کشف و رده بندی وقایع صوتی محیطی با استفاده از نگاشت سگمنت بر دیکشنری در نمایش تنک

مراد درخشان[۱]، دانشجوی دکترا؛ حسین مروی[۲]، دانشیار

۱- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران – mderakhshan@shahroodut.ac.ir
۲- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران – h.marvi@shahroodut.ac.ir

چکیده: در اینجا کشف و رده بندی وقایع صوتی محیطی با استفاده از نگاشت سگمنت بر دیکشنری در نمایش تنک مورد پژوهش قرار گرفته است. یکی از مشکلات رویکردهای مبتنی بر دیکشنری، نبود کنترل لازم در نحوه نگاشت سگمنت‌های ورودی بر بردارهای پایه دیکشنری جهت شناسایی صدای موجود در سگمنت است. این عدم کنترل، سبب تجزیه سگمنت به تعدادی از اصوات کوچک که بخش‌های سگمنت هستند می‌شود. در اینجا الگوریتمی جهت کنترل صریح نگاشت هر سگمنت بر بردارهای پایه دیکشنری پیشنهاد شده است که از طریق به‌کارگیری تابع انحراف-بتا و کنترل تنکی تجزیه نامنفی دنبال می‌شود و کل سگمنت را به عنوان یک واقعه صوتی شناسایی می‌کند. در عمل با کنترل میزان تنکی، هر سگمنت با مجموع چند بردار پایه تقریب زده می‌شود به‌طوری که کل سگمنت یکجا شناسایی شود. این الگوریتم در مراحل تست جهت شناسایی صداهای محیط اداری بکار رفته و نتایج به دست آمده گویای رشد نرخ شناسایی صداها می‌باشد که تاییدی بر مؤثر بودن روش پیشنهادی است.

واژه‌های کلیدی: کشف و رده بندی وقایع صوتی، تجزیه نامنفی، تولید دیکشنری، بردارهای پایه، تجزیه تنک، تابع انحراف بتا، نگاشت سگمنت، نمایش زمان-فرکانس.

Audio Event Detection Using the Mapping Segment on the Dictionary in Sparse Representation

Morad Derakhshan[1], PhD student; Hossein Marvi, Associate Professor[2]

1- Computer and IT Engineering Department, Shahrood University of Technology, Shahrood, Iran, Email: mderakhshan@shahroodut.ac.ir
2- Faculty of Computer and IT Engineering, Shahrood University of Technology, Shahrood, Iran, Email: h.marvi@shahroodut.ac.ir

Abstract: Audio event detection (AED) is addressed by using a segment mapping on the NMF dictionary in the sparse representation. One problem with dictionary methods is the lack of controls in the decomposition process of the input signal, so the process yields some unstructured sound pieces that are not the valid audio events. We proposed an algorithm which uses sparsity constraint and beta-divergence to decompose the input segments into the predefined dictionary atoms instead. Here, the sparsity control in each segment decomposes it into a linear combination of basis vectors thereby the segment is approximated into a hypothetical audio event. This method is applied to the recognition of variety live official sound events and has promising results.

Keywords: Audio event detection (AED), non-negative matrix factorization (NMF), dictionary creation, basis vectors, sparsity constraint, beta-divergence, mapping segment, time-frequency representation.

تاریخ ارسال مقاله: ۱۳۹۵/۱۰/۲۹
تاریخ اصلاح مقاله: ۱۳۹۵/۱۲/۲۳
تاریخ پذیرش مقاله: ۱۳۹۶/۰۴/۲۳
نام نویسنده مسئول: حسین مروی
نشانی نویسنده مسئول: ایران - شاهرود – دانشگاه صنعتی شاهرود - دانشکده مهندسی کامپیوتر و فناوری اطلاعات.

۱- مقدمه

کشف و شناسایی وقایع صوتی کاربرد فراوانی در سیستم‌های نظارت صوتی و بررسی محتوایی رکوردهای ضبط‌شده در محیط دارد. پس از کشف وقایع صوتی پردازش‌های ثانویه متعددی قابل انجام است. این پردازش‌ها شامل شناسایی محیط، تعیین فعالیت‌های رخداده در محیط، تولید شاخص‌ها در مورد محتوای یک فایل صوتی اعم از صداهای موجود در آن و مکان یابی محتوایی در رکورد ضبط‌شده می‌باشد. این نوع کسب اطلاعات به دلیل فراوانی رکوردهای صوتی و مزایایی که ضبط صدا نسبت به ضبط ویدیو دارد روز به روز در حال افزایش است. تا کنون روش‌های مختلفی جهت بهبود عملکرد چنین سیستم‌هایی پیشنهاد شده است اما هنوز جای مطالعات فراوانی وجود دارد. در سالیان اخیر الگوریتم‌های جداسازی کور منابع از جمله استفاده از تجزیه نامنفی ماتریس[1]سیگنال مشاهده مد نظر قرار گرفته که می‌توان به مقالات [۱-۷] اشاره نمود. اما به دلیل پایه‌ای نبودن این روش‌ها در شناسایی وقایع صوتی نتایج به دست 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 [۸]. در سیستم شناسایی گفتار صدادار پیشنهادشده در [۹] از ویژگی تابع خودهمبستگی پوش بهبود یافته[2]جهت حذف پیک‌های اشتباه یا مضارب پیک اصلی در مرحله برچسب گذاری واحدهای گفتاری استفاده شده است. این ویژگی منجر به شناسایی بهتر حداکثرهای اصلی تابع خودهمبستگی پوش پاسخ برای واحدهای گفتاری گردیده و درصد موفقیت شناسایی گفتار صدادار را تا حد زیادی تضمین نموده است اما عدم وجود پیک‌های مذکور در صداهای محیطی شناسایی آن‌ها را دشوار می‌کند. بهبود حاصل از روش تخمین قدرت سیگنال به نویز پیشنهادی در [۱۰] که بر اساس ماسک دودویی ایده‌آل یا تخمینی با فرضیات خاص است نیز بیشتر مناسب سیگنال‌های گفتار در محیط‌های نویزی است. 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. به‌طور 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 به دست 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. به‌طوری که افزوده شدن یا جایگزینی اجزای صوتی بجای یکدیگر می‌تواند منجر به خطای عدم همخوانی سیگنال بازتولیدی با سیگنال اصلی گردد [۱۱]. 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. از مطالعات مرتبط می‌توان به اعمال محدودیت تنک[3]در تجزیه نامنفی جهت مشاهده گام چندگانه موزیکال و شناسایی دستگاه‌های موسیقی در [۱۲، ۱۳] اشاره نمود. اما با وجود کنترل‌های صورت گرفته بر روند تجزیه بخش‌های حاصل فاقد ساختار است که کارآمدی روش را کم می‌کند. فرض 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 بر حسب 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.

