌بند دودویی است که دو کلاس را با استفاده از یک مرز خطی از هم جدا می‌کند. مرز تصمیم یک ماشین بردار پشتیبان طبق رابطه‌ی زیر تعریف می‌شود:

$$f(x) = b + \sum_{i=1}^{N} \alpha_i k(x_i, x) \qquad (۶)$$

دراین رابطه K تابع هسته و xi بردارهای پشتیبان نامیده می‌شود. بردارهای پشتیبان تعدادی نمونه آموزشی هستند که در محاسبه‌ی مرز تصمیم مورد استفاده قرار گرفته‌اند، برای یک نمونه‌ی ناشناخته‌ی x، f(x) را ارزش تصمیم می‌گویند. در صورتیکه f(x) منفی باشد، x به کلاس "۰" و در غیر اینصورت به کلاس "۱" تخصیص داده می‌شود. احتمال نمونه ناشناخته‌ی x به کلاس "۱" از رابطه‌ی زیر محاسبه می‌شود.

$$P_r(y = 1 | x) = \frac{1}{1 + \exp(Af + B)}, f = f(x) \qquad (۷)$$

در این رابطه A و B پارامترهای تابع سیگموید هستند که براساس حداکثر تفکیک نمونه‌های آموزشی محاسبه می‌شوند. در این مقاله ماشین بردار پشتیبان احتمالی به جای اختصاص نمونه ورودی به یک کلاس، احتمال تعلق آن به هر دو کلاس را محاسبه می‌کند. از ترکیب و جمع‌بندی این احتمال‌های دو به دو، احتمال نهایی تعلق یک نمونه به هر کدام از کلاس‌ها محاسبه می‌شود.

۲-۵- کلیشه تصمیم‌گیری

در مرحله آخر الگوریتم ارائه شده، همانطور که در شکل (۴) نشان داده است، یک سیستم هوشمند طراحی شده است که با توجه به ویژگی‌های استخراج شده از صوت، احساس درون آن را تشخیص و ارائه می‌دهد. این قسمت کلیشه تصمیم[1] نامیده می‌شود[۲۱]. این سیستم برای هر الگوی احساس، با توجه به خروجی طبقه بندها برای نمونه های آن الگو شکل می گیرد. کلیشه تصمیم طراحی شده برای این الگوریتم پیشنهادی، دارای یک خروجی است که می تواند شش الگوی احساسی را پوشش دهد. این الگوها عبارت از خشم، شادی، ترس، خستگی، نفرت و غم می‌باشند. ورودی‌های کلیشه تصمیم که در اینجا نتیجه طبقه بندها می باشد، با توجه به الگوهای مختلف احساس، تعداد متفاوتی دارد. ورودی ها از طبقه‌بند مخلوط گاوسی، حداکثر ۲۱ عدد و از ماشین بردار پشتیبان حداکثر ۱۶ عدد می باشد.

۳- پیاده‌سازی نرم‌افزاری و اعتبارسنجی

با هدف اعتبارسنجی الگوریتم ارائه شده، تمامی مراحل آن در نرم‌افزار MATLAB [2] پیاده سازی و شبیه سازی شده است. این شبیه سازی امکان بررسی نتایج حاصل از الگوریتم تشخیص احساس در گفتار گوینده را امکان پذیر می‌نماید. همچنین امکان مقایسه نتایج حاصل شده با سایر پژوهش‌های انجام شده در شرایط یکسان حاصل می‌شود. فرآیند انتخاب ویژگی در روش رپر با کلاس‌بند در ارتباط است، بنابراین فرآیند آموزش سیستم

