ارائه یک مدل پارامتریک تطبیقی جهت کشف و رده‌بندی وقایع صوتی در سیگنال‌های محیطی

مراد درخشان[۱]، دانشجوی دکتری؛ حسین مروی[۲]، دانشیار؛ حمید حسن پور[۳]، استاد
۱- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران– mderakhshan@shahroodut.ac.ir
۲- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران– marvi_hossein@yahoo.co.uk
۳- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران– h.hassanpour@shahroodut.ac.ir

چکیده: کشف وقایع صوتی در محیط کار و زندگی یک نیاز مدرن جهت گردآوری اطلاعات است. تاکنون بیشتر تحقیق‌ها بر واقعه صوتی خاص و یا تعداد محدودی از وقایع صوتی برجسته متمرکز بوده‌است. در اینجا یک مدل‌سازی جدید جهت کشف تمام وقایع صوتی رخداده در رکورد و تعیین محدوده زمانی برای هر یک از آن‌ها ارائه شده‌است. نوآوری شامل مدل‌سازی جدید همراه با پارامترهای تطبیقی در مدل است. پس از استخراج ویژگی‌ها و تعیین مقادیر دو پارامتر آلفا و بتا از دو قطعه‌بندی مجزا و ترکیب خروجی آن‌ها برای تعیین وقایع صوتی و محدوده زمانی آن‌ها استفاده شده‌است. این وقایع جهت رده‌بندی به الگوریتم KNN فرستاده می‌شوند. پارامترها امکان دقت بیشتر و یا میزان کشف حداکثری را ممکن می‌سازند. وقایع صوتی آزمایش‌شده شامل ۱۶ نوع صدای اتاق کار اداری هستند که برخی شبیه هم و بعضی نیز مشابه نویز محیط هستند. در سنجش عملکرد برحسب واقعه، میزان درستی کشف ۷۰/۱ درصد، فراخوانی ۷۵/۸ درصد و میزان F1، ۷۲/۸ درصد بوده‌است. همچنین میزان F1 برحسب فریم ۸۰/۶ درصد حاصل شد. مقدار F1 برحسب واقعه، نسبت به قبل ۱۰/۸٪ بهبود داشته‌است که مویدکارآمدی مدل پیشنهادی است.

واژه‌های کلیدی: کشف وقایع صوتی، صداهای محیطی، الگوریتم‌های یادگیری بدون نظارت، سیستم‌های پارامتریک تطبیقی، سیستم‌های نظارت صوتی، سیستم‌های کسب اطلاعات مبتنی بر صدا.

Providing an Adaptive Model with two Adjustable Parameters for Audio Event Detection and Classification in Environmental Signals

Morad Derakhshan, PhD student[1]; Hossein Marvi, Associate professor[2]; Hamid Hassan poor, professor[3]
1- Computer and IT Engineering Department, Shahrood University of Technology, Shahrood, Iran, mderakhshan@shahroodut.ac.ir
2- Computer and IT Engineering Department, Shahrood University of Technology, Shahrood, Iran, marvi_hossein@yahoo.co.uk
3- Computer and IT Engineering Department, Shahrood University of Technology, Shahrood, Iran, h.hassanpour@shahroodut.ac.ir

Abstract: Audio event detection (AED) is a modern way to collect data about human activities in the workplace or in other life environments. We proposed a novel adaptable model based on using two parameters, α and β to detect all audio events that may be present in a given record accompanied by their time limits in which they occur. After feature extraction and setting the values of the two key parameters, alpha and beta, the audio sequence will be sent into two distinct sub-systems for event detection. The outputs from the two sub-classifiers are then combined and necessary refinements are made on the event time limits. The final detected events are sent to the KNN classifier.The parameters serve as a trade-off tool between precision and recall expectation in the detection process. In the tests, 16 different audio events of an office room were detected, some being similar to each other and some have very similar characteristics to those of the background noise. At frame-based (FB) level, the precision rate was 70.1%, the rate of recall was 75.8%, and F1-measure was 72.8%. The F1-measure has increased by 10.8% suggesting promising applications of the model.

Keywords: Audio event detection (AED), environmental sounds, unsupervised learning, adaptable modeling systems, audio monitoring systems, audio-based acquisition systems.

تاریخ ارسال مقاله: ۱۳۹۶/۰۳/۲۳
تاریخ اصلاح مقاله: ۱۳۹۶/۰۵/۰۷ و ۱۳۹۶/۰۷/۰۵
تاریخ پذیرش مقاله: ۱۳۹۶/۰۸/۱۸
نام نویسنده مسئول: دکتر حسین مروی
نشانی نویسنده مسئول: ایران - شاهرود – دانشگاه صنعتی شاهرود - دانشکده مهندسی کامپیوتر و فناوری اطلاعات.

۱- مقدمه

جداسازی و شناسایی یک اتفاق صوتی در محدوده زمانی رخداده برای انسان، کاری آسان و تکراری است، اما انجام آن برای پردازش‌های ماشینی به‌عنوان یک چالش اساسی مطرح است و حل آن یکی از اولویت‌های اصلی انجمن‌های پردازش صوت می‌باشد. کاربردهایی همچون نظارت و دیده‌بانی صوتی در محیط‌های عمومی، داده‌کاوی صوتی، درک یک ابزار هوشمند مثل گوشی‌های تلفن سیار نسبت به محیط اطراف خود و همچنین مراقبت خودکار از افراد بیمار یا سالمند همگی از صوت به عنوان بستر اطلاعات استفاده می‌کنند.

از پژوهش‌های گذشته در این حوزه می‌توان به مطالعه [۱] اشاره نمود که سیستمی را مبتنی بر رده‌بندی One-class SVM برای شناسایی ناهنجاری‌هایی همچون صدای شلیک تفنگ، شکستن شیشه و صدای جیغ ارائه نموده‌اند. در [۲] یک مدل تطبیق‌پذیر برای کشف وقایع صوتی کوتاه‌مدت و درازمدت و همچنین برای صداهایی که در زمینه خود نویز مشابه با صدای اصلی دارند، پیشنهاد کردند. پژوهش [۳] به شناسایی مشخصه‌های صدای شلیک تفنگ پرداخته‌است. در [۴] به‌منظور دسته‌بندی صحنه‌های شنوایی، شناسایی و ایندکس کردن معنایی وقایع صوتی در هر صحنه شنوایی انجام شده‌است. در [۵] با استفاده از مدل نیمه‌نظارتی بیز وقایع صوتی در رکورد ورودی را رونویسی کرده و وقایع چندگانه همزمان را مورد پردازش قرار داده‌است. پژوهش [۶] یک مدل احتمالاتی محدود زمانی را برای کشف وقایع استفاده نموده‌است. در [۷] صحنه‌های شنوایی را از طریق شناسایی صداهای موجود در آن تشخیص می‌دهند. در رویکرد دسته‌فریمی [۸] برای هر صدا یک مدل GMM آموزش داده‌می‌شود آنگاه تخمینی از مجموع GMMهای تمام فریم‌ها محاسبه‌شده و رده‌ای که دارای بالاترین شباهت باشد، انتخاب می‌شود. در [۹، ۱۰] رویکرد ویژگی‌های دسته‌ای پیشنهاد شده‌است که هیستوگرام حاصل از ویژگی‌های خوشه‌بندی‌شده را برای رده‌بندی استفاده می‌کند. تحقیق [۱۱] از روش جداسازی سیگنال‌ها برای تجزیه سیگنال ورودی به وقایع استفاده کرده‌است. تجزیه نامنفی ماتریس در مدل‌های دیگری از جمله [۱۲، ۱۳] استفاده شده‌است. در آنجا سیگنال ورودی مشاهده به دو ماتریس نامنفی تجزیه شده و برای سیگنال آزمایش از یک دیکشنری وقایع استفاده می‌کنند. در [۱۴] وصله‌های طیفی از پیش استخراج‌شده برای تجزیه سیگنال به وقایع استفاده می‌شود و از تعداد زیادی ویژگی حوزه زمان و فرکانس استفاده نموده‌است.

سیستم‌های متعددی عمدتا بر اساس شبکه‌های عصبی CNN، DNN و RNN به‌منظور شناسایی صداهای محیطی و شناسایی صحنه‌های صوتی در DCASE_2016 پیشنهاد شده‌است که برروی پایگاه داده یکسان حاوی ۱۱ صدای محیطی بسط داده‌شده‌اند [۱۵]. ازجمله در [۱۶] شبکه‌های عصبی عمیق DNN همراه با مدل‌سازی حذف نویز برای شناسایی وقایع صوتی بکار گرفته‌شده و از ویژگی‌های انرژی باندهای فرکانسی مل استفاده کرده‌است. در [۱۷] از شبکه عصبی LSTM-RNN دوسویه به همراه HMM و ۱۰۰ ویژگی از نوع فیلتر بانک فرکانس مل برای هر فریم استفاده نموده‌اند. در [۱۸] از شبکه‌های حسگر با چندین محل اخذ صدای محیط و سپس ترکیب و امتزاج اطلاعات و بهره‌گیری از پشته‌سازی رده‌بند برای شناسایی میکروفون و محل ارسال صدا استفاده شده‌است. در [۱۹] از ترکیب شبکه‌های مصنوعی GRNN با LDA و تعریف یک تابع هدف براساس تابع هزینه ترکیبی از MSE علاوه بر شناسایی وقایع، صحنه صوتی را نیز شناسایی کرده‌اند.