ادامه مقاله به صورت زیر تنظیم شده است. در بخش ۲ روش تجزیه نامنفی ماتریس مشاهدات معرفی شده و توسعه آن به مدل‌های انحراف بتا و تجزیه نامنفی تنک انجام شده است. در بخش ۳ دیاگرام کلی سیستم پیشنهادی برای کشف وقایع صوتی ارائه شده است. بخش۴ به آزمایشات و نتایج با فرمول بندی تجزیه نامنفی همراه با کنترل میزان تنکی و انحراف بتا تمرکز دارد. بخش ۵ نتیجه گیری و پیشنهادات برای توسعه سیستم پیشنهادی را در بر دارد.

۲- تجزیه نامنفی ماتریس

تکنیک کاهش رتبه ماتریس، از جمله تجزیه نامنفی با داشتن ماتریس نامنفی V_{n*m} و یک عدد صحیح $r \le min(n,m)$ ماتریس V را به صورت رابطه (۱) تجزیه می‌کند:

$$V_{n*m} \approx W_{n*r} * H_{r*m} \qquad (۱)$$

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 (۲) محاسبه می‌شود.

$$v_j \approx \sum_{k=1}^{r} w_k h_{kj} \qquad (۲)$$

هر یک از h_j, w_k ستون j ام و k ام در ماتریس هستند، W دیکشنری و ستون‌های آن پایه‌ها؛ الگوها یا اتم‌ها هستند. ماتریس H ضرایب یا فعال کننده اتم‌های دیکشنری است. در زمان تصویر سازی، هر ستون ماتریس H کد گذاری ستون متناظر در ماتریس V می‌باشد.

از آنجا که رابطه (۱) تجزیه تقریبی به فرم $V \approx W*H$ است، لذا از طریق بهینه سازی یک تابع هزینه انجام می‌شود. اگر $\Lambda = W*H$ باشد و تابع هزینه $C(V,\Lambda)$ نامیده شود آنگاه تجزیه به صورت یک مسئله بهینه سازی محدودشده به صورت زیر خواهد بود:

$$\arg\min_{W \in R_+^{n*r},\, H \in R_+^{r*m}} C(V,\Lambda) \qquad (۳)$$

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

$$C(A,\Lambda) = \frac{1}{2}\| V-\Lambda \|_F^2 = \frac{1}{2}\Sigma_{i,j}(V_{ij}-\lambda_{ij})^2 \qquad (۴)$$

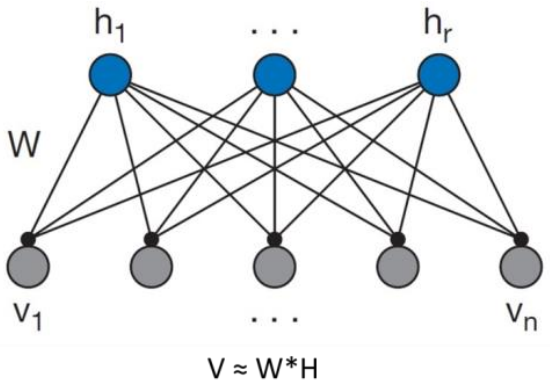
از نظر آماری رابطه (۴) در واقع حداکثر سازی یک تابع شباهت[۵]توسط کمینه نمودن میزان خطا بر حسب فاصله اقلیدسی است.

تابع بهینه سازی تا زمان همگرایی یا به تعداد تکرار مشخص اجرا می‌شود. در تجزیه نامنفی استاندارد برای بهنگام‌سازی عامل‌های W,H از الگوریتم ضربی که در [۲، ۱۳، ۱۴] آمده است استفاده می‌شود که نوعی گرادیان نزولی[۶]همراه با مراحل تطبیقی رابطه (۵) است:

$$H \leftarrow H \otimes \frac{|W^T V|}{W^T WH} \quad , \quad W \leftarrow W \otimes \frac{|VH^T|}{WHH^T} \qquad (۵)$$

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 معمولاً از مدل‌های تصحیح‌شده، محدودیت‌های تصحیح‌شده مثل افزودن عامل تنکی، توابع هزینه تصحیح‌شده همچون استفاده از واگراها[۷]یا افزودن عامل‌های جبران‌کننده استفاده می‌نمایند [۱۵-۱۸]. 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 ۱ 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 مقاله این دیدگاه دنبال شده است.



شکل ۱: تجزیه نامنفی $V \approx W*H$ به صورت یک مدل احتمالاتی با متغیرهای پنهان H و متغیر مشاهده شده V همراه با ماتریس انتقال W قابل تعریف و مدلسازی است.

۱-۲- تجزیه نامنفی ماتریس و بهنگام‌سازی با انحراف بتا

انحراف بتا یکی از انواع توابع پارامتریک مغایرت[۹]نوعی تصحیح کننده 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 زیر تعریف کرد[۲]:

$$d_\beta(a,b) = \frac{1}{\beta(\beta-1)}(a^\beta + (\beta-1)y^\beta - \beta a b^{\beta-1}) \qquad (۶)$$

اگر از رابطه (۶) نسبت به β حد گرفته شود آنگاه به ازای $\beta=0$ انحراف ایتاکورا-سایتو و به ازای $\beta=1$ انحراف کولبک لیبلر به دست می‌آید:

$$d_{\beta=0}(a,b) = d_{IS}(a,b) = (a/b) - log(a/b) \qquad (۷)$$

$$d_{\beta=1}(a,b) = d_{KL}(a,b) = alog(a/b) + b - a \qquad (۸)$$

برای حالت $\beta=2$ تعریف فوق نصف فاصله مربع اقلیدسی است:

$$d_{\beta=2}(a,b) = d_E(a,b) = \frac{1}{2}(a-b)^2 \qquad (۹)$$

انحراف بتا مطابق رابطه (۹) 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[۱۰] $\lambda \in \mathrm{R}_{++}$ رابطه (۱۰) وجود دارد:

$$d_\beta(\lambda a, \lambda b) = \lambda^\beta d_\beta(a,b) \qquad (۱۰)$$

انحراف عددی رابطه (۶) را می‌توان در پردازش سیگنال به یک ماتریس انحراف تحت عنوان انحراف تفکیکی[۱۱]تبدیل نمود. این کار با جمع عنصر-محور انحراف‌ها مانند زیر قابل انجام است:

$$\mathrm{D}_\beta(V,\Lambda) = \sum_{i,j} d_\beta(\upsilon_{ij}, \lambda_{ij}) \qquad (۱۱)$$

بنابراین در پردازش سیگنال تجزیه نامنفی با تابع هزینه انحراف بتا منجر به مسئله بهینه‌سازی مقیدشده رابطه (۱۲) می‌گردد که هم ارز رابطه (۳) است.

$$\arg\min_{W \in R_+^{n*r},\, H \in R_+^{r*m}} \mathrm{D}_\beta(V, WH) \qquad (۱۲)$$

