بهبود نرخ تشخیص احساس از روی گفتار با استفاده از محدودیت تفکیک جنسیتی گویندگان

علی حریمی[۱،*]، خشایار یغمائی[۲]

اطلاعات مقاله

دریافت مقاله: ۱۳۹۱/۰۶/۱۴
پذیرش مقاله: ۱۳۹۴/۰۹/۲۳

واژگان کلیدی:
تشخیص احساس،
احساس در زنان و مردان،
الگوهای طیفی،
ویژگی‌های هارمونیکی.

چکیده

تشخیص احساس از روی سیگنال گفتار یکی از شاخه‌های نسبتاً جدید در پردازش گفتار می‌باشد که می‌تواند در تعامل انسان و روبات نقش مهمی ایفا کند. در این مقاله ضمن استفاده از دو نوع ویژگی طیفی جدید به منظور افزایش نرخ بازشناسی به بررسی تاثیر جنسیت گویندگان در تشخیص احساس پرداخته شده است. ویژگی‌های یاد شده با استفاده از روش‌های پردازش تصویر، از تصویر طیف‌نگاره سیگنال گفتار استخراج می‌شوند . در این تحقیق به منظور جداسازی احساس‌های مختلف از یکدیگر از طبقه‌بند مرتبه ای استفاده شده است. به منظور بهینه سازی ساختار این طبقه‌بند، ابتدا جداپذیرترین کلاس‌ها از هم جدا می‌شوند تا خطای ایجاد شده در مراحل اولیه طبقه‌بندی حداقل بوده و این خطا در الگوریتم منتشر نشود. سیستم پیشنهادی بر روی پایگاه داده‌ی آلمانی برلین آزمایش شده است. بر اساس نتایج بدست آمده نرخ تشخیص برای گویندگان مختلط ٪۴۳/۴ می‌باشد که این مقدار پس از تفکیک گویندگان بر اساس جنسیت به ٪۸۲/۸۶ افزایش پیدا می‌کند. نرخ تشخیص برای گویندگان زن ٪۸۳/۰۵ و برای مردان ٪۸۲/۶۱ بدست آمده است.

۱- مقدمه

با توجه به پیشرفت روز افزون تکنولوژی و نیاز ارتباط گسترده انسان و ماشین، تلاش‌های زیادی برای افزایش دقت و کیفیت این سیستم ها صورت گرفته است. از آنجا که گفتار یکی از مهمترین روش‌های ارتباط بین انسان ها می‌باشد، تحقیقات زیادی برای ارتباط انسان و ماشین از طریق گفتار انجام شده است [۱]. از دیدگاه مخابرات، انتقال اطلاعات بوسیله گفتار را می‌توان یک سیستم مخابراتی در نظر گرفت که در آن گوینده اطلاعات را کد نموده و اطلاعات از طریق کانال مخابراتی به شنونده منتقل می‌شود. این کانال مخابراتی را می‌توان در دو بخش جداگانه بررسی نمود. بخش اول مربوط به انتقال اطلاعات زبانی[۳] گفتار است و بخش دوم مربوط به انتقال اطلاعات فرازبانی[۴] می‌باشد. اطلاعات زبانی مربوط به آن دسته از اطلاعاتی است که به محتوای گفتار و اینکه چه مطالبی بیان شده است مربوط می‌باشند اما اطلاعات فرا زبانی به چگونگی و کیفیت ارسال این اطلاعات نظیر سن، جنس، احساس و بیماری‌های احتمالی گوینده که می‌توان از نحوه ی گفتار گوینده به آن‌ها پی برد مربوط می‌شوند [۲]. در رابطه با انتقال اطلاعات زبانی بین انسان و ماشین (سیستم‌های تشخیص

*. پست الکترونیک نویسنده مسئول: a.harimi@gmail.com
۱. دانشگاه آزاد اسلامی، واحد شاهرود، دانشکده برق، شاهرود، ایران
۲. دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران

[3] Linguistic information
[4] Paralinguistic information

گفتار) تلاش‌های بسیار زیادی صورت گرفته است، اما با توجه به پیچیده بودن موضوع در مورد اطلاعات فرازبانی محققان هنوز با چالش‌هایی در این زمینه مواجه می‌باشند. در این میان تشخیص احساس[۵] و حالات روحی گوینده یکی از جالب ترین مواردی است که نظر بسیاری از محققین را به خود جلب‌کرده است. ارتباط مؤثرتر انسان و ماشین، افزایش نرخ تشخیص در سیستم‌های تشخیص گفتار و استفاده در روانپزشکی از کاربردهای تشخیص احساس بوسیله سیگنال گفتار می‌باشند.

مساله تشخیص احساس از روی گفتار را از دیدگاه تشخیص الگو می‌توان به سه بخش استخراج ویژگی، انتخاب ویژگی و طبقه‌بندی تفکیک‌کرد. مهم‌ترین چالش پیش‌رو در این زمینه مربوط به مرحله استخراج ویژگی می‌باشد. در این راستا پژوهشگران ویژگی‌های زیادی را معرفی کرده‌اند. اما هنوز اتفاق نظر واحدی در این زمینه بدست نیامده و گزارش‌های متناقضی وجود دارد که بیانگر نیاز به انجام تحقیقات بیشتر در این راستا می‌باشد. با توجه به وابستگی احساس به زبان، فرهنگ و گوینده [۳]، استخراج ویژگی‌هایی که مستقل از این موارد در تفکیک احساس‌های مختلف کارآمد باشند مساله‌ای حائز اهمیت است. این ویژگی‌ها را می‌توان در چهار گروه مورد بررسی قرارداد. اولین گروه ویژگی‌های عروضی[۶] هستند که عمدتاً از آنالیز سیگنال در حوزه زمان بدست آمده و اغلب بر پایه ی فرکانس گام و انرژی می‌باشند. این ویژگی ها که تا حد زیادی مرتبط با لحن گفتار می‌باشند از مهمترین ویژگی‌هایی هستند که در مقالات مربوطه مورد استفاده قرار گرفته‌اند [۱۷-۳]. گروه دوم ویژگی‌های طیفی هستند که از محتویات فرکانسی سیگنال استخراج می‌شوند، از اینرو بعنوان مکمل ویژگی‌های عروضی مورد استفاده قرار می‌گیرند [۱]. در گروه‌های سوم و چهارم، ویژگی‌های مرتبط با کیفیت گفتار[۷] و برخی دیگر از ویژگی ها[۸] قرار می‌گیرند که در مقالات مختلف بعنوان ویژگی‌هایی مؤثر در بازشناسی احساس معرفی شده‌اند [۱، ۵ و ۱۷]. پژوهش‌هایی نیز در زمینه تأثیر روش‌های انتخاب ویژگی و طبقه‌بندی در عملکرد سیستم بازشناسی احساس انجام شده است. بر اساس تحقیقات انجام شده، LDA[۹] [۴] و SFS[۱۰] [۲، ۴، ۱۱، ۱۶ و ۱۸] دو نمونه از مهمترین روش‌های انتخاب ویژگی هستند که در مقالات مرتبط مورد استفاده قرارگرفته‌اند. در این مقالات الگوریتم‌های طبقه‌بندی مورد استفاده نیز عمدتاً بر پایه‌ی SVM[۱۱] [۲، ۴، ۶، ۷، ۱۰و ۱۱] و HMM[۱۲] [۳] طراحی شده‌اند.

یکی از موضوعات جالب توجه پژوهشگران تحقیق در مورد تفاوت‌ها و چگونگی بیان احساس در زنان و مردان می‌باشد. بر اساس گزارشات اطلاعات جنسیتی گویندگان در نتایج طبقه‌بندی مؤثر می‌باشد [۲]. در این تحقیق ضمن استفاده از ویژگی‌های طیفی جدید به منظور ارتقاء سیستم به بررسی تأثیر جنسیت گویندگان در نرخ بازشناسی احساس خواهیم پرداخت.

در ادامه‌ی مقاله، در بخش ۲ به مدولاسیون احساس در گفتار می‌پردازیم. پس از آن در بخش ۳ پایگاه داده‌ی[۱۳] مورد استفاده در این مقاله را معرفی‌خواهیم‌کرد. بخش‌های ۴ و ۵ به چگونگی استخراج و انتخاب ویژگی اختصاص داده شده‌اند. پس از آن در بخش ۶ الگوریتم طبقه‌بند معرفی می‌شود. نتایج در بخش ۷ آورده شده‌اند و در بخش ۸ نیز به نتیجه‌گیری پرداخته شده است.

۲- مدولاسیون احساس در گفتار

احساس‌های مختلف انسان را می‌توان در دو گروه احساس‌های اولیه[۱۴] و احساس‌های اجتماعی[۱۵] بررسی کرد. احساس‌های اولیه احساس‌هایی ذاتی به شمار می‌آیند و بدون اینکه انسان در مورد استفاده از آن‌ها آموزشی دیده

[5] Emotion recognition
[6] Prosodic features
[7] Voice quality features
[8] Other features
[9] Linear Discriminant Analysis
[10] Sequential Forward Selection
[11] Support Vector Machine
[12] Hidden Markov Model
[13] Database
[14] Primitive (basic) emotions
[15] Social emotions

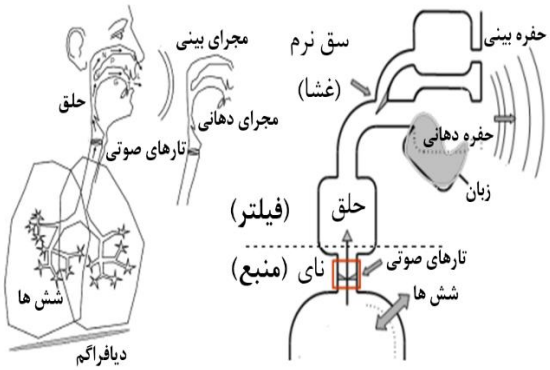
باشد، در زندگی روزمره از آن‌ها استفاده می‌کند. این احساس‌ها عموماً نوعی ابزار دفاعی بشمار آمده و در جوامع مختلف به شکل تقریباً مشابهی بروز می‌کنند [۲۶-۱۹]. ترس، عصبانیت، ناراحتی، انزجار و خوشحالی در زمره ی این احساس ها به شمار می‌آیند. از سوی دیگر احساس‌های اجتماعی احساس‌هایی هستند که اکتسابی بوده و نحوه‌ی بروز آن‌ها به آداب و رسوم و فرهنگ یک جامعه وابسته است [۸، ۱۹، ۲۷ و ۲۸]. این دسته از احساس‌ها را می‌توان بصورت ترکیبی از احساس‌های اولیه درنظرگرفت [۲۹ و ۳۰]، از اینرو در اغلب پژوهش‌های مرتبط محققان در پی بازشناسی احساس‌های اولیه بوسیله ی سیگنال گفتار می‌باشند.

