بسمه‌تعالی

دانشگاه سمنان

دانشکده برق و کامپیوتر

سمینار کارشناسی ارشد

مهندسی کامپیوتر- گرایش هوش مصنوعی

عنوان سمینار:

پردازش صوت و موسیقی با استفاده از شبکه های عصبی و یادگیری عمیق

Sound and Music Processing By Using Neural Network and Deep Learning

توسط:

نگار رضائی

۹۵۱۱۹۲۰۰۰۶

استاد راهنما:

دکتر رحمانی‌منش

آذر ۹۶

چکیده

این سمینار به بررسی دو بخش، روش های پردازش سیگنال موسیقی و دسته‌بندی آن به وسیله ی شبکه ی عصبی[1] و یادگیری عمیق[2] می پردازد. امروزه به دلیل حجم گسترده اطلاعات موسیقیایی، پردازش و دسته‌بندی آنها در اولویت داده کاوی[3] و بازیابی اطلاعات[4] قرار گرفته است. دسته‌بندی سیگنال موسیقیایی در مواردی همچون تفکیک ژانرها، دسته‌بندی سازها، تفکیک مود ها و... کاربرد دارد. در این حوزه چالش هایی نظیر چه ویژگی هایی استخراج شوند، پیچیدگی محاسباتی ابعاد ویژگی استخراج شده چگونه بهبود یابند،کدام دسته بند مناسب تر است و دقت بالاتری دارد وجود دارد. در این سمینار روش شبکه عصبی و یادگیری عمیق به منظور دسته‌بندی مورد بررسی قرار گرفته و از آنجایی‌که سیگنال موسیقی خود نوعی سیگنال صوتی است، تحقیقات انجام شده در زمینه دسته بندی سیگنال های صوتی نیز، مورد مطالعه قرار گرفته است.

[1] Neural Network
[2] Deep Learning
[3] Data Mining
[4] Information retrieval

فهرست مطالب

فصل اول

استخراج ویژگی های صوتی

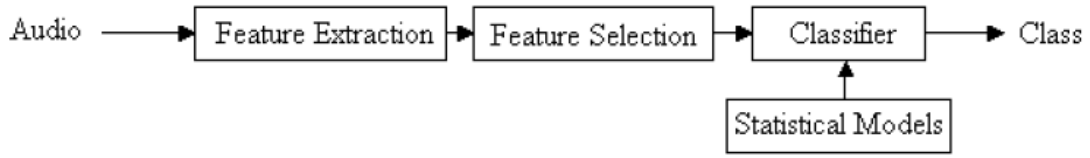
۱-۱ مقدمه

با توجه به رشد و توسعه اینترنت و تکنولوژی اطلاعات در سال های اخیر، بازیابی اطلاعات فایلهای صوتی [۷] و سازمان‌دهی آن بر اساس نوع صوت برای اهداف گوناگون به چالشی مهم در داده کاوی تبدیل شده است [۸]. داده های صوتی اغلب شامل بخشهای متنوعی شامل گفتار و موسیقی هستند [۱۴]. بنابراین تفکیک این بخشهای گوناگون یکی از چالشهای بنیادین در پردازش صوت می باشد [۴] .

هدف پردازش سیگنال های صوتی شامل دسته بندی اصوات موسیقی،دسته بندی نوع موسیقی ، آوا نویسی ابزار موسیقی، تقسیم بندی موسیقی، تشخیص گوینده، تشخیص زبان، بازیابی صدا و تشخیص مفهوم... می باشد. در این راستا دسته بندی موسیقی از لحاظ جغرافیایی، تاریخی، ارزشی، نوع ساز، تئوری موسیقی و نوع کاربرد انجام می شود.

استخراج ویژگی اولین گام در بازیابی اطلاعات صوتی و دسته بندی است. استخراج ویژگی از سیگنال خام به دلیل حذف نوفه داده‌ها، فروکاهی ابعاد، افزایش بعد، جداسازی اجزای مستقل داده ها انجام می شود. نوع ویژگی در تجزیه و تحلیل سیگنال صوت از آن جهت حائز اهمیت است که برخی از ویژگی ها در محیط‌های نویزی نتیجه بهتری می دهند، برخی از ویژگی ها حجم محاسبات کمتری را می طلبند، برخی از ویژگی ها در کل سیگنال صوت و برخی در یک پنجره از سیگنال صوتی محاسبه می شوند. برخی از ویژگی ها یک عدد و برخی بردار هستند.

سیستم های دسته بندی موسیقی از مدل آنالیز بر اساس فریم پیروی می کنند. مشکل اصلی در بازیابی اطلاعات موسیقی، دسته بندی موسیقی بر پایه صدا است. یک آهنگ کلی به فریم هایی که هر بردار ویژگی از آن استخراج می شود تقسیم می شود. سیگنال صوتی و موسیقی را می توان به وسیله ی یک مجموعه بردار ویژگی نشان داد. شکل ۱ مراحل کلی دسته بندی سیگنال موسیقی را نشان می دهد.



شکل ۱.مراحل کلی دسته بندی صدای موسیقیایی

مطالعات پیشین از ویژگی ها و مدل های محتمل که محاسبات عددی بسیار پیچیده ای برای حل دارند استفاده کرده اند. در سال های اخیر علاقه به استفاده از یادگیری ویژگی ها و معماری عمیق افزایش یافته است،که در نتیجه از محاسبات پیچیده می کاهد و دقت بالاتری را نشان می دهد[۱۰].

مفاهیم اولیه ویژگی های صوتی در موسیقی

۲,۱,۱. نواک[1]

به **زیروبمی، زیرایی صدا** که باعث تفکیک یک صدا از صدای دیگر می شود نواک گویند. برای مثال تفاوت صدای کودکان و بزرگسالان و یا صدای آقایان و بانوان نواک را تعریف می کند. به بیانی دیگر به فرکانس های بالا و پایین تولید شده به وسیله ی تعداد لرزش (ارتعاش) از یک وسیله صدایی مانند زه گفته میشود.

صدایی که شنیده می شود شامل فرکانس پایه و سری های هارمونیک است. به بیان دیگر پیچ، فرکانس اصلی صدا است . با توجه به ادراک انسان تعداد ارتعاشات بیشتر شامل فرکانس های بالاتر و صدای روشن تر خواهد بود.

محتوای نواکی می‌تواند ویژگی مفیدی برای تشخیص سبک ارائه دهد. ویژگی های محتوای زیر و بمی صدا با محاسبه هیستوگرام نواک بدست می آید. اصولا در محاسبه هیستوگرام نواک در ویژگی هایی که شامل چندین ملودی هماهنگ هستند، از الگوریتم شناسایی نواک چندگانه استفاده می شود[۳]. هیستوگرام نواک یا پیچ، نمایش آماری کلی از محتوای تغییرات فرکانسی یک قطعه موسیقی ارائه می دهد و ویژگی های استخراج شده ی حاصل از آن می تواند در دسته‌بندی موسیقی مورد استفاده قرار گیرد.

[1] Pitch

۲,۱,۲.دیرند[1]

به **کشش صدا** و میزان **امتداد** صدا دیرند گفته می شود. اینکه صدا به چه میزان، از واحد زمان به گوش می رسد و چه زمانی پایان می یابد دیرند نامیده می شود.

۲,۱,۳. رنگ[2]

طنین یا شیوش به شخصیت یک صدا گفته می شود. ویژگی های شیوشی ویژگیهایی هستند که برای تمایز مابین گفتار و موسیقی مورد استفاده قرار می گیرند که امکان تشخیص تفاوتهای بین صدا های دارای فرکانسهای نزدیک که با آلات موسیقی مختلف و یا افراد مختلف تولید می شود را ایجاد می کند [۹] .

ساز ها از مواد مختلفی تولید می شوند به همین دلیل ارتعاشات متفاوتی تولید می کنند که همین ارتعاشات متفاوت خود هارمونیک ها و فرکانس اصلی ساز را ایجاد می کند. از آنجا که هر ژانر موسیقی از سازهای مختلف تشکیل شده است، هر ژانر ویژگی های رنگی (شخصیتی) متمایز را نشان می دهد[۱۳].

شیوش در موسیقی با تعداد و قدرت نسبی هارمونیک های تولید شده توسط آلات موسیقی مختلف در ارتباط است [۸]. به منظور استخراج ویژگی های شیوشی، لازم است تا موسیقی به فواصل زمانی کوچک تقسیم شده و به بعد فرکانس برده شود تا از وابستگی زمانی خارج شده تا ثابت و بدون تغییر در نظر گرفته شود[۱].

۲,۱,۴. شدت[3]

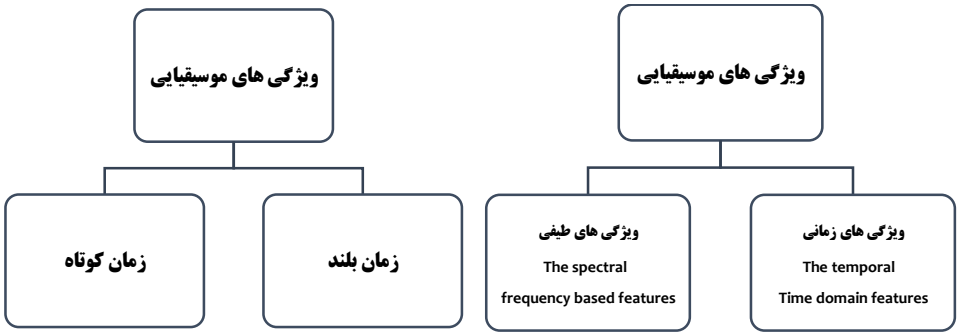
در آکوستیک شدت به بلندی صدا و میزان صدا و یا انرژی برمی گردد. شدت واضح ترین ویژگی است. به طور معمول در یک فریم در یک دامنه ی زمانی اندازه گیری می شوند . دامنه های بالاتر دارای شدت بالاتر هستند. با واحد دسی بل نمایش داده می شود. در موسیقی ژانر های دارای صدای قوی (مانند ژانر هوی متال)دارای شدت بالاتر هستند . برای محاسبه شدت محلی یک فریم از مربع میانگین ریشه دامنه RSM[4] استفاده می شود.

