
بسم الله الرحمن الرحيم

دانشکده مهندسی برق و رباتیک

بازشناخت مقاوم گفتارفارسی با استفاده از ضرایب مل-کپستروم بهبود یافته و شبکه عصبی

دانشجو: دانیال دارابیان

استاد راهنما:

دکتر حسین مروی

استاد مشاور:

دکتر حسین خسروی

پایان نامه جهت اخذ درجه کارشناسی ارشد

شهریور ۹۲

شماره : ۱۰۸۷/آ.ت.ب
تاریخ : ۹۲/۰۶/۲۶
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 :

دانیـــال دارابیـــان رشته : **بـــــرق** گرایش : **الکترونیک**

تحت عنوان : **بازشناخت مقاوم گفتار فارسی با استفاده از ضرایب مل – کپستروم بهبود یافته و شبکه عصبی**

که در تاریخ **۹۲/۰۶/۲۶** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح زیر است :

قبول (با درجه : عالی) ☑ امتیاز (۱۹/۲) 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 استـاد راهنمـا	حمید سرداری	استادیار	
۲ـ استـاد مشـاور	حسین خسروی	استادیار	
۳ـ نماینده شورای تحصیلات تکمیلی	غلامرضا الهی یزدی	استادیار	
۴ـ استـاد ممتحـن	یونس زاهدی	استادیار	
۵ ـ استـاد ممتحـن	سید رضا سید قاضی	استادیار	

رئیس دانشکده :

تقدیم به پدر و مادر و همسرم که در این مسیر مرا همراه و یاور بودند.

با تشکر از

دکتر مروی به پاس راهنمایی و همراهی بی دریغشان

و دکتر خسروی که در این مسیر همراه من بودند.

تعهد نامه

اینجانب **دانیال دارابیان** دانشجوی دوره کارشناسی ارشد رشته **الکترونیک سیستم** دانشکده **مهندسی برق و رباتیک** دانشگاه صنعتی شاهرود نویسنده پایان نامه با عنوان :

بازشناخت مقاوم گفتار فارسی با استفاده از ضرایب مل-کپستروم بهبود یافته و شبکه عصبی

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

- تحقیقات در این پایان نامه توسط اینجانب انجام شده 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 پایان نامه 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 می باشد 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 پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده 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:

بازشناخت اتوماتیک گفتار[1] در واقع مساله تشخیص گفتار برای یک نمونه گفتار صحبت شده توسط سخنگوی نامعلوم است. هر سیستم اتوماتیک تشخیص گفتار نیازمند استخراج ویژگی است که به وسیله-ی ویژگی استخراج شده بتواند گفتار ورودی را متمایز نماید و سپس تشخیص دهد از این دسته ویژگی ها می توان به ضرایب مل-کپستروم[2] اشاره نمود که بعدها به یکی از فراگیرترین ویژگی‌های مورد استفاده در پردازش گفتار تبدیل شدند. مشکل عمده در استفاده از این ضرایب زمانی ظاهر می شود که تشخیص گفتار در محیط‌های همراه با نویز صورت بپذیرد. عملکرد این الگوریتم در حضور نویز، با سرعت زیاد و به شدت کاهش می‌یابد.

برای دست‌یابی به یک عملکرد رضایت بخش تحت شرایط نویزی در سیستم‌های تشخیص گفتار نیاز به مصون سازی این الگوریتم در برابر نویز می‌باشد. در این پایان نامه یک روش مقاوم در برابر نویز، جهت استخراج ویژگی ضرایب مل-کپستروم معرفی شده است.

ویژگی‌های استخراج شده از سیگنال گفتار را به وسیله ی یک شبکه عصبی طبقه بندی می کنیم. در این پایان نامه از یک شبکه عصبی پرسپترون با یک لایه ورودی، دو لایه میانی یا مخفی و یک لایه خروجی جهت کلاسه بندی نتایج استفاده شده است. جهت ارزیابی الگوریتم پیشنهادی تعداد ۴۰ لغت[3] متفاوت، در هر لغت ۲۰ تکرار توسط ۲۰ گوینده ی متفاوت شامل زن و مرد بزرگسال از نمونه های صوتی پایگاه اطلاعاتی FARS-DAT استفاده شده است.

تاکنون روش های مختلفی در مصون سازی این الگوریتم به کار رفته است که هر یک به یک بلوک خاص پرداخته اند و یا بلوکی مکمل به الگوریتم پایه افزوده اند در حالی که در روش پیشنهادی علاوه بر در نظر

[1] Automatic Speech Recognition (ASR)
[2] MFCC (Mel Frequency Cepstral Coefficients)
[3] Isolated Word

گرفتن اکثر تغییرات مهم اعمال شده در سایر الگوریتم‌ها، هر یک را در مکان مناسب استفاده و نیز با روش‌های مناسب ترکیب کرده ایم و نیز بلوک‌هایی به الگوریتم پایه افزوده ایم.

نتایج آزمایش های تشخیص گفتار فارسی در نرخ سیگنال به نویزهای متفاوت، نشان گر افزایش معنادار نرخ تشخیص و نیز مصونیت بیشتر الگوریتم پیشنهادی نسبت به الگوریتم پایه و همچنین برخی از الگوریتم‌های مطرح در این خانواده همچون الگوریتم‌های :

AMFCC، GMFCC، ROOT-MFCC، CMN-SMN-MFCC و نیز **RAS-MFCC** است.

کلمات کلیدی:

بازشناخت گفتار- ضرایب مل-کپستروم- تابع خود همبستگی- فیلتربانک گوسی- تفریق میانگین گفتار- جبرانگر لگاریتم

لیست مقاله های استخراج شده :

باز شناسی مقاوم گفتار فارسی با استفاده از ضرایب مل-کپستروم بهبود یافته و شبکه عصبی (پذیرفته شده در شانزدهمین کنفرانس سیستم های هوشمند)

Improving the performance of MFCC for Persian Robust Speech Recognition

(Submmit شده در Journal of Intelligent Automation systems دانشگاه صنعتی شاهرود)

بازشناخت گفتار فارسی در محیط نویزی با استفاده از ضرایب مل-کپستروم بهبودیافته

(Submmit شده در نشریه سیستم های هوشمند در مهندسی برق دانشگاه اصفهان، علمی-پژوهشی)

فهرست مطالب

فصل اول: مقدمه

فصل دوم: بازشناخت گفتار در محیط نویزی

فصل سوم: روش های متداول استخراج ویژگی

فصل چهارم: مروری بر کارهای انجام شده

فصل پنجم: روش پیشنهادی

فصل ششم: نتایج آزمایش های انجام شده

فهرست شکل ها:

فصل اول:

فصل دوم:

فصل سوم:

فصل چهارم:

فصل پنجم:

فهرست جدول ها :

فهرست علائم و اختصارات

ASR	**Automatic Speech Recognition**
AR	**Autocorrelation Regressive**
AMFCC	**Autocorrelation Mel Frequency Cepstral Coefficient**
CMN-SMN MFCC	**Cepstral mean normalization Spectral Mean Normalization Mel Frequency Cepstral Coefficient**
DRHOASS- MFCC	**Differentiated Higher Order Relative Autocorrelation Sequence**
GMFCC	**Gaussian Mel Frequency Cepstral Coefficient**
GMM	**Gaussian Mixture Model**
HMM	**Hidden Morkov Model**
IMFCC	**Inverse Mel Frequency Cepstral Coefficient**
IGMFCC	**Inverse Gaussian Mel Frequency Cepstral Coefficient**
LPC	**Linear Prediction Cepstral**
LPCC	**Linear Prediction Cepstral Coefficient**
MLP	**Multy Layer Perceptron**
NN	**Neural Network**
PLP	**Perceptual Linear Prediction**
WGN	**White Gaussian Noise**

فصل اول

مقدمه

فصل اول: مقدمه

۱.۱) بیان مسئله:

یکی از زیرشاخه‌های علم پردازش سیگنال پردازش گفتار است که در روند پیشرفت‌های صورت گرفته در زمینه‌ی پردازش سیگنال، بسیار مورد توجه قرار گرفته است و تحقیقات گسترده ای را متوجه خود ساخته است. پردازش گفتار شامل سه بخش اصلی می باشد: تبدیل متن به گفتار ، بازشناسی گفتار و بهسازی گفتار.

بازشناسی گفتار توسط سیستم های اتوماتیک تشخیص گفتار(ASR) انجام می گیرد. هدف از انجام این پایان‌نامه ایجاد بهبود در عملکرد این سیستم با تکیه بر بهبود یکی از پرکاربردترین الگوریتم های استخراج ویژگی در حوزه‌ی پردازش گفتار می باشد .

هر سیستم تشخیص گفتار نیازمند استخراج ویژگی است که بوسیله‌ی آن بتواند گفتار ورودی را به طور صحیح تشخیص دهد. مقاوم سازی این ویژگی ها و بطور خاص ویژگی ضرایب مل-کپستروم، به عنوان رایج‌ترین آن‌ها در حوزه‌ی پردازش گفتار، در برابر نویز از اهمیت بسزایی در بهبود نرخ تشخیص، در سیستم‌های اتوماتیک تشخیص گفتار برخوردار است.

آنچه همواره متخصصان علم پردازش سیگنال با آن روبه رو و در صدد رفع و یا کاستن اثر نامطلوب آن هستند، سیگنال مزاحم نویز در انواع مختلف آن و به طور خاص نویز سفید است. حذف نویز سفید یکی از مهمترین انواع حذف نویز است، چرا که نویز سفید فاقد محدوده ی زمانی یا فرکانسی خاص است از این رو در بسیاری از شبیه سازی‌ها و کارهای تحقیقاتی نویز را سفید در نظر می گیرند.

مصون سازی در برابر نویز یکی از چالش برانگیزترین مسائل در سیستم‌های تشخیص گفتار می باشد. هدف از استخراج ویژگی مصون شده در برابر نویز، بهبود نرخ بازشناسی در سیستم های ASR است که زمینه ی بسیاری از تحقیقات در پردازش سیگنال و هوش مصنوعی بوده است[۱].

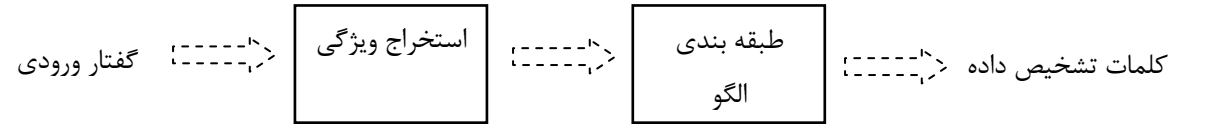
۱.۲) مختصری در مورد پیشینه ی تاریخی سیستم های اتوماتیک تشخیص گفتار:

اولین سیستم ASR در سال ۱۹۵۰ میلادی در آزمایشگاه بل ساخته شد. این اولین بار بود که هر چند به صورت ابتدایی یک ماشین می توانست گفتار انسان را تشخیص دهد. تحقیقات در سال های ۱۹۶۰ و ۱۹۷۰ با ظهور تکنولوژی اعداد باینری و با تاکید بر زمینه‌ی پردازش سیگنال و تشخیص الگو ادامه پیدا نمود. بیشترین کمک به بهبود سیستم های ASR را تبدیل فوریه، آنالیز کپسترال و برخی روش های نو-ظهور نمودند که جانشین روش‌های منسوخ گذشته گردیدند. بعضی از روش های تشخیص آماری الگو مانند شبکه های عصبی و مدل مخفی مارکوف نیز به سیستم تشخیص گفتار اضافه شدند و موفقیت خود را در بهبود عملکرد آن در بسیاری از موارد به ثبت رساندند.

سال های ۱۹۸۰ و اوایل ۱۹۹۰ تحقیقات بر بسط و توسعه‌ی توانایی‌های سیستم‌های تشخیص گفتار در انجام وظایف پیچیده تر شامل تشخیص گوینده ، تشخیص با دامنه ی لغات بیشتر، و مصون سازی ویژگی ها نسبت به نویز پرداختند. پیشرفت ها با روش های مختلف پردازش سیگنال برای استخراج ویژگی های متفاوت ادامه یافت و داده های آماری سازمان یافته تری برای آزمایش نمودن ویژگی ها، طراحی و میان

محققین به اشتراک گذاشته شد . متداولترین روش استخراج ویژگی از سیگنال صوت الگوریتمی بود به نام ضرایب مل-کپستروم (در اختصار ضرایب مل) که در سال ۱۹۸۰ توسط محققی به نام مرمل استاین معرفی گردید و در سال ۱۹۸۸ توسط شخصی به نام اسلانی به جعبه ابزار نرم افزار متلب اضافه شد و امروزه همچنان در تمام سیستم های ASR به عنوان یکی از فراگیرترین ویژگی ها مورد استفاده است.