---

[1] Decision template
[2] Matlab

در بخش انتخاب ویژگی انجام می‌شود. انتخاب بهترین روش آموزش و بهترین روش کاهش و انتخاب ویژگی، نیاز به بررسی عملکرد روش‌های موجود داشته است. بنابراین روش‌های موجود در برنامه MATLAB پیاده-سازی شد و برای انتخاب بهترین روش آموزش و بهینه‌سازی از بین روش‌های موجود از تمامی روش ها به همراه طبقه‌بندها تست گرفته شد و نتیجه آنها با یکدیگر مقایسه شد و روشی که بهترین نتیجه را داشت انتخاب شد. زمان آموزش و آزمون بعضی از روش‌ها بیش از ۷۲ ساعت به طول انجامید. نتایج حاصل از این آزمایشات در جدول زیر قابل مشاهده است. در هر یک از روش‌ها می‌توان، از یک یا چند مخروط گاوسی برای مدل کردن هر گوینده استفاده کرد. به تعداد این مخروط‌ها مرتبه روش گفته می‌شود. هر چه تعداد این مخروط‌ها بیشتر شود، دقت و زمان پردازش الگوریتم بیشتر می‌شود، همچنین به نمونه‌های بیشتری برای ساخت هر مدل احتیاج است. مرتبه هر یک از الگوریتم‌های آموزش به صورت عددی در جدول (۱) نشان داده شده است.

همان طور که در جدول (۱) نشان داده شده است، الگوریتم خفاش بهترین عمل کرد را دارد، بنابراین برای انتخاب ویژگی از الگوریتم خفاش استفاده شده است. پارامترهای مهم برای تشخیص بهترین روش بهینه‌سازی، بهترین درصد تشخیص احساس و بهینه‌ترین روش از نظر زمان بوده است.

بنابراین به دلیل زمان زیادی که الگوریتم خفاش-۴ نسبت به خفاش-۱ نیاز دارد تا به جواب مورد قبول برسد، از الگوریتم خفاش-۱ استفاده شد.

برای ارزیابی عملکرد الگوریتم پیشنهادی، نتیجه اعتبار سنجی الگوریتم با نتیجه الگوریتم های دیگر بر روی دو پایگاه داده مقایسه شد. به دلیل وابسته نبودن این الگوریتم به ویژگی‌های آوایی و دامنه و فرکانس سیگنال-های صوتی، این الگوریتم بر روی دو پایگاه داده به زبان‌های مختلف پیاده سازی شد که هر کدام از زبان ها از نظر آوایی و فرکانس و دامنه با زبان دیگر متفاوت بوده. به همین دلیل نتیجه پیاده‌سازی الگوریتم روی هر پایگاه‌داده با پایگاه‌داده دیگر مقایسه نشده است، بلکه چند الگوریتم روی هر پایگاه‌داده پیاده سازی شده است و با هم مقایسه شده اند.

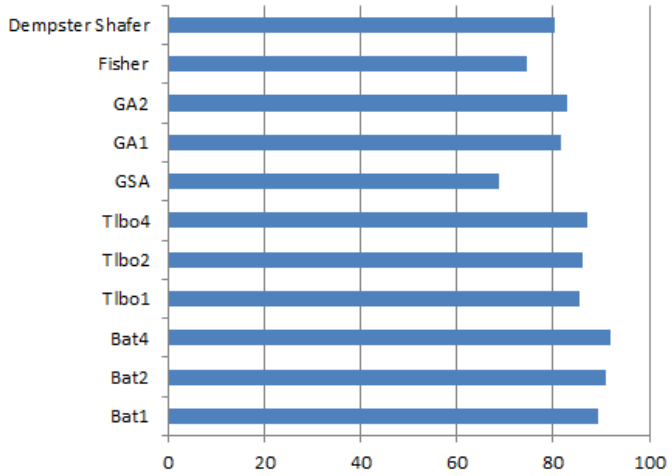
در اولین بررسی از پایگاه داده برلین[1] استفاده شده است[۲۴]. این پایگاه داده به دلیل اعتبار و کیفیت اطلاعات به عنوان پایگاهی ویژه بین محققان این حوزه مطرح است و مقایسه تحقیقات مختلف را ممکن می‌سازد. در این پایگاه ده بازیگر (۵ مرد و ۵ زن) هر یک ده جمله از جملات روزمره زبان آلمانی (۵ جمله کوتاه و ۵ جمله بلند با طول‌ای بین ۱,۵ تا ۴ ثانیه) بیان می‌کنند. جملات به گونه‌ای انتخاب شده‌اند که طبیعی بوده و به سادگی قابل بیان به فرم هفت احساس مختلف می‌باشند. گفتار ضبط شده دارای دقت ۱۶ بیت و نرخ نمونه برداری ۱۶ کیلو هرتز می‌باشد. پایگاه داده خام شامل تقریبا ۸۰۰ جمله با ۷ احساس شامل ۱۰ جمله و توسط ۱۰ بازیگر می‌باشد. فایل‌های صوتی توسط ۲۰ گوینده برای اطمینان از قابل تشخیص و طبیعی بودن آنها مورد آزمایش قرار می‌گیرند. سپس تنها فایل‌هایی که نرخ تشخیص بالای ۸۰ درصد داشته باشند و توسط بیش از ۶۰ درصد شنوندگان طبیعی تشخیص داده شده‌اند، انتخاب گشته‌اند. تعداد جملات برای هر دسته از احساسات در این پایگاه داده به شرح زیر است: خشم(۱۲۷)، خستگی(۵۰)، تنفر(۵۰)، ترس(۶۹)، خوشحالی(۷۱)، طبیعی(۷۹) و غم(۶۲).

