بسم الله الرحمن الرحيم

دانشکده فنی و مهندسی رسانه

پایان نامه برای دریافت درجه کارشناسی ارشد

رشته مهندسی صدا

تشخیص خودکار سیگنالهای موسیقی و گفتار فارسی

نگارنده

فرزانه عزیزخانی

استاد راهنما

دکتر محمد بهدادفر

استاد مشاور

مهندس محمد ابراهیم صادقی

زمستان ۱۳۹۱

تقدیم به:

پدر و مادر دلسوز و مهربانم که همواره مشوق و پشتیبان من بودند.

تشکر و قدردانی:

با سپاس ویژه از زحمات استاد گرامی جناب آقای دکتر محمد بهداد فر

و

همچنین جناب آقای مهندس محمد ابراهیم صادقی.

چکیده:

با توجه به رشد روز افزون محتوای چند رسانه‌ای، زمینه‌های تحقیقاتی جدیدی همچون تحلیل محتوای صوتی جهت پردازش، تشخیص و دسته‌بندی داده‌ها بر اساسِ محتوا معرفی شده است. مسئله دسته‌بندی سیگنال‌هایِ گفتار و موسیقی که زیر مجموعه‌ای از این تحقیقاتِ جدید است، توسط محققان زیادی مورد توجه و بررسی قرار گرفته است. با این وجود، تاکنون الگوریتمی که بتواند در لحظه همانند سیستمِ شنوایی انسان عمل کرده و سیگنال‌های صوتی دریافتی را با دقتی برابر دقت سیستم شنوایی طبقه‌بندی نماید ارائه نشده است. هر تحقیق جدیدی در این زمینه یا ویژگی‌های جدیدی را معرفی می‌کند یا ویژگی‌هایی که در مقالات قبلی از آنها استفاده شده را بهبود می‌بخشد. در این پایان‌نامه نیز سعی شده است که از مجموعه ویژگیهایی همچون درصد فریم‌های انرژی پایین وانحراف معیار شار طیفی که در تحقیقات انجام شده توانایی مناسبی در تفکیک از خود نشان داده‌اند، استفاده شود. همچنین، دو ویژگی رویدادِ عبور نکردن سیگنال مثبت از صفر و عبور نکردن سیگنال منفی از صفر در این پایان‌نامه مورد بررسی قرار گرفته‌اند که بر اساس یافته‌های ما، تاکنون در تحقیقات داخل کشور برای تشخیص گفتار فارسی از موسیقی مورد بررسی قرار نگرفته‌اند. بر این اساس، چهار ویژگی فوق از نمونه‌های صوتی با عرض ۲۰ میلی‌ثانیه از گفتار فارسی و موسیقی استخراج و جهت آموزش و آزمایش درستی تشخیص سیگنال گفتار و موسیقی به دو نوع شبکه عصبیMLP و RBF و یک روش آماری K-NN ارائه شده است. نتایج این الگوریتمها بر اساس ویژگی‌های به‌کار رفته مورد ارزیابی و مقایسه قرار گرفته است.

واژه‌های کلیدی: دسته‌بندی خودکار سیگنال، استخراج ویژگی، تشخیص سیگنال گفتار و موسیقی، شبکه های عصبی.

فهرست مطالب:

عنوان: **صفحه**

فهرست جداول ..... د

فهرست شکل‌ها ..... ه

فصل اول: کلیات تحقیق ۱

۱-۱- مقدمه ۱

۱-۲- طرح مسئله ۱

۱-۳- ضرورت و اهمیت تحقیق ۲

۱-٤- اهداف تحقیق ۳

۱-۵- ساختار تحقیق ٦

فصل دوم: مبانی نظری تحقیق ۷

۲-۱- مقدمه ۸

۲-۲- استخراج ویژگیها ۸

۲-۲-۱- تعریف بردار و فضای ویژگی ۹

۲-۲-۲- سطوح استخراج ویژگی ۱۰

۲-۳- انواع ویژگی‌ها ۱۱

۲-۳-۱- ویژگی‌های حوزه زمان ۱۳

۲-۳-۱-۱- مقدار مؤثر ۱٤

۲-۳-۱-۲- انرژی زمان کوتاه ۱٤

۲-۳-۱-۳- درصد فریم‌های انرژی پایین ۱۵

۲-۳-۱-٤- نرخ عبور از صفر ۱٦

۲-۳-۲- ویژگی‌های حوزه فرکانس ۱۸

۱-۲-۳-۲- نقطه انحراف طیفی ۱۹
۲-۲-۳-۲- شار طیفی ۲۰
۳-۲-۳-۲- مرکز ثقل طیفی ۲۰
٤-۲-۳-۲- ضرایب کپسترال مبتنی بر مقیاس مل ۲۱
٥-۲-۳-۲- ضرایب پیشگویی خطی ۲۲
۳-۳-۲- خواص آماری ویژگیها ۲۳
٤-۳-۲- رویدادهای عبور نکردن از صفر ۲٤
٤-۲- طبقه‌بندی ۲۹
۱-٤-۲- شبکه‌های عصبی زیستی ۳۰
۲-٤-۲- شبکه های عصبی مصنوعی ۳۱
۱-۲-٤-۲- نرون مصنوعی ۳۱
۲-۲-٤-۲- ساختار شبکه‌های عصبی ۳۲
۳-۲-٤-۲- شبکه عصبی پرسپترونی چند لایه (MLP) ۳۳
٤-۲-٤-۲- شبکه های عصبی با توابع پایه شعاعی (RBF) ۳٤
۳-٤-۲- الگوریتم آماریِ K نزدیکترین همسایه ٤۱
٥-۲- بررسی تحقیقات پیشین ٤۲
فصل سوم : روش تحقیق ٥۲
۱-۳- مقدمه ٥۳
۲-۳- مجموعه نمونه‌های آموزش و آزمایش ٥۳
۳-۳- مرحله پیش پردازش ٥٥
٤-۳- مرحله استخراج ویژگی ٥٥
٥-۳- مرحله بهنجارسازی ٥٦
٦-۳- فضای ویژگی ٥٦

۳-۷- الگوریتم‌های طبقه‌بندی ۵۸
۱-۷-۳- ساختار شبکه MLP ۵۸
۲-۷-۳- ساختار شبکه RBF ۶۱
۱-۲-۷-۳- الگوریتم یادگیری شبکه newrb ۶۲
۲-۲-۷-۳- ساختار شبکه newrb ۶۳
۳-۷-۳- ساختار الگوریتم K-NN ۶۵
فصل چهارم: یافته‌های تحقیق ۶۷
۱-٤- ارزیابی شبکه RBF ۶۸
۲-٤- ارزیابی الگوریتمِ K-NN ۷۱
۳-٤- بررسی تأثیر حذف یک ویژگی ۷۳
٤-٤- بررسی تأثیر حذف دو ویژگی ۷۶
٥-٤- بررسی تأثیر ویژگی‌ها به صورت جداگانه ۷۸
٦-٤- بررسی و مقایسه ویژگی انحراف معیار نرخ عبور از صفر ۸۰
۷-٤- بررسی دقت سیستم بر اساس نوع داده ۸۲
فصل پنجم: نتیجه‌گیری و پیشنهادهای ۸۶
۱-٥- نتیجه‌گیری ۸۷
۲-٥- پیشنهادهای تحقیق ۹۱
فهرست مراجع ۹۳

فهرست جدول‌ها

جدول ۱-۲- بررسی تحقیقات پیشین. ٤٠
جدول ۱-٤- تغییرات کارایی و درصد تشخیص شبکه نسبت به تغییرات پهنای توابع شعاعی و خطای هدف ٦٨
جدول ۲-٤- تغییرات K نزدیکترین همسایگی و درصد تشخیصِ الگوریتم. ٧١
جدول ۳-٤- بررسی تأثیر حذف یک ویژگی بر درصد تشخیص شبکه RBF ٧٤
جدول ٤-٤- بررسی تأثیر حذف یک ویژگی بر درصد تشخیص الگوریتم K-NN ٧٥
جدول ٥-٤- بررسی حذف دو ویژگی بر درصد تشخیص شبکه RBF. ٧٦
جدول ٦-٤- بررسی حذف دو ویژگی برای الگوریتم K_NN. ٧٧
جدول ۷-٤- بررسی اثر تک تک ویژگیها بر درصد تشخیص شبکه RBF. ٧٩
جدول ۸-٤- بررسی اثر تک تک ویژگیها بر درصد تشخیص K_NN. ٧٩
جدول ۹-٤- بررسی و مقایسه ویژگی انحراف معیار نرخ عبور از صفر. ٨١
جدول ۱۰-٤- بررسی دقت سیستم بر اساس نوع داده‌ها. ٨٢
جدول ۱۱-٤- میزان تشخیص هر ویژگی برای هر نوع نمونه. ٨٣

