بسمه‌تعالی

دانشگاه سمنان

دانشکده برق و کامپیوتر

سمینار کارشناسی ارشد

مهندسی کامپیوتر- گرایش هوش مصنوعی

عنوان سمینار:

نشانه‌گذاری سیگنال‌های صوتی

Tagging of Audio Signals

توسط:

زهراخسروانی

9411920006

استاد راهنما:

دکتر رحمانی‌منش

فروردین 96

چکیده

این سمینار به بررسی دو مساله دسته‌بندی صحنه‌های صوتی و همچنین نشانه‌گذاری سیگنال‌های صوتی بر اساس چالش‌های ترتیب داده شده در وب‌سایت DCASE2016 [1] می‌پردازد. دسته‌بندی صحنه‌های صوتی نسبت دادن یک برچسب معنایی به یک جریان صوتی برای مشخص کردن محیط آن است. نشانه گذاری سیگنال صوتی طبقه‌بندی فایل صوتی در یکی از دسته های از پیش تعیین شده می‌باشد. این کلاس‌ها می‌تواند صدای انسان، صدای پرنده ،صدای اتوموبیل و ... باشد. هدف از این گزارش بررسی هم زمان این دو مساله با الهام از وب سایت DCASE2016است. پژوهش‌های انجام‌شده در این دو حوزه جهت افزایش دقت و کارایی، بررسی شده اند. از آنجایی‌که دسته‌بندی صحنه‌های صوتی یک عمل پایه جهت نشانه‌گذاری سیگنال‌های صوتی است، بیشتر تحقیقات انجام شده در زمینه دسته‌بندی صحنه‌های صوتی و حوزه‌های مرتبط مرورشده اند.

[1] Detection and Classification of acoustic scenes and events

فهرست مطالب

فصل اول
نشانه گذاری سیگنالهای صوتی

1-1 مقدمه

از زمانی‌که تشخیص گفتار خودکار (ASR) [1] در سیستم‌های صنعتی به کار گرفته شد، چشم‌انداز دست‌یابی به الگوریتم‌هایی که می‌توانند تمام حالات صدا را توصیف کند، روشن‌تر شده است [1]. در ASR محققان به ارتقای کیفیت تشخیص در شرایط صوتی نامناسب مثل اصوات دور با نویز پس‌زمینه می‌پردازند [2]. درجای دیگر، پیشرفت‌ها در بازیابی اطلاعات موسیقی (MIR)[2]، سیستم‌هایی را برای ما به ارمغان آورده‌اند که می‌توانند نت‌ها و رکوردهای موسیقی را رونویسی کنند [3]، یا از یک قطعه بی‌کیفیت صوتی عنوان آلبوم و خواننده آن را تشخیص دهند [4]. با این وجود، گفتار و موسیقی تنها دو گونه از گونه‌های فراوان صوت هستند که می‌توانند در محیطهای داخلی یا خارجی شنیده شوند. بیشتر و بیشتر، ماشین‌های مستقر در محیط‌های متنوع مانند گوشی‌های موبایل، ابزارهای کمکی شنوایی، یا روبات‌های خودکار می‌توانند بشنوند، اما آیا آن‌ها می‌توانند حس کنند چه چیزی را می‌شنوند ؟ [34]

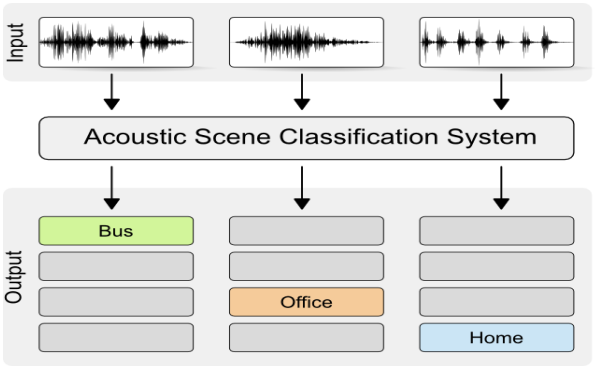
صدا اغلب یک مکمل برای محتواهایی مثل ویدیو است که حاوی اطلاعاتی از زمان حال است. با این تفاوت که صوت به روش‌های آسان‌تری، مثلا با استفاده از یک گوشی موبایل، می‌تواند جمع‌آوری شود. اطلاعاتی که از یک آنالیز صوتی معنایی جمع‌آوری می‌شود می‌تواند برای پردازش‌های بعدی مثل مسیریابی روبات، هشدار به کاربر، یا آنالیز و پیش‌بینی الگوهای یک اتفاق مفید باشد [5]. فراتر از ابزارهای شنوایی، تکنولوژی‌های مشابهی در جستجو و فهرست‌گذاری خودکار آرشیوهای صوتی که به طور فزاینده در سال‌های اخیر رشد کرده‌اند، کاربرد دارد [6]. آرشیوهای صوتی اغلب شامل یک تنوع غنی ازگفتار، موسیقی، صدای حیوانات، فضای صدای شهری، صدای ضبط‌شده قومیتی وغیره می‌باشد. هم‌اکنون دسترسی به این آرشیوها از آرشیوهای متنی عقب مانده است.

[1] Automatic speech recognition
[2] Music information retrieval

برای شبیه‌سازی پژوهش‌ها درشنوایی ماشین برای محیط‌های صوتی عمومی، یک چالش تحقیقاتی در سالهای 2012 و 2013 تحت نظارت کمیته تکنیکی پردازش سیگنال صوتی و آکوستیکی[1] IEEE (IEEE AASP) به صورت چالش آشکارسازی ودسته‌بندی صحنه‌های صوتی و رویدادها (DCASE)[2] ترتیب داده شد. این چالش روی دو حوزه مرتبط اما نسبتا کلی که یک سیستم عمومی شنوایی ماشین انجام می‌دهد، تمرکز کرده است: تشخیص نوع محیط (صحنه صوتی)، و آشکارسازی و دسته‌بندی اتفاقاتی که دریک صحنه می‌افتد [35].

این کارها که ما آن‌ها را به عنوان کارهای" شنوایی ماشین" توصیف می‌کنیم، همچنین می‌توانند تحت موضوع کلی آنالیز محاسباتی صحنه شنوایی (CASA) [3] در نظر گرفته شوند [7]. این نامگذاری به کارهای تاثیرگذار برگمن در حوزه توانایی "آنالیز صحنه شنوایی" بشر برمی‌گردد [8]. بنابراین CASA اغلب به عنوان ارائه راه‌حلی برای تقلید مراحل شنوایی انسان به کار می‌رود [7]. اولین نتایج چالش مذکور درکنفرانس IEEE WASPAA سال 2013 منتشر شد [9].