محققان بر این باورند که احساس‌های اولیه و اجتماعی بترتیب بوسیله نیمکره‌های راست و چپ مغز کنترل می‌شوند [۲۶]. از طرفی سیستم اعصاب نباتی [۱۶](ANS) نیز از نیمکره راست مغز فرمان می‌گیرد [۳۱]، بدین ترتیب می‌توان ارتباطی بین احساس‌های اولیه و سیستم اعصاب نباتی در نظر گرفت. سیستم اعصاب سمپاتیک و سیستم اعصاب پاراسمپاتیک دو بخش مهم سیستم اعصاب نباتی می‌باشند که بسیاری از فعالیت‌های حیاتی انسان را کنترل می‌کنند. بطور مثال ضربان قلب، فشار خون، شش ها، بسیاری از غدد مانند غدد ترشح کننده مایعات مخاطی و بزاق از جمله اندام‌هایی هستند که بوسیله‌ی سیستم اعصاب نباتی کنترل می‌شوند.

سیستم اعصاب سمپاتیک که در مواقع اضطراری وارد عمل می‌شود تحت تأثیر احساس‌هایی نظیر عصبانیت، ترس، هیجان و نظایر آن‌ها فعال می‌شوند. در این هنگام ضمن ترشح آدرنالین، ضربان قلب، فشار خون، تحرکات شش ها، میزان ترشح مایعات مخاطی و بزاق افزایش یافته و در حالت‌های بحرانی‌تر لرزش اندام‌های بدن را در پی خواهد داشت [۳۲]. گفتار تولید شده در این حالت نیز از تغییرات ایجاد شده در اندام‌های بدن تأثیر می‌پذیرد. در این حالت انرژی سیگنال گفتار و نرخ تغییرات آن افزایش پیدا می‌کند. همچنین در اثر تغییرات ایجاد شده در اندام‌های تولید گفتار نظیر تارهای صوتی، فرکانس ارتعاش نیز تحت تأثیر قرار می‌گیرد. در چنین شرایطی افزایش بزاق و مایعات مخاطی باعث ایجاد تغییرات در جنس مجرای دهانی و مجرای بینی و در پی آن گفتار تولید شده خواهد شد. گفتار تولید شده در این حالت بلند و سریع بوده و فرکانس گام آن ضمن اینکه تغییرات بیشتری دارد، مقدار متوسط بیشتری هم خواهد داشت. این عوامل باعث افزایش قدرت مؤلفه‌های هارمونیکی بالاتر در سیگنال گفتار حاصله خواهد شد [۱].

در مقابل سیستم اعصاب سمپاتیک، سیستم اعصاب پاراسمپاتیک باعث ایجاد آرامش شده و شرایط را برای استراحت آماده می‌کند [۳۲]. در چنین شرایطی انرژی سیگنال گفتار کمتر بوده و فرکانس گام آن نیز تغییرات کمتری خواهد داشت. همچنین انرژی سیگنال در این حالت بیشتر در مؤلفه‌های پایین تر متمرکز خواهد بود [۱].

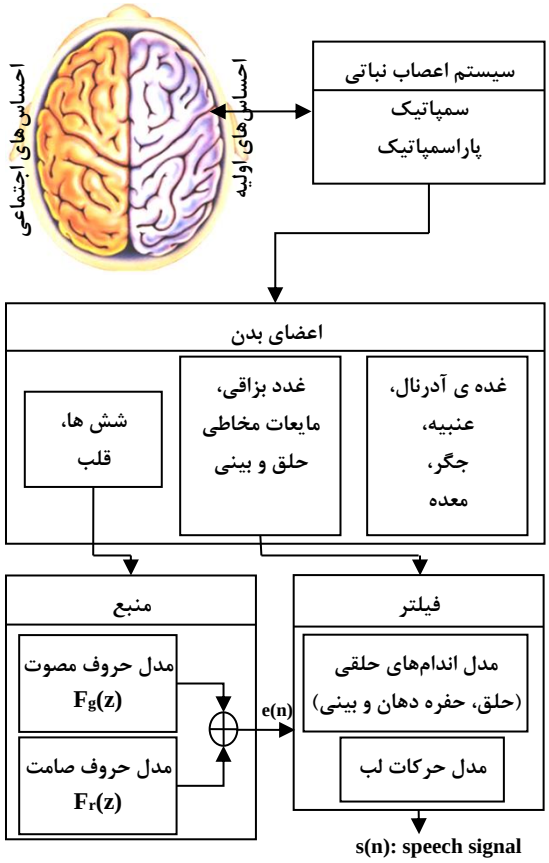
شکل ۱ سیستم تولید گفتار انسان را نشان می‌دهد. همانطور که در این شکل مشاهده می‌شود سیستم تولید گفتار از یک منبع و یک فیلتر تشکیل شده است. منبع شامل شش ها و تارهای صوتی می‌باشد که هوای مرتعش را به حنجره می‌دمد. فیلتر نیز شامل حلق، مجرای دهانی، مجرای بینی و متعلقات آن‌ها می‌شود.



شکل ۱. سیستم تولید گفتار انسان و اجزای آن.

[16] Autonomic Nervous System

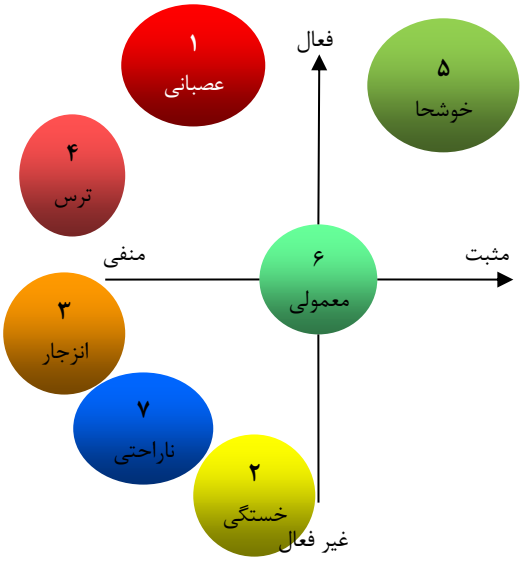
شکل ۲ نیز چگونگی تأثیر احساس‌های مختلف را بر اعصاب سمپاتیک و پاراسمپاتیک و نیمکره‌های مغز و در پی آن تأثیر آن بر اجزای تولید گفتار را نشان می‌دهد.



شکل ۲. ارتباط بین احساس‌های مختلف، مغز و اجزای سیستم تولید گفتار

احساس‌های مختلف را می‌توان توسط مدل دو بعدی نشان داده شده در شکل ۳ بیان نمود. در این شکل محور افقی درجه مثبت یا منفی بودن یک احساس و محور عمودی میزان فعالیت انجام شده برای بروز آن را نشان می‌دهد. بر اساس این مدل و مطالبی که در این بخش بیان شد می‌توان گفت که احساس‌های فعال مانند عصبانیت و ترس باعث تحریک اعصاب سمپاتیک شده، احساس‌های غیر فعال نظیر خستگی و ناراحتی اعصاب پاراسمپاتیک را تحریک می‌کنند.

تحقیقات زیادی در مورد تفاوت‌های زنان و مردان در ابراز احساسات انجام شده است. بر اساس گزارش ارائه شده در [۳۳] زنان در درک و ابراز احساسات محرک تر از مردان عمل می‌کنند. آن‌ها احساس خود را با شدت و فرکانس بالاتری بیان می‌کنند. از طرفی مردان احساسات خود را با کنترل بیشتری ابراز می‌کنند.



شکل ۳. مدل دو بعدی احساس

۳- پایگاه داده‌های مربوطه

پایگاه داده‌های مربوط به بازشناسی احساس بوسیله ی سیگنال گفتار را می‌توان در دو گروه مختلف بررسی کرد. پایگاه داده‌های مصنوعی[17] که معمولاً توسط بازیگران حرفه ای تهیه و پایگاه داده‌های طبیعی[18] که از ارتباط و گفتار روزمره مردم عادی تشکیل شده است. در پایگاه داده‌های مصنوعی از آنجا که گوینده با هدف گنجاندن یک احساس خاص در گفتار جمله را ادا می‌کند، نرخ تشخیص معمولاً بالاتر بوده و راحت تر می‌توان احساس گوینده را تشخیص داد. اما در پایگاه داده‌های طبیعی که از مکالمات روزمره مردم عادی جمع آوری می‌شود، نرخ تشخیص پایین تر است. متأسفانه دسترسی به اینگونه پایگاه داده ها به دلیل رعایت حریم خصوصی افراد معمولاً میسر نمی‌باشد [۸].

در [۱] لیست کاملی از پایگاه داده‌های رایجی که در این زمینه مورد استفاده قرار می‌گیرند گردآوری شده است. در این تحقیق از پایگاه داده برلین استفاده شده است [۳۴].

[17] Acted database

[18] Natural database

این پایگاه داده که به زبان آلمانی تهیه شده است شامل ۵۳۵ جمله می‌باشد که توسط ۱۰ بازیگر حرفه ای (۵ مرد و ۵ زن) با ۷ احساس مختلف بیان شده‌اند. اطلاعات مربوط به این پایگاه داده در جدول ۱ ارائه شده است.

