
بسم الله الرحمن الرحيم

دانشکده: مهندسی برق و رباتیک

گروه: الکترونیک

تشخیص لهجه‌های مختلف فارسی بر اساس شکل موج گفتار

دانشجو: مجتبی شریف نوقابی

استاد راهنما:

دکتر حسین مروی

پایان‌نامه ارشد جهت اخذ درجه کارشناسی ارشد

ماه و سال انتشار: بهمن ۱۳۹۲

شماره : ۱۱۲۴ /ا.ت.ب
تاریخ : ۹۲/۱۱/۰۸
ویرایش : ——

بسمه تعالی

مدیریت تحصیلات تکمیلی
فرم شماره (۶)

فرم صورتجلسه دفاع پایان نامه تحصیلی دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (عج) جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای :

مجتبی شریف نوقابی رشته : **برق** گرایش : **الکترونیک (سیستم)**

تحت عنوان : **تشخیص لهجه های مختلف فارسی بر اساس شکل موج گفتار**

که در تاریخ **۹۲/۱۱/۰۸** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح زیر است :

قبول (با درجه : عالی امتیاز ۱۹/۸۷) ☑ دفاع مجدد ☐ مردود ☐

۱ـ عالی (۲۰ ـ ۱۹)
۲ـ بسیار خوب (۱۸/۹۹ ـ ۱۸)
۳ـ خوب (۱۷/۹۹ ـ ۱۶)
۴ـ قابل قبول (۱۵/۹۹ ـ ۱۴)
۵- نمره کمتر از ۱۴ غیر قابل قبول

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضـــاء
۱ـ استاد راهنما	[illegible]	استادیار	
۲ـ استاد مشاور	——	—	
۳ـ نماینده شورای تحصیلات تکمیلی	مهدی بانوادر	استادیار	
۴ـ استاد ممتحن	امیر نوروزی	استادیار	
۵ ـ استاد ممتحن	حسین حسنی	استادیار	

رئیس دانشکده :

تقدیم به

پدر و مادر مهربان

و

همسر عزیزم

که همواره مشوق و پشتیبان من در تمام عرصه ها بوده اند

با تشکر و سپاس از استاد محترم

جناب آقای دکتر مروی

و

تمام اساتید و دانشجویانی

که با کمک ها و راهنمایی های بی دریغشان در این مسیر من را یاری کردند

تعهد نامه

اینجانب **مجتبی شریف نوقابی** دانشجوی دوره کارشناسی ارشد رشته **مهندسی برق – الکترونیک**

دانشکده **برق و رباتیک** دانشگاه صنعتی شاهرود نویسنده پایان‌نامه

تشخیص لهجه‌های مختلف فارسی بر اساس شکل موج گفتار

تحت راهنمایی آقای **دکتر حسین مروی** متعهد می شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام‌شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده‌شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده‌شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده:

ترکیب طبقه بندها استفاده‌شده است. برای انجام آزمایش‌های مختلف،داده‌های این پایان‌نامه یک سیگنال گفتار علاوه بر متن گفته شده حاوی اطلاعات زیادی از جمله سن و جنسیت احساسات و استرس، لهجه و گویش و سلامتی گوینده می‌باشد. یکی از مواردی که ممکن است باعث کاهش بازدهی یک سیستم تشخیص گفتار گردد تغییر لهجه آن است. به طوری که اگر یک سیستم با یک لهجه خاص آموزش دیده باشد و سپس با لهجه‌ای غیر از لهجه‌ای که با آن آموزش دیده است آزمایش شود، شاهد کاهش نسبتاً زیادی در بازدهی سیستم تشخیص گفتار خواهیم بود.

با روشن شدن اهمیت مسئله تشخیص لهجه‌ها، اهمیت تدوین این پایان‌نامه نیز روشن می‌شود. در این تحقیق تعدادی ویژگی جدید مانند فرکانس مرکزی طیفی و دامنه مرکزی طیفی جهت یک سیستم تشخیص لهجه زبان فارسی پیشنهاد شده‌اند تا در کنار سایر ویژگی-هایی که در تحقیقات گذشته استفاده‌شده‌اند از سیگنال گفتار لهجه دار استخراج شده و موجب افزایش کارآیی این سیستم گردند. علاوه بر این تعدادی ویژگی مقاوم به نویز به منظور تشخیص لهجه‌ها در محیط نویزی نیز معرفی گردیده‌اند. در مرحله طبقه‌بندی علاوه بر طبقه-بندهای متداول از شبکه توابع بنیادی شعاعی نیز استفاده‌شده است و یک پیشنهاد به منظور بهبود عملکرد طبقه بند ماشین بردار پشتیبان ارائه گردیده است. به عنوان آخرین روش پیشنهادی در مرحله طبقه‌بندی از روش از پایگاه داده FARSDAT انتخاب‌شده‌اند. نتایج آزمایش‌ها، نشان‌دهنده بهبود عملکرد سیستم تشخیص لهجه‌های زبان فارسی در اغلب روش-های پیشنهادی است.

کلمات کلیدی

لهجه فارسی، تشخیص گفتار لهجه دار، تشخیص مقاوم گفتار، فرکانس مرکزی طیفی، دامنه مرکزی طیفی، ماشین بردار پشتیبان، شبکه توابع بنیادی شعاعی، ضرایب مل-کپستروم بهبودیافته، ترکیب طبقه بندها.

لیست مقالات مستخرج:

- **بازشناخت مقاوم گفتار فارسی با استفاده از ضرایب مل-کپستروم بهبودیافته و شبکه عصبی** (پذیرفته‌شده در یازدهمین کنفرانس سیستم-های هوشمند)
- **افزایش کارایی سیستم تشخیص لهجه‌های گفتار زبان فارسی با استفاده از ماشین بردار پشتیبان** (پذیرفته‌شده در دوازدهمین کنفرانس سیستم‌های هوشمند)
- **بازشناخت مقاوم لهجه‌های زبان فارسی با استفاده از ویژگی فرکانس مرکزی طیفی** (پذیرفته‌شده در نوزدهمین کنفرانس انجمن کامپیوتر ایران)

فهرست مطالب

عنوان.....صفحه

فهرست شکل‌ها

شماره و عنوان شکل .....صفحه

فهرست جدول‌ها

شماره و عنوان جدول ..... صفحه

فهرست علائم و اختصارات

MFCC: Mel-Frequency Cepstral Coefficients

SDC: Shifted Delta Cepstral coefficients

LPC: Linear Prediction Cepstral

PLP: Perceptual Linear Prediction

MLP: Multi Layer Perceptron

KNN: K-Nearest Neighbor

RBF: Radial Basis Function

PNN: Probabilistic Neural Network

HMM: Hidden Markov Model

GMM: Gaussian Mixture Model

SVM: Support Vector Machine

SCM: Spectral Centroid Magnitude

SCF: Spectral Centroid Frequency

PRLM: Phone Recognition followed by Language Modeling

AMFCC: Autocorrelation Mel-Frequency Cepstral Coefficients

GMFCC: Gaussian Mel-Frequency Cepstral Coefficients

AGMFCC: Autocorrelation Gaussian Mel-Frequency Cepstral Coefficients

SNR: Signal-to-Noise Ratio

فصل اول:

مقدمه

۱- مقدمه

بازشناسی خودکار گفتار یکی از مسائل و مشکلاتی است که از سال‌ها قبل مطرح بوده است و با گذشت زمان روش‌های متنوعی برای انجام این کار و بهبود نتایج حاصله پیشنهاد شده است.

به طور کلی از یک سیگنال گفتار در دسترس معمولاً برای سه منظور استفاده می‌شود:

۱- تشخیص گفتار[1] ۲- تشخیص زبان[2] ۳- تشخیص گوینده[3]

یک سیگنال گفتار علاوه بر متن گفته‌شده حاوی اطلاعات زیادی از جمله سن و جنسیت گوینده، احساسات و استرس، لهجه و گویش و سلامتی گوینده می‌باشد. در میان این ویژگی‌ها دو ویژگی جنسیت و لهجه به ترتیب بیشترین تأثیر را در کاهش بازدهی سیستم‌های تشخیص گفتار دارد.

اگر یک سیستم تشخیص گفتار با یک لهجه و یا گویش خاصی آموزش دیده باشد و سپس با لهجه و یا گویشی غیر از آنچه که به آن شناسانده شده است مورد امتحان قرار گیرد مشاهده خواهیم کرد که دقت سیستم به طور چشم‌گیری کاهش می‌یابد. بنابراین برای داشتن یک سیستم کارا و مقاوم در برابر این‌گونه تغییرات باید ابتدا لهجه یا گویش گوینده را تشخیص دهیم.

برای تشخیص جنسیت از مدل‌های مخصوص به خود این مسئله استفاده می‌شود و ما در این تحقیق، روی مسئله تشخیص لهجه‌های فارسی از روی سیگنال گفتار تمرکز خواهیم کرد و سعی خواهد شد با ارائه مدل‌های مختلف به حل این مسئله کمک کنیم.

---

[1] Speech recognition
[2] Language recognition
[3] Speaker recognition

ابتدا به بیان تعریفی از لهجه و تفاوت آن با گویش می‌پردازیم.

۱-۱ تعریف لهجه و تفاوت آن با گویش

قبل از آغاز تعاریف این بخش بهتر است اشاره‌ای داشته باشیم به دو آیه از آیات قرآن کریم که در آن‌ها به مسئله تفاوت زبان‌ها و قومیت‌ها پرداخته‌شده است.

الف) ای مردم ما شما را از یک مرد و زن آفریدیم و شما را تیره‌ها و قبیله‌ها قراردادیم تا یکدیگر را بشناسید. (سوره حجرات آیه ۱۳)

ب) از نشانه‌های او آفرینش آسمان‌ها و زمین و تفاوت زبان‌ها و رنگ‌های شماست.در این نشانه-هایی است برای اهل دانش. (سوره روم آیه ۲۲)

زبان از نظر علم زبان‌شناسی شامل تعداد محدودی قاعده‌ی آوایی، معنایی و دستوری است که همراه تعداد محدودی واژه می‌تواند بی‌نهایت جمله بسازد و این جمله‌ها از طریق دستگاه گفتار آدمی تولید می‌شوند و واسطه‌ی ارتباط میان افراد می‌گردند.

ولی زبان از نظر گویش[1] شناسی تعریف دیگری دارد. در یک محدوده‌ی سیاسی آن چیزی که ما به آن زبان می‌گوییم باید دو ویژگی داشته باشد؛ نخست این که زبان رسمی یک مملکت باشد، یعنی قدرت سیاسی داشته باشد. مانند زبان فارسی در ایران، و دوم این که نسبت به زبان‌ها و گویش‌های اطراف خود، زبان مادری دیگری داشته باشد.

گویش‌ها شاخه‌هایی از یک زبان واحد هستند، مثلاً گویش‌های: فارسی، تاتی، کردی، بلوچی، مازندرانی، گیلکی و ... گویش‌های گوناگون یک زبان ایرانی هستند.

به انواع هر گویش، لهجه[2] می‌گویند، برای مثال گویش فارسی دارای لهجه‌های تهرانی،

[1] dialect
[2] accent

اصفهانی، شیرازی، کرمانی و غیره است.گویش گیلکی دارای لهجه‌های رشتی، لاهیجانی، رودسری، آستانه‌ای و ... و گویش کردی دارای لهجه‌های مهابادی، سنندجی، کرمانشاهی، ایلامی و غیره است. گویش‌ها از نظر آوایی، واژگانی و دستوری باهم تفاوت‌های بسیاری دارند و فهم آن‌ها نیاز به آموزش دارد، ولی لهجه‌های هر گویش معمولاً ،فقط تفاوت‌های آوایی و واژگانی دارند و فهم آن‌ها نیاز به آموزش چندانی ندارد. به عنوان مثال یک نفر اصفهانی با یک تهرانی یا شیرازی به راحتی می‌تواند هم صحبت شود، ولی همین فرد اصفهانی وقتی با گویشور گیلکی یا مازندرانی و یا بلوچی روبرو می‌شود، اگر نخواهد از فارسی که برای آنان زبان میانجی به شمار می‌آید استفاده کند، دچار مشکل می‌شود.

هر لهجه دارای گونه‌های زبانی نیز هست که وابسته به شغل، تحصیل، سن و جنس گویشور است. مثلاً یک مرد اصفهانی تحصیل‌کرده با یک مرد بی‌سواد اصفهانی تفاوت لهجه دارد. و یا یک جوان اصفهانی در مقابل افراد مسن، لهجه متفاوتی دارد. شغل و جنس نیز در ایجاد گونه‌های زبانی تأثیر می‌گذارد.

زبان فارسی افزون بر لهجه‌هایی مانند تهرانی، اصفهانی، شیرازی و جز آن، دارای گونه‌ای است که به آن فارسی معیار می‌گویند. فارسی معیار دارای دو گونه ی نوشتاری و گفتاری است، این دو گونه از نظر واژگان و ساختار تفاوتی باهم ندارند و در واقع یکی هستند. تفاوت آن‌ها در این است که فارسی معیار نوشتاری را می‌نویسیم و فارسی معیار گفتاری را به همان شکل اصلی از روی نوشته می‌خوانیم، آن چه با فارسی معیار متفاوت است فارسی گفت و گویی (محاوره‌ای) است، یعنی همان فارسی که با آن حرف می‌زنیم. این نوع فارسی از نظر آوایی، واژگانی و ساخت دستوری تفاوت‌هایی با فارسی معیار گفتاری دارد.

از آنچه که در بالا عنوان شد نتیجه می‌گیریم که بین لهجه و گویش تفاوت ساختاری وجود دارد و معمولاً در یک گویش خاص تغییر در لهجه‌های آن باعث کاهش سیستم بازشناسی

گفتار می‌گردد.

۲-۱ تاریخچه پژوهش‌های صورت گرفته

اگر بخواهیم تاریخچه‌ی کلی تحقیقات صورت گرفته در مورد تشخیص لهجه‌ها را بیان کنیم می‌توان آن را همزمان با اولین تحقیقات انجام‌شده در حوزه پردازش گفتار، یعنی زمانی که اولین سیستم خودکار تشخیص گفتار در سال ۱۹۵۰ در آزمایشگاه بل ساخته شد، دانست. اما اگر بخواهیم تاریخچه‌ای از پژوهش‌هایی که صرفاً در مورد موضوع مشخص «تشخیص لهجه-ها» صورت گرفته است بیان کنیم می‌توان گفت تاکنون تحقیقات مختلفی در این زمینه انجام‌شده است که آن‌ها را به دو حوزه لهجه‌های فارسی و خارجی تقسیم‌بندی می‌کنیم. در حوزه لهجه‌های خارجی تحقیقات وسیع‌تری به ویژه روی لهجه‌های انگلیسی که از مناطق مختلف دنیا بوده‌اند انجام‌شده است. در سال ۱۹۹۶یکی از اولین و مهم‌ترین تحقیقات در این مورد صورت گرفت که به نتایج خوبی نیز منتهی شد. پژوهش‌های این حوزه در سال‌های مختلف و با اجرای روش‌های مختلف ادامه یافت که سال ۲۰۰۷ را می‌توان سال اوج این پژوهش‌ها چه از نظر کمی و چه از نظر کیفی نامید. البته پایان‌نامه‌هایی در مقاطع کارشناسی ارشد و دکترا در دانشگاه‌های مختلف نیز در سال‌های ۲۰۱۱ و ۲۰۱۲ به بررسی این موضوع پرداخته‌اند و روش‌های جدیدی را برای بهبود تشخیص لهجه‌ها پیشنهاد داده‌اند[1] و [2].

در حوزه لهجه‌های فارسی مشاهدات ما نشان می‌دهد که تاکنون سه تحقیق و پژوهش در این موضوع انجام‌شده است. یکی از این تحقیقات در سال ۱۳۸۹ هجری شمسی و دوتای دیگر در سال ۱۳۹۱ انجام پذیرفته‌اند که هر کدام با اجرای روش‌های مختلف به نوعی به بهبود این مسئله کمک نموده‌اند.

در بخش‌های ۲-۴ و ۲-۵ این پایان‌نامه به بررسی و ذکر جزئیات بیشتری از این تحقیقات و بیان روش‌های به‌کاررفته در آن‌ها خواهیم پرداخت.

۱-۳ اهداف و ساختار پایان‌نامه

با توجه به مطالب ذکرشده در صفحات قبل و با پی بردن به اهمیت تشخیص لهجه‌ها در افزایش بازدهی سیستم‌های تشخیص گفتار، بنا داریم در این پایان‌نامه بـه مطـرح کـردن ایـن موضوع و اجرای روش‌هایی برای رسیدن به آن بپردازیم.

در این تحقیق سعی شده است با الگو گرفتن از تحقیقات قبلی انجام‌شده و گاها استفاده و ترکیب روش‌های آن‌ها و همچنین پیشنهاد روش‌های جدید به سمت بهبود مسئله تشخیص لهجه‌های زبان فارسی در مراحل مختلف آن حرکت کنیم.

این پایان‌نامه در پنج فصل تدوین گردیده است. فصل اول با بیان مقدمـه و تاریخچـه‌ای درباره موضوع مورد نظر این تحقیق گذشت. در فصل دوم روش کلـی یـک سیسـتم تشـخیص لهجه بیان خواهد شد. در این فصل مراحل مختلف یک سیستم تشخیص لهجه شـامل پـیش-پردازش، استخراج ویژگی و طبقه‌بندی به طور کامل تشریح خواهند شد. لازم به ذکر است که قسمت‌های پایانی این فصل به ذکر جزئیات بیشتری از پژوهش‌های انجام‌شده قبلی اختصـاص یافته است.

در فصل سوم روش پیشنهادی این پایان‌نامه بیان خواهد شد. در این فصل در دو حـوزه استخراج ویژگی و طبقه‌بندی پیشنهادهایی ارائه گردیده است. در مرحله استخراج ویژگی چهار ویژگی فرکانس مرکزی طیفی، دامنه مرکزی طیفی، تبدیل Zak و ویژگی ضرایب مل-کپستروم بهبودیافته و روش استخراج آن‌ها معرفی می‌شوند. در مرحله طبقه‌بندی، شبکه RBF به عنوان یک طبقه بند جدید در زمینه تشخیص لهجه‌ها معرفی‌شده است و یـک پیشـنهاد بـه منظـور بهبود عملکرد طبقه بند SVM ارائه خواهد شد. در آخرین روش پیشنهادی برای مرحله طبقه-بندی، روش ترکیب طبقه بندها معرفی‌شده است. ترکیب طبقه بندها به صورت دوتایی و سه-

تایی انجام‌شده است.

در فصل چهارم با اجرای روش‌های قبلی و روش‌های پیشنهادی روی داده‌های انتخاب‌شده برای این تحقیق، آزمایش‌هایی در مورد تشخیص لهجه‌های فارسی انجام‌شده است که نتایج آن در جداول مختلفی آورده شده است و نتایج نیز با یکدیگر مقایسه شده‌اند. مقایسه میانگین نتایج به طور کلی نشان‌دهنده بهبود سیستم تشخیص لهجه‌های زبان فارسی با روش-های پیشنهادی می‌باشد.

نهایتاً در فصل پنجم با مشاهده نتایج آزمایش‌های حاصل از روش‌های پیشنهادی و روش‌های قبلی به بیان نتیجه‌گیری نهایی خواهیم پرداخت و با ارائه روش‌ها و پیشنهادهایی برای کارهای آینده این فصل به پایان خواهد رسید.

امید است با انجام این تحقیق بتوانیم چراغ کوچکی در مسیر پیشرفت علمی دانشمندان این مرز و بوم روشن نماییم.

فصل دوم:

سیستم تشخیص لهجه

و

مراحل مختلف آن

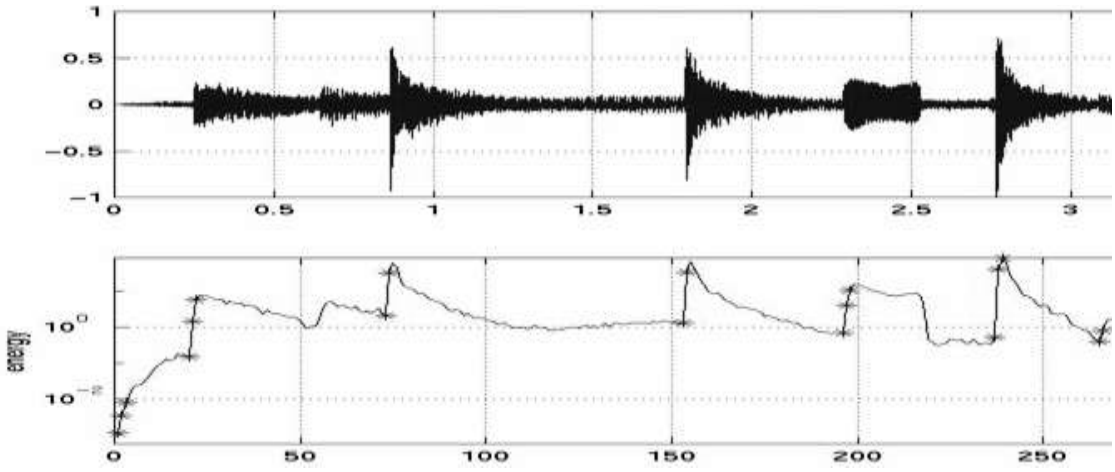
۲-سیستم تشخیص لهجه و مراحل مختلف آن

سیستم تشخیص لهجه همانند هر سیستم تشخیص گفتاری شامل سه مرحله است. این سه مرحله شامل پیش‌پردازش، استخراج ویژگی و طبقه‌بندی است. در ادامه به تفصیل در مورد هر کدام از این مراحل بحث خواهد شد.

۲-۱ پیش‌پردازش

معمولاً سیگنال‌های گفتاری که برای انجام آزمایش‌های پردازش گفتار استفاده می‌شوند شامل قسمت‌هایی هستند که وجود آن‌ها باعث کاهش بازدهی سیستم می‌شود. این قسمت‌ها می‌توانند شامل قسمت سکوت[۱] و یا صامت[۲] باشد. برای حذف این قسمت‌ها از سیگنال گفتار عملیاتی انجام می‌شود که به آن پیش‌پردازش می‌گویند. تعیین آستانه انرژی و نرخ عبور از صفر سیگنال دو فرآیندی هستند که به طور معمول برای حذف قسمت سکوت از گفتار انجام می‌شوند. همان طور که می‌دانیم انرژی قسمت گفتار یک سیگنال از انرژی قسمت‌های سکوت و صامت بیشتر است. بنابراین با محاسبه انرژی قسمت‌های مختلف سیگنال و تعیین یک آستانه می‌توانیم قسمت‌های غیر مفید را حذف کنیم. شکل (۲–۱) سیگنال گفتاری را نشان می‌دهد که انرژی قسمت‌های مختلف آن محاسبه شده است. همان طور که این شکل نشان می‌دهد انرژی قسمت‌های گفتار کاملاً از سایر قسمت‌ها بیشتر است و به راحتی با قرار دادن یک عدد به عنوان آستانه می‌توان بین آن‌ها تمایز برقرار کرد. نکته‌ی دیگری که از آن به عنوان راه‌کاری برای جدا کردن قسمت گفتار از سایر قسمت‌ها استفاده می‌شود، نرخ عبور از صفر است. می‌توان گفت قسمت سکوت به علت انرژی کم و نزدیک به صفر، تقریباً روی محور صفر است و حرکتی ندارد و بنابراین نرخ عبور از صفر آن نیز بسیار کم است. اما قسمت صامت در عین

[1] silence
[2] unvoiced



شکل (۲–۱): انرژی قسمت‌های مختلف یک سیگنال گفتار

حالی که انرژی آن کم است ولی دارای فرکانس زیادی است و حول محور صفر تحرک زیادی دارد که این خود باعث بالا رفتن نرخ عبور از صفر آن می‌گردد. نرخ عبور از صفر قسمت گفتار از قسمت سکوت بیشتر و از قسمت صامت کمتر است در نتیجه با دانستن این نکته می‌توان بخش‌های مختلف سیگنال گفتار را از هم جدا کرد.

علاوه بر دو روشی که بیان شد برای حذف قسمت صامت از روش محاسبه ضرایب خودهمبستگی[1] سیگنال نیز استفاده می‌شود. اگر ضرایب خودهمبستگی یک سیگنال گفتار را محاسبه و رسم کنیم علاوه بر اینکه پریودیک بودن سیگنال گفتار را به طور واضح‌تر نشان خواهد داد مشاهده خواهیم کرد که قسمت‌های گفتار دارای پیک‌هایی هستند درحالی‌که قسمت‌های unvoiced پیک ندارند و بنابراین می‌توانیم این دو قسمت را از هم جدا کنیم.