در دهه‌های اخیر تعدادی راه‌حل برای ASC [4] پیشنهاد شده‌است. با این وجود، عدم وجود مجموعه داده های معیار حس می‌شود. کمیته IEEE AASP اولین آشکارسازی و دسته‌بندی صحنه‌های صوتی و رویدادها را در سال 2013 ترتیب داد، و پس از آن چالش DCASE در سال 2016 با یک مجموعه داده ASC گسترش داده شده است. هدف از این گزارش بررسی مسائل دسته‌بندی صحنه‌های صوتی با توجه به قالب ارائه شده در چالش سال 2016 در این وب سایت [5] است [34].



شکل 1- نمای کلی سیستم دسته‌بندی صحنه های صوتی

[1] The IEEE Audio And Acoustic Signal Processing Technical Committee
[2] Detection and classification of acoustic scene and events
[3] computational auditory scene analysis
[4] Acoustic scene classification
[5] http://www.cs.tut.fi/sgn/arg/dcase2016/task-acoustic-scene-classification

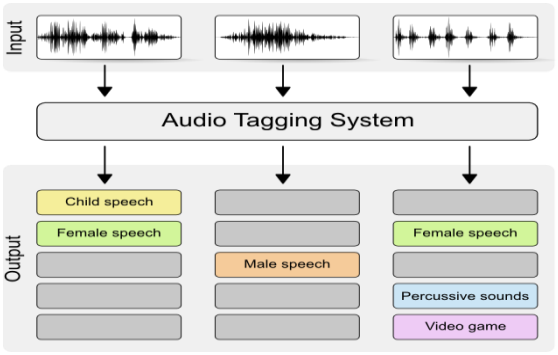
2-1 مفاهیم پایه

در حوزه شنوایی ماشین، تشخیص محیط یک مساله مهم است. دسته‌بندی صحنه صوتی (ASC) ابزارها را قادر می‌کند که محیط را درک کنند و پایه بسیاری از کاربردها است. ASC به عمل نسبت دادن یک برچسب معنایی به یک جریان صوتی گفته می‌شود که محیطی را که صوت در آن تولید شده است را مشخص می‌کند. هدف دسته‌بندی صحنه صوتی، دسته‌بندی یک صدای ضبط شده در یکی از کلاس‌های از پیش تعیین شده است که محیطی را که صدا در آن ضبط شده را مشخص می‌کند، برای مثال پارک، خانه، دفتر و غیره. در شکل 1 نمای کی یک سیتم دسته بندی صحنه صوتی نشان داده شده است. سیگنال ورودی پس از بررسی در یکی از دسته‌های از پیش تعیین شده قرار می‌گیرد.

آشکارسازی رویدادهای صوتی یک حوزه تحقیقاتی بسیار مرتبط با ASC است. یک صحنه صوتی احتمالا به عنوان یک مجموعه از رویدادهای صوتی مثل سرعت، ترمز، اعلان برای مسافران، صداهای بازشدن در و ... در نظر گرفته می‌شود، درحالیکه صدای موتور و اشخاص در پس‌زمینه هستند. بعضی راه‌حل‌ها برای ASC روش‌های آشکارسازی رویداد را به کار می‌برند [34] .

ازسوی دیگر، چالش نشانه‌گذاری سیگنال‌های صوتی[1] بر اساس صداهای ضبط‌شده‌ای است که دریک محیط داخلی تولید شده است. هدف این چالش انجام دسته‌بندی چندبرچسبه[2] (چند برچسب برای یک فایل صوتی ورودی) روی صدای ضبط شده با مدت زمان 4 ثانیه است (نسبت دادن صفر یا بیشتر برچسب به هر قطعه[3] صوتی). انگیزه‌ی راه اندازی این چالش، کاربردهایی مثل سیتم‌های نظارت بر فعالیت‌های انسان[4] است، در این سیستم‌ها تعیین مرز (زمان شروع وپایان) های دقیق رویدادهای صوتی در مقایسه با مشخص کردن خود رویدادها در صحنه صوتی در اولویت دوم قراردارد. به‌علاوه، هنگام به دست آوردن تفسیر برای این چالش مشاهده شد که نشانه‌گذاری دستی قطعه‌های صوتی از لحاظ زمانی مقرون به صرفه‌تر از مکان‌یابی دستی مرزهای رویدادها است. ما اعتقاد داریم که راه‌حل انتخابی پتانسیل زیادی را برای کاهش زمان و درنتیجه بهبود انعطاف حاشیه‌نویسی دستی پایگاه داده بزرگ صوتی فراهم می‌کند.در شکل 2 نمای کلی یک سیستم نشانه‌گذاری سیگنال‌های صوتی نشان داده شده است. همان‌طور که مشاهده می‌شود ممکن است برای یک فایل صوتی چندین دسته در نظر گرفته شود.

[1] Audio Signals Tagging
[2] multi-label classification
[3] chunk
[4] Human activity monitoring



شکل 2- مرور کلی سیستم نشانه‌گذاری صوت

3-1 کاربردها

ابزارهایی همچون گوشی‌های هوشمند، ابزارهای اینترنت اشیا[1] (IoT)، تجهیزات پوشیدنی و روبات‌های مجهز به هوش مصنوعی، و سیستم‌های مسیریابی روبات همه از ASC سود می‌برند. علاوه بر آن، دستیار شخصی هوشمند[2] (IPA) یک حوزه دیگر است که می‌تواند از ASC سود ببرد. IPA ها عامل‌های نرم افزاری توصیه‌گر هستند که با استفاده از تجزیه و تحلیل داده‌های مختلف شامل صوت، تصویر، ورودی کاربر، یا اطلاعات زمینه‌ای مثل مکان، آب وهوا و برنامه شخصی به طور خودکار، توصیه‌هایی را برای کاربران فراهم می‌کنند. سرویس‌های IPA مثل سرویس Google Now مربوط به شرکت گوگل، سرویس Cortana مایکروسافت، و سرویس Siri شرکت اپل، استفاده زیادی از داده‌های صوتی ورودی می‌کنند [1].