۲,۲. بازیابی اطلاعات از سیگنال موسیقی

برای بازیابی اطلاعات از سیگنال موسیقی، می توان دو دسته کلی ویژگی استخراج کرد. دسته ی اول ویژگی هایی که بر پایه ی دامنه

[1] Duration
[2] Timbre
[3] Dynamics
[4] root mean square

(زمانی / فرکانسی) بدست می آیند. دسته دوم شامل ویژگی هایی هستند که بر پایه ی مدت زمان سیگنال صوتی استخراج می شوند.شکل شماره ۲ نشان دهنده دسته بندی کلی ویژگی های صوتی موسیقیایی است.



شکل ۲. دسته بندی ویژگی های موسیقیایی

۲,۲,۱. ویژگی های زمانی

ویژگی های زمانی[۱] بر پایه ی دامنه ی زمانی استخراج می شوند. این دسته از ویژگی ها از دینامیک نسبتا بلندمدت یک سیگنال موسیقی، انتقال زمانی و ویژگی های ریتمیک بدست می آیند. ویژگی هایی شامل نرخ عبور از صفر[۲]، توسعه زمانی[۳] و هیستوگرام سرعت اجرا[۴] شامل ویژگی های زمانی می باشند.

۲,۲,۱,۱. نرخ عبور از صفر

نرخ عبور از صفر زمانی رخ می دهد که نمونه های پی در پی سیگنال دیجیتال، دارای علامت های مختلف (مثبت/ منفی) باشند[۶]. نرخ عبور از صفر میزان نویزی بودن سیگنال را بررسی می کند و بر خلاف ویژگی های طیفی یک ویژگی حوزه زمان است. منحنی نرخ عبور از صفر موسیقی به دلیل اینکه در یک دوره زمانی مشخص نسبت به گفتار پایدار تر است، تغییرات دامنه کمتری دارد. البته این موضوع در گویش های مختلف متفاوت است .

۲,۲,۱,۲. توسعه زمانی

هر سیگنال در طی گذر زمان با تغییرات فرکانسی همراه است. این تغییرات توسط نمودار توسعه زمانی نشان داده می شود. در این نمودار تغییرات بر اساس فرکانس در واحد زمان مشاهده می شود .

۲,۲,۱,۳. هیستوگرام سرعت اجرا

در ابتدا به وسیله ی تمپوگرام ، میزان سرعت اجرا بر حسب ضرب در ثانیه بدست می آید سپس هیستوگرام کلی سیگنال موسیقی بر حسب ضرب در دقیقه رسم می شود. پیک بدست آمده در هیستوگرام نمایانگر بالاترین سرعت اجرای موجود در سیگنال موسیقی است .

[1]The temporal features
[2] Zero-crossing rate (ZCR)
[3] Temporal envelope
[4] Tempo histogram

۲,۲,۲. ویژگی های طیفی

ویژگی های طیفی [1] بر پایه ی دامنه ی فرکانسی، استخراج می شوند.تبدیل سیگنال موسیقی به یک شکل تصویری برای انجام سایر پردازش ها طیف سیگنال را در اختیار قرار میدهد. تغییراتی که بر اساس ویژگی های مختلف سیگنال در شکل ایجاد می شود سبب ایجاد تفاوت مابین سیگنال های مختلف می شود. این دسته از ویژگی ها وابسته به زمان کوتاه هستند و از فریم هایی کوتاه از سیگنال صوتی استخراج می شوند. ضرایب طیفی یا کپسترال از فرمول شماره ۱ محاسبه می شوند.

$$C(k) = IDFT\{log|DFT\{x(n)\}|\} \quad (1)$$

از M ضریب اول $Cepstral$ به عنوان ویژگی استفاده می شود. در صورت استفاده از کلیه ضرایب، طیف به صورت دقیق به دست می آید. دقت مدل سازی با توجه به تعداد ضرایب تعیین می شود. هرچه تعداد ضرایب مورد استفاده بیشتر باشد دقت بالاتر می رود.

درک ویژگی های طیفی را می توان به صورت رنگ یا درون مایه در موسیقی بیان کرد. ویژگی های شامل مرکز ثقل طیفی[2]، پخش طیفی[3] ، شار طیفی[4]، میزان صافی طیفی[5]، ضرایب فرکانسی مل[6] و کما[7] شامل ویژگی های طیفی می باشند .

۲,۲,۲,۱. مرکز ثقل طیفی

این ویژگی به عنوان نقطه تعادل دامنه طیفی معرفی می شود. وضوح طیفی را اندازه گیری می کند. هر چه مقدار مرکز ثقل طیفی بالاتر باشد بافت سیگنال موسیقی واضح تر می باشد. در این صورت سیگنال موسیقی فرکانس های بالای بیشتری دارد [۲].فرمول شماره ۲ نشان دهنده محاسبه مرکز ثقل طیفی است .

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \quad (2)$$

۲,۲,۲,۲. Roll-off طیفی

معیار کجی شکل طیفی است. با R معرفی می شود، هر چه R بزرگتر باشد صدا واضح تر است. این ویژگی شامل فرکانس هایی است که بیش از ۸۵ درصد انرژی ، در آنها ذخیره شده باشد[۱۳]. در فرمول شماره ۳ نحوه بدست آمدن $Roll-off$ مشاهده می شود.

$$\sum_{k=1}^{M} |X_r[k]| = 0.85 \sum_{k=1}^{N/2} |X_r[k]| \quad (3)$$

۲,۲,۲,۳.شار طیفی

به صورت مربع تفاوت بین دامنه های نرمالیزه شده ی توزیع طیفی متوالی تعریف می شود. شارطیفی اندازه گیری مقدار تغییرات محلی طیفی است[۵]. فرمول ۴ نمایانگر استخراج ویژگی شار طیفی است.

---

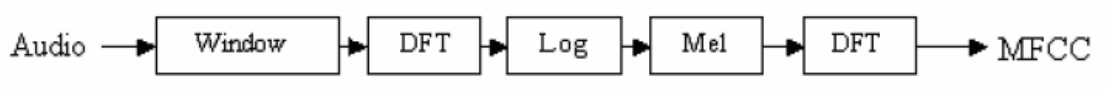
[1] The temporal
[2] Spectral Centroid
[3] Spectral spread
[4] Spectral flux
[5] Spectral flatness measure
[6] Mel-frequency cepstral coefficients (MFCCs)
[7] Chroma

$$F_r = \sum_{k=1}^{N/2} (|X_r[k]| - |X_{r-1}[k]|)^2 \quad (4)$$

۲,۲,۲,۴.ضرایب طیفی فرکانس مل [۱و۵و۷و۱۲]

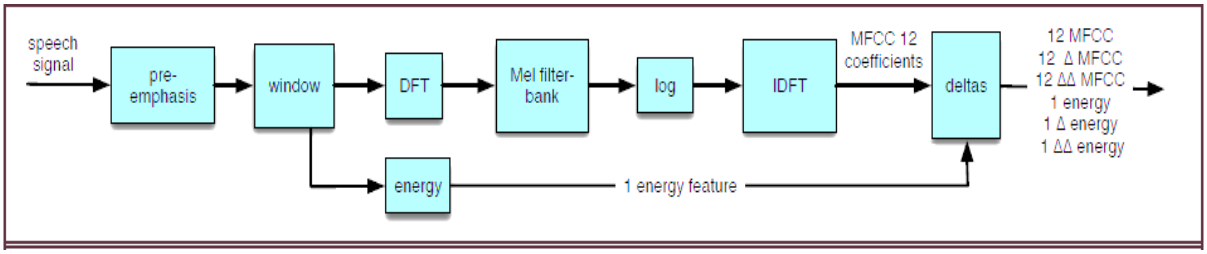
درک انسان از محتوای فرکانسی از یک توزیع لگاریتمی پیروی می‌کند. همانگونه که در فرمول ۵ نشان داده شده است، ضرایب طیفی فرکانسی مل نمایش فشرده شده ای از طیف ویژگی صوتی بر مبنای درک لگاریتمی گوش انسان از صدا است که با مقیاس مل نشان داده می شود.شکل ۳ نمایانگر شمای کلی استخراج ویژگی MFCC است.

$$Mel\,(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right) \quad (5)$$



شکل ۳. مراحل کلی استخراج ویژگی MFCC

ضرایب فرکانسی مل معمولا در سامانه های تشخص گفتار مورد استفاده قرار می گیرند و نتایج پژوهش ها نشان داده است که این ویژگی نسبت به سایر ویژگی ها بهتر می تواند بخش ادراکی طیف را نمایش دهد. مخصوصا این ویژگی در سامانه تشخیص سبک موسیقی با موفقیت استفاده شده است. این ضرایب نوع بهبود یافته از ضرایب cepstral می باشند. به دلیل وجود درجه وضوح بالای این ویژگی تغییرات جزئی طیف صوتی کاملا دیده می شود. MFCC باعث خلاصه سازی داده ها می شود، همبستگی بین ویژگی ها را از بین می برد، عملیات دسته بندی را بهبود می بخشد و یک بردار پیوسته از ویژگی ها را بدست می آورد. شکل شماره ۴ نشان دهنده مراحل استخراج ۳۹ ویژگی از طریق ضرایب فرکانسی مل است.



