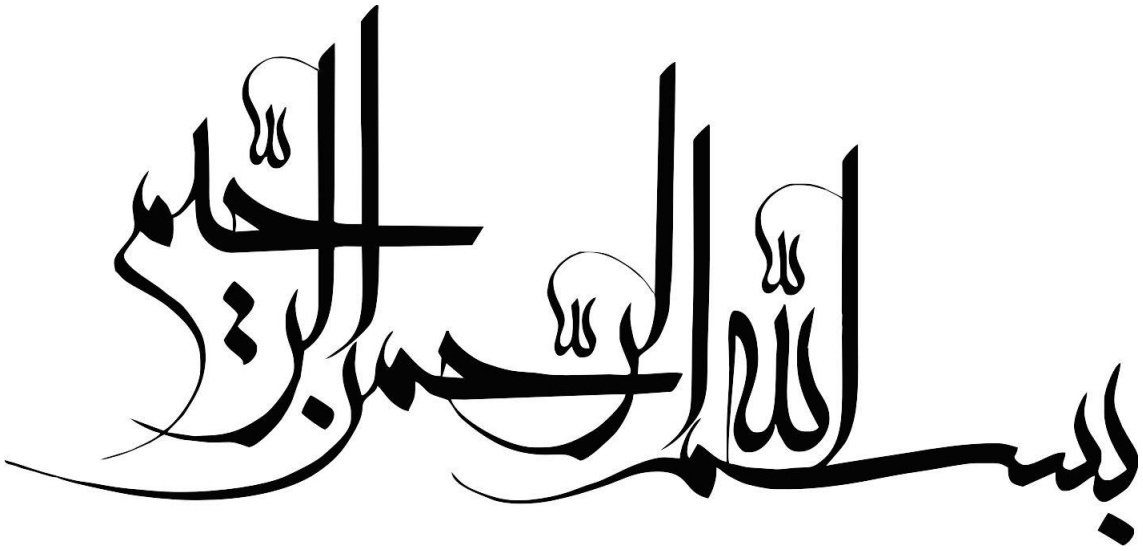

بسم الله الرحمن الرحيم

دانشکده مهندسی برق و رباتیک

رشته مهندسی برق گرایش الکترونیک

پایان‌نامه کارشناسی ارشد

تشخیص وقایع صوتی بر اساس ویژگی‌های MP (Matching Pursuit)

نگارنده: رقیه بهمنی

استاد راهنما

دکتر حسین مروی

تیر ۱۳۹۵

دانشکده: مهندسی برق و رباتیک
گروه: الکترونیک

پایان نامه کارشناسی ارشد خانم رقیه بهمنی به شماره دانشجویی ۹۲۰۳۴۸۴ تحت عنوان

تشخیص وقایع صوتی بر اساس ویژگی‌های MP (Matching Pursuit)

در تاریخ ۱۳۹۵/۴/۳۰ توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد در مهندسی برق-الکترونیک مورد ارزیابی و با درجه‌ی عالی مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
دکتر حسین مروی			

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
دکتر امیدرضا معروضی		دکتر احسان رحیمی	
دکتر هادی گرایلو			

تقدیم

تقدیم به پدر و مادر عزیزم.

سپاس‌گزاری

از استاد گرامی، جناب آقای دکتر مروی سپاس‌گزارم.

تعهدنامه

اینجانب رقیه بهمنی دانشجوی دوره کارشناسی ارشد رشته مهندسی برق-الکترونیک دانشکده‌ی مهندسی برق و رباتیک دانشگاه صنعتی شاهرود نویسنده پایان‌نامه تشخیص وقایع صوتی بر اساس ویژگی های MP (Matching Pursuit) تحت راهنمائی دکتر حسین مروی متعهد می‌شوم:

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

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه های رایانه ای، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

در سال‌های اخیر تحقیقاتی روی شناسایی صداهای محیطی و وقایع صوتی صورت گرفته است اما حجم پژوهش‌ها در مقایسه با زمینه‌هایی همچون گفتار و موسیقی بسیار ناچیز است. هدف این پایان-نامه گسترش روش‌های استخراج ویژگی برای شناسایی وقایع صوتی محیط اداری است. پایگاه داده‌ی مورد استفاده تحت عنوان D-CASE از ۱۶ کلاس وقایع صوتی مربوط به محیط اداری تشکیل شده است که برخی از این صداها آلوده به نویزهای با نرخ سیگنال به نویز متفاوت می‌باشد. برای این مسئله، دو روش استخراج ویژگی معرفی شده است. روش اول، استفاده از ویژگی‌های استخراج شده توسط الگوریتم پیگیری انطباق در ترکیب با ویژگی‌های متداول MFCC، به عنوان بردارهای ویژگی و استفاده از طبقه‌بند نزدیکترین همسایگی است. نرخ شناسایی ۶۹/۶۷ درصد برای این روش بدست آمده، که نسبت به استفاده‌ی تنها از ویژگی‌های MFCC، ۶ درصد افزایش داشته است. در روش دوم، از ضرایب کپسترال فرکانس بارک به عنوان ویژگی و طبقه‌بند GMM استفاده شده است. این روش، نرخ شناسایی ۸۰/۰۵ درصد را بدست داده که در مقایسه با بسیاری از روش‌های موجود برای این پایگاه داده بهبود داشته است.

واژگان کلیدی: وقایع صوتی، استخراج ویژگی، پیگیری انطباق، ضرایب کپسترال فرکانس بارک (BFCC).

مقالات مستخرج از پایان‌نامه

[۱] بهمنی، رقیه و مروی، حسین، "شناسایی وقایع صوتی با استفاده از الگوریتم پیگیری تطبیق"، اولین کنفرانس بین المللی دستاوردهای نوین پژوهشی در مهندسی برق و کامپیوتر، دانشگاه صنعتی امیرکبیر، ۱۳۹۵.

[۲] بهمنی، رقیه و مروی، حسین، "پیشنهاد یک سیستم شناسایی وقایع صوتی با استفاده از ترکیب ویژگی های طیفی سیگنال"، اولین کنفرانس بین‌المللی دستاوردهای نوین پژوهشی در مهندسی مکانیک، مکاترونیک و بیومکانیک، دانشگاه صنعتی امیرکبیر، ۱۳۹۵.

فهرست عنوان‌ها

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱ مقدمه

۱-۱- مقدمه

انسان معمولاً برای ارتباط با دیگران، درک محیط اطراف و نشان دادن واکنش مناسب، به طور همزمان از هر دو قوه‌ی بینایی و شنوایی استفاده می‌کند؛ توانایی‌ای که ماشین هنوز از داشتن آن در سطح انسان، محروم است. در مواردی که بینایی ماشین دچار مشکل می‌شود مانند موقعیتی که نور محیط کم است، سیگنال‌های شنیداری می‌توانند اطلاعات مفیدی را ارائه دهند. علاوه بر این، عملیات ذخیره‌سازی و محاسبات مربوط به سیگنال‌های شنیداری نسبت به سیگنال‌های بینایی هزینه‌ی کم-تری دارد [۱]. از این‌رو تلاش برای ارتقاء سطح شنوایی ماشین از جمله موضوعاتی است که در سال-های اخیر مورد توجه پژوهشگران بوده است [۲, ۳].

با این‌که بیش از پنجاه سال از توسعه‌ی سیستم‌های شناسایی صدا می‌گذرد، شناسایی گفتار انسان و یا موسیقی، حجم عمده‌ی پژوهش‌های این زمینه را به خود اختصاص داده و موضوع شناسایی وقایع صوتی[1] کم‌تر موردتوجه بوده است. زیرا تصور می شده که گفتار و موسیقی، ارزشمندترین اطلاعات موجود در یک فایل صوتی هستند. اما از آن‌جایی که صداهای محیطی می‌توانند حامل اطلاعات بسیار مهمی باشند، در سال‌های اخیر مطالعه روی صداهای محیطی و تحلیل آن‌ها در راستای ارتقاء سطح شنوایی ماشین مورد توجه قرار گرفته و افزایش یافته است [۴-۷].

۱-۲- انگیزه ی تحقیق

اگرچه انواع بسیاری از صداهای محیطی وجود دارد، اما هنوز دسته‌بندی استانداردی برای آن‌ها ارائه نشده است. به طور کلی در شناسایی صداهای محیطی با دو مسئله مواجه هستیم: شناسایی محیط، شناسایی وقایع یک محیط. در مسئله‌ی شناسایی محیط، هدف تنها شناسایی محیط مربوطه، صرف نظر از نوع وقایعی است که در آن محیط رخ می‌دهد. در این راستا ممکن است از برخی مولفه‌های آهنگین صدا محیط شناسایی شود مثل صدای باران یا از طریق یافتن برخی وقایع خاص محیط

[1] Audio Events

شناسایی شود.اما در این روش تمرکز دقیقی روی تمام وقایع یک محیط صورت نمی‌گیرد و در واقع تنها مهمترین و مشخص‌ترین وقایع محیط به عنوان مشخصه‌های آن محیط در نظر گرفته می‌شوند. به طور مثال محیطی که در آن صدای بوق اتومبیل شنیده شود به احتمال زیاد خیابان یا بزرگراه است. در این زمینه تا کنون کارهای موفقیت آمیزی صورت گرفته است[۸]. اما در مسئله‌ی شناسایی وقایع صوتی، هدف یافتن نوع واقعه‌ی رخ داده در محیطی خاص است. وقایع که در زمان کوتاهی رخ می‌دهند معمولاً ماهیت ضربه‌ای دارند مانند صدای بسته شدن در، سرفه، صاف کردن گلو و غیره. شناسایی این دسته از صداها که با عنوان «وقایع صوتی» از آن‌ها یاد می‌شود، برای ماشین بسیار دشوار بوده و تاکنون نتایج چندان مناسبی در این زمینه حاصل نشده است. از این‌رو با توجه به جذابیت و باز بودن زمینه-های پژوهشی این موضوع، در این پایان‌نامه این دسته از صداها را مورد بررسی قرار می‌دهیم.

ازآنجاکه وقایع صوتی به درک انسان از محیط مجاور کمک می‌کنند، توصیفگرهای خوبی برای محیط هستند. واقعه‌ی صوتی، برچسبی است که برای شرح یک رویداد قابل‌تشخیص در ناحیه‌ای از صوت استفاده می‌شود. چنین برچسبی معمولاً به ماشین اجازه می‌دهد که مفهوم کلی نهفته در صوت را بفهمد و آن را در بین وقایع شناخته‌شده‌ی دیگر جای دهد [۹]. از این‌رو شناسایی صداهای محیطی یک مسئله‌ی اساسی در پردازش سیگنال است که نقش مهمی را در تکامل شنوایی ماشین ایفا می‌کند.

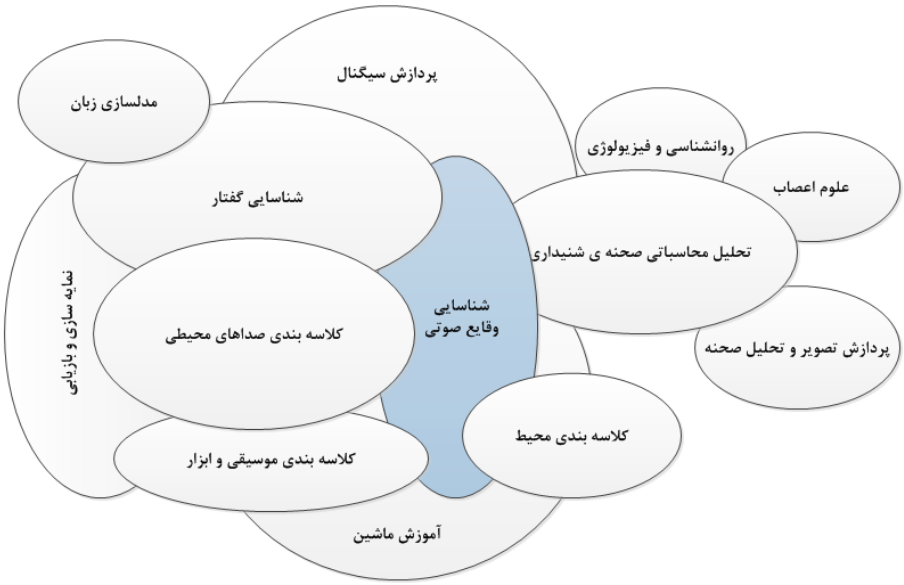
وقایع صوتی، ساختاری نویزمانند دارند. به عبارت دیگر، نویز، مورد خاصی از یک رویداد صوتی است؛ زیرا دارای خواص منحصر به فردی مانند مدت زمان طولانی و محتوای طیفی ثابت در زمان می‌باشد. صداهایی مانند کلیک کردن، صفحه کلید، قدم زدن و یا بسته شدن در، که در حوزه‌ی تشخیص گفتار اغلب به عنوان نویز ضربه‌ای در نظر گرفته می‌شوند در این‌جا وقایع صوتی‌ای هستند که برخلاف نویز، دارای مدت زمان محدودی هستند.

آشکارسازی وقایع صوتی[1] (AED)، آشکارسازی و طبقه‌بندی[2] رویدادهای صوتی موجود در یک سیگنال صوتی احتمالاً نویزی و پرانعکاس[3] است که معمولاً در سیستم‌های کنترل خودکار انجام می‌شود. می‌توان گفت که بسیاری از پژوهش‌ها مربوط به مجموعه‌ی محدودی از صداهای محیطی بوده است [۱۰-۱۲]. به عنوان مثال، برای تشخیص تیراندازی [۱۳]، پایش صدای تنفس [۱۴]، تشخیص رویداد جویدن [۱۵] و تشخیص آژیر [۱۶]. شناسایی خودکار وقایع صوتی می‌تواند در زمینه‌های مختلفی از قبیل مراقبت‌های بهداشتی، کاربردهای نظامی، تشخیص وقایع غیرگفتاری در محیط اتاق ملاقات، شناسایی محتوای صدا، قطعه‌بندی صدا و برچسب زدن خودکار [۱۷] ، هدایت ربات [۳]، ربات‌های کمک (مراقب)، و دیگر سرویس‌های مبتنی بر وسایل متحرک که در آن‌ها آگاهی از فضا، اغلب مطلوب یا مورد نیاز است کاربرد داشته باشد [۱۸]. شناسایی وقایع صوتی می‌تواند در محیط نظارت خانگی، برای کمک به افراد سالخورده‌ای که در خانه تنها زندگی می‌کنند [۱۹]، و یا برای خانه‌های هوشمند [۲۰] استفاده شود. هم‌چنین می‌تواند برای شناسایی انواع حیوانات و پرندگان به‌وسیله‌ی تمایز صداهای آن‌ها به‌کار گرفته شود [۲۱]. کاربردهای دیگر شامل مواردی مانند طراحی یک تلفن همراه هوشمند می‌شود که می‌تواند به طور خودکار، وضعیت اطلاع‌رسانی را بر اساس آگاهی از محیط اطراف کاربر تغییر دهد، مانند تغییر دادن به حالت بی‌صدا در تئاتر یا در کلاس درس [۳]. هم‌چنین می‌تواند اطلاعات مکانی کاربر را مهیا کند [۲۲].

شناسایی وقایع صوتی ارتباط تنگاتنگی با دیگر زمینه‌های مورد بررسی در پردازش صوت دارد. شکل ۱-۱ چگونگی این ارتباط را با دیگر زمینه‌های موجود، مانند طبقه‌بندی موسیقی یا محیط طبیعی نشان می‌دهد. با توجه به این ارتباط می‌توان ادعا کرد که روش‌های مربوط به زمینه‌های گفتار و موسیقی نیز در طبقه‌بندی صداهای محیطی قابل استفاده خواهد بود. زمینه های اساسی نشان داده

[1] Audio Event Detection
[2] Classification
[3] Reverberant

شده، پردازش سیگنال، یادگیری ماشین و تجزیه و تحلیل محاسباتی صحنه‌ی شنیداری[1] (CASA) می‌باشند.



شکل ۱-۱: نمای کلی از چگونگی ارتباط شناسایی وقایع صوتی با زمینه های دیگر [۲۳].

۱-۳- هدف پایان نامه

هدف ما در این پایان‌نامه مطالعه‌ی وقایع صوتی محیط اداری بدون ساختار در یک مفهوم کلی تر و نیز تلاش برای یافتن یک مجموعه ویژگی قوی برای شناسایی کلاس‌های مختلف صدا می‌باشد. در این کار ما یک تجزیه و تحلیل تجربی روی ویژگی‌های مختلف برای توصیف وقایع صوتی انجام می‌دهیم و استفاده از دو روش استخراج ویژگی پیگیری انطباق (MP)[2] و ضرایب کپسترال فرکانس بارک (BFCC)[3] را برای داشتن یک مجموعه ویژگی قوی پیشنهاد می‌کنیم.

[1] Computational Auditory Scene Analysis
[2] Matching Pursuit
[3] Bark Frequency Cepstral Coefficients

۱-۴- ساختار پایان نامه

این پایان نامه دارای شش فصل است.در فصل ۱ اهمیت شناسایی وقایع صوتی و برخی از کاربردهای آن به‌صورت مختصر بیان گردید. در فصل ۲ مروری بر کارهای انجام شده در زمینه‌ی شناسایی صداهای محیطی و به طور خاص شناسایی وقایع صوتی ارائه شده و آخرین دستاوردهای پژوهشی در زمینه‌ی تشخیص و شناسایی وقایع صوتی مورد بررسی قرار می‌گیرد. فصل ۳ شرح مختصری از اصول کلی مربوط به یک سیستم شناسایی وقایع صوتی را که شامل بخش‌های تشخیص، استخراج ویژگی و طبقه بندی است ارائه می‌دهد. در فصل ۴ جزئیات روش پیشنهادی برای تشخیص و استخراج ویژگی از پایگاه داده شرح داده می‌شود. در فصل ۵ ابتدا در مورد پایگاه داده‌ی استفاده شده در این پایان‌نامه توضیح داده می‌شود و در ادامه، نکات تنظیمات عملی آزمایش‌ها و در نهایت، نتایج شبیه‌سازی ها در قالب نمودارهای گرافیکی و جدول‌ها ارائه شده و با تحلیل نتایج تکمیل می‌شود. و در نهایت فصل ۶ خلاصه‌ای از نتایج این پایان‌نامه و پیشنهادها و برنامه‌های آینده را برای این موضوع بیان می‌کند.

فصل ۲ مروری بر پیشینه‌ی شناسایی وقایع صوتی

۲-۱- ویژگی های عمومی وقایع صوتی

با توجه به تنوع منابع صوتی مختلفی که صداها را ایجاد می‌کنند خلاصه کردن ویژگی‌های یک رویداد صوتی کار دشواری است. برخلاف گفتار که به صداهای تولید شده به‌وسیله‌ی دستگاه صوتی انسان و زبان محدود می‌شود یک رویداد صوتی ممکن است از انواع فعل و انفعالات مختلف تولید شود. به‌عنوان مثال صدای تولیدشده به‌وسیله‌ی صفحه‌کلید رایانه، باتوجه به تنوع کابران می‌تواند متفاوت باشد. جدول ۲-۱ مقایسه‌ی ویژگی‌های گفتار، موسیقی و صداهای محیطی را نشان می‌دهد [۲۴]. می‌توان دید در حالی که گفتار و موسیقی تعاریف دقیقی با توجه به این ویژگی ها دارند، صداهای محیطی یا تعریف نشده هستند و یا می‌توانند طیف گسترده‌ای از ویژگی ها را شامل شوند. از اینرو معمولاً یک محدوده‌ی کوچک، مانند یک نوع خاص از صداها برای مسئله تعریف می‌شود، تا حداقل برخی از مشخصه‌ها قابل تعریف باشد.

جدول ۲-۱- مشخصات گفتار، موسیقی و صداهای محیطی

	گفتار	موسیقی	صداهای محیطی
مشخصات آکوستیکی	تعداد واج ها	تعداد تون‌ها	نامشخص
طول پنجره‌ها	کوتاه (ثابت)	بلند (ثابت)	نامشخص
طول شیفت زمانی	کوتاه (ثابت)	بلند (ثابت)	نامشخص
پهنای باند	باریک	نسبتاً باریک	پهن/باریک
هارمونی‌ها	واضح	واضح	واضح/غیرواضح
ایستایی (در زمان کوتاه)	ایستان	ایستان	غیرایستان/ایستان

۲-۲- مقایسه ی شناسایی وقایع صوتی با شناسایی گفتار/ گوینده

شناسایی گفتار و گوینده، در مقایسه با زمینه‌ی وسیع‌تر شناسایی وقایع آکوستیکی، موضوعات نسبتاً تکامل‌یافته‌ای هستند. این موضوعات، مبتنی بر مفاهیم پردازش سیگنال مشابهی هستند که در آن ابتدا یک آشکارساز برای بخش‌بندی جریان صوتی پیوسته استفاده می‌شود و سپس ویژگی از بخش استخراج می‌شود. پس از آن برای ارائه‌ی اطلاعات در مورد بخش، خوشه‌بندی و یا مدل‌سازی انجام می‌شود. با این‌حال، اصول زمینه‌های مختلف یکی نیست و نیازمند روش‌های مختلف برای مسئله است. این زمینه‌ها عبارتند از:

- تشخیص گفتار: تبدیل گفتار پیوسته به متن، با طبقه‌بندی واج با توجه به نمونه‌های آموزش داده‌شده‌ی قبلی انجام می‌شود. جنبه‌های دیگر که به تنوع بیان کمک می‌کنند نیز جالب هستند، مانند سرعت صحبت کردن، احساسات و لهجه [۲۵].

- شناسایی گوینده: این زمینه شامل چندین زیرموضوع ازجمله تأیید گوینده برای زیست‌سنجی و تفکیک گوینده است و تعیین می‌نماید که چه کسی در چه زمانی صحبت می‌کند. این مسئله نیازمند ایجاد مدلی است که منحصراً گفتار یک فرد را شناسایی کند، و لزوماً به شناسایی واج یا کلمات فرد نیازی نیست.

- تشخیص زبان: تشخیص زبان موضوعی است که معمولاً برای ترجمه‌ی ماشینی استفاده می‌شود و مشابه شناسایی گوینده می‌باشد. در این سیستم‌ها، ممکن است به‌جای تشخیص کلمات گفتار، تغییرپذیری زبان‌های مجزا مدل شود.

- شناسایی وقایع آکوستیکی: در این‌جا هدف، شناسایی و طبقه‌بندی وقایع آکوستیکی به گروه صوتیِ صحیح می‌باشد. به‌عنوان مثال، گفتار، موسیقی، نویز، پارس سگ و یا زنگ زنگوله، و غیره. تعداد گروه‌های صدا می‌تواند بسیار بزرگ باشد؛ ازاین‌رو در مسئله‌ی مورد نظر،

این گروه‌ها اغلب دارای محدودیت هستند، مانند صداهای یک اتاق جلسه. تفاوت‌های بین گروه‌ها به دامنه‌ی مسئله‌ی موردنظر مربوط می‌شود.

در گفتار، واج‌های مجزا، حوادث آکوستیکی جدا نیستند. بنابراین، استخراج ویژگی‌های آکوستیکی به‌صورت قاب به قاب و استفاده از مدل مخفی مارکوف[۱] (HMM) برای پیدا کردن محتمل‌ترین دنباله از واج‌ها برای ویژگی‌های داده‌شده، متداول است [۲۶]. در مورد تشخیص گوینده، ویژگی‌های آکوستیکی همانند گفتار استخراج می‌شوند، اما قاب‌ها به دنباله‌ای از کلمات، رمزنگاری نمی‌شوند و استفاده از روش‌های خوشه‌بندی برای شناسایی گویندگان مختلف، متداول است؛ به‌خصوص هنگامی‌که تعداد گویندگان از پیش مشخص نباشد.

برای شناسایی وقایع صوتی، اساس کار مشابه است و معمولاً از همان ویژگی‌های آکوستیکی و سیستم‌های تشخیص الگوی به‌کار رفته در شناسایی گفتار و صدا می‌توان استفاده کرد. بااین‌حال، دامنه‌ی وقایع آکوستیکی بسیار گسترده‌تر است؛ زیرا شامل صداهای محیطی می‌شود که دارای طیف وسیع‌تری از مشخصات هستند. علاوه بر این، محیط‌هایی که در آن‌ها وقایع آکوستیکی رخ می‌دهد بدون ساختار در نظر گرفته می‌شوند؛ به این معنی که ممکن است نویز پس‌زمینه و یا منابع متعدد صدا و طنین[۲] وجود داشته باشد که باعث می‌شود شناسایی بسیار سخت‌تر گردد. بنابراین سیستم‌های شناسایی وقایع صوتی مبتنی براین اصول هستند، و اغلب روش‌های مختلف را ترکیب می‌کنند.

