
بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

رشته مهندسی کامپیوتر - گرایش هوش مصنوعی

رساله دکتری

کشف وقایع صوتی غیرهمزمان در سیگنال‌های محیطی و رده‌بندی آنها

نگارنده: مراد درخشان

استاد راهنما

دکتر حسین مروی

استاد مشاور

پروفسور حمید حسن‌پور

شهریور ۱۳۹۶

تقدیم به

پدر و مادر عزیز و مهربانم به خاطر همه‌ی تلاش‌ها و فداکاری‌های بی‌دریغ و ستودنی آنها که در دوران زندگی خود نثار من داشتند، همچنین **برادران و خواهرانم** که به‌تمام معنا دوستدار آنها هستم، و **همسر عزیز و فرزندان دلبندم نیلوفر و امیرحسین** که چشم و چراغ زندگی من هستند.

تشکر و قدردانی

سپاس و ستایش مخصوص الله است. اوست تنها آفریننده و خالق هر آنچه در زمین و آسمانها است. احسن‌الخالقین جز بر ذات او، فقط از جانب آن توانا است. هر آنچه او بیافریند، مختوم به وقت معلوم و به جز ذات ازلی و ابدی او که باقی است، باقی همه‌چیز فانی است. مخلوق گر نرسد به‌حد عالی، باشد اثرش همیشه باقی. حمد و سپاس آن خدایی، که سزد بر او تشکر و قدردانی. الحمدالله‌رب‌العالمین.

نگارنده این اثر از اساتیدی که کمک و راهنمایی ایشان در انجام این کار پژوهشی همراه بوده‌است، ازجمله استاد راهنمای محترم این رساله **جناب‌آقای‌دکتر حسین مروی** که از راهنمایی و کمک‌های ایشان بهره‌مند گردیدم، **جناب‌آقای‌پروفسور حمید حسن‌پور** که سمت مشاوره این کار را برعهده داشتند، همچنین **جناب‌آقای‌دکتر امیدرضا معروضی و جناب‌آقای‌دکتر مرتضی زاهدی** که داوری این اثر را انجام داده‌اند، تشکرمی‌نماید.

چکیده

تشخیص و رده‌بندی وقایع صوتی[1] در محیط کار و زندگی به صورت یک نیاز در دنیای مدرن قرن ۲۱ تبدیل شده‌است، به‌نحوی که شاخه‌ای از پژوهش در هوش مصنوعی و به تبع آن در زیرشاخه محاسبات آنالیز صحنه شنیداری[2] از حوزه پردازش سیگنال دیجیتال به این کار اختصاص یافته‌است.

این رساله، دو پیشنهاد مهم در خصوص کشف و رده‌بندی صداهای محیطی ارائه می‌کند. در راهکار اول یک مدل با چندین جزء مستقل کوچک‌تر پیشنهاد می‌شود که هرکدام دارای نقطه قوتی برای کل مدل است. بیش‌ترِ مدل‌های موجود، همه انتظارات از مدل‌سازی را در یک تعریفِ منسجم و آماری می‌بینند، اما در اینجا، مدل اصلی از ترکیب موازی چند جزء ساخته می‌شود تا بتواند حداکثر اطلاعات را از داده‌های خام استخراج و آن‌ها را به‌نحو مطلوب مدل کند و به‌علاوه بین اجزاء، به‌خصوص ویژگی‌ها، تاثیر جانبی مثبت را جایگزین تاثیر جانبی منفی نماید. معیار F1 در این روش برای شناسایی فریم و واقعه صوتی به ترتیب ۸۰/۶ درصد و ۷۲/۸ درصد بدست آمده‌است. راهکار دوم مبتنی بر تجزیه نامنفی[3] سیگنال مشاهده جهت ذخیره بافت فرکانسی نمونه از صداها به صورت اتم‌ها یا وصله‌های‌صوتی در یک دیکشنری و سپس بازسازی سیگنال آزمایش از طریق تولید یک نمایش تنک[4] از اتم‌های دیکشنری در بازنمایی سگمنت‌های آن سیگنال است. این روش در واقع مبتنی بر جداسازی و تفکیک یک سیگنال به اجزای اولیه یعنی وقایع صوتی است. معیار F1 در این روش برای شناسایی فریم و واقعه صوتی به ترتیب ۵۶/۵ درصد و ۴۹/۲ درصد بدست آمده‌است. این نتایج برتری روش‌های پیشنهادی نسبت به روش‌های پیشین بر روی پایگاه داده DCASE2013 را نشان می‌دهد.

کلمات کلیدی: شناسایی و رده‌بندی وقایع صوتی، سگمنت‌بندی، استخراج ویژگی، جداسازی منابع صوت، تجزیه نامنفی ماتریس، نمایش تنک سیگنال، آنالیز صحنه‌های شنیداری، الگوریتم‌های یادگیری ماشینی، اسپکتروگرام، نظارت و دیده‌بانی صوتی.

[1]- Audio Event Detection and classification (AED)
[2]- Computational Auditory Scene Analysis (CASA)
[3]- Non-negative Matrix Factorization (NMF)
[4]- Sparse representation

فهرست مقالات مستخرج از رساله

الف: مقالات چاپ‌شده در مجلات علمی پژوهشی داخلی

[۱] مراد درخشان و حسین مروی، «کشف و رده‌بندی وقایع صوتی محیطی با استفاده از نگاشت سگمنت بر دیکشنری در نمایش تنک،» مجله مهندسی برق دانشگاه تبریز، جلد ، شماره ، صفحات ، ۱۳۹٦.

[۲] مراد درخشان، حسین مروی و حمید حسن‌پور، «ارائه یک مدل پارامتریک تطبیقی جهت کشف وقایع صوتی در سیگنال‌های محیطی،» مجله مهندسی برق دانشگاه تبریز، جلد ، شماره ، صفحات ، ۱۳۹٦.

ب: مقالات ارائه شده به‌صورت سخنرانی و پوستر در کنفرانس‌های داخلی

[۱] مراد درخشان، حسین مروی، حمید حسن‌پور، "کشف وقایع صوتی محیطی بر اساس مدل پارامتریک بدون نظارت"، سومین کنفرانس بین‌المللی پژوهش‌های کاربردی در مهندسی کامپیوتر و فن‌آوری اطلاعات، دانشگاه تربیت‌مدرس، بهمن ۱۳۹٤.

[۲] مراد درخشان و حسین مروی، "ارائه یک الگوریتم بدون‌نظارت جهت کشف وقایع صوتی محیطی با تاکید بر تعیین دقیق زمان رخداد وقایع"، دومین کنفرانس بین‌المللی پردازش سیگنال و سیستم‌های هوشمند، دانشگاه صنعتی امیرکبیر، آذرماه ۱۳۹۵

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

علائم اختصاری

ACF	Auto-Correlation Function
AED	Audio Event Detection
ANN	Artificial Neural Network
ASC	Audio Spectrum Centroid
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
BFCC	Bark-Frequency Cepstral Coefficients
BIC	Bayesian Information Criteria
BMP	Basic Matching Pursuit
CASA	Computational Auditory Scene Analysis
CELP	Code Excited Linear Prediction
CHIL	Computer in the Human Interaction Loop
CLEAR	Classification of Event, Activities, and Relationships
CQP	Convex Quadratic Program
CWT	Continuous Wavelet Transform
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
EM	Expectation-Maximization
F0	Fundamental frequency
FFBE	Frequency Filtered Band Energies
FFT	Fast Fourier Transform
FPR	False Positive Rate
GMM	Gaussian Mixture Model
HCC	Homomorphic Cepstral Coefficients
HMM	Hidden Markov Model
HZCRR	High Zero-Crossing Rate Ratio
ICA	Independent Component Analysis
IEEE	Institute of Electrical and Electronics Engineers
KL	Kullback-Leibler
KNN	K-Nearest Neighbor
LPC	Linear Prediction Coefficients
LPCC	Linear Predictive Cepstral Coefficient
LSP-VQ	Line Spectral Pair Vector Quantization
LSTER	Low Short-Time Energy Ratio
MAP	Maximum A Posterior Estimation
MFCC	Mel-Frequency Cepstral Coefficients
MIR	Music Information Retrieval
MP	Matching Pursuit
NASE	Normalized Audio Spectrum Envelope
NB-ACFF	Narrow-Band Auto Correlation Function Features
NMF	Non-negative Matrix Factorization
NRAF	Noise-Robust Auditory Features
PCA	Principal Component Analysis

PSD	Power Spectral Density
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RT	Rich Transcription
SBER	Sub-Band Energy Ratio
SDF	Spectral Dynamic Features
SNR	Signal-to-Noise Ratio
SR	Sparse Representation
STE	Short-Time Energy
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TM	Template Matching
TPR	True-Positive Rate
VQ	Vector Quantization
ZCR	Zero-Crossing Rate

فصل اول

۱- مقدمه

مراقبت و نظارت بر فعالیت‌های انسان هیچ‌گاه به اندازه امروزه همه‌گیر و وسیع نبوده است. میلیون‌ها حس‌گر مراقبتی در اغلب نواحی شهرها، در محیط‌های حساس و تأسیسات تجاری یا صنعتی در بعد متنوعی از کاربردها نصب شده‌اند و وقایع صوتی را ضبط و جهت دیده‌بانی و نظارت بر محیط به دستگاه‌های پردازشگر صدا می‌فرستند. متعاقب آن مطالعات روی نظارت و مراقبت خودکار رشد خیلی وسیعی داشته و به نصب هزاران الگوریتم در دستگاه‌های تجاری مختلف منجر شده است. در نسل‌های اولیه سیستم‌های نظارتی، مراقبت از طریق یک کاربر انسانی برای شناسایی وضعیت‌های بغرنج و غیرعادی انجام می‌شد اما سیستم‌های پیشرفته امروزی از طریق بینایی ماشین و شناسایی الگو این کار را انجام می‌دهند.

سیستم‌های خودکار اولیه معمولاً از یک یا چند دوربین ویدیویی استفاده می‌کردند. اما مشکلاتی ازجمله کارایی کم دوربین‌ها در شرایط آب و هوایی متفاوت، حساسیت نسبت به تغییرات ناگهانی روشنایی، انعکاس‌های نوری و سایه‌ها و همچنین نقایص نور کم در شب‌ها در این دستگاه‌ها وجود دارد. لذا توجه و اقبال به استفاده از اطلاعات صوتی برای نظارت و مراقبت از اماکن عمومی و خصوصی رو به رشد گذاشته است [۱]. به دلیل اینکه ضبط صدا با هزینه‌های کم‌تر در مکآن‌های عمومی و خصوصی قابل انجام است اخیر در بانک‌ها [۲]، در آسانسورها [۳]، وسایل نقلیه عمومی [۴]، ایستگاه قطار[۵]، در مزرعه [۶]، شناسایی و ردیابی وقایع صوتی در محیط‌های هوشمند [۷] از این نوع نظارت استفاده‌شده است. غیرازاینکه کشف وقایع صوتی برای سیستم‌های نظارت و مراقبت لازم است جهت آنالیز شنیداری محیط نیز مفید هستند که در بحث شناسایی صحنه شنوایی کاربرد دارند.

اگر صداها را به‌طورکلی به چهار دسته گفتار، موزیک، صداهای محیطی و نویز تقسیم نماییم در این تقسیم‌بندی صداهای محیطی بعد خیلی گسترده‌ای را در برمی‌گیرند و هر صدایی غیر از سه گروه دیگر را شامل می‌شوند. لذا منظور از صداهای محیطی صداهای طبیعی و مصنوعی روزمره بدون در نظر گرفتن موزیک و گفتار است. از نظر محل وقوع معمولاً صداهای محیطی به دو گروه اصلی تقسیم

می‌شوند. شاخه اول صداهای محیطی داخلی[1](زیر سقف) و شاخه دوم صداهای محیطی بیرونی[2] نامیده می‌شوند. صداهای محیطی در مواردی مثل نظارت‌های خانگی ، اتاق ملاقات، خانه‌های هوشمند از نوع داخلی هستند و صداهای موجود در خیابان، مکآن‌های عمومی، حیات‌وحش و همانند این‌ها از نوع صداهای بیرونی می‌باشند. بنابراین بین موزیک و گفتار از یک طرف و صداهای محیطی از سوی دیگر تفاوت‌های چشمگیری وجود دارد. به‌طوری‌که استخراج ویژگی‌های مناسب صوتی برای صداهای محیطی متفاوت از موزیک و گفتار هست. به دلیل ساختار فونتیکی متفاوت صداهای محیطی ویژگی‌هایی همچون ضرایب ال پی سی سی[3]ال پی سی[4]ام اف سی سی[5]یا تیمبر[6] و ریتم[7]اف صفر[8]و گام[9]که در شناسایی موزیک و گفتار و همچنین شناسایی گوینده مکرراً استفاده می‌شود چندان مناسب صداهای محیطی نیستند. همچنین تحقیقاتی که در زمینه‌ی صداهای محیطی انجام‌شده است بیش‌تر متمرکز بر شناسایی و رده‌بندی تعداد محدودی از وقایع صوتی بوده‌اند لذا لازم است تا در زمینه صداهای محیطی پژوهش‌های جدید و متنوعی انجام شود.

وقایع صوتی محیطی در حقیقت صحنه‌های شنوایی پیرامون ما را تشکیل می‌دهند لذا جداسازی و شناسایی یک اتفاق صوتی در محدوده زمانی رخ‌داده برای انسان کاری آسان و تکراری است اما انجام چنین اموری برای پردازش‌های ماشینی هنوز به‌عنوان یک چالش اساسی و حل‌نشده مطرح است به‌طوری‌که تحقیق و پژوهش جهت حل این معضل یکی از اولویت‌های اصلی انجمن‌های پردازش صوت هست. به‌طورکلی می‌توان گفت که کشف و شناسایی وقایع صوتی در زمینه‌های مختلفی از جمله نظارت و دیده‌بانی صوتی در محیط‌های عمومی، داده‌کاوی صوتی، تجزیه‌وتحلیل محتوایی رکوردهای صوتی، نمایه کردن و مکان‌یابی زمانی وقایع صوتی، درک یک ابزار هوشمند مثل گوشی-های تلفن سیار نسبت به محیط اطراف خود، پیگیری یک واقعه صوتی خاص، مراقبت از افراد بیمار و

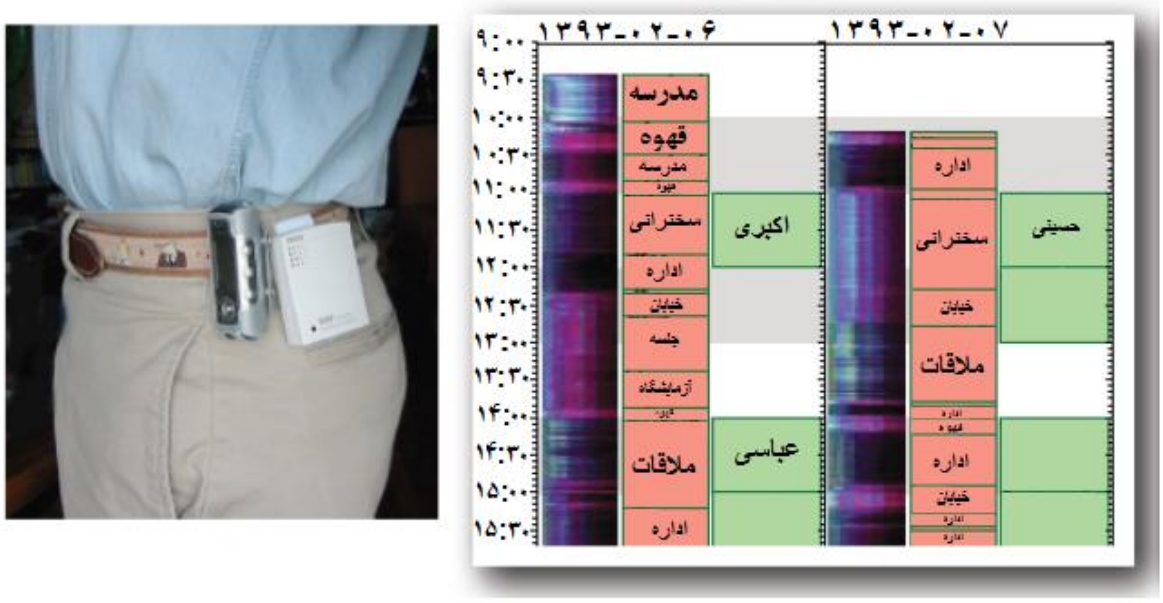
[1]-Indoor Environmental Sounds
[2]- Outdoor Environmental Sounds
[3]- linear prediction cepstral coefficients(LPCC)
[4]- Linear predictive coding (**LPC**)
[5]- Mel-frequency cepstral coefficients (MFCCs)
[6]- Timbre
[7]- Rhythm
[8]- fundamental frequency(F0)
[9]- Pitch

سالمند، وقایع‌نگاری صوتی روزانه و تعداد زیادی از کاربردهای دیگر که از صوت به‌عنوان بستر اطلاعات استفاده می‌کنند همگی نیازمند پردازش‌های هوشمندی است که قابلیت تشخیص واقعه صوتی، تعیین محدوده شروع و پایان آن‌ها و همچنین شناسایی و رده‌بندی وقایع صوتی را داشته باشند. درعین‌حال کسب اطلاعات از صداهای موزیکال و موسیقی و همچنین سیگنال‌های گفتار پیشرفت‌های محسوسی داشته است و کاربردهای زیادی نیز داشته‌اند اما در حوزه صداهای محیطی تحقیقات اندک و موردی بوده است و نیازمند توجه بیش‌تر در مورد صداهای محیطی و کسب اطلاعات در مورد محیط از طریق کشف وقایع صوتی محیطی کاملاً احساس می‌گردد.

علاوه بر این با وجود دامنه گسترده برای کاربردهای مبتنی بر صوت هنوز چارچوب‌های منسجم در چگونگی به دست آوردن اطلاعات از صداهای ضبط‌شده از جمله اینکه از چه نوع ویژگی‌های صوتی و از چه نوع الگوریتم‌هایی استفاده شود وجود ندارد و نمی‌توان شناسنامه مشخصی از ویژگی‌های مهم و الگوریتم‌های آزمایش‌شده برای انواع کاربردهای صدا به ترتیب درجه اهمیت لیست کرد. یک عامل این کمبودها نبودن پایگاه داده جامع برای بررسی ویژگی‌ها به‌طور عام است و دیگری چندبخشی شدن کاربردهای صوت در حوزه‌هایی همچون پردازش سیگنال صوت، شناسایی الگو و چندرسانه‌ای است. از سوی دیگر تحقیقات در شناسایی صوت عمدتاً بر روی سیگنال‌های گفتار و موزیک متمرکزشده است و طی دهه اخیر مطالعات در خصوص مدل‌سازی و شناسایی این نوع صداها افزایش چشمگیری داشته است ولی مطالعات در زمینه صداهای محیطی که بیش‌ترین ابعاد کاربردی را دارند و انواع متعددی از صداها را در برمی‌گیرند چندان پیشرفت مهمی نداشته است. با توجه به ابعاد گسترده وجود صداهای محیطی، نقش و جایگاه شناسایی آن‌ها در راستای تکمیل ماشین‌های شنیداری از اهمیت ویژه‌ای برخوردار است. نمونه‌ای از این نیاز را می‌توان در افزایش روزافزون تقاضا برای جستجو بر اساس نمونه صوتی نام برد. برای مثال جستجوی ویدیو و تصویر بر اساس محتوا و یا به‌طورکلی هر نوع کاربردی که جستجوی صوتی را در خود داشته باشد از جمله موارد جستجو برحسب نمونه صوتی فرض می‌شود. شناسایی صداهای محیطی در موارد متعدد دیگری نیز می‌تواند مورد استفاده قرار بگیرد. از

جمله می‌توان به مواردی همچون، برچسب‌گذاری خودکار فایل‌های صوتی به‌منظور بازیابی آن‌ها برحسب کلمه کلیدی [۸]، مراقبت خانگی، دیده‌بانی و کمک به سالمندانی که تنها در منزل خودشان نگهداری می‌شوند و همچنین کاربرد آن در اتوماسیون خانه‌های هوشمند اشاره نمود [۹]. از سایر زمینه‌های کاربردی برای شناسایی صداهای محیطی می‌توان به آنالیز همزمان تصویر و صوت در کاربردهای نظارتی و خدمات پزشکی از جمله جراحی نام برد و همچنین می‌توان از آن در تشخیص گونه‌های مختلف حیوانات و پرندگان استفاده کرد [۱۰]. یکی از نمونه‌های عملیاتی استفاده از پردازش صداهای محیطی استفاده از امکان پردازش صوت در هدایت صحیح روبات‌ها در محیط و پاسخگویی مناسب نسبت به محرک‌های صوتی است که این امکان میزان هوشمندی و قدرت جهت‌یابی این‌گونه روبات‌ها را بسیار بالا خواهد برد. یک سیستم کاربردی دیگر برای پردازش صداهای محیطی وقایع‌نگاری از زندگی روزانه[۱] با استفاده از صوت است. در این کاربرد شخص یک سیستم ضبط صدا همراه خود دارد که این سیستم تمام صداهای محیط را طی فعالیت روزانه او ضبط می‌کند و سپس از خروجی این سیستم برای کشف اتفاقات صوتی مختلف که در طی روز کاری ضبط‌شده است استفاده می‌شود و با این امکان در واقع می‌توان یک واقعه‌نگاری از زندگی روزمره شخص انجام داد. همچنین می‌توان از وقایع صوتی کشف‌شده از خروجی این سیستم مشخص کرد که در طول روز در چه مکآن‌ها یا محیط‌هایی حضورداشته است و یا در چه موقعیت‌هایی قرارگرفته است. نمونه مکآن‌ها می‌تواند به‌صورت : منزل- خیابان – مدرسه- کافی‌شاپ- کلاس درس- سخنرانی - جلسه گروهی – خوردن قهوه – آزمایشگاه – رستوران و .... باشد که این‌ها کارها یا اتفاقات مختلف روزمره‌ای است که با شناسایی وقایع صوتی مختلف به ترتیب رخداد زمانی لیست شده است. شکل ۱-۱ نمایی از این نوع کاربرد را نشان می‌دهد.

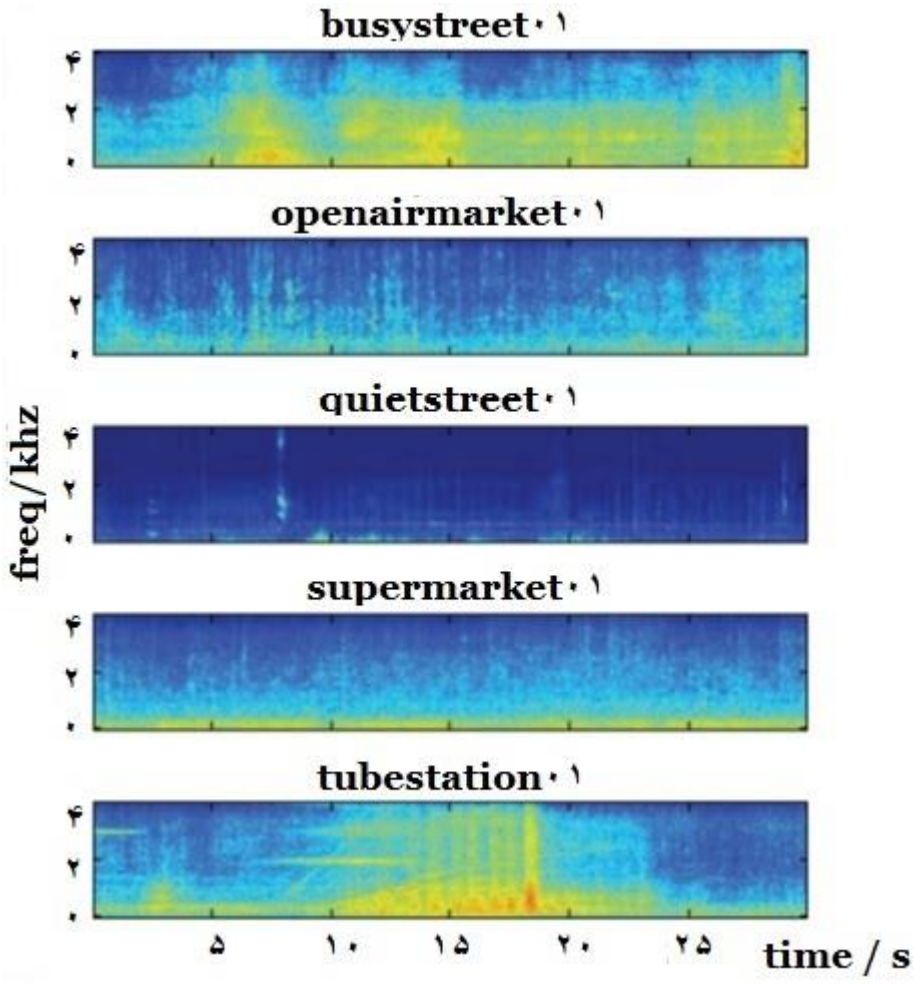
[۱]- Audio Lifelog Diarization



شکل ۱-۱: نمونه‌ای از سیستم وقایع‌نگاری از زندگی روزانه توسط ضبط صداهای محیطی.

کاربرد عملی دیگری که می‌توان برای تشخیص وقایع و درک صحنه‌های صوتی نام بـرد تنظیمـات خودکار گزینه‌ها و انتخاب مدهای مختلف کاری در ابزارها و واسط‌های شنوایی و شـنیداری برحسـب محیط کاری به‌منظور بهبود کارایی و عملکرد این‌گونه ابزارها است. برای مثال در یـک نگـاه جزئـی‌تـر می‌توان به کاربردهای مختلفی که به کشف و رده‌بندی وقایع صوتی نیاز دارند به موارد زیر اشاره نمود: نمایه کردن و بازیابی پایگاه داده‌های چندرسانه‌ای بر اساس زمینه صوتی ، پایش و نظارت در مراقبت-های پزشکی، نظارت‌های کلی در محیط، کاربردهای نظامی، شناسایی صحنه صـوتی (بـرای نمونـه در شکل ۱-۲ تعداد ۵ صحنه صوتی با وقـایع صـوتی مختلـف نمـایش داده‌شـده اسـت)، برچسـب‌گذاری خودکار رکوردهای صوتی، قطعه‌بندی صوتی، حفظ امنیت مکآن‌هـای عمـومی همچـون بانـک‌هـا یـا زیرگذرها و فرودگاه‌ها، شناسایی و دسته‌بندی خودکار موزیک‌ها برحسـب نـوع و دسـتگاه‌هـای تولیـد موزیک، شناسایی و آنالیز حالت‌های غیر نرمال و همچنین خلاصه‌سازی رکوردهای صـوتی بـر اسـاس تعیین وقایع مهم موجود در آن‌ها از جمله کاربردها است. همان‌طور که در شـکل ۱-۲ دیـده می‌شـود تعداد وقایع صوتی اتفاق افتاده در صـحنه‌های busystreet , tubestation نسـبت بـه سـایر صـحنه‌هـا بیش‌تر است در زمآن‌های مختلفی ناهنجاری صـوتی مشـاهده می‌شـود. ایـن در حـالی اسـت کـه در

quietstreet وقـایع به‌نـدرت رخ‌داده‌انـد و در supermarket بـروز وقـایع یکنواخـت، دائمـی و فاقـد ناهنجاری صوتی است و تا حدودی مشابه نـویز زمینـه اسـت. در صـحنه openairmarket نیـز وقـایع صوتی با فواصل تقریباً منظم و متعدد بدون ناهنجاری صوتی رخ‌داده‌اند.



شکل ۱-۲: نمونه‌ای از ۵ رکورد صوتی مختلف که هر کدام مربوط به یکی از صحنه‌های صوتی رایج است.

یک سؤال اساسی ممکن است در اینجا وجود داشته باشـد و آن این‌کـه چـرا بـا وجـود امکانـات و تجهیزات وسیع‌تر و پیشرفته‌تر در حوزه پردازش تصویر و ویدیو و همچنین وجود الگوریتم‌های راحت-تر و گسترده‌تر در آن ما به صدا و پردازش صداهای محیطی متوسل می‌شویم؟ چندین دلیل برای این کار وجود دارد. ابتدا اینکه مزیت عمده استفاده از پردازش صوت به‌جای پردازش داده‌های ویـدیویی و تصویری در این است که در بسیاری از موارد که نظـارت‌هـای ویـدیویی بـه حجـم زیـاد ذخیره‌سـازی

نیازمند است داده‌های صوتی حجم نسبتاً کم‌تری را نیاز دارد. دوم اینکه در مکان‌ها و مواقعی که استفاده از ویدیو سبب نقض حریم خصوصی افراد شده و به‌نوعی برای آنان ایجاد مزاحمت محسوب می‌شود استفاده از صوت به لحاظ حفظ حریم‌های شخصی و خصوصی اولویت می‌یابد. سومین دلیل این است که در مکان‌هایی که تحت پوشش جهت‌های مختلف جغرافیایی مدنظر باشد می‌توان از کاربرد صوت به‌جای ویدیو استفاده نمود. چهارم اینکه می‌دانیم یکی از راه‌های مهم کشف واقعیت‌ها و احراز هویت‌ها و شخصیت‌ها از طریق داده‌های صوتی انجام می‌شود و گاهی اوقات اطلاعاتی که از صدا به دست می‌آید منحصر به صوت است و هیچ‌گاه از تصویر و ویدیو حاصل نمی‌گردد. بنابراین داده‌های صوتی به همان اندازه که تصویر اهمیت دارد مهم و کاربردی هستند. مزید بر این‌ها می‌توان گفت که وابستگی داده‌های ویدیویی به داده‌های سیگنال صوت نیز بر کسی پوشیده نیست.

۱-۱- اهمیت و جایگاه شناسایی وقایع صوتی در تعامل با انسان

واسط‌ها و کامپیوترهای هوشمندی که در محیط‌های ارتباطی با انسان همکاری می‌کنند لازم است قدرت شناسایی و توصیف فعالیت‌های پیرامون خود را داشته باشند و باید طوری طراحی شوند که یک حد لازم از آگاهی را نسبت به کاربران داشته باشند. بنابراین در چنین محیط‌هایی واسط‌های کاربری دارای ادراک باید به‌گونه‌ای طراحی شوند که علاوه بر مالتی مودال[۱] و مقاوم بودن بتوانند از حسگرهای غیر مزاحم[۲] و از جمله ورودی‌های صوتی نیز استفاده کنند. نمونه‌ای از این نوع محیط‌ها و چالش‌های مربوط به آن طراحی اتاق‌های هوشمند است. اتاق هوشمند یک فضا و محیط بسته است که مجهز به تعدادی میکروفون و دوربین است و تعدادی عملکرد تعریف‌شده وجود دارد که به‌منظور کمک به انسان و تکمیل فعالیت‌های او طراحی می‌شود. در این محیط‌ها یا واسط‌های هوشمند در خصوص پردازش و درک گفتار تکنولوژی‌های خاصی به کار گرفته می‌شوند که عمدتاً شامل شناسایی فعالیت گفتاری[۳]، تشخیص خودکار گفتار، شناسایی و تصدیق گوینده و مکان‌یابی گوینده هست. از

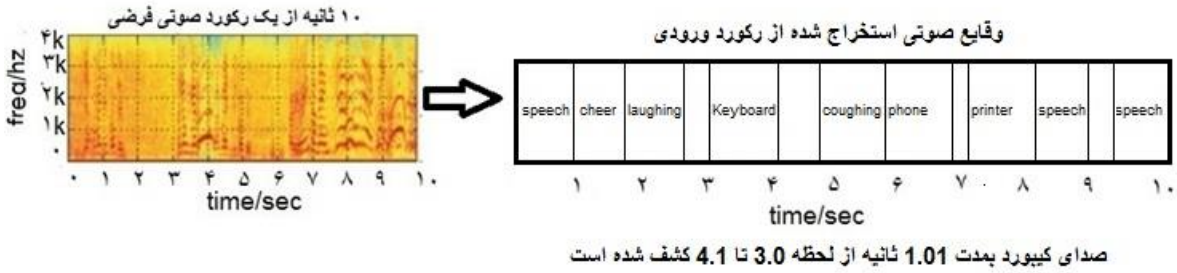
---

[۱]- Multimodal سیستم‌هایی هستند که همزمان این امکان را دارند تا با کاربر به چندین روش مختلف ارتباط برقرار کنند. مثلا از طریق صفحه کلید، صدا یا ویدیو

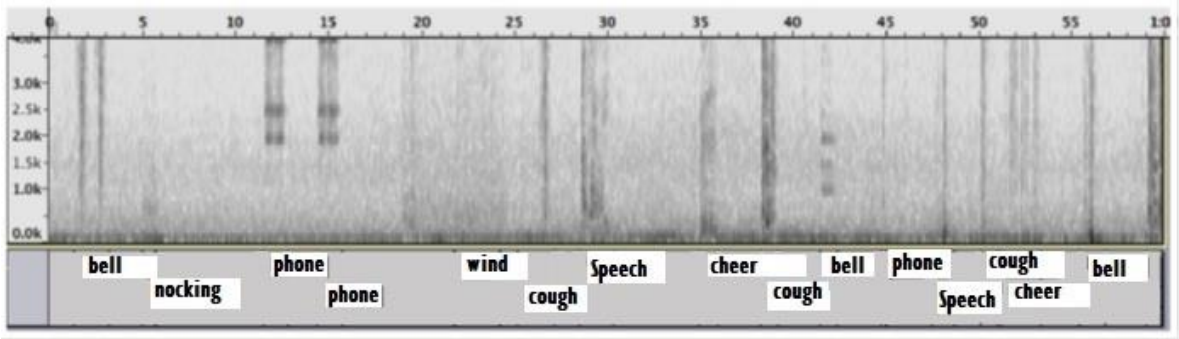
[۲]- unobtrusive

[۳]- speech activity detection

طرفی در چنین محیط‌هایی فعالیت انسان در بعد وسیع‌تری از وقایع صوتی منعکس می‌شود که منحصر به گفتار نیست. این صداهای محیطی یا توسط خود انسان تولید می‌شوند و یا حاصل جابجایی اشیا توسط انسان هستند. لذا مجموعه این وقایع صوتی می‌تواند در زمینه رمزگشایی فعالیت‌های انسان بکار گرفته شود. برای مثال کف زدن یا خندیدن حین گفتار، خمیازه طولانی در طی یک سخنرانی، حرکت صندلی یا صدای درب زمانی که یک نشست در حال انجام است نمونه وقایع صوتی تولیدشده می‌باشند. بنابراین گفتار تنها رویداد صوتی نیست که حاوی اطلاعات است بلکه سایر صداهای محیطی نیز اطلاعات ارزشمندی را با خود حمل می‌کنند که برای شناسایی فعالیت‌ها و درک محیط این نوع صداها نیز باید تشخیص داده شده و رمزگشایی شوند. در شکل ۱-۳ تعدادی از وقایع صوتی موجود در یک سیگنال صوتی ضبط‌شده از یک محیط مفروض نشان داده شده است که در آن تعدادی از وقایع به صورت همزمان رخ داده و با هم تداخل و همپوشانی دارند اما در عمل کشف این وقایع همزمان عملی نیست و چنانچه واقعه‌ای در این حالت گزارش شود مربوط به واقعه‌ای است که حالت برجسته‌تر و چیره‌تری دارد ولی در عین حال گوش شنونده می‌تواند صداهای مختلف در محدوده گزارش شده را در صورت امکان درک نماید. بنابر این معمولا وقایع در وضعیت غیرهمزمان قابل گزارش است و محدوده‌هایی که احتمال وقوع همزمان دو یا بیش‌تر واقعه در آن‌ها باشد را باید با تکنیک‌های جداسازی سیگنال‌های توام تشخیص و جداسازی نمود. لذا خروجی پروسه کشف وقایع چیزی شبیه شکل ۱-۴ است که در اینجا وقایع چیره در بازه زمانی مشخص گزارش شده‌اند. علاوه بر درک محیط بر اساس صداهای غیر گفتاری می‌توان گفت که کشف و شناسایی وقایع صوتی غیر گفتاری می‌تواند در مقاوم سازی و بهبود کیفیت در سیستم‌های تشخیص اتوماتیک گفتار نیز نقش مهمی را ایفا نماید.



شکل ۱-۳: یک سیگنال صوتی مفروض و وقایع صوتی موجود در آن



شکل ۱-۴: وقایع صوتی استخراج شده از یک رکورد صوتی که از نظر زمانی غیرهمزمان تشخیص داده می‌شوند.

با توجه به مقدمه ذکر شده و درجه اهمیت خیلی زیادی که برای کاربرد شناسایی و تشخیص صداهای محیطی وجود دارد و به لحاظ انجام تحقیقات و مطالعات کم‌تر در این خصوص نسبت به گفتار و اینکه وقایع صوتی اطلاعات کاملی از شرایط محیط و درک شرایط و صحنه‌ها را برای انسان و ماشین هوشمند امکان پذیر می‌سازد و از سوی دیگر تشخیص اتفاقات صوتی محیطی یک نیاز کاملا تعریف شده و هدف گذاری شده برای سیستم‌های هوشمند تعاملی است لذا هدف اصلی در این رساله کشف و رده‌بندی وقایع صوتی غیرهمزمان در رکوردهای صوتی محیطی زیر سقف است. این رساله تحت عنوان " کشف وقایع صوتی غیرهمزمان در سیگنال‌های محیطی و رده‌بندی آن‌ها " تحقیق و بررسی می‌شود و حاصل آن نرم افزاری است که ورودی آن یک رکورد صوتی ضبط شده از محیط خاص زیر سقف است و خروجی آن لیستی از وقایع کشف‌شده غیرهمزمان از رکورد ورودی است. این خروجی زمان شروع و زمان پایان و برچسب مناسب را برای هر اتفاق کشف شده مشخص می‌نماید.

مدل را برای کشف و رده‌بندی اتفاقات صوتی موجود در یک محیط کاری بکـار گرفتـه و پیاده‌سـازی می‌کنیم و با توجه به اینکه یکی از اهداف سیستم‌های پـردازش صـوت در شـاخه آنـالیز صـحنه‌هـای شنیداری محاسباتی شبیه سازی مدل درک و تصور سیستم شنوایی در انسان است این پردازش شامل تفکیک وقایع صوتی از هم و تعیین محدوده زمانی آن‌هـا و نهایتـا تبـدیل وقـایع صـوتی مختلـف بـه برچسب‌های متناظر با ادراک شنونده خواهد بود. همچنین طول زمانی هـر یـک از وقـایع اکتشـافی را برای استدلال‌های بعدی و استفاده از وقایع کشف شده مشخص می‌کنیم. همانطور که در مقدمه ذکـر شد جایگاه این نوع پردازش صوت شاخه‌ای از آنالیز صحنه‌های شنیداری محاسباتی[1] است کـه شـامل پردازش سیگنال‌های صوتی و تبدیل وقایع صوتی مختلف در آن به اسامی شناخته شـده بـرای انسـان است. در دهه‌های اخیر تلاش‌های زیادی بـرای جداسـازی سـیگنال‌های همزمـان و تولیـد مفـاهیم و برچسب‌های معنا دار از سیگنال‌های صوتی موجود در رکوردهای صوتی صورت گرفته است که درصـد موفقیت‌های حاصل چشمگیر نبوده است و به همین علت شاخه آنالیز صحنه‌های شنیداری محاسباتی در صدد رسیدن به روش‌هایی است که با دست یافتن به مکانیزم ادراک صوت توسط سیستم شـنوایی انسان راهی به سوی حل معضل درک صدا در سیستم‌های هوشمند بـاز کنـد. مسـلما درک محـیط و صحنه صوتی نیازمند درک وقایع صوتی است و یکی از راه‌هـای رسـیدن بـه اسـرارموجود در سیسـتم شنوایی انسآن‌ها رسیدن به روش‌های مطمئن برای درک وقایع صوتی مجزا و برچسـب دار در محـیط است.

برای رسیدن به توسعه چنین سیستم‌هایی نیاز به پیاده‌سازی الگوریتم‌هایی در خصـوص اسـتخراج ویژگی و همچنین نیاز به الگوریتم‌های مناسب برای مدل سازی وقایع و رده‌بندی آن‌ها اسـت. در هـر سیستم رده‌بندی مطالعه انواع مختلف ویژگی‌ها دارای اهمیت زیادی است. از جملـه تحقیـق در مـورد ارتباط ویژگی‌های عادی که در بحث شناسایی اتوماتیک گفتار استفاده می‌شود با این موضوع اهمیـت زیادی دارد. لازم است انواع ویژگی‌ها را بررسی نمود و در نهایت مناسب‌ترین گروه از ویژگی‌ها را بـرای

[1]- Computational Auditory Scene Analysis (CASA)

این کار انتخاب نمود. همچنین بررسی ویژگی‌های ادراکی در این خصوص دارای اهمیت زیادی است. لازم است میزان مفید بودن و درجه اهمیت آن‌ها برای کشف و شناسایی وقایع صوتی را بررسی نمود.