شکل ۴. مراحل استخراج ۳۹ ویژگی MFCC

همانطور که در جدول شماره۱ مشاهده می شود از طریق این ویژگی می توان ۳۹ ویژگی از سیگنال سیگنال استخراج کرد.

جدول 1. ۳۹ ویژگی MFCC

13	Absolute	Energy (1) and MFCCs (12)
13	Delta	First-order derivatives of the 13 absolute coefficients
13	Delta-Delta	Second-order derivatives of the 13 absolute coefficients
39	Total	Basic MFCC Front End

۲,۲,۳.ویژگی های زمان کوتاه

ویژگی های زمان کوتاه ویژگی هایی هستند که می توانند بر پایه دامنه زمانی یا دامنه فرکانسی بدست بیایند، تنها تفاوت آنها این است که از درون فریم های تقسیم شده از یک سیگنال بزرگ کشف و استخراج می شوند. این ویژگی ها بر مبنای کمیت های مشتق شده از طیف یک سیگنال در یک بازه زمانی کوتاه به شکل قاب بدست می آید. ویژگی های زمان کوتاه شامل ضرایب فرکانسی کپسترال مل MFCC ، مرکز ثقل طیفی و نرخ عبور از صفراست .

۲,۲,۴. ویژگی های زمان بلند

ویژگی های زمان بلند بر خلاف ویژگی های زمان کوتاه از کل سیگنال استخراج میشوند. این ویژگی ها شامل تغییرات شکل طیفی در طول سیگنال موسیقی، اکسترمم های سیگنال در یک بازه زمانی بزرگ، ضرایب موجک[1]و ویژگی انرژی پایین[2] می باشند. در بازیابی اطلاعات موسیقی از هر دو ویژگی زمان کوتاه و زمان بلند استفاده می شود .

۲,۲,۴,۱. ویژگی انرژی پایین

ویژگی انرژی پایین از ویژگی های زمان بلند است. این ویژگی میزان انرژی را در طول کل سیگنال بررسی می کند و برخلاف سایر ویژگی ها بر مبنای پنجره بافت است. ویژگی انرژی پایین به صورت درصد بیان می شود. این ویژگی فریمی که انرژی کمتری در کل سیگنال نسبت به سایر فریم ها دارد را متمایز می‌کند [۶]. به عنوان مثال یک موسیقی آوازی که همراه با سکوت است مقدار انرژی پایین بیشتری به یک رشته موسیقیایی پیوسته دارد [۱۴].

۲,۲,۵.ویژگی های ریتمیک

گرفتن ریتم برای نوازندگان و آهنگسازان به صورت ضربه زدن با پا، کار بسیار آسانی است، اما توضیح این عمل برای سیستم های اتوماتیک کار بسیار دشواریست. ساختار ریتمیک و وزنی یک آهنگ یک مشخصه بسیار عالی برای شناخت سبکهای مختلف موسیقی از یکدیگر است [۱]. دوره تناوب پایه ی شناسایی ویژگیهای ساختار ریتم است. برای این منظور در ابتدا با استفاده از تبدیل موجک گسسته بخش انتخابی به تعدادی باند فرکانسی اکتاوی تجزیه می شود و بدنبال آن قله‌ی دامنه‌ی حوزه‌ی زمانی

[1] DWCH: Daubechies Wavelet Coefficient Histograms
[2] RSM

در هر باند به صورت جداگانه استخراج می گردد. سپس بخش متوسط حذف می گردد و قله‌ی هر باند با همدیگر جمع می‌شود. پس از آن همبستگی [1]حاصل از مجموعه ی قله ها محاسبه می گردد. تابع همبستگی قله‌های غالب به صورت نمودار ستونی ضرب انباشته می‌شود که هر ستون متناظر با دوره تناوب بر حسب ضرب در دقیقه می‌باشد [۱۱].

لازم به ذکر است،بر اساس موارد ذکر شده ویژگی های زیادی استخراج می شوندکه بار محاسباتی بالایی دارند. بنابراین برای کاهش ابعاد ویژگی های استخراج شده از روش هایی شامل PCA و LDA،... استفاده می شود .

۲,۳.فریمورک های استخراج ویژگی

به تازگی، محققان چندین فریمورک برای استخراج ویژگی های مفید پردازش سیگنال موسیقی توسعه داده اند. که شامل چهار فریمورک مارسیاس ،MPEG-7، MIRToolbox، jAudio می باشد .

۲,۳,۱. فریمورک مارسیاس[2]

توسط Tzanetakis و Cook برای کمک به دانشجویان و محققین ایجاد شده و برای پروژه های تحقیقاتی دانشگاهی و صنعتی مورد استفاده قرار گرفته است. مارسیاس یک فریمورک اپن سورس بوده است که برای پردازش صوت با تاکید مخصوص بر روی بازیابی اطلاعات موسیقی مورد استفاده قرار می گیرد. شامل ۳۰ ویژگی توصیفی ریتم، فرکانس و رنگ برای دسته بندی موسیقی است .

۲,۳,۲.فریمورک MPEG-7

صدای توصیفی توسعه داده شده با استاندارد MPEG ، ۱۷ ویژگی توصیفی صوت شامل شدت، فرکانس و رنگ را برای بازیابی اطلاعات موسیقیایی فراهم می کند.

۲,۳,۳ .MIRToolbox

توسط Lartillot و Toiviainen در نرم افزار متلب و نرم افزار jAudio توسط McEnnis و همکارانش توسعه داده شده است. که این دو فریمورک چهارچوبی از ویژگی های یکپارچه از پنج ویژگی آکوستیک، از جمله ویژگی های توصیفی موجود در Marsyas و MPEG-7 ارائه می دهند.

[1] correlation
[2] Marsyas

فصل دوم

مطالعات پیشین

لیدی و رابر [۱] در پژوهشی با مورد اهمیت قرار دادن تحولات روانشناختی برای استفاده صحیح ویژگیهای صوتی برای دسته بندی ژانر های موسیقی از توصیفگرهای طیف سنجی و ویژگی های هیستوگرام ریتم استفاده کردند. ارزیابی بر روی هر دو مجموعه ویژگی به صورت تکی و ترکیبی از طریق یک الگوریتم دسته بندی ژانر موسیقی، بر روی سه دیتاست GTZAN ISMIRrhythm, ISMIRgenre نشان داد که با استفاده از این دو ویژگی می توان دسته بندی دقیقی را در موسیقی انجام داد .

در مرحله پیش پردازش سه گام طی شده است. در گام اول پیش پردازش، داده صوتی از فرمت wav یا mp3 به داده خام دیجیتالی صوتی تبدیل شده است. در گام دوم تعداد کانال های داده صوتی بدست می آید که به طور متوسط هر داده صوتی دارای یک کانال می باشد. در گام سوم ۶ ثانیه از داده صوتی استخراج شده است، از بخش های صوتی استخراج شده تبدیل سریع فوریه با پنجره بندی ۲۳ میلی ثانیه ای و ۵۰ درصد همپوشانی برای از دست نرفتن جزئیات اطلاعات صوتی گرفته شده است. در نهایت در این پژوهش با استفاده از دسته بند SVM و GMM[1] به دسته بندی دیتاست ها پرداختند.

هیلورر، مندریک و کونکلین [۲] مجموعه ای از ۳۳۶۷ آهنگ محلی ملل شامل شش منطقه جغرافیایی را دسته‌بندی کردند. بدین منظور از روش تبدیل خط ملودی برای استخراج یک بردار ویژگی شامل ویژگیهای فرکانس پیچ، فاصله ملودیکی، طول کشش نت ها، میانگین فرکانس پیچ ... استفاده کردند و سپس بردار ویژگی تولید شده را به وسیله ی چهار روش " Naive Bayes"، درخت تصمیم، "SVM" و"kNN (k=20) " مورد ارزیابی قرار دادند. نتایج نشان داد که مدل رویداد باید به عنوان مدل پیش فرض انتخاب ویژگی برای دسته بندی موسیقی های محلی باشد.

رابر و فروورس [۳] برای به دست آوردن دسته‌بندی داده های موسیقی و ایجاد یک سیستم کتابخانه دیجیتال از SOMLib استفاده کردند. در این پژوهش برای پیش پردازش و استخراج ویژگی در ابتدا از سیستم XMMS[2] برای تقسیم کردن به چند باند فرکانسی که مقدار هر کدام ۲۰-۲۵ میلی ثانیه بوده است، استفاده شده است. سپس از مجموعه ی ۱۷ زیر باند به طول زمانی ۵ ثانیه تعداد ۴۳۵۲ ویژگی با استفاده از FFT به نمایندگی استخراج و از آن برای آموزش شبکه عصبی SOM استفاده

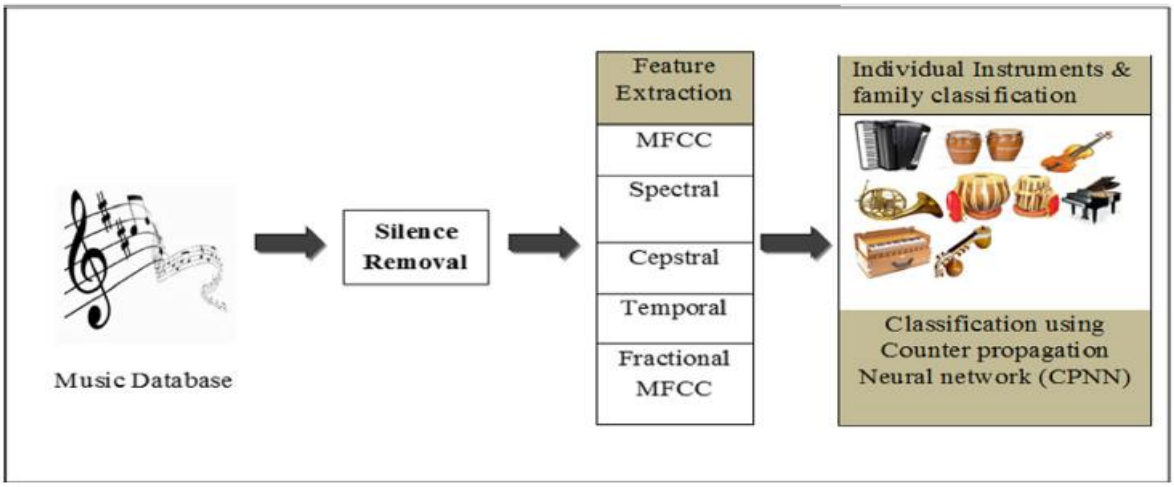
[1] Gaussian Mixture Models
[2] X Multimedia System

شده است. پس از استخراج طیف فرکانسی، بخش هایی از سیگنال موسیقی به طول ۵ ثانیه خوشه بندی می شوند، به طوری که توسط شبکه عصبی SOM بخش هایی با ویژگی های صوتی مشابه با یکدیگر در یک بخش نزدیک به هم قرار می گیرند.