فهرست اشکال:

شکل ۱-۲- بلوک دیاگرام سیستم تفکیک کننده گفتار و موسیقی. ۸

شکل ۲-۲- فضای ویژگی ۳ بعدی ۱۰

شکل ۳-۲- یک فایل صوتی که شامل N نمونه است به M فریم تقسیم شده که هر فریم شامل R نمونه می‌باشد. ۱۱

شکل ٤-۲- ویژگیهای حوزه زمان و حوزه فرکانس سیگنال صوتی ۱۳

شکل ۵-۲- مقدار مؤثر هر فریم برای یک نمونه گفتار ۱۵

شکل٦-۲- مقدار مؤثر هر فریم برای یک نمونه موسیقی ۱٦

شکل ۷-۲- ویژگی عبور از صفر برای یک نمونه گفتار ۱۷

شکل ۸-۲- ویژگی عبور از صفر برای یک نمونه موسیقی. ۱۸

شکل ۹-۲- بلوک دیاگرام محاسبه MFCC ۲۲

شکل ۱۰-۲- مدل گفتار آنالیز پیشگویی خط. ۲۳

شکل ۱۱-۲- نمایش یک سیگنال گسسته در زمان. ۲۵

شکل ۱۲-۲- رویدادهای سری زمانی دو نمونه متوالی یک سیگنال گسسته در زمان. ۲٦

شکل ۱۳-۲- نمودار احتمال انواع رویدادهای ممکن برای گفتار و موسیقی ۲۸

شکل ١٤-۲- مرزهای تصمیم در طبقه‌بندی. ۲۹

شکل ۱۵-۲- نمایش یک سلول عصبی.. ۳۰

شکل ١٦-۲- نمایش ریاضی از یک نرون. ۳۲

شکل ۱۷-۲- نمایش شبکه عصبی RBF. ۳۵

شکل ۱۸-۲- نمایش $\phi(x)$ بر حسب x در تابع گوسین ۳۵

شکل ۱۹-۲- روند کاهش گرادیان خطا. ۳۹

شکل ۲۰-۲- نمایش نحوه قرارگیری یک نمونه در یک دسته توسط الگوریتم K-NN. ٤۱

شکل ۱-۳- فضای ویژگی سه بعدی از سه ویژگی نمونه‌های مجموعه آموزش و آزمایش. ۵۷

شکل ۲-۳- انتخاب مرز مناسب تصمیم. ٦۳

شکل ۳-۳- توابع فعالیت خطی و شعاعی استفاده شده در شبکه newrb. ٦٤

شکل ٤-۳- ساختار شبکه عصبی newrb. ٦٤

شکل ۱-٤- روند آموزش شبکه. ۷۰

شکل ۲-٤- نمایش شماتیک شبکه. ۷۱

فصل اول

کلیات تحقیق

۱-۱- مقدمه:

همانگونه که می‌دانیم شنوایی یکی از مهمترین حواس بشری است که او را در درک جهان پیرامون یاری می‌کند. همانگونه که اختراع میکروسکوپ تحول عظیمی در شناخت جهان هستی بوجود آورد، افزودن قابلیت شنوایی به ماشین نیز می‌تواند در درک بهتر صداها به ما کمک کند و ما قادر خواهیم بود که راجع به جهان بیشتر از آنچه که می‌دانستیم بدانیم، زیرا چنین سیستم‌هایی این قابلیت را دارند که بتوانند بهتر از انسان بشنوند.

۱-۲- طرح مسئله:

با توجه به حجم وسیع ذخیره و بازیابی اطلاعات چندرسانه‌ای مانند متن، عکس، ویدئو و صدا در سازمانها و مؤسسات مختلف از جمله سازمانهای برودکست، نیاز به دسته‌بندی آنها برای کاربردهایی همچون پردازش بر اساس محتوا، جستجوی اطلاعات خاص بر اساس محتوا و سهولت دسترسی، امری اجتناب ناپذیر می‌نماید. در این میان اطلاعات صوتی حاصل از شبکه‌های پخش، پایگاههای اطلاعات رایانه‌ای، شبکه جهانی اینترنت، لوح‌های فشرده و موارد مشابه دیگر، بخش قابل توجهی از مجموعه اطلاعات چندرسانه‌ای را به خود تخصیص داده‌اند.