۲-۱- طبقه‌بندی اصوات[۱]

تحقیقات بر روی رده‌بندی صداها پیش از این برای تعداد محدودی از رده‌ها انجام شده است. مانند

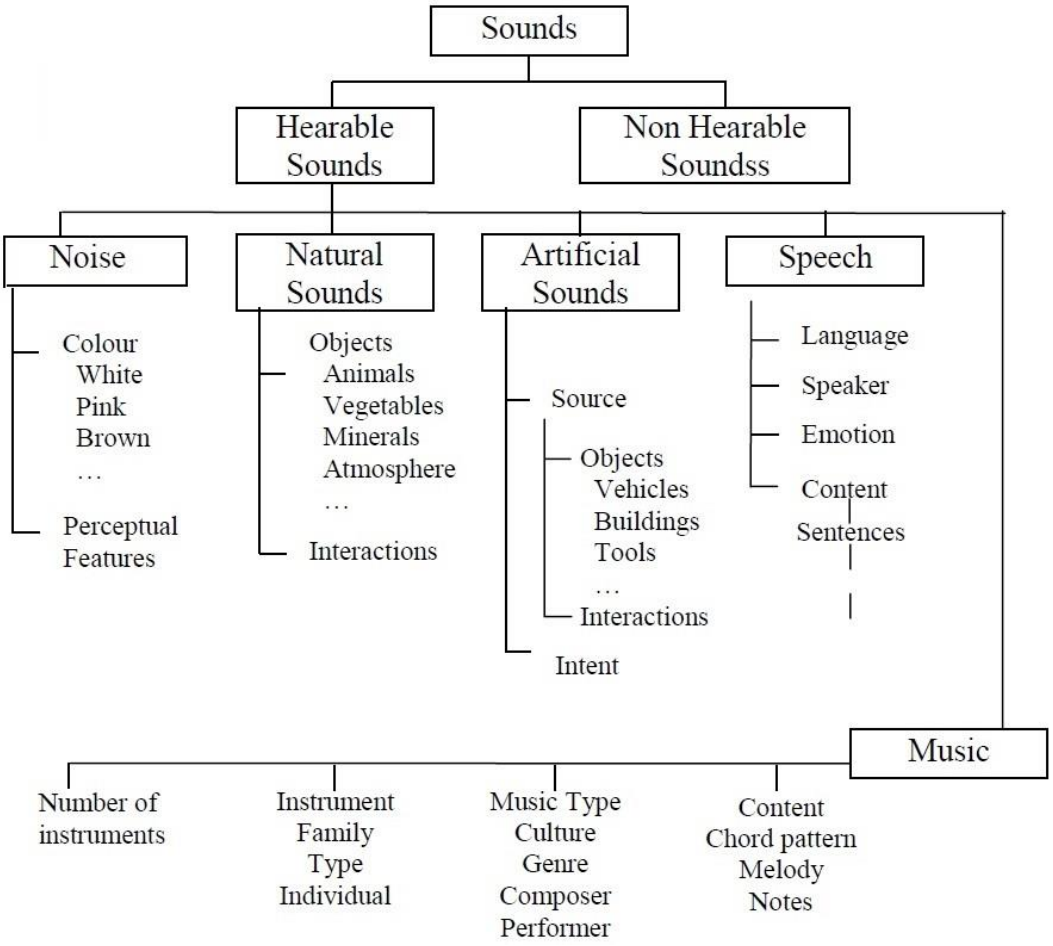
speech/music, music/song , music/speech/other.

رده‌بندی صداها را می‌توان در سطوح مختلف معنایی سازمان داد. مثلا رده‌بندی صداهای محیط-های خاص، رده‌بندی صداهای مخصوص یک فعالیت مورد نظر، رده‌بندی صداهای آلات موسیقی (مثلا صدای ویولون، گیتار و ...). طبقه‌بندی اصوات نیز بر پایه تنوع صداها و منشا تولید آن‌ها ، طبیعی یا مصنوعی بودن و ... انجام می‌شود. یک نوع از طبقه‌بندی صدا مانند شکل ۱-۵ در [۱۱] پیشنهاد شده است. این ساختار روش‌های پردازش و رویکردهای مختلفی مدل‌سازی را برای انواع صداها متذکر می‌گردد. از جمله به تفاوت بین پردازش گفتار [۱۲] و پردازش موسیقی [۱۳] که در جنبه‌های صوتی و سیگنالی و ساختار اطلاعاتی متفاوت هستند می‌توان اشاره نمود.

در پروژه CHIL از دو نوع طبقه‌بندی استفاده شده است. یکی عمومی و دیگری مفهومی. در طبقه‌بندی عمومی مثلا برای صدا دسته‌بندی‌های عمومی در نظر گرفته شده است. در طبقه‌بندی مفهومی صداهای موجود برای کاربرد خاصی مثل اتاق ملاقات درنظر گرفته شده است.

مطالب این رساله در ۵ فصل سازماندهی شده است. در فصل ۱ مقدمه و آشنایی با موضوع پژوهش آورده شده است. مطالب فصل ۲ در رابطه با مبانی نظری و پیشینه موضوع پژوهش است. در فصل ۳ روش پیشنهادی بر اساس مدل‌های تطبیق‌پذیر برای کشف و رده‌بندی وقایع صوتی شرح داده شده و نتایج آورده شده‌اند. در فصل ۴ یک روش پیشنهادی بر اساس مدل جداسازی سیگنال‌ها جهت شناسایی وقایع صوتی محیطی همراه با نتایج مربوطه آورده شده است. در فصل ۵ نتیجه‌گیری و پیشنهادهای مربوطه آورده شده است.

[۱]- Sounds Taxonomy



شکل ۱-۵: طبقه‌بندی صداها ارائه شده در [۱۱].

فصل

۲- مبانی نظری و پیشینه پژوهش

مقدمه

در این فصل، هدف ما بررسی مبانی نظری و آشنایی با ابزارها و روش‌های بکار رفته در شناسایی و کشف وقایع صوتی و همچنین مروری بر کارهای انجام شده در پیشینه این نوع پژوهش در حوزه پردازش سیگنال است که به نوعی مرتبط با بحث شناسایی و رده‌بندی صداها می‌باشند. در راستای این هدف ضمن اشاره به کارهای پژوهشی مرتبط، هدف و روش مورد استفاده نویسندگان ارائه می‌شود و چنانچه نکته مهمی در ارتباط با موضوع بوده است جهت آشنایی با مبانی نظری و معرفی ابزارها و روش‌ها اشاره خواهد شد. مطلب مهمی که در اکثریت پژوهش‌ها بچشم می‌آید عدم وجود یک خط سیر مشخص و برنامه منسجم در پیشبرد مسائل و کاربردهای حوزه پردازش سیگنال نه در یک موضوع بلکه در کلیه سطوح و موضوعات می‌باشد. لذا تحقیقات دامنه‌دار و منسجم برای رفع مشکلات و کاستی‌ها در حوزه‌های مشخص دیده نمی‌شود و بسیاری از پژوهش‌ها در سطح اولیه و بدون رسیدن به نقطه مشخص ناتمام مانده است. لذا تا جایی که امکان دارد ما از حوزه‌های متعدد کاربردها را بررسی و گزارشی از مطالعات در هر زمینه ارائه خواهیم کرد.

۱-۲- طرح های مختلف پردازش صداها با روش‌های فریمی

قبل از هر گونه پردازشی بر روی یک سیگنال صوتی لازم است طرحی را برای پردازش آن انتخاب نمود. این طرح می‌تواند نحوه پرداختن به جزئیات سیگنال صوت و روش بدست آوردن ویژگی‌های موثر و کارآمد را تعیین کند. هر طرح می‌تواند نوع مدل‌سازی و نحوه تصمیم گیری درمورد رده بندی سیگنال را تحت تاثیر قرار دهد. با توجه به بررسی های انجام گرفته به طور کلی می‌توان سه طرح پایه و مهم را برای پردازش سیگنال صوت به صورت زیر تعریف نمود:

۱) طرح پردازش مبتنی بر فریم‌بندی سیگنال.

۲) طرح پردازش مبتنی بر زیر فریم.

۳) طرح پردازش متوالی سیگنال صوت.

در ادامه هر یک از طرح‌های فوق را با توجه به نیاز این تحقیق بررسی و توضیح می‌دهیم.

۱-۱-۲- پردازش مبتنی بر فریم

در این طرح ابتدا سیگنال صوت با استفاده از پنجره های هنینگ یا همینگ به تعدادی فریم تقسیم می‌شود سپس ویژگی‌های صوتی هر فریم استخراج شده و برای آموزش یا آزمایش استفاده می‌شود. در این روش رده بندی هر فریم جداگانه انجام می‌شود و هر فریم ممکن است متعلق به یک کلاس خاص باشد. معضل اصلی این نوع طرح آن است که هیچ روشی برای تعیین یک طول ایده‌آل برای پنجره فریم‌ها که برای تمام رده‌ها مناسب باشد وجود ندارد. بعضی از اتفاقات صوتی لحظه ای و کوتاه هستند مثل صدای شلیک تفنگ و بعضی از آن‌ها مانند صدای رعد و برق طولانی مدت هستند. اگر طول پنجره خیلی کوچک باشد، تغییرات طولانی مدت سیگنال از طریق ویژگی‌های استخراج شده قابل دسترسی نیست. بنابراین روش فریم‌بندی ممکن است وقایع را به تعدادی فریم خورد کند. از طرف دیگر اگر طول پنجره خیلی طولانی باشد تفکیک مرز قطعه‌بندی بین وقایع همجوار مشکل خواهد بود و امکان وجود بیش از یک واقعه صوتی درون هر فریم وجود دارد. در این طرح مشخصات غیر ساکن سیگنال فقط از روی ویژگی‌های استخراج شده امکان پذیر است چون در این مدل امکان استفاده از روش‌های یادگیری متوالی وجود ندارد.

۲-۱-۲- پردازش زیرفریمی

در مدل زیر فریمی، هر فریم به زیر فریم‌های جدیدی که معمولا تا حدی رویهم افتادگی دارند تقسیم می‌شوند و ویژگی‌های هر زیر فریم استخراج می‌شود. برای آموزش رده‌بندها ویژگی‌های استخراج شده یا به هم متصل شده و تشکیل یک بردار ویژگی بزرگ‌تر را می‌دهند و یا اینکه با روش متوسط گیری آن‌ها را به یک بردار ویژگی جدید تبدیل می‌کنند. روش دیگر این است که رده‌بند را برای هر زیر فریم آموزش می‌دهند و سپس برای رده‌بندی فریم از روش تصمیم گیری اشتراکی(یا جمعی) روی برچسب کلاس‌های تمام زیر فریم‌ها استفاده می‌کنند. این روش را قانون رای گیری حداکثری می‌گویند. این مدل امکان استفاده همزمان از ویژگی‌های غیر ساکن و رده‌بندهای متوالی را میسر

می‌سازد. این طرح پردازشی حتی با استفاده از یک رده‌بند غیر متوالی نیز می‌تواند نمایش بهتری از یک فریم ارائه کند چون توزیع اشتراکی روی تمام زیر فریم‌ها امکان می‌دهد تا مشخصات درون فریمی را با میزان صحت بالاتری مدل سازی کنیم. یکی از ویژگی‌های دیگر این روش آن است که انعطاف پذیری بیش‌تری را در قطعه‌بندی وقایع صوتی همجوار بر اساس برچسب‌های کلاس‌های زیر فریم‌ها فراهم می‌سازد.

۳-۱-۲- پردازش متوالی

در این روش باز هم سیگنال صوتی را به قسمت‌های کوچک‌تری بنام قطعه تقسیم می‌کنند که طول هر قطعه در حدود ۳۰ میلی‌ثانیه و میزان اشتراک قطعه‌ها ۵۰ درصد است. رده‌بند تصمیمات خود را برحسب برچسب کلاس‌ها و قطعه‌بندی صورت گرفته انجام می‌دهد و برای این کار از ویژگی‌های استخراج شده از قطعه‌ها استفاده می‌کند. در مقایسه با دو روش قبل این روش با استفاده از مدل‌سازی متوالی سیگنال مثلا با بکارگیری HMM[۱] در بدست آوردن همبستگی درون قطعه‌ها و مدل کردن تغییرات طولانی مدت صداهای محیطی خیلی بهتر عمل می‌کند.

الگوریتم‌های شناسایی صداهای محیطی، ابتدا یکی از سه طرح مورد اشاره فوق را با کمی تغییرات در پیش‌پردازش‌ها و انتخاب طرح‌هایی برای کاهش ویژگی‌ها دنبال می‌کنند. در پیش‌پردازش برای مثال می‌توان از فیلتر برای تقویت محتوای فرکانس‌های بالا یا برای هماهنگی بلندی صداها استفاده نمود.

۲-۲- پردازش صداها با روش‌های تجزیه و آنالیز داده‌های مشاهده

یک گروه کلی از روش‌های دیگر برای پردازش سیگنال صدا را می‌توان از نوع شیوه‌های تجزیه و کاهش مرتبه ماتریس داده‌های مشاهده در مورد سیگنال صوت دسته‌بندی نمود. در این روش‌ها معمولا، هدف تجزیه ماتریس به اجزای پایه و ضریب‌یابی متناسب برای پایه‌ها بر اساس انواع موج‌های درون سیگنال مانند روش ICA[۲] است و یا هدف تولید و استخراج پایه‌ها و اتم‌های اصلی از یک

[۱]- Hidden Markov Model (HMM)
[۲]- Independent Component Analysis (ICA)

سیگنال است مانند آنچه در روش‌های ساخت دیکشنری اتم‌ها یا کدبوک‌ها انجام می‌شود و یا ممکن است تجزیه به منظور تولید ترکیبی یک سیگنال مشاهده از روی جمع پایه‌های وزنی از پیش استخراج‌شده مانند روش تجزیه نامنفی ماتریس (NMF[1]) می‌باشد. روش‌هایی مانند ICA در جداسازی منابع کور (BSS[2]) و کوکتل پارتی و حذف نویز و بهبود کیفیت سیگنال به طور گسترده استفاده شده‌اند. از روش‌های ساخت دیکشنری یا کدبوک‌ها نیز در عملیات کدگذاری، فشرده‌سازی صدا و تصویر، استخراج ویژگی از صوت یا تصویر استفاده شده است. کاربرد روش NMF نیز در شناسایی چهره، آنالیز معنایی متون در اسناد، آنالیز صوت، حذف نویز در صوت و تصویر و همچنین سیستم‌های هوشمند رتبه‌بندی و سیستم‌های پیشنهاد بوده است. نمایش تنک[3] و به طور کلی رسیدن به روش‌های نمایش داده‌ها همراه با کاهش رتبه ماتریس مشاهدات از اهداف مهم NMF می‌باشد.

۱-۲-۲- تجزیه ماتریس مشاهده به روش NMF

تجزیه نامنفی ماتریس یک نوع تکنیک تقریب رتبه پایین برای تجزیه داده‌های چند متغییری محسوب می‌شود. در این تجزیه با داشتن ماتریس حقیقی مثبت V با ابعاد n*m و با در نظر گرفتن عدد صحیح مثبت $r < \min(n,m)$ هدف رسیدن به تجزیه‌ای است که V را به دو ماتریس مثبت پایه‌های W با ابعاد n*r و ماتریس مثبت ضرایب H با ابعاد r*m تجزیه نماید به طوری که تقریب رابطه (۲-۱) در مورد تجزیه مذکور برقرار باشد.

$$V \approx WH \qquad (۲-۱)$$

در این تجزیه هر ستون یا بردار مشاهده v_j از طریق عوامل تجزیه بصورت تقریبی از رابطه (۲-۲) بدست می‌آید.

$$v_j \approx Wh_j = \sum_i h_{ij} w_i \qquad (۲-۲)$$

در اینجا w_i و h_j بترتیب ستون iام و ردیف jام از ماتریس مربوطه می‌باشند.

[1]- Non-Negative Matrix Factorization (NMF)
[2]- Blind Source Separation (BSS)
[3]- Sparse Representation

از آنجا که پردازش تجزیه فوق بصورت تکراری انجام می‌شود و از مقادیر اولیه برای عوامل W و H استفاده می‌کند لذا نتایج حاصل تقریبی هستند و همواره با درصدی اختلاف نسبت به V بدست می‌آیند. بنابراین جهت بدست آمدن بهترین تقریب از یک تابع و معیار بهینه‌سازی به همراه تجزیه استفاده می‌کنند. این تابع هزینه باید اختلاف حاصل تجزیه یعنی WH را با ماتریس مشاهدات V به حداقل برساند. اگر مینیمم‌سازی تابع هزینه را معیار بهینه‌سازی قرار دهیم در اینصورت تابع هزینه (یا تابع ارزش) بصورت رابطه (۲-۳) تعریف می‌شود.

$$C(W,H)=\frac{1}{2}\sum_{j}\| v_j - Wh_j \|_2^2=\frac{1}{2}\sum_{i,j}(v_{ij}-[WH]_{ij})^2 \qquad (۲-۳)$$

در رابطه (۲-۳) تابع $C(W,H)$ تابع ارزش یا معیار بهینه‌سازی در تجزیه NMF نامیده می‌شود و می‌توان مینیمم‌سازی را بصورت رابطه (۲-۴) خلاصه نمود.

$$\underset{W\in\mathbb{R}_+^{n*r},\,H\in\mathbb{R}_+^{r*m}}{arg\,min}\ \frac{1}{2}\| V-WH \|_F^2 \qquad (۲-۴)$$

همان طور که اشاره شد تجزیه فوق در یک پردازش تکراری انجام می‌شود و در هر تکرار لازم است تا عامل‌های W و H قبل از شروع تکرار بعدی بهنگام‌سازی شوند. این بهنگام‌سازی با روش‌های مختلفی انجام می‌شود. یکی از روش‌ها حداقل خطای مربعات است که بصورت متناوب بین عامل‌های W و H با هدف کم شدن تدریجی مقدار تابع بهینه‌سازی به صورت روابط (۲-۵) و (۲-۶) انجام می‌شود.

$$H\leftarrow\underset{H\in\mathbb{R}_+^{r*m}}{arg\,min}\ \| V-WH \|_F^2 \qquad (۲-۵)$$

$$W\leftarrow\underset{W\in\mathbb{R}_+^{n*r}}{arg\,min}\ \| V-WH \|_F^2 \qquad (۲-۶)$$

۲-۲-۲- محدودیت تنک روی تجزیه نامنفی ماتریس

واژه تنک (sparse) یک خاصیت قابل اندازه‌گیری از بردارها است و با تعداد عناصر غیرصفر در یک بردار در ارتباط است. معمولا نرم صفر (Zero-norm) را بصورت L0 نمایش داده و برای اندازه‌گیری تنک بودن یک بردار استفاده می‌کنند. در عبارات ریاضی نرم صفر(L0) یک بردار به صورت $\|x\|_0$ نمایش داده می‌شود. نرم ۱ (L1) برابر است با جمع قدر مطلق مقادیر درون یک بردار و بصورت رابطه (۲-۷) محاسبه می‌شود.

$$l1 = \| x \|_1 = \sum_{i=1}^{n} | x_i | \qquad (۲-۷)$$

نرم ۲ (L2) یک بردار به صورت $\|X\|_2$ نمایش داده می‌شود و معمولا بصورت فاصله اقلیدسی مربوط به یک بردار از مرکز مختصات طبق رابطه (۲-۸) نمایش داده می‌شود.

$$l2 = \| x \|_2 = (\sum_{i=1}^{n} | x_i |^2)^{1/2} \qquad (۲-۸)$$

اگر در یک رابطه تجزیه تقریبی مثل $Ax \approx b$ بخواهد تنک بودن لحاظ شود باید این شرط روی بردار ضرایب x گذاشته شود. لذا بردار تنک x را طوری پیدا می‌کند که رابطه (۲-۹) صحیح باشد.

$$\| Ax - b \|_2 < \varepsilon \ , \ \varepsilon > 0 \qquad (۲-۹)$$

دراینجا انعطاف بیشتری برای تعین بردار x هست چون مقدار کمی خطا در حل مساله $Ax \approx b$ پذیرفته شده است. به طوری که اگرx حاوی مقادیر خیلی کوچک باشد می‌توان آن‌ها را صفر کرد.

بنابراین تنک بودن روش‌های محاسباتی در تجزیه ماتریس‌های مشاهده در پردازش سیگنال به معنی این است که ضرایب بیشتری در نمایش صفرشوند تا راه حل مذکور با پیچیدگی و فضای داده کمتری روبرو باشد.

۲-۳- انواع رده‌بندی

۲-۳-۱- روش K نزدیکترین همسایه

روش رده‌بندی K-NN که از نوع رده‌بندهای بدون پارامتر و اصطلاحا بر پایه نمونه (instance based) است و بخوبی برای رده‌بندی هر دو حالت باینری و چند کلاسه طراحی شده است. از

ویژگی‌های برجسته آن این است که نیازی به یک مرحله آموزشی در معنی دقیق نیست. بلکه نمونه‌های آموزشی به طور مستقیم توسط الگوریتم رده‌بند در مرحله رده‌بندی به طور مستقیم استفاده می‌شود. ایده کلیدی این رده‌بندی، این است که اگر مجموعه‌ای از الگوهای (بردار ویژگی‌های ناشناخته)، X، داده شود ما تعداد k نزدیکترین همسایه را در مجموعه آموزش مشخص کرده و تعداد تعلق هر یک از همسایه‌ها به هر کلاس را تعیین می‌کنیم. در پایان، بردار ویژگی به کلاسی اختصاص داده می‌شود که دارای بیشترین تعداد همسایه‌ها از بین انتخاب شده‌ها باشد. بنابراین، جهت استفاده از الگوریتم k-NN تعیین موارد زیر لازم است:

۱- مجموعه داده‌ای از نمونه‌های برچسب شده، که یک مجموعه آموزشی از بردارهای ویژگی بهمراه برچسب‌های داده‌های مربوطه است.

۲- یک عدد صحیح $k \geq 1$.

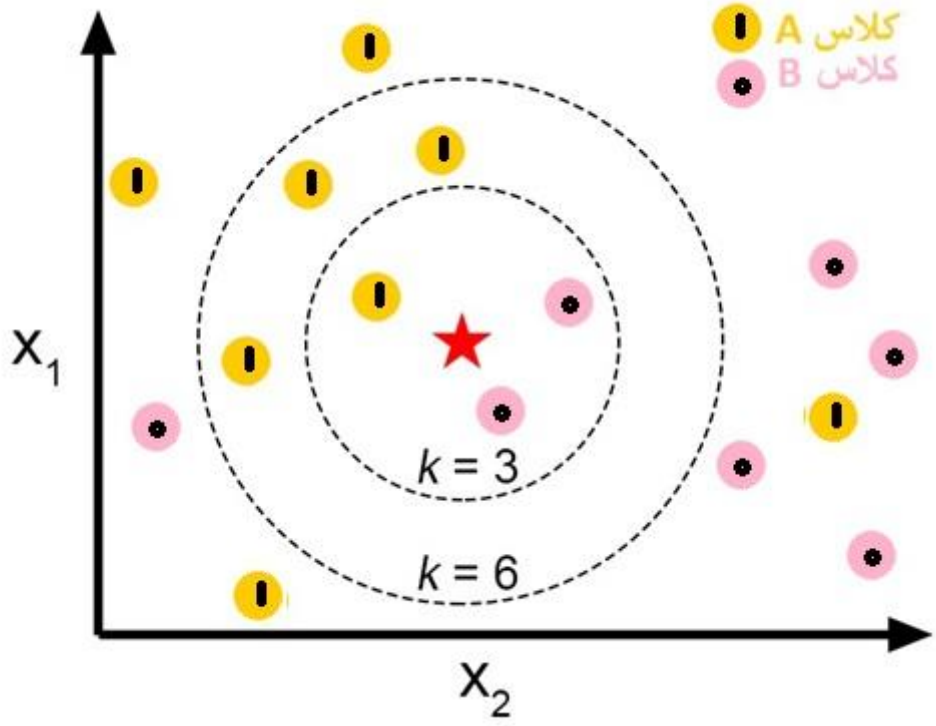
۳- یک تابع جهت محاسبه فاصله.

در ابتدا الگوریتم فاصله‌ی $d(X,V_i)$ بین بردار ناشناخته X و هر بردار $V_i \quad , \quad i=1,...,M$ از مجموعه داده‌های آموزشی که M نمونه از آن در دست است را محاسبه می‌کند. از توابع مختلف فاصله بین بردارها می‌توان به روش اقلیدسی، فاصله کسینوسی، مینکواسکی، چای‌اسکویر، ماهالانوبیس و مانهتن نام برد. روش معروف اقلیدسی در رابطه (۲-۱۰) آورده شده است. عدد D بعد فضای بردارها را مشخص می‌کند.

$$d(X,V_i)=\sqrt{\sum_{j=1}^{D}(X(j)-V_i(j))^2} \qquad (۲-۱۰)$$

پس از آنکه فاصله $d(X,V_i)$ برای هر بردار V_i محاسبه شد فاصله‌های بدست آمده را بصورت صعودی مرتب می‌کنند. در این حالت k مقدار اول لیست مربوط به k همسایه نزدیک به بردار ناشناخته محسوب می‌شوند. اگر k_i را تعداد بردارهای آموزشی از میان k همسایه X که متعلق به کلاس iام هستند و $i=1,...,N_c$ را در نظر بگیریم آنگاه بردار ناشناخته متعلق به کلاسی است که با

حداکثر k_i در ارتباط است. در شکل ۲-۱ نمونه‌ای از رده‌بندی باینری با k-NN نشان داده شده است. نمونه ناشناخته (★) با مقدار k=3 متعلق به کلاس B و با مقدار k=6 متعلق به کلاس A است. نکته مهم در مورد رده‌بندی k-NN این است که به سادگی قابل تعمیم به حالت چند کلاسی می‌باشد. با افزایش نمونه‌های آموزشی برای Mهای بسیار بزرگ و k بزرگ به طوری که $k \ll M$ باشد این رده‌بندی حالت ایده‌آل خود را دارد. اما در عمل داشتن تعداد خیلی زیاد از نمونه تمام کلاس‌ها عملی نیست.



شکل ۲-۱: رده‌بندی باینری در روش k-NN. با k=3 نمونه آزمایش متعلق به کلاس B و با k=6 متعلق به کلاس A است.

۲-۳-۲- رده‌بندی مدل مخلوط گوسی (GMM)

مدل مخلوط گوسی GMM در رده‌بندی کلاس‌های مختلف صوتی استفاده می‌شود و نمونه‌ای از

رده‌بندی پارامتری است. این مدل برای زمانی که ویژگی‌های نمونه‌ها شامل چندین جزء گاوسی است که هر کدام برای مدل کردن ویژگی‌های سیگنال‌های صوتی بکار گرفته شده باشد یک رویکرد خیلی خوب محسوب می‌شود.

در رده‌بندی، هر کلاس توسط یک GMM نشان داده می‌شود و به مدل مربوطه به خود اشاره دارد. بنابراین پس از آموزش GMMهای هر کلاس، می توان از آنها برای پیش‌بینی اینکه یک نمونه مشاهده شده متعلق به چه کلاسی است استفاده کرد.

توزیع احتمال بردارهای ویژگی با شیوه‌های پارامتری یا غیرپارامتری مدل‌سازی می‌شود. معمولا مدل‌هایی که شکل چگالی احتمال را در نظر بگیرند پارامتریک نامیده می‌شوند در حالی که مدل‌سازی غیرپارامتری بدون در نظر گرفتن نحوه توزیع چگالی احتمال بردارهای ویژگی انجام می‌شود. یک مخلوط گاوسی ذاتا برای نمایش یک گروه از کلاس‌های صوتی توسط اجزای گوسی مجزا به طوری که شکل طیفی کلاس‌های صوتی را با بردار میانگین‌ها و ماتریس کوواریانس نشان دهد ظرفیت بالایی دارد و با داشتن مجموعه داده آموزشی حجیم این ظرفیت باز هم تقویت می‌شود.

۱-۲-۳-۲- مدل‌سازی با GMM

از لحاظ ریاضی، GMM به عنوان یک مجموع وزنی از تراکم اجزای گوسی است و مانند رابطه (۲-۱۱) تعریف می‌شود.

$$p(x|\lambda)=\sum_{i=0}^{K} w_i g(x|\mu_i,\zeta_i) \qquad (۲-۱۱)$$

در اینجا، x یک بردار ویژگی با ابعاد D است و $w_i\ ,\ i=0,...,K$ وزن‌های مخلوط است. همچنین چگالی‌های گوسی اجزا با $g(x|\mu_i,\zeta_i)\ ,\ i=0,...,K$ نشان داده شده است. هر جزء چگالی یک تابع گوسی D متغییری به فرم رابطه (۲-۱۲) است.

$$g(x|\mu_i,\zeta_i)=\frac{1}{(2\pi)^{D/2}|\zeta_i|^{1/2}}\exp\{-\frac{1}{2}(x-\mu_i)\zeta^{-1}(x-\mu_i)\} \qquad (۲-۱۲)$$

بردار میانگین‌ها با μ_i و ماتریس کوواریانس با ζ_i نشان داده شده است. جمع وزن‌های مخلوط مجموعا به عدد ۱ می‌رسند لذا شرط رابطه (۲-۱۳) برقرار است.

$$\sum_{i=1}^{K} w_i = 1 \qquad (۲-۱۳)$$

بنابراین مدل مخلوط گوسی به طور کلی با بردارهای میانگین، ماتریس‌های کوواریانس و وزن‌های تمام چگالی‌ها تعریف می‌شود. مجموعه این پارامترها را بصورت رابطه (۲-۱۴) به عنوان مدل لاندا λ نمایش می‌دهند.

$$\lambda = \{w_i, \mu_i, \zeta_i\} \; i = 1, ..., K \qquad (۲-۱۴)$$

هدف آموزش در GMM، محاسبه پارامترهای GMM با استفاده از بردارهای ویژگی آموزشی از طریق مدل پیش‌بینی است. روش‌های مختلفی را می‌توان برای برآورد استفاده کرد اما رایج‌ترین و احتمالا قوی‌ترین برآورد روش تخمین حداکثر احتمال (ML) است. هدف ML یافتن بهترین پارامترهای مدل GMM است و به همین منظور محاسبات شباهت GMM قبلی را تصحیح می‌کند. برای این منظور معمولا از الگوریتم تکراری به حداکثررسانی انتظارات (EM) Expectation Maximization یا حداکثر برآورد پسین Maximum A Posterior Estimation (MAP) جهت برآورد پارامترها برای هر جزء گوسی و وزن مخلوط استفاده می‌شود.

در زمان رده‌بندی کردن نمونه جدید x می‌توان از رابطه (۲-۱۵) احتمال تعلق x به کلاس C_i را مشخص کرد و در بین مقادیر بدست آمده طبق رابطه (۲-۱۶) حداکثر را انتخاب نمود.

$$p(C_i \mid x) = \frac{p(x \mid C_i)}{\sum_{j=1}^{K} p(x \mid C_j)\, p(C_j)} \qquad (۲-۱۵)$$

$$x \in C_i \;\; if \;\; p(C_i \mid x) \; is \; the \; maximum \;\; for \; \{p(C_n \mid x),\, n = 1, ..., K \;\; , \;\; K \geq i\} \qquad (۲-۱۶)$$

۲-۴- تکنیک‌های مختلف پردازش صداهای محیطی

علاوه بر اینکه جهت پردازش یک سیگنال صوتی نیاز به انتخاب یکی از طرح‌های پردازشی مورد اشاره در فوق داریم لازم است تا نوع الگوریتم پردازش خود را نیز با توجه به نوع صداها (برای مثال سیگنال گفتار/ موزیک/ سیگنال محیطی) و نوع کاری که قرار است با داده‌های صوتی سیگنال انجام دهیم انتخاب نماییم. با این نوع رویکرد و لزوم اتخاذ الگوریتمهای کاراتر برای پردازش صداهای محیطی و با مطالعات صورت گرفته به این جمع بندی می‌رسیم که تکنیک‌های مختلف پردازش صداها را می‌توان در سه گروه مجزا به صورت زیر دسته‌بندی نمود:

۱) تکنیک‌های پایه‌ای پردازش صداهای محیطی.

۲) تکنیک‌های پردازش صداهای صوتی ساکن.

۳) تکنیک‌های پردازش صداهای غیر ساکن.

۲-۴-۱- تکنیک‌های پایه‌ای پردازش صداهای محیطی

همانطور که اشاره شد در بین انواع مختلف سیگنال‌های صوتی، بر روی موزیک و گفتار به طور گسترده و جامع مطالعه صورت گرفته است. به همین دلیل الگوریتم‌های پایه ای پردازش صداهای محیطی نیز بیش‌تر مشابه روش‌ها و الگوریتم‌های بکار رفته در پردازش سیگنال‌های موزیک و گفتار هستنند. اما با توجه به ویژگی‌های بسیار غیرساکن صداهای محیطی استفاده از این الگوریتم‌ها در پایگاه داده‌های خیلی بزرگ عملا ناکارآمد است [۸]. به عنوان مثال در شناسایی گفتار از ساختار فونوم‌ها که بلاک‌های اصلی سازنده گفتار هستند استفاده می‌شود و وجود این ساختارها این امکان را ایجاد می‌کند تا کلمات پیچیده گفتاری را به فونوم‌های اولیه تجزیه کرده و هر کدام را با یک HMM مدل سازی نمود [۱۲]. اما صداهای محیطی اغلب فاقد زیر ساختارهایی همچون فونوم‌ها هستند و حتی در صورتی که بتوان یک دیکشنری از واحدهای اولیه (شبیه فونوم‌ها در گفتار) را تشکیل داده و آموزش داد، باز هم مدل‌سازی تغییرات زمانی آن‌ها با HMM مشکل است چرا که وقوع زمانی آن‌ها کاملا اتفاقی است و این بر خلاف سیگنال گفتار است که فونوم‌ها یک ترتیب و توالی از پیش تعیین

شده ای دارند. همچنین در مقایسه با صداهای موزیکال و ریتم‌های موسیقایی، صداهای محیطی فاقد نمونه‌های ثابت معناداری همچون ملودی، هارمونی و ریتم هستند [۱۳]. این موضوع سبب می‌شود تا صداهای محیطی را نتوان با تکنیک‌ها و الگوریتم‌های مورد استفاده در صداهای موزیک مورد پردازش قرار داد. با این وجود تحقیق بر روی شناسایی صداهای محیطی یک الزام برای سیستم‌های مختلف در محدوده کاربردهای شنوایی هوشمند است که با وجود تعاملات زیاد بین انسان و محیط پرداختن به آن‌ها ضروری و اجتناب ناپذیر است و تمرکز بر شناسایی و تفسیر اتوماتیک آن‌ها نیازمند الگوریتم‌ها و رویکردهای خاص است. لذا اکثر مطالعات اخیر بر روی جنبه‌های غیر ساکن صداهای محیطی متمرکز است و سعی بر این است تا با تعریف ویژگی‌های جدید محتوای اطلاعاتی مربوط به جنبه‌های طیفی و زمانی سیگنال‌های محیطی را تا حد امکان افزایش دهند هرچند که عدم قطعیت نیز در نتیجه کار نقش دارد. در اغلب صداهای دنیای واقعی این ویژگی‌ها حتی در پریودهای زمانی طولانی وضعیت غیر ساکن بودن خود را حفظ می‌کنند علاوه بر این، روش‌های یادگیری متوالی برای بدست آوردن تغییرات طولانی مدت صداها مورد استفاده قرار گرفته است. لذا برای استخراج این تغییرات دراز مدت روش‌های یادگیری متوالی بکار گرفته می‌شود. مشخص است که روش‌های شناسایی صداهای محیطی علاوه بر اینکه باید ویژگی‌های غیر ساکن صداها را مدل سازی کنند باید با وجود رسته‌های خیلی زیاد صداهای محیطی مقیاس پذیر و مقاوم هم باشند.

از تحقیقات گذشته در این حوزه می‌توان به مطالعه [۱۴] اشاره نمود که سسیستمی را مبتنی بر رده‌بندی One-SVM[۱] برای شناسایی اتفاقات صوتی ناهنجاریهایی همچون صدای شلیک تفنگ، شکستن شیشه و صدای داد و فریاد(جیغ) ارائه نموده‌اند. در [۱۵] یک مدل تطبیق پذیر برای کشف وقایع صوتی کوتاه مدت و دراز مدت و همچنین برای صداهایی که در زمینه خود نویز مشابه با صدای اصلی دارند پیشنهاد کردند. همچنین [۱۶] در یک مدل پیوسته دو پشته[۲] مزایای مدل‌سازی متوالی را با قابلیت‌های جداسازی پرسپترون چند لایه ترکیب نموده و با اجرای یک مرحله رتبه‌بندی مجدد

[۱]- Support Vector Machine(SVM)
[۲]- Tandem connectionist model

(همانند آنچه در بازشناسی گوینده انجام می‌شود) کارایی رده‌بندی را از طریق بکارگیری مدل GMM-SVM تقویت کرده است.

در تحقیق [۱۷] به شناسایی مشخصات صدای شلیک تفنگ پرداخته شده و اطلاعاتی را جهت ارائه به مسئولین از واقعه صوتی رخ داده مهیا ساخته است. در [۱۸] از ترکیب اطلاعات ویدیو و تصویر با بکارگیری مدل زوج HMM یکی برای اطلاعات تصویر و دیگری برای صدا جهت شناسایی وقایع صوتی در یک اتاق ملاقات مجهز به تجهیزات چند رسانه‌ای استفاده شده است. اما اخیرا روش‌هایی بر پایه مدل پیوسته سرهم بندی دو مرحله‌ای پیشنهاد شده است. از جمله در [۱۹] با هدف استخراج ویژگی‌های دارای قابلیت جدا کنندگی بالا رویکرد بوستینگ[۱] پیشنهاد شده است و نتایج حاصل نیز نسبت به ویژگی‌های ادراکی سنتی همچون ضرایب مل فرکانس[۲] و لگاریتم فرکانس بانک‌های فیلتر[۳] بهتر بوده است. این نوع مدل‌سازی را آماری هرمی[۴] می‌گویند.

در مقاله [۲۰] که در زمان خود بهترین نتایج را داشته است استفاده از روش انتخاب ویژگی بر پایه آدابوست[۵] با کمک معیار فاصله کولبک لیبلر[۶] برای اندازه‌گیری قابلیت جداسازی هر ویژگی در هر اتفاق صوتی را پیشنهاد داده است و پس از آن HMM را استفاده نموده است. همچنین در [۲۱] به منظور رده‌بندی صحنه‌های شنوایی اقدام به شناسایی و ایندکس کردن معنایی وقایع صوتی در هر صحنه شنوایی نموده‌اند و با استفاده از الگوریتم BIC[۷] برای تخمین اتوماتیک تعداد خوشه‌ها و تخصیص یک واقعه صوتی به یک صحنه شنوایی استفاده کرده‌اند. در تحقیق [۲۲] که به منظور شناسایی ۹ نوع از صداهای محیطی برای کاربرد نظارتی و امنیتی انجام گرفت از ویژگی‌های بر پایه ویولت[۸] و از مدل چند کلاسه سازی SVM[۹] به همراه یک معیار عدم شباهت استفاده کردند.

۱- Boosting
۲- Mel-frequency Cepstral Coefficients (MFCC)
۳- Log frequency filterbank
۴- Leveraging statistical models
۵- Adaboost
۶- Kullback-Leibler
۷- Bayesian information criterion (BIC)
۸- Wavelets
۹- SVM-based multiclass classification approach

در [۲۳] با استفاده از یک تخمینگر بیز[۱]بهترین مجموعه ویژگی‌ها را شامل ۷۸ ویژگی (۲۶*۳) برای کشف وقایع صوتی با استفاده از رده‌بندی HMM مورد استفاده قرار داده‌اند. تحقیق [۲۴] از یک مدل سلسله مراتبی برای کشف وقایع صوتی بمنظور نظارت بر محیط استفاده کرده است. آن‌ها ابتدا فریم‌ها را به حنجره‌ای[۲]و غیرحنجره‌ای[۳]و در مرحله بعدی به وقایع نرمال و آشفتگی تقسیم می‌کنند. موضوع شناسایی صداها در یک رشته صوتی جهت خلاصه سازی ویدیو و استخراج صحنه‌های برجسته در [۲۵] انجام شده است. در تحقیق [۲۶] رکوردهای صوتی مربوط به فوتبال را برای شناسایی ۶ نوع از صحنه‌های خاص در فوتبال با استفاده از مدل بیز و ویژگی‌های MFCC[۴]به همراه HMM انجام داده‌اند. در [۲۷] آنالیز محتوایی جهت قطعه‌بندی و رده‌بندی رکوردهای صوتی بکار گرفته شده است.

در روش مقاله [۲۸] ابتدا سیگنال را توسط الگوریتم KNN[۵]و LSP-VQ[۶] به قطعه‌های گفتار و غیرگفتار تقسیم کرده و در مرحله بعد با یک الگوریتم مبتنی بر قوانین[۷]رده‌بندی قطعه‌های غیرگفتار را به موزیک- صداهای‌محیطی-سکوت انجام داده‌اند. در [۲۹] با استفاده از مدل نیمه نظارتی بیز وقایع صوتی در رکورد ورودی را رونویسی کرده و وقایع چندگانه همزمان را مورد پردازش قرار داده است. تحقیق [۳۰] از روش جداسازی سیگنال‌ها برای تجزیه سیگنال ورودی به وقایع استفاده کرده است. مدل های دیگری از جمله [۳۱، ۳۲] استفاده شده است تجزیه نامنفی ماتریس مشاهدات است که سیگنال ورودی را به دو ماتریس نامنفی تجزیه نموده و برای سیگنال آزمایش از یک دیکشنری وقایع استفاده می‌کند.