۲-۳- کاربردهای شناسایی وقایع صوتی

در این بخش، کاربردهای شناسایی وقایع صوتی مورد بحث قرار می‌گیرد. با توجه به شکل ۱-۱ هر یک از این کاربردها خود یک موضوع متفاوت پژوهشی در زمینه‌ای وسیع‌تر است. اکثر کاربردهای شناسایی وقایع صوتی مبتنی بر عمل طبقه‌بندی است. به این صورت که با یک قطعه‌ی صوتی[۳] کوتاه، سیستم

[1] Hidden Markov Model
[2] Timbre
[3] Audio Clip

شناسایی باید تعیین کند که کدام واقعه‌ی صوتی در پایگاه داده‌ی آموزش‌داده‌شده، نزدیک‌ترین مورد به صوت جدید است. در یک کاربرد کمی متفاوت‌تر، هدف، شناسایی سبک موسیقی و یا به طور معادل، محیط پس‌زمینه به جای وقایع خاص است. در این کار، به طور معمول نیاز به یک قطعه‌ی صوتی طولانی‌تر می‌باشد. کاربرد دیگر، نمایه‌سازی و بازیابی است که در آن یک رویداد صوتی را می‌توان با محتوای صوتی‌اش جست‌جو کرد. این کاربردها در بخش‌های زیر به صورت مفصل‌تر بیان شده اند.

۲-۳-۱ طبقه بندی صداهای محیطی

صداهای محیطی اغلب شامل حوادث آکوستیکی عمومی می‌شود و گاهی اوقات، گفتار و موسیقی را در بر نمی‌گیرد. با این‌حال، ترجیح داده می‌شود که این صداها، به معنای واقعی کلمه به عنوان "صداهایی که ممکن است در یک محیط مشخص شنیده شود" در نظر گرفته شود. هرچند صداهای محیطی، گفتار و یا اصوات موسیقی را می‌تواند شامل شود، اما نمی‌توان محتوای آن‌ها را با این سیستم شناسایی تفسیر نمود. در حقیقت، محیط مورد نظر، دامنه‌ی مسئله‌ی شناسایی را تعیین می‌کند. عملکرد کلی یک سیستم شناسایی وقایع صوتی به بخش های اصلی آن یعنی مرحله‌ی استخراج ویژگی صوتی و مرحله‌ی طبقه‌بندی، و نیز ترکیبی از آن‌ها وابسته است. طبقه‌بندی به طور معمول به‌وسیله‌ی الگوریتم های آماری یادگیری ماشین انجام می‌شود. مانند HMM ها و یا ماشین‌های بردار پشتیبان[1] (SVM ها) که توزیع ویژگی کلاس خاصی را مدل می‌کنند و پارامترهای سیستم جداساز را تخمین می‌زنند. قدرت و عملکرد در شرایط نامساعد آکوستیکی تا حد زیادی به‌وسیله‌ی استخراج ویژگی ها تعیین می‌شود. ویژگی‌هایی که اطلاعات مربوط به طبقه‌بندی را حتی زمانی که نویز، وجود داشته باشد حفظ می‌کنند. ویژگی های مورد استفاده اغلب مجموعه‌ای از ویژگی‌های متداولی مانند ضرایب انرژی، نرخ عبور از صفر[2] (ZCR)، ضرایب کپسترال فرکانس مل[3] (MFCCs) و یا ترکیبی از آن‌ها هستند. بسیاری از ویژگی‌ها مانند MFCC از مؤلفه‌های طیفی در طول پنجره‌های زمان-کوتاه

[1] Support Vector Machine
[2] Zero Crossing Rate
[3] Mel Frequency Cepstral Coefficients

میانگین‌گیری می‌کنند. اطلاعات زمانی نیز از طریق الحاق مشتقات مرتبه اول و مرتبه‌دوم در ضرایب MFCC گنجانده می‌شود. این ویژگی های دلتا، شروع و پایان پویایی سیگنال را نشان می‌دهند.

نمونه‌هایی از کاربردهای شناسایی صداهای محیطی در نظارت را می‌توان در [۱۲] و [۱۳] یافت. مرجع [۱۲] تشخیص وقایع فریاد و شلیک گلوله را در محیط‌های نویزی، با استفاده از دو مدل مخلوط گوسی[1] (GMM) موازی برای تمایز هر دو صدا از نویز درنظر گرفته است. نویسندگان از روش رتبه‌بندی و انتخاب ویژگی برای پیدا کردن یک بردار ویژگی بهینه استفاده کرده‌اند که بازده‌ی دقت ۹۰ درصد و نرخ رد اشتباه[2] (FRR) ۸ درصد را نتیجه داده است. مرجع [۱۳] سیستمی برای نظارت در محیط‌های نویزی را با تمرکز بر تشخیص صدای شلیک و نرخ رد کاذب پایین -که در شرایط امنیتی مهم می‌باشد- ارائه داده است. این پژوهش بیان می‌کند که سطح نویز پایگاه داده‌ی آموزش، تأثیر قابل‌توجهی بر نتایج دارد، زیرا اطلاع از سطح نویز می‌تواند تشخیص نادرست را کاهش دهد.

در [۲۷]، ارزیابی سیستم‌های تشخیص و طبقه‌بندی در محیط اتاق جلسه انجام گرفته است. کار برای طبقه‌بندی رویدادها، فعالیت‌ها و روابط کارگاه[3] (CLEAR) که بخشی از پروژه‌ی کامپیوترها در حلقه تعامل انسان[4] (CHIL) می‌باشد، ترتیب داده شده است [۲۸]. در این کار، وقایع آکوستیکی مورد نظر عبارتند از قدم زدن، تایپ کردن صفحه کلید، کف زدن، سرفه وخنده. گفتار در طول جلسات نادیده گرفته شده است. چندین سیستم مختلف، به کارگرفته شده که یکی مبتنی بر روش جداسازی SVM است و دوتای دیگر مبتنی بر روش تشخیص گفتار معمولی با استفاده از HMM هستند. در این پژوهش مشاهده شده که عمل تشخیص یا تقسیم‌بندی، سخت‌ترین بخش کار است، درحالی‌که طبقه‌بندیِ بخش‌های تشخیص‌داده‌شده، دقت مناسبی به همراه داشته است. هم‌چنین روش‌هایی که به طور مستقیم از تشخیص گفتار گرفته شده‌اند به خوبی عمل می‌کنند. به همین دلیل، این روش‌ها می‌توانند

[1] Gaussian Mixture Model
[2] False Rejection Ratio
[3] Classification of Events, Activities and Relationships
[4] Computers in the Human Interaction Loop

بدون هیچ گونه تغییراتی بهترین روش موجود برای شناسایی وقایع صوتی در نظر گرفته شوند. از این‌رو در اولین کارهای صورت‌گرفته روی طبقه‌بندی و تشخیص وقایع صوتی، نمایش‌های پارامتری سیگنال‌های صوت مورداستفاده، به شدت بر روش‌های پیشین که برای پردازش گفتار و کارهای مرتبط مانند شناسایی گفتار و گوینده توسعه داده شده بودند مبتنی بود. از آنجا که این پارامترهای آکوستیکی معمولاً قاب به قاب استخراج می‌گردند معمولاً به عنوان ویژگی‌های زمان-کوتاه شناخته می‌شوند. MFCC متداول، انرژی‌های بانک فیلتر لگاریتمی، پیش‌بینی ادراکی خطی[1] (PLP)، انرژی لگاریتمی، شار طیفی[2]، آنتروپی و نرخ عبور از صفر مثال‌های خوبی در این زمینه هستند. ترکیب برخی از این ویژگی‌های زمان-کوتاه به بردارهای آکوستیکی با ابعاد بالاتر نیز ، هم‌زمان با استفاده الگوریتم‌های انتخاب ویژگی از این استخرهای بزرگ از مشخصات، به منظور کاهش ابعاد آن‌ها، با دقت مطالعه شده است [۱۸].

کار بسیاری از مراجع دیگر، اغلب به محیط‌های خاص متمرکز شده است. برای مثال، در [۱۱]، تجزیه و تحلیل صدای مته در طول عمل جراحی ستون فقرات در نظرگرفته شده است، زیرا اطلاعات مربوط به بافت را فراهم می‌کند و می‌تواند گذار بین مناطق با تراکم‌های مختلف استخوان را تشخیص دهد. مثال دیگر، شناسایی صدای پرندگان را در نظر گرفته و یک شبکه عصبی همراه با ویژگی‌های MFCC پیشنهاد می‌کند [۱۰].

با این‌حال، همان‌گونه که در [۱۸] اشاره شده است، بسیاری از این ویژگی‌های آکوستیکی مرسوم، لزوماً برای اهداف AED مناسب نیستند زیرا بسیاری از آن‌ها براساس مشخصات طیفی گفتار طراحی شده‌اند و کاملاً متفاوت از ساختار طیفی وقایع صوتی می‌باشند. به علاوه، برخی از انواع وقایع آکوستیکی یک ساختار زمانی ارائه می‌دهند. برای مثال، الگوی متناوب صدای زنگ تلفن که باید برای

[1] Perceptual Linear Prediction
[2] Spectral Flux

بهبود نمایش ویژگی و توانایی‌های جداسازی، به طریقی بهره برده شود. به این دو دلیل، پژوهش‌های اخیر بر یافتن مجموعه‌ای از ویژگی‌ها که وقایع صوتی را به خوبی نمایش دهند متمرکز شده است.

برای مقابله با مشکل اول، پارامترهای جدید صوتی مانند ضرایب کپسترال توان نرمالیزه شده[۱] (PNCC) [۲۹] و پارامترهای به دست آمده از بانک‌های فیلتر گاماتون[۲] [۳۰] و یا گاماچیرپ[۳] [۳۱] ارائه شده است. در پژوهش‌های دیگر برای کشف ساختار نهفته‌ی داده‌های صوتی با استفاده از فاکتورگیری ماتریس نامنفی[۴] (NMF) یا تجزیه به K-مقدار منحصر به فرد[۵] (K-SVD) روی اسپکتروگرام‌های صوتی [۳۲] تلاش شده است. در یک روش دیگر [۳۳]، از تحلیل مشخصات طیفی وقایع صوتی، اهمیت فرکانس متوسط و بالا برای جداسازی بین وقایع مختلف صوتی نتیجه گرفته شده و منجر به طراحی جدید مبتنی بر فیلترینگ بالا گذر سیگنال‌های صوتی گردیده است. این کار در شرایط تمیز و نویزی به نتایج خوبی دست یافته است [۳۴]. باید توجه داشت که تمام این روش‌ها به عنوان تغییرات مختلف از بانک فیلتر شنوایی مرسوم مقیاسِ مل[۶] هستند که به اسپکتروگرام‌های صوتی در فرآیند استخراج ویژگی کوتاه مدت اعمال می‌شود.

۲-۳-۲ طبقه بندی محیط

طبقه‌بندی محیط، عمل شناخت محیط‌های فعلی مانند خیابان، آسانسور و یا ایستگاه راه‌آهن از یک قطعه‌ی صوتی کوتاه است و گاهی اوقات به عنوان تشخیص صحنه از آن نام برده می‌شود. اطلاعات به‌دست‌آمده از محیط، برای دستگاه‌های حساس به محتوا قابل‌استفاده است و می‌تواند اطلاعات ارزشمندی در مورد محل و فعالیت کاربر ارائه کند.

[1] Power-Normalized Cepstral Coefficients
[2] Gammatone Filter Banks
[3] Gammachirp
[4] Non-negative Matrix Factorization
[5] K-Singular Value Decomposition
[6] Mel-scaled

در [۳۵]، نویسندگان، یک طبقه‌بند محیط صوتی مبتنی بر HMM توسعه داده‌اند که شامل مدل‌سازی سلسله مراتبی و یادگیری انطباقی است. آن‌ها یک مدل سلسله مراتبی پیشنهاد کرده‌اند که ابتدا برای مطابقت نویز پس‌زمینه با محیط تلاش می‌کند، اما اگر نمره اطمینان پایین ثبت شود، بخش، برخلاف منابع مشخصی که ممکن است در محیط وجود داشته باشند دسته‌بندی می‌شود. در آزمایش‌ها تنها با استفاده از دسته‌ی محیط‌های پس‌زمینه، سیستم آن‌ها میانگین دقت ۹۷ درصد را می‌دهد که از شنوندگان انسان که تنها ۳۵ درصد پاسخ صحیح داده‌اند نتیجه‌ی بهتری دارد.

مثال دیگر در [۲] یافت می‌شود، که تشخیص صحنه برای ربات‌های متحرک را درنظر می‌گیرد. نویسندگان این گزارش از مجموعه‌ای ترکیبی از ویژگی‌ها همراه با یک الگوریتم انتخاب ویژگی استفاده کرده‌اند و عملکرد با طبقه‌بندهای K-نزدیکترین همسایه[1] (KNN)، SVM و GMM سنجیده شده است. بهترین سیستم کلی طبقه‌بندی کننده KNN است، که دقت طبقه‌بندی محیط ۹۴/۳ درصد با استفاده از ۱۶ ویژگی بدست می‌دهد، و دارای زمان اجرای سریعتر نسبت به روش‌های SVM و GMM است.

۲-۳-۳ شناسایی موسیقی

دو عمل اصلی در شناسایی موسیقی، تشخیص ابزار و طبقه‌بندی سبک موسیقی است. شباهت‌هایی با صداهای محیطی وجود دارد که به عنوان مشکل اول، طبقه‌بندی صداهایی را که از ترکیب منابع چندگانه حاصل می‌شوند درگیر می‌سازد. از سوی دیگر، شناسایی سبک، معادل با طبقه‌بندی محیط است.

شناسایی ابزار، نیازمند شناخت ابزارهای مورد استفاده در یک قطعه‌ی مشخص می‌باشد، که مستلزم مطالعات مجزا روی ابزارها است. این درحالی‌است که اخیراً موسیقی مورد پسند جوامع، چندابزاری[2] است. در چنین حالتی سیستم شناسایی باید صداهای دارای هم‌پوشانی و ابزارهایی را که می‌توانند

[1] K-Nearest Neighbor
[2] Polyphonic

هم‌زمان نوت‌های چندگانه اجرا کنند، جدا کند. از این‌رو سیستم شناسایی با مشکل مواجه است. در [۳۶]، نویسندگان، سیستمی با الهام از ایده‌های تجزیه و تحلیل محاسباتی صحنه‌ی شنیداری و از تقسیم‌بندی تصویر توسط افراز گراف توسعه داده‌اند. آنها از فاصله اقلیدسی بین یک نمونه‌ی درون‌یابی شده و خوشه‌ی قاب ورودی برای اندازه‌گیری شباهت طنین با ابزارهای آموزش‌داده‌شده استفاده می‌کنند. در ابزارهای مجزا، سیستم به دقت طبقه‌بندی ۸۳ درصد دست یافته، اگرچه برای کلارینت و ویولن ضعیف عمل کرده است. در مخلوط‌هایی از چهار نوت، سیستم قادر به تشخیص ۵۶ درصد رخداد بادقت شناسایی ۶۴ درصد است. جدیدترین کار همان گروه در [۳۷] یافت می‌شود که برای برای مشخصه‌یابی موسیقی چندابزاری تلاش کرده‌اند.

وظیفه‌ی تعیین نوع موسیقی یک قطعه‌ی کوتاه، طبقه‌بندی سبک موسیقی مانند کلاسیک، بلوز، جاز، کانتری، راک، متال، رگی، هیپ-هاپ و پاپ می‌باشد. شاید معروف‌ترین کارها در این زمینه [۳۸] و دیگر مقالات منتشر شده پس از آن از جمله [۳۹, ۴۰] باشند. مرجع [۳۸] تعدادی از ویژگی‌های خاص را برای جدا کردن سبک‌های مختلف موسیقی توسعه داده‌است. این کار شامل ویژگی‌های بافت طنینی و ویژگی‌های محتوای ریتمیک مانند تشخیص اوج می‌باشد. در این پژوهش دقت طبقه‌بندی ۶۱ درصد بدست آمده و نویسندگان را قادر می‌سازد تا آن را با شنوندگان انسان که دقت ۷۰ درصد به دست آورده‌اند مقایسه کنند.

۲-۳-۴ نمایه سازی و بازیابی

دسترسی آسان به صوت ذخیره شده در پایگاه داده‌ای که از انواع منابع بدست می‌آید یک توانایی مورد نیاز است. در این رابطه سه موضوع اصلی برای پژوهش وجود دارد: بخش‌بندی، برای تعیین نقاط شروع و پایان واقعه؛ نمایه‌سازی، که ذخیره‌سازی اطلاعات تشخیص آن از انواع دیگر وقایع است؛ و بازیابی، که درآن، کاربر همه‌ی وقایع از یک نوع را جستجو می‌کند [۱۷].

مهم‌ترین جنبه، توسعه‌ی مجموعه‌ای از ویژگی است که به طور منحصر به فرد یک نوع رویداد را تعریف می‌کند و می‌تواند با طیف وسیعی از توصیفات، مانند ویژگی‌های فیزیکی، شباهت با سایر صداها، توصیفات ذهنی، یا محتوای معنایی مانند متن یا نمره‌ی گفته شده جستجو شود. این چیزی است که آن را از یک مسئله طبقه‌بندی ساده متفاوت می‌کند.

تعدادی از پژوهش‌ها به تازگی روی صداهای محیطی و موسیقی متمرکز شده‌اند. در [۴۱]، نویسندگان با اعمال تکنیک‌هایی که می‌تواند شباهت معنایی کلماتی مانند "خرخر" و "میو" را مقایسه کند تلاش کرده‌اند نتایج بازیابی را بهبود دهند. در [۴۲]، مجموعه‌ای از "توصیفات مورفولوژیکی" که توصیفات شکل، ماهیت و تغییرات صدا هستند، در نظر گرفته شده تا یک ویژگی برای نمایه‌سازی و بازیابی شکل دهد. یک کار دیگر در [۴۳]، سیستم "پرس‌وجوی متن" را ارائه کرده که می‌تواند بر اساس حاشیه‌نویسی معنایی از محتوا، از جمله موسیقی و جلوه‌های صوتی، بازیابی مناسب آهنگ‌ها را انجام دهد.

۲-۴- کاربرد روش های شناسایی گفتار/موسیقی برای شناسایی وقایع صوتی

محبوبترین تکنیک‌های شناسایی گفتار اغلب از ترکیب ویژگی‌های MFCC با طبقه‌بند HMM استفاده می‌کند [۲۵]. اگرچه ویژگی‌ها و طبقه‌بندهای دیگری مانند شبکه‌های عصبی مصنوعی[1] نیز می‌تواند استفاده شود. با این حال، روش محبوب MFCC-HMM اغلب به خوبی عمل می‌کند؛ زیرا به‌واسطه‌ی MFCCها، نمایشی فشرده از طیف فرکانسی را با یک طبقه‌بند ترکیب می‌کند که می‌تواند تغییرات زمانی یک رویداد صوتی را از طریق انتقال بین حالت‌های مختلف یک مدل، مدل‌سازی کند. نشان داده شده است که این روش با صداهای محیطی، به همان خوبی که روی گفتار نتیجه داده، عمل می‌کند [۲۳]. در [۳۵]، طبقه‌بندی محیط‌های صوتی با استفاده از ویژگی‌های MFCC و طبقه‌بندی HMM، به طور متوسط ۹۲ درصد است. برای رویدادهای صوتی مجزا، با استفاده از یک پایگاه داده متشکل از

[1] Artificial Neural Network

۱۰۵ عمل، برخورد و صداهای مشخصه، دقت کلی ۸۵ درصد است. نمونه‌های دیگر در [۴۴] دوباره نشان می‌دهد که ترکیب MFCC و HMM برای طبقه‌بندی صدای محیطی به خوبی عمل می‌کند.

با این‌حال، [۴۵] بیان می‌کند که تکنیک‌های مبتنی بر HMM، با توجه به عدم وجود یک "الفبای صدای محیطی"، برای شناسایی وقایع صوتی مناسب نیستند. در [۲۳] ادعا شده است که این بیان درست نیست چون اگر الفبای صدا وجود نداشته باشد آموزش مدل‌های HMM برای هر طبقه از رویدادهای صوتی، به جای تعریف واحدهای فرعی صدا که در سراسر صداهای مختلف رایج است، ساده‌تر می‌باشد. شباهت با شناسایی گفتار، آموزش یک مدل HMM برای هر کلمه به جای آموزش یک مدل برای هر واحد زیرکلمه که تنها با وجود یک الفبای ثابت امکان‌پذیر است، می‌باشد. بنابراین، برای شناسایی صدای محیطی، یک HMM برای هر کلاس صدا آموزش داده می‌شود و برای رمزگشایی صداهای ناشناخته در محیط استفاده می‌گردد.

مرجع [۴۵] یک تحلیل از دو روش دیگر شناسایی گفتار ارائه می‌دهد: کوانتیزه‌کردن بردار خطی[1] (LVQ) و شبکه های عصبی مصنوعی. روش LVQ مبتنی بر تولید یک نمونه‌ی اولیه برای هر کلاس در طول آموزش می‌باشد و با نزدیک کردن نمونه‌ی اولیه به هر نمونه‌ی برنده، مرحله‌ی آزمون، نزدیک‌ترین نمونه‌ی اولیه را به نمونه‌ی مشاهده شده می‌یابد. این موضوع به متریک فاصله‌ی مورداستفاده وابسته است، ازاین‌روست که محبوبیت رویکرد HMM را به‌دست نیاورده است. نتایج نشان می‌دهدکه LVQ نتایج مشابه در هر دو آزمون گفتار و غیرگفتار دارد، درحالی‌که شبکه عصبی مصنوعی برای آزمون گفتار به خوبی عمل می‌کند، اما در آزمون غیرگفتار ضعیف است. مرجع [۴۵] اشاره می‌کند که این به‌علت شباهت کلاس‌های صدای مورد استفاده در این آزمایش است و LVQ یک طبقه‌بند بهتر برای انواع مشابه صدا می‌باشد. با این‌حال، این نتیجه باید به دقت تفسیر شود، زیرا معروف است که شبکه‌های عصبی مصنوعی با تعداد کافی از نرون‌های مخفی، تخمین‌زننده‌های جامعی هستند و باید قادر به نشان دادن مرز دلخواه بین دو طبقه باشند[۲۳].

[1] Learning Vector Quantization

نویسندگان دیگر، شبکه‌های عصبی مصنوعی را برای طبقه‌بندی رویدادهای صوتی با موفقیت استفاده کرده‌اند. در [۴۶]، بردارهای ویژگی MFCC با استفاده از کوانتیزه‌سازی بردار، پیش‌پردازش شده و برای طبقه‌بندی به یک شبکه عصبی داده می‌شوند. نتایج، دقت متوسط ۷۳ درصد را برای ۱۰ کلاس از صداهای محیطی نشان می‌دهد. مثال دیگر در [۱۰] مشاهده می‌شود که شناسایی گونه‌های پرندگان بر اساس ویژگی‌های MFCC با طبقه‌بندی شبکه عصبی ترکیب می‌شود. نویسندگان استفاده از ویژگی‌های MFCC را در شناسایی گفتار تأیید می‌نمایند، اما بیان می‌کنند که با توجه به کیفیت ادراکی فیلتر مل، خواص، کاهش پیدا می‌کند، و این‌که می‌توانند هر دو سیگنال نامتناوب و متناوب را که برای صدای پرندگان مناسب است توصیف کنند. نویسندگان دقت شناسایی ۸۷ درصد را برای صدای ۱۴ گونه پرنده گزارش داده‌اند.

۲-۵- مرور آخرین دستاوردها در زمینه ی شناسایی وقایع صوتی

برخی از معتبرترین پژوهش‌های صورت گرفته در زمینه‌ی شناسایی وقایع صوتی محیط اداری در این بخش ارائه می‌شود. در سال ۲۰۱۳ پژوهشگران [۴۷] ابتدا با استفاده از فیلتر وینر[1] اقدام به کاهش نویز تک کاناله برای پاک‌سازی نویز ایستان پس‌زمینه از فایل‌های صوتی نمودند. سپس با استفاده از روش استخراج ویژگی MFCC، ۱۳ ضریب استخراج نموده و میانگین این ضرایب را برای آموزش به طبقه‌بند SVM اعمال کردند.در نهایت نرخ F-Score برابر با ۱۱/. برای شناسایی بر مبنای رویداد بدست آمده است.