در تجزیه و تحلیل سیگنال گفتار ویژگی های مختلفی استفاده می شود که انتخاب ویژگی مورد نظر بسته به کاربرد صورت می گیرد. این ویژگی ها که با الگوریتم های خاص خودشان بدست می آیندگاه یک عددند نظیر انرژی یا میانگین و گاه یک بردارند مانند بردار ضرایب مل-کپستروم که از آن ها با عنوان بردار ویژگی، یاد می کنند. بلوک دیاگرام یک سیستم ASR به صورت بسیار کلی، در شکل (۱-۱) نمایش داده شده است :



شکل (۱- ۱) بلوک دیاگرام پایه در یک سیستم ASR [۲]

دنباله به صورت : $y=\{y_1,y_2,y_3,\ldots,y_t\}$ باشد، وظیفه ی سیستم یافتن بردار w است. داریم: $w=\{w_1,w_2,w_3\ldots,w_t\}$ که w کلمات تشخیص داده شده اند، چنانچه بردار مشاهدات w را با $\hat{w}$ نشان دهیم و داشته باشیم : $\hat{w}=\{\hat{w}_1,\hat{w}_2,\hat{w}_3,\ldots\hat{w}_t\}$

هدف سیستم ASR برقراری معادله ی (۱-۱) می باشد :

$$\hat{w}=\arg\max pr(y|w)pr(w) \qquad (۱-۱)$$

در این معادله، y بردار ویژگی ها و $\hat{w}$ بردار مشاهدات w می باشد[۱] [۲].

۱.۳) چالش ها در سیستم های ASR:

سیستم‌های تشخیص گفتار امروزی کاربردهای بسیار گسترده ای دارند اما انتقال این سیستم ها از شرایط کاملا کنترل شده به دنیای بیرونی و بدون کنترل، یک چالش بزرگ محسوب می شود. سیستم‌های تشخیص گفتار متداول در محیط های کنترل شده عملکرد قابل قبولی دارند، آنچنان که بتوان آن ها را در محصولات تجاری به کاربرد اما در محیط های نویزی عملکرد آن‌ها به سرعت و شدت زیادی کاهش می یابد، هدف از این نوشتار مرور روش هایی بر پایه ی ضرایب مل-کپستروم است که تا کنون در بالابردن مصونیت این سیستم‌ها در برابر نویز بوسیله استخراج ویژگی مصون شده در برابر نویز، تاثیر بسزایی داشته -اند[۲] [۳].

۱.۴)ساختار پایان نامه:

در این پایان نامه ابتدا ودر فصل دوم به بررسی اجمالی پیرامون بازشناخت گفتار در محیط نویزی خواهیم پرداخت و اینکه چه فاکتورهایی در تغییر گفتار موثرند و سپس روی نویز به عنوان یکی از اصلی ترین عوامل تمرکز می کنیم، الگوریتم پایه‌ی مل-کپستروم معرفی خواهد شد و در فصل سوم به روش های متداول استخراج ویژگی از سیگنال گفتار به همراه مراحل استخراج هر یک و نیز مهمترین آن ها یعنی ضرایب مل-کپستروم، خواهیم پرداخت و در فصل چهارم روش های برجسته ی پیشین در مصون سازی این الگوریتم در برابر نویز بررسی می گردد و هر یک به تفصیل بررسی خواهد شد سپس و در فصل پنجم روش پیشنهادی معرفی و بخش‌های مختلف آن تشریح می گردد و در نهایت و در فصل ششم آزمایش های تشخیص گفتار صورت گرفته به انضمام جداول و اشکال مربوط به نتایج، بیان خواهد شد.

فصل دوم

بازشناخت گفتار در محیط نویزی

فصل دوم بازشناخت گفتار در محیط نویزی

۲.۱)مقدمه:

یک سیستم بازشناسی گفتار با چالش های فراوانی روبروست. از جمله مهمترین این چالش ها می‌توان به وجود نویز، انتخاب مجموعه ویژگی‌های مناسب، انتخاب مدل آکوستیکی مناسب، تنوع زبان، تنوع جنسیت،حالت گوینده ومشکل لهجه در بازشناسی گفتار اشاره نمود.

در مورد زبان‌های رایج مانند انگلیسی کارهای زیادی در جهت مقابله با این چالش‌ها انجام شده‌است اما در مورد زبان فارسی هنوز راه زیادی در پیش است. امروزه با داشتن میکروفن و کارت صوتی در کامپیوتر و بکارگیری نرم‌افزار تشخیص گفتار می‌توان دستورات یا کلمات را به صورت صوتی به کامپیوتر وارد کرد. حتی در بعضی از گوشی‌های تلفن همراه از این سیستم‌ها جهت دریافت دستورات بصورت صوتی استفاده می‌شود.

در این فصل در ادامه به تغییرات نامطلوب سیگنال گفتار اشاره می کنیم و سپس به آموزش سیستم های تشخیص گفتار و نیز روش های اصلی در مصون سازی سیستم های تشخیص گفتار در برابر این چالش ها اشاره خواهیم نمود و در نهایت به طبقه بندهای رایج در تشخیص عملکرد سیستم های تشخیص گفتار در محیط نویزی به صورت مختصر خواهیم پرداخت.

در روش پیشنهادی از شبکه ی عصبی به عنوان طبقه بند استفاده شده است که به آن نیز پرداخته شده است.

۲.۲) تغییرات نامطلوب در سیگنال گفتار:

آنچه در یک سیستم تشخیص گفتار سبب اختلال در تشخیص صحیح می گردد تغییرات ناخواسته در سیگنال گفتار می باشد که در جهت شناخت صحیح و سپس رفع آن با استفاده از روش های موجود باید کوشید.

بعضی علت های رایج در تغییر گفتار به شرح زیر می باشند:

۱. آلودگی با نویز (شامل دو نوع رایج : نویز اضافی شونده[۱] و نیز کانوالوشن[۲])
۲. حالات گفتار[۳]
۳. تغییرات درونی گوینده[۴]

۲.۲.۱) آلودگی با نویز:

نویز شامل دو نوع، نویز اضافی شونده و نیز نویز کانوالوشن محور مصون سازی در الگوریتم ضرایب مل-کپستروم در این پایان نامه خواهد بود که به بررسی و معرفی آن می پردازیم.

نویز اضافی شونده ناشی از مواردی چون تغییر مکان میکروفون، صدای وسایل نقلیه، صدای جابه جا شدن اجسام و اشخاص، صدای وسایل گرمایشی و سرمایشی و ... که به سیگنال گفتار اضافه می شوند.