از آنجاییکه موتورهای جستجو معمولاً کار فراخوانی و بازیابی داده‌ها را بر اساس کلمات کلیدی موجود در فایل‌ها انجام می‌دهند، در اکثر موارد بدلیل تطابق نداشتن کلمات کلیدی با محتوای فایل های ذخیره شده کار فراخوانی با مشکل مواجه می‌شود، از طرفی تشخیص و دسته‌بندی این حجم از اطلاعات توسط نیروی انسانی پر هزینه و بسیار وقت گیر است. بنابراین مسئله تشخیص و دسته‌بندی فایل‌ها بر اساس محتوا بسیار پراهمیت است. یکی از زیرمجموعه‌های دسته‌بندی اطلاعات

بر اساس محتوا، چنانچه در ادامه این پایان‌نامه نیز اشاره خواهد شد، مسئله تشخیص گفتار از موسیقی فارسی است که با توجه به حجمِ اندک تحقیقات صورت گرفته در داخل کشور تحقیق در این زمینه بسیار مفید می‌نماید. بر این اساس، هدف از انجام این تحقیق، بسترسازی اولیه‌ای جهت پردازشهای بر مبنای محتوا می‌باشد که در قالب تشخیص موسیقی از گفتار فارسی به عنوان یکی از قدمهای اولیه در این راستا، در نظر گرفته شده است.

۳-۱- ضرورت و اهمیت تحقیق:

با توجه به اهمیت روز افزون پردازش سیگنالهای صوتی و با توجه به اینکه اینگونه داده‌ها امروزه بخش بسیار وسیعی از اطلاعات را به خود اختصاص داده‌اند، امروزه تلاش‌ها و مطالعات فراوانی در زمینه تشخیص و دسته‌بندی این نوع از داده‌ها صورت گرفته است. از آنجایی که معمولاً منابع مختلف صوتی همچون پخش رادیویی، برنامه‌های زندهٔ تلویزیونی و یا فیلم‌ها، شامل قسمت‌هایی از موسیقی یا گفتار هستند، بنابراین یک پردازش مهم برای چنین اطلاعاتی، تشخیص و دسته‌بندی آنها به گفتار یا موسیقی است که غالباً برای کاربردهایی که نیاز به تعیین محدودهٔ زمانی سیگنال گفتار یا موسیقی دارند، مفید و مؤثر واقع می‌شود. از جمله این کاربردها، پردازشهای خودکار بر روی موسیقی یا گفتار است، مانند تغییر ضریب تقویت، اعمال همسان سازها[1] و افکت‌های فضایی نظیر تغییر زمان واخنش[2] و شبیه‌سازی سراند[3]. همچنین با نگاهی دیگر، کاربردهایی نظیر رونویسی خودکار اخبار برودکست یا رونویسی نت‌های موسیقی، بازشناسی[4] گفتار و یا گوینده و همچنین دسته‌بندی موسیقی

[1] Equalizer
[2] Reverberation Time
[3] Surround
[4] Recognition

بر اساس ژانر[1] یا سبک[2] را می‌توان نام برد. علاوه بر این، می‌توان برای دسته‌بندی دیگر داده‌ها مانند دسته‌بندی سیگنالهای صوتی_تصویری بر اساس صوت که بسیار ساده‌تر از دسته‌بندی بر اساس تصویر می‌باشد این سیستم یا نسخه‌های تکمیل شده آن را به کار گرفت.

در این پایان‌نامه که بر دسته‌بندی سیگنال صوتی به موسیقی و گفتار فارسی تأکید دارد، کاربردهای اصلی را می‌توان در موارد زیر خلاصه کرد:

۱. به کارگیری این مفهوم در سیستمهای ذخیره سازی یا فراخوانی اطلاعات بر اساس محتوا.

۲. کاهش فضای جستجو در سیگنالهای صوتی.

۳. تشخیص بخشی از فایل صوتی که حاوی گفتار است، که از خروجی آن می‌توان بعنوان ورودی سیستم‌های بازشناسی خودکار گفتار مانند تبدیل گفتار به متن در رونویسی اخبار برودکست، تبدیل گفتار به گفتار در ترجمه از زبانی به زبانی دیگر و تبدیل گفتار به فرمان استفاده نمود.

٤. ایجاد زمینه‌ای برای پیاده سازی سیستم‌های دسته‌بندی موسیقی بر اساس سبک یا ژانر.

