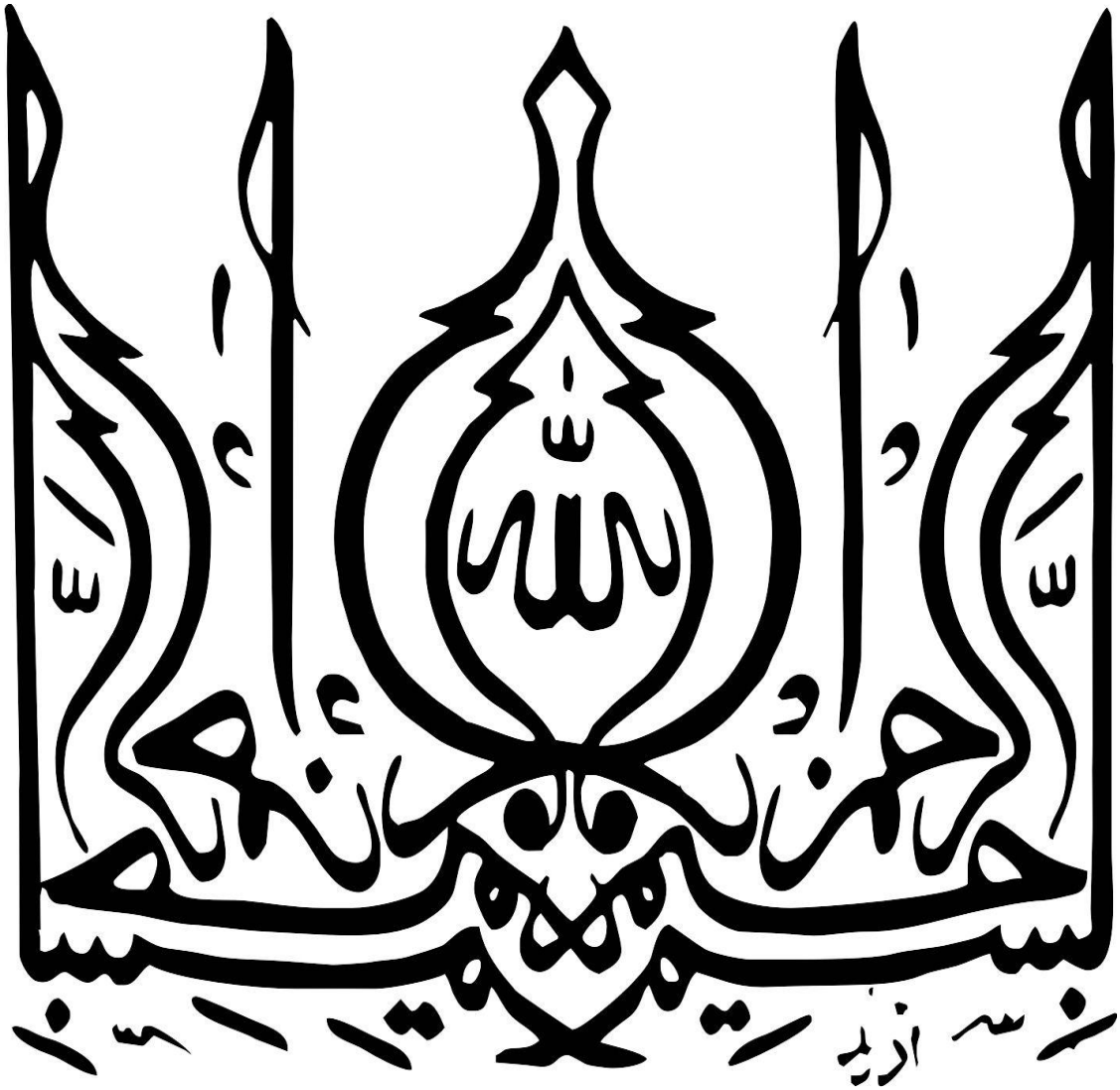

بسم الله الرحمن الرحيم

دانشکده مهندسی برق ورباتیک

پایان نامه دوره کارشناسی ارشد مهندسی برق-الکترونیک

تشخیص نوع زبان گفتاری به کمک ویژگی perceptual linear predictive و شبکه عصبی

نگارش:

محمد احدنژاد

استاد راهنما:

دکتر حسین مروی

استاد مشاور:

دکتر حسین خسروی

دی ۹۱

تقدیم به :

مادرم که با عاطفه سرشارش،روحم را از تنهایی ونومیدی رهایی می بخشد و پدرم که وجود پر مهرش در سردترین دوران،تسلی بخش جانهاست

و

همه کسانی که در راه علم آموزی مشوق من بوده اند

و

معلمانی که دانش اندکم را مدیون آن ها هستم

سپاس و قدردانی:

سپاس و حمد پروردگار را که توان اندیشیدن به ما عطا فرمود.اکنون که در این مقطع از تحصیلات به درجه کارشناسی ارشد نائل آمده‌ام بر خود لازم می‌دانم از کسانی که مرا در این مسیر یاری نموده اند تشکر و قدر دانی نمایم.از معلمانی که مرابا الفبای علم آشنا نمودند وبه من قدرت تفکر وتعمق آموختند.

از اساتید دوره کارشناسی ارشد علی الخصوص دکتر مروی به پاس راهنمایی‌های بی‌دریغشان و دکتر خسروی که به عنوان اساتید راهنما و مشاور همکاری نزدیکی با بنده داشته ومرا از نظرات سودمندشان بهره مند ساختند کمال تشکر و سپاس را دارم .

چکیده:

تشخیص خودکار زبان فرایندی است که طی آن سیستم، زبان مربوط به گفتار دیجیتال شده را تشخیص می‌دهد. در واقع وظیفه‌ی اصلی شناسایی زبان شناسایی دقیق و سریع زبان گفتاری است. یکی از مسائلی که می‌تواند به ارتباط بین مردم نواحی مختلف کمک کند و کاربردهای مختلفی در امور گردشگری، تجارت، شرکت در همایش‌ها و غیره داشته باشد، تشخیص زبان‌های مختلف از یکدیگر می‌باشد. بنابراین طراحی و ساخت سیستم هوشمندی که بتواند زبان‌ها را از یکدیگر تشخیص دهد از اهمیت ویژه‌ای برخوردار است.

در این پایان‌نامه از روش‌های مختلف استخراج ویژگی و کلاسه‌بندی جهت بهبود دقت سیستم شناسایی زبان استفاده شده است. عملکرد هرکدام با یکدیگر مقایسه و در آخر این نتایج با نتایج بدست آمده در مقالات دیگر مقایسه شده اند. دیتا بیس مورد استفاده در این پایان‌نامه OGI-TS می‌باشد. نمونه‌های صوتی این دیتابیس که ۱۱ زبان گوناگون را شامل می‌شود دارای مدت زمانی بین ۳ تا۴۵ ثانیه می‌باشند که در این پایان‌نامه از نمونه‌های ۵و ۱۰ ثانیه‌ای این دیتابیس استفاده شده است. همچنین زبان‌ها، به صورت دو، چهار و شش تایی با یکدیگر مقایسه شده‌اند. با استفاده از ویژگی‌های سطح پایین صوتی، یعنی ویژگی‌های آکوستیکی که شامل روش‌های MFCC, LPC, PLP و تبدیل فوریه‌ی بسل می‌شود ویژگی‌های مورد نظر استخراج شده و با کلاسه‌بندهای مختلف آزمایش شده‌اند.نتایج آزمایشگاهی بیانگر آن است که روش تبدیل فوریه‌ی بسل که به عنوان یک روش جدید استخراج ویژگی در تشخیص زبان مورد استفاده قرار گرفت بالاتر از روش LPC عمل می‌کند و همینطور دقت نزدیکی نسبت به بقیه‌ی روش‌ها دارد. کلاسه بندهای مورد استفاده در این پایان‌نامه، شبکه‌های MLP، RBF و شبکه‌ی عصبی جدید WRBF است. مشاهده می‌شود که استفاده از شبکه‌ی RBF و WRBF بهبود قابل توجهی در دقت

سیستم به وجود می‌آورد. همچنین روش SDC نیز بر روی روش‌های استخراج ویژگی، اعمال شده که سبب بهبود درصد صحت سیستم می‌شود.

کلمات کلیدی: تشخیص زبان، تبدیل فوریه بسل، PLP، MFCC، WRBF

فهرست عناوین:

فهرست اشکال:

فهرست جداول:

فصل اول

مقدمه

۱-مقدمه

۱-۱ زبان

زبان یکی از مهمترین ابزارها برای برقراری ارتباط میان انسان‌هاست. طبق محاسبه‌ای که زبان‌شناسان در سال ۲۰۰۰ کردند، در دنیا بیش از شش‌هزار زبان وجود دارد که اکثر آن‌ها را ده‌هزار نفر یا کمتر بکار می برند. از دیدگاه زبان شناسی دو شیوه در دسته بندی زبان مرسوم است[۱]:

- شیوه‌ی وابسته به گونه‌شناسی: در این شیوه، زبان‌ها بر پایه ساختارشان دسته‌بندی می‌شوند. به عنوان مثال، در زبان‌های چینی و انگلیسی از واژه‌پردازی "فاعل-فعل-مفعول" پیروی می‌کنند که باعث می‌شود این دو زبان از این حیث در یک گروه قرار گیرند[۱].
- شیوه‌ی زادگانی: این شیوه بسیار دشوارتر و پیچیده‌تر و علمی‌تر از روش وابسته به گونه‌شناسی است. در واقع این شیوه، زبان‌ها را بر پایه‌ی تاریخی دسته‌بندی می کند. در شیوه‌ی زادگانی، بعضی دانش‌ها مانند باستان‌شناسی، تاریخ، مردم‌شناسی و... توسط زبان‌شناسان به خدمت گرفته می‌شوند[۱].

زبان‌های مختلف یک وجه اشتراک با یکدیگر دارند و آن اشتراک چیزی نیست جز این که همگی آن‌ها از کنار هم قرار گرفتن یک سری آواها بوجود می‌آیندکه مجموعه‌ی آواها و نحوه‌ی قرارگیری آن‌ها در کنار یکدیگر و خصوصیات نوایی نظیر آهنگ و ریتم از یک زبان به زبان دیگر متفاوت است.

۲-۱ تشخیص خودکار زبان

تشخیص خودکار زبان[1] فرایندی است که طی آن سیستم، زبان مربوط به گفتار دیجیتال شده را تشخیص می‌دهد. در واقع وظیفه‌ی اصلی شناسایی زبان[2](LID)، شناسایی دقیق و سریع زبان گفتاری است[۲و۳]. معمولا انسان‌ها دقیق‌ترین سیستم‌های شناسایی زبان در جهان هستند. عموما انسان‌ها با یک تمرین کوتاه مدت با شنیدن تنها چند ثانیه از مکالمه قادر به شناسایی زبان مربوطه هستند[۴و۵]. حتی در صورتی که زبانی برای آن‌ها ناآشنا باشد، قادرند مبتنی بر تشابهات با زبان‌هایی که قبلا از آن‌ها شناخت دارند قضاوت درستی از نوع زبان مورد نظر داشته باشند. فرایند شناسایی خودکار زبان که به وسیله‌ی ماشین انجام می‌شود فواید زیادی دارد. بعضی از این فواید شامل کاهش هزینه‌ها و دوره‌های آموزش کوتاه‌تر می‌باشد.

با زیاد شدن ارتباطات اجتماعی و اقتصادی و غیره نیاز به سیستم‌های تشخیص زبان محسوس‌تر می‌شود. نیاز به سیستم‌هایی از این دست را می‌توان با مثال‌هایی از قبیل رزرو هتل، ترتیب دادن یک جلسه، گرفتن اطلاعات آب و هوا و ترافیک و راه برای سفر معرفی کرد. گرفتن این اطلاعات برای گویندگان غیر بومی می‌تواند دشوار باشد. در صورتی‌که شرکت‌های مخابراتی و تلفن‌های گویا مجهز به سیستم‌های تشخیص زبان گوینده باشند، می‌توانند خدمات مربوطه را به صورت آسان‌تری ارائه کنند. این در حالی است که نیاز به تجهیز سیستم‌های تشخیص زبان در شرایط بحرانی نمود بیشتری پیدا می‌کند. موارد زیادی گزارش شده است که متصدیان اورژانس قادر به فهمیدن زبان فرد مضطربی که پشت تلفن است، نیستند که در این موارد تشخیص سریع زبان و ترجمه‌ی آن حتی می‌تواند باعث نجات زندگی فرد شود. در امر سرویس‌دهی شناسایی زبان‌ها، در صورتی‌که تعداد زبان‌ها زیاد باشد، مسلما تعداد افراد زیادی

[1] -Language Identification Automatic
[2] - Language Identification

برای شناسایی درست این زبان‌ها نیاز می‌باشد، در حالی‌که سیستم‌های شناسایی زبان با یک بار آموزش به طور همزمان بر روی چندین ماشین جواب می‌گیرند. بنابراین نیاز به سیستم شناسایی زبان ایده‌آلی داریم که بتواند با دقت لحظه به لحظه نسبت به اطلاعات گفتار و ویژگی‌های برجسته گویش‌ها و لهجه‌ها، زبان مورد نظر را از میان خیل عظیمی از زبان‌ها شناسایی کند[۶]. می‌توان کاربردهای تشخیص زبان را در دو دسته کلی تقسیم بندی کرد. پیش‌پردازش برای سیستم‌های ماشینی و پیش‌پردازش برای شنوندگان انسانی[۷]. به عنوان مثال در مرکز مخابراتی کشور سنگاپور که دارای چهار زبان رسمی است، سیستم بازیابی اطلاعات چندزبانه که توسط صدا کنترل می شود، مورد استفاده قرار می‌گیرد. سیستم باید توانایی این را داشته باشد که اگر از گفتار به عنوان ورودی استفاده شود، زبان گوینده را قبل یا در هنگام دریافت دستورات گفتاری تشخیص دهد. تعیین زبان در حین تشخیص، تعداد زیادی بازشناس گفتار را می طلبد(برای هر زبان یک بازشناس) که به طور موازی اجرا شوند. از آنجاییکه در بعضی موارد نیاز به پشتیبانی بیش از چهار زبان است، هزینه‌ی تهیه‌ی سخت افزار چنین سیستمی بسیار بالاست. راه حل ساده‌تر این است که قبل از بازشناسی گفتار، فرایند تشخیص زبان انجام شود. در این صورت سیستم تشخیص زبان بعضی از زبان‌هایی را که احتمال کمتری می‌رود که زبان مربوطه می‌باشد را حذف می‌کند و در نتیجه تعداد کمتری از زبان‌ها که احتمال بالاتری نسبت به بقیه دارند را بر روی سخت‌افزار بار گذاری کرده و اجرا می‌کند[۳].

در فرایند شناسایی زبان، ویژگی‌های گفتار را می‌توان به دو سطح عمده تقسیم کرد: سطح گویش و سطح لغات، که ویژگی‌های گویش خود شامل صوت‌شناسی، آواشناسی و واج‌آرایی و همین‌طور علم عروضی است که می‌تواند از گفتار خام بدست آید. در حالیکه مهمترین تفاوت در سطح لغت این است که آن‌ها از کلمات گوناگون استفاده می‌کنند(دایره‌ی لغات گوناگون). هر زبان کلمات اصلی مربوط به فرهنگ

لغت و در نتیجه ویژگی‌های سطح لغت، شامل ریخت‌شناسی، علم نحو و اطلاعات گرامری مربوط به خودش را دارد.

همان‌طور که پیشتر اشاره شد، برای تشخیص خودکار زبان نیاز به سیتمی دقیق می‌باشد. ویژگی اصلی یک سیستم ایده‌آل این است که نسبت به هیچ زبانی(یا دسته‌ای از زبان‌ها) تحت تاثیر قرار نگیرد و یا تمایل به زبان خاصی نداشته باشد، محاسبات زمانی کوتاهی داشته باشد، افزایش تعداد زبان‌های هدف یا کاهش طول گفتار، عملکرد سیستم را ضعیف نکند و همچنین در برابر گوینده و نویزهای دیگر مقاوم باشد. با وجود پژوهش‌های مستمر در سیستم‌های تشخیص زبان و پیشرفت‌های سریع در این زمینه و بهبود قابل توجه عملکرد این سیستم‌ها طی ۱۰ الی ۲۰ سال اخیر، هنوز مشکلات زیادی در این زمینه وجود دارد. به عنوان مثال، تعداد محدود داده‌های گفتار چند زبانه قابل دسترسی برای آموزش سیستم‌ها، تعداد محدود زبان‌هایی که سیستم‌های کنونی قادر به شناسایی آنها هستند(معمولا ۱۰ تا ۱۵ زبان از مجموع ۶۹۰۰ زبان زنده‌ی دنیا) و همچنین گنجاندن لهجه‌های مختلف یک زبان در داخل آن زبان برای شناسایی زبان مشکلاتی هستند که برطرف کردن آن‌ها نیاز به کار وسیعی دارد. ضعف قابل توجه دیگر در بیشتر سیستم‌های رایج کنونی این است که این سیستم‌ها برای نمونه‌های بین ۳۰ الی ۴۵ ثانیه عملکرد خوبی دارند، اما معمولا هنگامی که زمان نمونه‌ها به ۳ الی ۱۰ ثانیه تنزل پیدا می‌کند عملکرد ضعیفی از خود نشان می‌دهند. این در حالی است که در موقعیت‌های اضطراری(مثل تماس با اورژانس)، زمان شناسایی کوتاه‌تر ترجیح داده می‌شود.

در این پایان‌نامه ویژگی‌های صوتی مانند MFCC[1] و PLP[2] و LPC[3] و تبدیل فوریه‌ی بسل از نمونه‌های۵ و ۱۰ ثانیه‌ای دیتابیس OGI-TS استخراج می‌شوند. سپس درصد صحت آن‌ها با استفاده از

[1] -Mel-frequency cepstral coefficients
[2] -Perceptual Linear Predictive
[3] -Linear Prediction Cepstral

شبکه‌های عصبی MLP[1] و RBF[2]و WRBF[3] سنجیده می‌شود. همچنین ویژگی SDC[4] که یکی از رایج ترین ویژگی‌های آکوستیکی در تشخیص زبان است مورد استفاده قرار گرفته و بررسی می‌شود.

در فصل دوم تاریخچه‌ی تشخیص زبان مرور می‌شود و نگاهی گذرا به روش‌های انجام شده در این راستا خواهد شد. در فصل سوم ضمن آشنایی با فرایند تولید گفتار، نگاهی اجمالی به سیستم‌های شناسایی گفتاری نیز شده و روش‌های استخراج ویژگی توضیح داده می‌شود. همچنین در فصل چهارم به بررسی شناساگرهای مختلف زبان و تشریح شبکه‌های عصبی بکار رفته در این پایان نامه پرداخته می‌شود. پس از آن در فصل پنجم پیاده سازی و نتایج حاصل از آن‌ها را شرح داده و در نهایت، در فصل ششم جمع‌بندی از کارهای انجام شده صورت می‌گیرد و پیشنهاداتی برای کارهای آتی داده می‌شود.

[1] -Multilayer Perceptron
[2] -Radial Basis Function
[3] -Weighted RBF
[4] -Shifted Delta Cepstrum

فصل دوم

مرور تاریخچه

۲- مرور تاریخچه

در این فصل سیر تکامل سیستم‌های تشخیص زبان و تکنیک‌هایی که محققان در جهت بهبود دقت این سیستم‌ها ارائه داده‌اند ذکر شده است. اکثرا تنها به ذکر نام سیستم‌های تشخیص زبان یا توضیحی مختصر در رابطه با آنها اکتفا شده است.

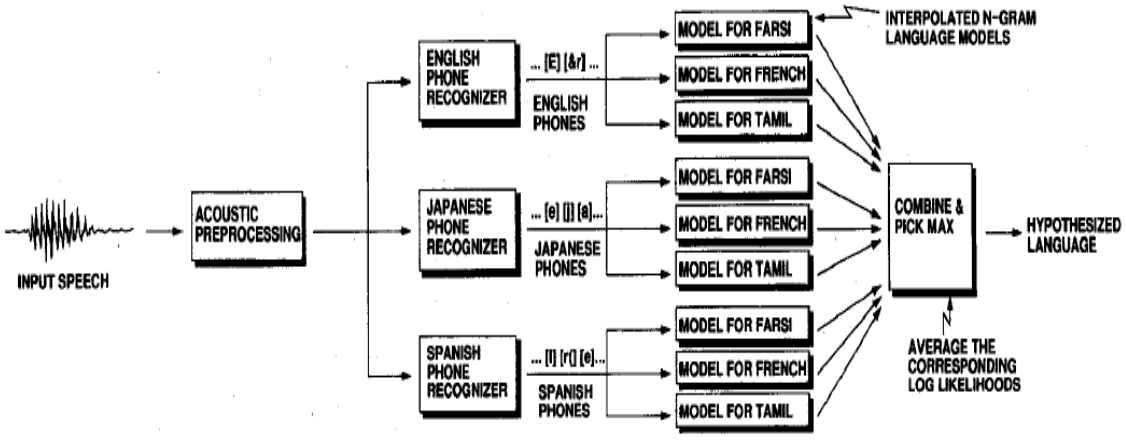
تشخیص خودکار زبان، زمینه‌ای است که طی چندین سال اخیر مورد توجه قرار گرفته و از حدود ۳۰ سال پیش تحقیقات وسیعی در این مورد انجام شده. شروع این فعالیت‌ها در دهه ۱۹۷۰ بود. به نحوی که در سال ۱۹۷۴ لئونارد[۱] و دودینگتون[۲] کار تشخیص زبان را با روش فیلترسازی صوتی[۳] شروع نمودند[۸]. در سال ۱۹۷۷ هوس[۴] و نئوبورگ[۵] با استفاده از روش واج‌آرایی یکی از تاثیرگذارترین افراد در زمینه‌ی تشخیص خودکار زبان بودند[۹]. در این قسمت برخی از مقالات آورده شده است. در این مقالات به بعضی روش‌های متفاوت کاربردی برای تشخیص خودکار زبان اشاره شده است.

در مقاله‌ای که در سال ۱۹۹۶ توسط آقای زیسمن منتشر شد، چهار روش پایه برای تشخیص زبان به صورت کامل شرح داده و از نظر کارایی با یکدیگر مقایسه شد. این روش‌ها شامل روش‌های PRLM[۶]، PPRLM[۷]، PPR[۸] و GMM[۹] بود. این سیستم‌ها بر روی مجموعه داده‌ی OGI-TS مورد ارزیابی قرار گرفتند و روش PPRLM نسبت به سایر روش‌ها کارایی بالاتری از خود نشان داد و برای داده‌های تست با طول ۴۵ ثانیه و ۱۰ ثانیه و برای تشخیص بین ۱۰ زبان ، به ترتیب به خطای ۲۱ و ۳۷ درصد

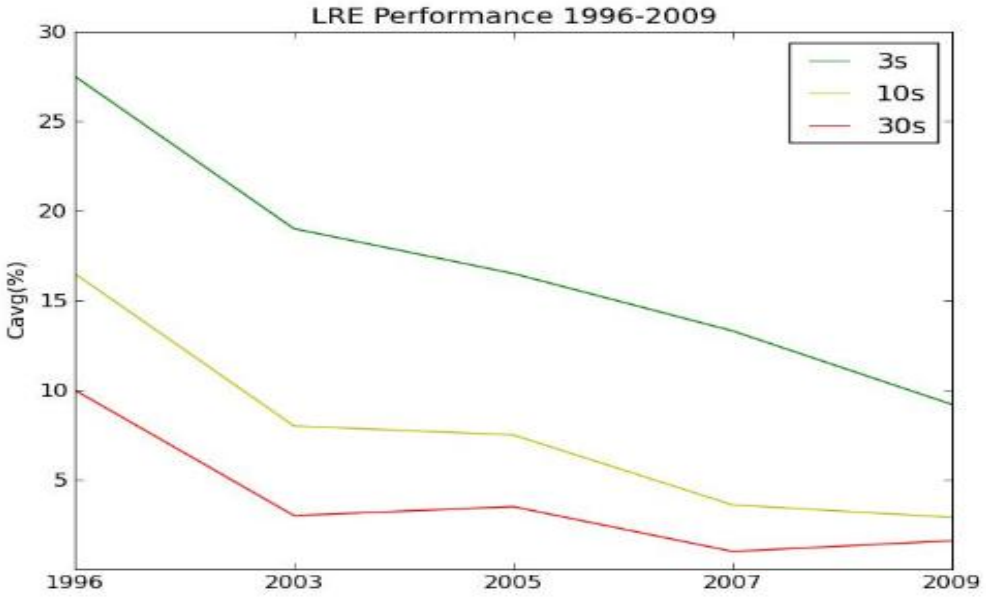
[1] -Leonard
[2] -Doddington
[3] -Acoustic filter bank
[4] -House
[5] -Neuberg
[6] -Phone Recognition followed by Language Modeling
[7] -Parallel Phone Recognition followed by Language Modeling
[8] - Parallel Phone Recognition
[9] -Gaussian Mixture Model

دست یافت. سیستم PPRLM مورد استفاده در این آزمایش از شش بازشناس آوایی به زبان‌های انگلیسی، آلمانی، هندی، ژاپنی، ماندارین و اسپانیایی تشکیل شده بود. در این مقاله برای بهبود سیستم PPRLM از بازشناس‌های آوایی وابسته به جنسیت استفاده شد. مدل‌سازی آکوستیکی وابسته به جنسیت، یک روش مشهور برای بهبود کارایی شناسایی گفتار است. در این صورت می‌توان با استفاده از بازشناس‌های آوایی وابسته به جنسیت، دنباله‌ی آوایی قابل اعتمادتری را تولید نمود و در نتیجه مدل‌های زبانی بهبود پیدا خواهند کرد. این سیستم برای بازشناس آوایی زبان انگلیسی در شکل ۲-۱ نشان داده شده است. در این حالت سه بازشناس آوایی با استفاده از سیگنال‌های گفتار زن، مرد و ترکیب آن‌ها، برای هر زبانی که داده‌ی برچسب‌خورده برای آن موجود است، آموزش داده می شود[۱۰].

کارهای دیگری که پس از انتشار این مقاله صورت گرفت، معمولا کارهایی بود که کارایی خود را با سیستم‌های تشخیص زبانی که در این مقاله ارائه شده، مقایسه نمودند و سعی کردند که این روش‌ها را بهبود بخشند. همچنین مسابقاتی که جهت ارزیابی سیستم‌های تشخیص زبان از سال ۱۹۹۶ تا کنون هر دو سال یک‌بار توسط NIST برگزار می‌شود، بستری برای بهبود سیستم‌های تشخیص زبان بود. طبق نتایج ارائه شده توسط این موسسه، کارایی سیستم‌های تشخیص زبان هر دوره بهبود پیدا می‌کند. این روند بهبود در شکل ۲-۲ نشان داده شده است. در ادامه برخی از کارهایی که در زمینه بهبود سیستم‌های تشخیص زبان صورت گرفته، معرفی شده است.



شکل۲-۱: یک مثال از سیستم PPRLM با بازشناس های آوایی وابسته به جنسیت (در این شکل تنها بازشناس زبان انگلیسی نشان داده شده است)[۱۰]



شکل۲- ۲: کارایی بهترین سیستم تشخیص زبان در دوره های مختلف NIST LRE [۱۱]

بهبود سیستم‌های تشخیص زبان، از جنبه‌های مختلفی صورت گرفته است. این بهبودها را می‌توان از نقطه نظر ویژگی‌های آکوستیکی و عروضی مورد استفاده برای تشخیص زبان، بازشناس‌های

آوایی، مدل‌های زبانی، نحوه‌ی ترکیب نتایج حاصل از مدل‌های زبانی، استفاده از طبقه‌بندهای انتهایی، استفاده از طبقه‌بندهای تمایزی و ترکیب سیستم‌های تشخیص زبان مختلف مورد بررسی قرار داد. در بحث ویژگی‌های آکوستیکی و عروضی مورد استفاده برای تشخیص زبان کارهای مختلفی صورت گرفته است. در این کارها برای تشخیص زبان از ویژگی MFCC و مشتقات آن استفاده می‌شد.

در سال ۲۰۰۱ آقای فاریناس[1] از ویژگی عروضی ریتم[2] استفاده کرد و برای مدل سازی آن از الگوهای شبه‌هجایی CV استفاده کرد و آن‌ها را به بردارهای ویژگی سه بعدی، برای مدل کردن توسط GMM با ۱۲ مولفه‌ی مخلوط گاوسی تبدیل کرد. بعد اول این بردار ویژگی مجموع دیرش همخوان‌ها، بعد دوم آن دیرش واکه و بعد سوم را تعداد همخوان‌های موجود در الگوی شبه هجایی در نظر گرفت[۱۲].

آقای لین[3] از اطلاعات کانتور پیچ برای تمایز بین زبان‌ها استفاده کرد و کانتور پیچ را با استفاده از چند جمله‌ای‌های لژاندر تخمین زد و ضرایب این چند جمله‌ای‌ها را به عنوان ویژگی، مورد استفاده قرار داد و این بردارها را با استفاده از GMM مدل نمود. طبق نتایج بدست آمده در این مقاله، دو یا سه ضریب اول این چندجمله‌ای‌ها، برای رسیدن به نتایج خوب، کافی است. همچنین از طول کانتور پیچ نیز می‌توان به عنوان ویژگی جهت بهبود کارایی استفاده کرد[۱۳].

آقای تیموشنکو الگوهای شبه‌هجا را از سیگنال گفتار استخراج کرد و دیرش آن‌ها را به عنوان ویژگی مورد استفاده قرار داد[۱۴].