[1] Additive Noise
[2] Convolution Noise
[3] Speaking Style
[4] Inter Speaker Variation

سیستم شنوایی انسان نسبت به این گونه نویز بسیار مصون عمل می نماید بدین ترتیب که در بیشتر موارد فرد متوجه وجود این‌گونه نویز در اطراف خود نیست و به راحتی به روند شنیداری و گفتاری خویش ادامه می دهد.

نویز اضافی شونده در حوزه زمان و فرکانس به سیگنال اضافی می گردد. معادله های (۲-۱) و (۲-۲) نحوه ی ترکیب این نوع نویز را در دو حوزه‌ی زمان و فرکانس نمایش می دهند.

در این معادله ها که یکی در حوزه‌ی زمان و دیگری در حوزه‌ی فرکانس می باشد، سیگنال ورودی را با s، سیگنال نویز را با d و سیگنال حاصل از ترکیب این دو را را با x نمایش می دهیم :

$$X(n)=s(n)+d(n) \qquad (۲-۱)$$

$$X(w)=s(w)+d(w) \qquad (۲-۲)$$

نویز کانوالوشن نیز، ناشی از تغییر سیگنال گفتار در طول مسیر از منبع (دهان) تا زمانی که به فرم باینری در می آید، می‌باشد. این تغییرات علل متفاوتی دارد، از آن جمله می توان به تغییرات در اثر برخورد با دیوارهای اتاق و یا انتقال ناقص صدا[1] توسط میکروفن تلفن و... اشاره نمود.

نویز کانالوشن در حوزه ی فرکانس در سیگنال ضرب و در حوزه ی زمان با سیگنال کانوال می گردد. به عنوان یک مثال ارزشمند از نویز کانوالوشن، می توان به انعکاس صدا ناشی از دیوارهای اتاق اشاره نمود. میکروفن علاوه بر دریافت صدای فرد انعکاس هایی از صدای او را نیز دریافت می کند. این انعکاس ها با قدری تاخیر و اعوجاج به محل تولید باز می گردند بنابراین سیگنال بازگشتی به میکروفن، مجموع سیگنال اصلی و چندین کوپی تاخیر یافته از آن است که به مرور زمان دامنه ی این کپی ها کم خواهد

[1] Imperfect Transduction

شد. این نویز نه تنها مستقل از سیگنال نیست بلکه دقیقا به آن وابسته است. حذف این نوع نویز به علت همین وابستگی، با روش های استخراج ویژگی مصون شده در برابر نویز امکان پذیر نمی باشد، زیرا فرض های به کار رفته در این الگوریتم ها که یکی از مهم ترین آن ها عدم وابستگی سیگنال به نویز می باشد صادق نیست و برای حذف این اثر از روش هایی چون به کاربردن میکروفن با فاصله ی بسیار کم نسبت به دهان و نیز قرار دادن آرایه ای از میکروفن ها در محل ضبط صدا صورت می گیرد.

اکثر نویز های موجود در محیط های واقعی ایستان نیستند. منظور از نویز ایستان، نویزی است که در آن چگالی طیف توان ثابت است و با زمان تغییر نمی‌کند و منظور از نویز غیر ایستان نویزی است که مشخصات آماری آن با زمان تغییر می‌کند مثالی از نویز غیرایستان نویز همهمه است که در مواردی نظیر مهمانی‌ها رخ می دهد این نویز با سیگنال اصلی گفتار همبستگی دارد و سیستم های ASR باید این دسته از انواع نویز را نیز در نظر بگیرند. توجه می‌کنیم که مصون سازی در برابر این نوع نویز موضوع کار در این پایان نامه نمی باشد. در طول این پایان نامه در کلیه‌ی روش‌های بهبود و مصون‌سازی الگوریتم مل-کپستروم و نیز روش پیشنهادی، نویز ناهمبسته و ایستان در نظر گرفته شده است.

همان‌طور که در مقدمه ذکر شد در آزمایش‌های تشخیص گفتار از نویز سفید به عنوان نویز آزمایش، استفاده می شود نویز سفید سیگنالی است که طیف توان آن هموار است، همه‌ی محدوده‌های فرکانسی را دربر می گیرد و میانگین صفر دارد. نویز هایی که محدوده های خاص فرکانسی را شامل می شوند نویز-های رنگی نام دارند. در عمل اکثر نویز ها موجود رنگی‌اند، به عنوان نمونه می توان به نویز صورتی که نویزی رنگی می باشد اشاره نمود که ماهیتی پایین گذر نیز دارد همچنین نویز های ناشی از فن کامپیوتر و صدای اتومبیل که با این نویز تخمین زده می شوند.

نویز هایی با ماهیت فرکانسی پایین (۴۰۰ هرتز تا ۵ کیلوهرتز) قابلیت درک گفتار را کاهش می دهند. این در حالی است که نویز با مولفه‌ی فرکانسی بالا تاثیر کمی بر قابلیت فهم گفتار دارد و بیشتر باعث

خستگی شنونده می شوند. همان‌طور که در مقدمه ذکر شد نویز سفید فاقد محدوده‌ی زمانی یا فرکانسی خاص است از این رو در بسیاری از شبیه سازی ها و کارهای تحقیقاتی نویز را سفید در نظر می‌گیرند زیرا حذف این‌گونه نویز بعلت عدم تعلق به یک محدوده‌ی فرکانسی خاص پیچیده تر و حائز اهمیت می - باشد[۳][۴].

۲.۲.۲) حالات گفتار:

سیستم تشخیص گفتار گاه در محیط های شلوغ مورد استفاده قرار می گیرد. در این محیط ها فرد برای رساندن صدای خود به دیگران نیاز به فریاد زدن و گاه تغییر صدا دارد و نیز افراد در نقاط مختلف با لهجه -های مختلف سخن می‌گویند که سبب تغییر در یک کلمه‌ی واحد در دو مکان متفاوت می‌گردد که این گونه تغییرات باید توسط سیستم در نظر گرفته شود[۴].

۲.۲.۳) تغییرات درونی گوینده:

به عواملی چون بیماری، غم و اندوه، خستگی، شادی، سن، هیجان و.. اطلاق می شود و پارامترهایی است که در راه رسیدن به یک تشخیص گفتار صحیح، سیستم تشخیص گفتار باید این پارامترها را در نظر بگیرد و نیز نسبت به آن‌ها مجهز گردد[۴].