اما اخیرا روش‌هایی بر پایه مدل پیوسته سرهم‌بندی دومرحله‌ای پیشنهاد شده‌است. ازجمله در [۲۰] با هدف استخراج ویژگی‌های دارای قابلیت جداکنندگی بالا رویکرد بوستینگ[۱] پیشنهاد شده‌است و نتایج حاصل نیز نسبت به ویژگی‌های ادراکی سنتی همچون ضرایب فرکانس[۲] مل و لگاریتم بانک‌های فیلتر[۳] فرکانس مل بهتر بوده‌است. این نوع مدل‌سازی را آماری هرمی[۴] می‌گویند. همچنین [۲۱، ۲۲] در یک مدل پیوسته دوپشته،[۵] مزایای مدل‌سازی متوالی را با قابلیت‌های جداسازی پرسپترون چندلایه ترکیب نموده و با اجرای یک مرحله رتبه‌بندی مجدد (همانند آنچه در بازشناسی گوینده انجام می‌شود) کارایی رده‌بندی را از طریق بکارگیری مدل GMM-SVM تقویت کرده‌است.

ایده اصلی در پژوهش حاضر، یک مدل‌سازی دارای دو فاز مجزا یکی جهت قطعه‌بندی و دیگری جهت کشف و رده‌بندی وقایع در سیگنال ورودی به همراه انجام قطعه‌بندی در دو زیرسیستم و داشتن پارامترهای مجزا برای آن‌ها است. با این روش مزیت استفاده حداکثری از ویژگی‌ها و همچنین مدل‌سازی به صورت تفکیک اجزا بجای مدل‌سازی یکپارچه و غیرقابل تفکیک دیده شده‌است. دو زیرسیستم مجزا قابلیت تطبیق الگوریتم با نوع ویژگی‌های استخراجی را دارد. مزیت اصلی این مدل استفاده هر چه بیشتر از ظرفیت ویژگی‌های مورد استفاده و ایجاد فضا برای تصمیم‌گیری چندشاخه برای تعیین قطعات و همین طور استفاده از قدرت هم‌افزایی دو زیرسیستم بجای یک سیستم یکپارچه می‌باشد.

ادامه مقاله به‌صورت زیر است: بخش۲ مدل پیشنهادی و اجزای آن. بخش۳ استخراج ویژگی‌ها جهت الگوریتم کشف وقایع. بخش۴ قطعه‌بندی و رده‌بندی بر حسب مدل پیشنهادی. بخش۵ اجرای مدل و نتایج آن. بخش۶ نتیجه‌گیری. بخش ۷ چند پیشنهاد برای کار آینده.

۲- مدل پیشنهادی و اجزای آن

کشف وقایع صوتی معمولا در دو فاز انجام می‌شود. در فاز نخست ابتدا سیگنال قطعه‌بندی می‌شود، به‌طوری که بخش‌ها یا فریم‌هایی از سیگنال ورودی که حاوی صداهای محیطی هستند شناسایی شوند. فاز دوم کشف وقایع است که طی آن قطعه‌های جداازهم در قالب وقایع صوتی دارای محدوده زمانی شروع و پایان تعیین می‌شوند و رده‌بندی

قطعات نیز انجام می‌شود. دراین‌حالت چنانچه قطعه‌بندی به خوبی صورت نگیرد کل عملیات کشف وقایع و رده‌بندی با خطا و دقت پایین همراه خواهدبود. بنابراین میزان موفقیت الگوریتم بستگی به مدل‌سازی مناسب، تعیین درست پارامترهای مدل و همچنین انتخاب ویژگی‌هایی دارد که بتواند به‌خوبی ابعاد تغییرپذیر داده‌ها در رده‌های صوتی مختلف را نمایش دهد. چنانچه تعداد صداهای موجود در کشف وقایع زیاد باشد می‌توان استدلال کرد که ویژگی‌های خاص در صداهای مشخصی سبب تفکیک‌پذیری بهتر می‌شوند، لذا داشتن دو زیر سیستم مجزا ایده منطقی به‌نظر می‌رسد.

در این مدل ابتدا عملیات اولیه نرمال‌سازی، سفیدکردن داده‌ها و حذف نویز از طریق فیلترینگ بر روی سیگنال اصلی انجام می‌شود. سپس از سیگنال ورودی $X(n)$ بردارهای ویژگی موردنظر استخراج می‌شوند. با داشتن بردارهای ویژگی آن‌ها به دو دسته جداگانه تبدیل می‌شوند. این تقسیم‌بندی به دلیل آن است که استفاده همزمان از تمام ویژگی‌ها می‌تواند در کشف تعداد زیاد صداها نقش منفی ایفا کنند، به‌طوری‌که امکان اینکه ویژگی‌ها اثر متمایزکننده یکدیگر را در الگوریتم خنثی نمایند زیاد است، لذا تبدیل ویژگی‌ها به دو دسته جهت شناسایی بیشتر صداها و ایجاد تمایز بین صداها بهتر عمل می‌کند. سپس عمل قطعه‌بندی بر روی سیگنال ورودی با دو دسته ویژگی متفاوت انجام می‌شود که این دو نوع قطعه‌بندی می‌توانند دو الگوریتم متفاوت و یا یکسان باشند. الگوریتم قطعه‌بندی با به‌کارگیری پارامتر آلفا و بتا می‌تواند بین کشف دقیق‌تر و یا کشف بیشتر وقایع مصالحه‌ای را بوجودآورد.

پس از اتمام دو نوع قطعه‌بندی نتایج آن‌ها برای کشف وقایع و یکپارچه‌سازی محدوده هر اتفاق صوتی و بدست‌آوردن زمان شروع و پایان هر واقعه صوتی مورد استفاده قرار می‌گیرد. در پایان مرحله قطعه‌بندی لیستی از وقایع به همراه زمان رخداد هر کدام از آن‌ها تولید شده‌است که در رشته $X_e(n)$ ذخیره می‌شوند.

یک مکانیزم دیگر برای فراهم نمودن شرایط محلی (local) برای یافتن قطعه‌های احتمالی از وقایع که بخاطر تصمیمات برحسب شرایط global کشف نشده‌اند اجرا می‌گردد. جهت شروع این مرحله سیگنال مرحله جدید را از طریق رابطه $X_1(n)=X(n)-X_e(n)$ بدست می‌آوریم و عملیات مرحله نخست با شرایط محلی در این مرحله با سیگنال باقیمانده اجرا می‌گردد. اجرای این مکانیزم با هدف کشف بیشتر و دقیق‌تر وقایع انجام می‌شود. رشته $X_e(n)$ لیست وقایعی است که با این مکانیزم کشف شده‌اند. نتایج مراحل اول و دوم با هم جمع شده، لیست نهایی وقایع و زمان‌های آن‌ها بدست می‌آید. اجزای مدل پیشنهادی در شکل ۱ آورده شده‌است.

۲-۱- نقش پارامترها در مدل پیشنهادی

در مدل از دو پارامتر آلفا (α) و بتا (β) جهت تنظیم عملکرد قطعه‌بندی استفاده می‌شود. الگوریتم‌های موجود قطعه‌بندی عموما از تعیین یک مقدار آستانه که بتواند شرایط متفاوتی را در انتخاب یا رد فریم‌های صوتی بکار گیرد و یا اینکه میزان کشف و دقت کشف را نسبت به هم کنترل نماید، دچار معضل هستند. ازسوی‌دیگر، وجود کمیت‌های مختلف برای آستانه که بتوانند اعداد صحیح، اعشاری و یا مختلط باشند، لزوم استفاده از پارامترهای انعطاف‌پذیر و متفاوت را توجیه‌پذیر می‌کند. برای مثال در فاز پیاده‌سازی این تحقیق پارامتر آلفا از نوع عددی صحیح و پارامتر بتا از نوع عددی اعشاری است. کارکرد این دو پارامتر می‌تواند سبب تراز بین میزان کشف و میزان دقت وقایع گردد. وجود این دو پارامتر انعطاف لازم به هر یک از الگوریتم‌های قطعه‌بندی را می‌دهد تا بتوانند در مورد کمیت آستانه و همچنین مقدار بهینه مورد نظر برای آستانه اقدام کنند.

۳- استخراج ویژگی‌ها جهت الگوریتم کشف وقایع

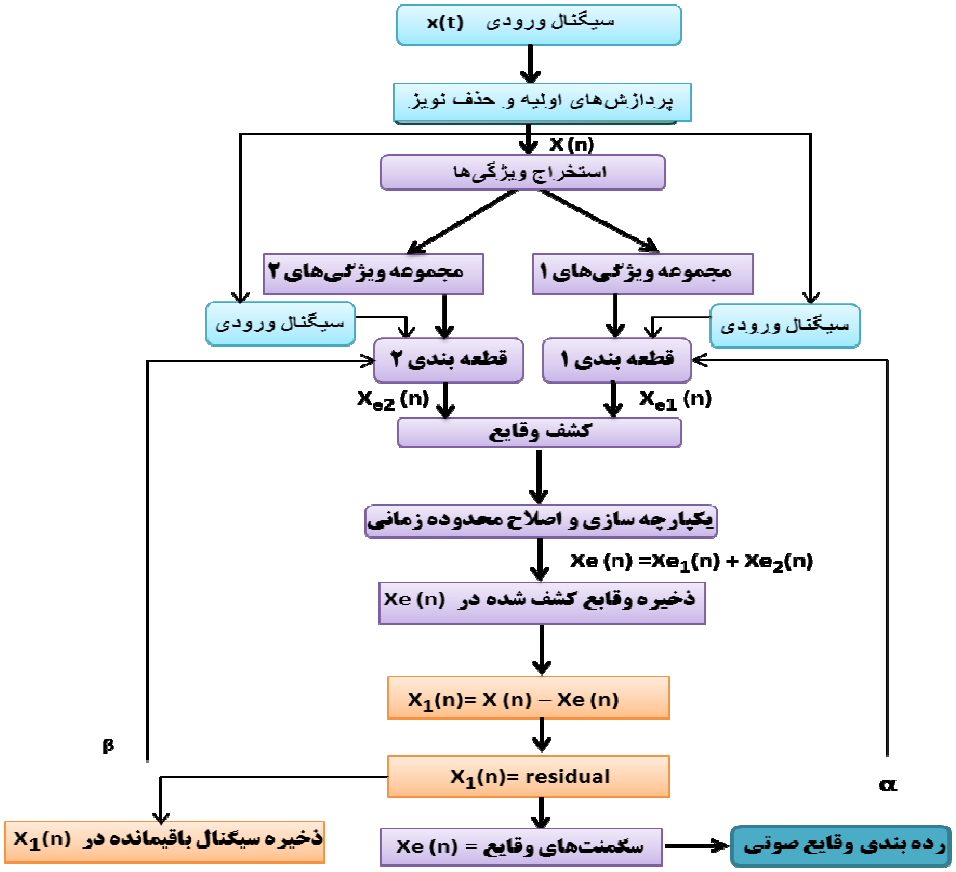
به دلایلی ازجمله سنجش توانایی ویژگی‌های حوزه زمان در کشف تعداد متنوع وقایع صوتی، تحقیق در نحوه تولید دو گروه از ویژگی‌های متفاوت با نوع پارامتر متفاوت و بعلاوه وجود تغییرهای بسیار سریع مشخصه‌های زمانی در وقایع صوتی جهت قطعه‌بندی حوزه زمان انتخاب شده‌است. پایگاه داده مورد استفاده از تعداد ۱۶ نوع واقعه صوتی مختلف تشکیل شده است که لیست این وقایع در جدول ۱ آمده است.