پس از آن در [۴۸] ابتدا فایل‌های صوتی فراخوانی شده و ضرایب MFCC متناظر و مشتقات مرتبه اول و مرتبه‌دوم این ضرایب محاسبه می‌شود. برچسب‌زنی ویژگی‌های استخراج شده به ویژگی‌های وقایع حقیقی و ویژگی‌های نویز در دو مرحله صورت می‌گیرد. ابتدا، از دو مفسر متفاوت برای مکان‌یابی ویژگی‌ها استفاده شده است. نقطه شروع اولیه‌ی هر دو مفسر به عنوان شروع و نزدیکترین نقطه پایان

[1] Wiener

بعدی هر دو به عنوان پایان استفاده شده است. این کار، با هدف کاهش احتمال برچسب‌زنی ویژگی‌های وقایع به عنوان پس‌زمینه صورت گرفته است. سپس یک آستانه قبل از ضرایب MFCC برای حذف سکوت داخل رویداد (به طور مثال سکوت بین دو زنگ تلفن) اعمال شده است. برای آموزش و طبقه‌بندی داده‌ها نیز از طبقه‌بند GMM استفاده شده است. نتیجه‌ی این مراحل، رسیدن به F-Score برابر ۰/۵۶ است که نسبت به پژوهش‌های مرتبط پیشین افزایش چشم‌گیری داشته است.

در [۴۹] برای استخراج ویژگی از قاب‌های به طول ۸۰ میلی‌ثانیه معادل ۳۵۰۰ نمونه در فرکانس ۴۴/۱ کیلو هرتز استفاده شده است که طول نسبتاً زیادی است. ویژگی‌های مورد استفاده طیف وسیعی از انواع ویژگی‌های زمانی و فرکانسی را شامل می‌شود. هم‌چنین برای خوشه‌بندی و طبقه‌بندی داده‌ها از یک HMM مرتبه‌ای استفاده شده است. درنهایت بهترین پاسخ سیستم برای ترکیب ویژگی‌های MFCC و LPC و استفاده از HMM حاصل شده که F-Score برابر ۰/۴۴ است.

در [۵۰] یک سیستم شناسایی وقایع صوتی مبتنی بر مدل مخفی مارکوف ،که از رمزگشایی ویتربی[1] برای یافتن محتمل‌ترین دنباله از وقایع استفاده می‌کند، پیشنهاد شده است. دراین سیستم از پنجره‌هایی به طول ۱۰ میلی‌ثانیه استفاده شده است. از این قاب‌ها ضرایب MFCC استخراج شده و برای شناسایی به HMM داده می‌شود. در پایان، پارامتر F-Score برای این سیستم، برابر ۰/۵۵ بدست آمده است.

نمونه‌ای دیگر در [۵۱] ارائه شده که یک روش نمونه‌محور مبتنی بر فاکتورگیری ماتریس نامنفی، برای آشکارسازی وقایع صوتی پیشنهاد داده است. بنابر آخرین کارهای صورت گرفته در شناسایی خودکار گفتار مقاوم به نویز، وقایع را به عنوان ترکیب خطی از اتم‌های واژه‌نامه، و مخلوط‌ها را به عنوان ترکیب خطی وقایع دارای هم‌پوشانی مدل کرده‌اند. وزن اتم‌های فعال شده در یک مشاهده، به طور مستقیم به عنوان مدرک برای کلاس‌های واقعه‌ی مربوطه به کار برده می‌شود. اتم‌های واژه‌نامه شامل چند قاب

[1] Viterbi Decoding

هستند و از طریق استخراج تمام نمونه‌های طول-ثابتِ ممکن از داده‌های آموزشی ایجاد می‌شوند. در این کار از نمونه‌های ۲۰۰ میلی ثانیه معادل ۲۰ قاب به طول ۲۵ میلی ثانیه که دارای هم‌پوشانی ۱۰ میلی‌ثانیه‌ای در فرکانس نمونه‌برداری ۴۴/۱ کیلوهرتز هستند استفاده شده است. واژه‌نامه‌ی حاصل شده در نهایت دارای ۱۰۶۲۱ اتم است. برای غلبه بر مشکل کم بودن طول داده‌های آموزشی، استفاده از چرخش زمانی خطی در فضای ویژگی برای بسط داده‌های آموزشی پیشنهاد شده است. در نهایت، F-Score برابر ۰/۴۷ حاصل شده است.

در [۵۲] ابتدا از طریق یک مرحله کاهش نویز، کیفیت سیگنال ارتقاء داده شده است. در مرحله‌ی بعدی، استخراج ویژگی از قاب‌های به طول ۳۲ میلی‌ثانیه با هم‌پوشانی ۱۶ میلی‌ثانیه‌ای در فرکانس نمونه‌برداری ۴۴/۱ کیلوهرتز، با استفاده از بانک فیلتر گابور صورت می‌گیرد. و مرحله‌ی طبقه‌بندی با استفاده از مدل مخفی مارکوف دو لایه انجام می‌شود. پارامتر F-Score در این سیستم برابر ۰/۵۱ بدست آمده است.

در [۴] که جزو آخرین کارهای صورت گرفته در زمینه‌ی شناسایی وقایع صوتی می‌باشد، نویسندگان اقدام به تحلیل اسپکتروگرام‌های وقایع صوتی در طیف فرکانسی مل و در بازه‌هایی به طول ۱۰۵ میلی‌ثانیه کرده‌اند که از این طریق عمل استخراج ویژگی صورت گرفته است. برای طبقه‌بندی داده‌ها نیز از طبقه‌بند مدل مخفی مارکوف استفاده شده است و پارامتر F-Score مبتنی بر رویداد، برابر ۰/۶۲ بدست آمده است.

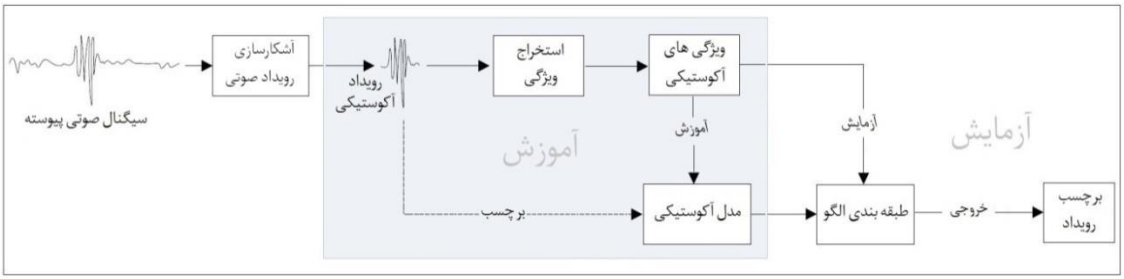
فصل ۳ اصول شناسایی وقایع صوتی

۳-۱- ساختار سیستم شناسایی

وظیفه‌ی سیستم شناسایی، تشخیص وقایع صوتی از یک نمونه‌ی صوتی و تعیین برچسب مناسب برای آن رویداد، بر اساس آموزش انجام شده بر روی نمونه‌های مشابه است. چنین سیستمی می‌تواند به صورت "برخط" یا "برون‌خط" طبقه‌بندی شود. سیستم برخط باید وقایع آکوستیکی را به محض رخ دادن تشخیص دهد و آن‌ها را به‌صورت بلادرنگ پردازش کند تا نزدیک‌ترین نظیر را بیابد. سیستم برون‌خطی، صوت را در دسته‌ها پردازش می‌کند و با دنباله‌ای از صداها سر وکار دارد.

معمولاً سیستم‌های"زنده"ی برخط چالش‌برانگیزتر هستند، زیرا نیازمند تصمیم‌گیری در قطعه‌های صوتی نسبتاً کوتاه برای کاربردهایی مانند نظارت و یا طبقه‌بندی محیط ربات می‌باشند. علاوه بر این، به‌طورمعمول محاسبات سنگین مقبول نیست، زیرا چنین سیستم‌هایی در دستگاه‌های قابل حمل با توان محاسباتی پایین استفاده می‌شوند.

یک سیستم معمولی شناسایی برخط از فرآیندهای نشان داده شده در شکل ۳-۱ تشکیل می‌شود. همان‌گونه که در این شکل نشان داده‌شده است یک سیگنال صوتی پیوسته گرفته می‌شود و قطعه‌های صوتی کوتاه که هرکدام شامل یک واقعه‌ی صوتی است استخراج می‌گردد. سپس اطلاعات مفید برای ایجاد ویژگی آکوستیکی جهت طبقه‌بندی از این رویداد استخراج می‌شود. در طول آموزش، این ویژگی‌های صوتی برای آموزش مدل آکوستیکی استفاده می‌شود، که اطلاعات را برای انواع مختلف رویدادهای صوتی می‌گیرد. در حین آزمایش، طبقه‌بندی الگو، برای مطابقت ویژگی‌های صوتی ناشناخته با مدل آکوستیکی، با هدف تولید یک برچسب برای این رویداد صورت می‌گیرد. در بخش‌های بعدی، هر یک از مراحل به صورت مفصل شرح داده می‌شود.



شکل ۳-۱: ساختار یک سیستم تشخیص وقایع صوتی.

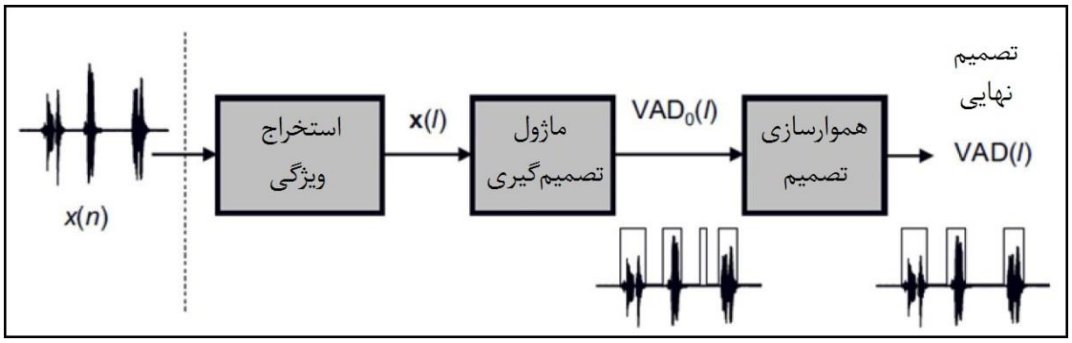
۳-۱-۱ آشکارسازی رویداد صوتی

آشکارسازی، به پیداکردن نقاط شروع و پایان حوادث آکوستیکی در یک جریان پیوسته‌ی صوتی اطلاق می‌شود، به‌طوری‌که سیستم استخراج ویژگی تنها با بخش‌های فعال سر و کار دارد. این فرآیند، به طور خاص در مورد تقسیم‌بندی گفتار/غیرگفتار، آشکارسازی واقعه‌ی صوتی یا آشکارسازی فعالیت صوتی[1] (VAD) نیز گفته می‌شود. آشکارسازی، تنها در مورد صدای زنده که ورودی، یک جریان صوتی پیوسته می‌باشد مورد نیاز است. با این‌حال، هم‌زمان‌که توانایی سیستم در کار بلادرنگ مهم است، به‌طور معمول در روند توسعه‌ی مطالعات برای ارزیابی سیستم شناسایی وقایع آکوستیکی، قطعه‌های صوتی شامل وقایع صوتی مجزا هستند. در این مورد، نیازی به آشکارسازی وقایع صوتی نیست.

طرح کلی یک سیستم معمولی آشکارسازی، در شکل ۳-۲ نشان داده شده است که در آن ابتدا ویژگی‌ها از سیگنال پیوسته استخراج شده، سپس تصمیم گرفته می‌شود و پس از آن برای هموار کردن خروجی آشکارساز، پس‌پردازش صورت می‌گیرد. الگوریتم‌های واحد تصمیم‌گیری اغلب به دو دسته تقسیم می‌شوند: قاب-آستانه-آشکارسازی، یا آشکارسازی-به‌وسیله‌ی-طبقه‌بندی [۵۳]. اولی، براساس ویژگی در سطح قاب تصمیم می‌گیرد که قاب موردنظر شامل واقعه است یا نویز. بخش تصمیم‌گیری به طور ساده یک آستانه است؛ بدین‌صورت که اگر خروجی ویژگی بیش‌تر از یک مقدار معین باشد، تصمیم مثبت خواهد بود. آشکارساز در این روش، برخلاف روش آشکارسازی با طبقه‌بندی، عمل طبقه‌بندی سیگنال را انجام نمی‌دهد و به عنوان جایگزین، بخش‌های فعال را از

[1] Voice Activity Detection

جریان صوتی پیوسته استخراج می‌کند تا به سیستم طبقه‌بندی بدهد. ساده‌ترین ویژگی برای چنین سیستمی می‌تواند سطح توان قاب باشد. بدین‌صورت که اگر توان کل در یک قاب مشخص، بیش از یک آستانه‌ی معین باشد، آن قاب، فعال شناخته می‌شود. با این‌حال، چنین عملی بسیار ساده و مستعد خطا در نویز غیرایستان است. ویژگی‌های ممکن دیگر، شامل تخمین پیچ، نرخ عبور از صفر یا آمار مرتبه‌ی بالاتر است. هم‌چنین بهبود عملکرد برای زمانی که ویژگی‌ها از پنجره‌ی زمانی طولانی‌تر استفاده کنند گزارش شده است [۵۴]. این کار مزایایی چون هزینه‌ی محاسباتی کم و پردازش بلادرنگ دارد. اما معایبی نیز وجود دارد. مانند انتخاب آستانه که بسیار مهم می‌باشد و ممکن است در طول زمان تغییر کند.



شکل ۳-۲: نمودار بلوکی آشکارسازی رویداد صوتی [۲۳].

روش‌های آشکارسازی با طبقه‌بندی چنین مشکلاتی ندارند، زیرا برای برچسب زدن بخش به عنوان نویز و یا غیرنویز، از یک طبقه‌بند به جای استفاده از آستانه بهره می‌گیرند. در این پیکربندی، یک پنجره‌ی متحرک بر روی سیگنال عبور می‌کند و مجموعه‌ای از ویژگی‌ها از هر پنجره استخراج می‌شود. این ویژگی‌ها به طبقه‌بندی که برای جداسازی نویز از وقایع غیرنویز آموزش دیده است داده می‌شوند. طبقه‌بند باید یک مرحله آموزش را بگذراند تا یاد بگیرد که چه ویژگی‌هایی معرف نویز هستند. سیستم تصمیم بدون نظارت ساده می‌تواند مبتنی بر خوشه‌بندی و GMM باشد. یک بخش کوتاه صوتی، شامل نویز و غیرنویز، در فاز آموزش خوشه‌بندی می‌شود؛ بنابراین یک خوشه باید شامل نویز و خوشه‌ی دیگر شامل سایر وقایع باشد. مدل مخلوط گوسی سپس می‌تواند به توزیع هر خوشه منطبق شود به‌طوری‌که قاب‌های آینده، با هر GMM مقایسه خواهد شد و محتمل‌ترین‌شان به عنوان برچسب انتخاب می‌شود.

البته طرح‌های طبقه‌بندی پیشرفته‌تر را نیز می‌توان استفاده کرد، مانند ترکیب درخت‌های تصمیم‌گیری و انتخاب ویژگی که می‌تواند عملکرد را بهبود بخشد [۵۵].

۳-۱-۲ استخراج ویژگی

هدف از استخراج ویژگی، فشرده‌سازی سیگنال‌های صوتی در یک بردار است تا این بردار، نماینده‌ی کلاس رویداد صوتی‌ای باشد که می‌خواهد آن را توصیف کند. یک ویژگی خوب باید نسبت به عوامل خارجی مانند نویز و یا محیط حساس نبوده و قادر به تأکید بر تفاوت میان کلاس‌های مختلف صدا باشد و تغییرات یک کلاس صدای معین را کوچک نگاه دارد. این کار، طبقه‌بندی را ساده‌تر می‌کند، زیرا جدا کردن کلاس‌های مختلف صدا که قابل جداسازی هستند آسان است. یکی از مشکلات اصلی در ساخت یک سیستم خودکار تشخیص صدا، انتخاب ویژگی‌های مناسب سیگنال است که بتواند منجر به تفکیک مؤثر بین محیط‌های مختلف صدا شود. وقایع صوتی داده‌های بدون ساختار متشکل از منابع مختلف هستند و برخلاف موسیقی یا گفتار، در مورد تکرار قابل پیش‌بینی و یا ساختار هارمونیکی در سیگنال هیچ فرضی نمی‌توان داشت. به دلیل همین ماهیت داده‌های بدون ساختار، ایجاد یک تعمیم برای بیان کمّیِ آن‌ها دشوار است. با توجه به تنوع ذاتی برای توصیف سیگنال‌های صوتی، ویژگی‌های بسیاری وجود دارد که می‌تواند استفاده شود. انتخاب مناسب این ویژگی‌ها در ایجاد یک سیستم آشکارسازی قوی بسیار مهم است.

به طور کلی، دو روش برای استخراج ویژگی‌ها وجود دارد که با توجه به وسعت زمان تحت پوشش ویژگی‌ها تغییر می‌کند [۵۶]. استخراج ویژگی یا سراسری است، که در آن توصیف‌گر، از کل سیگنال تولید می‌شود، یا محلی است، که در آن از هر قاب زمانی کوتاه درحدود ۳۰ تا ۶۰ میلی‌ثانیه از طول سیگنال، یک توصیفگر تولید می‌شود. روش دوم، مرسوم‌ترین روش استخراج ویژگی است و اغلب رویکرد مجموعه‌ای از قاب‌ها نامیده می‌شود.

از آنجا که شناسایی گفتار، زمینه‌ی غالب در شناسایی الگوهای صوتی می‌باشد استفاده‌ی مستقیم از ویژگی‌های توسعه‌داده‌شده برای گفتار، برای وقایع صوتی نیز متداول است. محبوب‌ترین ویژگی، ضرایب کپسترال فرکانس مل است، اگرچه موارد دیگر مانند ضرایب کپسترال پیش‌بینی خطی (LPCC) نیز استفاده می‌شود. هنوز، راه‌های بسیاری برای تجزیه و تحلیل سیگنال وجود دارد، از این‌رو طیف گسترده‌ای از ویژگی‌های دیگر که برای به‌دست آوردن اطلاعات موجود در سیگنال توسعه یافته‌اند وجود دارد. اما به طور کلی، ویژگی‌های صوتی را می‌توان به سه دسته گروه‌بندی کرد: حوزه‌ی زمان (یا ویژگی‌های زمانی)، حوزه‌ی فرکانس (یا ویژگی‌های طیفی) و حوزه‌ی زمان-فرکانس. تعدادی از این ویژگی‌ها که در این پایان‌نامه استفاده شده شرح داده می‌شود. دو مقیاس حوزه‌ی زمان که بسیار استفاده می‌گردد در زیر آورده شده است.

انرژی کوتاه‌مدت:

$$E_n = \frac{1}{N}\sum_m [x(m)\omega(n-m)]^2 \qquad (۱-۳)$$

که در آن $x(m)$ سیگنال صوتی گسسته، n شاخص زمان انرژی کوتاه مدت و $\omega(m)$ پنجره‌ای به طول N می‌باشد. انرژی کوتاه‌مدت، نمایشی مناسب از تغییر دامنه در طول زمان فراهم می‌کند.

نرخ عبور از صفر کوتاه مدت متوسط (ZCR):

$$Z_n = \frac{1}{2}\sum_m |sgn[x(m)] - sgn[x(m-1)]|\omega(n-m) \qquad (۲-۳)$$

که در آن

$$sgn[x(n)] = \begin{cases} 1, & x[n] \geq 0, \\ -1, & x[n] < 0. \end{cases}$$

عبور از صفر وقتی که نمونه های متوالی علامت مختلف دارند رخ می‌دهد، و ZCR، متوسط تعداد دفعاتی است که علامت سیگنال در پنجره زمان کوتاه تغییر می‌کند.

به طور مشابه، تعدادی از ویژگی‌های طیفی نیز ارائه شده است. این ویژگی‌ها به‌طور معمول توسط یک بار اعمال تبدیل فوریه (به صورت تبدیل سریع فوریه یا FFT) به بخش‌های پنجره کوتاه مدت از سیگنال

های صوتی به دست می‌آیند که با پردازش بیش‌تر برای به‌دست آوردن ویژگی‌های مطلوب دنبال می‌شود. برخی از این ویژگی‌ها که معمولاً مورداستفاده قرار می‌گیرند عبارتند از:

نرخ باند انرژی: نسبت انرژی در یک باند فرکانسی خاص به کل انرژی است. در این پایان‌نامه از ۲۳ باند فرکانسی (با دانه‌بندی ریزتر برای زیر ۱۰۰۰ هرتز) استفاده می شود.

ضرایب پیشگویی خطی(LPC): ایده‌ی اصلی تحلیل LPC این است که نمونه‌ی صوت را می‌توان به عنوان ترکیبی خطی از نمونه‌های قبلی تخمین زد به‌طوری‌که یک ویژگی قوی در برابر تغییرات ناگهانی فراهم کند. از طریق حداقل‌سازی مجموع مربع تفاضلات (در یک بازه‌ی محدود) بین صوت واقعی و نمونه‌ای که پیش‌گویی خطی شده، یک مجموعه‌ی منحصر به فرد از ضرایب پیش‌گویی تعیین می‌شود.ضرایب فیلتر LPC توسط یک مدل تمام قطب توصیف می‌شود. در این پایان‌نامه از الگوریتم LPC در جعبه‌ابزار متلب استفاده شده است.

روش ادراکی، بر تبدیل فرکانس متناظر با دریافت ذهنی سیستم شنوایی انسان که از مقیاس خطی پیروی نمی‌کند مبتنی است. در این پژوهش، مقیاس‌های ادراکی مانند مل و بارک استفاده شده است. در این پایان‌نامه از ضرایب کپسترال فرکانس مل ، ضرایب پیشگویی خطی ادراکی، و ضرایب پیشگویی خطی ادراکی اصلاح شده، برای استخراج ویژگی استفاده شده است. طرح کلی استخراج ویژگی با استفاده از این سه روش در شکل (۳-۳) آمده است [۵۷].

ضرایب کپسترال فرکانس مل: بسیاری از سیستم‌های شناسایی گفتار از MFCC استفاده می‌کنند. طراحی آن‌ها از مشخصات کپسترال غیرخطی گوش انسان الهام گرفته شده است و باعث می‌شود تحت شرایط مختلف کارآمد و قوی باشند. روند استخراج MFCC شامل گام‌های زیر است. ابتدا سیگنالِ کلی به قاب‌ها تقسیم می‌شود. سپس بر روی هر قاب یک پنجره‌ی همینگ اعمال شده و ضرایب تبدیل فوریه محاسبه می‌شود. سپس تخمین تابع چگالی طیف توان محاسبه شده و با استفاده از توابع مثلثی

دارای هم‌پوشانی میانگین گرفته می‌شود. طراحی توابع فیلتر مثلثی شامل فرکانس مل، توسط معادله‌ی زیر ارائه شده است:

$$Mel(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right) \quad (۳-۳)$$

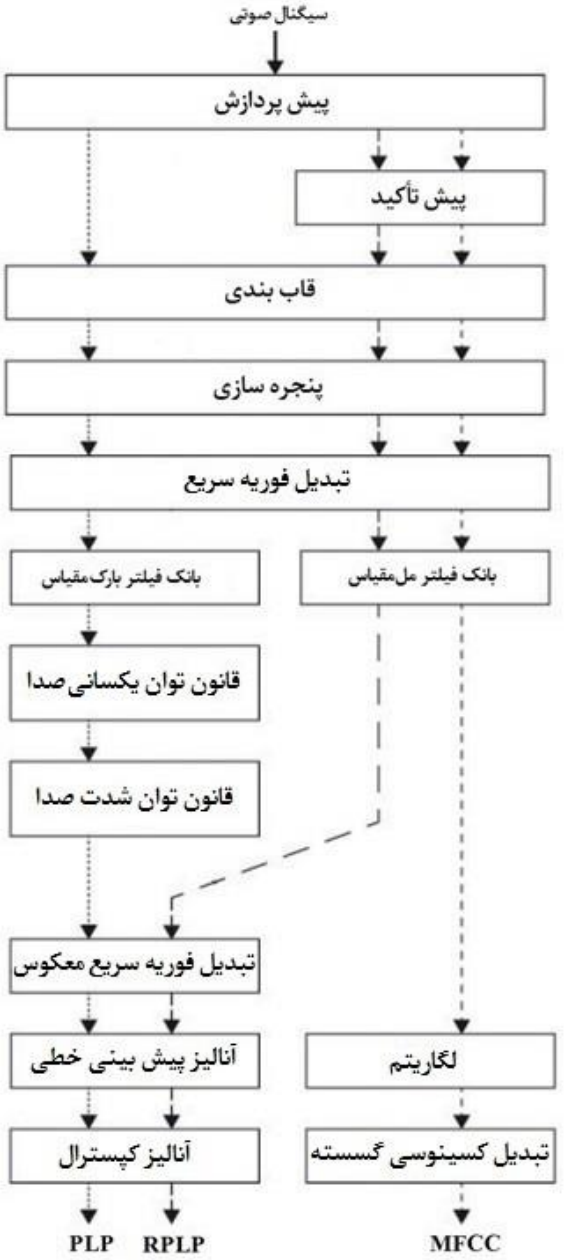
گام آخر، محاسبه‌ی تبدیل کسینوسی گسسته از لگاریتم تخمین، با استفاده از معادله‌ی زیر است:

$$c_k = \sqrt{\frac{2}{L}}\sum_{l=0}^{L-1} \ln\tilde{S}(l)\cos\left(\frac{\pi k}{L}\left(l + \frac{1}{2}\right)\right), \mathrm{k} = 0, 1, \ldots, \mathrm{q} - 1 \quad (۴-۳)$$

که در آن L تعداد توابع وزن و q تعداد ضرایب مل است. به طور معمول ۱۳ ضریب MFCC با استفاده از گرفتن تبدیل کسینوسی گسسته بدست می‌آید. MFCC پوشش زمانی را نیز دریافت می‌کند که از طریق استفاده از مؤلفه‌های دلتا و شتاب (دلتا دو) برآورد می‌شود. اگرچه، تغییرات زمانی فقط به صورت محلی طی چند قاب گرفته می‌شود.