[1] -Farinas
[2] -Rhythm
[3] -Lin

رواس اطلاعات کانتور گام، انرژی و دیرش را با استفاده از مدل‌های زبانی آماری، مدل نمود. به این‌ترتیب که پس از تقسیم‌بندی کانتور گام به واحدهای شبه هجا، به ازای هر شبه هجا یک خط با استفاده از روش رگرسیون خطی تخمین زده شد. سپس مینیمم‌های محلی هر شبه هجا بر اساس خط تخمین زده شده و مشخص و با استفاده از یک سری خطوط به یکدیگر وصل شدند. در مرحله‌ی نهایی نیز هر شبه هجا بر اساس شیب خطوط حاصل به توالی از U(شیب مثبت)، D(شیب منفی) یا # (شیب صفر) کد شد. کد کردن کانتور انرژی نیز به همین روش صورت گرفت. در مورد دیرش، ابتدا دیرش میانگین روی کل قطعه‌ی گفتاری محاسبه شد و بر اساس کمتر یا بیشتر بودن دیرش هر شبه هجا نسبت به دیرش میانگین، قطعه‌ی گفتاری به توالی از s(کمتر از میانگین) و l (بیشتر از میانگین) کد شد[۱۵].

در مقاله‌ای دیگر از میان ۶۳ ویژگی عروضی مختلف که روی یک یا توالی از شبه هجا تخمین زده شدند، ۲۰ ویژگی که کارایی بالاتری در تشخیص زبان از خود نشان دادند انتخاب شدند و یک روش با نام PAM[1] برای مدل کردن ویژگی‌های عروضی ارائه شد[۱۶].

آقای تورس[2] از ویژگی SDC برای آموزش مدل‌های مخلوط گاوسی استفاده کرد و به کارایی بالاتری دست یافت. با استفاده از ویژگی MFCC، سیستم‌های تشخیص زبانی که تنها از اطلاعات آکوستیکی استفاده می‌کردند، کارایی پایین‌تری نسبت به سیستم‌های تشخیص زبانی که بر مبنای اطلاعات آوایی عمل می‌کردند، داشتند. اما با استفاده از ویژگی SDC کارایی این سیستم‌ها در حد سیستم‌های تشخیص زبان آوایی قرار گرفت[۱۷]. آلن در تحقیقاتش یک روش جدید برای محاسبه‌ی SDC ارائه داد و آن را ویژگی SDC تغییر یافته بر مبنای درون‌یابی نامید و یک سیستم تشخیص زبان آکوستیکی با استفاده از GMM توسعه داد و قدرت تفکیک‌کنندگی ویژگی‌های MFCC ,PLP و

[1] - prosodic attribute model
[2] -Toress

MODGDF[1] و ترکیب آن‌ها را مورد بررسی قرار داد. همچنین از تکنیک feature warping برای نرمال‌سازی ویژگی‌های حاصل استفاده کرد[۱۸].

همچنین بعد از آن ژیرونگ یانگ[2] نیز با تحقیقی در مورد شناسایی اتوماتیک زبان‌ها و با مقایسه‌ی سه روش PRLM و PPRLM و GMM به نتیجه‌ای مشابه با زیسمن رسید که در آن برتری روش PPRLM را ثبت کرد. به این ترتیب که در داده‌های با طول صفر تا ۱۰ ثانیه روش PPRLM بهبود ۱۶درصدی را نسبت به روش PRLM داشت و این بهبود برای داده‌های با طول ۱۰ تا ۳۰ ثانیه ای ۱۲درصد و برای داده های ۳۰ تا ۴۵ ثانیه ای ۱۴درصد بود. او همچنین بررسی کرد که GMM با ۴۰ مؤلفه، بهبودی بین یک تا پنج در صدی را برای داده‌های با طول صفر تا ۶۰ ثانیه نسبت به GMM با ۱۶ مؤلفه دارد[۱۹].

تانگ ترکیب پنج ویژگی در سطوح انتزاع مختلف را برای تشخیص زبان مورد بررسی قرار داد. ویژگی‌های طیفی، دیرش، پیچ، n تایی آوایی و Bag-of-Sound و به این نتیجه دست یافت که سطوح مختلف اطلاعات، مکمل یکدیگرند و اطلاعات آکوستیکی در حالتی که طول بازه‌ی گفتار کوتاه‌تر باشد، موثرتر است. برای طول بازه‌های طولانی‌تر، اطلاعات آوایی کارایی بالاتری دارند[۲۰].

در مبحث استخراج ویژگی آقای مگنوس راسل[3] مقاله‌ای را در سال ۲۰۰۶ تحت عنوان "مقدمه‌ای بر پیش‌پردازش و ویژگی‌های صوتی برای تشخیص اتوماتیک گفتار" منتشر کرد که در آن مراحل مختلف استخراج ویژگی‌های PLP و MFCC بررسی شد. سپس با مقایسه‌ای که برگرفته از مقاله‌ی مینلر[4] بین این دو روش انجام داد شباهت‌های بین این دو الگوریتم را مطرح کرد و این ایده را با

[1] - Modified group delay function
[2] -Zhirong Yang
[3] -Magnus Rosell
[4] -Minler

پیاده سازی ادامه داد. دست آخر نتیجه‌ی این مقاله را این‌طور مطرح کرد که ویژگی MFCC با فیلتر RASTA بهترین نتیجه را می‌دهد[۲۱].

در همین سال بویین[1] پیشنهاد جدیدی در مورد ترکیب ویژگی‌های طیفی و ویژگی‌های عروضی در شناسایی زبان عنوان کرد که این پیشنهاد بر روی یک سیستم شناسایی زبان مبتنی بر GMM-UBM، بهبودی قابل توجهی به‌وجود آورد. این تکنیک در واقع از SDC پیشرفته وترکیب ویژگی‌هایی استفاده کرده است که درصد صحت را به میزان ۱۲درصد بهبود می‌بخشد[۲۲].

از دیگر مقاله‌هایی که در سال ۲۰۰۶ به چاپ رسید مقاله‌ی آقای یو مین زنگ[2] بود که درآن موضوع کلاسه‌بند وابسته به جنسیت افراد(مرد یا زن)، با استفاده از GMM بر روی پارامترهای پیچ[3] و RASTA-PLP گفتار بحث شده است. عملکرد این ایده‌ی جدید که همان کلاسه‌بند وابسته به جنسیت است بسیار عالی می‌باشد. این سیستم در مقابل نویز بسیار قدرتمند است. درصد صحت این کلاسه‌بند برای داده‌های تمیز(بدون نویز)، بالای ۹۸درصدو برای بیشتر داده‌های نویزی ۹۵درصد است[۲۳].

آقای فاریس در تشخیص زبان از مدل GMM استفاده کرد، با این تفاوت که در مرحله‌ی آموزش مدل‌ها از بعضی داده‌های آموزشی استفاده کرد. بدین صورت که در آغاز یک مجموعه را آموزش داد. بعد از آن کل نمونه‌های مجموعه آموزشی را ارزیابی کرده و با توجه به معیار آنتروپی یا احتمال نسبی، آن‌هایی را که بر اساس میزان عدم قطعیت تعیین می‌شود، مرتب کرده و درصدی از این لیست را به مجموعه آموزشی پیشین اضافه کرده و بر مبنای این مجموعه آموزشی جدید تا مادامی‌که دقت مجموعه‌ی ارزیابی بالا نرفت، مدل ها را آموزش داده و این کار را ادامه دهند[۲۴].

[1] -Bo Yin
[2] -Yu-Min Zeng
[3] -pitch

از جمله مقاله‌های دیگری که در این سال یعنی ۲۰۰۸منتشر شد مقاله‌ی آقای سو بود که این پژوهشگر نیز از روش PPRLM در مقاله‌ی خود سود برد. در این سیستم ابتدا بردارهای ویژگی ۳۹ بعدی از داده‌های ورودی که شامل بردارهای ویژگی پیش‌گویی خطی ادراکی در معیار فرکانس مل است را استخراج کرده، سپس مشتق اول و دوم این ویژگی‌ها به یک سری بازشناس‌های آوایی داده شده و توالی آوایی حاصل از این بازشناس‌ها با استفاده از مدل‌های زبانی، مدل‌سازی می‌شوند که این مدل‌ها به تفکیک جنسیت آموزش داده شدند. سه طبقه‌بند مدل مخلوط گاوسی، ماشین بردار پشتیبان و شبکه‌ی عصبی برای تصمیم‌گیری نهایی بر اساس خروجی‌های این مدل زبانی به کار برده شده و بهترین نتیجه، از سیستم PPRLM ای که از ماشین بردار پشتیبان در طبقه‌بندی نهایی خود استفاده کرده حاصل شد[۱۱].

در سال۲۰۱۰ آقای ژانگ[1] ویژگی TFC[2] را به عنوان ویژگی جدید برای تشخیص زبان ارائه داد، نحوه‌ی محاسبه‌ی این ویژگی مشابه SDC است، با این تفاوت که از روش DCT[3] به عنوان روشی برای فشرده‌سازی توالی بردارها استفاده می‌کند. نتایج حاصل نشان داد که دقت این ویژگی بالاتر از SDC است[۲۶].

یکی از جدیدترین مقاله‌هایی که در زمینه شناسایی زبان ارائه داده شده با عنوان "پیشنهادی برای استخراج ویژگی مستقل از محتوا برای تشخیص زبان‌های گفتاری چینی و کره ای"[۲۷] است که توسط لی شی دان[4] در سال ۲۰۱۱ نوشته شده. در روش جدیدی که این مقاله پیشنهاد کرده است گفته شده که ابتدا سیگنال گفتار به فریم‌های متوالی تقسیم می‌شود. نسبت بین تعداد عبور از صفر در بازه‌ی زمانی کوتاه و انرژی در همان بازه‌ی زمانی محاسبه می‌شود. در این صورت نرخ انرژی فرکانس زمان

[1] -Zhang
[2] -Time Frequency Cepstral
[3] -Discrete Cosine Transform
[4] -LU Shi-Dan

کوتاه[1](STFER) محاسبه شده و میانگین STFER در هر فریم به عنوان ویژگی کلاسه‌بند، برای پیاده‌سازی شناسایی زبان کره‌ای وچینی استفاده می‌شود. سرانجام آستانه، از اطلاعات بهره تعیین می‌شود. نتایج آزمایش‌های انجام شده در این روش پیشنهادی نشان می‌دهد که سادگی این روش بیشتر از پارامترهای MFCC است و برای ویژگی‌های با پیچیدگی کمتر قدرت شناسایی بیشتری دارد[۲۷].

در ایران و در دانشگاه پردازش هوشمند گفتار دانشگاه صنعتی امیرکبیر، کانتور گام با استفاده از چند جمله‌ای‌های لژاندر تخمین زده شد و سپس ضرایب حاصل با استفاده از مدل‌های GMM مدل شده و این مدل‌ها در کنار مدل‌های GMM حاصل از ضرایب SDC در قالب سیستم GMM-PSK-SVM مورد استفاده قرار گرفت. ارزیابی‌های صورت گرفته بر روی دادگان OGI نشان داد که ترکیب ویژگی‌های عروضی در کنار ویژگی‌های SDC سبب افزایش دقت این سیستم می‌شود[۲۸].

از جمله کارهای دیگری که در ایران شده توسط آقای ضیایی انجام شده است که از سه ویژگی جدید مرکز ثقل طیفی، آنتروپی شانون[2] و رنیی[3] در کنار ویژگی‌های MFCC، گام و شدت استفاده و یک بردار ویژگی ۱۴ بعدی ایجاد شد و سپس از SDC برای بدست آوردن بردار ویژگی که تغییرات زبانی را بهتر مدل کند، استفاده شد. با ترکیب این ویژگی، دقت تشخیص زبان بر روی دادگان OGI به میزان تقریبا پنج درصد نسبت به حالتی که از ویژگی MFCC، گام و شدت استفاده شود، بهبود پیدا کرد[۲۹].

آقای مرتضی نیا از مدل GMM برای بازشناسی زبان‌های رایج در ایران چون ترکی، عربی استفاده کرد و تاثیر پارامترهای مختلف از جمله حجم داده‌های آموزشی و تعداد دوره‌های الگوریتم آموزش بر کارایی این سیستم را مورد بررسی قرار داد[۳۰].

[1] -Short-time-frequency-energy-ratio
[2] -Shanon
[3] -Renyi

آقای حسینی سیستم ARLM را ارائه کرد که در آن از یک روش رای‌گیری برای تبدیل دنباله‌های آوایی تولید شده به یک دنباله‌ی آوایی واحد استفاده شده است. این سیستم نسبت به سیستم PPRLM بر روی دادگان OGI، ۱/۳ درصد بهبود داشت[۳۱].

در مقاله‌ای دیگر توسط آقای حسینی سیستم تشخیص زبان GPPRLM ارائه شد که نسبت به سیستم PPRLM و APRLM از دقت بالاتری برخوردار است و در آن از یک بازشناس مستقل از زبان که با استفاده از کلیه‌ی داده‌ها آموزش داده شده، استفاده شده است[۳۲].

مهارلویی در دانشگاه صنعتی شاهرود نیز به تحقیق برروی طراحی سیستم شناسایی زبان گفتاری پرداخت. روشی که در این پایان‌نامه کار شده، مستقل از اطلاعات واج‌آرایی بوده است. مهارلویی از تبدیل موجک و تبدیل کپسترال نمونه‌های صوتی استفاده کرد که بدون نیاز به اطلاعات زبان‌شناسی، بر روی زبان‌های گوناگون قابل استفاده باشد. در این روش برای استخراج ضرایب موجک ابتدا این ضرایب را محاسبه و پس از انجام کاهش ابعادی، ویژگی‌های مفیدتر را استخراج نمود و آن‌ها را به برنامه‌ی اصلی داد. در مورد ضرایب کپسترال عنوان شده که به علت تعداد محدود آن نیاز به کاهش ابعادی نمی‌باشد. سپس این ضرایب را با استفاده از الگوریتم بازگشتی و تکنیک گسسته سازی چند بازه‌ای نام‌گذاری کرد. همچنین با استفاده از تکنیک رتبه‌بندی ویژگی‌ها برای بهبود درصد صحت تشخیص اقدام به تلفیق ویژگی‌های اولیه و ویژگی‌ها به صورت ترکیبی کرد. در ادامه با ارزیابی برنامه و مقایسه‌ی آن با دو روشی که توسط گومینز و رواس انجام شد، به نتیجه‌ای بهتر از این دو روش می‌رسد، به صورتی‌که در نمونه‌های ۱۰ ثانیه‌ای ضرایب PLP با حدود ۷۵ درصد بالاترین صحت را دارد و پس از آن و در اکثر موارد به ترتیب، LPC، موجک و در نهایت MFCC ضریب صحت بهتری دارند، همچنین در مقایسه با روش گومینز و رواس در بهترین مورد تا حدود ۱۵ درصد نیز دقت تشخیص ارتقا می‌یابد[۳۳].

فصل سوم

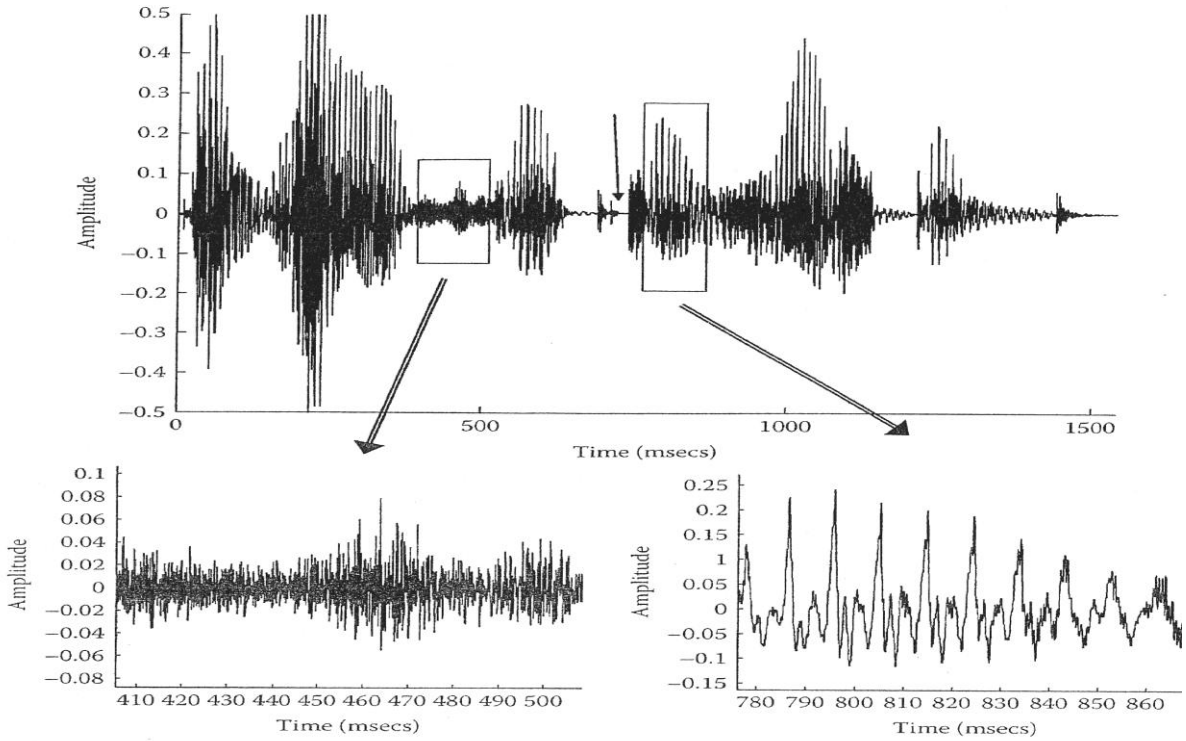
نگاهی اجمالی به سیستم‌های شناسایی زبان گفتاری

۳- نگاهی اجمالی به سیستم‌های شناسایی زبان گفتاری

۳-۱ پیش زمینه‌های تئوری سیگنال گفتار

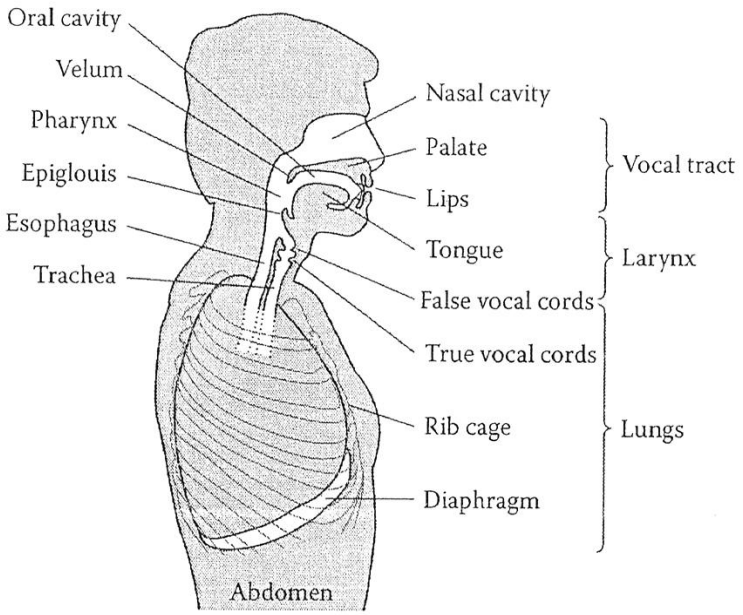
قبل از شروع تشریح الگوریتم‌های تشخیص زبان، مهم است که با سیگنال گفتار و مخصوصا با فرآیند تولید گفتار، بیشتر آشنا شویم و اشاره‌ای به تنوع آکوستیکی که به درک گفتار مربوط می‌شود، داشته باشیم.

سیگنال گفتار یک سیگنال غیرایستان است که توان دوم آن(طیف توان) در طول زمان تغییر می‌کند. هنگامی که بررسی دقیق‌تری می‌کنیم در دوره‌های زمانی کوچک، مشخصات آن تقریباً ایستان هستند. شکل ۳-۱ مثالی از شکل موج زمانی یک جمله به زبان انگلیسی را نشان می‌دهد.



شکل۳-۱: شکل موج زمانی جمله "The wife helped her husband" که به وسیله یک مرد تولید شده است[۳۴]

شکل موج سیگنال به تعدادی قاب[1] یا فریم تقسیم شده که متناظر با تعداد کلمات گفته شده و نوع صدا می‌تواند متفاوت باشد. همانطور که در شکل ۳-۱ نشان داده شده است برخی از قسمت‌های گفتار شبه متناوب هستند. برای نمونه شکل موج سیگنال در طول هجای "er" در"her " در قسمت‌هایی می‌تواند غیرمتناوب شبه نویزی باشد و یا در طول تولید صامت "f" در "wife" قسمت‌هایی نیز می‌تواند شامل سکوت یا فاصله بین جملات باشد(برگشت در شکل ۳-۱). برخی از قسمت‌ها دارای شدت زیاد هستند(صدای "i" در "wife") و از سوی دیگر برخی از قسمت‌ها شدت کمتری دارند(مثل صامت "f" در "wife"). به صورت کلی دوره‌ی زمانی، شدت و طیف هر قسمت می‌تواند بسته به گوینده و در طول صحبت تغییر کند. قسمت‌های بعدی گفتار، شبه نویزی بدون صوت است که به صورت عادی در گفتار سلیس یافت می‌شود و شدت و طول و مشخصات فرکانسی آن متغیر می‌باشد. در ادامه شرح می‌دهیم که این قسمت‌های سیگنال گفتار چگونه به وسیله سیستم تولید گفتار انسان ساخته می‌شود.



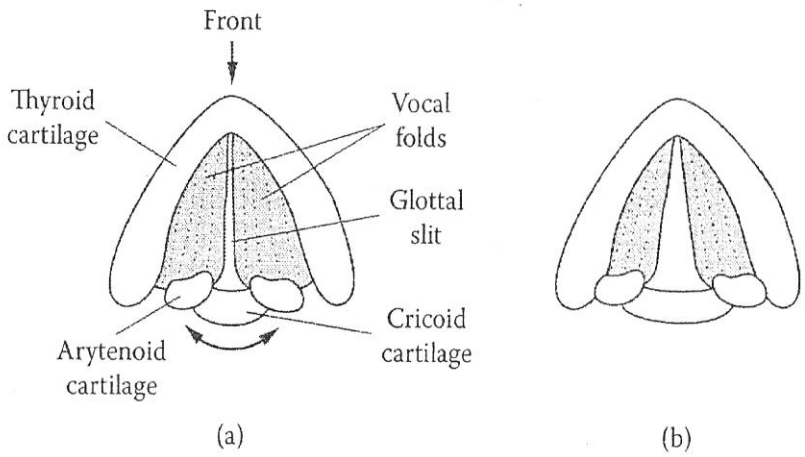
شکل۳-۲: یک تصویر مقطعی از آناتومی تولید صوت[۳۴]

[1] -Frame

شکل ۳-۲ یک تصویر مقطعی از آناتومی تولید گفتار را نشان می‌دهد. همان‌طور که نشان داده شده است، تولید گفتار تعدادی از اعضا و عضلات مانند حنجره را در بر می‌گیرد.

۳-۱-۱-حنجره

حنجره متشکل از عضلات، رباط‌ها و غضروف‌هایی است که عمل تارهای صوتی را انجام می‌دهد. این عضو متشکل از دو توده رباط وعضلات فشرده شده از قسمت عقب به جلوی حنجره است.



شکل ۳-۳: طرحی از تارهای صوتی که پایین حنجره وجود دارند.(a) حالت صدادار، (b) حالت بی صدا[۳۴]

قسمت عقبی تارها به دو غضروف متصل شده که با غضروف جنبی حرکت می‌کند (شکل ۳-۳) تارهای صوتی را می‌توان در سه حالت فرض نمود:

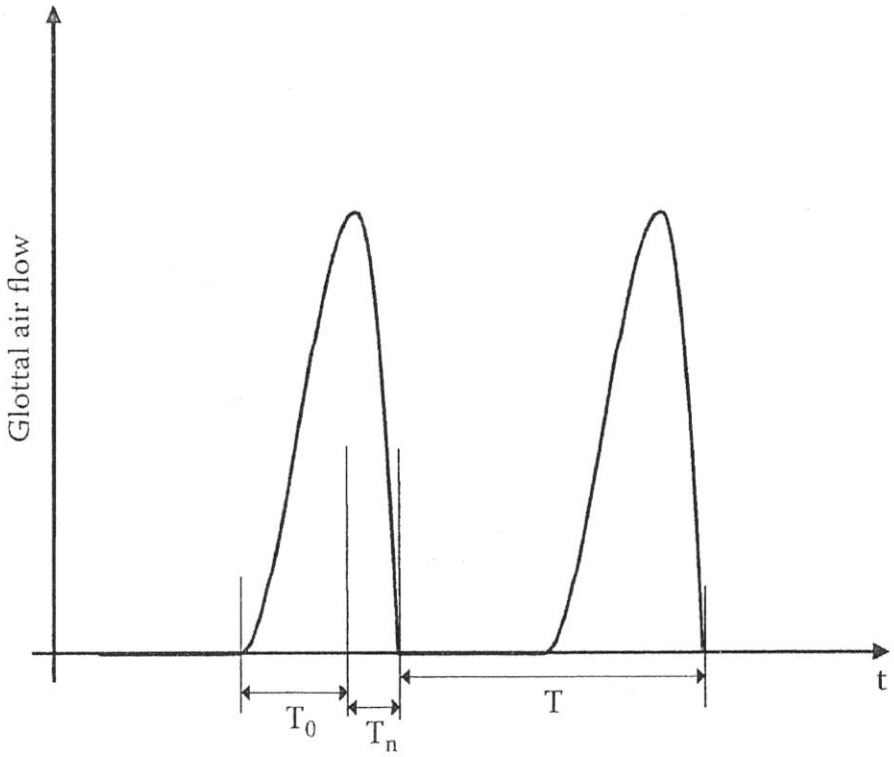
- تنفس
- گفتار صدادار
- گفتار بی‌صدا

درحالت تنفس که هوا از ریه به صورت آزادانه به سمت دهانه‌ی حنجره حرکت می‌کند (که کاملا باز است) مقاومت قابل توجهی از تارهای صوتی به آن وارد نمی‌شود. در حالت گفتار صدادار، در طول تولید حروف صدادار (مثل"aa") غضروف‌ها به سمت همدیگر حرکت می‌کنند و تارهای صوتی به همدیگر نزدیک‌تر می‌شوند. افزایش یا کاهش فاصله بین تارها سبب کاهش و افزایش فشار در دهانه حنجره می‌شود که همین باعث باز و بسته شدن متناوب تارهای صوتی می‌شود. باز وبسته شدن تارهای صوتی به صورت متناوب را می‌توان با استفاده از اصل برنولی در دینامیک سیالات توضیح داد[۳۵].

دو فشار برای ارتعاش تارهای صوتی مورد نیاز است:

۱. فشار هوایی که به قسمت پایین تارها وارد شده و آن‌ها را از هم جدا می‌کند.

۲. فشار منفی که سبب عبور هوا از بین تارها می‌شود (اثر برنولی).



شکل۳-۴: شکل موج هوای نای. دوره پیچ با T نشانه گذاری شده است[۳۴]

اگر سرعت عبور جریان هوا در حنجره را تابعی از زمان بدانیم، به یک شکل موج شبیه شکل ۳-۴ می‌رسیم. هنگامی که تارها بسته می‌شوند، جریان هوا به آرامی شروع می‌شود و به مقدار حداکثر می‌رسد، سپس یک‌دفعه به صفر کاهش پیدا کرده و در آن نقطه است که تارها بسته شده‌اند. هنگامی که تارها بسته شدند، هیچ جریان هوایی از کانال صوتی عبور نمی‌کند و در این فاز حنجره بسته است. دوره‌ای که در طول آن تارهای صوتی باز است را فاز حنجره باز گویند. دوره زمانی یک سیکل حنجره را دوره پیچ[۱] گویند و معکوس دوره‌ی پیچ را فرکانس بنیادی[۲] گویند (برحسب هرتز) که در آن تارهای صوتی ارتعاش می‌کند (فرکانس پیچ). پس دوره‌ی پیچ در اولویت اول تحت تاثیر جرم و الاستیسیته‌ی تارها می‌باشد.

اگر جرم تار زیاد باشد(اصطلاحاً تنبل بوده)، طول دوره‌ی پیچ طولانی‌تر می‌شود(فرکانس بنیادی، کوچک‌تر می‌شود). مردان معمولا فرکانس بنیادی کم‌تری نسبت به زنان دارند زیرا تارهای صوتی آن‌ها طولانی‌تر از زنان و سنگین‌تر نیز است. فرکانس بنیادی، محدوده‌ی ۶۰-۱۵۰ هرتز برای گویندگان مرد و ۲۰۰- ۴۰۰ هرتز برای زنان و کودکان دارد[۳۶].