رابطه (۱۲) در پردازش ماتریس‌های مشاهده V و دیکشنری W و فعال کننده‌های اتم‌های دیکشنری H به صورت رابطه (۱۳) قابل بیان است:

$$\mathrm{D}_\beta(V,WH) = \sum_{n=1}^{N}\sum_{m=1}^{M} d_\beta([V]_{nm} \mid [WH]_{nm}) \qquad (۱۳)$$

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 (۵) یک بهنگام‌سازی ضربی اکتشافی[۱۲]برای عامل‌های W,H می‌تواند دارای فرم زیر باشد:

$$H \leftarrow H \otimes \frac{W^T (V \otimes (WH)^{\beta-2})}{W^T (WH)^{\beta-1}}$$
$$W \leftarrow W \otimes \frac{(V \otimes (WH)^{\beta-2}) H^T}{(WH)^{\beta-1} H^T} \qquad (۱۴)$$

تابع هزینه انحراف بتای کولبک لیبلر در رابطه (۱۵) آمده است:

$$C\left(A,\Lambda\right)=D_{KL}\left(V,\Lambda\right)=\sum_{i,j}v_{ij}log\,\frac{v_{ij}}{\lambda_{ij}}+\lambda_{ij}-v_{ij} \quad (۱۵)$$

این تابع با جایگزینی در رابطه (۱۳) به‌صورت رابطه (۱۶) است:

$$\mathrm{D}_{\beta=1}(V,WH)=D_{KL}(V,WH)=\sum_{n=1}^{N}\sum_{m=1}^{M}(V_{nm}\log\frac{V_{nm}}{[WH]_{nm}}+[WH]_{nm}-V_{nm}) \quad (۱۶)$$

با تغییر تابع هزینه، بهنگام‌سازی ماتریس‌های W,H نیز طبق روابط موجود در [۱۳، ۱۴] به رابطه (۱۷) تغییر می‌یابند:

$$H\leftarrow H\otimes\frac{W^{T}(V\otimes(WH)^{-1})}{W^{T}WH}$$
$$W\leftarrow W\otimes\frac{(V\otimes(WH)^{-1})H^{T}}{WHH^{T}} \quad (۱۷)$$

به ازای $\beta=1$ و $\beta=2$ بهنگام‌سازی ضربی به ترتیب معادل کولبک لیبلر و اقلیدسی است. هر چند هزینه در بهنگام‌سازی‌ها به ازای $0\le\beta\le2$ به‌طور یکنواخت کم می‌شود اما این وضعیت برای سایر مقادیر β صدق نمی‌کند [۱۹]. در اینجا شناسایی صداها با روش تجزیه انحراف بتا و روش محدودشده با قید تنک انجام شده‌اند.

۲-۲- تجزیه نامنفی ماتریس با محدودیت تنک[۳] جهت کشف وقایع صوتی

تجزیه نامنفی رابطه (۱) که بسط آن در روابط (۲) تا (۵) آمده است با یک معضل اساسی روبرو است و آن نرسیدن به پاسخ‌های انحصاری و ثابت است به‌طوری‌که 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 غیر صفر نمایش داده می‌شود. لذا محدود بودن تعداد بردارهای پایه سبب نتایج انحصاری در نمایش سیگنال می‌شود. بسط‌های مختلفی از جمله افزودن عامل جریمه تنکی[۴] و یا استفاده از بهنگام‌سازی ضربی همگرا همراه با عامل جریمه نرم l_1 و تصویرسازی گرادیان نزولی به‌منظور 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 [۲۰] به‌صورت رابطه (۱۸) معرفی شده است:

$$arg\,min\frac{1}{2}h^{T}(W^{T}W+\lambda_2 I)h+(\lambda_1 e-W^{T}v)^{T}h\ ,\quad -Ih\le0\ ,\ \lambda_1,\lambda_2\ge0\ ,\ h\in R_{+}^{r} \quad (۱۸)$$

λ_1,λ_2 پارامترهای تنظیم کننده هستند. این رابطه یک بهنگام‌سازی درجه دوم کوژ[۵] [۲۱] است که با مسئله کمترین خطای مربعات نامنفی به صورت رابطه (۱۹) قابل بیان است:

$$arg\,min\frac{1}{2}\|v-Wh\|_2^2+\lambda_1\|h\|_1+\frac{\lambda_2}{2}\|h\|_2^2 \quad (۱۹)$$

در رابطه (۱۹) نرم l_1 بردارهای کم تنک را جریمه می‌کند، نرم l_2 یک حالت خاص از تصحیح کننده تیکونوو[۶] 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 [۲۰]. 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 استفاده می‌شود. محدودیت تنک شامل ماتریس W و ماتریس ضرایب H می‌شود [۱۵]:

$$arg\,min\frac{1}{2}\|V-WH\|_F^2+\frac{\eta}{2}\|W\|_F^2+\frac{\lambda}{2}\sum_{i=1}^{n}\|h_i\|_1^2 \quad (۲۰)$$
$$subject\ to\ W,H\ge0$$

h_i ستون iام در ماتریس H است. از دید بیز فرمول فوق احتمال لگاریتم پیشینی[۷] 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) به فرم حداقل سازی خطای مربعات استفاده شده است و ضمن ثابت نگه داشتن یک ماتریس، بهنگام‌سازی ماتریس دیگر انجام می‌گیرد به‌طوری‌که 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.

در پردازش صداها، ماتریس مشاهدات V معمولاً 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} به دست می‌آید. در این تجزیه هر اتم یا بردار پایه w_k از دیکشنری الگوی طیفی kام را می‌سازد. ضریب تأثیر 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 [۲۲] 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 به‌طوری که هر جزء پنهان[۸] بیانگر یک منبع است. در این مدل‌ها تخمین شباهت ماگزیمم[۹] 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.

۳- معماری کلی سیستم پیشنهادی

در این بخش سیستم پیشنهادی برای کشف و رده بندی وقایع صوتی توضیح داده می‌شود. در شکل ۲ 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 (قطعه بندی) و رده بندی (کلاس بندی) فرض می‌شود. سیگنال ورودی ابتدا قطعه بندی شده و قطعات s_1 تا s_n هر کدام در بازه‌های زمانی مشخص t_1 تا t_n حاصل می‌گردد سپس در فاز کلاس بندی، نوع رده یا اتفاق صوتی رخ داده در هر قطعه مشخص می‌شود. وقایع صوتی رده بندی شده e_1 تا e_n نامیده شده است.