در [۳۳] وصله‌های طیفی از پیش استخراج شده را برای تجزیه سیگنال به وقایع مورد استفاده قرار داده است و همچنین از تعداد زیادی ویژگی حوزه زمان و فرکانس استفاده کرده است. پژوهش [۳۴] یک مدل احتمالاتی محدود زمانی را برای کشف وقایع استفاده نموده است. در [۳۵] صحنه‌های

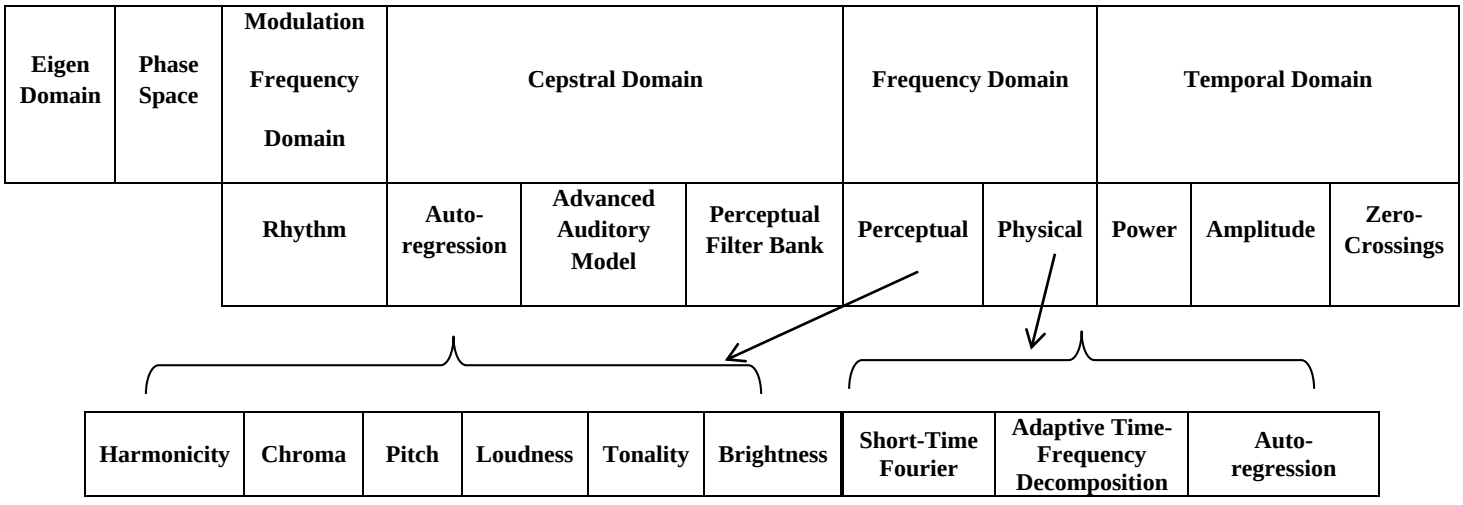
۱- Bayes Estimator
۲- Vocal
۳- Non-vocal
۴- Mel-Frequency Cepstral Coefficients (MFCCs)
۵- k-Nearest Neighbors algorithm (k-NN)
۶-Line Spectral Pair Vector Quantization(LSPVQ)
۷- Rule-based

شنوایی را از طریق شناسایی صداهای موجود در آن تشخیص میدهند. در [۳۶] برای شناسایی وقایع از شبکه های عصبی عمیق همراه با مدل‌سازی جهت حذف نویز استفاده شده است. همچنین [۳۷] در یک مدل پیوسته دو پشته مزایای مدل‌سازی متوالی را با قابلیت‌های جداسازی پرسپترون چند لایه 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 کرده است.

در رویکرد دسته فریمی [۳۵] برای هر صدا یک مدل GMM آموزش داده می‌شود آنگاه تخمینی از مجموع GMMهای تمام فریم‌ها محاسبه شده و رده‌ای که دارای بالاترین شباهت باشد انتخاب می‌شود. رویکرد ویژگی‌های دسته‌ای نیز در [۳۸، ۳۹] پیشنهاد شده است که هیستوگرامی متشکل از ویژگی‌های خوشه بندی شده را برای رده بندی استفاده می‌کند.

۲-۴-۲- روش‌های شناسایی صداهای محیطی ساکن

برای شناسایی صداهای محیطی ساکن به طور سنتی از ویژگی‌های بسط یافته برای کاربردهای گفتار/موزیک استفاده شده است. این ویژگی‌ها اغلب بر اساس خواص صوتی/روانی صداها همچون بلندی، گام، طنین و غیره پایه ریزی شده است. یک نوع رده‌بندی ویژگی‌های صوتی را که در [۴۰] معرفی شده است می‌توان مطابق شکل ۲-۱ در نظر گرفت.



شکل۲-۱: رده‌بندی ویژگی‌های صوتی معرفی شده در [۴۰].

۱- Tandem Connectionist Model

محاسبه ویژگی‌هایی همچون ZCR[1], STE[2], SBER[3], SF[4] اندازه‌های کلی در مورد محتوای فرکانسی و زمانی سیگنال‌های صوتی را مشخص می‌کند.

ویژگی‌های کپسترال شامل (MFCC), (ΔMFCC[5], ΔΔMFCC), HCC[6], BFCC[7] بر اساس شبیه سازی سیستم شنوایی انسان کار می‌کنند و در کاربردهای گفتار و موزیک به خوبی مورد استفاده قرار گرفته‌اند. همانطور که قبلا اشاره شد بدلیل فقدان پایگاه داده استاندارد برای شناسایی صداهای محیطی، لذا محققان از MFCC برای محک زدن الگوریتم‌های خود استفاده می‌کنند. یکی از روش‌های بهبود نتایج استفاده همزمان ویژگی‌های MFCC با سایر ویژگی‌های پیشنهادی است [۴۱، ۴۲].

ویژگی‌های مبتنی بر MPEG-7[8] نیز کاربرد گسترده‌ای در کاربردهای گفتار/موزیک دارند. این ویژگی‌ها دارای پیچیدگی محاسباتی کم بوده و شامل خواص ادراکی و روانی صداها نیز هستند. ونگ و همکاران در [۴۳] استفاده از توصیفگرهای صوتی سطح پایین همچون ASC[9] و ASF[10] با پیوندی از رده‌بندهای متشکل از SVM و KNN را پیشنهاد دادند. آن‌ها خروجی‌های رده‌بندهای SVM و KNN را به مقدارهای احتمالی تبدیل کرده و با آمیختن آن‌ها به یکدیگر میزان درستی رده‌بندی را افزایش دادند. در [۴۴] ترکیب چندین توصیفگر سطح پایین MPEG-7 و MFCC را بکار گرفتند و از نسبت جداساز فیشر[11] برای حذف ویژگی‌های نامرتبط استفاده نمودند. آن‌ها نشان دادند که هر چند ویژگی‌های MPEG-7 نسبت به MFCC بهتر عمل می‌کنند اما استفاده همزمان از توصیفگرهای MPEG-7 و MFCC یکدیگر را کامل کرده و میزان صحت رده‌بندی را افزایش می‌دهند.

---

[1]- Zero-Crossing Rate (ZCR)
[2]- Short-Time Energy (STE)
[3]-Sub-band Energy Ratio (SBER)
[4]- Spectral Flux (SF)
[5]- Delta-MFCC
[6]- Homomorphic Cepstral Coefficients (HCC)
[7]- Bark-Frequency Cepstral Coefficients (BFCC)
[8]- Multimedia content description interface (The Moving Picture Experts Group (MPEG))
[9]- Audio Spectrum Centroid (ASC)
[10]- Audio Spectrum Flatness (ASF)
[11]- F-Ratio

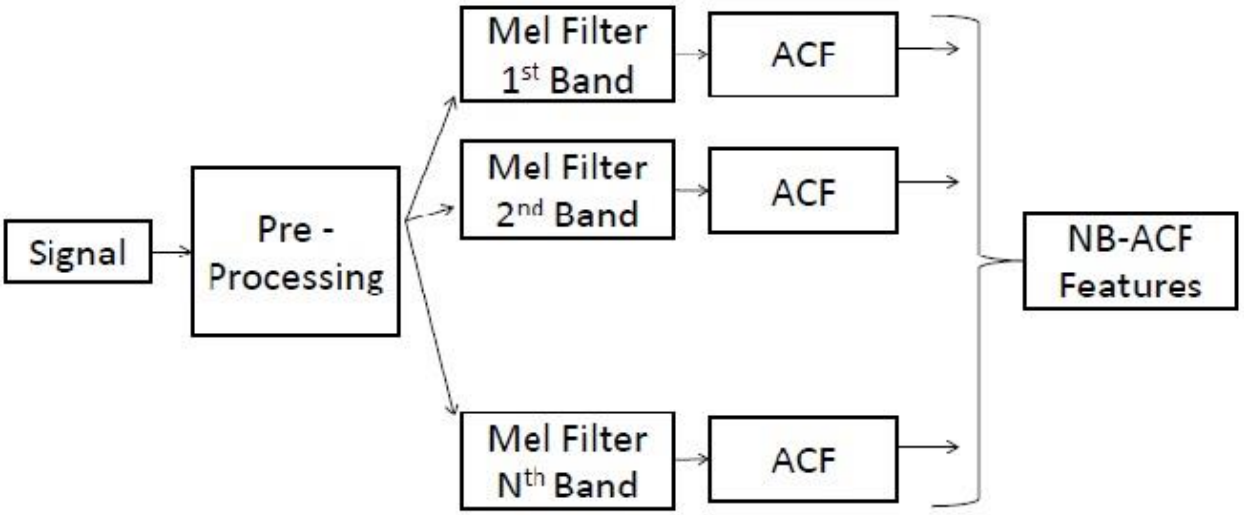
بکارگیری ویژگی‌های مبتنی بر Auto-regression بخصوص LPC[1]در پردازش سیگنال‌های گفتار رایج بوده و سابقه دیرینه دارند. ضرایب کپسترال پیشبینی خطی (LPCC[2]) که نمایش دیگری از LPC هستند نیز به طور گسترده‌ای استفاده شده‌اند. با این وجود LPC و LPCC برای مدل فیلتر- منبع گفتار[3] استفاده می‌شوند و چندان مناسب شناسایی صداهای محیطی نیستند. تسا و همکاران در [۴۵] ویژگی‌های مبتنی بر CELP[4]به همراه LPC, pitch و pitch gain را پیشنهاد دادند. از آنجا که CELP از کدبوک ثابتی برای تحریک مدل فیلتر- منبع استفاده می‌کند، لذا از LPC قوی‌تر و مقاوم‌تر است. نتایج آن‌ها نشان دهنده کارایی بهتر نسبت به MFCC است. استفاده همزمان CELP و MFCCسبب افزایش دقت رده‌بندی می‌شود. بخصوص این بهبودی در مورد کلاس‌های صوتی همچون rain, stream, thunder که شناسایی آن‌ها مشکل است بیش‌تر دیده شده است.

الگوریتم‌های شناسایی صداهای محیطی مبتنی بر طرح پردازش زیر-فریم معمولا مدل سیگنال را در زیرفریم یاد می‌گیرند و لذا از ساختار زمانی بهره برداری نمی‌کنند. جهت به خدمت گرفتن ساختار زمانی، لازم است که مدل سیگنال را بر اساس ویژگی‌های تمام زیر-فریم‌های مرتب شده در قالب مدل‌هایی همچون HMM بدست آورد [۴۶، ۱۹] مدل دیگری که اخیرا در [۴۷] پیشنهاد شده است سعی دارد تا تغییرات زمانی در طول زیر-فریم‌ها را توسط مجموعه ای از ویژگی‌های جدید بنام (SDF[5]) بدست آورد. نشان داده شده است که تحت سه نوع رده‌بندی KNN, GMM[6], SVM ویژگی‌های SDF نسبت به چندین ویژگی متداول از جمله ZCR, LPC, MFCC کارایی بهتری دارد. همچنین مشخص شده است که حداکثر کارایی ویژگی‌های استاتیک وقتی بدست می‌آید که ترکیبی از ویژگی‌های MFCC, ΔMFCC استفاده شده باشد به طوری که افزودن سایر ویژگی‌های متداول تاثیری در افزایش کارایی ندارد. با استفاده از ویژگی‌های دینامیک SDF یک بهبود کیفیت ۱۰ الی ۱۵ درصدی نسبت به حداکثرهای نوع استاتیک بدست آمده است.

---

[1]- Linear predictive coding (LPC)
[2]- Linear. Prediction Cepstral Coefficient (LPCC)
[3]- Source-filter model for Speech
[4]- Code Excited Linear Prediction
[5]- Spectral Dynamic Features
[6]- Gaussian-Mixture Model (GMM)

از بانک‌های فیلتر اغلب برای استخراج ویژگی‌های خاص باندهای کوچک استفاده می‌شود و خواص طیفی را به طور موثری جمع آوری می‌کنند. از سوی دیگر، تابع همبستگی (ACF[1]) نمایش تکامل زمانی را در خود دارد و یک ارتباط مانوس‌تری با چگالی طیف توان (PSD[2]) سیگنال تحت بررسی دارد. والرو و آلیاس در [۴۸] ویژگی‌های جدیدی بنام توابع همبستگی باند-باریک (NB-ACF[3]) پیشنهاد دادند. نحوه استخراج ویژگی‌های NB-ACF در شکل ۲-۲ توضیح داده شده است.



شکل ۲-۲: عملیات مختلف مورد نیاز برای استخراج ویژگی‌های NB-ACF [۴۸]

در اینجا ابتدا سیگنال از یک بانک فیلتر عبور داده می‌شود که دارای N=۴۸ باند فرکانسی است و مراکز باندهای فرکانسی با اندازه‌های مل مطابقت داده شده‌اند. پس از آن ACF سیگنال فیلتر شده برای باند iام محاسبه و به صورت $\Phi i(\tau)$ نشان داده می‌شود. می‌توان چهار نوع از ویژگی‌های NB-ACF را برحسب هر ACF به صورت زیر محاسبه نمود:

$\Phi i(\cdot)$ انرژی در $\tau=0$: این ویژگی مقیاسی است برای میزان فشار صدای درک شده در باند iام.

$\tau i1$: میزان تاخیر اولین قله که بیانگر فرکانس چیره در iامین باند است.

$\Phi\ i1\ (\tau i1)$: مقدار ACF نرمال شده اولین قله مثبت را مشخص می‌کند. این پارامتر با دوره تناوب سیگنال مرتبط است و گام سیگنال فیلتر شده را در باند iام مشخص می‌کند.

[1]- Auto-Correlation Function
[2]- power spectral density
[3]- Narrow-Band Auto Correlation Function features

τ_{ie} : دوره ماندگاری موثر پوشش ACF نرمال شده را تعیین می‌کند. میزان طنین دار بودن سیگنال فیلتر شده در باند iام را مشخص می‌کند.

در رابطه با طول زمانی فریم و زیر-فریم آن‌ها پیشنهاد داده‌اند که زیر-فریم‌ها با طول ۵۰۰ میلی‌ثانیه و با میزان شیفت ۱۰۰میلی‌ثانیه و طول فریم ۴ ثانیه مناسب است. تصمیم‌گیری در هر زیر-فریم نیز توسط رده‌بندهای KNN و SVM انجام شده است. کارایی ویژگی‌هایNB-ACF با MFCC و DWT[۱] بر روی مجموعه داده‌هایی از ۱۵ صحنه محیطی مقایسه گردیده است. صحنه‌هایی که تغییرات دینامیکی بیش‌تری دارند مثل office, library, Classroom از ویژگی‌های NB-ACF بیش‌تر استفاده کرده‌اند.

۳-۴-۲- روش‌های شناسایی صداهای محیطی غیرساکن

۱-۳-۴-۲- روش‌های مبتنی بر ویولت

در [۴۹] شناسایی سه نوع داده گفتار، موزیک و صداهای محیطی توسط رده‌بندهایی همچون LVQ[۲]، GMM، DTW[۴]، ANN[۳] و با استفاده از ویژگی‌های متعارفی مانند MFCC, HCC[۵]، STFT[۶]، DWT[۷]، CWT[۸] انجام شده و نتایج مقایسه گردیده است. آن‌ها همچنین از طرح فریمی برای شناسایی تعداد ۸ رده از صداهای محیطی استفاده کرده‌اند. گزارش شده است که در مورد رده‌بندی صداهای محیطی، بهترین کارایی از ویژگی‌های CWT (که نمایش زمان-فرکانس را ارائه می‌کند) به همراه رده‌بندی DTW بدست آمده است و این نتایج با بکارگیری ویژگی‌های MFCC به همراه رده‌بندی DTW قابل مقایسه است. اما در مورد DWT و STFT کارایی مناسبی گزارش نشده است. آن‌ها همچنین گزارش کرده‌اند که به علت کوچک و محدود بودن پایگاه داده مورد استفاده امکان بررسی وجود تفاوت معناداری بین کارایی MFCC و CWT وجود نداشته است. چنانچه عوامل دیگر را

۱- Discrete Wavelet Transform
۲- Learning Vector Quantization
۳- Artificial Neural Networks (ANN)
۴- Dynamic Time Warping (DTW)
۵- Homomorphic Cepstral Coefficients
۶- Short-Time Fourier Transform (STFT)
۷- Discrete Wavelet Transform
۸- Continuous Wavelet Transform

مساوی فرض کنیم بکارگیری ویژگی‌های MFCC به دلیل پیچیدگی محاسباتی کم‌تری که دارند نسبت به ویژگی‌های CWT برتری دارند. از بین رده‌بندهای مختلف نیز DTW نسبت به سایرین برتری محسوسی داشته است.

در [۵۰] توابع اصلی گاماتون را با ویولت تطبیق داده و مربع مجموع گاماتون‌های سیگنال را به عنوان ویژگی استفاده کرده‌اند و آن‌ها را ویژگی‌های ویولت گاماتون نامیده‌اند. این ویژگی‌ها برتری محسوسی نسبت به ویژگی‌های DWT داشتند. ویژگی‌های گاماتون در مورد رده‌های صوتی همچون footsteps و gunshots بهتر عمل می‌کنند و دلیل آن هم قابلیت این ویژگی‌ها در مشخص نمودن حالت گذرای صداها است. جدای از اینکه ویولت ویژگی‌های زمان-فرکانس را ارائه می‌کند، کارایی ویژگی‌های ویولت بهتر از ویژگی‌های MFCC نبوده است بلکه این دو نوع ویژگی‌ها با هم قابل مقایسه بوده و در یک سطح هستند.

۲-۳-۴-۲- روش‌های مبتنی بر تطابق الگو

در [۵۱] پیشنهاد استفاده از ویژگی‌های MP[۱] در شناسایی صداهای محیطی را عنوان کردند. الگوریتمMP پایه که BMP[۲] نامیده می‌شود یک الگوریتم نوع حریصانه برای بدست آوردن نمایش تنک یک سیگنال بر اساس اتم‌های یک دیکشنری است. آنچه می‌توان از بکارگیری این الگوریتم نتیجه گیری کرد آن است که ویژگی‌های MP که قابلیت استخراج اطلاعات از نمایش زمان-فرکانس با تفکیک‌پذیری بالا را دارند کارایی شناسایی را افزایش می‌دهند و همچنین کارایی در زمانی که مدل‌های یادگیری HMM بکار گرفته می‌شوند افزایش می‌یابد.

۳-۳-۴-۲- روش‌های مبتنی بر انرژی طیف [۳]

اسپکتروگرام اطلاعات مفیدی را در قالب بعد زمان و نواحی فرکانس ارائه می‌کند و به همین دلیل ابزاری است که مشخصه های گذرا و متغییر صداهای محیطی را می‌تواند مستقیما در معرض نمایش قرار دهد. اما ابعاد زیاد ویژگی‌های اسپکتروگرام بکارگیری آنرا در آموزش مدل‌ها سخت می‌کند.

۱- Matching Pursuit
۲- Basic Matching Pursuit
۳- Power-Spectrum

مطالعات خونارسال و همکاران در [۵۲] حکایت از برتری ویژگی‌های اسپکتروگرام نسبت به MFCC, LPC دارند و نتایج آن با ویژگی‌های MP-Gabor قابل مقایسه و تقریبا یکسان است. از طرف دیگر ترکیب ویژگی‌های اسپکتروگرام، LPC و MP-Gabor بهترین نتایج را تولید کرده‌اند اما نتایج رده‌بندی در سایر ترکیبات نیز با بهترین حالت قابل مقایسه بوده‌اند که این موضوع بیانگر وجود ویژگی‌های اضافه و هرز در ترکیبات است. مطالعات دیگری نیز در زمینه کشف وقایع از محیط‌های ترکیبی که ویدیو و صدا همزمان حضور دارند انجام شده است. نمونه ای از این تحقیقات در مقاله [۵۳] اشاره شده است که در آنجا از اطلاعات تصویری برای تکمیل شناسایی استفاده شده است. در [۵۴] بر اساس واحدهای مشخص صوتی که سیستم آموزش دیده است می‌تواند اتفاقات همخوان با آن واحدها را شناسایی و کشف کند. در [۵۵] کشف از طریق اطلاعات و ویژگی‌های دو بعدی که از حوزه زمان –فرکانس بدست می آید انجام می‌شود. در [۳۰] از تکنیک‌های موجود در جداسازی منابع سیگنال برای کشف اتفاقات صوتی استفاده شده است. در مطالعه [۵۶] پیوند میان تکنیک‌های مقاوم سازی در سیستم‌های تشخیص گفتار و استفاده از آن‌ها در شناسایی اتفاقات صوتی بررسی شده است. در مقاله [۵۷] شناسایی صحنه صوتی از طریق رده‌بندی صداهای مختلف در محیط انجام شده است.

۵-۲- کاربرد شناسایی صداهای محیطی در آنالیز صحنه‌های شنیداری

یکی از کاربردهای مهم و بدیهی صداهای محیطی استفاده از آن‌ها برای پردازش و آنالیز صحنه‌های شنیداری است. این نوع پردازش عمدتا شامل دو نوع عملیات کلی است که عبارتند از :

۱-کشف وقایع صوتی در صحنه.

۲- تعیین هویت و رده‌بندی وقایع اکتشافی به برچسب‌های مجزا.

هر یک از عملیات‌های فوق نیاز به الگوریتم‌ها و ابزارهای ویژه خود دارد. در مورد رده‌بندی می‌توان از رده‌بندهای موجود مانند SVM, KNN, GMM, HMM (جایی که تعداد نمونه‌های آموزشی کم است) و یا سایر موارد استفاده کرد و چنانچه تعداد ویژگی‌های متغییر داشته باشیم می‌توان عمل

خوشه‌بندی را با ماتریس ترکیب[1] انجام داد [۵۸]. از آنجا که کار با بردارهای ویژگی طول ثابت از محدودیت‌های SVM است، می‌توان به جای آن از DTW در حوزه زمان استفاده کرد و از جمله ابتکاراتی که در مرحله رده‌بندی می‌توان بکار گرفت ترکیب نمودن خروجی چندین سیستم رده‌بندی مختلف با روش‌های خاص تلفیق است.

اما کشف وقایع صوتی، نسبت به رده‌بندی آن‌ها موضوع پیچیده‌تر و حساس‌تری است چون در اینجا علاوه بر نیاز به مدل سازی وقایع باید محدوده زمانی رخ داد هر یک از وقایع نیز دقیقا مشخص شود بنابراین اطلاعات یک اتفاق صوتی بخشی در حوزه فرکانس و بخش دیگر در حوزه زمان بدست می‌آید. تفاوت قابل توجه دیگر آن است که اگر وقایع مشخص شده باشند رده‌بندی صوتی باز هم ساده‌تر می‌شود و این به معنای آن است که کشف وقایع کار رده‌بندی را تا حد زیادی به پیش می برد. یک نوع رده‌بندی دیگر هم در مورد سیگنال صوت امکان پذیر است و آن رده‌بندی یک سیگنال صوتی به صورت یک عملیات کلی و یکباره بدون کشف وقایع درون آن است که همراه با عملیات فریم گذاری و دنبال کردن متوالی تغییرات و به هم پیوند دادن آن‌ها در طول سیگنال صوتی انجام می‌شود که این روش را حتی برای جداسازی دو سیگنال از هم نیز می‌توان بکار برد. این در حالی است که کشف وقایع با جزئیات و بخش‌های خیلی کوتاه سیگنال صوتی در ارتباط است و همین امر عملیات را بسیار پیچیده‌تر می‌کند [۵۹].

نمونه‌ای از تلاش‌های متمرکز برای طرح و حل مسئله کشف وقایع و رده‌بندی آن‌ها، یک گردهمایی ارزیابی بین المللی تحت عنوان CLEAR[2] است که در سال‌های ۲۰۰۶ و ۲۰۰۷ برگزار گردید و در آن همایش پایگاه داده‌های مختلفی در اختیار شرکت کنندگان قرار گرفت. در آنجا از دو نوع پایگاه داده مختلف برای اتفاقات صوتی مجزا و یک پایگاه داده برای سمینارهای حضوری که شامل تعداد زیادی اتفاقات صوتی مورد نظر بود استفاده شد. در آنجا یکی از اتفاقات مهمی که در سیگنال‌های صوتی مورد شناسایی قرار گرفته است تعیین بخش‌های گفتاری سیگنال از سایر بخش‌ها

[1]- Confusion matrix
[2]- Classification of Event, Activities, and Relationships(CLEAR)

است . نتایج این سیستم‌ها با پایگاه داده‌های RT[1] که با اسامی RT01,RT02,… معرفی شده‌اند مورد ارزیابی قرار گرفته است. یک نمونه طراحی شده بر پایه SVM که با پایگاه داده RT06 مورد ارزیابی قرار گرفته است کارایی بهتری نسبت به نمونه GMM داشته است [۶۰]. علاوه بر این چارجوب‌های کاری مشخصی نیز مانند پروژه CHIL EU[2] وجود دارد که طراحان سیستم بر اساس این چارچوب-ها سیستم‌های خود را طراحی می‌کنند. نکته مهم دیگر در طراحی این سیستم‌ها آن است که برخلاف استقبال روز افزون از این زمینه تحقیقاتی هنوز پایگاه داده مشخصی برای این زمینه وجود ندارد و به همین دلیل محک زدن الگوریتم‌های جدید به صورت مقایسه ای و در زمینه خاص امکان پذیر است [۶۱].

۶-۲- کاربردهای عملیاتی از پردازش صدا

۱-۶-۲- ایندکس‌گذاری و بازیابی صوتی

در [۶۲] رده‌بندی صدای حیوانات به روش بازیابی شنیداری-مفهومی انجام شده است. در اینجا ابتدا فضاهای مفهومی و صوتی خوشه بندی شده و سپس میزان احتمال پیوند بین مدل‌ها برقرار شده است. خوشه بندی صوتی با ویژگی‌های بدست آمده از ضرایب مل کپسترال انجام شده است و هر خوشه را با یک مدل مخلوط گوسی مشخص کرده است. روش دیگری که برای فراگیری ارتباط بین مدل صوتی و مفهومی بکار گرفته شده است، مخلوط احتمال است.

سیستم رده‌بندی ، جستجو و بازیابی بر اساس محتوای شنوایی نیز توسط [۶۳] ارائه شده است. این سیستم یکی از اولین سیستم‌های رده‌بندی شنوایی محسوب می‌شود و به نام سیستم ماهی صدفی ثبت شده است. رده‌بندی بر اساس فاصله اقلیدسی میان بردارهای ویژگی که متشکل از میانگین، واریانس و ضرایب خودهمبستگی بین فریم‌ها بودند انجام می‌شود.

در [۶۴] کارمشابهی روی همان پایگاه داده به طور موثرتری توسط طرح درخت باینری که هر گره آن یک SVM است انجام شده و بازیابی بر اساس فاصله از حاشیه ادراکی انجام گردیده است. در این

[1]- NIST Rich Transcription
[2]- Computers in the Human Interaction Loop(CHIL)

سیستم از تکنیک الحاق ویژگی‌های کپسترال و مفهومی و همچنین رده‌بندی با SVM استفاده شده است.

در [۶۵] نویسندگان از دو تکنیک رده‌بندی SVM و GMM برای ایندکس صوتی استفاده کرده اند. آن‌ها تمایز بین گفتار و موزیک در برنامه های رادیویی و همچنین تمایز بین صداهای محیطی خنده و تشویق (کف زدن) را در برنامه های تلوزیونی انجام دادند.

در [۶۶] برای اکتشاف و مقوله بندی محتوای مفهومی در یک جریان شنوایی ترکیبی از یک رویکرد غیر نظارتی استفاده کردند. آن‌ها ابتدا خوشه بندی طیفی را به منظور کشف خوشه‌های صداهای مفهومی طبیعی انجام دادند. آنگاه بر اساس اجزای شنوایی استخراج شده صحنه‌های شنیداری را رسته‌بندی کردند.

۲-۶-۲- شناسایی شنیداری در محیط‌های خاص

اخیرا تمایل زیادی برای شناسایی و رده‌بندی صداهای مربوط به یک محیط خاص بوجود آمده است. مثلا اتاق ملاقات، اتاق سخنرانی، کلینیک یا بیمارستان، استادیوم ورزشی، پارک‌ها، آشپزخانه، کافی شاپ و غیره. در [۶۷] شناسایی خنده در ملاقات‌ها توسط استخراج ویژگی‌های MFCC و بکارگیری SVM انجام شده است و مشخص شد که ۶ ضریب اولیه کپسترال بیش‌ترین اطلاعات را برای رده‌بندی در بر می‌گیرند.

در [۶۸] شناسایی یک تاکید[۱] به منظور رده‌بندی ملاقاتهای ضبط شده پیشنهاد گردید. رویکرد این سیستم فقط بر اساس اطلاعات گام برای تعیین سخنرانیهای مورد علاقه بنا شده است. به غیر از محیط های ملاقات، رده‌بندی صداها برای محیط‌های پزشکی نیز انجام شده است. در [۶۹] از یک سیستم رده‌بندی برای آنالیز صدای مته ها در زمینه جراحی ستون فقرات استفاده کرده اند. برای اینکه جراح دقت لازم را داشته باشد، این سیستم اطلاعاتی را در رابطه با فشردگی و استحکام استخوان‌ها بر اساس آنالیز صوتی ارائه می‌کند. در این سیستم برای رده‌بندی از شبکه های عصبی،

[۱]- emphasis

SVM و HMM به همراه ویژگی‌های مختلفی از جمله نرخ عبور از صفر، فرکانس میانه[۱]، انرژی زیر-باند[۲]، MFCC و پریود گام استفاده شده است. در [۷۰] یک سیستم مانیتورینگ راه دور به همراه حسگرهای شنوایی باهوش به منظور انجام پزشکی از راه دور طراحی شده است. حسگر شنوایی مجهز به میکروفون‌هایی است که اتفاقات صوتی همچون نویز غیرمعمول و یا هر گونه درخواست کمک را شناسایی می‌کند. در این سیستم مقایسه LPC، MFCC به همراه ترکیب آن‌ها با مشتقات زمانی و بعضی از ویژگی‌های ادراکی در نظر گرفته شده است. همین نویسندگان در [۷۱] تکنیکی را بر اساس مدل‌های گذرا و ضرایب ویولت برای رده‌بندی صداها در کلینیک مراقبت از راه دور[۳]پیشنهاد کردند. در این پروژه فعالیت‌های بیمار، رفتارهای روانی و استرس‌های احتمالی او آنالیز صوتی می‌شود. از بین مدل‌های رده‌بندی ، GMM دارای کم‌ترین پیچیدگی است. معیار اطلاعاتی بیزین (BIC)[۴]برای تعیین تعداد بهینه گوسی‌ها بکارگرفته شده است. همچنین در [۷۲] رده‌بندی صداها در مدل‌های مختلف نسبت سیگنال به نویز جهت نظارت از راه دور در پزشکی مورد مطالعه و تحقیق قرا گرفت. در [۷۳] استخراج صحنه‌های برجسته در بازی‌های فوتبال، گلف و بیسبال مورد مطالعه قرار گرفته است و طی آن نویسندگان به مقایسه بردارهای ویژگی MPEG-7 با ویژگی‌های MFCC پرداخته‌اند. استخراج ویژگی‌های MPEG-7 شامل موارد زیر است: نرمال سازی پوشش طیف صوتی (NASE)[۵]، الگوریتم تجزیه به اجزا اولیه (SVD یا ICA)[۶] و تصویرسازی پایه‌ای طیفی[۷]که این کار از طریق ضرب NASE در توابع پایه‌ای استخراج شده انجام می‌شود همچنین از HMM به همراه الگوریتم‌های آموزشی entropy prior و حداکثر شباهت برای رده‌بندی استفاده شده است. نویسندگان نتایج امید بخشی را از طریق بکارگیری تکنیک‌های پیش و پس پردازش و بهره گیری از دانش عمومی ورزشی کسب نموده اند.

۱- median
۲- sub-band
۳- clinic telesurvey
۴- Bayesian Information Criteria (BIC)
۵- Normalized Audio Spectrum Envelope (NASE)
۶- Singular Value Decomposition (SVD)
۷- spectrum basis projection

در [۷۴] درک و تصور صداهای پیش‌زمینه و پس‌زمینه را تعریف کرده و همچنین از طریق پیگیری یک پردازش مرحله ای و نسلی[۱]کار شناسایی و تطبیق تغییرات را حین پردازش انجام می-دهند. هدف آن‌ها مدل سازی صدای پس‌زمینه به منظور شناسایی صداهای مشکوک یک آسانسور بود. در مقاله [۷۵] ارائه شده در پروژه CHIL ، از یک روش بدون نظارت که بر اساس الگوریتم اصلاح شده ای از EM [۲]است برای قطعه‌بندی صوتی به منظور شناسایی وقایع صوتی مجزا در اتاق ملاقات استفاده شده است و نتایج بدست آمده در مقایسه با BIC رضایت بخش تر بوده است.

۳-۶-۲- پردازش صداهای عمومی

این نوع پردازش مربوط به صداهایی است که محدود به محیط خاصی نیستند. در [۷۶] نویسنده دو رویکرد را برای شناسایی و رده‌بندی زنگ هشدار بکار برده است. رویکرد اول روش ANN است و رویکرد دوم یک تکنیک طراحی شده برای کشف ساختار زنگ‌های هشدار است که تاثیر نویز زمینه را به حداقل می‌رساند. نویسنده همچنین تاثیر یک سری از ویژگی‌های عمومی انواع مختلف نویزها را بر روی تعداد محدودی از زنگ‌های هشدار تحقیق نموده است.

در [۷۷] نویسندگان خوشه بندی مدل K-means را جهت تعیین برچسب "برنامه"[۳]یا "تجاری"[۴] برای یک قطعه صوتی بکار گرفته‌اند. بر خلاف سایر سیستم‌های موجود، در اینجا هیچ فرضی برای محتوای برنامه صورت نگرفته است و این ناشی از بکارگیری روش تطبیق محتوا[۵]است. در [۷۸] شناسایی صدای گونه‌های مختلف پرندگان صورت گرفته است. این کار از طریق مقایسه نمایش سینوسی سیلاب‌های مجزای صدای تعداد محدودی پرنده انجام شده است. فرض آن‌ها این است که تعداد زیادی از سیلاب‌های آواز پرندگان را می‌توان با پالس‌های سینوسی محدودی که در اندازه و فرکانس متفاوت هستند تقریب زد.

۱- generative process
۲- Maximization Expectation (EM)
۳- program
۴- commercial
۵- content-adaptive

در [۲۴] با یک رویکرد سلسله مراتبی تشخیص وقایع صوتی را به منظور نظارت انجام دادند. فریم صوتی ابتدا به vocal یا non-vocal و بعد به "عادی"[۱] یا "بر آشفته"[۲] تفکیک می‌شود. این رویکرد را با رده‌بندی GMM بر روی ویژگی‌های LPC انجام داده‌اند.

در [۷۹] مقایسه ویژگی‌های MFCC و Mpeg7 را انجام دادند. همچنین سه رویکرد گزینش ویژگی (یا کاستن از فضای ویژگی‌ها) که PCA, ICA و NMF بودند را انجام دادند. آن‌ها از HMM پیوسته برای رده‌بندی استفاده کردند. در آنالیز بهره وری به این نتیجه رسیدند که برای شناسایی صداهای عمومی تحت محدودیت‌های خاص ویژگی‌های MFCC از ویژگی‌های MPEG-7 بهتر هستند. علاوه بر این مشخص شد که بکارگیری PCA[۳]بر روی کاهش ویژگی‌های MPEG-7 بهترین نتیجه را داشته است. آن‌ها همچنین در مقاله [۸۰] به مقایسه رده‌بندی یک سطحی و سلسله مراتبی بر پایه HMM و ویژگی‌های MPEG-7 که از طریق ICA پیش‌پردازش شده بودند پرداختندکه در تحلیل‌های صورت گرفته بهترین نتیجه از مدل سلسله مراتبی به همراه اطلاعات اضافی[۴]بدست آمده است.

در [۸۱] هدف طراحی سیستمی بود که در شرایط نویز عملکرد بهتری داشته باشد و همچنین میزان انرژی کم‌تری مصرف کند. لذا به طور جداگانه از ویژگی‌های MFCC و NRAF[۵] برای رده‌بندی چهار دسته از صداها با استفاده از GMM استفاده گردید و نتایج هرکدام با هم مقایسه شدند. از آنجا که MFCC در حضور نویز عملکرد مناسبی ندارد لذا مشخص گردید که NRAF جایگزین مقاوم‌تری برای آن است.

در [۸۲] سیستم رده‌بندی صوتی چندین کلاسه را بر اساس ویژگی‌های مربوط به انرژی باندهای فرکانسی از جمله پهنای باند، قله‌ها، بلندی[۶]و ... پیشنهاد داده‌اند. آن‌ها از چندین درخت تصمیم

۱- normal
۲- excited
۳- Principal Component Analysis (PCA)
۴- Hint
۵- Noise-Robust Auditory Features (NRAF)
۶- loudness

استفاده کردند و به دلیل تنوع صداها از رده‌بندی آبشاری[1] استفاده شده است تا خطاهای خاصی که در مراحل قبلی رده‌بندی رخ داده است اصلاح شود.

۴-۶-۲- رده‌بندی محیط از روی صداها

علاوه بر این که لازم است سیستم‌هایی برای شناسایی صداهای یک محیط از قبل تعیین شده طراحی شود، شناسایی محیط از روی چند صدای داده شده نیز باید امکان پذیر باشد. در [۸۳] صدای جمع آوری شده افراد را در قالب شناسایی محیط خوشه بندی می‌کنند. نوارهای صوتی ضبط شده را در قالب ۱۶ کلاس مختلف رده‌بندی می‌کنند و برای این کار از انرژی فرکانسی bark-scaled و spectral entropy استفاده نموده‌اند.

در [۸۴] رده‌بندی بر اساس HMM برای شناسایی محیط‌های مختلف گفتاری انجام شده است. محیط‌هایی همچون گفتار در محیط ساکت، در ترافیک، در هیاهو و جنجال و غیره و هدف تقویت وسایل کمک شنوایی است که بتوانند محیط گفتگو را برای تنظیمات اتوماتیک تشخیص دهند.

در [۲۷] از آنالیز محتوایی برای رده بندی و قطعه‌بندی سیگنال صوتی بر اساس نوع صدا یا هویت گوینده استفاده نموده است. رویکرد مورد اشاره قابلیت رده بندی و قطعه‌بندی سیگنال را به گفتار/موزیک/صدای محیطی/سکوت دارد. رده‌بندی صوتی را در دو مرحله انجام می‌دهد. در مرحله اول با استفاده از KNN و LSP-VQ وضعیت گفتار/غیرگفتار را تعیین می‌کند. در مرحله دوم قطعه‌های غیرگفتار را توسط یک رویکرد مبتنی بر قانون به موزیک/صدای محیطی/سکوت رده بندی می‌کند.

۷-۲- رویکردهای تجزیه ماتریس و جداسازی منابع

در سالیان اخیر الگوریتم‌های جداسازی کور منابع از جمله استفاده از تجزیه نامنفی ماتریس (NMF) سیگنال مشاهده مد نظر قرار گرفته که می‌توان به مقالات [۲۹-۳۵] اشاره نمود. اما بدلیل پایه‌ای

[1]- cascading

نبودن این روش‌ها در شناسایی وقایع صوتی نتـایج بدسـت آمـده در مقایسـه بـا روش‌هـای مرسـوم و مدل‌سازی‌های آماری راضی کننده نیست.

هرچند که روش‌های پایه نیز در شرایط خاص نتایج خوبی داشته‌انـد امـا در حالـت عمـومی هنـوز پاسخگویی لازم را ندارند [۸۵]. لذا در شناسایی وقایع صوتی محیطی تعریف نوعی از همبسـتگی بـین اجزای جریان صوتی ورودی با الگوهای طیفی از پیش تعیین شده مناسـبت بیش‌تـری دارد. بـه طـور نمونه، تجزیه نامنفی ماتریس، جریان صوتی ورودی را به ترکیبی خطی از اتم‌های دیکشنری با ضرایب خاص تبدیل می‌کند. در این حالت ماتریس ضرایب فعال کننده به عنوان ویژگی‌هـای بدسـت آمـده از سیگنال ورودی جهت رده‌بندی صـداها اسـتفاده می‌شـود. از چالش‌هـای مطـرح در ایـن نـوع تجزیـه می‌توان به یکسان نبودن سیگنال بازسازی‌شده با سیگنال اولیه اشاره نمود. به طوری که افزوده شـدن یا جایگزینی اجزای صوتی بجای یکدیگر می‌تواند منجر به خطای عدم همخوانی سیگنال بازتولیدی بـا سیگنال اصلی گردد [۸۶].

فرض ما ایـن اسـت کـه اعمـال کنتـرل بـر نحـوه تجزیـه مـاتریس مشـاهده و چگـونگی نگاشـت سگمنت‌های ورودی بر دیکشنری وقایع سبب شناسایی بهتر وقایع صوتی می‌گردد. از مطالعات مرتبط می‌توان به اعمال محدودیت تنک[۱] در تجزیه نامنفی جهت مشاهده گام چندگانه موزیکـال و شناسـایی دستگاه‌های موسیقی در [۸۷، ۸۸] اشاره نمود. اما با وجود کنترل‌های صورت گرفتـه بـر رونـد تجزیـه بخش‌های حاصل فاقد ساختار است که کارآمدی روش را کم می‌کند. فرض ما این است کـه سـیگنال ورودی دارای اجزای کوچک‌تر سازمان‌یافته است. برای مثال در رونویسـی موزیـک پلی‌فونیـک اجـزای سیگنال تعدادی نت یا وقایع صوتی در حال نواختن فرض می‌شوند. در این حالت شناسایی هر نـت یـا اتفاق صوتی اهمیت دارد و شناسایی اجزای کوچک‌تر مد نظر نیست.

[۱]- sparsity constraint

فصل

۳- روش پیشنهادی اول: مدل‌های تطبیق‌پذیر برای کشف و رده‌بندی وقایع صوتی محیطی

مقدمه

در این فصل، هدف ما ارائه ساختارهای انعطاف‌پذیری است که بتواند با شرایط استخراج ویژگی‌ها در حوزه زمان یا فرکانس و یا ترکیبی از آن‌ها و تنظیم پارامترهایی برای شرایط خاص همچون کنترل میزان فراخوانی و دقت در کشف وقایع نتایج مطلوبی را برحسب تنوع صداهای محیطی از خود نشان دهد. با معرفی چندین روش برای ایجاد انعطاف در قطعه‌بندی سیگنال از جمله نحوه تعیین آستانه[1] روی انتخاب فریم، در نظر گرفتن مکانیزم‌هایی جهت کنترل سخت یا نرم در گزینش فریم بصورت پارامترهایی در مدل، سوق دادن عملیات قطعه‌بندی بسمت پردازش چند منظوره و بدست آوردن نتیجه‌ای مشترک از آن‌ها و در نهایت هدف قرار دادن استفاده از اطلاعات پیشینی از بخش‌های صوت و نویز سیگنال سعی داریم تا بهترین نتیجه از قطعه‌بندی که فاز اصلی در تفکیک و شناسایی صداها هست را انجام دهیم.