٤-١- اهداف تحقیق:

مسئله جداسازی و دسته‌بندی گفتار و موسیقی تاکنون از جنبه‌های مختلف مورد بررسی قرار گرفته است نحوه نگرش به مسئله، مدلسازیها و شبیه‌سازیهای متفاوت باعث ایجاد دقت‌های مختلف در دسته‌بندی شده است.

[1] Genre
[2] Mode

هدف از این مطالعه امکان‌سنجی دسته‌بندی خودکار گفتار فارسی و موسیقی ایرانی با استفاده از ترکیب جدیدتری از ویژگیها و همچنین بررسی دقتِ دو شبکه عصبیMLP [۱] و RBF [۲] و همچنین الگوریتم آماری K-NN [۳] می‌باشد. از آنجاییکه سیگنالهای صوتی مورد بررسی در این پایان‌نامه، به صورت گفتار فارسی و موسیقی سنتی ایرانی خواهند بود و با توجه به اینکه بر اساس یافته‌های ما گفتار فارسی و موسیقی سنتی ایرانی در تحقیقات انجام شدهٔ خارج از کشور تاکنون مورد بررسی قرار نگرفته است و از طرف دیگر از آنجاییکه نمونه‌های گفتاری که در این تحقیق استفاده خواهد شد علاوه بر گفتار زن و مرد شامل نمونه‌های آواز نیز می‌باشد و در تحقیقات داخل کشور به آن پرداخته نشده است، این تحقیق در نوع خود جدید محسوب می‌شود و نتایج آن دست کم در کاربردهای داخل کشور قابل استفاده و استناد خواهد بود. همچنین در تحقیقاتی که تاکنون در این زمینه انجام شده از ترکیبهای مختلفی از ویژگیها نظیر مرکز ثقل طیفی (SC)[۴]، تغییرات طیفی یا شار طیفی (SF)[۵]، نقطه انحراف طیفی یا نقطهٔ قطع طیف (SR)[۶]، نرخ عبور از صفر (ZCR)[۷]، نسبت فریم‌های زمان کوتاه با انرژی کم (LEF)[۸]، ضرایب کپسترال در مقیاس مل (MFCC)[۹]، ضرایب پیشگویی خط (LPCC)[۱۰]، انرژی زمان کوتاه (STE)[۱۱]، پهنای باند (BW)[۱۲] و ... استفاده شده است. با این وجود، تا-

[1] Multi Layers Perceptron
[2] Radial Basic Function
[3] K Nearest Neighbour
[4] Spectral Centroid
[5] Spectral Flux
[6] Spectral Roll Off Point
[7] Zero Crossing Rate
[8] Low Energy Frame Rate
[9] Mel Frequency Cepstral Coefficients
[10] Linear Predictive Coefficients
[11] Short Time Energy
[12] Band Width

کنون از ترکیب ویژگیهای انحراف معیار تغییرات طیفی (VSF)[1]، درصد فریم‌های انرژی پایین (LEF)، رویداد عبور نکردن سیگنال مثبت از صفر (NZCEP)[2] و رویداد عبور نکردن سیگنال منفی از صفر (NZCEN)[3] تحقیق چندانی به منظور تشخیص موسیقی و گفتار فارسی صورت نگرفته است. همچنین لازم به ذکر است که دو ویژگی آخر بر اساس یافته‌های ما تاکنون در تحقیق‌های داخل کشور مورد استفاده قرار نگرفته است. در این تحقیق سعی بر این است که با استفاده از این ترکیب جدیدِ ویژگیها، دقتِ سیستم تشخیص گفتار فارسی از موسیقی با استفاده از طبقه‌بندی کننده‌های خودکاری مانند شبکه‌های عصبی مصنوعی و الگوریتم‌های آماری مورد بررسی قرار گیرد. در این تحقیق، از دو نوع شبکه عصبی MLP و RBF استفاده شده است. شبکه MLP بدلیل کاربرد فراوان در مسائل مربوط به طبقه‌بندی و شبکه RBF بدلیل پیچیدگی کمتر نسبت به انواع دیگر شبکه‌های عصبی انتخاب شده‌اند. همچنین از یک الگوریتم آماریِ K-NN که یک الگوریتم آماری ساده است، استفاده شده که در نهایت نتایج آنها مورد مقایسه و ارزیابی قرار گرفته است. همچنین لازم به ذکر است که تا زمان انجام این تحقیق بر اساس یافته‌های ما، از شبکه عصبی RBF به منظور تشخیص موسیقی و گفتار فارسی در تحقیقات داخل کشور استفاده نشده و اکثراً از سیستمهای طبقه‌بندی خودکاری نظیر شبکه‌های عصبی MLP، شبکه‌های عصبی فازی، الگوریتم‌های طبقه‌بندی آماری نظیر K نزدیکترین همسایگی (K-NN)، مدل مخفی مارکوف (HMM)[4]، مدل مخلوط گاوسی (GMM)[5]، ماشین بردار پشتیبان (SVM) [6]و... استفاده شده است.