پس از حذف قسمت‌های سکوت و صامت، به منظور جبران دامنه در فرکانس‌های بالا یک فیلتر بالا گذر که پیش تاکید نیز نامیده می‌شود به سیگنال گفتار اعمال می‌شود[۵]. این فیلتر را با رابطه (۲-۱) نشان می‌دهند.

$$P(z) = 1 - az^{-1} \qquad (۲-۱)$$

[1] Auto correlation

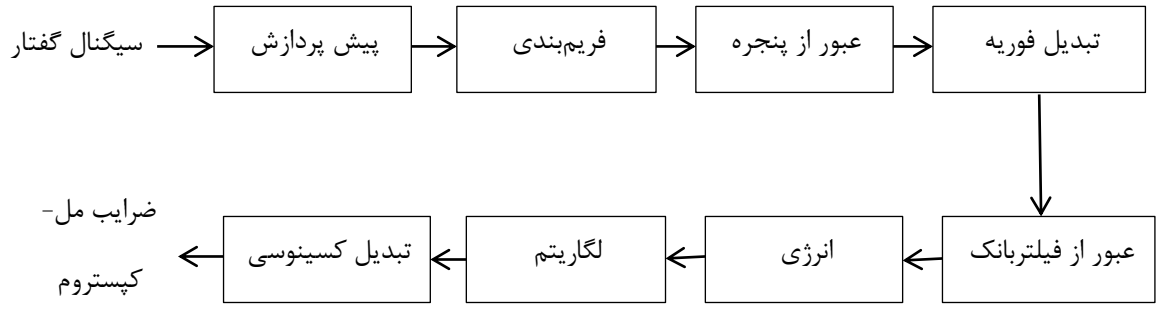
که در این رابطه a ضریبی است که معمولاً آن را بین 0.9 تا 1 در نظر می‌گیرند.

۲-۲استخراج ویژگی

مرحله استخراج ویژگی و استخراج یک ویژگی مناسب و کارآمد به منظور تشخیص گفتار و تشخیص لهجه یکی از مهم‌ترین مراحل در یک سیستم تشخیص گفتار است. با توجه به این که در این پایان‌نامه تمرکزمان روی مسئله تشخیص لهجه‌ها از روی شکل موج گفتار است باید به دنبال ویژگی باشیم که بتواند تمایز خوبی بین لهجه‌های مختلف ایجاد نماید. این ویژگی‌ها می‌توانند در دو حوزه فرکانس و زمان باشند. برخی از ویژگی‌هایی که به طور معمول در این زمینه استفاده می‌شوند در مسائل عادی تشخیص گفتار نیز استفاده می‌شوند و برخی ویژگی‌ها بیشتر مخصوص کارهایی همچون تشخیص لهجه، گویش، احساسات و سلامتی گوینده هستند. در زیر بخش‌های بعدی به معرفی روش‌های متداول استخراج ویژگی می‌پردازیم.

۲-۲-۱ ضرایب مل-کپستروم و مشتقات آن

یکی از ویژگی‌هایی که معمولاً در بیشتر سیستم‌های تشخیص خودکار گفتار (ASR) استفاده می‌شود ضرایب مل-کپستروم (MFCC) است که مربوط به ویژگی‌های حوزه فرکانس می‌باشد. نمودار محاسبه این ضرایب در شکل (۲–۲) نشان داده‌شده است.



شکل (۲–۲): نمودار بلوکی محاسبه ضرایب مل-کپستروم [3]

مرحله اول پیش‌پردازش است که در بخش ۲-۱ توضیح داده شد. در مرحله بعد سیگنال را فریم بندی می‌کنیم. دلیل اصلی این عمل، نا ایستا بودن سیگنال گفتار است که به علت کوچک شدن طول سیگنال گفتار در فریم‌ها می‌توان آن را تقریباً ایستان فرض کرد. البته این عمل به منظور کم کردن حجم داده ورودی و صرفه‌جویی در زمان و همچنین جزئی‌تر نگاه کردن به سیگنال نیز انجام می‌شود. به طور معمول تعداد و طول فریم‌ها را متناسب با فرکانس سیگنال گفتار در نظر می‌گیرند. در برخی از موارد سیگنال‌ها را به گونه‌ای فریم بندی می‌کنند که فریم‌ها با همدیگر هم‌پوشانی داشته باشند که گاهی این همپوشانی به حدود ۵۰ درصد نیز می‌رسد. در مرحله بعد به منظور حذف ناپیوستگی موجود در مرز فریم‌ها هر فریم را در یک پنجره ضرب می‌کنیم. پنجره‌ای که در محاسبه این ضرایب استفاده می‌شو یک پنجره‌ی - همینگ است که با رابطه (۲-۲) به دست می‌آید.

$$w(n) = 0.54 - 0.46 * \cos\left(\frac{2k\pi}{N-1}\right) \quad k = 0,1,\dots.,N-1 \qquad (۲-۲)$$

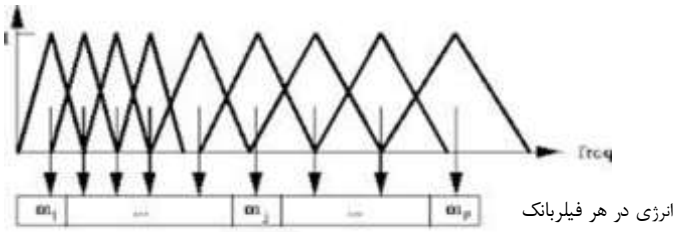
در این رابطه N طول پنجره می‌باشد که با طول فریم‌ها برابر است. در مرحله چهارم از فریم‌های پنجره بندی شده تبدیل فوریه گسسته می‌گیریم.

ایده‌ی اصلی ضرایب مل-کپستروم الهام گرفته از خواص شنیداری گوش انسان است. عملکرد گوش انسان به گونه‌ای است که فرکانس دریافتی را به همان اندازه فیزیکی آن درک نمی‌کند بلکه آن را به صورت لگاریتمی و طبق رابطه (۲-۳) درک می‌کند.

$$f_{mel} = 2595\log(1 + \frac{f}{700}) \qquad (۲-۳)$$

در این رابطه f فرکانس بر حسب هرتز و f_{mel} تبدیل یافته فرکانس‌ها از حوزه خطی به حوزه مل هستند. این رابطه همچنین بیانگر دقت بالای گوش انسان در درک فرکانس‌های پایین و دقت کم آن در درک فرکانس‌های بالا است. برای محاسبه ضرایب مل-کپستروم به منظور تبدیل فرکانس‌ها از مقیاس هرتز به مل از یک مجموعه فیلتر بانک استفاده می‌شود. به

طور معمول در این مرحله یک فیلتربانک مثلثی به کار می‌رود. این فیلتربانک‌ها به گونه‌ای هستند که در فرکانس‌های بالاتر پهنای باند فیلترهای مثلثی شکل زیاد است که نشان‌دهنده حساسیت کمتر گوش انسان نسبت به تغییر فرکانس در فرکانس‌های بالاتر نسبت به فرکانس‌های پایین‌تر است. شکل (۲–۳) این نوع فیلتربانک را نشان می‌دهد.



شکل (۲–۳): فیلتربانک مثلثی شکل

سپس انرژی هر کدام از فیلتربانک‌ها را محاسبه می‌کنیم. پس از آن به منظور کوچک کردن اعداد به دست آمده از مقادیر حاصل از انرژی با رابطه (۲-۴) لگاریتم گرفته می‌شود.

$$X'(m) = \log(X_1(m)) \qquad (۲-۴)$$

در نهایت با استفاده از رابطه (۲-۵) تبدیل کسینوسی ضرایب حاصل را به دست می‌آوریم.[4] .

$$Ceps_{MFCC}(l) = \sum_{m=1}^{M} X'(m).\cos(l\frac{\pi}{m}.(m-\frac{1}{2})) \qquad (۲-۵)$$

که در این رابطه M طول هر فریم و l شماره فیلتربانک است. آنچه که در نهایت با استفاده از رابطه (۲-۵) به دست می‌آید همان ضرایب مل-کپستروم هستند که معمولاً به ازای هر فریم ۱۳ و یا ۱۴ ضریب به دست می‌آورند.

علاوه بر ضرایب مل-کپستروم از مشتقات این ضرایب نیز به عنوان یک ویژگی استفاده می‌شود. به طور معمول مشتق اول و دوم این ضرایب در استخراج ویژگی محاسبه می‌شوند. از رابطه (۲-۶) برای به دست آوردن مشتق اول استفاده می‌شود[2].

$$\Delta MFCC = D[n] = C[n+m] - C[n-m] \quad (۶-۲)$$

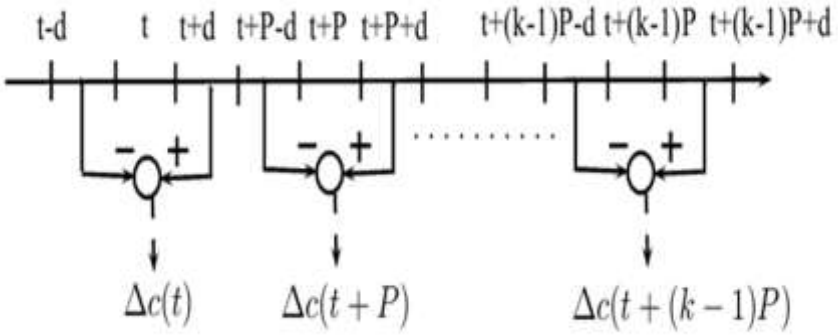
و رابطه (۲-۷) برای محاسبه مشتق دوم به کار گرفته می‌شود.

$$\Delta\Delta MFCC = DD[n] = D[n+m] - D[n-m] \quad (۷-۲)$$

در تحقیقات متعددی که در مورد ضرایب مل-کپستروم انجام‌شده است مشاهده‌شده است که با تغییر در برخی از بلوک‌های محاسبه این ضرایب می‌توان به نتایج بهتری به خصوص در محیط‌های نویزی دست یافت که در فصل سوم برخی از این تغییرات بیان خواهد شد.

۲-۲-۲ ضرایب کپسترال شیفت‌یافته (SDC)

یکی دیگر از ویژگی‌هایی که در مورد تشخیص گفتار و لهجه از سیگنال گفتار استخراج می‌شود و با ضرایب مل-کپستروم نیز مرتبط است، ضرایب کپسترال شیفت‌یافته یا همان SDC است. برای محاسبه این ضرایب از نمودار شکل (۲–۴) استفاده می‌شود.



شکل (۲–۴): نمودار محاسبه SDC [2]

در این شکل D زمان پیشرفت و تأخیر برای محاسبه دلتا و P زمان شیفت بین بلوک‌های متوالی و K تعداد بلوک‌های ضرایب دلتا است که به هم متصل می‌شوند تا بردار ویژگی نهایی را بسازند. اساساً برای ساخت هر بردار ویژگی SDC به اندازه K*N پارامتر استفاده می‌شود که N تعداد ضرایب مل-کپستروم در هر فریم است. به عنوان مثال بردار نهایی در زمان t را می‌توان از رابطه (۲-۸) محاسبه کرد[2].

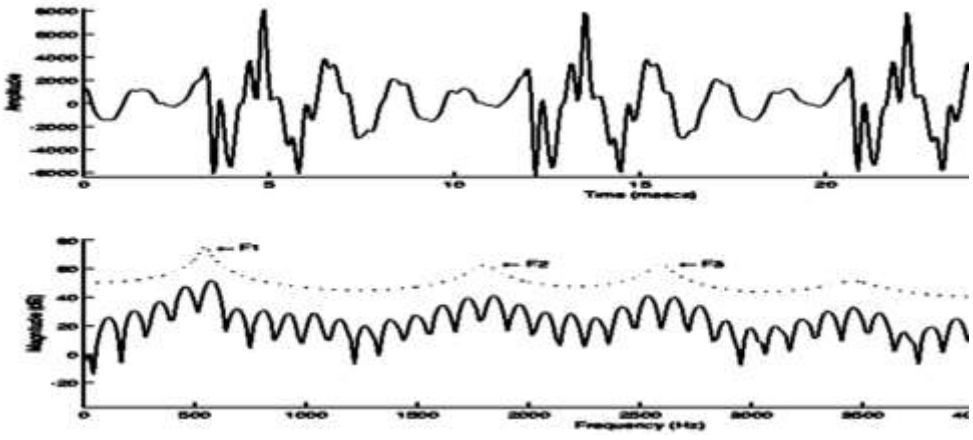
$$\Delta C(t) = C(t + ip + d) - C(t + ip - d) \qquad (۸-۲)$$

در این رابطه i از 0 تا k-1 تغییر می‌کند.

۲-۲-۳ فرکانس‌های فرمنت

کوچک‌ترین واحد در نظام آوایی یک زبان واج نامیده می‌شود که به دو دسته واکه (مصوت) و همخوان (صامت) تقسیم‌بندی می‌شود. واکه‌ها در زبان فارسی شامل ای (/i/)، آ(/a/)، او (/u/)، اِ(/e/)، اَ(/æ/) و اُ(/o/) می‌باشد. فرکانس‌های فرمنت نشان‌دهنده فرکانس‌های تشدید دستگاه گفتار می‌باشد. فر منت‌ها پارامترهای مربوط به لوله صوتی[1] می‌باشند که برای هر آوایی مقدار فرکانس فرمنت متفاوتی در مقایسه با آواهای دیگر وجود دارد که مشخصه همان آوا به شمار می‌رود. پایین‌ترین قله در طیف فرکانسی به عنوان اولین فرمنت و دومین قله به عنوان دومین فرمنت و بقیهٔ فرمنت‌ها به این ترتیب مشخص می‌شوند. برای به دست آوردن طیف فرکانسی یک سیگنال گفتار باید آن را از حوزه زمان به حوزه فرکانس منتقل کنیم. پوش طیف فرکانسی نشان‌دهنده تابع تبدیل لوله صوتی می‌باشد.

شکل (۲–۵) سه فرکانس فرمنت اول(F_1، F_2 و F_3) را برای یک سیگنال گفتار نشان می‌دهد. با توجه به اینکه فرکانس‌های فرمنت برای هر واج و از یک فرد به فرد دیگر متغیر



شکل (۲–۵): به دست آوردن فرکانس‌های فرمنت از روی پوش طیف فرکانسی سیگنال

[1] -Vocal Tract

۲-۲-۴ ضرایب کپسترال پیشگوی خطی (LPC)

آنالیز LPC روی قطعاتی از سیگنال صحبت انجام می‌گردد و اساس آن بر پیشگویی خطی است. این ویژگی همانند ضرایب مل-کپستروم مبتنی بر پردازش کپسترال است. این آنالیز به دلیل توانایی‌اش در محاسبه نسبتاً دقیق پارامترهای مربوط به صوت و همچنین سرعت بالای محاسبات آن اهمیت زیادی در پردازش صوت دارد و نقش مهمی در به دست آوردن پارامترهای مربوط به سیگنال گفتار مانند فرکانس‌های فرمنت و دوره تناوب اصلی سیگنال دارد. در مدل پیشگویی خطی، فرض می‌گردد که سیگنال صحبت یک فرایند بازگشتی است. با استفاده از رابطه‌ی (۲-۹) می‌توان این ضرایب را به دست آورد.

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (۲-۹)$$

در این رابطه H(z) بیانگر تابع تبدیل لوله صوتی، S(z) تبدیل یافته سیگنال گفتار از حوزه زمان به حوزه فرکانس، U(z) سیگنال تحریک، G پارامتر بهره، P مرتبه پیشگو کننده یا همان تعداد ضرایب و a_k همان ضرایب LPC هستند[6].

در تحلیل LPC صحبت، پارامترهای مربوط به هر دو مدل تحریک و مدل تولید صحبت توسط سیگنال ورودی تخمین زده می‌شود. روش‌های متفاوتی جهت به دست آوردن ضرایب پیش‌گو وجود دارد. از جمله مشهورترین این روش‌ها، روش همبستگی و روش کواریانس است که هر دو از تکنیک‌های پایه در حوزه زمان می‌باشند و بر اساس معیار کمترین مجذور عمل می‌نمایند.

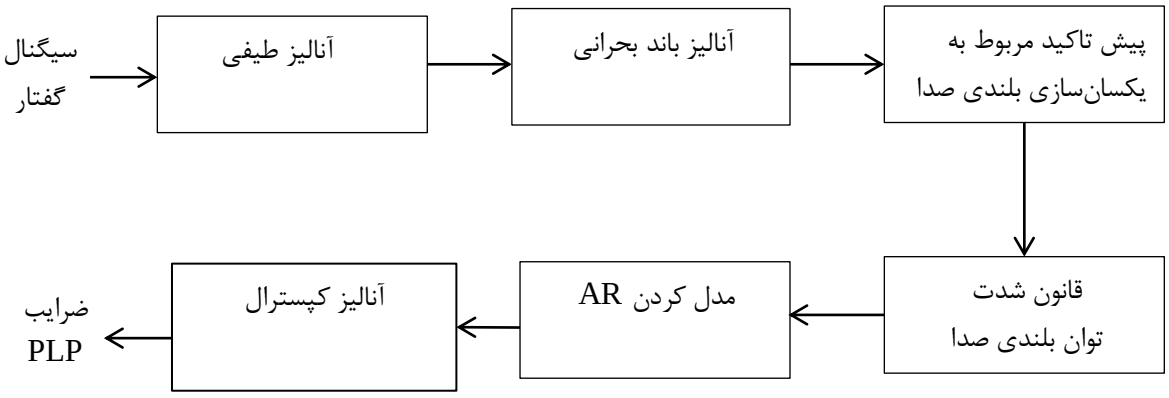
LPC سعی می‌کند که سیستم تولید گفتار صدای انسان را توسط یک فیلتر تمام قطب مدل‌سازی کند. بسته به اینکه هر فریم سیگنال گفتار، صدا باشد یا چیزی غیر از صدا و سکوت (unvoiced)، ورودی فیلتر می‌تواند یک قطار ضربه‌ی متناوب و یا یک نویز سفید باشد. دوره

تناوب سیگنال گفتار نیز توسط قطار ضربه تعیین می‌شود.

تعداد ضرایبی که پس از این فرآیند به دست می‌آید نامحدود است. اما معمولاً انتخاب ۱۲ تا ۲۰ ضریب اول می‌تواند در رسیدن به نتایج مطلوب کافی باشد.

۲-۲-۵ پیشگویی خطی مبتنی بر درک انسان (PLP)

یکی از معایب آنالیز LPC آن است که فرآیند و نحوه عملکرد سیستم شنوایی انسان را در محاسبه ویژگی‌ها منظور نمی‌کند. به عبارت دیگر LPC در تمامی فرکانس‌ها سیگنال گفتار را به یک صورت تخمین می‌زند که این مطابق با سیستم شنوایی انسان نیست. برای حل این مشکل و به منظور هماهنگ کردن روش پیشگویی خطی با سیستم شنوایی انسان، هرمانسکی[1] [7] در سال ۱۹۸۹ روش پیشگویی خطی مبتنی بر درک انسان را پیشنهاد داد[7]. او آنالیز طیفی را به گونه‌ای انجام داد که بعضی از نواحی حساس‌تر از بقیه قسمت‌ها شوند. برای رسیدن به این هدف، هرمانسکی به جای استفاده از مقیاس مل از یک مقیاس جدید به نام بارک استفاده کرد. برای به دست آوردن این ویژگی، مراحل شکل (۲–۶) استفاده می‌شود.



شکل (۲–۶): مراحل محاسبه ویژگی ضرایب PLP [7]

پس از اینکه سیگنال فریم بندی می‌شود و از پنجره عبور می‌کند در مرحله اول یعنی مرحله آنالیز طیفی، طیف توان زمان کوتاه مربوط به هر فریم با استفاده از تبدیل فوریه به

[1] Hermansky

دست می‌آید. پس از این مرحله فرکانس‌ها توسط رابطه (۲-۱۰) از مقیاس خطی به مقیاس بارک[1] نگاشته می‌شوند.

$$\Omega(w) = 6 * \ln(\frac{w}{1200\pi} + [(\frac{w}{1200\pi})^2 + 1]^{0.5}) \qquad (۲-۱۰)$$

در این رابطه w فرکانس زاویه‌ای بر حسب رادیان بر ثانیه است.

پس از انتقال طیف توان به مقیاس جدید آن را با نمودار باند بحرانی پوشاننده ($\Psi(\Omega)$) کانوال می‌کنیم. این قسمت تقریباً شبیه آنالیز کپسترال در محاسبه ضرایب مل-کپستروم است. نمودار باند بحرانی را با رابطه (۲-۱۱) نشان می‌دهند.

$$\Psi(\Omega) = \begin{cases} 0 & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & -1.3 \leq \Omega < -0.5 \\ 1 & -0.5 \leq \Omega < 0.5 \\ 10^{-0.1(\Omega-0.5)} & 0.5 \leq \Omega < 2.5 \\ 0 & \Omega \geq 2.5 \end{cases} \qquad (۲-۱۱)$$

در مرحله بعد نمودار مربوط به یکسان‌سازی بلندی صدا به نتیجه حاصل از کانوال مرحله قبل اعمال می‌شود. تابعی که در این مرحله استفاده می‌شود تخمینی از حساسیت غیریکنواخت شنوایی انسان در فرکانس‌های مختلف می‌باشد که حساسیت شنوایی انسان را در حدود ۴۰ دسی‌بل شبیه‌سازی می‌کند.

با توجه به اینکه میزان احساس بلندی صدا در گوش انسان با ریشه سوم انرژی آن متناسب است در مرحله بعد به منظور تخمین قانون توان مربوط به سیستم شنوایی انسان ریشه سوم را به طیف موجود اعمال می‌کنیم.

در مرحله مدل کردن AR ابتدا از نتایج حاصل از مرحله قبل تبدیل فوریه معکوس گرفته می-شود تا رشته خودهمبستگی به دست آید. سپس توسط معادله‌های یول-واکر، که یکی از روش‌های محاسبه ضرایب پیشگویی خطی هستند، فیلتر تمام قطب مدل می‌شود و

[1] Bark

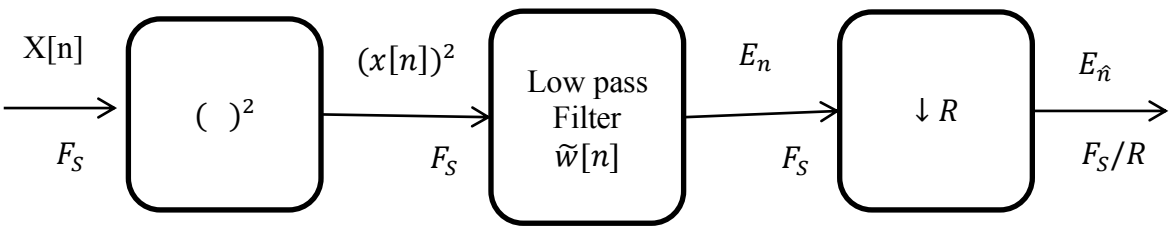
ضرایب AR به دست می‌آید. پس از محاسبه ضرایب AR آنالیز کپسترال همانند محاسبه LPC انجام‌شده و نهایتاً ضرایب پیش‌گویی خطی مبتنی بر درک انسان به دست می‌آیند.

آنالیز باند بحرانی در محاسبه ضرایب PLP تا حد زیادی شباهت به آنالیز فیلتربانک مل در محاسبه‌ی ضرایب مل-کپستروم دارد. برخی از آزمایش‌های انجام‌شده بیانگر این مسئله هست که روش MFCC نسبت به PLP دارای نتایج بهتری در تشخیص گفتار است[8]. اما روش PLP نسبت به تغییرات تعداد ضرایب و تعداد فیلترهای مورد استفاده در محاسبه‌ی ضرایب دارای نتایج پایدارتری نسبت به MFCC می‌باشد[9]. علاوه بر این ویژگی PLP در برابر نویز مقاومت بیشتری از خود نشان می‌دهد و مصون‌تر است.

به طور کلی این آنالیز به دلیل کم بودن پیچیدگی محاسباتی آن یک ویژگی کارا است و سیگنال گفتار را با بعدی کمتر از خود سیگنال نمایش می‌دهد. و این ویژگی‌ها باعث شده است که از این آنالیز در تشخیص گفتار مستقل از گوینده بیشتر استفاده گردد[7].

۲-۲-۶ انرژی سیگنال در هر فریم

آخرین ویژگی که در این بخش بررسی خواهد شد، ویژگی انرژی سیگنال در هر فریم است که معمولاً آن را به همراه یک یا دو ویژگی دیگر در یک بردار ویژگی قرار داده و استفاده می‌کنند. برای به دست آوردن این ویژگی، سیگنال گفتار مراحل شکل (۲–۷) را طی می‌کند.