دسته بندی توسط شبکه عصبی SOM در دو مرحله انجام شده است. در مرحله اول، فریم های کوتاه موسیقی به منظور نمایش دادن جزئیات دقیق از شباهت های قطعی، خوشه بندی شدند. در فرایند خوشه بندی سطح دوم، خوشه بندی بر روی کل سیگنال موسیقی بر اساس یک ویژگی توسط شبکه عصبی SOM انجام شده است. در این پژوهش قطعات موسیقی با ویژگیهای صوتی مشابه در صفحه نمایش دو بعدی در کنار هم نگاشت می‌شوند.

پلوا و کاستک [۶] برای نمایش گرافیکی مود آهنگها از شبکه عصبی SOM استفاده کردند. ۱۵۳ ویژگی که شامل Temporal Centroid، Spectral Centroid، Mel-Frequency، MFCC و... از موسیقی‌ها استخراج شده است. سپس با استفاده از همبستگی ما بین ویژگی ها بر اساس مقیاس چند بعدی، دسته‌بندی انجام شده است و موسیقی هایی با حالت مشابه در صفحه نمایش دو بعدی در کنار هم قرار گرفته و نقشه موسیقیایی ایجاد می شود.

بهالک، راوو و برومن [۱۲] به منظور دسته بندی آلات موسیقی از تبدیل فوریه بر پایه ویژگی های ضرایب فرکانسی مل و شبکه عصبی بک پروپگیشن استفاده کردند. در مرحله ی پیش پردازش به منظور کاهش پیچیدگی سیستم قسمت سکوت را از سیگنال موسیقی حذف کردند. در مرحله استخراج ویژگی، ویژگی های آکوستیک سیگنال موسیقی شامل ویژگی های زمانی، طیفی و ضرایب فرکانسی کپسترال مل بر پایه ضرایب تبدیل فوریه کسری استخراج شده است. برای دسته بندی آلات موسیقی از شبکه عصبی شمارنده بکپروپگیشن استفاده کردند. شکل شماره ۵ شمای کلی پژوهش صورت گرفته را نشان می دهد.



شکل ۵. مراحل کلی دسته بندی آلات موسیقی بر پایه تبدیل فوریه کسری

در این پژوهش ۱۹ ساز از چهار گروه موسیقی مختلف از جمله ساز های برنجی، زهی، ضربی و چوبی و محدوده ی مختلفی از نت ها به صورت جداگانه برای آموزش و آزمایش مورد بررسی قرار گرفته شده است. از این بین ۷۰ درصد برای آموزش و ۳۰ درصد از نت ها برای آزمایش انتخاب شدند.

از پایگاه داده MUMS[1] که دارای نمونه هایی شامل ۳ دی وی دی که نمونه برداری آنها با نرخ ۴۴۱۰۰ هرتز در محدوده فرکانسی انجام شده، برای آزمایش مورد استفاده قرار گرفته است.

نتایج نشان دادند که قابلیت تفکیک پذیری در بین دسته ها حداکثر و در درون دسته ها حداقل شده است. در زمینه دقت و استحکام دسته بندی نتایج نشان دهنده این مورد بوده است که ویژگی های پیشنهادی بهبود قابل توجهی در برابر نویز سفید گوسی افزودنی AWGN[2] نسبت به سایر ویژگی های متعارف نشان داده اند.

پژوهشگران این مقاله دلیل استفاده از شبکه عصبی CPNN را آموزش سریع تر، پیش بینی بهتر، انعطاف پذیری بالا در داده های پیچیده غیر خطی و ریسک پایین تر نسبت به سایر شبکه های عصبی بیان کردند.

نتایج بررسی قابلیت تفکیک پذیری ویژگی ها، توسط متوسط خطای مربع (MSE) در بین ویژگی های حاصل از تبدیل فوریه در درون دسته ها و بین دسته ها انجام شده است. نتایج نشان دادند که MSE در بین آلات موسیقی داخل یک دسته نسبت به MSE بین دسته ها بسیار کم است.

برای بهبود قابلیت تفکیک پذیری ویژگی های MFCC، از تبدیل فوریه کسری گسسته به جای تبدیل فوریه گسسته در MFCC استفاده شده است. نتایج حاصل از میانگین خطای مربع (RMSE) نشان می دهد که ویژگی های پیشنهادی، تغییرات مابین دسته ها را به حداکثر کرده و تغییرات درون دسته ها را به حداقل می رساند و قابلیت های تفکیک پذیری ویژگی ها را افزایش می دهد.

شبکه عصبی CPNN[3] :

شبکه عصبی CPNN دو بخشی است. بخش اول شبکه عصبی SOM کوهنن بدون نظارت است. برای خوشه بندی بردار های ورودی و مشخص کردن وزن نود های خوشه بند از SOM استفاده می شود. به بیان دیگر CPNN توسعه یافته ی SOM است، که پس از لایه های SOM لایه ی دیگری به آن افزوده شده است. CPNN توانایی بالایی در خوشه بندی بدون نظارت دارد و خطای موجود در خروجی مطلوب در یادگیری نظارت شده را کاهش می دهد. CPNN شامل ۳ لایه ی ورودی

[1] McGill University Master Sample
[2] Additive White Gaussian Noise
[3] Counter propagation neural network

، لایه ی کوهنن یا SOM و لایه ی خروجی است. ویژگی های استخراج شده در لایه ورودی نمایش داده می شوند . لایه ی کوهنن بردارهای ورودی را با استفاده از فاصله آموزش و وزن نود های خوشه بند خوشه بندی می کند.

مراحل برای آموزش CPNN به صورت زیر است :

- **در مرحله اول:** X بعد بردار ورودی و Yبعد بردار خروجی به شبکه داده می شود .
- **در مرحله دوم:** فاصله مابین ورودی ها محاسبه می شود و وزن ها به لایه ی کوهنن داده می شود .
- **در مرحله سوم:** نود های برنده محاسبه می شوند .
- **در مرحله چهارم:** وزن همه ی نرون ها مرتب می شود.
- **در محله پنجم:** به مرحله دوم برمی گردد تا همه ی ورودی ها آپدیت شوند.

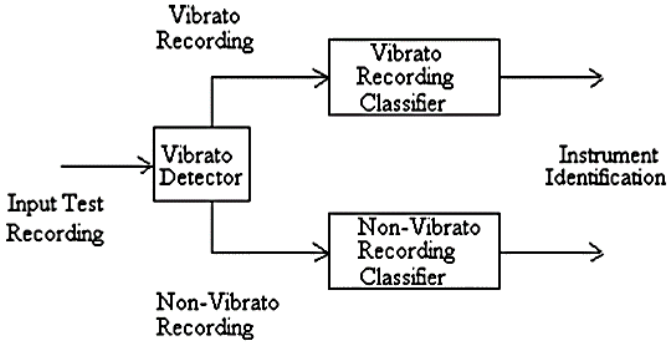
در این پژوهش توپولوژی شبکه به صورت مربعی بوده است. سایز شبکه ۱۰*۱۰ و تعداد اپک بهینه ۲۰۰ بدست آمده است. نرخ یادگیری ۰,۱در نظر گرفته شده است. آموزش شبکه از طریق Batch صورت گرفته است.

کامینسکی و کزاسزسکو [۱۳] به منظور تشخیص خودکار موسیقی تک صدای آلات موسیقی ازشبکه عصبیkNNC استفاده کردند. در این پژوهش دسته بندی بر روی ۱۹ آلت موسیقی انجام شده و از ۶۰۴ تک صدای ضبط شده از بانک اطلاعاتی MUMS استفاده شده است. تک صداها شامل صدای گیتار، ویلن، ویلنسل، کنترباس، پیانو، فلوت، آکاردن، کلارینت، ساکسیفون، ارگان، فرنچ هورن، تورمبون، باسون، ترومپت، زیلوفون و ... می باشد. از سیگنال های موسیقی موجود شش ویژگی استخراج شده است: ۱.ضرایب کپسترال[1] ۲.تبدیل طیف فرکانس ثابت Q[2] ۳ . نقشه برداری مسیرهای چند بعدی (MSA)[3] ۴. میانگین مربع توسعه دامنه[4] ۵. مرکز طیف[5] ۶. حضور لرزش[6].

به منظور محدود کردن پیچیدگی سیستم محدوده ی سه اکتاو از "نت دو" اکتاو سوم تا "نت دو" اکتاو ششم برای فاز آموزش و آزمایش در نظر گرفته شده است. همه ی آلات موسیقی نمی توانند صدایی در این رنج تولید کنند اما این رنج کمترین میزان از دست رفتگی صداها را در دسترس قرار می دهد. در این پژوهش توانستند به دقت ۹۳ درصدی در تشخیص آلت موسیقی، ۹۷ درصدی در تشخیص آلات موسیقی هم خانواده و دقت ۱۰۰ درصدی در آلات موسیقی دارای کشش (غیر ضربه ای) برسند.