همان‌طورکه قبلا گفته شد هنگامی که تارهای صوتی درحالت مصوت هستند، مصوت‌هایی مانند "aa"و "iy" تولید می‌شوند. به همین دلیل مصوت‌ها را آوای صدادارگویند. هنگامی که تارهای صوتی درحالت صامت قرار دارندصداهایی مانند "y" تولید می‌شوند. این حالت شبیه حالت تنفس در تارهای صوتی است که تارهای صوتی ارتعاش نمی‌کنند. تارها تشدید شده و به یکدیگر نزدیک می‌شوند و بنابراین اجازه می‌دهند که جریان هوا عبور کند. این کار سبب تولید اغتشاش[۳] درحنجره می‌شود. درگفتار عادی

[1] Pitch
[2] Fundemantal frequency
[3]- Turbulence

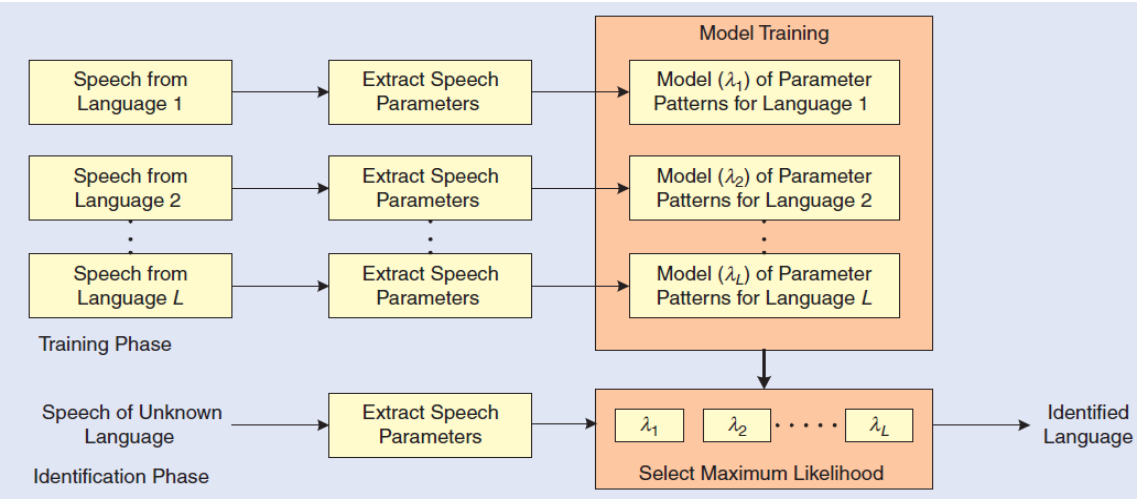
اغتشاش وقتی اتفاق می‌افتد که صامت هایی مانند "h" در "house" یا هنگامی که پچ پچ می‌کنیم تولید شود. هنگامی که تارهای صوتی در حالت صامت هستند نیز اغتشاش تولید می‌کنند مانند "s"،"f"، "h" [۳۶].

حال پس از آشنایی با بعضی از اصطلاحات پردازش صوت همانند پیچ صدا و فرکانس بنیادی و همچنین چگونگی در نظر گرفتن سیگنال صوتی به‌صورت یک سیگنال ایستان، به تشریح سیستم‌های شناسایی زبان می‌پردازیم.

۳-۲-نگاهی اجمالی به سیستم‌های شناسایی زبان گفتاری

صورت کلی عملکرد یک سیستم شناسایی زبان در شکل ۳-۵ آمده است. سیستم شناسایی خودکار زبان به دو بخش تقسیم می‌شود[۳۷]:

- پیش‌پردازش که ویژگی‌های مورد نظر را استخراج می‌کند.
- پس‌پردازش که شامل مدل‌های زبانی می‌شود، $\{\lambda_1 | l = 1,2,...,L\}$ که L در اینجا تعداد کل زبان‌ها است.



شکل ۳-۵: بلوک دیاگرام کلی یک سیستم تشخیص خودکار زبان[۳۷]

هدف اصلی پیش‌پردازش این است که بهترین و مناسب‌ترین اطلاعات مربوط و متناسب از شکل موج استخراج گردد و اطلاعات بیهوده و حشو تا حد امکان حذف گردد. معمولا این فرایند به‌صورت فریم به فریم انجام می‌شود و هر فریم از گفتار به یک بردار ویژگی N بعدی که مقادیری کمتر از تعداد نمونه‌های هر فریم دارد تبدیل می‌شود. می‌توان به وضوح پیش‌بینی کرد که عمدتا این عمل باعث کاهش حجم داده‌ها می‌شود، بنابراین پردازش مورد نیاز سیستم پیش‌پردازش کمتر می‌شود. در این فرایند شکل موج گفتار به مجموعه‌ی متوالی از بردارها تبدیل می‌شود، $X = [x_1, x_2, ..., x_k]$ ،که k در اینجا اندیس مربوط به هر فریم و x_k ، بردار N بعدی است.

تکنیک استخراج ویژگی ایده‌آل نسبت به نویز مقاوم و مستقل از گوینده است و به مولفه‌هایی از شکل موج گفتار که در بین زبان‌های مختلف متمایز است توجه بیشتری دارد. یکی از رایج‌ترین تکنیک‌های استخراج ویژگی مورد استفاده، ضرایب کپسترال فرکانس مل (MFCC) است. نخست گفتار، توسط فرایند پیش‌پردازش به رشته‌ای از بردارهای ویژگی تبدیل می‌شود(X) سپس از پس‌پردازش سیستم عبور می‌کند(شکل۳-۵). لازم به ذکر است که وظیفه‌ی پس‌پردازش، آموزش مدل‌ها و شناسایی زبان‌هاست. در فاز آموزش بردارهای ویژگی استخراج شده از قسمت پیش‌پردازش برای آموزش مدل زبانی λ_l برای هر زبان L ، به وسیله‌ی سیستم تشخیص داده می‌شوند. نحوه‌ی مدل‌ها و روش‌های آموزش مورد استفاده در هر سیستم با سیستم دیگر می‌تواند متفاوت باشد. در مرحله‌ی شناسایی، ویژگی‌های گفتار از گفتارهای ناشناخته استخراج می‌شوند، سپس مجموعه‌ی ویژگی‌ها با مجموعه‌ی مدل‌ها مقایسه می‌شوند، $\{\lambda_l | l = 1,2,,...,L\}$ که در اینجا L تعداد زبان‌هایی است که سیستم قادر به شناسایی آن‌هاست. در این مرحله سیستم باید تعیین کند که کدام زبان L بیشترین شباهت را با بردار ویژگی X دارد. این انتخاب توسط فرمول زیر تعیین می‌شود:

$$\hat{l} = \arg\max P(\lambda_l | X) \qquad (۱-۳)$$

در عمل فرمول ۳-۱ را به صورت ۳-۲ نیز می‌توان بیان کرد:

$$\hat{l} = \arg\max \frac{P(X|\lambda_1)}{P(X)} \qquad (۲-۳)$$

فرض می‌کنیم که هر مدل زبانی کاملا مشابه و P(X) ها یکسان است. می‌توان با صرفنظر از مدل زبانی، برای شناسایی زبان از فرمول (۳-۳) نیز بهره برد:

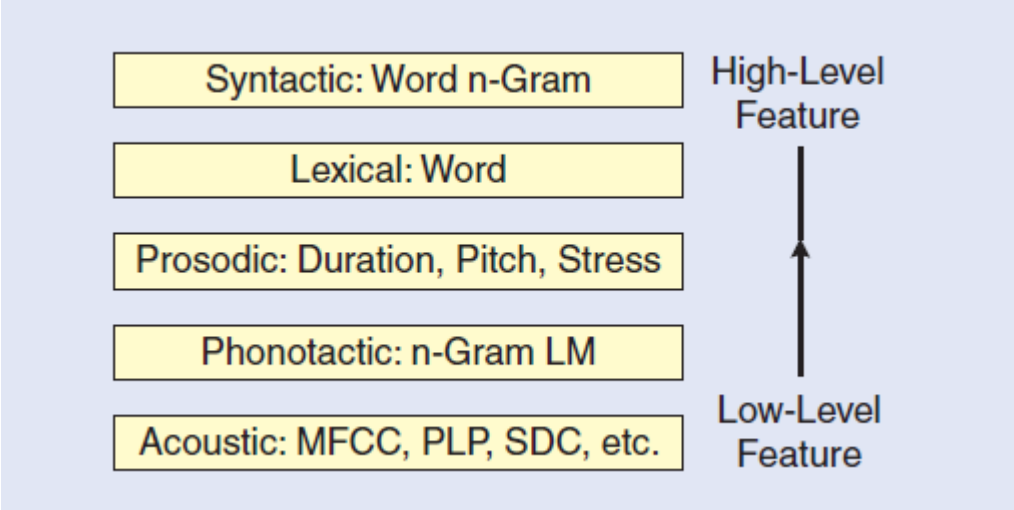
$$\hat{l} = \arg\max P(X|\lambda_l) \qquad (۳-۳)$$

با توجه به رابطه‌ی (۳-۳) می‌توان نتیجه گرفت که برای یافتن زبانی که بیشترین شباهت را به زبان گفتاری داده شده دارد، می‌توان مدل زبانی را یافت که با بردار X بیشترین احتمال شباهت را دارا می‌باشد[۳۷].

۳-۲-۱-اطلاعات گفتاری برای شناسایی زبان

اطلاعات گوناگونی برای تشخیص و متمایز کردن زبان‌ها از یکدیگر وجود دارد که انسان‌ها و ماشین‌ها می‌توانند از آن برای شناسایی زبان استفاده کنند. در سطوح پایین‌تر، ویژگی‌های گفتار مانند ویژگی‌های آکوستیکی، آوایی، واج‌آرایی و اطلاعات عروضی به صورت گسترده در زمینه‌ی تشخیص خودکار زبان مورد استفاده قرار می‌گیرند. اما در سطوح بالاتر، تفاوت بین زبان‌ها را می‌توان بر پایه‌ی

ریخت شناسی[1] و علم نحو[2] نشان داد. شکل ۳-۶ سطوح مختلف ویژگی‌های گفتار را که بـرای شناسـایی زبان سودمند است نشان می‌دهد.



شکل۳-۶: سطوح ویژگی‌های گفتاری[۳۷]

همانطور که در شکل، در سطح پایین ویژگی‌ها آمده است، ویژگی آکوستیکی گفتار نمایشی فشرده از گفتار خام است و می‌توان آن را با ویژگی‌های کپسترال، مانند MFCC مدل کرد[۳۸]. واج آرایی بر الگوسازی صدا در یک نوع زبان خاص دلالت دارد که همان ترکیب مجاز آواها در یک زبان خاص است. در یک سطح بالاتر، از مدل زبانی N-gram (LM) در ویژگی‌های واج‌آرایی استفاده می‌شود. خواص عروضی مانند دیرش، گام و تکیه‌ی گفتار مولفه‌هایی هستند که نوعا به گرامر مربوط نمی‌شوند، به عنوان مثال حالت هیجان گوینده یا نوع بیان جمله که مثلا سئوالی است یا توضیحی. در سطوح بالاتر ریخت‌شناسی لغوی موضوعی است که به ساختار درونی کلمات مرتبط می‌شود و قواعد صرف و نحو تحلیل لغاتی هستند که مجموع آن‌ها یک کلمه، جمله و یا یک عبارت را می‌سازد. هنگامی‌که سطوح ویژگی‌های گفتار را با یکدیگر مقایسه می‌کنیم به این نتیجه می‌رسیم که بدست آوردن ویژگی‌های سطوح

[1] -morphology
[2] -syntax

پایین آسان‌تر صورت می‌گیرد. در مقابل ویژگی‌های سطوح بالاتر مانند ویژگی‌های صرف و نحو، به اطلاعات بیشتری از زبان مربوطه نیاز دارند[۳۹]. در حالی‌که هر چه به سطوح پایین‌تر نزدیک می‌شویم نیاز به اطلاعات زبانی برای شناسایی زبان کمتر می‌شود. شکل ۳-۶ نگاهی اجمالی به تفاوت بین سطوح مختلف ویژگی و همچنین تفاوت بین آن‌ها انداخته است. سطوح مختلف به صورت کامل‌تر در زیر بحث شده است.

۳-۲-۱-۱-اطلاعات آکوستیکی

معمولا اطلاعات آکوستیکی به عنوان اولین سطح تحلیل برای تولید گفتار معرفی می‌شود[۲۷]. صدای انسان موجی است که از فشار به حنجره ایجاد می‌شود و تفاوت بین این صداها به مولفه‌های فرکانسی و دامنه‌ی موج بستگی دارد[۴۰]. اطلاعات آکوستیکی یکی از ساده‌ترین اشکال اطلاعاتی است که در طول استخراج ویژگی می‌توان مستقیما از گفتار خام استخراج کرد. همچنین اطلاعات گفتار سطح بالاتر نظیر واج‌آرایی و اطلاعات کلمات را نیز می‌توان از اطلاعات آکوستیکی استخراج کرد. پیشگویی خطی، ضرایب کپسترال فرکانس مل(MFCC) و پیشگویی خطی ادراکی(PLP) و همچنین ضرایب کپسترال پیشگویی خطی (LPCC)، رایج‌ترین تکنیک‌های مورد استفاده در استخراج ویژگی هستند[۴۱و۵].

سابقا تکنیک‌های استخراج ویژگی‌های آکوستیکی به تنهایی استفاده می‌شد، اما بعدا ویژگی‌های دیگری به هر بردار ویژگی اضافه شدند و یک بردار ویژگی جدید ساختند. دلتا کپستروم شیفت یافته(SDC) از جمله این ویژگی‌هاست[۴۲].

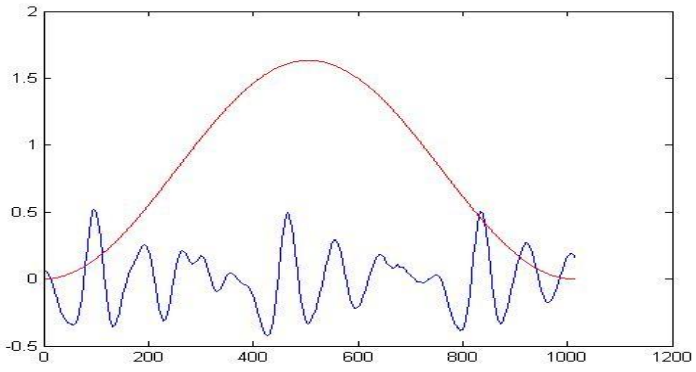
۳-۲-۱-۱-۱-ویژگی MFCC

یکی از معروفترین سیستم ها برای تشخیص زبان MFCC می باشدکه بخش های اصلی فرایند استخراج ویژگی MFCC به صورت زیر می باشد[۲۷].

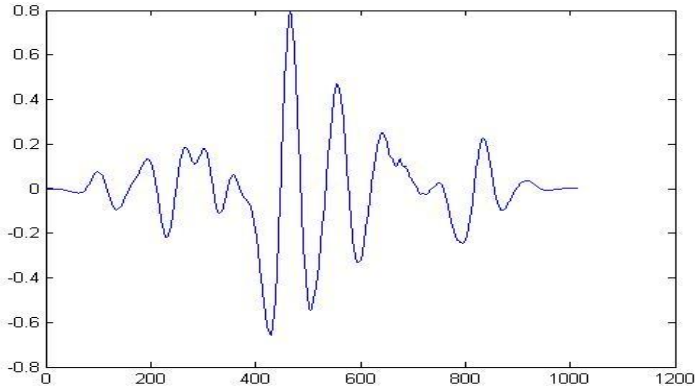
سیگنال صحبت را به فریم های درهم رونده به اندازه ۲۵ میلی ثانیه تقسیم می کنیم . فرض کنید x(n) یک فریم ۲۵ میلی ثانیه از سیگنال صحبت به عنوان سیگنال ورودی باشد. سیگنال را از پنجره W که در زیر تعریف شده است عبور می‌دهیم ، تا اثر ناپیوستگی لبه را کاهش دهیم .

$$w(n) = \beta_w \left(.5 - .5 \cos\left(\frac{2\pi n}{N_w - 1}\right)\right) \qquad 0 \le n \le N_w \qquad (۴-۳)$$

ضریب نرمالیزاسیون می‌باشدو طوری β_w تعداد نمونه‌های معادل با ۳۰ میلی‌ثانیه می‌باشد . N_w تعریف می‌شود که ریشه میانگین مربعات مقادیر نمونه‌های پنجره برابر با یک شود.



شکل ۳- ۷: سیگنال x(n) به همراه پنجره که هنوز با هم ضرب نشده اند[۳۷]



شکل ۳- ۸: سیگنال x(n) بعد از عبور از پنجره W)[۳۷]

طیف سیگنال را به صورت زیر محاسبه می‌کنیم :

$$\tilde{x}(k) = \sum_{n=0}^{N_w-1} x(n)\, W(n)\, e^{-j2\pi nk/N_w} \qquad (۵-۳)$$

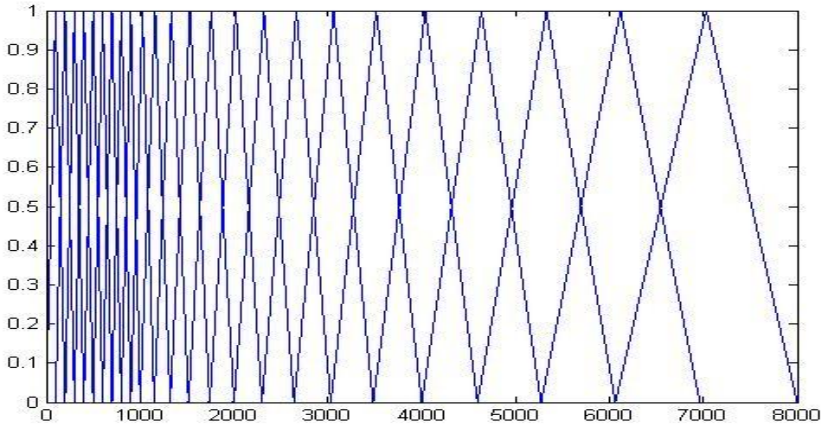
طیف انرژی را به صورت رابطه‌ی ۳-۶ تعریف می‌کنیم :

$$X_k = |\tilde{x}(k)|^2 \qquad (۶-۳)$$

که k برابر با $\frac{N_w}{2}$ می باشد، زیرا فقط نیمی از طیف در نظر گرفته می شود.

از ۱ هرتز تا ۸ کیلو هرتز را به وسیله ۲۴ فیلتر به صورتی که در شکل ۳-۹ مشاهده می کنیم ، تقسیم‌بندی می‌کنیم . برای نمایش، همه فیلترها را با دامنه‌ی ۱ نرمالیزه کرده‌ایم. در همه این فیلترها شرط زیر برقرار می‌باشد.

$$\sum_{k=0}^{k-1} \varphi_j(k) = 1 \qquad (۷-۳)$$



1-200	100-300	200-400	300-500	400-600	500-700	600-800	700-900
800-1000	900-1149	1000-1320	1149-1516	1320-1741	1516-2000	1741-2297	2000-2639
2297-3031	2639-3482	3031-4000	3482-4595	4000-5272	4595-6063	5272-6964	6063-8000

شکل ۳- ۹: فیلتر با بازه های فرکانسی متفاوت در شکل ها دامنه فیلتر ها نرمالیزه شده است[۲۷]

سیگنال را از هر کدام از فیلتر ها عبور می‌دهیم و انرژی سیگنال را در هر زیر باند به دست می‌آوریم.

$$E_j = \sum_{k=0}^{k-1} \varphi_j(k)\, X(k) \qquad (۸-۳)$$

پارامتر MFCC را به صورت زیر محاسبه می‌کنیم:

$$c_m = \beta_c \sum_{j=0}^{J-1} \cos\left(m \frac{\pi}{J}(j+.5)\right) \log(E_j) \qquad (۹-۳)$$

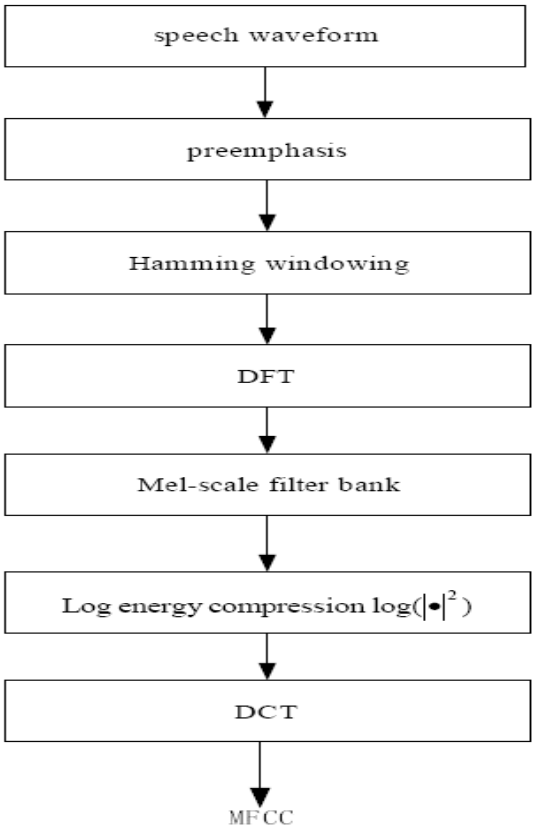
معادله آخر را می توان به عنوان ضرب داخلی لگاریتم بردار انرژی طیف و بردار V_m در نظر گرفت که V_m به صورت زیر می‌باشد.

$$V_m = \left\{ \left(\cos\left(m \frac{\pi}{J} (j + .5) \right) \;\middle|\; 0 \le j \le J \right) \right\} \qquad (۱۰-۳)$$

که به معادله زیر منجر می شود ; که معادل با DCT گرفتن از $\log(E_j)$ می‌باشد . و هدف آن کاهش ابعاد بردار ویژگی می باشد.

$$c_m = \beta_c \sum_{j=0}^{J} V_{m,j} \log(E_j) \qquad (۱۱-۳)$$

مقدار β_c وابسته به مقدار β_w می باشد که معمولا برابر ۲۰۰ می باشد و با تغییر m تغییر نمی‌کند . ما معمولا فقط از ۱۵ مقدار c_m استفاده می کنیم . فرایند استخراج ویژگی MFCC در شکل ۳-۱۰ نمایش داده شده است .

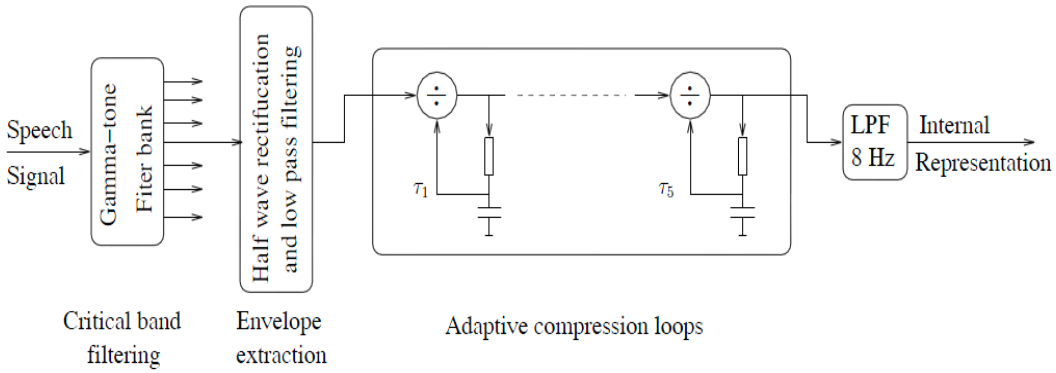


شکل ۳- ۱۰: فرآیند استخراج ویژگی MFCC[۲۷]

۳-۲-۱-۱-۲-ویژگی gMFCC[1]

شخصی به نام شاترجی[۴۳] براساس مدل شنیداری طیفی و طیفی-زمانی، ویژگی gMFCC که بر پایه‌ی ویژگی MFCC است را ارائه داده است. ویژگی MFCC یک ویژگی استاتیک است که به طور مستقل، فریم به فریم مقایسه می‌شود. در ویژگی gMFCC تغییرات طیف در طی زمان نیز در نظر گرفته می‌شود. این ویژگی بر اساس مدل شنیداری DAM[2] ارائه شده است که در شکل ۳-۱۱ نشان داده شده است.

[1] -Generalized MFCC
[2] -Dau Auditory model



شکل ۳-۱۱: مدل شنیداریDAM[۴۳]

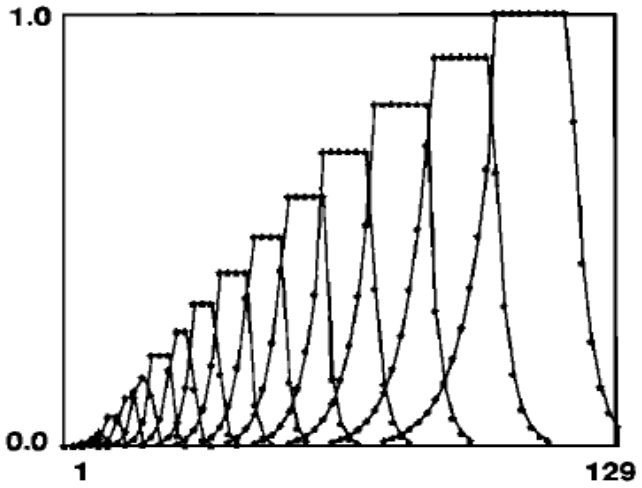
برای مدل کردن غشای پایینی گوش، ابتدا سیگنال گفتار توسط بانک فیلتر گاماتون به باندهای بحرانی تجزیه می‌شود. برای محاسبه‌ی ضرایب gMFCC در مرحله‌ی بعد مدل سلول مویی شنوایی در هر کانال بحرانی بیان می‌شود. در این مرحله سیگنال خروجی هر یک از فیلترهای گاماتون را با استفاده از یکسوساز نیم موج یکسو کرده، سپس با استفاده از فیلتر پایین گذر مرتبه‌ی اول با فرکانس قطع ۱ کیلو هرتز فیلتر می‌شود تا پوش دامنه‌ی سیگنال در خروجی هر فیلتر گاماتون بدست آید. در اینجا هر کانال باند فرکانسی بحرانی شامل اطلاعاتی درباره‌ی تغییرات دامنه‌ی سیگنال در آن کانال است. پس از این مرحله، پوش سیگنال خروجی هر فیلتر با استفاده از مدار تطبیقی شامل پنج حلقه‌ی فشرده سازی تطبیقی غیرخطی فشرده می‌شود. هرکدام از این مدارها از یک تقسیم‌کننده و یک فیلتر پایین‌گذر IIR مرتبه اول تشکیل شده است. ثابت‌های زمانی مربوط به این فیلترها در هر یک از حلقه‌ها به ترتیب برابر $\tau_5 = 562ms, \tau_4 = 253ms, \tau_3 = 129ms, \tau_2 = 50ms, \tau_1 = 5ms$ و فرکانس‌های قطع آن‌ها برابر ۳۲ هرتز، ۳/۲ هرتز، ۱/۲۳ هرتز، ۰/۶۲ هرتز، ۰/۳۲ هرتز است. حلقه‌ی تطبیقی سیگنال ورودی بر سیگنال خروجی از فیلتر پایین‌گذر تقسیم می‌شود. در نتیجه‌ی این فشرده‌سازی، انتقال‌های ناگهانی در پوش خروجی فیلترهای باند بحرانی که نسبت به ثابت زمانی سریع‌تر هستند بدلیل تغییرات آهسته در خروجی

فیلترهای پایین‌گذر در خروجی نهایی به صورت ثابت می‌شوند، در حالی‌که بخش‌های با تغییر کم پوش فشرده می‌شوند. با توجه به خصوصیات این تبدیل، تغییرات در سیگنال ورودی مانند onset ها و offset ها تقویت می‌شوند، در حالی‌که بخش‌های حالت پایدار فشرده می‌شوند. توابع غیر خطی ACL همانند یک حافظه‌ی بلند مدت ذاتی در مدل عمل می‌کنند و کمک می‌کنند تا مدل بتواند ساختار زمانی پویا در پاسخ شنیداری را در نظر بگیرد. مرحله‌ی آخر پردازش در این مدل یک فیلتر پایین‌گذر (LPF) از مرتبه‌ی اول با فرکانس قطع ۸ هرتز است. این فیلتر پایین‌گذر به عنوان یک فیتر مدولاسیون عمل می‌کند و تغییرات ناگهانی و سریع در پوش سیگنال خروجی هر کانال باند فرکانس بحرانی را تضعیف می‌کند[۴۳].

۳-۲-۱-۱-۳- ویژگی PLP

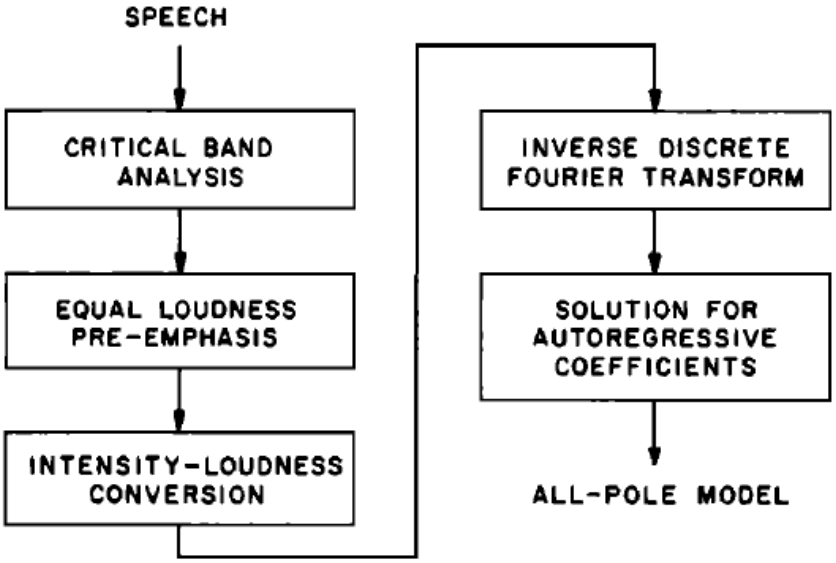
پیشنهادات بسیار دیگری از سیگنال گفتار زمان کوتاه در مورد MFCC شد. یکی از این پیشنهادات ضرایب پیشگویی خطی ادراکی(PLP) بوده است. این پیشنهاد که توسط هرمانسکی داده شد[۴۴]،ترکیب سه مفهوم اساسی از فیزیولوژی صوتی گوش بوده است. دقت طیفی باند بحرانی، منحنی یکسان‌سازی بلندی صدا و قانون توان شدت صدا سه مفهوم بیان شده توسط هرمانسکی بود. در اینجا باند بحرانی مشابه با فیلتر ملی است که در استخراج ویژگی توسط روش MFCC ارائه می‌شود، با این تفاوت که به جای مقیاس مل از مقیاس بارک استفاده می‌شود و منحنی‌های ذوزنقه‌ای شکل به جای فیلترهای مثلثی شکل بکار می‌رود. مدل منحنی یکسان‌سازی بلندی صدا حساسیت غیرخطی از مدل گوش انسان در فرکانس‌های مختلف است و در نهایت قانون شدت توان یک رابطه‌ی غیر خطی بین شدت صدا و بلندی ادراکی صدا پیدا می‌کند. شکل های۳-۱۲ و ۳-۱۳ فیلتر بانک بارک و مراحل انجام این روش استخراج ویژگی را نشان می‌دهند. در مورد PLP از تراکم دامنه‌ی ریشه‌ی سوم صدا استفاده می‌شود.