جدول ۱- تعداد جملات مربوط به هر احساس در پایگاه داده برلین

احساس	گویندگان زن	گویندگان مرد
عصبانیت	۶۷	۶۰
خستگی	۴۶	۳۵
انزجار	۳۵	۱۱
ترس	۳۲	۳۷
خوشحالی	۴۴	۲۷
عادی	۴۰	۳۹
ناراحت	۳۷	۲۵

۴- استخراج ویژگی

در این مقاله از انرژی هارمونیک‌های سیگنال گفتار و الگوهای طیفی[19] به عنوان مکملی برای ویژگی‌های عروضی استفاده نموده ایم. جدول ۲ ویژگی‌هایی را که در این مقاله مورد استفاده قرار گرفته اند نشان می‌دهد.

در این تحقیق به منظور استخراج ویژگی‌های مربوط به پارامترهای مختلف سیگنال از ۱۱ معیار آماری استفاده کرده ایم. این معیارها عبارتند از: میانگین، انحراف معیار، واریانس، میانه، کشیدگی[20]، چولگی[21]، حداقل، حداکثر، تفاضل حداکثر و حداقل، ممان سوم[22] و لرزش[23]. این ۱۱ تابع به منحنی فرکانس گام[24] و انرژی و مشتق اول آن‌ها اعمال شده و ۴۴ ویژگی استخراج می‌شوند. نسبت زمانی آواهای مصوت به صامت هم به عنوان ۴۵ امین ویژگی عروضی مورد استفاده قرار گرفته است. منحنی مربوط به ۱۲ ضریب اول MFCC و مشتق اول آن‌ها را محاسبه نموده و از ۱۱ معیار آماری مطرح شده برای استخراج ویژگی استفاده می‌شود که در پی آن ۲۶۴ ویژگی حاصل می‌شود.

[19] Spectral patterns
[20] Kurtosis
[21] Skewness
[22] Third moment

۲۹ ویژگی نیز مربوط به انرژی هارمونیک ها و ۳۰۶ ویژگی مربوط به الگوهای طیفی است که در ادامه بصورت مفصل بیان خواهند شد.

جدول ۲- ویژگی‌های مورد استفاده در این مقاله.

گروه ویژگی	تعداد ویژگی	نوع ویژگی
ویژگی‌های عروضی	۲۲	فرکانس گام
	۲۲	انرژی
	۱	زمان بندی
ویژگی‌های طیفی	۲۶۴	ضرایب MFCC
	۲۹	انرژی هارمونیک ها
	۳۰۶	الگوهای طیفی
سایر ویژگی ها	۲۲	نرخ عبور از صفر
	۲۲	شیب سیگنال در عبور از صفر
	۲۲	عملگر انرژی تیگر

منحنی‌های مربوط به نرخ عبور از صفر، شیب سیگنال هنگام عبور از صفر، عملگر انرژی تیگر[25] و مشتق اول آن‌ها محاسبه شده و از ۱۱ معیار آماری بیان شده برای استخراج ویژگی از این منحنی ها استفاده می‌کنیم. بدین ترتیب ۶۶ ویژگی نیز در این مرحله حاصل خواهد شد.

اغلب ویژگی‌های یاد شده در پژوهش‌های مرتبط مورد استفاده قرار گرفته اند. اما در این میان ویژگی‌های مربوط به انرژی هارمونیک ها و الگوهای طیفی ویژگی‌هایی می‌باشند که در آن‌ها را در [۳۵] معرفی نموده و کارآمد بودن آن‌ها را نشان داده ایم.

۴-۱- الگوهای طیفی

همانطور که در بخش ۲ اشاره شد، احساس گوینده در بسیاری از پارامترهای سیگنال گفتار تأثیرگذار می‌باشد. بطور مثال در احساس‌هایی با سطح فعالیت بالا مقدار میانگین و همچنین نوسانات فرکانس گام افزایش پیدا

[23] Shimmer
[24] Pitch frequency
[25] Teager energy operator

می‌کند. علاوه بر آن در چنین شرایطی انرژی هارمونیک‌های بالای سیگنال بیشتر از حالت‌های دیگر می‌باشد. از طرفی در احساس‌های با سطح فعالیت پایین انرژی سیگنال بیشتر در هارمونیک‌های پایین تر متمرکز خواهد شد و تغییرات منحنی فرکانس گام و انرژی کمتر می‌شود. مجموعه عوامل یاد شده باعث می‌شوند که تأثیر احساس گوینده در تصویر طیف‌نگاره سیگنال گفتار نمایان شود. تصویر طیف‌نگاره توزیع انرژی سیگنال گفتار را در راستای محور زمان و فرکانس نشان‌می‌دهد. در این تصویر هر پیکسل در راستای افقی بیانگر یک فریم از سیگنال گفتار می‌باشد. ارزش فرکانسی هر پیکسل هم از تقسیم بازه فرکانسی بر تعداد سطرهای تصویر قابل محاسبه است. شکل ۴ تصویر طیف‌نگاره مربوط به یک جمله را که با سه احساس مختلف توسط یک گوینده بیان شده است نشان می‌دهد.

به منظور استخراج ویژگی از تصویر طیف‌نگاره، هر کدام از این تصاویر را به ۶ تصویر تجزیه می‌کنیم که هر کدام از این تصاویر بیانگر نوع خاصی الگو در تصویر اصلی می‌باشند. الگوهای مورد نظر خطوطی با زوایای ۰، ۴۵، ۶۳، ۹۰، ۴۵- ، ۶۳- درجه می‌باشند. با محاسبه توزیع آماری این الگوها در باندهای فرکانسی مختلف در تصویر ویژگی‌های مختلف را استخراج می‌کنیم. بدین منظور در مرحله پیش پردازش مقدار متوسط سیگنال حذف شده و مؤلفه‌های فرکانس بالا با استفاده از فیلتر پیش تأکید تقویت می‌شوند. رابطه ی فیلتر پیش تأکید را می‌توان بوسیله ی رابطه ی زیر بیان نمود [۳۶].

$$s'_n = s_n - \alpha s_{n-1}, 0.96 \le \alpha \le 0.99 \qquad (۱)$$

که در آن s_n و s'_n، بترتیب nامین نمونه از سیگنال اصلی و سیگنال فیلتر شده را نشان می‌دهند.$\propto$ نیز معمولاً مقداری نزدیک به یک و کوچکتر از آن دارد.

سپس تصویر طیف‌نگاره برای سیگنال مربوطه بدست آمده و برای مشخص‌تر کردن باندهای انرژی در این تصویر کنتراست آن را افزایش می‌دهیم.

افزایش کنتراست تصویر باعث می‌شود مکان‌هایی از تصویر که انرژی بیشتری دارند نمود بیشتری پیدا کرده و در محاسبات مؤثرتر واقع شوند و در مقابل اثر مناطقی که انرژی کمتری دارند کم‌تر شده یا حذف شود.

در مرحله بعد به منظور واضح‌تر کردن الگوهای تشکیل دهنده‌ی تصویر از فیلترهایی با ابعاد ۳×۳ استفاده می‌کنیم. این فیلترها در شکل ۵ نشان داده شده‌اند.

فیلتر H1 تا H8 بترتیب برای مشخص تر کردن الگوهای ۰، ۴۵، ۹۰، ۴۵-، ۶۳، ۶۳- و ۶۳- استفاده می‌شوند. بدین ترتیب با اعمال این ۸ فیلتر به تصویر طیف‌نگاره ۸ تصویر مختلف بدست خواهد آمد که در هر کدام از این تصاویر حضور یک الگوی خاص پر رنگ‌تر نمایان می‌کند. تصاویر بدست آمده را باینری نموده و با استفاده از عملگرهای مرفولوژی نویز تصویر را حذف می‌کنیم. سپس با استفاده از روش تطبیق الگو، در هر کدام از تصاویر بدست آمده به جستجوی الگوی مورد نظر می‌پردازیم. در صورت تطبیق الگو، در مکان متناظر با آن یک و در صورت عدم تطبیق الگو، در آن مکان صفر در نظر گرفته می‌شود. الگوهای مورد نظر در شکل ۶ نشان داده شده اند.

همانطور که در این شکل مشاهده می‌شود برای هر کدام از زوایای ۶۳ و ۶۳- درجه دو الگوی مختلف وجود دارد. در این تحقیق ما با استفاده از عملگر OR این دو تصویر را با هم ادغام می‌کنیم. بدین ترتیب ۶ تصویر باینری داریم که هر کدام بیانگر وجود یک نوع الگوی خاص در تصویر می‌باشند. در آخرین مرحله این تصویرها را بر اساس فرکانس‌های بحرانی گوش انسان به ۱۷ باند فرکانسی مختلف تقسیم‌بندی می‌کنیم. در نتیجه ۱۰۲ تصویر باینری مختلف خواهیم داشت که هر کدام مربوط به یک الگوی خاص در یک باند فرکانسی خاص هستند. ویژگی‌های مورد نظر از این ۱۰۲ تصویر استخراج می‌شوند.

سه نوع ویژگی از این تصاویر استخراج می‌شود:

۱) میانگین تعداد الگوها در یک فریم: در هر کدام از تصاویر تعداد الگوهای منطبق شده را محاسبه نموده و آن‌ها را بر

تعداد ستون‌های تصویر (تعداد فریم‌های سیگنال اصلی) تقسیم می‌کنیم.

۲) تعداد نسبی هر کدام از الگوها به کل الگوها: تعداد الگوهای منطبق شده را در هر کدام از تصاویر محاسبه کرده و مقدار حاصل را به تعداد الگوهای منطبق شده در همه‌ی تصاویر تقسیم می‌کنیم.

۳) تعداد نسبی هر کدام از الگوها در یک باند فرکانسی خاص به تعداد همان الگو در تمام باند فرکانسی: تعداد الگوهای منطبق شده را در هر کدام از تصاویر محاسبه نموده و آن را بر مجموع تعداد الگوهای منطبق شده مربوط به یک الگوی خاص در تمام باندهای فرکانسی مختلف تقسیم می‌کنیم. چگونگی انجام این پروسه در شکل ۷ نشان داده شده است.