شکل (۲–۷): مراحل محاسبه انرژی سیگنال در هر فریم

در این شکل، X[n] سیگنال فریم بندی شده است که ابتدا آن را به توان دو می‌رسانیم و سپس از فیلتر پایین گذر $\tilde{w}[n]$ عبور داده و آنگاه با کاهش نرخ نمونه‌برداری به اندازه R از خروجی فیلتر نمونه‌برداری کرده و خروجی $E_{\hat{n}}$ را به عنوان یک بردار ویژگی استفاده می‌کنیم که هر بردار ویژگی نشان‌دهنده انرژی سیگنال در هر فریم است.

۲-۳ طبقه‌بندی

با پایان یافتن مرحله استخراج ویژگی و تشکیل بردار ویژگی‌های مناسب به منظور جدا کردن کلاس‌های مختلف و به عبارت دیگر طبقه‌بندی کردن داده‌ها باید از یک طبقه بند[1] مناسب استفاده کرد. در بیشتر موارد طبقه بندها در این مرحله به صورت تکی استفاده می‌شود اما گاهی اوقات برای رسیدن به نتایج بهتر دو یا چند طبقه بند با یکدیگر ترکیب می‌شوند. در ادامه به معرفی چند طبقه بند که در زمینه تشخیص لهجه‌ها بیشتر استفاده‌شده‌اند خواهیم پرداخت.

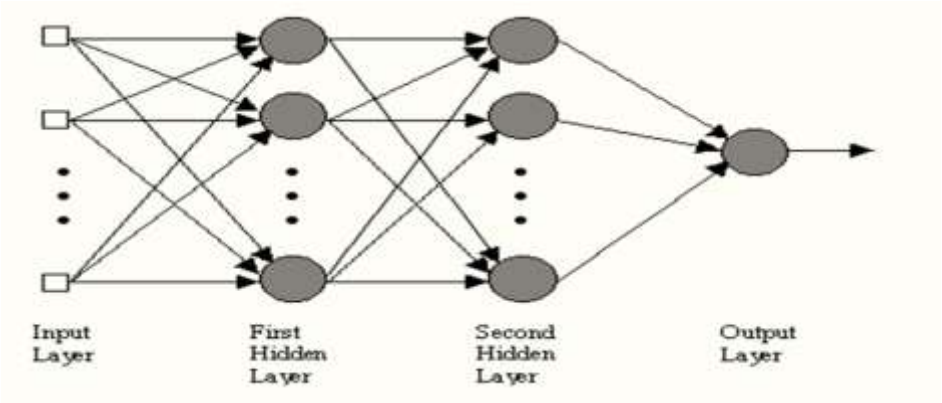
۲-۳-۱ شبکه عصبی پرسپترون چند لایه (MLP)

این شبکه عصبی اولین بار در سال ۱۹۵۷ توسط آقای روزنبلات معرفی شد. در ابتدا این شبکه قادر به تشخیص دو کلاس بود و نیاز به گسترش داشت که در نتیجه شبکه پرسپترون چند لایه به وجود آمد[10] . شبکه MLP شامل سه لایه به نام‌های ورودی، مخفی و خروجی است که تعداد سلول‌های هر لایه به روش سعی و خطا مشخص می‌گردد. البته تعداد لایه‌های مخفی می‌تواند بیش از یکی باشد که به منظور بهبود عملکرد شبکه این کار انجام می‌شود. در این شبکه سیگنال‌های ورودی به وسیله‌ی ضریب‌های به هنجار کننده به مقدار ۱ نرمالیزه شده و بعد از محاسبات، خروجی به مقدار واقعی برگردانده می‌شود.همچنین مقادیر اولیه وزن‌ها به صورت اتفاقی در نظر گرفته می‌شوند. این شبکه بر مبنای الگوریتم پس انتشار خطا آموزش

[1] classifier

می‌بیند. بدین ترتیب که خروجی‌های واقعی با خروجی‌های دلخواه مقایسه می‌شوند و وزن‌ها به وسیله‌ی الگوریتم پس انتشار به صورت تحت نظارت تنظیم می‌گردند تا الگوی مناسب به وجود آید. وزن‌ها باهدف کاهش خطا به روش گرادیان نزولی تنظیم می‌گردند و به طور مکرر برای تمام الگوهای یادگیری به‌روز درآورده می‌شوند. روند یادگیری هنگامی متوقف می‌شود که مجموع کل خطا از مقدار آستانه تعیین‌شده کمتر شود و یا تعداد کل دوره آموزش به پایان برسد[11].

شکل (۲–۸) نمایی کلی از این شبکه را نشان می‌دهد.



شکل (۲–۸): نمایی کلی از یک شبکه عصبی پرسپترون چند لایه

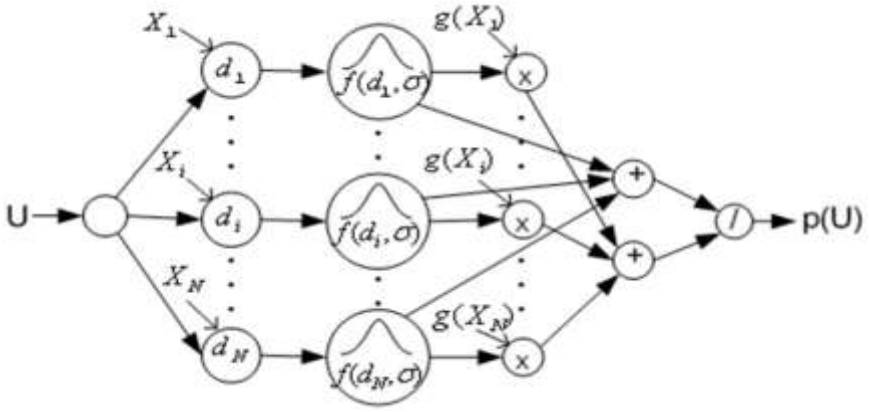
لازم به ذکر است شبکه عصبی نشان داده‌شده در شکل (۲–۸) دارای یک لایه ورودی، دو لایه مخفی و یک لایه خروجی است که تعداد نرون[1] هر لایه با توجه به ابعاد بردار ویژگی ورودی به این شبکه مشخص می‌شود. تحقیقات نشان می‌دهد که کارایی شبکه عصبی پرسپترون چند لایه و بیشتر طبقه بندهای دیگر نیز با افزایش تعداد کلاس کاهش می‌یابد و عملکرد بهینه و مطلوب این شبکه را تا زمانی که تعداد کلاس‌ها از عدد ۵ بیشتر نشود می‌دانند و بعد از آن با کاهش کارایی این طبقه بند روبه‌رو خواهیم شد[25] .

[1] neuron

۲-۳-۲شبکه عصبی احتمالاتی (PNN)

این طبقه بند از نوع شبکه‌های عصبی توابع بنیادی شعاعی(RBF) می‌باشد و بر پایه تابع چگالی احتمال نمایی و قانون تصمیم‌گیری Bayes عمل می‌کند. PNN یک شبکه پیش‌رونده سه لایه بوده و از یادگیری با نظارت استفاده می‌کند. لایه اول ورودی‌ها را دریافت می‌کند، لایه میانی بردار احتمال را بر پایه توزیع هر کلاس تعیین می‌کند و لایه آخر مقدار بیشینه‌ی احتمال را انتخاب کرده و کلاس مربوطه را تعیین می‌کند. این شبکه به صورت موازی محاسبات را انجام می‌دهد، آموزش ساده‌ای دارد و به دلیل سرعت بالا در کاربردهای بی‌درنگ مورد استفاده قرار می‌گیرد[۵].

شکل (۲–۹) ساختار شبکه عصبی احتمالاتی را نشان می‌دهد.



شکل (۲–۹): ساختار شبکه عصبی احتمالاتی[۵]

۲-۳-۳ماشین بردار پشتیبان (SVM)

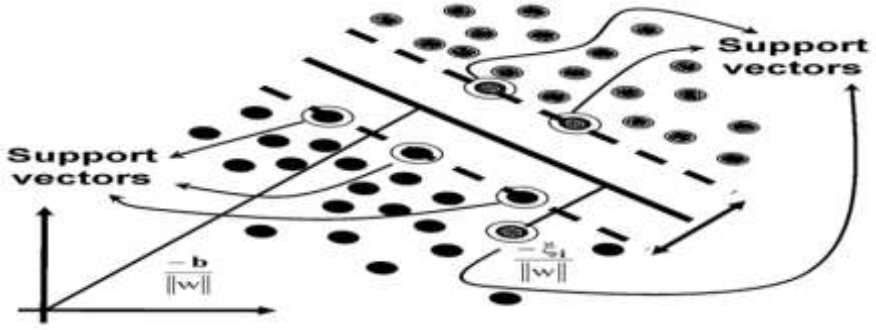
SVM در واقع یک طبقه‌بندی کننده دودویی است که دو کلاس را با استفاده از یک مرز خطی از هم جدا می‌کند. در این روش با استفاده از تمامی باندها و یک الگوریتم بهینه‌سازی، نمونه‌هایی که مرزهای کلاس‌ها را تشکیل می‌دهند به دست می‌آورند. این نمونه‌ها را بردارهای پشتیبان گویند. تعدادی از نقاط آموزشی که کمترین فاصله تا مرز تصمیم‌گیری را دارند می-

توانند به عنوان زیرمجموعه‌ای برای تعریف مرزهای تصمیم‌گیری و به عنوان بردار پشتیبان در نظر گرفته شوند[12]. البته قبل از تقسیم خطی باید دقت کرد که برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به وسیله تابع Ø به فضایی با ابعاد بالاتر منتقل می‌کنیم. برای محاسبه مرز تصمیم‌گیری دو کلاس کاملاً جدا از هم از روش حاشیه بهینه استفاده می‌شود. در این روش مرز خطی بین دو کلاس به گونه‌ای محاسبه می‌شود که:

۱- تمام نمونه‌های کلاس ۱+ در یک طرف مرز و تمام نمونه‌های کلاس ۱- در طرف دیگر مرز قرار می‌گیرند.

۲- مرز تصمیم‌گیری به گونه‌ای باشد که فاصله نزدیک‌ترین نمونه‌های آموزشی هر دو کلاس از یکدیگر در راستای عمود بر مرز تصمیم‌گیری تا جایی که ممکن است حداکثر شود.

شاید به گونه‌ای بتوان محبوبیت کنونی روش ماشین بردار پشتیبان را با محبوبیت شبکه‌های عصبی در دهه گذشته مقایسه کرد. علت این قضیه نیز قابلیت استفاده این روش در حل مسائل گوناگون می‌باشد. در شکل (۲-۱۰) این طبقه بند و نحوه عملکرد آن و همچنین بردارهای پشتیبان حاصل از اجرای آن نشان داده‌شده است.



شکل (۲–۱۰): ساختار و نحوه عملکرد طبقه بند SVM

۲-۳-۴طبقه بند آماری K نزدیک‌ترین همسایه (KNN)

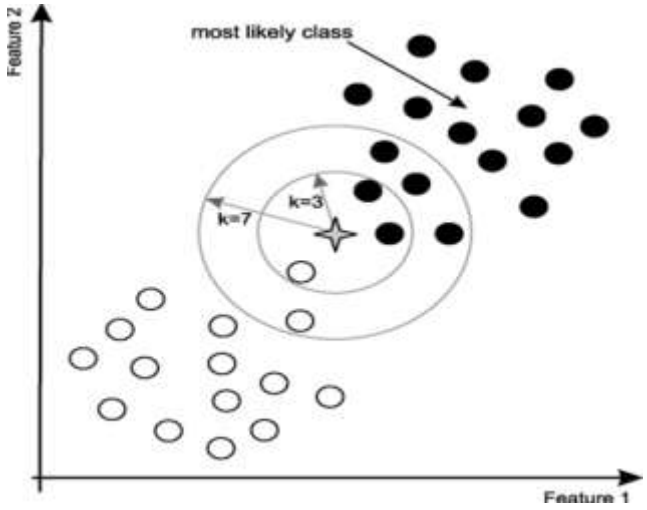
KNN یک الگوریتم آموزش با سرپرستی است. در حالت کلی از این الگوریتم به دو منظور استفاده می‌شود: برای تخمین تابع چگالی توزیع داده‌های آموزش و برای طبقه‌بندی داده‌های آزمایش بر اساس الگوهای آموزش. برای تخمین p(x) از روی n نمونه‌ی آموزش توسط الگوریتم KNN می‌توانیم یک سلول به مرکزیت x ایجاد کرده و اجازه دهیم شعاع این سلول تا حدی گسترش پیدا کند که kn نمونه‌ی آموزش را در بر گیرد. این نمونه‌ها kn نزدیک‌ترین همسایه‌های x هستند. در حالت کلی k را به صورت kn در نظر می‌گیریم که kn تابعی تعریف‌شده از n است. اگر چگالی نقاط آموزش اطراف x زیاد باشد سلول کوچک می‌شود و بنابراین نتیجه‌ی به دست آمده نتیجه‌ی بهتری است و در صورتی که چگالی نقاط آموزش اطراف x کم باشد سلول بزرگ می‌شود. در حالت کلی چگالی توزیع به ازای هر نقطه x توسط رابطه‌ی (۲-۱۲) محاسبه می‌شود.

$$P_n(x) = \frac{k_n/n}{V_n} \qquad (۱۲-۲)$$

اگر با رشد n، k_n نیز افزایش پیدا کند به طوری که با رفتن n به سمت بی‌نهایت k_n نیز به بی‌نهایت میل کند، آنگاه می‌توان مطمئن بود که k_n/n یک تخمین خوب از این احتمال است که یک نقطه در سلولی به حجم v_n قرار بگیرد. بنابراین دو شرط بیان‌شـده در رابطـه (۲-۱۳) شروط لازم و کافی برای این هستند که $p_n(x)$ در بی‌نهایت به p(x) همگرا شود.

$$\lim_{n\to\infty} k_n = \infty \ , \ \lim_{n\to\infty} k_n/n = 0 \qquad (۱۳-۲)$$

شکل (۲–۱۱) نمایی از نحوه عملکرد طبقه بند KNN را نشان می‌دهد.



شکل (۲–۱۱): نمایی از نحوه عملکرد طبقه بند KNN

لازم به ذکر است که بالاترین بازدهی طبقه بند KNN زمانی است که تعداد کـلاس‌هـا دوتا باشد.

۲-۳-۵ مدل مخلوط گوسی (GMM)

GMM توانایی قابل‌قبولی در مدل کردن داده‌های نامنظم دارد و روشی کاملاً پایدار برای نشان دادن ویژگی‌های آوایی گوینده است. این مدل، ترکیبی از چند مدل گوسی است. به طور ساده می‌توان گفت که قله‌های گوسی در چگالی طیف این مدل، همان محل تجمع بردارهای مربوط به یک آوای خاص است. این موضوع یکی از دلایل عمده‌ای است که باعث می‌شود از GMM برای بیان فضای آوایی گوینده استفاده شود. در آموزش مدل گوسی هدف این است که پارامترهای مدل با استفاده از داده‌های آموزشی موجود تخمین زده شود تا بهترین تطبیق بر روی بردارهای ویژگی گوینده به دست آید.

در سیستم‌های مبتنی بر مدل مخلوط گوسی توزیع احتمال بردارهای ویژگی یک سیگنال گفتار $X=\{X_t, 1\leq t \leq T\}$ به صورت یک ترکیب خطی از K مخلوط گوسی به صورت رابطه‌ی (۲-۱۴) بیان می‌شود[13] .

$$P(X_t|\lambda) = \sum_{k=1}^{K} C_k N(X_t, \mu_k, \Sigma_k) \qquad (۱۴-۲)$$

در رابطه‌ی بالا λ یک علامت اختصاری برای پارامترهای مدل مخلوط گوسی است و به صورت رابطه‌ی (۲-۱۵) تعریف می‌شود.

$$\lambda = \{C_k, \mu_k, \Sigma_k\} \qquad 1 \leq k \leq K \qquad (۱۵-۲)$$

همچنین $N(X_t, \mu_k, \Sigma_k)$ یک تابع چگالی احتمال با بردار میانگین μ_k و ماتریس کواریانس Σ_k است که با رابطه‌ی (۲-۱۶) نشان داده می‌شود..

$$N(X_t, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \times exp\left[-\frac{1}{2}(x_t - \mu_k)'{\Sigma_k}^{-1}(x_t - \mu_k)\right] \qquad (۱۶-۲)$$

که در این رابطه منظور علامت $(')$، ترانهاده‌ی یک ماتریس و D تعداد بعد بردار X_t است. در[13] گفته‌شده است که معمولاً از ماتریس کواریانس قطری استفاده می‌شود. دلیل این امر را نیز در ویژگی‌های کپسترال دانسته است. زیرا ویژگی‌های کپسترال تقریباً غیر همبسته هستند. انتخاب تعداد مخلوط‌ها وابسته به تعداد داده‌های آموزشی است. البته بایستی به اندازه‌ای باشند که بتوانند تغییرات صوتی گویندگان را مدل کنند. به عبارت دیگر تعداد مخلوط‌های گوسی باید به اندازه‌ای باشد که با تعداد موجود از داده‌های آموزشی بتوان پارامترهای آن را تخمین زد [۱۴].

۲-۳-۶ مدل مخفی مارکوف (HMM)

HMM یک مدل آماری برای مدل‌سازی سیگنال است که در اواخر دهه ۱۹۶۰ میلادی معرفی گردید و در حال حاضر به سرعت در حال گسترش دامنه کاربردها می‌باشد. دو دلیل مهم برای این مسئله وجود دارد ؛ اول اینکه از لحاظ ریاضی بسیار قدرتمند است و به همین دلیل مبانی نظری بسیاری از کاربردها را شکل داده است. دوم اینکه مدل مخفی مارکوف اگر به

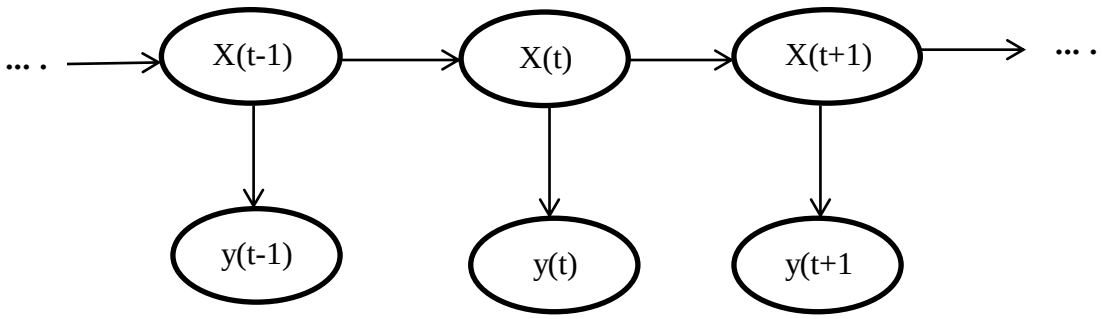
صورت مناسبی ایجاد شود می‌تواند برای کاربردهای بسیاری مورد استفاده قرار گیرد. دلیل اطلاق کلمه مخفی به این مدل این است که درباره‌ی مسائلی صحبت می‌کنیم که روش انجام آن‌ها از دید ما پنهان است و البته ماهیت پارامتری آماری دارد؛ یعنی اینکه نه تنها نمی‌دانیم نتیجه چه خواهد بود، بلکه نوع اتفاق و احتمال آن اتفاق نیز باید از پارامترهایی که در دسترس است نتیجه‌گیری شود.

پارامترهای اصلی مدل مارکوف عبارت‌اند از:

- مجموعه حالت‌هایی که ممکن است اتفاق بیفتد.
- مجموعه تصمیماتی که می‌توان در حالت‌های مختلف گرفت.
- مجموعه نتایجی که ممکن است به دنبال هر تصمیم‌گیری به دست آید.
- منافع و ارزش افزوده این تصمیم‌گیری در مقایسه با تصمیمات ممکن دیگر.

با توجه به موارد بالا باید با گرفتن مناسب‌ترین تصمیم، بهترین راه حل برای مسئله مطرح‌شده را تشخیص داده و به بهترین حالت بعدی ممکن رسید.

بسته به کاربرد مورد نظر برای مدل مارکوف ممکن است آن را به صورت‌های مختلف استفاده کنیم. نمودار شکل (۲–۱۲) ساختار کلی این مدل را نشان می‌دهد.



شکل (۲–۱۲): نمودار ساختار کلی مدل مخفی مارکوف

در این نمودار هر شکل بیضی بیانگر یک متغیر تصادفی است که مقادیری را می‌پذیرد.

x(t) مقدار متغیر تصادفی است که مقدار تغییرپذیرش در واحد زمان مخفی است. y(t) مقدار متغیر تصادفی است که مقدارش در زمان t قابل مشاهده است. از نمودار مشخص است که مقدار x(t) به مقدار x(t-1) وابسته است که این خاصیت را مارکوف می‌نامند. به طور مشابه مقدار y(t) نیز به x(t) وابسته است.

۲-۳-۷ تشخیص زبان با مدل‌سازی صدا (PRLM)

PRLM یک روش واج‌آرایی[1] است که اساساً در تشخیص زبان استفاده می‌شود. در این روش بعد از استخراج ویژگی از سیگنال گفتار، یک سیستم بازشناسی واج را آموزش می‌دهیم تا هر گفتار را به یک رشته از واج‌های آن تبدیل کند. اگر فرض کنیم هدف ما تشخیص لهجه باشد پس از اعمال مرحله آموزش به سیگنال گفتار دارای لهجه و تشکیل رشته واجی آن یک مدل زبانی آوایی برای آن لهجه می‌سازیم و این کار را برای همه داده‌های آموزش انجام می‌دهیم. سپس اگر یک سیگنال گفتار به عنوان داده تست به سیستم وارد شود ابتدا رشته واج‌های تشکیل‌دهنده آن را به دست می‌آوریم و فاصله هر کدام از مدل‌هایی را که قبلاً ساخته‌شده‌اند را با داده تست محاسبه می‌کنیم. مدلی که کمترین فاصله را داشته باشد، لهجه‌ی داده‌ی تست مربوط به آن مدل می‌شود و بدین ترتیب نوع لهجه مشخص می‌گردد[15]

تشخیص زبان با مدل‌سازی صدا به صورت موازی (PPRLM) طبقه بند دیگری است که عملکرد آن همانند PRLM است، با این تفاوت که در این طبقه بند ابتدا همه‌ی داده‌های مربوط به آموزش به صورت موازی به آن واردشده و عمل آموزش سیستم بازشناسی واج برای همه‌ی آن‌ها به طور همزمان انجام می‌شود و سپس رشته واج‌های تشکیل‌دهنده هر گفتار به دست می‌آید در نتیجه سرعت کاری و بازدهی بیشتری خواهیم داشت.

همان طور که در ابتدای این بخش عنوان شد علاوه بر استفاده از طبقه بندهای

[1] - Phonotactic

معرفی‌شده به صورت تکی، گاهی با ترکیب دو طبقه بند با یکدیگر نیز طبقه بندهای جدیدی ساخته می‌شود که در برخی مواقع بازده بهتری از خود نشان می‌دهند. در یکی از این موارد به ترکیب دو طبقه بند GMM و PRLM پرداخته‌شده است که باعث بهبود نرخ بازشناسی گشته است[15]

۲-۴ مروری بر تحقیقات انجام‌شده درباره تشخیص لهجه‌های خارجی

هرچند هدف ما در این پایان‌نامه تشخیص لهجه‌های فارسی است. اما نگاهی مختصر به پژوهش‌های صورت گرفته در حوزه لهجه‌های خارجی می‌تواند روش‌ها و مسیرهای جدیدی به منظور رسیدن به هدفمان در اختیار ما قرار دهد. همان طور که در بخش مقدمه بیان شد بیشتر تحقیقات صورت گرفته در زمینه لهجه‌های خارجی مربوط به زبان انگلیسی است که در ادامه به ذکر جزئیاتی از برخی از این تحقیقات می‌پردازیم.لازم به ذکر است در این بخش از آوردن جدول‌های نتایج به دلیل عدم اهمیت آن‌ها در این پایان‌نامه صرف‌نظر شده است.

در سال ۱۹۹۶ ارسلان و هانسن[16] با ارائه یک سیستم کلاس‌بندی لهجه برای زبان انگلیسی یکی از برجسته‌ترین و اولین تحقیقات انجام‌شده در این زمینه را انجام دادند. این تحقیق نشان داد که هر چه طول و تعداد کلمات گفته‌شده بیشتر باشد، دقت سیستم کلاس-بندی گفتار بیشتر می‌شود. این مرجع همچنین نشان داد بازدهی در سیستم تشخیص گفتار با قابلیت دسته‌بندی لهجه‌ها به طور چشم‌گیری بیشتر از سیستم تشخیص گفتار بدون شناخت لهجه است.