چنانچه هدف کشف انواع مختلفی از وقایع صوتی محیطی در رکوردهای ضبط شده همراه با تعیین زمان شروع و پایان هر کدام از آن‌ها باشد لازم است نوع مدل‌سازی تطبیق‌پذیری با ویژگی‌های استخراجی را داشته باشد و همچنین امکان استفاده از پارامترهایی جهت کنترل عملکرد وجود داشته باشد.

ما در ادامه این فصل با توجه به دیدگاه‌های مختلفی که در بحث قطعه‌بندی و شناسایی صداهای محیطی وجود دارد مدلی که بتواند در قطعه‌بندی و سپس رده‌بندی صداهای محیطی موفقیت نسبی داشته باشد را پیشنهاد می‌کنیم. باید توجه داشت که پژوهش بر روی یک یا تعداد محدودی از صداها کار را تا حدودی آسان می‌کند اما زمانی که تنوع صداها زیاد می‌شود حتی با اضافه شدن یک کلاس از صداهای جدید، سختی کار دو چندان می‌شود. ما رویکرد خود را با توجه به زیاد بودن نوع این کلاس وقایع و همچنین در دست نبودن داده‌های آموزش زیاد برای هر کلاس ارائه کرده‌ایم و مدل پیشنهادی در زمان آزمایش نیز از کارایی خوبی برخوردار بوده است که در بررسی و مقایسه نتایج این

[1]- Threshold

امر مشخص گردید. جهت توضیحات کامل مدل، ما در بخش‌های مختلف این فصل تمام جزئیات را آورده‌ایم. ابتدا مدل‌های مختلف سگمنت کردن و همچنین روش‌های تعیین مقادیر آستانه توضیح داده شده است. سپس مدل پیشنهادی جهت قطعه‌بندی به همراه نقش پارامترها و اثرات نویز و به دنبال آن استخراج ویژگی‌ها و تشکیل مدل و اجرای قطعه‌بندی و همچنین عملیات رده‌بندی و آزمایشات مربوطه تا حصول نتایج و مقایسه آنها با روش‌های دیگری که از پایگاه داده مشابه استفاده نموده‌اند را در بر می‌گیرد.

۱-۳- مدل سگمنت‌بندی از طریق فریم‌بندی

ابتدا در مورد حداقل طول فریم آزمایشات متعددی انجام شد و مشخص گردید که اگر طول فریم کم‌تر از ۱۰۰ میلی‌ثانیه باشد در آن صورت موفقیت ویژگی‌هایی مثل ZCR در سگمنت‌بندی خیلی پایین است چون ابتدا و انتهای سگمنت‌ها که تغییر دینامیکی واضحی رخ می‌دهد قابل ردیابی نیستند و علت عمده هم آن است که تغییرات دینامیکی در حوزه زمان در فاصله زمانی طولانی‌تری قابل مشاهده است. این موضوع بخصوص در مورد شناسایی فریم‌های پایانی صداها بیش‌تر ایجاد اشکال می‌نمود. به همین دلیل با انجام آزمایش‌های متعدد روی صداهای مختلف، طول فریم ۱۲۵ میلی‌ثانیه انتخاب شد تا تغییرات در حوزه زمان بهتر قابل رهگیری باشد.

۱-۱-۳- سه راهکار برای سگمنت‌بندی سیگنال

این روش‌ها با استفاده از ویژگی‌های زمانی همچون ZCR و STE توسعه و آزمایش گردیده است و نتایج مطلوبی داشته‌اند.

۱-۱-۱-۳- استفاده از ویژگی ZCR به تنهایی و تکمیل آن با سایر ویژگی‌ها

این شیوه از دو راه زیر قابل اجرا است:

۱- روش فیلتر کردن فریم‌ها بر اساس مقدار ویژگی ZCR می‌تواند نام مناسبی برای این روش باشد. ابتدا فریم‌ها بر اساس مقدار ZCR بصورت صعودی به نزولی مرتب می‌شوند آنگاه درصدی از فریم‌های اولیه لیست به عنوان فریم‌های صدا انتخاب می‌گردند. آزمایشات متعددی

که توسط ما در پایگاه داده مورد استفاده انجام گرفت نشان داد که انتخاب ۲۵ درصد از فریم‌های بالای لیست قابل قبول بوده و نتایج خیلی خوبی در بحث سگمنت‌بندی با این روش دارد. این روش دارای مزیت گزینش سریع و یکجای فریم‌های صدا در جاهای مختلف رکورد ورودی است و نیاز به پردازش اولیه فریم‌ها به صورت سریالی را مرتفع می‌کند. در مراحل بعدی می‌توان در مورد فریم‌های گزینش نشده که بین فریم‌های انتخاب شده هستند تصمیم‌گیری نمود. مزیت دیگر این روش آن است که به طور قطعی می‌توان تعداد زیادی از فریم‌های انتهای لیست را به عنوان فریم‌های غیرصدا در نظر گرفت. در فایل ورودی script03.wav بیش از ۶۰ درصد سگمنت‌بندی با همین یک روش بدرستی انجام شد که در مقایسه با روش‌های آماری پیچیده و مدل‌سازی‌های سخت میزان موفقیت خیلی بالایی بحساب می‌آید.

۲- در روش دوم، تمام فریم‌ها به شیوه سریالی از اول تا انتها مورد پردازش قرار می‌گیرند و بر اساس تعیین مقادیر آستانه عمومی و پس از آن در صورت لزوم تعیین آستانه محلی اقدام به گزینش فریم‌های صدا می‌شود. در این روش راهکار عملی این است که از فریم اول شروع کرده و از آن‌هایی که مقادیر کم و بیش یکسانی دارند عبور نموده تا به اولین فریم با مقدار ZCR بالا و متفاوت برسیم. این فریم به عنوان نقطه خیزش یک صدا در نظر گرفته می‌شود. آنگاه چند فریم قبل و بعد از فریم خیزش صدا را انتخاب کرده و میانگین آن‌ها را به عنوان آستانه گزینش فریم‌های بعدی انتخاب می‌کنیم.

۲-۱-۱-۳- استفاده از ویژگی انرژی به تنهایی و تکمیل آن با سایر ویژگی‌ها

این روش دقیقا مانند استفاده از روش ZCR است اما ویژگی مورد استفاده در اینجا انرژی کوتاه-مدت فریم‌ها است. با تحقیقات ما مشخص شد انتخاب ۳۰ درصد فریم‌های بالایی لیست مرتب شده به عنوان صدا و همچنین ۳۰ درصد از فریم‌های پایین لیست به عنوان غیرصدا نتایج خوبی در پی دارد.

۳-۱-۱-۳- استفاده از ویژگی ZCR + STE و تکمیل آن با روش‌های اصلاحی

در این روش سگمنت‌بندی هم از طریق ویژگی عبور از صفر و هم از طریق ویژگی انرژی استفاده می‌کند و نتایج سگمنت‌بندی با هر دو ویژگی با یکدیگر ترکیب می‌شوند.

۳-۲- روش‌های اصلاحی جهت تعیین محدوده سگمنت

صداهایی هستند که ZCR پایین دارند اما دارای انرژی بالاتر از حد متوسط هستند. این صداها با ویژگی ZCR قابل تفکیک از نویز نیستند اما در هنگام کشف از طریق انرژی پیدا می‌شوند. معمولا این فریم‌ها درون یک سگمنت قرار می‌گیرند و انرژی بالاتر از حد متوسط هم دارند. برای کشف صدا از این فریم‌ها باید بخش‌هایی از این قطعه که قبلا توسط ویژگی ZCR پیدا شده‌اند را از پردازش مجدد کنار گذاشت. اما آن بخش‌هایی که با ویزگی ZCR گزارش نشده‌اند را به صورت یک بردار در نظر گرفت. آنگاه میانگین انرژی این بردار را بدست آورد و مقادیر بردار را از میانگین کم نمود. برحسب نتیجه اگر انرژی مثبت باشد فریم را ۱ و در غیر این صورت فریم صفر در نظر گرفته می‌شود. در نهایت قسمت‌های ۱ شده به همراه بخشی که توسط ZCR از قبل ۱ شده بودند را در کنار هم بررسی کرده و محدوده آن‌ها را به عنوان یک اتفاق صوتی در نظر می‌گیریم.

در رکورد نمونه script01.wav بین فریم‌های ۲۵۴ تا ۳۳۷ دقیقا وضعیت گفته شده بروز کرد. برحسب ویژگی ZCR در فریم‌های ۳۱۴-۳۰۷ و ۳۳۸-۳۳۳ دو سگمنت صوتی جدا از هم دیده می‌شد. به طوری که سایر فریم‌های مجاور دارای ویژگی عبور از صفر کم‌تر از میانگین بودند. لذا وضعیت فریم‌های ۳۰۶-۲۵۴ و ۳۳۲-۳۱۵ را با ویژگی انرژی بررسی کردیم. در واقع وضعیت فواصل بین دو بخش که صدا تشخیص داده شده بودند نامشخص بود. لذا مراحل زیر بررسی و نتیجه دور از انتظار بدست آمد. ۱- هر یک از دو بخش نامشخص بصورت یک بردار جدید تعریف شد. V1=frames (۲۵۴:۳۰۶); و V2=frames (۳۱۵:۳۳۲); آنگاه مراحل زیر برای هر کدام از بردارهای V1, V2 اجرا گردید. ۲- محاسبه میانگین ZCR در بردار انجام شد. این مقدار در واقع محلی است چون در چند فریم خاص انجام می‌شود. ۳- مقادیر ZCR هر فریم از میانگین محلی کم شد. ۴- نتیجه هر فریم که

مثبت بود آن فریم مربوطه ۱ و سایر فریم‌ها صفر می‌شوند. ۵- جهت همواره ساختن سگمنت‌های حاصل در جایی که بین دو فریم انتخاب شده حداکثر دو فریم انتخاب نشده بودند آن‌ها را نیز انتخاب می‌کنیم. ۶- در نهایت هر جا رشته‌ای پیوسته از فریم‌های انتخاب شده وجود داشت آنرا سگمنت صوتی در نظر می‌گیریم. با این تکنیک در فریم‌های بردار V1 صدای clear threat و در بخش فریم‌های V2 صدای door knock کشف گردید. بر همین اساس ما موارد کلی که سبب بهبود سگمنت‌بندی می‌شود را با استفاده از دو ویژگی ZCR و STE بررسی کردیم و چهار شیوه اصلاحی زیر را بدست آوردیم که در نتایج سگمنت‌بندی تاثیر چشم‌گیری داشتند.

۱- ابتدا سگمنت‌بندی با استفاده از ویژگی ZCR انجام می‌شود تا بخش‌هایی که در این ویژگی غالب هستند تعیین شوند. مسلما فریم‌هایی هستند که تحت تاثیر کلی آستانه تعیین شده در این روش قرار گرفته و از انتخاب شدن باز می‌مانند. این فریم‌ها ممکن است مربوط به صدای جداگآن‌های باشند که دارای یک مقدار ZCR محلی بالا باشند. لذا لازم است این بخش‌ها را با یک روش محلی بررسی کنیم. به همین علت میانگین ZCR محلی را برای این بخش‌ها بکار می‌گیریم تا فریم‌های غالب پیدا شوند. برای اطمینان از این انتخاب، ویژگی دوم یعنی انرژی فریم‌ها را نیز بکار می‌گیریم. هر گاه انرژی این فریم‌ها بالاتر از متوسط بود گزینش اولیه را قطعی می‌کنیم.

۲- در روش دوم ابتدا از طریق ویژگی انرژی اقدام می‌کنیم و فریم‌های دارای انرژی بالاتر از متوسط را گزینش می‌کنیم. چنانچه بین بخش‌های انتخابی فریم‌ها یا نواحی پیوسته انتخاب نشده وجود دارد آن‌ها را با ویژگی انرژی محلی بررسی می‌کنیم و چنانچه ویزگی ZCR محلی بالایی نیز داشتند در گزینش نهایی آورده می‌شوند.

۳- روش دیگر که استفاده شد و موثر بود انتخاب فریم با بزرگ‌ترین مقدار ZCR است آنگاه تمام فریم‌های اطراف آن که دارای ZCR با یک فاصله مشخص (آستانه) از این فریم دارند نیز

انتخاب می‌شوند. در نهایت فریم‌های انتخاب نشده بین دو فریم انتخاب شده گزینش می‌شوند تا یک بخش پیوسته حاصل گردد.

۴- هر گاه بین فریم‌های انتخاب شده در مرحله اصلی فریم‌های انتخاب نشده‌ای هستند که ZCR پایین ولی میزان انرژی بالا دارند یا برعکس ZCR بالا و انرژی پایین دارند را به عنوان فریم‌های خاصی که با متعلق به صدای متفاوتی هستند و یا اینکه تحت تاثیر جانبی روش اصلی قرار گرفته‌اند را به عنوان فریم‌های صوتی انتخاب می‌کنیم.

۳-۳- معرفی یک روش‌جدید جهت تعیین مقدار آستانه

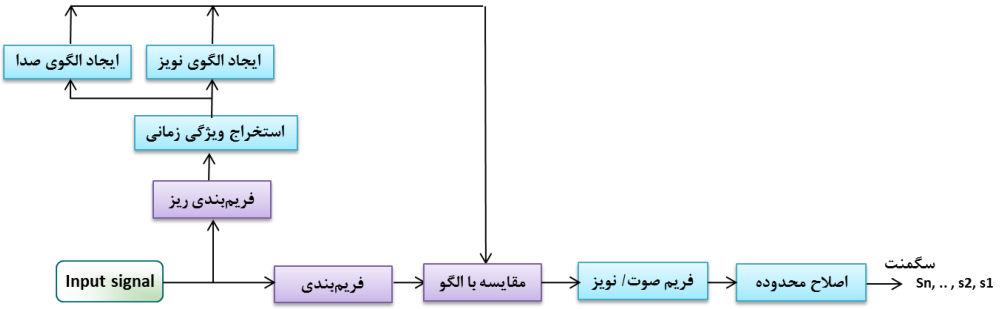
یک راهکار جدید برای تعیین مقدار آستانه برای سیگنال آن است که ابتدا یک ویژگی که چند مقدار بالای آن با احتمال خیلی زیاد بیان کننده فریم‌های صوت است را مورد استفاده قرار داد. با این ویژگی انتخاب شده، تعدادی فریم اولیه پیدا می‌شود که با فرض اولیه مربوط به صدا هستند. آنگاه می‌توان فریم‌های مجاور این فریم‌های انتخابی که گزینش نشده‌اند را با فرض غیرصدا بودن مورد استفاده قرار داد و از آن‌ها یک میزان آستانه برحسب مقدار میانگین تعیین نمود. علاوه بر اینکه می‌توان آستانه را تعیین کرد از فریم‌های غیرصوتی که در ابتدا تعیین شده‌اند می‌توان برای تخمین سیگنال نویز و پس‌زمینه استفاده نمود. ما از این روش برای تعیین مقدار آستانه عمومی اولیه استفاده کردیم.

۴-۳- روش مقایسه الگو[۱]در سگمنت‌بندی

یکی از مشکلات اصلی سگمنت‌بندی تداخل نویز با سیگنال اصلی و مشکلات بازشناسی آن‌ها از یکدیگر است. لذا پیشنهاد ما بهره‌گیری از ایجاد الگوی زمانی برای نویز بر اساس مدل‌سازی اولیه بخش‌های نویزی سیگنال است. چنانچه بخش خیلی کوچک از ابتدای سیگنال (مثلا ۵۰۰ میلی‌ثانیه) انتخاب شود و پس از فریم‌بندی در ابعاد کوچک، ویژگی‌های حوزه زمان را برای فریم‌های آن محاسبه نمود آنگاه الگویی از فریم‌های غیرصوتی که بیش‌تر مربوط به نویز هستند قابل تعیین است. تعیین این الگو نوعی فرمول‌بندی نویز بحساب می‌آید و طبق آن می‌توان بخش‌های پس‌زمینه را تعیین نمود.

[۱]- Template Matching

انتخاب فریم‌های خیلی کوچک منجر به شناخت بیش‌تر از جزئیات سیگنال می‌شود و رفتار سطح پایین سیگنال را نمایندگی می‌کند و بعلاوه به نوعی رزولوشن زمانی پردازش را بالا می‌برد. با این روش ماتریسی متشکل از چند فریم غیرصدا ایجاد می‌شود. در اینجا ۵۰۰ میلی‌ثانیه به ۲۰ فریم تقسیم گردید. از آن ماتریس یک الگو برای فریم‌های غیرصدا ایجاد می‌شود که قابل مقایسه با هر بخش دیگری از سیگنال خواهد بود. الگوی دیگری بر اساس همین روش برای بخش‌های صدا ایجاد می‌شود که آنهم تا حد زیادی نماینده بخش‌های صدا در سیگنال است. سپس فریم‌های سیگنال با دو الگوی صدا و غیرصدا مقایسه می‌شود. میزان فاصله در هر کدام کم‌تر شد فریم متعلق به آن دسته است. مراحل این الگوریتم در شکل ۳-۱ آمده است.



شکل ۳-۱: تهیه الگوی پیشنهادی برای شناسایی قطعه‌های نویز در پیش‌پردازش جهت استفاده در قطعه‌بندی سیگنال صوت

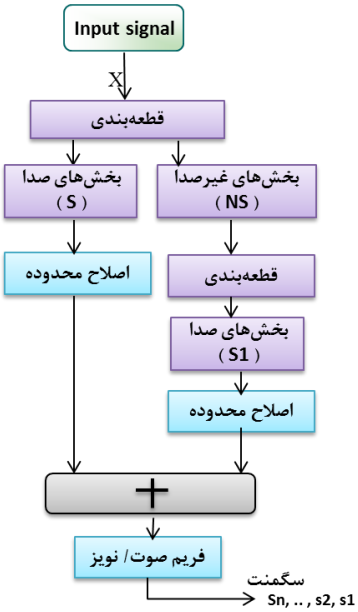
۳-۵- ایجاد یک دیکشنری از الگوی نویز

در این روش لازم است فریم‌هایی از انواع نمونه نویز را برداشته و آن‌ها را به عنوان الگوهای نویز در یک دیکشنری قرار داد. برای این منظور هر یک از نمونه‌های نویز در داده‌های آموزش به صورت بردار در یک فایل ذخیره می‌شود که بصورت یک سیگنال متشکل از نویز دیده می‌شود. آنگاه با دقت ۱۲۵ میلی‌ثانیه آن‌ها را فریم‌بندی نموده و الگوهای هر نویز را در کنار هم قرار داده و با توجه به داده‌های مورد پردازش یک دیکشنری با ابعاد مشخص ایجاد می‌شود. برای نمونه با داده‌هایی که در دست بود یک دیکشنری نویز با نام noizeDictionary بصورت یک ماتریس با ابعاد ۱۶۵۲*۵۵۱۲ ایجاد گردید.

در زمانی که فریم‌های سیگنال به صدا/غیرصدا تقسیم‌بندی می‌شوند میزان تابع همبستگی داده این دیکشنری با داده فریم محاسبه می‌گردد. هر گاه میزان تابع همبستگی از یک حد مشخص بالاتر باشد فریم سیگنال ورودی از نوع نویز محسوب می‌شود.

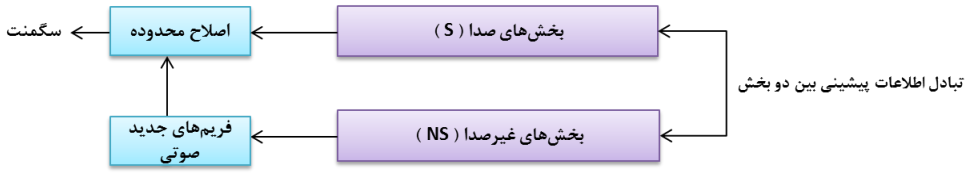
۶-۳- استفاده از residual سیگنال قطعه‌بندی شده در بهبود قطعه‌بندی

زمانی که قطعه‌بندی یک سیگنال تمام می‌شود دو چیز قطعی است. اول بخش‌های شناسایی شده به عنوان صوت و دیگری بخش‌هایی که به عنوان زمینه یا غیرصدا تشخیص داده شده‌اند. اینکه دقت کار تا چه میزان هست قطعیت ندارد اما می‌توان مطمئن شد که بخش‌هایی از غیرصدا متعلق به صداهای تشخیص داده نشده می‌باشد. در اینجا می‌توان بخش غیرصدا را باقیمانده یا residual عملیات قطعه‌بندی در نظر گرفت و تمهیداتی جهت استخراج بخش‌های غیرزمینه از آن را بکار گرفت. یکی از این ایده‌ها اجرای دوباره قطعه‌بندی در بخش غیرصدا یا همان بخش باقیمانده است. این ایده در اینجا بکار گرفته شد و تعداد قابل توجهی از صداهای قبلا کشف شده تکمیل گردیدند و حتی بعضی از بخش‌هایی که اصلا شناسایی نشده بودند در این دوره شناسایی شدند. ایده از آنجا قوت می‌گیرد که فریم‌های متعددی در سیگنال اولیه در تاثیرپذیری از الگوریتم اولیه به جمع غیرصداها پیوسته است. این الگوریتم فرصت بازیابی اینگونه فریم‌ها را مهیا می‌کند. به شکل ۳-۲ مراجعه کنید.



شکل ۳-۲: مدل پیشنهادی برای شناسایی قطعه‌بندی سیگنال در دو فاز هم‌افزا. شناسایی هم‌زمان بخش‌های صوت و بخش‌های نویز یا زمینه

در سمت راست شکل ۳-۲ که مربوط به پردازش مجدد قسمت‌های غیرصدای مرحله اول است مواردی قابل اشاره است. این قسمت بیش‌تر حاوی نویز است و احتمالا بخش‌هایی از صداها یا حتی صداهای کاملی را نیز شامل می‌شود. این بخش‌ها تکه‌هایی از سیگنال اولیه هستند که در اینجا می‌توان الگوریتم را بصورت محلی اجرا نمود. می‌توان ویژگی‌های متفاوتی نسبت به قبل از این بخش‌ها استخراج نمود و لزومی ندارد همان ویژگی‌های قبلی باشند. از این قسمت‌ها می‌توان اطلاعات پیشینی جهت الگوریتم جدید بدست آورد. بعلاوه در این مرحله ما با دو نوع اطلاعات روبرو هستیم که می‌توانند به طور متقابل از یکدیگر استفاده نمایند. از یک طرف بخش‌های حاوی صداها مشخص شده‌اند که می‌تواند اطلاعاتی را برای استخراج صداهای کشف نشده در بخش دوم در اختیار الگوریتم بگذارد. از سوی دیگر در بخش‌های غیرصدا هم می‌توان اطلاعاتی بدست آورد که توسط آن‌ها بتوان اصلاحاتی را در بخش صداها که قبلا کشف شده است اعمال نمود. این تبادل اطلاعات در شکل ۳-۳ نشان داده شده است.



شکل ۳-۳: بین فاز شناسایی صدا و فاز شناسایی نویز (یا زمینه) تبادل اطلاعات بصورت پیشینه انجام می‌گیرد. در بخش تعیین صدا بصورت اصلاح محدوده و در بخش غیرصدا بصورت ایجاد فرصت جهت شناسایی فریم‌های صدا ظاهر می‌شود.

۳-۷- معرفی مدل پیشنهادی جهت قطعه‌بندی سیگنال

مسئله کشف وقایع صوتی معمولا در دو فاز انجام می‌شود. در فاز نخست ابتدا سیگنال را قطعه‌بندی می‌کنند به طوری که قطعه‌ها یا فریم‌هایی از سیگنال ورودی که حاوی صداهای محیطی هستند شناسایی می‌شوند. فاز دوم مربوط به کشف وقایع می‌شود که طی آن فریم‌ها و قطعه‌های جدا از هم در قالب یک واقعه صوتی دارای محدوده زمانی شروع و پایان تعیین می‌شوند. چنانچه قطعه‌بندی به خوبی صورت نگیرد کل عملیات کشف وقایع با خطا و دقت پایین همراه خواهد بود. بنابراین میزان موفقیت الگوریتم نیازمند یک مدل‌سازی مناسب به همراه تعیین درست پارامترهای مدل و همچنین انتخاب ویژگی‌هایی که بتواند تا حد زیادی ابعاد تغییر پذیر داده‌ها در رده‌های صوتی مختلف را نمایش دهد بستگی مستقیم دارد.

چنانچه تعداد صداهای موجود در کشف وقایع زیاد باشد می‌توان چنین استدلال کرد که ویژگی‌های خاص در صداهای مشخصی سبب تفکیک‌پذیری بهتر می‌شوند، از اینرو اعمال یک الگوریتم کشف یکسان بر روی تمام صداها منطقی و کارساز نیست و داشتن دو زیرسیستم مجزا رده‌بندی ایده بسیار منطقی بنظر می‌رسد. بدین جهت باید دو نکته کلیدی را مد نظر داشت. یکی داشتن مدل‌سازی مناسب و دیگری ویژگی‌های با خاصیت تفکیک کنندگی بالا بین فریم‌های صوتی و غیرصوتی است.

در مدل پیشنهادی توجه خاصی به امکان قطعه‌بندی کاربرد-محور برحسب صداها و همچنین دقت و میزان کشف بالا برای وقایع صورت پذیرفته است. به طوری که الگوریتم‌های پیاده‌سازی

می‌توانند در تمام مراحل انعطاف مورد نیاز در کاربردی نمودن هر چه بیش‌تر مـدل و تطبیـق پـذیری آنرا اعمال نمایند.

در این مدل ابتدا عملیات اولیه و حذف نویز بر روی سیگنال اصلی انجام می‌شود. سپس از سیگنال ورودی X(n) بردارهای ویژگی مورد نظر استخراج می‌شـوند. ویژگی‌هـای اسـتخراج شـده می‌تواننـد از حوزه زمان، از طیف فرکانسی سیگنال، از نوع ویژگی‌های ادراکـی و یـا ترکیبـی از ایـن انـواع باشـند. می‌توان ویژگی‌های مناسب را برحسب نوع صداها مورد تحقیق و آزمایش قـرار داد و ویژگی‌هـایی کـه بتوانند صداها را بهتر شناسایی کنند استخراج نمود.

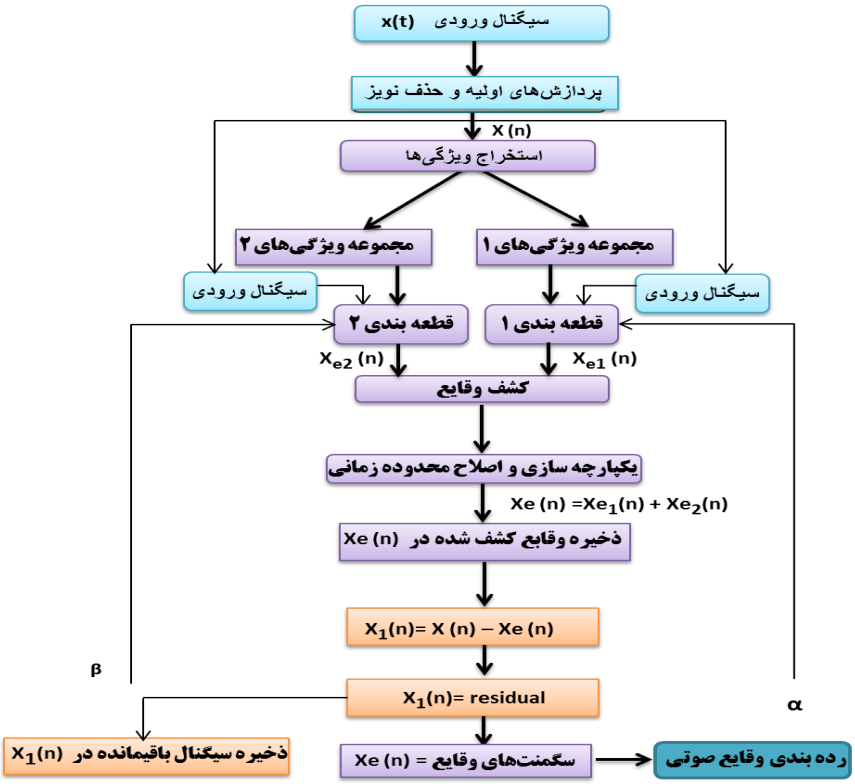
در اینجا جهت پیاده‌سازی الگوریتم ابتدا ویژگی‌های مناسب در حوزه زمان بررسی شده‌انـد. امـا در هر نوع پیاده‌سازی دیگری می‌توان بستگی به تعداد و نوع صداها و همچنـین نـوع کـاربرد ویژگی‌هـای مورد نظر را از قبل شناسایی نمود.

با داشتن بردارهای ویژگی آن‌ها را به دو دسته جداگانه تبدیل می‌کنیم این تقسیم‌بندی بـه دلیـل آن است که استفاده همزمان از تمام ویژگی‌ها می‌تواند در کشف تعداد زیاد صـداها نقـش منفـی ایفـا کنند به طوری که امکان اینکه ویژگی‌ها اثر متمایز کننده یکدیگر را در الگوریتم خنثـی نماینـد زیـاد است لذا تبدیل ویژگی‌ها به دو دسته جهت شناسایی بیش‌تر صداها و ایجـاد تمـایز بـین صـداها بهتـر عمل می‌کند. سپس عمل قطعه‌بنـدی بـر رو.ی سـیگنال ورودی بـا دو دسـته ویژگـی متفـاوت انجـام می‌شود که ایـن دو نـوع قطعه‌بنـدی می‌تواننـد دو الگـوریتم متفـاوت و یـا یکسـان باشـند. الگـوریتم قطعه‌بندی با بکارگیری پارامتر آلفا و بتا می‌تواند بین کشف دقیق‌تر و یا کشف بیش‌تر وقایع مصـالحه-ای را بوجود آورد. پس از اتمام دو نوع قطعه‌بندی نتایج آن‌هـا بـرای کشـف وقـایع و یکپارچـه سـازی محدوده هر اتفاق صوتی و بدست آوردن زمان شـروع و پایـان هـر واقعـه صـوتی مـورد اسـتفاده قـرار میگیرد.

در پایان مرحله قطعه‌بندی، لیستی از وقایع به همراه زمان رخداد هر کدام از آن‌ها تولید شده است که در رشته $X_e(n)$ ذخیره می‌شوند. اجزای مدل پیشنهادی در شکل ۳-۴ آورده شده است.

۱-۷-۳- 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. از سوی دیگر وجود کمیت‌های مختلف برای آستانه که بتوانند اعداد صحیح، اعشاری و یا مختلط باشند لزوم استفاده از پارامترهای انعطاف پذیر و متفاوت را توجیه پذیر 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. اگر دقت زیاد مورد نظر باشد افزایش آلفا در زیرسیستم اول و همچنین افزایش بتا در زیرسیستم دوم سبب افزایش میزان دقت کشف می‌گردد. چنانچه بخواهیم میزان کشف را در قبال از دست دادن اندکی دقت در حداکثر خود قرار دهیم بهتر است پارامترهای آلفا و بتا را کاهش دهیم. 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.



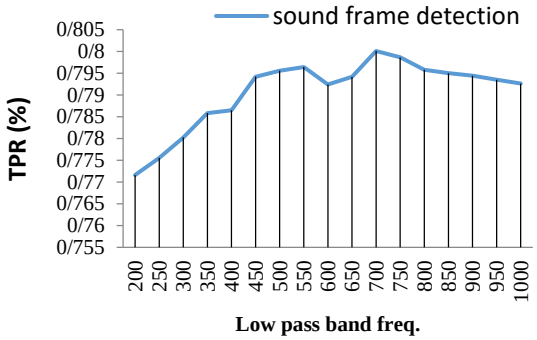
شکل۳-۴: بلاک دیاگرام مدل پیشنهادی کشف وقایع صوتی.

۲-۷-۳- بررسی اثرات نویز و حذف آن

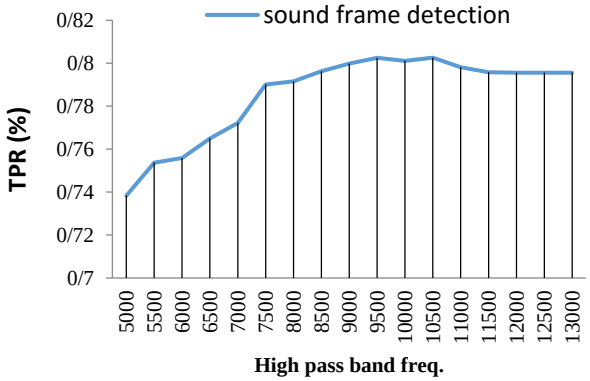
بدون داشتن اطلاعات اولیه از ساختار نویز استفاده از روش‌های برآورد نویزهای ضربه‌ای و حذف آن‌ها با الگوریتم وفقی کار ساده و موثری نخواهد بود. همچنین بهبود کیفیت سیگنال از طریق روش‌های زیرفضا که برای نویزهای سفید پیشنهاد می‌شود نیازمند تجزیه سیگنال به مقادیر منفرد کسری ادراکی (PCQSVD) است که به پردازش‌های 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 باتروث به عنوان 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 بدست آمده نشان داد که استفاده از فیلتر میان‌گذر باتروث[۱] 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 ۳-۵ و ۳-۶ برحسب میزان True-

[۱]- butterworth filter

positive rate (TPR) برای تعیین وضعیت صدا/غیرصدا در فریم‌های آزمایش آورده شده است. لـذا در عملیات پیش‌پردازش از فیلتر باتروورث میان‌گذر مرتبه ۷ با فرکاس پایین ۷۰۰ هرتـز و فرکـانس قطـع بالای ۱۰۵۰۰ هرتز استفاده شد. نویز موجود در خارج از ایـن پهنـای بانـد تـاثیر منفـی بـر شناسـایی صداهای پایگاه داده داشت.



شکل ۳-۵: پاسخ فیلتر باتروورث (butterworth) مرتبه ۷. تعیین فرکانس پایین جهت بیشترین بازدهی در شناسایی فریم‌های صدا. در فرکانس ۷۰۰ هرتز نرخ شناسایی درست فریم‌ها حداکثر است.



شکل ۳-۶: پاسخ فیلتر باتروورث (butterworth) مرتبه ۷. تعیین فرکانس بالا جهت بیشترین بازدهی در شناسایی فریم‌های صدا. در فرکانس ۱۰۵۰۰ هرتز نرخ شناسایی درست فریم‌ها حداکثر است.

۳-۷-۳- پیاده‌سازی مدل در یک الگوریتم کشف وقایع

انتخاب هر یک از ویژگی‌های حوزه‌های زمانی، فرکانسی و ادراکی یا ترکیبی از این انـواع در طراحـی الگوریتم امکان پذیر است. در اینجا به دلیل سنجش توانایی ویژگی‌های حـوزه زمـان در کشـف تعـداد متنوع وقایع صوتی در ساختار مدل پیشنهادی، تحقیق در نحوه تولید دو گروه از ویژگی‌هـای متفـاوت با نوع پارامتر متفاوت، سریع تر بودن محاسبات و پایین بودن هزینه عملیاتی و بعلاوه وجـود تغییـرات

بسیار سریع مشخصه‌های زمانی در وقایع صوتی حوزه زمان را در الگوریتم پیاده‌سازی انتخاب کرده‌ایم. پایگاه داده مورد استفاده از ۱۶ نوع واقعه صوتی مختلف تشکیل شده است که لیست این وقایع در جدول ۳-۱ آمده است.

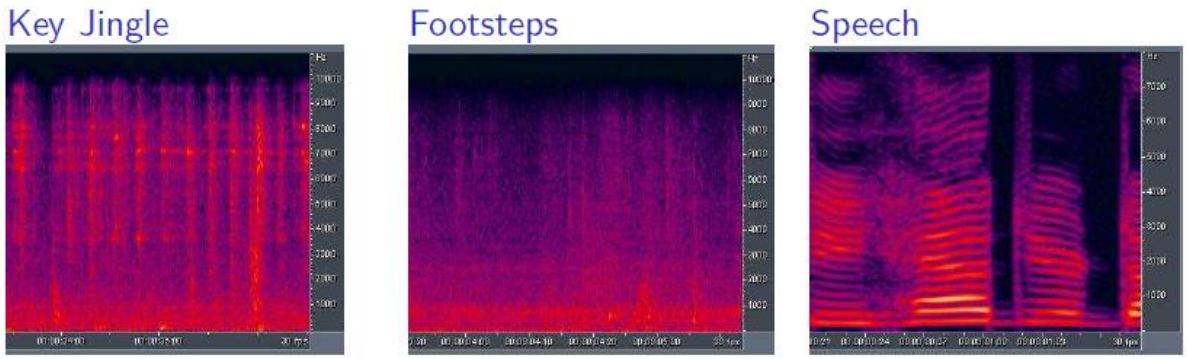
۳-۷-۴- استخراج ویژگی‌ها

در بخش مقدمه تفاوت بین صداهای حنجره‌ای (گفتاری) و صداهای محیطی ذکر شد. شکل ۳-۷ نمونه‌ای از تفاوت اطلاعات طیفی بین این دو دسته از صداها را نشان می‌دهد. همانطور که مشخص است از صداهای گفتاری به راحتی می‌توان ساختار فرمنت‌ها و گام را شناسایی و استخراج کرد اما صداهای محیطی فاقد ساختار طیفی مشخص هستند. لذا در استخراج ویژگی‌ها لازم است از ویژگی‌هایی که بیش‌تر با ساختار صداهای محیطی همخوانی دارند استفاده شود. برای مثال ویژگی‌هایی چون ضرایب MFCC که در شناسایی گفتار نهادینه شده‌اند چندان مناسب شناسایی صداهای محیطی فاقد ساختار طیفی نیستند. در اینجا جهت بررسی ویژگی‌های مناسب آزمایش‌های فراوانی انجام شده است و ویژگی‌ها برحسب قابلیت جداسازی انتخاب شده‌اند. در ادامه نحوه محاسبه هر یک از ویژگی‌ها و مناسبت استفاده از آن‌ها آمده است.

۳-۷-۴-۱- ویژگی عبور از صفر

مقدار ویژگی عبور از صفر در سیگنال صوتی از رابطه (۳-۱) محاسبه می‌شود:

$$ZCR=\frac{1}{N}\sum_{n=1}^{N-1}|\mathrm{sgn}(x_n)-\mathrm{sgn}(x_{n+1})| \qquad (۳-۱)$$



شکل۳-۷: وجود ساختار مشخص طیفی از جمله فرمنت‌ها و گام در صدای گفتاری و عدم وجود آنها در صداهای محیطی.