جدول ۱: انواع وقایع صوتی مورد پردازش در این پژوهش (پایگاه‌داده ارائه‌شده توسط انجمن تخصصی پردازش سیگنال صوت در موسسه بین‌المللی IEEE)

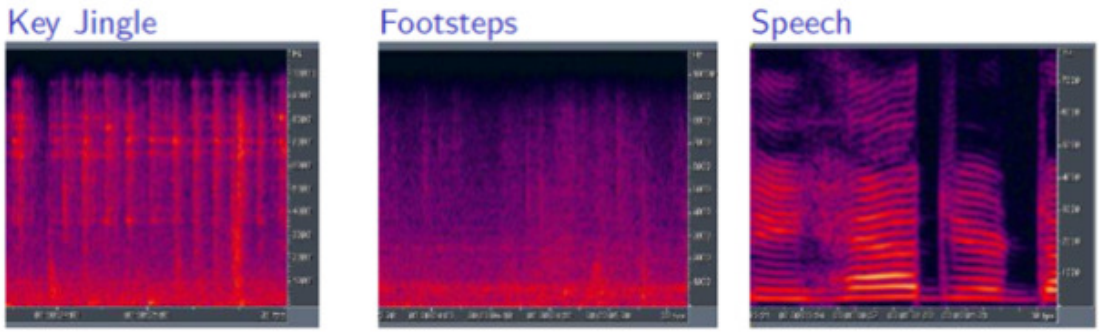
ردیف	نام واقعه صوتی	توضیحات
۱	alarm	(short alert (beep) sound)
۲	clearthroat	(clearing throat)
۳	cough	
۴	doorslam	(door slam)
۵	Drawer	
۶	keyboard	(keyboard clicks)
۷	keys	(keys put on table)
۸	knock	(door knock)
۹	laughter	
۱۰	mouse	(mouse click)
۱۱	pageturn	(page turning)
۱۲	pendrop	(pen, pencil, or marker touching table surfaces)
۱۳	phone	
۱۴	printer	
۱۵	speech	
۱۶	Switch	

تفاوت بین صداهای حنجره‌ای (گفتاری) و صداهای محیطی بحث مهمی است که در اینجا به آن اشاره می‌شود. شکل ۲ نمونه‌ای از تفاوت اطلاعات طیفی بین این دو دسته از صداها را نشان می‌دهد. همان‌گونه‌که مشخص است از صداهای گفتاری به‌راحتی می‌توان ساختار فرمنت‌ها و گام را شناسایی و استخراج کرد. اما صداهای محیطی فاقد ساختار طیفی مشخص هستند. لذا در بحث استخراج

ویژگی‌ها لازم است از ویژگی‌هایی که بیشتر با ساختار صداهای محیطی همخوانی دارند، استفاده شود. برای مثال می‌توان گفت ویژگی‌هایی همچون ضرایب MFCC که در شناسایی گفتار نهادینه شده‌اند چندان مناسب بحث ما نیستند. لذا در اینجا ویژگی‌های پیشنهادی بر حسب میزان و قابلیت جداسازی آن‌ها انتخاب شده‌اند.



شکل ۱: بلاک دیاگرام مدل پیشنهادی کشف وقایع صوتی.

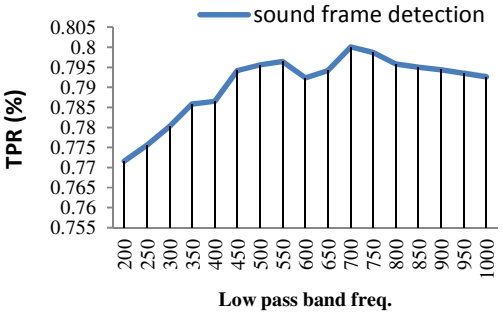


شکل ۲: وجود ساختار مشخص طیفی از جمله فرمنت‌ها و گام در صدای گفتاری و عدم ساختارطیفی در صداهای محیطی.

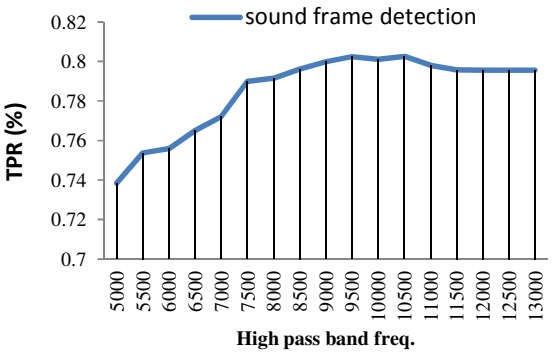
۳-۱- بررسی اثرات نویز و حذف آن

بدون داشتن اطلاعات اولیه از ساختار نویز استفاده از روش‌های برآورد نویزهای ضربه‌ای و حذف آن‌ها با الگوریتم وقتی کار ساده و موثری نخواهد بود [۲۳]. همچنین بهبود کیفیت سیگنال از طریق روش‌های زیرفضا که اخیرا برای نویزهای سفید پیشنهاد شده‌است [۲۴] نیازمند تجزیه سیگنال به مقادیر منفرد کسری ادراکی (PCQSVD) است که به پردازش‌های اولیه تخمین نویز وابسته است. با توجه به عدم وجود اطلاعات مذکور، در اینجا جهت بررسی نویز و همچنین اثرات حذف نویز بر روی پردازش‌های بعدی ابتدا از طریق یک آزمایش از فیلتر باترورث به‌عنوان پیش‌پردازش جهت حذف نویز استفاده گردید. با آزمایش‌های زیاد مقادیر ایده‌آل برای فرکانس‌های پایین و بالای فیلتر میان‌گذر و همچنین مرتبه فیلتر تعیین شد. نتایج بدست‌آمده نشان داد که استفاده از فیلتر میان‌گذر باترورث مرتبه ۷ با فرکانس پایین

۷۰۰ هرتز و فرکانس بالای ۱۰۵۰۰ هرتز بیشترین اثر مثبت را در شناسایی فریم‌های صدا از فریم‌های غیرصدا دارد. نتایج این آزمایش در شکل‌های ۳ و ۴ بر حسب میزان True-positive rate (TPR) برای تعیین وضعیت صدا/غیرصدا در فریم‌های آزمایش آورده شده‌است. لذا در عملیات پیش‌پردازش از فیلتر باتروورث میان‌گذر مرتبه ۷ با فرکانس پایین ۷۰۰ هرتز و فرکانس قطع بالای ۱۰۵۰۰ هرتز استفاده شد.



شکل ۳: پاسخ فیلتر باتروورث مرتبه ۷. تعیین فرکانس پایین جهت بیشترین بازدهی در شناسایی فریم‌های صدا. در فرکانس ۷۰۰ هرتز نرخ شناسایی درست فریم‌ها حداکثر است.



شکل ۴: پاسخ فیلتر باتروورث مرتبه ۷. تعیین فرکانس بالا جهت بیشترین بازدهی در شناسایی فریم‌های صدا. در فرکانس ۱۰۵۰۰ هرتز نرخ شناسایی درست فریم‌ها حداکثر است.

نویز موجود در خارج از این پهنای باند تاثیر منفی بر شناسایی صداهای پایگاه داده داشت.