در همان سال کارلوس تکسیرا[16] و همکارانش بر روی لهجه‌های انگلیسی که از شش کشور اروپایی بودند به تحقیق پرداختند. آن‌ها در این تحقیق با جداسازی کلمات از داخل

جملات ویژگی مورد نظر را استخراج کردند و در مرحله طبقه‌بنـدی از مـدل HMM اسـتفاده کردند.

در سال ۱۹۹۸ برکلینگ، زیسمن[18] و دو تن دیگر از همکارانشان در زمینه تشخیص لهجه‌ی انگلیسی استرالیایی تحقیقاتی انجام دادند. آن‌ها با استفاده از ویژگی‌های مربوط به زبان مانند ویژگی‌های آوایی و واجی و همچنین با ترکیب روش‌های قبلی پیشنهاد شده برای این کار توانستند بهبود قابل‌ملاحظه‌ای در نرخ بازشناسی لهجه‌های مورد تحقیقشان ایجاد کنند.

در سال ۱۹۹۹فانگ و کت[19] یک روش ترکیبی را هم بـر اسـاس ویژگـی و هـم بـر اساس مدل ارائه کردند. برای کلاس‌بندی سریع لهجه‌ها با داده‌های کـم بـه جـای اسـتفاده از HMM بر اساس آوا، این مقاله آموزش HMM های کلاس- آوایی را پیشنهاد می‌کند.

در سال ۲۰۰۱ چن[20] به همراه همکارانش در مرکز تحقیقات مایکروسافت چـین، بـر روی تشخیص لهجه‌های چینی تحقیقاتی انجام دادند. آن‌ها بدین منظور یک نوع مدل مخلوط گوسی استفاده کردند که در آن ابتدا تعدادی مدل آمـوزش مـی‌دادنـد و سـپس مـدلی را کـه بیشترین شباهت به ورودی این طبقه بند داشت به عنوان لهجه آن گفتار در نظر می‌گرفتند.

فاریا[21] در سال ۲۰۰۵ تلاش دیگری بـرای کـلاس‌بنـدی لهجـه‌هـا روی پایگـاه داده Fisher با روش‌های GMM و SVM انجام داد. در این تحقیق علاوه بر ویژگی‌های صوتی گفتار از ویژگی‌های کلامی گفتار نیز استفاده‌شده است.

در همان سال هنگ [22] و همکارانش با ارائه یک روش جدید که ترکیبی از روش‌هـای قبلی بود توانست به بهبود تشخیص لهجه‌های چینی کمک کنـد. پدرسـن و دیـدریچ[23] در سال ۲۰۰۷ کلاس‌بندی لهجه را با کمک SVM و یادگیری درخت تصمیم برای دو لهجه عربی و هندی از زبان انگلیسی انجام داد. در این مطالعه کلاس‌بندی بر اساس متوسط ویژگی‌ها روی

چند فریم متوالی انجام شد. به منظور از بین بردن تفاوت گفتارها و برای یکسان‌سازی از تفاضل متوسط کپسترال و نرمال سازی انرژی در این تحقیق استفاده‌شده است.

اولا و کاری[24] نیز در سال ۲۰۰۷ کلاس‌بندی دیگری از گفتار بر اساس یک معیار فاصله انجام دادند. ایده‌ی این روش از آنجا گرفته‌شده است که زمانی که یک گوینده غیربومی شروع به یادگیری زبان دومی کند تلفظ آواهای زبان مادری خود را به زبان جدید القا می‌کند. این جایگذاری منجر به ابهاماتی بین مرزهای آوا می‌گردد و شباهت بین آواهای مختلف را افزایش می‌دهد. این روش از این اطلاعات برای انتقال نقاط داده به فضایی که فاصله اقلیدسی بین شباهت‌ها حداقل و بین تفاوت‌ها حداکثر است استفاده می‌کند.

۲-۵ مروری بر تحقیقات انجام‌شده درباره تشخیص لهجه‌های فارسی

تحقیقاتی که در حوزه لهجه‌های فارسی انجام‌شده است محدود است. آنچه که ما در این زمینه مشاهده کردیم شامل سه تحقیق می‌شود که در ادامه به بیان جزئیات آن‌ها خواهیم پرداخت.

مرجع [۵] با انتخاب پنج لهجه اصفهانی، تهرانی، آذری، کردی و مازندرانی، روش کلی شامل مراحل ذکرشده در بخش‌های قبل یعنی پیش‌پردازش، استخراج ویژگی و طبقه‌بندی را برای تشخیص لهجه‌ها انجام داده است. در مرحله پیش‌پردازش قسمت سکوت و گفتار از یکدیگر جدا می‌شوند و به منظور جبران دامنه در فرکانس‌های بالا یک فیلتر پیش تاکید به سیگنال گفتار اعمال می‌شود. در مرحله استخراج ویژگی، ویژگی‌هایی شامل ضرایب مل-کپستروم و مشتقات اول و دوم آن و همچنین انرژی هر فریم استخراج‌شده‌اند. طول فریم‌ها ده میلی‌ثانیه است و همپوشانی ندارد. پس از این مرحله با به کار بردن دو طبقه بند شبکه عصبی

احتمالاتی و ماشین بردار پشتیبان نتایج نهایی به دست آمده‌اند. جدول (۲-۱) نتایج حاصل از این تحقیق را که شامل سه مورد از بالاترین نرخ‌های بازشناسی به ازای جمله‌ها و طول واج‌های مشخص است نشان می‌دهد.

جدول (۲–۱): نتایج نرخ بازشناسی مرجع [۵]

شماره جمله	نرخ طبقه‌بندی PNN	نرخ طبقه‌بندی SVM	طول جمله (s)	طول واج (ms)
۲۱	۶۸	۵۲	۰.۵	۵۵
۱۷	۶۵	۵۸	۱	۸۵
۱۶	۶۳	۵۵	۱	۴۵ ،۴۰

مرجع[15] از میان لهجه‌های مختلف فارسی پنج لهجه اصفهانی، تهرانی، آذری، شمالی و جنوبی را برای انجام آزمایش‌های خود برگزیده است. در این مقاله ویژگی‌های ضرایب مل-کپستروم و مشتق اول و دوم آن، انرژی هر فریم، سه فرکانس فرمنت اول و SDC از سیگنال گفتار دارای لهجه استخراج‌شده است. هدف اصلی این مقاله ترکیب دو طبقه بند GMM و PRLM است که با انجام این کار نرخ تشخیص لهجه‌های مختلف نسبت به حالتی که این طبقه بندها به طور تکی استفاده می‌شوند به اندازه مطلوبی افزایش می‌یابد.

جدول (۲-۲) نتایج نرخ بازشناسی را به وسیله طبقه بندهای مختلف و در حالت‌های مختلف نشان می‌دهد.

جدول (۲–۲): نرخ بازشناسی به ازای طبقه بندهای مختلف مرجع[15]

طبقه بند / لهجه	PRLM	GMM	PRLM+GMM
اصفهانی	22.86	46.14	67
جنوبی	22.86	61	73.57
شمالی	22.86	38	50.42
تهرانی	28.57	39.14	48
آذری	25.71	40.42	41
میانگین	24.57	44.94	56.08

مرجع [25] به منظور کلاس‌بندی لهجه‌های مختلف زبان فارسی طبقه بندهای مختلفی را به‌کاربرده است. در این تحقیق سه لهجه تهرانی، اصفهانی و کرمانشاهی برای انجام آزمایش‌های مختلف انتخاب‌شده‌اند. همانند بیشتر مقالات، این مقاله نیز در مرحله استخراج ویژگی از ویژگی‌های متداولی همچون ضرایب مل-کپستروم، فرکانس‌های فرمنت و انرژی استفاده کرده است. در مرحله طبقه‌بندی سه طبقه بند مختلف شامل KNN، SVM و MLP به کار گرفته‌شده است. هدف این مقاله نشان اثبات این نکته است که شبکه‌های عصبی عملکرد بهتری در زمینه کلاس‌بندی لهجه‌ها دارند. نتایج حاصل نیز نشان می‌دهد که طبقه بند MLP عملکرد بهتری دارد. جدول (۲-۳) نتایج حاصل را نشان می‌دهد.

جدول (۲–۳): نرخ بازشناسی لهجه‌های مختلف با طبقه بندهای متفاوت مرجع[25]

طبقه بند	لهجه	تهرانی	اصفهانی	کرمانشاهی	میانگین
MLP		٪۷۴.۵	٪۸۶.۹۵	٪۸۱.۹۹	٪ ۸۱.۱۵
SVM		٪ ۵۲.۱۵	٪۴۹.۲۴	٪۳۹.۷۱	٪۴۷.۰۳
KNN1		٪۳۵.۸۶	٪ ۵۲.۹۳	٪۵۲.۸۷	٪۴۷.۲۲
KNN2		٪۳۲.۳۳	٪۵۳.۲۸	٪۵۲.۴۴	٪۴۶.۰۲

آنچه که در بخش۲-۵ بیان شد، خلاصه‌ای از تحقیقات صورت گرفته و نتایج حاصل از آن‌ها در مورد تشخیص لهجه‌های زبان فارسی بود.

در ادامه و در فصل سوم روش پیشنهادی این تحقیق در زمینه تشخیص لهجه‌های زبان فارسی بیان خواهد شد.

فصل سوم:

روش پیشنهادی

برای

تشخیص لهجه‌های

زبان فارسی

۳- روش پیشنهادی برای تشخیص لهجه‌های زبان فارسی

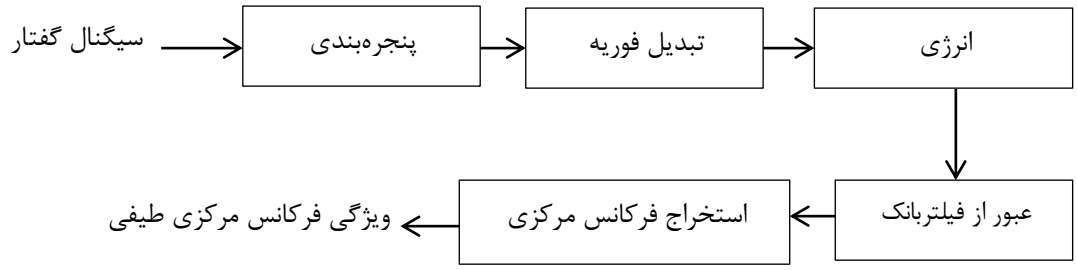
در این فصل از پایان‌نامه سعی داریم با افزودن یک یا چند روش جدید بـه هـر کـدام از مراحل سیستم تشخیص لهجه گامی در جهت بهبود نتایج و یا بهبود سیستم از لحـاظ کیفـی برداریم. این فصل از چند زیر بخش تشکیل شده است. در زیر بخش اول روش‌های پیشنهادی به منظور استخراج ویژگی بیان‌شده است و علاوه بر این به موضوعی پرداخته‌شده که تقریباً در تحقیقات مشاهده‌شده به آن توجهی نشده است و آن تشخیص لهجه در محیط نویزی است که چند ویژگی مناسب این محیط‌ها نیز معرفی‌شده است. در زیر بخش دوم به مسئله طبقه‌بندی و تغییرات و نوآوری‌هایی که می‌توان در این حوزه ایجاد کرد پرداخته‌شده است.

۳-۱ پیشنهادات مرحله استخراج ویژگی

در این پایان‌نامه از میان ویژگی‌های معرفی‌شده در فصـل ۲ چنـد ویژگـی متـداول کـه دارای بازدهی بهتری نسبت به سایر ویژگی‌ها هستند انتخاب و از سیگنال گفتـار دارای لهجـه استخراج‌شده‌اند. این ویژگی‌ها شامل ضرایب مل-کپستروم، مشتق اول و دوم آن و دو فرکـانس فرمنت اول می‌باشد. البته در استخراج این ویژگی‌ها یک نکته جدید رعایت شده است و آن هم این است که با توجه به اینکه ممکن است هر فریم به تنهـایی شـامل یـک آوا نباشـد و بـرای تشخیص لهجه معمولاً وجود چند آوا در کنار هم لازم است، از بردارهای ویژگی استخراج‌شده از هر هفت فریم متوالی میانگین‌گیری می‌شود و یک بردار ویژگی بـه ازای ایـن هفـت فـریم در خروجی خواهیم داشت که می‌تواند به تمایز و تشخیص بهتر لهجه‌ها از هم کمک نماید.لازم به ذکر است هر هفت فریم کنار هم به عنوان یک بردار ویژگی از کلاس مربوطـه وارد طبقـه‌بنـد می‌شود.در کنار این ویژگی‌ها چند ویژگی دیگر نیز از سیگنال گفتار لهجـه دار اسـتخراج‌شـده است که برخی از آن‌ها برای محیط‌های معمولی و برخی برای محیط‌های نویزی مناسب‌اند.

۳-۱-۱ پیشنهاد اول، فرکانس مرکزی طیفی (SCF)

تعدادی از مراحل محاسبه ویژگی SCF مشابه ضرایب مل-کپستروم است، نمودار شکل (۳–۱) مراحل محاسبه آن را نشان می‌دهد.



شکل (۳–۱): مراحل محاسبه ویژگیSCF[26]

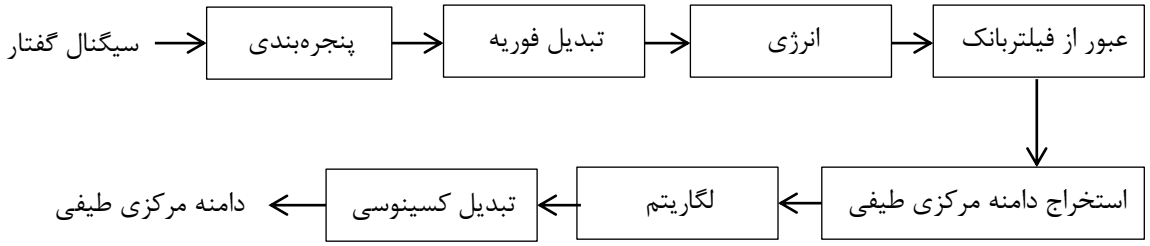
پس از مرحله عبور از فیلتربانک با اعمال رابطه (۳-۱) می‌توانیم فرکانس‌های مرکزی را استخراج کنیم.

$$F_k = \frac{\sum_{f=l_k}^{u_k} f|S[f]\omega_k[f]|}{\sum_{f=l_k}^{u_k}|S[f]\omega_k[f]|} \qquad (۱-۳)$$

در این رابطه l_k فرکانس لبه پایین و u_k فرکانس لبه بالای هر فیلتربانک مثلثی شکل می‌باشد. S[f] تبدیل یافته هر فریم از سیگنال گفتار از حوزه زمان به حوزه فرکانس و $w_k[f]$ نشان‌دهنده هر کدام از فیلتربانک‌های مثلثی شکل هستند[26] فرکانس‌های مرکزی طیفی هستند که نشان‌دهنده‌ی فرکانس میانگین وزن‌دار برای هر زیرباند است. با توجه به اینکه این ویژگی مرکز ثقل هر باند را تعیین می‌کند می‌توان از روی آن‌ها مکان تقریبی فرکانس‌های فرمنت را به دست آورد. فرکانس مرکزی طیفی معمولاً با نام مرکز طیفی نیز بیان می‌گردد. این ویژگی با توجه به اینکه با ویژگی‌های لوله صوتی انسان مرتبط است به عنوان یک ویژگی مناسب برای تشخیص لهجه می‌تواند به‌کاربرده شود. لازم به ذکر است که این ویژگی علاوه بر بازدهی خوب در محیط معمولی، در محیط‌های آلوده به نویز نیز دارای کارایی مناسبی است.

۳-۱-۲ پیشنهاد دوم، دامنه مرکزی طیفی (SCM)

این ویژگی نیز مراحلی همانند محاسبه ضرایب مل-کپستروم دارد که شکل (۳–۲) آن را نشان می‌دهد.



شکل (۳–۲): نمودار محاسبه ویژگی SCM[26]

همان طور که شکل (۳–۲) نشان می‌دهد تنها تفاوت محاسبه این ضریب با نحوه محاسبه MFCC استخراج دامنه مرکزی طیفی پس از مرحله عبور از فیلتربانک است. برای این منظور پس از عبور فریم‌ها از فیلتر بانک توسط رابطه (۳-۲) دامنه مرکزی طیفی را به دست می‌آوریم.

$$M_k = \frac{\sum_{f=l_k}^{u_k} f|S[f]\omega_k[f]|}{\sum_{f=l_k}^{u_k} f} \qquad (۳-۲)$$

در این رابطه نیز l_k فرکانس لبه پایین و u_k فرکانس لبه بالای هر فیلتربانک مثلثی شکل می‌باشد. S[f] تبدیل یافته هر فریم از سیگنال گفتار از حوزه زمان به حوزه فرکانس و $w_k[f]$ نشان‌دهنده هر کدام از فیلتربانک‌های مثلثی شکل هستند[26] . M_k دامنه مرکزی طیفی است که به ازای هر فیلتر بانکی یک ضریب به دست می‌آید، سپس با اعمال لگاریتم و تبدیل کسینوسی ویژگی نهایی را استخراج می‌کنیم. SCM تقریبی از توزیع انرژی در هر زیرباند به دست می‌دهد. این ویژگی نیز دارای عملکرد نسبتاً خوبی برای تشخیص لهجه در محیط معمولی است.

همانند ضرایب مل-کپستروم از مشتق اول ($\Delta SCM,\ \Delta SCF$) ویژگی‌های اخیر یعنی SCF و SCM نیز می‌توان به عنوان یک ویژگی برای تشخیص لهجه‌ها استفاده کرد.

۳-۱-۳ پیشنهاد سوم، ضرایب مل-کپستروم بهبودیافته

همان طور که در ابتدای این بخش بیان شد موضوع تشخیص لهجه در محیط نویزی موضوعی است که کمتر روی آن کار شده است. معمولاً ضرایب مل-کپسترومی که به طور متداول استخراج و استفاده می‌شوند دارای عملکرد خوبی در محیط‌های نویزی نیستند و باید برای مصون کردن آن‌ها در برابر نویز، تغییراتی در نحوه محاسبه‌شان ایجاد کرد. تحقیقات زیادی در این زمینه انجام‌شده است که به برخی از آن‌ها اشاره می‌شود.

در سال ۱۹۹۹ کاربرد ضرایب خودهمبستگی در بهبود این الگوریتم مطرح شد، اساس این روش این است که تابع خود همبستگی نویز در بسیاری از موارد می‌تواند در طول زمان ثابت فرض شود بنابراین اثر نویز سیگنال ورودی پس از استخراج ضرایب خود همبستگی با عبور از یک فیلتر بالا گذر مناسب تا حدودی خنثی می‌شود[27] . در ادامه این نوع ضریب را با نماد AMFCC به کار خواهیم برد.

در سال ۲۰۰۶ ضرایب خود همبستگی نویز، مجدداً مورد توجه قرار گرفت و با فرض نا همبسته بودن نویز و سیگنال اصلی، نشان داده شد که تخریب سیگنال توسط نویز در ضرایب مرتبه پایین تابع خودهمبستگی، بیشتر از ضرایب مرتبه بالا است بنابراین با حذف آن‌ها، که آستانه‌ی آن به صورت سعی و خطا مشخص می‌شد، اثر نویز تا حدود زیادی کاسته شد[28] .

در سال ۲۰۰۹ به جای فیلتر بانک مثلثی از توابع گوسی استفاده شد که به علت ایجاد همبستگی بیشتر بین فریم‌ها موجب بهبود در نرخ تشخیص گشت. از این ضریب با علامت GMFCC استفاده خواهیم کرد. در همان سال طرحی بیان شد که در آن به جای اینکه

مثلث‌های به‌کاررفته در فیلتربانک از عرض کم به عرض زیاد تشکیل شوند به عکس، از عرض زیاد به عرض کم تشکیل می‌شدند، که در ترکیب با فیلتر بانک اصلی به علت توجه به اطلاعات مکمل نادیده گرفته‌شده در فرکانس‌های بالا عملکرد بهتری در کاربرهای تشخیص گفتار از خود نشان می‌داد[29] .

در سال ۲۰۱۰ ایده‌ی استفاده از ضرایب مرتبه بالای تابع خود همبستگی با یک تفریق فرکانسی تکمیل گشت و نوع دیگری از ضرایب مل-کپسترم مصون شده در برابر نویز را به وجود آورد. [30].

در سال ۲۰۱۲ علاوه بر بخش نرم‌افزاری به بخش سخت‌افزاری در پیاده‌سازی، این الگوریتم نیز توجه گردید به نحوی که با ایجاد تغییراتی در الگوریتم پایه و کم نمودن محاسبات ضرب و جمع در آن، دروازه‌های منطقی برای پیاده‌سازی این الگوریتم کاهش چشم‌گیری یافت [31].

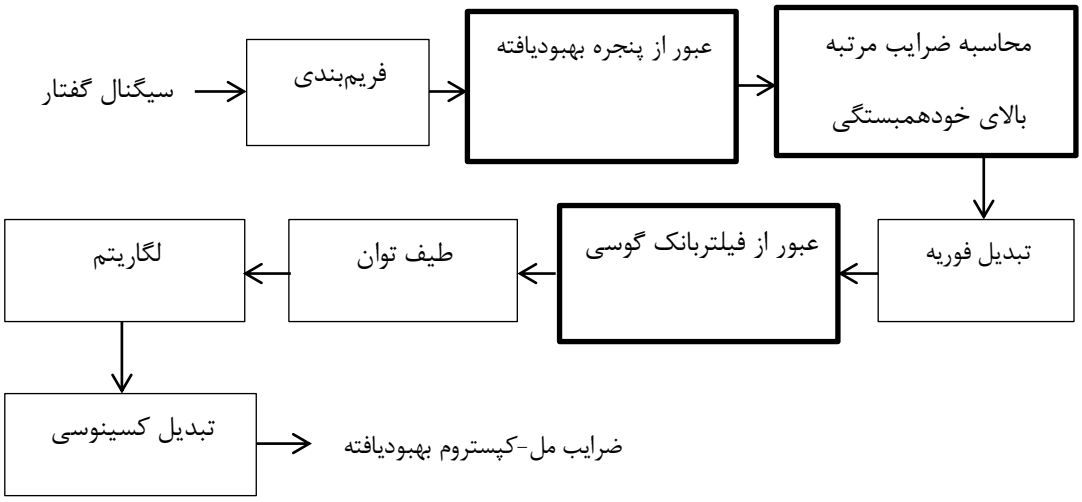
آنچه که بیان شد تنها بخشی از پیشرفت‌های حاصله بود و این روند بهبود با سرعت چشمگیری در حال پیشروی است. تغییر در الگوریتم پایه محاسبه ضرایب مل-کپستروم به منظور بهبود آن معمولاً در سه حوزه زیر انجام می‌گیرد.

- مدل های بهبودیافته شامل تغییر در بلوک‌های پایه‌ی این الگوریتم
- مدل های بهبودیافته شامل یک بلوک تکمیل‌کننده که به الگوریتم پایه اضافه گردیده است.
- بهبود در پیاده‌سازی سخت‌افزاری این الگوریتم توسط کاستن محاسبات ضرب و جمع در الگوریتم پایه.

روش‌هایی که در این پایان‌نامه مورد استفاده قرارگرفته‌اند در دو حوزه اول این تقسیم-

بندی قرار می‌گیرند.

به منظور مصون کردن ضرایب مل-کپستروم در برابر نویز یک روش که از ترکیب روش-های قبلی به دست می‌آید پیشنهاد می‌شود. شکل (۳–۳) نحوه به دست آمدن این ضرایب را نشان می‌دهد. در این شکل بلوک‌هایی که نسبت به الگوریتم پایه تغییر کرده‌اند و یا افزوده‌شده‌اند با ضخامت بیشتر و اندازه بزرگ‌تر متمایز شده‌اند.



شکل (۳–۳): تغییرات روش پیشنهادی نسبت به الگوریتم پایه محاسبه MFCC [۳۲]

با توجه به تغییرات ایجادشده در محاسبه ضرایب مل-کپستروم در ادامه ضرایب حاصل از روش پیشنهادی را با نام AGMFCC به کار خواهیم برد [۳۲]. در ادامه به توضیح بیشتر در مورد تغییرات ایجادشده می‌پردازیم.