ضرایب پیش‌گویی خطی ادراکی(PLP): تکنیک PLP از سه مفهوم روان‌فیزیک شنوایی استفاده می‌کند: رزولوشن طیفی باند بحرانی، منحنی برابری بلندی صدا و قانون توان شدت بلندی صدا. آنالیز PLP مشابه آنالیز فرکانس مل است. تفاوت‌ها در استفاده از مقیاس بارک به جای مقیاس مل و استفاده از مدل‌سازی تمام قطب خودکاهشی به جای تبدیل کسینوسی گسسته در حین یافتن ضرایب نتیجه است. ضرایب PLP کنونی بعد از MFCC بیش‌ترین کاربرد را در مشخصه‌یابی گفتار دارد.

پیش‌گویی خطی ادراکی اصلاح شده (RPLP): این مورد یک روش ترکیبی استخراج ویژگی است که توسط [۵۷] به منظور شناسایی زبان پیشنهاد شده است. برخلاف PLP، برای محاسبه‌ی RPLP از بانک فیلتر مل استفاده می‌شود.



شکل ۳-۳: فرآیند استخراج ویژگی با روش‌های مختلف [۵۷].

۳-۱-۳ طبقه بندی الگو

پس از استخراج ویژگی، الگوها باید به یکی از کلاس های ارائه شده که در طول آموزش طبقه‌بندی شده است تعلق یابد و یک برچسب به بخش صوتی اعمال شود. همان‌طور قبلاً ذکر شد، بسیار مهم است که ویژگی های طبقات مختلف متمایز و جداپذیر باشند، زیرا برای یک طبقه‌بند، تمایز بین کلاس‌های صدا که با هم تداخل دارند امکان‌پذیر نمی‌باشد. مهم‌ترین روش‌های طبقه‌بندی از مدل مخفی مارکوف،

مدل مخلوط گوسی، ماشین بردار پشتیبان، K-نزدیک‌ترین همسایه و شبکه‌های عصبی مصنوعی استفاده می‌کنند. این روش‌ها در زیر با جزئیات بیشتری مورد بحث قرار گرفته‌اند.

الف) مدل مخفی مارکوف: یک مدل مارکوف متشکل از حالت‌های متصل به هم می‌باشد که در آن گذار بین حالت‌های مختلف توسط مجموعه‌ای از احتمالات تعیین می‌شود و حالت بعدی تنها به حالت کنونی وابسته است. برای یک مدل مخفی مارکوف، تنها خروجی یا مشاهده از هر حالت برای ناظر قابل رؤیت است. از این‌رو، با دانستن گذار و توزیع احتمال دنباله‌ای از مشاهدات، نیازمند محاسبه‌ی محتمل‌ترین دنباله از حالت‌ها که می‌تواند برای مشاهدات حساب شود هستیم. پیاده‌سازی مدل مخفی مارکوف در عمل دارای پیچیدگی‌های بسیاری است؛ بنابراین سعی می‌شود که از طبقه‌بندهای دیگر با پیچیدگی محاسباتی کمتر استفاده شود.

ب) مدل مخلوط گوسی: امروزه یکی از رایج‌ترین روش‌های طبقه‌بندی داده، مدل مخلوط گوسی است که در آن توزیع بردارهای ویژگی یك کلاس با استفاده از یك توزیع احتمالاتی به شکل آمیزه‌ای از توابع توزیع گوسی مدل مي‌شود. پارامترهای مدل در GMM عبارتند از بردارهای میانگین، ماتریس‌های کواریانس و وزن‌های هر یك از توابع گوسی که در طی یك فرایند آموزشی تکراری مبتنی بر بیشینه‌سازی امید ریاضی[1] (EM) تخمین زده می‌شوند.

ج) ماشین بردار پشتیبان: ماشین بردار پشتیبان یک طبقه‌بندی دودویی است که ابرصفحه‌ی جداساز بین دو خوشه از نقاط در یک فضای با ابعاد بالا را محاسبه می‌کند. ماشین بردار پشتیبان متعارف جدایی خطی بین دو کلاس را در نظر می‌گیرد، با این‌حال، مدل‌های اصلاح شده، داده‌های دارای هم‌پوشانی، نگاشت‌های کرنل غیرخطی و راه حل مسئله‌های چندکلاسه را نیز پشتیبانی می‌کند. با توجه به عملکرد سریع طبقه‌بندی، این طبقه‌بند اغلب در کاربردهای برخط استفاده می‌شود. هرچند

[1] Expectation Maximization

که برای طبقه بندی گفتار مرسوم نیست زیرا مانند HMM، تحول زمانی سیگنال را به طور طبیعی مدل نمی‌کند.

د) k-نزدیکترین همسایگی: k-NN الگوریتمی ساده است که به الگوی آزمایش، با استفاده از رأی اکثریت k-نزدیکترین الگوهای آموزش یک برچسب کلاس اختصاص می‌دهد. این روش، اغلب به عنوان الگوریتمی کند شناخته می‌شود زیرا تمام محاسبات در مرحله‌ی آزمایش انجام می‌شود و از این‌رو می‌تواند عملکردی آهسته برای تعداد زیادی از نمونه‌های آموزشی داشته باشد.

ه) شبکه‌های عصبی مصنوعی: این روش، که هم‌زمان به پرسپترون چندلایه (MLP) نیز اشاره دارد، یک مدل محاسباتی الهام گرفته از نرون‌های مغز است. در تئوری، با تعداد کافی از نرون‌های مخفی، شبکه عصبی که یک تخمین‌گر عمومی شناخته می‌شود می‌تواند یک طبقه‌بند عالی باشد، اما به دلیل این‌که مانند یک جعبه سیاه بوده و تفسیر تابع عملکرد هر نرون در شبکه دشوار است، مورد انتقاد است. هم‌چنین این روش در آموزش مشکلاتی دارد، زیرا روش مرسوم پس‌انتشار به احتمال زیاد در کمینه‌ی محلی به دام می‌افتد.

۳-۲- معیار مقایسه

برای سنجش عملکرد هر کدام از مراحل آشکاسازی یا شناسایی وقایع صوتی، یک معیار مقایسه‌ی استاندارد معرفی شده است.

در مسئله‌ی آشکارسازی وقایع صوتی، به طور معمول از متریک F-Score به عنوان معیار مقایسه استفاده می‌شود. این متریک، که نماینده‌ای از نتایج مرحله‌ی آشکارسازی است به شکل رابطه‌ی (۳-۵) تعریف می‌شود:

$$F - Score = \frac{2 \times P \times R}{P + R} \qquad (۵-۳)$$

که R (فراخوانی[1]) و P (دقت[2]) به ترتیب به صورت زیر تعریف می‌شوند:

$$R = \frac{N_{Corr}}{N_e} \quad (۶-۳)$$

$$P = \frac{N_{Corr}}{N_d} \quad (۷-۳)$$

که N_{Corr}، تعداد قطعه‌های صوتی است که به درستی به عنوان رویداد صوتی آشکارسازی شده و نیز به درستی در کلاس مربوطه طبقه‌بندی شده‌اند، N_e، تعداد رویدادهای صوتی موجود در سیگنال پیوسته‌ی صوت است و N_d، تعداد تمام قطعه‌های صوتی آشکارسازی شده به عنوان رویداد صوتی است.

برای مرحله‌ی شناسایی نیز میانگین نرخ شناسایی که حاصل تقسیم تعداد نمونه‌های درست طبقه‌بندی شده به تمام نمونه‌ها می‌باشد استفاده می‌شود.

[1] Recall
[2] Precision

فصل ۴ روش های پیشنهادی استخراج ویژگی

۴-۱- مقدمه

در مسئله‌ی شناسایی وقایع صوتی، مانند بسیاری از سیستم‌های تشخیص الگو، انتخاب ویژگی‌های مناسب، کلید عملکرد مؤثر سیستم است. سیگنال‌های صوتی به طور معمول توسط ضرایب MFCC و یا برخی از نمایش‌های زمان-فرکانس دیگر مانند تبدیل فوریه‌ی زمان کوتاه و تبدیل موجک مشخصه‌یابی شده است. بانک فیلترهای مورد استفاده برای محاسبات MFCC، برخی از خصوصیات مهم سیستم شنوایی انسان را تخمین می‌زند. نشان داده شده است که ضرایب MFCC برای صداهای دارای ساختار، مانند گفتار و موسیقی خوب عمل می‌کنند، اما عملکرد آن‌ها در حضور نویز تنزل می‌یابد. ضرایب MFCC هم‌چنین در تجزیه و تحلیل سیگنال‌های نویزمانند که دارای یک طیف تخت هستند ناکارآمد است [۸]. وقایع صوتی که صداهای شبه نویزی هستند ممکن است با ضرایب MFCC به طور مؤثر مدل نشوند. از این‌رو در این پایان‌نامه، دو روش برای استخراج ویژگی وقایع صوتی محیط اداری پیشنهاد شده است: الف) الگوریتم پیگیری انطباق[1](MP)، ب) ضرایب کپسترال فرکانس بارک[2](BFCC)، که ابتدا استفاده از الگوریتم MP را برای تجزیه و تحلیل وقایع صوتی بررسی می‌کنیم و در ادامه استفاده از BFCC را مورد بررسی قرار خواهیم داد.

پیگیری انطباق، یک راه مناسب برای استخراج ویژگی‌های حوزه زمان-فرکانس فراهم می‌کند و می‌تواند هنگامی که استفاده‌از ویژگی‌های حوزه‌ی فرکانس شکست می‌خورد (به عنوان مثال ضرایب MFCC) به سیستم طبقه‌بندی کمک کند. روند کار شامل پیدا کردن تجزیه‌ی سیگنال از یک واژه‌نامه از اتم‌هاست که بهترین مجموعه از توابع را برای ایجاد یک نمایش تقریبی بدست می‌دهد.

الگوریتم MP در انواع کاربردها مانند رمزگذاری ویدئو و تشخیص نوت موسیقی استفاده شده است. پیگیری انطباق هم‌چنین در طبقه‌بندی سبک موسیقی و طبقه‌بندی خروجی‌های صوتی یک سیستم نظارت و نیز طبقه‌بندی نوع محیط پیرامون، مورد استفاده قرارگرفته است. در روش پیشنهادی ما، MP

[1] Matching Pursuit
[2] Bark Frequency Cepstral Coefficients

برای استخراج ویژگی در زمینه‌ی وقایع صوتی محیط اداری استفاده می‌شود. در این پایان‌نامه نشان داده شده است که ویژگی‌های مبتنی بر MP را می‌توان برای تکمیل ویژگی‌های سنتی حوزه‌ی فرکانس (MFCC) مورد استفاده قرار داد تا دقت شناسایی خودکار بالاتر برای وقایع صوتی حاصل شود.

هدف، استفاده از MP برای یادگیری ساختارهای ذاتی هر نوع از صداها به عنوان یک راه برای تشخیص کلاس‌های مختلف صدا می‌باشد. در این کار، ما یک تجزیه و تحلیل تجربی ویژگی را برای توصیف واقعه‌ی صوتی انجام داده و استفاده از MP را برای به‌دست آوردن ویژگی‌های مؤثر زمان-فرکانس بررسی می‌کنیم. نشان خواهیم داد که استفاده از MP این نمایش را ممکن خواهد ساخت. مزایای استفاده از این نمایش، توانایی گرفتن ساختار ذاتی هر نوع سیگنال و نگاشت یک سیگنال بزرگ و پیچیده به یک فضای ویژگی کوچک و ساده می‌باشد. مهم‌تر آن‌که، این روش به طور قابل توجهی نسبت به نویز پس زمینه مقاوم است و می‌تواند جایی که MFCC ناتوان است مشخصات سیگنال را بگیرد.

۴-۲- نمایش سیگنال با استفاده از پیگیری انطباق

ایده‌ی اصلی الگوریتم MP استفاده از ساختارهای نهفته‌ی موجود در سیگنال‌های مربوط به هر نوع صدا برای استخراج ویژگی‌های آن‌هاست. انواع مختلف صداها که دارای ویژگی‌های منحصر به فرد هستند باعث می‌شوند تجزیه به مجموعه‌ای از بردارهای پایه به طرز قابل توجهی متفاوت از دیگری‌ها باشد. با استفاده از یک واژه‌نامه که متشکل از طیف گسترده‌ای از توابع است، MP راهی کارآمد برای انتخاب مجموعه‌ی کوچکی از بردارهای پایه فراهم می‌کند و ویژگی‌های معنادار و هم‌چنین نمایش‌های انعطاف‌پذیر برای توصیف یک واقعه‌ی صوتی ایجاد می‌کند.

برای دستیابی به یک نمایش کارآمد، به دنبال حداقل تعداد بردارهای پایه برای نشان دادن یک سیگنال هستیم تا منجر به تقریب پراکنده شود. با این حال این یک مسئله‌ی NP[1]-کامل است. روش‌های تقریب

[1] Non-Deterministic Polynomial

تطبیقی متفاوتی در پژوهش‌های مختلف برای به دست آوردن چنین نمایشی از سیگنال پیشنهاد شده است؛ ازجمله پیگیری پایه[1] (BP) ، پیگیری انطباق (MP)، و پیگیری انطباق متعامد[2] (OMP). همه‌ی این روش‌ها از مفهوم واژه‌نامه استفاده می‌کنند که تجزیه‌ی یک سیگنال را با انتخاب بردارهای پایه از یک واژه‌نامه‌ی معلوم برای پیدا کردن بهترین زیرمجموعه انجام می‌دهد.

پیگیری اساس، چارچوبی فراهم می کند که نُرم یکِ ضرایب موجود در نمایش را به حداقل می‌رساند، اما منجر به هزینه‌ی بیش‌تر در محاسبات می‌شود و اگرچه نمایش خوبی فراهم می‌کند به لحاظ محاسباتی سنگین است. با استفاده از یک واژه‌نامه که متشکل از طیف گسترده‌ای از شکل موج‌های پایه‌ای است،MP برای یافتن تجزیه‌ی تنک یک سیگنال به نحو بهینه تلاش می‌کند. الگوریتم پیگیری انطباق بهینه نیست زیرا ممکن نیست تنک‌ترین پاسخ را بدهد. معمولاً عناصر یک واژه‌نامه‌ی معلوم، با به حداکثر رساندن انرژی حذف شده از سیگنال باقیمانده در هر مرحله انتخاب می‌شوند. حتی در فقط چند قدم، الگوریتم می‌تواند یک تقریب معقول با چند اتم به‌دست دهد، و تجزیه، تفسیری از ساختار سیگنال فراهم خواهد کرد. ما در این تحقیق روش MP کلاسیک را برای تولید ویژگی‌های صوتی اتخاذ کرده‌ایم.

الگوریتم MP اولین بار در [۵۸] برای تجزیه‌ی سیگنال‌ها در یک واژه‌نامه‌ی فراکامل از توابع معرفی شد که یک بسط خطی پراکنده از شکل موج‌ها فراهم می‌کرد. تا زمانی که واژه‌نامه فراکامل است، تضمین می‌شود که بسط به راه‌حلی همگرا شود که در آن سیگنال باقیمانده انرژی صفر دارد. توضیحات زیر از الگوریتمMP مبتنی بر [۸] می‌باشد.

اگر واژه‌نامه‌ی D مجموعه‌ای از شکل موج‌های پارامتری φ_γ باشد:

$$D = \{\varphi_\gamma : \gamma \in \Gamma\} \qquad (۱-۴)$$

[1] Basis Pursuit
[2] Orthogonal Matching Pursuit

که در آن Γ مجموعه‌ی پارامترها و φ_γ یک اتم نامیده می‌شود آن‌گاه تجزیه‌ی تخمینی یک سیگنال را می‌توان به صورت زیر نوشت:

$$s = \sum_{i=1}^{m} \alpha_{\gamma_i} \varphi_{\gamma_i} + R^{(m)} \qquad (۲-۴)$$

که $R^{(m)}$ باقیمانده‌ی سیگنال می‌باشد. با s، m و D معلوم، هدف، یافتن اندیس‌های γ_i و محاسبه‌ی α_{γ_i}، در شرایط $i = 1, 2, \dots, m$ و حداقل کردن $R^{(m)}$ می‌باشد. با شروع از مقدار اولیه‌ی $s^{(0)} = 0$ و $R^{(0)} = s$ الگوریتم MP یک زنجیره از تخمین پراکنده را می‌سازد.

الگوریتم MP در ابتدا ضرب داخلی سیگنال s را با تمامی اتم‌های واژه‌نامه‌ی D محاسبه می‌کند. اتمی که بزرگترین اندازه‌ی ضرب داخلی را داشته باشد یعنی φ_{γ_0} به عنوان اولین المان انتخاب می‌شود. از این‌رو معیار انتخاب اتم را می‌توان به شکل زیر نوشت:

$$\left|\langle \boldsymbol{s}, \varphi_{\gamma_0} \rangle\right| \geq \left|\langle \boldsymbol{s}, \varphi_{\gamma} \rangle\right|, \qquad \forall\, \boldsymbol{\gamma} \in \boldsymbol{\Gamma} \qquad (۳-۴)$$

بعد از گام اول، اتم φ_{γ_0} از سیگنال s کسر می‌شود تا باقیمانده‌ی $R^{(0)}$ حاصل شود. به طور کلی در مرحله‌ی $k = 1, 2, \dots$، الگوریتم MP، اتمی که بیشترین همبستگی را با باقیمانده دارد تعیین می‌کند، سپس ضریب اسکالری از آن اتم را با تخمین فعلی جمع می‌کند.

$$s^{(k)} = s^{(k-1)} + \alpha_k \varphi_{\gamma_k} \qquad (۴-۴)$$

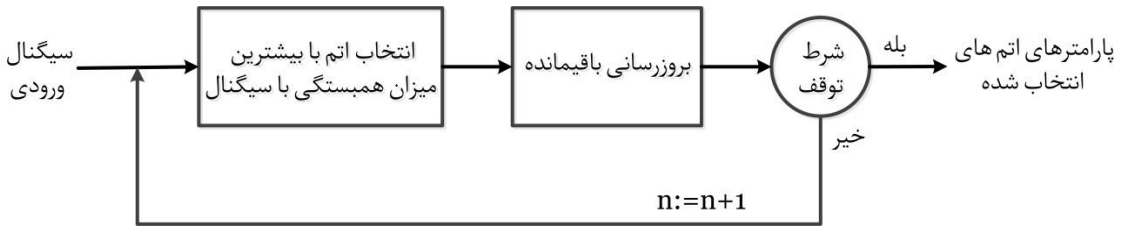
که در آن

$$\alpha_k = \langle R^{(k-1)}, \varphi_{\gamma_k} \rangle \qquad (۵-۴)$$

و

$$R^{(k)} = s - s^{(k)} \qquad (۶-۴)$$

بعد از m مرحله، مجموعه‌ای از اتم‌های نماینده‌ی سیگنال و باقیمانده‌ی $R = R^{(m)}$ داریم. شکل ۴-۱ خلاصه‌ی این مراحل را نشان می‌دهد. شبه کد الگوریتم مربوطه نیز در زیر آورده شده است.



شکل ۴-۱: نمودار بلوکی الگوریتم پیگیری انطباق.

Algorithm matching pursuit

Input: signal s, dictionary D

Return: List of coefficients α for $(\alpha^k, \phi_{\gamma_k})$

Initialize: $s^{(0)} \leftarrow s$

Repeat

Find ϕ_{γ_k} with maximum inner product $\langle s^{(k)}, \phi_{\gamma_k} \rangle$

$\alpha_k \leftarrow \langle s^{(k)}, \phi_{\gamma_k} \rangle$

$s^{(k+1)} = s^{(k)} - \alpha_k \phi_{\gamma_k}$

$k \leftarrow k + 1$

until either $||s^{(k)}|| < threshold$ or certain k is reached

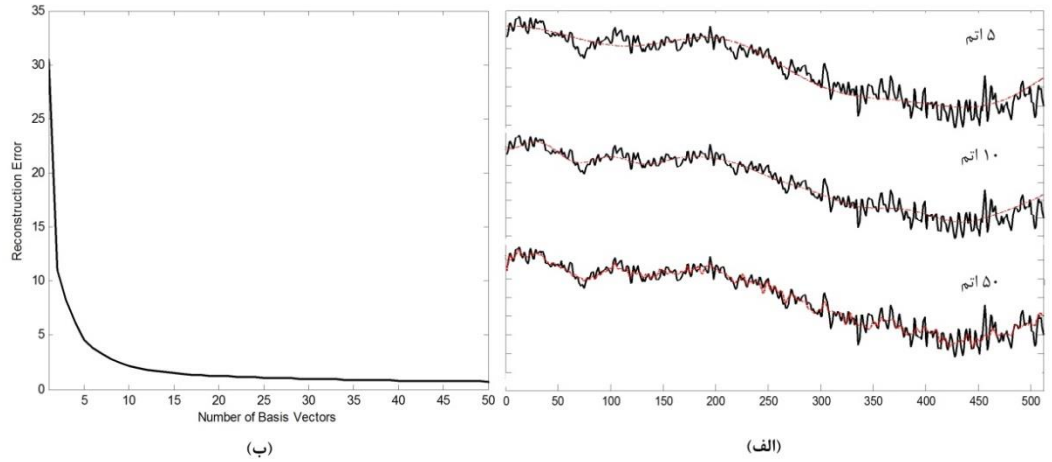
واژه‌نامه‌های بسیاری برای استفاده با MP پیشنهاد شده است از جمله: موجک‌ها، بسته‌های موجک، بسته‌های کسینوسی، واژه‌نامه‌های گابور، واژه‌نامه‌های گابور چندمقیاس، چیرپلت‌ها و غیره. اغلب واژه‌نامه‌ها کامل یا فراکامل هستند و روش‌های تخمین، مانند MP، اجازه‌ی ترکیب واژه‌نامه‌های مختلف را دارند. نمونه‌هایی از برخی واژه‌نامه‌های پایه به این‌صورت هستند: ۱) فرکانس(مانند توابع فوریه)، ۲) زمان-مقیاس، (مانند موجک‌های هار) و ۳) زمان-فرکانس،(مانند توابع گابور). برای گرفتن مشخصات غیرایستان سیگنال‌های صوت، ما از واژه‌نامه‌ی اتم‌های گابور استفاده می‌کنیم تا به یک نمایش زمان-فرکانس قابل جداسازی دست یابیم. در بخش ۴-۲-۲ به جزئیات بیشتری در این باره خواهیم پرداخت.