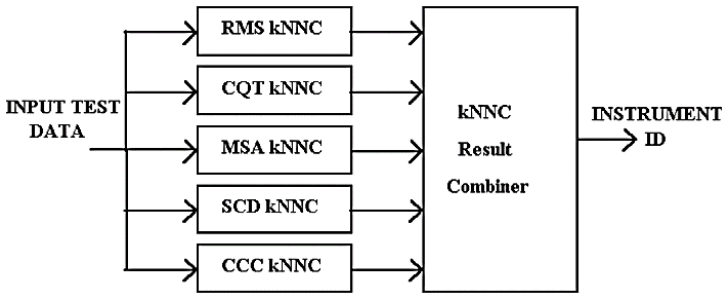
[1] Cepstral coefficients
[2] constant Q transform(CQT)frequency spectrum
[3] Multidimensional scaling analysis(MSA) trajectories
[4] Root mean square(RMS) amplitude envelope
[5] spectral centroid
[6] presence of vibrato

در مرحله اول آشکار ساز لرزش، ساز های دارای لرزش (ارتعاش) را از سایر ساز ها جدا می کند. این مرحله دسته بندی را تسریع می کند. در صورتی که هیچ ارتعاشی وجود نداشته باشد دسته بندی وارد ساز های غیر ارتعاشی می شود.



شکل ۶. مرحله پیش دسته بندی بر اساس وجود لرزش

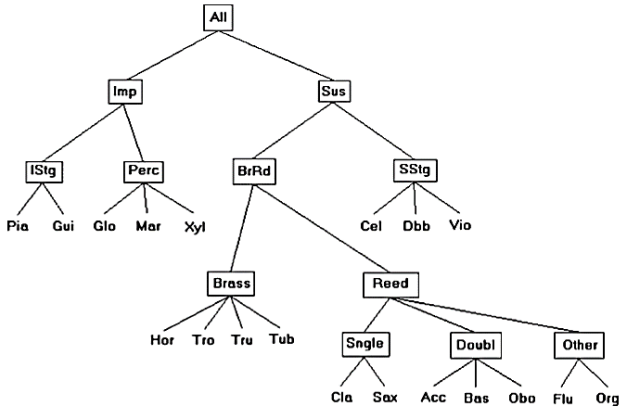
با توجه به تعداد زیاد آلات موسیقی در دسته غیر ارتعاشی ، سه معماری دسته بندی متفاوت مورد بررسی قرار گرفته است: (الف) تک مرحله، (ب) سلسله مراتبی و (ج) ترکیبی. ساده ترین، دسته بندی تک مرحله ای است که در شکل ۷ نشان داده شده است و از پنج دسته استفاده می کند.



شکل ۷. دسته بندی تک مرحله ای ، در هرمرحله بر اساس فقط یک ویژگی

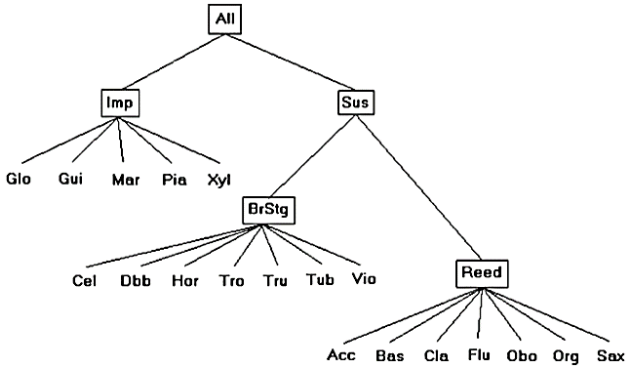
ویژگی های استخراج شده به صورت مستقیم یا پیش پردازش شده با استفاده از تجزیه و تحلیل مولفه اصلی (PCA) مورد استفاده قرار گرفتند. نتایج نشان دادند دسته بندی دقیق مستلزم تعیین ارزش های ارزیابی شده توسط PCA است. همچنین نتایج نشان دادند که برای کسب بهترین نتایج ممکن باید نقاط قوت و ضعف ویژگی های دسته های فردی در نظر گرفته شود. همچنین استفاده از ماتریس سردرگمی برای رسیدن به نتایج بهینه ی مناسب است. یکی از راه های بدست آوردن تفاوت های بین دسته ها، استفاده از ماتریس سردرگمی است.

برای بهبود نتایج با استفاده از دسته بندی تک مرحله، ساختار سلسله مراتبی ارزش گذاری شده است. هر دسته ی تک مرحله ای یک سطح در ساختار سلسله مراتبی را تشکیل می دهد. شکل ۸ نشان دهنده شمای ساختار سلسله مراتبی است.



شکل 8 .دسته بندی سلسله مراتبی

اساس ساختار دسته بندی کننده ترکیبی، سازش بین سرعت دسته بندی و عملکرد بوده است که شامل برخی از ویژگی های دسته تک مرحله و سلسله مراتبی است. شکل ۹ نمایانگر شمای کلی ساختار دسته بندی کننده ترکیبی است.



شکل ۹. شمای کلی ساختار دسته بندی کننده ترکیبی

همانطور که در جدول ۲ مشاهده می شود، روش سلسله مراتبی نتایج بهتری را در دسته بندی نشان می دهد .

جدول ۲. نتایج دسته بندی: ساده، ترکیبی و سلسله مراتبی

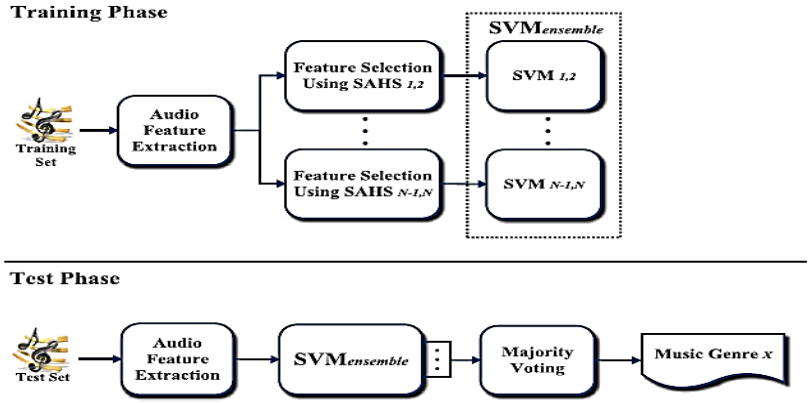
Non-vibrato recording classifier architecture	Impulsive vs sustain identification accuracy (%)	Family identification accuracy (%)	Instrument identification accuracy (%)
Single stage	100	94.6	88.6
Hybrid	100	95.5	89.5
Hierarchic	100	96.1	91.7

هوانگ، لین، وو و لی [۱۴] یک سیستم اتوماتیک دسته بندی ژانر موسیقی بر اساس انتخاب ویژگی های محلی با استفاده از الگوریتم جستجوی هماهنگ سازگار (SAHS) پیشنهاد دادند. در ابتدا پنج خصوصیت آکوستیک شامل فرکانس، رنگ صدایی، تونالیته، شدت و ریتم برای تولید یک مجموعه ویژگی استخراج شدند. برای انتخاب ویژگی در هر ژانر از الگوریتم SAHS استفاده کردند و مجموعه ویژگی های مربوطه محلی را به دست آورند. پس از استخراج ویژگی ها با استفاده از SVM، شیوه یک در برابر یک (ویژگی) و روش رای گیری اکثریت دسته بندی هر اثر ضبط شده موسیقی را انجام دادند. دیتاست مورد استفاده

در مرحله آموزش و آزمایش GTZAN بوده است. نتایج نشان داد که استراتژی انتخاب محلی ویژگی ها نتایج دقیق تری را بدست می دهد.

شرح سیستم

(N-1) / 2 مجموعه ویژگی محلی برای N ژانر تولید می شود سپس الگوریتم SAHS برای بدست آوردن ویژگی های مناسب از نسبت همبستگی درونی بین ویژگی های انتخاب شده استفاده می کند. SVM مربوط به هر ویژگی با استفاده از مجموعه ویژگی های محلی آموزش می بیند. در نهایت، تمام SVM های آموزش دیده ترکیب می شوند تا مدل یک گروه SVM که در مرحله آزمایش استفاده می شود را، تشکیل دهند. شکل شماره ۱۱ نمایانگر شمای کلی عملکرد این پژوهش است.



شکل ۱۰. عملکرد کلی سیستم بر پایه الگوریتم SHAS

در مرحله آزمایش، ابتدا همه ی ویژگی ها استخراج و هر SVM را با ویژگی های بدست آمده تغذیه و از روش رای گیری اکثریت برای تعیین ژانر استفاده کردند. برای استخراج ویژگی فریمورک های توصیفگر MIRToolbox، jAudio و MPEG-7 به کار برده شد. مجموعه ویژگی های اصلی شامل ۲۶۵ ویژگی از ۳۲ توصیف کننده صوتی بدست آمد. اکثر توصیفگرها در واحد فریم یا پنجره ۲۳ میلی ثانیه ایجاد شدند.

در این پژوهش به جای استفاده از مقادیر ویژگی به طور مستقیم، پارامترهای توزیع گاوسی چند مورد استفاده قرار گرفتند. به طور خاص، میانگین و انحراف معیار تمام فریم ها در یک بافت پنجره ۳۰ ثانیه برای ایجاد دو ویژگی آماری محاسبه شد. علاوه بر توصیفگرهای اصلی، میانگین و انحراف معیار برای تفاوت های بین فریم های مجاور نیز برای ایجاد دو ویژگی جدید محاسبه شد.