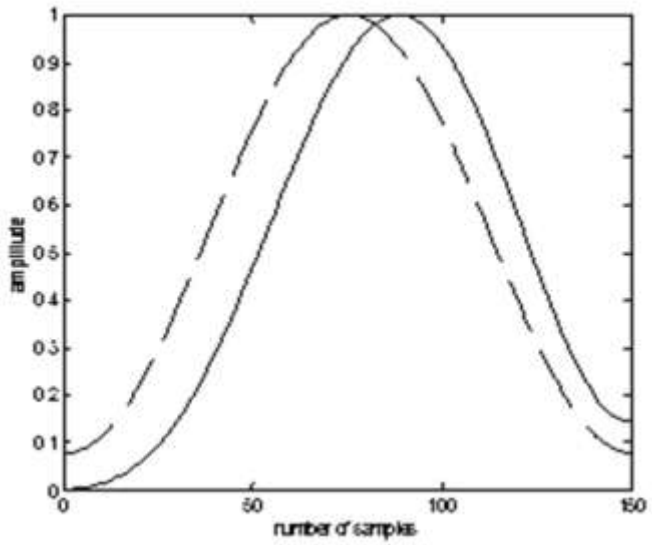
پس از فریم بندی سیگنال را از پنجره‌ای عبور می‌دهیم که نسبت به پنجره همینگ ساده دارای تغییراتی هست[33] . این پنجره را با رابطه (۳-۳) نشان می‌دهیم.

$$w_{new}(n) = nw(n) \qquad (۳-۳)$$

که w(n) همان پنجره همینگ ساده بیان‌شده در رابطه (۲-۳) با طولی متناسب با طول فریم‌ها است.در پنجره‌ی به کار گرفته‌شده در این روش، سه پارامتر مهم پارامتر پراکندگی،

همگرایی بخش‌های جانبی[1] و عرض بخش اصلی[2] پنجره در نظر گرفته‌شده است.

در این پنجره نسبت به یک پنجره‌ی همینگ ساده فاکتور پراکندگی طیفی و نیز عرض وجه اصلی افزایش و فاکتور همگرایی وجه‌های جانبی کاهش می‌یابد که دو مورد اول تغییراتی مطلوب و مورد آخر نامطلوب می‌باشد[33] تغییرات، حاکی از بهبود نتایج می‌باشد و بنابراین از عیب ایجادشده در مقابل دو مزیت فوق چشم‌پوشی می‌کنیم. شکل (۳–۴) این دو پنجره را نشان می‌دهد.



شکل (۳–۴): تفاوت دو پنجره همینگ ساده (- - - -) و تغییریافته (——)[33]

تغییر دوم ایجادشده حذف ضرایب مرتبه پایین خودهمبستگی سیگنال است. نویز اضافه‌شده به سیگنال را می‌توان با رابطه (۳-۴) بیان کرد.

$$X(n) = S(n) + d(n) \qquad (۴-۳)$$

که در آن S(n) سیگنال ورودی و d(n) نویز اضافی شونده به سیگنال است.

ویژگی این نویز با فرض نا همبسته بودن نسبت به سیگنال اصلی این است که تابع خود همبستگی مربوط به آن تا حدود زیادی نسبت به زمان بدون تغییر و نزدیک به صفر است.

[1] Side lob
[2] Main lob

رابطه (۳-۵) روابط بین تابع خودهمبستگی را نشان می‌دهد.

$$R_{XX}(m,k) = R_{SS}(m,k) + R_{dd}(m,k) \qquad (۵-۳)$$

که در آن k شماره‌ی فریم می‌باشد. با توجه به نکته یادشده که تغییرات تابع خودهمبستگی سیگنال نویز نسبت به زمان بسیار ناچیز و گاه نزدیک به صفر است می‌توان رابطه (۳-۵) را به صورت رابطه (۳-۶) نمایش داد.

$$R_{XX}(m,k) = R_{SS}(m,k) + R_{dd}(m) \qquad (۶-۳)$$

رابطه‌ی (۳-۶) این نکته را خاطرنشان می‌کند که تابع خودهمبستگی نویز نا همبسته به سیگنال، مستقل از فریم می‌باشد و رابطه (۳-۵) به رابطه‌ی (۳-۶) تبدیل خواهد شد که در آن نویز مستقل از فریم می‌باشد و با توجه به ناچیز بودن آن می‌توان با این روش اثر نویز را کاهش داد. تخریب سیگنال توسط نویز در ضرایب مرتبه پایین تابع خودهمبستگی، بیشتر از ضرایب مرتبه بالا است بنابراین با حذف آن‌ها، از اثر نویز کاسته خواهد شد.

خروجی حاصل از این مرحله وارد بلوک تبدیل فوریه می‌شود و نتیجه این تبدیل از فیلتربانک گوسی گذرانده می‌شود.

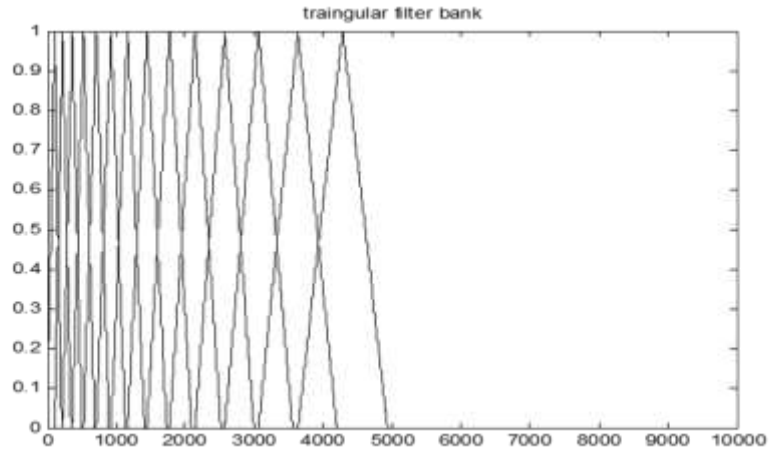
در الگوریتم پایه محاسبه ضرایب مل معمولاً از یک فیلتر بانک مثلثی استفاده می‌گردد. در این نوع فیلتربانک اطلاعات بخش‌هایی از فریم که در نقاط ابتدایی و انتهایی و خارج از زیر بخش‌ها، قرار می‌گیرند از دست می‌روند زیرا مثلث‌ها در خارج از زیر باندها وزنی ندارند. اما اگر به جای این فیلتربانک از یک فیلتربانک گوسی که متقارن نیز می‌باشد استفاده کنیم به دلیل وجود وزن در خارج از زیر باندهای آن، مانع از دست رفتن اطلاعات در این بخش‌ها می‌گردد. مزیت دیگر آن نسبت به فیلتر بانک مثلثی شروع و اتمام آن با شیب کمتر و ملایم‌تر می‌باشد. این فیلتر بانک با ایجاد همبستگی بیشتر بین زیر باندها اطلاعات از دست رفته در مرزها را کاهش داده و در بهبود الگوریتم و متعاقباً بالا بردن نرخ بازشناسی سیستم خودکار پردازش

گفتار موثر خواهد بود. شکل (۳–۵) این دو نوع فیلتربانک را نشان می‌دهد.

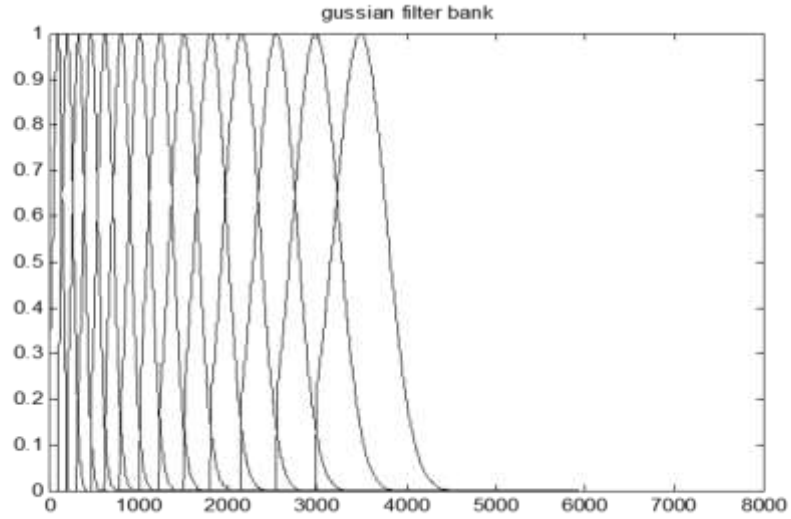
برای ایجاد فیلتربانک گوسی ابتدا توسط رابطه (۲-۳) فرکانس‌ها را به حوزه مـل انتقـال می‌دهیم. سپس به وسیله رابطه (۳-۷)، Δ_{mel} را محاسبه می‌کنیم.

$$\Delta_{mel}=\frac{f_{max}(mel)}{i+1} \qquad (۳-۷)$$

در این رابطه i نشان‌دهنده شماره متعلق به فیلتربانک‌ها است.



(الف)



(ب)

شکل (۳–۵): الف: فیلتربانک مثلثی، ب: فیلتربانک گوسی

سپس با رابطه (۳-۸) kb ها که در فیلتر بانک مثلثی نقاط مرزی را مشخص می‌کنند و در فیلتر بانک گوسی در تعیین پارامتر سیگما کاربرد دارند و به بیان دیگر در تعیین پراکندگی

هر گوسی مشخص‌کننده هستند به دست می‌آیند.

$$kb_i = (i+1).\Delta_{mel} \quad (۸-۳)$$

پس از آن با رابطه (۳-۹)، σ_i را که واریانس هر زیر بخش از فیلتر بانک است محاسبه می‌کنیم.

$$\sigma_i = \frac{kb_i - kb_{i-1}}{2} \quad (۹-۳)$$

نهایتاً با رابطه (۳-۱۰) معادله نهایی فیلتربانک گوسی را و در نتیجه خود فیلتربانک‌ها را به دست می‌آوریم.

$$\Psi_i = e^{\frac{-(k-kb_i)}{2\sigma_i^2}} \quad (۱۰-۳)$$

پس از عبور سیگنال از فیلتربانک، همانند سایر الگوریتم‌ها طیف توان آن محاسبه و از آن لگاریتم گرفته می‌شود و نهایتاً با اعمال تبدیل کسینوسی AGMFCC به دست خواهد آمد.

علاوه بر روشی که در اینجا پیشنهاد شد روش‌های متنوع دیگری نیز وجود دارد که می‌توان به تنهایی و یا با ترکیب کردن با سایر روش‌ها، از آن‌ها به منظور مصون کردن سیستم در برابر نویز استفاده کرد.

۳-۱-۴ پیشنهاد چهارم، تبدیل Zak

یکی دیگر از ویژگی‌هایی که می‌توان از سیگنال گفتار لهجه دار استخراج کرد تبدیل Zak است. این تبدیل قادر است ضرایبی از سیگنال گفتار هم از حوزه زمان و هم از حوزه فرکانس استخراج کند. رابطه (۳-۱۱) نحوه محاسبه این تبدیل را نشان می‌دهد[34]

$$(Z\,f)_T(t,v) = T^{1/2} \sum_{k=-\infty}^{\infty} f(t+kT)e^{-2\pi jkvT} \quad (۱۱-۳)$$

در این رابطه f تابعی است که می‌خواهیم تبدیل Zak را به آن اعمال کنیم که در اینجا f

همان سیگنال گفتار است. T دوره تناوب سیگنال است. $0 \le t \le T$ و $0 \le v \le T^{-1}$ است. با اعمال این تبدیل به هر فریم سیگنال گفتار یک ماتریس در خروجی خواهیم داشت که شامل N سطر و M ستون خواهد بود. N تعداد ضرایب حاصل از تبدیل Zak در حوزه زمان و M تعداد ضرایب حاصل در حوزه فرکانس می‌باشد.

پس از بیان پیشنهادات مرحله استخراج ویژگی در بخش بعد به بیان پیشنهادهایی در مورد مرحله طبقه‌بندی پرداخته می‌شود.

۳-۲ پیشنهادات مرحله طبقه‌بندی

در زمینه استفاده از طبقه بندها در سیستم تشخیص لهجه خوب است به این نکته اشاره‌کنیم که همان طور که استخراج یک ویژگی مناسب و کارآمد برای بازدهی مطلوب این سیستم لازم است، به کار بردن یک طبقه بند مناسب نیز باعث افزایش کارایی سیستم تا حد مطلوبی خواهد شد. در این بخش، ابتدا یک طبقه بند جدید که تاکنون در سیستم‌های تشخیص لهجه استفاده نشده است و سپس پیشنهادی به منظور تغییر در یک طبقه بند برای افزایش بازدهی بیان خواهد شد.

۳-۲-۱ پیشنهاد اول، شبکه توابع بنیادی شعاعی (RBF)

RBF روشی برای تقریب توابع است و یادگیری با آن ارتباط نزدیکی با شبکه‌های عصبی مصنوعی دارد.

در این روش فرضیه یاد گرفته‌شده به صورت رابطه (۳-۱۲) می‌باشد.

$$\hat{f}(x) = w_0 + \sum_{u=1}^{k} w_u K_u\big(d(x_u, x)\big) \qquad (۳-۱۲)$$

w_0 نشان‌دهنده بردار بایاس و w_u بردار وزن‌های به‌روز شده است. x خروجی مطلوب، x_u

ورودی تابع و d فاصله بین ورودی و بردار نرون مخفی مطلوب است. و در نهایت $\hat{f}(x)$ خروجی شبکه RBF را به ازای هر ورودی نشان می‌دهد.در این روش از تعداد k تابع کرنل برای تقریب تابع استفاده می‌شود.تابع کرنل معمولاً به صورت تابع گوسی رابطه (۳-۱۳) با واریانس σ_u^2 انتخاب می‌شود.

$$K_u\big(d(x_u,x)\big) = e^{\frac{1}{2\sigma_u^2}d(x_u,x)^2} \qquad (۱۳-۳)$$

نشان داده‌شده است در صورتی که تعداد کافی تابع کرنل گوسی انتخاب شود با استفاده از این شبکه می‌توان هر تابعی را با خطای نسبتاً کمی تقریب زد. در صورت داشتن مجموعه‌ای از مثال‌های آموزشی ، آموزش RBF در دو مرحله صورت می‌گیرد:

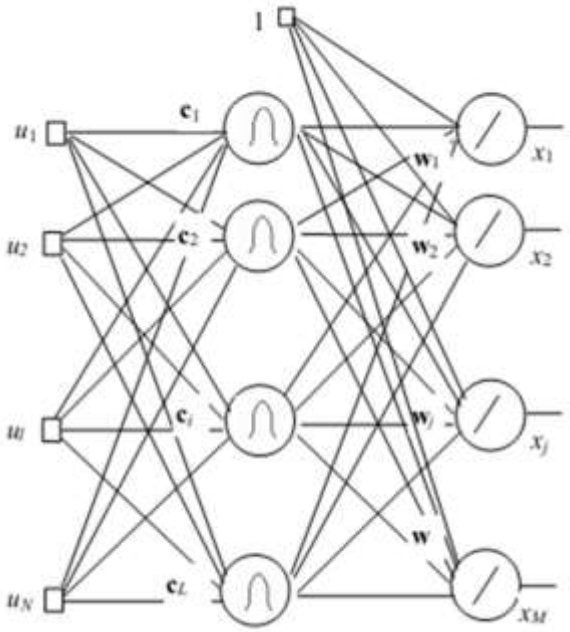
در مرحله اول، تعداد توابع کرنل انتخاب می‌شود. به عبارت دیگر با انتخاب مقداری برای K, مقادیر x_u و σ_u^2 برای هر تابع کرنل تعیین می‌گردد.

و در مرحله دوم، وزن‌های شبکه طوری انتخاب می‌شوند که شبکه با داده‌های آموزشی منطبق گردد. این‌کار با استفاده از رابطه خطای کلی (۳-۱۴) انجام می‌شود.

$$E(\vec{w}) = \frac{1}{2}\sum_{x\epsilon D}(f(x) - \hat{f}(x))^2 \qquad (۱۴-۳)$$

E نشان‌دهنده بردار خطا در هر مرحله است. به ازای هر مثال آموزشی وزن‌ها طوری حساب می‌شوند تا در خروجی شبکه رابطه $f(x) = \hat{f}(x)$ برقرار باشد. در نهایت شبکه RBF به طور کامل با مثال‌های آموزشی انطباق پیدا خواهد کرد.

شکل (۳–۶) ساختار استاندارد این شبکه را نشان می‌دهد.



شکل (۳–۶): ساختار استاندارد شبکه RBF

شبکه RBF یک تقریب کلی از تابع را با استفاده از مجموع تقریبات محلی محاسبه می-کند.

از مزایای این شبکه می‌توان به آموزش آسان‌تر آن نسبت به شبکه‌های عصبی معمولی، که از روش پس انتشار استفاده می‌کنند، نام برد. این شبکه با شبکه MLP تفاوت‌های ساختاری دارد که برخی از آن‌ها عبارت‌اند از:

- MLP یک یا چند لایه مخفی دارد ولی RBF تنها یک لایه مخفی دارد.
- تعداد عصب‌های شبکه RBF معمولاً بیشتر است.
- عدم حساسیت به ترتیب داده‌ها در آموزش شبکهRBF
- سرعت بالاتر آموزش در شبکه RBF
- MLP مانند یک جعبه سیاه است درحالی‌که RBF همانند یک جعبه سفید شفافیت عملکرد دارد.

۳-۲-۲ پیشنهاد دوم، افزایش کارایی طبقه بند SVM

یکی از طبقه بندهای مورد استفاده در این پایان‌نامه و بسیاری از تحقیقات دیگر طبقه بند SVM است. این طبقه بند به منظور جداسازی داده‌ها و طبقه‌بندی مطلوب نیاز به بردار ویژگی‌هایی دارد که اعداد این بردارها به طور آشکاری از هم متمایز و بزرگ نیز باشند تا مرز تشکیل شده بین هر کدام از بردارها به گونه‌ای باشد که قادر به جداسازی صحیح و دارای خطای کمتر بین بردار ویژگی‌ها باشد. با دانستن این نکته می‌توان با ضرب کردن یک ضریب مانند $\propto$ در بردارهای ویژگی به دست آمده آن‌ها را قبل از ورود به SVM به بردارهایی با اعداد مناسب‌تری تبدیل کنیم. آزمایش‌های انجام گرفته در این پایان‌نامه نشان می‌دهد که ضریب $\propto$ را تا یک حد معین می‌توان زیاد کرد تا بازدهی SVM افزایش یابد و بعد از آن دوباره نرخ بازشناسی توسط این طبقه بند نزول پیدا می‌کند.

۳-۲-۳ پیشنهاد سوم، ترکیب طبقه بندها

هدف نهایی طراحی یک سیستم شناسایی الگو رسیدن به بهترین عملکرد طبقه‌بندی برای مسئله موجود می‌باشد[11] . از آنجایی که هیچ طبقه‌بندی به طور کامل قادر نیست تمام مسائل را حل کند، ترکیب طبقه بندها به عنوان روشی برای کاهش خطای طبقه‌بندی و افزایش کارایی سیستم‌ها پیشنهاد شده است.

ایده اصلی این طرح، ایجاد طبقه بندهای گوناگون با ناحیه‌های خطای متفاوت است تا سیستم ترکیبی با بهره‌گیری از نقاط قوت تک‌تک طبقه بندها و رفع خطای ایجادشده‌ی یک طبقه بند، با اطلاعات به دست آمده از طریق طبقه بندهای دیگر، خطای کل سیستم را کاهش دهد. بنابراین مهم‌ترین نیاز ترکیب طبقه بندها ایجاد تعدادی طبقه بند با کارایی قابل‌قبول و مستقل در تصمیم‌گیری می‌باشد، در غیر این صورت طبقه بندها نتایج یکسانی برای تمام حالات خواهند داد که این به هیچ‌وجه برای ترکیب مفید نیست.[35-38] به این نکته نیز باید

توجه کرد که طبقه‌بندهایی که قرار است با یکدیگر تر کیب شوند باید دارای خطـای متفـاوت باشد.

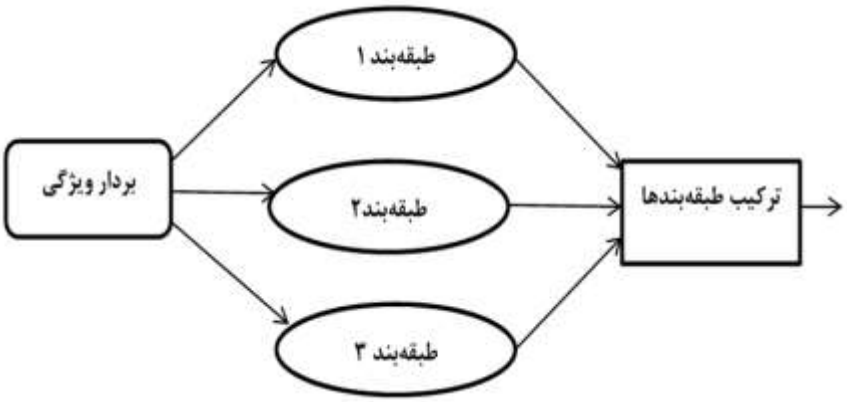
برای ترکیب طبقه بندها روش‌ها و طرح‌های مختلفی پیشنهاد شده است که از آن جمله می‌توان روش‌هـای سـری، مـوازی و ترکیبـی را نـام بـرد. در ایـن پایـان‌نامـه از روش مـوازی استفاده‌شده است. در این روش همه طبقه بندها به طور همزمان باهم ترکیب می‌شوند. یکی از ساده‌ترین فرآیندهایی که در مرحله ترکیب استفاده می‌شود، روش میانگین‌گیری است، بـدین صورت که نتیجه حاصل از تمام طبقه بندهای مورد نظر برای ترکیب کردن باهم جمع می‌شوند و سپس بر تعداد طبقه بندهای ترکیب‌شده تقسیم می‌شوند.

روش دیگر این است که برای هر کدام از طبقه بندها یـک وزن خـاص تعیـین کنـیم و سپس این وزن را در نتیجه نهایی هر کدام از طبقه بندها ضرب کنیم. وزن‌ها باید بـه گونـه‌ای باشند که مجموع وزن‌های به‌کاررفته برابر ۱ شود. بهینه‌ترین وزن ممکن برایی هر طبقه بند را از روش سعی و خطا به دست می‌آوریم. پس از ضرب کردن وزن در نظر گرفته‌شده بـرای هـر طبقه بند در طبقه بند مورد نظر، نتیجه حاصل را باهم جمع می‌کنیم و سپس نرخ بازشناسی را با ترکیب جدید به دست می‌آوریم.

در این پایان‌نامه از هر دو روش مذکور استفاده‌شـده اسـت. روش میـانگین‌گیـری بـرای زمانی که سه طبقه بند را باهم ترکیب کرده‌ایم و روش تعیین وزن هنگامی که دو طبقـه بنـد با یکدیگر ترکیب ترکیب‌شده است استفاده می‌شود.

در ترکیب طبقه بندها، بردار ویژگی که هر کدام از طبقه بندها با آن آموزش‌دیده‌اند هم می‌تواند متفاوت باشد و هم کاملاً مشابه هم باشند. در این پایان‌نامه ویژگی استخراج‌شده برای آموزش هر کدام از طبقه بندها باهم یکسان است.

شکل(۳-۷) روش استفاده‌شده در این پایان‌نامه برای ترکیب طبقه بندها را نشـان مـی-دهد.



شکل (۳–۷): روش استفاده‌شده برای ترکیب طبقه بندها

نتایج حاصل، نشان‌دهنده بهبود مطلوب نرخ بازشناسی با این روش هستند.

فصل چهارم:

نتایج

آزمایش‌های انجام‌شده

۴-نتایج آزمایش‌های انجام‌شده

در این فصل نتایج تعدادی آزمایش آورده شده است که با به کار بردن روش‌های قبلی و روش‌های پیشنهادی روی لهجه‌های گفتار زبان فارسی به دست آمده‌اند.

۴-۱ پایگاه داده استفاده‌شده

برای انجام آزمایش‌های مختلف نیاز به داشتن تعدادی جمله که دارای لهجه از پیش مشخص‌شده باشند داریم. به همین دلیل یا باید اقدام به تولید و جمع‌آوری یک پایگاه داده[1] شود و یا از پایگاه‌های داده‌ای که قبلاً تهیه‌شده است استفاده شود. در این پایان‌نامه ما از پایگاه داده FARSDAT که معتبرترین پایگاه دادگان گفتار زبان فارسی است استفاده کرده‌ایم. این پایگاه داده شامل ۶۰۸۰ جمله است که توسط ۳۰۴ نفر گوینده ایرانی با ده لهجه مختلف بیان‌شده است. از هر گوینده ۲۰ جمله ضبط گردیده است. در این پایگاه داده گویندگان از لحاظ سن، جنس، سطح تحصیلات و لهجه‌هایشان متمایز شده‌اند که شرایط خوبی را برای انجام تحقیقات مختلف فراهم کرده است. عملیات تولید این پایگاه داده در آزمایشگاه زبان-شناسی دانشگاه تهران انجام پذیرفته است.