۴-۲- انرژی هارمونیک ها

با توجه به سیستم تولید گفتار که در شکل ۱ نشان داده شده است، هوا از سمت شش‌ها به سمت حلق دمیده می‌شود. تارهای صوتی که در مسیر قرار دارند در این حالت شروع به ارتعاش می‌کنند. این ارتعاش بر روی محتوای فرکانسی گفتار حاصله تأثیر گذار خواهد بود بطوریکه فرکانس اصلی نوسانات در سیگنال گفتار که با فرکانس گام شناخته می‌شود و هارمونیک اصلی گفتار را تشکیل می‌دهد مقدراش برابر با فرکانس نوسان تارهای صوتی می‌باشد. بر اساس مدل نشان داده شده در شکل ۲ احساس هر فرد از طریق سیستم اعصاب اتوماتیک و بوسیله تارهای صوتی در گفتار شخص مدوله می‌شود [۳۷]. بدین ترتیب می‌توان گفت که فرکانس ارتعاش تارهای صوتی تحت تأثیر احساس گوینده می‌باشند.

با توجه به اینکه فرکانس ارتعاش تارهای صوتی را به عنوان هارمونیک اصلی در نظر می‌گیریم، سایر هارمونیک‌های سیگنال گفتار ضرایب این فرکانس خواهند بود. بر اساس آنچه که در بخش ۲ بیان شد، انرژی هارمونیک ها با احساس گوینده مرتبط هستند. در روش متداول برای محاسبه انرژی هارمونیک ها از یک بانک فیلتری متغیر با زمان استفاده می‌شود. فرکانس مرکزی فیلترهای تشکیل دهنده این بانک فیلتری منطبق بر ضرایب فرکانس گام بوده و پهنای باند هر فیلتر نیز برابر فرکانس گام در نظر گرفته‌می‌شود. در این روش مشخصات فیلترها در هر فریم به روز رسانی می‌شود. شکل ۸ چگونگی محاسبه انرژی h هارمونیک اول را به این روش نشان می‌دهد.

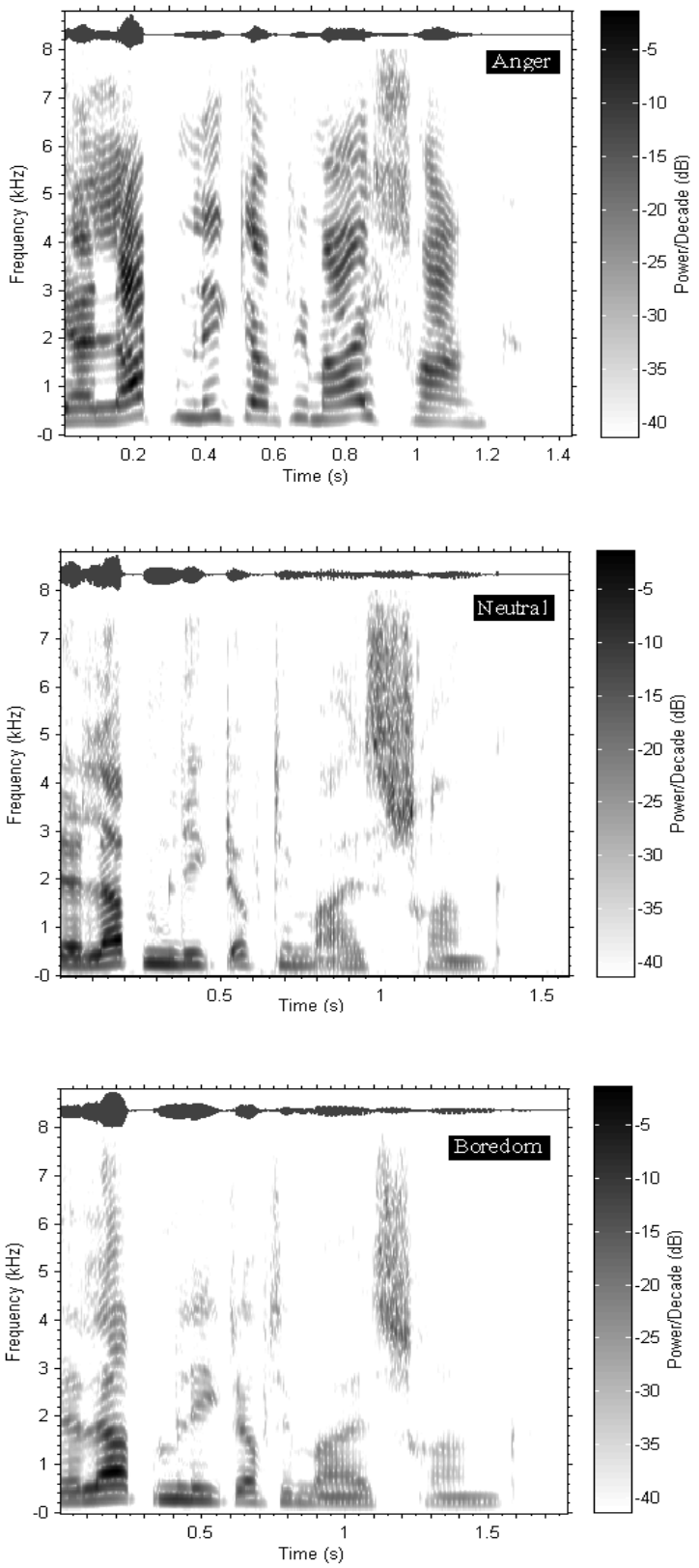
این تحقیق برای محاسبه انرژی هارمونیک ها روش جدیدی ارائه شده است که در کاهش حجم محاسبات بسیار مؤثر می‌باشد. در این روش انرژی هارمونیک ها را با در استفاده از تصویر طیف‌نگاره که در مرحله قبل نیز مورد استفاده قرار گرفته بود محاسبه می‌کنیم.

بدین منظور در تصویر طیف‌نگاره موقعیت مربوط به فرکانس مرکزی فیلترها و پهنای باند هر فیلتر را محاسبه نموده و سطح روشنایی پیکسل‌هایی را که در محدوده پهنای باند هر فیلتر هستند با هم جمع می‌کنیم. بدین ترتیب انرژی هر هارمونیک محاسبه می‌شود. لازم به ذکر است که در تصویر طیف‌نگاره سطح روشنایی پیکسل‌های تصویر متناسب با انرژی سیگنال اصلی در زمان و فرکانس متناظر می‌باشد. شکل ۹ چگونگی انجام این پروسه را بر روی تصویر طیف‌نگاره نشان می‌دهد.

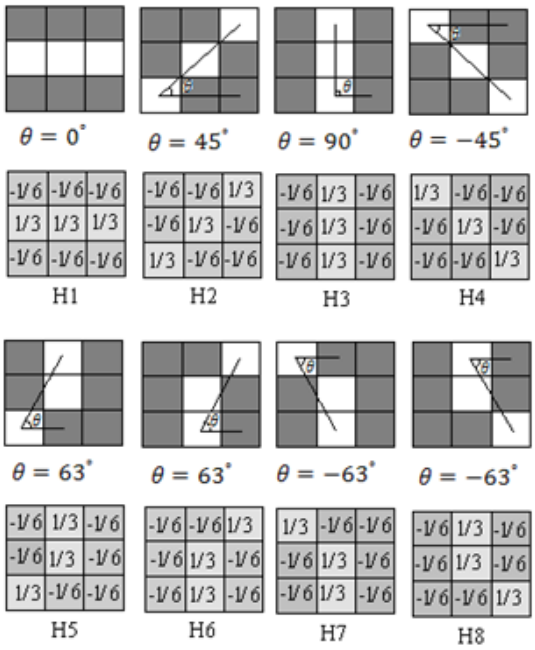
در این شکل فرکانس مرکزی فیلترهای اول، پنجم، نهم و سیزدهم با خط چین نشان داده شده است و محدوده ی پهنای باند نیز اطراف همین خط چین با خطوط ممتد مشخص شده است. لازم به ذکر است که این محاسبات فقط برای فریم‌های مصوت انجام می‌شود. برای بدست آوردن مکان هر کدام از این فرکانس ها در تصویر ابتدا باید ارزش فرکانسی هر پیکسل در راستای محور فرکانس محاسبه شود. بدین منظور$I(n\times m)$ را بعنوان تصویر طیف‌نگاره در نظر می‌گیریم.

در این تصویر ارزش فرکانسی هر پیکسل را می‌توان توسط رابطه زیر محاسبه نمود.

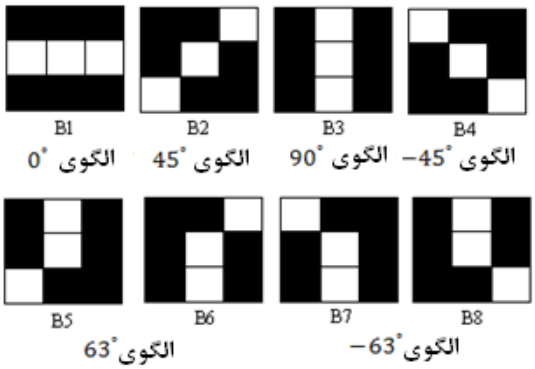
$$PFV = \frac{f_s / 2}{n} \qquad (۲)$$



شکل ۴. تصویر طیف‌نگاره مربوط به یک جمله که توسط یک زن با سه احساس (الف) عصبانیت، (ب) معمولی و (ج) خستگی بیان شده است



شکل ۵. فیلترهایی که برای نمایان تر کردن الگوها استفاده می‌شوند



شکل ۶. الگوهایی که در تصاویر باینری جستجو می‌شوند

که در این رابطه n تعداد سطرهای تصویر و f_s فرکانس نمونه برداری سیگنال گفتار می‌باشد. در این تحقیق فرکانس نمونه برداری 16khz می‌باشد. فرکانس مرکزی هر کدام از هارمونیک ها نیز از رابطه زیر محاسبه می‌شود.

$$Fc_h(i) = h \times F0(i) \times PFV, 1 \le i \le m \qquad (۳)$$

که در این رابطه $Fc_h(i)$ شماره سطر مورد نظر را در فریم یا ستون i ام برای هارمونیک hام مشخص می‌کند. $F0(i)$ نیز فرکانس گام را در فریم i ام نشان می‌دهد. بنابراین پیکسل‌های مشخص کننده فرکانس مرکزی در تصویر را می‌توان با مختصات $I(Fc_h(i).i)$ نشان داد که در این رابطه i شماره فریم سیگنال اصلی یا بعبارتی شماره ستون تصویر است $1 \le i \le m$ و h شماره هارمونیک مورد نظر را مشخص می‌کند. بدیهی است که در این رابطه داریم: $1 \le Fc_h(i) \le n$.