جدول۳-۱: انواع وقایع صوتی مورد پردازش در این پژوهش از پایگاه داده انجمن تخصصی پردازش سیگنال صوت IEEE

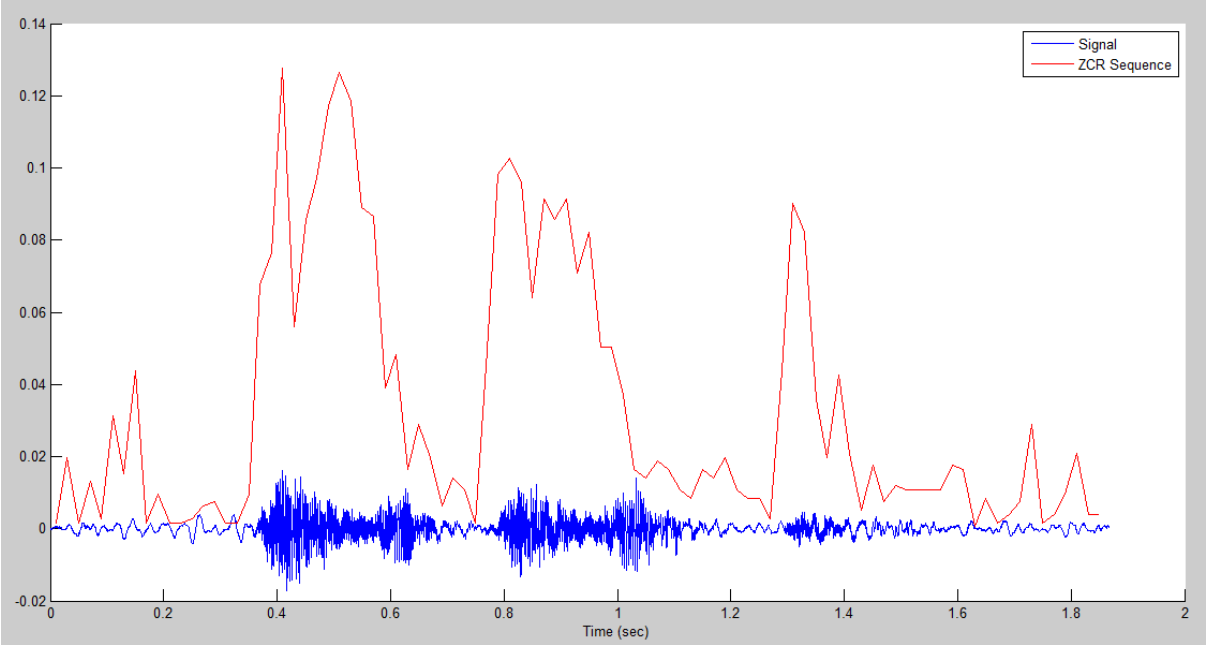
ردیف	نام واقعه صوتی	توضیحات
۱	alarm	(short alert (beep) sound)
۲	clearthroat	(clearing throat)
۳	cough	
۴	doorslam	(door slam)
۵	Drawer	
۶	keyboard	(keyboard clicks)
۷	keys	(keys put on table)
۸	knock	(door knock)
۹	laughter	
۱۰	mouse	(mouse click)
۱۱	pageturn	(page turning)
۱۲	pendrop	(pen, pencil,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

متغییرهای x_n و x_{n+1} نمونه‌های پشت سر هم 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 از جمله صداهای Switch, Printer, Keys, Page-turn, Keyboard Alert, دارای مقدار بالا با تفاوت معنادار نسبت به سایر صداها می‌باشد که از این نظر این صداها را می‌تواند تا حد زیادی از سایر صداها تفکیک نماید. همچنین مشخص شد، در بعضی از صداها از جمله صداهای keys, cough, door slam, Clear throat مقدار ویژگی عبور از صفر در فریم‌های ابتدایی زیاد است و بالا بودن این ویژگی می‌تواند برای تشخیص ابتدای صداها موثر باشد. علاوه بر اینها 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 صداهای

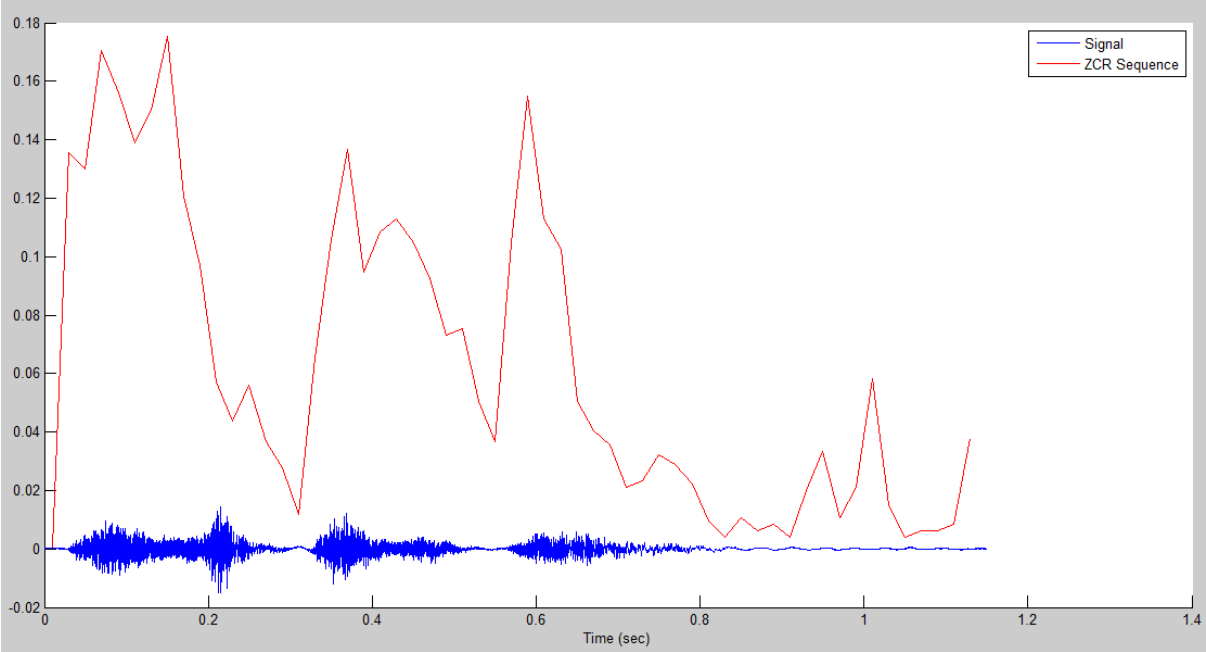
مختلفی این ویژگی میتواند در بحث قطعه‌بندی و شناسایی ابتدا و انتهای آنها مفید باشد. شکل ۳-۸ اثر این ویژگی در صدای نمونه cough دیده می‌شود. همانطور که مشخص است در جایی که سیگنال افزایش ناگهانی انرژی یا سقوط انرژی را دارد ویژگی عبور از صفر نیز تغییر توزیع آماری از خود نشان می‌دهد و سبب آشکار شدن کنش‌های دینامیکی در حوزه زمان می‌شود.

در صدای نمونه phone که در شکل ۳-۹ نمایش آن آورده شده است تغییرات این ویژگی بیش از هر ویژگی دیگر می‌تواند محدوده کامل یک صدای متعلق به تلفن را مشخص سازد. همانطور که مشاهده می‌شود از زمان بالا رفتن مقدار این ویژگی که از یک مقدار نزدیک صفر شروع می‌شود تا انتهای صدای تلفن، مقدار عبور از صفر فریم‌های سیگنال به نزدیک صفر نرسیده و همواره بالای مقدار نرمال شده ۰,۰۱۵ می‌ماند و فقط زمانی که فریم‌های صدای مذکور تمام می‌شوند افت ناگهانی ویژگی بروز می‌کند. اینگونه موارد استعداد این ویژگی را برای شناسایی صداها تا حدی زیادی افزایش می‌دهد. همچنین در شکل ۳-۱۰ نموداری از مقادیر عبور از صفر ۹۵۰ فریم از صداهای مختلف را نمایش داده‌ایم. همانطور که مشخص است مقدار این ویژگی در تعداد زیادی از فریم‌ها دارای مقدار بالا و قابل تفکیک است. در فریم‌های ۹۳ تا ۱۰۴ یک اتفاق صوتی مرکب قابل تشخیص است و همچنین در فریم‌های ۶۲۳ تا ۶۶۵ دو واقعه صوتی مجزا قابل تفکیک و شناسایی هستند. مناسب بودن این ویژگی را با رسم نمودار راک[1]تحقیق و بررسی کردیم و مشخص شد این ویژگی قابلیت تفکیک و شناسایی صداهای مورد نظر را تا حد زیادی دارد. راک نموداری است که می‌تواند میزان کارایی یک رده‌بند دوگانه را با تغییر مقدار آستانه نمایش دهد [۸۹].

[1]- Receiver Operating Characteristic (ROC)



شکل۳-۸: تغییرات دینامیکی زمانی در سیگنال نمونه cough و همزمان ظاهر شدن تغییر در توزیع آماری ویژگی عبور از صفر.



شکل۳-۹: شناسایی محدوده صدای نمونه phone از طریق توزیع آماری ویژگی عبور از صفر. در اینجا نرخ عبور از صفر فریمها از ابتدا تا انتهای صدا همواره بالاتر از مقدار نرمال شده ۰٫۰۱۵ است و خارج از محدوده صدا مقدار مذکور به نزدیک صفر می‌رسد.

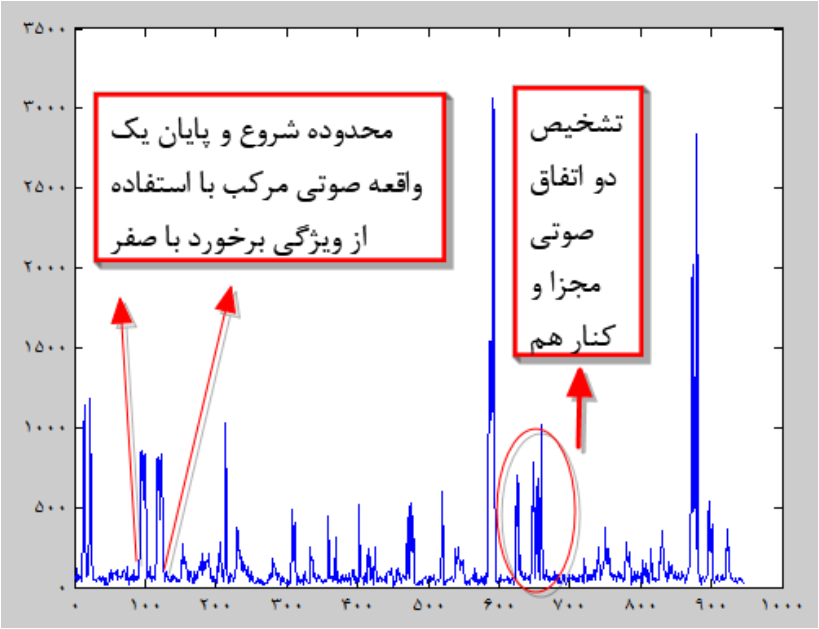
۲-۴-۷-۳- ویژگی نسبت بالای نرخ عبور از صفر

دومین ویژگی مورد استفاده که بر اساس تغییرات مقادیر عبور از صفر تعریف می‌شود HZCRR[1] است. این ویژگی برحسب میزان تغییرات عبور از صفر در یک پنجره ۱ ثانیه‌ای با تعداد N فریم محاسبه می‌شود و نسبت تعداد فریم‌هایی که مقدار عبور از صفر آنها بیشتر از ۱٫۵ برابر متوسط نرخ میزان عبور از صفر در یک پنجره ۱ ثانیه‌ای می‌باشد را نشان می‌دهد و طبق رابطه (۲-۳) تغییرات ویژگی عبور از صفر را در بازه زمانی طولانی‌تری مد نظر قرار می‌دهد. متغییر n اندیس فریم‌ها و avZcr میانگین عبور از صفر در طول یک پنجره ۱ ثانیه‌ای است و N نیز تعداد کل فریم‌های دورن پنجره پردازش است. تابع sgn را قبلا توضیح دادیم. 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 مشخص شد اکثر صداهای محیطی دارای ویژگی عبور از صفر با میزان تغییرپذیری بالا و متغییر هستند که این تغییر پذیری کمک می‌کند تا فریم‌های یک صدای مشخص از سایر صداها قابل تشخیص باشد.

$$HZCRR=\frac{۱}{۲N}\sum_{n=۰}^{N-۱}[\mathrm{sgn}(Zcr(n)-۱.۵\,avZcr)+۱] \qquad (۳-۲)$$

طی بررسی که انجام دادیم مقادیر این ویژگی در سیگنال گفتار عموما از یک خط پایه و حداقلی مشخص برخوردار است که تقریبا ثابت بوده و نزدیک به عدد ۵۰ است. در مورد صدای های دیگر تغییرات یا به طور یکنواخت و 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 تغییرات کم تکرار می‌شوند که در این حالت واریانس کمی دارند و بعلاوه 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.

[1]- High Zero-Crossing Rate Ratio (HZCRR)



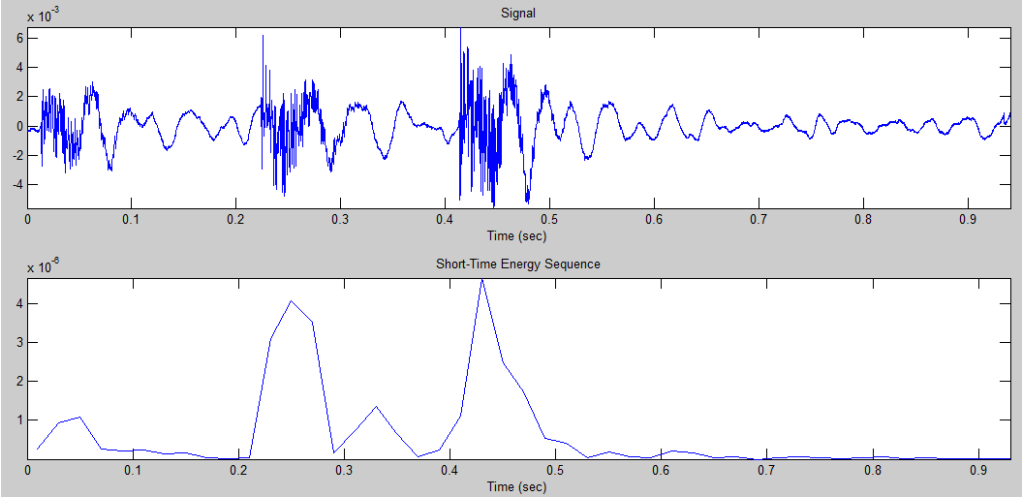
شکل۳-۱۰: نمودار بررسی قابلیت ویژگی عبور از صفر. محور افقی اندیس فریم‌ها و محور عمودی مقدار این ویژگی در فریم را نشان می‌دهد. مقدار این ویژگی در ۹۵۰ فریم نشان می‌دهد که نقاط شروع و پایان برخی از وقایع قابل شناسایی هستند.

۳-۴-۷-۳- ویژگی انرژی کوتاه-مدت[1] فریم‌های صوتی

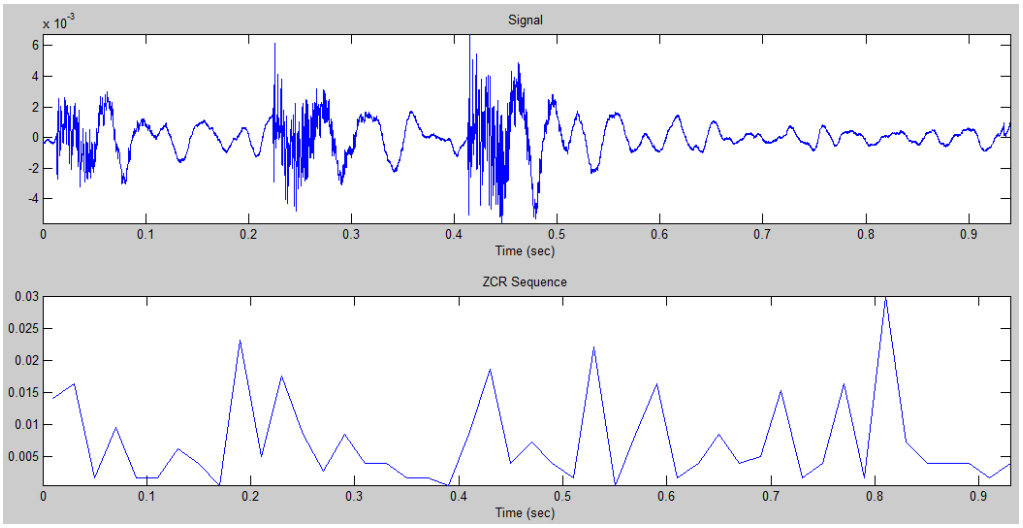
انرژی یک فرم ازحاصل جمع مربعات اندازه هر سمپل درون فریم بصورت $STE = \sum_{n=1}^{n=N} x_n^{\,2}$ بدست می‌آید. در این محاسبه N تعداد کل سمپل‌های درون فریم و x_n بزرگی سمپل n ام را نشان می‌دهد. هر چند طبق آزمایش‌هایی که انجام شد مشخص گردید که این ویژگی در شرایط محیط نویزی چندان ممیز نیست اما پس از حذف نویز و استفاده از روش‌های پیش‌پردازش مشاهده گردید که انرژی فریم‌ها می‌تواند در مورد صداهای خاص اطلاعات مفیدی تولید کند. صداهای کوبه‌ای مانند door knock، door slam، drawer، pen drop و یا صداهای دماغی و صامت در گفتار دارای میزان انرژی بالایی بودند. در سیگنال‌های نمونه مورد آزمایش تعداد خطای شناسایی فریم‌های صداها قبل از پیش‌پردازش توسط ویژگی انرژی در حدود FPR[2]≅۰/۵۸ بود. اما پس از انجام مراحل حذف نویز این میزان به FPR=۰/۴۶ کاهش یافت که بیانگر قابلیت شناسایی صداها توسط این ویژگی است. مزیت

[1]- Short-Time Energy (STE)
[2]- False Positive Rate

خاصی که این ویژگی نسبت به میزان ویژگی عبور از صفر دارد این است که حساسیت آن نسبت به تغییرات دینامیکی در سیگنال خیلی کمتر است لذا جایی که ویژگی عبور از صفر اعوجاج شدیدی را نشان می‌دهد اعوجاج‌ها در این ویژگی کمتر است لذا می‌توانند



شکل۳-۱۱: شناسایی محدوده صدای نمونه knock از طریق توزیع آماری ویژگی انرژی کوتاه مدت در فریم‌ها. در مقایسه با ویژگی عبور از صفرکه در شکل۳-۱۲ آمده است ویژگی انرژی توانایی بهتری در این نوع صداها دارد.



شکل۳-۱۲: تعیین محدوده صدای نمونه knock از طریق توزیع آماری ویژگی نرخ عبور از صفر در فریم‌ها مشکل است. در مقایسه با ویژگی انرژی کوتاه مدت فریم‌ها در شکل۳-۱۱ ، در اینجا اعوجاج های موجود شناسایی را مشکل می‌کند.

تکمیل کننده یکدیگر باشند. در شکل ۳-۱۱ نمایش ویژگی انرژی زمان کوتاه سیگنال صدای knock نمایش داده شده است و امکان شناسایی این صدا با ویژگی مذکور بوضوح دیده می‌شود. همین سیگنال را به همراه

ویژگی‌های عبور از صفر در شکل ۳-۱۲ آورده شده است و اعوجاج‌های موجود کاملا مشخص می‌کند که کار شناسایی با ویژگی عبور از صفر مشکل است.

۳-۷-۴-۴- ویژگی نسبت پایین انرژی کوتاه مدت

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

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

اندیس فریم توسط متغییر n نشان داده می‌شود، میانگین انرژی در فریم‌های مجاور به طول ۱ ثانیه avSTE است که در اینجا هر ۱ ثانیه شامل ۸ فریم ۱۲۵ میلی‌ثانیه است. تعداد کل فریم‌ها N است. از این ویژگی برای شناسایی صداهایی همچون door knock, drawer, pen drop استفاده کرده‌ایم. مقادیر انرژی فریم‌ها در این صداها علاوه بر بالا بودن، تغییرات درونی زیادی نیز دارند. این تغییرات سبب می‌شود نسبت فریم‌های دارای انرژی کم‌تر از متوسط زیاد باشد و به همین علت صداهایی که مرتبا انرژی آن‌ها کم و زیاد می‌شود مقدار این ویژگی برای آن‌ها زیاد است و تا حد زیادی قابل شناسایی هستند.

۳-۸- تشکیل مدل برحسب الگوریتم

در آزمایش‌های مقدماتی مشخص شد که بعضی از صداها فقط با ویژگی عبور از صفر و تغییرات عبور از صفر در طول یک ثانیه قابل شناسایی هستند و در همان حال این فریم‌ها دارای انرژی کمی بودند. علاوه بر این شناسایی فریم‌های بعضی از صداها توسط انرژی و تغییرات آن در طول یک ثانیه امکان پذیر بود که در همان حال این فریم‌ها دارای ویژگی عبور از صفر کوچکی بودند. نکته سوم این بود که در برخی از صداها میزان انرژی و تغییرات آن و میزان عبور از صفر و تغییرات آن همگی یا کم و یا زیاد بودند. در چنین وضعیتی بردارهای ویژگی را به دو گروه زیر تقسیم نمودیم. گروه نخست شامل

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

ویژگی‌های (ZCR, HZCRR) (عبور از صفر و تغییرات آن) و گروه دوم شامل ویژگی‌های (STE, LSTER) (انرژی و تغییرات آن) است. هر گروه از این ویژگی‌ها در زیرسیستم رده‌بندی جداگانه استفاده می‌شود. ابتدا بر روی سیگنال ورودی x(t) پیش‌پردازش‌های لازم که شامل نرمال‌سازی و سپس حذف نویز است انجام می‌شود. استخراج ویژگی در دو مدل متفاوت برای هر کدام از زیرسیستم‌ها صورت می‌گیرد.

زیرسیستم اول از ویژگی‌های گروه اول استفاده می‌کند. این ویژگی‌ها با توجه به آزمایش‌های انجام شده و با انتخاب مناسب طول بین‌ها به اندازه ۵ میلی‌ثانیه امکان شناسایی نقاط خیزش[۱] را در صداهای مختلفی در سیگنال میسر ساخته است. قطعه‌بندی در این زیرسیستم دارای پارامتر α است که امکان یک تصمیم سخت[۲] یا تصمیم نرم[۳] قابل انعطاف را برای انتخاب نقاط کاندید خیزش مهیا می‌سازد. مقدار پارامتر α در اینجا عدد صحیح است که هر چه آلفا بزرگ‌تر باشد میزان دقت کشف وقایع بیش‌تر شده و همزمان میزان فراخوانی به سمت کم شدن میل می‌کند. پس از قطعه‌بندی مرحله کشف وقایع نیز انجام می‌شود که طی آن نقاط خیزش پراکنده توسط الگوریتم تبدیل به قطعه‌های پیوسته دارای زمان شروع و زمان پایان می‌شوند. بخش پیوسته از طریق ادغام شدن فریم‌های ۱۲۵ میلی‌ثانیه‌ای مجاور هم که کاندید بوده‌اند و یا فریم‌های طرفین آن‌ها جزء کاندیدها هستند تشکیل می‌شود. همچنین یک فریم اضافی در انتهای هر بخش پیوسته در نظر گرفتیم تا خاصیت میرایی صداها که معمولا به طور ناگهانی قطع نمی‌شوند را در نظر گرفته باشیم و این کار در اکثر قطعه‌های تشکیل شده تاثیر مثبت داشته است.

زیرسیستم دوم از ویژگی‌های گروه دوم استفاده می‌کند. این دو ویژگی با توجه به آزمایش‌های انجام شده و با انتخاب طول فریم به اندازه ۱۲۵ میلی‌ثانیه تا حد زیادی امکان شناسایی نقاط خیزش در سیگنال را مشخص می‌کنند. این ویژگی‌ها به همراه پارامتر β جهت قطعه‌بندی سیگنال به

۱-On-set point
۲-Hard(sharp) decision
۳-Soft(flexible) decision

زیرسیستم دوم داده می‌شود. قطعه‌بندی این زیرسیستم نقاط خیزش را تعیین می‌کند و پارامتر β که در اینجا عددی اعشاری است امکان می‌دهد انتخاب نقاط خیزش با انعطاف یا سخت‌گیرانه انتخاب شوند. نقاط کاندید برای خیزش معمولا از یک انرژی بیش‌تر و ناگهانی نسبت به فریم‌های مجاور در طول ۱ ثانیه برخوردار می‌باشند. نتیجه قطعه‌بندی این زیرسیستم جهت کشف وقایع ارسال می‌شود. الگوریتم کشف وقایع از نقاط کاندید خیزش که معمولا در نقاط پراکنده‌ای از هم گسترده شده‌اند یک بخش پیوسته با زمان شروع و پایان مشخص بدست می‌آورد و همان روال ادغام موجود در زیرسیستم اول نیز انجام می‌گردد. کلیت مدل شکل ۳-۴ به‌طور خلاصه در شکل ۳-۱۳ نشان داده شده است.

الگوریتم طبق شکل ۳-۱۳ اجرا شده و نتیجه اجرا طبق رابطه (۳-۴) است.

$$X۱(n) = X(n) - Xe(n) \qquad (۳-۴)$$

در رابطه (۳-۴)، $X(n)$ سیگنال ورودی اولیه است. $Xe(n)$ بخش‌هایی از سیگنال اولیه است که به‌عنوان قطعات اولیه وقایع صوتی بدست آمده‌اند و $X۱(n)$ سیگنال باقیمانده است.

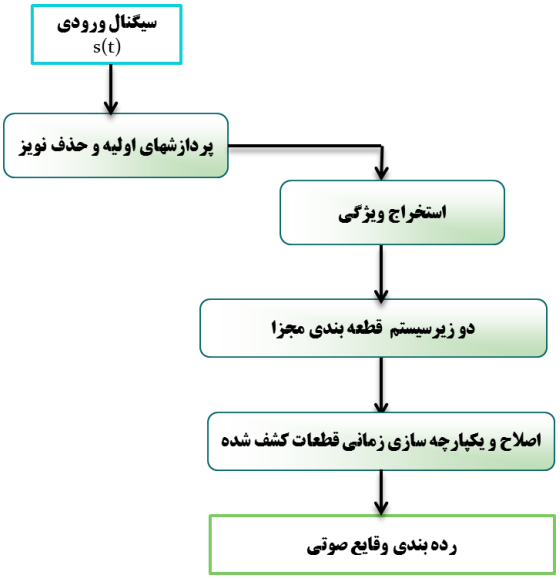
پس از اتمام الگوریتم، قطعه‌های یافته‌شده طبق رابطه (۳-۵) تجمیع شده و در فاز نهایی ترکیب می‌شوند.

$$Xe(n) = Xe۱(n) + Xe۲(n) \qquad (۳-۵)$$

بخش‌ها و فریم‌های کاندید شده برای قطعه‌های وقایع صوتی در زیرسیستم اول و دوم به ترتیب با $Xe1(n)$ و $Xe2(n)$ نشان داده شده‌اند. متغییر Xe(n) خروجی دو زیرسیستم قطعه‌بندی در شکل ۳-۴ است.

آنچه در فاز نهایی کشف وقایع انجام می‌شود این است که خروجی هر دو زیرسیستم قطعه‌بندی را ترکیب می‌نماید و وقایع نهایی را پس از اصلاح زمان‌های شروع و پایان استخراج می‌نماید. چون اعتماد به عملکرد اولیه دو زیرسیستم بالا است لذا فریم‌هایی که در هر کدام از زیرسیستم‌ها یا هر دو انتخاب شده باشند در لیست نهایی خواهند بود. فریمی که زمان شروع زودتری دارد زمان شروع اتفاق

صوتی و فریمی که زمان شروع دیرتری دارد زمان خاتمه واقعه را مشخص می‌کند. این الگوریتم می‌تواند وقایع با فاصله ۱۲۵میلی‌ثانیه از هم را کشف کند.



شکل۳-۱۳: بلاک دیاگرام مراحل مختلف استخراج ویژگی، قطعه‌بندی و کشف وقایع صوتی محیطی در مدل پیشنهادی.

۳-۸-۱- تعیین مقادیر پارامترهای مدل

تعیین بازه موثر برای هر یک از پارامترهای آلفا و بتا یک موضوع کلیدی محسوب می‌شود. لازم است الگوریتم را به ازای کم‌ترین حد و بیش‌ترین حد هر یک از پارامترها اجرا کرده و بازه آن‌ها را مشخص نماییم. لازم به ذکر است که بازه‌های بدست آمده با توجه به داده‌های آموزشی این تحقیق است و بدیهی است که برای کاربرد متفاوت دیگر لازم است بازه معنادار پارامترها بررسی و اصلاح گردند. اما سعی بر این بوده که حتی‌المقدور یک بازه عمومی را تعریف کنیم تا در کاربردهای متعدد دیگری نیز قابلیت استفاده داشته باشند.

در آزمایش‌های متعدد مشخص گردید که مقدار پارامتر α از صفر تا ۷ با درجه آزادی ۱ قابل تغییر است ($\alpha \leq ۰ \leq ۷$) که دارای ۸ درجه متفاوت است. علت محدوده ۰ تا ۷ آن است که به ازای α کم‌تر از صفر عملا فریم‌هایی انتخاب می‌شوند که مقدار ویژگی‌های آن فریم‌ها کم‌تر از حداقل norm مجاز خواهد بود و در این صورت سبب کاهش دقت و افزایش خطای کشف وقایع خواهد شد. حد بالای ۷ به این دلیل انتخاب شد که در آزمایش‌های مکرر مشخص شد به ازای α بزرگ‌تر از ۷ عملا هیچ فریم

جدیدی که متعلق به یک واقعه صوتی باشد انتخاب نخواهد شد و هر فریم انتخابی با این شرایط از فریم‌های غیرصوتی خواهد بود و به همین دلیل خطای کشف را اضافه می‌کند. همچنین در مرحله تعیین بازه پارامتر بتا مشخص شد β را می‌توان از ۱/۰۰ تا ۱/۶۰ با درجه آزادی ۰/۰۱ تنظیم نمود ($۱/۰۰ \leq \beta \leq ۱/۶۰$). تفسیری که در مورد بازه تعیین شده برای α آورده شده در مورد β نیز صدق می‌کند با این تفاوت که این دو پارامتر در دو سیستم جداگانه و با دو مجموعه ویژگی مختلف در ارتباط هستند (مراجعه به شکل‌های ۳-۴و۳-۷). نقش پارامترها در شبه‌کدهای (۳-۶، ۳-۷) که در بخش بعد آمده بیش‌تر مشخص شده است. مقادیر بهینه بدست آمده برای دو پارامتر در بخش آزمایش‌ها و نتایج آورده شده است.

۳-۹- عملیات قطعه‌بندی

جهت کشف وقایع، ابتدا رکورد ورودی را خوانده و پس از نرمال سازی، داده‌ها را حذف نویز نموده و پس از آن سیگنال به دو زیرسیستم اول و دوم ارسال می‌شود. هر کدام از این زیرسیستم‌ها به شیوه خاص خود ویژگی‌ها را استخراج می‌کنند. زیرسیستم اول سیگنال را به فریم‌های مجاور هم به طول زمانی ۱۲۵ میلی‌ثانیه تقسیم می‌کند. با بررسی انجام شده مشخص شد که طول‌های کم‌تر از ۱۲۵ میلی‌ثانیه برای فریم‌ها به علت تغییرات زیاد صداها در بازه‌های کوتاه زمانی و اینکه در فریم‌های کوچک‌تر تشخیص مرز بین صداها به دلیل تنوع کاندید امکان پذیر نبود بازدهی کم‌تری داشتند. برای هر کدام از دو زیرسیستم یک مقدار آستانه عمومی[۱] و ثابت جداگانه تعریف شده است. شبه‌کدهای (۳-۶) و (۳-۷) نقش هر یک از پارامترهای آلفا و بتا را در انتخاب فریم در زیرسیستم‌ها نشان میدهند.

$$IF\ (\ ZC_k >= (\text{norm\_zc} + \alpha)\ .OR.\ HZCRR_k >= ۰/۵۰)$$
then select Frame k as a candidate framebeing in an event. (۳-۶)
Where $(\text{norm\_zc and } \alpha$ *have a predefined values*$)$.

[۱]- Global Threshold

$$IF\left(STE_k >= \left(norm\_En^* \ \beta \right) .OR. LSTER_k <= ۰/۵۰ \right)$$
$$\textit{then select Frame k as a candidate frame being in an event} \qquad (۳-۷)$$
$$\textit{Where } \left(norm\_En \text{ and } \beta \textit{ have a predefined values} \right).$$

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

در زیرسیستم اول ویژگی عبور از صفر به تنهایی قادر است فریم‌های برخی از صداها را مشخص کند اما به دلیل وجود نویز در فریم‌ها و مشابهت صداهای خاصی از جمله Drawer, Phone, Printer به فریم‌های نویز، تعدادی از فریم‌ها بدرستی شناسایی نمی‌شوند لذا از ویژگی میزان عبور از صفر و تغییرات عبور از صفر در این رده‌بند استفاده کردیم. تمام فریم‌هایی که مقدار عبور از صفر آن‌ها بالاتر از $norm+\alpha$ باشد و یا اینکه میزان تغییرات عبور از صفر در طی ۱ ثانیه بیش از ۰,۵ باشد به عنوان کاندید در یک واقعه صوتی انتخاب می‌شوند. نقش پارامتر α کلیدی است به طوری که 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 به صورت شناسایی نقاط خیزش[۱] و افول[۲] استفاده شده است. فریم‌هایی که پس از یک دوره پایداری در انرژی سیگنال به یکباره حداقل به اندازه $norm^*\beta$ افزایش انرژی دارند به عنوان 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 نوسانات آن مجددا انرژی سیگنال به حالت پایدار و با تغییرات کم‌تر از $norm^*\beta$ می‌رسد که این نقطه شروع پایداری مجدد را نقطه افول و عبور از صدای محیطی می‌دانیم. پارامتر β می‌تواند میزان فریم‌هایی که کاندید در صدای اکتشافی می‌شوند را کم یا زیاد کند به طوری که با زیاد شدن β انتخاب سخت‌تر می‌شود.

۱-Onset
۲-Offset

۱-۹-۳- یکپارچه سازی و اصلاح زمانی

هر دو زیرسیستم رده‌بندی فوق در نهایت می‌توانند فریم‌های پراکنده‌ای که با احتمال بالا از فریم‌های درون یک اتفاق صوتی باشند را شناسایی کنند اما محدوده دقیق وقایع هنوز قطعی نیست و ما بخش پیوسته‌ای را نداریم. در فاز بعدی عمل یکپارچه سازی و تعیین دقیق محدوده زمانی وقایع انجام می‌شود. فرآیند تعریف شده در این فاز فریم‌هایی که در نزدیکی هم باشند را در یک واقعه صوتی دانسته و تا ۰/۵ ثانیه فریم‌های گزینش نشده بین دو فریم انتخاب شده را انتخاب کرده و در محدوده واقعه صوتی قرار می‌دهد. در اکثر مواقع فریمی که انتخاب شده است با شرایط خاصی از طرفین خودش قابل بسط است. با توجه به بررسیهایی که انجام شد فریم‌هایی که در جوار یک فریم انتخابی باشند چنانچه مقدارشان حداقل ۹۰ درصد فریم انتخاب شده باشد قابل انتخاب هستند. نکته اصلاحی بعدی انتخاب یک فریم اضافی در انتهای هر واقعه کشف شده است و دلیل آن هم این است که معمولا انتهای صداها مقدار ویژگی‌ها حالت میرایی و رو به زوال دارند لذا افزودن یک فریم به انتهای واقعه صوتی توجیه خیلی بالایی دارد و در آزمایش‌ها مشخص شد که این ایده خیلی موثر است.

فاز نهایی الگوریتم کشف وقایع ترکیب نتایج[۱]دو زیرسیستم رده‌بندی است. در مرحله ترکیب هر گاه واقعه‌ای توسط یکی از زیرسیستم‌ها یا هر دو زیرسیستم کشف شده باشد در لیست نهایی وقایع کشف شده آورده می‌شود.

۱۰-۳- 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 استفاده شده در این مرحله و تعداد هر کدام عبارتند از: عبور از صفر (۱)، انرژی کوتاه مدت (۱)، مرکز ثقل فرکانسی (۱)، بهنای باند فرکانسی (۱)، لگاریتم انرژی بانک

[۱]- Fusion

فیلترهای فرکانسی مل (۱۶)، ضرایب دلتا و دلتای دوم بانک فیلترهای مـل (۳۲)، انـرژی زیربانـدهای چهارگانه (۴)، اسپکترال فلکس هر زیرباند (۴) و آنتروپی یا پیچیدگی درون فریم (۱).

در انتخاب ویژگی‌ها دو جهت کاملا مجزا و مکمل سیگنال یعنی حوزه ادراکی و حوزه تبدیل طیف لحاظ گردید و از ویژگی‌ها مستقل اسـتفاده گردیـد. تعـداد ۱۶ ویژگـی تبـدیل طیـف کـه از لگـاریتم بانک‌های فیلتر مل گرفته شده است بر مبنای درک فرکـانس توسـط سیسـتم شـنوایی انسـان اسـت. مشتقات اول و دوم زمانی آن‌ها تکمیل‌کننده اطلاعات دینامیکی در گذشت زمان هستند کـه در کنـار ویژگی‌های قبلی مکمل فرض می‌شوند. بعلاوه ویژگی‌های دیگری همچون میزان عبور از صفر یا انرژی نیز توصیف کننده رفتار زمانی و مکمل حوزه فرکانس هستند. دلیل انتخاب ما علاوه بـر مکمـل بـودن این ویژگی‌ها و مستقل از هم بودن آن‌ها وجود صداهای متنوع از جمله گفتار، خنده، سرفه از یک سـو و از سوی دیگر صدای تلفن، چاپگر، کلید، آلارم و غیره است که گروه اول وکال یا گلوتال (حنجـره‌ای) هستند و گروه دوم (یعنی صداهای محیطی) بیشتر بدون نظم و شکل خاص هسـتند کـه هـم از نظـر زمانی باید مدل شوند هم از نظرفرکانس.

طول فریم‌ها ۲۵ میلی‌ثانیه و میزان همپوشانی آن‌ها ۶۰ درصد انتخاب گردید. نمایش فرکانسی هر فریم m که آنرا $X_k(m)$, می‌نامیم با استفاده از تبدیل گسسته فوریـه و اعمـال پنجـره هنینـگ $W_b(n)$ بترتیب مطابق رابطه (۸-۳) و رابطه (۹-۳) انجام شده است.

$$X_k(m)=\sum_{n=-\infty}^{\infty} x(n).w_b(n-m).e^{\frac{-j۲\pi kn}{N}} \quad , \quad ۰\le k\le N-۱ \qquad (۳-۸)$$

$$w_b(n)=\begin{cases} ۰.۵-۰.۵.\cos(\frac{۲\pi n}{b}) & , \quad ۰\le n\le b \\ ۰ & , \quad otherwise \end{cases} \qquad (۳-۹)$$

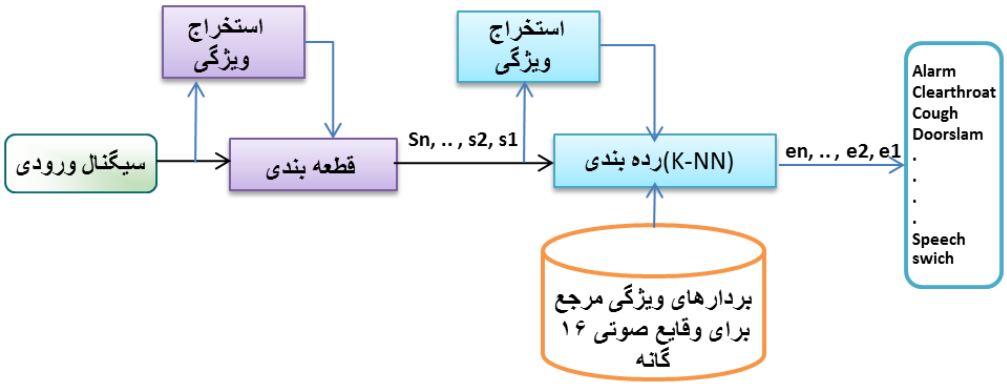
در روابط (۸-۳)، (۹-۳) متغیرهای N, n, k, b به ترتیب چپ به راست طول DFT، انـدیس حـوزه زمان، اندیس حوزه فرکانس و طول پنجره هنینگ می‌باشند. محاسبه مرکز ثقل فرکانسی، اسـپکترال فلکس و انتروپی در رابطه‌های (۳-۱۰ الی ۳-۱۲) بر اساس مراجع [۹۰، ۹۱] آورده شده است.

$$Centroid(m)=\frac{\sum_{k=0}^{N-1}(k+1).|X_k(m)|}{\sum_{k=0}^{N-1}|X_k(m)|} \qquad (۳-۱۰)$$

$$Flux(m)=\sum_{k=0}^{N-1}(|X_k(m)|-|X_k(m-1)|)^2 \qquad (۳-۱۱)$$

$$Entropy(m)=-\sum_{k=0}^{N-1}|X_k(m)|^2.\log_2(|X_k(m)|^2) \qquad (۳-۱۲)$$

پس از فریم‌بندی وقایع، استخراج ویژگی‌ها انجام و تمام بردارهای ویژگی استخراجی از فریم‌ها نرمال‌سازی شده و سپس از الگوریتم K-Nearest Neighbors algorithm (KNN) همانند شکل ۳-۱۴ برای رده بندی هر فریم استفاده می‌شود. پس از رده‌بندی فریم‌های هر سگمنت یک تصمیم گیری نهایی بر پایه رای‌گیری و کسب بیش‌ترین رای برای تعیین کلاس مربوط به واقعه صوتی انجام می‌شود. برای تخمین دقیق‌تر مقدار K آزمایش‌های متعددی بر روی مقادیر K =۱، ۳، ۵، ۷، ۹، ۱۱ انجام شد که درموارد زیادی مقدار K=۷ بهترین نتیجه را در بر داشت.



شکل ۳-۱۴: پس از قطعه‌بندی از الگوریتم K-NN با مقدار K=۷ برای رده بندی استفاده شده است.

۳-۱۱- آزمایش‌ها و نتایج

مدل پیشنهادی را در قالب الگوریتم شکل ۳-۱۳ با آزمایش‌های مکرر آزمایش کرده و نتایج را بدست آوردیم. جهت آزمایش‌ها از پایگاه داده تهیه شده در انجمن بین المللی تخصصی پردازش سیگنال صوت[۱] که داده‌های مرتبط به صحنه‌های شنیداری مختلف از جمله صداهای موجود در اتاق کار یا

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

جلسه که حاوی ۱۶ نوع اتفاق صوتی مطابق جدول ۳-۱ را شامل می‌شود استفاده کردیم. 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 متغییر ضبط شده‌اند لذا صداها دارای شرایط محیط زنده[۱]هستند و صداهای مصنوعی[۲] محسوب نمی‌شوند. به طور کلی از تعداد ۶۰ رکورد مختلف استفاده کردیم و جهت اجرای سریع‌تر آزمایش‌ها هر بار یک گروه ۱۰ رکوردی را برای آزمایش انتخاب کرده و با آن گروه نتایج را بدست آوردیم. انتخاب رکوردهای هر گروه شامل انتخاب ۱۰ رکورد به صورت تصادفی و غیر تکراری از بین ۶۰ رکورد موجود می‌باشد. لذا در آزمایش‌ها نتایج گروه‌های ۱ تا ۶ را جداگانه بدست آورده و نهایتا میانگین گروه‌ها را به عنوان نتیجه کلی آزمایش ۶۰ رکورد در نظر گرفتیم. الگوریتم را با معیارهای دقت[۳] میزان فراخوانی[۴] و میانگین هارمونیک وزنی[۵] (یا معیار کارایی) مورد سنجش قرار دادیم. تعریف این معیارها با توجه به نیاز الگوریتم در رابطه‌های (۳-۱۳) تا (۳-۱۵) معرفی شده‌اند.

$$\text{Precision}=\frac{\text{تعداد کل وقایع صوتی که الگوریتم بدرستی آن‌ها را شناسایی کرده است}}{\text{تعداد کل وقایعی که الگوریتم شناسایی کرده است}} \quad (۳-۱۳)$$

$$\text{Recall}=\frac{\text{تعداد کل وقایع صوتی که الگوریتم بدرستی آن‌ها را شناسایی کرده است}}{\text{تعداد واقعی وقایع در رکورد صوتی ورودی}} \quad (۳-۱۴)$$

$$\text{F1-measure}=\frac{۲*(Precision*Recall)}{(Precision+Recall)} \quad (۳-۱۵)$$

منظور از معیار دقت در این الگوریتم تعداد وقایع درست نسبت به کل وقایع کشف شده توسط الگوریتم است. معیار فراخوانی نشان می‌دهد الگوریتم چه درصدی از کل وقایع موجود صوتی در رکوردهای مورد پردازش را توانسته به درستی کشف کند. معیار F1-measure میانگین وزنی دو معیار دیگر است لذا از اهمیت بیش‌تری برخوردار است و به تنهایی می‌تواند ملاک سنجش کارایی الگوریتم باشد.