الگوریتم‌های محاسباتی که سعی می‌کنند به طور خودکار ASC را انجام دهند از روش‌های یادگیری ماشین و پردازش سیگنال استفاده می‌کنند. کار درحوزه ASC با حوزه‌های تحقیقاتی مرتبط زیادی همراه شده است. به عنوان مثال الگوریتم‌هایی برای تشخیص منبع صوتی به کار می‌روند که سعی در مشخص کردن منبع رویدادهای صوتی در یک صدای ضبط شده دارند و با تکنیک‌های آشکارسازی رویداد ارتباط نزدیکی دارند. همچنین نوع دیگری از الگوریتم‌ها برای تشخیص رویداد استفاده می‌شوند. هدف از این

[1] Internet-of-Things
[2] Intelligent Personal assistant

الگوریتم‌ها مشخص کردن و برچسب‌گذاری محدوده‌های زمانی است که شامل اتفاقات منفرد مربوط به یک کلاس خاص است و به طور خاص در سیستم‌های نظارتی، کمک‌رسانی به سالمندان و آنالیز گفتار با قطعه‌بندی صحنه‌های صوتی به کار می‌رود. به علاوه الگوریتم‌های آنالیز معنایی جریان‌های صوتی که بر تشخیص یا دسته‌بندی رویدادهای صوتی متکی هستند، برای آرشیو شخصی، قطعه‌بندی صوت و بازیابی به کار رفته‌اند [2].

4-1 مجموعه داده

کمیته IEEE AASP اولین آشکارسازی و دسته‌بندی صحنه‌های صوتی و رویدادها را در سال 2013 ترتیب داد، و پس از آن چالش DCASE در سال 2016 با یک مجموعه داده ASC گسترش داده شده است. مجموعه داده‌ها برای مساله دسته‌بندی صحنه صوتی شامل دو مجموعه Development و Evaluation است که مجموعه Development شامل 9 ساعت و 75 دقیقه صدا است و مجموعه Evaluation شامل 3 ساعت و 15 دقیقه صدا است. برای این کار از **TUT Acoustic scenes 2016** استفاده می‌شود. این مجموعه داده شامل صداهای ضبط شده از صحنه‌های صوتی مختلف است که همگی مکان‌های ضبط متفاوتی دارند. برای هر مکان، یک فایل 3-5 دقیقه‌ای ضبط شد. سپس فایل اصلی به قطعه‌های 30 ثانیه‌ای برای چالش تقسیم شد.

کلاس‌ها برای چالش دسته‌بندی صحنه‌های صوتی (15 صحنه) در ادامه آمده‌اند:

- اتوبوس- درحال مسافرت با اتوبوس در شهر (اتوموبیل)
- رستوران/کافه- رستوران/کافه کوچک (داخلی)
- ماشین- راننده یا مسافر، در شهر(اتوموبیل)
- مرکزشهر(خارجی)
- مسیرجنگل (خارجی)
- مغازه عتیقه فروشی- مغازه عتیقه فروشی متوسط (داخلی)
- خانه (داخلی)
- کنارساحل(خارجی)
- کتابخانه(داخلی)
- ایستگاه مترو (داخلی)
- دفترکار- چندین شخص، یک روز عادی کاری(داخلی)
- منطقه مسکونی (خارجی)
- ایستگاه قطار(با قطار یا با اتوموبیل)
- تراموا (تراموا یا اتوموبیل)
- پارک شهری (خارجی)

این مجموعه داده در فنلاند توسط دانشگاه تمپر [1] در سالهای 2015-2016 جمع‌آوری شده است.

[1] Tampere University of Technology

فصل دوم

تحقیقات انجام‌شده

2-1 دسته بندی صحنه های صوتی

هدف از دسته‌بندی صحنه صوتی مشخص کردن محیط یک جریان صوتی با انتخاب یک برچسب معنایی برای آن است [10]. این امر می‌تواند به عنوان یک مساله شنوایی ماشین دسته‌بندی تک‌برچسبه[1] درنظرگرفته شود، که در آن یک مجموعه ازبرچسب‌ها گردآوری می‌شود و سیستم باید دقیقا یک برچسب را برای ورودی انتخاب نماید [11] . بنابراین شباهت‌هایی با کارهای دسته‌بندی صدا مثل تشخیص ژانر موسیقی [12] یا تشخیص گوینده [13] و یا کارهای دسته‌بندی در رسانه‌های دیگر مثل ویدیو دارد.

در [12] روش‌های مختلف استخراج خصیصه در بازیابی اطلاعات برای المان‌های مختلف موسیقی مرور شده است. سپس با داشتن این ویژگی‌ها سه روش سیستم‌های خبره، دسته‌بندی با ناظر، و دسته‌بندی بدون ناظر برای دسته‌بندی ژانر صدا به کار گرفته شده است. علاوه بر آن فیلدها و تکنیک‌های تحقیقاتی جدیدی که ژانر موسیقی را تخمین می‌زنند معرفی می‌شوند. درنهایت نتیجه این است که تکنیک‌های یادگیری ماشین و پردازش سیگنال نقش کلیدی درحل این مساله دارند.

[13] پارامترهای صوتی را بررسی کرده است که درتشخیص گوینده نقش کلیدی دارند. این مقاله یک روش کارامد برای انتخاب این پارامترها ارائه کرده است که الهام‌گرفته از ارتباط بین سیگنال صوت و اشکال مختلف رشته صوتی است. یک شبیه‌سازی از سیستم تشخیص گوینده به وسیله قرادادن دستی رویدادهای گفتاری درسخن و استفاده از پارامترهای انتخاب شده در این مکان‌ها انجام

[1] machine-learning task within the widespread single-label classification paradigm

شده است. ویژگی‌های فرکانس پایه، ویژگی‌های حروف صدادار، طیف همخوان خیشومی[1]، تقریب طیف منبع حلقوی، طول کلمه، و زمان شروع صدا برای استخراج پارامترهای مفید به کار رفته اند.