۲.۳) آموزش سیستم های تشخیص گفتار:

سیستم‌های ASR امروزی تحت محدوده‌ی وسیعی از گوینده ها آموزش داده می شوند به نحوی که باعث عملکردآن‌ها، مستقل از گوینده می‌گردد اما بین ویژگی‌های آماری سیگنال گفتار دریافتی در محیط واقعی با سیگنال گفتاری که برای آموزش (تخمین پارامتر های آماری) استفاده می شود تفاوت وجود دارد. این تفاوت را در اصطلاح عدم تطابق بین شرایط آموزش و تست می نامند.[1] یکی از راه حل‌های موجود برای برطرف نمودن این عدم تطابق جمع آوری داده های آموزشی در همان شرایط داده های

[1] Mismatch Between Train and Test Condition

تست است که مانع از ایجاد عدم تطابق می گردد. اما آیا این امر در عمل امکان پذیر است؟ فاکتورهای زیادی در دنیای واقعی باعث ایجاد تغییر می گردند. غالبا این فاکتورها مستقل از هم هستند و برای آنکه این فاکتورهای تغییر را، در هر دو شرایط به صورت دقیق کنار هم آوریم، نیاز به هزینه‌ی بالا و سختی بسیار است و حجم وسیعی از داده ها مورد نیاز است. اما این نکته را تا حدودی می توان رعایت کرد مثلا سیستم های تشخیص گفتار امروزی در حالت چند وضعیتی آموزش داده می شوند بدین معنا که نویز در چندین نرخ سیگنال به نویز متفاوت به سیگنال اضافه می گردد (به عنوان مثال ۵ ،۱۰ ،۱۵ و ۲۰ دسی بل) .

راه دیگر برای کاهش عدم تطابق بین داده های آموزش و آزمایش، وفق پذیر نمودن سیستم تشخیص گفتار است که در ایده آل ترین حالت به معنای آگاهی سیستم تشخیص گفتار به شرایطی است که باید در آن عمل نماید. این روش خود به دو طریق انجام می گیرد :

- روش دانش بنیان
- روش داده محور

منظور از روش دانش بنیان مفروض قرار دادن بعضی از ویژگی های نویز و نحوه‌ی ترکیب شدن آن باسیگنال است.

منظور از روش های داده محور استفاده از داده هایی است که از همان شرایط آزمایش و تشخیص بدست آمده اند.

امروزه در بیشتر تشخیص دهنده ها ترکیبی از تطابق داده های آموزشی و وفق پذیری را برای بهبود تشخیص سیستم مورد استفاده قرار می دهند.[۵].

۲.۴)روش های مصون سازی سیستم های اتوماتیک تشخیص گفتار در برابر نویز:

در سیستم های تشخیص، تلاش برای برطرف نمودن مشکل عدم تطابق بین داده‌های آموزش و داده‌های آزمایش در سه گروه کلی زیر صورت می پذیرد :

- مدل های وفق پذیر گفتار
- بهسازی گفتار
- استفاده از ویژگی های ذاتا مقاوم و یا ویژگی های مصون شده در برابر نویز

بهسازی گفتار به عنوان یک مرحله‌ی پیش پردازش در تشخیص گفتار مطرح می باشد تکنیک‌های بهسازی گفتار در بازیابی پارامترهای سیگنال نویزی به کار می‌رود. این تکنیک‌ها مستقیما به دقت سیستم های تشخیص گفتار بستگی ندارند. اصل این تکنیک‌ها از تلاش برای بالا بردن قابلیت درک سیگنال ناشی می شود و جهت بالابردن کیفیت سیگنال می باشد تا شناسایی گفتار، اگر چه آن‌ها به عنوان یک پیش پردازش سیگنال در سیستم های تشخیص به کار می روند.

در ادامه به چند مورد از آن ها اشاره ای کوتاه می نماییم :

- روش تفریق طیفی
- فیلتر وینر
- روشهای وفقی
- زیرفضای سیگنال و...

بهسازی با استفاده از فیلتر وینر و نیز بهسازی با استفاده از فیلتر های وفقی سایر روش های مطرح در این زمینه می باشد که در این پایان نامه به روش تفریق میانگین گفتار که برگرفته از ایده‌ی استفاده

شده در روش تفریق طیفی می‌باشد توجه شده است و در بخشی از الگوریتم پیشنهادی از آن استفاده شده است[۶] .

همچنین انتخاب ویژگی های مناسب مهمترین مرحله در طراحی سیستم های تشخیص گفتار مقاوم در برابر نویز است که مستقیما عملکرد سیستم را بهبود می بخشد. فرایند ایجاد بردارهای ویژگی از روی سیگنال گفتار را، استخراج ویژگی می نامند هدف از استخراج ویژگی در تشخیص گفتار، ایجاد یک نمایش فشرده از شکل موج گفتار می باشد که دارای بیشترین اطلاعات مربوط به گفتار است.ویژگی‌های استخراج شده باید تا حد امکان نسبت به کانال مخابراتی، نویز و تغییرات ناشی از گوینده های متفاوت، مقاوم باشند.

در انتخاب ویژگی مناسب باید موارد زیر را در ذهن داشت:

۱. حاوی بیشترین اطلاعات لازم برای تشخیص گفتار باشند.
۲. عدم حساسیت زیاد نسبت به حالت گوینده ، نویز پیش زمینه ، اعوجاج کانال و...
۳. توانایی بدست آوردن وتخمین دقیق آن ها
۴. داشتن مفهومی فیزیکی (مانند ضرایب مل-کپستروم که الهام گرفته از سیستم شنوایی انسان و نیز سیستم صوتی انسان هستند)

واضح است که انتخاب مجموعه ای از ویژگی ها که تمام این ملزومات را در نظر داشته و رعایت نماید، دشوار است اما همواره سعی بر رعایت آن ها است[۱][۲][۳][۷].

هنگامی که یک ویژگی انتخاب می شود وظیفه‌ی سیستم تشخیص گفتار استخراج آن از سیگنال گفتار است.غالبا سیستم این وظیفه را با تقسیم نمودن سیگنال گفتار به فریم هایی که با یکدیگر هم پوشانی

دارند انجام می دهد سپس مقادیر ویژگی را برای هر فریم استخراج می نماید پیچیدگی سیستم تشخیص به پیچیدگی ویژگی انتخاب شده بستگی دارد.