شماره سطر مربوط به فرکانس قطع پایین و قطع بالای فیلتر ها در ستون iام در تصویر نیز از روابط زیر قابل محاسبه می‌باشد.

$$Fs_{h,1}(i) = Fc_h(i) - \frac{(F0(i) \times PFV)}{2} \qquad (۴)$$

و

$$Fs_{h,2}(i) = Fc_h(i) + \frac{(F0(i) \times PFV)}{2} \qquad (۵)$$

بنابراین مختصات مربوط به پیکسل‌های مشخص کننده فرکانس قطع پایین $I(Fs_{h,1}(i), i)$ و مختصات مربوط به پیکسل‌های مشخص‌کننده فرکانس قطع بالا $I(Fs_{h,2}(i), i)$ خواهندبود. بدین ترتیب انرژی هارمونیک h ام در فریم i ام از رابطه زیر قابل محاسبه است.

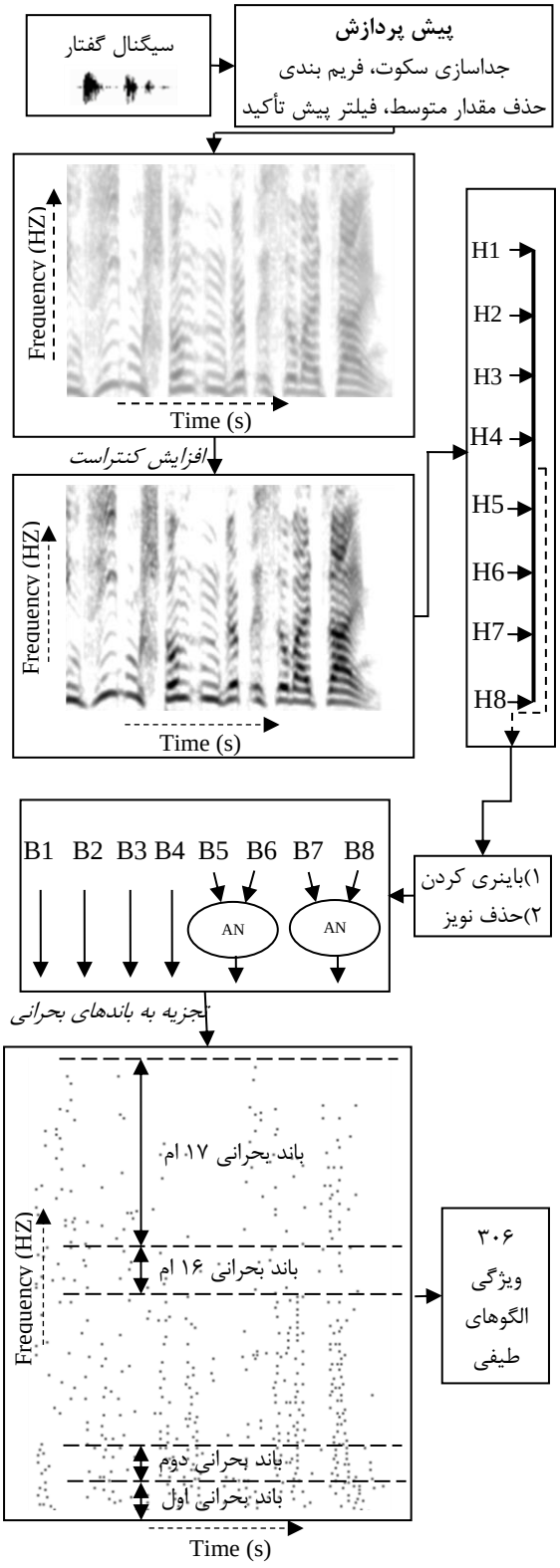
$$E_h(i) = \sum_{j=Fs_{h,1}(i)}^{Fs_{h,2}(i)} I(i, j), 1 \le i \le m \qquad (۶)$$

و انرژی متوسط هارمونیک hام نیز از رابطه زیر محاسبه می‌شود.

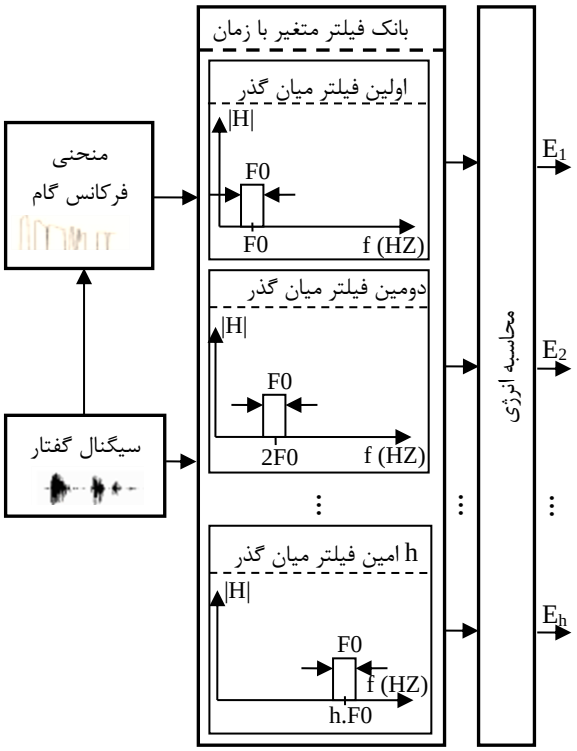
$$NE_h = \frac{1}{m} \sum_{i=1}^{m} E_h(i) \qquad (۷)$$

در این تحقیق با بدست آوردن انرژی متوسط ۱۳ هارمونیک اول یک بردار ویژگی شامل ۲۹ ویژگی مختلف طراحی شده است که توسط رابطه زیر قابل بیان است.

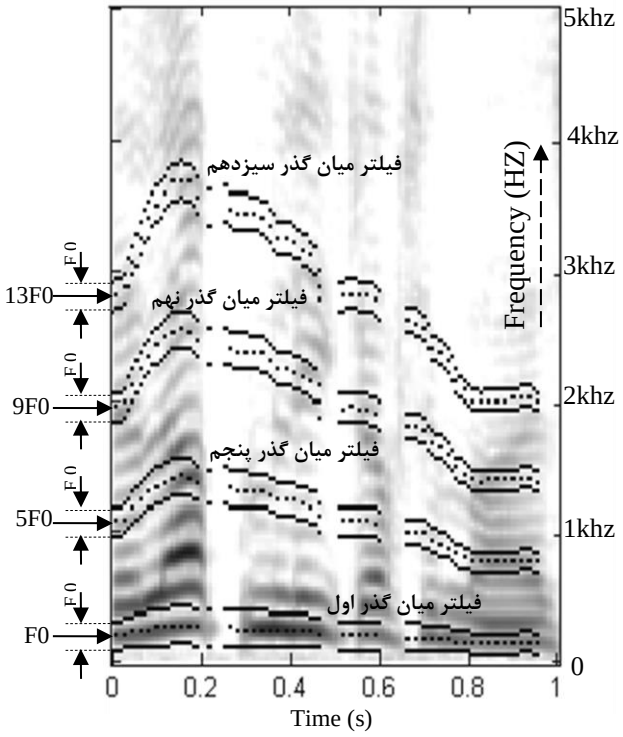
$$HFV = [NE_1, NE_2, ..., NE_{13}, \frac{NE_2}{NE_1}, \frac{NE_3}{NE_2}, ..., \frac{NE_{13}}{NE_{12}}, \frac{NE_1 + NE_2 + NE_3}{NE_4 + NE_5 + NE_6}, \frac{NE_4 + NE_5 + NE_6}{NE_7 + NE_8 + NE_9}, \frac{NE_7 + NE_8 + NE_9}{NE_{10} + NE_{11} + NE_{12} + NE_{13}}, \frac{NE_1 + NE_2 + NE_3}{NE_{10} + NE_{11} + NE_{12} + NE_{13}}] \qquad (۸)$$



شکل ۷. شماتیک استخراج ویژگی‌های مبتنی بر الگوهای طیفی بر اساس روش پیشنهادی



شکل ۸. پروسه ی محاسبه انرژی h هارمونیک اول سیگنال گفتار بوسیله ی بانک فیلتری متغیر با زمان



شکل ۹. چگونگی محاسبه انرژی هارمونیک ها بوسیله ی تصویر طیف‌نگاره

۵- انتخاب ویژگی

المان‌های بردار ویژگی که در این تحقیق استفاده شده است را می‌توان در جدول ۲ مشاهده نمود. مناسب‌ترین ویژگی‌ها در این میان ویژگی‌هایی هستند که کمترین فاصله ی درون کلاسی و بیشترین فاصله ی بین کلاسی را ایجاد نمایند. به بیان دیگر ویژگی‌های خوب ویژگی‌هایی هستند که با استفاده از چنین ویژگی‌هایی داده‌های یک کلاس تا حد امکان به یکدیگر شبیه بوده و در عین حال داده‌های دو کلاس مختلف فاصله کافی با یکدیگر داشته باشند. بسیاری از ویژگی‌هایی که از سیگنال مورد نظر استخراج شده‌اند ممکن است این شرایط را نداشته باشند. از طرف دیگر تعداد زیاد ویژگی ها در بردار ویژگی ممکن است باعث پیچیدگی مساله شده، کار بهینه سازی را در الگوریتم طبقه‌بند مختل کنند. این امر باعث کاهش نرخ طبقه‌بندی می‌شود. بنابراین انتخاب ویژگی‌های مؤثر در راستای تفکیک احساس‌های مختلف امری ضروری می‌باشد. بدین منظور FDR بعنوان معیاری برای سنجش مؤثر بودن یک ویژگی در طبقه‌بندی با رابطه ی زیر بیان می‌شود [۴].

$$FDR(u)=\frac{2}{c(c-1)}\sum_{c_1}\sum_{c_2}\frac{(\mu_{c_1,u}-\mu_{c_2,u})^2}{\sigma_{c_1,u}^2+\sigma_{c_2,u}^2},\quad 1\le c_1<c_2\le C \qquad (۹)$$

که در این رابطه $\mu_{c_i,u}$ و $\sigma_{c_i,u}^2$ بترتیب میانگین و واریانس ویژگی uام برای کلاس iام می‌باشند. C نیز در این رابطه تعداد کل کلاس ها را مشخص می‌کند. همانطور که در این رابطه پیداست، تفکیک‌پذیری کلاس‌ها دو به دو بررسی شده و میانگین آن‌ها به عنوان معیاری برای مناسب بودن ویژگی مورد نظر بیان می‌شود. در این تحقیق FDR برای همه ی ۷۱۰ ویژگی محاسبه شده از آن میان فقط ۳۰۰ ویژگی که دارای بالاترین مقادیر FDR باشند انتخاب می‌شوند.