علاوه بر پایگاه داده نام‌برده شده اخیراً یک پایگاه داده نیز مخصوص لهجه‌های مختلف زبان فارسی در دانشگاه صنعتی سهند تدوین شده است که با نام SAS شناخته می‌شود.

در تحقیقات خارجی نیز از یک پایگاه داده جامع مشابه FARSDAT استفاده می‌شود که در آن گویندگان در زمینه‌های مختلف از جمله سن و جنس و لهجه از هم تفکیک‌شده‌اند. نام این پایگاه داده TIMIT است.

در این پایان‌نامه با توجه به اینکه هدف ما تشخیص لهجه‌های زبان فارسی است، تعدادی

[1] Database

جمله لهجه دار از پایگاه داده FARSDAT انتخاب کرده‌ایم. تعداد لهجه‌های انتخاب‌شده در این تحقیق ۵ لهجه است، که سعی گردیده است طوری انتخاب شود که از لحاظ جغرافیایی اکثر مناطق کشورمان را شامل شود. به همین منظور لهجه‌های تهرانی، اصفهانی، ترکی، شمالی و جنوبی انتخاب‌شده‌اند. البته لازم به ذکر است که برخی از لهجه‌ها در این پایگاه داده از جمله لهجه یزدی و یا بلوچی، جامعه آماری پایینی داشتند که برای انجام آزمایش‌های مد نظر ما مطلوب نبودند و بنابراین در میان لهجه‌های انتخابی قرار نگرفتند. در انتخاب جملات تلاش شده است تا جملات انتخابی از طولانی‌ترین جملات این پایگاه داده باشند که دارای کلمات مختلفی از لحاظ آوایی می‌باشند.

۲-۴ نتایج آزمایش‌های انجام‌شده با روش‌های متداول

در این بخش با استخراج برخی از ویژگی‌های استفاده‌شده در تحقیقات قبلی از سیگنال-های گفتار تهیه‌شده برای این پایان‌نامه و همچنین به کار بردن طبقه بندهای مختلف، نتایج گوناگونی به دست آمده است که در جدول‌هایی که در ادامه می‌آیند قرارگرفته‌اند.

در آزمایش اول ویژگی‌های ضرایب مل-کپستروم، مشتق اول و دوم آن و فرکانس‌های فرمنت استخراج گردیده‌اند، سپس با تشکیل ۲ بردار ویژگی شامل «MFCC» و «$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$» ، ابتدا نرخ بازشناسی با طبقه بندهای MLP، PNN و KNN درحالی‌که لهجه‌ها را به صورت دوتا دوتا باهم مقایسه می‌کنیم به دست می-آوریم. منظور از 2F دو فرکانس فرمنت اول است. سپس با افزودن دو بردار ویژگی «$MFCC + \Delta MFCC$» و «$MFCC + \Delta MFCC + \Delta\Delta MFCC$» و با همان طبقه بندها نتایج را درحالی‌که که تمام لهجه‌ها به طور همزمان باهم مقایسه شده‌اند محاسبه می‌کنیم. لازم به ذکر است که در بیشتر آزمایش‌های انجام‌شده ۷۰ درصد داده‌ها را به عنوان آموزش و ۳۰ درصد را به عنوان آزمایش در نظر گرفته‌ایم و در برخی از آزمایش‌ها این مقادیر به ترتیب به ۳۵ و ۶۵

تغییر پیدا کرده است. آنچه که در جدول‌ها بیان‌شده است نتایج مربوط به مرحله آزمایش است. جدول (۲–۱) نرخ بازشناسی با ویژگی MFCC و طبقه بندهای KNN، MLP و PNN را نشان می‌دهد. در این آزمایش لهجه‌ها دوتا دوتا مقایسه شده‌اند. لازم به ذکر است که برای طبقه‌بنـد MLP برای ویژگی MFCC تعداد نرون‌های لایه مخفی برابر ۱۰ و برای سایر ویژگی‌ها برابر ۱۵ در نظر گرفته شده است.

جدول (۴–۱): نرخ بازشناسی لهجه‌ها با ویژگی MFCC و طبقه بندهای مختلف

لهجه	طبقه بند	MLP	PNN	KNN	میانگین
اصفهانی و تهرانی		80.05	**80.4**	30.46	63.63
تهرانی و ترکی		81.4	**83.3**	62.1	**75.6**
ترکی و اصفهانی		80.86	**81.65**	32.04	64.85
شمالی و جنوبی		**83.55**	82.34	46.26	70.71
تهرانی و جنوبی		81.26	**81.79**	35.91	66.32
تهرانی و شمالی		80.05	**80.55**	40.68	67.09
ترکی و جنوبی		**80.99**	80.96	35.03	65.66
ترکی و شمالی		**79.78**	78.62	34.9	64.43
اصفهانی و شمالی		81.13	**81.37**	55.27	72.59
اصفهانی و جنوبی		81.67	**82.34**	55.03	73.01

همانطور که اعداد و ارقام این جدول نشان می‌دهد به طور میانگین لهجه تهرانی و ترکی بهتر از یکدیگر تشخیص داده می‌شوند و در میان طبقه‌بندهای به کار رفتـه طبقـه‌بنـد PNN بالاترین و KNN کمترین بازدهی را دارد.

جـدول (۴–۲) نـرخ بازشناسـی لهجـه‌هـا را بـه صـورت دوتـایی بـه ازای ویژگـی «$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$» و طبقه بندهای مختلف نشان مـی‌دهـد. جـدول (۴–۳) به جای مقایسه لهجه‌ها به صورت دوتا دوتا ، میانگین نرخ بازشناسی تمام لهجه‌هـا بـه طور همزمـان و در حضـور ۴ بـردار ویژگـی «MFCC»، «$MFCC + \Delta MFCC$»، «$MFCC +$ $\Delta MFCC + \Delta\Delta MFCC$»و «$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$» و بـا طبقـه بنـدهای مختلف مقایسه شده‌اند. به عبارت دیگر در مرحله قبل طبقه بندها با ۲ کلاس روبـرو بـوده‌انـد ولی در این مرحله ۵ کلاس وارد طبقه بندها می‌شود و مسـلماً شـاهد کـاهش بـازدهی آن‌هـا خواهیم بود.

جدول (۴–۲): نرخ بازشناسی با ویژگی «$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$» وطبقه‌بندهای مختلف

لهجه	طبقه بند	MLP	PNN	KNN	میانگین
اصفهانی و تهرانی		77.94	**80.68**	32.16	63.59
تهرانی و ترکی		**83.15**	80.9	58.9	**74.31**
ترکی و اصفهانی		79.11	**79.77**	35.07	64.65
شمالی و جنوبی		**81.67**	80.91	47.39	69.99
تهرانی و جنوبی		**81.26**	80.34	33.13	64.91
تهرانی و شمالی		**82.2**	82.05	40.92	68.39
ترکی و جنوبی		**82.47**	81.48	37.57	67.17
ترکی و شمالی		**82.47**	81.71	39.55	67.91
اصفهانی و شمالی		**81.4**	81.14	54.98	72.50
اصفهانی و جنوبی		80.59	**81.82**	56	72.80

جدول (۴–۳): میانگین نرخ بازشناسی پنج لهجه مختلف در حضور ویژگی‌ها و طبقه بندهای مختلف

ویژگی	طبقه بند	MLP	PNN	KNN	میانگین
MFCC		55.04	71.41	**22.07**	49.5
$MFCC + \Delta MFCC$		68.09	70.08	21.11	53.09
$MFCC + \Delta MFCC + \Delta\Delta MFCC$		**73.46**	71.45	19.53	54.81
$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$		71.86	**73.38**	19.53	**54.92**
میانگین		67.11	**71.58**	20.56	

همان طور که نتایج جدول‌های بالا نشان می‌دهد در میان لهجه‌های مختلف دو لهجه تهرانی و ترکی به صورت بهتری از یکدیگر تشخیص داده می‌شوند. همچنین در میان طبقه بندهای مختلف، طبقه بند PNN دارای میانگین عملکرد بالاتری نسبت به دو طبقه بند KNN و MLP است. در میان بردار ویژگی‌های مختلف، نیز بردار ویژگی «$MFCC + \Delta MFCC +$ $\Delta\Delta MFCC + 2F$» میانگین عملکرد بهتر و مناسب‌تری است. نتیجه دیگری که می‌توان به دست آورد این است که طبقه بند KNN در حضور دو کلاس، عملکرد نسبتاً خوبی دارد اما با افزایش تعداد کلاس‌ها عملکرد این طبقه بند تا حد زیادی کاهش پیدا می‌کند و با استفاده از آن نتایج قابل‌قبولی حاصل نخواهد شد.

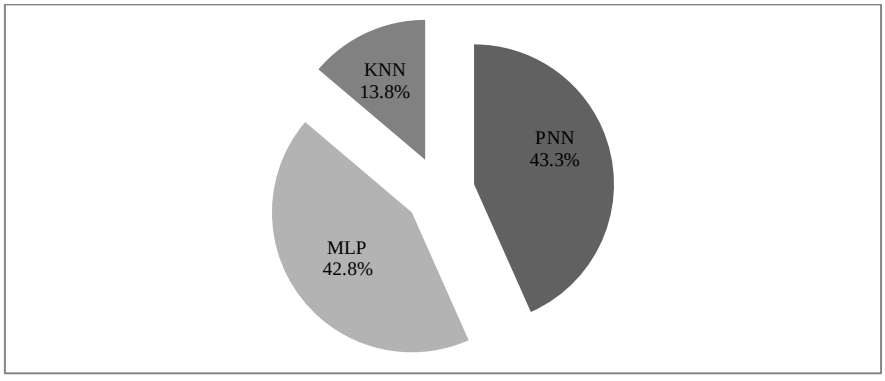
به منظور روشن‌تر شدن عملکرد طبقه بندها در مورد لهجه‌های مختلف و در حضور بردارهای ویژگی متفاوت جدول (۴-۴) رسم شده است. در این جدول با توجه به نتایج جدول (۴–۳) برای بردار ویژگی که به ازای آن طبقه بند مربوطه بهترین عملکرد را داشته است نرخ بازشناسی هر کدام از لهجه‌ها آورده شده است.

جدول (۴–۴): نرخ بازشناسی هر کدام از لهجه‌ها با طبقه بندهای مختلف با روش‌های قبلی

لهجه	طبقه بند	MLP	PNN	KNN	میانگین
تهرانی		**72.31**	71.96	13.35	52.54
ترکی		63.33	**70.73**	22.89	52.31
اصفهانی		**73.16**	71.38	19.17	54.57
شمالی		**74.96**	69.8	15.54	53.43
جنوبی		98.34	**99.41**	31.8	**76.51**

همان طور که نتایج این جدول نشان می‌دهد در میان تمامی لهجه‌ها، لهجه جنـوبی در حضور طبقه بندهای مختلف نرخ تشخیص بالاتری را به خود اختصاص داده است و بعـد از آن به ترتیب لهجه‌های اصفهانی، شمالی، تهرانی و سپس ترکی قرار دارد که این مسئله مربوط بـه وضوح لهجه‌ها در جملات بیان‌شده است.

برای روشن‌تر شدن عملکرد سه طبقه بند به‌کاررفته در تشخیص لهجه‌هـا یـک نمـودار دایره‌ای رسم شده است که در آن درصد تأثیر هر کدام از طبقه بنـدها در لهجـه جنـوبی کـه بیشترین نرخ بازشناسی را دارد به نمایش گذاشته شده است. شکل (۴–۱) این نمودار را نشان می‌دهد.



شکل (۴–۱): نمودار دایره‌ای مقایسه عملکرد سه طبقه بند MLP، PNN و KNN

این نمودار دایره‌ای نیز نتیجه حاصل از جدول (۴–۳)، یعنی عدم عملکرد مناسب طبقه بند KNN را تأیید می‌کند. علاوه بر جدول‌هایی که بیان شد برای اینکه ببینیم در تشخیص لهجه‌ها هر کدام از لهجه‌ها با کدام یک از لهجه‌های دیگر به طور نادرست تشخیص داده می‌شود، ماتریسی به نام سردرگمی نتایج رسم می‌شود. در این ماتریس جدول مانند، قطر ماتریس نشان‌دهنده نرخ بازشناسی مطلوب ماست که در بهترین حالت باید اعداد روی قطر اصلی ماتریس ۱۰۰ و سایر اعداد آن صفر باشد. جدول (۴–۵) ماتریس سردرگمی را برای طبقه بند PNN نشان می‌دهد.

جدول (۴–۵): ماتریس سردرگمی نتایج برای طبقه بند PNN با روش‌های متداول (اعداد به درصد)

	تهرانی	ترکی	اصفهانی	شمالی	جنوبی
تهرانی	**71.96**	25.81	1.19	0.68	0.34
ترکی	16	**70.73**	12.42	0.15	0.68
اصفهانی	0.16	13.97	**71.38**	13.39	1.08
شمالی	0	0.17	15.64	**69.8**	14.39
جنوبی	0.09	0.09	0.19	0.19	**99.41**

جدول (۴-۶) ماتریس سردرگمی نتایج را برای طبقه بند MLP نشان می‌دهد.

جدول (۴–۶): ماتریس سردرگمی نتایج به ازای طبقه بند MLP با روش‌های متداول (اعداد به درصد)

	تهرانی	ترکی	اصفهانی	شمالی	جنوبی
تهرانی	**72.31**	26.49	0.71	0.23	0.23
ترکی	19.24	**63.32**	16.8	0.24	0.36
اصفهانی	0.24	11.19	**73.16**	14.56	0.84
شمالی	0	0.14	13.02	**74.96**	11.88
جنوبی	0.44	0	0.58	0.58	**98.38**

جدول (۴–۷) نشان‌دهنده ماتریس سردرگمی نتایج برای طبقه بند KNN نشان می‌دهد.

جدول (۴–۷): ماتریس سردرگمی نتایج با طبقه بند KNN با روش‌های متداول (اعداد به درصد)

	تهرانی	ترکی	اصفهانی	شمالی	جنوبی
تهرانی	**13.35**	23.56	24.48	18.55	20.03
ترکی	11.54	**22.89**	22.36	22.1	21.09
اصفهانی	17.8	8.21	**19.17**	47.94	6.84
شمالی	14.23	21.74	18.72	**15.54**	29.77
جنوبی	8.31	15.38	22.24	22.24	**31.8**

از آنچه که از ماتریس‌های سردرگمی نتایج، به دست می‌آید نکات زیر قابل توجه است:

- لهجه تهرانی بیشتر با لهجه ترکی اشتباه گرفته می‌شود.
- لهجه ترکی بیشتر با لهجه تهرانی اشتباه گرفته می‌شود.
- لهجه اصفهانی با لهجه‌های ترکی و شمالی اشتباه گرفته می‌شود.

- لهجه شمالی با لهجه‌های اصفهانی و جنوبی اشتباه گرفته می‌شود.
- لهجه جنوبی دارای کمترین خطا است و اگر هم مقداری خطا رخ دهد بیشتر با لهجه-های اصفهانی و شمالی اشتباه گرفته می‌شود.
- طبقه بند KNN آن‌قدر عملکرد ضعیفی دارد که در برخی حالات، اعداد قرارگرفته روی قطر ماتریس سردرگمی که باید بیشترین مقادیر را داشته باشند بسیار کمتر از اعداد قسمت‌های دیگر ماتریس هستند.

در ادامه نتایج حاصل از روش‌های پیشنهادی را بررسی می‌کنیم.

۴-۳ نتایج آزمایش‌های انجام‌شده با روش‌های پیشنهادی

در این بخش با اجرای روش‌های بیان‌شده در فصل سوم این پایان‌نامه بر روی پایگاه داده موجود، نرخ بازشناسی به ازای لهجه‌های مختلف به دست آمده است.

۴-۳-۱ نتایج استخراج ویژگی‌های پیشنهادی برای محیط‌های معمولی

اولین پیشنهادات در حوزه استخراج ویژگی بود. ابتدا به ازای ویژگی‌های SCM و مشتق اول آن و تبدیل Zak و سه طبقه بند KNN، MLP و PNN نرخ بازشناسی را در حالتی که لهجه‌ها را دوتا دوتا مقایسه می‌کنیم به دست می‌آوریم.

جدول (۴-۸) نتایج را به ازای بردار ویژگی «SCM» نشان می‌دهد.

جدول (۴–۸): نتایج نرخ بازشناسی به ازای ویژگی SCM و طبقه بندهای مختلف

لهجه	طبقه بند	MLP	PNN	KNN	میانگین
اصفهانی و تهرانی		76.47	**83.31**	52.56	70.78
تهرانی و ترکی		78.98	**80.41**	53.93	71.10
ترکی و اصفهانی		77.93	**79.86**	55.43	71.07
شمالی و جنوبی		79.38	**80.27**	55.15	71.6
تهرانی و جنوبی		73.24	**80.41**	54.70	69.45
تهرانی و شمالی		73.80	**81.79**	56.28	70.62
ترکی و جنوبی		78.25	**79.86**	54.14	70.75
ترکی و شمالی		76.07	**80.68**	53.57	70.10
اصفهانی و شمالی		77.44	**83.03**	59.59	**73.35**
اصفهانی و جنوبی		74.69	**81.79**	59.27	71.91

آنچه که این جدول نشان می‌دهد این است که دو لهجه اصفهانی و شمالی به طور میانگین بهتر تشخیص داده می‌شوند و طبقه‌بندPNN بالاترین و KNN کمترین بازدهی را دارد.

جدول (۴-۹) نتایج نرخ بازشناسی را برای ویژگی «$SCM + \Delta SCM$» نشان می‌دهد.

جدول (۴–۹): نتایج حاصل از به‌کارگیری ویژگی «$SCM + \Delta SCM$» با طبقه بندهای مختلف

لهجه	طبقه بند	MLP	PNN	KNN	میانگین
اصفهانی و تهرانی		77.68	**81.24**	49.69	69.53
تهرانی و ترکی		76.63	**79.58**	53.05	69.75
ترکی و اصفهانی		80.43	**81.51**	58.70	73.54
شمالی و جنوبی		80.19	**82.89**	55.95	73.01
تهرانی و جنوبی		**79.70**	79.31	48.08	69.03
تهرانی و شمالی		79.30	**79.44**	49.09	69.27
ترکی و جنوبی		79.95	**81.79**	57.05	72.93
ترکی و شمالی		79.06	**80.82**	56.52	72.63
اصفهانی و شمالی		77.36	**80.41**	58.58	72.11
اصفهانی و جنوبی		79.62	**82.2**	60.12	**73.98**

در نهایت با اعمال تبدیل Zak روی سیگنال‌های گفتار دارای لهجه نرخ بازشناسی را درحالی‌که لهجه‌ها به صورت دوتا دوتا مقایسه می‌شوند به دست می‌آوریم. با اعمال این تبدیل به هر فریم سیگنال گفتار با توجه به اینکه طول هر فریم تقریباً ۲۵ میلی‌ثانیه و بدون همپوشانی است، ماتریسی در خروجی با ۵۰ سطر و ۵۰ ستون خواهیم داشت که سطرها ویژگی‌هایی در حوزه زمان و ستون‌ها ویژگی‌هایی در حوزه فرکانس هستند. نتایج حاصل از این آزمایش در جدول (۴-۱۰) نشان داده‌شده است.

جدول (۴–۱۰): نرخ بازشناسی به وسیله ضرایب حاصل از تبدیل Zak

لهجه	طبقه بند	MLP	PNN	KNN	میانگین
اصفهانی و تهرانی		72.51	**88.80**	55.94	72.41
تهرانی و ترکی		69.98	**90.18**	58.33	72.83
ترکی و اصفهانی		74.85	**89.26**	59.35	**74.48**
شمالی و جنوبی		73	**89.41**	53.07	71.82
تهرانی و جنوبی		75.24	**88.95**	52.82	72.33
تهرانی و شمالی		67.34	**89.57**	53.21	70.04
ترکی و جنوبی		74.26	**89.72**	58.23	74.07
ترکی و شمالی		71.15	**91.87**	58.47	73.83
اصفهانی و شمالی		73	**90.49**	58.67	74.05
اصفهانی و جنوبی		73.09	**90.33**	57.40	73.60

همانطور که مقادیر این جدول نشان می‌دهد ویژگی تبـدیل Zak در حضـور طبقـه‌بنـد PNN بازدهی خوبی دارد. پس از مقایسه لهجه‌ها به صورت دوتایی، در آزمایش بعدی میانگین نرخ بازشناسی تمامی لهجه‌ها را در حضور ویژگی‌های پیشنهادی و طبقه بندهای مختلـف بـه دست می‌آوریم. جدول (۴–۱۱) نتایج حاصل از این آزمایش را نشان می‌دهد.

جدول (۴–۱۱): میانگین نرخ بازشناسی پنج لهجه در حضور ویژگی‌های پیشنهادی و طبقه بندهای مختلف

ویژگی	طبقه بند	MLP	PNN	KNN	میانگین
SCM		53.42	69.13	26.93	49.82
$SCM + \Delta SCM$		64.08	69.42	26.33	**53.27**
Zak Transform		45.6	71.54	27.09	48.07
میانگین		54.36	**70.03**	26.78	

در جدول (۴–۱۲) نرخ تشخیص هر کدام از لهجه‌ها به طور جداگانه آمده است. مقـادیر این جدول بر اساس بالاترین نرخ بازشناسی که برای هر طبقه بند و به ازای یک بردار ویژگـی خاص به دست آمده است، محاسبه و لحاظ شده‌اند.

جدول (۴–۱۲): نرخ بازشناسی هر کدام از لهجه‌ها با طبقه بندهای مختلف با روش‌های پیشنهادی

لهجه	طبقه بند	MLP	PNN	KNN	میانگین
تهرانی		43.72	**71.22**	26.74	47.22
ترکی		**68.15**	67.89	31.8	55.94
اصفهانی		**77.51**	75.53	30.14	**61.06**
شمالی		55.20	**70.62**	31.31	52.37
جنوبی		**81.48**	74.91	23.37	59.92

برای روشن‌تر شدن عملکرد طبقه بندها با لهجه‌های مختلف در جدول (۴–۱۳)ماتریس سردرگمی نتایج به ازای طبقه بند PNN نشان داده‌شده است.

جدول (۴–۱۳): ماتریس سردرگمی نتایج برای طبقه بند PNN با روش پیشنهادی (اعداد به درصد)

	تهرانی	ترکی	اصفهانی	شمالی	جنوبی
تهرانی	**71.22**	4.25	6.37	7.30	10.73
ترکی	6.80	**67.89**	7.75	8.16	9.38
اصفهانی	7.06	4.10	**75.53**	6.89	6.40
شمالی	7.65	6.73	5	**70.62**	10
جنوبی	7.62	5.93	5.76	5.72	**74.91**

جدول (۴–۱۴) ماتریس سردرگمی نتایج را برای طبقه بند MLP نشان می‌دهد.

جدول (۴–۱۴): ماتریس سردرگمی نتایج برای طبقه بند MLP با روش پیشنهادی (اعداد به درصد)

	تهرانی	ترکی	اصفهانی	شمالی	جنوبی
تهرانی	**43.72**	34.18	8.83	10.46	2.79
ترکی	7.58	**68.15**	21.26	2.73	0.24
اصفهانی	0.25	14.68	**77.51**	11.62	1.91
شمالی	0.12	0.63	23.25	**55.20**	15.78
جنوبی	0.13	3.78	7.29	7.29	**81.48**

جدول (۴–۱۵) ماتریس سردرگمی نتایج را در حضور طبقه بند KNN نشان می‌دهد.

جدول (۴–۱۵): ماتریس سردرگمی نتایج برای طبقه بند KNN با روش پیشنهادی (اعداد به درصد)

	تهرانی	ترکی	اصفهانی	شمالی	جنوبی
تهرانی	**26.74**	17.24	16.02	18.54	21.43
ترکی	15.81	**31.8**	15.81	20.91	15.64
اصفهانی	14.90	18.98	**30.14**	17.98	17.98
شمالی	24.21	15.52	15	**31.31**	13.94
جنوبی	20.15	19.72	18.37	18.37	**23.37**

همان طور که از ماتریس‌های سردرگمی نتایج مشاهده می‌شود تقریباً همان نکاتی که در مورد ویژگی‌های متداول بیان شد، درباره ویژگی‌های پیشنهادی نیز البته با مقداری تفاوت صدق می‌کند. یکی از تفاوت‌هایی که وجود دارد این است که طبقه بند KNN در حضور ویژگی پیشنهادی عملکرد بهتری دارد. اعداد قرارگرفته روی قطر اصلی ماتریس مربوط به این طبقه

بند این موضوع را به طور واضح‌تری نشان می‌دهند.