۲.۵) طبقه بند ها:

در طی بررسی روش های به کاررفته در بهبود الگوریتم ضرایب مل-کپستروم و دسته بندی آن ها نیاز به روشی جهت ارزیابی و کلاسه بندی نتایج می باشد. عمده ی روش های به کار رفته در ارزیابی الگوریتم های بهبود یافته شامل مواردی است که در پیرامون هریک توضیح مختصری خواهیم داد سه طبقه بند رایج در تشخیص گفتار عبارتند از:

۱.شبکه عصبی[۱]

۲.مدل مخفی مارکوف[۲]

۳.مدل مخلوط گوسی مرحله ی طبقه بندی[۳]

۲.۵.۱) شبکه ی عصبی:

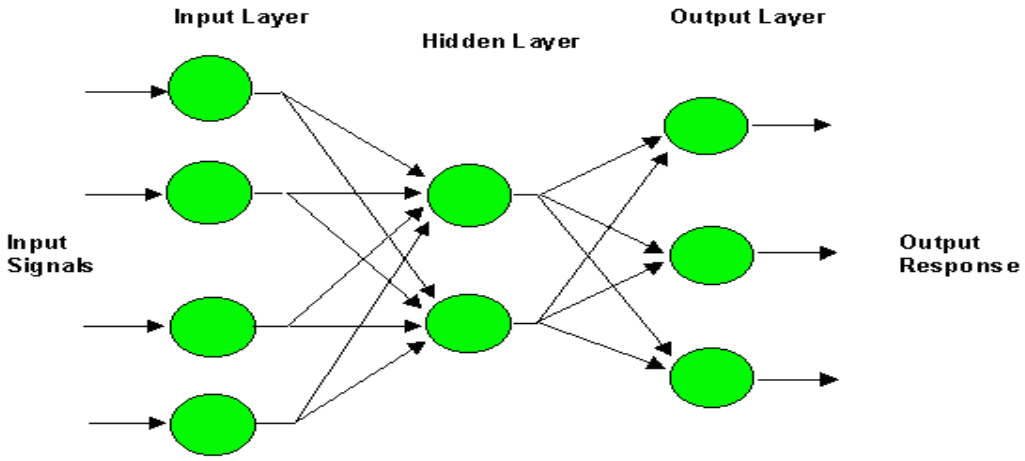
در مرحله‌ی کلاس بندی نتایج در سیستم‌های تشخیص گفتار در واقع به دنبال مطابقت ویژگی های ورودی به سیستم تشخیص گفتار و ویژگی های استخراجی ذخیره شده در سیستم می باشیم تا به وسیله -ی این معیار به نرخ بازشناسی جهت رسیدن به قضاوت صحیح و مقایسه‌ی روش استخراج ویژگی به کار رفته برسیم.

شبکه های عصبی وسیله ای هستند که به منظور ایجاد این تطابق در سیستم های ASR مورد استفاده قرار می گیرند. یکی از پرکاربردترین شبکه های عصبی مورد استفاده در بازشناسی گفتار، شبکه عصبی

[1] Neural Network (NN)
[2] Hidden Morkov Model (HMM)
[3] Gaussian Mixture Model (GMM)

پرسپترون می باشد که اولین بار در سال ۱۹۵۷ توسط آقای روزنبلات معرفی شد. در ابتدا این شبکه تنها قادر به تشخیص دو کلاس بود بنابراین نیاز به گسترش داشت که نتیجه ی آن یک شبکه‌ی پرسپترون چند لایه بود[۳۷].

یک شبکه‌ی پرسپترون چند لایه شامل یک لایه ورودی، یک یا چند لایه‌ی مخفی و یک لایه‌ی خروجی می باشد. الگوریتم شبکه مورد نظر در شکل (۲-۱) نمایش داده شده است:



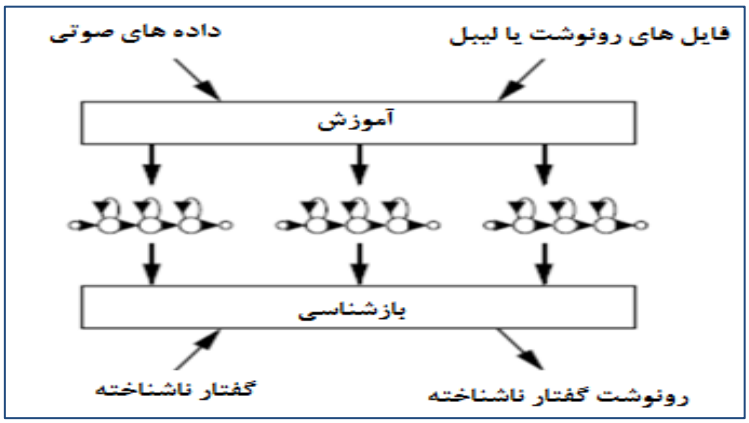
شکل (۲-۱) شبکه ی عصبی پرسپترون دولایه [۳۷]

در مرحله آزمایش توسط این شبکه به بررسی این مقوله خواهیم پرداخت که آیا الگوریتم ضرایب مل-کپستروم بهبود یافته در تشخیص گفتار نسبت به الگوریتم های پیشین عملکرد بهتری خواهد داشت یا خیر.

ویژگی های استخراج شده از سیگنال گفتار را توسط الگوریتم های بهبود یافته‌ی مل-کپستروم به وسیله -ی یک شبکه عصبی طبقه بندی می کنیم و به بررسی نرخ تشخیص حاصل می پردازیم و به دنبال ایجاد بهبود در آن‌ها خواهیم بود.

۲.۵.۲) مدل مخفی مارکوف:

مدل مخفی مارکوف یک مدل آماری است که در آن فرض می شود که سیستم، یک فرایند مارکوف با حالت‌های غیرقابل مشاهده باشد. خروجی هر یک از این حالت‌ها قابل مشاهده است و هر حالت یک توزیع احتمالاتی روی خروجی های ممکن دارد. نمای ساده از یک مدل مخفی مارکوف دوحالته در شکل (۲-۲) آورده شده است:



شکل (۲-۲): روند آموزش و بازشناسی در مدل مخفی مارکوف[۴۶]

یک مدل مخفی مارکوف از تعدادی حالت پایدار تشکیل شده است. حرکت از یک حالت پایدار به حالت پایدار دیگر با توجه به یک احتمال انجام می گیرد. احتمال حرکت از حالت پایدار i به حالت پایدار j با مقدار a_{ij} نمایش داده می شود این احتمالات در یک ماتریس به نام A ذخیره می شوند. حرکت از یک