۳-۲- ویژگی عبور از صفر

مقدار ویژگی عبور از صفر به صورت رابطه (۱) قابل محاسبه است:

$$ZCR = \sum_{n=1}^{N-1} |\mathrm{sgn}(x_n) - \mathrm{sgn}(x_{n+1})| \qquad (1)$$

متغیرهای x_n و x_{n+1} نمونه‌های پشت‌سرهم در سیگنال ورودی هستند. تابع sgn برای مقادیر مثبت عدد ۱ و برای مقادیر منفی عدد ۱- تولید می‌کند. با بررسی صداهای مورد پردازش مشخص شد که مقدار این ویژگی در صداهای خاصی ازجمله *switch, printer, keys, page - turn, alert, keyboard* دارای مقدار بالا با تفاوت معنادار نسبت به سایر صداها می‌باشد و همچنین مشخص شد در بعضی از صداها از جمله *keys, cough, door slam, clear throat* مقدار این ویژگی در فریم‌های ابتدایی زیاد است و بالا بودن این ویژگی می‌تواند برای تشخیص ابتدای صداها موثر باشد. علاوه‌براین‌ها مشخص شد که فریم‌های انتهایی برخی از صداها از جمله *Page - turn, phone, cough, laughter* دارای مقدار عبور از صفر بالا هستند و به خوبی قابل تشخیص می‌باشند. در شکل ۵ نموداری از مقادیر عبور از صفر ۹۵۰ فریم از صداهای مختلف آمده است. همانطور که مشخص است مقدار این ویژگی در تعداد زیادی از فریم‌ها دارای مقدار بالا و قابل تفکیک است. در فریم‌های ۱۰۴-۹۳ یک اتفاق صوتی مرکب قابل تشخیص است و همچنین در فریم‌های ۶۶۵-۶۲۳ دو واقعه صوتی مجزا قابل تفکیک و شناسایی هستند. مناسب بودن این ویژگی با رسم نمودار راک[۶] بررسی شد و مشخص گردید این ویژگی قابلیت تفکیک و شناسایی صداهای مورد نظر را تاحدزیادی دارد. راک نموداری است که می‌تواند میزان کارایی یک رده‌بند دوگانه را با تغییر مقدار آستانه نمایش دهد[۲۵].

۳-۳- ویژگی نسبت بالای نرخ عبور از صفر

یک ویژگی مهم تغییرات مقادیر عبور از صفر یا $HZCRR$ در تعداد N فریم در یک پنجره ۱ ثانیه‌ای است که در رابطه (۲) آمده‌است.

$$HZCRR = \frac{1}{2N} \sum_{n=0}^{N-1} [\mathrm{sgn}(Zcr(n) - 1.5 * avZcr) + 1] \qquad (2)$$

متغیر n اندیس فریم‌ها و $avZcr$ میزان میانگین عبور از صفر در طول یک پنجره ۱ ثانیه‌ای است و N نیز تعداد کل فریم‌های درون پنجره پردازش است. مقدار ویژگی $HZCRR$ همواره عددی از صفر تا حداکثر ۱ است. هر چه تعداد فریم‌های با مقدار عبور از صفر بالا بیشتر باشند مقدار این ویژگی به عدد ۱ نزدیکتر می‌شود.

طبق بررسی اکثر صداهای محیطی دارای تغییرپذیری بالا و متغیر در ویژگی عبور از صفر هستند که می‌توانند فریم‌های یک صدای مشخص را از سایر صداها قابل تشخیص نماید. همچنین مشاهده شد که مقادیر این ویژگی در سیگنال گفتار عموما از یک خط پایه و حداقلی برخوردار است که تقریبا نزدیک به عدد ۵۰ است.

در مورد صداهای دیگر تغییرها یا به طور یکنواخت سریع افزایش یافته و سپس با همان شیب کم می‌شوند، مانند $Page - turn$، یا اینکه حالت پریودیک با نقطه اوج و یک نقطه سقوط دارند و قبل از هر اوج یک سری مقادیر با تغییرهای کم تکرار می‌شوند و واریانس کمی دارند. به‌علاوه دامنه مقادیر خیلی نزدیک به عدد مشخصی بین ۲۵۰ تا ۳۰۰ است که برای مثال صدای $Phone$ از این گروه هست. در مورد صدای $printer$ هم مقادیر کاملا بدون نظم و ساختار بوده و تصادفی می‌باشند. لذا وجود چنین خواصی در ویژگی‌های یک سری از صداها امکان شناسایی آن‌ها را ممکن می‌سازد.

۳-۴- ویژگی انرژی فریم‌های صوتی

طبق آزمایش‌ها مشخص شد که این ویژگی در شرایط محیط نویزی چندان ممیز نیست اما پس از حذف نویز و استفاده از روش‌های پیش‌پردازش مشاهده شد که انرژی فریم‌ها می‌تواند اطلاعات مفیدی در مورد صداهای خاص تولید کند. صداهای کوبه‌ای مانند *door - knock, door - slam, drawer, pendrop* و یا صداهای دماغی و صامت در گفتار دارای میزان انرژی بالایی بودند. جهت نمونه در ۱۰ سیگنال آزمایش تعداد خطای شناسایی فریم‌های صداها قبل از پیش‌پردازش بطور متوسط در حدود $FPR = ۰.۵۸$ بود. اما پس از انجام مراحل حذف نویز در همان نمونه‌ها این میزان به $FPR = ۰.۴۶$ کاهش یافت که بیانگر قابلیت شناسایی صداها توسط این ویژگی است.

۳-۵- ویژگی نسبت پایین انرژی کوتاه‌مدت

یکی از ویژگی‌های سیگنال در حوزه زمان تغییرات انرژی در فریم‌های مجاور به‌طول ۱ ثانیه است که آنرا نسبت پایین انرژی کوتاه‌مدت[۷] (LSTER) می‌نامند. این ویژگی طبق رابطه (۳) تعریف می‌شود.

$$LSTER = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(0.5avSTE - STE(n)+1] \qquad (۳)$$

متغیر n اندیس فریم است، $avSTE$ میانگین انرژی در فریم‌های مجاور به طول ۱ ثانیه است. در اینجا هر ۱ ثانیه شامل ۸ فریم ۱۲۵ میلی‌ثانیه است. تعداد کل فریم‌ها N است. در اینجا این ویژگی برای شناسایی صداهایی همچون *door - knock, drawer, pen - drop* استفاده شده‌است. مقادیر انرژی فریم‌ها در این صداها علاوه‌بر بالا بودن، تغییرهای درونی زیادی نیز دارند. این تغییرها سبب می‌شود نسبت فریم‌های دارای انرژی کمتر از متوسط زیاد باشد و به‌همین دلیل صداهایی که مرتبا انرژی آن‌ها کم و زیاد می‌شود مقدار این ویژگی برای آن‌ها زیاد است و تاحدزیادی قابل شناسایی هستند.

۴- پیاده‌سازی مدل برحسب الگوریتم پیشنهادی

در فاز قطعه‌بندی بردارهای ویژگی به دو گروه تقسیم می‌شوند. گروه نخست شامل ویژگی‌های *(Zc, HZCCR)* (عبور از صفر و تغییرات آن). گروه دوم ویژگی‌های *(STE, LSTER)* (انرژی و تغییرات آن) است. زیرسیستم اول از ویژگی‌های گروه اول استفاده می‌کند. این ویژگی‌ها با انتخاب مناسب طول bin ها به اندازه ۵ میلی‌ثانیه امکان شناسایی نقاط خیزش را در صداهای مختلفی در سیگنال میسر می‌سازد. قطعه‌بندی در این زیرسیستم دارای پارامتر α است که امکان یک تصمیم سخت یا تصمیم نرم قابل انعطاف را برای انتخاب نقاط کاندید خیزش مهیا می‌سازد. پارامتر α در اینجا عدد صحیح است که هر چه آلفا بزرگتر باشد، میزان دقت کشف وقایع بیشتر است.

زیرسیستم دوم از ویژگی‌های گروه دوم استفاده می‌کند. طول فریم در اینجا ۱۲۵ میلی‌ثانیه تعیین شده‌است. این ویژگی‌ها همراه با پارامتر β که عدد اعشاری است، انتخاب نقاط خیزش با انعطاف یا سخت‌گیرانه جهت قطعه‌بندی سیگنال را مهیا می‌سازند. نقاط کاندید برای خیزش معمولا از یک انرژی بیشتر و ناگهانی نسبت به فریم‌های مجاور در طول ۱ ثانیه برخوردار می‌باشند.

الگوریتم طبق شکل ۶ اجرا شده و نتیجه اجرا طبق رابطه (۴) است.

$$X۱(n) = X(n) - Xe(n) \qquad (۴)$$

در رابطه (۴)، $X(n)$ سیگنال اولیه است. $Xe(n)$ بخش‌هایی از سیگنال اولیه است که به‌عنوان قطعات اولیه وقایع صوتی بدست آمده‌اند و $X۱(n)$ سیگنال باقیمانده است. پس از اتمام الگوریتم، قطعه‌های یافته‌شده طبق رابطه (۵) تجمیع شده و در فاز نهایی ترکیب می‌شوند.

$$Xe(n) = Xe۱(n) + Xe۲(n) \qquad (۵)$$

متغیر $Xe(n)$ خروجی دو زیرسیستم قطعه‌بندی در شکل ۱ است.

آنچه در فاز نهایی قطعه‌بندی انجام می‌شود این است که خروجی هر دو زیرسیستم را ترکیب می‌نماید و وقایع نهایی را پس از اصلاح زمان‌های شروع و پایان استخراج می‌کند. در اصلاح زمانی، فریمی که زمان شروع زودتری دارد زمان شروع اتفاق صوتی و فریمی که زمان پایان دیرتری دارد زمان خاتمه واقعه را مشخص می‌کند.