برای انجام یک مقایسه بین کارایی ویژگی‌های متداول و ویژگی‌های پیشنهادی در مورد تشخیص لهجه‌ها چند جدول رسم شده است. در جدول (۴-۱۶) ستون میانگین جدول‌هایی که در آن‌ها لهجه‌ها به صورت دوتایی مقایسه شده‌اند در کنار یکدیگر قرارگرفته‌اند.

جدول (۴–۱۶): مقایسه کارایی ویژگی‌های متداول و پیشنهادی در تشخیص لهجه‌ها به صورت دوتایی

لهجه	طبقه بند	میانگین جدول (۴-۱) (متداول)	میانگین جدول (۴-۲) (متداول)	میانگین جدول(۴-۸) (پیشنهادی)	میانگین جدول (۴-۹) (پیشنهادی)	میانگین جدول (۴-۱۰) (پیشنهادی)
اصفهانی و تهرانی		63.63	63.59	70.78	69.53	**72.41**
تهرانی و ترکی		**75.6**	74.31	71.10	69.75	72.83
ترکی و اصفهانی		64.85	64.65	71.07	73.54	**74.48**
شمالی و جنوبی		70.71	69.99	71.6	**73.01**	71.82
تهرانی و جنوبی		66.32	64.91	69.45	69.03	**72.33**
تهرانی و شمالی		67.09	68.39	**70.62**	69.27	70.04
ترکی و جنوبی		65.66	67.17	70.75	72.93	**74.07**
ترکی و شمالی		64.43	67.91	70.10	72.63	**73.83**
اصفهانی و شمالی		72.59	72.50	73.35	72.11	**74.05**
اصفهانی و جنوبی		73.01	72.80	71.91	**73.98**	73.60

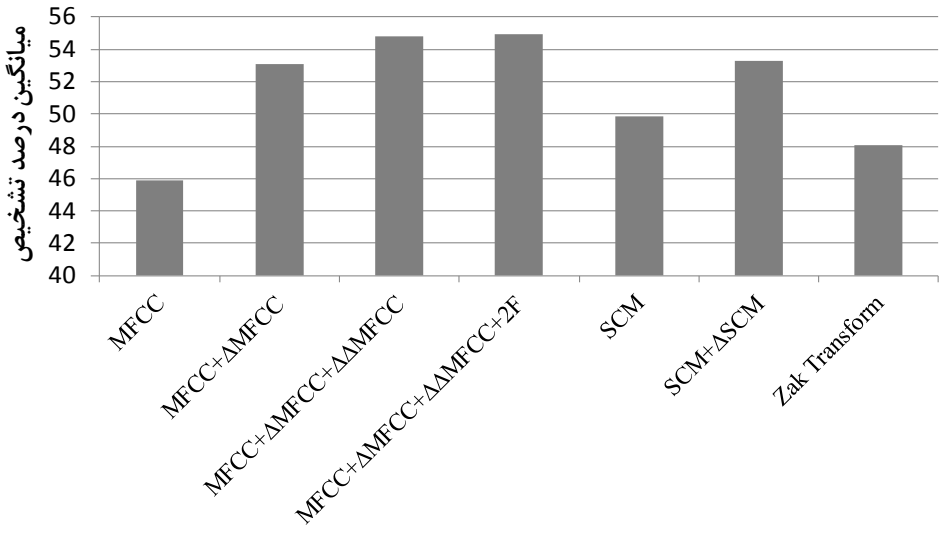
همان طور که مقادیرجدول (۴–۱۶) نشان می‌دهند غیر از دو لهجه تهرانی و ترکی که ویژگی‌های متداول در مورد تشخیص آن‌ها عملکرد بهتری دارند ، سایر لهجه‌ها در حضور ویژگی‌های پیشنهادی به طور مطلوب‌تری بازشناسی می‌شوند که این نکته حاکی از کارایی بهتر ویژگی‌های پیشنهادی است.

درجدول (۴–۱۷) میانگین عملکرد ویژگی‌های مختلف متداول و پیشنهادی در حضور تمامی لهجه‌ها مقایسه شده است. در این قسمت تعداد کلاس‌ها از دو کلاس به پنج کلاس افزایش یافته است.

جدول (۴–۱۷): مقایسه عملکرد ویژگی‌های متداول و پیشنهادی در حضور پنج لهجه مختلف

ویژگی	میانگین
MFCC (متداول)	49.5
$MFCC + \Delta MFCC$ (متداول)	53.09
$MFCC + \Delta MFCC + \Delta\Delta MFCC$ (متداول)	54.81
$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$ (متداول)	**54.92**
SCM (پیشنهادی)	49.82
$SCM + \Delta SCM$ (پیشنهادی)	**53.27**
Zak Transform (پیشنهادی)	48.07

مقادیر جدول نشان می‌دهند که ویژگی‌های متداول در حالتی که پنج کلاس داریم عملکرد مناسب‌تری نسبت به ویژگی‌های پیشنهادی دارند. هر چند که بازدهی بردار ویژگی پیشنهادی $SCM + \Delta SCM$ نزدیک به بازدهی و عملکرد ویژگی‌های متداول است. برای مقایسه بهتر و نمایان‌تر مقادیر این جدول، نمودار ستونی شکل (۴–۲) رسم شده است.



شکل (۴–۲): نمودار ستونی مقایسه عملکرد ویژگی‌های متداول و پیشنهادی

۴-۳-۲نتایج استفاده از پیشنهادات مرحله طبقه‌بندی

در ادامه آزمایش‌ها به سراغ طبقه بندها و پیشنهادات مربوط به آن‌ها می‌رویم. با توجه به فصل سوم این پایان‌نامه، در حوزه طبقه‌بندی، شبکه RBF به عنوان یک طبقه بند جدیدی که تاکنون در تشخیص لهجه‌ها استفاده نشده است پیشنهاد شد. همچنین پیشنهادهایی به منظور بهبود عملکرد ماشین بردار پشتیبان (SVM) با ضرب کردن ضریبی مانند $\propto$ در بردارهای ویژگی ارائه شد. لازم به ذکر است در این پایان‌نامه با توجه به اینکه تعداد کلاس‌ها از دوتا بیشتر است از برنامه LIBSVM استفاده‌شده است برای اجرای طبقه بند SVM استفاده‌شده است.

جدول (۴-۱۸) نتایج حاصل از نرخ بازشناسی لهجه‌ها را به صورت دوتایی و با ویژگی MFCC نشان می‌دهد.

جدول (۴–۱۸): نرخ بازشناسی لهجه‌ها با طبقه بندهای پیشنهادی و ویژگی MFCC

لهجه	طبقه بند	RBF	SVM ($\propto=1$)	$\propto=2$	$\propto=3$	$\propto=4$	$\propto=5$	$\propto=10$	$\propto=30$	میانگین
اصفهانی و تهرانی		72.21	65.16	71.09	**80.12**	79.87	79.74	79.61	79.61	74.33
تهرانی و ترکی		74.95	68.25	75.09	81.16	82.45	82.32	**82.96**	82.83	77.12
ترکی و اصفهانی		70.59	66.45	74.32	80	80.90	81.41	81.41	**81.54**	74.29
شمالی و جنوبی		**76.57**	66.83	75.48	80.77	**82.19**	**82.19**	81.67	81.67	**77.62**
تهرانی و جنوبی		71.24	66.45	74.19	80.25	81.67	81.80	82.06	**82.19**	74.8
تهرانی و شمالی		69.79	65.16	72	78.58	**79.61**	79.35	79.35	79.22	72.98
ترکی و جنوبی		74.31	67.48	73.67	78.32	82.19	82.7	82.96	**83.09**	76.47
ترکی و شمالی		72.21	62.96	74.19	78.19	79.09	80	80.51	**80.77**	74.37
اصفهانی و شمالی		75.12	65.54	76.64	80.51	81.54	**81.80**	**81.80**	**81.80**	76.81
اصفهانی و جنوبی		71.08	67.87	78.19	81.54	**82.58**	**82.58**	82.19	82.32	75.34

از اعداد و ارقام این جدول متوجه می‌شویم که دو لهجه شمالی و جنوبی به طور میانگین بهتر تشخیص داده می‌شوند و با افزایش ضریب $\propto$ شاهد بهبود نرخ بازشناسی هستیم. برای این حالت که دو لهجه داریم عملکرد طبقه‌بندهای پیشنهادی بهتر است.

در جدول (۴-۱۹) عملکرد طبقه بندهای پیشنهادی به ازای یکی از ویژگی‌های پیشنهادی یعنی «$SCM + \Delta SCM$» نشان داده‌شده است.

جدول (۴–۱۹): نرخ بازشناسی لهجه‌ها با طبقه بندهای پیشنهادی و ویژگی پیشنهادی $SCM + \Delta SCM$

لهجه	طبقه بند	RBF	SVM ($\propto = 1$)	$\propto = 3$	$\propto = 4$	$\propto = 5$	$\propto = 10$	$\propto = 15$	$\propto = 30$	میانگین
اصفهانی و تهرانی		74.63	65.41	69.41	73.41	75.61	82.45	**83.48**	**83.48**	75.4
تهرانی و ترکی		74.95	68.12	73.16	77.80	80.25	**81.80**	**81.80**	**81.80**	76.38
ترکی و اصفهانی		77.38	64.25	69.54	75.09	78.58	80.90	81.03	**81.16**	76.58
شمالی و جنوبی		**80.45**	64.25	70.32	76.25	81.54	**81.67**	81.29	81.29	**78.55**
تهرانی و جنوبی		77.70	64.12	68	72.64	75.74	81.93	**82.06**	**82.06**	76.46
تهرانی و شمالی		73.18	65.41	68.25	75.22	75.34	**80**	**80**	**80**	74.03
ترکی و جنوبی		75.44	64.90	73.29	79.22	82.7	**82.83**	**82.83**	**82.83**	76.9
ترکی و شمالی		73.82	66.83	71.48	77.29	80.64	82.58	**82.83**	**82.83**	75.8
اصفهانی و شمالی		76.57	63.22	69.16	76.64	74.96	80.64	**81.29**	**81.29**	75.94
اصفهانی و جنوبی		68.33	66.32	73.29	76.51	78.45	78.70	**78.83**	**78.83**	72.08

ضرایب $\propto$ به روش سعی و خطا انتخاب می‌شوند. به عنوان مثال در جدول (۴-۱۸)، چون به ازای $\propto = 2$ تغییر زیادی در نرخ بازشناسی صورت می‌گیرد آن را در جدول قرار داده-ایم، اما در جدول (۴-۱۹) چون به ازای $\propto = 2$ تغییر زیادی نسبت به $\propto = 1$ رخ نمی‌دهد آن را قرار نداده‌ایم، اما به جای آن ضریب $\propto = 15$ را چون باعث ایجاد تغییر زیاد می‌شود در جدول قرار داده‌ایم.

برای مقایسه عملکرد طبقه بندهای پیشنهادی با طبقه بندهای متداول ، ستون میانگین جدول‌های به دست آمده از مراحل قبل را کنار هم قرار می‌دهیم. جدول (۴–۲۰) این مقایسه را نشان می‌دهد. لازم به ذکر است در این جدول به جای کلمه طبقه بند از حروف مختصر «ط.ب» استفاده‌شده است.

جدول (۴–۲۰): مقایسه میانگین عملکرد طبقه بندهای متداول و پیشنهادی با حضور دو لهجه و ویژگی‌های مختلف

لهجه	طبقه بند	«ط.ب» متداول MFCC	«ط.ب» پیشنهادی MFCC	«ط.ب» متداول $SCM + \Delta SCM$	«ط.ب» پیشنهادی $SCM + \Delta SCM$
اصفهانی و تهرانی		63.63	74.33	69.53	**75.4**
تهرانی و ترکی		75.6	77.12	69.75	**76.38**
ترکی و اصفهانی		64.85	74.29	73.54	**76.58**
شمالی و جنوبی		70.71	77.62	73.01	**78.55**
تهرانی و جنوبی		66.32	74.8	69.03	**76.46**
تهرانی و شمالی		67.09	72.98	69.27	**74.03**
ترکی و جنوبی		65.66	76.47	72.93	**76.9**
ترکی و شمالی		64.43	74.37	72.63	**75.8**
اصفهانی و شمالی		72.59	**76.81**	72.11	75.94
اصفهانی و جنوبی		73.01	**75.34**	73.98	72.08

همان طور که نتایج جدول (۴–۲۰) نشان می‌دهند طبقه بندهای پیشنهادی در حضور یک بردار ویژگی پیشنهادی یعنی «$SCM + \Delta SCM$» بهترین عملکرد را در تشخیص لهجه‌ها به صورت دوتایی دارند. در دو موردی نیز که یکی از بردار ویژگی‌های قبلی یعنی «MFCC» نرخ بالاتری دارد در حضور طبقه بندهای پیشنهادی بوده است.

در آزمایش بعدی بازدهی طبقه بندهای پیشنهادی را در حضور همه لهجه‌های مورد بررسی در این پایان‌نامه اندازه می‌گیریم.

جدول (۴–۲۱) نرخ بازشناسی پنج لهجه را در حضور طبقه بندهای پیشنهادی و ویژگی‌ها متداول نشان می‌دهد.

جدول (۴–۲۱): میانگین بازدهی طبقه بندهای پیشنهادی در حضور ویژگی‌های متداول با پنج لهجه

ویژگی / طبقه بند	RBF	SVM $\propto = 1$	$\propto = 2$	$\propto = 4$	$\propto = 5$	$\propto = 10$	$\propto = 15$	$\propto = 30$	میانگین
MFCC	43.07	34.94	55.79	74.31	**74.85**	74.76	74.80	74.45	63.33
$MFCC + \Delta MFCC$	38.33	39.10	62.63	75.25	75.21	**75.52**	74.89	74.72	64.45
$MFCC + \Delta MFCC + \Delta\Delta MFCC$	**68.37**	46.39	69.78	**75.66**	**75.66**	75.52	75.30	75.30	**70.24**
$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$	66.88	43.75	68.72	**74.36**	74.04	73.10	71.85	71.67	68.04

برای بررسی عملکرد ویژگی‌های پیشنهادی با طبقه بندهای پیشنهادی به جای ضـریب $\propto = 2$ ، ضریب $\propto = 20$ را قرار می‌دهیم که این با روش سعی و خطا به دست می‌آید. جدول (۴–۲۲) نتایج مربوطه را نشان می‌دهد.

جدول (۴–۲۲): میانگین بازدهی طبقه بندهای پیشنهادی در حضور ویژگی‌های پیشنهادی با پنج لهجه

ویژگی / طبقه بند	RBF	SVM $\propto = 1$	$\propto = 4$	$\propto = 5$	$\propto = 10$	$\propto = 15$	$\propto = 20$	$\propto = 30$	میانگین
SCM	36.04	28.32	47.06	58.12	73.69	**73.95**	73.91	73.73	61.47
$SCM + \Delta SCM$	43.51	31	60.26	67.65	74.45	**74.58**	74.40	74.22	**62.5**
Zak Transform	**54.16**	27.11	40.24	43.61	60.06	69.78	**74.29**	69.87	54.87

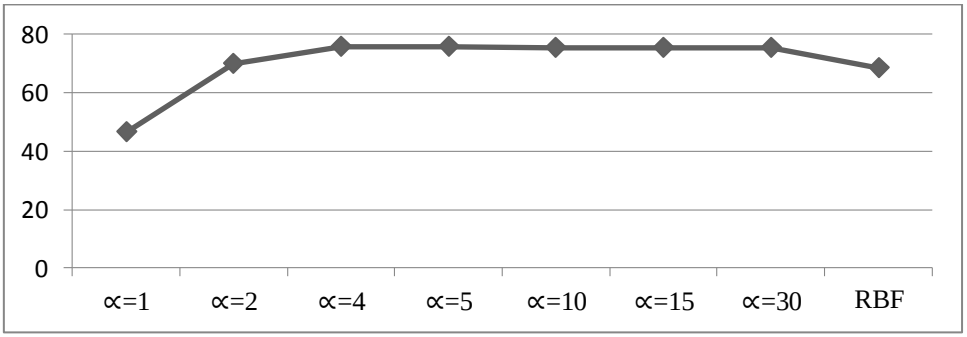
برای مقایسه عملکرد طبقه بندهای پیشنهادی با طبقه بندهای متداول، ستون میانگین جدول‌های قبل را در کنار هم قرار می‌دهیم. جدول (۴-۲۳) این مقایسه را نشان می‌دهد.

جدول (۴–۲۳): مقایسه عملکرد طبقه بندهای پیشنهادی با متداول در حضور ویژگی‌های مختلف

ویژگی	میانگین «ط.ب» قبلی	میانگین «ط.ب» پیشنهادی
MFCC	49.5	63.33
$MFCC + \Delta MFCC$	53.09	64.45
$MFCC + \Delta MFCC + \Delta\Delta MFCC$	54.81	**70.24**
$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$	**54.92**	68.04
SCM(پیشنهادی)	49.82	61.47
$SCM + \Delta SCM$(پیشنهادی)	**53.27**	**62.5**
Zak Transform(پیشنهادی)	48.07	54.87

همان طور که جدول‌های نتایج نشان می‌دهند طبقه بند SVM بـا افـزایش ضـریب $\propto$ عملکردش بهبود می‌یابد به طوری که گاهی افزایش نرخ بازشناسی به ۴۰ درصد نیز می‌رسـد. البته افزایش این ضریب تا یک حدی می‌تواند انجام بشود و بعد از آن شاهد کاهش عملکرد آن خواهیم بود. برای نمایش بهتر این نکته دو نمودار خطی رسم شده است.

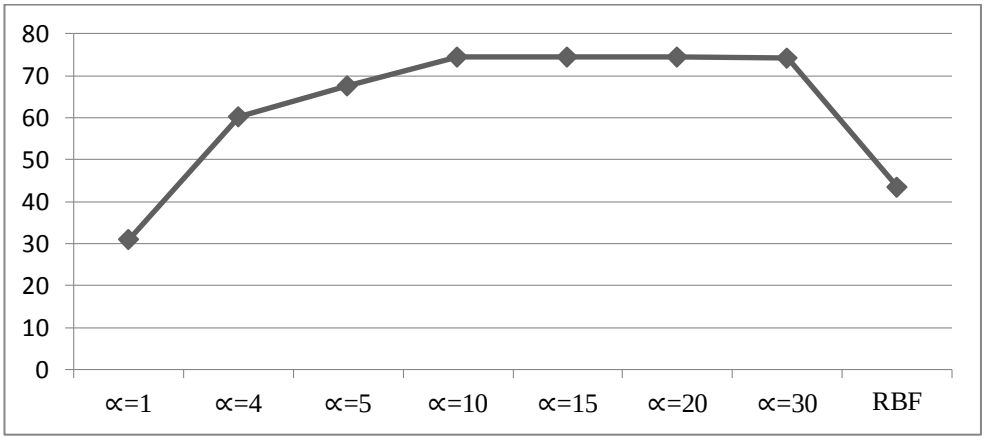
شکل (۴–۳) نمودار خطی عملکرد شبکه RBF و همچنین طبقـه بنـد SVM را بـرای ضرایب مختلف و ویژگی «$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$» نشان می‌دهد.



شکل (۴–۳): نمودار خطی عملکرد طبقه بند SVM با ضرایب مختلف و ویژگی‌های متداول

شکل (۴–۴) نمودار خطی عملکرد شبکه RBF و همچنین طبقـه بنـد SVM را بـرای

ضرایب مختلف و ویژگی پیشنهادی «$SCM + \Delta SCM$» نشان می‌دهد.



شکل (۴–۴): نمودار خطی عملکرد طبقه بند SVM با ضرایب مختلف و ویژگی‌های پیشنهادی

به طور کلی از جدول‌های نتایج حاصل از این بخش می‌توان این نکته را بیان کرد که با طبقه بندهای پیشنهادی شاهد بهبود سیستم تشخیص لهجه‌ها خواهیم بود. اما سؤالی که ممکن است پیش آید این است که آیا ضرب کردن ضریب ∝ در بردار ویژگی‌ها روی عملکرد سایر طبقه بندها تأثیری دارد یا خیر؟ برای پاسخ دادن به این سؤال آزمایشی انجام‌شده است که در آن تعدادی ضریب در بردار ویژگی MFCC ضرب و سپس به طبقه بند KNN واردشده‌اند. دلیل انتخاب طبقه بند KNN این است که همانند SVM در ضریب $\propto = 1$ بازدهی کمتری دارد. نتایج در جدول (۴–۲۴) نشان داده‌شده است.

جدول (۴–۲۴): بررسی عملکرد طبقه بند KNN به ازای ضرایب مختلف

ویژگی	طبقه بند	KNN $\propto = 1$	$\propto = 2$	$\propto = 4$	$\propto = 5$	$\propto = 10$	$\propto = 15$	$\propto = 30$
MFCC		22.07	18.5	19.68	**19.01**	18.8	18.81	19.31

همان طور که نتایج جدول نشان می‌دهند با تغییر ضریب ∝ تغییر مطلوبی در نرخ بازشناسی KNN رخ نمی‌دهد. در بخش بعد به بررسی تشخیص لهجه‌ها در محیط نویزی خواهیم پرداخت.

۴-۳-۳ نتایج آزمایش‌های حاصل از ترکیب طبقه بندها

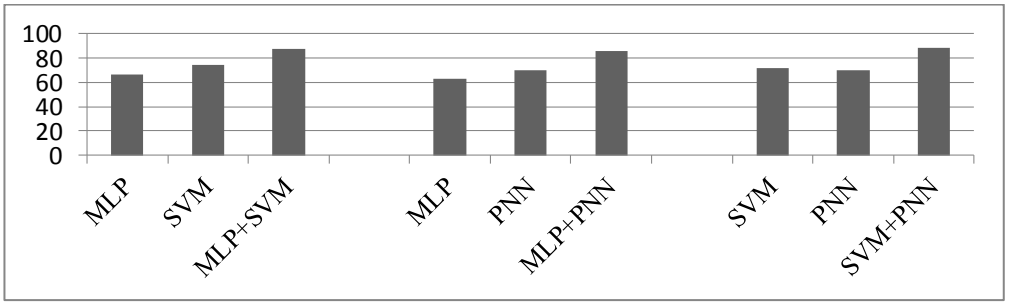
آخرین روش پیشنهادی که در مرحله طبقه‌بندی به کار گرفته‌شده است روش ترکیب طبقه بندها است. در این قسمت سه طبقه بند MLP، PNN و SVM که دارای بالاترین بازدهی در بین طبقه بندهای به‌کاررفته در این پایان‌نامه بوده‌اند به منظور ترکیب کردن استفاده‌شده است. طبقه بندها در مرحله اول به صورت دوتایی باهم ترکیب‌شده‌اند و سپس همه آن‌ها را با یکدیگر به طور همزمان ترکیب کرده‌ایم. از میان بردار ویژگی‌ها آن‌هایی که بهترین عملکرد را در بین ویژگی‌های قبلی و پیشنهادی داشته‌اند انتخاب‌شده‌اند. جدول (۴–۲۵) نرخ بازشناسی را با ترکیب طبقه بندها به صورت دوتایی نشان می‌دهد. در این حالت از روش وزن دهی به هر کدام از طبقه بندها استفاده‌شده است. در آزمایش‌های انجام‌شده در این قسمت برای طبقه بند SVM ضریب $\propto$ طوری انتخاب‌شده است که بالاترین بازدهی را داشته باشد.

جدول (۴–۲۵): نرخ بازشناسی پنج لهجه مختلف با ترکیب دوتایی طبقه بندها

ویژگی	طبقه بند	MLP+SVM	MLP+PNN	SVM+PNN	میانگین بدون ترکیب	میانگین با ترکیب	میزان بهبود
$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$		89.97	87.75	88.65	70.79	**88.79**	**18%**
$SCM + \Delta SCM$		85.5	83.27	88.36	67.17	**85.71**	**18.5%**
میانگین		87.73	85.38	**88.50**	68.98	**87.20**	**18.22**

دلیل متفاوت بودن مقادیر میانگین بدون ترکیب با مقادیر جدول‌های قبل این است که با هر بار اجرای طبقه بندهای مختلف به دلیل وجود وزن‌های تصادفی نتیجه‌ای متفاوت اما نزدیک به هم به دست می‌آید. جزئیات بیشتری از نتایج حاصل از ترکیب طبقه بندها و مقایسه آن با حالتی که هر طبقه بند به تنهایی استفاده می‌شود در نمودار شکل (۴–۵) نشان داده‌شده

است. در این نمودار از میانگین نتایج استفاده‌شده است.