هنگام دسته‌بندی رسانه زمان‌بندی شده، یک موضوع کلیدی آنالیز داده با ساختار زمانی برای تولید برچسبی است که به درستی سیگنال را بازنمایی کند. دو راهبرد اصلی دراین زمینه وجود دارد. راهبرد اول استفاده از یک مجموعه خصیصه های سطح پایین است که با صحنه به عنوان یک شی منفرد رفتار می‌کند و هدفش این است که صحنه را تحت یک توزیع آماری طولانی و یک مجموعه از خصیصه های طیفی محلی بازنمایی نماید. خصیصه برتر درمیان ویژگی‌های دیگر برای این راه‌حل، استفاده از ضرایب کپسترال فرکانس مل (MFCC) [2] است که خوب عمل کرده است [10]. فوت [14] یک نمونه اخیر است که توزیع MFCC را با چندی‌سازی بردار مقایسه کرده است. این مقاله یک سیستم برای بازیابی مدارک صوتی از طریق شباهت‌های صوتی ارائه می‌دهد. مقیاس شباهت بیشتر بر اساس آماری است که عمدتا از تفکیک‌کننده بردار نظارتی به دست آمده است تا تطابق ساده ویژگی‌های طیفی و گامی. از آن زمان، ساختن یک مدل مختلط گوسی (GMM)[3] برای هر کلاس به عنوان یک راه‌حل برای مقایسه توزیع‌ها ارائه می‌شود [10]. راهبرد دیگر ارائه یک نمایش میانی به عنوان پیش‌پردازش پیش از دسته‌بندی است که صحنه را با استفاده از یک مجموعه ویژگی‌های سطح بالاتر مدل می‌کند. این ویژگی‌ها معمولا با یک دیکشنری یا "اتم های صوتی"[4] ثبت می‌شوند. این اتم‌ها معمولا رویدادهای صوتی یا جریانات در داخل صحنه را بازنمایی می‌کنند که لزوما از قبل شناخته شده نیستند و بنابراین تحت یک شیوه غیرناظر از داده‌ها آموخته می‌شوند. یک مثال استفاده از فاکتورگیری غیرمنفی ماتریس (NMF)[5] است که برای استخراج پایه‌هایی است که بعدا به MFCC تبدیل می‌شوند و برای دسته‌بندی یک مجموعه داده صحنه‌های ایستگاه قطار در [15] استفاده می‌شود. دراین مقاله یک توجه ویژه به کارایی که با استفاده از فاکتورگیری ماتریس غیرمنفی حاصل می‌شود، وجود دارد. شهود این بوده است که اگر منابع داخل صحنه‌های صوتی را آشکار کنیم، یک دسته‌بندی خوب می‌تواند به دست آید. درمثال‌های کوچک مصنوعی ثابت می‌شود که درنظرگرفتن غیرایستا بودن محتوای طیف منبع صوت می‌تواند آشکارسازی منبع را بهبود دهد. درنهایت، متد دسته‌بندی آن‌ها روی مجموعه‌ای ازصداهای ایستگاه قطار پیاده شده و نتایج با روش‌های قبلی مقایسه شده است. نتایج نشان داده اند که این روش فاکتورگیری، دسته‌بندی را بهبود می‌دهد.

[1] nasal
[2] Mel-frequency cepstrum coefficient
[3] Gaussian mixture model
[4] Acoustic atoms
[5] non-negative matrix factorization

براساس این راه‌حل نویسندگان [16] از آنالیز احتمالاتی اجزاء پنهان نامتغیر نسبت به شیفت[1] (SI-PLCA) با محدودیت‌های زمانی ازطریق مدل‌های مخفی مارکوف[2] (HMM) برای رونویسی خودکار موسیقی که تبدیل یک تکه آهنگ به علائم موسیقی است، استفاده کرده‌اند. SI-PLCA می‌تواند تکامل زمانی نت‌ها را مدل کند و همچنین از مدولاسیون فرکانس در نت‌های تولید شده، پشتیبانی می‌کند. سیستم پیشنهادی نسبت به بسیاری از روش‌های پیشرفته رونویسی کارآیی بالاتری را نشان می‌دهد. درنهایت، مدل رونویسی موسیقی پیشنهادی دریک زمینه وسیع‌تر، به نام مدل‌سازی صحنه‌های صوتی به کارگرفته شده است.

[17] از الگوریتم پیگیری انطباق (MP) [3] برای به دست آوردن یک روش کارامد برای انتخاب خصیصه زمان فرکانس استفاده می‌کند که از آن‌ها به عنوان تقویت‌کننده MFCC برای دسته‌بندی صداهای محیطی استفاده کرده است. در این مقاله یک آنالیز خصیصه تجربی برای مشخص‌کردن محیط صوتی انجام شده و برای به دست آوردن ویژگی‌های موثر زمان فرکانس استفاده از الگوریتم پیگیری انطباق (MP)[4] پیشنهاد شده است. MP از دیکشنری اتم‌ها که یک مجموعه قابل تفسیر فیزیکی، انعطاف‌پذیر و قابل‌درک از ویژگی‌ها را نتیجه می‌دهد، برای انتخاب خصیصه استفاده می‌کند. این مجموعه برای تقویت ویژگی‌های MFCC برای بدست آوردن دقت تشخیص بیشتر صداهای محیطی به کارمی‌رود. کارایی سیستم پیشنهادی قابل مقایسه با شنونده‌های انسانی است.

2-2 آشکارسازی رویدادهای صوتی

هدف آشکارسازی رویداد صوتی برچسب‌گذاری محدوده‌های زمانی یک صدای ضبط شده است که یک توصیف نمادین را نتیجه می‌دهد که هر توصیف، زمان شروع، زمان پایان و یک برچسب برای نمونه واحد از یک اتفاق خاص را نشان می‌دهد. این به رونویسی موزیک[3]، و هم‌چنین شناسایی گوینده مربوط می‌شود که به طور مشابه یک نشانه‌گذاری ساختاری را از قطعه‌های زمانی پوشش می‌دهد، و بیشتر بر چرخش‌ها تمرکز دارد تا اتفاقات منفرد[18]. در[18] یک مرور کلی بر روش‌هایی که در حیطه کلیدی حاشیه‌نویسی صوتی، به نام حاشیه‌نویسی گفتار استفاده می‌شود، انجام شده است و درباره ویژگی‌ها و محدودیت‌های آنان بحث شده است. کارایی تکنیک‌های مختلف درقالب چالش ارزیابی رونویسی DARPA EARS Rich بررسی شده است. هم‌چنین شیوه معرفی هر روش به سیستم پخش اخبار واقعی و احتمال حضور آن‌ها در دیگر حوزه ها وکارها مثل ملاقات‌ها و شناسایی گوینده بررسی

[1] shift-invariant probabilistic latent component analysis

[2] Hidden markov model

[3] Matching pursuit

شده است. در آشکارسازی رویداد با سیگنال صوت به عنوان یک سیگنال یک کاناله با یک رویداد در هر زمان رفتار می‌شود [19]،[20].