۴-۱- عملیات قطعه‌بندی

جهت قطعه‌بندی، برای هر کدام از دو زیرسیستم یک مقدار آستانه عمومی ثابت و جداگانه تعریف شده‌است. شبه کدهای (۶) و (۷) نقش هر یک از پارامترهای آلفا و بتا را در انتخاب فریم در زیرسیستم‌ها نشان می‌دهند.

$$\begin{aligned} &IF\ (\ ZC_k >= (\text{norm\_zc} + \alpha)\ .OR.\ HZCRR_k >= ۰.۵۰) \\ &select\ Frame\ k\ as\ a\ candidate\ frame\ being\ in\ an\ event. \\ &Where\ (\text{norm\_zc and } \alpha\ \ have\ a\ predefined\ values). \end{aligned} \qquad (۶)$$

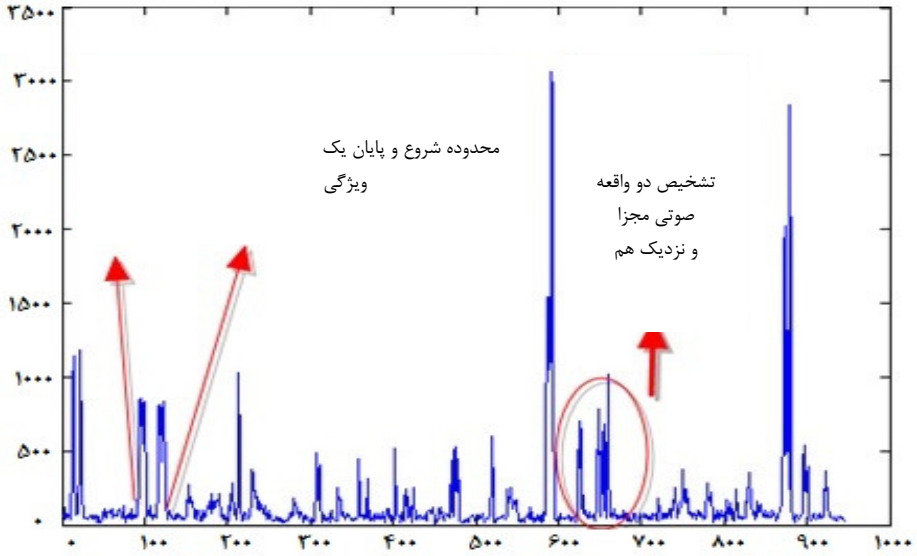
$$\begin{aligned} &IF\left(STE_k >= \ (norm\_En * \beta)\ .OR.\ LSTER_k <= \ ۰.۵۰\ \right) \\ &select\ Frame\ k\ as\ a\ candidate\ frame\ being\ in\ an\ event \\ &Where\ (norm\_En \text{ and } \beta\ have\ a\ predefined\ values). \end{aligned} \qquad (۷)$$

در شبه‌کدها، k اندیس فریم است و $norm_{zc}$ و $norm_{en}$ به ترتیب متعلق به ویژگی عبور از صفر و ویژگی انرژی هستند.

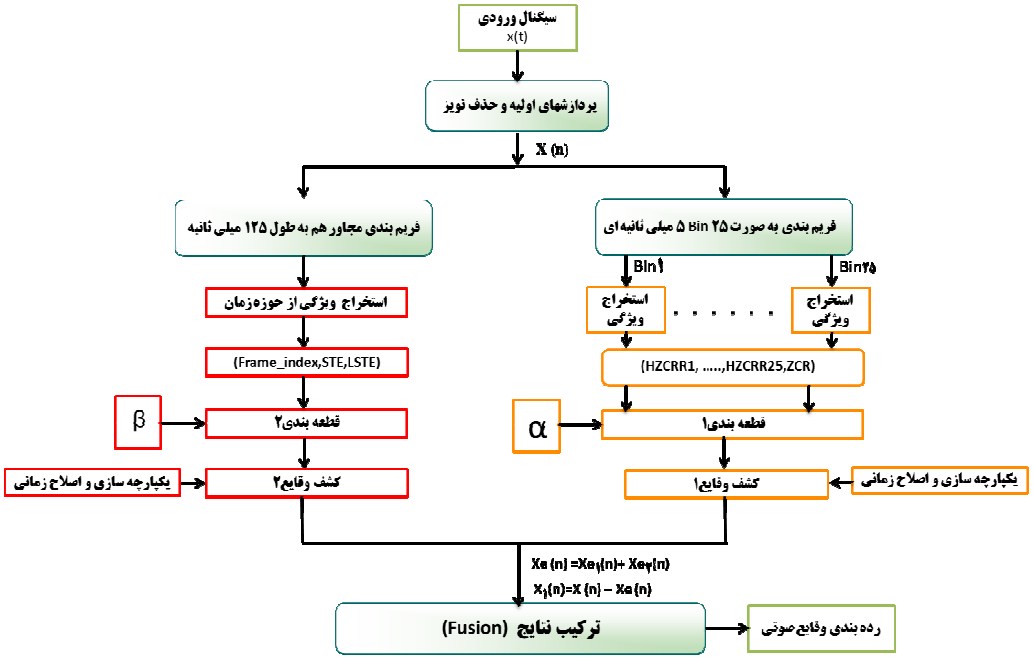
در زیرسیستم اول به دلیل وجود نویز در فریم‌ها و مشابهت صداهای خاصی مثل *drawer, phone, printer* به فریم‌های نویز، از ویژگی میزان عبور از صفر به همراه تغییرات عبور از صفر استفاده شده‌است. تمام فریم‌هایی که مقدار عبور از صفر آن‌ها بالاتر از $norm + \alpha$ است و یا اینکه میزان تغییرات عبور از صفر در طی ۱ ثانیه بیش از ۰/۵ است، به‌عنوان کاندید در یک واقعه صوتی انتخاب می‌شوند. نقش پارامتر α کلیدی است به‌طوری‌که با تغییر α عملا می‌توان نقش ویژگی عبور از صفر را کم یا زیاد کرد و در مورد صداهای متفاوت می‌توان مقدار مختلفی را برای α بدست آورد.

در زیرسیستم دوم از ویژگی انرژی و تغییرات یک ثانیه‌ای آن به‌صورت شناسایی نقاط خیزش و افول استفاده شده‌است. فریم‌هایی که پس از یک دوره پایداری در انرژی سیگنال به یکباره حداقل به اندازه $\beta * norm$ افزایش انرژی دارند، به‌عنوان نقاط خیزش صداهای محیطی شناسایی می‌شوند و نقاطی که این حداقل انرژی را پس از خیزش اولیه حفظ کنند، در محدوده صدا در نظر گرفته می‌شوند. پس از سپری شدن نقطه خیزش و نوسان‌های آن مجددا انرژی سیگنال به حالت پایدار و با تغییرات کمتر از $\beta * norm$ می‌رسد که این نقطه شروع پایداری مجدد، نقطه افول و عبور از صدای محیطی فرض می‌شود. پارامتر β می‌تواند میزان فریم‌هایی که در صدای اکتشافی کاندید می‌شوند را کم یا زیاد کند به‌طوری که با زیاد شدن β انتخاب سخت‌تر می‌شود.

فاز نهایی ترکیب نتایج دو زیرسیستم رده‌بندی است. در مرحله ترکیب هر گاه واقعه‌ای توسط یکی از زیرسیستم‌ها یا هر دو زیرسیستم کشف شده‌باشد در لیست نهایی وقایع کشف‌شده آورده می‌شود.



شکل ۵: نمودار بررسی قابلیت ویژگی عبور از صفر. محور افقی اندیس فریم‌ها و محور عمودی مقدار این ویژگی در فریم را نشان می‌دهد. مقدار این ویژگی در ۹۵۰ فریم نشان می‌دهد که نقاط شروع و پایان برخی از وقایع قابل شناسایی هستند.



شکل ۶: بلاک دیاگرام مراحل مختلف استخراج ویژگی، قطعه‌بندی و کشف وقایع صوتی محیطی در مدل پیشنهادی.

۴-۲- عملیات رده‌بندی

ابتدا تعداد ۶۱ ویژگی از فریم‌های وقایع اکتشافی استخراج می‌شود. همین ویژگی‌ها در مورد داده‌های آموزشی برای هر واقعه صوتی نیز بصورت مجزا استخراج می‌شود و از آن‌ها بردارهای نمونه هر صدا ایجاد می‌گردد. ویژگی‌های استفاده‌شده در این مرحله و تعداد هر کدام عبارتند از: عبور از صفر (۱)، انرژی کوتاه‌مدت (۱)، مرکز ثقل فرکانسی (۱)، بهنای باند فرکانسی (۱)، لگاریتم انرژی بانک فیلترهای فرکانسی مل (۱۶)، ضرایب دلتا و دلتای دوم بانک فیلترهای مل (۳۲)، انرژی زیرباندهای چهارگانه (۴)، اسپکترال فلکس هر زیرباند (۴) و انتروپی یا پیچیدگی درون فریم (۱).

در انتخاب ویژگی‌ها دو جهت کاملا مجزا و مکمل سیگنال یعنی حوزه ادراکی و حوزه تبدیل طیف لحاظ گردید. این ویژگی‌ها پیش از این در کارهای پژوهشی متعدد استفاده شده‌است و مستقل هستند [۲۹-۲۶]. تعداد ۱۶ ویژگی تبدیل طیف که از لگاریتم بانک‌های فیلتر مل گرفته شده‌است، برمبنای درک فرکانس توسط سیستم شنوایی انسان است. مشتقات اول و دوم زمانی آن‌ها تکمیل‌کننده اطلاعات دینامیکی در گذر زمان هستند که در کنار ویژگی‌های قبلی مکمل فرض می‌شوند. به‌علاوه ویژگی‌های دیگری همچون میزان عبور از صفر یا انرژی نیز توصیف‌کننده رفتار زمانی و مکمل حوزه فرکانس هستند. دلیل انتخاب ما علاوه بر مکمل بودن این ویژگی‌ها و مستقل از هم بودن آن‌ها، وجود صداهای متنوع از جمله گفتار، خنده، سرفه ازیک‌سو و ازسوی‌دیگر صدای تلفن، چاپگر، کلید، آلارم و غیره است. گروه اول وکال یا گلوتال (حنجره‌ای) است و گروه دوم (یعنی صداهای محیطی) بیشتر بدون‌نظم و شکل‌خاص است که باید ازنظر زمانی و ازنظر فرکانسی مدل شوند.