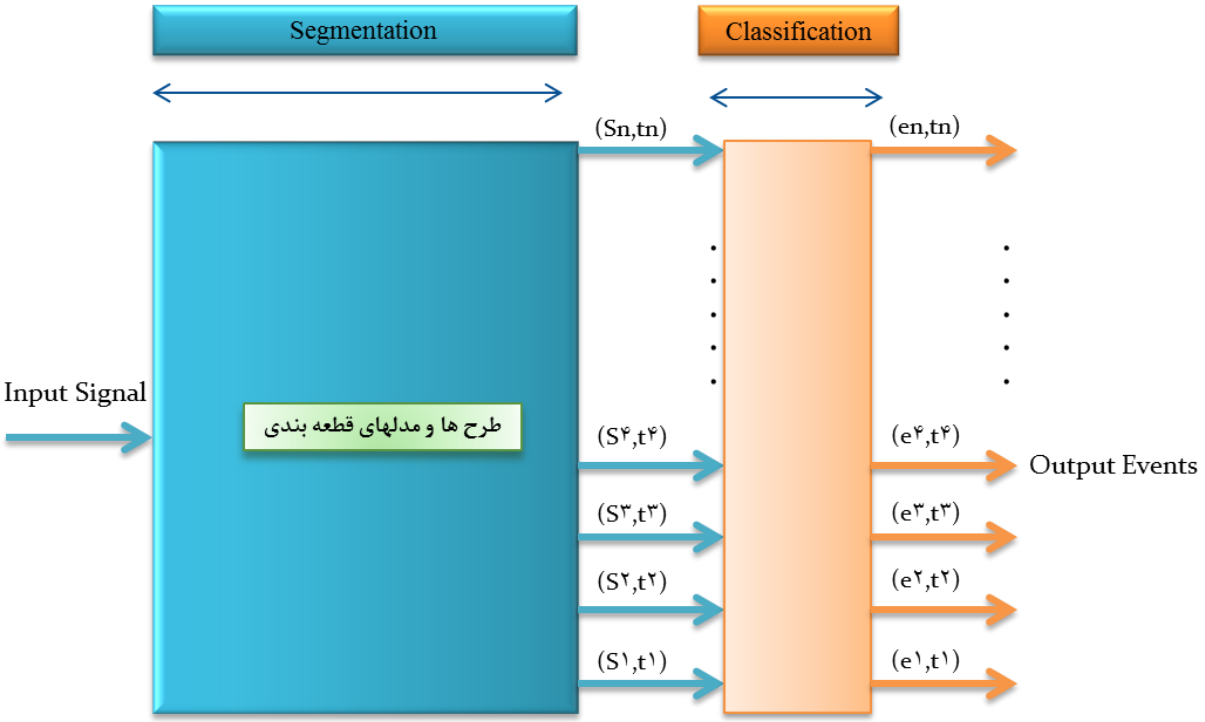
از آنجا که مدل پیشنهادی بر پایه نگاشت سگمنت بر دیکشنری در نمایش تنک است لذا کنترل بر نحوه تجزیه و نگاشت بردار مشاهده بر دیکشنری جهت اجرای فازهای شکل ۲ به صورت یک مدل توسعه یافته در شکل ۳ با فازهای آموزش و تست پیشنهاد شده است.

در شکل ۴ مراحل مختلف تجزیه نامنفی یک ماتریس ورودی v نشان داده شده است. در این الگوریتم نمونه ورودی v پس از چندین تکرار الگوریتم و بهنگام‌سازی مکرر W,H تجزیه می‌گردد 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 عامل‌ها را از بین می‌برد.

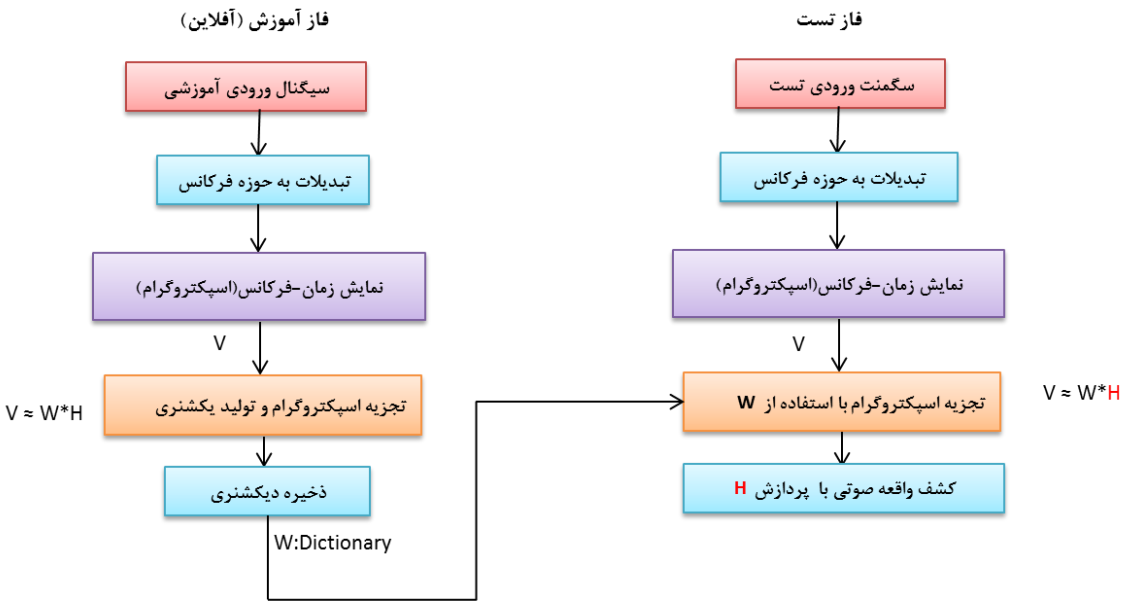
۳-۱- آموزش الگوها

در پایان مرحله آموزش دیکشنری W تولید می‌شود که اتم‌های مجزا برای وقایع صوتی مورد نظر را در ستون‌های خود دارد. این دیکشنری به‌صورت 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. در شکل ۵ مراحل مقدماتی برای هر نمونه آموزشی قبل از تولید ماتریس پایه‌های W نشان داده شده است.

در شکل ۵ برای هر نمونه صوتی i ابتدا آنرا به فریم‌های کوتاه مدت تبدیل نموده و پس از تبدیل در حوزه فرکانس انرژی طیف را به دست آورده و از فریم‌های تبدیل یافته ماتریس $V^{(i)}$ ایجاد می‌شود به‌طوری که ستون V^i_j نمایش فرکانسی فریم زمانی j ام است. سپس تجزیه نامنفی روی $V^{(i)}$ با مرتبه $r=20$ مطابق مراحل شکل ۶ انجام می‌شود تا عامل‌های این ماتریس به دست آید. برای بهنگام‌سازی مدل ضربی از رابطه (۴) استفاده نموده و ماتریس ضرایب نیز نرمالیزه می‌شوند و در انتها الگوی واقعه صوتی iام در بردار ستونی $W^{(i)}$ قرار می‌گیرد.



شکل ۲: سیستم کشف و رده‌بندی صداهای محیطی در دو فاز کشف (قطعه‌بندی) و رده‌بندی (کلاس‌بندی) انجام می‌شود. در فاز کشف قطعات $s1$ تا sn که هر کدام حاوی یک اتفاق صوتی هستند با محدوده‌های زمانی $t1$ تا tn به دست می‌آید. در فاز رده‌بندی هر واقعه صوتی که در هر قطعه رخ داده شناسایی می‌شود و وقایع رده‌بندی شده $e1$ تا en نامیده می‌شوند.