[19] یک سیستم برای آشکارسازی رویداد دریک فایل ضبط شده از زندگی واقعی ارائه کرده است. رویدادها با استفاده ازیک شبکه مدل مخفی مارکوف، مدل‌سازی شده اند. اندازه و توپولوژی براساس مطالعه تشخیص رویدادهای منفرد درصداهای ضبط‌شده از زندگی واقعی انجام شده است. هم‌چنین آن‌ها تاثیر نویز پیش‌زمینه اضافه شده روی کارایی دسته‌بند رویداد را بررسی کرده‌اند. برای تشخیص رویداد، سیستم عمل تشخیص وجای‌گذاری زمانی یک دنباله از رویدادها را انجام می‌دهد. یک دقت 24٪ از دسته‌بندی رویدادهای صوتی مختلف به 61 کلاس به دست آمده است. این مقدار با دقت دسته‌بندی بین 61 رویداد وقتی نویز با نرخ 0db سیگنال به نویز، به آن افزوده می‌شود، منطبق است .درآشکارسازی رویداد، سیستم قادر است تقریبا یک سوم رویدادها را صحیح تشخیص دهد، و جای‌گذاری زمانی رویداد در 84٪ اوقات صحیح نیست.

[20] یک راه‌حل برای آشکارسازی ومدل‌سازی رویدادهای صوتی که به طورمستقیم زمینه زمانی را توصیف می‌کند، با استفاده از فاکتورگیری ماتریس غیرمنفی کانوالوی (NMF)، ارائه کرده است. NMF برای پیداکردن تجزیه‌های براساس قسمت‌بندی داده مفید است. اینجا این تکنیک برای برای کشف یک مجموعه از پایه های بسته های مکانی-زمانی که به بهترین شکل داده را توصیف کند، استفاده شده است. آن‌ها ویژگی‌ها را با استفاده ازفعال‌سازی این بسته‌ها به دست آورده‌اند و روش خود را با روش پایه که ازویژگی‌های MFCC به دست آمده، مقایسه کرده‌اند و یک سیستم براساس رویداد ساخته اند که درحضور نویز از سیستم پایه MFCC مقاوم‌تر است. درآخر ترکیب دو سیستم از هر دوی آن‌ها بهترکار کرده است.

در کل، تشخیص صحنه‌های صوتی و اتفاقاتی که هم‌زمان می‌افتند، و بنابراین آشکارسازی رویداد چندآوایی(با مجازبودن هم‌پوشانی محدوده رویدادها) ایده‌ال است. به هرحال، رویدادهای برجسته می‌توانند به طور نسبتا پراکنده اتفاق بیفتند و حتی جستجوی رویداد تکی هم ارزش خود را دارد. تعدادی کار درسیسم‌های گسترش داده شده برای آشکارسازی چندآوایی انجام شده است [21] .

در [21] یک سیستم آشکارسازی رویداد صوتی برای محیط‌های چندمنبعه طبیعی، با استفاده از جداسازی منابع پیشنهادشده است. هدف تشخیص‌دهنده، آشکارسازی رویدادهای صوتی از منابع متفاوت هرروزه است. صدا با استفاده ازفاکتورگیری ماتریس غیرمنفی پیش‌پردازش و به چهارسیگنال منفرد تقسیم شده است. هرکلاس رویداد صوتی با یک مدل مخفی مارکوف که با ضرایب طیفی مخفی فرکانس مل آموزش دیده‌اند، بازنمایی شده است. هرسیگنال جداشده به طور جداگانه برای استخراج خصیصه و سپس قطعه‌بندی ودسته‌بندی رویدادهای صوتی با استفاره از الگوریتم ویتربی، استفاده شده است. جداسازی سبب آشکارسازی ماکزیمم

چهار رویداد هم‌پوشان شده است. سیستم پیشنهادشده افزایش قابل توجهی را در دقت آشکارسازی رویداد درمقایسه با خروجی رویدادهای تک دنباله‌ای نشان داده است..

آشکارسازی رویداد شاید بیشتر از دسته‌بندی صحنه مورد تقاضا باشد، اما درعین حال بسیاردرهم تنیده هستند. برای مثال، اطلاعات از دسته‌بندی صحنه می‌تواند اطلاعات زمینه تقویت کننده‌ای را برای آشکارسازی رویداد فراهم کند [22].

[22]، این مساله را بررسی کرده است که چه طور اطلاعات زمینه‌ای می‌تواند در تشخیص خودکار رویداد صوتی استفاده شود. بشر، اطلاعات زمینه‌ای را برای پیش‌بینی دقیق‌تر رویدادهای صوتی و مسلط شدن بر رویدادهای ناخواسته زمینه‌ای، استفاده می‌کند. نتایج، یک بهینه‌سازی مشابه اطلاعات زمینه‌ای در مرحله آشکارسازی رویداد خودکار پیشنهاد کرده است. روش پیشنهادی ترکیب دو مرحله است: مرحله تشخیص زمینه خودکار و مرحله آشکارسازی رویداد. زمینه‌ها با مدل مخلوط گوسی و رویدادهای صوتی با استفاده ازمدل مخفی مارکوف سه حالته چپ به راست مدل شده اند. درمرحله اول، زمینه صوتی سیگنال تست تشخیص داده شده است. براساس زمینه تشخیص داده شده، یک مجموعه کلاس براساس زمینه رویدادهای صوتی برای مرحله آشکارسازی(رویدادصوتی) انتخاب شده است. مرحله آشکارسازی رویداد صوتی نیز، ازمدل‌های صوتی وابسته به زمینه و زوج‌های رویداد براساس شمارش استفاده شده است. دو روش جایگزین آشکارسازی رویداد مطالعه شده اند. دراولین روش، یک دنباله رویداد تک‌آوایی با آشکارسازی برجسته‌ترین رویدادهای صوتی در هر نمونه زمانی با استفاده ازالگوریتم ویتربی خارج شده است. راه‌حل دوم یک روش جدید برای تولید دنباله رویداد چندآوایی با آشکارسازی چند رویداد صوتی هم‌پوشان با استفاده از چند گذر محدود شده ویتربی معرفی کرده است. یک معیار جدید برای کارایی ارزیابی آشکارسازی رویداد صوتی با چندین سطح ازچندآوایی معرفی شده است. این معیار دقت آشکارسازی و خطای درشت اندازه زمانی را دریک معیار، ترکیب کرده ومقایسه کارایی الگوریتم‌های آشکارسازی رویداد را ساده‌تر ساخته است.