۱- Live Environment
۲- Synthesis Sounds
۳- Precision
۴- Recall
۵- F1measure

در اولین آزمایش میزان عملکرد الگوریتم بدون در نظر گرفتن پارامترهای آلفا و بتا که معادل α=۰ و β=۱/۰۰ است را بدست آورده و در جدول ۳-۲ ارائه کرده‌ایم. میانگین میـزان دقـت ۴۴/۹ درصـد و میانگین فراخوانی ۴۷/۶ درصد و میزان کارایی میانگین ۴۶/۱ درصد است. این شرایط اولیـه الگـوریتم است و می‌توان با دخالت دادن پارامترها عملکـرد سیسـتم را بهبـود داد. بعـلاوه در آزمـایش دیگـری پارامترهای آلفا و بتا را در محدوه کامل مقادیرشان بکـار گـرفتیم و نتـایج آنـرا در جـدول ۳-۳ نشـان داده‌ایم. طبق آنالیز بخش ۳-۳ محدوده کامـل بـرای آلفـا (۰:۱:۷) و بـرای پـارامتر بتـا (۱/۶۰: ۰/۰۱: ۱/۰۰) است. محدوده کامل پارامترها به این صورت اجرا می‌شود که هر پارامتر را به ازای تمام مقـادیر پارامتر دیگر اجرا و نتایج را جمع نموده و میانگین می‌گیریم. همانطور که از مقایسه جدول‌های (۲-۳) و (۳-۳) مشخص است استفاده از پارامترها عملکرد سیستم را به طور محسوسی افـزایش داده اسـت و این بیانگر دینامیک عمل کردن سیستم با تغییر پارامترها است. میانگین دقت به میزان ۱۰/۵ درصـد، فراخوانی به میزان ۶/۵ درصد و کارایی نیز ۸/۲ درصد بهبود داشته است. بهبود بیش‌تـر میـزان دقـت، بیانگر تـاثیر بیش‌تـر و مسـتقیم پارامترهـا در مـدل مـذکور بـر کـم شـدن خطـای کشـف اسـت. در آزمایش‌های بعدی مقدار آلفا و بتاهای حداکثر عملکرد را بدست آورده‌ایم.

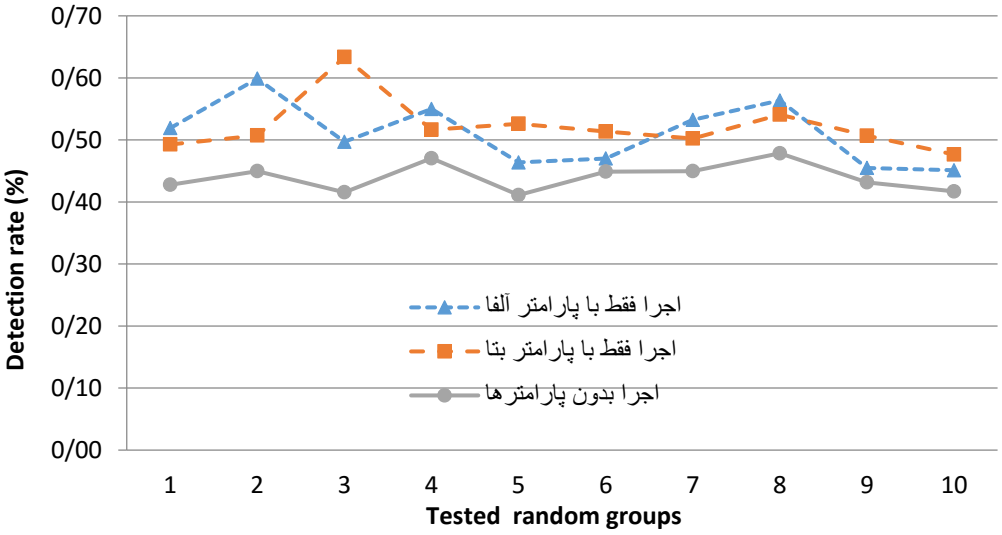
جدول ۳-۲: عملکرد الگوریتم بدون اعمال پارامترها (هر Group شامل ۱۰ رکورد از بین ۶۰ رکورد آزمایش و بدون تکرار است)

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

جدول ۳-۳: عملکرد الگوریتم با در نظر گرفتن محدوده کامل برای پارامترهای آلفا و بتا

Precision	Recall	F1-measure	Group	Averaging on:
۰/۶۰۴	۰/۶۱۵	۰/۶۰۹	۱	
۰/۵۷۱	۰/۶۰۲	۰/۵۸۶	۲	
۰/۵۱۶	۰/۵۳۵	۰/۵۲۵	۳	α= (۰:۱:۷)
۰/۵۴۳	۰/۶۱۷	۰/۵۷۸	۴	
۰/۵۹۶	۰/۳۸۳	۰/۴۶۶	۵	
۰/۴۹۶	۰/۴۸۹	۰/۴۹۲	۶	β=(۱/۰۰ :۰/۰۱ :۱/۶۰)
۰/۵۵۴	۰/۵۴۰	۰/۵۴۳	میانگین	

یکی از اهداف بکارگیری پارامترها افزایش قابلیت شناسایی صداها است. جهت بررسی تاثیر هر یک از پارامترها، ابتدا ۱۰ گروه هر کدام شامل ۶ رکورد با انتخاب تصادفی جهت آزمایش مشخص شد. با هر گروه سه اجرای مختلف انجام شد. اجرای نخست بدون استفاده از پارامترها، اجرای دوم فقط پارامتر آلفا و در اجرای سوم فقط پارامتر بتا اثر داده شد. در هر اجرا از نتایج هر گروه جداگانه میانگین گرفته شد و درصد معیار F1 از میانگین محاسبه گردید. این نتایج در شکل ۳-۱۵ آورده شده است. همانطور که مشخص است میزان شناسایی در اجراهای دوم و سوم در تمام گروه‌ها بالاتر از اجرای اول در همان گروه است.



شکل ۳-۱۵: 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 به طور قابل ملاحظه‌ای افزایش دارد. در بعضی موارد پارامتر آلفا و در بعضی دیگر پارامتر بتا تاثیر بیشتری در افزایش نرخ شناسایی دارد.

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 انجام شده است که هدف یافتن بهترین آلفا و بتا برای حداکثرسازی یکی از معیارهای سه گانه بود. نتایج حداکثرسازی معیار میزان دقت در جدول ۳-۴ نشان داده شده است. ستون‌های α و β مقادیری را نشان می‌دهند که به ازای آن‌ها میزان دقت بدست آمده در رکوردهای ورودی حداکثر شده است. سطر آخر جدول ۳-۴ میانگین را بیان می‌کند. حداکثر دقت ۶۶/۲ درصد و پارامترهای میانگین (β =۱/۴۱ و α =۵) می‌باشد. میانگین مقادیر پارامترها از جدول ۳-۴ را به عنوان مقادیر بهینه برای حداکثر کردن دقت در سیستم در نظر گرفته و سیستم را با این دو مقدار پارامتر اجرا کردیم. حاصل اجرا با مقادیر بهینه برای دقت در جدول ۳-۵ آمده است. از آنجا که نتیجه جدول ۳-۵ با میانگین جدول ۳-۴ کاملا مطابقت دارد می‌توان آنرا اینگونه تفسیر نمود که امید

ریاضی برای میزان دقت در سیستم حداکثر ۶۶/۱ درصد است و می‌توان سیستم را با این پارامترها در وضعیت کشف حداکثری قرار داد.

$$E(\text{حداکثر دقت}) = ٪\ ۶۶/۱ \quad \alpha=۵ \text{ و } \beta = ۱/۴۱ \qquad (۳-۱۶)$$

در گام بعدی آزمایش‌های متعددی جهت یافتن مقدار پارامترها برای داشتن میزان فراخوانی بالا با هدف حداکثرسازی تعداد وقایع کشف شده صورت پذیرفت که نتایج در جدول ۳-۶ نشان داده شده است. سطر آخر میانگین را نشان می‌دهد. مقدار بهینه پارامترها در وضعیت فراخوانی حداکثری برابر (β=۱/۱۶ و α=۵) می‌باشد.

جدول ۳-۴: حداکثرسازی میزان دقت برحسب انتخاب مقدار بهینه برای پارامترهای α و β در سیستم

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

جدول ۳-۵: میانگین عملکرد الگوریتم به ازای مقادیر بهینه حداکثرسازی دقت.

Precision	Recall	F1-measure	α = ۵
۰/۶۶۱	۰/۴۰۳	۰/۵۰۱	β =۱/۴۱

در ادامه آزمایش مجددی را با استفاده از مقادیر میانگین پارامترها در وضعیت فراخوانی حداکثری را اجرا نمودیم که نتایج جدول ۳-۷ بدست آمد. این نتایج نشان از همگرایی مدل با پارامترهای مشخص شده برای کشف حداکثری دارد و با نتیجه میانگین جدول ۳-۶ نیز مطابقت دارد. بنابراین امید ریاضی میزان فراخوانی سیستم برابر ۵۸/۴ درصد می‌باشد:

$$E(\text{حداکثر فراخوانی}) = ٪\ ۵۸/۴ \quad \alpha =۵ \text{ و } \beta = ۱/۱۶ \qquad (۳-۱۷)$$

جدول ۳-۶: مقدار بهینه پارامترهای α و β جهت حداکثرسازی میزان فراخوانی در سیستم

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

جدول ۳-۷: میانگین عملکرد الگوریتم به ازای مقادیر بهینه حداکثرسازی فراخوانی

Precision	Recall	F1-measure	α=۵
۰/۵۴۴	۰/۵۸۴	۰/۵۶۳	β=۱/۶۱

می‌توان میزان عملکرد سیستم را با معیار کارایی[۱] ارزیابی نمود. دلیل اهمیت این معیار محاسبه وزنی و هارمونیکی آن توسط دو معیار دیگر است. در ادامه آزمایش‌های مقادیر معیار کارایی را ارزیابی نموده و آنرا توسط پارامترها بهینه‌سازی کردیم. با نگاه به جدول ۳-۲ می‌بینیم که این معیار بدون دخالت پارامترها ۴۶/۱ درصد بوده است. نتایج حداکثرسازی میزان کارایی را در جدول ۳-۸ نشان داده‌ایم.

علاوه بر این یک آزمایش کلی مجدد نیز با مقادیر میانگین پارامترها که در جدول ۳-۸ بدست آمد انجام دادیم که نتایج آن در جدول ۳-۹ نشان داده شده است. همانطور که مشخص است نتایج جدول ۳-۹ با میانگین جدول ۳-۸ بسیار نزدیک است و کاملا مطابقت دارد. بنابراین می‌توان امید ریاضی میزان کارایی سیستم را با پارامترهای بدست آمده برابر۵۷/۳ درصد دانست:

$$E(\text{حداکثر کارایی}) = ٪\ ۵۷/۳ \qquad \alpha = ۶ \text{ و } \beta = ۱/۲۱ \tag{۳-۱۸}$$

در جدول ۳-۱۰ مقادیر معیارها بر اساس پارامترهای بهینه α و β به صورت مقایسه ای آورده شده است.

[۱]-F1measure

طبق مقادیر جدول ۳-۱۰ میزان دقت بر اساس میانگین برابر ۶۰/۶ درصد، مقـدار فراخـوانی برابـر ۵۰/۹ درصد و مقدار کارایی ۵۴/۶ درصد است و مقدار پارامترها برابر $\beta = ۱/۲۶$ و $\alpha=۵$ است که این مقادیر را می‌توان امید ریاضی کل سیستم تلقی نمود.

$$E(\text{عملکرد کلی سیستم}) = (۶۰٪/۶ \quad ۵۰٪/۹ \quad ۵۴٪/۶) \quad \alpha = ۵, \ \beta = ۱/۲۶ \qquad (۳-۱۹)$$

جدول ۳-۸: مقدار بهینه پارامترهای α و β جهت حداکثرسازی میزان کارایی در سیستم

Precision	Recall	F1-measure:max	α	β	Group
۰/۶۷۰	۰/۵۶۸	۰/۶۱۵	۶	۱/۲۰	۱
۰/۶۳۳	۰/۵۲۷	۰/۵۷۵	۵	۱/۳۱	۲
۰/۵۴۲	۰/۵۳۴	۰/۵۳۸	۷	۱/۱۶	۳
۰/۶۲۸	۰/۶۰۴	۰/۶۱۶	۶	۱/۲۲	۴
۰/۵۹۷	۰/۳۵۹	۰/۴۴۸	۷	۱/۱۵	۵
۰/۶۱۲	۰/۶۱۳	۰/۶۱۲	۷	۱/۲۵	۶
۰/۶۱۴	۰/۵۳۴	۰/۵۶۷	۶	۱/۲۱	میانگین

جدول ۳-۹: میانگین عملکرد الگوریتم به ازای مقادیر بهینه حداکثرسازی کارایی

Precision	Recall	F1-measure	α=۶
۰/۶۱۲	۰/۵۳۹	۰/۵۷۳	β=۱/۲۱

جدول ۳-۱۰: مقایسه مقادیر معیارها بر اساس پارامترهای بهینه

Precision	Recall	F1-measure:max	α	β	
۰/۶۶۱	۰/۴۰۳	۰/۵۰۱	۵	۱/۴۱	
۰/۵۴۴	۰/۵۸۴	۰/۵۶۳	۵	۱/۱۶	
۰/۶۱۲	۰/۵۳۹	۰/۵۷۳	۶	۱/۲۱	
۰/۶۰۶	۰/۵۰۹	۰/۵۴۶	۵	۱/۲۶	میانگین

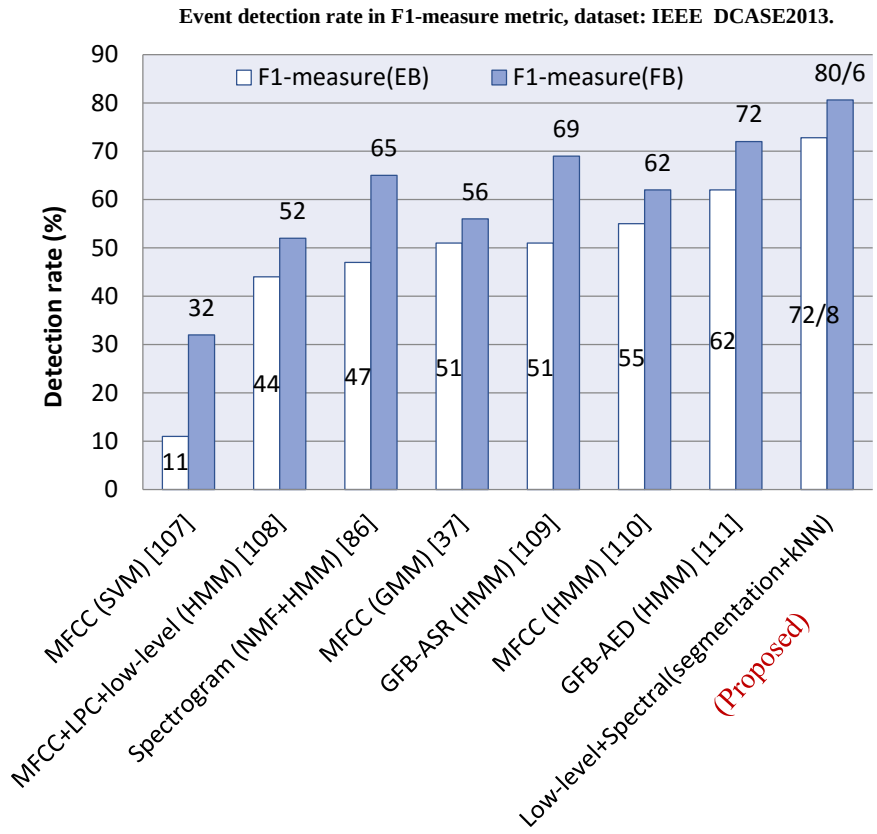
جهت مقایسه نتایج با کارهای مشابه ابتدا لازم اسـت نتـایج معیـار F1 کـه در جـداول ۳-۲ تـا ۳-۱۰برحسب کلیپ (یا رکورد) و میانگین بین رکوردها محاسبه شده است و اصـطلاحا record-based (RB)می‌باشد را بصورت یکپارچه و بر اساس فـریم Frame-based (FB) و سـپس بـر اسـاس واقعـه event-based (EB)محاسبه نماییم تا بتوان نتایج را با موارد مشابه مقایسه نمود. نتایج پردازش فریمی

و برحسب واقعه در جـدول ۳-۱۱ آورده شـده اسـت. بهبـودی نتـایج در سـطح فریمـی (FB) بـدلیل تشخیص بهتر اطلاعات در فریم نسبت به سطح بالاتر واقعه (EB) است و طبیعاتا شناسایی یک واقعـه در حد چند ثانیه که متشکل از تعداد زیادی فریم است کار مشکل‌تر و همراه با خطـای بیشـتر اسـت. 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 هـر گروه است که نهایتا لحاظ نمودن نتایج در رکوردها به طور مستقل و میانگین بین آن‌ها، سـبب نتـایج ضعیف‌تر RB نسبت به EB گردیده است. اما آنچه در نتایج مورد قیاس قرار می‌گیرد میزان معیـار F1 بر اساس FB, EB است.

جدول ۳-۱۱: مقادیر معیار ارزیابیF1 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 اجرای مدل پیشنهادی با پژوهش‌های مرتبط در شکل ۳-۱۶ نشان داده شـده اسـت. نتایج مشخص می‌کند که مدل پیشنهادی بیش از ۱۰/۸٪ بهبودی نتایج را داشته است و این به دلیـل مدل‌سازی بهتر نسبت به روش‌های قبلی است.



شکل۳-۱۶: درصد شناسایی وقایع برحسب معیار F1 در سیستم پیشنهادی (proposed) در دو مقیاس فریم (FB) و واقعه (EB). مجموعه داده‌های استفاده شده مربوط به IEEE AASP (۲۰۱۳) در شاخه AED است که شامل ۱۶ نوع صدای محیطی اداری می‌باشد.

فصل

۴- روش پیشنهادی دوم: جداسازی سیگنال‌ها برای کشف و رده‌بندی وقایع صوتی محیطی

مقدمه

کشف و شناسایی وقایع صوتی از طریق پردازش داده‌های فضای دوبعدی زمان-فرکانس همواره مورد توجه پژوهشگران قرار داشته است. دلیل این استقبال بدست آمدن دینامیک و تغییرات لحظه‌ای و دراز مدت از طریق شکل طیف سیگنال در قالب اسپکتروگرام می‌باشد. به طوری که در کنار روش‌های سنتی پردازش سیگنال که در قالب حوزه زمان یا حوزه فرکانس و معمولا با مکانیزم فریم‌بندی و استخراج ویژگی‌ها در قالب زمآن‌های خیلی کوتاه انجام می‌شود توجه به شکل طیف فرکانسی سیگنال در گذر زمان و اتفاقاتی که در انرژی باندهای فرکانسی مختلف بروز می‌کند روز به روز در حال افزایش است. از طرفی اگر نگاه دقیقی به شکل طیف فرکانسی سیگنال در اسپکتروگرام داشته باشیم الگوهای متعددی از انرژی دیده می‌شود که در کنار هم قرار گرفتن این الگوها سبب تولید شکل کلی اسپکتروگرام خواهد شد. لذا پژوهش‌گران بسیاری از خصوصیات اسپکتروگرام به عنوان راهکاری برای استخراج ویژگی استفاده نموده و آن را مورد مطالعه جدی قرار داده‌اند. به بیان واضح‌تر چنانچه الگوهای موجود در اسپکتروگرام استخراج شده و به عنوان اجزای اصلی سیگنال نگه‌داری شوند دو اتفاق مهم رخ می‌دهد. اول آنکه یک سیگنال با مقایسه الگوهای طیفی آن با یک سری الگوهای پایه و اصلی قابل شناسایی است و دوم اینکه از کنار هم قرار دادن الگوها می‌توان یک سیگنال را دوباره‌سازی نمود و همچنین در بازسازی آن سیگنال از این الگوها استفاده نمود. لذا پس از مشخص شدن اینکه شکل طیف فرکانسی سیگنال در گذر زمان اطلاعات زیادی به همراه دارد زمینه‌های بکارگیری این اطلاعات در چند موضوع گسترش پیدا کرد و الگوریتم‌های متعددی پیشنهاد شد. از کاربردهای مهم این الگوهای پایه می‌توان به موارد متعددی اشاره نمود. از جمله می‌توان حذف نویز از سیگنال، بهبود کیفیت سیگنال اصلی، تفکیک صداهای همزمان از یک سیگنال ترکیب‌شده، خلاصه‌سازی و فشرده کردن سیگنال‌ها از طریق نگه‌داری داه‌های مهم و حذف داده‌های فرعی یا پس‌زمینه، شناسایی صداها و محیط‌های صوتی، استخراج ویژگی از سیگنال، تولید دیکشنری از الگوهای مهم موجود در یک سیگنال را از مهمترین این کاربردها برشمرد. همچنین در راستای پیاده‌سازی این کاربردها می‌توان به

الگوریتم‌های متعددی از جمله ویولت، جداسازی مولفه‌های مستقل (ICA)، تجزیه یک سیگنال به ماتریس‌های عامل نامنفی (NMF) اشاره نمود. از ویولت تا کنون در کاربردهای بسیاری در حوزه تجزیه و تحلیل سیستم‌ها که با سیگنال در ارتباط هستند استفاده شده است. یکی از موارد مهم این کاربردها تعیین بافت و ساختار یا تغییرات دینامیکی پدیده‌های صوتی و یا تصویری در حوزه زمان-فرکانس است. علاوه بر اینها از ویولت در فشرده‌سازی و استخراج ویژگی‌ها نیز به طور وسیعی استفاده شده است.

از الگوریتم ICA به خصوص در جداسازی منابع تولید سیگنال در یک سیگنال ترکیبی به کرار استفاده شده است و روش‌هایی همچون جداسازی کور منابع (BSS) که در مورد یک سیگنال ترکیبی انجام می‌شود در همین راستا بوجود آمده است.

الگوریتم‌های تجزیه به عوامل نامنفی یا NMF نیز در حوزه‌های متعددی از جمله داده‌کاوی، پیش بینی در بورس، سیستم‌های پیشنهاد[1] سیستم‌های رتبه بندی اتوماتیک، حذف نویز از تصویر، فشرده‌سازی داه‌های صوتی و تصویری، پردازش سیگنال و استخراج ویژگی صوتی یا تصویری، تولید دیکشنری‌های تنک و به طورکلی در کدنویسی تنک کاربرد دارد. نقاط قوت NMF باعث شده است که سمت و سوی مطالعات زیادی در حوزه پردازش سیگنال متوجه آن شده و با گسترش مفاهیم آن کاربردهای متعدد آن نیز بسط پیدا کند. از جمله کاربردهای NMF می‌توان به جداسازی صداها در محیط‌های صوتی واقعی که تعداد متعددی از وقایع صوتی همراه با گفتار وجود دارند اشاره نمود. مثلا در یک اتاق سخنرانی غیر از گفتار سایر وقایع صوتی همچون موزیک پس‌زمینه، صدای همهمه و خنده شنوندگان و غیره نیز وجود دارد. در این حالت چنانچه هدف شناسایی گفتار و متن سخنرانی باشد، قبل از شروع عملیات شناسایی اتوماتیک گفتار (ASR)[2] توسط الگوریتم‌های مربوطه لازم است هر واقعه صوتی غیر از گفتار در پیش‌پردازش‌ها حذف شوند که این کار از طریق شناسایی وقایع صوتی امکان‌پذیر است و در اینجا NMF می‌تواند در کنار سایر ابزارها و الگوریتم‌ها بکار گرفته شود. لذا در

[1]- Recommendation systems
[2]- Automatic Speech Recognition (ASR)

محیط‌های صوتی مرکب یکی از ابزارهای کمکی برای سیستم‌های ASR می‌تواند شناسایی صداهای دیگر و به معنای دیگر شناسایی و کشف وقایع صوتی غیر از گفتار باشد که همزمان کمک بزرگی نیز به سیستم ASR می‌کند. از اینرو طراحی الگوریتم‌های موثر برای کشف وقایع صوتی (AED[1]) که با هدف تبدیل یک جریان صوتی به رده‌های معنایی وقایع و تعیین زمان وقوع هر کدام از آن‌ها انجام می‌شود نه تنها به طور کلی در مواردی همچون آنالیز محتوایی صوتی و تفکیک و ارزیابی اطلاعات مفید است بلکه می‌تواند در کمک به سیستم‌های ASR نیز مفید باشد.

با اینکه کشف وقایع صوتی و سیستم‌های ASR می‌توانند مکمل هم باشند اما وجه متفاوت AED و ASR آن است که در AED ساختارهای ذاتی موجود در گفتار همچون فونوم‌ها و کلمات که در الگوریتم‌های ASR از آن‌ها در مدل‌سازی HMM یا SVM استفاده می‌شود وجود ندارد. لذا نگاشت مستقیم نمایش فریمی به رده‌های مفهومی سبب بروز خطاهای بزرگ در مدل‌سازی می‌شود. برای رفع این مشکل ما با در نظر گرفتن ساختار زمان-فرکانسی سازمان‌یافته که در تعدادی از فریم‌های پیوسته بصورت یک وصله طیفی در وقایع صوتی محیطی وجود دارند، یک نمایش تنک ویژگی بر اساس اتم‌ها (وصله‌های طیفی) برای AED در نظر می‌گیریم که ذاتا از رده‌بندی الگوهای تصویری الهام گرفته شده است. در این روش ابتدا یک دیکشنری صوتی آموزش داده می‌شود به طوری که اتم‌های صوتی در این دیکشنری یک گروه از وصله‌های طیفی صداهای مورد پردازش خواهند بود. سپس یک نمایش تنک با استفاده از یک مقیاس شباهت بین سیگنال ورودی با دیکشنری صوتی استخراج می‌گردد. از این ویژگی‌ها معمولا جهت آموزش مدل‌هایی همچون SVM یا HMM استفاده می‌شود ولی ما از همین دیکشنری در فاز آزمایش برای بازشناسی سیگنال ورودی استفاده می‌کنیم و معمولا از مقیاس شباهت که ممکن است ار نوع فاصله اقلیدسی، انحراف کولبک لیبلر[2]و یا ایتاکورا سایتو[3] است بین سیگنال صوتی و دیکشنری صوتی استفاده می‌شود. تنکی ویژگی‌ها نیز یا از طریق

[1]- Audio Event Detevtion (AED)
[2]- Kullback–Leibler
[3]- Itakura saito

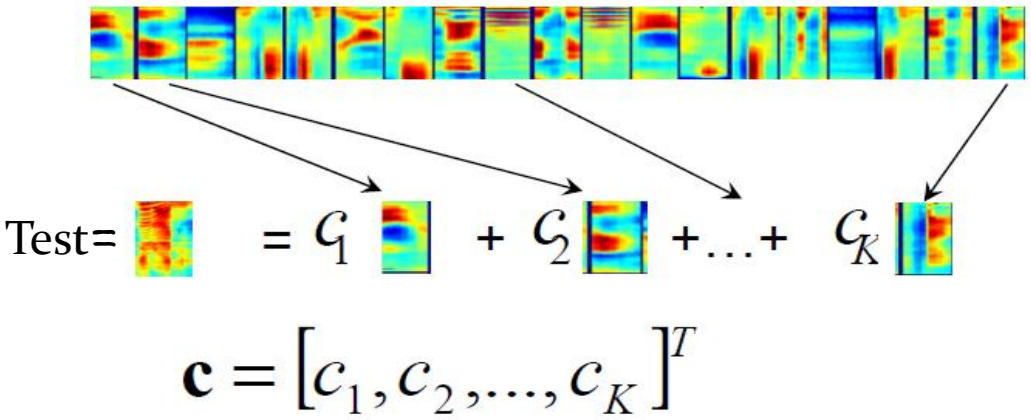
تنظیم دستی یک آستانه برای صفر کردن مقادیر کوچک و یا از طریق شرط ضمنی برای تابع بهینه هدف که میزان تنکی و خطای بازسازی را با هم در نظر می‌گیرد اعمال می‌شود. تا کنون روش‌های مختلفی جهت بهبود عملکرد سیستم‌ها از این طریق پیشنهاد شده است اما هنوز جای مطالعات فراوانی وجود دارد. در سالیان اخیر از ایده الگوریتم‌های جداسازی کور منابع و با استفاده از الگوریتم‌های تجزیه نامنفی ماتریس، پردازش سیگنال مشاهده مد نظر قرار گرفته است که می‌توان به مقالات [۲۹-۳۵] اشاره نمود. اما بدلیل پایه‌ای نبودن این روش‌ها در شناسایی وقایع صوتی نتایج بدست آمده در مقایسه با روش‌های مرسوم و مدل‌سازی‌های آماری راضی کننده نیست. هرچند که روش‌های پایه نیز در شرایط خاص نتایج خوبی داشته‌اند اما در حالت عمومی هنوز پاسخگویی لازم را ندارند [۸۵]. لذا در شناسایی وقایع صوتی محیطی تعریف نوعی از همبستگی بین اجزای جریان صوتی ورودی با الگوهای طیفی از پیش تعیین شده مناسبت بیش‌تری دارد. به طور نمونه، تجزیه نامنفی ماتریس، جریان صوتی ورودی را به ترکیبی خطی از اتم‌های دیکشنری با ضرایب خاص تبدیل می‌کند. در این حالت ماتریس ضرایب فعال کننده به عنوان ویژگی‌های بدست آمده از سیگنال ورودی جهت رده‌بندی صداها استفاده می‌شود. از چالش‌های مطرح در این نوع تجزیه می‌توان به یکسان نبودن سیگنال بازسازی‌شده با سیگنال اولیه اشاره نمود. به طوری که افزوده شدن یا جایگزینی اجزای صوتی بجای یکدیگر می‌تواند منجر به خطای عدم همخوانی سیگنال بازتولیدی با سیگنال اصلی گردد [۸۶]. یک فرض ما این است که اعمال کنترل بر نحوه تجزیه ماتریس مشاهده و چگونگی نگاشت سگمنت‌های ورودی بر دیکشنری وقایع سبب شناسایی بهتر وقایع صوتی می‌گردد. فرض دیگر ما این است که سیگنال ورودی دارای اجزای کوچک‌تر سازمان‌یافته است. برای مثال در رونویسی موزیک پلی‌فونیک اجزای سیگنال تعدادی نت یا وقایع صوتی در حال نواختن فرض می‌شوند. در این حالت شناسایی هر نت یا اتفاق صوتی اهمیت دارد و شناسایی اجزای کوچک‌تر مد نظر نیست.

بنابراین ارائه تکنیک‌های قابل انعطاف جهت کنترل بر چگونگی تجزیه برحسب تغییرات جاری در سیگنال لازم است. بعلاوه اعمال کنترل بر نحوه نگاشت سیگنال صوتی ایده دیگری برای تجزیه

سیگنال ورودی به اجزای کوچک‌تر معنادار می‌باشد. فرض ما این است که چنین کنترل‌هایی بتواند اهمیت بیش‌تری برای اجزای مهم و خاص در تجزیه در نظر بگیرد. اگر در رکوردهای صوتی محیطی، وقایع صوتی اجزای مهم هستند پس اعمال این نوع کنترل‌ها می‌تواند سبب بهبود شناسایی این اجزا در صداها شود. لذا در اینجا، هدف اهمیت دادن به سگمنت حاوی یک واقعه صوتی است.

در شکل ۴-۱ نمونه‌ای از سیستم انتخاب K اتم یا وصله صوتی به عنوان ویژگی جهت نمایندگی از سیگنال اصلی آورده شده است. سیگنال آزمایش ورودی با استفاده از یک ضریب خطی از اتم‌ها بازسازی شده است.

اسپکتروگرام سیگنال و K اتم استخراجی به نمایندگی از آن



$$\mathbf{c} = [c_1, c_2, \ldots, c_K]^T$$

شکل ۴-۱: یک تقریب خطی از سیگنال ورودی Test با استفاده از K اتم استخراجی از اسپکتروگرام. جهت نمایش نمونه Test، ضرایب C در اتم‌ها ضرب شده و حاصل جمع زده می‌شود.

در مسائل نمایش سیگنال ورودی با اتم‌های یک دیکشنری معمولا از یک تابع بهینه‌سازی جهت یافتن ضرایب ترکیب اتم‌ها استفاده می‌شود. این تابع بصورت یک انحراف در بازسازی سیگنال ورودی از روی اتم‌های دیکشنری تعریف می‌شود و معمولا شرایط اضافی مثل میزان تنکی بر بهینه‌سازی افزوده می‌شود. در رابطه (۴-۱) نوعی از تجزیه با شرایط تنک بودن آورده شده است.

$$h^* = \arg\min \frac{1}{2}\|t - Wh\|_2 + \lambda|h|_1 \qquad (۴-۱)$$

در رابطه (۱-۴) t یک نمایش برداری از اسپکتروگرام سیگنال ورودی است، W دیکشنری است که مجموعه‌ای از اتم‌های طیفی را در خود دارد که معمولا کدبوک هم گفته می‌شود، h بردار پارامتر است که به عنوان نمایش t روی W استفاده می‌شود. در رابطه (۴-۱) ترم اول میزان درستی تقریب را بدست می‌آورد، ترم دوم با میزان تنکی نمایش مرتبط است و پارامتر لاندا (λ) به منظور مصالحه میان میزان درستی تقریب و میزان تنکی نمایش استفاده می‌شود. اگر λ=۰ باشد مسئله به یک رگراسیون عادی تبدیل می‌شود. وجود λ بمعنی این است که اندکی خطای بازسازی سیگنال یا نمایش سیگنال نسبت به سیگنال اصلی قابل پذیرش است.

۱-۴- اندازه‌گیری میزان تنکی و کنترل آن در نمایش ویژگی

تنکی روش‌ها راه مناسبی برای کاستن از حجم فضای مورد نیاز پردازش‌ها در تجزیه نامنفی در مسائل پیچیده است. در خصوص شناسایی صداها نیاز به کنترل میزان تنکی، در طرح‌های بهینه‌سازی پیچیده و سخت کاملا احساس می‌شود. در اینجا مسئله کنترل تنکی از بعد تئوری آن مورد بررسی قرار گرفته و تعدادی از روش‌های اندازه‌گیری میزان تنکی را مرور می‌کنیم. نیاز به کنترل تنکی در تجزیه نامنفی ماتریس را بیان نموده و دو تکنیک بهینه‌سازی آن یعنی تصویر گرادیان[۱] و برنامه نویسی مخروطی درجه دوم را معرفی می‌کنیم. با اعمال این روش‌ها تجزیه نامنفی با کنترل تنکی بسط می‌یابد. از تطبیق الگوریتم کاهش گرادیان تکنیک بهینه‌سازی دیگری بنام برنامه ریزی کوژ درجه دوم بدست می‌آید.

یک تعریف ساده از میزان تنکی[۲] یک بردار، تعداد عناصر صفر یا پوچ در آن بردار است. با این وجود توافق کلی بر تعریف و نحوه محاسبه تنکی بردار حاصل نشده است لذا روش‌های متعددی در این خصوص پیشنهاد شده است. ایده اصلی برای تنک بودن یک بردار یا یک نمایش از اسپکتروگرام این است که نمایش متراکم نباشد بلکه تعداد کمی از اجزا در بر گیرنده بخش اعظم انرژی آن

۱- projected gradient optimization

۲- sparseness (or sparsity)

اسپکتروگرام باشند. اندازه‌گیری میزان تنکی بدنبال این است که از نظر عددی مشخص کند چه مقدار از انرژی در تعداد اندکی از اجزا ذخیره شده است. کد نویسی تنک نیز در پی آن است کـه تعـداد هـر چه کم‌تر از اجزا را طوری پیدا کند که بیش‌ترین اطلاعات یا بیش‌ترین بار اطلاعاتی آن مجموعه را در خود داشته باشند طوری که این تعداد اندک نماینده‌های کامل برای نمایش آن مجموعه کامل باشـند. این بحث مستقیما به استخراج ویژگی، به موضوع PCA، با اهداف SVD در ارتباط است.

یک برآورد ساده از تنکی یک بردار x بطول n تعداد عناصر غیرصـفر آن اسـت کـه بـه نـرم صـفر ℓ_0 (Zero-norm) معروف است که رابطه (۴-۲) آنرا بیان نموده است.

$$sp(x)=\frac{\|x\|_0}{n}=\frac{card\{i:x_i\neq 0\}}{n} \qquad (۴-۲)$$

تعریف مشتق‌پذیر دیگر برای میزان تنکی بـردار روی محـدوده $\mathbb{R}^n_{++}$ از طریـق نـرم p (ℓ_p) در محدوده $0<p\leq 1$ صورت رابطه (۴-۳) قابل تعریف است:

$$\mathrm{sp}(x)=\frac{\|x\|_p^p}{n}=\frac{\sum_i |x_i|^p}{n} \qquad (۴-۳)$$

در تجزیه نامنفی یک روش اندازه‌گیری میزان تنکی بصورت رابطه (۴-۴) تعریف شده است:

$$\mathrm{sp}(x)=\frac{\sqrt{n}-\|x\|_1/\|x\|_2}{\sqrt{n}-1}=\frac{\sqrt{n}-\sum_i |x_i|/\sqrt{\sum_i x_i^2}}{\sqrt{n}-1} \qquad (۴-۴)$$

در رابطه با افزایش تنکی در بردار x مقـدار $\mathrm{sp}(x)$ نیـز افـزایش می‌یابـد. ایـن رابطـه همچنـین مستقل از مقیاس مقادیر درون بردار است (یعنی اگر تمام مقادیر بردار به یک نسبت تغییر یابند مقدار $\mathrm{sp}(x)$ حاصل تغییری نمی‌کند و ثابت است). رابطه فوق برای بردار پوچ تعریـف نشـده اسـت و بـرای برداری که فقط یک مقدار غیرصفر دارد برابر ۱ است.

۴-۲- تجزیه نامنفی ماتریس

تکنیک کاهش رتبه ماتریس، از جمله تجزیه نامنفی با داشتن ماتریس نامنفی V_{n*m} و یک عدد صحیح $r\leq \min(n,m)$ ماتریس V را به صورت رابطه (۴-۵) تجزیه می‌کند:

$$V_{n*m}\approx W_{n*r} * H_{r*m} \qquad (۴-۵)$$

در یادگیری ماشین ستون‌های ماتریس V مشاهدات از یک پدیده در محیط و ردیف‌ها نیز ویژگی‌های مختلف هر نمونه هستند. در این صورت هر بردار مشاهده v_j از طریق رابطه (۴-۶) محاسبه می‌شود.

$$v_j \approx \sum_{k=1}^{r} w_k h_{kj} \qquad (۴-۶)$$

هر یک از h_j, w_k ستون j ام و k ام در ماتریس هستند، W دیکشنری و ستون‌های آن پایه‌ها[۱] الگوها یا اتم‌ها هستند. ماتریس H ضرایب یا فعال کننده اتم‌های دیکشنری است. در زمان تصویرسازی، هر ستون ماتریس H کدگذاری ستون متناظر در ماتریس V می‌باشد.

از آنجا که رابطه (۴-۵) تجزیه تقریبی به فرم $V \approx W*H$ است، لذا از طریق بهینه‌سازی یک تابع هزینه انجام می‌شود. اگر $\Lambda = W*H$ باشد و تابع هزینه $C(V,\Lambda)$ نامیده شود آنگاه تجزیه به صورت یک مسئله بهینه‌سازی محدودشده به صورت رابطه (۴-۷) خواهد بود:

$$\arg\min_{W \in R_+^{n*r}, H \in R_+^{r*m}} C(V,\Lambda) \qquad (۴-۷)$$

جهت کمینه شدن فاصله اقلیدسی، هزینه از طریق نرم فروبنیوس طبق رابطه (۴-۸) محاسبه می‌شود:

$$C(A,\Lambda) = \frac{۱}{۲}\|V-\Lambda\|_F^۲ = \frac{۱}{۲}\sum i,j\,(V_{ij}-\lambda_{ij})^۲ \qquad (۴-۸)$$

از نظر آماری رابطه (۴-۸) در واقع حداکثرسازی یک تابع شباهت[۲] توسط کمینه نمودن میزان خطا برحسب فاصله اقلیدسی است.

تابع بهینه‌سازی تا زمان همگرایی یا به تعداد تکرار مشخص اجرا می‌شود. در تجزیه نامنفی استاندارد برای بهنگام‌سازی عامل‌های W,H از الگوریتم ضربی که در [۳۱، ۸۸، ۹۲] آمده است استفاده می‌شود که نوعی گرادیان نزولی[۳] همراه با مراحل تطبیقی رابطه (۴-۹) است:

۱- basis vectors

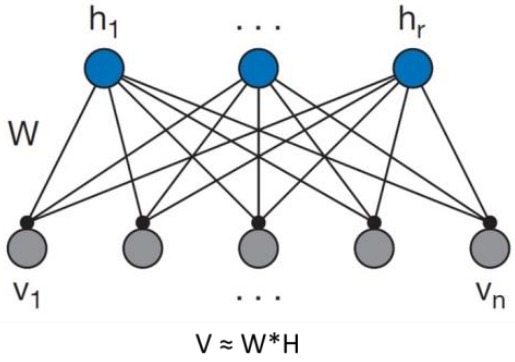
۲- similarity function

۳- gradient descent

$$H \leftarrow H \otimes \frac{|W^T V|}{W^T WH} \quad , \quad W \leftarrow W \otimes \frac{|VH^T|}{WHH^T} \tag{۴-۹}$$

بهنگام‌سازی طوری است که در پایان اجرا تقریب $V \approx W*H$ دارای کم‌ترین انحراف یا فاصله است. با توجه به شرط کمینه شدن تابع هزینه، بهنگام‌سازی در هر تکرار انجام می‌شود. مقادیر W,H همواره مثبت هستند و مقدار تابع هزینه نیز باید به طور یکنواخت کاهش یابد.