در این پژوهش ۸ ویژگی شدت، ۱۷ ویژگی از پیچ (فرکانس اصلی) در نظر گرفته شد. برای برآورد پیچ یک فریم، چندی از قوی ترین فرکانس ها با استفاده از چهار روش مختلف شامل اتوکرولیشن، نرخ عبور از صفر، مرکز طیفی و حداکثر FFT و نرخ ناهماهنگی فرکانس هایی که مضرب پیچ در یک پنجره بافت نیستند محاسبه شد.

کالاپاتاپو و همکاران [۱۵] مطالعه ای بر روی انتخاب ویژگی و تکنیک های دسته بندی در موسیقی هند انجام دادند. در این مقاله، اثر چهار الگوریتم انتخاب ویژگی شامل الگوریتم ژنتیک[۱]، انتخاب ویژگی جلو[۲] ، بدست آوردن اطلاعات[۳] و کرولیشن[۴] بر اساس چهار دسته بند متفاوت شامل درخت تصمیم گیری C4.5 ، K-نزدیکترین همسایگان، شبکه عصبی و ماشین بردار پشتیبان را بررسی کردند .

مجموعه ویژگی ها در این مقاله با استفاده از MIR Toolbox در نرم افزار متلب استخراج شده است. ویژگی های استخراج شده شامل ریتم، پیچ (فرکانس اصلی)، تونالیته (درون مایه) و ویژگی های پویا بوده است. بردارهای ویژگی از ۳۰ ثانیه اول سیگنال موسیقی و سی ثانیه آخر سیگنال موسیقی استخراج شدند.

آزمایشات بر روی سه دسته موسیقی اصلی هندی شامل کارناتیک[۵] ، هینداستانی[۶] و بالیوود[۷] انجام شد. دیتاست مورد مطالعه شامل ۲۹۰ آهنگ کوچک بوده است. نتایج این پژوهش نشان داد که استفاده از *الگوریتم بدست آوردن اطلاعات*، دسته بندی بهتر و پایداری بیشتر نسبت به سایر الگوریتم های انتخاب ویژگی فراهم می کند. همچنین دسته بند شبکه های عصبی و SVM، مناسب ترین دسته بند برای مجموعه داده های موسیقی هند بوده است.

فو و همکاران [۱۹] مدل Bag-Of-Features (BOF) برای دسته بندی موسیقی پیشنهاد دادند. مدل BOF بر اساس مدل Bag-Of-Words (BOW) ساخته شده است که برای دسته بندی سند مورد استفاده قرار می گیرد. در این مقاله بر روی ضریب های MFCC تمرکز کردند. مدل BOF در زمینه دسته بندی موسیقی از دو مرحله استفاده می کند که شامل ۱. ساخت Codebook ۲. تخصیص می باشد.

۱. ساخت Codebook

در ابتدا ویژگی های محلی هر موسیقی از مجموعه داده آموزشی استخراج می شود. توسط کوانتیزه سازی، بردارهای ویژگی جمع آوری می شوند و خوشه بندی K-means بر روی آنها اعمال می شود.

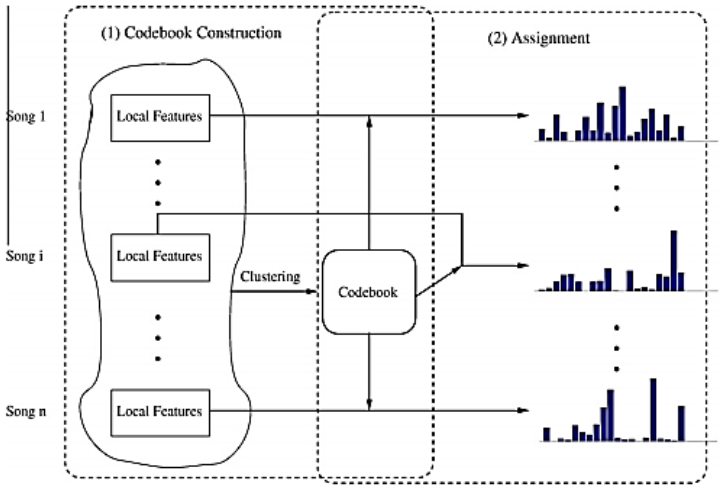
با توجه به اینکه تعداد ویژگی های محلی در مجموعه داده های آموزشی زیاد است، مجموعه ی ویژگی ها جمع آوری شدند و به طور تصادفی یک زیر مجموعه ای از بردارهای ویژگی از فریم های مختلف انتخاب شدند. تعداد مراکز خوشه ها به منظور ثابت نگه داشتن اندازه codebook ثابت است. نتیجه Codebook یک نمایش مختصر و کامل از تمام ویژگی های محلی

[1] genetic algorithm
[2] Forward feature selection
[3] information gain
[4] correlation
[5] Carnatic
[6] Hindustani
[7] Bollywood

است، که یک فرهنگ لغت " کلمات صوتی " است. هر مرکز خوشه یک بردار از Codebook است که هر کلمه صوتی را نشان می دهد . Codebook فقط یک بار در مرحله آموزش ساخته می شود.

۲. تخصیص

در مرحله تخصیص، مدل BOF با استفاده از codebook آموخته شده، ویژگی های هر آهنگ را به صورت یک بردار codebook و محاسبه هیستوگرام نرمال برای فرکانس متناظر هر بردار codebook نشان می دهد. شکل ۱۱ نمایانگر شمای کلی مدل BOF است.



شکل ۱۱. شمای کلی مدل BOF

در پیاده سازی، سیگنال های صوتی به فریم های ۲۳ میلی ثانیه با ۵۰٪همپوشانی بین فریم های مجاور تقسیم بندی شدند. برای استخراج ویژگی MFCC در هر فریم از ۳۶ فیلتر مقیاس Mel در محدوده طیفی ۰ Hz تا قانون فرکانسی Nyquist با ۵۰٪ همپوشانی بین فیلترهای مجاور استفاده شده است. در نهایت، برای استخراج 20 بعد از ویژگی MFCC برای هر فریم محلی DCT به خروجی فیلتر اعمال شده و تنها ۲۰ ضریب DCT بالا (به جز مولفه DC) نگهداری می شود.

از فاصله اقلیدسی برای محاسبه فاصله بین ویژگی های MFCC محلی در هر دو روش خوشه بندی K-means و محاسبات ویژگی BOF استفاده شده است. برای همساز کردن ویژگی ها با رنج های مختلف مقداری، مقدار ویژگی ها را در رنج [0,1] قرار دادند.

دو دیتاست GTZAN برای دسته بندی ژانر ها و دیتاستPOP2002 برای معرفی کردن هنرمند برای آموزش و آزمایش استفاده کردند. تمامی سیگنال های موسیقی در فرمت wav بوده و از تمامی داده ها ویژگی های MFCC استخراج شده است. برای هر دیتاست دسته بندی ۱۰ بار با بخش بندی مختلف در آموزش و آزمایش انجام شده است .

برای هر دو دیتاست از دسته بند k-NN و دسته بند SVM بر پایه ی RBF استفاده شده است. در نهایت پژوهشگران به این نتیجه رسیدند مدل BOF می تواند برای چند لیبل شدن موسیقی مورد استفاده قرار بگیرد و این مشکل را در برابر دسته بند های دستهیک که سیگنال موسیقی را تنها به یک دسته متعلق می دانند حل کند.

لیو و همکاران [۱۷] سیستمی را پیشنهاد کردند که موسیقی را بر اساس دوره زمانی دسته بندی می کند. سیستم پیشنهادی از ژن های موسیقی برای دسته بندی استفاده می کند. ژن های موسیقی از یک فایل XML[1] استخراج می شوند. با توجه به تئوری موسیقی، موسیقی دستهیک بیان خود را در یک ملودی بلندتر نسبت به یک موسیقی مدرن بیان می کند. ژن ها در این پژوهش نشان دهنده ی بردار ویژگی موسیقیایی از دو دوره دستهیک اروپایی و مدرن اروپایی هستند. استخراج ژن ها در دو مرحله انجام می شود:

در ابتدا فرکانس های موجود در هر ژن بدست آمده و اختلاف آنها نسبت به فرکانس بعد خود بدست می آید. بدین صورت فواصل فرکانسی بدست می آیند. در مرحله دوم دسته بندی برای بدست آوردن الگو های مشابه صورت می گیرد.

برای دسته بندی بندی فایل های XML از SVM استفاده شده است. برای ارزیابی روش پیشنهادی از ۱۵۶ آهنگ دارای ترانه از دو دوره مختلف زمانی استفاده کردند. طبق نتایج بدست آمده از این پژوهش، ژن های موسیقی یکپارچگی ملودی را حفظ می کنند.

پراسانا وخونگلا [۲۱] استفاده از ویژگی های خاص گفتار را برای دسته بندی گفتار/ موسیقی پیشنهاد دادند. این ویژگی ها شامل:

۱. توان در قله (پیک) از فرکانس صفر تا نرخ طرفین در اتوکرولیشن نرمال شده
۲. لگاریتم انرژی MFCC که نشان دهنده اطلاعات صوتی است.
۳. طیف مدولاسیون،که نشان دهنده تغییرات زمانی بوده و مربوط به نرخ سیگنال گفتار است.

سیگنال گفتار در زمان ایجاد شدن از یک تشدیدگر عبور می کند. این تشدیدگر در فرکانس صفر قرار داشته و انرژی سیگنال را در اطراف فرکانس صفر نگه می دارد و اطلاعات دیگر را به طور قابل توجهی کاهش می دهد. این امر به علت رزونانس های صوتی آوازی رخ می دهد.