۴-۲-۱ استخراج ویژگی با استفاده از پیگیری انطباق (MP)

انواع مطلوب ویژگی‌ها باید مقاوم، پایدار ، ساده و واضح، هم‌چنین با نمایش تنک و قابل تفسیر به صورت فیزیکی باشد. نشان داده شده است که استفاده از MP این نمایش را امکان‌پذیر می‌کند [۸]. مزیت استفاده از این نمایش، توانایی گرفتن ساختار ذاتی هر نوع سیگنال و نگاشت از یک سیگنال بزرگ و پیچیده به یک فضای ویژگی کوچک و ساده می‌باشد. از همه مهم‌تر این‌که، به طور قابل توجهی

نسبت به نویز پس زمینه مقاوم است و می‌تواند جایی که MFCC ناتوان است مشخصات سیگنال را بگیرد.

هدف ما استفاده ازMP به عنوان یک ابزار برای استخراج ویژگی به منظور طبقه‌بندی، و نه لزوماً برای بازیابی یا تقریب سیگنال اصلی برای فشرده‌سازی است. با این وجود، MP یک راه عالی برای به انجام رساندن هر کدام از این وظایف نیز فراهم می‌کند. پیگیری انطباق روشی مطلوب برای فراهم کردن نمایش تنک و کاهش انرژی باقی‌مانده با چند اتم است. تجزیه از طریق MP ما را به توصیفی از ساختار سیگنال مجهز می‌کند. استراتژی استخراج ویژگی بر این فرض استوار است که مهم‌ترین اطلاعات یک سیگنال، نهفته در اتم‌های با بالاترین انرژی است و منجر به نمایشی ساده از ساختار مزبور می‌شود. از آنجا که MP اتم را به‌وسیله‌ی حذف بزرگ‌ترین انرژی باقی‌مانده انتخاب می‌کند، توجه خود را به انتخاب مفیدترین اتم‌ها، حتی پس از چند بار تکرار معطوف می‌نماید. اثربخشی استفاده ازMP با توابع گابور در شکل ۴-۲ نشان داده شده است. در شکل ۴-۲-الف می‌توان دید که تنها با استفاده از ۱۰ اتم اول انتخابی از واژه‌نامه می‌توان بازسازی معقولی از سیگنال داشت و استفاده از ۵۰ اتم اول، تقریبی بسیار شبیه به سیگنال اصلی ایجاد می‌کند. هم‌چنین در شکل ۴-۲-ب مشاهده می‌شود که بزرگ‌ترین افت در خطای باقی‌مانده، در چند جمله‌ی اول اتفاق می افتد.



شکل ۴-۲: نمونه‌هایی از بازسازی سیگنال با استفاده از MP با تعداد بردارهای پایه‌ی متفاوت از واژه‌نامه‌ی گابور.

الگوریتمMP اتم ها را به صورت گام به گام از میان مجموعه‌ای از شکل موج‌ها در یک واژه‌نامه که بیش‌ترین همبستگی را با ساختار سیگنال دارند انتخاب می‌کند. تکرار می‌تواند زمانی که ضریب مرتبط با انتخاب اتم به زیر یک آستانه رسید یا زمانی که تعداد معینی از اتم‌های کلی انتخاب شده باشد متوقف گردد. معیار توقف متداول دیگر استفاده از نسبت انرژی سیگنال به باقی‌مانده است.

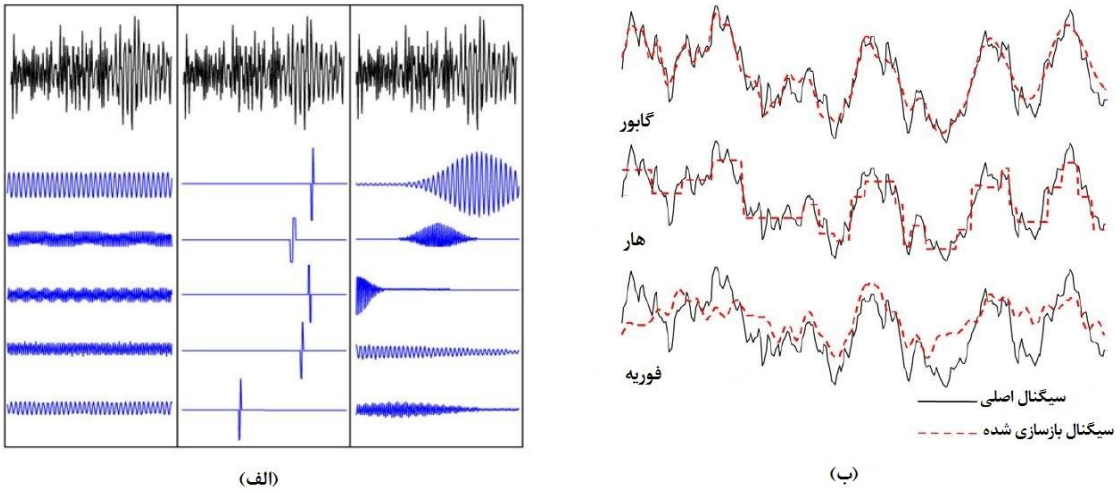
تجزیه‌ی سیگنال های مختلف از یک کلاس صدا، ممکن نیست دقیقاً از همان اتم‌ها یا همان ترتیب تشکیل شود. با این حال، از آن‌جا که ما میانگین پارامترها را به عنوان ویژگی می‌گیریم، نظم توالی اتم‌ها نادیده گرفته شده و قدرت این ویژگی‌ها با میانگین‌گیری افزایش می‌یابد. با استفاده از این پارامترهای اتم به عنوان ویژگی، از جزئیات ریزتر دور شده و بر ویژگی های مشخص متمرکز می‌شویم.

هنگام نگاشت یک فضای مسئله‌ی بزرگ به فضای ویژگی، تنها چند ویژگی مهم در نظر گرفته می‌شود که ما را قادر به نادیده گرفتن بقیه می‌کند. اطلاعات مهم در توصیف یک سیگنال می تواند در چند بردار پایه با بالاترین انرژی یافت شود و این روند که در آن MP این بردارها را انتخاب می‌کند دقیقاً در جهت حذف بزرگ‌ترین انرژی باقی‌مانده است. این بدان معنی است که حتی چند اتم اول پیدا شده توسط MP به طور طبیعی حاوی بیش‌ترین اطلاعات خواهند بود که آن‌ها را به ویژگی های قابل توجهی تبدیل می‌کند. این موضوع هم‌چنین ما را قادر می‌سازد که هر سیگنال را از یک فضای مسئله‌ی بزرگ‌تر به یک نقطه در فضای ویژگی کوچک‌تر نگاشت دهیم. داده‌ها، به همان میزان که در فضای ویژگی اولیه دارای نمایش مشابه یا تقریباً نزدیک هستند، در فضای جدید نیز مشابه هستند.

۴-۲-۲ انتخاب واژه نامه ی MP

نمونه هایی از تجزیه‌ی MP با استفاده از واژه‌نامه‌های مختلف در شکل۴-۳ مقایسه شده است. پنج اتم اول به‌دست‌آمده از تجزیه‌ی MP با واژه‌نامه‌های فوریه، هار و گابور در شکل ۴-۳-الف نشان داده شده اند. از آنجا که نمایش فوریه توسط برهم‌نهی سیگنال‌های غیرمحلی تشکیل می‌شود، نیازمند تعداد زیادی از اتم‌ها است تا منجر به یک شکل موج محلی شود. در مقابل، نمایش گابور توسط یک سیگنال

باند محدود زمان‌محدود تشکیل می‌شود، بنابراین برای سیگنال‌های زمان-فرکانس محلی مناسب تر است.



شکل ۴-۳: (الف) تجزیه‌ی سیگنال‌ها با استفاده از MP (۵ اتم اول) با واژه نامه های فوریه (چپ)، هار (وسط) و گابور (راست)؛ (ب) بازسازی سیگنال با استفاده از ۱۰ اتم اول از الگوریتم MP با واژه نامه های گابور (بالا)، هار (وسط) و فوریه (پایین) [۸].

اثربخشی بازسازی یک سیگنال با استفاده از تنها تعداد کمی از اتم‌ها در شکل ۴-۳-ب مقایسه شده است که در آن ۱۰ اتم استفاده می شود. اتم‌های گابور، در مقایسه با تبدیل‌های هار یا فوریه با استفاده از همان تعداد ضرایب، منجر به کم‌ترین خطای بازسازی می‌شود [۸]. با توجه به ماهیت غیرهمگن وقایع صوتی، استفاده از ویژگی‌های با خصوصیات گابور، برای سیستم طبقه‌بندی مفید خواهد بود. بر اساس مشاهدات فوق، ما در این تحقیق از تابع گابور استفاده می‌کنیم.

توابع گابور، توابع گوسی مدوله‌شده‌ی سینوسی هستند که در آن‌ها هر اتم به صورت زیر می باشد:

$$g_{s,u,\omega,\theta} = \frac{K_{s,u,\omega,\theta}}{\sqrt{s}} e^{-\pi(n-u)^2/s^2} \cos[2\pi\omega(n-u)+\theta] \qquad (۴-۷)$$

که در آن $s \in \mathbb{R}^+$، $u, \omega \in \mathbb{R}$ و $\theta \in [0,2\pi]$ می‌باشند. $K_{s,u,\omega,\theta}$ ضریب نرمال‌سازی می‌باشد، به گونه‌ای که $\|g_{s,u,\omega,\theta}\|^2 = 1$ برقرار باشد. ما از $\gamma = (s,u,\omega,\theta)$ برای نشان دادن پارامترهای تابع گابور استفاده می‌کنیم که s، u، ω و θ به‌ترتیب نشان‌دهنده‌ی مقیاس، زمان، فرکانس و فاز اتم

هستند. واژه‌نامه‌ی گابور در [۵۸] با پارامترهای اتم انتخاب‌شده از توالی دوتایی از اعداد صحیح اجرا شده است. مقیاس s که مربوط به عرض زمانی اتم است از توالی دوتایی $s = 2^p$، $(1 \leq p \leq m)$ حاصل شده است و اندازه‌ی اتم $N = 2^m$ است.

۴-۳- استخراج ویژگی با استفاده از ضرایب کپسترال فرکانس بارک

ضرایب کپسترال فرکانس بارک روش دیگری برای استخراج ویژگی از فایل صوتی است. در شکل ۴-۴ مراحل مختلف این روش نشان داده شده است. همانطور که مشاهده می‌شود مراحل به شرح زیر است:

۱) پیش‌پردازش: در این مرحله مولفه‌ی ثابت سیگنال حذف می‌شود.

۲)قاب‌بندی: مانند اکثر الگوریتم‌های استخراج ویژگی از سیگنال صوت، عمل قاب‌بندی برای رسیدن به یک سیگنال ایستان صورت می‌گیرد.

۳) پنجره‌سازی: در این مرحله برای سیگنال گفتار معمولاً از پنجره‌ی همینگ استفاده می‌شود.

۴) تبدیل فوریه سریع (FFT)

۵) بانک فیلتر بارک-مقیاس: BFCC مشابه MFCC است با این تفاوت که برای طیف توان در امتداد محور فرکانس، فرکانس بارک که در معادله زیر آورده شده استفاده می‌شود:

$$Bark = 26.81/\big(1 + (1960/f)\big) - 0.53 \qquad (۸-۴)$$

مقیاس بارک، طیف شنیداری را به ۱۲ باند بحرانی تقسیم می‌کند تا پاسخ فرکانسی گوش انسان را شبیه‌سازی کند.

۶) قانون توان یکسانی بلندی صدا: احساس یکسان نبودن بلندی صدا در فرکانس‌های مختلف را جبران می‌کند.

$$E(f) = \left(\frac{f^2}{f^2 + 1.6e5}\right)^2 . \frac{f^2 + 1.44e6}{f^2 + 9.61e6} \qquad (۹-۴)$$

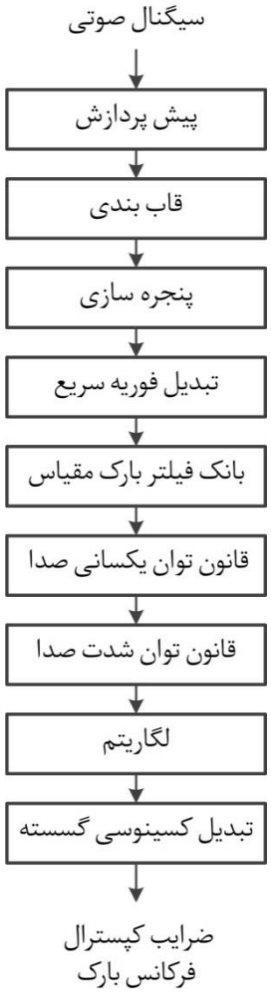
۷) قانون توان شدت بلندی صدا: رابطه‌ای غیرخطی بین شدت صدا و بلندی صدای دریافت شده شبیه‌سازی می‌کند.

$$\varphi(f) = \psi(f)^{0.33} \qquad (۱۰-۴)$$

۸) لگاریتم

۹) تبدیل کسینوسی گسسته

این الگوریتم پردازش PLP طیف و تبدیل کسینوسی را برای به‌دست آوردن ضرایب کپسترال ترکیب می‌کند. به جای استفاده از بانک فیلتر مل، فیلترهای بارک و پیش‌تأکید برابری بلندی صدا همراه با قانون توان شدت بلندی صدا به ویژگی‌های مشابه MFCC اعمال شده است.



شکل ۴-۴: استخراج ویژگی با استفاده از ضرایب کپسترال فرکانس بارک (BFCC) [۵۷].

فصل ۵ شبیه سازی ها و نتایج

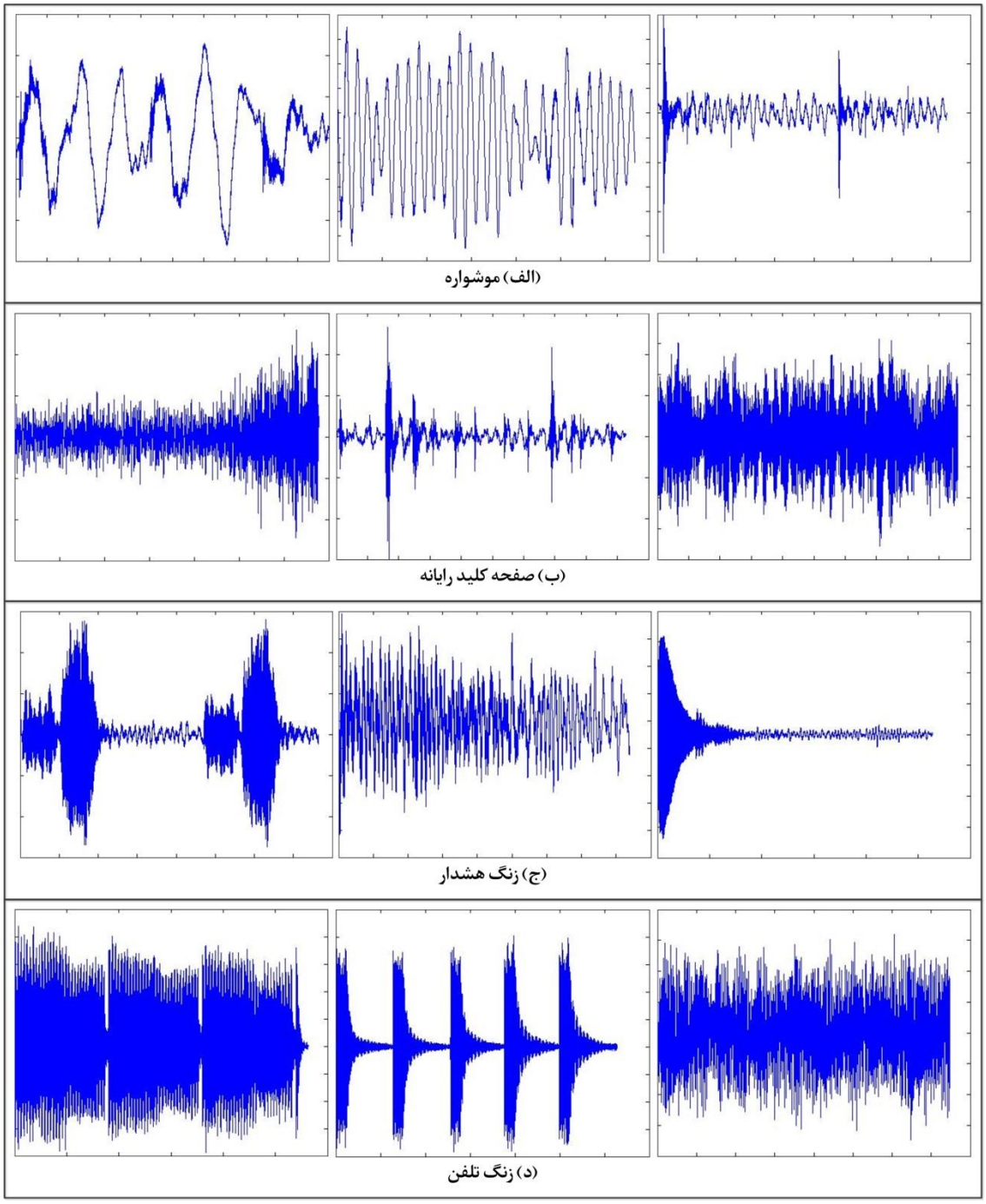
۵-۱- تنظیمات آزمایش

در این بخش، مشخصات مربوط به هر یک از مرحله های انجام آزمایش بیان می شود. ابتدا پایگاه داده معرفی شده و پس از آن، روش های مورد استفاده برای تشخیص فعالیت، استخراج ویژگی و طبقه‌بندی بیان خواهد شد.

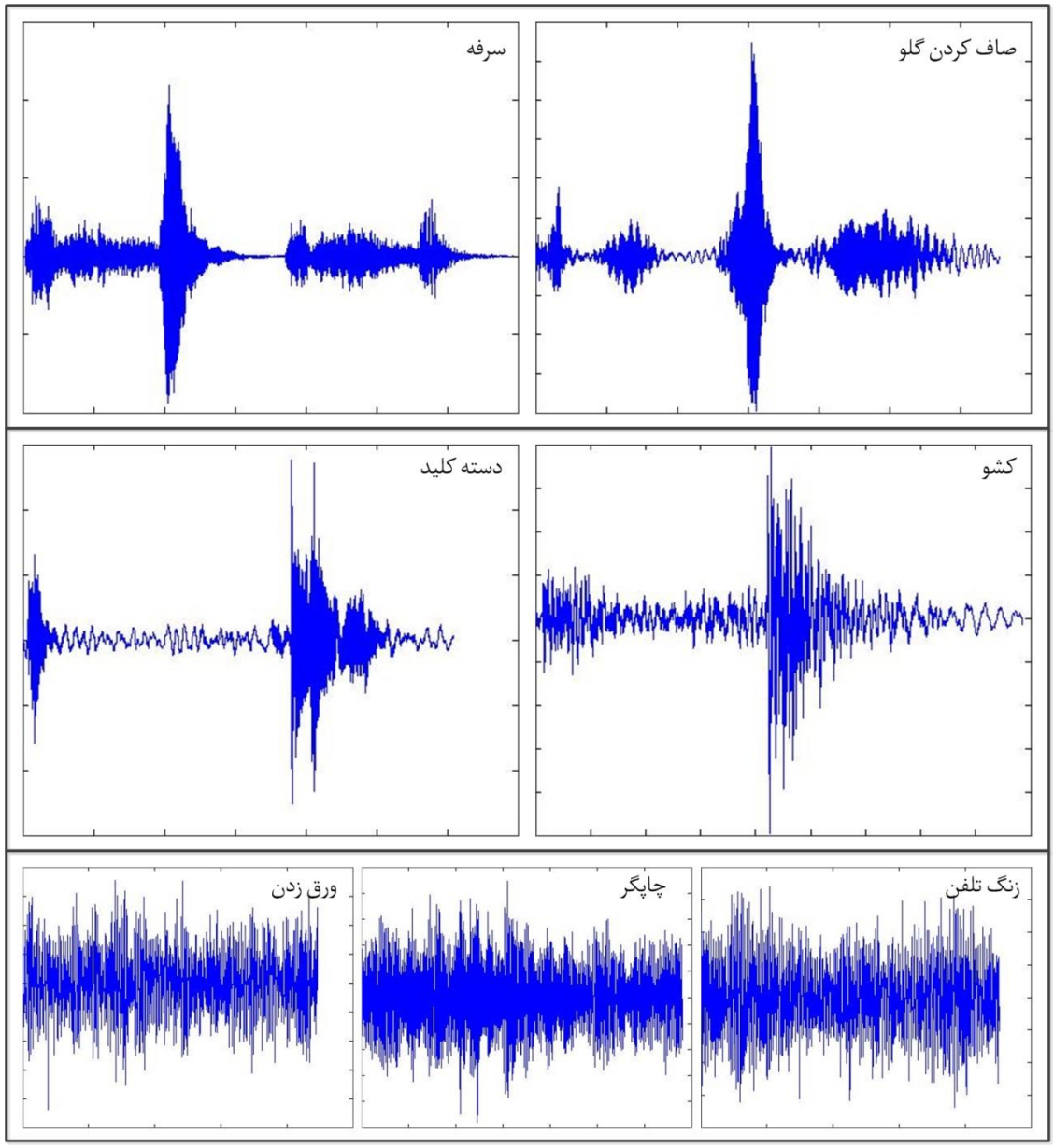
۵-۱-۱ پایگاه داده

در این پایان‌نامه از پایگاه داده‌ی ارائه شده در [۵۹] تحت عنوان D-CASE استفاده شده است. این پایگاه داده یک مجموعه آموزش دارد که شامل ۱۶ کلاس داده از وقایع صوتی مربوط به محیط اداری به نام‌های زنگ هشدار، صاف کردن گلو، سرفه، بستن در، کشو، صفحه کلید رایانه، دسته کلید، در زدن، خنده، موشواره، ورق زدن، افتادن خودکار (قلم)، زنگ تلفن، چاپگر، گفتگو و کلید می‌باشد و هر کلاس دارای ۲۰ فایل صوتی است. مجموعه‌ی آموزش برای یافتن بهترین ویژگی‌های معرفی کننده‌ی سیگنال‌ها و بهترین طبقه‌بند برای جداسازی این وقایع استفاده می‌شود. این پایگاه داده یک مجموعه‌ی توسعه نیز در بردارد که شامل سه فایل صوتی پیوسته از انواع وقایع صوتی است. از این مجموعه نیز در مرحله‌ی آشکارسازی وقایع صوتی استفاده می‌شود. برخی از فایل‌های مجموعه‌ی آموزش و نیز تمامی فایل‌های مجموعه‌ی توسعه، آلوده به نویز با نرخ سیگنال به نویز متفاوت است. از این‌رو استفاده از روش‌های مقاوم در برابر نویز برای استخراج ویژگی لازم است. تمامی داده‌های مربوط به هر دو مجموعه، فایل‌های صوتی دوکاناله هستند که از میانگین دو کانال استفاده شده است. فایل‌ها به صورت غیرفشرده در قالب «wav. » و با نرخ نمونه‌برداری ۴۴/۱ کیلوهرتز هستند. طول‌های زمانی مجموعه‌ی آموزش متغیر بین ۱ ثانیه تا ۲۹ ثانیه و طول زمانی فایل‌های مجموعه‌ی توسعه بین ۹۴ تا ۱۱۸ ثانیه می‌باشند. شناسایی صداهای این پایگاه داده به دلیل داشتن صداهایی با شباهت‌های بین کلاسی و تفاوت‌های بسیارِ درون کلاسی، یک مسئله‌ی دشوار است. در شکل‌های ۵-۱ و ۵-۲ نمونه‌هایی از شکل زمانی سیگنال‌های مختلف موجود در پایگاه داده ترسیم شده است. با مشاهده‌ی

این شکل‌ها می‌توان به تفاوت‌های درون‌کلاسی و شباهت‌های بین کلاسی در پایگاه داده‌ی مذکور پی برد.



شکل ۵-۱: نمونه هایی از تفاوت های درون کلاسی در شکل زمانی قطعه های صوتی مربوط به چند کلاس.



شکل ۵-۲: نمونه هایی از شباهت های بین کلاسی در شکل زمانی قطعه های صوتی مربوط به چند کلاس.