سپس طیف شنیداری به‌وسیله‌ی مدل خود برگشتی تمام قطب تخمین زده می‌شود. ونگ[1][۶]روش PLP و MFCC و چندین روش استخراج ویژگی دیگر را بر روی سیستم تشخیص زبان مقایسه می‌کند و در می‌یابد که روش PLP، بهترین عملکرد را با یک سیستم تشخیص زبان توسط GMM ی که از ضرایب کاهش میانگین کپسترال(CMS) دلتا و دلتادلتا بهره می‌برد دارد.



۳-۱۲: فیلتر بانک بارک[۴۴]

[1] -wong



شکل ۳-۱۳: بلوک دیاگرام مراحل مختلف روش استخراج ویژگی PLP[۴۴]

۳-۲-۱-۱-۴-کپسترال دلتا و دلتا دلتا(Δ و $\Delta\Delta$)

در حالی‌که بردارهای ویژگی ایستا مانند MFCC ها تخمین خوبی از طیف‌های محلی می‌زنند، اما از نظر دینامیکی مدلی از گفتار انسان را که باعث تمایز بین زبان‌ها می‌شود را به خوبی نمایش می‌دهند. با اضافه کردن مشتقات زمان به پارامترهای ایستا، عملکرد سیستم‌های پردازش گفتار بسیار بهتر می‌شود. مدلی از این نوع، کپسترال دلتا و دلتا دلتا است‌که تخمینی از مشتقات زمانی از کپستروم گفتار است و به عنوان تخمین مربعی کوچکی از شیب محلی می‌باشد. مشتق مرتبه‌ی اول که مربوط به ضرایب دلتا می‌شود را می‌توان با استفاده از فرمول ۳-۱۲ بدست آورد:

$$\Delta C_i(n) = \frac{\sum_{k=-N}^{N} k C_i(n+k)}{\sum_{k=-N}^{N} k^2} \qquad (۱۲-۳)$$

که در اینجا $\Delta C_i(n)$ ضرایب دلتای محاسبه شده در n امین فریم برایiامین کپسترال C_i است و N برای تعیین شماره‌ی فریم‌هایی که کپستروم دلتا محاسبه شده استفاده می‌شود. معمولا N مقادیر مختلفی را می‌تواند به خود اختصاص دهد، اما معمولا مقداری بین ۲ تا ۴ می‌گیرد.

به طریقی مشابه، می‌توان مشتق مرتبه‌ی دوم(ضرایب دلتادلتا)را با معادله‌ای مانند معادله‌ی بالا بر روی ضرایب دلتا به جای ضرایب کپسترال اصلی استفاده کرد. چون معادلاتی که در بالا آمده است مربوط به زمان‌های مختلف ویژگی‌های طیفی می‌باشد نیازمند برخی اصلاحات در شروع و پایان گفتار هستیم که این مشکلات با استفاده از معادله‌ی دیفرانسیل مرتبه‌ی اول در شروع و پایان گفتار به وسیله‌ی T فریم قابل حل است:

$$\Delta C_i(n) = C_i(n+1) - C_i(n), n < N \qquad (۱۳-۳)$$

$$\Delta C_i(n) = C_i(n) - C_i(n-1), n \geq T - N \qquad (۱۴-۳)$$

در بسیاری از مواقع ضرایب دلتا و دلتادلتا برای N=1 (در معادله‌ی ۳-۱۲) محاسبه می‌شوند که با صرفنظر از مخرج به صورت زیر در می‌آید[۱۰]:

$$\Delta C_i(n) = C_i(n+1) - C_i(n-1) \qquad (۱۵-۳)$$

۳-۲-۱-۱-۵-ویژگی SDC استاندارد

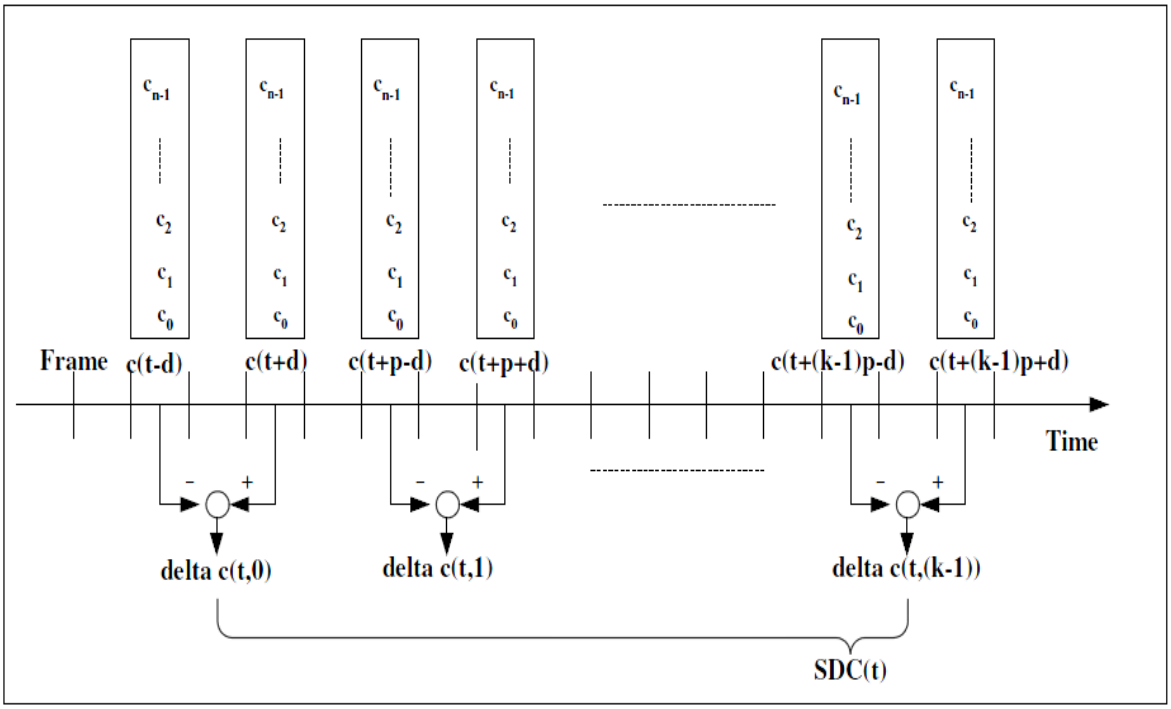
ویژگی SDC استاندارد توسعه یافته‌ی ویژگی MFCC است. ویژگی SDC به صورتی که در شکل ۳-۱۴ نشان داده شده است، محاسبه می‌شود. این ویژگی بر مبنای ۴ پارامتر عمل می‌کند که معمولا به صورت N-d-p-k نوشته می‌شود. برای هر فریم N ضریب اول از ضرایب MFCC انتخاب می‌شود: $C_0, C_1, C_2, ..., C_{N-1}$ و به این ترتیب سیگنال گفتار به توالی از بردارهای MFCC با بعد N تبدیل

می‌شود[۲۵]. هر بردار SDC در زمان t از کنار هم قرار دادن K بردار N بعدی تشکیل می‌شود.هر کدام از این بردارهای N بعدی از اختلاف بردارهایی که با فاصله‌ی d*۲ بر روی نقاط زمانی که به فاصله‌ی P از یکدیگر قرار دارند به صورت ۳-۱۶ محاسبه می‌شود:

$$\Delta c(t,i) = c(t+ip+d) - c(t+ip-d) \qquad (۳-۱۶)$$

در این فرمول، i امین بردار N بعدی است. در نهایت بردار ویژگی SDC مربوط به زمان t از کنار هم قرار دادن این K بردار به صورت زیر حاصل می‌شود:

$$SDC(t) = \left[\Delta c(t,0)^t \Delta c(t,1)^t ... \Delta c(t,k-1)^t\right]^t \qquad (۳-۱۷)$$



شکل ۳-۱۴: ویژگی SDC استاندارد[۲۵]

۳-۲-۱-۱-۶-ویژگی SDC تغییر یافته بر مبنای درون یابی

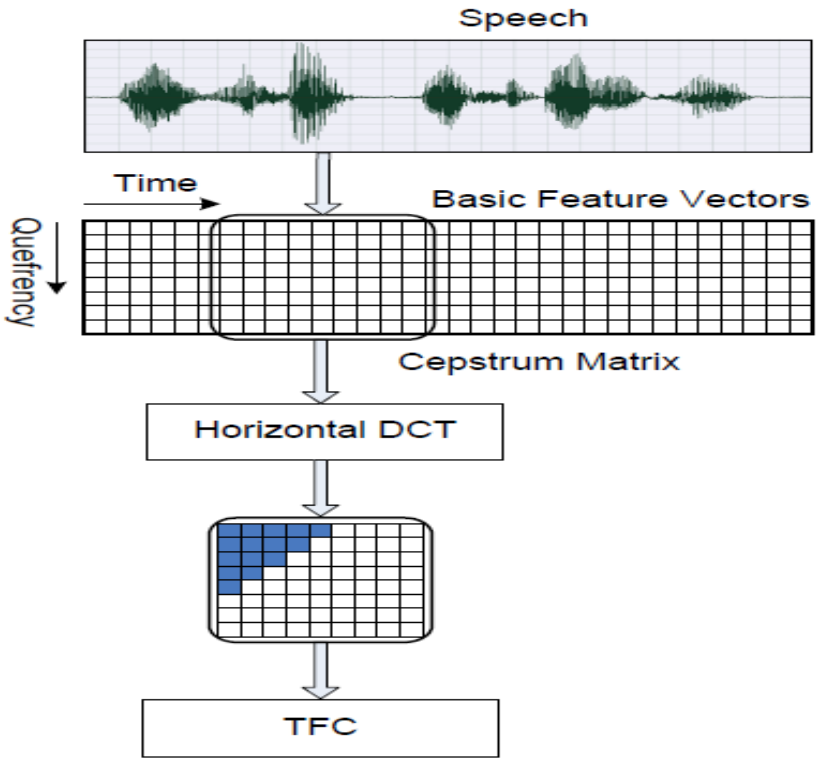
در مرجع[۲۵] در نحوه‌ی محاسبه‌ی SDC استاندارد تغییری ایجاد کرد و آن را برای کاربرد تشخیص زبان بر روی دادگان OGI مورد ارزیابی قرار داد و به نتیجه‌ی بهتری دست یافت. در ویژگی SDC تغییر یافته بر مبنای درون یابی، پارامترها مشابه SDC استاندارد است و تنها نحوه‌ی محاسبه‌ی $\Delta c(t,i)$ متفاوت است که از رابطه‌ی (۳-۱۸) محاسبه می‌شود:

$$\Delta c(t,i) = \frac{\sum_{d=-D}^{D} dc(t+iP+d)}{\sum_{d=-D}^{D} d^2} \qquad (۳-۱۸)$$

۳-۲-۱-۱-۷-ویژگی کپستروم زمان-فرکانس(TFC)

ویژگی کپستروم زمان-فرکانس در مقاله‌ی مرجع[۴۵] آورده شده است. نحوه‌ی محاسبه‌ی این ویژگی در شکل ۳-۱۵ آمده است. در این ویژگی ابتدا بردارهای ویژگی پایه استخراج می‌شود که برای تشخیص زبان، ویژگی کپسترال در نظر گرفته می‌شود، سپس TFC با استفاده از این ویژگی‌ها محاسبه می‌شود. TFC برمبنای ۳ پارامتر N,M,D عمل می‌کند که پارامتر اول تعداد ضرایبی است که از هر بردار پایه مورد استفاده قرار گرفته است، پارامتر دوم تعداد بردارهای پایه‌ای را مشخص می‌کند که TFC بر روی آن‌ها اعمال می‌شود و پارامتر سوم تعداد ابعاد بردار ویژگی نهایی را مشخص می‌کند. در این ویژگی، یک پنجرا به اندازه‌ی M مورد استفاده قرار می‌گیرد که هر بار یک واحد به جلو شیفت داده می‌شود. ابتدا یک تبدیل DCT اولیه به ماتریس N*M دیگر تبدیل می‌شود که در آن مولفه‌های فرکانس پایین در درایه‌های بالاتر ماتریس و مولفه‌های فرکانس بالا که اطلاعات کمتری را شامل می‌شوند در درایه‌های

پایینی ماتریس قرار می‌گیرند. با توجه به پارامتر D، درایه‌های ماتریس حاصل به صورت زیک زاک در کنار یکدیگر قرار گرفته و بردار ویژگی نهایی را ایجاد می‌کنند.



شکل ۳-۱۵: روند استخراج ویژگی TFC [۴۴]

۳-۲-۱-۱-۸-تبدیل فوریه-بسل

در این بخش علاقه‌مند به بسط یک سیگنال گفتار به سری‌های فوریه-بسل هستیم. یکی از خواص سری‌های فوریه این است‌که محاسبات در پردازش سیگنال را ساده‌تر می‌کند، از سویی دیگر ضرایب سری فوریه با استفاده از تبدیل فوریه‌ی گسسته محاسبه می‌شوند که ضرایب آن توسط تبدیل فوریه‌ی سریع(FFT) بدست می‌آید.

توابع بسل با حل معادله دیفرانسیل ۳-۱۹ به دست می آیند:

$$t^2y'' + ty' + (t^2 - n^2)y = 0 \quad n > 0 \qquad (۱۹-۳)$$

که معادله دیفرانسیل بسل نامیده می شود. حل کلی این معادله جواب زیر را می‌دهد:

$$y = C_1J_n(t) + C_2Y_n(t) \qquad (۲۰-۳)$$

که$J_n(x)$ را تابع بسل نوع اول مرتبه n و $Y_n(x)$ را تابع بسل نوع دوم مرتبه n می نامند. توابع بسل قابل ارائه به فرم سری هستند. برای مثال $J_n(x)$ را می توان این گونه نوشت:

$$J_n(t) = \sum_{r=0}^{\infty} \frac{(-1)^r (t/2)^{n+2r}}{r!\,\Gamma(n+r+1)} \qquad (۲۱-۳)$$

و در حالت خاص:

$$J_0(t) = 1 - \frac{t^2}{2^2} + \frac{t^4}{2^2 4^2} - \frac{t^6}{2^2 4^2 6^2} + \cdots \qquad (۲۲-۳)$$

می‌توان نشان داد توابع بسل با در نظر گرفتن تابع وزن‌دهی[1] t متعامد هستند. به این ترتیب می توان دید:

$$\int_0^a tJ_n(\alpha t)J_n(\beta t)dt = \frac{a[\beta J_n(a\alpha)J'_n(a\beta) - \alpha J_n(a\beta)J'_n(a\alpha)]}{\alpha^2 - \beta^2} \qquad \alpha \neq \beta \qquad (۲۳-۳)$$

و اگر $\beta \to \alpha$ به وسیله قاعده هوپیتال خواهیم داشت:

$$\int_0^a xJ_n^2(\alpha t)dt = {a^2}/{2}\left[J'^2_n(a\alpha) + (1 - \frac{n^2}{\alpha^2})J_n^2(a\alpha)\right] \qquad (۲۴-۳)$$

حال اگر α و β ریشه‌های مختلف معادله‌ی $J_n(at) = 0$ باشند می توان نوشت:

$$\int_0^a tJ_n(\alpha t)J_n(\beta t)dt = 0 \quad \alpha \neq \beta \qquad (۲۵-۳)$$

و در نتیجه $J_n(\alpha t)$ و $J_n(\beta t)$ با در نظر گرفتن تابع وزن دهی t متعامد هستند. همچنین با توجه به اینکه $J_n(a\alpha) = 0$ می‌توان رابطه‌ی ۳-۲۴ را به صورت زیر نوشت:

$$\int_0^a tJ_n^2(\alpha t)dt = {a^2}/{2}\left[J'^2_n(a\alpha)\right] \qquad (۲۶-۳)$$

[1] Weighting function

با در نظر گرفتن معادله ۳-۲۷:

$$J'_n(t) = \frac{1}{2}(J_{n-1}(t) - J_{n+1}(t)) \qquad (۲۷-۳)$$

در نهایت می‌توان رابطه‌ی ۳-۲۴ را به صورت رابطه‌ی ۳-۲۸ نوشت:

$$\int_0^a tJ_n^2(\alpha t)dt = \frac{a^2}{8}(J_{n-1}(a\alpha) - J_{n+1}(a\alpha))^2 \qquad (۲۸-۳)$$

حال با دانستن این مطلب می توان هر تابع دلخواه روی بازه‌ی $(0, a)$ را به صورت جملاتی از توابع بسل نوشت.

$$f(t) = \sum_{m=1}^{\infty} C_m J_n(\lambda_m t) \qquad (۲۹-۳)$$

که در رابطه‌ی ۳-۲۹، λ_1 ، λ_2 و... ریشه‌های مثبت معادله‌ی $J_n(at) = 0$ هستند. برای به دست آوردن ضرایب C_m طرفین رابطه ۳-۲۹ را در $tJ_n(\lambda_m t)$ ضرب می کنیم. با توجه به خاصیت تعامد، همه‌ی جملات سمت راست رابطه، به جز یک m خاص صفر می شوند. در نتیجه داریم:

$$tf(t)J_n(\lambda_m t) = C_m J_n^2(\lambda_m t) \qquad (۳۰-۳)$$

با فرض اینکه $f(t)$ روی بازه‌ی $(0, a)$ تعریف شده است، از رابطه‌ی ۳-۲۲ در این بازه انتگرال می‌گیریم.

$$\int_0^a tf(t)J_n(\lambda_m t) = C_m \int_0^a J_n^2(\lambda_m t) \qquad (۳۱-۳)$$

با استفاده از رابطه‌ی ۳-۲۸ داریم:

$$\int_0^a tf(t)J_n(\lambda_m t) = \frac{a^2}{8} C_m [J_{n-1}(a\lambda) - J_{n+1}(a\lambda)]^2 \qquad (۳۲-۳)$$

با ساده کردن این رابطه، ضرایب C_m به صورت زیر به دست می‌آیند.

$$C_m = \frac{8}{a^2 C_m [J_{n-1}(a\lambda_m) - J_{n+1}(a\lambda_m)]^2} \int_0^a tJ_n(\lambda_m t) f(t)dt \qquad (۳۳-۳)$$

در حالت خاص اگر بخواهیم $f(t)$ را روی هر بازه‌ی دلخواه $(0, a)$ با استفاده از تابع بسل مرتبه صفر بسط دهیم:

$$f(t) = \sum_{m=1}^{\infty} C_m J_0(\lambda_m t) \quad , \quad 0 < t < a \qquad (۳۴-۳)$$

با در نظر گرفتن معادله ۳-۳۵:

$$J_{-n}(\mathrm{t}) = (-1)^n J_n(t) \qquad (۳۵-۳)$$

ضرایب C_m به صورت زیر محاسبه می شوند:

$$C_m = \frac{2\int_0^a tf(t)J_0(\lambda_m t)dt}{a^2[J_1(\lambda_m a)]^2} \qquad (۳۶-۳)$$

و λ_m به ازای $m=1,2,\ldots$ ریشه‌های مثبت معادله‌ی $J_0(a\lambda_m) = 0$ می باشد که به ترتیب صعودی مرتب شده‌اند.

۳-۲-۲- اطلاعات واج‌آرایی

مجموعه‌ی محدودی از صداهای معناداری که به وسیله‌ی انسان‌ها تولید می‌شود وجود دارد. همه‌ی این صداها در تمام زبان‌ها وجود ندارد، بلکه زیرمجموعه‌ی محدود معناداری از این صداها برای هر زبان بکار می‌رود. صداشناسی بررسی سیستم صدایی یک زبان خاص یا مجموعه‌ای از زبان‌هاست و واج‌آرایی شاخه‌ای از صداشناسی است که با الگوهای صدای صحیح و معتبر در زبان مورد نظر سروکاردارد. به عنوان مثال ترکیب مجاز آواها(که جزئی از واحدهای صوت یک زبان هستند که آنها را معنادار می‌کنند)که شامل دو گروه حروف بی‌صدا و صدادار هستند. این ترکیب مجاز قوانین واج‌آرایی را تشکیل می‌دهند[۴۶]. در حوزه‌ی واج‌آرایی اختلافات متعددی در زبان‌های مختلف وجود دارد. برای مثال در گروه ترکیب صدادار، /st/ به طور رایج در زبان انگلیسی بکار می‌رود، در حالی‌که این ترکیب در زبان ژاپنی مجاز نیست و یا استفاده از ترکیب صدادار/sr/ در زبان انگلیسی غلط است، اما در زبان تامیل این ترکیب رایج است. همینطور در زبان ژاپنی دو حرف بی‌صدا نمی‌توانند پشت هم بیایند، اما در زبان دانمارکی و سوئدی این کار انجام می‌شود. بنابراین اطلاعات واج‌آرایی سودمندی بیشتری نسبت به اطلاعات مصوت‌ها یا همان حروف صدادار دارند. بنابراین اطلاعات واج‌آرایی برای استخراج ویژگی مطلوب هر زبان مناسب‌تر است.

۳-۲-۳-اطلاعات عروضی

عروض یکی از مولفه‌های کلیدی در درک شنیداری انسان است. نوا[1]، تکیه و طول مدت هجا و ریتم اجزای اصلی عروض را تشکیل می‌دهند. معمولا گام(فرکانس پایه)برای نمایش نوا استفاده می‌شود. شدت[2] برای نشان دادن تکیه استفاده می‌شود و همینطور طول مدت هجا برای ریتم بکار می‌رود. در زبان‌های گوناگون بعضی از واکه‌ها یا حروف صدادار مشترکند و در مبحث آواشناسی مشخصات طول مدت هجایشان یکسان است. آهنگ[3] را نیز می‌توان به صورت تغییر گام در هنگام صحبت کردن تعریف کرد. در تمام زبان‌ها برای تعجبی کردن یا سئوالی کردن و یا طنز کردن جملات از گام نیز استفاده می‌کنند. به عنوان مثال در زبان انگلیسی می‌توان یک جمله‌ی خبری را با بلند کردن نوا در آخر جمله به جمله‌ی سئوالی تبدیل کرد. معمولا تغییرات گام در نوا در شناسایی زبان‌های آسیایی مانند چینی ماندارین یا ویتنامی استفاده می‌شود که در آنجا آهنگ کلمه معانی را تعیین می‌کند[۴۷]. در بعضی از زبان‌ها نقش و چگونگی بکارگیری تکیه معنی کلمات را تغییر می‌دهد. برای مثال یک اسم در زبان انگلیسی با تغییر مکان تکیه می‌تواند نقش یک فعل را بازی کند. با دانستن این نکته که تکیه در بعضی زبان‌ها مانند فرانسوی معمولا در آخر کلماتشان بکار می‌رود یا در زبان مجارستانی که بیشتر در اول کلماتشان به‌کار می‌رود، می‌توان از تکیه در تشخیص دادن بین زبان‌ها سود برد[۴۵].

در سال ۱۹۹۸ محققان طی آزمایشاتی متوجه شدند که نوزادان زیر ۶ ماه به تغییرات زبانی واکنش نشان می‌دهند. چندین فرضیه می‌تواند برای توجیه این که چگونه نوزادان قادر به تشخیص تغییر زبان هستند به‌کار رود. یکی از محتمل‌ترین این فرضیه‌ها این است که نوزادان به ویژگی‌های عروضی حساس هستند. بر اساس این فرضیه نوزادان اطلاعات عروضی، خصوصا اطلاعات ریتمی جملات را دریافت

[1] -Tone
[2] -Intensity
[3] -Intonation

می‌کنند و بر اساس آن تفاوت بین آن‌ها را تشخیص می‌دهند[۴۸]. این فرضیه محققان را تشویق کرد که از اطلاعات عروضی در سیستم‌های تشخیص زبان استفاده نمایند.

سیستم‌های LID بر مبنای اطلاعات عروضی تاکنون از سیستم‌هایی که بر مبنای اطلاعات واج‌آرایی کار می‌کنند، ضعیف‌تر عمل کرده است، دلیل آن نیز عدم وجود روشی موثر برای مدل کردن خصوصیات عروضی است[۴۹].

۳-۲-۳-۱ وزن طبیعی گفتار(ریتم)

زبان‌ها بر اساس خاصیتی به نام تقارن[1] که نشان دهنده‌ی این است که چقدر می‌توان یک گفتار را به قطعات مساوی یا معادل از نظر دیرش تقسیم کرد به ۳ دسته تقسیم می‌شوند[۵۰]:

۱. زبان‌های تکیه-زبانی[2](انگلیسی و آلمانی)
۲. زبان‌های هجایی زبانی[3](فرانسوی و اسپانیایی)
۳. زبان‌های مورا-زبانی[4](ژاپنی)

در زبان‌های هجایی-زبانی، خاصیت تقارن میان دیرش هجاها وجود دارد، در حالی که در زبان‌های تکیه-زمانی این تقارن میان واحددهای تکیه‌ای قرار دارد. در زبان‌های مورا-زمانی به جای هجا، واحد مورا را داریم که تقارن بین این واحدها وجود دارد[۱۵]. مورا به یکی از ۳ شکل زیر ظهور پیدا می‌کند[۵۱]:

[1] -isochrony
[2] -stress-timed
[3] -syllable-timed
[4] -mora-time

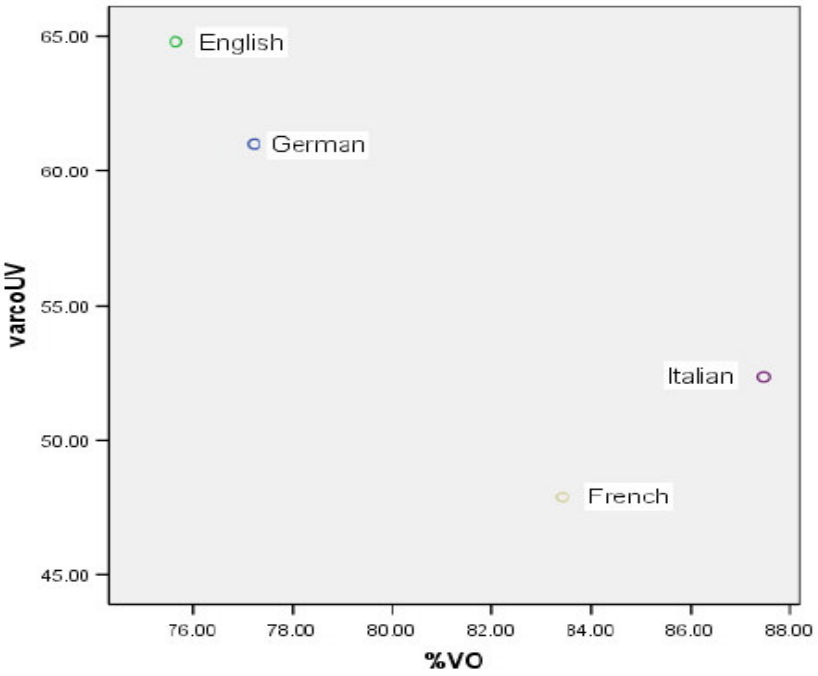
- یک واکه تنها یا به همراه یک همخوان (C)V
- بخش اول یک همخوان کشیده
- پایان کلمه یا پس از یک واکه

می‌توان دو دسته‌ی اول از زبان‌ها را بر اساس درصدی از گفتار که صدادار است (VO%) و انحراف معیار قسمت‌های بی‌صدا(ΔUV) از یکدیگر جدا کرد. وجود پدیده‌ای به نام کاهش صوتی[1] در زبان‌های تکیه-زمانی سبب می‌شود که درصد قسمت‌های صدادار در گفتار نسبت به زبان‌های هجایی-زمانی کم شود. همچنین ساختارهای همخوان در زبان‌های تکیه زمانی پیچیده تر است که این سبب می‌شود تغییرات بیشتری در دیرش این بازه‌ها وجود داشته باشد. در شکل ۳-۱۶ بر اساس این دو معیار زبان‌ها به این دو دسته تقسیم شده‌اند. نوزادان قبل از اینکه دانشی در مورد ساختار آوایی زبان‌ها پیدا کنند، قادرند با استفاده از این دو معیار تفاوت بین زبان‌ها را تشخیص دهند[۶۶].