شکل ۳: سیستم کشف و رده‌بندی صداهای محیطی در مدل تجزیه نامنفی. فاز آموزش در قسمت سمت چپ شکل به‌صورت 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 $V \approx W * H$

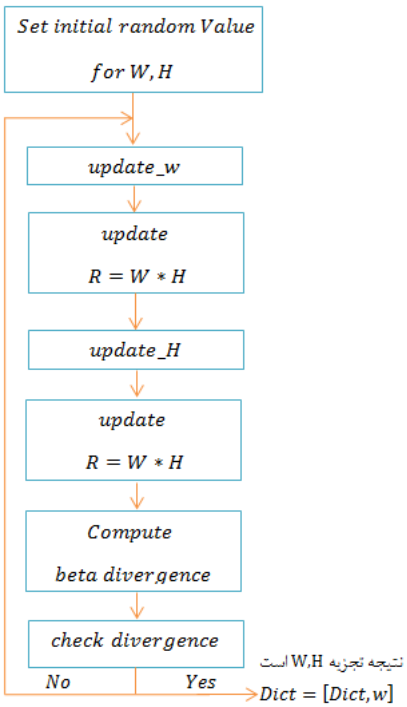
Updating rules:

$R = W * H$

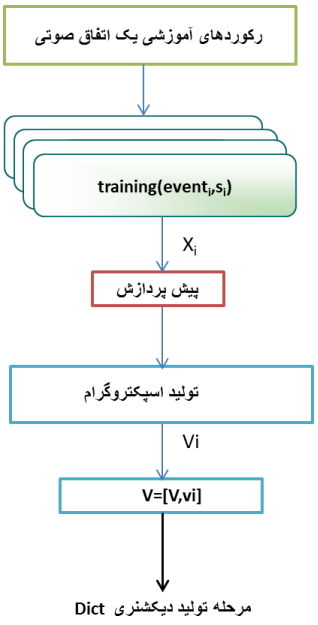
$W = W.*(((R.\wedge(\beta-2).*V)*H^T./\max(R.\wedge(\beta-1)*H^T, eps)$

$H = H.*((W^T*(R.\wedge(\beta-2).*V))./\max(W^T*R.\wedge(\beta-1), eps))$

$eps = 1e-20$



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



شکل ۵: 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 ۶ اجرا می‌گردد.

۲-۳- تجزیه سیگنال ورودی تست

دیکشنری W از کنار هم قرار گرفتن بردارهای ستونی $W^{(i)}$ تولید می‌شود. برای تجزیه یک جریان صوتی ورودی به عنوان تست، 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 تصویر سازی می‌شود. در این فاز که کشف و شناسایی وقایع صوتی است از تجزیه نامنفی برای به دست 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 پیوسته مانند شکل ۷ دیده می‌شوند.

در فاز تست از بردارهای به دست آمده h_j به عنوان فعال کننده‌های 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.

۴- آزمایشات و نتایج

در اینجا جهت ارزیابی کمی سیستم پیشنهادی قابلیت‌های آنرا در موضوع شناسایی صداها بررسی می‌کنیم. در ارزیابی از متریک‌های استاندارد استفاده شده است. این متریک‌ها

$$Accuracy(A),\ Recall(R),\ Precision(P),\ F-measure(F)$$

هستند که بر اساس سگمنت‌های شناسایی شده محاسبه می‌شوند. همچنین نتایج شناسایی را بر اساس هر یک از صداهای ۱۶ گانه به دست می‌آوریم. در این تحقیق از پایگاه داده تهیه‌شده در انجمن بین‌المللی تخصصی پردازش سیگنال صوت[۲۱] 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. همچنین استفاده از کنترل‌های انعطاف‌پذیر در تجزیه را طبق پارامترها بکار گرفته‌ایم.

۴-۱- شناسایی وقایع صوتی محیطی

در اینجا جهت شناسایی صداهای محیطی از صحنه‌هایی شامل ۱۶ نوع واقعه صوتی مختلف استفاده شده است. صداهای مورد پردازش عبارتند از:

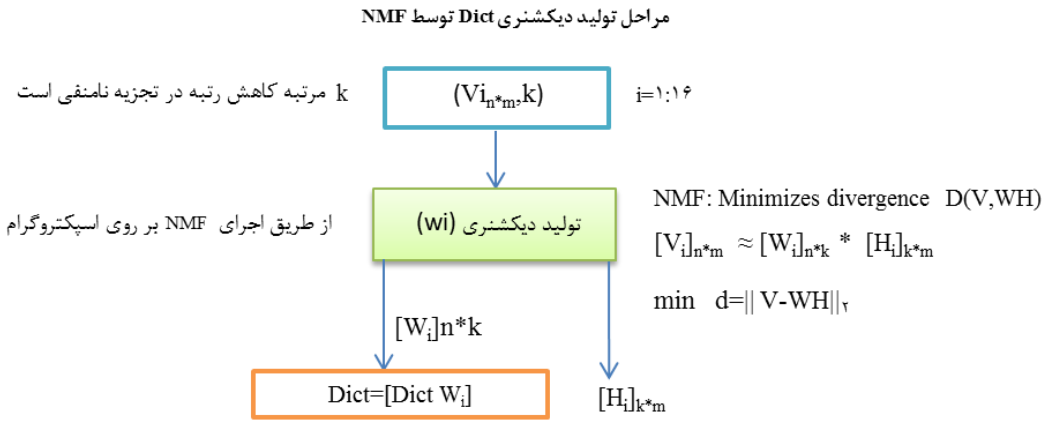
alarm(short alert beep sound), clear throat, cough, door slam, drawer, keyboard(keyboard clicking), keys(keys put on table), knock(door knocking), laughter, mouse, page turn(turning pages back & forward), pen drop(pen, pencil, or marker touching table surfaces), phone(ringing phone), printer, speech, 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. به‌طور نمونه صدای printer از حدود ۹ ثانیه تا حداکثر ۲۲ ثانیه است. صدای pen drop از ۲۵۰ میلی ثانیه تا حداکثر ۱۲۰۰ میلی ثانیه است. صدای alert از ۳۵۰ میلی ثانیه تا ۳ ثانیه است.