شکل (۴–۵): نمودار ستونی مقایسه نرخ بازشناسی حاصل از ترکیب طبقه بندها به صورت دوتایی

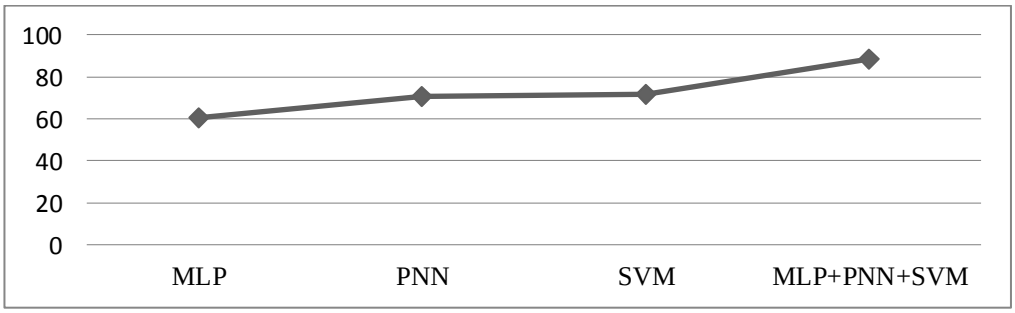
در حالت بعد هر سه طبقه بند را با روش میانگین‌گیری با یکدیگر ترکیب می‌کنیم.

جدول (۴–۲۶) نتایج حاصل را به ازای دو ویژگی برتر نشان می‌دهد.

جدول (۴–۲۶): نرخ بازشناسی پنج لهجه مختلف با ترکیب سه طبقه بند مختلف

ویژگی	طبقه بند	MLP	SVM	PNN	میانگین	MLP+SVM+PNN	میزان بهبود
$MFCC + \Delta MFCC + \Delta\Delta MFCC + 2F$		66.5	70.29	72.24	69.67	88.83	**19.1%**
$SCM + \Delta SCM$		54.21	70.62	71.21	65.34	87.58	**22.2%**
میانگین		60.35	70.45	71.72	67.5	**88.2**	**20.7%**

برای مقایسه بهتر نتایج، نمودار خطی شکل (۴–۶) رسم شده است.



شکل (۴–۶): نمودار خطی عملکرد ترکیب طبقه بندها به صورت سه‌تایی در حضور پنج لهجه

نمودارها بیانگر بهبود نتایج با ترکیب کردن طبقه بندها در حالت‌های مختلف هستند.

۴-۳-۴تشخیص لهجه‌ها در محیط نویزی

همان طور که در فصل سوم عنوان شد برخی از ویژگی‌ها هستند که برای محیط‌های نویزی کاربرد بیشتری دارند و به همین منظور دو ویژگی جدید AGMFCC و SCF پیشنهاد شد.

در این بخش قصد داریم با انجام چند آزمایش عملکرد این دو ویژگی را بررسی کنیم. به همین منظور به جملات لهجه دار، مقداری نویز سفید گوسی با نسبت سیگنال به نویزهای (SNR) مختلف اضافه کرده‌ایم. در آزمایش اول با در نظر گرفتن تمام لهجه‌ها عملکرد ویژگی‌های قبلی شامل MFCC، AMFCC، GMFCC و ویژگی‌های پیشنهادی شامل SCF، «$SCF + \Delta SCF$»، AGMFCC ، SCM و ضرایب تبدیل Zak را در شرایط نویزی و در حضور طبقه بند PNN که در میان طبقه بندهای قبلی بهترین عملکرد را داشته است بررسی می‌کنیم. جدول (۴–۲۷) نتایج حاصل را نشان می‌دهد.

جدول (۴–۲۷): میانگین نرخ بازشناسی لهجه‌ها در شرایط نویزی با طبقه بند PNN

ویژگی / SNR	Without Noise	25dB	10dB	5dB	0dB	-5dB	میانگین
MFCC	71.41	71.05	70.6	70.76	72.23	71.25	71.21
AMFCC	71	71.38	71.11	69.55	69.73	69.41	70.36
GMFCC	72.61	71.25	70.64	71.18	70.47	72.32	71.41
AGMFCC	72.14	71.65	70.33	72.30	70.49	72.41	**71.55**
SCF	70.44	70.29	70.98	70.62	68.88	71.02	70.37
$SCF + \Delta SCF$	70.65	69.26	70.80	69.28	70.31	69.7	70
Zak Transform	71.54	70.69	70.72	69.12	67.83	67.05	69.49
SCM	69.13	71.38	72.08	71.16	69.06	63.17	69.17

در آزمایش دوم همان بردارهای ویژگی را در حضور طبقه بند SVM که در بین طبقه بندهای پیشنهادی بهترین عملکرد را داشته است، در نظر می‌گیریم و میانگین نرخ بازشناسی

گفتارهای لهجه دار زبان فارسی را که با نویز سفید ترکیب‌شده‌اند به دست می‌آوریم. لازم به ذکر است که ضریب $\propto$ برای هر ویژگی طوری در نظر گرفته‌شده است که بالاترین نرخ بازشناسی را داشته باشیم. جدول (۴-۲۸) نتایج به دست آمده را نشان می‌دهد.

جدول (۴–۲۸): میانگین نرخ بازشناسی لهجه‌ها در شرایط نویزی با طبقه بند SVM

ویژگی / SNR	Without Noise	25dB	10dB	5dB	0dB	-5dB	میانگین
MFCC ($\propto= 5$)	74.85	74.13	73.73	71.54	66.97	58.52	69.95
AMFCC ($\propto= 10$)	75.21	74.13	74.22	73.02	71.45	68.94	72.82
GMFCC ($\propto= 15$)	74.40	73.28	74.13	74.15	73.28	73.73	73.82
AGMFCC ($\propto= 25$)	74.63	75.21	73.69	73.95	75.92	72.12	**74.25**
SCF ($\propto= 1$)	74.31	74.31	73.78	73.06	73.10	73.55	73.68
$SCF + \Delta SCF$ ($\propto= 1$)	75.16	74.76	74.72	75.66	74.22	75.21	**74.95**
Zak Transform ($\propto= 20$)	74.29	74.84	72.14	74.23	71.10	68.58	72.53
SCM ($\propto= 15$)	73.95	75.75	74.18	75.16	71.94	69.97	73.49

همان طور که نتایج دو جدول قبل نشان می‌دهند دو ویژگی پیشنهادی AGMFCC و SCF و مشتق اول SCF یعنی ΔSCF در میان سایر ویژگی‌ها و در حضور هر دو طبقه بند PNN و SVM بهترین میانگین عملکرد را دارند و بازدهی سیستم تشخیص گفتار لهجه دار را در محیط نویزی افزایش می‌دهند. البته اگر بخواهیم عملکرد طبقه بندها را نیز با یکدیگر مقایسه کنیم مشاهده خواهیم کرد که طبقه بند SVM همانند مراحل قبل دارای کارایی مطلوب‌تری است.

ترکیب طبقه بندها در شرایط نویزی نیز می‌تواند باعث بهبود نرخ بازشناسی شود. در جدول (۴–۲۹)، به ازای ویژگی AGMFCC و نسبت سیگنال به نویزهای مختلف، عملکرد طبقه بند ترکیبی SVM+PNN نمایش داده‌شده است.

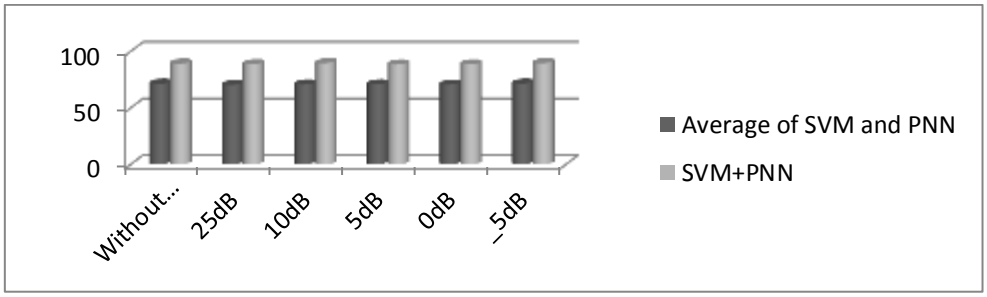
جدول (۴–۲۹): تأثیر ترکیب طبقه بندها در تشخیص لهجه‌ها در محیط نوبزی

طبقه بند / SNR	Without Noise	25dB	10dB	5dB	0dB	-5dB
SVM	70.74	70.09	70.37	71.4	69.95	71.25
PNN	71.7	70.29	70.82	70.04	70.98	71.32
میانگین	71.22	70.19	70.59	70.72	70.46	71.28
SVM+PNN	89.05	88.61	**89.44**	88.32	88.40	**89.44**
میزان بهبود	17.8%	18.4%	**18.8%**	17.6%	17.9	18.1%

همان طور که در بخش قبل بیان شد، علت تغییر نرخ بازشناسی طبقه بندهای SVM و PNN در جدول (۴-۲۹)، نسبت به جدول (۴-۲۷) و جدول (۴-۲۸)، تعیین وزن‌ها به صورت تصادفی هستند که در هر بار اجرای این طبقه بندها نرخ بازشناسی مقداری افزایش و یا کاهش پیدا می‌کند.

جدول (۴-۲۹) نشان می‌دهد که ترکیب دو طبقه بند SVM و PNN می‌تواند عملکرد سیستم را در محیط نویزی تا حد خوبی افزایش دهد. بنابراین در کنار ویژگی مناسب AGMFCC استفاده از ترکیب طبقه بندها نیز در شرایط نویزی توصیه می‌شود.

نمودار شکل (۴–۷) مقایسه میانگین عملکرد طبقه بندها را در SNRهای مختلف نشان می‌دهد.



شکل (۴–۷): نقش ترکیب طبقه بندها در تشخیص لهجه‌ها در محیط نویزی

فصل پنجم:

نتیجه‌گیری

و

پیشنهادات

۵-نتیجه‌گیری و پیشنهادات

۵-۱ نتیجه‌گیری کلی پایان‌نامه

در سیستم خودکار تشخیص گفتار اگر لهجه گفتار با آنچه که سیستم با آن آموزش دیده است تفاوت کند شاهد کاهش کارایی سیستم خواهیم بود به همین دلیل در این پایان-نامه به موضوع تشخیص لهجه‌های زبان فارسی از روی شکل موج گفتار پرداختیم. ابتدا الگوریتم کلی یک سیستم تشخیص گفتار و تشخیص لهجه و مراحل آن بیان شد که شامل سه مرحله پیش‌پردازش، استخراج ویژگی و طبقه‌بندی بود. سپس به بیان روش‌های پیشنهادی پرداختیم که دو حوزه استخراج ویژگی و طبقه‌بندی را شامل می‌شد.

در مرحله استخراج ویژگی، سه ویژگی SCM، «$SCM + \Delta SCM$» و ضرایب تبدیل Zak برای محیط‌های معمولی و سه ویژگی SCF، «$SCF + \Delta SCF$» و AGMFCC برای محیط‌های نویزی پیشنهاد شد. که استفاده از این ویژگی‌ها تا حدودی عملکرد سیستم تشخیص لهجه را بهبود بخشید. برای مرحله طبقه‌بندی سه پیشنهاد مطرح شد. یکی شبکه RBF بود که به عنوان یک طبقه بند جدید در موضوع تشخیص لهجه‌ها به کار گرفته شد و بازدهی بهتری نسبت به طبقه بند KNN و گاهی MLP داشت. در پیشنهاد دوم قبل از ورود بردارهای ویژگی به طبقه بند SVM آن‌ها را دریک ضریبی به نام $\propto$ ضرب کردیم که سبب افزایش نرخ بازشناسی تا حدود ۴۰ درصد نیز شد و این در حالی بود که اعمال این ضرایب به بردارهای ویژگی روی عملکرد سایر طبقه بندها تأثیری نداشت. پیشنهاد سوم، انجام عمل ترکیب طبقه بندها بود. این کار باعث بهبود نرخ بازشناسی در حضور پنج لهجه مختلف بین ۱۵ تا ۲۰ درصد نیز شد.از آنچه که در جدول‌های نتایج مختلف در حضور ویژگی‌ها و طبقه بندهای قبلی و پیشنهادی به دست آمد نتایج زیر قابل برداشت است.

- در میان ویژگی‌های متداول بردار ویژگی «$MFCC + \Delta MFCC +$ $\Delta\Delta MFCC + 2F$» در حضور طبقه بندهای متداول و بردار ویژگی «$MFCC + \Delta MFCC + \Delta\Delta MFCC$» در حضور طبقه بندهای پیشنهادی بهترین عملکرد را داشتند. که این بردارهای ویژگی در میان تمام بردارهای ویژگی متداول و پیشنهادی نیز زمانی که پنج لهجه را باهم بررسی می‌کنیم و یا به عبارتی پنج کلاس داریم بازدهی بالاتری دارند.
- در میان ویژگی‌های پیشنهادی ویژگی «$SCM + \Delta SCM$» در حضور طبقه بندهای متداول و پیشنهادی بهترین عملکرد را داشت. البته این ویژگی برای زمانی که دو کلاس داریم در بین تمام ویژگی‌ها بازدهی بالاتری دارد.
- در بین طبقه بندهای متداول طبقه بند PNN بالاترین بازدهی را داشته است و طبقه بند KNN کمترین بازده را دارد.
- در بین طبقه بندهای پیشنهادی طبقه بند SVM بهترین عملکرد را داشت که این طبقه بند در بین تمامی طبقه بندهای متداول و پیشنهادی نیز مطلوب‌ترین عملکرد را داشته است.
- در محیط نویزی در بین ویژگی‌های متداول ویژگی GMFCC و در بین ویژگی‌های پیشنهادی ویژگی AGMFCC نرخ بازشناسی بالاتری را فراهم کردند که ویژگی اخیر یعنی AGMFCC در میان تمام ویژگی‌های قبلی و پیشنهادی بالاترین عملکرد را داشت. لازم به ذکر است که در محیط نویزی نیز طبقه بند SVM در بین تمام طبقه بندها بهترین بازدهی را از آن خود کرده است.
- ترکیب طبقه بندها باعث بهبود خوبی در نرخ بازشناسی چه در محیط‌های معمولی و چه در محیط‌های نویزی می‌گردد. این امر زمانی به اوج خود می‌رسد

که سه طبقه بند MLP، SVM و PNN با یکدیگر ترکیب می‌شوند.

۵-۲ پیشنهادهایی برای ادامه کار

برای تحقیقاتی که ممکن است در حوزه تشخیص لهجه‌ها در آینده انجام شود چند پیشنهاد به شرح زیر وجود دارد:

- تهیه یک پایگاه داده جامع مخصوص لهجه‌های زبان فارسی
- افزایش تعداد لهجه‌های انتخابی و به عبارت دیگر تعداد کلاس‌ها و هم‌چنین انجام آزمایش‌هایی با قائل شدن تفکیک جنسیتی بین زنان و مردان.
- استخراج ویژگی‌های کاراتر از سیگنال گفتار که قابلیت تمایز بیشتری بین لهجه‌ها ایجاد کند مانند ویژگی SMFCC.
- قرار دادن تبدیل‌های جدید به جای تبدیل کسینوسی در الگوریتم محاسبه ضرایب مل-کپستروم مانند تبدیل میلین و تبدیل بسل و همچنین قرار دان تبدیل موجک به جای تبدیل فوریه در این الگوریتم.
- استفاده از طبقه بند KNN به دلیل بازدهی پایین توصیه نمی‌شود.
- در این پایان‌نامه از ساده‌ترین روش‌های ترکیب طبقه بندها استفاده شد. برای افزایش بیشتر کارایی ترکیب طبقه بندها استفاده از روش‌های دیگری مانند Bagging، Boosting و ECOC پیشنهاد می‌شود.
- به دلیل بازدهی کم طبقه بندهای GMM و PRLM در تحقیقاتی که قبلاً درباره این موضوع انجام‌شده است، استفاده از این طبقه بندها به تنهایی توصیه نمی‌شود اما از ترکیب آن‌ها با یکدیگر می‌توان استفاده کرد.
- مقایسه خطای ماشین با خطای انسانی در مورد تشخیص لهجه‌ها می‌تواند به عنوان یک آزمایش و مقایسه جالب مورد توجه قرار گیرد.

مراجع

[1] Biadsy F., (2011), PhD. thesis, “Automatic Dialect and Accent Recognition and its Application to Speech Recognition”. Columbia university,

[2] Behrava H., (2012), Master’s thesis, “Dialect and Accent Recognition”. School of Computing, Eastern Finland university.

[3] W.Junqin, Y. Junjun,. (2011). “An improved arithmetic of MFCC in speech recognition system”. In Electronics, Communications and Control (ICECC), International Conference on (pp. 719-722). IEEE.

[4] O.Grigeor, C.Grigeor, V.Velican,” Impaired Speech Evaluation using Mel-Cepstrum Analysis” International Journal Of Circuit, Systems And Signal Processing.

[۵] قلی‌پور ع، صداقی م، شمسی م،(۱۳۹۱)، " طبقه‌بندی برخی از لهجه‌های زبان فارسی با استفاده از شبکه عصبی احتمالاتی "، بیستمین کنفرانس مهندسی برق ایران، دانشگاه تهران.

[6] J.Makhoul. (1975) “Linear Prediction: A tutorial review” Proceedings of the IEEE 63(4) , pp 561-580.

[7] H.Hermansky (1990) “Perceptual Linear Predictive (PLP) analysis of speech” J. Acoust. Soc. Am. Volume 87, Issue 4, pp. 1738-1752.

[8] B.Milner,(2004)"A Comprison of Front-End Configrations for Robust Speech Recognition"presented at Acoustic Speech, and Signal Processing, IEEE Internatioanl Conference on (ICASSP)

[9] J.Psutka,L.Muller and J.V.Psutka,(2001) " Comprison of Mfcc and PLP Parameterizations in the Speaker Indipendant Continues Speech Recognition Task",Eurospeech, Scandinavia,

[10] F.Phan, M.T.Evangelia.sideman,(2000)"Speaker Identification using Nerula Network and Wavelets" IEEE Engineering in medicin and Biology Magazine, vol.191, pp.92-101

[11] Simon Haykin (1999), “Neural Networks”, Macmillan College Publishing Company.

[12] V. Vapnik and A. Chervonenkis,(1991) “The necessary and sufficient conditions for consistency in the empirical risk minimization method,” Pattern Recognition and Image Analysis, vol. 1, no. 3, pp. 283-305,

[13] Q.Y.Hong, S.Kwong,(2005) “ A genetic classification method for speaker recognition” , Engineering Applications of Artificial Intelligence , vol. 18, pp.13-19,

[۱۴] حاج احمدی ا ، همایون‌پور م و احدی م (۱۳۸۶) " بهبود مدل مخلوط گوسی با استفاده از خوشه بندی C- میانگینه‌ی فازی وزن‌دار مقاوم" ، سومین کنفرانس فناوری اطلاعات و دانش ، دانشگاه فردوسی مشهد.

[15] S.Jalalvand, A.Akbari, and B.Nasersharif,(2012) “A classifier combination approach for Farsi accents recognition” 20th Iranian Conference on Electric Engineering, pp.716-720, Tehran, Iran,

[16] Arslan M. L. and Hansen H.L. J (1996)., "Language Accent Classification in American English", Speech Communication 18(4): 353-367 .

[17] C.Teixeira, I.Trancoso, A.Seerralheiro, (1996), “Accent Identification”, Spoken Language, ICSLP 96. Fourth International Conference on (Vol. 3, pp. 1784-1787). IEEE.

[18] K.Berkling, M. Zissman, J. Vonwiller, and C. Cleirigh.(1998) "Improving accent identification through knowledge of English syllable structure." In ICSLP (Vol. 98, pp. 89-92).

[19] Fung P. and Kat L. W.,(1999) "Fast Accent Identification and Accented Speech Recognition", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, pp.221 – 224.

[20] Chen, T., Huang, C., Chang, E., & Wang, J. (2001, December). Automatic accent identification using Gaussian mixture models. In Automatic Speech Recognition and Understanding,. ASRU'01. IEEE Workshop on (pp. 343-346). IEEE.

[21] Faria A,(2005) "Accent Classification for Speech Recognition",

Machine Learning for Multimodal Interaction, pp.285-293, Springer Berlin Heidelberg.

[22] Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y. & Yoon, S. Y. (2005).” Accent detection and speech recognition for Shanghai-accented Mandarin”. In Interspeech (pp. 217-220).

[23] Pedersen C. and Diederich J.,(2007) "Accent Classification Using Support Vector Machines", 6th IEEE/ACIS International Conference on Computer and Information Science , ICIS.

[24] Ullah S. and Karray F.,(2007) "Speaker Accent Classification Using Distance Metric Learning Approach", IEEE International Symposium on Signal Processing and Information Technology.

[25] A. Rabiee and S. Setayeshi,(2010) “Persian Accents Identification Using an Adaptive Neural Network”, 2th.Int. Conf. on Education Technology and Computer Science, pp. 7-11.

[26] J.M.Karen Kua, T.Thiruvaran, M.Nosratighods, E.Ambikairajah and J.Epps.(2010) “Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition”, In Odyssey, pp 007.

[27] K.Hwei You and H.Chuan wang (1999) “Roubust features for noisy speech recognition based on temporal trajectory filtering of short time autocorrelation sequences” speech communication 28,pp 13-24,.

[28] B. j.Shannon, K.k.Paliwal,(2006) “ feature extraction from higher lag autocorrelation coefficient for robust speech recognition” 48,pp 1458-1481.

[29] S.Chakroborty and G.Saha, (2009)”improved text-independent speaker identification using fused mfcc and imfcc feature sets based on Gaussian filter” International Journal of Signal Processing 5, no. 1.

[30] A. Devand and P. Bansal,(2010) “ Roubust feature extraction for noisy speech recognition from magnitude spectrum of higher order autocorrelation coefficients” , international jurnal of computer application(0975-8887),Vo.10,No.8.

[31] C. Xie and X. Cao and Lingling He “Algorithm of Abnormal Audio Recognition Based on Improved MFCC”international workshop on information and electronic engineering (IWIEE),pp 731-737.

[۳۲] مروی ح، دارابیان د و شریف نوقابی م (۱۳۹۱) " بازشناخت مقاوم گفتار فارسی با استفاده از ضرایب مل-کپستروم بهبودیافته و شبکه عصبی".یازدهمین کنفرانس سیستم‌های هوشمند، دانشگاه خوارزمی تهران.

[33] M. Sahidullah and G. Saha,(2012), “A novel windowing technique for efficient computation of mfcc for speaker recognition”Arxiv, pp 1206-2437 v1.

[34] L. Auslander, I. Gertner, R. Tolimieri,(1991) “The Discrete Zak Transform Application to Time-Frequency Analysis and Synthesis of Nonstationary Signals” IEEE Trans. onSignal Proc., Vol. 39, No. 4, pp. 825-835.

[35] T. G. Dietterich and G. Bakiri (1995), “Solving multiclass learning problems via error correcting output codes”, J. of Artificial Intelligence rsearch 2, pp263-286.

[36] Y. Freund and R. Schapire.(1997) “ A decision-theoretic generalization of on-line learning and an application to boosting”. Journal of Computer and System Sciences, 55, 1, pp. 119-139.

[37] L. Breiman. Bagging predictors (1996). Machine Learning, 26, 2, pp. 123-140.

[38] N. Hatami and R. Ebrahimpour (2007), “Combining Multiple Classifiers: Diversify with Boosting and Combining byStacking”, IJCSNS International Journal of Computer Science and Network Security, vol.7 No.1, January.

Abstract

A speech signal contains lots of information such as age, gender, emotion and stress, health and accent.One of the challenges in ASR system is varying in accent. It means that if an ASR system train with a special accent and then test it with some different ones the speech recognition ratio degrade drastically.

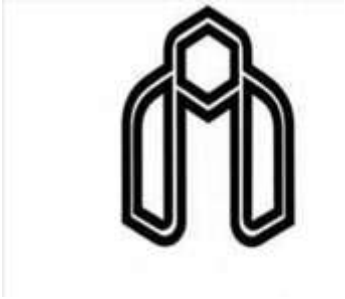
In this thesis we present some novel feature extraction methods for improving the efficiency of a Farsi accent recognition system such as Spectral Centroid Magnitude (SCM) and Spectral Centroid Frequency (SCF) Furthermore we introduce some robust noise feature extraction methods.

To classify the results, we use Radial Basis Function (RBF). In addition, we suggest a novel method to improve the efficiency of SVM classifier and a combination of different classifiers.

To evaluate the results we use Farsdat data base.exprimental result show promising improvement in recognition ratio.

Key Word:

Farsi accent, accented Speech recognition, Spectral centroid frequency, Spectrl centroid magnitude, Support vector machine, Radial basis function, Improved mel-frequency cepstral coefficient, combination of classifiers.



Shahrood University of Technology

Faculty of Electrical and Robotic Engineering

Different Farsi Accents Recognition Based on Speech Signal

Mojtaba Sharif Noughabi

Supervisor:

D.r Hossein Marvi

January 2014