---

[1] Berlin database

جدول۱– مقایسه نتیجه الگوریتم های مختلف انتخاب ویژگی

الگوریتم آموزش	زمان آموزش و آزمون	درصد تشخیص مرد	درصد تشخیص زن	درصد تشخیص میانگین
Bat1	۱۲ ساعت	۹۲	۸۷	۸۹/۵
Bat2	۲۲ ساعت	۹۲	۹۰	۹۱
Bat4	۳۶ ساعت	۹۳	۹۱	۹۲
Tlbo1	۱۶ ساعت	۸۶	۸۵	۸۵/۵
Tlbo2	۲۴ ساعت	۸۷	۸۵	۸۶
Tlbo4	۷۲ ساعت	۸۸	۸۶	۸۷
GSA	۲۴ ساعت	۶۹/۵	۶۸	۶۸/۷
GA1	۱۲ ساعت	۸۲	۸۱	۸۱/۵
GA2	۲۰ ساعت	۸۴	۸۲	۸۳
Fisher	۱۲ ساعت	۷۵/۵	۷۴	۷۴/۷
Dempster-Shafer	۱۸ ساعت	۸۰	۷۹	۸۰/۵



شکل۵ – نمودار مقایسه نتیجه الگوریتم های مختلف انتخاب ویژگی

در دومین بررسی از پایگاه داده ای به زبان فارسی با نام پایگاه داده احساسی فارسی درام[1] استفاده شد[۲۵]. این پایگاه داده از جملات موجود در نمایش های رادیویی برگرفته شده است، قطعه های موجود در نمایش های رادیویی با احساس های مختلف توسط نرم افزار از یکدیگر جدا شده و ذخیره شده اند. فرکانس نمونه برداری سیگنال ها برابر ۴۴ کیلو هرتز می باشد. هر چند که پایگاه داده مصنوعی به شمار می آید، عواملی نظیر نویز زمینه شامل صدای بوق ماشین، سر و صدای موجود در محیط و موسیقی متن باعث ایجاد شرایط طبیعی در این پایگاه داده شده است. تعداد جملات برای هر دسته از احساسات در این پایگاه داده به شرح زیر است: خشم(۱۷۷)، خستگی(۲۹)، تنفر(۱۷)، ترس(۶۳)، خوشحالی(۸۴)، طبیعی(۱۹۴) و غم(۱۴۵).