۵-۱-۲ آشکارسازی وقایع صوتی

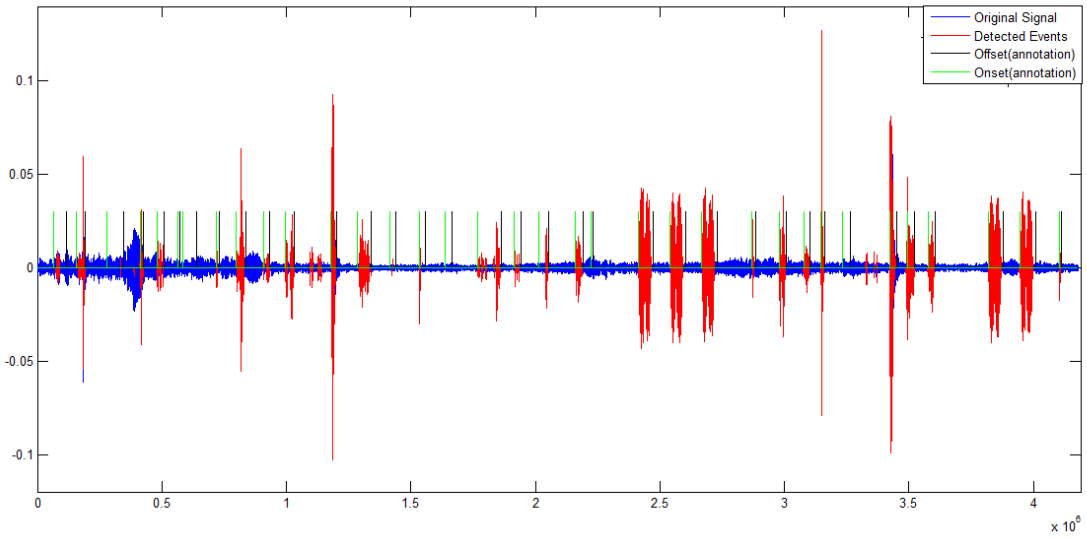
با توجه به این‌که فایل‌های صوتی پایگاه داده‌ی مورد استفاده، آلوده به نویز با نرخ سیگنال به نویز متفاوت است و نویز مربوطه می‌تواند دارای مولفه‌های فرکانسی بالا یا پایین باشد، در مرحله‌ی تشخیص فعالیت، ابتدا از یک فیلتر باترورث میان‌گذر مرتبه دو با باند گذر $\omega = [0.013\ 0.5]$ برای کاهش نویز استفاده شده است. بازه‌ی انتخاب شده برای ω، از طریق آزمایش‌های متوالی و سعی و خطا بدست

آمده است. در ادامه، با تعریف یک مقدار آستانه برمبنای بیشینه‌ی سیگنالِ حذف نویز شده(Sn) (طبق رابطه‌ی زیر)، وقایع صوتی آشکارسازی شده است.

$$TH = 0.03 \times Max(Sn) \quad (۱-۵)$$

شکل ۵-۳ نتیجه‌ی اجرای این مراحل آشکارسازی روی یک نمونه از فایل‌های صوتی است. نتایج، در جدول ۵-۸ آورده شده است.

-



شکل ۵-۳: یک نمونه از نتایج مرحله ی آشکارسازی وقایع صوتی موجود در یک فایل صوتی پیوسته.

۵-۱-۳ استخراج ویژگی

در این مرحله، تعدادی از ویژگی‌های صوتی را مورد بررسی قرار داده و یک ارزیابی تجربی روی آن‌ها انجام می‌دهیم. ما از ۱۳ ضریب MFCC، ۴ ضریب MP، ۱۲ ضریب BFCC، ۱۲ ضریب LPC، ۱۲ ضریب PLP، ۱۲ ضریب RPLP ، ۲۳ ضریب BER، انرژی کوتاه مدت و نرخ عبور از صفر استفاده می‌کنیم.

برای بدست آوردن ویژگی‌های MP، بر اساس آزمایشات تجربی و نتایج پژوهش‌های پیشین،از واژه‌نامه‌ی گابور با پارامترهای زیر استفاده شده است: $(1 \le p \le 9)\ s = 2^p$، $(0 \le Z \le 15)\ u = 32Z$ ، $(1 \le i \le 35, K = 0.5 \times 35^{-2.6}\ \omega = Ki^{2.6}$ که در این شرایط، ω نرمالیزه شده بین ۰ و

۰/۵می‌باشد. از آنجا که در کارهای قبلی قید شده است که پارامتر θ تاثیر چندانی ندارد، در این پایان‌نامه نیز این پارامتر برابر صفر در نظر گرفته شده است تا اندازه‌ی واژه‌نامه کوچک نگه داشته شود. با تغییر فاز، به عنوان مثال $\theta = \{0, \pi/4, \pi/2, \dots\}$، هر بردار پایه، تنها کمی تغییر می‌کند. از آنجا که ما از چند اتم بالا برای ایجاد ویژگی های MP استفاده می‌کنیم لازم نیست که بردارهای پایه‌ی شیفت یافته با فاز را وارد مسئله کنیم.

طول اتم برابر با ۵۱۲ گرفته شده و در نتیجه واژه‌نامه حاصل شده به تعداد $5040 = 9 \times 35 \times 16$ اتم خواهد داشت. با بیش‌تر شدن تعداد اتم‌های واژه‌نامه بار محاسباتی بیش‌تری به سیستم تحمیل می‌شود در عوض دقت تخمین سیگنال با تعداد اتم‌های ثابت بالاتر می‌رود.

مقیاس لگاریتمی فرکانس برای فراهم کردن وضوح بالاتر در منطقه فرکانس پایین تر و رزولوشن پایین تر در منطقه فرکانس بالاتر استفاده می شود. توان ۲/۶ برای ω از [۱۱] گرفته شده است. هدف، داشتن دانه‌بندی ریزتر برای زیر ۱۰۰۰ هرتز و هم‌چنین قدرت توصیفی کافی در فرکانس بالاتر است. دلیل دانه‌بندی ریزتر در فرکانس‌های پایین تر این است که بیش‌تر انواع صداها در این محدوده رخ می‌دهند و ما می‌خواهیم تفاوت‌های ریزتر بین آن‌ها را بیابیم.

با توجه به این‌که در [۱۰] بیان شده است که ۵ اتم بهترین اطلاعات را در اختیار قرار می‌دهد، در این پایان‌نامه نیز، ما ۵ اتم رابه عنوان معیار توقف برای تکرار انتخاب می‌کنیم. ویژگی هایMP توسط فرآیند زیر انتخاب شده است.

از یک پنجره مستطیل شکل از ۵۱۲ نمونه با هم‌پوشانی ۵۰ درصد استفاده شده است. این مربوط به اندازه‌ی پنجره‌ی مورد استفاده برای تمام استخراج ویژگی‌ها می‌باشد. ما تجزیه‌ی هر بخش ۵۱۲ نمونه‌ای را با استفاده از MP با یک واژه‌نامه از اتم‌های گابور که آن‌ها نیز به طول ۵۱۲ نمونه هستند انجام می‌دهیم. روند MP بعد از بدست آوردن ۵ اتم متوقف می‌شود. پس از آن، پارامترهای فرکانس و

مقیاس برای هر یک از این ۵ اتم، ثبت شده ومیانگین و انحراف معیار مربوط به هر پارامتربه طور جداگانه پیدا می‌شود، تا در نهایت منجر به ۴ مقدار ویژگی شود.

۵-۱-۴ طبقه بندی

در فضای ویژگی از روش طبقه‌بندی مدل مخلوط گوسی (GMM) استفاده شده است. با GMMها، هر کلاس از داده‌ها به شکل مخلوطی از چند خوشه‌ی گوسی (در اینجا ۲ یا ۳ خوشه) مدل می‌شود. هر مؤلفه‌ی مخلوط، یک گوسی نمایش داده شده توسط میانگین و ماتریس کواریانسِ داده است. هنگامی که مدل ایجاد شد، احتمال شرطی با استفاده از رابطه‌ی زیر محاسبه می‌شود:

$$p(x|X_k) = \sum_{j=1}^{m_k} p(x|j)P(j) \qquad (۲-۵)$$

که X_k نقاط داده برای هر کلاس می‌باشد، m_k تعداد مؤلفه‌ها، $P(j)$ احتمال پیشین این است که داده‌ی x بوسیله‌ی مولفه‌ی j ایجاد شده است، و $p(x|j)$ چگالی مولفه‌ی مخلوط است. پس از آن الگوریتم بیشینه‌سازی امیدریاضی برای پیدا کردن حداکثر احتمال پارامترهای هر کلاس مورد استفاده قرار می‌گیرد. همچنین روش طبقه بندی k-نزدیکترین همسایه نیز بررسی شده است. k-NN، یک الگوریتم یادگیری با نظارت ساده است که در آن یک نمونه‌ی جدید بر اساس کلاس اکثریت k نزدیک‌ترین همسایگان خود طبقه‌بندی می‌شود. معیار فاصله متداول، فاصله اقلیدسی است:

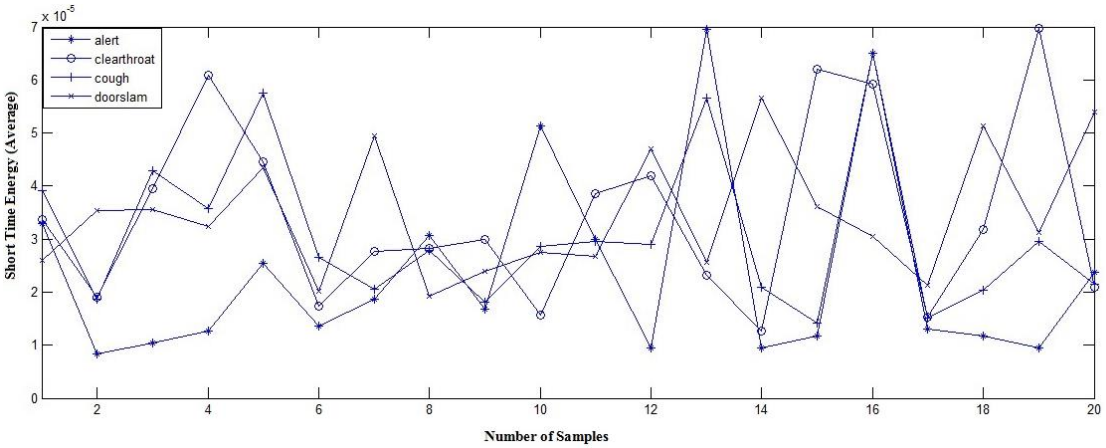
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (۳-۵)$$

در آزمایش‌های ما، مشاهده شد که به ازای k=1 بیش‌ترین نرخ شناسایی حاصل می‌شود از این‌رو از 1-NN استفاده شده است. فایل‌های جداگانه برای مجموعه آموزش و آزمون استفاده شده و برای نمایش نتایج، هر آزمایش ۱۰۰ بار تکرار شده است.

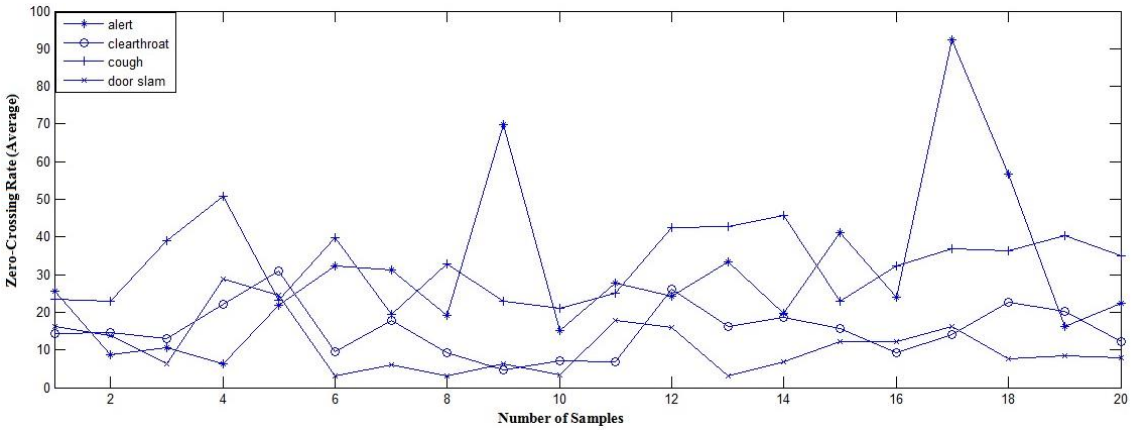
۵-۲- نتایج شبیه سازی ها

۵-۲-۱ ویژگی های متداول حوزه ی زمان

ابتدا متداول‌ترین ویژگی‌های حوزه‌ی زمان یعنی انرژی کوتاه مدت و نرخ عبور از صفر کوتاه مدت، مورد بررسی قرار می‌گیرد. همان‌طور که در شکل‌های ۵-۴ و ۵-۵ مشاهده می‌شود به دلیل شبه‌نویزی بودنِ صداهای پایگاه داده‌ی مورد استفاده و وجود شباهت‌های بین‌کلاسی داده‌های این مجموعه، ویژگی‌های حوزه‌ی زمان توانایی چندانی در جداسازی کلاس‌های پایگاه داده‌ی مربوطه ندارند. از این‌رو می‌توان گفت که این ویژگی‌های زمانی برای شناسایی وقایع صوتی محیط اداری مناسب نیستند.



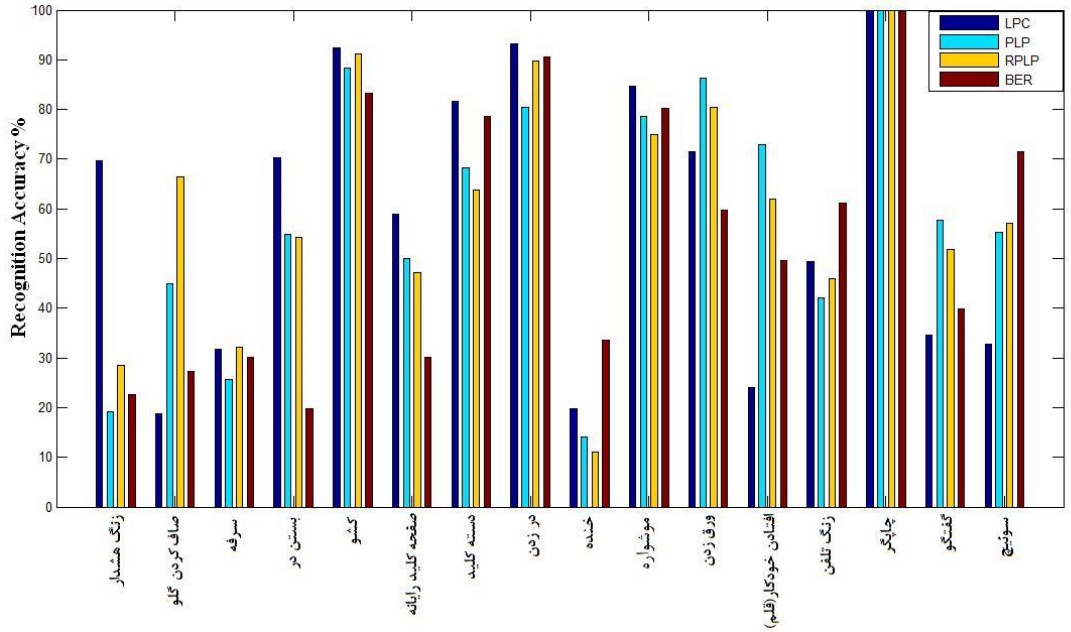
شکل ۵-۴: میانگین انرژی کوتاه مدت نمونه های مختلف برای ۴ کلاس "زنگ هشدار"، "صاف کردن گلو"، "سرفه" و "بستن در".



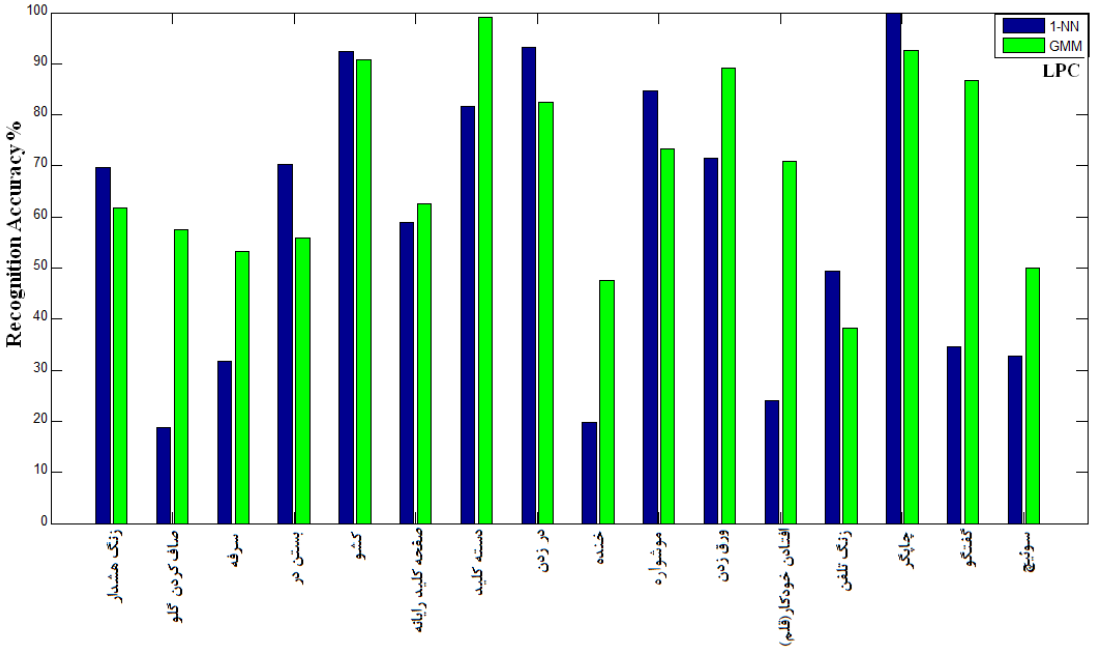
شکل ۵-۵: میانگین نرخ عبور از صفر کوتاه مدت نمونه های مختلف برای ۴ کلاس "زنگ هشدار"، "صاف کردن گلو"، "سرفه" و "بستن در".

۵-۲-۲ ویژگی های متداول فرکانسی

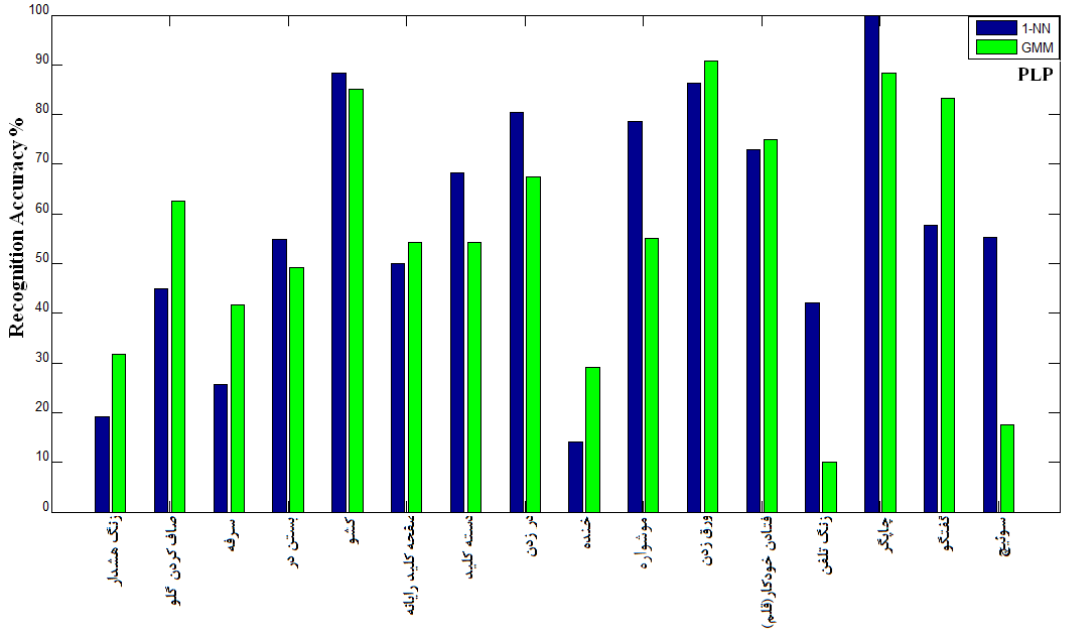
در این بخش نتایج حاصل از شبیه‌سازی‌ها با استفاده از ویژگی‌های LPC، PLP،RPLP و BER برای ۱۶ کلاس از وقایع صوتی مربوط به محیط اداری با استفاده از طبقه‌بند نزدیکترین همسایه در شکل ۵-۶ آورده شده است. همچنین نرخ شناسایی سیستم به ازای هر کلاس با استفاده از طبقه‌بندهای نزدیکترین همسایه و مدل مخلوط گوسی برای ویژگی‌های LPC، PLP، RPLP، BER و MFCC به ترتیب در شکل‌های ۵-۷ تا ۵-۱۱ ارائه شده است. نرخ شناسایی کلی سیستم برای مقایسه کارآیی این ویژگی‌ها در جدول ۵-۱ آمده است. نرخ شناسایی کلی سیستم در استفاده‌ی تنها یا دوبه‌دوی این مجموعه ویژگی‌ها در جدول ۵-۲ گزارش شده است. در جدول ۵-۲ مشاهده می‌شود که نرخ شناسایی کلی سیستم با استفاده از طبقه‌بند نزدیکترین همسایگی برای ترکیب ویژگی‌های BER و RPLP بیشترین نرخ، برابر ۶۵ درصد را به دست داده است. در جدول ۵-۱ می‌توان دید که استفاده از طبقه‌بند GMM همراه با ویژگی‌های MFCC یا LPC منجر به بهبود نسبتاً خوب در نرخ شناسایی کلی سیستم می‌شود اما ما به دنبال بهبود این نتیجه هستیم.



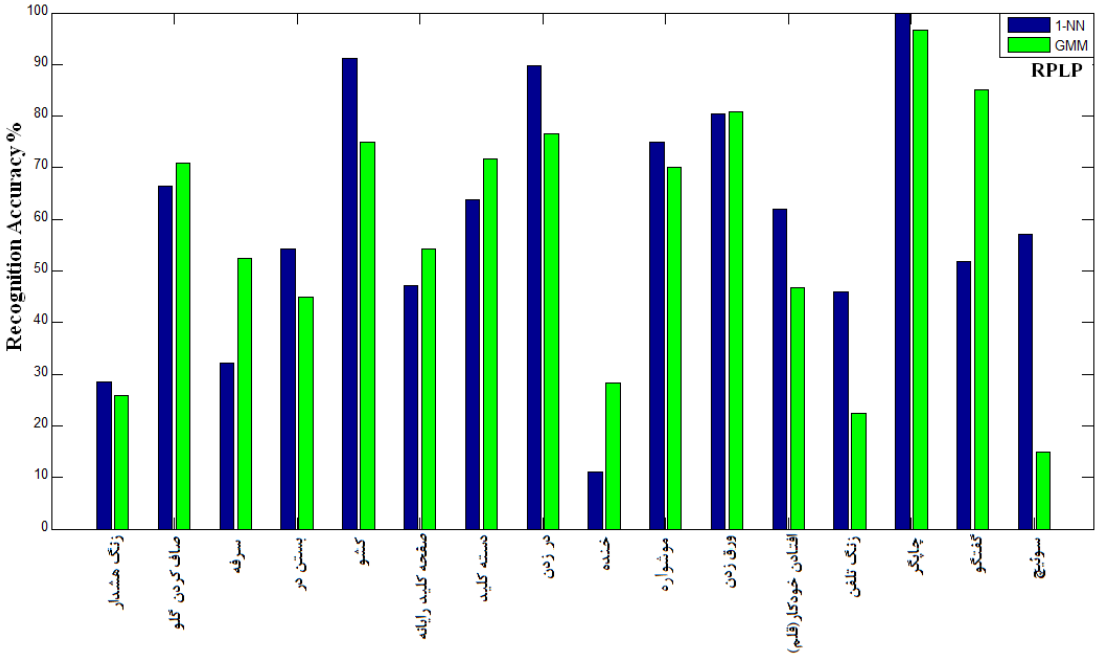
شکل ۵-۶: مقایسه ی نرخ شناسایی سیستم(1-NN) با استفاده از LPC، PLP، RPLP و BER.