همان‌طور که قبلا اشاره شد، می‌توان تفاوت‌های ریتمی بین زبان‌ها را با استفاده از ساختار هجایی و وجود و عدم وجود کاهش واکه[2] نشان داد. به عبارت دیگر ریتم، وابسته به دیرش برخی از واحدهای گفتاری است که این واحدهای گفتاری در زبان‌های هجایی-زمانی عموما هجا در نظر گرفته می‌شوند. به عبارت دیگر می‌توان ریتم یک گفتار را با استفاده از دیرش واحدهای هجایی متوالی مدل نمود. هجاها به منظور داشتن یک سیستم مستقل از زبان مناسب نیستند، چرا که می‌بایست به ازای هر زبان جدید که قصد داریم به سیستم اضافه کنیم، مجموعه هجاها را مشخص نماییم(مجموعه هجاها از یک زبان به زبان دیگر متفاوت است) . از همه مهم‌تر این که تشخیص محدوده هجاها کار دشواری است. به همین دلیل به جای هجا از شبه هجا استفاده می‌شود که برای اولین بار در مرجع[۱۲] معرفی شد. بر اساس این مرجع

[1] -vocallic reduction
[2] -vowel reduction

ساختار هجایی که در اکثر زبان‌های جهان وجود دارد، ساختار CV است که در آن C یک همخوان و V یک واکه است. شبه هجا به صورت الگوی C^nV تعریف می‌شود که در آن n عدد صحیحی است که می‌تواند مقدار صفر را نیز اختیار کند. بنابراین مسئله تبدیل به یافتن مکانیزمی برای استخراج الگوهای همخوان-واکه می‌شود. در این روش نیازی به داشتن دانش در مورد ساختار هجایی زبان‌ها نداریم و مرز قسمت‌ها در مکان هایی است که واکه‌ها قرار دارند[۵۰].



شکل ۳-۱۶: دسته بندی زبان‌های stress-time و syllable-time بر اساس معیارهای%VO و (ΔUV) [۵۲]

۳-۲-۳-۲- استخراج دیرش شبه هجاها

برای مدل کردن ریتم، بایستی ابتدا شبه هجاها را بدست آورد. ایده‌ی اولیه به این صورت است که نقاط زمانی پایانی هر واکه را بدست آورده و زمان بین نقاط پایانی دو واکه متوالی را اندازه‌گیری نماییم. برای تشخیص واکه‌ها می‌توان از روش مبتنی بر سیگنال یا روش مبتنی بر HMM استفاده کرد.

در روش اول طیف انرژی مربوط به سیگنال گفتار استخراج شده و از پیک‌های انرژی برای تشخیص مکان واکه‌ها استفاده می‌شود. روش دوم با استفاده از HMM است که نه تنها توالی آواها را تشخیص می‌دهد بلکه توالی زمانی را نیز به عنوان خروجی بر می‌گرداند. روش دوم دقت بالاتری در تشخیص مکان واکه‌ها دارد[۵۰].

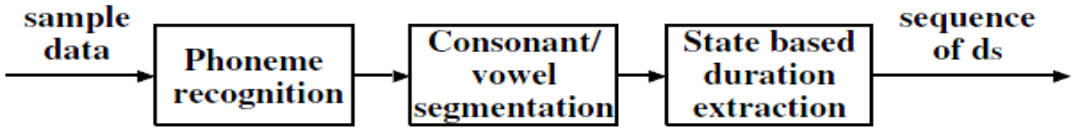
روند استخراج ویژگی با استفاده از یک HMM چند زبانه در شکل ۳-۱۷ نشان داده شده است[۵۰].

- سیگنال ورودی با استفاده از یک بازشناس آوا که با استفاده از یک HMM چند زبانه پیاده‌سازی شده است به توالی از آواها و دیرش متناظر با هر آوا تبدیل می‌شود.
- این توالی به الگوهای C^nV تقسیم شده و در نتیجه توالی از شبه هجاها بدست می‌آید. برای توالی شبه هجاهای بدست آمده، دیرش متناظر با آن، d_s محاسبه می‌شود.

به این ترتیب ریتم با استفاده از توالی دیرش شبه هجاهای استخراج شده مدل می‌شود و مقادیر دیرش حاصل می‌تواند به عنوان ویژگی در تشخیص زبان مورد استفاده قرار گیرد[۵۰]. در مرجع [۱۲]، برای مدل کردن ریتم، ابتدا گفتار ورودی بخش بندی شده شده که برای این کار از روش واگرایی رو به جلو و عقب[1] استفاده شده است. در نتیجه گفتار ورودی به یک سری بخش‌های گذر صداها[2] و بخش‌های بزرگ‌تر(قسمت‌های پایدار صوت[3]) تقسیم می‌شود. سپس از یک الگوریتم ردیابی فعالیت صوتی به منظور تشخیص بخش‌های سکوت استفاده شده است. در مرحله‌ی بعد با بررسی طیف سیگنال، با استفاده از یک الگوریتم ردیابی واکه‌ها، قسمت‌های واکدار سیگنال تشخیص داده می‌شود. در شکل ۳-۱۸ خروجی این مراحل بر روی یک سیگنال گفتار ورودی نشان داده شده است. در نتیجه‌ی این مراحل سیگنال ورودی به توالی از قسمت‌های سکوت، واکدار و همخوان تبدیل می‌شود. سپس دنباله‌ی بدست آمده پویش شده تا

[1] -Forward-backward divergence
[2] -transient parts of voiced sounds
[3] -steady parts

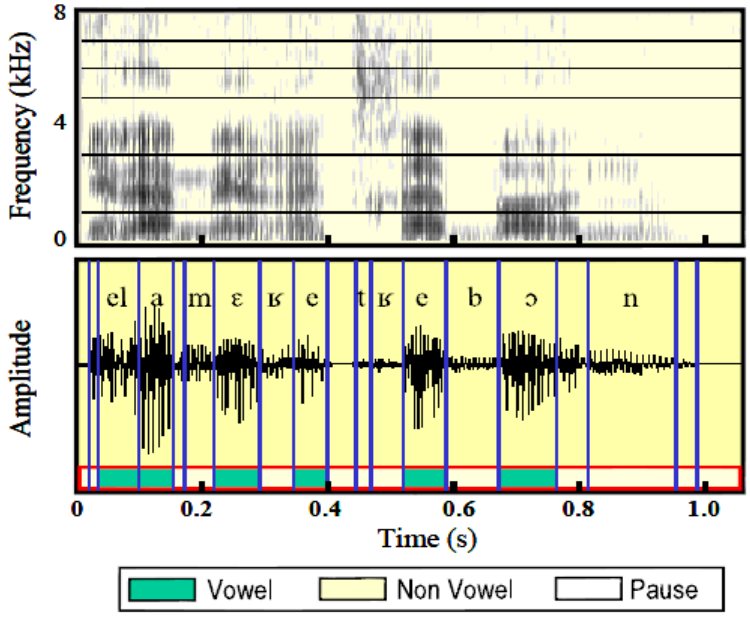
الگوهای C^nV از آن استخراج شود. برای مثال دنباله‌ی سیگنال صوتی شکل ۳-۱۸ به‌صورت (CCVV.CCV.CV.CCCV.CV.CCC) است. قسمت‌هایی که هیچ واکه‌ای در آن‌ها نیست، حذف شده و در صورتی که دو واکه پشت سر هم آمده باشند، با یکدیگر ادغام می‌شوند. دنباله‌ی شبه هجاها به‌صورت (CCV.CCV.CV.CCCV.CV) در می‌آید. به این ترتیب می‌توان ریتم را در یک گفتار ورودی مدل نمود. مزیت استفاده از شبه هجاها این است که نیازی به داده‌ی برچسب خورده نیست و همچنین در مورد ساختار هجایی زبان‌ها نیاز به دانش نداریم.



شکل ۳-۱۷: استخراج دیرش شبه هجاها[۵۰]

پس از استخراج دنباله‌ی شبه هجاها می‌توان یک شبه هجا را با استفاده از یک آرایه سه بعدی به صورت زیر نشان داد:

که در آن d_{c1} و d_{c2} دیرش همخوان‌های C_1 و C_2 (همخوان‌های اول و دوم از سمت واکه) و d_v دیرش واکه V و N_c تعداد همخوان‌های موجود در شبه هجا است. به این ترتیب می‌توان شبه هجاها را به یک سری بردارهای ۳ بعدی تبدیل نمود.



شکل ۳-۱۸: مثالی از بخش بندی خودکار سیگنال گفتار به قسمت‌های سکوت، واکدار و بی‌واک[۱۲]

۳-۲-۳-۳-فرکانس پایه

تغییرات گام یا به عبارت دیگر زیر و بمی، آهنگ را به وجود می‌آورد که از یک زبان به زبان دیگر متفاوت است. در مرجع [۴۹] از اطلاعات کانتور گام برای LID استفاده شده است. مدل کردن گام در این مرجع به این صورت است که پس از استخراج کانتورهای گام، کانتورهایی که از نظر بازه‌ی زمانی طولانی هستند، به قسمت‌های کوچکتر تقسیم می‌شوند، سپس هر قسمت کانتور گام توسط چند جمله‌ای لژاندر تخمین زده می‌شود. این ضرایب به همراه طول کانتور گام به عنوان بردار ویژگی مورد استفاده قرار می‌گیرد.

۳-۲-۳-۴- استخراج کانتور گام

برای استخراج کانتور گام در این مرجع از روش ارایه شده توسط برسما استفاده شده است. این روش از تابع اتوکرولیشن برای تشخیص قسمت‌های صدادار و نقاط کاندید کانتور گام استفاده می‌کند، سپس از الگوریتم ویتربی برای یافتن مناسب‌ترین کانتور گام استفاده می‌نماید. برخی از پارامترهای معمول برای استخراج کانتور گام در جدول ۳-۱ نشان داده شده است[۴۹].

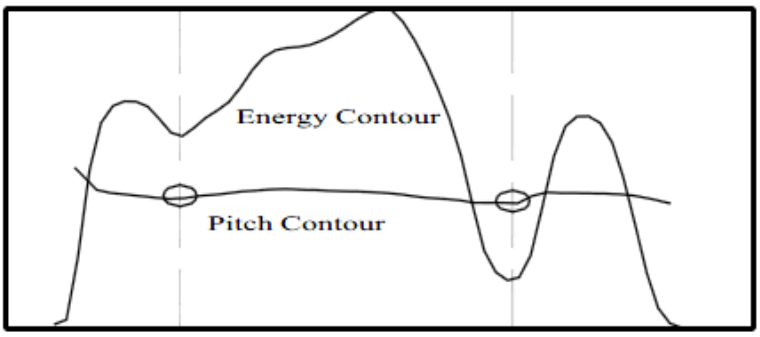
جدول ۳-۱: پارامترهای مورد استفاده برای استخراج کانتور گام[۴۹]

Analysis window length(ms)	**30**
Analysis window time step(ms)	**10**
Pitch floor (Hz)	**50**
Pitch ceiling (Hz)	**500**
Max. number of candidates	**5**
Silence threshold	**0.03**
Voicing threshold	**0.6**
Octave cost	**0.01**
Octave-jump cost	**0.6**
Voiced/unvoiced cost	**0.14**

۳-۲-۳-۴-قسمت بندی کانتور گام

برخی از بخش‌های استخراج شده کانتور گام بسیار طولانی هستند. به منظور تقسیم کردن این قسمت‌های طولانی به قسمت‌های کوچکتر می‌توان از کانتور انرژی استفاده کرد. همانطور که در شکل

۱۹-۳ مشاهده می‌شود، ابتدا کانتور گام توسط کانتور انرژی علامت‌گذاری و یک سری نقاط کاندید مشخص می‌شود. نقاط کاندید نقاطی هستند که در گودی‌ها(دره‌ها) کانتور انرژی قرار دارند. به منظور اجتناب از بدست آوردن قسمت‌های که طول خیلی کوتاه دارند، یک محدودیت طول نیز گذاشته شده است که در این مرجع ۵۰ میلی ثانیه در نظر گرفته شده است. همانطور که در شکل ۳-۱۹ مشاهده می‌شود، ۲ نقطه به عنوان نقاط کاندید در نظر گرفته شده است که تنها نقطه‌ی سمت راست به عنوان نقطه ای برای تقسیم بندی انتخاب شده است، چرا که نقطه‌ی کاندید اول سبب می‌شود که یک قسمت با طول خیلی کوتاه(کمتر از ۵۰ میلی ثانیه) بدست آید[۴۹].



شکل ۳-۱۹: قسمت بندی کانتور گام[۴۹]

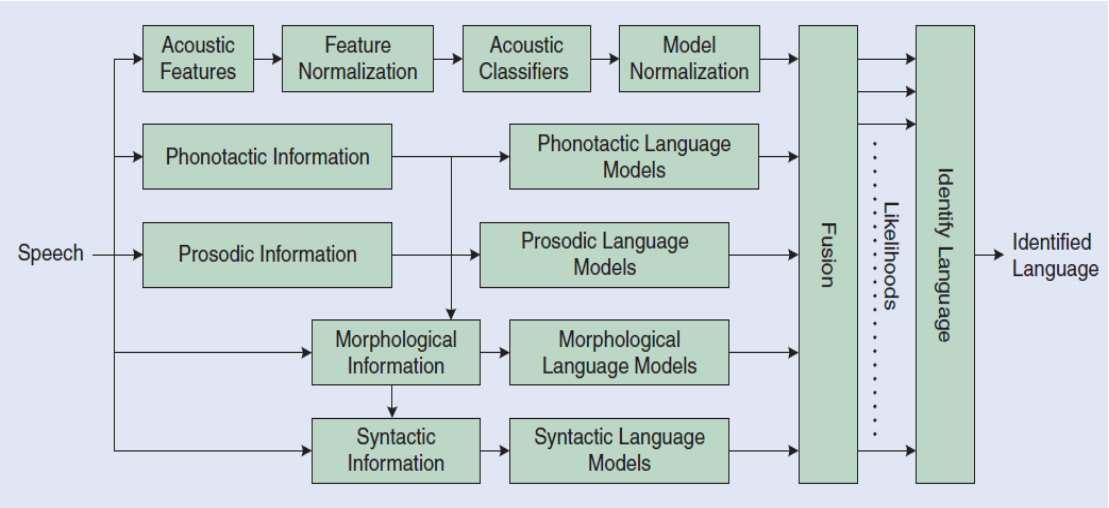
۳-۲-۴-ریخت شناسی

ریخت شناسی رشته‌ای از زبان‌شناسی است که بر روی ساختار داخلی کلمه مطالعه می‌کند[۵۳]. معمولا کوچکترین واحد در علم نحو کلمات هستند و در اکثر زبان‌ها می‌توان با استفاده از قانون ریخت‌شناسی آن‌ها را با یکدیگر مرتبط ساخت. ریشه‌ی کلمات و معنی لغات معمولا از زبانی به زبانی دیگر متفاوت هستند. نتیجتا در بحث شناسایی خودکار زبان، کلمات با بررسی مشخصات وساختار شکلیشان شناسایی می‌شوند.

۳-۲-۵-علم نحو

مطالعه بر روی قواعد و اصول چگونگی قرار گرفتن کلمات در جمله را علم نحو گویند[۵۴]. الگوی جملات در هر زبانی متفاوت است. ولی در مواردی تک کلمات در دو زبان یکسان هستند. برای مثال ممکن است در زبان انگلیسی و آلمانی، مجموعه‌ی کلماتی که در پس یا پیش از کلمه‌ی bin استفاده می‌شود متفاوت باشد[۱۰]. واضح و مبرهن است در صورتی‌که یکپارچگی و مجموع کلمات، مبتنی بر لغات و گرامر و استفاده‌ی آن‌ها از طریق ریخت شناسی و اطلاعات نحوی باشد سیستم تشخیص زبان عملکرد بهتری دارد و تلاش برای استفاده از چنین اطلاعاتی پیشرفت‌هایی را نیز به همراه دارد. به هر حال هنگامی‌که دو سطح آوا با یکدیگر مقایسه می‌شوند، نیاز به گردآوری فرهنگ لغت و کلمات مبتنی بر گرامر برای تشخیص زبان محسوس تر می‌شود. در حال حاضر سیستم‌هایی که از اطلاعات ریخت‌شناسی و علم نحو استفاده می‌کنند خیلی رایج نیستند. معمولا سیستم‌های تشخیص زبان از چند زیرسیستم یا سیستم‌های کوچکتر تشکیل می‌شوند که از بعضی یا تمام انواع اطلاعات ذکر شده در بالا برای تخمین یا بدست آوردن بعضی از مقادیر شبیه‌سازی در زبان‌های مختلف استفاده می‌کنند.

شکل ۳-۲۰ بلوک دیاگرام کلی سیستم تشخیص زبانی را نشان می‌دهد که از تمام سطوح اطلاعات استفاده می‌کند.



شکل۳-۲۰: بلوک دیاگرام کلی یک سیستم تشخیص اتوماتیک زبان[۳۷]

هر چند لازم نیست که یک سیستم تشخیص زبان تمام این روش‌ها را انجام دهد و در حقیقت اکثر سیستم‌های تشخیص زبان تمام این کارها را انجام نمی‌دهند. در واقع بیشتر سیستم‌ها از اطلاعات صوتی و واج‌آرایی استفاده می‌کنند.

فصل چهارم

شناساگرهای زبان

۴-شناساگرهای زبان

سیستم‌های تشخیص زبان مبتنی بر اطلاعات آوایی تاکنون جزو بهترین سیستم‌های تشخیص زبان بوده اند که در این فصل اجزای این سیستم‌ها شرح داده خواهد شد.

۴-۱- مدل‌های زبانی

یک مدل زبانی آماری با استفاده از یک توزیع احتمالاتی، به توالی از m واحد زبانی(در بحث تشخیص زبان این واحدها می‌توانند واحدهای آوایی باشند) یک احتمال $P(w_1, w_2, ..., w_m)$ را نسبت می‌هد. مدل‌های زبانی در کاربردهای ترجمه‌ی ماشینی، بازیابی اطلاعات، تشخیص گفتار، بازشناسی زبان و ... کاربرد دارند.در یک مدل زبانی n تایی[1]، احتمال مشاهده‌ی توالی $w_1, w_2, ..., w_m$ به صورت زیر تخمین زده می‌شود.

$$P(w_1, ..., w_m) = \prod_{i=1}^{m} P(w_i | w_1, ..., w_{i-1}) = \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, ..., w_{i-1}) \qquad (۱-۴)$$

که در آن مقدار احتمال شرطی $P(w_i | w_{i-(n-1)}, ..., w_{i-1})$ به صورت رابطه‌ی۴-۲ محاسبه می‌شود:

$$P(w_i | w_{i-(n-1)}, ..., w_{i-1}) = \frac{Count(w_{i-(n-1)}, ..., w_{i-1}, w_i)}{Count(w_{i-(n-1)}, ..., w_{i-1})} \qquad (۲-۴)$$

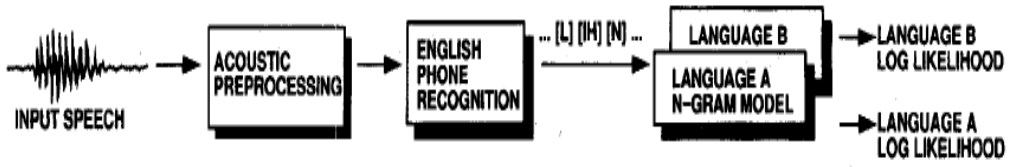
که در این تساوی $Count(w_{i-(n-1)}, ..., w_{i-1}, w_i)$ تعداد دفعاتی که توالی $w_{i-(n-1)}, ..., w_{i-1}, w_i$ در پیکره‌ی آموزشی تکرار شده است را نشان می‌دهد. در مدل‌های زبانی هرچقدر که اندازه‌ی پیکره‌ی آموزشی بزرگتر باشد، مقادیر احتمالات بدست آمده قابل اطمینان‌تر است. از طرف دیگر هرچقدر که n

[1] -N.gram language model

بزرگ‌تر باشد از آن‌جایی که تعداد توالی‌هایی که بایستی برای آن‌ها این میزان احتمال را بدست آوریم، بیشتر می‌شود، بنابراین نیاز به پیکره بزرگتری داریم که به اندازه‌ی کافی رخدادهای n تایی توالی واحدهای زبانی را پوشش می‌دهند.

۴-۲- سیستم PRLM

همانطور که در شکل ۴-۱ می‌بینید، سیستم PRLM از یک بازشناس آوایی و متعاقبا از یک سری مدل‌های زبانی تشکیل شده است. در این روش واحدهای آوایی سیگنال گفتار توسط واحد بازشناس آوا استخراج شده و توالی واحدهای آوایی برای آموزش مدل زبانی n تایی برای هر زبان مورد استفاده قرار می‌گیرد. در طول فاز تشخیص، سیگنال گفتار آزمایشی توسط بازشناس آوا به واحدهای آوایی تجزیه شده و احتمال تعلق این توالی به هر یک از مدل‌های n تایی زبان محاسبه می‌شود. سیگنال گفتار متعلق به زبانی است که مدل مربوط به آن بیشترین احتمال را برگرداند[۱۰].



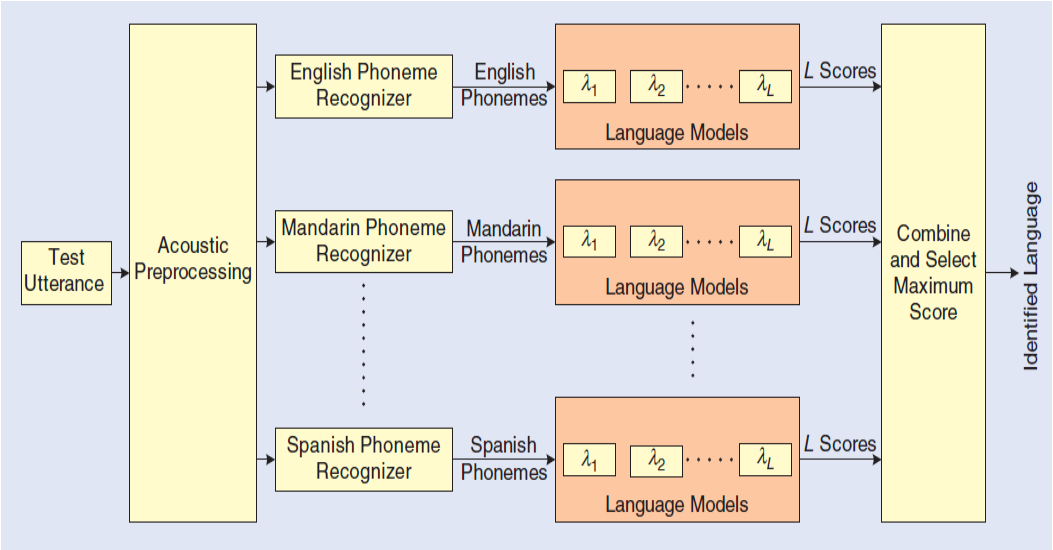
شکل۴-۱: بلوک دیاگرام سیستم PRLM [۱۰]

این سیستم از دو قسمت ابتدایی و انتهایی تشکیل شده است. قسمت ابتدایی که شامل یک بازشناس آوا است. گرچه بازشناس آوا در سیستم PRLM می‌تواند با هر یک از این‌ها آموزش داده شود، اما به‌دلیل وجود سیگنال‌های گفتار برچسب‌دار بیشتری که برای زبان انگلیسی وجود دارد، معمولا از تشخیص‌دهنده‌ی آوای زبان انگلیسی استفاده می‌شود[۱۰].

قسمت انتهایی این سیستم شامل مجموعه‌ای از مدل‌های زبانی است. برای ساخت مدل زبانی n تایی برای هریک از زبان‌ها به این صورت عمل می‌شود که ابتدا سیگنال‌های گفتار آموزشی مربوط به آن زبان به عنوان ورودی به تشخیص دهنده‌ی آوا داده می‌شود و یک مدل آوایی برای توالی آواهای خروجی محاسبه می‌شود. برای مثال می‌توان از یک مدل ۲ تایی برای مدل کردن توالی آواها استفاده کرد[۱۰].

۴-۳- سیستم PPRLM

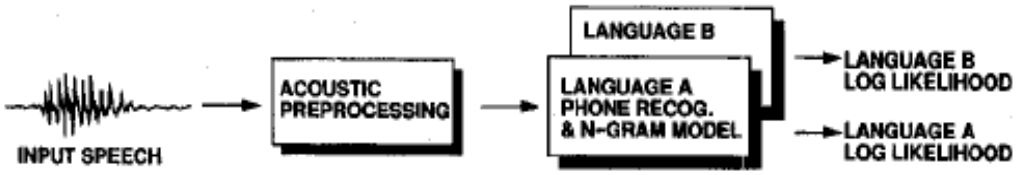
گرچه سیستم PRLM می‌تواند به عنوان یک سیستم موثر برای تشخیص زبان مورد استفاده قرار گیرد، اما ممکن است گاهی اوقات صداهای زبانی که قصد تشخیص آن‌را داریم، در زبانی که برای آموزش تشخیص دهنده‌ی آوا استفاده کرده‌ایم، وجود نداشته باشد، بنابراین طبیعی است که به‌دنبال روشی برای استفاده از آواهای چندزبانه در تشخیص زبان باشیم. برای مثال می‌توان از آواهای چندزبانه برای آموزش تشخیص‌دهنده‌ی آوا استفاده نمود یا می‌توان از چند تشخیص‌دهنده‌ی آوا به‌صورت همزمان استفاده کرد. این روش مستلزم این است که داده‌های آموزشی برچسب‌دار برای بیش از یک زبان در دسترس باشد. یک نمونه از چنین سیستم تشخیص زبان موازی در شکل ۴-۲ برای سه زبان انگلیسی، ماندارین و اسپانیایی نشان داده شده است. قابل توجه است که این سیستم را می‌توان به هر تعداد سیستم PRLM موازی تعمیم داد. تنها محدودیت تعداد زبان‌هایی است که داده‌های آموزشی برچسب‌دار برای آن‌ها موجود است. همچنین از نظر زمانی نیز کندتر از روش PRLM است[۱۰].



شکل ۴-۲- بلوک دیاگرام یک سیستم PPRLM [۳۷]

۴-۴-سیستم PPR

همانطور که در قسمت‌های قبل گفته شد، در سیستم‌های PRLM و PPRLM، پس از بدست آوردن دنباله‌های آوایی با استفاده از مدل‌های زبانی برای زبان‌های مورد نظر، یک آنالیز واج‌آرایی با استفاده از مدل‌های زبانی صورت می‌گیرد. این روش زمانی منطقی است که داده‌های گفتاری بر چسب‌دار برای هر زبانی که قصد تشخیص آن‌را داریم در دسترس نباشد. در صورت وجود داده‌های برچسب‌دار برای تمامی زبان‌های مورد نظر می‌توان از استراتژی PPR استفاده کرد. بلوک دیاگرام این سیستم در شکل ۴-۳ برای دو زبان مختلف نشان داده شده است. در این روش بر خلاف روش‌های PRLM و PPRLM که ابتدا توالی واج‌ها را تشخیص داده و سپس محدودیت‌های واج‌آرایی مربوط به هر زبان را با استفاده از مدل‌های زبانی اعمال می‌کنند، این محدودیت‌ها در حین جستجو برای پیدا کردن بهترین دنباله‌ی واجی اعمال می‌شود[۱۰].



شکل ۴-۳- بلوک دیاگرام سیستم PPR [۱۰]

۴-۵-سیستم TOPT[1]

این سیستم که توسط آقای تنگ ارائه شده است مشابه سیستم PPRLM است، با این تفاوت که با داشتن یک مجموعه از بازشناس‌های آوایی، مجموعه‌ی جدیدی از بازشناس‌ها که قدرت بیشتری در بازشناسی زبان‌ها از یکدیگر دارند را تولید می‌نماییم. به این صورت که هر بار مجموعه‌ای از آواهای هر بازشناس آوایی به ازای هر زبان هدف، برای تولید بازشناس‌های آوایی جدید مورد استفاده قرار می‌گیرد. این روش باعث بهبودی کارایی سیستم بدون نیاز به داده‌های آموزشی زیادتر می‌شود. همچنین در این روش برای ساخت بازشناس‌های آوایی جدید، مجموعه آواهایی که قدرت تمایزکنندگی بیشتری دارند انتخاب می‌شوند، در نتیجه اندازه‌ی مجموعه آوای بازشناس‌های آوایی، کوچکتر است و می‌توان مدل‌های آوایی آماری با مرتبه‌ی بالاتر را تخمین زد. روش کار بدین صورت است که برای هر بازشناس آوا و به ازای هر زبان هدف، ابتدا احتمالات یکتایی[2] مجموعه‌ی آواها محاسبه شده و در نتیجه برداری بصورت $[x_1, x_2, ..., x_{vf}]$ با ابعاد vf بدست می‌آید. در مرحله‌ی بعد مجموعه بردارهای حاصل برای آموزش مدل‌های SVM به ازای هر زبان، مورد استفاده قرار می‌گیرند. پس از اتمام آموزش، هر SVM یک ابر صفحه به عنوان خروجی بر می‌گرداند که به‌صورت رابطه‌ی۵-۳ است:

[1] -Target Oriented Phone Tokenizer
[2] -bigram(ngram,n=1)

$$f(x) = a^T \varphi(x) + b \qquad (۳-۴)$$

که در آن درایه‌ی iام از بردار a (a_i) وزن ویژگی x_i از بردار ویژگی x است. حاشیه‌ی ابر صفحه با $\|a_i\|$ رابطه‌ی عکس دارد، بنابراین هرچه $|a_i|$ بزرگتر باشد، در تعیین ابر صفحه تاثیر گذارتر است. آوای متناظر با ویژگی‌هایی که وزن بیشتری دارند به عنوان آوا برای ساخت بازشناس‌های آوایی جدید انتخاب می‌شوند[۲۰].