حالت پایدار به حالتی دیگر در زمان های گسسته انجام می پذیرد که با t = ۴،۳،۲،۱،... نشان می دهند. در هر لحظه t ، مدل در یکی از حالات پایدار خواهد بود . در لحظه ی t+۱ بسته به مقادیر موجود در ماتریس A، یا به همان حالت یا به یک حالت دیگر تغییر خواهد کرد و در همان حال یکی از M علامت مربوط به الفبای خاص را تولید خواهد کرد. در اصطلاحات مربوط به HMM این علامت را مشاهده می-نامند. این علامت بسته به مقادیر احتمالی، تولید می شود. این احتمالات را در ماتریسی به نام B نمایش می دهند. با توجه به اینکه حرکت از یک حالت به حالت دیگر با استفاده از مقادیر احتمال صورت می-پذیرد، بنابراین مجموعه احتمالات گذار از یک حالت به سایر حالات باید برابر با یک باشد. در مدل کردن سیگنال با مدل مخفی مارکوف، پس از تبدیل سیگنال به پنجره هایی که فرض می کنیم سیگنال در آن ها ایستان است، هر کدام از این پنجره ها را نظیر یک حالت مدل مارکوف در نظر می گیریم و نیز تغییر وضعیت اندام گفتاری از حرکت از یک حالت پایدار به حالت پایدار دیگر را مدل می کنیم. بسته به کاربرد مورد نظر برای مدل مارکوف، ممکن است آن را به صورت های مختلف استفاده کنیم[۳۸][۳۹][۴۰] .

۲.۵.۳) مدل مخلوط گوسی:

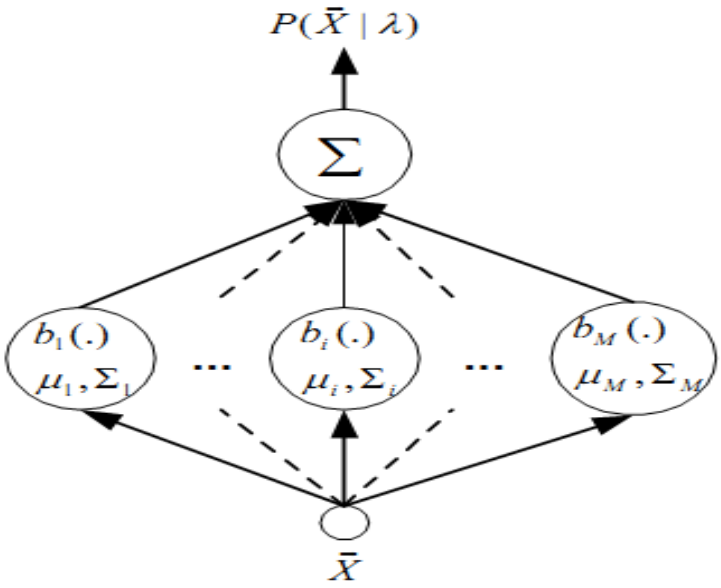
در واقع مدل های مخلوط گوسی یکی از مهمترین و پرکاربردترین روش‌های به کار رفته در تشخیص گفتار و یا تشخیص گوینده می باشد در این مدل سعی بر آن است که تابع چگالی احتمال کلاس ها را که یا به صورت تشخیص گوینده و یا تشخیص گفتار می باشد، مدل نماید.

این مدل کردن را با ترکیب خطی تعدادی تابع گوسی انجام می دهد به همین دلیل به آن مدل مخلوط گوسی گفته می شود و عناصر مخلوط ها معمولا به صورت D بعدی تعریف می شوند رابطه ی یک توزیع گوسی D بعدی بصورت زیر است:

$$b_i(x)=\frac{1}{(2\pi)^{D/2}\sigma_i^{1/2}}\ \exp\{-1/2(x-\mu_i)\sigma i^{-1}(x-\mu_i)\} \qquad (۳-۲)$$

در معادله ی فوق x بردار D بعدی ویژگی های گفتار ورودی می باشد و μ_i بردار میانگین و Σi ماتریس کواریانس می باشد.

یک مخلوط گوسی از مجموع وزن دار M توزیع گوسی مطابق شکل زیر بدست می آید که در آن x بردار ویژگی است:



شکل(۲-۳):مدل مخلوط گوسی[۴۴]

در این روش برای هر گفتار و کلاس یک GMM ،λ_l ایجاد شده است. تحت فرض GMM، احتمال بردار خاصیت x_k استخراجی از مدل λ_l به نمایندگی از یک جمع وزنی چگالی گوسین چند متغیره می باشد:

$$P(x_k \mid \lambda_l) = \sum_{I=1}^{M} w_i . b_i(x_k) \qquad (۴-۲)$$

که $b_i(x_k)$ چگالی ترکیب مولفه ها و w_i وزن های ترکیبی و M تعداد گوسی ها، می باشند.

در رابطه ی فوق p_i ها وزن های عناصر مخلوط هستند و به منظور اینکه شرایط احتمال برقرار باشد باید داشته باشیم :

$$\sum_{I=1}^{M} p_i=1 \qquad (۵-۲)$$

که مدل گفتار λ_l به این صورت بیان شده است:

$$\lambda_l= \{ w_i,\mu_i,\sigma_i \} \qquad (۶-۲)$$

مدت زمان شناسایی یک متن گفتار ناشناخته توسط یک رشته بردارهای خاصیت بیان شده است. سپس لگاریتم احتمال L_l به این صورت تعریف می شود:

$$L_l=\sum_{k=1}^{K} \log\ P(vk \mid \lambda) \qquad (۷-۲)$$

که k شاخص مرز گام و K تعداد کل مرزهای گام در یک گفتار است. نهایتا یک کلاسه کننده گفتاری را که بیشترین احتمال را داشته باشد به عنوان کلاس انتخابی بیان می کند:

$$\hat{l}= \arg\max L_l \qquad (۸-۲)$$

این ضرایب همراه با طول منحنی گام تکه ای، برای ساخت یک بردار خاصیت استفاده شده اند، سپس این بردار خاصیت برای آزمایش یک مدل مخلوط گوسی برای هر کلاس گفتار مورد استفاده است. به عنوان

مثال در تشخیص گوینده هر عنصر مخلوط گوسی بیانگر PDF یک کلاس آوایی است که در مجموع کل PDF ها توزیع چگالی احتمال گوینده را تخمین می زند[۴۱]تا [۴۶].