طول فریم‌ها ۲۵ میلی‌ثانیه و میزان هم‌پوشانی آن‌ها ۶۰ درصد انتخاب شد. نمایش فرکانسی هر فریم m، $X_k(m)$ نامیده می‌شود. با استفاده از تبدیل گسسته فوریه و اعمال پنجره هنینگ $W_b(n)$ طبق رابطه‌های (۸)، (۹) انجام شده‌است.

$$X_k(m)=\sum_{n=-\infty}^{\infty} x(n).w_b(n-m).e^{\frac{-j2\pi kn}{N}} \quad ,0\le k\le N-1 \qquad (۸)$$

$$w_b(n)=\begin{Bmatrix} 0.5-0.5.\cos(\frac{2\pi n}{b}) & ,0\le n\le b \\ 0 & otherwise \end{Bmatrix} \qquad (۹)$$

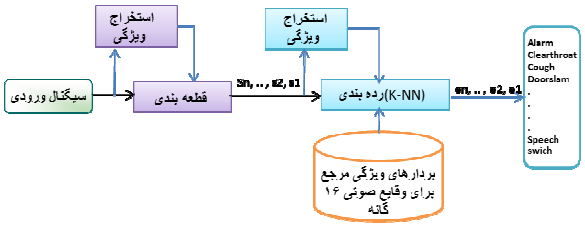
در رابطه‌های (۸)، (۹) متغیرهای N, n, k, b به‌ترتیب از چپ‌به‌راست طول DFT، اندیس حوزه‌زمان، اندیس حوزه‌فرکانس و طول پنجره‌هنینگ می‌باشند. محاسبه مرکزثقل‌فرکانسی، اسپکترال‌فلکس و انتروپی در رابطه‌های (۱۰) تا (۱۲) بر اساس مراجع [۲۶، ۲۷] آورده شده‌است.

$$Centroid(m)=\frac{\sum_{k=0}^{N-1}(k+1).|X_k(m)|}{\sum_{k=0}^{N-1}|X_k(m)|} \qquad (۱۰)$$

$$Flux(m)=\sum_{k=0}^{N-1}(|X_k(m)|-|X_k(m-1)|)^2 \qquad (۱۱)$$

$$Entropy(m)=-\sum_{k=0}^{N-1}|X_k(m)|^2.\log_2(|X_k(m)|^2) \qquad (۱۲)$$

پس از فریم‌بندی و استخراج ویژگی‌ها از وقایع کشف‌شده تمام بردارهای ویژگی استخراجی از فریم‌ها نرمال‌سازی شده و سپس از الگوریتم k-nearest neighbors algorithm (k-NN) همانند شکل ۷ برای رده‌بندی هر فریم استفاده می‌شود. پس از رده بندی فریم های هر سگمنت، یک تصمیم‌گیری نهایی بر پایه رای‌گیری و کسب بیشترین رای برای تعیین کلاس مربوط به واقعه‌صوتی انجام می‌شود. برای تخمین دقیق‌تر مقدار K آزمایش‌های متعدد بر روی مقادیر ۱۱، ۹، ۷، ۵، ۳، ۱ = K انجام شد که مقدار K=۷ بهترین‌نتیجه را داشت. جهت فاصله‌یابی ابتدا روش‌های اقلیدسی، فاصله‌کسینوسی، مینکواسکی و چای‌اسکویر بررسی و محاسبه شدند که نتایج حاصل از روش چای‌اسکویر بهتر از روش‌های دیگر بود و انتخاب گردید.



شکل ۷: پس از قطعه‌بندی از الگوریتم K-NN با مقدار K=۷ برای رده‌بندی استفاده شده‌است.

۵- اجرای مدل و نتایج

در این مرحله مدل پیشنهادی در قالب الگوریتم شکل ۱ بصورت مکرر آزمایش شد و نتایج جمع‌آوری گردیدند. در آزمایش‌ها از پایگاه داده تهیه‌شده در انجمن بین‌المللی تخصصی پردازش سیگنال صوت[۸] استفاده شده‌است. این داده‌ها مرتبط با صحنه‌های شنیداری مختلف ازجمله صداهای موجود در اتاق کار یا جلسه است که حاوی ۱۶ نوع اتفاق صوتی مطابق جدول ۱ است. طول زمانی هر یک از رکوردهای مورد پردازش ۳ دقیقه و به طور متوسط هر رکورد شامل ۳۵ اتفاق صوتی می‌باشد. رکوردها در شرایط وجود نویز محیطی متغیر ضبط شده‌اند، لذا صداها دارای شرایط محیط زنده هستند و صداهای مصنوعی[۹] محسوب نمی‌شوند. در مجموع از تعداد ۶۰ رکورد مختلف استفاده شد و جهت اجرای سریع‌تر آزمایش‌ها هر بار یک گروه ۱۰ رکوردی برای آزمایش انتخاب و نتایج آن گروه یادداشت شد. انتخاب رکوردهای هر گروه شامل انتخاب ۱۰ رکورد به صورت تصادفی و غیرتکراری از بین ۶۰ رکورد موجود می‌باشد. لذا در آزمایش‌ها نتایج گروه‌های ۱ تا ۶ را جداگانه بدست‌آورده و نهایتا میانگین گروه‌ها

به‌عنوان نتیجه کلی آزمایش ۶۰ رکورد درنظر گرفته‌شد. الگوریتم با معیارهای دقت، میزان فراخوانی و میانگین هارمونیک وزنی[10] (معیار کارایی) مورد سنجش قرار گرفته‌است. این معیارها با توجه به نیاز الگوریتم در رابطه‌های (۱۳) تا (۱۵) معرفی شده‌اند.

$$\text{Precision}=\frac{\text{تعداد کل وقایع صوتی که الگوریتم بدرستی آنها را شناسایی کرده است}}{\text{تعداد کل وقایعی که الگوریتم شناسایی کرده است}} \quad (۱۳)$$

$$\text{Recall}=\frac{\text{تعداد کل وقایع صوتی که الگوریتم بدرستی آنها را شناسایی کرده است}}{\text{تعداد واقعی وقایع در رکورد صوتی ورودی}} \quad (۱۴)$$

$$\text{F1-measure}=\frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (۱۵)$$

در تمام آزمایش‌های انجام‌شده بهترین نتایج مربوط به رکوردها در ارزیابی در نظر گرفته شده‌اند.

در اولین آزمایش میزان عملکرد الگوریتم بدون در نظر گرفتن پارامترهای آلفا و بتا که معادل $\alpha=۰$ و $\beta=۱/۰$ است بررسی گردید که نتایج آن در جدول ۲ ارائه شده‌است. میانگین میزان دقت ۴۴/۹ درصد و میانگین فراخوانی ۴۷/۶ درصد و میزان کارایی میانگین در بین رکوردهای آزمایش‌شده که آنرا F1 بر اساس کلیپ یا رکورد record-based (RB) نامیده و با F1(RB) نشان می‌دهیم ۴۶/۱ درصد است. این شرایط اولیه الگوریتم است و می‌توان با دخالت دادن پارامترها عملکرد سیستم را بهبود داد. در آزمایش‌های بعدی با استفاده از پارامترهای مذکور، عملکرد سیستم به طور محسوسی افزایش داشته است و این بیانگر دینامیک عمل کردن سیستم با تغییر پارامترها است. همچنین می‌توان گفت که بهبود بیشتر میزان دقت، بیانگر تاثیر بیشتر و مستقیم پارامترها بر کم شدن خطای کشف است.

جدول ۲: عملکرد الگوریتم بدون اعمال پارامترهای α و β (هر Group شامل ۱۰ رکورد از بین ۶۰ رکورد آزمایش و بدون تکرار است)

Group	F1(RB)	Recall	Precision
۱	۰/۴۵۸	۰/۵۰۱	۰/۴۲۲
۲	۰/۴۷۳	۰/۵۱۴	۰/۴۳۸
۳	۰/۵۱۸	۰/۵۴۳	۰/۴۹۵
۴	۰/۴۵۷	۰/۴۷۱	۰/۴۴۳
۵	۰/۴۱۱	۰/۳۹۸	۰/۴۲۵
۶	۰/۴۴۸	۰/۴۲۶	۰/۴۷۳
میانگین	۰/۴۶۱	۰/۴۷۶	۰/۴۴۹

جهت بررسی تاثیر پارامترها، ابتدا ۱۰ گروه ۶ رکوردی به‌طور تصادفی انتخاب شد. با هر گروه سه اجرا به‌صورت عدم استفاده از پارامترها، فقط پارامتر آلفا و فقط پارامتر بتا انجام و میانگین هر اجرا گرفته شد. سپس درصد معیار F1 از میانگین محاسبه و نتایج در شکل ۸ آورده شد. همانطور که مشخص است میزان شناسایی در اجراهای دوم و سوم در تمام گروه‌ها بالاتر از اجرای اول در همان گروه است. در برخی اجراها از جمله ۲، ۴، ۷، ۸ تاثیر آلفا به تنهایی بیشتر از بتا بوده و در اجراهای دیگر این نتیجه معکوس است. نتایج این آزمایش به وضوح موثر بودن هر دو پارامتر آلفا و بتا را مستقل از هم تایید می‌کند.

در ادامه آزمایش‌هایی جهت تعیین بهترین آلفا و بتا برای حداکثرکردن معیارهای سه‌گانه انجام شد. نتایج حداکثرسازی میزان دقت در جدول ۳ آمده است. ستون‌های α و β مقادیری را نشان می‌دهند که به ازای آن‌ها میزان دقت در رکوردهای ورودی حداکثر است. به‌طور میانگین، برای پارامترهای $\alpha=۵$ و $\beta=۱/۴۱$ حداکثر دقت ۶۶/۲ درصد است.