الگوریتم‌های دیگری برای بهبود تجزیه نامنفی وجود دارد که معمولا از مدل‌های تصحیح‌شده، محدودیت‌های تصحیح‌شده مثل افزودن عامل تنکی، توابع هزینه تصحیح‌شده همچون استفاده از واگراها[۱] یا افزودن عامل‌های جبران‌کننده[۲] استفاده می‌نمایند [۹۳-۹۶]. در تئوری اطلاعات تجزیه نامنفی $V \approx W*H$ بیانگر یک مدل احتمالاتی با متغیرهای پنهان h است که با داشتن متغیر مشاهدات V و فضای تبدیل W در پی تعیین یک توزیع تخمینی برای H است که چگونگی نگاشت توزیع V بر توزیع W را تعیین نماید. این دیدگاه مانند شکل ۴-۲ با رویت بردار مشاهده V و داشتن ماتریس W بهترین تجزیه $V \approx W*H$ همراه با تعیین ضرایب فعال‌سازی بردارهای پایه در دیکشنری را اجرا می‌کند. در رویکرد ما برای حل موضوع، این دیدگاه دنبال شده است.



شکل ۴-۲: تجزیه نامنفی $\boldsymbol{V \approx W * H}$ به صورت یک مدل احتمالاتی با متغیرهای پنهان $\boldsymbol{H}$ و متغیر مشاهده شده $\boldsymbol{V}$ همراه با ماتریس انتقال $\boldsymbol{W}$ قابل تعریف و مدل‌سازی است.

[۱]- divergences

[۲]- penalty terms

۴-۳- تجزیه نامنفی ماتریس و بهنگام‌سازی با انحراف بتا

انحراف بتا یکی از انواع توابع پارامتریک مغایرت[۱]نوعی تصحیح‌کننده برای تابع هزینه محسوب می‌شود. برای هر $\beta \in \mathrm{R}$ و هر نقطه $a,b \in \mathrm{R}_{++}$ انحراف بتا از a به b را می‌توان مانند رابطه (۴-۱۰) تعریف کرد [۳۱].

$$d_{\beta}(a,b) = \frac{۱}{\beta(\beta-۱)}(a^{\beta} + (\beta-۱)y^{\beta} - \beta a b^{\beta-۱}) \qquad (۴-۱۰)$$

اگر از رابطه (۴-۱۰) نسبت به β حد گرفته شود آنگاه به ازای $\beta = 0$ انحراف ایتاکورا-سایتو رابطه (۴-۱۱) و به ازای $\beta = 1$ انحراف کولبک لیبلر رابطه (۴-۱۲) بدست می‌آید:

$$d_{\beta=0}(a,b) = d_{IS}(a,b) = (a/b) - log(a/b) \qquad (۴-۱۱)$$

$$d_{\beta=1}(a,b) = d_{KL}(a,b) = a log(a/b) + b - a \qquad (۴-۱۲)$$

برای حالت $\beta = 2$ تعریف فوق مانند رابطه (۴-۱۳) نصف فاصله مربع اقلیدسی است:

$$d_{\beta=۲}(a,b) = d_{E}(a,b) = \frac{۱}{۲}(a-b)^{۲} \qquad (۴-۱۳)$$

انحراف بتا مطابق رابطه (۴-۱۳) همواره نامنفی است مگر در حالت $a=b$ که نتیجه صفر خواهد شد.

یک ویژگی انحراف بتا در ارتباط با تجزیه سیگنال‌ها این است که برای هر ضریب بزرگنمایی[۲] $\lambda \in \mathrm{R}_{++}$ رابطه (۴-۱۴) وجود دارد:

$$d_{\beta}(\lambda a, \lambda b) = \lambda^{\beta} d_{\beta}(a,b) \qquad (۴-۱۴)$$

انحراف عددی رابطه (۴-۱۰) را می‌توان در پردازش سیگنال به یک ماتریس انحراف تحت عنوان انحراف تفکیکی[۳]تبدیل نمود. این کار با جمع عنصر-محور انحراف‌ها مانند رابطه (۴-۱۵) قابل انجام است:

$$\mathrm{D}_{\beta}(V,\Lambda) = \sum_{i,j} d_{\beta}(\upsilon_{ij}, \lambda_{ij}) \qquad (۴-۱۵)$$

---

[۱]- contrast function

[۲]- scaling factor

[۳]- separable divergence

بنابراین در پردازش سیگنال تجزیه نامنفی با تابع هزینه انحراف بتا منجر به مسئله بهینه‌سازی مقیدشده رابطه (۴-۱۶) می‌گردد که هم ارز رابطه (۴-۷) است.

$$\arg\min \mathrm{D}_{\beta}(V,WH) \qquad (۴-۱۶)$$
$$W \in R_{+}^{n*r}, H \in R_{+}^{r*m}$$

رابطه (۴-۱۶) در پردازش ماتریس‌های مشاهده V و دیکشنری W و فعال‌کننده‌های اتم‌های دیکشنری H به صورت رابطه (۴-۱۷) قابل بیان است:

$$\mathrm{D}_{\beta}(V,WH) = \sum_{n=1}^{N}\sum_{m=1}^{M} d_{\beta}([V]_{nm}\,|\,[WH]_{nm}) \qquad (۴-۱۷)$$

منظور از $[X]_{nm}$ عنصر واقع در سطر n و ستون m ماتریس X است و d_{β} نیز تابع هزینه نوع انحراف بتا می‌باشد.

تا کنون الگوریتم‌های بهنگام‌سازی ضربی متعددی برای حل تجزیه نامنفی با تابع هزینه انحراف بتا با بسط‌های مختلف معرفی شده‌اند. با اقتباس از رابطه (۴-۹) یک بهنگام‌سازی ضربی اکتشافی[۱] برای عامل‌های W,H می‌تواند مانند رابطه (۴-۱۸) باشد:

$$H \leftarrow H \otimes \frac{W^{T}(V \otimes (WH)^{\beta-2})}{W^{T}(WH)^{\beta-1}} \quad , \quad W \leftarrow W \otimes \frac{(V \otimes (WH)^{\beta-2})H^{T}}{(WH)^{\beta-1}H^{T}} \qquad (۴-۱۸)$$

تابع هزینه انحراف بتای کولبک لیبلر در رابطه (۴-۱۹) آمده است:

$$C(A,\Lambda) = D_{KL}(V,\Lambda) = \sum_{i,j} v_{ij} log \frac{v_{ij}}{\lambda_{ij}} + \lambda_{ij} - v_{ij} \qquad (۴-۱۹)$$

این تابع با جایگزینی در رابطه (۴-۱۷) بصورت رابطه (۴-۲۰) است:

$$\mathrm{D}_{\beta=1}(V,WH) = D_{KL}(V,WH) = \sum_{n=1}^{N}\sum_{m=1}^{M}(V_{nm}\log\frac{V_{nm}}{[WH]_{nm}} + [WH]_{nm} - V_{nm}) \qquad (۴-۲۰)$$

با تغییر تابع هزینه، بهنگام‌سازی ماتریس‌های W,H نیز طبق روابط موجود در [۸۸، ۹۲] به رابطه (۴-۲۱) تغییر می‌یابند:

---

[۱]- heuristic

$$H \leftarrow H \otimes \frac{W^T (V \otimes (WH)^{-1})}{W^T WH} \quad , \quad W \leftarrow W \otimes \frac{(V \otimes (WH)^{-1}) H^T}{WHH^T} \qquad (۴-۲۱)$$

به ازای $\beta=۱$ و $\beta=۲$ بهنگام‌سازی ضربی به ترتیب معادل کولبک لیبلر و اقلیدسی است. هر چند هزینه در بهنگام‌سازی‌ها به ازای $۰\leq\beta\leq۲$ به طور یکنواخت کم می‌شود اما این وضعیت برای سایر مقادیر β صدق نمی‌کند [۹۷]. در اینجا شناسایی صداها با روش تجزیه انحراف بتا و روش محدودشده با قید تنک انجام شده‌اند.

۴-۴- تجزیه نامنفی ماتریس با محدودیت تنک جهت کشف وقایع صوتی

تجزیه نامنفی رابطه (۵-۴) که بسط آن در روابط (۶-۴) تا (۹-۴) آمده است با یک معضل اساسی روبرو است و آن نرسیدن به پاسخ‌های انحصاری و ثابت است به طوری که با هر بار اجرا نتایج متفاوتی حاصل می‌شود. جهت رفع این معضل محدودیت‌هایی از جمله تجزیه تنک و بخصوص تجزیه با محدودیت نرم l_1 پیشنهاد شده است. با اعمال محدودیت تنک انتظار می‌رود ماتریس‌های حاصل از تجزیه دارای مقادیر زیادی صفر در سطرها و ستون‌ها باشند. در چنین حالتی سیگنال تجزیه‌شده با تعداد کم‌تری از اتم‌های دیکشنری با ضرایب غیرصفر نمایش داده می‌شود. لذا محدود بودن تعداد بردارهای پایه سبب نتایج انحصاری در نمایش سیگنال می‌شود. بسط‌های مختلفی از جمله افزودن عامل جریمه تنکی[۱] و یا استفاده از بهنگام‌سازی ضربی همگرا همراه با عامل جریمه نرم l_1 و تصویرسازی گرادیان نزولی بمنظور اطمینان از نامنفی بودن و سپس نرمال سازی l_1 سایر عامل‌ها پیشنهاد شده است. شرط تنک بودن می‌تواند روی دیکشنری W یا ضرایب H و یا روی هر دو ماتریس اعمال شود. یک رابطه بهینه‌سازی برای تجزیه نامنفی با محدودیت تنک در [۹۸] بصورت رابطه (۲۲-۴) معرفی شده است:

[۱]- sparsity penalty

$$arg\,min \frac{1}{2} h^T (W^T W + \lambda_2 I) h + (\lambda_1 e - W^T v)^T h \ , \quad -Ih \le 0 \quad , \quad \lambda_1, \lambda_2 \ge 0 \quad , \quad h \in R_+^r \tag{۴-۲۲}$$

λ_1, λ_2 پارامترهای تنظیم کننده هستند. این رابطه یک بهنگام‌سازی درجه دوم کوژ[۱][۹۹] است که با مسئله کم‌ترین خطای مربعات نامنفی به صورت رابطه (۴-۲۳) قابل بیان است:

$$arg\,min \frac{1}{2} \| v - Wh \|_2^2 + \lambda_1 \| h \|_1 + \frac{\lambda_2}{2} \| h \|_2^2 \tag{۴-۲۳}$$

در رابطه (۱۹) نرم l_1 بردارهای کم تنک را جریمه می‌کند، نرم l_2 یک حالت خاص از تصحیح‌کننده تیکونوو[۲]است که سبب مثبت شدن ماتریس $P = W^T W + \lambda_2 I$ برای هر $\lambda_2 > 0$ می‌شود و این موضوع کوژ بودن مسئله را تضمین می‌کند [۹۸]. به دلیل هزینه محاسباتی بالای مسائل کوژ درجه دوم در اینجا برای اعمال محدودیت تنک در تجزیه نامنفی حالت خاصی از این مسئله که به فرم حداقل خطای مربعات است در رابطه (۴-۲۴) استفاده می‌شود. محدودیت تنک شامل ماتریس W و ماتریس ضرایب H می‌شود [۹۳]:

$$arg\,min \frac{1}{2} \| V - WH \|_F^2 + \frac{\eta}{2} \| W \|_F^2 + \frac{\lambda}{2} \sum_{i=1}^{n} \| h_i \|_1^2 \quad subject\,to\ W, H \ge 0 \tag{۴-۲۴}$$

h_i ستون iام در ماتریس H است. از دید بیز فرمول فوق احتمال لگاریتم پیشینی[۳]تحت شرایط بررسی خطای گوسی است که در آن بردارهای پایه W دارای توزیع گوسی و بردارهای ضرایب H توزیع لاپلاسی دارند. در اینجا جهت تجزیه نامنفی از بهینه‌سازی مدل رابطه (۲۰-۲۴) به فرم حداقل سازی خطای مربعات استفاده شده است و ضمن ثابت نگه داشتن یک ماتریس، بهنگام‌سازی ماتریس دیگر انجام می‌گیرد به طوری که محدودیت تنک در هر دو ماتریس دیکشنری و ماتریس فعال کننده اعمال گردد. در واقع بهنگام‌سازی اعمال‌شده حالت خاصی از معادله درجه دوم کوژ را در نظر می‌گیرد.

۱- convex quadratic program(CQP)

۲- Tikhonov regularization

۳- log-posterior

در پردازش صداها، ماتریس مشاهدات V معمولا یک نمایش زمان-فرکانس از صداهای مورد پردازش را در خود ذخیره می‌کند. ردیف‌ها باریکه‌های فرکانسی مختلف و ستون‌ها فریم‌های زمانی پیاپی هستند. در این حالت تجزیه $v_j \approx \sum_{k=1}^{r} w_k h_{kj}$ از مجموع حاصل ضرب بردارهای پایه w_k در ضرایب تجزیه h_{kj} بدست می‌آید. در این تجزیه هر اتم یا بردار پایه w_k از دیکشنری الگوی طیفی kام را می‌سازد. ضریب تاثیر h_{kj} وزن الگوی kام در فریم مشاهده شده v_j در زمان jام را تعیین می‌کند. از یک نظر می‌توان رویکرد مدل‌های آماری با متغیرهای پنهان [۱۰۰] را با تکنیک تجزیه نامنفی مقایسه کرد. در آنجا داده‌های نامنفی به صورت یک توزیع گسسته فرض می‌شود که با یک مدل مخلوط گوسی تجزیه می‌گردد به طوری که هر جزء پنهان[۱] بیانگر یک منبع است. در این مدل‌ها تخمین شباهت حداکثر[۲] پارامترهای مخلوط، منجر به یک تجزیه نامنفی با تابع هزینه واگرایی کولبک لیبلر می‌شود. در واقع الگوریتم EM متناظر با الگوریتم بهنگام‌سازی ضربی در تجزیه نامنفی است.

۵-۴- مشخصات پایگاه داده

در اینجا پایگاه داده و نوع صداهای مورد پردازش توضیح داده می‌شود. همانطور که در فصل ۳ اشاره شد، در این رساله از پایگاه داده تهیه‌شده در انجمن بین‌المللی تخصصی پردازش سیگنال صوت[۳] استفاده شده است. این داده‌ها مرتبط به صحنه‌های شنیداری مختلف از جمله صداهای موجود در اتاق کار یا جلسه است که حاوی ۳۲۰ نوع صدا در قالب ۱۶ نوع واقعه صوتی می‌باشد. صداهای مورد پردازش عبارتند از

[۱]- latent component

[۲]- maximum likelihood estimation

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

$alarm(short\ alert\ beep\ sound),\ clear\_throat,\ cough,\ door\_slam,$
$drawer,\ keyboard(keyboard\ clicking),\ keys(keys\ put\ on\ table),$
$knock(door\ knocking),\ laughter,\ mouse,\ page\_turn$
$(turning\ pages\ back\ \&\ forward),\ pen\_drop$
$(pen,\ pencil,\ or\ marker\ touching\ table\ surfaces),$
$phone(ringing\ phone),\ printer,\ speech,\ switch.$

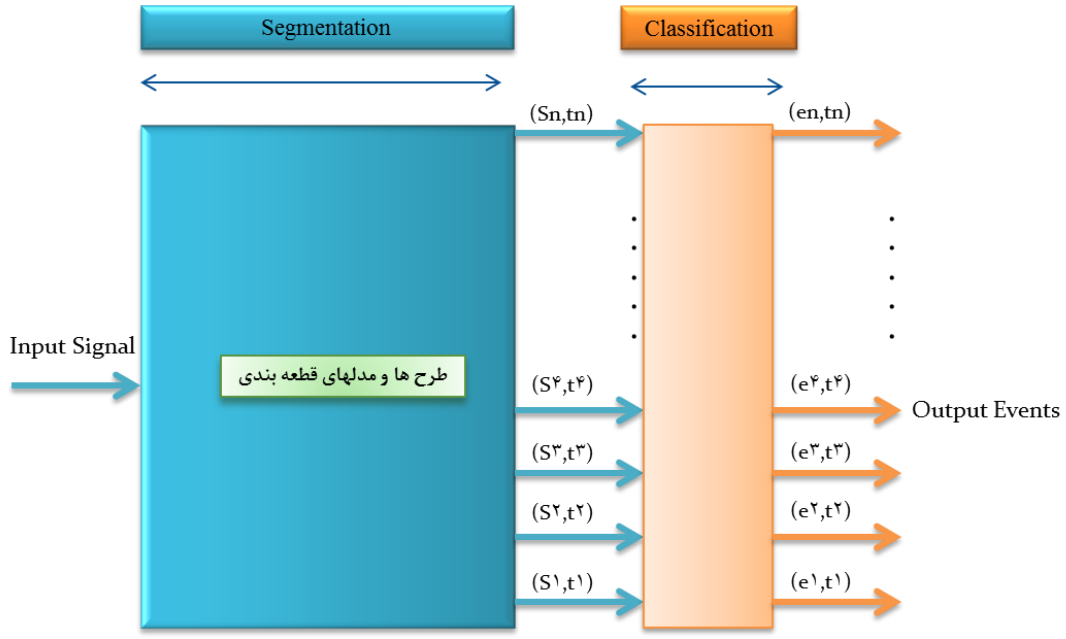
صداهای مذکور با نرخ نمونه برداری ۴۴۱۰۰ هرتز دو کاناله ضبط شده‌اند. طول صداها متنوع بوده و حتی نمونه‌های مختلف از یک صدا نیز طول یکسان ندارند. به طور نمونه صدای printer از حدود ۹ ثانیه تا حداکثر ۲۲ ثانیه است. صدای pen_drop از ۲۵۰ میلی‌ثانیه تا حداکثر ۱۲۰۰ میلی‌ثانیه است. صدای alert از ۳۵۰ میلی‌ثانیه تا ۳ ثانیه است. این صداها در محتوای فرکانسی و شکل متفاوت هستند و اغلب آن‌ها از جهات مختلفی متغیر محسوب می‌شوند. برای مثال صداهای drawer, printer, switch, phone and page_turn دارای طیفی با یکنواختی طولانی، اجزای نویزی مهم و همراه با ترکیب طیفی در سطح لحظه‌ای و میکرونی هستند. صداهای switch و phone از خشنی خاصی برخوردار هستند. صداهای laughter، speech،clear_throat، cough، keys و door_slam دارای طول متوسط بوده اما همگی دارای الگویی متغیر لحظه‌ای در طیف خود هستند. صداهای alarm، mouse، keyboard، knock و pen_drop حالت ایستایی بیش‌تری دارند و دارای طیفی با حمله‌های ساکن و ادامه‌دار می‌باشند.

۶-۴- آزمایشات در سیستم پیشنهادی

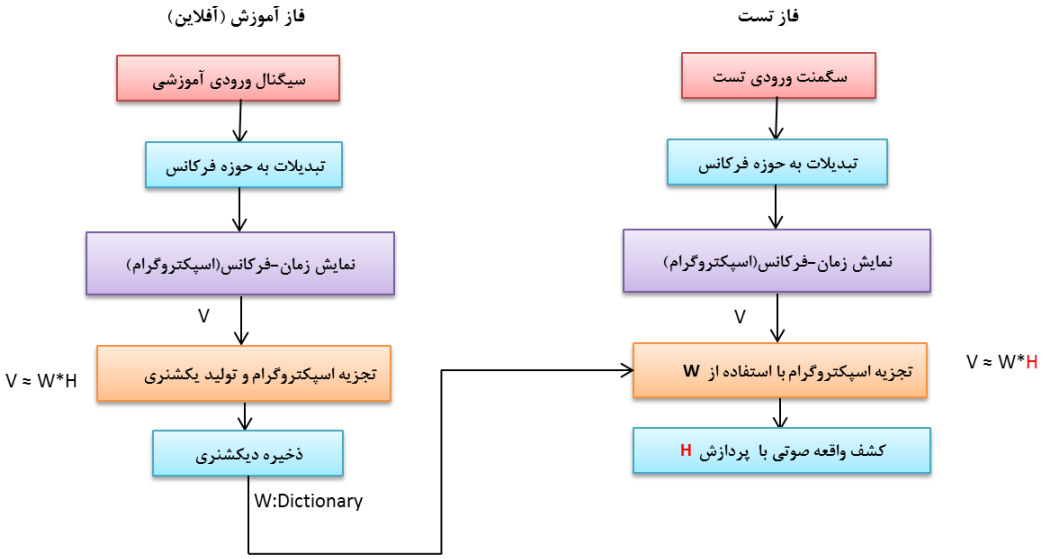
در این بخش سیستم پیشنهادی برای کشف و رده‌بندی وقایع صوتی توضیح داده شده و آزمایش می‌شود. در شکل ۴-۳ موضوع به صورت دو فاز مختلف در یک سیستم یادگیری ماشین نشان داده شده است. در این سیستم کشف و رده بندی صداهای محیطی در دو فاز کشف (قطعه‌بندی) و رده‌بندی (کلاس‌بندی) فرض می‌شود. سیگنال ورودی ابتدا قطعه‌بندی شده و قطعه‌های s_1 تا s_n هر کدام در بازه‌های زمانی مشخص t_1 تا t_n حاصل می‌گردد. سپس در فاز کلاس‌بندی، نوع رده یا اتفاق صوتی رخ داده در هر قطعه مشخص می‌شود. وقایع صوتی رده‌بندی شده e_1 تا e_n نامیده شده است.

از آنجا که مدل پیشنهادی بر پایه نگاشت سگمنت بر دیکشنری در نمایش تنک است لذا کنترل بر نحوه تجزیه و نگاشت بردار مشاهده بر دیکشنری جهت اجرای فازهای شکل ۴-۳ به صورت یک مـدل توسعه یافته در شکل ۴-۴ با فازهای آموزش و آزمایش پیشنهاد شده است.

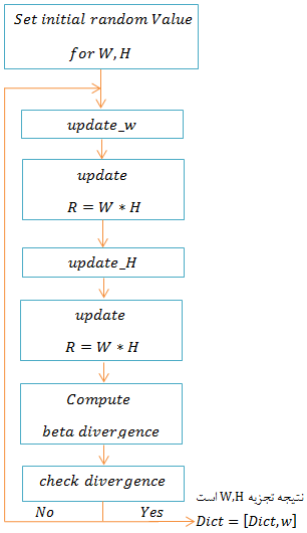
در شکل ۴-۵ مراحل مختلف تجزیه نامنفی یک ماتریس ورودی v نشان داده شده اسـت. در ایـن الگوریتم نمونه ورودی v پس از چندین تکرار الگوریتم و بهنگام‌سازی مکرر W,H تجزیه می‌گـردد و نتیجه این تجزیه ماتریس W است که به دیکشنری $Dict$ اضافه می‌گردد. متغییر استفاده شده eps با مقدار مثبت بسیار نزدیک صفر امکان وقوع خطای تقسیم بر صفر در مرحله بهنگام‌سازی عامل‌هـا را از بین می‌برد.



شکل ۴-۳: سیستم کشف و رده‌بندی صداهای محیطی در دو فاز کشف (قطعه‌بندی) و رده‌بندی (کلاس‌بندی) انجام می‌شود. در فاز کشف قطعه‌های $s1$ تا sn که هر کدام حاوی یک اتفاق صوتی هستند با محدوده‌های زمانی $t1$ تا tn بدست می آید. در فاز رده‌بندی هر واقعه صوتی که در هر قطعه رخ داده شناسایی می‌شود و وقایع رده‌بندی شده $e1$ تا en نامیده می‌شوند.



شکل ۴-۴: سیستم کشف و رده‌بندی صداهای محیطی در مدل تجزیه نامنفی. فاز آموزش در قسمت سمت چپ شکل بصورت آفلاین انجام می‌شود و منجر به تولید یک دیکشنری از وقایع مورد نظر می‌گردد. در بخش سمت راست شکل با استفاده از عملیات تجزیه نامنفی سگمنت ورودی بر دیکشنری تصویرسازی شده و واقعه صوتی استخراج می‌شود.



$$NMF\ \ decomposition\ \ for\ \ V \approx W * H$$
$$Updating\ \ rules:$$
$$R = W * H$$
$$W = W.*(((R.\wedge(\beta-2).*V)*H^T./\max(R.\wedge(\beta-1)*H^T, eps)$$
$$H = H.*(((W^T*(R.\wedge(\beta-2).*V)./\max(W^T*R.\wedge(\beta-1), eps))$$
$$eps = 1e-20$$

شکل ۴-۵: تجزیه نامنفی $\boldsymbol{V \approx W * H}$ جهت تولید بردارهای پایه در قالب ماتریس $(\boldsymbol{W})$ که مرتبط با یک نمونه آموزشی ورودی است. ماتریس $\boldsymbol{W}$ در پایان عملیات به عنوان یک اتم به دیکشنری $\boldsymbol{Dict}$ اضافه می‌شود $\boldsymbol{Dict = [Dict\ W]}$

۴-۶-۱- آموزش الگوها

در پایان مرحله آموزش، دیکشنری W تولید می‌شود که اتم‌های مجزا برای وقایع صوتی مورد نظر را در ستون‌های خود دارد. این دیکشنری بصورت آفلاین توسط یک تجزیه نامنفی بر روی ماتریس مشاهدات تولید می‌شود. تعیین مرتبه r در تجزیه مذکور اختیاری است و مقدار پیش فرض در اینجا $r=20$ انتخاب شده است. قبل از اجرای فاز آموزش لازم است نمونه‌های مجزایی برای هر یک از

وقایع صوتی آماده گردد که از روی آن‌ها اتم‌های مورد نظر تولید شده و آموزش داده شوند. در اینجا برای هر واقعه صوتی از ۲۰ نمونه آموزشی استفاده شده است. استخراج وصله‌های طیفی (اتم‌ها) معمولا از باندهای فرکانسی مل صورت می‌پذیرد. یک روش، انتخاب تمام وصله‌های فرکانسی بصورت تصادفی از مکآن‌های زمانی مختلف در اسپکتروم داده‌های آموزشی انتخاب است تا ویژگی‌های استخراجی نماینده مناسبی برای صدای مذکور باشد. اما روش تجزیه نامنفی از این نظر که اتم‌های مهم‌تر را استخراج می‌کند گزینش مناسب‌تری بنظر می‌رسد. در نهایت بردارهای ویژگی از اتم‌های استخراج شده از نمونه‌های آموزشی ایجاد می‌شوند. در اینجا از یک مرحله دوم پیش‌پردازش جهت نرمال‌سازی کنتراست در اتم‌ها استفاده شده است. این کار جهت حذف اختلاف محدوده دینامیک ایجاد شده توسط چگالی مطلق در طول اتم انجام می‌شود. نرمال‌سازی کنتراست از طریق رابطه (۴-۲۵) انجام می‌شود.

$$\tilde{x} = \frac{x - mean(x)}{\sqrt{\mathrm{var}(x) + \varepsilon_c}} \qquad (۴-۲۵)$$

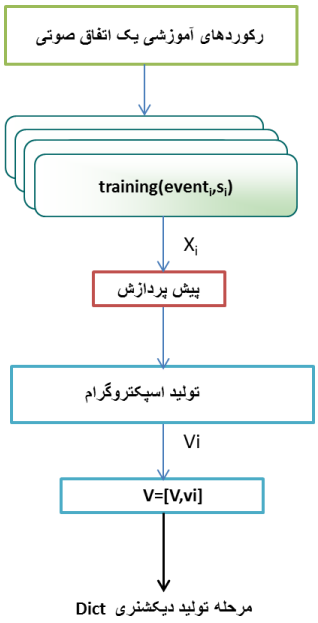
که در آن x بردار داده مربوط به اتم‌ها است، mean(.) و var(.) اپراتورهای میانگین و واریانس هستند، ϵ_c یک پارامتر تنظیم کننده است که تقویت نشدن ساختار نویزی را تضمین می‌کند و مقدار آن در اینجا ۱ انتخاب شد. چنانچه میزان همبستگی در داده‌های آموزشی هر یک از نمونه‌های طیفی بالا باشد احتمال یادگیری این ساختار همبستگی زیاد است لذا قدرت نمایندگی نمونه‌های آموزش داده شده کم خواهد بود. جهت جلوگیری از این نقص و بالا بردن گستردگی قدرت نمایش نمونه‌ها عمل سفید کردن داده‌ها (whitening) در مرحله بعدی انجام می‌گیرد تا ساختارهای آماری بالاتر داده‌ها در آموزش شرکت داده شوند. سفید کردن داده‌ها از روش‌های مختلف از جمله استفاده از PCA قابل انجام است. وقتی از PCA استفاده شود بدلیل انتقال و نگاشت صورت گرفته داده‌های جدید سفید شده در فضای تصویرسازی شده توسط PCA قابل دستیابی است. لذا برای هر چه شبیه شدن داده‌های سفید شده به همان داده‌ها در فضای اولیه یک آنالیز فاز-صفر (ZCA) Zero-phase component analysis (که گاهی اوقات تبدیل ماهالانوبیس Mahalanobis transformation هم

گفته می‌شود) روی داده‌ها مطابق رابطـه (۴-۲۶) انجـام می‌گیـرد [۱۰۱، ۱۰۲]. در رابطـه (۴-۲۶) در واقع مولفه‌های اصلی داده جدید بر ریشه دوم مقادیر ویژه تقسیم می‌شوند.

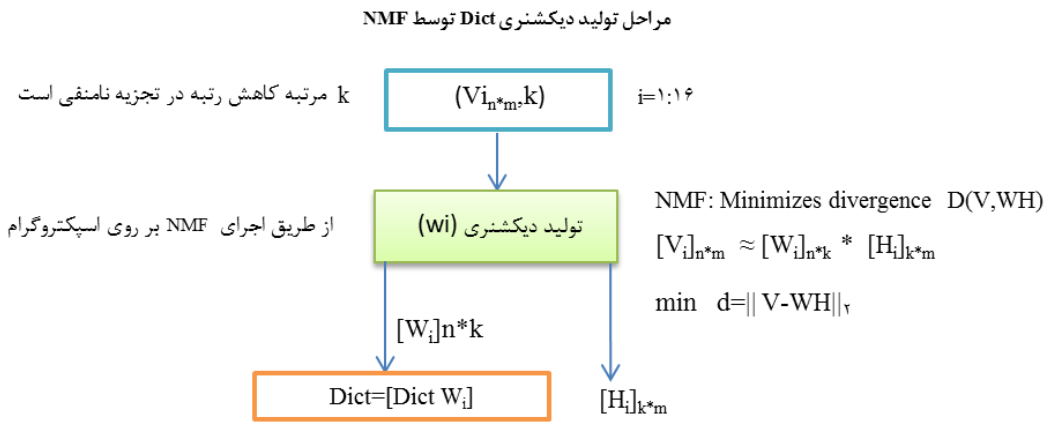
$$\hat{x} = V(D+\varepsilon_W I)^{-1/2} V^T x_C \qquad (۴-۲۶)$$

در رابطه (۴-۲۶)، V و D بترتیب بردار ویژه و ماتریس مقادیر ویژه مرتبط بـا مـاتریس کوواریـانس $\mathrm{cov}(\tilde{x}_C)$ است، $\tilde{x}_C$ برداری مرکز صفر شده برای بردار $\tilde{x}$ است، ε_W پـارامتر تنظیمـی در پـردازش سفید کردن است که به منظور کم کردن تاثیر بردارهای ویژه دارای مقادیر کوچک اسـتفاده می‌شـود. در اینجا مقدار این پارامتر ۰/۱ انتخاب شد. از اتم‌های طیفی سفید شده برای آموزش مجموعه نمونه‌ها استفاده می‌شود. الگوریتم‌های متفاوتی برای آموزش نمونه‌ها وجود دارد از جمله K-SVD در [۱۰۳] و الگوریتم تصویرسازی شـیب (projected gradient) در [۱۰۴] و همچنـین NMF نیـز اسـتفاده شـده است. مدل آموزش K-SVD تعمیم یافته روش خوشه بندی K-means است که متناوبا بین کدگذاری تنک داده‌های ورودی بر اساس دیکشنری موجود و بهنگام‌سـازی اتم‌هـا در دیکشـنری تعامـل برقـرار می‌کند تا اینکه داده‌هـا بهتـر نمـایش داده شـوند. ایـن کـار تـا حـدودی شـبیه بکـارگیری الگـوریتم خوشه‌بندی K-means در کوانتیزه کردن بردارهـا vector quantization (VQ) اسـت. روشـی کـه در اینجا استفاده شده است آموزش بر اساس NMF است. در شکل ۴-۶ مراحل مقدماتی برای هـر نمونـه آموزشی قبل از تولید ماتریس پایه‌های W نشان داده شده است.

در شکل ۴-۶ برای هر نمونه صوتی i ابتدا آنرا به فریم‌های کوتـاه مـدت تبـدیل نمـوده و پـس از تبدیل در حوزه فرکانس انرژی طیف را بدست آورده و از فریم‌های تبـدیل یافتـه مـاتریس $V^{(i)}$ ایجـاد می‌شود به طوری که ستون V_j^i نمایش فرکانسی فریم زمانی j ام اسـت. سـپس تجزیـه نـامنفی روی $V^{(i)}$ با مرتبه $r=۲۰$ مطابق مراحل شکل ۴-۷ انجام می‌شود تا عامل‌های ایـن مـاتریس بدسـت آیـد. برای بهنگام‌سازی مدل ضربی از رابطه (۴-۸) استفاده نموده و ماتریس ضرایب نیز نرمالیزه می‌شـوند و در انتها الگوی واقعه صوتی i ام در بردار ستونی $W^{(i)}$ قرار می‌گیرد.



شکل ۴-۶: مراحل مقدماتی برای هر نمونه آموزشی قبل از ارسال جهت تولید ماتریس پایه‌های (W) نشان داده شده است. پس از تولید v که نمایش اسپکتروگرام ورودی است جهت تجزیه نامنفی $V \approx W * H$ مراحل ذکر شده در شکل ۴-۷ اجرا می‌گردد.

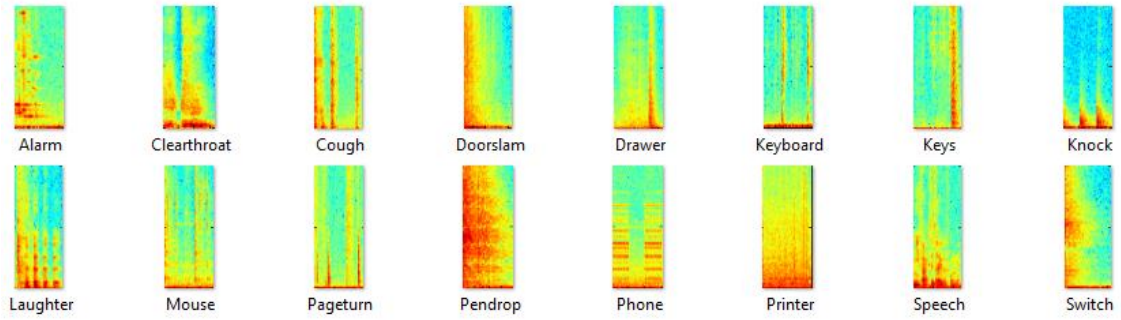


شکل ۴-۷: مراحل لازم جهت تولید ماتریس پایه‌های W برای هر نمونه آموزشی توسط تجزیه نامنفی $W * H \approx V$ نشان داده شده است.

۴-۶-۲- تجزیه سیگنال ورودی آزمایش

دیکشنری W از کنار هم قرار گرفتن بردارهای ستونی $W^{(i)}$ تولید می‌شود. برای تجزیه یک جریان صوتی ورودی به عنوان آزمایش، ابتدا جهت حذف نویز از فیلتر میان گذر باتروورث مرتبه ۷ با پهنای فرکانس ۵۰۰ تا ۱۲۰۰۰ هرتز و سپس عملیات سفید کردن استفاده شده است. پس از تبدیل سیگنال

ورودی به اسپکتروگرام v_j این نمایش بر دیکشنری W تصویرسازی می‌شود. در این فاز که کشف و شناسایی وقایع صوتی است از تجزیه نامنفی برای بدست آوردن $W^{(i)}$ در رابطه $v_j \approx W^{(i)} h_j$ استفاده می‌شود. پس از تولید دیکشنری $Dict$ بردارهای پایه نماینده هر واقعه صوتی به صورت یک تکه یا وصله[1] پیوسته مانند شکل ۴-۸ دیده می‌شوند.



شکل ۴-۸: وصله‌های استخراج‌شده از نمونه‌های آموزشی وقایع صوتی ۱۶ گانه. هر واقعه در دیکشنری $\boldsymbol{Dict}$ دارای وصله مربوط به خود است.

در فاز آزمایش از بردارهای بدست آمده h_j به عنوان فعال‌کننده‌های وقایع مختلف صوتی که در صحنه شنوایی هستند استفاده می‌شود. تا اینجا فعالیت وقایع صوتی مختلف در سطح فریم تعیین می‌شود. تعدادی پردازش نهایی جهت استخراج اطلاعات بیش‌تر درباره وجود هر یک از صداها در سطح سگمنت انجام می‌شود. این پردازش‌ها شامل آستانه گذاری روی ضرایب فعال کننده، شناسایی نقاط خیزش، مدل‌سازی زمانی و هموارسازی می‌باشد.

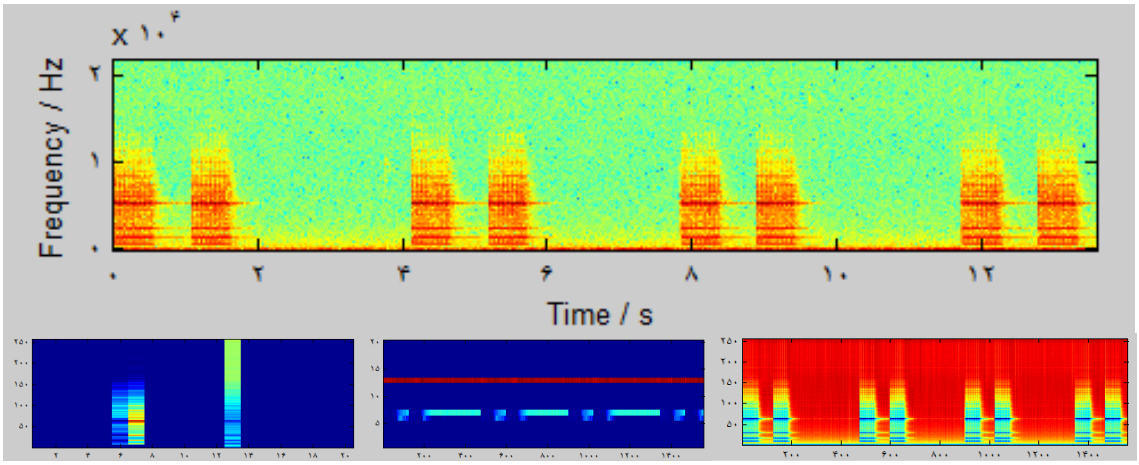
۳-۶-۴- شناسایی وقایع صوتی و بررسی نتایج

در اینجا جهت ارزیابی کمی سیستم پیشنهادی قابلیت‌های آن در موضوع شناسایی صداها بررسی می‌شود. در ارزیابی از متریک‌های استاندارد استفاده شده است. این متریک‌ها $F-measure\,(F)$ $Accuracy\,(A)$, $Recall(R)$, $Precision\,(P)$, هستند که بر اساس سگمنت‌های شناسایی شده محاسبه می‌شوند. همچنین نتایج شناسایی را بر اساس هر یک از صداهای ۱۶ گانه بدست می‌آوریم.

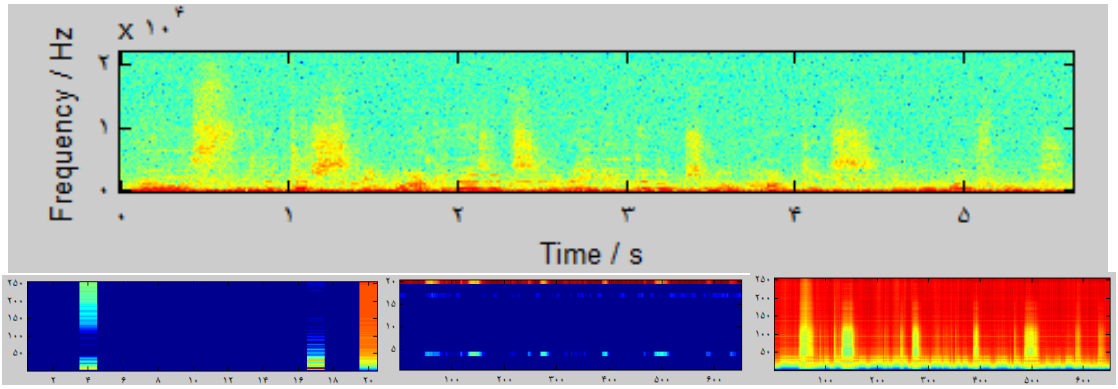
[1]- patch

جهت شناسایی وقایع صوتی ابتدا مطابق فاز آموزش در شکل ۴-۴، به صورت آفلاین یک دیکشنری توسط تجزیه نامنفی همراه با محدودیت تنک از نمونه‌های وقایع صوتی ساخته می‌شود و برای هر واقعه صوتی تعداد $r=20$ بردار پایه در دیکشنری در نظر گرفته می‌شود. تعداد بردارها اختیاری است اما با توجه به نوع صداها در آزمایش‌ها مقدار ۲۰ نتایج بهتری داشت. از این دیکشنری به صورت ثابت در فاز آزمایش مطابق شکل ۴-۴ استفاده می‌شود. نگاشت سگمنت بر دیکشنری از طریق فعال کننده انجام می‌شود. ابتدا عمل نرمال‌سازی ردیف‌های ماتریس فعال کننده که کلاس واقعه صوتی را مشخص می‌کنند انجام می‌شود آنگاه عمل باینری کردن ماتریس به ۰ یا ۱ انجام می‌گیرد تا ضرایب نزدیک به صفر حذف شده و ضرایب قوی‌تر به ۱ تبدیل شوند. مقدار آستانه باینری کردن از طریق تجربی $1/n$ تعیین شد که n تعداد فریم‌های سگمنت مورد پردازش می‌باشد. چنانچه حداقل ۲۰ فریم متوالی ۱ شده باشند آن‌ها را به عنوان بخش صدا و در غیر این صورت به عنوان زمینه در نظر می‌گیرد. در شکل ۴-۹ نمودارهای مربوط به تجزیه نمونه صدای تلفن آورده شده است. در قسمت بالای شکل سیگنال اصلی قبل از تجزیه دیده می‌شود. قسمت پایین شکل تجزیه $W*H \approx V$ می‌باشد که در آن بردارهای پایه و بردارهای ضرایب تجزیه مشخص هستند. اثر نویز در سیگنال ورودی در بخش فعال کننده (ماتریس H) به دو صورت دیده می‌شود. نویز با فرکانس‌های بالا در قالب یک خط قرمز پیوسته و نویز در فرکانس‌های پایین در رنگ فیروزه‌ای پیوسته دیده می‌شود. به طور مشابه در شکل ۴-۱۰ تجزیه سیگنال نمونه صحبت دیده می‌شود.