با استفاده از این روش، در صورتی که F بازشناس آوایی و L زبان هدف وجود داشته باشد، به تعداد F*L بازشناس آوایی جدید بدست خواهد آمد و از نظر علمی بایستی زیر مجموعه‌ای از این بازشناس‌ها را انتخاب نمود. برای انتخاب به ازای هر بازشناس آوایی بدست آمده یک بردار کد تولید شده و بازشناس‌هایی که بردار کد متمایزی داشته باشند انتخاب خواهند شد. برای l امین بازشناس آوایی جدید حاصل از f امین بازشناس آوا، یک بردار کد به‌صورت زیر محاسبه می‌شود:

$$[C_1^l, C_2^l, ..., C_i^l, C_{vf}^l] C_i^l \in [0,1] \qquad (۴-۴)$$

در صورتی که i امین آوا موجود در بازشناس f ام در این بازشناس وجود داشته باشد $C_i^l = 1$ و در غیر این صورت برابر صفر است. برای هر یک از L بردار کد بدست آمده مرتبط به بازشناس f ام، فاصله همینگ هر بردار تا L-1 بردار کد دیگر جمع شده و به عنوان امتیاز نهایی آن بازشناس در نظر گرفته می‌شود. بازشناس‌های آوایی که بالاترین امتیاز را داشته باشند به عنوان بازشناس‌های جدید، در نظر گرفته خواهند شد[۲۰].

در این بخش چند نمونه از سیستم‌های تشخیص زبان مبتنی بر اطلاعات آوایی که در تحقیقات اخیر ارائه شده است، شرح داده شد. این سیستم‌ها جزو بهترین سیستم‌هایی هستند که تا کنون در کاربرد تشخیص زبان ارائه شده است. قدیمی‌ترین این سیستم‌ها، سیستم PPRLM است که هنوز هم از محبوبیت قابل توجهی برخوردار است. سیستم TOPT جزو سیستم‌هایی هستند که اخیرا معرفی شده‌اند و در قسمت‌های قبل به تفصیل شرح داده شد.

۴-۷-شبکه‌های عصبی به عنوان شناساگرهای زبان

امروزه کلاسه‌بندی یکی از المان‌های کلیدی در حوزه‌ی یادگیری ماشینی است. هدف از کلاسه‌بندی، تمایز بین دو یا چند کلاس از اشیایی است که ویژگی‌های متفاوتی دارند. در یادگیری ماشینی، الگوریتم‌هایی را برای شناسایی اشیاء توسعه می‌دهیم. الگوریتم‌هایی بر مبنای توصیف‌های ریاضی،ساختاری و یا آماری. شبکه‌های عصبی مصنوعی با الهام از شبکه‌های عصبی زیستی در حوزه‌ی کلاسه بندی اشیاء در چند دهه‌ی اخیر توسعه‌ی چشم‌گیری یافتند. با این وجود مدل‌های موجود تقلید ناقصی از سیستم عصبی انسان هستند. شروع کار شبکه‌های عصبی به سال ۱۹۴۳، هنگامی‌که pitts و McCulloch اولین مدل ریاضی را برای نورون‌های زیستی مطرح کردند می‌رسد. شبکه‌های عصبی پس از وقفه‌ای ۱۵ ساله دوباره از دهه‌ی ۸۰ با طرح روش آموزش پس‌انتشار خطا برای شبکه‌های چندلایه جان تازه‌ای گرفت و از آن پس پیشرفت چشمگیری داشت. هم اکنون از شبکه‌های عصبی در زمینه‌های مختلفی همچون شناسایی الگو، شناسایی سیستم، خوشه‌یابی، تخمین توابع و زمینه‌های متعدد دیگر استفاده می‌شود. در بحث شناسایی الگو عموما از شبکه‌های MLP و RBF استفاده می‌شود. شبکه‌هایی که در بین شبکه‌های عصبی از گستردگی استفاده و محبوبیت بیشتری برخوردار است. در این پایان‌نامه نیز از این دو شبکه‌ی عصبی و همچنین از شبکه‌ی عصبی جدید دیگری با نام WRBF به عنوان کلاسه

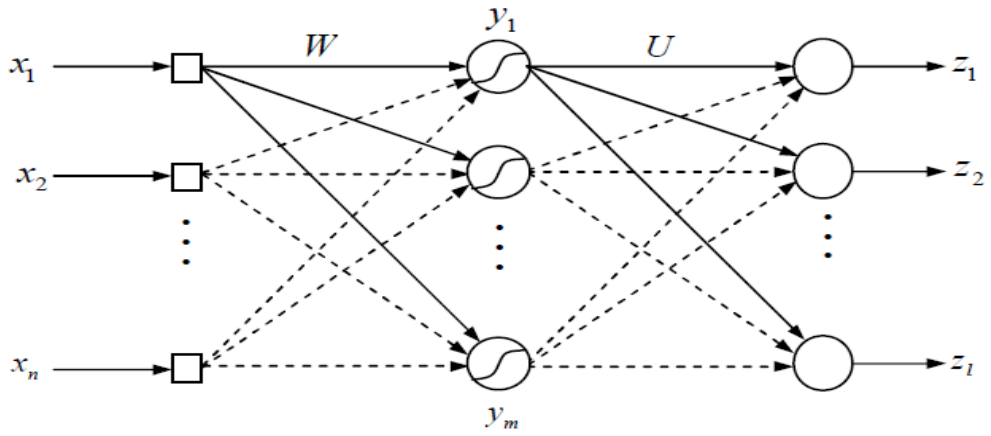
بند استفاده شده است که در فصل بعد نتایج استفاده از این شبکه‌های عصبی به تفصیل مورد ارزیابی قرار خواهد گرفت.

۴-۷-۱- شبکه‌ی MLP

برای حل مسائل ساده که به صورت خطی قابل حل هستند می‌توانیم از شبکه های عصبی تک لایه (پرسپترون) استفاده کنیم. اما گاهی اوقات که مسائل پیچیده‌تر می شوند دیگر این شبکه‌ها پاسخگو نیستند و باید یک یا چند لایه‌ی دیگر با تعدادی نرون در این لایه به شبکه اضافه کرد که به این لایه‌ها، لایه‌های مخفی می‌گویند(لایه‌های بین لایه‌های ورودی و خروجی). شبکه‌های عصبی چند لایه به طور وسیعی در زمینه‌های متنوعی از قبیل طبقه بندی الگوها، پردازش تصاویر، تقریب توابع و غیره مورد استفاده قرار گرفته است. الگوریتم یادگیری پس‌انتشار خطا[1](BP)، یکی از رایج‌ترین الگوریتم‌ها جهت آموزش شبکه‌های عصبی چند لایه می‌باشد. از قانون یادگیری پس انتشار خطا (BP)، برای آموزش شبکه‌های عصبی چند لایه‌ی پیش‌خور که عموما شبکه‌های چند لایه پرسپترون (MLP) هم نام می‌گیرند، استفاده می‌شود. به عبارتی توپولوژی شبکه‌های MLP، با قانون یادگیری پس‌انتشار خطا تکمیل می‌شود. بطور خلاصه، فرایند پس انتشار خطا از دو مسیر اصلی تشکیل می شود. مسیر رفت[۴] و مسیر برگشت [۵] .در مسیر رفت، یک الگوی آموزشی به شبکه اعمال می شود و تأثیرات آن از طریق لایه های میانی به لایه‌ی خروجی انتشار می یابد تا این‌که نهایتا خروجی واقعی شبکه MLP، به دست می‌آید. در این مسیر، پارامترهای شبکه(ماتریس‌های وزن و بردارهای بایاس)، ثابت و بدون تغییر در نظر گرفته می‌شوند.

[1] - Back-Propagation Algorithm

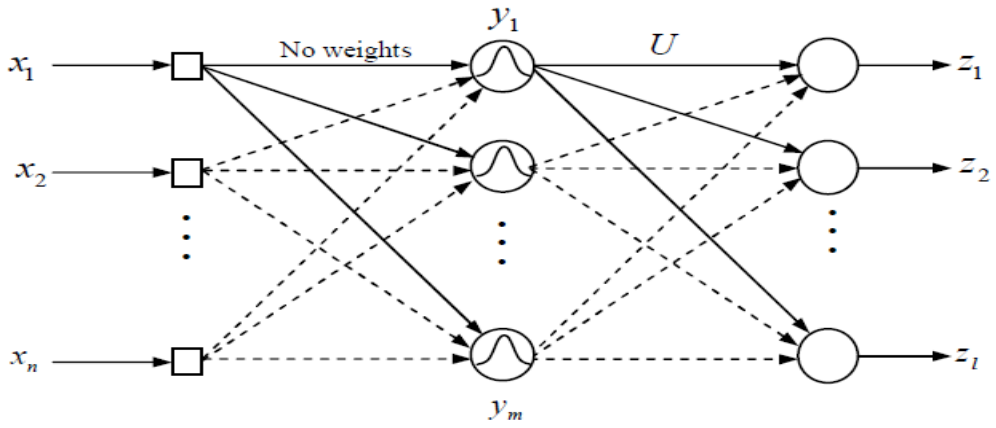
در مسیر برگشت، برعکس مسیر رفت، پارامترهای شبکه‌ی MLP تغییر و تنظیم می گردند. این تنظیمات بر اساس قانون یادگیری اصلاح خطا انجام می‌گیرد. سیگنال خطا، در لایه‌ی خروجی شبکه تشکیل می‌گردد. بردار خطا برابر با اختلاف بین پاسخ مطلوب و پاسخ واقعی شبکه می باشد. مقدار خطا، پس از محاسبه، در مسیر برگشت از لایه‌ی خروجی و از طریق لایه‌های شبکه به سمت پاسخ مطلوب حرکت می‌کند. در شبکه‌های MLP، هر نرون دارای یک تابع تحریک غیرخطی است که از ویژگی مشتق پذیری برخوردار است. در این حالت، ارتباط بین پارامترهای شبکه و سیگنال خطا، کاملا پیچیده و غیرخطی می‌باشد، بنابراین مشتقات جزئی نسبت به پارامترهای شبکه به راحتی قابل محاسبه نیستند. جهت محاسبه مشتقات از قانون زنجیره‌ای معمول در جبر استفاده می‌شود. شکل زیر ساختار یک شبکه عصبی MLP را نشان می‌دهد.



۴-۴: ساختار یک شبکه عصبی MLP [۵۵]

۴-۷-۲-شبکه‌ی RBF

شبکه توابع شعاعی پایه یکی از انواع شبکه‌های عصبی است. در دهه‌های اخیر از این شبکه به عنوان کلاسه بند نیز استفاده شده و نتایج درخور توجهی از آن بدست آمده است. برای اولین بار در سال ۱۹۶۴ بشکیرف از شبکه‌یRBF برای طبقه‌بندی استفاده کرد[۵۶]. پس از آن از این شبکه در مسائل شناسایی الگو استفاده شد، تا آنجا که امروزه این شبکه به‌عنوان یک کلاسه‌بند قابل اطمینان بشمار می‌رود[۶۰و۵۹و۵۷و۵۸]. شبکه‌ی RBF یک شبکه‌ی سه لایه است؛ لایه‌ی اول لایه‌ی ورودی است که ویژگی‌ها به آن وارد می‌شوند، لایه‌ی دوم یا لایه‌ی مخفی شامل توابع شعاعی پایه است و نام شبکه از همین توابع گرفته شده است و لایه‌ی سوم لایه‌ی خروجی است که شامل توابع فعالسازی شبه سیگموید است[۵۷]. شکل زیر ساختار ساده‌ی یک شبکه‌ی RBF را نشان می‌دهد.



شکل۴-۵: شبکه عصبی RBF [۵۵]

لایه‌ی ورودی، ویژگی‌های استخراج شده از نمونه‌ها را دریافت کرده و بدون تغییر به لایه‌ی میانی می‌دهد. نرون‌های لایه‌ی میانی توابع شعاعی هستند که باید چهار خاصیت را دارا باشند:

- نسبت به مرکز متقارن باشند.
- بیشترین مقدار را در مرکز داشته باشند.
- با فاصله گرفتن از مرکز مقدار تابع کم شود.
- مشتق پذیر باشند.

نوعا از توابع گاوسی برای این لایه استفاده می‌شود[۵۵].

$$y_m = \phi_m(X - \mu_m) = \exp(-\frac{\|X - \mu_m\|^2}{2\sigma_m^2} \quad (۷-۴)$$

در این تابع X بردار ورودی شبکه عصبی، m اندیس نرون مخفی، μ بردار میانگین تابع گاوسی با ابعادی برابر با ابعاد بردار X و σ پارامتر پراکندگی آن است. بدلیل زیاد بودن تعداد نمونه‌ها در اکثر مسائل شناسایی الگو تعداد نرون‌های لایه‌ی میانی کمتر از تعداد نمونه‌های آموزش انتخاب می‌شود. اما از طرفی چون رفتار شبکه‌ی RBF و MLP با یکدیگر متفاوت است و هر نرون محدوده‌ای از داده‌های ورودی را می‌بایست پوشش دهد، عموما تعداد نرون‌های خیلی بیشتری نسبت به MLP نیاز است[۶۱و۶۲].

لایه‌ی خروجی در شبکه‌ی عصبی RBF تماما شبیه شبکه‌ی MLP بوده و نرون‌های این لایه می‌توانند هر یک از توابع فعالسازی سیگموید، خطی، تانژانت هیپربولیک و یا توابع مرسوم دیگر را استفاده کنند[۶۳]. وزن‌هایی نیز بین لایه‌ی میانی و خروجی وجود دارد که آموزش‌پذیر می‌باشند. در نهایت ورودی نرون‌های لایه‌ی خروجی مجموع وزن یافته‌ی توابع شعاعی پایه هستند. در بحث آموزش شبکه‌های RBF عموما از روش پس‌انتشار خطا استفاده می‌شود که این نوع آموزش در شبکه‌ی MLP نیز مرسوم است. در این پایان نامه ضمن استفاده از این شبکه، عملکرد دو شبکه‌ی RBF و شبکه‌ی جدید WRBF که در بخش بعدی توضیح داده می‌شود نیز پرداخته خواهد شد.

۴-۷-۳-شبکه WRBF

شبکه‌ی WRBF شبکه‌ای جدید است که در آن از شبکه‌های RBF و MLP الهام گرفته شده است. همانطور که در قبل ذکر شد دو شبکه‌ی مذکور در بخش‌های قبل، دارای یک لایه‌ی ورودی یک لایه‌ی خروجی و یک لایه‌ی مخفی یا میانی می‌باشند که در شبکه‌ی MLP لایه‌ی میانی بیشتر از یکی هم می‌تواند باشد. لازم به ذکر است که یک لایه‌ی مخفی نیز برای حل هر مسئله‌ی غیر خطی شناسایی الگو کفایت می‌کند[۶۴]. همچنین دو شبکه‌ی RBF و MLP شباهت‌های دیگری نیز با یکدیگر دارند و از جمله‌ی آن‌ها این که اتصالات بین لایه‌ی میانی و خروجی شامل وزن‌هایی هستند که در طول فرایند آموزش به روز می‌شوند[۵۵]. اما دو تفاوتی که این دو شبکه با یکدیگر دارند در وزن‌های بین لایه‌ی میانی و ورودی و همچنین نوع توابع فعالسازی در لایه‌ی میانی می‌باشد. به این صورت که در شبکه‌ی RBF در لایه‌ی میانی از تابع فعالسازی گاوسی استفاده می‌شود، در صورتی‌که در شبکه‌ی MLP توابع شبه سیگموید استفاده می‌شود. از طرفی دیگر در شبکه‌ی MLP بین لایه‌ی ورودی و میانی وزن‌هایی آموزش‌پذیر وجود دارد که این وزن‌ها در شبکه‌ی RBF وجود ندارد.

در شبکه WRBF بردار وزنی به نام W درساختار RBF اضافه شده است که سبب می‌شود هر کدام از گره‌های ورودی، وزن منحصر به فردی متفاوت از مقدار ۱ داشته باشند[۵۵]. با این کار توابع شعاعی به جای این‌که روی ورودی‌ها(بردارX) عمل کنند، روی نگاشت خطی از ورودی‌ها(W.X) عمل خواهند کرد:

$$y_m = f_m(X) = \exp(-\frac{\|W_m.X - \mu_m\|^2}{2\sigma_m^2}) \qquad (۸-۴)$$

در اینجا $W_m = \{w_{1m}, w_{2m}, ...w_{nm}\}$ یک بردار وزن بین ورودی و نرون مخفی m است که ابعاد آن با ابعاد بردار ورودی برابر است. با افزودن این بردار به نرون‌های ورودی اجازه داده می‌شود وزن‌های متفاوتی داشته باشند و به تبع آن تاثیر متفاوتی در شناسایی شبکه داشته باشند[۵۵]. در فصل پایانی

خواهیم دید که این شبکه‌ی جدید عملکرد بهتری را از خود در برابر شبکه‌هایRBF و MLP ارائه خواهد کرد.

فصل پنجم

آزمایش‌ها

۵-آزمایش‌ها

۵-۱-مقدمه

در این فصل آزمایشات انجام شده در زمینه‌ی تشخیص نوع زبان آمده است. ویژگی‌های مورد استفاده در این آزمایشات عبارتند از MFCC، PLP، LPC و تبدیل فوریه‌ی بسل.

در دو مرحله از آزمایشات ابتدا یکی از انواع ویژگی‌های مذکور از سیگنال گفتار استخراج شده است. سپس این ویژگی‌ها به شبکه‌های عصبی RBF،MLP و WRBF جهت تشخیص نوع زبان داده شده و نتایج حاصله از این آزمایشات با یکدیگر مقایسه شده‌اند.

۵-۲- دادگان OGI-TS

پایگاه داده‌ی مورد استفاده در این پایان‌نامه دارای ۱۰ زبان گفتاری تلفنی است که در ابتدا توسط آقای موسوسامی جمع آوری شد. این گفتار شامل ۹۰۰ تماس تلفنی بود که زبان‌های انگلیسی، ماندارین، فارسی، اسپانیایی، فرانسوی، تامیل، آلمانی، ویتنامی، ژاپنی و کره‌ای را دارا بود. بعد از آن و در چند سال

بعد این دادگان با استفاده از افزایش ضبط‌ها به ازای هر یک از ۱۰ زبان و افزودن زبان هندی به مجموعه زبان‌ها، گسترش یافت. این دادگان یکی از دادگان‌های استانداردی است که NIST برای ارزیابی سیستم‌های تشخیص زبان در نظر گرفته است. در این مجموعه دادگان یک سری سئوالات توسط هر تماس گیرنده مطرح می‌شود که جواب هر یک از این سئوالات به عنوان دادگان جمع‌آوری شده است[۶۵]. در زیر بعضی از این سئوالات آورده شده است:

- اکثرا به چه زبانی صحبت می‌کنید؟(۳ثانیه)
- زبان بومی شما چیست؟(۳ثانیه)
- بیشترین غذاهای سرو شده در چند وقت گذشته توسط شما چه بوده؟(۵ثانیه)
- روزهای هفته را نام ببرید.(۸ثانیه)
- اعداد از ۰ تا ۱۰ را به ترتیب بیان کنید.(۱۰ثانیه)
- در مورد آب و هوای شهر خود صحبت کنید.(۱۰ثانیه)
- در مورد چیزهایی که در شهر خود به آن‌ها علاقه‌مند هستید صحبت کنید(۱۰ثانیه)
- اتاقی که در آن هستید را شرح دهید.(۱۲ثانیه)
- در مورد یک موضوع به دلخواه صحبت کنید.(۴۵ثانیه)

۵-۳-استخراج ویژگی

همانطور که در قبل آمده است دیتا بیس مورد استفاده‌ی ما یعنی دیتا بیس OGI-TS دارای نمونه‌هایی با زمان‌های مختلف است. در ابتدا این دیتا بیس را از حیث تفاوت زمانی دسته بندی می‌کنیم. انتخاب نمونه‌ها و تعیین مرد یا زن بودن گوینده‌ها نیز به صورت اتفاقی انجام شد. در این بین بعضی از این دسته بندی‌ها تنوع بیشتری در محتوای صحبتی دارد. نمونه‌های ۵ و ۱۰ ثانیه‌ای که این مزیت را

دارند را انتخاب کرده و آزمایشات را بر روی آن‌ها انجام می‌دهیم. همچنین زبان‌های مختلفی که در این دیتا بیس موجود می‌باشد را به طور دوتایی، چهار تایی و شش تایی با یکدیگر مقایسه می‌کنیم.

برای استخراج ویژگیMFCC ابتدا سیگنال ورودی را به فریم‌های ۲۵ میلی ثانیه‌ای با ۱۰ میلی ثانیه همپوشانی تقسیم کرده و سپس از هر فریم ۸ ضریب MFCC به همراه انرژی، مشتق اول و مشتق دوم این ضرایب را استخراج نمودیم. در نتیجه نمونه‌های گفتاری ما به بردارهای ویژگی ۲۴ بعدی تبدیل می‌شود. در روش بعدی یعنی روش PLP نیز به همان صورت، یعنی فریم‌های ۲۵ میلی ثانیه‌ای با ۱۰ میلی ثانیه همپوشانی مورد نظر ماست. سپس در مرحله‌ی نهایی ۹ ضریب از ضرایب PLP را جدا کرده و در نتیجه خروجی برنامه توالی از بردارهای ویژگی ۹ تایی می‌باشد. نتایج حاصله از انتخاب تعداد ضرایب نشان می‌دهد که تعداد ۹ ضریب از PLP وLPC و ۸ ضریب از MFCC بهترین نتیجه را برای این پروژه رقم می‌زند. در نهایت برای ویژگی تبدیل فوریه‌ی بسل تعداد ۲۰۰ ویژگی و همپوشانی ۵۰ درصد بهترین نتیجه را به ما می‌دهد. در جداول زیر نتایج حاصل از بازشناسی۲ زبانه، توسط شبکه‌ی عصبی MLP به ازای تغییر پارامترهای موثر در فرایند استخراج ویژگی نشان داده شده است.

جدول۵-۱-تاثیر تعداد ضرایب در استخراج ویژگی MFCC

تعداد ویژگی	۲۱	۲٤	۲۷	۳۰
درصد صحت	۸۸/۵۰	۹۰/۴۱	۹۰/۰۱	۸۹/۰۹

جدول۵-۲- تاثیر تعداد ضرایب در استخراج ویژگی PLP

تعداد ویژگی	۷	۸	۹	۱۰
درصد صحت	۸۴/۲۳	۸۵/۵۴	۸۶/۳۰	۸۵/۵۲

جدول۵-۳-تاثیر تعداد ضرایب در استخراج ویژگی LPC

تعداد ویژگی	۷	۸	۹	۱۰
درصد صحت	۷۰/۷۰	۷۳/۷۱	۷۴/۹۰	۷۳/۲۸

جدول۵-۴- تاثیر تعداد ضرایب در استخراج ویژگی بسل

تعداد ویژگی	۵۰	۱۰۰	۱۵۰	۲۰۰
درصد صحت	۷۶/۹۳	۸۱/۱۹	۸۲/۳۰	۸۳/۰۰

همان‌طور که از جداول بالا مشخص است، تعداد ضرایب MFCC و PLP و LPC و تبدیل فوریه‌ی بسل که بهترین نتایج را به ما می‌دهد به ترتیب ۲۴و۹و۹و۲۰۰ است، اما از آنجا که در تبدیل فوریه‌ی بسل تعداد بالاتر ضرایب سرعت ما را پایین می‌آورد، برای وجود موازنه‌ای بین سرعت و دقت ما تعداد ۱۵۰ ضریب را انتخاب می‌کنیم.

۵-۳-۱- روش SDC

زبان‌ها از نظر آواهایی که برای تولید گفتار استفاده می‌کنند، تشابهات زیادی دارند، لیکن تفاوت زبان‌ها را می‌توان از نظر فرکانس تکرار آواها و همچنین آواهایی که می‌توانند در یک زبان در کنار یکدیگر قرار گیرند مقایسه کرد. یک روش برای مدل کردن اطلاعات واج آرایی این است که اطلاعات واج‌آرایی را در سطح ویژگی مدل کنیم. یعنی بردار ویژگیی استخراج کنیم که بتواند تغییرات آوایی را مدل کند و

بازه‌ی زمانی طولانی‌تری از گفتار که تغییرات آوایی را در بر می‌گیرد را در نظر بگیرد. با رجوع به کارهای مختلفی که در دوره‌های قبل در زمینه‌ی تشخیص زبان ارائه شده است، دو ویژگی SDC و TFC قادر به مدل کردن اطلاعات واج آرایی در سطح ویژگی می‌باشند. این دو ویژگی در فصل سوم مورد بررسی قرار گرفتند. ما در این قسمت از یکی از این ویژگی ها یعنی SDC با مولفه های ۷-۳-۱-۷ به عنوان ویژگی استفاده می‌کنیم و آن را با حالتی که به تنهایی از ویژگی‌های دیگر استفاده می شود مقایسه می‌کنیم. نتیجه‌ی دقت تشخیص زبان برای حالتی که به تنهایی از ویژگی‌های MFCC و PLP و LPC و تبدیل فوریه‌ی بسل استفاده شده است و حالتی که از روش SDC نیز به همراه آن‌ها استفاده شده است در شکل‌های آتی آمده است. لازم به ذکر است که روش SDC ارائه شده در این بخش توسط شبکه‌ی عصبی MLP اجرا شده است.

۵-۴-شبکه‌های عصبی

در این بخش به توضیح تاثیر تغییر پارامترهای شبکه‌های مورد استفاده در این پایان‌نامه می‌پردازیم. همچنین هر کدام از روش‌های استخراج ویژگی را با استفاده از شبکه‌های MLP و RBF و WRBF برای نمونه‌های ۵ و ۱۰ ثانیه‌ای مورد استفاده قرار داده و با یکدیگر مقایسه می‌کنیم. ۲۵٪ از کل داده‌های موجود که به‌صورت تصادفی انتخاب شده اند را به عنوان مجموعه‌ی آموزش و ۷۵٪ از داده‌ی باقیمانده را به عنوان مجموعه‌ی آموزش در نظر می‌گیریم. نتایج این آزمایشات در جداول و شکل‌های این بخش آمده است.

۵-۴-۱- شبکه MLP

ما در این پایان‌نامه از شبکه عصبی MLP بهره می‌بریم که دارای یک لایه‌ی ورودی یک لایه‌ی میانی و یک لایه‌ی خروجی می‌باشد همانطور که در بخش۴ نیز ذکر شد می توان از بیش از یک لایه نیز

در این بخش استفاده کرد لیکن اثبات شده است که یک لایه‌ی مخفی برای هر مسئله‌ی غیر خطی شناسایی الگو کفایت می‌کند[۶۴]، پس در اینجا منظور ما از شبکه MLP همان شبکه با یک لایه نرون مخفی می‌باشد. همچنین از تابع فعالسازی سیگموید و روش آموزش گرادیان نزولی[1] در آزمایشات خود بهره می‌بریم. با توجه به آزمایشات انجام شده تعداد نرون‌های لایه‌ی میانی که بهترین عملکرد را در این پروژه داشته است ۲۰ نرون بوده که نتایج این آزمایش بر روی دو زبان انگلیسی و ژاپنی برای ویژگی PLP در جدول زیر آورده شده است.

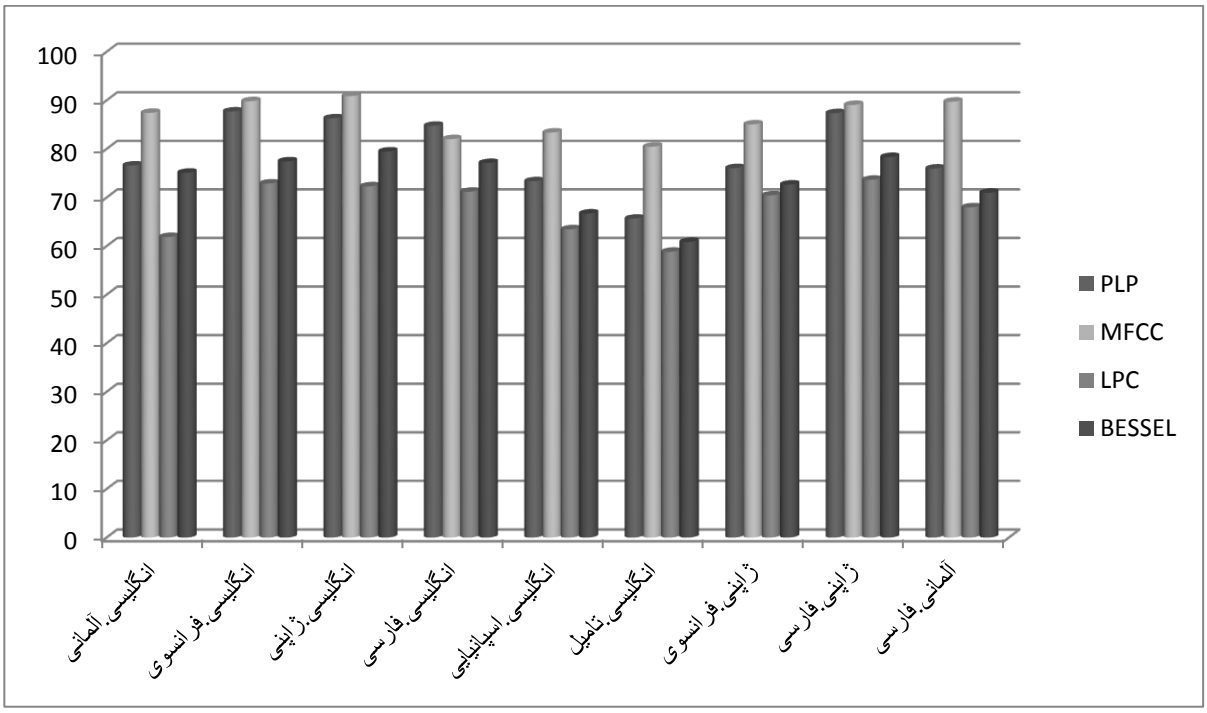
جدول ۵-۵-تاثیر تعداد نرون‌های لایه‌ی میانی شبکه‌ی MLP در دقت تشخیص زبان

تعداد نرون	۱۵	۱۷	۲۰	۲۳	۲۶	۳۰
درصدصحت	۸۵/۸۶	۸۶/۰۸	۸۶/۳۰	۸۶/۰۹	۸۵/۳۹	۸۵/۲۸

همان‌طور که در جدول بالا آمده است تعداد ۲۰ نرون لایه‌ی میانی بهترین نتایج را در این آزمایش می‌دهد. لذا در قسمت‌های دیگر آزمایشات نیز از ۲۰ لایه‌ی مخفی در شبکه‌ی عصبی MLP استفاده می‌نماییم.