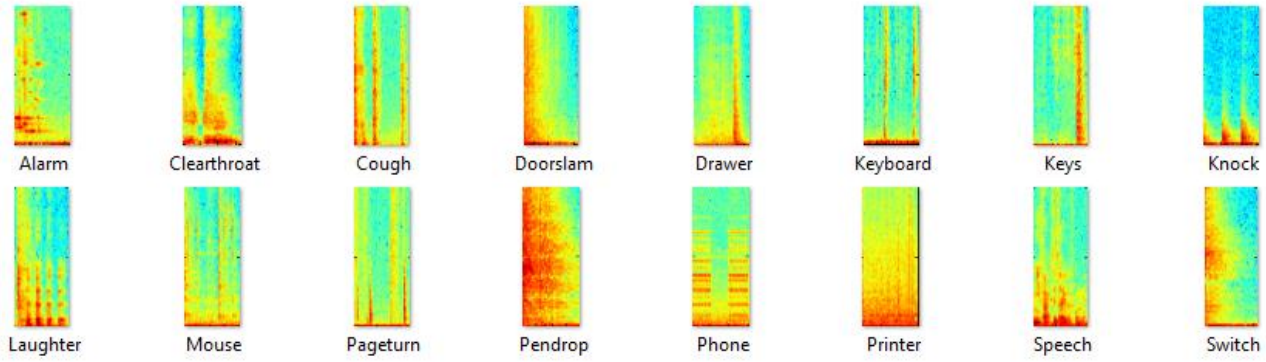
این صداها در محتوای فرکانسی و شکل متفاوت هستند و اغلب آنها از جهات مختلفی متغیر محسوب می‌شوند. برای مثال صداهای
drawer, printer, switch, phone and page turn
دارای طیفی با یکنواختی طولانی، اجزای نویزی مهم و همراه با ترکیب طیفی در سطح لحظه‌ای و میکرونی هستند. صداهای switch و phone از خشنی خاصی برخوردار 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. در شکل ۸ 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. در شکل ۹ تجزیه سیگنال نمونه صحبت دیده می‌شود.

۴-۲- شناسایی وقایع صوتی در رکورد ورودی

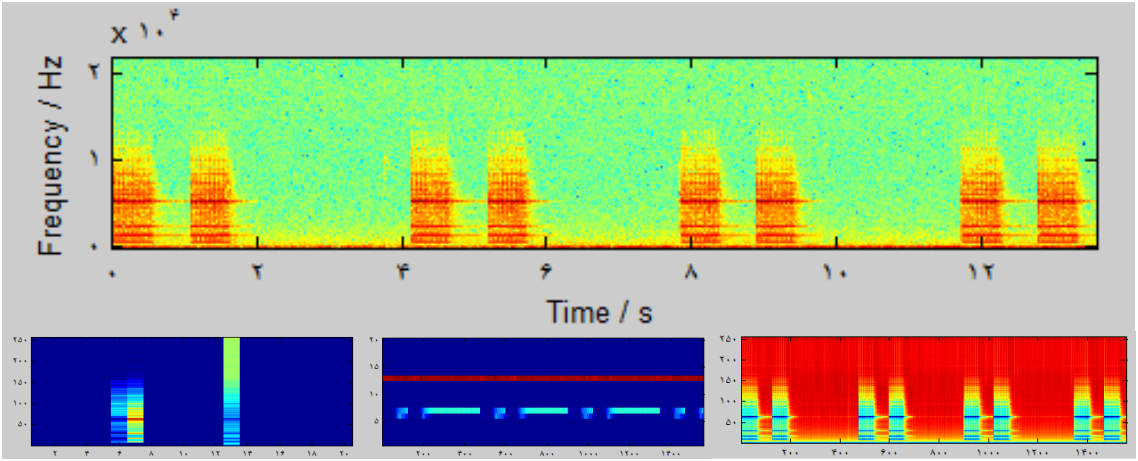
در شناسایی وقایع صوتی ابتدا مطابق فاز آموزش در شکل ۳، به صورت آفلاین یک دیکشنری توسط تجزیه نامنفی همراه با محدودیت تنک از نمونه‌های وقایع صوتی ساخته می‌شود و برای هر واقعه صوتی تعداد $r=20$ بردار پایه در دیکشنری در نظر گرفته می‌شود. تعداد بردارها اختیاری است اما با توجه به نوع صداها در آزمایشات مقدار ۲۰ نتایج بهتری داشت. از این دیکشنری به صورت ثابت در فاز تست مطابق شکل ۳ 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 ۱۰ تجزیه یک رکورد ورودی دارای ۳۷ واقعه صوتی مختلف توسط بردارهای پایه و ضرایب مربوطه نمایش داده شده است. در شکل ۱۱ تجزیه نامنفی رکورد ورودی دیگری با ۳۵ واقعه صوتی مختلف با محدودیت تنک توسط بردارهای پایه و ضرایب مربوطه نمایش داده شده است.



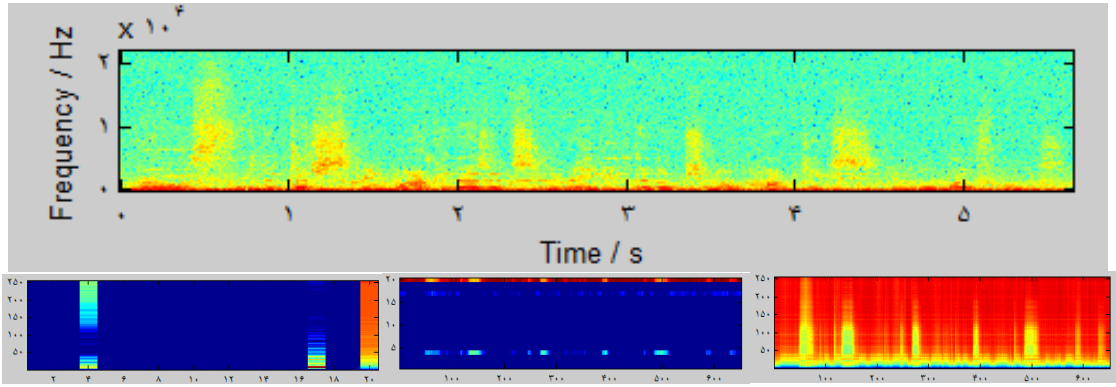
شکل ۶- مراحل لازم جهت تولید ماتریس پایه‌های W برای هر نمونه آموزشی توسط تجزیه نامنفی $W * H \approx V$ نشان داده شده است.