تمامی نتایج گزارش شده بر اساس روش اعتبار سنجی متقابل ارزیابی شده است. در این روش مجموعه‌ی نمونه‌های گفتار به صورت تصادفی به دسته‌های آموزش و تست تقسیم می‌شوند. این فرآیند چندین بار تکرار شده و دقت طبقه‌بندی به صورت تعداد نمونه‌هایی که به صورت صحیح تشخیص داده شده‌اند، به کل نمونه‌ها بدست می‌آید. سپس میانگین این دقت‌ها برای تعداد دفعات تکرار محاسبه و به عنوان دقت نهایی گزارش می‌شود. نتایج حاصل در جدول زیر آورده شده است. برای کاهش ابعاد بردارهای ویژگی استخراج شده از سیگنال، از روش لفاف با الگوریتم تکاملی خفاش استفاده شده که نتایج برای هر یک از ویژگی ها در جدول (۱) ارائه شده است. در جدول (۲) نتایج حاصل از سیستم پیشنهادی و الگوریتم پیشنهادی به تفکیک جنسیت نشان داده شده است. همان طورکه در جدول (۲) نشان داده در تمامی کلاس‌های احساس، جنسیت در تشخیص احساس موجود در صدا تاثیر دارد. دقت و عملکرد سیستم پیشنهادی را برای ارزیابی کیفیت و صحت الگوریتم پیشنهادی با نتیجه آزمایشات دانشگاه برلین بر روی تشخیص احساس از روی گفتار توسط کاربران[۲۴] و همچنین توسط الگوریتم‌های مشابه[۲۶،۲۷] بر روی پایگاه داده مشترک مقایسه شد. مقایسه نتایج در جدول (۳) نشان داده شده است.

جدول۲ – نتایج به تفکیک جنسیت و کلاس احساس

	مرد	زن	میانگین دقت کل
خشم	٪۸۴	٪۸۰	٪۸۲
شادی	٪۹۲	٪۸۵	٪۸۸/۵
ترس	٪۸۶	٪۸۲	٪۸۴
خستگی	٪۹۶	٪۹۲	٪۹۴
نفرت	٪۸۹	٪۸۳	٪۸۶
غم	٪۱۰۰	٪۱۰۰	٪۱۰۰
نرمال	٪۹۴	٪۹۰	٪۹۲
دقت کل	٪۹۲	٪۸۷	٪۸۹،۵

[1] Farsi Emotion database

شکل ۶ – نتایج به تفکیک جنسیت و کلاس احساس

جدول ۳– نتیجه آزمایشات پایگاه داده دانشگاه برلین

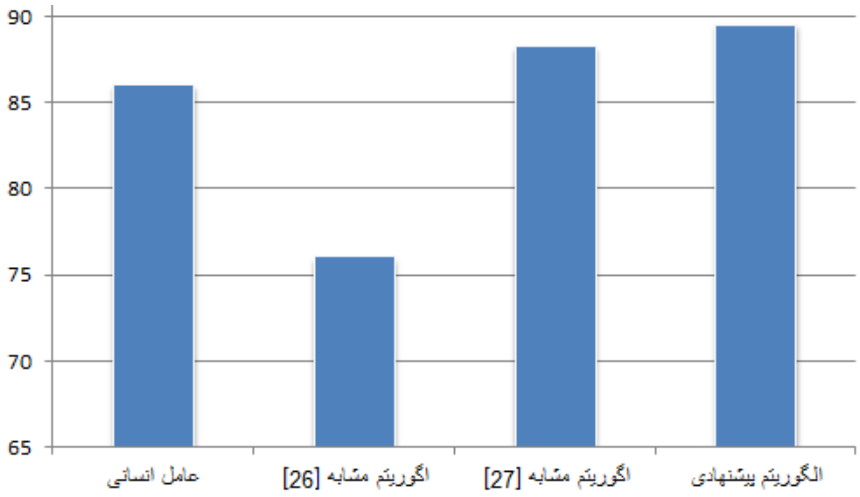
	درصد تشخیص توسط الگوریتم پیش‌نهادی	درصد تشخیص توسط الگوریتم مشابه[۲۶]	درصد تشخیص توسط الگوریتم مشابه[۲۷]	درصد تشخیص توسط عامل انسانی
خشم	٪۸۲	٪۹۰	٪۷۴	٪۹۶/۹
شادی	٪۸۸/۵	٪۷۰	٪۶۲	٪۸۸/۲
ترس	٪۸۴	٪۸۷	٪۶۶	٪۸۷/۳
خستگی	٪۹۴	-	٪۷۶	٪۸۶/۲
نفرت	٪۸۶	-	-	٪۸۳/۷
غم	٪۱۰۰	٪۱۰۰	٪۹۶	٪۸۰/۷
نرمال	٪۹۲	٪۹۴	٪۸۲	٪۷۹/۶
دقت کل	**٪۸۹/۵**	**٪۸۸/۲**	**٪۷۶**	**٪۸۶**

همچنین نتیجه پیاده سازی الگوریتم پیشنهادی بر روی پایگاه داده احساسی درام و مقایسه آن با الگوریتم مشابه [۲۵] در جدول (۴) نشان داده شده است.