ZFFS [۲] فیلتری است که فرکانس های صفر را حذف کرده و اطلاعاتی را درباره مکان اصلی دوره های تناوب را در سیگنال های گفتار در اختیار قرار می دهد.

[1] Sheet-MusicXML
[2] zero frequency filtered signal

ZFFS در بلوک های ۳۰ میلی ثانیه ای با ۱ میلی ثانیه شیفت در هر فریم پردازش می شود. در این پژوهش نسبت قله در طرفین[۱]، توسعه هیلبرت[۲] و پیشگویی خطی[۳] مورد بررسی قرار گرفت. آنالیز LP یک متد است که بخش گفتاری و اطلاعات را از یک سیگنال گفتار استخراج می کند. انرژی در رنج پایین فرکانسی را می توان با انرژی های فیلتر بانک مل نمایش داد. به منظور استخراج ویژگی انرژی MFCC بر روی هر بلاک ۳۰ میلی ثانیه ای یک DFT ۵۱۲ نقطه ای محاسبه شده است.

با استفاده از دسته بندهایی مانند مدل های ترکیبی Gaussian (GMM[۴])و ماشین های بردار پشتیبانی ویژگی های استخراج شده برای دسته بندی مورد ارزیابی قرار گرفتند. دیتاست های مورد بررسی شامل GTZAN، S&Sو Indian broadcast news بوده است. نتایج نشان دادند که عملکرد دسته بندی با ویژگی های خاص مطرح شده نسبت به ویژگی های مطالعات پیشین بهتر است.

دلیمن و اسچراوو [۱۸] در پژوهشی بررسی کردند که آیا یادگیری عمیق را می توان به طور مستقیم به سیگنال های صوتی خام اعمال کرد؟ بسیاری از ویژگی های سطح بالا در صدا مربوط به انرژی در باند فرکانس های مختلف است. شکل ۱۲ نشان دهنده ی شبکه ی عصبی کانولوشن این پژوهش است. اندازه فیلتر را با (↔) و تعداد سلول ها با (#) نشان داده شده است. در این پژوهش سه رویکرد در نظر گرفته شده است:

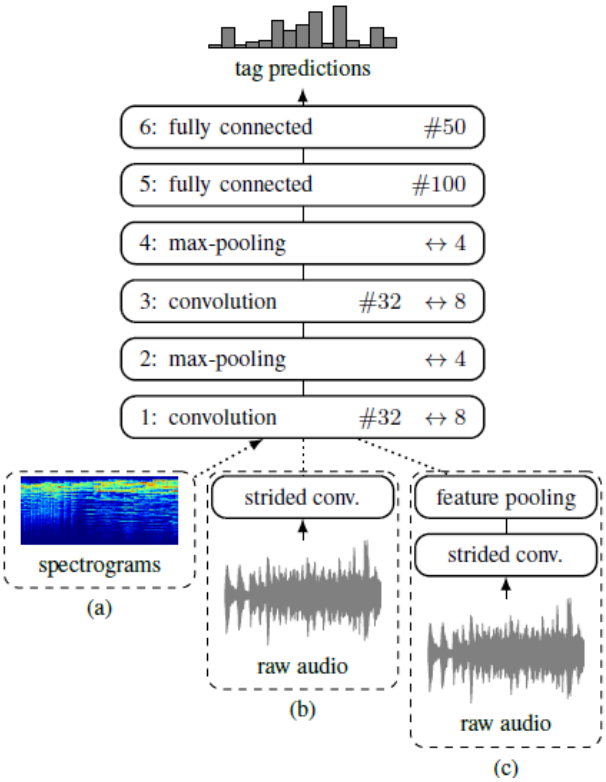
a. ورودی طیف اسپکتگرام باشد.

b. صدای خام با یک لایه اضافی کانولوشن

c. صدای خام با جمع آوری ویژگی ها

برای مقایسه کردن روش یادگیری end-to-end با روش سنتی بازیابی اطلاعات موسیقی شبکه عمیق CNN را برای اتوماتیک تگ زدن بر روی دیتاست Magnatagatune آموزش دادند. دیتاست مورد بررسی شامل ۲۵۸۶۳ کلیپ ۲۹ ثانیه ای با نرخ نمونه برداری ۱۶ کیلو هرتز بوده است. این دیتاست از آثار ۲۳۰هنرمند که با ۱۸۸ تگ، تگ نویسی شده، تشکیل شده است. در این پژوهش فقط از ۵۰ تگ فرکانسی پر تکرار استفاده شده است.

[۱] The peak-to-sidelobe ratio
[۲] Hilbert envelope
[۳] linear prediction
[۴] Gaussian mixture models



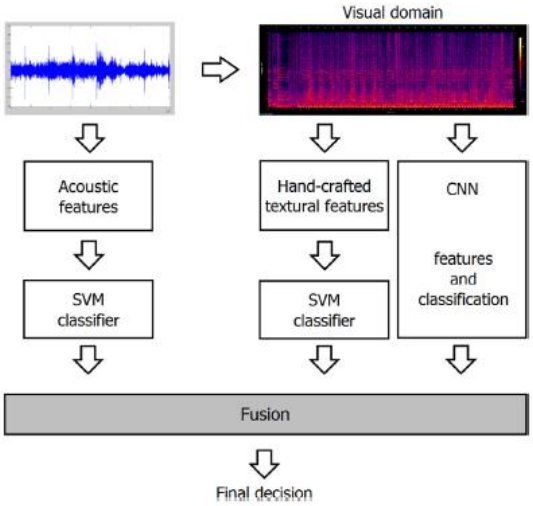
شکل ۱۲. شبکه عصبی کانولوشن برای دسته بندی سیگنال صوتی

پردازش در ۱۶ بخش انجام شده است: ۱۲ مرحله اول برای آموزش و ۱۳ امین مرحله برای اعتبار سنجی و ۳ مرحله باقی مانده برای تست مورد استفاده قرار گرفته است. نتایج آزمایشات نشان داد که سطح کارایی بالایی بر پایه ی رویکرد اسپکتگرام بدست نمی آید. همچنین نتایج نشان دادند که شبکه های یادگیری عمیق قادرند که ویژگی های مهم را از صدای خام یادبگیرند و قادرند به صورت خودگردان فرکانس ها را کشف و تجزیه کنند. در نهایت زمانی که یک لایه ویژگی استخراج شده نیز به آن افزوده شود قادرند تا ویژگی های فازی و غیر متعارف صدا را نیز کشف کنند.

یاندر،کوستا، اولیوریاب و سیلا [۱۹] برخلاف استفاده از ویژگی ها متعارف برای دسته بندی ژانرهای موسیقی از ویژگی های نامتعارفی شامل الگو های باینری محلی، فاز محلی دیجیتالی شاری کردن و فیلتر گابور برای یادگیری شبکه عمیق استفاده کردند. در این پژوهش آنها نتایج شبکه عصبی کانولوشن را با نتایج استخراج ویژگی به صورت دستی و دستهبند SVM مقایسه کردند.

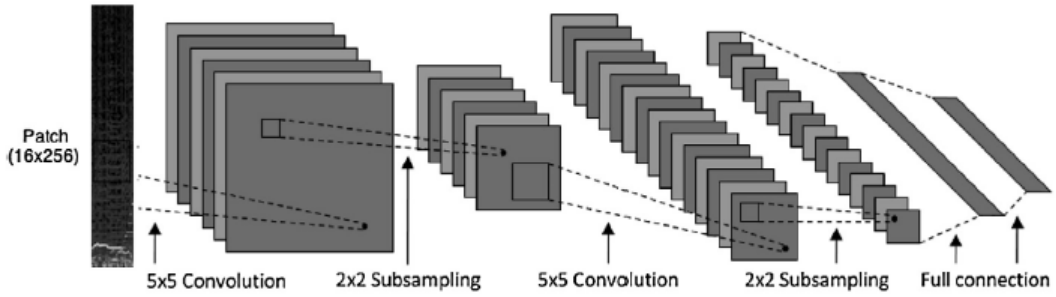
آزمایشات خود را بر روی سه بانک اطلاعاتی موسیقی غرب (ISMIR 2004 Database) و بانک اطلاعاتی موسیقی آمریکای لاتین (LMD database) و مجموعه ای از موسیقی های قومی آفریقایی انجام دادند.

آزمایشات نشان دادندکه CNN در چندین سناریو نسبت به دسته بندی های دیگر مناسب تر است و از این رو، جایگزینی بسیار جالب برای شناخت ژانر موسیقی است. در پایگاه داده LMD ترکیب کردن CNN و الگوریتم الگوی قوی دودویی محلی بهترین نتیجه را با ۹۲٪ در تشخیص بدست آورد. شکل ۱۳ نشان دهنده شمای کلی پژوهش صورت گرفته است.



شکل ۱۳. شمای کلی مقایسه سه روش دسته بندی موسیقی

در این پژوهشCNN با SGD[1] آموزش داده شده و توسط بک پروپگیشن با ۸۰ اپک به نتیجه مطلوب دست پیدا کرده است. نرخ یادگیری بر روی 10^{-3} تنظیم شده است تا در شروع، وزن ها به سرعت ثابت شوند، سپس این مقدار هر بار تا 5×10^{-4} کاهش پیدا می کند. شکل ۱۴ بهترین ساختار بدست آمده برای شبکه کانولوشن را نشان می دهد.