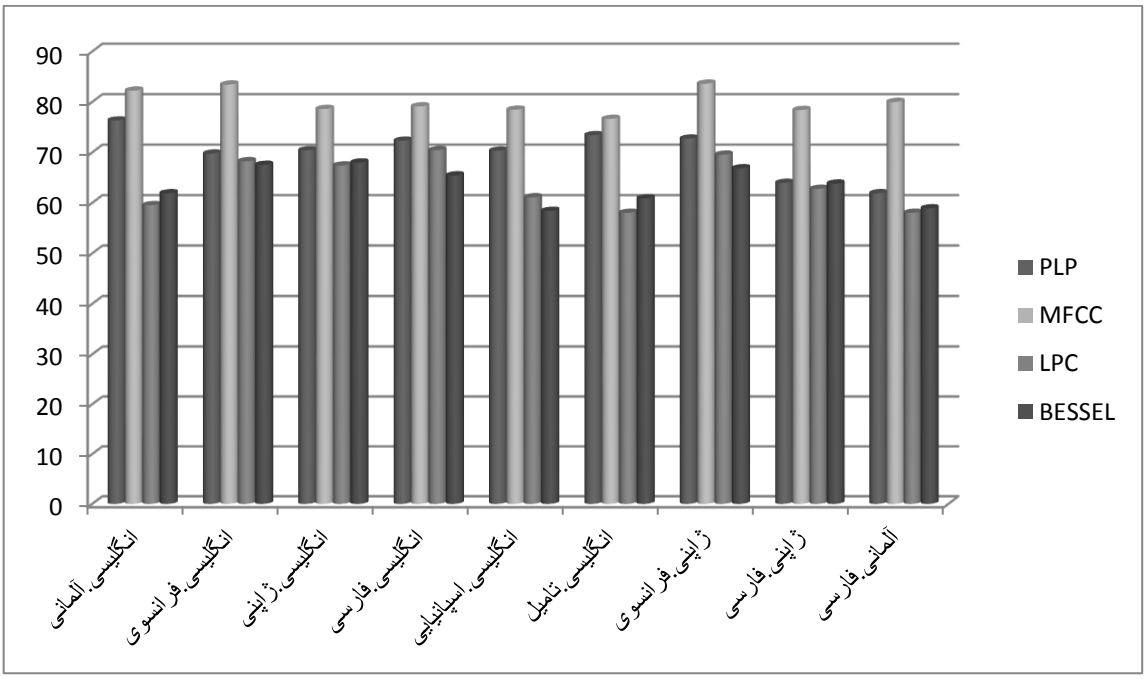
حال پس از بررسی میزان تاثیر تعداد نرون‌های لایه‌ی میانی در دقت تشخیص زبان به سراغ مقایسه‌ی تاثیر روش‌های استخراج ویژگی در دقت تشخیص زبان می‌رویم.

[1] - Gradient Descent



شکل۵-۱: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی MLP برای نمونه‌های صوتی ۱۰ ثانیه ای

همان‌طور که از شکل ۵-۱ مشخص است ضرایب MFCC بیشترین درصد دقت را در میان دیگر ضرایب دارا می‌باشند. البته یکی از دلایلی که موجب شد میزان این دقت بیشتر از ضرایب PLP باشد این است که ما در مراحل بدست آوردن ضرایب MFCC از انرژی و مشتق اول و دوم این ضرایب استفاده کردیم. همچنین با توجه به نزدیکی درصد دقت ضرایب تبدیل فوریه‌ی بسل به ضرایب کپستروم و این که در همه‌ی این موارد دقت روش تبدیل فوریه‌ی بسل از دقت روش LPC بیشتر است می‌توان این نتیجه را گرفت که استفاده از این روش در تشخیص زبان امیدوار کننده می‌باشد و با انجام تحقیقات بیشتر روی این ضرایب شاید دقت این روش حتی بیشتر از روش‌های MFCC و PLP نیز شود. در شکل ۵-۲ مقایسه‌ای نیز بین همین ویژگی‌ها برای نمونه‌های ۵ ثانیه‌ای از دیتا بیس OGI-TS شده است.



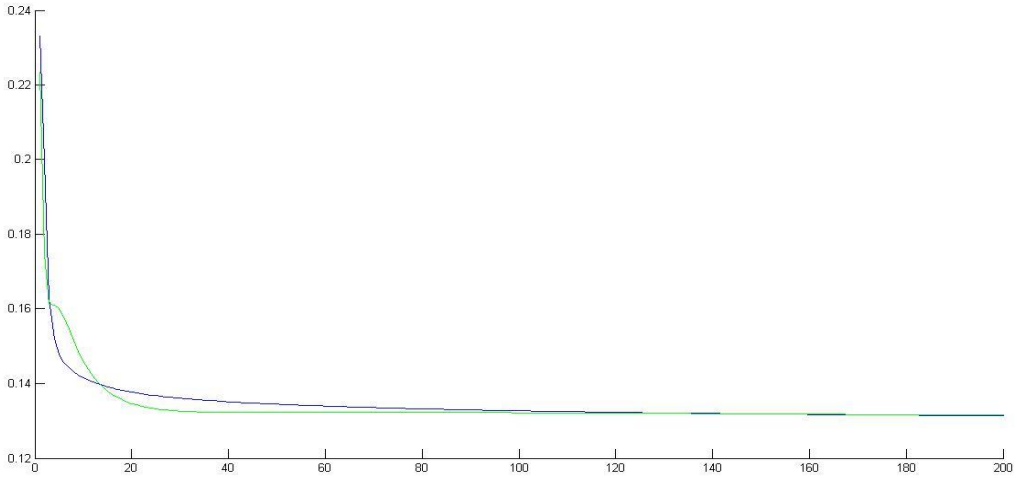
شکل۵-۲: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی MLP برای نمونه‌های صوتی ۵ ثانیه ای

آنچه از دوشکل ۵-۱ و ۵-۲ مشخص است این است که با کم شدن زمان نمونه‌های مورد آزمایش، دقت تشخیص نیز به شدت پایین می‌آید. این درحالی است‌که ضرایب MFCC به نسبت دیگر ضرایب میزان افت کمتری داشته است. در جدول ۵-۶ میزان بهبود دقت هر کدام از جفت زبان‌ها در نمونه‌های ۱۰ ثانیه ای نسبت به نمونه‌های ۵ ثانیه ای آورده شده است. قابل پیش بینی است که با زیاد شدن زمان نمونه‌ها دقت بیشتری در خروجی داشته باشیم. مسلما هنگامی‌که نمونه‌های ورودی به شبکه بیشتر باشد شبکه آموزش بهتری را می‌بینند.

جدول ۵-۶- میزان بهبود دقت هر کدام از جفت زبان‌ها در نمونه‌های ۱۰ ثانیه ای نسبت به نمونه‌های ۵ ثانیه ای

	انگلیسی. آلمانی	انگلیسی. فرانسوی	انگلیسی .ژاپنی	انگلیسی. فارسی	انگلیسی. اسپانیایی	انگلیسی. تامیل	ژاپنی. فرانسوی	ژاپنی .فارسی	آلمانی. فارسی
Plp	۰/۲۸	۱۷/۹۶	۱۵/۸۹	۱۲/۴۳	۳/۰۲	۷/۷۵	۳/۳۲	۲۳/۴۳	۱۴/۰۶
MFCC	۵/۲۱	۵/۴	۱۲/۳۲	۲/۹۳	۴/۹۵	۳/۸۳	۱/۴۵	۱۰/۶۹	۹/۲۶
LPC	۲/۳۵	۴/۶۵	۴/۹۱	۱/۷۷	۲/۳۷	۰/۸۱	۰/۹۲	۱۰/۹	۹/۹۶
BESSEL	۱۲/۸	۱۴/۵۴	۵/۸۳	۲۳/۰	۸/۳	۱۱/۶۹	۱۱/۴۹	۹/۹۲	۱۳/۲۱

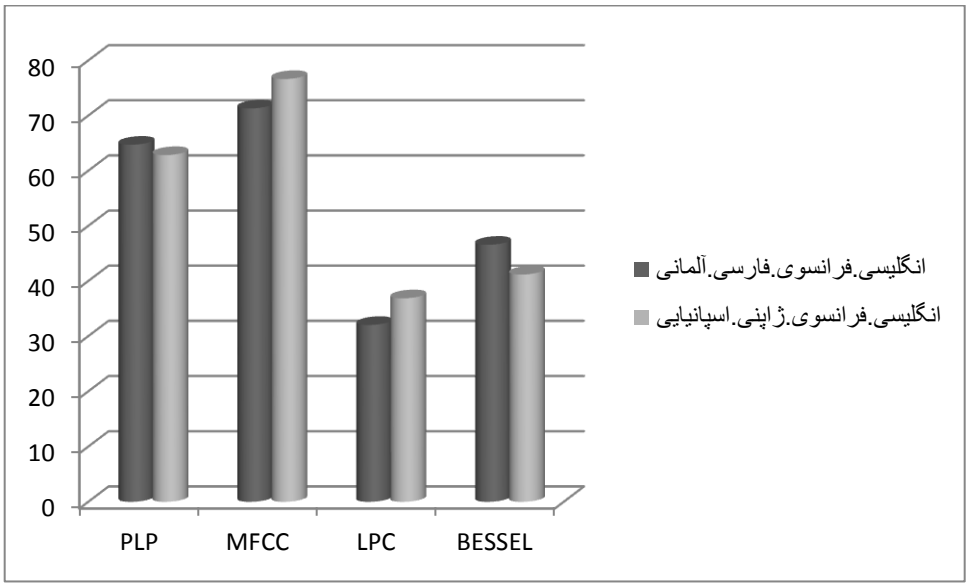
همچنین نمودار خطای MSE برای مقایسه بین دو زبان انگلیسی و آلمانی مشاهده می‌شود.



شکل۵-۳: نمودار خطای MSE برای نمونه‌ی ۱۰ ثانیه‌ای زبان‌های انگلیسی-آلمانی

با توجه به این نمودار دو منحنی در شکل وجود دارد که یکی مربوط به داده‌های آموزش است و دیگری مربوط به داده‌های ولیدیشن می‌باشد.

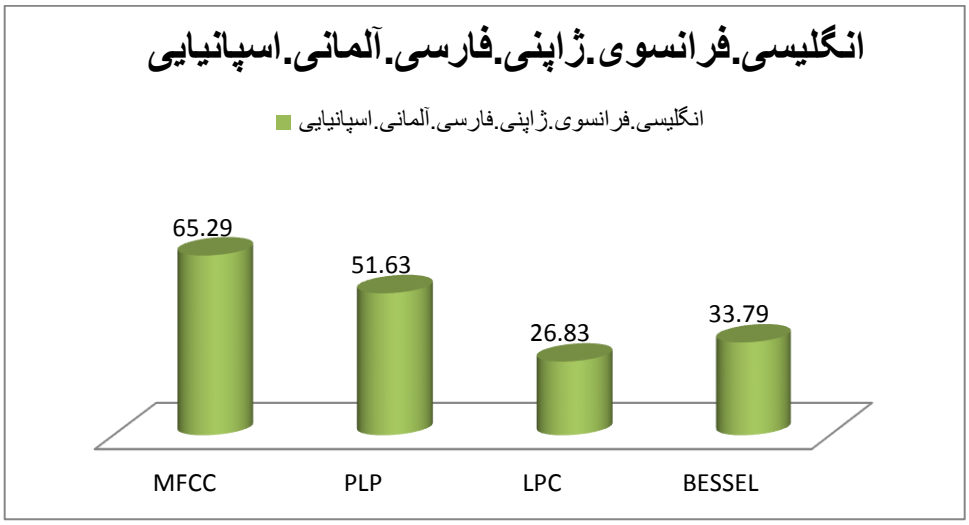
همچنین در اشکال بعدی مقایسه ای بین چهار و شش زبان از مجموعه دیتا بیس OGI-TS شده است.



شکل۵-۴: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی MLP برای چهار زبان

درشکل ۵-۴ یک‌بار مقایسه ای بین چهار زبان انگلیسی، فرانسوی، فارسی و آلمانی و یک‌بار دیگر بین چهار زبان انگلیسی، فرانسوی، ژاپنی و اسپانیایی انجام شده است. مقایسه نشان می‌دهد بهترین نتایج به ترتیب با ۷۱ و ۷۶ درصد، به ویژگی MFCC تعلق دارد و طبق روال اشکال قبل روش تبدیل فوریه‌ی بسل بالاتر از روش LPC جواب گرفته است.

در شکل ۵-۵ نتیجه‌ی مقایسه‌ی بین شش زبان با نمونه‌های ۱۰ ثانیه‌ای با استفاده از شبکه‌ی عصبی MLP شده است.



شکل۵-۵: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی MLP برای شش زبان

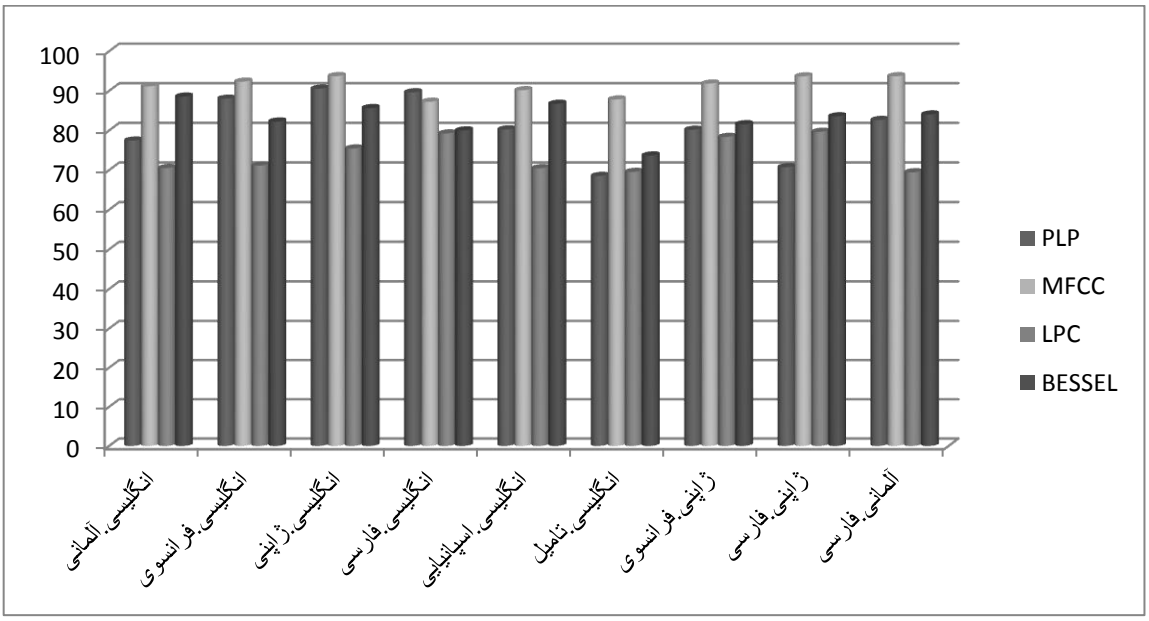
همچنین در جدول ۵-۷ نتایج چند زبانه برای ویژگی PLP به صورت ماتریسی بیان شده است.

جدول ۵-۷- دقت درستی مقایسه شش زبانه برای نمونه های ۱۰ ثانیه‌ای ویژگی PLP

	انگلیسی	فرانسوی	ژاپنی	فارسی	آلمانی	اسپانیایی
انگلیسی	۴۸/۶۱	۷/۸	۶/۱۷	۸/۴۵	۱۵/۶	۱۳/۵۵
فرانسوی	۸/۱	۴۸/۲۳	۹/۱۶	۱۱/۴۷	۱۲/۱۵	۱۰/۸۹
ژاپنی	۵/۴	۶/۷	۵۸/۱۲	۱۱/۴	۸/۷	۹/۶۸
فارسی	۱۳/۷	۷/۶	۹/۹۴	۵۶/۵۲	۷/۲۳	۵/۰۱
آلمانی	۱۱/۸	۷/۶۱	۱۳/۱۹	۱۰/۸۱	۵۲/۵۴	۴/۰۵
اسپانیایی	۱۵/۰۲	۸/۹۱	۱۱/۹۸	۷/۴۶	۱۰/۸۷	۴۵/۷۶

در جدول بالا برای نمونه‌های ۱۰ ثانیه ای مقایسه‌ای بین ۶ زبان انجام شد. درصد درستی هر زبان را به صورت مجزا از شبکه‌ی عصبی بدست آورده و وارد کردیم. برای نمونه در سطر آخر جدول، ۴۵٪ نمونه‌ها در زبان اسپانیایی به درستی تشخیص داده شدند و ۵۵٪ بقیه‌ی داده‌ها اشتباه تشخیص داده شدند که این مقادیر اشتباه در بین ۵ زبان دیگر تقسیم شدند.

در ادامه به ارائه‌ی نتایج حاصل از پیاده‌سازی روش SDC به همراه ویژگی‌های MFCC، LPC، PLP و تبدیل فوریه‌ی بسل می‌پردازیم. این نتایج را برای نمونه‌های ۱۰ ثانیه‌ای در شکل زیر می‌توان مشاهده نمود:



شکل۵-۶: نتایج حاصل از ضریب SDC به همراه ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی MLP

در این قسمت نیز نتایج نشان می‌دهد که استفاده از ضرایب SDC تاثیر زیادی در بهبود دقت سیستم خواهد داشت، چنانکه در مواردی تا ۱۵ درصد نیز در بهبود دقت سیستم به ما کمک می‌کند.

۵-۴-۲-شبکه RBF

در شبکه‌ی RBF مورد استفاده‌ی ما پارامترهای مختلفی از جمله تعداد نرون‌های مخفی، تعداد دوره‌ی آموزش، توابع فعالسازی لایه‌ی خروجی، نرخ آموزش اولیه‌ی وزن‌های خروجی و غیره در دقت تشخیص زبان، موثر می‌باشند. از این‌رو باید به‌دنبال انتخاب مناسب‌ترین پارمترهایی که بهترین درصد تشخیص زبان را به ما می‌دهد باشیم. در ابتدا با انتخاب ۲۵۰ دوره‌ی آموزش و تخصیص ۲۵٪ از داده‌ها به عنوان مجموعه آموزش و ۷۵٪ از داده‌ها به عنوان مجموعه‌ی آزمایش شروع می‌کنیم. سپس به سراغ تعیین تعداد نرون‌های لایه‌ی مخفی می‌رویم. جدول ۵-۸ تاثیر تعداد نرون‌های لایه‌ی میانی را بر روی دو زبان انگلیسی و ژاپنی برای ویژگی MFCC نشان می‌دهد.

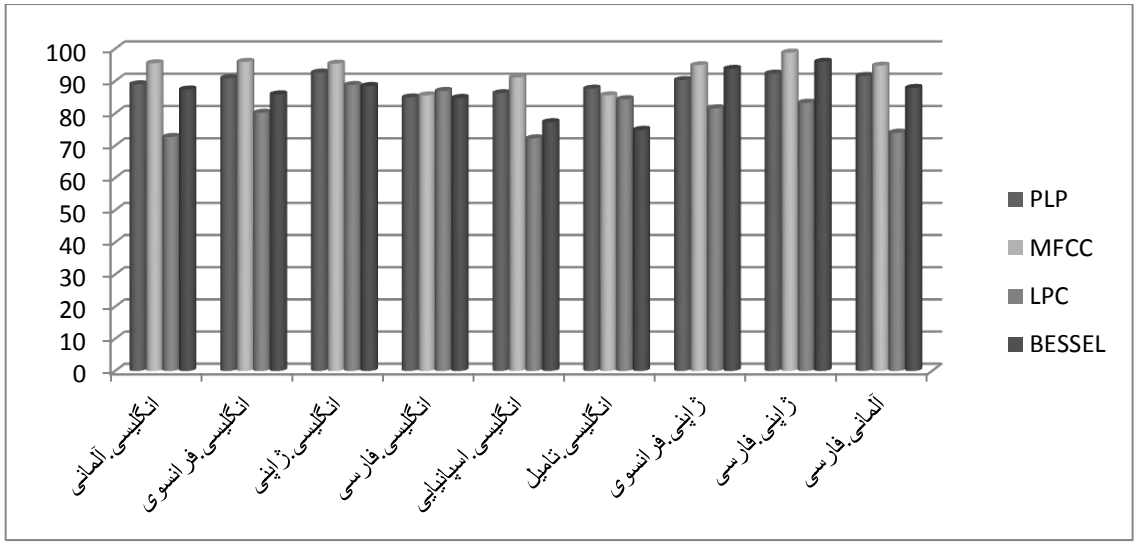
جدول ۵-۸-تاثیر تعداد نرون‌های لایه‌ی میانی شبکه‌ی RBF در دقت تشخیص زبان

تعداد نرون	۲۰	۲۵	۳۰	۳۵	٤۰	٤۵
درصدصحت	۹۳/۶۵	۹۴/۲۸	۹۴/۴۲	۹۴/۹۰	۹۵/۳۹	۹۴/۴۲

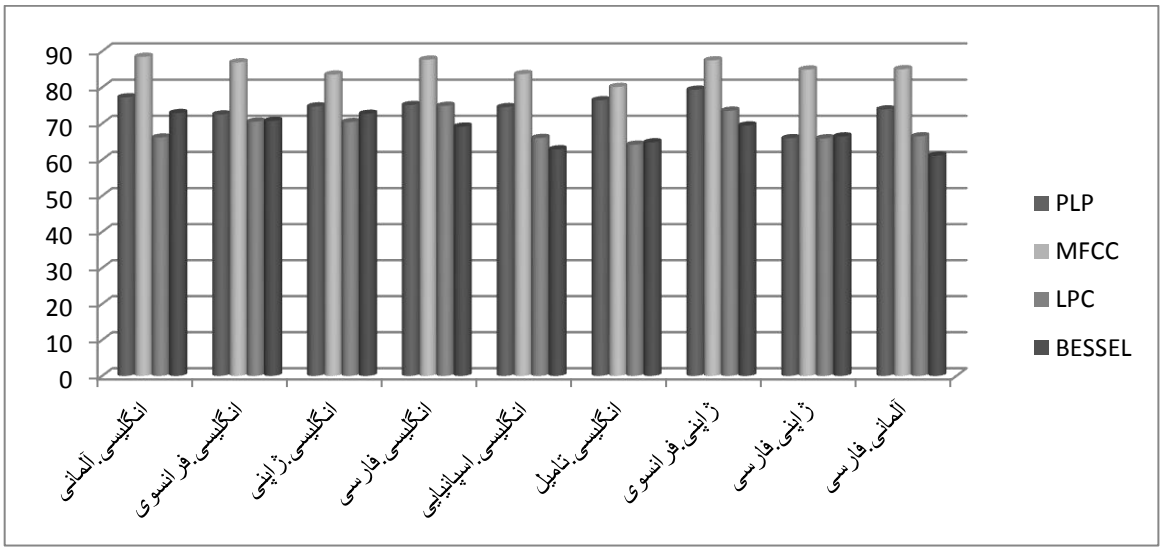
با توجه به جدول بالا می‌توان دریافت که با ۴۰ نرون در لایه‌ی مخفی بهترین درصد درستی بدست می‌آید. شایان ذکر است که تعیین بقیه‌ی پارامترهای موثر در بهبود دقت تشخیص در همان آزمایش مورد بررسی قرار گرفته و بهترین نتیجه در آن شکل یا جدول مورد نظر آورده می‌شود. یعنی مثلا در آزمایشی ممکن است تابع فعالسازی سیگموید بهتر از تابع خطی عمل کند و در آزمایشی دیگر، تابع

فعالسازی خطی بهتر عمل کند. به بیانی دیگر بقیه‌ی پارامترها ممکن است در هر آزمایش تغییر کند و نمی‌توان معیار مشخصی را برای مقدار پارامترها در همه‌ی آزمایشات در نظر گرفت.

بعد از بررسی پارامترهای موثر، به سراغ مقایسه‌ی زبان‌ها می‌رویم. نخست تاثیر روش‌های استخراج ویژگی در دقت تشخیص زبان را با استفاده از شبکه‌ی RBF مقایسه می‌نماییم.

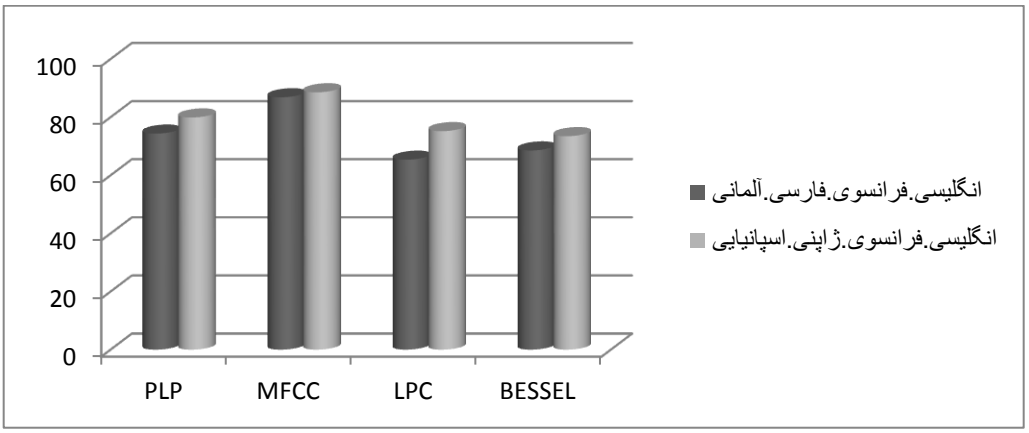


شکل۵-۷: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی RBF برای نمونه‌های صوتی ۱۰ ثانیه ای

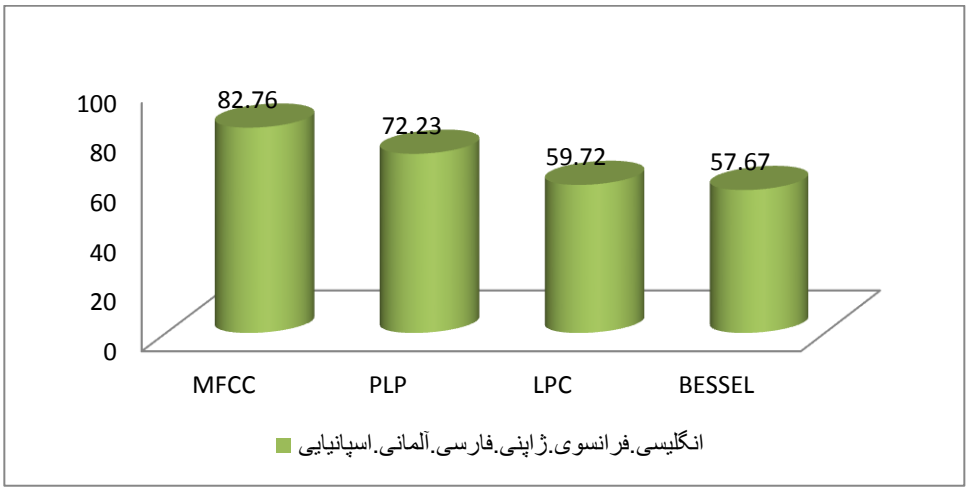


شکل۵-۸: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی RBF برای نمونه‌های صوتی ۵ ثانیه ای

با مقایسه‌ی دو شکل ۵-۷ و ۵-۱ ،همچنین ۵-۸ و ۵-۲ در می‌یابیم که جز در چند مورد، در بقیه‌ی آزمایش‌ها هنگامی که از شبکه‌ی RBF استفاده کردیم به نتیجه‌ی بهتری نسبت به حالتی که از شبکه‌ی MLP استفاده کردیم، رسیده‌ایم. اشکال ۵-۹ و ۵-۱۰ استفاده از شبکه‌ی RBF را برای نمونه‌های صوتی ۱۰ ثانیه‌ای چهار و شش زبانه نشان می‌دهد.