همان طور که در جدول (۳) نشان داده شده است، با اینکه تعداد کلاس های احساسی که مورد ارزیابی قرار گرفته، توسط الگوریتم پیشنهادی این مقاله بیشتر از تعداد کلاس های احساسی دو الگوریتم دیگر است اما نتیجه بهتری نسبت به آنها داشته است.

این نکته قابل ذکر است که به دلیل وجود نویز و صداهای زمینه در پایگاه داده فارسی، درصدتشخیص درست احساس درتمامی کلاس های احساس پایین تر از پایگاه داده برلین شده است.

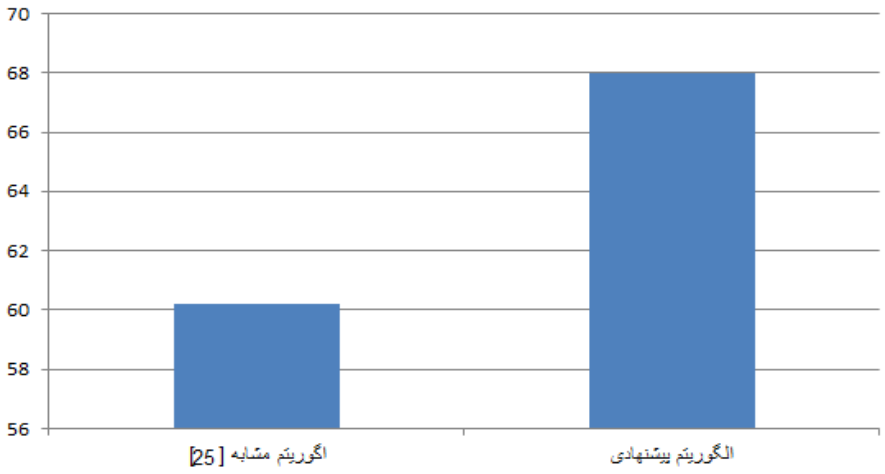
همان‌طور که در شکل (۷) و شکل (۸) و همچنین جدول (۳) و جدول (۴) نشان داده شده، الگوریتم پیشنهادی در این مقاله عملکرد قابل قبولی داشته است و اگر چه سیستم پیش‌نهادی در تشخیص برخی از الگوهای احساس دقت کمتری نسبت به الگوریتم های مشابه در تشخیص احساس دارد، ولی در تنوع تعداد احساسات و در میانگین دقت تشخیص صحیح ، عملکرد دقیق‌تر و بهتری دارد.



شکل ۷– نمودار مقایسه نتیجه الگوریتم‌های مختلف بر روی پایگاه‌داده دانشگاه برلین

جدول ۴– نتیجه آزمایشات پایگاه داده فارسی

درصد تشخیص توسط الگوریتم پیش‌نهادی	درصد تشخیص توسط الگوریتم مشابه[۲۵]	
٪۷۱	٪۷۲/۴	خشم
٪۶۳	٪۴۲/۵	شادی
٪۷۰	٪۴۶/۴	ترس
٪۶۸	٪۴۹/۱	غم
٪۶۵	٪۶۹/۴	نرمال
٪۶۸	**٪۶۰/۲**	دقت کل