در میان این ویژگی ها ممکن است برخی از آن‌ها تکراری بوده و یا اینکه وابستگی زیادی به هم داشته باشند، در اینصورت حتی اگر مقدار FDR برای این ویژگی ها عدد بزرگی باشد، عملاً تعداد زیاد آن‌ها نه تنها کمکی به حل مساله نخواهد کرد بلکه بر درجه پیچیدگی مساله خواهد افزود. بدین منظور از یک الگوریتم مکمل با نام جستجوی ترتیبی رو به جلو (SFS) بهره خواهیم برد. در این روش از میان ویژگی‌هایی که در مرحله قبل انتخاب شده اند، بهترین ویژگی که بیشترین مقدار FDR را به خود اختصاص داده است را انتخاب می‌کنیم. سپس سایر ویژگی ها را یکی یکی به بردار ویژگی می‌افزاییم. در صورتی که ویژگی اضافه شده باعث افزایش نرخ تشخیص شود، این ویژگی باقی خواهد ماند در غیر این صورت آن را حذف می‌کنیم. این پروسه برای همه ۳۰۰ ویژگی انتخاب شده از مرحله قبل انجام می‌شود.

لازم به ذکر است که می‌توانستیم از ابتدا فقط از الگوریتم SFS استفاده کنیم، اما با توجه به پیچیده‌تر بودن این الگوریتم نسبت به محاسبه ی FDR، به منظور کاهش حجم محاسبات ابتدا ویژگی‌های نویزی را بوسیله الگوریتم FDR حذف و در نهایت مؤثرترین ویژگی ها در بالا بردن نرخ تشخیص بوسیله ی SFS انتخاب شدند.

۶- طبقه‌بندی

طبقه‌بندی آخرین مرحله سیستم بازشناسی احساس است. در این مرحله بهترین ویژگی ها برای بازشناسی احساس انتخاب شده اند و فقط باید احساس مورد نظر با توجه به ویژگی‌های استخراج شده تعیین شود. طبقه‌بندی شامل یک مرحله آموزش و یک مرحله تست می‌باشد. در مرحله آموزش با توجه به ویژگی‌های انتخاب شده و برچسب کلاس‌ها، مرزهای بین کلاس‌ها مشخص می‌شوند و در مرحله تست با توجه به اینکه ویژگی‌های استخراج شده در کدام ناحیه قرار می‌گیرند کلاس مورد نظر (احساس مورد نظر) مشخص می‌شود. در این تحقیق ۸۰٪ جملات پایگاه داده بصورت تصادفی انتخاب شده و برای آموزش الگوریتم به کار گرفته شده اند. ۲۰٪ باقی مانده نیز تا آخرین مرحله برای تست الگوریتم دست نخورده باقی می‌ماند. با توجه به اینکه در بسیاری از پژوهش‌های مرتبط طبقه‌بند SVM بعنوان یک طبقه‌بند کارآمد معرفی شده است ما نیز در این

تحقیق طبقه‌بندی مرتبه ای مبتنی بر SVM طراحی می‌کنیم. در این روش از چند کلاس بند SVM باینری بهره می‌بریم. بدین صورت که ابتدا پایگاه داده را به دو قسمت تقسیم ، سپس هر کدام از این دو قسمت دوباره به دو قسمت تقسیم می‌شوند. این پروسه تا آنجا ادامه پیدا می‌کند که همه کلاس ها از هم تفکیک شوند. با توجه به اینکه در پایگاه داده برلین که در این تحقیق بررسی می‌کنیم ۷ احساس مختلف وجود دارد، با یک مساله ی ۷ کلاسه مواجه هستیم. این ۷ کلاس ابتدا به یک کلاس ۴ تایی و یک کلاس ۳ تایی تقسیم می‌شوند و این پروسه تا تفکیک همه ی کلاس ها از هم ادامه خواهد یافت. حال اینکه کدام ۴ احساس در یک گروه و کدام ۳ احساس در گروه دیگر قرار گیرند مساله ای است که باید حل شود. از آنجا که خطای این مرحله در کل الگوریتم انتشار می‌یابد، دقت این مرحله از طبقه‌بندی بسیار حائز اهمیت است، پس بهتر است کلاس ها را به گونه‌ای دسته‌بندی کنیم که بهترین تفکیک پذیری را داشته باشند. در بسیاری از مقالات ساختار مؤثر برای طبقه‌بند بصورت تجربی پیشنهاد شده است [۱۳]. در این مقاله به منظور بهینه سازی الگوریتم طبقه‌بند، از FDR بعنوان معیاری برای تفکیک پذیری کلاس‌ها استفاده‌کرده‌ایم. بدین ترتیب طبقه‌بند به گونه‌ای طراحی شده است که تفکیک پذیر ترین کلاس ها در گام‌های اول از هم جدا شوند تا از پخش شدن خطا در ساختار طبقه‌بند جلوگیری شود.

در بخش ۷ بصورت دقیق تر در مورد ساختار طبقه‌بندهای طراحی شده و نتایج بدست آمده بحث خواهیم کرد.

۷- نتایج آزمایش‌ها

از آنجا که یکی از اهداف مهم این تحقیق بررسی تأثیر تفکیک جنسیتی گویندگان در نرخ تشخیص احساس می‌باشد، سیستم مورد نظر را در دو آزمایش مختلف ارزیابی می‌کنیم. در آزمایش اول نتایج تفکیک احساس را برای تلفیق گویندگان زن و مرد ارائه می‌کنیم. در دومین آزمایش پس از تفکیک گویندگان زن و مرد و طراحی سیستم بازشناسی احساس برای این گویندگان نتایج مربوطه ارائه شده‌اند. در [۳۸] الگوریتمی برای تفکیک گویندگان زن و مرد در پایگاه داده برلین با دقت ۱۰۰٪ ارائه شده است. می‌توان از این الگوریتم در اولین مرحله طبقه‌بندی به منظور تفکیک جنسیتی گویندگان استفاده نمود و پس از آن طبقه‌بندهای جداگانه را برای گویندگان زن و مرد به کار برد.

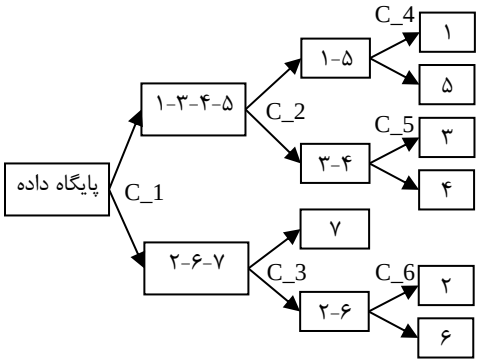
در آزمایش اول، در اولین گام طبقه‌بندی هدف تفکیک ۷ احساس مختلف به دو کلاس است، یک کلاس شامل ۴ احساس و کلاس دیگر شامل ۳ احساس. در این راستا $\binom{7}{3} = 35$ انتخاب متفاوت خواهیم داشت. FDR را برای همه ی ویژگی ها در هر ۳۵ حالت ممکن محاسبه می‌کنیم. از بین این ۳۵ انتخاب، حالتی که بیشترین مقدار FDR را به خود اختصاص داده باشد بر می‌گزینیم. لازم به ذکر است که در هر حالت برای محاسبه FDR با یک مساله دو کلاسه مواجه می‌باشیم که هر کلاس تلفیقی از چند احساس می‌باشد. در جدول ۳، سه انتخاب برتر که بالاترین مقدار FDR برای آن‌ها بدست آمده است آورده شده است. از این پس بطور قرار دادی مطابق با آنچه در شکل ۳ نشان داده شده است، هر احساس را با یک کد معرفی می‌کنیم به این صورت که از اعداد یک تا هفت به ترتیب برای احساس‌های عصبانیت، خستگی، انزجار، ترس، خوشحالی، عادی و ناراحتی استفاده خواهد شد.

جدول ۳- مقادیر حداکثر FDR برای سه انتخاب برتر برای اولین مرحله طبقه‌بندی

شماره انتخاب	کلاس ۱	کلاس ۲	مقدار حداکثر FDR
انتخاب اول	۱، ۳، ۴ و ۵	۲، ۶ و ۷	۲/۷۵
انتخاب دوم	۲، ۳، ۶ و ۷	۱، ۴ و ۵	۲
انتخاب سوم	۲، ۴، ۶ و ۷	۱، ۳ و ۵	۱/۹

همانطور که از نتایج پیداست اولین انتخاب معرفی شده در جدول که بیشترین مقدار FDR را به خود اختصاص داده مربوط به حالتی است که احساس‌های عصبانیت، انزجار، ترس و خوشحالی در یک کلاس و احساس‌های عادی،

خستگی و ناراحتی در کلاس دوم قرارگیرند. با توجه به مدل نشان داده شده در شکل ۳، احساس‌های کلاس اول در سطح بالایی از فعالیت قرار دارند. در حالی‌که احساس‌های کلاس دوم در سطح پایینی از فعالیت قرار گرفته اند. به بیان دیگر می‌توان گفت که در اولین مرحله طبقه‌بندی، تفکیک کلاس ها بر اساس میزان فعالیت صورت گرفته است. بسیاری از مقالاتی که ساختاری تجربی برای طبقه‌بند مذکور ارائه کرده اند نیز، ترکیب مشابهی برای اولین مرحله طبقه‌بندی ارائه نموده اند [۱۳]. ساختار کامل طبقه‌بند طراحی شده در شکل ۱۰ نشان داده شده است.