روش دومرحله‌ای که برای بهبود نتایج پیدا شد، با سیستم معیار غیروابسته به زمینه مقایسه شده است. درسطح بلوکی، دقت آشکارسازی می‌تواند با استفاده از آشکارساز رویداد وابسته به زمینه پیشنهادشده دوبرابر شود. بسیاری از راه‌حل‌های پیشنهادی دراین زمینه را می‌توان یافت که از آن جمله می‌توان به تکنیک‌های فاکتورگیری اسپکتروگرام اشاره کرد که یک انتخاب منطقی به نظرمی‌رسد.

[23] یک سیستم آنالیز معنایی پنهان احتمالاتی (PLSA) [1] که راه‌حلی نزدیک به NMF است برای آشکارسازی رویدادهای صوتی هم‌پوشان پیشنهاد کرده است. هم‌زمان اتفاق افتادن رویدادها با درجه هم‌پوشانی یک قطعه چندآوایی بازنمایی می‌شود. درمرحله آموزش، PLSA برای یادگیری ارتباطات بین رویدادهای مختلف استفاده شده است. درآشکارسازی، مدل PLSA به طورمداوم احتمالات رویدادها را با توجه به تاریخچه رویدادهای آشکارشده تاکنون، تنظیم کرده است. احتمالات رویداد که با مدل تامین‌شده، در یک سیستم آشکارساز رویداد صوتی جمع‌آوری شده اند. مدل یک بازنمایی بسیارخوب از داده، که سرگشتگی پایینی روی داده‌های تست دارد، پیشنهاد داده است. استفاده از PLSA برای تخمین احتمالات پیشین روی رویدادها دقت آشکارسازی رویداد را تا 35٪ درمقایسه با 30٪ استفاده ازاحتمالات یکنواخت برای رویدادها افزایش می‌دهد. سطوح مختلفی ازافزایش کارایی درزمینه‌های مختلف صوتی وجود دارد که تعداد کمی از زمینه‌ها بهبود قابل توجهی را نشان می‌دهند.

در[20] یک الگوریتم پیچشی NMF، روی ضرایب MFCC اعمال شده است که روی آشکارسازی رویدادهای صوتی ناهم‌پوشان تست شده است. درآخر، تعدادی ازسیستم‌های پیشنهادی که روی آشکارسازی ودسته‌بندی رویدادهای صوتی خاص ازصحنه‌های صوتی مثل گفتار[24]، صدای پرنده [25] ابزارالات موسیقی و دیگرصداهای هارمونیک[26] ، صداهای پورنوگرافی [27] یا رفتارهای پرخطر[28] تمرکز کرده‌اند .

[24] یک تئوری آماری تشخیص گوینده در حضور صحنه‌های دیگر صوتی ارائه کرده است. برخلاف راه‌حل‌های قبلی براساس مدل، چهارچوب پیشنهادی، فرضی درباره نویز پس‌زمینه نمی‌کند، هرچند که این فرض می‌تواند اطلاعاتی را به دست دهد. لازم نیست که مدل منابع پس‌زمینه یا یک تخمین از تعداد آن‌ها را بداند، روش جدید ASR خودکار را با معرفی مدل تفکیکی مدل صوتی ومدل زبانی مرسوم گسترش داده است. درحالی‌که مشکل ASR آماری مرسوم، پیدا کردن شبیه‌ترین دنباله مدل گوینده که یک دنباله مشاهده شده را تولید کرده است، به علاوه راه‌حل جدید شبیه‌ترین مجموعه سیگنال را که سیگنال صوت را تولیدکرده، جستجو می‌کند. روش جدید خطای کلمه را درشرایط نویز ساختگی از 50٪ به 22٪ کاهش داده است.

در [25] بازنمایی پیشنهادشده از قطعه‌بندی زمان فرکانس دوبعدی سیگنال صوت استفاده کرده است که می‌تواند صداهای پرنده را که در طول زمان هم‌پوشانی دارد جدا کند. آزمایشاتی که از داده ای که شامل 13 گونه جمع‌شده با میکروفن‌های گیرنده امواج بدون جهت در جنگل آزمایشی اچ .جی .اندرس استفاده می‌کنند، نشان داده اند که روش پیشنهادی دقت بالایی را به دست آورده است (96.1% درست-غلط صحیح).

[1] probabilistic latent semantic analysis

[26] روشی براساس ویژگی‌های طیفی مکانی و تکنیک‌های ویژگی‌های غایب برای تشخیص صداهای هم‌ساز در سیگنال‌های ترکیبی ارائه داده است. یک الگوریتم تخمین ماسک برای تشخیص محدوده‌های طیفی که شامل اطلاعات قابل اطمینان از هر منبع صوت است، و سپس حاشیه‌سازی محدود به کار گرفته شده برای برخورد با المآن‌های بردار خصیصه که غیرقابل اعتماد ارزیابی شده‌اند، پیشنهاد شده است. روش ارائه شده روی صداهای ابزارآلات موسیقی با توجه به دسترسی وسیع داده آزمایش شده است، اما می‌تواند روی صداهای دیگر نیز به کار برده شود (مثلا صداهای حیوانات، صداهای محیط)، چرا که اینها هم هم‌ساز هستند. در شبیه سازی‌ها روش پیشنهادی به وضوح بهتر از روش معیار برای سیگنال‌های ترکیبی عمل کرده است.

مساله چندآوایی، مربوط به هردو حیطه توضیح داده شده است. چراکه کلا صحنه‌های صوتی چندآوایی (چندمنبعی) هستند. همین‌طور راجع به موسیقی، ممکن است که آنالیزهایی روی سیگنال صوتی درکل بدون توجه به چندآوایی انجام دهیم، با این وجود ممکن است درنظرگرفتن مراجع اجزا سیگنال صوت سودمند باشد. این آنالیز براساس اجزائی مشابه *جریان شنیداری* است که در مدل برگمن از شنوایی انسان اتفاق افتاده است [8]. درکاربردهای تشخیص گفتار اغلب می‌توان فرض کرد که یک منبع غالب وجود دارد که برای تحلیل باید روی آن تمرکز شود [24]، اما این فرض درمورد صحنه‌های کلی صوتی صادق نیست. یک راهبرد که سیگنال‌های چندآوایی را مدیریت می‌کند، جداسازی منابع صوتی و تحلیل هرکدام ازمنابع به طورجداگانه است [29]،[21].