شکل۸- نمودار مقایسه نتیجه الگوریتم های مختلف بر روی پایگاه داده فارسی

۴- جمع بندی

در این مقاله الگوریتم جدیدی برای تشخیص احساس از روی صدای انسان ارائه شد. این الگوریتم با استفاده از ویژگی‌هایی که از سیگنال صدا استخراج می‌شود و انتخاب زیرمجموعه‌ای از ویژگی‌ها بر اساس بهینه بودن سرعت و دقت موردنیاز و در ادامه با ترکیب طبقه‌بندها و عملیات طبقه‌بندی نوع احساس را مشخص می‌کند. نتایج حاصل از شبیه‌سازی و پیاده‌سازی این الگوریتم بر روی دو پایگاه داده به زبان‌های فارسی و آلمانی با نتایج الگوریتم‌های دیگر برای همان پایگاه داده ها مقایسه شد. عملکرد الگوریتم پیشنهادی برای پایگاه داده فارسی ۶۸٪ و برای پایگاه داده آلمانی ۸۹٪ بدست آمده است. نتایج حاصل از مقایسه عملکرد الگوریتم ارائه شده با سایر پژوهش‌ها در شرایط یکسان نشان می‌دهد، بنابراین می توان از این الگوریتم در تشخیص احساسات گوینده درطراحی سیستم های کنترلی و تعاملی بین انسان و ماشین استفاده نمود.

مراجع

[1] Nicholson, J., Takahashi, K., and Nakatsu, M., "Emotion Recognition in Speech using Neural Networks", Neural Computation, Vol. 1, pp. 9290–9296, (2000).

[2] Cowie, E., Campbell, N., Cowie, R., and Roach, P., "Emotional Speech: Towards a New Generation of Data Bases Original Research Article", Speech Communication, Vol. 40, pp. 33-60, (2003).

[3] Banse, R., and Scherer, K., "Acoustic Profiles in Vocal Emotion Expression", J. Pers. Soc. Psychol. Vol. 70, No. 3, pp. 614–636, (1996).

[4] Hozjan, V., and Kacic, Z., "Context-independent Multi Lingual Emotion Recognition from Speech Signal", Int. J. Speech Technol, Vol. 6, pp. 311–320, (2003).

[5] Kleinginna, Jr., and Kleinginna, A.M., "A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition", Motivation Emotion, Vol. 5, No. 4, pp. 345–379, (1981).

[6] Fernandez, R., "A Computational Model for the Automatic Recognition of Affect in Speech", Ph.D. Thesis, Massachusetts Institute of Technology, MIT Media Arts and Science, February, (2004).

[7] Bradley, M. M., and Lang, P. J., "Measuring Emotion: The Self-assessment Manikin and the Semantic Differential", Journal of Behavior Therapy & Experimental Psychiatry, Vol. 25, No. 1, pp. 49-59, (1990).

[8] Schubiger, M., "*English in to Nation: its form and Function*", Niemeyer, Tubingen, Germany, (1958).

[9] O'Connor, J., and Arnold, G., "*Intonation of Colloquial English*", Seconded. Longman, London, UK, (1973).

[10] Kim, D.H., "Fuzzy Rule Based Voice Emotion Control for user Demand Speech Generation of Emotion Robot", Computer Applications Technology (ICCAT), Germany, (2013).

[11] Sudhkar, R., "Analysis of Speech Features for Emotion Detection: A Review", International Conference on Computing Communication Control and Automation, Germany, (2015).

[12] Gharsellaoui, S., Selouani, S., and Dahmane, A., "Automatic Emotion Recognition using Auditory and Prosodic Indicative Features", Proceeding of the IEEE 28th Canadian Conference on Electrical and Computer Engineering, Halifax, Canada, (2015).

[13] Bosse, T., and Zwanenburg, E., "There's Always Hope: Enhancing Agent Believability Through Expectation-based Emotions", ACII 2009, 3rd International Conference, Amsterdam, Netherlands, pp. 1–8, (2009).

[14] Hudlicka, E., and Broekens, J., "Foundations for Modeling Emotions in Game Characters: Modelling Emotion Effects on Cognition", ACII 2009, 3rd International Conference, (2009).

[15] Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., and McOwan, Peter, W., "It's All in the Game: Towards an Affect Sensitive and Context Aware Game Companion", ACII 2009, 3rd International Conference, Amsterdam, Netherlands, pp. 1–8, (2009).