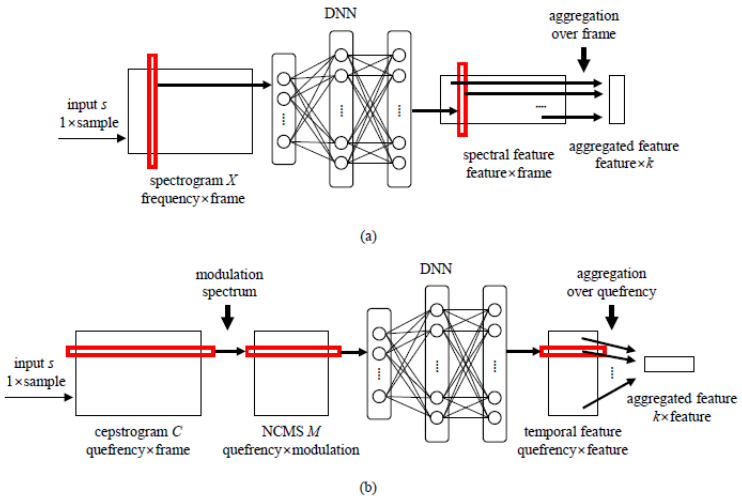
شکل ۱۴. بهترین ساختار بدست آمده برای شبکه کانولوشن

برای انتخاب بهترین مدل برای یادگیری، مجموعه داده را به دو دسته آموزش (۷۰٪) و اعتبار سنجی (۳۰٪) تقسیم کردند. در این پژوهش برای پیاده سازی، از فریم ورک Caffe بر روی TeslaC2050 GPU استفاده شده است.

[1] Stochastic Gradient Descent

جونگ و لی [۲۰] چارچوبی برای دسته بندی ژانر موسیقی بر اساس یادگیری ویژگی های صوتی توسط شبکه عصبی عمیق معرفی کردند. بدین منظور، آنها ویژگی های زمانی صوت و دامنه طیفی مدولاسیون صوت را به عنوان ویژگی برای یادگیری عمیق مورد استفاده قرار دادند.

در این پژوهش با تجمع زمانی و معکوس کردن ویژگی های طیفی یک باند طیفی برای ویژگی های زمانی ایجاد کردند. از دو ویژگی حاصل برای سیستم دسته بندی ژانر استفاده کردند.



شکل ۱۵. شمای کلی شبکه یادگیری عمیق برای دسته بندی موسیقی

در این پژوهش به واسطه ی PCA بردار های ویژگی فشرده سازی شدند. رنج ویژگی های زمانی و طیفی با یکدیگر متفاوت بوده و به منظور استفاده از دو ویژگی به طور همزمان نرمال سازی بر روی آنها صورت گرفته است .

آموزش به وسیله ی مینی بیت گرادیان نزولی با نرخ یادگیری ۰,۰۱ انجام پذیرفته است. الگوریتم پیشنهادی برای هر کانولوشن سایز بیت ۱۰۰ را بدست آورده است. عملیات بهینه شده بعد از ۲۰۰ اپک ایجاد می شود. نتایج آموزش و آزمایش در دیتاست GTZAN نشان می دهد که ویژگی های زمانی می توانند دقت دسته بندی بالاتری نسبت به ویژگی های طیفی آموخته شده بدست آورند.

تفاوت mini-batch و batch و SGD

SGDیک تقریب از GD یا Gradient Descent است. شیوه عملکرد SGD بصورت تکراری است. در هر مرحله یک نمونه تعیین شده و محاسبات بر روی آن انجام می شود، سپس تغییرات در شبکه اعمال می شود. SGD به شیوه online معروف است[۱۰].

در حالت استاندارد یا بچ گرادیان، باید گرادیان تمامی Training set برای هر آپدیت محاسبه شود. در حالت batch کل دیتا باید خوانده شود. در گام اول گرادیان ها محاسبه می شوند سپس میانگین گیری انجام می شود. در مرحله آخر پارامترها بروز می شوند. مشکل بچ گرادیان این است که سربار زیادی ایجاد می کند.

در Mini-batch گرادیان به ازای چند نمونه محاسبه می شود که در این صورت گرادیان پایدارتری بدست می آید و امکان پیاده سازی برداری سریع و بهینه ایجاد می شود[۹].

نتیجه گیری

با مطالعه پژوهش های پیشین نتایج نشان می دهد که استفاده از ویژگی هایی نظیر MFCC و سایر ویژگی های طیفی برای دسته بندی موسیقی بهتر عمل می کنند. به دلیل حجم بالای ویژگی های استخراج شده کاهش ابعاد پیچیدگی سیستم توسط الگوریتم هایی نظیر PCA و LPA ، ... یکی از مراحل مهم در حوزه دسته بندی می باشد.

روش های متنوعی برای دسته بندی از جمله KNN ، درخت تصمیم گیری، SVM، شبکه های عصبی و ... وجود دارد. در این بین دسته بندی به وسیله ی شبکه های عصبی دقت بیشتر و نتایج بهتری کسب کرده است.

در حوزه یادگیری عمیق و دسته بندی صوت پژوهش هایی صورت گرفته است البته این حوزه از هوش مصنوعی در بخش پردازش صوت و موسیقی نوپا بوده و هنوز از چالش های زیادی برخوردار است.

مراجع

1. Lidy, Thomas&Rauber, Andreas. (2005).EVALUATION OF FEATURE EXTRACTORS AND PSYCHO-ACOUSTIC TRANSFORMATIONS FOR MUSIC GENRE CLASSIFICATION. International Conference on Music Information Retrieval (ISMIR), London, UK; 11.09.2005 - 15.09. in:"Proceedings of the Sixth International Conference on Music Information Retrieval", (2005), ISBN: 0-9551179-0-9; S. 34 - 41
2. Hillewaere, Ruben. Manderick, Bernard& Conklin, Darrell. (2009). GLOBAL FEATURE VERSUS EVENT MODELS FOR FOLK SONG CLASSIFICATION. 10th International Society for Music Information Retrieval Conference.
3. Rauber, Andreas & Fruhwirth, Markus. (2001). Automatically Analyzing and Organizing Music Archives. Springer-Verlag Berlin Heidelberg. ECDL2001, LNCS 2163, pp. 402–414.
4. Singh, Inderjeet & G.Koolagudi, Shashidhar. (2017). Classification of Punjabi Folk Musical Instruments Based on Acoustic Features. Springer Science+Business Media Singapore.
5. McKinney, Martin F& Breebaart, Jeroen. (2003).Proceedings of the International Symposium on Music Information Retrieval.pp 151-158.
6. PLEWA, Magdalena& KOSTEK, Bożena. (2015). Music Mood Visualization Using Self-Organizing Maps. Archives of Acoustics, 40, 4, pp. 513–525.
7. Humphrey, Eric J. P. Bello, Juan & LeCun, Yann. (2013). Feature learning and deep architectures: new directions for music informatics. Journal of Intelligent Information Systems. Volume 41, Issue 3, pp 461–481.

8. M. Schedl & D. Schnitzer,(2014) "Location-Aware Music Artist Recommendation," in MultiMedia Modeling (C. Gurrin, F. Hopfgartner, W. Hurst, H. Johansen, H. Lee, and N. O'Connor, eds.), vol. 8326 of Lecture Notes in Computer Science, pp. 205--213, Springer International Publishing.
9. Kereliuk, Corey. Sturm, Bob L, & Larsen, Jan. (2015). Deep Learning and Music Adversaries. IEEE Transactions on Multimedia, Volume: 17, Issue: 11, Nov, pp 2059 – 2071.
10. Huang, Allen& Wu, Raymond. (2016). Deep Learning for Music. Stanford CS224D.
11. Ayu Vida, Gst. Giri, Mastrika& Harjoko, Agus. (2016). Music Recommendation System Based on Context Using Case-Based Reasoning and Self Organizing Map. Indonesian Journal of Electrical Engineering and Computer Science Vol. 4, No. 2, November 2016, pp. 459 – 464.
12. Bhalke, D.G., Rao, C.B.R. & Bormane, D.S. (2016). Automatic musical instrument classification using fractional Fourier transform based- MFCC features and counter propagation neural network. Journal of Intelligent Information Systems. Volume 46, Issue 3, pp 425–446
13. Kaminskyj, I. & Czaszejko, T.(2005). Automatic Recognition of Isolated Monophonic Musical Instrument Sounds using kNNC. Journal of Intelligent Information Systems, Volume 24, Issue 2–3, pp 199–221.
14. Huang, Yin-Fu. MinLin, Sheng. YuWu, Huan. Siou, LiYu. (2014). Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. Data & Knowledge Engineering. Volume 92, pp 60-76.
15. Kalapatapu, Prafulla. Goli, Srihita. Arthum, Prasanna. Malapati, Aruna. (2016). A Study on Feature Selection and Classification Techniques of Indian Music. The 7th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2016). Procedia Computer Science 98, pp125 – 131.
16. Fu, Zhouyu. Lu, Guojun. Ming Ting, Kai. Zhang, Dengsheng. (2011). Music classification via the bag-of-features approach. Pattern Recognition Letters. Volume 32, Issue 14, pp1768-1777.
17. Liu, Yang. Li, Xiongfei. Te, Rigen. Pan, Chi. Zang, Xuebai. (2014). Extracting music genes for era classification. Expert Systems with Applications. Volume 41, Issue 11, pp 5520-5525.
18. Dieleman, Sander. Schrauwen, Benjamin. (2014). END-TO-END LEARNING FOR MUSIC AUDIO. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP).
19. M.G, Yandre. Luiz S, Costaa. Oliveirab, N. Silla Jr, Carlos (2017). An evaluation of Convolutional Neural Networks for music classification using spectrograms. Applied Soft Computing. Volume 52, pp 28–38.
20. IL, Young Jeong & Lee, Kyogu. (2016).LEARNING EMPORAL FEATURES USING A DEEP NEURAL NETWORK AND ITS APPLICATION TO MUSIC GENRE CLASSIFICATION. Proceedings of the 17th ISMIR Conference, New York City, USA, August 7-11.