جدول ۳: حداکثرسازی میزان دقت بر حسب انتخاب مقدار بهینه برای پارامترهای α و β در سیستم

Group	β	α	F1(RB)	Recall	Precision:max
۱	۱/۵۳	۵	۰/۵۰۴	۰/۴۰۴	۰/۶۷۱
۲	۱/۴۰	۴	۰/۵۵۲	۰/۴۵۵	۰/۷۰۲
۳	۱/۲۴	۵	۰/۴۴۸	۰/۳۶۳	۰/۵۸۶
۴	۱/۴۴	۶	۰/۴۹۶	۰/۳۷۹	۰/۷۱۶
۵	۱/۴۸	۳	۰/۴۳۱	۰/۳۳۴	۰/۶۰۸
۶	۱/۳۵	۷	۰/۵۳۰	۰/۴۳۰	۰/۶۸۹
میانگین	۱/۴۱	۵	۰/۴۹۴	۰/۳۹۴	۰/۶۶۲

نتایج حداکثرسازی تعداد وقایع کشف‌شده (میزان فراخوانی) در جدول ۴ آمده‌است. مقدار بهینه پارامترها $\alpha=۵$ و $\beta=۱/۱۶$ است. اجرا با مقادیر میانگین پارامترها از جدول ۴ میزان فراخوانی را حداکثر به ۵۸/۴ درصد رسانید.

جدول ۴: مقدار بهینه پارامترهای α و β در حداکثرسازی میزان فراخوانی

Group	β	α	F1(RB)	Recall:max	Precision
۱	۱/۱۳	۶	۰/۶۰۱	۰/۶۴۴	۰/۵۶۳
۲	۱/۱۵	۵	۰/۵۶۵	۰/۵۹۷	۰/۵۳۶
۳	۱/۱۲	۷	۰/۵۳۰	۰/۵۵۱	۰/۵۱۰
۴	۱/۲۲	۶	۰/۶۱۶	۰/۶۰۴	۰/۶۲۸
۵	۱/۱۳	۵	۰/۴۸۸	۰/۵۰۵	۰/۴۷۲
۶	۱/۲۰	۳	۰/۵۶۱	۰/۵۷۳	۰/۵۵۰
میانگین	۱/۱۶	۵	۰/۵۶۰	۰/۵۷۹	۰/۵۴۳

در جدول ۵ مقادیر معیارها بر اساس پارامترهای بهینه α و β به صورت مقایسه‌ای آورده شده‌است.

جدول ۵: مقایسه مقادیر معیارها بر اساس پارامترهای بهینه

	β	α	F1(RB)	Recall	Precision
	۱/۴۱	۵	۰/۵۰۱	۰/۴۰۳	**۰/۶۶۲**
	۱/۱۶	۵	۰/۵۶۳	**۰/۵۸۴**	۰/۵۴۴
	۱/۲۱	۶	**۰/۵۷۳**	۰/۵۳۹	۰/۶۱۲
میانگین	۱/۲۶	۵	۰/۵۴۶	۰/۵۰۹	۰/۶۰۶

جهت مقایسه نتایج با کارهای مشابه ابتدا نتایج معیار F1 که در جداول ۵-۱ برحسب کلیپ (یا رکورد) و میانگین بین رکوردها محاسبه شده و record-based (RB) می‌باشد به‌صورت یکپارچه براساس فریم Frame-based (FB) و سپس براساس واقعه event-based (EB) محاسبه می‌شود. نتایج پردازش فریمی و برحسب واقعه در جدول ۶

آورده شده‌است. بهبودی نتایج در سطح فریمی (FB) بدلیل تشخیص بهتر اطلاعات در فریم نسبت به سطح بالاتر واقعه (EB) است و طبیعاتا شناسایی یک واقعه درحد چندثانیه که متشکل از تعداد زیادی فریم است کار مشکل‌تر و همراه با خطای بیشتر است. تفاوت نتایج سطح واقعه با سطح رکورد یا کلیپ نیز ناشی از تفاوت تعداد کلی وقایع در بین رکوردها و همچنین چگالی متفاوت برای هر واقعه در رکورد و میانگین‌گیری صورت‌گرفته در بین رکوردهای هر گروه است که نهایتا لحاظ نمودن نتایج در رکوردها به‌طور مستقل و میانگین بین آن‌ها، سبب نتایج ضعیف‌تر RB نسبت به EB گردیده‌است. اما آنچه در نتایج مورد قیاس قرار می‌گیرد میزان معیار F1 بر اساس FB, EB است.

جدول ۶: مقادیر معیار ارزیابی F1 بر اساس واقعه (EB) و فریم (FB) بر روی مجموعه داده‌های آزمایش.

Maximization	F1(EB)	F1(FB)
precision	۰/۷۰۱	۰/۷۷۳
recall	۰/۷۵۸	۰/۸۳۷
F1-measure	۰/۷۲۸	۰/۸۰۶

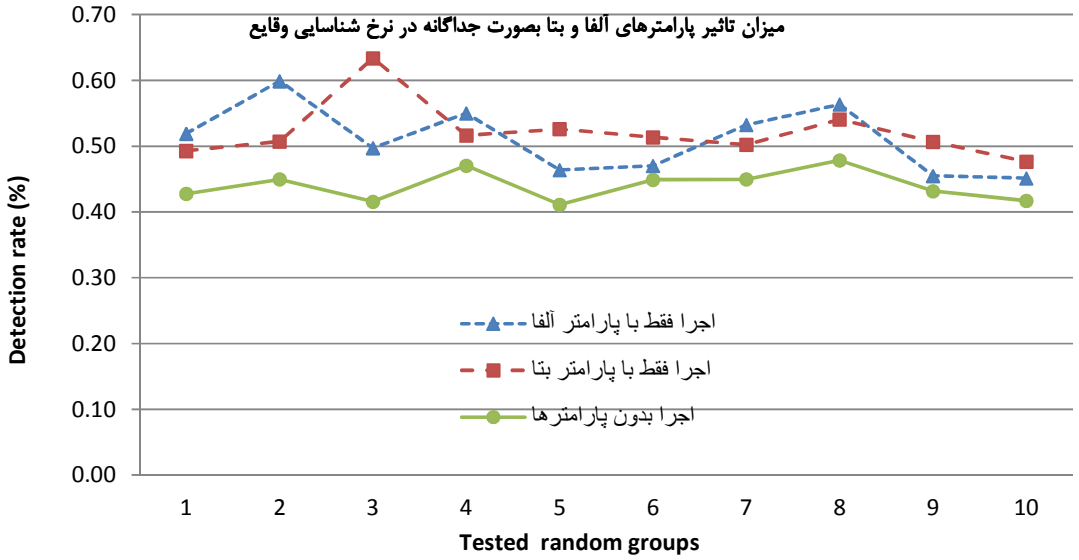
مقایسه نتایج با پژوهش های مرتبط در شکل ۹ آمده است. نتایج مشخص می‌کند که مدل پیشنهادی بیش از ۱۰/۸% بهبودی داشته‌است و دلیل آن مدل‌سازی بهتر نسبت به قبل است.

۶- نتیجه‌گیری

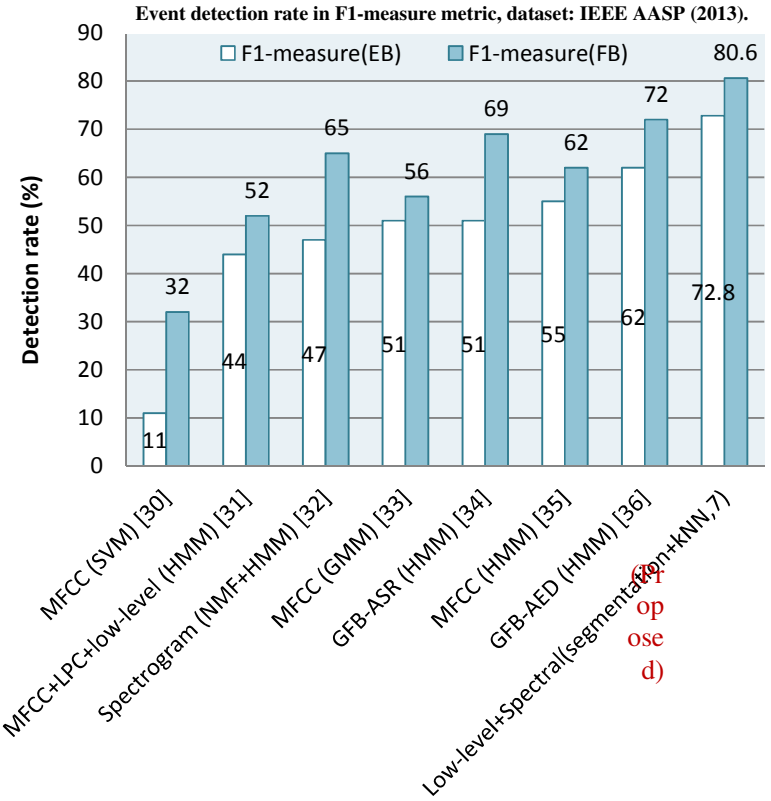
در این پژوهش یک روش جدید مدل‌سازی تطبیق‌پذیر با دو پارامتر جهت کشف وقایع صوتی در سیگنال‌های محیطی ارائه شد. مدل از دو پارامتر کلیدی α و β استفاده می‌کند که تاحدزیادی سبب بهبود عملکرد اولیه سیستم می‌گردد و همچنین انعطاف‌پذیری و قابلیت تنظیم سیستم بین دقت بالاتر یا فراخوانی بالاتر را ممکن می‌سازد. نتایج دیگر تحقیق این است که استفاده از دو زیرسیستم قطعه‌بندی و کشف وقایع و ترکیب نتایج سبب استفاده هر چه بهتر از ویژگی‌ها شده و امکان بالارفتن قابلیت جداسازی ویژگی‌ها و کشف بیشتر وقایع را میسر می‌سازد. به‌علاوه در آزمایش‌ها مشخص شد که تفکیک ویژگی‌ها به دو گروه، شناسایی صداهای خاص توسط ویژگی خاص را عملی می‌سازد. نتایج بدست‌آمده تاثیر قابل ملاحظه پارامترها در مقایسه با حالت اولیه مدل را به‌خوبی نشان می‌دهد. در آزمایش‌ها مقادیر پارامترها برای حداکثر نمودن معیارهای دقت، فراخوانی و کارایی محاسبه و نتایج ارائه شد. مشخص شد، با تعیین مقادیر پارامترها معیارهای مورد نظر تا حد امکان بهینه و حداکثر می‌شوند. مشخص شد، ترکیب نتایج دو زیرسیستم و استفاده از روش‌های یکپارچه‌سازی و اصلاح نهایی محدوده وقایع در میزان دقت نهایی اثر دارد.