[16] Ling, He., Lech, M., Maddage, N., and Allen, N., "Emotion Recognition in Speech of Parents of Depressed Adolescents", Bioinformatics and Biomedical Engineering, ICBBE 2009, 3rd International Conference, Beijing, China, pp. 1–4, June (2009).

[17] Ververidis, D., Constantine, K., "Automatic Speech Classification to Five Emotional States Based on Gender Information", Signal Processing Conference, 12th European, pp. 341-344, (2004).

[18] Morrison, D., Wang, R., and Silva, L., "Ensemble Methods for Spoken Emotion Recognition in Call-centres Original Research Article", Speech Communication, Vol. 49, No. 2, pp. 98-112, (2007).

[19] Pao, T.L., Liao, V.Y., Chen, Y.T., and Yeh, J.H, "Comparison of Several Classifiers for Emotion Recognition from Noisy Mandarin Speech", Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kaohsiung City, Taiwan, November 26-28, pp. 21-26, (2007).

[20] Yang, X.S., "A New Metaheuristic Bat-inspired Algorithm", Nature Inspired Cooperative Strategies for Optimization (NICSO), Studies in Computational Intelligence, Vol. 284, pp. 65-74, (2010).

[21] Khan, K., and Sahai, A., "A Comparison of BA, GA, PSO, BP and LM for Training Feed forward Neural Networks in E-Learning Context", I.J. Intelligent Systems and Applications, Vol. 7, pp. 23-29, (2012).

[22] Kuncheva, L.I., "*Combining Pattern Classifiers: Methods and Algorithms*", John Wiley & Sons, (2004).

[23] Krogh, A., and Vedelsby, J., "Neural Network Ensembles, Cross Validation and Active Learning", Advances in Neural Information Processing Systems, Vol. 7, pp. 85-91, (1995).

[24] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B., "A Database of German Emotional Speech", Proceedings Inter Speech, Lissabon, Portugal, (2005).

[25] Mervi, H., and Esmailian, Z., "Showing up New Data Base for Detect Emotion from Speech", (2013).

[26] Esmailian, Z., and Marvi, H., "Recognition of Emotion in Speech using Variogram Based Features", Malaysian Journal of Computer Science, Vol. 27, No. 3, pp. 21-27, (2014).

[27] Ayadi, M., Kamel, M.S., and Karray, F., "Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models", IEEE, (2007).

فهرست نمادهای انگلیسی

A: پارامترهای تابع سیگموید

B: پارامترهای تابع سیگموید

C(i): ضرایب فرکانس مل

D: بعد فضای ورودی‌ها

F: فرکانس سیگنال مل

F(x): مرز تصمیم

K: تابع هسته

M: تعداد مخلوط

N: تعداد ضرایب

P: احتمال تعلق بردار

X_j: خروجی فیلتر

Θ: ضریب فیلتر

μi: بردار میانگین

$\sum i$: ماتریس کوواریانس

Abstract

The easiest method of communication between humans and machines is through speech and one of the essentials aspects of this relationship, is perception of humanistic sentiments by machine. As a result, getting speech's patterns and creating a system based on this model has been a challenge for researchers in recent years. Although the emotion shown in speech could ranges in a very divergent spectrum, because of pander to accent, culture and environment, but a fixed patterns could be found in feelings of people's speech.
In this paper, a new algorithm to detect emotion in human voice is presented.
In proposed algorithm, features are extracted from the audio signal, inspired by human hearing. and then the optimal features are chosen from the extracted features with the aim of increasing the speed and accuracy of diagnosis with an intelligent method.
In addition, classifying is done by combining a set classifiers subsequently, patterns of anger, joy, fear, boredom, disgust and sadness is distinguished by the designed intelligent system. Results of the simulations of the implemented algorithm is presented with two databases, Farsi and Germany and then compared with the outcomes of other algorithms with the same databases. Results indicate that the proposed algorithm could predict emotions of anger, joy, fears, boredom, disgust and sadness with good accuracy. This algorithm could be used in designing control systems and robot guidance. In addition, emotion recognition system could be utilized in psychology, medicine, and behavioral science and security applications such as polygraph.