دانشکده **فنی و مهندسی**

درس یادگیری ماشین

دوره **کارشناسی ارشد**

رشته **مهندسی کامپیوتر** گرایش **هوش مصنوعی**

**عنوان**

# سیستم تشخیص رویدادهای صوتی

# Non speech signal recognition

استاد راهنما

**دکتر نوشین ریاحی**

دانشجویان

**عادله تمسکنی زاهدی – فاطمه سادات علمی موسوی**

**خرداد ۱۳۹۷**


بسم الله الرحمن الرحیم

**عنوان:** سیستم تشخیص  رویدادهای صوتی

**نام و نام خانوادگی:** عادله تمسکنی زاهدی – فاطمه سادات علمی موسوی

**رشته تحصیلی:** مهندسی کامپیوتر گرایش هوش مصنوعی

**استاد راهنما:** دکتر نوشین ریاحی

**چکیده:**

در این نوشته، ابتدا مقدمه ای درباره ی تجزیه و تحلیل سیگنالهای غیرگفتاری، کاربردهای آن ، تقاوت و شباهت های سیگنال های غیرگفتاری و گفتاری بیان می شود. سپس الگوریتم کلی مورد استفاده در سیستم های تشخیص سیگنالهای غیرگفتاری و آزمایشات و نتایج پیاده سازی صورت گرفته توضیح داده می شود.

## فهرست مطلب ها

فصل ۱ :

# مقدمه

## ۱-۱- مقدمه

علاوه بر مهمترین سیگنال صوتی یعنی گفتار انسانی، تجزیه و تحلیل سایر سیگنالهای صوتی غیرگفتاری(مانند موسیقی، صداهای محیطی) مهم و مهم تر می شوند.

تشخیص سیگنالهای صوتی غیرگفتاری تکنیکی مورداستفاده در پردازش گفتار است که در آن وجود یا عدم وجود گفتار انسانی تشخیص داده می شود.

نام های دیگر این تکنیک عبارتند از : **Speech , Voice activity detection (VAD) Activity Detection(SAD)**.

استفاده مهم VAD در کدگذاری گفتار و تشخیص گفتار است. این تکنیک می تواند پردازش گفتاری را تسهیل کند و همچنین می تواند برای غیرفعال کردن برخی فرآیندها در بخش غیرگفتاری یک اتصال صوتی استفاده شود. این می تواند از کدگذاری غیر ضروری بسته های سکوت در برنامه های کاربردی Voice over Internet Protocol جلوگیری کند و در محاسبات و پهنای باند شبکه نیز صرفه جویی کند.

VADفناوری مهمی برای انواع برنامه های کاربردی مبتنی بر گفتار است. بنابراین، الگوریتم های مختلف VADتوسعه یافته اند که ویژگی های متنوع ،و سازگاری بین زمان تاخیر، حساسیت، دقت و هزینه محاسبات را فراهم می کنند. بعضی از الگوریتم های VADهمچنین تجزیه و تحلیل های بیشتری را ارائه می دهند، مثلا اینکه آیا گفتار بیان شده، بیان نشده و یا تقویت شده است. تشخیص فعالیت صدا (VAD) معمولا مستقل از زبان است.

## ۱-۲- کاربردهای سیستم تشخیص سیگنالهای غیر گفتاری

VAD بخش جدایی ناپذیر از سیستم های مختلف ارتباط گفتاری مانند زیر است:

کنفرانس های صوتی[1]

[1] audio conferencing

لغو اکو[1]

تشخیص گفتار[2]

رمزگذاری گفتار[3]

تشخیص گوینده[4]

تلفن های همراه[5]

در زمینه کاربردهای چندرسانه ای،VAD استفاده ی همزمان داده و صوت را ممکن می سازد.

به طور مشابه در سیستم جهانی ارتباطات تلفن همراه(UMTS)، نرخ بیت متوسط را کنترل و کاهش می دهد و کیفیت کلی کدگذاری گفتار را افزایش می دهد.

در کاربردهای پردازش گفتار، VAD نقش مهمی را ایفا می کند زیرا فریم های غیرگفتاری اغلب دور انداخته می شوند.

با این حال، بهبود به طور عمده بستگی به درصد مکث ها در طول گفتار و قابلیت اطمینان VADمورد استفاده برای شناسایی این فواصل زمانی دارد. از یک طرف، برای داشتن درصد کمی در فعالیت گفتاری سودمند است. از سوی دیگر، قطع کردن، یعنی از دست دادن میلی ثانیه های گفتار ، باید برای حفظ کیفیت به حداقل برسد. این یک مشکل اساسی برای الگوریتم VADدر شرایط نویزی است.

## ۱-۳- رویداد های غیرگفتاری/صوتی

به طور کلی ویژگی های رویدادهای صوتی متفاوت از گفتار است به طوری که آنها بازه ی گسترده تری از فرکانس و زمان را پوشش می دهند. با این وجود در یک چشم انداز مشابه هستند.

گفتار در معرض ساختار زمانی قرار می گیرد. به عنوان مثال، ممکن است کلمات را به واج های تشکیل دهنده ی آن تجزیه کند. به علاوه، یک رویداد صوتی می تواند به واحدهای اتمی صدا تجزیه

[1] echo cancellation
[2] speech recognition
[3] speech encoding
[4] speaker recognition
[5] hands-free telephony

شود. به عنوان مثال، صدای استفاده از شیر آب ممکن است بیشتر به صدای جریان آب در شیر آب، سپس پرتاب شدن به هوا و درنهایت ریختن در سینک تجزیه شود.