شکل ۱۰. ساختار طبقه‌بند طراحی شده برای تلفیق گویندگان زن و مرد

اندازه ماکزیمم FDR و دقت طبقه‌بندی برای هر کدام از طبقه‌بندهای باینری C_1 تا C_3 در جدول ۴ نشان داده شده است.

جدول ۴- مقادیر حداکثر FDR و دقت تشخیص در سه طبقه‌بند باینری اول در ساختار نشان داده شده در شکل ۱۰

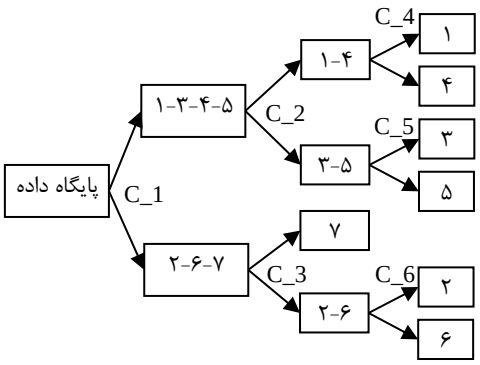
طبقه‌بند	C_1	C_2	C_3
مقدار ماکزیمم FDR	۲/۷۵	۱/۶۲	۶/۷
دقت تشخیص	۷۴٪/۵۳	۷۷٪/۴۲	۵۸٪/۱۴

ماتریس تداخل برای این طبقه‌بند در جدول ۵ نشان داده شده است. همانطور که از نتایج پیداست بیشترین نرخ تشخیص مربوط به احساس عصبانیت و کمترین نرخ تشخیص مربوط به احساس ناراحتی می‌باشد. بیشترین نرخ خطا در این طبقه‌بند مربوط به تشخیص ناراحتی می‌باشد. جملاتی که با ناراحتی بیان شده اند با نرخ ۷۵٪ انزجار تشخیص داده شده اند. نرخ تشخیص متوسط بدست آمده در این حالت برابر ۴۳/۴٪ می‌باشد.

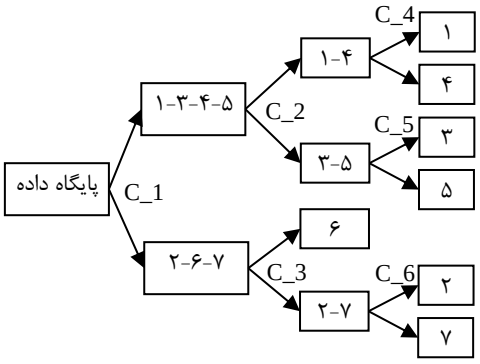
جدول ۵- ماتریس تداخل برای طبقه‌بند نشان داده شده در شکل ۱۰ (تلفیق گویندگان زن و مرد)

احساس	۱	۲	۳	۴	۵	۶	۷	نرخ تشخیص
۱	۱۸	۰	۱	۰	۶	۰	۰	٪۷۲
۲	۱	۴	۶	۰	۰	۰	۵	٪۲۵
۳	۲	۱	۴	۰	۲	۰	۰	٪۴۴/۴
۴	۵	۰	۰	۷	۲	۰	۰	٪۵۰
۵	۶	۰	۰	۰	۳	۵	۰	٪۲۱/۴
۶	۱	۰	۲	۰	۱	۸	۴	٪۵۰
۷	۰	۰	۹	۱	۰	۰	۲	٪۱۶/۷

در دومین آزمایش، طبقه‌بندهای جداگانه ای برای گویندگان زن و مرد طراحی شده است. ساختار مربوط به این طبقه‌بندها در شکل‌های ۱۱ و ۱۲ نشان داده شده است.



شکل ۱۱. ساختار طبقه‌بند طراحی شده برای گویندگان زن



شکل ۱۲. ساختار طبقه‌بند طراحی شده برای گویندگان مرد

اندازه ماکزیمم FDR برای هر کدام از طبقه‌بندهای باینری C_1 تا C_3 برای گویندگان زن و مرد در جدول ۶ نشان داده شده است.

جدول ۶- مقادیر حداکثر FDR و دقت تشخیص برای بهترین انتخاب ها در سه طبقه‌بند باینری اول در ساختار نشان داده شده در شکل‌های ۱۱ و ۱۲

طبقه‌بند		C_1	C_2	C_3
مقدار ماکزیمم FDR	زن	۵/۳	۴	۲
دقت تشخیص		٪۹۴/۹	٪۸۷/۵	٪ ۸۸/۵۷
مقدار ماکزیمم FDR	مرد	۳/۵	۲/۱	۲/۷
دقت تشخیص		٪۹۳/۴۷	٪۸۵	٪۱۰۰

ماتریس‌های تداخل که گویای نتایج این دو طبقه‌بند می‌باشند در جداول ۷ و ۸ نشان داده شده است. نرخ تشخیص متوسط بدست آمده برای گویندگان زن برابر ٪۸۳/۰۵ می‌باشد. نرخ تشخیص متوسط بدست آمده برای گویندگان مرد نیز برابر ٪۸۲/۶۱ می‌باشد.

جدول ۷- ماتریس تداخل برای طبقه‌بند نشان داده شده در شکل ۱۱ (گویندگان زن)

احساس	۱	۲	۳	۴	۵	۶	۷	نرخ تشخیص
۱	۱۳	۰	۰	۰	۰	۰	۰	٪۱۰۰
۲	۰	۷	۱	۰	۱	۰	۰	٪۷۷/۷۸
۳	۰	۰	۷	۰	۰	۰	۰	٪۱۰۰
۴	۰	۰	۰	۲	۴	۰	۰	٪۳۳/۳۳
۵	۰	۰	۰	۰	۹	۰	۰	٪۱۰۰
۶	۰	۱	۰	۰	۰	۷	۰	٪۸۷/۵
۷	۰	۲	۱	۰	۰	۰	۴	٪۶۶/۶۷

جدول ۸- ماتریس تداخل برای طبقه‌بند نشان داده شده در شکل ۱۲(گویندگان مرد)

احساس	۱	۲	۳	۴	۵	۶	۷	نرخ تشخیص
۱	۱۲	۰	۰	۰	۰	۰	۰	٪۱۰۰
۲	۰	۶	۰	۰	۰	۰	۱	٪۸۵/۷۱
۳	۰	۱	۰	۱	۰	۰	۰	۰
۴	۰	۰	۰	۷	۰	۰	۰	٪۱۰۰
۵	۱	۰	۰	۰	۴	۰	۰	٪۸۰
۶	۰	۲	۰	۱	۰	۵	۰	٪۶۲/۵
۷	۰	۰	۱	۰	۰	۰	۴	٪۸۰

شکل ۱۳. نرخ تشخیص به تفکیک احساس و جنسیت گویندگان

نرخ تشخیص برای هر احساس و نرخ تشخیص متوسط برای تلفیق گویندگان زن و مرد، گویندگان زن و گویندگان مرد برای مقایسه بهتر در شکل ۱۳ نشان داده شده است. کمترین نرخ تشخیص برای گویندگان زن در تشخیص احساس ترس رخ داده است که الگوریتم به اشتباه در ٪۶۶/۶۷ مواقع به جای ترس احساس مورد نظر را خوشحالی تشخیص می‌دهد. همچنین کمترین نرخ تشخیص برای گویندگان مرد مربوط به احساس انزجار می‌باشد که الگوریتم به اشتباه در ٪۵۰ مواقع این احساس را خستگی و در ٪۵۰ موارد این احساس را ترس تشخیص می‌دهد. البته با اندکی دقت در جدول خواهیم دید تعداد جملات تست برای این احساس تنها ۲ جمله بوده است. متأسفانه در پایگاه داده برلین تعداد جملات برای احساس‌های مختلف یکسان نبوده و برای برخی از حالات تعداد جملات به اندازه ای کم است که ارزیابی الگوریتم را با مشکل مواجه می‌کند. به همین دلیل برخی از پژوهشگران فقط به طبقه‌بندی برخی از احساس‌های موجود در این پایگاه داده پرداخته اند. در این تحقیق هیچ یک از قسمت‌های پایگاه داده برلین را حذف نشده است.

با توجه به اینکه تعداد کل جملات تست برای گویندگان زن و مرد به ترتیب ۵۹ و ۴۵ جمله می‌باشند، میانگین نرخ تشخیص برای این الگوریتم ٪۸۲/۸۶ محاسبه می‌شود که

۳۹/۴۶٪ افزایش نرخ تشخیص را نسبت به تلفیق گویندگان زن و مرد نشان می‌دهد. این درحالی است که در تست انسانی نرخ تشخیص برای این پایگاه داده ۸۷/۴٪ بدست آمده است [۳۴]. هر چند که با توجه به شرایط متفاوت آزمایش ها در پارامترهایی نظیر چگونگی تقسیم بندی نمونه‌های مورد استفاده در آموزش و تست الگوریتم، مقایسه عددی بین مقالات مختلف عادلانه نمی‌باشد، اما می‌تواند یک دید کلی در این زمینه ارائه نماید. در [۱۱] نتایج برای طبقه‌بند مبتنی بر مدل مخفی مارکوف بر روی همین پایگاه داده ۶۸/۵۷٪ و برای طبقه‌بند مرتبه ای ۷۱/۷۵٪ گزارش شده است. در [۴] و [۸] دستیابی به نرخ تشخیص ٪۸۵/۶ و ۸۸/۸۹٪ با استفاده از یک گروه ویژگی جدید و ساختار طبقه‌بند مرتبه ای میسر شده است. با حذف احساس خستگی و فقط بررسی ۶ احساس دیگر نتیجه ٪۸۴/۸ در [۹] بدست آمده است. [۳۹] و [۲] نیز نرخ تشخیص ٪۷۲/۵ و ۷۵٪ را برای طبقه بندی ۶ احساس گزارش نموده اند.