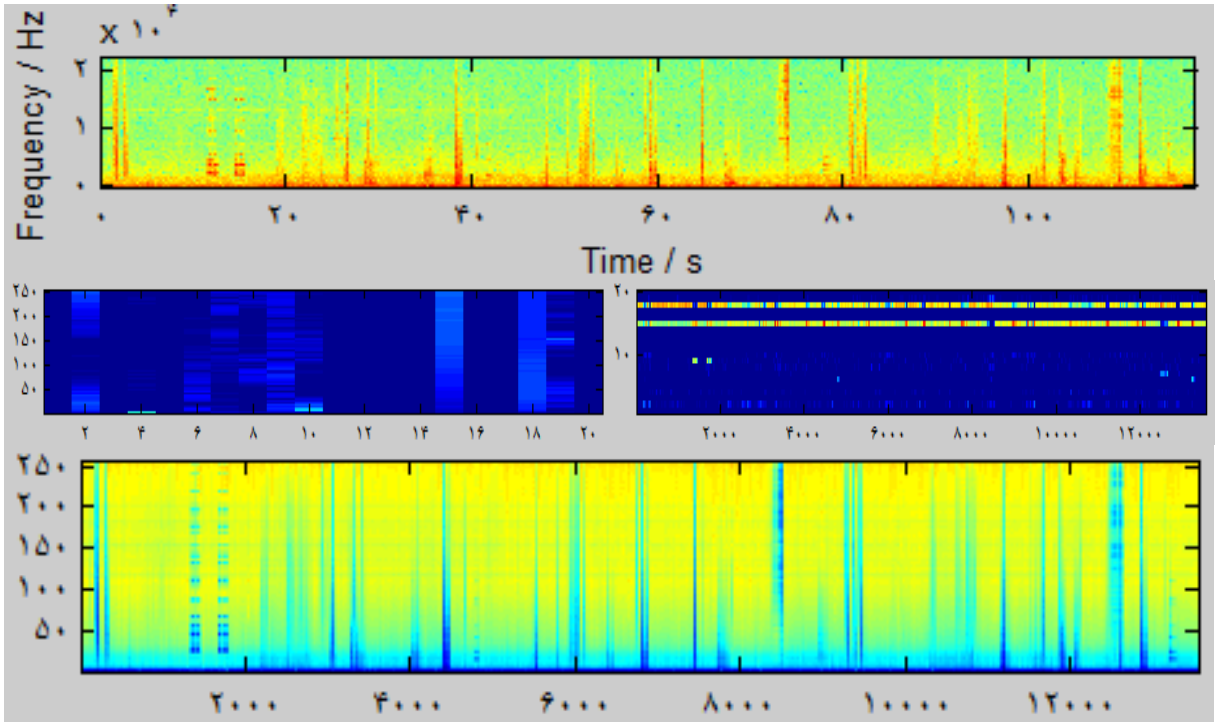
شکل ۷- وصله‌های استخراج‌شده از نمونه‌های آموزشی وقایع صوتی ۱۶ گانه. هر واقعه در دیکشنری $Dict$ دارای وصله مربوط به خود است.



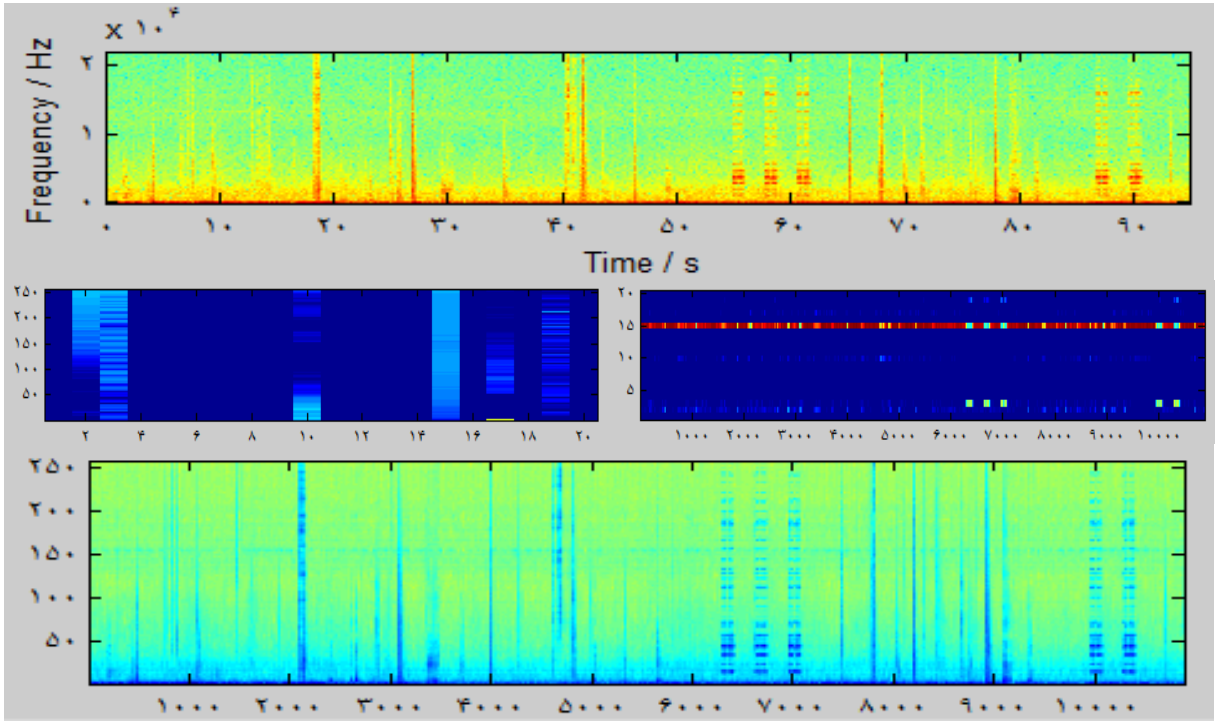
شکل ۸- تجزیه نامنفی صدای تلفن. سیگنال اصلی در بالای شکل و تجزیه به صورت $W * H \approx V$ در پایین شکل آمده است. تنک بودن ماتریس‌های W, H در تصویر مشخص است.



شکل ۹- تجزیه نامنفی صدای صحبت. سیگنال اصلی در بالای شکل و تجزیه به صورت W*H≈V در پایین شکل آمده است.



شکل ۱۰- تجزیه رکورد ورودی با ۳۷ واقعه صوتی مختلف. سیگنال اصلی در بالای شکل و نتیجه تجزیه در پایین شکل دیده می‌شود. دیکشنری استخراجی و فعال کننده‌های 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 ۱۱- 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 فعال کننده‌های 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 ۱۰ 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 $F=46.7\%$ است. همانطور که در شکل ۱۰ 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 کلاً نویز زمینه، سطح فعالیت الگوهای ۱۵ و ۱۸ را بالا نگه داشته است. این خطاها به‌طور آشکاری در شکل به‌صورت نقاط قرمز پیوسته دیده می‌شود. تجزیه نامنفی با انحراف بتا با مقدار $\beta=0$ (انحراف بتای ایتاکورا سایتو) و $\beta=0.5$ و $\beta=1$ انحراف بتای کولبک لیبلر) نیز انجام شد که نتایج در مورد انحراف بتا با مقدار $\beta=1$ 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 $F=44.0\%$ به دست 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 ۱۱ 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 ۱۰ است. یکی از علت‌ها تنوع بیشتر نوع صداها در رکورد تست شده در شکل ۱۰ می‌باشد. اثرات نویز زمینه به‌صورت پایدار در شکل ۱۱ 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 شده است. میزان وقایع شناسایی شده در این رکورد $F=49.5\%$ است. تجزیه نامنفی با انحراف بتا نیز انجام شد که معیار ارزیابی برابر $F=44.7\%$ به دست آمد. نتایج به دست آمده از پردازش رکوردهای ورودی با تعداد وقایع مختلف صوتی بر حسب شناسایی فریم صوتی در جدول ۱ آورده شده است. بر حسب نوع صداها و نویز زمینه در هر یک از رکوردها میزان شناسایی متفاوت است.