در [29] یک پژوهش روی آنالیز محاسباتی صحنه‌های شنیداری برای به دست آوردن تعامل بین انسان و ربات با تشخیص اطلاعات شنیداری انجام شده است. هدف این پژوهش فهم یک صدای ترکیبی دلخواه شامل صداهای غیرگفتاری و موسیقی به خوبی گفتار صادار است که توسط گوش‌های ربات (یا میکروفن‌هایی که در ربات جای‌گذاری شده) به دست می‌آید. موضوعات اصلی مکان‌یابی منبع صدا، جداسازی و تشخیص در سطوح پردازش سیگنال و تبدیل سیگنال به نماد در سطح رابط با سطوح پردازش نماد است. مورد دوم، در جامعه توسعه‌یافته مهم است چرا که آن‌ها در حال توسعه یک سیستم تشخیص خودکار صداواژه هستند. این مقاله یک مرور از شنوایی روبات به ویژه فیلترهای مسیرگذرفعال (ADPF)[1] که منابع صوتی را که ازیک مسیر خاص سرچشمه می‌گیرند، ازطریق انتگرال‌گیری مکان‌یابی منبع صوت و پردازش بینایی جدا می‌کند. ADPF روی سه نوع از روبات‌ها اجرا شده جداسازی و تشخیص سه گفتار با یک جفت میکروفون را نشان داده است.

[1] active direction-pass filter

به هرحال، با توجه به اینکه مدل‌سازی محاسباتی جریان صوتی لزوما نیاز به بازسازی سیگنال‌های صوت جدا ازهم ندارد- برگمن ادعا نکردکه انسان‌های شنونده چنین کاری انجام می‌دهند- می‌تواند با یک بازنمایی سطح میانی مثل مدل چندمنبعی احتمالاتی کار کند [30].

در [30]، یک کار استنتاجی که درآن یک مجموعه ازمشاهدات رویداد مهرزمان شده که باید دریک تعداد نامعلوم دنباله‌های زمانی بانرخ‌های مستقل و متغیر مشاهدات، دسته‌بندی شوند، انجام شده است. راه‌حل‌های موجود مختلف برای ردیابی چند شی فرض می‌کنند که یک تعداد مشخصی از منابع و یک نرخ ثابت مشاهدات وجود دارد. دراین مقاله یک راه‌حل را برای استنتاج کردن ساختار در داده مهرزمان شده تولیدشده با ترکیب از روند مشابه تکراری مارکوف، گسترش داده شده است. استنتاج به طور هم‌زمان سیگنال را از نویز تشخیص می‌دهد و مشاهدات سیگنال را در یک جریان منبعی جدا دسته‌بندی می‌کند. آن‌ها تکنیکی را ازطریق آزمایشات معنایی توضیح داده اند که یک آزمایش ترکیب آواز پرندگان را ردیابی می‌کند.

جداسازی منبع برای یک صدای همه منظوره هنوز راه طولانی تا حل کامل مساله دارد[31].در[31] نتیجه سه جنبش اخیر ارزیابی حوزه جداسازی منبع صوتی و پزشکی ارائه شده است. این جنبش‌ها گواه یک رونق در حوزه کاربردهای سیستم‌های جداسازی منبع درسال‌های اخیر هستند، همان‌طورکه درتعداد رو به افزایش مجموعه داده از 1 به 9 و تعداد رو به افزایش مقالات ثبت‌شده از 15 به 34 نشان داده می‌شود. آن‌ها ابتدا روی تاثیر تعریف روش ارزیابی مرجع به همراه پایگاه‌های داده و نرم‌افزارها بحث کرده اند. سپس نتایج کلیدی که تقریبا درتمام مجموعه داده های به دست آمده ارائه داده اند. درآخر با پیشنهادکردن مسیرها برای تحقیقات آینده وارزیابی‌ها، براساس خصوصیات ایده‌هایی که درپنل مباحثه در 19 امین کنفرانس جداسازی سیگنال و آنالیز متغیرهای پنهان (LVA/ICA 2010)[1] مطرح شد، نتیجه‌گیری کرده اند.

برای مثال، ارزیابی که در چالش اخیر برای "تشخیص گفتار در یک محیط چندمنبعی" انجام گرفته است، ازالگوریتم‌های ثبت‌شده این انتظار را نداشت که جداسازی منابع را انجام دهند: ارزیابی روی خروجی رونوشت‌شده گفتار انجام گرفته است. الگوریتم های ثبت‌شده شامل مرحله جداسازی منبع نبوده اند، خیلی ازآن‌ها از سرکوب نویز مکانی یا طیفی برای تمرکز برمنبع استفاده کرده اند تا جداسازی همه منابع [32].

[1]Latent Variable Analysis

[32] نتایج دومین چالش 'CHiME' را گزارش داده است. یک ابتکار که برای ارزیابی کارایی سیستم‌های ASR درمحیط‌های داخلی انجام داده اند. آن‌ها منطق چالش‌ها را بررسی کرده اند و سپس یک خلاصه از سیستم‌های معیار وکارها و مجموعه داده فراهم کرده اند. مقاله سیستم‌هایی را که در حوزه چالش واژگان کوچک با گوینده‌های متفاوت و واژگان متوسط با گوینده ثابت است را مرور کرده است.

درشنوایی ماشین، ارزیابی عمومی و معیارگذاری سیستم‌ها نقش مهمی را ایفا می‌کند. این ارزیابی، موجب مقایسه هدف‌مند بین انواع سیستم‌های پیشنهادشده شده، و هم‌چنین می‌تواند برای مطالعه بهبود کیفیت در طول زمان استفاده شود. بسیاری ازاین قبیل چالش‌ها درزمینه گفتار برگزار شده است. برای مثال، ارزیابی رونویسی قدرتمند DARSA EARS (2002-2009) روی کارهای قطعه‌بندی گفتار، روی اخبار عمومی و همین‌طور ملاقات‌های ضبط‌شده اعمال شده است [18]. چالش MIREX[1] (2005 تاکنون) سیستم های MIR برای کارایی‌شان روی کارهای موسیقی ویژه مثل رونویسی ملودی یا ردیابی ریتم [33]. چالش SiSEC (2007 تا کنون) روی الگوریتم‌های جداسازی منابع تمرکز دارد.