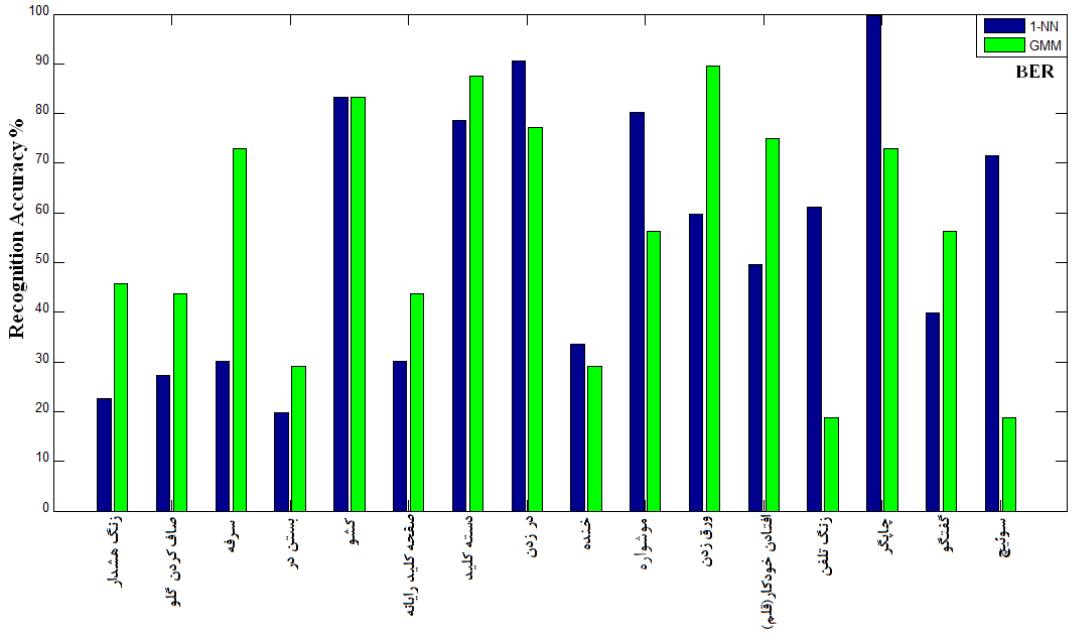
شکل ۵-۷: مقایسه ی نرخ شناسایی سیستم برای کلاس های مختلف با طبقه بندهای 1-NN و GMM و ویژگی LPC.



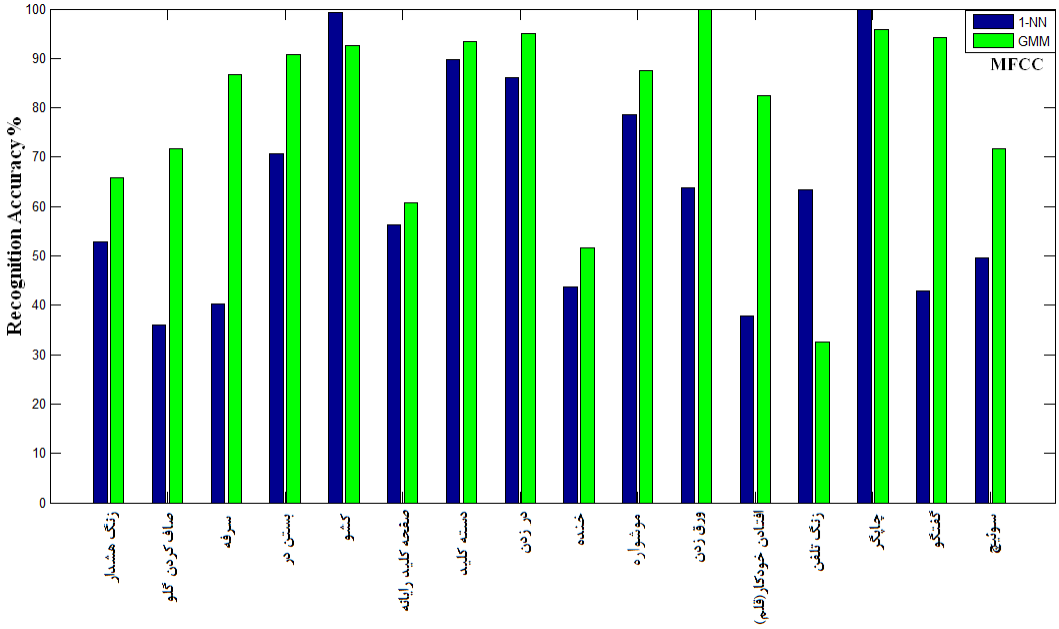
شکل ۵-۸: مقایسه ی نرخ شناسایی سیستم برای کلاس های مختلف با طبقه بندهای 1-NN و GMM و ویژگی PLP.



شکل ۵-۹: مقایسه ی نرخ شناسایی سیستم برای کلاس های مختلف با طبقه بندهای 1-NN و GMM و ویژگی .RPLP



شکل ۵-۱۰:مقایسه ی نرخ شناسایی سیستم برای کلاس های مختلف با طبقه بندهای 1-NN و GMM و ویژگی .BER



شکل ۵-۱۱: مقایسه ی نرخ شناسایی سیستم برای کلاس های مختلف با طبقه بندهای 1-NN و GMM و ویژگی MFCC.

جدول ۵-۱- نتایج نرخ شناسایی کلی سیستم با استفاده از 1-NN و GMM برای ویژگی های LPC، PLP، RPLP، BER و MFCC (%).

	MFCC	LPC	PLP	RPLP	BER
1-NN	۶۳/۴۹	۵۸/۲۰	۵۷/۳۳	۶۰/۱۲	۵۴/۶۶
GMM	**۷۴/۷۸**	**۶۹/۴۸**	۵۵/۹۴	۵۷/۲۹	۵۷/۴۲

جدول ۵-۲- نتایج نرخ شناسایی کلی سیستم با استفاده از 1-NN برای ترکیب دو به دوی مجموعه ویژگی ها (%).

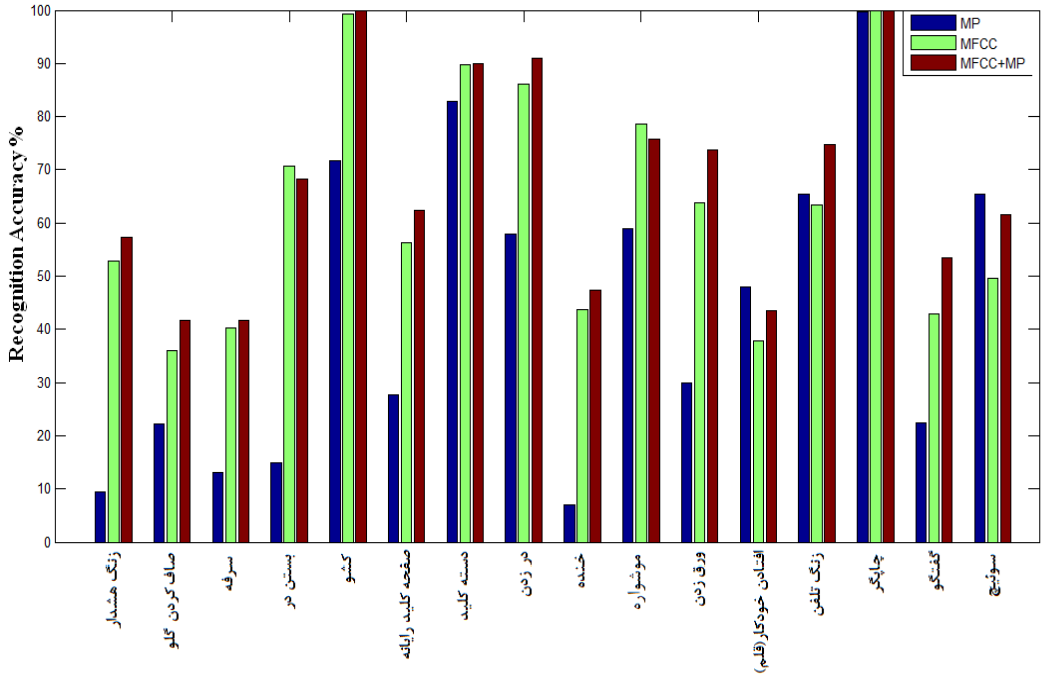
	MFCC	LPC	PLP	RPLP	BER
MFCC	------	------	------	------	------
LPC	۶۴/۸۰	------	------	------	------
PLP	۶۴/۵۸	۶۳/۰۸	------	------	------
RPLP	۶۲/۶۳	۶۱/۲۸	------	------	------
BER	۶۱/۴۲	۶۲/۱۱	۶۳/۴۶	**۶۵/۰۶**	------

۵-۲-۳ ویژگی های پیشنهادی

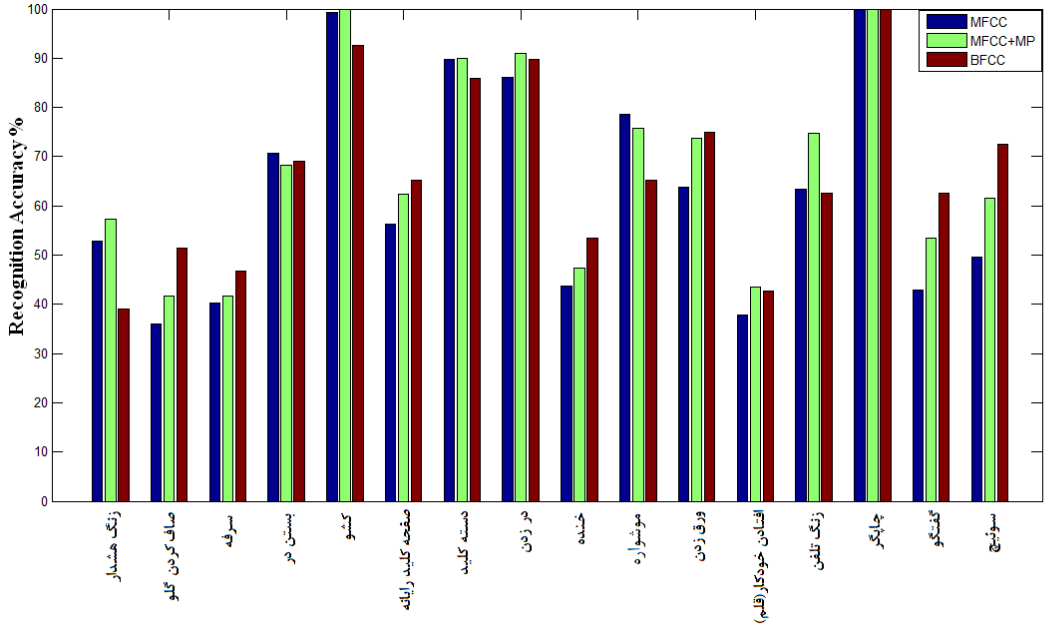
نتایج حاصل از شبیه‌سازی‌ها برای طبقه‌بندی ۱۶ کلاس از وقایع صوتی اداری، با استفاده از طبقه‌بند نزدیکترین همسایه و ویژگی‌های MFCC، MP و ترکیب ویژگی‌های MFCC و MP در شکل ۵-۱۲ آورده شده است. همانطور که مشاهده می‌شود ویژگی‌های MP تنها برای ۳ کلاس از ۱۶ کلاس نسبت به MFCC نتایج بهتری به دست داده است اما برای داده‌های ۱۲ کلاس، ترکیب این دو مجموعه ویژگی منجر به بهبود نسبی نرخ شناسایی سیستم شده است. در شکل ۵-۱۳ نیز نتایج مربوط به استفاده از BFCC با MFCC و ترکیب MFCC و MP مقایسه شده است. می‌توان مشاهده کرد که برای اغلب کلاس‌ها BFCC نسبت به MFCC نتیجه‌ی بهتری به دست داده است. در شکل ۵-۱۴ نتایج کلی سیستم شناسایی برای استفاده‌ی تنها از هرکدام از ویژگی‌ها و ترکیب هریک با ویژگی‌های MP با استفاده از طبقه‌بند نزدیکترین همسایه آورده شده است. مشاهده می‌شود که ترکیب هر مجموعه ویژگی با MP منجر به بهبود نرخ شناسایی کلی سیستم شده است. جدول ۵-۳ نیز اعداد مربوط به این شکل را نشان می‌دهد. می‌توان دید که ترکیب ویژگی‌های MFCC و MP تا اینجا بالاترین نرخ شناسایی را بدست داده است. در شکل‌های ۵-۱۵ و ۵-۱۶ نیز به ترتیب برای ویژگی‌های MFCC+MP و BFCC، برای کلاس‌هاس مختلف، نتایج استفاده از طبقه‌بندهای نزدیکترین همسایگی و مدل مخلوط گوسی مقایسه شده است.

جدول ۵-۳- مقایسه ی نرخ شناسایی سیستم (1-NN) به ازای استفاده از هریک از ویژگی ها به صورت مجزا و ترکیب با MP (%).

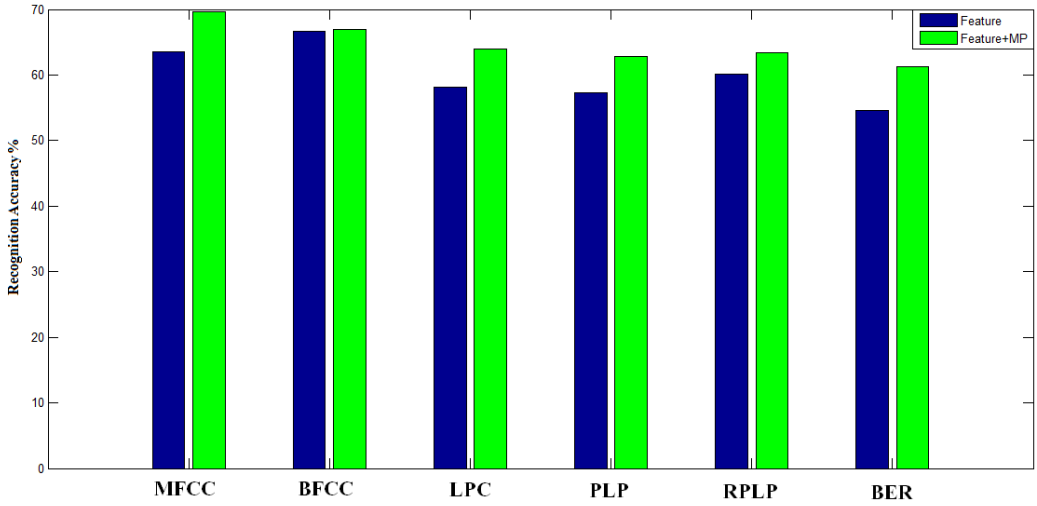
	MFCC	BFCC	LPC	PLP	RPLP	BER
Feature	۶۳/۴۹	۶۶/۷۰	۵۸/۲۰	۵۷/۳۳	۶۰/۱۲	۵۴/۶۶
Feature+MP	**۶۹/۶۷**	۶۶/۹۶	۶۳/۹۴	۶۲/۸۲	۶۳/۴۴	۶۱/۲۸



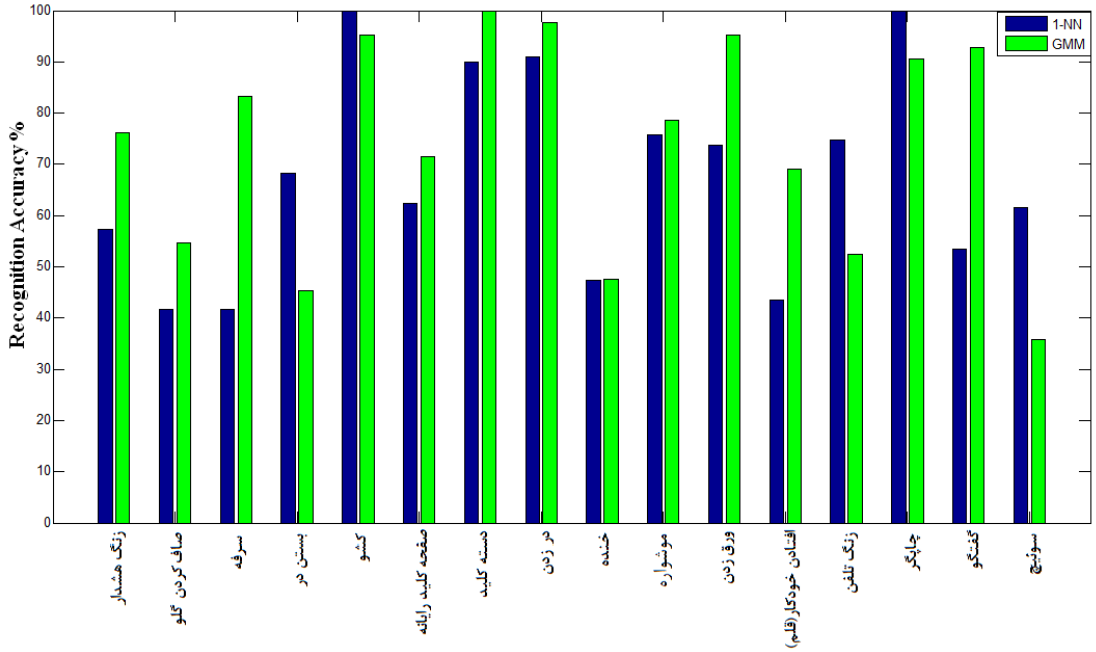
شکل ۵-۱۲: مقایسه ی نرخ شناسایی سیستم (1-NN) با استفاده از MFCC تنها، MP تنها، و ترکیب MFCC و MP.



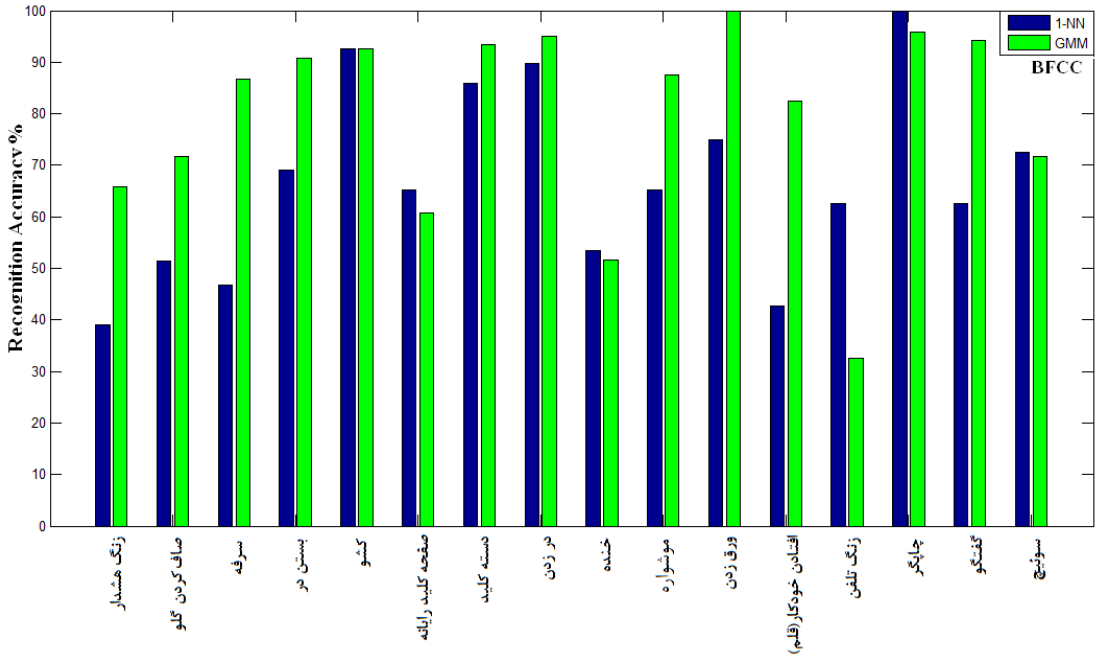
شکل ۵-۱۳: مقایسه ی نرخ شناسایی سیستم (1-NN) با استفاده از MFCC تنها، ترکیب MFCC+MP و BFCC تنها.



شکل ۵-۱۴: مقایسه ی نرخ شناسایی سیستم (1-NN) به ازای استفاده از هریک از ویژگی ها به صورت مجزا و ترکیب با MP.



شکل ۵-۱۵: مقایسه ی نرخ شناسایی سیستم برای کلاس های مختلف با طبقه‌بندهای 1-NN و GMM و ترکیب ویژگی‌های MFCC و MP.



شکل ۵-۱۶: مقایسه ی نرخ شناسایی سیستم برای کلاس های مختلف با طبقه بندهای 1-NN و GMM و ویژگی BFCC.

در جدول ۵-۴ نتایج کلی تمامی شبیه‌سازی‌ها برای دو طبقه‌بند نزدیکترین همسایه و مدل مخلوط گوسی با استفاده از انواع ویژگی‌ها آورده شده است. مشاهده می‌شود که ترکیب ویژگی‌های MFCC و MP در استفاده از طبقه‌بند نزدیکترین همسایگی منجر به بهبود نرخ شناسایی کلی سیستم شده اما در استفاده از طبقه‌بند مدل مخلوط گوسی نرخ شناسایی نه تنها افزایش نداشته بلکه به مقدار کمی کاهش نشان داده است. اما استفاده از ویژگی‌های BFCC برای طبقه‌بند مدل مخلوط گوسی منجر به نرخ شناسایی کلی برابر ۸۰/۰۵ درصد شده که بالاترین میزان است.

جدول ۵-۴- مقایسه ی نرخ شناسایی سیستم (1-NN) به ازای استفاده از هریک از ویژگی‌ها به صورت مجزا و ترکیب با MP (%).

	MFCC	LPC	PLP	RPLP	BER	MFCC+MP	BFCC	BFCC+MP
1-NN	۶۳/۴۹	۵۸/۲۰	۵۷/۳۳	۶۰/۱۲	۵۴/۶۶	۶۹/۶۷	۶۶/۷۰	۶۶/۹۶
GMM	۷۴/۹۶	۶۹/۴۸	۵۵/۹۴	۵۷/۲۹	۵۷/۴۲	**۷۴/۱۱**	**۸۰/۰۵**	۷۴/۵۵

۵-۳- ماتریس پراکندگی

در جدول‌های ۵-۵، ۵-۶ و ۵-۷ ماتریس پراکندگی نرخ شناسایی سیستم به ازای کلاس‌های مختلف برای طبقه‌بند GMM به ترتیب با استفاده از ویژگی‌های MP + MFCC ، BFCC و MP + BFCC آورده شده است. با دقت در این ماتریس‌ها می‌توان پی برد که کدام کلاس‌ها بیش‌ترین شباهت را داشته‌اند و منجر به افت نرخ شناسایی کلی سیستم شده‌اند. به طور مثال در جدول ۵-۵ می‌توان دید که صدای "خنده" در موارد زیادی با "گفتار" یا "سرفه" که هردو مانند "خنده" از حنجره‌ی انسان تولید می‌شوند اشتباه گرفته شده است، همچنین صداهای "سرفه" و "صاف کردن گلو" و "گفتار" نیز به همین دلیل در برخی موارد اشتباه گرفته شده‌اند. صدای "بستن در" و "کشو" نیز دارای شباهت بوده‌اند. از دیگر صداهای دارای شباهت می‌توان به "افتادن قلم" و "سوئیچ" اشاره کرد که هر دو دارای ماهیت ضربه‌ای هستند. اما مسئله‌ی مهمی که از دقت در ماتریس‌های پراکندگی برداشت می‌شود این است که در هر سه ماتریس، بیش‌ترین میزان پراکندگی متعلق به کلاس "زنگ تلفن" است. دلیل این مسئله این است که در این پایان‌نامه در مرحله‌ی استخراج ویژگی، برای صداهایی که دارای سکوت داخل واقعه هستند مانند "زنگ تلفن" ، "گفتار" یا "صفحه کلید رایانه"، عمل حذف سکوت صورت نگرفته است. ازاین‌رو ویژگی‌های استخراج شده از قاب‌های حاصل از این سکوت داخل وقایع، باعث ایجاد شباهت بین این کلاس‌ها و کاهش نرخ شناسایی کلی سیستم شده است. به احتمال زیاد، با حذف سکوت داخل وقایع، و تمرکز روی نواحی دارای اطلاعات سیگنال‌های وقایع، می‌توان انتظار داشت که نرخ شناسایی سیستم بهبود داشته باشد.

جدول ۵-۵- ماتریس پراکندگی نرخ شناسایی(%) کلاس های مختلف با استفاده از ترکیب ویژگی های MFCC و MP و طبقه بند GMM (خانه های خالی جدول مربوط به اعداد کوچکتر از یک هستند).