جدول ۱- میزان $F-measure(F)$ بر حسب شناسایی فریم صوتی در رکوردهای تست حاوی n واقعه با دو نوع الگوریتم تجزیه نامنفی

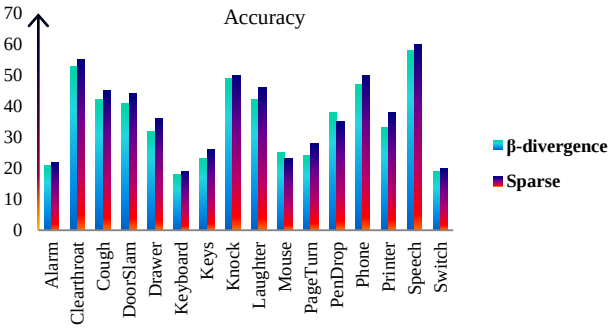
الگوریتم	n=۳۵	n=۳۷	n=۴۱	n=۴۷	میانگین
β-divergence	۵۲/۳	۵۰/۸	۵۲/۳	۵۱/۶	۵۱/۸
sparse	۵۸/۶	۵۴/۷	۵۶/۶	۵۶/۳	۵۶/۵

یکی دیگر از ارزیابی‌های صورت گرفته میزان شناسایی وقایع بر حسب سگمنت‌های صوتی استخراج شده است. در این حالت هر سگمنت حاوی یک واقعه صوتی است. نتایج به دست آمده بر حسب واقعه در جدول ۲ آورده شده است. میانگین شناسایی وقایع در الگوریتم انحراف بتا $F=45.0\%$ و برای الگوریتم تنک $F=49.2\%$ به دست آمد. در اینجا نتایج نسبت به وضعیت فریمی که در جدول ۱ آمده است اندکی کاهش نشان می‌دهد که ناشی از حذف فریم‌هایی است که مجموعاً در محدوده یک سگمنت اصلی نبوده‌اند لذا به‌عنوان واقعه صوتی شناسایی شده محسوب نشده‌اند.

جدول ۲- میزان $F-measure(F)$ بر حسب شناسایی واقعه صوتی در رکوردهای تست حاوی 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 ۱۲ 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 تأثیر نویز زمینه می‌باشد.



شکل ۱۲- 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 تأثیر پذیر از نویز زمینه کمتر است.

مقایسه معیار F روش پیشنهادی با سه رویکرد دیگر در جدول ۳ آمده است. روش‌های ارزیابی frame-based(FB) وevent-based(EB) [۲۴، ۲۵] هستند.

جدول ۳- مقایسه عملکرد الگوریتم با سایر رویکردها

روش ارزیابی/رویکرد		F (FB)	F (EB)
Stowell et al.[7]		۱۰/۷	۷/۴
Gemmeke et al.[11]		۴۹/۷	۳۱/۲
Vuegen et al.[23]		۴۳/۴	۳۰/۸
روش پیشنهادی	انحراف بتا	۵۱/۸	۴۵/۰
	تنک	۵۶/۵	۴۹/۲

۵- نتیجه‌گیری و پیشنهادها

در این مقاله، کشف و رده بندی وقایع صوتی محیطی با استفاده از نگاشت سگمنت بر دیکشنری در نمایش تنک ارائه گردید. برای پرداختن به این موضوع، یک سیستم بر حسب 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 به دست 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 تأثیر مثبت دارد. به‌طوری که بخش‌های دارای فرکانس بالا با انرژی پایین برای جداسازی بین وقایع صوتی مختلف مهم تلقی می‌شوند.

یک پیشنهاد این است که معمولاً در روش‌های تجزیه نامنفی با وجود ساکن نبودن سیگنال‌ها، الگوها ساکن هستند اما در مواردی که صداها دارای دینامیک و تغییر پذیری زمانی زیاد هستند لازم است حالت ساکن نبودن سیگنال در نظر گرفته شود. در این مورد می‌توان در نمایش اولیه سیگنال تغییر پذیری را در بازه‌های زمانی کوتاه لحاظ نمود یا از ترکیب طیف نمونه‌های مختلف یک صدا استفاده کرد. پیشنهاد دیگر در نظر گرفتن وضعیت زمانی الگوها به‌طور مستقیم در الگوهای تجزیه نامنفی است. طوری که بتوان ترکیب تجزیه نامنفی با یک نمایش حالت از صداها را در نظر گرفت. علاوه بر مدلسازی زمانی وقایع، فاز یادگیری الگوها نیز قابل بهبود است. یک امتیازخاص در تجزیه نامنفی فرمول‌بندی فاز آموزش در طرح‌های بسط یافته مختلفی است که می‌توان از آنها برای آموزش یک الگو یا بیشتر برای هر نوع صدا استفاده نمود. این کار با تعیین یک رتبه r برای تجزیه عملی می‌شود لذا می‌توان در انتخاب طرح مناسب برای آموزش الگوها مطالعات بیشتری انجام داد.

در اینجا جهت نمایش اولیه سیگنال از اسپکتروگرام استفاده شده است. در حالی که در صداهای محیطی استفاده از فرکانس غیر خطی همچون تبدیل Q ثابت[۲۲]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 تست بتوان از آن‌ها استفاده نمود. همچنین جهت کم کردن تأثیر نویز پیشنهاد می‌گردد از بهنگام‌سازی غیر ثابت برای بردارهای پایه دیکشنری استفاده گردد تا بدین وسیله بتوان از تقویت بردارهای نویزی جلوگیری کرد.

مراجع

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] مسعود گراوانچی زاده و صنم ایمانی شاملو، «جداسازی تک گوشی گفتار صدادار مبتنی بر روشهای جدید انتخاب واحدهای زمان -فرکانس در فرکانسهای پایین و بالا،» مجله مهندسی برق دانشگاه تبریز، جلد ۴۳، شماره ۱، صفحات ۶۱-۵۱، ۱۳۹۲.

[10] مسعود گراوانچی زاده و پریا دادور، «تخمین SNR ورودی با استفاده از ماسک باینری در سیستمهای مبتنی بر آنالیز ترکیب شنیداری محاسباتی»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۶، شماره ۲، صفحات ۱۹۶-۱۸۷، ۱۳۹۵.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

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

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

زیرنویس‌ها

[1] Non-negative matrix factorization (NMF)

[2] Enhanced envelope autocorrelation function (EEACF)

[3] sparsity constraint

[4] basis vectors

[5] similarity function

[6] gradient descent

[7] divergences

[8] penalty terms

[9] contrast function

[10] scaling factor

[11] separable divergence

[12] heuristic

[13] Non-negative matrix factorization with sparsity constraint

[14] sparsity penalty

[15] convex quadratic program (CQP)

[16] Tikhonov regularization

[17] log-posterior

[18] latent component

[19] maximum likelihood estimation

[20] patch

[21] IEEE Audio and Acoustic Signal Processing (AASP)

[22] constant-Q transform (CQT)