۷- پیشنهادها جهت کارهای آینده

یکی از معضلات اصلی وجود نویز بالا و متغیر در رکوردهای صوتی است. روش‌های بهتری برای مقاوم نمودن الگوریتم در برابر نویز پیشنهاد می‌شود. بعضی از صداهای محیطی شبیه نویز زمینه هستند لذا معرفی ویژگی‌های زمانی که صداهای نویز-مانند را دقیق‌تر کشف نماید پیشنهاد می‌شود. جهت مقابله با نویز در داده‌ها، مدل‌سازی نویز محیط به‌صورت یک سطح مجزا پیشنهاد می‌شود. استفاده از ویژگی‌های حوزه‌فرکانس و ادراکی در یک زیرسیستم و ویژگی‌های حوزه‌زمان در زیرسیستم دیگر در قطعه‌بندی پیشنهاد می‌شود.



شکل ۸: بررسی عملکرد پارامترهای آلفا و بتا به صورت جداگانه در شناسایی وقایع. اجرای بدون پارامترها کمترین میزان شناسایی وقایع را دارد. با اعمال هر یک از پارامترها دیده می‌شود که نرخ شناسایی بطور قابل ملاحظه‌ای افزایش دارد. در بعضی موارد پارامتر آلفا و در بعضی دیگر پارامتر بتا تاثیر بیشتری در افزایش نرخ شناسایی دارد.



شکل ۹: درصد شناسایی وقایع برحسب معیار F1 در سیستم پیشنهادی (proposed) در دو مقیاس فریم (FB) و واقعه (EB). مجموعه داده‌های استفاده‌شده مربوط به IEEE AASP (2013) در شاخه AED است که شامل ۱۶ نوع صدای محیطی اداری می‌باشد.

مراجع

[1] F. Aurino, M. Folla, F. Gargiulo, V. Moscato, A. Picariello, and C. Sansone, "One-class SVM-based approach for detecting anomalous audio events," *International Conference on Intelligent Networking and Collaborative Systems*, Salerno, Italy, pp. 145-151, 2014.

[2] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," *Advanced Video and Signal Based Surveillance, 10th IEEE International Conference on*, Krakow, Poland, pp. 81-86, 2013.

[3] R. Maher, "Acoustical modeling of gunshots including directional information and reflections," in *131st Audio Engineering Society Convention,* New York, NY, 2011.

[4] R. Cai, L. Lu, and A. Hanjalic, "Co-clustering for Auditory Scene Categorization," in *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 170-177, 2008.

[5] Y. Ohishi, D. Mochihashi, T. Matsui, M. Nakano, H. Kameoka, T. Izumitani, and K. Kashino, "Bayesian semi-supervised audio event transcription based on markov indian buffet process," *IEEE (ICASSP)*, Vancouver, Canada, pp. 3163–3167, 2013.

[6] E. Benetos, G. Lafay, M. Lagrange, and M. Plumbley, "Detection of overlapping acoustic events using a temporally constrained probabilistic model," *IEEE (ICASSP)*, shanghai, china, pp. 6450–6454, 2016.

[7] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, " Detection and classification of acoustic scenes and events " *IEEE Transactions on Multimedia* vol. 17 no. 10 pp. 1733 – 1746, 2015.

[8] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011.

[9] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Interspeech*, Portland, Oregon, USA, 2012.

[10] A. Plinge, R. Grzeszick, and G. Fink, "A Bag-of-Features approach to acoustic event detection," in *IEEE (ICASSP),*Florence, Italy, May 2014.

[11] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Proc. CHiME*, Florence, Italy, pp. 36–40, 2011.

[12] R. Hennequin, R. Badeau and B. David, "NMF with Time–Frequency Activations to Model Nonstationary Audio Events," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744-753, 2011.

[13] T. Komatsu, Y. Senda, and R. Kondo, "Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation," *IEEE (ICASSP)*, shanghai, china, pp. 2259–2263, 2016.

[14] X. Lu, Y. Tsao, S. Matsuda and C. Hori, "Sparse representation based on a bag of spectral exemplars for acoustic event detection," *IEEE (ICASSP)*, Florence, Italy, pp. 6255-6259, 2014.

[15] IEEE DCASE 2016 Challenge, *http://www.cs.tut.fi/sgn/arg/dcase2016/*, 2016.

[16] I. Choi, K. Kwon, S. Hyun Bae, and N. Soo Kim, "DNN-based sound event detection with exemplar-based approach for noise

reduction," in *Proc. IEEE (DCASE)*, Budapest, Hungary, September 2016.

[17] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection," in *Proc. IEEE (DCASE)*, Budapest, Hungary, September 2016.

[18] J. Kurby, R. Grzeszick, A. Plinge, and G A. Fink, "Bag-of-features acoustic event detection for sensor networks," in *Proc. IEEE (DCASE)*, Budapest, Hungary, September 2016.

[19] M. Zohrer, and F. Pernkopf, "Gated recurrent networks applied to acoustic scene classification and acoustic event detection," in *Proc. IEEE (DCASE)*, Budapest, Hungary, September 2016.

[20] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[21] E. Miquel, F. Masakiyo, S. Daisuke, O. Nobutaka, and S. Shigeki, "A tandem connectionist model using combination of multi-scale spectro-temporal features for acoustic event detection". in *Proc. IEEE (ICASSP)*, Kyoto, Japan, pp. 4293–4296, 2012.

[22] L. Vuegen, B. Van Den Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, , and H. Van hamme, "An MFCC-GMM approach for event detection and classification," *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA),* New Paltz, NY, USA, Oct. ۲۰۱۳.

[۲۳] مجتبی حاجی آبادی، عباس ابراهیمی مقدم و حسین خوش‌بین، «حذف نویز مبتنی بر یک الگوریتم وفقی نوین،» مجله مهندسی برق دانشگاه تبریز، جلد ۴۶، شماره ۳، صفحات ۱۴۷-۱۳۹، ۱۳۹۵.

[۲۴] مسعود گراوانچی‌زاده و ساناز قائمی سردرودی، «بهبود کیفیت گفتار مبتنی بر بهینه‌سازی ازدحام ذرات با استفاده از ویژگی‌های ماسک‌گذاری سیستم شنوائی انسان»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۶، شماره ۳، صفحات ۲۹۷-۲۸۷، ۱۳۹۵.

[25] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Researchers," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 882–891, 2004.

[26] J. T. Geiger, B. Schuller, and G. Rigoll, "Recognizing acoustic scenes with large-scale audio feature extraction and SVM," *TUM, technical report*, 2013.

[27] D. Li, J. Tam, and D. Toub, "Auditory scene classification using machine learning techniques," *technical report*, 2013.

[28] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "HMM-based acoustic event detection with AdaBoost feature selection," in *Multimodal Technologies for Perception of Humans*, *Springer*-verlag Berlin, Heidelberg, pp. 345-353, 2008.

[29] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proceedings of the 18th European Signal Processing Conference, Eusipco 2010,* Aalborg, Denmark, pp. 1267-1271, August 2010.

[30] 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 Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[31] M. E. Niessen, T. L. M. V. Kasteren, and A. Merentitis, "Hierarchical modeling using automated sub-clustering for sound event recognition," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[32] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. V. hamme, "An exemplar-based NMF approach to audio event detection," *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA),* New Paltz, NY, USA, Oct. 2013.

[33] L. Vuegena, B. V. D. Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, and H. V. hamme, "An MFCC-GMM approach for event detection and classification," *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA),* New Paltz, NY, USA, Oct. 2013.

[34] 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," *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA),* New Paltz, NY, USA, Oct. 2013.

[35] A. Diment, T. Heittola, and T. Virtanen, "Sound event detection for office live and office synthetic AASP challenge," *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA),* New Paltz, NY, USA, Oct. 2013.

[36] J. Schroder, S. Goetze, and J. Anemuller, "Spectro-Temporal Gabor Filterbank Features for Acoustic Event Detection," in *IEEE/ACM Transactions on Audio, and Language Processing,* vol. 23, no. 12, pp. 2198-2208, 2015.

زیرنویس‌ها

[1] - Boosting

[2] - Mel-frequency Cepstral Coefficients (MFCC)

[3] - Log frequency filter bank

[4] - Leveraging statistical models

[5] - Tandem connectionist model

[6] - Receiver Operating Characteristic (ROC)

[7] - Low short-Time Energy Ratio (LSTER)

[8] - IEEE Audio and Acoustic Signal Processing (AASP)

[9] - Synthesis Sounds

[10] - F1-measure