۱۶ کلاس: ۱) زنگ هشدار، ۲) صاف کردن گلو، ۳) سرفه، ۴) بستن در، ۵) کشو، ۶) صفحه کلید رایانه، ۷) دسته کلید، ۸) در زدن، ۹) خنده، ۱۰) موشواره، ۱۱) ورق زدن، ۱۲) افتادن خودکار(قلم)، ۱۳) زنگ تلفن، ۱۴) چاپگر، ۱۵) گفتار، ۱۶) سوئیچ.

	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶
۱	۶۷												۱۰			
۲	۳	۶۰	۷						۸							
۳	۵	۱۱	۸۵	۲					۲۳			۸	۳			
۴		۳		۵۳					۵	۳						۳
۵		۷		۲۵	۹۸				۲				۶	۵		
۶	۵	۵		۶		۶۰				۱۲			۱۸			۳
۷							۹۸					۱۳				۷
۸								۹۷	۳				۵			
۹	۴	۳	۴						۳۹						۳	
۱۰	۲						۳			۸۵						۲
۱۱			۳	۶		۳۹			۳		۹۸	۸	۳			۸
۱۲	۵			۸								۶۵				۳۲
۱۳	۵												۴۷			
۱۴														۹۵		
۱۵	۴	۱۲						۳	۱۸				۷		۹۶	
۱۶											۲	۴				۴۶

جدول ۵-۶- ماتریس پراکندگی نرخ شناسایی(%) کلاس های مختلف با استفاده از ویژگی های BFCC و طبقه-بند GMM (خانه های خالی جدول مربوط به اعداد کوچکتر از یک هستند).

۱۶ کلاس: ۱) زنگ هشدار، ۲) صاف کردن گلو، ۳) سرفه، ۴) بستن در، ۵) کشو، ۶) صفحه کلید رایانه، ۷) دسته کلید، ۸) در زدن، ۹) خنده، ۱۰) موشواره، ۱۱) ورق زدن، ۱۲) افتادن خودکار(قلم)، ۱۳) زنگ تلفن، ۱۴) چاپگر، ۱۵) گفتار، ۱۶) سوئیچ.

	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶
۱	۶۲												۱۶			
۲	۳	۶۴							۳				۲		۳	
۳	۹	۱۳	۸۵	۳		۳			۲۱			۶	۱۰			۳
۴				۸۳				۳	۳	۳						
۵		۵		۸	۹۷					۲				۸		
۶		۳		۲	۳	۵۸							۵			
۷				۲			۹۳					۵				
۸								۹۳		۳			۸			
۹	۵	۲	۸						۵۵				۵		۴	
۱۰							۶	۲		۸۹						۳
۱۱			۴			۳۶					۹۹	۸	۴			۲
۱۲			۳			۳						۸۱				۲۰
۱۳	۳												۲۹			
۱۴														۹۲		
۱۵	۱۷	۱۳						۲	۱۸				۲۰		۹۳	
۱۶										۳						۷۱

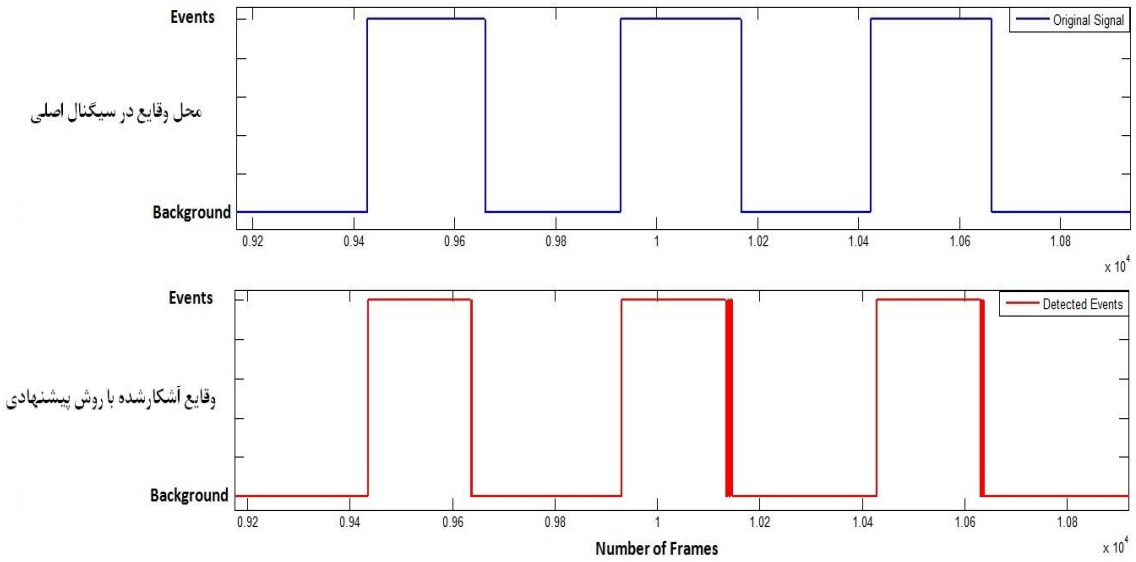
جدول ۵-۷- ماتریس پراکندگی نرخ شناسایی(%) کلاس های مختلف با استفاده از ترکیب ویژگی های BFCC و MP و طبقه بند GMM (خانه های خالی جدول مربوط به اعداد کوچکتر از یک هستند).

۱۶ کلاس: ۱) زنگ هشدار، ۲) صاف کردن گلو، ۳) سرفه، ۴) بستن در، ۵) کشو، ۶) صفحه کلید رایانه، ۷) دسته کلید، ۸) در زدن، ۹) خنده، ۱۰) موشواره، ۱۱) ورق زدن، ۱۲) افتادن خودکار(قلم)، ۱۳) زنگ تلفن، ۱۴) چاپگر، ۱۵) گفتار، ۱۶) سوئیچ.

	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶
۱	۵۶												۱			
۲		۷۷	۶						۵				۲			
۳	۱۲	۴	۸۵	۸					۲۰			۱۲				۱
۴				۸۰	۴			۲		۵		۱				۱
۵				۷	۹۳			۱						۵		
۶		۵		۲	۴	۵۶				۴			۴			۲
۷						۱	۹۶			۷		۵				
۸								۹۰					۱۰			
۹	۲	۱	۴						۵۰				۵	۲		
۱۰							۱	۲		۷۵						
۱۱			۶			۴۲	۲				۱۰۰	۱۱	۵			۱۰
۱۲										۴		۷۰				۱۵
۱۳	۲			۱									۵۰			
۱۴														۹۵		
۱۵	۲۶	۱۳		۱		۱		۴	۲۵	۱			۲۴		۹۸	
۱۶	۱									۵		۱				۷۰

۵-۴- محاسبه و مقایسه ی پارامتر F-Score

با توجه به نتایج بخش طبقه‌بندی، تصمیم گرفته شد تا بار دیگر آزمون آشکارسازی، این بار با روش آشکارسازی بر مبنای طبقه‌بندی، امتحان شود. برای این منظور، پس از یک مرحله حذف نویز که در بخش ۵-۱-۲ شرح داده شد، از روش طبقه‌بندی مدل مخلوط گوسی و ویژگی‌های ضرایب BFCC که در مرحله‌ی طبقه‌بندی بهترین نتیجه را بدست داد استفاده شده است. پس از آشکارسازی برای محاسبه‌ی دقت، وقایعی که دارای خطای مجموع ۲۰۰ میلی ثانیه برای شروع و پایان واقعه بودند به عنوان مورد صحیحِ آشکارسازی انتخاب شدند. در این روش آشکارسازی، واقعه‌ی "در زدن" هرگز شناسایی نمی‌شود. هم‌چنین وقایع "کشو" و "سوئیچ" مواردی هستند که به سختی شناسایی می‌شوند. در شکل ۵-۱۷ نمونه‌ای از این آشکارسازی را می‌توان ملاحظه کرد. نتایج نهایی مربوط به این روش آشکارسازی نیز همراه با نتیجه‌ی آشکارسازی از طریق آستانه‌گذاری، در جدول ۵-۸ در مقایسه با نتایج روش‌های پیشین آورده شده است.



شکل ۵-۱۷: نمونه ای از آشکارسازی وقایع در فایل پیوسته ی صوت با استفاده از ویژگی‌های BFCC و طبقه بند مدل مخلوط گوسی.

جدول ۵-۸- مقایسه ی نتایج روش های پیشنهادی با دیگر سیستم های مشابه در این زمینه.

	ویژگی	طبقه‌بند	F-Score (مبتنی بر رویداد)
[۴۷]	MFCC	SVM	۰/۱۱
[۴۸]	MFCC	GMM	۰/۵۶
[۴۹]	LowLevel+MFCC+LPC	HMM	۰/۴۴
[۵۰]	MFCC	HMM	۰/۵۵
[۵۱]	Spectrogram	NMF/HMM	۰/۴۷
[۵۲]	GFB-ASR	HMM	۰/۵۱
[۴]	GFB-AED	HMM	۰/۶۲
روش‌های پیشنهادی	**آشکارسازی با استفاده از آستانه‌گذاری**		**۰/۴۲**
	BFCC	**GMM**	**۰/۵۸**

در جدول فوق مشاهده می‌شود که روش پیشنهادی آشکارسازی بر مبنای آستانه‌گذاری، نسبت به اکثر روش‌های آشکارسازی بر مبنای طبقه‌بندی ضعیف‌تر عمل کرده است. اما روش پیشنهادی آشکارسازی بر مبنای طبقه‌بندی، برخلاف مورد قبلی، نتیجه‌ی بهتری نسبت به اکثر روش‌های موجود داشته و تنها نسبت به روش ارائه شده در [۴] نتیجه‌ی ضعیف‌تری دارد. البته با توجه به میزان اختلاف کم نتایج و در نظر گرفتن این مسئله که روش پیشنهادی در مقایسه با روش مذکور به لحاظ محاسباتی بسیار ساده‌تر می‌باشد می‌توان گفت که نتیجه‌ی قابل قبولی حاصل شده است.

فصل ۶ نتیجه گیری و پیشنهادها

۶-۱- نتیجه گیری

وظیفه‌ی سیستم شناسایی وقایع صوتی، آشکارسازی و برچسب‌زنی به صدایی است که از منبعی ناشناس در یک محیط که نویز یا عوامل مزاحم دیگر ممکن است وجود داشته باشند رخ می‌دهد. در این پایان‌نامه، پیش‌زمینه‌ای در مورد پژوهش‌های صورت گرفته در این زمینه ارائه شد و مهم‌ترین مشخصات وقایع صوتی مورد بحث قرار گرفت. رایج‌ترین و مهم‌ترین سیستم‌ها در این زمینه مبتنی بر سیستم‌های شناسایی خودکار گفتار هستند که به طور معمول در محیط‌های فاقد نویز به خوبی عمل می‌کنند اما در شرایط نویزی، عملکرد آن‌ها به شدت تضعیف می‌شود. در این پایان‌نامه دو روش استخراج ویژگی برای غلبه بر مشکل تضعیف نرخ شناسایی در شرایط نویزی پیشنهاد شد. هر دو روش عملکرد نسبتاً مناسبی از خود نشان دادند.

- **پیگیری انطباق در ترکیب با MFCC**: روش پیگیری انطباق به تنهایی نرخ شناسایی چندان مناسبی ارائه نداد اما در ترکیب با ویژگی‌های متداول MFCC پاسخ نسبتاً مناسبی حاصل شد، البته این بهبود، در استفاده از طبقه‌بند نزدیکترین همسایه مشاهده شد اما در استفاده از طبقه‌بند مدل مخلوط گوسی هیچ‌گونه بهبودی برای نرخ شناسایی سیستم با استفاده از این مجموعه ویژگی دیده نشد. از این‌رو می‌توان گفت که روش پیگیری انطباق به دلیل شباهت بین کلاسی داده‌ها، در نمایش این داده‌ها چندان کارآمد نبوده است.
- **ضرایب کپسترال فرکانس بارک: از** این روش استخراج ویژگی که برگرفته از روش PLP است و شباهت زیادی با این الگوریتم دارد برای شناسایی وقایع صوتی مربوط به محیط اداری نتیجه‌ی خوبی حاصل شد.

۶-۲- پیشنهادها

با اینکه نشان داده شد که روش‌های پیشنهادی در این پایان‌نامه در شرایط نویزی خوب عمل می‌کنند، هنوز جنبه‌های بسیاری وجود دارد که نیاز به پژوهش بیش‌تر دارند. فهرست زیر پیشنهاداتی برای پژوهش‌های آینده دربردارد:

- **حذف سکوت داخل برخی از وقایع:** در بخش ۵-۳ توضیح داده شد که نواحی سکوت داخل برخی از وقایع مانند "زنگ تلفن"، "گفتار" یا "زنگ هشدار" باعث ایجاد شباهت بین وقایع می‌شود و جداپذیری بین کلاس‌ها را کاهش می‌دهد، از این‌رو با حذف نواحی مربوط به سکوت داخل وقایع، می‌توان انتظار بهبود در نرخ شناسایی کلی سیستم داشت.

- **استفاده از واژه‌نامه‌های جدید در الگوریتم پیگیری انطباق:** این الگوریتم در مسئله‌های دیگر پاسخ خوبی بدست داده است، از این‌رو شاید با ارائه‌ی یک واژه‌نامه‌ی جدید (یا حتی یک واژه‌نامه ساخته شده از خود داده‌های آموزشی) بتوان پاسخ را بهبود داد.

- **کار روی وقایع دارای هم‌پوشانی زمانی:** در دنیای واقعی در هر محیطی، احتمال این‌که وقایع صوتی تولید شده توسط منابع مختلف دارای هم‌پوشانی زمانی باشند بسیار زیاد است، از اینرو پس از رسیدن به نتیجه‌ای مناسب در مورد وقایع صوتی مجزا، نیاز به پژوهش بیشتر در زمینه‌ی وقایع صوتی دارای هم‌پوشانی می‌باشد تا سیستم شناسایی دارای قدرت بیش‌تری باشد. هرچند این زمینه بسیار مشکل به نظر می‌رسد ولی از آنجا که گوش انسان توانایی تشخیص تمامی این موارد را دارد پس با تلاش بیش‌تر می‌توان به مجموعه ویژگی‌هایی که سیستم شنوایی انسان را بهتر مدل کنند دست یافت.

- **استفاده از روش‌های پردازش تصویر برای استخراج ویژگی از اسپکتروگرام‌های وقایع صوتی:** هرچند که اسپکتروگرام‌های قطعه‌های صوتی پایگاه داده‌ی مورد استفاده در این پایان‌نامه دارای شباهت بین کلاسی بسیاری هستند ولی به دلیل پیشرفت الگوریتم‌های

پردازش تصویر کنونی، می‌توان امید داشت که استفاده از اسپکتروگرام‌ها نیز منجر به پاسخ مناسبی برای مسئله‌ی شناسایی شود.

مراجع

[1] T. Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4. pp. 441–457, 2001.

[2] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," *2006 IEEE International Conference on Multimedia and Expo*. pp. 885–888, 2006.

[3] A. Waibel, H. Steusloff, and R. Stiefelhagen, "Chil—Computers in the human interaction loop," in *Proc. WIAMIS*, 2004, and the CHIL Project Consortium.

[4] J. Schröder, S. Goetze, and J. Anemüller, "Spectro-temporal Gabor filterbank features for acoustic event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2198–2208, 2015.

[5] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Process. Lett.*, vol. 18, no. 2, pp. 130–133, 2011.

[6] J. Dennis, H. D. Tran, and E. S. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 21, no. 2, pp. 367–377, 2013.

[7] J. Dennis, H. D. Tran, and H. Li, "Combining robust spike coding with spiking neural networks for sound event classification," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 176–180.

[8] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[9] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 1, pp. 1–13, 2013.

[10] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor network for the monitoring of ecosystem: Bird species recognition," in *Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on*, 2007, pp. 293–298.

[11] I. Boesnach, M. Hahn, J. Moldenhauer, T. Beth, and U. Spetzger, "Analysis of drill sound in spine surgery," in *Perspective in image-guided surgery: proceedings of the Scientific Workshop on Medical Robotics, Navigation, and Visualization: RheinAhrCampus Remagen, Germany, 11-12 March*, 2004, p. 77.

[12] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," in *Signal Processing Conference, 2007 15th European*, 2007, pp. 1216–1220.

[13] C. Clavel, T. Ehrette, and G. Richard, “Events detection for an audio-based surveillance system,” in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 1306–1309.

[14] F. Jin, F. Sattar, and S. Krishnan, “Log-frequency spectrogram for respiratory sound monitoring,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 597–600.

[15] S. Päßler and W.-J. Fischer, “Food intake monitoring: Automated chew event detection in chewing sounds,” *IEEE J. Biomed. Heal. informatics*, vol. 18, no. 1, pp. 278–289, 2014.

[16] J. Schröder, S. Goetze, V. Grutzmacher, and J. Anemüller, “Automatic acoustic siren detection in traffic noise by part-based models.,” in *ICASSP*, 2013, pp. 493–497.

[17] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, “Segmentation, indexing, and retrieval for environmental and natural sounds,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 18, no. 3, pp. 688–707, 2010.

[18] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, “Real-world acoustic event detection,” *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1543–1551, 2010.

[19] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, “Bathroom activity monitoring based on sound,” in *International Conference on Pervasive Computing*, 2005, pp. 47–61.

[20] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, “Robust environmental sound recognition for home automation,” *IEEE Trans. Autom. Sci. Eng.*, vol. 5, no. 1, pp. 25–31, 2008.

[21] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, “Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring,” *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1524–1534, 2010.

[22] J. Mantyjarvi, P. Huuskonen, and J. Himberg, “Collaborative context determination to support mobile terminal applications,” *IEEE Wirel. Commun.*, vol. 9, no. 5, pp. 39–45, 2002.

[23] J. W. Dennis, “Sound Event Recognition and Classification in Unstructured Environments,” Ph.D. dissertation, Nanyang Tech. Univ., 2011.

[24] N. Yamakawa, T. Kitahara, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, “Effects of modelling within-and between-frame temporal variations in power spectra on non-verbal sound recognition,” in *INTERSPEECH*, 2010, pp. 2342–2345.

[25] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, and C. Ris, “Automatic speech recognition and speech variability: A review,” *Speech Commun.*, vol. 49, no. 10, pp. 763–786, 2007.

[26] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[27] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, “CLEAR evaluation of acoustic event detection and classification systems,” in *International Evaluation Workshop on Classification of Events, Activities and Relationships*, 2006, pp. 311–322.

[28] A. Waibel, R. Stiefelhagen, R. Carlson, J. Casas, J. Kleindienst, L. Lamel, O. Lanz, D. Mostefa, M. Omologo, and F. Pianesi, “Computers in the human

interaction loop," in *Handbook of Ambient Intelligence and Smart Environments*, Springer, 2010, pp. 1071–1116.

[29] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni, and F. Piazza, "An integrated system for voice command recognition and emergency detection based on audio signals," *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5668–5683, 2015.

[30] A. Plinge, R. Grzeszick, and G. A. Fink, "A bag-of-features approach to acoustic event detection," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3704–3708.

[31] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique," *Digit. Signal Process.*, vol. 29, pp. 147–157, 2014.

[32] W. Choi, S. Park, D. K. Han, and H. Ko, "Acoustic event recognition using dominant spectral basis vectors," in *International Speech and Communication Association*, 2015.

[33] J. Ludeña-Choez and A. Gallardo-Antolín, "NMF-based spectral analysis for acoustic event classification tasks," in *International Conference on Nonlinear Speech Processing*, 2013, pp. 9–16.

[34] J. Ludeña-Choez and A. Gallardo-Antolín, "Feature extraction based on the high-pass filtering of audio signals for Acoustic Event Classification," *Comput. Speech Lang.*, vol. 30, no. 1, pp. 32–42, 2015.

[35] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 2, pp. 1–22, 2006.

[36] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic Instrument Recognition Using Spectral Clustering," in *ISMIR*, 2007, pp. 213–218.

[37] J. J. Burred, A. Robel, and T. Sikora, "Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 173–176.

[38] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. speech audio Process.*, vol. 10, no. 5, pp. 293–302, 2002.

[39] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal feature integration for music genre classification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 15, no. 5, pp. 1654–1664, 2007.

[40] R. Tao, Z. Li, Y. Ji, and E. M. Bakker, "Music genre classification using temporal information and support vector machine," in *Proc. of the 16th Advanced School for Computing and Imaging Conf.(ASCI 2010)*, 2010.

[41] B. Mechtley, G. Wichern, H. D. Thornburg, and A. Spanias, "Combining semantic, social, and acoustic similarity for retrieval of environmental sounds.," in *ICASSP*, 2010, pp. 2402–2405.

[42] G. Peeters and E. Deruty, "Sound indexing using morphological description," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 18, no. 3, pp. 675–687, 2010.

[43] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[44] J. Xiang, M. F. McKinney, K. Fitz, and T. Zhang, "Evaluation of sound classification algorithms for hearing aid applications," in *2010 IEEE international conference on acoustics, speech and signal processing*, 2010, pp. 185–188.

[45] M. Cowling and R. Sitte, "Analysis of speech recognition techniques for use in a non-speech sound recognition system," in *Proc. Digital Signal Processing for Communication Systems*, 2002.

[46] F. Beritelli and R. Grasso, "A pattern recognition system for environmental sound classification based on MFCCs and neural networks," in *Signal Processing and Communication Systems, 2008. ICSPCS 2008. 2nd International Conference on*, 2008, pp. 1–4.

[47] W. Nogueira, G. Roma, and P. Herrera, "Automatic event classification using front end single channel noise reduction, MFCC features and a support vector machine classifier," *IEEE AASP Chall. Detect. Classif. Acoust. Scenes Events*, pp. 1–2, 2013.

[48] L. Vuegen, B. V. D. Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, and H. V Hamme, "An MFCC-GMM approach for event detection and classification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–3.

[49] M. E. Niessen, T. L. M. Van Kasteren, and A. Merentitis, "Hierarchical modeling using automated sub-clustering for sound event recognition," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[50] A. Diment, T. Heittola, and T. Virtanen, "Sound event detection for office live and office synthetic AASP challenge," *Proc. IEEE AASP Chall. Detect. Classif. Acoust. Scenes Events*, 2013.

[51] J. F. Gemmeke, L. Vuegen, P. Karsmakers, and B. Vanrumste, "An exemplar-based NMF approach to audio event detection," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[52] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "Acoustic event detection using signal enhancement and spectro-temporal feature extraction," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2013.

[53] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, "Acoustic event detection and classification," in *Computers in the Human Interaction Loop*, Springer, 2009, pp. 61–73.

[54] J. Ramirez, J. M. Górriz, and J. C. Segura, *Voice activity detection. fundamentals and speech recognition system robustness*. INTECH Open Access Publisher, 2007.

[55] D. Hoiem, Y. Ke, and R. Sukthankar, "SOLAR: Sound object localization and retrieval in complex audio environments," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 2005, vol. 5, pp. v–429.

[56] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," *CUIDADO IST Project Report*, pp. 1–25, 2004.

[57] D. Kamińska, T. Sapiński, and A. Pelikant, "Comparison of perceptual features efficiency for automatic identification of emotional states from speech," in *2013*

6th International Conference on Human System Interactions (HSI), 2013, pp. 210–213.

[58] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[59] "D-CASE: Detection and classification of acoustic scenes and events," 2013 [Online]. Available: http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/

Abstract

In recent years, significant researches has been done in the field of environmental sounds and audio events recognition, but these studies are very limited in comparison with other related areas such as speech and music. The aim of this thesis is the development of feature extraction methods for sound events recognition in office environment. The IEEE Audio and Acoustic Signal Processing Technical Committee Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) dataset is used for the task. The dataset consists of 16 classes of sound events in office environment, which some of these sounds are noisy with different SNR. For this research, two feature extraction methods are introduced. In the first method, features are extracted by Matching Pursuit (MP) algorithm in combination with common Mel Frequency Cepstral Coefficients (MFCC) as feature vectors by using nearest neighbour classifier. The recognition rate is 69.67 percent in this method that is increaded 6 percent in comparison with the case of using MFCC without MP. In the second method, Bark Frequency Cepstral Coefficients (BFCC) and GMM are employed as feature and classifier, respectively. In this case, recognition rate is achieved 80.08 percent that shows effectiveness of proposed method in comparison with the most of existing methods for used dataset.

Keywords: sound events, feature extraction, matching pursuit (MP), brak-frequency cepstral coefficients (BFCC).

Shahrood University of Technology

Faculty of Electrical and Robotic Engineering

MSc Thesis in Electronic Engineering

Sound Event Detection Based on MP Features

By: Roghaye Bahmani

Supervisor:
Dr. Hossein Marvi

July 2016