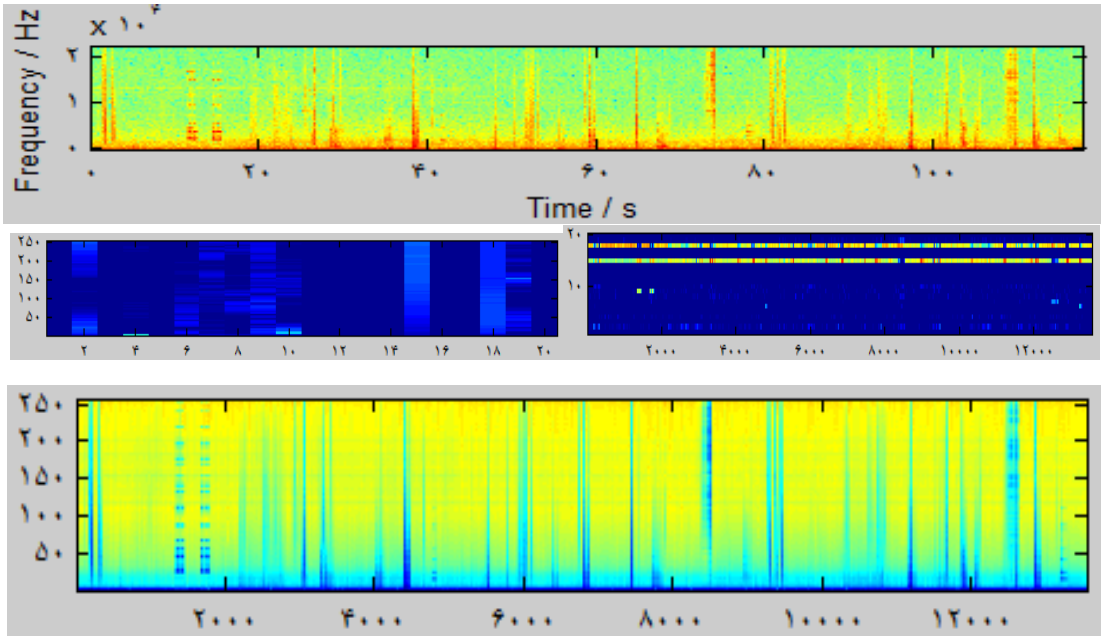
در شکل ۴-۱۱ تجزیه یک رکورد ورودی دارای ۳۷ واقعه صوتی مختلف توسط بردارهای پایه و ضرایب مربوطه نمایش داده شده است. در شکل ۴-۱۲ تجزیه نامنفی رکورد ورودی دیگری با ۳۵ واقعه صوتی مختلف با محدودیت تنک توسط بردارهای پایه و ضرایب مربوطه نمایش داده شده است.



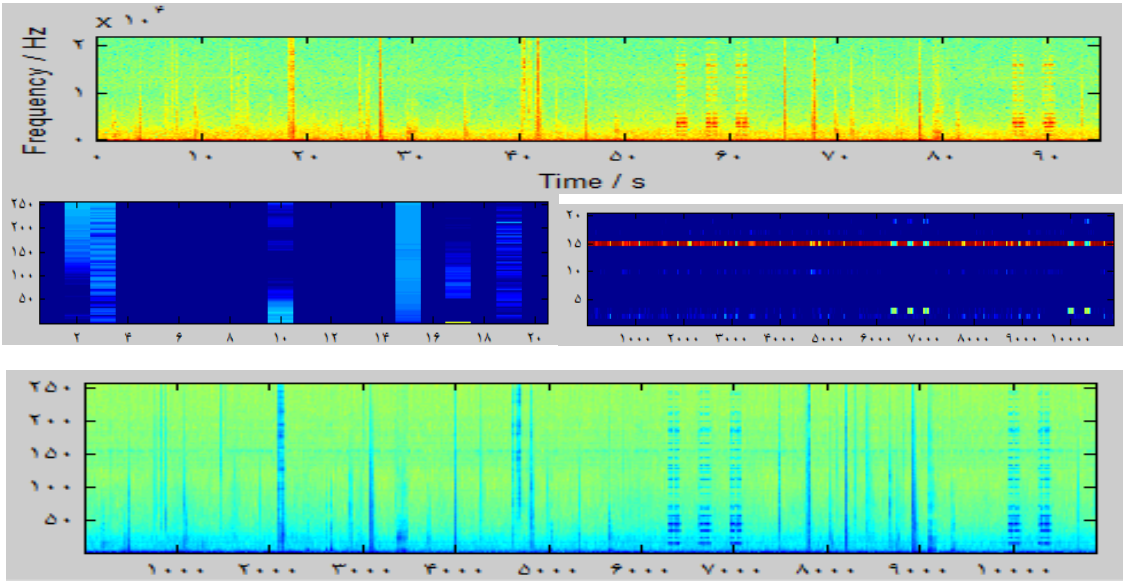
شکل ۴-۹: تجزیه نامنفی صدای تلفن. سیگنال اصلی در بالای شکل و تجزیه به صورت $\boldsymbol{W} * \boldsymbol{H} \approx \boldsymbol{V}$ در پایین شکل آمده است. تنک بودن ماتریس‌های $\boldsymbol{W}, \boldsymbol{H}$ در تصویر مشخص است.



شکل ۴-۱۰: تجزیه نامنفی صدای صحبت. سیگنال اصلی در بالای شکل و تجزیه به صورت W*H≈V در پایین شکل آمده است.



شکل ۴-۱۱: تجزیه رکورد ورودی با ۳۷ واقعه صوتی مختلف. سیگنال اصلی در بالای شکل و نتیجه تجزیه در پایین شکل دیده می‌شود. دیکشنری استخراجی و فعال‌کننده‌های آن در بخش میانی شکل آمده است. تنک بودن ماتریس‌های W, H تا حد زیادی دیده می‌شود. اثر نویز با قدرت‌های مختلف در کل طول سیگنال قابل مشاهده است. خط رنگی پیوسته با شدت‌های مختلف در فعال کننده بیانگر نویز موجود در سیگنال است.



شکل ۴-۱۲: تجزیه نامنفی یک رکورد ورودی با ۳۵ واقعه صوتی مختلف در شکل دیده می‌شود. سیگنال اصلی در بالای شکل آمده است و نتیجه تجزیه در پایین شکل دیده می‌شود. دیکشنری استخراجی و فعال‌کننده‌های آن در بخش میانی شکل آمده است. تنک بودن ماتریس‌های W, H تا حد زیادی دیده می‌شود. اثر نویز با قدرت‌های مختلف در کل طول سیگنال قابل مشاهده است. خط قرمز با شدت‌های مختلف در فعال کننده بیانگر نویز موجود در سیگنال است.

بخش میانی شکل ۴-۱۱ فعالیت هر یک از الگوها و تنکی راه حل برای تجزیه تنک در رکورد صوتی اول را نشان می‌دهد. میزان شناسایی وقایع صوتی در رکورد ورودی $F=۴۶/۷\%$ است. همانطور که در شکل ۴-۱۱ نیز قابل مشاهده است سیستم تمایل استفاده از الگوهای زیاد را دارد. صداهای برجسته و صداهای نویزی موجود در زمینه سبب فعالیت اشتباهی بعضی از الگوها از جمله phone، printer و alarm می‌گردد و کلا نویز زمینه، سطح فعالیت الگوهای ۱۵ و ۱۸ را بالا نگه داشته است. این خطاها به طور آشکاری در شکل بصورت نقاط قرمز پیوسته دیده می‌شود. تجزیه نامنفی با انحراف بتا با مقدار $\beta=۰$ (انحراف بتای ایتاکورا سایتو) و $\beta=۰/۵$ و $\beta=۱$ انحراف بتای کولبک لیبلر) نیز انجام شد که نتایج در مورد انحراف بتا با مقدار $\beta=۱$ بهتر بود. شاید دلیل این است که به ازای بتای صفر وزن نسبی یکسانی به ضرایب می‌دهد و این سبب تغییرات یکسانی در ضرایب کوچک و بزرگ برای تنظیمات تکرار بعدی می‌شود. معیار ارزیابی برای انحراف بتا $F=۴۴/۰\%$ بدست آمد. با استفاده از تعیین حداقل میزان تنکی تا حد کمی خطاها تضعیف می‌شود اما همزمان تعداد وقایع مفقودی افزایش می‌یابند. در چنین حالتی اگر میزان تنکی با تغییرات دینامیک سیگنال هماهنگ گردد می‌توان سیستمی مقاوم‌تر ایجاد نمود. در شکل ۴-۱۲ فعالیت هر یک از الگوها و تنکی راه حل برای تجزیه تنک در دومین رکورد صوتی نشان داده شده است. در اینجا میزان تنکی دیکشنری W (سمت چپ بخش میانی شکل) و H ضرایب فعال کننده بیش از رکورد قبل در شکل ۴-۱۱ است. یکی از علت‌ها تنوع بیش‌تر نوع صداها در رکورد آزمایش شده در شکل ۴-۱۱ می‌باشد. اثرات نویز زمینه بصورت پایدار در شکل ۴-۱۲ نیز دیده می‌شود. ضرایب پیوسته در ماتریس H بیانگر حضور نویز متمرکز در یکی از الگوها است که در اینجا الگوی ۱۵ در دیکشنری است. همچنین قدرت نویز زمینه نسبت به بعضی از صداها از جمله keyboard، mouse و keys قوی‌تر است لذا تشخیص چنین صداهایی با مشکل روبرو شده‌است. میزان وقایع شناسایی شده در این رکورد $F=۴۹/۵\%$ است. تجزیه نامنفی با انحراف بتا نیز انجام شد که معیار ارزیابی برابر $F=۴۴/۷\%$ بدست آمد. نتایج حاصل از پردازش رکوردهای ورودی با تعداد وقایع مختلف صوتی برحسب شناسایی

فریم صوتی در جدول ۴-۱ آمده‌است. برحسب نوع صداها و نویز زمینه در هر یک از رکوردها میزان شناسایی متفاوت است.

جدول ۴-۱: میزان $\boldsymbol{F-measure(F)}$ برحسب شناسایی فریم صوتی در رکوردهای آزمایش حاوی $\boldsymbol{n}$ واقعه با دو نوع الگوریتم تجزیه نامنفی

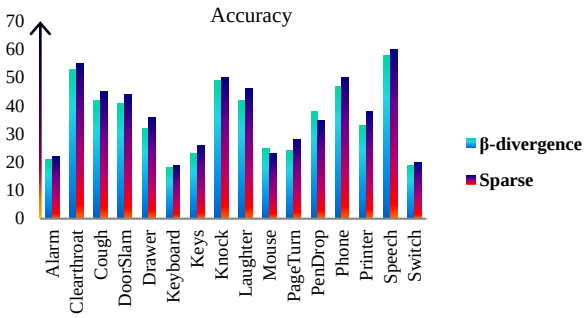
الگوریتم	n=۳۵	n=۳۷	n=۴۱	n=۴۷	میانگین
β-divergence	۵۲/۳	۵۰/۸	۵۲/۳	۵۱/۶	۵۱/۸
sparse	۵۸/۶	۵۴/۷	۵۶/۶	۵۶/۳	۵۶/۵

یک ارزیابی دیگر میزان شناسایی وقایع برحسب سگمنت‌های استخراج شده است. در این حالت هر سگمنت حاوی یک واقعه صوتی است و نتایج در جدول ۴-۲ آورده شده است. میانگین شناسایی وقایع در الگوریتم انحراف بتا $F=۴۵/۰\ \%$ و برای الگوریتم تنک $F=۴۹/۲\ \%$ بدست آمد. در اینجا نتایج نسبت به وضعیت فریمی که در جدول ۴-۱ آمده است اندکی کاهش نشان می‌دهد که ناشی از حذف فریم‌هایی است که مجموعا در محدوده یک سگمنت اصلی نبوده‌اند لذا به عنوان واقعه صوتی شناسایی شده محسوب نشده‌اند.

جدول ۴-۲: میزان $\boldsymbol{F-measure(F)}$ برحسب شناسایی واقعه صوتی در رکوردهای آزمایش حاوی$\boldsymbol{n}$ واقعه با دو نوع الگوریتم تجزیه نامنفی

الگوریتم	n=۳۵	n=۳۷	n=۴۱	n=۴۷	میانگین
β-divergence	۴۴/۷	۴۴/۰	۴۶/۳	۴۵/۱	۴۵/۰
sparse	۴۹/۹	۴۶/۷	۴۹/۸	۵۰/۴	۴۹/۲

میزان درستی شناسایی هر یک از وقایع صوتی به طور مجزا در دو الگوریتم تجزیه نامنفی در شکل ۴-۱۳ نشان داده شده است. درصد شناسایی برای صداهای ریز و کوتاه به میزان قابل توجهی کم‌تر از صداهای قوی و طولانی است. یکی از عوامل پایین بودن نرخ شناسایی درست مشابهت بعضی از صداها با یکدیگر است که سبب افزایش خطای شناسایی می‌شود. عامل مهم دیگری نیز وجود دارد که تاثیر نویز زمینه می‌باشد.



شکل ۴-۱۳: میزان درستی شناسایی وقایع صوتی به طور مجزا در دو الگوریتم تجزیه نامنفی نشان داده شده است. نرخ درستی برای صداهای ریز، صداهای مشابه و تاثیر پذیر از نویز زمینه کم‌تر است.

مقایسه معیار F روش پیشنهادی با سه رویکرد دیگر در جدول ۴-۳ آمده است. روش‌هـای ارزیـابی frame-based(FB) و event-based(EB) [۱۰۵، ۱۰۶] هستند.

جدول ۴-۳: مقایسه عملکرد الگوریتم با سایر رویکردها

روش ارزیابی/رویکرد		F (FB)	F (EB)
[۳۵]		۱۰/۷	۷/۴
[۸۶]		۴۹/۷	۳۱/۲
روش پیشنهادی	انحراف بتا	۵۱/۸	۴۵/۰
	تنک	۵۶/۵	۴۹/۲

فصل

۵- جمع‌بندی و پیشنهادها

در این رساله با هدف ارائه روش‌هایی جهت کشف و رده‌بندی وقایع صوتی محیطی پژوهش گسترده‌ای انجام شده است. روش‌ها و مدل‌های موجود بررسی شده و دو رویکرد جداگانه برای تحقق این هدف دنبال گردید. در رویکرد اول یک روش جدید مدل‌سازی تطبیق‌پذیر با دو پارامتر جهت کشف و رده‌بندی وقایع صوتی در سیگنال‌های محیطی ارائه گردید. این مدل از دو پارامتر مهم و کلیدی α و β استفاده می‌کند که تا حد زیادی سبب بهبود عملکرد اولیه سیستم می‌گردد. این میزان بهبودی یکی به دلیل وجود مکانیزم تنظیم عملکرد در زمان قطعه‌بندی و همچنین تکیه بر میزان هم‌افزایی اطلاعات در زمان قطعه‌بندی و اصلاح محدوده‌های قطعه‌های شناسایی‌شده می‌باشد. همچنین انعطاف‌پذیری و قابلیت تنظیم سیستم بین دقت بالاتر یا فراخوانی بالاتر توسط پارامترها قابل اعمال است. از مزایای مدل پیشنهادی می‌توان به استفاده از دو زیرسیستم قطعه‌بندی و کشف وقایع با امکان ترکیب نتایج اشاره نمود که سبب استفاده هر چه بهتر از ویژگی‌ها شده و امکان بالا رفتن قابلیت جداسازی ویژگی‌ها و کشف بیش‌تر وقایع را میسر می‌سازد. بعلاوه طی آزمایش‌ها مشخص گردید که تفکیک ویژگی‌ها به دو گروه می‌تواند شناسایی صداهای خاص توسط ویژگی خاص را عملی سازد. نتایج بدست آمده تاثیر قابل ملاحظه پارامترها در مقایسه با حالت اولیه مدل را بخوبی نشان می‌دهد. مقدار دقت حداکثر ۶۰/۶ درصدی (میانگین در جدول ۳-۱۰) در مقایسه با مدل‌های موجود بهبود ۱۰٪/۸۷ را نشان می‌دهد که کارایی مدل را تایید می‌کند. ما در آزمایش‌ها مقادیر پارامترها را برای حداکثر نمودن معیارهای دقت، فراخوانی و کارایی سیستم بدست آورده و نتایج را در جداول ارائه نمودیم و مشخص شد مدل ارائه شده می‌تواند با تعیین مقادیر پارامترها معیارهای مورد نظر را تا حد امکان بهینه و حداکثر نماید. یکی دیگر از مزایای مدل ترکیب نتایج و استفاده از روش‌های یکپارچه‌سازی و اصلاح نهایی محدوده وقایع اکتشافی است که تاثیر مهمی در میزان دقت نهایی داشته است. خاصیت پارامتریک بودن امکان افزایش میزان دقت به ازای کم شدن مقداری از میزان فراخوانی را می‌دهد و معکوس این نیز امکان پذیر است. نتایج این روش از لحاظ معیار ارزیابی F1 در حالت شناسایی درست فریم ۸۰/۶ درصد و در حالت شناسایی درست واقعه صوتی ۷۲/۸ درصد

می‌باشد. این نتایج در جدول ۳-۱۱ آورده شده است. مشروح مطالب مرتبط با این رویکرد و مدل ارائه شده نیز در فصل ۳ آمده است.

در رویکرد دوم جهت کشف و رده‌بندی وقایع صوتی محیطی، از مدل‌های جداسازی سیگنال‌های مخلوط که چند سالی است تحت عنوان جداسازی مولفه‌های مستقل یا جداسازی منابع کور مطرح شده است و الگوریتم‌های خاص آن الهام گرفته شده است. در اینجا با استفاده از مفاهیمی همچون نمایش تنک و تجزیه نامنفی ماتریس‌های مشاهده به حل موضوع پرداخته شده است. جهت دستیابی به این هدف روشی جهت نگاشت سگمنت‌های سیگنال مشاهده بر یک دیکشنری متشکل از اتم‌های پایه و در یک نمایش تنک ارائه گردید. در هنگام مدل‌سازی این روش، یک سیستم برحسب تکنیک تجزیه نامنفی ارائه شد و استفاده از محدودیت تنک جهت نگاشت سگمنت بر اتم‌های دیکشنری در تجزیه نامنفی سیگنال ورودی پیشنهاد گردید. الگوریتم با تکرار مشخصی به همگرایی رسیده و شامل کنترل بر میزان تنکی و ایجاد مصالحه فرکانسی در فرآیند تجزیه می‌باشد. الگوریتم پیشنهادی در شناسایی صداهای متعدد در محیط اداری بکار گرفته شد و مزیت‌های استفاده از چنین کنترل‌هایی در بهبود شناسایی صداهای محیطی شرح داده شد.

از نتایج بدست آمده مشخص گردید زمانی‌که با نویز زمینه و وقایع صوتی برجسته نامطلوب همراه با تداخل بالای محتوای فرکانسی سر و کار داریم کنترل تنکی سبب رشد مقاومت سیستم در کار شناسایی صداهای محیطی می‌شود. همچنین مشخص شد اگر سگمنت را ملاک شناسایی صدا قرار دهیم نرخ تشخیص درست وقایع افزایش می‌یابد. از سوی دیگر نشان داده شد که کنترل بر مصالحه فرکانسی حین تجزیه در کار شناسایی صداهای محیطی تاثیر مثبت دارد. به طوری که بخش‌های دارای فرکانس بالا با انرژی پایین برای جداسازی بین وقایع صوتی مختلف مهم تلقی می‌شوند. از طرفی چون به طور معمول در روش‌های تجزیه نامنفی با وجود ساکن نبودن سیگنال‌ها الگوها ثابت هستند مدل تنک پیشنهادی صداهایی که دارای دینامیک و تغییر پذیری زمانی زیاد هستند را بهتر مدل‌سازی می‌نماید. در این مورد علاوه بر استفاده از بازه‌های زمانی کوتاه در نمایش اولیه سیگنال،

ترکیب طیف نمونه‌های مختلف یک صدا نیز سبب نمایش بهتر تغییرپذیری شده است. معیار F1 در این روش طبق جدول ۴-۳ با اعمال محدودیت تنک بر تجزیه نامنفی و نگاشت سگمنت‌های وقایع بر دیکشنری تنک، برای حالت شناسایی فریم و واقعه صوتی به ترتیب به ۵۶/۵ درصد و ۴۹/۲ درصد رسیده‌است. مشروح مطالب مرتبط با این رویکرد و مدل ارائه شده در فصل ۴ آورده شده است.

۱-۵- پیشنهادها جهت ادامه تحقیقات

یکی از معضلات اصلی معمولا وجود نویز بالا و متغییر در رکوردهای صوتی است که لازم است روش‌های بهتری برای حذف نویز و مقاوم نمودن الگوریتم در برابر نویز ارائه شود. در مواردی نیز مشاهده شد که بعضی از صداهای محیطی ویژگی‌های زمانی متمایز از نویز زمینه ندارند لذا معرفی ویژگی‌های زمانی که صداهای نویز-مانند را دقیق‌تر کشف نماید پیشنهاد می‌شود. می‌توان در پژوهش‌های آینده جهت مقابله با نویز در داده‌ها نویز محیط را به صورت یک سطح مجزا مدل‌سازی نمود. علاوه بر اینها تحقیق جدید مبتنی بر استفاده توام از ویژگی‌های حوزه فرکانس و ادراک در یک زیرسیستم و ویژگی‌های حوزه زمان در زیرسیستم دیگر پیشنهاد می‌شود.

در رویکرد دوم جهت نمایش اولیه سیگنال از اسپکتروگرام استفاده شده است. در حالی که در صداهای محیطی استفاده از فرکانس غیر خطی همچون تبدیل Q ثابت[22] ممکن است سبب بهبود کارایی سیستم شود که لازم است این موضوع مورد تحقیق قرار گیرد.

افزایش مقاومت در برابر نویز و عمومیت دهی سیستم همواره یکی از اهداف سیستم‌های شناسایی است. پیشنهاد می‌گردد ضرایب فعال کننده در طول فرآیند آموزش الگوهای هر صدا کدگذاری و نگهداری شوند تا در فرآیند تجزیه فاز آزمایش بتوان از آن‌ها استفاده نمود. همچنین جهت کم کردن تاثیر نویز پیشنهاد می‌گردد از بهنگام‌سازی غیر ثابت برای بردارهای پایه دیکشنری استفاده گردد تا بدین وسیله بتوان از تقویت بردارهای نویزی جلوگیری کرد.

[22] - constant-Q transform (CQT)

پیشنهاد دیگر در نظر گرفتن وضعیت زمانی الگوها به طـور مسـتقیم در الگوهـای تجزیـه نـامنفی است. طوری که بتوان ترکیب تجزیه نامنفی با یک نمایش حالت از صداها را در نظر گرفـت. عـلاوه بـر مدل‌سازی زمانی وقایع، فاز یادگیری الگوها نیز قابل بهبود است. یـک امتیازخـاص در تجزیـه نـامنفی فرمول‌بندی فاز آموزش در طرح‌های بسط یافته مختلفی است که می‌توان از آن‌ها بـرای آمـوزش یـک الگو یا بیش‌تر برای هر نوع صدا استفاده نمود. این کار با تعیین یک رتبه r برای تجزیه عملی می‌شـود لذا می‌توان در انتخاب طرح مناسب برای آموزش الگوها مطالعات بیش‌تری انجام داد.

مراجع

[1] C. Pham, and P. Cousin, “Streaming the sound of smart cities: Experimentations on the smartsantander test-bed,” *In Green Computing and Communications (GreenCom),* 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, pp. 611–618, 2013.

[2] J. Kotus, K. Lopatka, A. Czy˙zewski, and G. Bogdanis, “Audio-visual surveillance system for application in bank operating room,” *In Multimedia Communications, Services and Security, Springer,* pp. 107–120, 2013.

[3] T. W. Chua, K. Leman, and F. Gao, “Hierarchical audio-visual surveillance for passenger elevators,” *In MultiMedia Modeling Springer,* pp. 44–55, 2014.

[4] Q. Pham, A. Lapeyronnie, C. Baudry, L. Lucat, P. Sayd, S. Ambellouis, D. Sodoyer, A. Flancquart, A. C Barcelo, F. Heer, F. Ganansia, and V. Delcourt , “Audio-video surveillance system for public transportation,” *In Image Processing Theory Tools and Applications (IPTA),* 2010 2nd International Conference on, pp. 47–53, 2010.

[5] W. Zajdel, J. Krijnders, T. Andringa, and D. Gavrila, “Cassandra: audio-video sensor fusion for aggression detection,” In Advanced Video and Signal Based Surveillance, AVSS 2007. IEEE Conference on, pages 200–205, 2007.

[6] Y. Chung, S. Oh, J. Lee, D. Park, H. Chang, and S. Kim, “Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems Sensors,” 13(10):12929–12942, 2013.

[7] J. Kotus, K. Lopatka, and A. Czyzewski, “Detection and localization of selected acoustic events in acoustic field for smart surveillance applications,” Multimedia Tools and Applications, 68(1):5–21, 2014.

[8] S. Duan, J. Zhang, P. Roe, and M. Towsey, “A survey of tagging techniques for music, speech and environmental sound,” Artificial Intelligence Review, pp. 1–25, 2012.

[9] Y.Peng, C.Lin, M.Sun, K.Tsai, “Healthcare audio event classification using HMM and HHMM,” in IEEE International Conference on Multimedia and Expo (ICME), IEEE Computer Society, New York, NY, USA, 2009.

[10] M. V. Ghiurcau, C. Rusu, J. Astola, R. Bilcu, “Towards an application for detecting intruders in wildlife regions,” in: Proceedings of the EUSIPCO, 2010.

[11] D. Gerhard, “Audio Signal Classification: History and Current Techniques”, Technical Report TR-CS 2003-07, 2003.

[12] 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.

[13] N. Scaringella, G. Zoia, and D. Mlynek, “Automatic genre classification of music content: a survey,” IEEE Signal Process. Mag., 2006.

[14] 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, Salerno, pp. 145-151, 2014.

[15] 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, pp. 81-86, 2013.

[16] 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”. ICASSP, page 4293-4296, 2012.

[17] R. Maher, “Acoustical modeling of gunshots including directional information and reflections,” in Audio Engineering Society Convention 131, 2011.

[18] P. Huang, X. Zhuang, M. Hasegawa-Johnson, “Improving Acoustic Event Detection using Generalizable Visual Features and Multi-modality Modeling”. In AED MultiStr CHMM ICASSP2011, 2011.

[19] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, “Real-world acoustic event detection”. Pattern recognition Letters, vol. 31, pp. 1543–1551, 2010.

[٢٠] X. Zhou, X. Zhuang, M. Liu,H. Tang, M. Hasegawa-Johnson, and T. S. Huang, “HMM-based acoustic event detection with adaboost feature selection”. in Multimodal Technologies for Perception of Humans, Springer Berlin / Heidelberg vol. 4625, pp. 345–353, 2008.

[21] 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.

[22] A. Rabaoui, M. Davy, S. Rossignol, and Z. Ellouze, “Using one-class SVMs and wavelets for audio surveillance”. IEEE Transactions on Information Forensics and Security, vol. 3, no. 4, pp. 763-775, 2008.

[23] X. Zhuang, X. Zhou, T. S. Huang, and M. A. Hasegawa-Johnson, “Feature analysis and selection for acoustic event detection”. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, (ICASSP ’08), pp.17–20.

[24] P. Atrey, N. Maddage, M. Kankanhalli, “Audio Based Event Detection for Multimedia Surveillance”, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006.

[25] C. Rui, L. Lie, Z. Hong-Jiang, and C. Liun-Hong, “Highlight sound effects detection in audio stream”. in ICME03, pp. 37–40, 2003.

[26] M. Baillie, and J. M. Jose, “Audio-based event detection for sports video”. Lecture Notes in Computer Science, vol. 2728, pp. 61–65, 2003.

[27] L. Lu, H. Zhang, H. Jiang, “Content analysis for audio classification and segmentation”. in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 7, pp. 504-516, 2002.

[28] J. Pinquier, “Robust speech/ music classification in audio document”. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP), pp. 2005–2008, 2002.

[29] 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.

[30] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, “Sound event detection in multisource environments using source separation,” in Proc. of CHiME, Munich, Germany, pp. 36–40, 2011.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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 on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, PP. 50-52, Oct 2013.

[38] S. Pancoast, and M. Akbacak, “Bag-of-audio-words approach for multimedia event classification,” in Interspeech, 2012.

[39] A. Plinge, R. Grzeszick, and G. Fink, “A Bag-of-Features approach to acoustic event detection,” in IEEE Int. Conf. on Acoustics, Speech, and Signal Proces., May 2014.

[40] D. Mitrovic, M. Zeppelzauer, and C. Breiteneder, “Features for contentbased audio retrieval,” Advances in computers, 2010.

[41] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, Audio context recognition using audio event histograms, in 18th European Signal Processing Conference, Aalborg, Denmark, 2010.

[42] B. Byun, I. Kim, S. M. Siniscalchi, and C.-H. Lee, “Consumerlevel multimedia event detection through unsupervised audio signal modeling,” in Proc. INTERSPEECH, 2012.

[43] J. C. Wang, J.-F. Wang, K. W. He, and C.-S. Hsu, “Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio lowlevel descriptor,” in Neural Networks, IJCNN’06. International Joint Conference on. IEEE, 2006.

[44] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda, “Environment recognition using selected MPEG-7 audio features and Mel Frequency Cepstral Coefficients,” in Digital Telecommunications (ICDT), 2010 Fifth International Conference on. IEEE, 2010.

[45] E. Tsau, S.-H. Kim, and C.-C. J. Kuo, “Environmental sound recognition with CELP-based features,” in Signals, Circuits and Systems (ISSCS), 2011 10th International Symposium on. IEEE, 2011.

[46] A. Vahdat, K. Cannons, H. Hajimirsadeghi, G. Mori, et al., “TRECVID 2012 GENIE: Multimedia event detection and recounting,” in Proc. NIST TRECVID Workshop, 2012.

[47] M. Karbasi, S. Ahadi, and M. Bahmanian, “Environmental sound classification using spectral dynamic features,” in Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on. IEEE, 2011.

[48] X. Valero and F. Alıas, “Classification of audio scenes using narrowband autocorrelation features,” in Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE, 2012.

[49] M. Cowling and R. Sitte, “Comparison of techniques for environmental sound recognition,” Pattern Recognition Letters, 2003.

[50] X. Valero and F. Alıas, “Gammatone wavelet features for sound classification in surveillance applications,” in Signal Processing Conference (EUSIPCO), 2012.

[51] S. Chu, S. Narayanan, and C.-C. J. Kuo, “Environmental sound recognition with time–frequency audio features,” IEEE Trans. Audio, Speech, Lang. Process., 2009.

[52] P. Khunarsal, C. Lursinsap, and T. Raicharoen, “Very short time environmental sound classification based on spectrogram pattern matching,” 2013.

[53] A. G. A. Perera, S. Oh, M. Leotta, I. Kim, B. Byun, et al., “2011 Multimedia Event Detection: Late- Fusion Approaches to Combine Multiple Audio-Visual features,” in Proc. NIST TRECVID Workshop, 2011.

[54] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj, “Audio event detection from acoustic unit occurrence patterns,” in Proc. ICASSP. IEEE, 2012.

[55] B. Ghoraani and S. Krishnan, “Time-frequency matrix feature extraction and classification of environmental audio signals,” Audio, Speech, and Language Processing, IEEE Transactions, 2011.

[56] J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and Y. Sun, “Toward a practical implementation of exemplar-based noise robust ASR,” in Proc. EUSIPCO, 2011.

[57] F. Su, L. Yang, T. Lu, and G. Wang, “Environmental sound classification for scene recognition using local discriminant bases and HMM,” in Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011.

[58] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. S. Spanias, “Segmentation, indexing, and retrieval of environmental and natural sounds,” IEEE Transactions on Audio, Speech and Language Processing, 2010.

[59] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, “Acoustic event detection in real-life recordings,” in 18th European Signal Processing Conference, (Aalborg, Denmark), 2010.

[60] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa and P. Soundararajan, “The CLEAR 2006 Evaluation”, (springer), 2006.

[61] A. Waibe1, H. Steusloff, R. Stiefelhagen and the CHIL Project Consortium, “CHIL: Computers in the Human Interaction Loop”,(CHIL project).

[62] M. Slaney, “Semantic–Audio Retrieval”, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.

[63] E. Wold, T. Blum, D. Keislar, J. Wheaton, “Content-Based Classification, Search, and Retrieval of Audio”, in Proc. IEEE Multimedia, Vol. 3, no. 3, pp. 27-36, 1996.

[64] G. Guo, Z. Li, “Content-based Audio Classification and Retrieval using Support Vector Machines”, IEEE Transactions on Neural Networks, Vol. 14, pp. 209-215, January, 2003.

[65] J. Arias, J. Pinquier, R. André-Obrecht, “Evaluation of classification techniques for audio indexing”, in Proc. European Signal Processing Conference, 2005

[66] R. Cai, L. Lu A. Hanjalic, “Unsupervised content discovery in composite audio”, in Proc. ACM International Conference on Multimedia, pp. 628-637, 2005.

[67] L. Kennedy, D. Ellis, “Laughter Detection in Meetings”, NIST Meeting Recognition Workshop, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004.

[68] L. Kennedy, D. Ellis, “Pitch-based emphasis detection for characterization of meeting recordings”, IEEE workshop on Automatic Speech Recognition and Understanding, 2003.

[69] I. Boesnach, M. Hahn, J. Moldenhauer, T. Beth, U. Spetzger, “Analysis of Drill Sound in Spine Surgery”, in Proc. International Workshop on Medical Robotics, Navigation and Visualization, 2004.

[70] M. Vacher, D. Istrate, L. Besacier, E. Castelli, J. Serignat, “Smart audio sensor for telemedicine”, in Proc. Smart Object Conference, 2003.

[71] M. Vacher, D. Istrate and J. Serignat , “Sound Detection and Classification through Transient Models using Wavelet Coefficient Trees”, in Proc. European Signal Processing Conference, pp. 1171-1174, 2004.

[72] M. Vacher, D. Istrate, L. Besacier, J. Serignat, E. Castelli, “Life Sounds Extraction and Classification in Noisy Environment”, in Proc. IASTED International Conference on Signal & Image Processing, 2003.

[73] Z. Xiong, R. Radhakrishnan, A. Divakaran, T. Huang, “Audio Events Detection Based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework”, in Proc. IEEE International Conference on Multimedia and Expo, pp. 401-404, 2003.

[74] R. Radhakrishnan, A. Divakaran, “Generative Process Tracking for Audio Analysis”, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006.

[75] T. Sainath, D. Kanevsky, G. Iyengar, “Unsupervised Audio Segmentation using Extended Baum-Welch Transformations”, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007.

[76] D. Ellis, “Detecting Alarm Sounds”, in Proc. Workshop on Consistent and Reliable Acoustic Cues, 2001.

[77] K. Goh, K. Miyahara, R. Radhakrishan, Z. Xiong, A. Divakaran, “Audio-Visual Event Detection Based on Mining of Semantic Audio-Visual Labels”, in Proc. SPIE Conference on Storage and Retrieval for Multimedia Databases, pp. 292-299, 2004.

[78] A. Härmä, “Automatic recognition of bird species based on sinusoidal modeling of syllables”, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[79] H. Kim, J. Burred, T. Sikora, “How efficient is MPEG-7 for General Sound Recognition”, in Proc. International AES Conference Metadata for Audio, 2004.

[80] H. Kim, N. Moreau, T. Sikora, “Audio Classification Based on MPEG-7 Spectral Basis Representations”, IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Audio and Video Analysis for Multimedia Interactive services, 2004.

[81] S. Ravindran, D. Anderson, M. Slaney, “Low Power Audio Classification for Ubiquitous Sensor Networks”, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004.

[82] D. Hoiem, Y. Ke, and R. Sukthankar, “SOLAR: Sound Object Localization and Retrieval in Complex Audio Environments”, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.

[83] D. Ellis, K. Lee, “Minimal-Impact Audio-Based Personal Archives”, in Proc. ACM workshop on Continuous Archiving and Recording of Personal Experiences, 2004.

[84] P. Nordqvist, Sound Classification in Hearing Instruments, PhD Thesis, Royal Institute of Technology, Stockholm, Sweden, 2004.

[85] 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. of IEEE (DCASE), Budapest, Hungary, pp. 16-19, September 2016.

[86] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, “An exemplar-based NMF approach to audio event detection,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, PP. 1-4, Oct 2013.

[87] A. Cont, “Realtime multiple pitch observation using sparse non-negative constraints,” International Symposium on Music Information Retrieval (ISMIR), Victoria, Canada, PP. 206-211, Aug 2006.

[88] A. Cont, S. Dubnov, D. Wessel, “Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints,” in Proc. of 10th Int. Conf. Digital Audio Effects (DAFx), Bordeaux, France, PP. 85-92, 2007.

[89] F. Tom, “ROC Graphs: Notes and Practical Considerations for Researchers”. Pattern Recognition Letters, vol. 27, no. 8, pp. 882–891, 2004.

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

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

[92] S. Innami and H. Kasai, “NMF-based environmental sound source separation using time-variant gain features,” Computers & Mathematics with Applications, vol. 64, no. 5, pp. 1333 – 1342, 2012.

[93] M. W. Berry, M. Browne, A. Langville, V. P. Pauca, and R. J. Plemmons, “Algorithms and applications for approximate nonnegative matrix factorization,” Comput. Stat. Data Anal. Vol. 52, no. 1, pp. 155–173, 2007.

[94] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, “Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation,” Wiley-Blackwell, 2009.

[95] C. Fevotte, and J. Idier, “Algorithms for nonnegative matrix factorization with the beta-divergence,” Neural Computation, vol. 23, no. 9, pp. 2421-2456, 2011.

[96] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, “Convergence guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence,” IEEE International Workshop on Machine Learning for Signal Processing, pp. 283–288, Finland, 2010.

[97] D. L. Sun, C. Fevotte, “Alternating direction method of multipliers for nonnegative matrix factorization with the β-divergence,” IEEE (ICASSP), Florence, Italy, pp. 6201-6205, 2014.

[98] S. Boyd, L. Vandenberghe: Convex Optimization, Cambridge University Press, Cambridge, 2004.

[99] F. Sha, Y. Lin, L. K. Saul, and D. D. Lee, “Multiplicative updates for nonnegative quadratic programming,” Neural Computation, Vol. 19, no. 8, pp. 2004–2031, 2007.

[100] M. Shashanka, B. Raj, P. Smaragdis, “Probabilistic latent variable models as nonnegative factorizations,” Comput. Intell. Neurosci., doi: 10.1155/2008/947438, May 11, 2008.

[101] A. J. Bell, T. J. Sejonwski, “The gIndependent Componentsh of Natural Scenes are Edge Filters,” Vision Res., vol. 37, no. 23, pp. 3327-3338, 1997.

[102] A. Coates, H. Lee, and A. Y. Ng, “An analysis of singlelayer networks in unsupervised feature learning,” in Proc. the 14-th International Conference on AI and Statistics, 215-223, 2011.

[103] M. Aharon, M. Elad and A. Bruckstein, K-SVD,“An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation,” IEEE Transactions on Signal Processing, vol 54, no. 11, pp. 4311-4322, 2006.

[104] J. Mairal, F. Bach, J. Ponce and G. Sapiro, “Online Dictionary Learning for Sparse Coding,” International Conference on Machine Learning, Montreal, Canada, 2009.

[105] Music Information Retrieval Evaluation eXchange (MIREX): Multiple Fundamental Frequency Estimation & Tracking. Available online: http://www.music-ir.org/mirex/, 2016.

[106] T. Heittola, M. Annamarie, sed_eval, Evaluation toolbox for online: https://github.com/TUT-ARG/sed_eval, 2016.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

Abstract

Audio Event Detection and classification (AED) in the workplace and life has become a need in the modern 21st century, so that a branch of research in artificial intelligence and, consequently, in the Computational Auditory Scene Analysis (CASA) group has been dedicated to this work.

This thesis offers two important approaches for audio event detection and classification in environmental audio signals. In the first approach, we proposed a model with several smaller independent components, each with a strong point for the entire model. Most existing models view all expectations of modeling in a coherent and statistical definition, but here the main model is made up of a multi-component parallel composition in order to utilize the maximum power of raw data and extract and manage their information desirable. In addition, such modeling replaces the negative side effect between the components, especially the features, with the positive side effect. The F1 measure obtained in this method was 80.6% and 72.8% for the frame-based and event-based audio detection, respectively. The second approach is based on the using non-negative matrix factorization (NMF) on the observation signal to capture the sample spectral tissue from sounds as atoms or patches in a dictionary, and then reconstruct the test signal by producing a sparse representation of the dictionary atoms in the representation of the segments of that signal. This method is based on the separating a signal into the primary components, that is, voice events. The F1 measure obtained in this method was ۵۶.۵٪ □□□ ۴۹.۲٪ for the frame-based and event-based audio detection, respectively. These results show the superiority of the proposed methods compared to previous methods done on the DCASE2013 database.

Keywords: audio event detection and classification (AED), segmentation, feature extraction, sound source separation, non-negative matrix factorization (NMF), sparse representation of signal, auditory scene analysis, machine learning algorithms, spectrogram, audio surveillance.

Shahrood University of Technology

Faculty of Computer Engineering and Information Technology

Ph.D Thesis in Artificial Inteligance

Non-Overlapped Event Detection and Classification for Environmental Audio Signals

By: **Morad Derakhshan**

Supervisor
Dr. Hossein Marvi

Advisor
Prof. Hamid Hassanpour

August 2017