[1] Variance Of Spectral Flux

[2] Non Zero Crossing Event positive

[3] Non Zero Crossing Event negative

[4] Hidden Model Markovf

[5] Gaussian Model Mixtures

[6] Support Vector Machine

۵-۱- ساختار تحقیق:

در این پایان نامه در فصل اول بطور مقدماتی به مفهوم و کاربردهای سیستم‌های تشخیص و دسته بندی سیگنال گفتار و موسیقی و سپس به موضوع و اهمیت و اهداف این تحقیق اشاره شد. در فصل دوم مروری بر یافته‌ها و نتایج تحقیقات پیشین در زمینه تشخیص سیگنال گفتار و موسیقی صورت خواهد گرفت. در فصل سوم به صورت مرحله به مرحله، روند پیاده‌سازی سیستمهای طبقه بندی سیگنال گفتار و موسیقی تشریح خواهد شد. در فصل چهارم داده‌های صوتی مورد استفاده و نحوهٔ استخراج آنها مورد بررسی قرار خواهد گرفت. پس از آن، کارایی سیستم طبقه‌بندی کننده بر اساس تغییر ویژگی‌های استفاده شده، انواع داده‌ها و تغییرات پارامترهای دو شبکه عصبی و الگوریتمِ آماری استفاده شده در این پایان‌نامه بررسی خواهد شد. در فصل پنجم ابتدا دستاوردها و نتایج را به صورت خلاصه ارائه و سپس پیشنهادات لازم جهت تحقیقات بعدی بیان می‌شود.

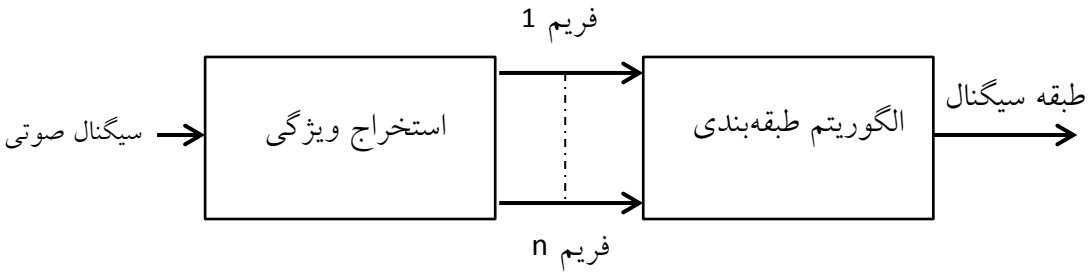
فصل دوم

مبانی نظری تحقیق

۱-۲- مقدمه:

دسته‌بندی سیگنالهای صوتی به موسیقی و گفتار، حالتی خاص از عنوان کلی "دسته‌بندی سیگنالهای صوتی" یا به اختصار ASC[1] است که در سالهای اخیر اهمیت و پیشرفت چشمگیری داشته است و مطالعات و تحقیقات گسترده‌ای دراین خصوص صورت گرفته است.

به طور کلی می‌توان بلوک دیاگرام زیر را برای یک سیستم تشخیص دهندهٔ گفتار و موسیقی در نظر گرفت که در آن ابتدا سیگنال صوتی به فریم‌هایی با طول کوچک تقسیم می‌شود و ویژگی‌ها از هر فریم، استخراج شده و سپس به الگوریتم طبقه‌بندی ارائه می‌شوند.



شکل ۱- ۲- بلوک دیاگرام سیستم تفکیک کننده گفتار و موسیقی.

۲-۲- استخراج ویژگی‌ها:

به منظور تشخیص سیگنال گفتار و موسیقی تنها در دست داشتن نمونه‌های[2] خام آنها کافی نیست زیرا این نمونه‌ها اطلاعات خاصی در موردِ نوع سیگنال بهمراه ندارند. اما از این نمونه‌های خام می توان ویژگی‌های[3] معینی را استخراج نمود. این ویژگیها را می‌توان بوسیله الگوریتم‌های ریاضی

[1] Audio Signal Classification

[2] Samples

[3] Features