بنابراین الگوهای توالی از واحدها می تواند برای تشخیص رویدادهای مختلف صوتی مورد استفاده قرار گیرد.

با این وجود، برخلاف واج ها در گفتار، روشن نیست که چگونه باید فرهنگ لغت واحد صدا برای رمزگذاری(encode) تمام رویدادهای صوتی طراحی و کشف شود. علاوه بر این، مقایسه ی الگوهای ترتیب تحت متغیرهای بین گروهی و درون گروهی رویدادهای صوتی، کار بی اهمیتی نیست.

بسیاری از کارها بر توسعه ی بازنمایی سیگنال متمرکز شده اند. اغلب آنها از بازنمایی های گفتار از قبیل زیر استفاده می کنند.

***Mel-scale filter banks***

***Log frequency filter bank***

***Time-frequency features***

با پیشرفت سریع یادگیری ماشین، یادگیری خودکار ویژگی ها رایج تر می شوند. اما هنوز یک روش کلی برای بازنمایی سیگنالهای صوتی نداریم.

## ۱-۴- الگوریتم کلی سیستم های تشخیص سیگنالهای غیرصوتی

اکثر رویکردهای موجود SAD، در دو مرحله انجام می شوند:

**۱. مرحله ی استخراج ویژگی ها**[1]

**۲. مرحله ی دسته بندی سیگنال های غیرصوتی/صوتی**[2]

در مرحله اول، از روشهای کلاسیک برای استخراج ویژگی ها استفاده می کنند مانند:

انرژی، نرخ عبور از صفر، یا تابع خودهمبستگی

خانواده ی ویژگی های پیچیده تر که با موفقیت مورد استفاده قرار گرفته اند:

[1] Feature extraction
[2] Speech/non speech classification

**MFFCC، Multi-Band Long-Term Signal Variability ، Multi-Resolution Cohleagram**

همچنین ویژگی های مبتنی بر استفاده از شبکه های باور عمیق(DBN) نیز پیشنهاد شده است.

در مرحله دوم، از الگوریتم های طبقه بندی مختلف می توان استفاده کرد مانند

**SVM ، Gaussian Mixture Models(GMM)**

در سالهای اخیر معماری های مختلف از شبکه عصبی عمیق نیز در حال استفاده شدن هستند مانند

**Fully connected feed-forward DNNs**

**CNNs**

**RNNs**

**رویکردهای پیچیده تر مانند jointed trained DNNs یا Boosted DNNs**

**ترکیبی از DNN و CNN**

در ادامه ی مطلب ، نحوه ی پیاده سازی سیستم تشخیص سیگنالهای غیرصوتی و دیتاست مورد استفاده پرداخته می شود.

# فصل ۲:

# پیاده سازی سیستم

## ۲-۱- دیتاست مورد استفاده

دیتاست **استاندارد ESC-50** ، مجموعه ای از **۲۰۰۰ صدای ضبط شده از محیط** مناسب برای آموزش و ارزیابی طبقه بند رویدادهای صوتی محیطی می باشد.

این دیتاست شامل صداهای ضبط شده با طول ۵ ثانیه ای است و نام هر فایل طوری است که شماره آخر ، شماره کلاس می باشد.

دارای **۵۰ کلاس مختلف ( با ۴۰ نمونه در هر کلاس)** که **به ۵ گروه اصلی** تقسیم می شوند:

حیوانات مانند سگ، خروس، گربه، گوسفند

صداهای طبیعی و صدای آب مانند باران، امواج دریا، آتش سوزی، باد، آب ریختن

صدای انسان(غیر گفتاری) مانند گریه بچه، عطسه کردن، کف زدن، نفس کشیدن

صداهای داخل خانه مانند در زدن، کلیک کردن موس، باز کردن در بطری نوشابه، جاروبرقی، زنگ ساعت، شکستن شیشه

صداهای بیرونی(حمل و نقلی) مانند هلی کوپتر، قطار، آژیر خطر، بوق ماشین، صدای آتش بازی

در این پیاده سازی بخشی از این دیتاست برای آموزش و تست سیستم استفاده می شود. دیتاست مورد استفاده شامل نمونه هایی از ۱۰ کلاس صداهای خروس، شیر، گاو، جیرجیرک،گربه، مرغ، زنبورها، گوسفند، کلاغ و شرشرآب می باشد.

## ۲-۲- پیاده سازی

برای پیاده سازی این سیستم ، از نرم افزار متلب ۲۰۱۴ و ابزارهای آن استفاده می شود. این پروژه شامل فایل های train.m ، mfcc.m ، melFilterBank.m ، vqCodeBook.m ، distance.m و test.m می باشد. در فایل train ، مسیر دیتاست مربوط به آموزش به سیستم داده می شود و سپس با فراخوانی تابع mfcc، مقادیر ویژگی های mfcc را برای هر کدام از نمونه های آموزشی محاسبه می کند و سپس بر اساس برچسب داده ها، کدبوکی از بردار ویژگی ها و کد

مربوط به هر کلاس ایجاد می کند. سپس با اجرای فایل test ، مسیر داده ی تست دریافت می شود و برنامه mfcc را برای آن محاسبه می کند و سپس بردار ویژگی های داده ی تست را با هر سطر از کدبوک مقایسه می کند و میزان شباهت داده ی تست را با هر کلاس با استفاده از فاصله اقلیدسی بدست می آورد و به هر کلاس که شباهت بیشتری دارد، برچسب آن کلاس را می دهد.