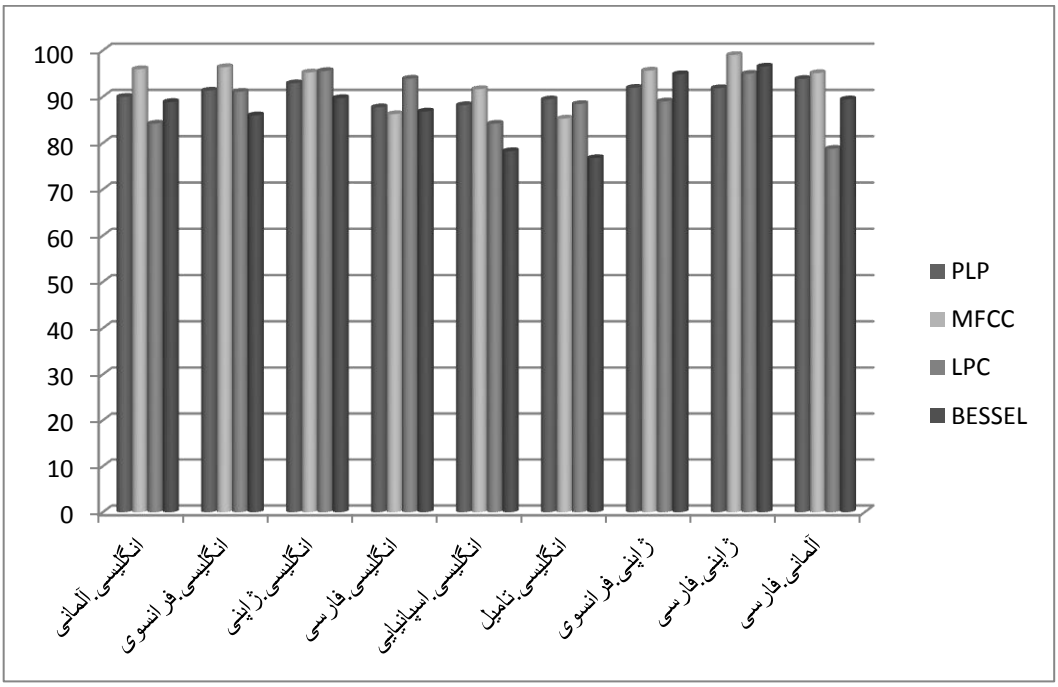
شکل۵-۹: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی RBF برای نمونه‌های صوتی ۱۰ ثانیه ای



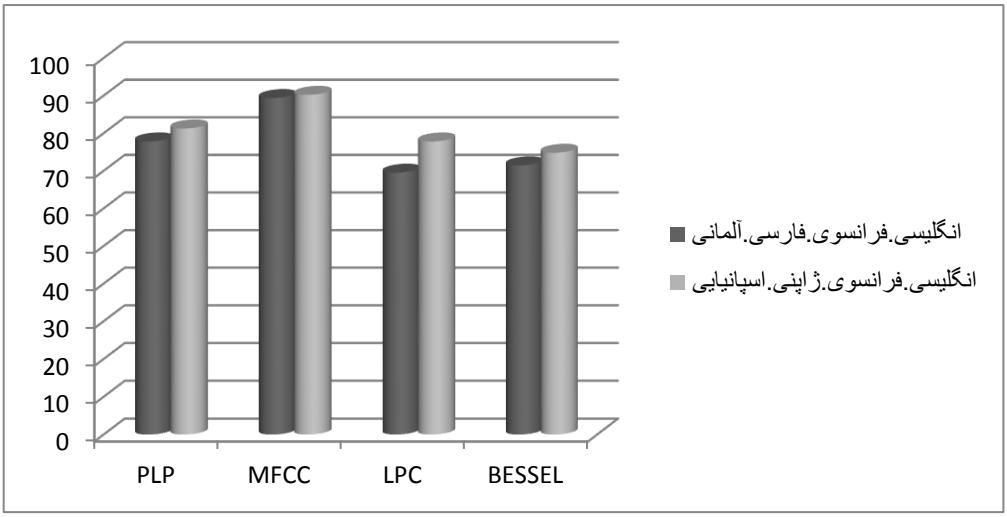
شکل۵-۱۰: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی RBF برای نمونه‌های صوتی ۱۰ ثانیه‌ای

۵-۴-۳-شبکه WRBF

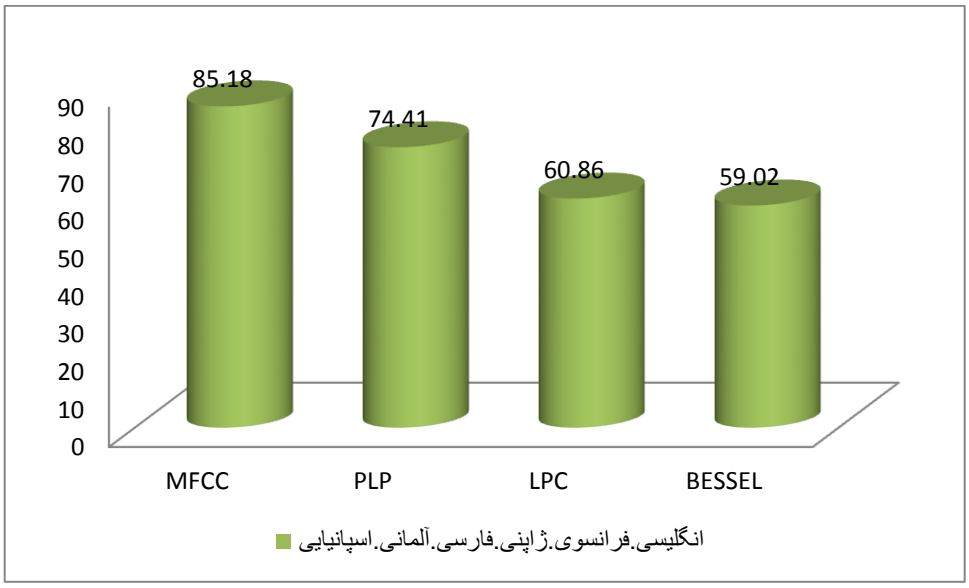
دقت شبکه‌ی WRBF همانند شبکه‌ی RBF استفاده شده در بخش قبل نیز تحت تاثیر پارامترهای مختلفی از جمله تعداد نرون‌های مخفی، تعداد دوره‌ی آموزش، توابع فعالسازی لایه‌ی خروجی، نرخ آموزش اولیه‌ی وزن‌های خروجی و ... می‌باشد. مطابق با شبکه‌ی RBF، این شبکه نیز با تعداد ۴۰ نرون در لایه‌ی مخفی خود بهترین نتیجه را می‌دهد. انتظار می‌رود این شبکه بهتر از شبکه‌های MLP و RBF استفاده شده در قسمت‌های قبل عمل نماید.



شکل۵-۱۱: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی WRBF برای نمونه‌های صوتی ۱۰ ثانیه ای



شکل۵-۱۲: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی WRBF برای نمونه‌های صوتی ۱۰ ثانیه ای

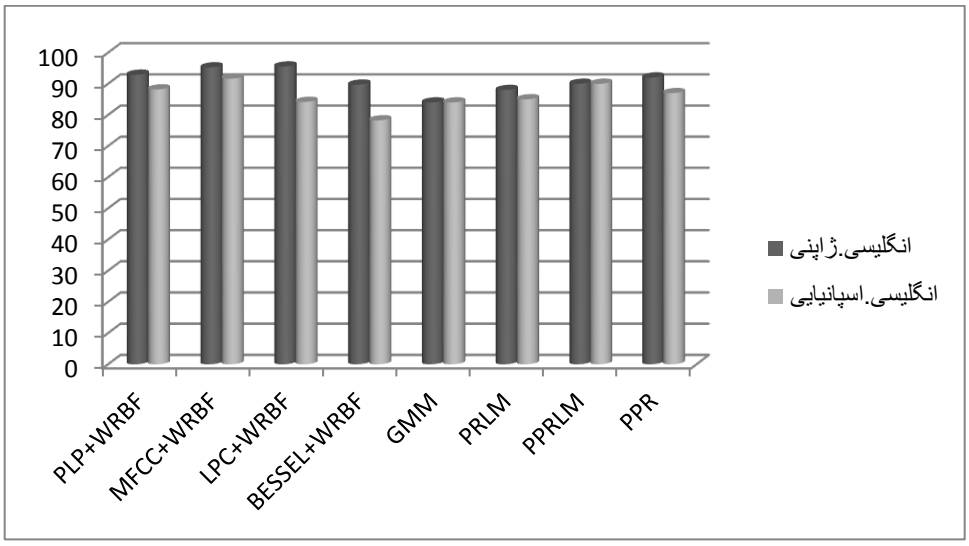


شکل۵-۱۳: مقایسه‌ی نتایج حاصل از ضرایب MFCC و PLP و LPC و BESSEL با استفاده از شبکه عصبی WRBF برای نمونه‌های صوتی ۱۰ ثانیه‌ای

همانطور که در شکل‌های۵-۱۱ تا ۵-۱۳ می‌بینید همه‌ی نتایج بدست آمده، نسبت به زمانی که از شبکه‌ی MLP استفاده شده است درصد صحت بیشتری دارد و همینطور در اکثر موارد، نتایج بدست آمده نسبت به حالتی که از شبکه‌ی RBF استفاده شده است بهتر است. همچنین مشاهده می‌شود هر چه تعداد زبان‌ها بیشتر می‌شود، بهبود دقت نیز بیشتر خودش را نشان می‌دهد.

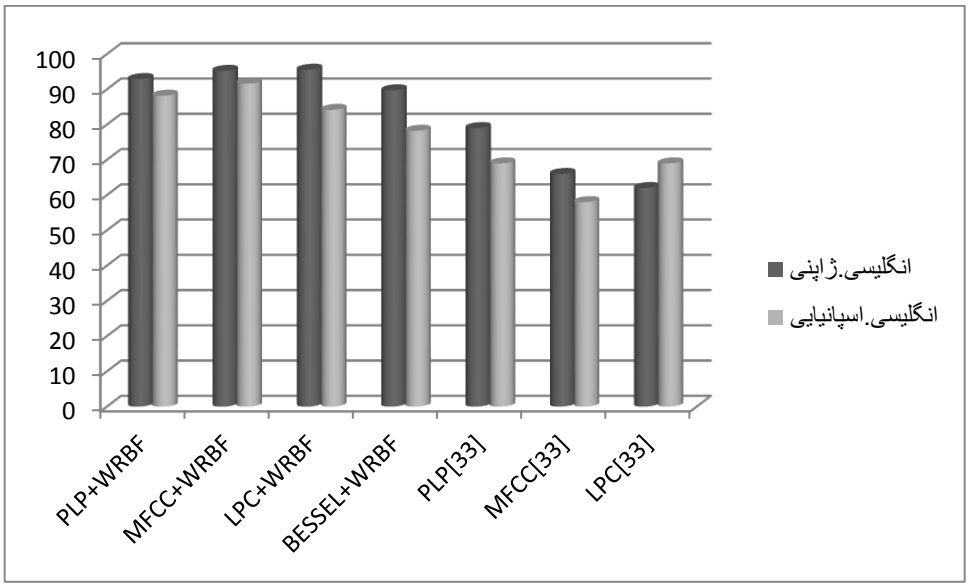
۵-۵-مقایسه

نتایج بدست آمده از این پایان‌نامه در بخش‌های پیشین به تفصیل نمایش داده شد و مورد بحث قرار گرفت. حال در قسمت پایانی این فصل این نتایج را با نتایج بدست آمده توسط مرجع[۱۰] مقایسه خواهیم کرد. دیتابیس مورد استفاده در این مرجع همان OGI-TS می‌باشد. نتایج حاصل از چهار روش مورد آزمایش در مقاله‌ی مورد نظر برای نمونه‌های ۱۰ ثانیه‌ای به همراه نتایج بدست آمده از شبکه‌ی WRBF این پایان‌نامه در شکل ۵-۱۴ آمده است.



شکل۵-۱۴: مقایسه‌ی نتایج این پایان‌نامه با روش مرجع[۱۰](زیسمن)

با توجه به شکل بالا اگر بخواهیم این نمودار را به دو گروه مقاله‌ی آقای زیسمن و نتایج بدست آمده در این پایان‌نامه تقسیم کنیم، خواهیم دید که در مقاله‌ی آقای زیسمن بهترین نتیجه مربوط به روش PPR، با ۹۲ درصد و بعد از آن PPRLM با ۹۰ درصد می‌باشد. این درحالی است‌که روش ارائه شده در این پایان‌نامه که استخراج ویژگی توسط چهار روش و سپس استفاده از شبکه‌ی عصبی WRBF می‌باشد دقت بهتری نسبت به روش آقای زیسمن دارد. همانطور که در نمودار بالا می‌توان دید استخراج ویژگی MFCC و LPC با ۹۵ درصد وسپس روش PLP با ۹۲ درصد و در آخر روش تبدیل فوریه‌ی بسل با ۸۹ درصد بهترین نتیجه را گرفتند.



شکل۵-۱۵: مقایسه‌ی نتایج این پایان‌نامه با روش مرجع[۳۳]

با توجه به شکل می‌توان مشاهده کرد که هرکدام از نتایج بدست آمده در این پایان‌نامه بهبود قابل توجهی نسبت به نتایج حاصل از روش مرجع[۳۳] دارد. مشاهده می‌شود که مثلا در روش MFCC ٪۳۷ افزایش دقت داشتیم. یکی از دلایلی که این ویژگی‌ها دقت بیشتری دارند این است که در روش ارائه شده در این پایان‌نامه از شبکه‌ی عصبی WRBF استفاده شده که نسبت به MLP به طور محسوسی دقت بالاتری را ارئه می‌کند.

فصل ششم

جمع‌بندی و پیشنهادات

۶- جمع‌بندی و پیشنهادات

۶-۱-جمع‌بندی

هدف از انجام این پروژه ارائه تکنیک‌هایی جهت بهبود سیستم‌های تشخیص زبان گفتاری مبتنی بر اطلاعات آکوستیکی بود. مزیت این دسته از سیستم‌ها تعمیم آن‌ها به تعداد زبان‌های بیشتر است. در این راستا یک سری روش‌ها ارائه شد که عبارتند از:

- ✓ بررسی تاثیر اطلاعات آکوستیکی در دقت تشخیص زبان
- ✓ ارائه روش استخراج ویژگی مبتنی بر تبدیل بسل
- ✓ بررسی تاثیر روش SDC بر روی ویژگی MFCC، PLP، LPC، و تبدیل فوریه‌ی بسل
- ✓ ارائه شبکه‌های عصبی مختلف همانند MLP، RBF و همینطور WRBF که برای اولین بار در زمینه‌ی تشخیص زبان استفاده شد.

نتایج ارزیابی‌های صورت گرفته با استفاده از دادگان OGI-TS نشان داد که:

- روش‌های ارائه شده در این پروژه قابلیت تعمیم به هر تعداد زبان مورد دلخواه را دارد، چنانکه در این پروژه، مقایسه بین دو، چهار و شش زبان مختلف صورت گرفته است.
- دقت تشخیص برای نمونه‌های ۱۰ ثانیه‌ای بهتر از نمونه‌های ۵ ثانیه‌ای است که نشان می‌دهد با افزایش زمان نمونه‌ها دقت سیستم بالا می‌رود.
- استفاده از روش MFCC بهترین دقت تشخیص زبان در میان دیگر روش‌ها را دارد.
- استفاده از روش SDC به تشخیص بهتر سیستم‌ها کمک می‌کند و دقت سیستم را به طور محسوسی بالا می‌برد.

- روش تبدیل فوریه‌ی بسل نتیجه‌ای بالاتر از روش LPC دارد و نزدیکی دقت این روش به روش‌های MFCC و PLP نشان از این دارد که می‌توان بسل را به عنوان روش استخراج ویژگی قدرتمند در نظر گرفت.
- استفاده از روش ارائه شده برای کلاسه بندی که همان شبکه‌ی RBF بود دقت تشخیص زبان را تا حدود زیادی بالا می‌برد و در برخی موارد، برای تشخیص دو زبانه این دقت به ۹۸٪ می‌رسد. این نتیجه نشان می‌دهد که از شبکه‌ی عصبی RBF می‌توان به عنوان کلاسه بند قدرتمندی در زمینه‌ی تشخیص زبان نام برد.
- ارائه‌ی روش جدید دیگری برای کلاسه‌بندی یعنی WRBF مجددا دقت تشخیص زبان را نسبت به کلاسه‌بند RBF تقویت کرد.

۶-۲-پیشنهادات

در این پروژه از ویژگی‌هایی استفاده شده است که مربوط به ویژگی‌های سطح پایین صوت می‌باشد. می‌توان از سایر ویژگی‌ها که در سطوح دیگر هستند نظیر خواص عروضی و بدیعی، یا از اطلاعات واج‌آرایی صوت ویا در سطوح بالاتر ریخت‌شناسی لغوی که به ساختار درونی کلمات مرتبط می‌شود و قواعد صرف و نحو استفاده کرد.همچنین می‌توان از ترکیبی از دو یا چند نوع ویژگی متفاوت جهت بهبود دقت سیستم استفاده کرد. بعد از عمل استخراج ویژگی نیز می‌توان جهت بهبود عملکرد سیستم و یا افزایش سرعت سیستم کارهایی را صورت داد. از جمله کارهایی که می‌توان در جهت افزایش سرعت سیستم انجام داد روش کاهش بعد PCA است که باید تاثیر این روش بر روی ویژگی‌های مختلف بررسی شود. بعد از استخراج ویژگی و پس‌پردازش‌هایی که بر روی آن‌ها انجام شد نوبت به شناساگرها می‌رسد. ما در این قسمت مستقیما از شبکه‌های عصبی به عنوان شناساگر استفاده کردیم و دیدیم که روش ارائه شده در این پایان‌نامه بهبود قابل توجهی در دقت عملکرد سیستم به وجود آورد. لیکن قبل از استفاده از

شبکه‌های عصبی می‌توان از شناساگرهای شناخته شده‌ای نظیر GMMو HMMو SVM ودیگر شناساگرها استفاده کرد که می‌توانند مدل‌های زبانی به ما بدهند که با استفاده از این مدل زبانی بتوان زبان سیستم را تشخیص داد. همچنین می‌توان از شبکه‌ی عصبی به عنوان طبقه بندی کننده‌ی نهایی در خروجی شناساگرمان نیز استفاده کرد. با توجه به افزایش دقتی که در این پروژه از شبکه‌های عصبی RBFو WRBF ایجاد شد انتظار می‌رود استفاده از این شبکه‌ها در طبقه بندی کننده‌ی نهایی عملکرد بهتری نسبت به سایر روش‌ها داشته باشد.

منابع

[۱]. "خانواده زبان های آسیایی " , شبکه ملی مدارس رشد ، قابل دسترس از آدرس:

http://daneshnameh.roshd.ir//mavara/mavara-index.php

[2]. T. Hazen and V. Zue, "**Recent improvements in an approach to segment-based automatic language identification**," in Proc. Int. Conf. Spoken Language Processing (ICSLP-94), 1994, pp. 1883–1886.

[3]. M. A. Zissman and K. M. Berkling, "**Automatic language identification**," Speech Communication, vol. 3 5, pp. 115-124,2001

[4]. Y. K. Muthusamy, N. Jain, and R. A. Cole, "**Perceptual benchmarks for automatic language identification**," in Proc. 1994 IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP94), 1994, vol. 1, pp.I/333–I/336.

[5]. D. Jurafsky and J. Martin, Speech and Language Processing: "**An Introduction to Natural Language**" Processing, Computational Linguistics, and Speech Recognition, 2nd ed. New Jersey: Prentice Hall, 2008.

[6]. E. Wong, "**Automatic spoken language identification utilizing acoustic and phonetic speech information**," Ph.D. dissertation, Speech and Audio Research Laboratory, Queensland Univ. Technol., 2004.

[7]. Zhirong Yang. ,"**Automatic Language Identification of Telephone Speech**". Laboratory of Computer Information Science Helsinki University of Technology

[8]. Leonardo R., Doddington G., "**Automatic language identification**", Technical report RADC-TR74-200, Air Force Rome Air Development Center, August, 1974.

[9]. House A.S., Neuborg E.P., "**Toward automatic identification of the language of an utterance. I. Preliminary methodological consideration**", Journal of the Acoustical Society of America,(JASA),Vol.62,no3,pp.708-713,Septamber,1977.

[10]. M.A.Zissman "**comparison of four approaches to automatic language identification of telephone speech**," IEEE Transaction on speech and Audio Prossesing, vol.4 , pp.31,44,1996

[11]. "Language Recognition Evauation" , available at : www.itl.nist.gov/iad/mig//test/lre

[12]. J. Farinas and F. Pellegrino, "**Automatic Rhythm Modeling For Language Identification**," Eurospeech, vol. 4,pp.2539-2542,2001

[13]. C. Y. Lin and H. C. Wang, "**Language Identification Using Pitch Contour Information**," ICASSP,pp.601-604,2005

[14]. E. Timoshenko, H, Hoge, "**Using Speech Rhythm for acoustic language identification**", interSpeech07, Antwerp, Belgium, pp. 182-185,2007

[15]. J. Rouas, "**Automatic Prosodic Variations Modeling for Language and Dialect Discrimination**", IEEE Transaction on Audio, Speech and Language Processing, pp. 1904-1911,2007.

[16]. W. M. Ng. Raymond, C. C. Leung, T. Lee, B. Ma, H. Li, "**Prosodic attribute model for spoken language identification**", ICASSP, pp. 5022-5025, 2010

[17]. P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr," **Approaches to language identification using Gaussian mixture models and shifted delta cepstral features**" , ICSLP ,pp. 89-92,2002

[18]. F. Allen, E. Ambikairajah, J, Epps, "**Warped magnitude and phase-based feature for language identification**" ICASSP, speech and signal processing, pp. 201-204,2006

[19]. Zhirong Yang., "**Automatic Language Identification of Telephone Speech**" Laboratory of Computer and University of Technology

[20]. R. Tong, M. Bin, D. Zhu, H. Li, and E. S. Chng, "**Integration acoustic, prosodic and phonotactic features for spoken language identification**," ICASSP, pp. 1205-1208,2006.

[21]. Magnus Rosell., "**An Introduction to Front-End Processing and Acoustic Features for Automatic Speech Recognition**". Term paper in Swedish National Graduate School of language Technology. January 17, 2006

[22]. Bo Yin; Ambikairajah, E. Fang Chen. "**Combining Cepstral and Prosodic Features In Language Identification**" Pattern Recognition, 2006. ICPR 2006 18th International Conference on Volume 4, 0-0 0 Page(s):254-257 Digital Object Identifier 10.1109/ICPR.2006.381

[23]. Yu-Min Zeng. , "**ROBUST GMM GENDER CLASSIFICATION USING PITCH AND RASTA-PLP PARAMETERS OF SPEECH**",. Proceedings of the fifth international conference on machine learning and cybernetics,Dalian,13-16 August 2006

[24]. D. Farris, C. White, and S. Khudanpur, "**sample selection for Automatic language identification**, " ICASSP,,pp. 4225-4228,2008

[25]. W, M. Campbell, E. Singer, P. A. Torres and D. A. Reynolds, "**Language Recognition with support Vector Machines,** " Proc, Odyssey, The speaker and Language Recognition Workshop, pp 41-44,2004

[26]. W.Q. Zhang, L. He, Y. Deng, J. Liu, M. T. Johnson, "**Time-Frequency Cepstral Features and Heteroscedastic Liniear Discriminant Analysis for Language Recognition**,", IEEE Transactions on Audio, Speech, and Language Processing,2010

[27]. S. B. Davis and P. Mermelstein. "**Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences**". IEEE Trans. Acoust., Speech, Signal Processing, 28(4):357{366, 1980.

[28]. S. A. Hosseini, M. M. Homayounpour, "**Using Probabilistic Characteristic Vector Based on Both Phonetic and Prosodic Features for Language Identification**" , IST , pp.,2010.

[29]. A. Ziaei, S.M. Ahadi, and H. Yeganeh, "**Enhanced spectral features for spoken language identification**," International Conference on Signal Processing Proceedings,(ICSP), pp. 1495-

[۳۰]. چ مرتضی نیا، " بازشناسی اتوماتیک برخی از زبان‌های رایج در ایران به کمک روش‌های آماری" دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، شهریور ۱۳۷۸

[31]. S. A. Hosseini, M. M. Homayounpour, "**Improvement of Language Identification Performance Using Aggregated Phone Recognizer**", USIPCO,pp. 1770-1773,2009

[32]. S. A. Hosseini, M. M. Homayounpour, "**Improvement of Language Identification Performance Using Generalized Phone Recognizer**", CSICC,pp. 596-600,2009

[۳۳]. پ. مهارلویی."طراحی و پیاده سازی سیستم شناسایی زبان گفتاری به صورت خودکار".پایان نامه کارشناسی ارشد. دانشکده برق و رباتیک دانشگاه صنعتی شاهرود.زمستان ۸۹.

[34]. P.C. Loizou," **Speech Enhancement: Theory and Practice**", CRC Press,Boca Raton, FL, 2007.section 3, p 46-490

[35]. Borden, G., Harris, K., and Raphael, L., "**Speech science Primier**", 3rd ed., Baltimore, MID: Williams and Wilkins

[36]. Rosenberg, A., "**Effect of glottal pulse shape on the quality of natural vowels**", J. Acoust. Soc. Am., 49(2), 583-588 (1971).

[37]. Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu." **Language Identification a Tutorial**". IEEE CIRCUITS AND SYSTEMS MAGAZINE. 2011 IEEE

[38]. S. Davis and P. Mermelstein, "**Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences**, "IEEE Trans. Acoustics, Speech Signal Processing, vol. 28, no. 4,pp. 357–366, 1980.

[39]. T. Schultz, I. Rogina, and A. Waibel, "**LVCSR-based language identification**," in Proc. 1996 IEEE Int. Conf. Acoustics, Speech, and Signal Processing(ICASSP-96), 1996, vol. 2, pp. 781–784.

[40]. J. Laver, "**Principles of Phonetics. Cambridge**", U.K.: Cambridge Univ. Press, 1994.

[41]. L. Rabiner and B. "**Juang, Fundamentals of Speech Recognition**". NewJersey: Prentice Hall, 1993.

[42]. B. Bielefeld, "**Language identification using shifted delta cepstrum**,"in Proc. 14th Annual Speech Research Symp., 1994.

[43]. S. Chatterjee, W. B. Klejin, "**Auditory model based modified MFCC features**" , ICASSP, PP 4590-4593, 2010

[44]. H. Hermansky, "**Perceptual linear predictive (PLP) analysis of speech**," J. Acoust. Soc. Amer., vol. 87, pp. 1738–1752, 1990.

[45]. W. Q. Zhang, L. He, Y. Deng, J. Liu, M. T. Johnson, "**Time-Frequency Cepstral Features and Heteroscedastic Linear Discriminant Analysis for Language Recognition**," ,IEEE Transactions on Audio, Speech, and Language Processing, 2010

[46]. T. Schultz and K. Kirchhoff, Multilingual "**Speech Processing**". NewYork: Academic, 2006.

[47]. M. Yip, Tone. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[48]. T. Nazzi, J. Bertoncini, and J. Mehler, "**Language Discriminantion by Newborns: Toward an Understanding of the Role of Rhythm**," Journal of Experimental Psychology: Human Perception and Performance, vol. 24, PP. 756-766, 1998

[49]. C. Y. Lin and H. C. Wang, "**Language Identification Using Pith Contour Information**," ICASSP, PP. 601-604,2005.

[50]. E. Timoshenko, H. Hoge, "**Using Speech Rhythm for acoustic language identification**", interspeech07, Antwerp, Belgium, pp. 182-185, 2007.

[۵۱]. ر. مقدم کیا، "مقایسه مشخصه های زبر زنجیری در زبان‌های فارسی و ژاپنی". پژوهش زبان‌های خارجی، سال دوازدهم، شماره ۳۳، صص ۱۸۳-۱۶۷،پاییز ۸۵

[52]. V. Dellwo, A. Fourcin, and E. Abberton, "**Rhythmical classification of language based on voice parameters**," 16th ICPhS, pp. 1129-1132,2007.

[53]. L. Bauer, "**Introducing Linguistic Morphology**". Georgetown Univ. Press, 2003.

[54]. A. Carnie, "**Syntax: A Generative Introduction, 2nd ed**". New York:Wiley-Blackwell, 2006.

[55]. H. Khosravi. "**A Novel Structure for Radial Basis Function Networks—WRBF**". Neural Process Lett DOI 10.1007/s11063-011-9210-0

[56]. Bashkirov, O.A., E.M. Braverman, and I.B. Muchnik," **Potential function algorithms networks for pattern recognition learning machines**". Automatic Remote Control, 1964: p. 629–631.

[57]. Hojjatoleslami, A., L. Sardo, and J. Kittler, "**An RBF based classifier**" for the detection of microcalcifications in mammograms with outlier rejection capability in International Conference on Neural Networks. 1997. p. 379-1384.
[58]. Chen, S., et al., "**Symmetric RBF Classifier for Nonlinear Detection in Multiple-Antenna-Aided Systems**". IEEE Transactions on Neural Networks, 2008. 19(5): p. 737 -745.
[59]. . Meng, K., et al.," **A Self-Adaptive RBF Neural Network Classifier for Transformer Fault Analysis**" IEEE Transactions on Power Systems, 2010. 25(3): p. 1350 - 1360.
[60]. Wang, D., et al., "**Protein Sequences Classification Using Radial Basis Function (RBF) Neural Networks**", in 9'th International Conference on Neural Information Processing.2002: Singapore. p. 764-769.
[61]. Wang, D., et al., "**Protein Sequences Classification Using Radial Basis Function (RBF) Neural Networks**", in 9'th International Conference on Neural Information Processing.2002: Singapore. p. 764-769.
[62]. King, J.L. and L. Reznik," **Topology Selection for Signal Change Detection in Sensor Networks: RBF vs MLP**". International Joint Conference on Neural Networks, 2006: p. 2529 - 2535.
[63]. Polat, G. and H. Altun, "**Evalutation of Performance of KNN, MLP and RBF Classifiers in Emotion Detection Problem**". IEEE 15th Conf. on Signal Processing and Communications Applications, 2007. 1: p. 1-4.
[64]. King, J.L. and L. Reznik," **Topology Selection for Signal Change Detection in Sensor Networks: RBF vs MLP**". International Joint Conference on Neural Networks, 2006: p. 2529 - 2535.
[65]. Izhikevich, E.M., "**Which model to use for cortical spiking neuron**". IEEE Trans. Neural Networks, 2004. 15(5): p. 1063-1070.
[66]. S. A. Hosseini, M. M. Homayounpour, "**Improvement of Language Identification Performance Using Aggregated Phone Recognizer**", USIPCO,pp. 1770-1773,2009