با توجه به نزدیک بودن نرخ تشخیص بدست آمده به نرخ تشخیص در تست انسانی این سؤال مطرح می‌شود که آیا ممکن است که نرخ تشخیص ماشین از انسان هم فراتر رود؟ به بیان دیگر آیا اطلاعاتی از احساس گوینده در سیگنال گفتار وجود دارد که توسط انسان قابل تشخیص نباشد اما ماشین قادر به تشخیص این اطلاعات باشد؟ این سؤال مهمی است که پاسخ آن مسیر آینده ی این شاخه از تحقیقات را تا حد زیادی مشخص می‌کند. به نظر می‌رسد با توجه به اینکه مبنای طراحی و تست الگوریتم ها برچسب‌هایی است که توسط انسان به پایگاه داده اختصاص داده می‌شود احتمالاً دستیابی به نرخ تشخیص بالاتر از تشخیص انسان مفهوم خاصی نداشته باشد.

۸- نتیجه‌گیری

علیرغم تلاش‌های زیادی که در تشخیص احساس بوسیله‌ی سیگنال گفتار انجام گرفته است هنوز این مساله با چالش‌های زیادی روبرو می‌باشد. برای حل مشکلات موجود محققان دو رویکرد به این مساله داشته اند: اولین رویکرد استفاده از اطلاعات جانبی نظیر چهره، متن، حرکات دست و بدن به منظور افزایش نرخ تشخیص می‌باشد. دومین راه حل نیز استفاده از الگوریتم‌های کارآمدتر برای تشخیص احساس بوسیله‌ی سیگنال گفتار می‌باشد. استفاده از ویژگی‌های مؤثر تر در تشخیص احساس، استفاده از روش‌های انتخاب ویژگی کارآمد و طراحی کلاسه‌بندهایی با دقت بیشتر از جمله مواردی است که محققان برای افزایش نرخ تشخیص به آن‌ها پرداخته اند. در آن میان استخراج ویژگی‌های مؤثر با توجه به اهمیت موضوع جایگاه خاصی را در تحقیقات انجام شده به خود اختصاص داده است. در این مقاله از دو ویژگی جدید مبتنی بر الگوهای طیفی و انرژی هارمونیک ها به منظور بهبود عملکرد الگوریتم استفاده شده است. در استخراج این ویژگی ها از ابزارهای پردازش تصویر به منظور آنالیز تصویر طیف‌نگاره سیگنال گفتار استفاده شده است. در روش پیشنهادی از الگوریتم طبقه‌بند مرتبه ای مبتنی بر طبقه‌بندهای باینری SVM استفاده نمودیم. به منظور بهینه سازی الگوریتم طبقه‌بند، از معیار FDR برای تعیین تفکیک پذیری کلاس‌های مختلف و قرار دادن جداپذیر ترین کلاس ها در مراحل ابتدایی طبقه‌بندی استفاده کردیم. این امر از بروز خطا در مراحل اولیه طبقه‌بندی و انتشار آن در سایر بخش‌های طبقه‌بند جلوگیری می‌کند. از آنجا که درک و بیان احساس زنان و مردان از نظر روانشناختی و فیزیولوژی متفاوت است به نظر می‌رسد الگوریتم‌های جداگانه ای برای بازشناسی احساس در زنان و مردان نیاز می‌باشد. آزمایشات حاکی از چشمگیر بودن اثر تفکیک جنسیتی گویندگان در بالا رفتن نرخ تشخیص می‌باشند.

در راستای تحقیقات آینده تهیه ی یک پایگاه داده فارسی جامع، می‌تواند گام مهمی در کمک به پژوهش در این راستا باشد. علاوه بر آن ارائه روش‌های استخراج ویژگی مؤثر با نرخ محاسباتی پایین که استفاده از آن‌ها در کاربردهای عملی و روبات ها میسر باشد نیز گام مهمی در جهت کاربردی کردن این تحقیقات خواهد بود. علاوه بر آن

استفاده از الگوریتم پیشنهادی به عنوان بخشی از یک سیستم بازشناسی گفتار ممکن است در افزایش نرخ تشخیص مؤثر باشد.

۹- تقدیر و تشکر

با تشکر فراوان از استاد گرانقدرم مرحوم دکتر خشایار یغمائی که این کار بخشی از ثمره زحمات بی دریغ اوست، روحش شاد و یادش گرامی باد.

۱۰- مراجع

[1] ElAyadi, M ., Kamel, M.S., Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44, pp 572–587.
[2] Yang, B., Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. Signal Processing 90, pp 1415–1423.
[3] Monti, G., Meletti, S. (2015). Emotion recognition in temporal lobe epilepsy: A systematic review. Neuroscience and Biobehavioral Reviews 55 .pp 280–293.
[4] Wu, S., Falk, T.H., Chan, W.Y. (2011). Automatic speech emotion recognition using modulation spectral features. Speech communication 53, pp 768–785.
[5] Harimi, A., AhmadyFard, A.R., Shahzadi, A., Yaghmaie, K. (2015). Anger or Joy? Emotion Recognition Using Nonlinear Dynamics of Speech. Applied Artificial Intelligence 29 .pp 675–696.
[6] Milton, A., Tamil Selvi, S. (2014). Class-specific multiple classifiers scheme to recognize emotions from speech signals. Computer Speech and Language 28. pp 727–742.
[7] Bitouk, D., Verma, R., Nenkova, A. (2010). Class-level spectral features for emotion recognition. Speech Communication 52, pp 613–625.
[8] Albornoz, E.M., Milone, D.H., Rufiner, H.L. (2011). Spoken emotion recognition using hierarchical classifiers. Computer Speech and Language 25, pp 556–570.
[9] Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T. (2008). Fear-type emotion recognition for future audio-based surveillance systems. Speech Communication 50, pp 487–503.
[10] Polzehl, T., Schmitt, A., Metze, F., Wagner, M. (2011). Anger recognition in speech using acoustic and linguistic cues. Speech Communication 53, pp 1198–1209.
[11] Pérez-Espinosa, H., Reyes-García, C.A., Villasenor-Pineda, L. (2011). Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model. Biomedical Signal Processing and Control 02.008.
[12] Kockmann, M., Burget, L., Cernocky, J. (2011). Application of speaker- and language identification state-of-the-art techniques for emotion recognition. Speech Communication 53, pp 1172–1185.
[13] Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach.Speech Communication 53, pp 1162–1171.
[14] Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., Elenius, K. (2011). Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. Computer Speech and Language 25, pp 84–104.
[15] Bozkurt, E., Erzin, E., Erdem, C.E., Erdem, A.T. (2011). Formant position based weighted spectral features for emotion recognition. Speech Communication 53, pp 1186–1197.
[16] Vayrynen, E., Toivanen, J., Seppanen, T. (2011). Classification of emotion in spoken Finnish using vowel-length segments: Increasing reliability with a fusion technique. Speech Communication 53, pp 269–282.
[17] Ooi, C.S., Seng, K.P., Ang, L.M., Chew, L.W. (2014). A new approach of audio emotion recognition. Expert Systems with Applications 41, pp 5858–5869.
[18] Ververidis, D., Kotropoulos, C. (2008). Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. Signal Processing 88, pp 2956–2970.
[19] Buck, R. (1988). Human motivation and emotion., Wiley, New York.

[20] Darwin, C. (1955). The expression of the emotions in man and animals. Philosophical Library Edition, London (reproduction of 1872 publication).
[21] Ekman, P., Friesen, W.V. (1975). Unmasking the face., Englewood Cliffs: Prentice Hall.
[22] Izard, C.E. (1977). Human emotions. Plenum Press, New York.
[23] Panksepp, J. (1988). Affective neuroscience: The foundations of human and animal emotion. Oxford University Press, New York.
[24] Pell, M.D., Monetta, L., Paulmann, S., Kotz, S.A. (2009). Recognizing emotions in a foreign language. Journal of Nonverbal Behavior. 33. pp 107–120.
[25] Ross, E.D., Prodan, C.I., Monnot, M., (2007). Human facial expressions are organized functionally across the upper-lower facial axis. Neuroscientist. 13. pp 433–446.
[26] Ross, E.D., Monnot, M. (2011). Affective prosody: What do comprehension errors tell us about hemispheric lateralization of emotions, sex and aging effects, and the role of cognitive appraisal Neuropsychologia. 49. pp 866–877.
[27] Lewis, W., Michalson, L. (1983). Children's emotions and moods: Developmental theory and measurement., Plenum Press, New York.
[28] Malatesta, C.Z., Kalnok, D. (1984). Emotional experience in younger and older adults. Journal of Gerontology. 39, pp 301–308.
[29] Engel, B. (2006). Healing Your Emotional Self. John Wiley & Sons, New Jersey.
[30] Engel, B. (2008). The Nice Girl Syndrome. John Wiley & Sons, New Jersey.
[31] Sanchez, F. (2006). A Thousand Moments of Solitude. Library of Congress Control Number: 2005911228, United States of America.
[32] Sherwood, L. (2010). HUMAN PHISIOLOGY From Cells to Systems, eight edition. Cengage Learning. Library of Congress Control Number: 2011939366.
[33] Whittle, S., Yücel, M., Yap, M.B.H., Allen, N.B. (2011). Sex differences in the neural correlates of emotion: Evidence from neuroimaging. Biological Psychology. 87. pp 319– 333.
[34] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B. (2005). A database of German emotional speech. Interspeech. pp 1517–1520.
[35] Shahzadi1, A., Ahmadyfard, A.R., Yaghmaie, K., Harimi, A. (2013). Recognition Of Emotion In Speech Using Spectral Patterns. Malaysian Journal of Computer Science 26(2), pp 140-158.
[36] Sreenivasa Rao, K., Ramu Reddy, V., Maity, S. (2015). Language Identification Using Spectral and Prosodic Features. New York, Springer.
[37] Mitsuyoshi, S., Monnma, F., Tanaka, Y., Minami, T., Kato, M., Murata, T. (2011). Identifying neural components of emotion in free conversation with fMRI. Defense Science Research Conference and Expo (DSR).
[38] Kotti, M., Kotropoulos, C. (2008). Gender classification in two Emotional Speech databases. 19th International Conference on Pattern Recognition (ICPR).
[39] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. Signal Processing Magazine, IEEE, vol. 18, pp. 32-80.