[1] The music information retrieval Evaluation exchange

نتیجه‌گیری

در این سمینار ما به بررسی تحقیقات انجام شده در حوزه دسته‌بندی صحنه‌های صوتی با توجه به چالش DCASE پرداختیم که نشانه‌گذاری سیگنال‌های صوتی را نیز پوشش می‌دهد. هدف از چالش DCASE شکل‌دهی یک مجموعه از کارهای همه‌منظوره شنوایی ماشین برای صداهای روزمره جهت محک‌زدن پیشرفته‌ترین روش‌ها و رشددادن جامعه تحقیقاتی در حوزه گفتار و موسیقی است. برای دسته‌بندی صحنه، سیستم‌ها نتایجی بهتر از سیستم‌های معیار به دست آوردند. با این وجود هنوز روش‌های زیادی برای ارتقای پیشرفته‌ترین سیستم‌ها وجود دارد. بهترین آن‌ها بزرگ‌تر کردن اندازه مجموعه داده برای نتیجه‌گیری بهتر در زمینه مقایسه کارایی سیستم‌هاست. یکی دیگر از راهکارها جداسازی منابع صداهای چندمنبعی است. هم‌چنین جداسازی نویز پس‌زمینه تاثیر زیادی در بهبود کارایی دارد. از انواع کاربردهای این چالش می‌توان به مسیریابی روبات، رونویسی موسیقی، تشخیص خواننده موسیقی، ابزارهای کمکی شنوایی و گوشی‌های موبایل می‌توان اشاره کرد.

مراجع

[1] L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition.Upper Saddle River, NJ, USA: PrenticHall, Inc., 1993.

[2] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, “The PASCAL CHiME speech separation and recognition challenge,” Computer Speech & Language, 2012.

[3] E. Benetos and S. Dixon, “Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model,” Journal of the Acoustical Society of America, vol. 133, p. 1727, 2013.

[4] A.Wang, “An industrial strength audio search algorithm,” in Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR ’03), pp. 7–13, Oct 2003.

[5] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, “Localization of simultaneous moving sound sources for mobile robot using a frequencydomain steered beamformer approach,” in Proceedings of the 2004 IEEE International Conference on Robotics and Automation, vol. 1, pp. 1033–1038, IEEE, 2004.

[6] R. Ranft, “Natural sound archives: Past, present and future,” Anais da Academia Brasileira de Ci ̂encias, vol. 76, no. 2, pp. 456–460, 2004.

[7] D. L. Wang and G. J. Brown, eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. New York: IEEE Press, 2006.

[8] A. S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, 1994.

[9] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events: an IEEE AASP challenge,” in Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),2013.

[10] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urbansoundscapes but not for polyphonic music," *J. Acoust. Soc. America*,vol. 122, no. 2, pp. 881–891, 2007.

[11] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA, USA: MorganKaufmann, 2005.

[12] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Process. Mag.*, vol. 23,no. 2, pp. 133–141, Mar. 2006.

[13] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J.Acoust. Soc. America*, vol. 51, pp. 2044–2056, Jun. 1972.

[14] J. Foote, "Content-based retrieval of music and audio," in *Proc. SPIE,Multimedia Storage Archiving Syst. II*, 1997, vol. 3229, pp. 138–147.

[15] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classification," M.S. thesis, ATIAM, ParisTech, Paris, France, Aug. 2011.

[16] E. Benetos, "Automatic transcription of polyphonic music exploiting temporal evolution," Ph.D. dissertation, School of Electron. Eng. And Comput. Sci., Queen Mary University of London, London, U.K., Dec.2012.

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

[18] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[19] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. Eur. Signal Process. Conf.*, Aug. 2010, pp. 1267–1271.

[20] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Proc. IEEE Workshop Appl. of Signal Process. Audio Acoust.* Oct. 2011, pp. 69–72.

[21] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Proc.Workshop Mach. Listening Multisource Environ.*, 2011, pp. 36–40.

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

[23] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection," in *Proc. Eur. Signal Process. Conf.*, 2011, pp.1307–1311.

[24] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Commun.*, vol. 45, no. 1, pp. 5–25, 2005.

[25] F. Briggs and B. Lakshminarayanan *et al.*, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach,"*J. Acoust. Soc. America*, vol. 131, pp. 4640–4650, 2012.

[26] D. Giannoulis, A. Klapuri, and M. D. Plumbley, "Recognition of harmonic sounds in polyphonic audio using a missing feature approach,"in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013,pp. 8658–8662.

[27] M. J. Kim and H. Kim, "Automatic extraction of pornographic contents using radon transform based audio features," in *Proc. 9th Int.Workshop Content-Based Multimedia Indexing*, Jun. 2011, pp. 205–210.

[28] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," *EURASIP J. Audio, Speech Music Process.*, vol. 2009, 2009, Art. ID 13.

[29] H. G. Okuno, T. Ogata, and K. Komatani, "Computational auditory scene analysis and its application to robot audition: Five years experience," in *Proc. 2nd Int. Conf. Informat. Res. Develop. Knowledge Soc. Infrastructure*, Jan. 2007, pp. 69–76.

[30] D. Stowell and M. D. Plumbley, "Segregating event streams and noise with a Markov renewal process model," *J. Mach. Learning Res.*, vol.14, pp. 1891–1916, 2013.

[31] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A.Ozerov, V. Gowreesunker, D. Lutter, and N. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Process.*, vol. 92, no. 8, pp. 1928–1936,2012.

[32] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M.Matassoni, "The second 'CHiME' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *Proc.IEEE Workshop Automat. Speech Recog. Understand.* Dec. 2013, pp.162–167.

[33] J. S. Downie, A. F. Ehmann, M. Bay, and M. C. Jones, "The musicinformation retrieval evaluation eXchange: Some observations and insights,"in *Advances in Music Information Retrieval*, ser. Studies inComputational Intelligence. New York, NY, USA: Springer, 2010 vol.274, pp. 93–115.

[34] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange and Mark D.Plumbley, "Detection and Classification of Acoustic Scenes and Events," IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 17, NO. 10, OCTOBER 2015.

[35] Daniele Barchiesi,Dimitrios Giannoulis,Dan Stowell, andMark D. Plumbley, "Detection and Classification Acoustic Scenes and Events," IEEE Signal Processing Magazine 32(3)(May2015) 16-34