فصل سوم

روش های متداول استخراج ویژگی

فصل سوم: روش های متداول استخراج ویژگی

۳.۱)مقدمه:

فرایند ایجاد بردارهای ویژگی از روی سیگنال ورودی به سیستم را، استخراج ویژگی می نامند. هدف از استخراج ویژگی در تشخیص گفتار، ایجاد یک نمایش فشرده از شکل موج گفتار می باشد که دارای بیشترین اطلاعات مربوط به گفتار است. این نمایش فشرده باید بیشترین جداسازی را بین صداهای مختلف ایجاد نماید. مرحله‌ی استخراج ویژگی یکی از مهمترین مراحل موجود در سیستم های تشخیص گفتار می باشد که بسیار تاثیر گذار بوده و دقت سیستم تا حد زیادی وابسته به آن است. اگر ویژگی های استخراج شده بتوانند نمایش خوبی از سیگنال داشته باشند و اطلاعات اساسی موجود در سیگنال گفتار که برای جداسازی بین صداهای مختلف ضروری می باشند، حذف نکنند دقت سیستم تشخیص تا حد زیادی بالا خواهد رفت. ویژگی‌های استخراج شده باید تا حد امکان نسبت به کانال مخابراتی، نویز و تغییرات ناشی از گوینده های متفاوت، مقاوم باشند.

در این فصل تعدادی از روش های استخراج ویژگی متداول در سیستم های بازشناسی گفتار، از جمله ضرایب مل-کپستروم که بهبود آن محور اصلی این پایان نامه می باشد بررسی خواهد شد و به تشریح مراحل آن خواهیم پرداخت.

۳.۲) ضرایب کپسترال پیشگوی خطی (LPCC[1]):

ضرایب کپسترال پیش گوی خطی و نیز ضرایب مل کپستروم هر دو مبتنی بر پردازش کپسترال هستند. برای استخراج ویژگی، معمولا از پاسخ فرکانسی مربوط به مجرای صوتی استفاده می شود و اطلاعات مربوط به سیگنال تحریک که برای اصوات صدادار، متناوب و برای اصوات بی صدا، نویز مانند است نادیده گرفته می شوند. دلیل این امر آن است که پاسخ فرکانسی مربوط به مجرای صوتی، بهترین جداسازی بین صداهای مختلف مربوط به گفتار را ایجاد می نماید.

برای جداسازی سیگنال تحریک و پاسخ فرکانسی مجرای صوتی، از پردازش کپسترال استفاده می کنیم. کپسترال یکی از انواع تبدیل های هوموورفیک[2] است که قادر به جداسازی اطلاعات مربوط به منبع از فیلتر می باشد. برای تولید گفتار از رابطه ی (۳-۱) استفاده می کنیم. داریم:

$$S(n)=Gu(n)*h(n) \qquad (۱-۳)$$

که u(n) سیگنال تحریک، h(n) پاسخ فرکانسی مجرای صوتی، s(n) سیگنال گفتار و G بهره می باشد.

اگر از رابطه ی (۳-۱) تبدیل فوریه بگیریم خواهیم داشت:

$$S(f)=GU(f).H(f) \qquad (۲-۳)$$

حال از دو طرف رابطه لگاریتم مختلط می گیریم. داریم :

$$Log(S(f)) = log(GU(f))+log(H(f)) \qquad (۳-۳)$$

[1] Linear Prediction Cepstral Coefficients
[2] Homomorphic Transformation

باتوجه به رابطه ی (۳-۳)، در حوزه ی لگاریتم اطلاعات مربوط به سیگنال تحریک و مجرای صوتی بر روی هم افتاده و با تکنیک های موجود در پردازش سیگنال، قابل جداسازی می باشند.

با گرفتن عکس تبدیل از رابطه ی (۳-۳) کپستروم مختلط مربوط به سیگنال s(n) بدست می آید در کپسترال، ضرایب مرتبه پایین تر مربوط به مجرای صوتی بوده و ضرایب مرتبه بالاتر مربوط به تحریک می -باشند بنابراین با حذف ضرایب مرتبه بالاتر و نگه داشتن ۱۲ یا ۱۳ ضریب مرتبه پایین تر، اطلاعات مربوط به مجرای صوتی از سیگنال تحریک جدا می شوند.

یکی از قویترین تکنیک های آنالیز صوت، آنالیز پیشگوی خطی است. این روش به یکی از روش های پرکاربرد در محاسبه پارامترهای مربوط به سیگنال صوت، مانند دوره تناوب اصلی و فرکانس های فورمنت تبدیل شده و کاربردهای زیادی در زمینه های مختلف پردازش صوت پیدا کرده است. آنالیز پیشگوی خطی به دلیل تواناییش در محاسبه نسبتا دقیق پارامترهای مربوط به صوت و همچنین سرعت بالا، اهمیت زیادی در پردازش صوت دارد. ایده ی اولیه ی مربوطه این است که می توان یک نمونه سیگنال را توسط ترکیبی از نمونه های قبلی تخمین زد.

مشکل اصلی در آنالیز پیشگوی خطی تعیین مجموعه ضرایب پیشگوی a_k به طور مستقیم از روی سیگنال گفتار است به گونه ای که تخمینی خوب از خواص طیفی سیگنال گفتار بدست آید.به دلیل خاصیت تغییر پذیری با زمان سیگنال گفتار، ضرایب پیشگوی باید از بخش های کوتاه سیگنال گفتار تخمین زده شوند. راهکار ابتدایی برای حل این مساله پیدا کردن مجموعه ای از ضرایب پیشگویی است که میانگین مربع خطای پیشگویی را در طول بخش کوچکی از مشکل موج گفتار، مینیمم کند. پارامترهای بدست آمده از این روش به عنوان پارامترهای تابع سیستم در مدل تولید گفتار فرض می شوند.

بعد از بدست آمدن ضرایب پیشگوی خطی ضرایب کپسترال پیشگوی خطی ، به کمک رابطه ی(۲-۶) محاسبه می شوند. همچنین از این رابطه دیده می شود که با تعدادی محدود از ضرایب پیشگوی خطی،

ضرایب LPCC بدست آمده نامحدود می باشند. به طور تجربی نشان داده شده است که معمولا استفاده از ۱۲ تا ۲۰ ضریب اول، برای بدست آوردن نتایج خوب، در تشخیص گفتار مفید است.

LPC در واقع سعی می نماید که سیستم تولید گفتار انسان را با استفاده از یک فیلتر تمام قطب مدلسازی کند بسته به این که صدای تولید شده صوت[1] و یا غیر صوت[2] باشد ورودی فیلتر می تواند یک قطار ضربه-ی متناوب و یا یک نویز سفید باشد.دوره تناوب سیگنال گفتار نیز توسط دوره تناوب قطار ضربه مشخص می شود.

$$C(n)=\begin{cases} 0 & n<0 \\ \ln G & n=0 \\ a+\sum_{k=1}^{n-1}\left(\frac{k}{n}\right)c(k)a_{n-k} & 0<n\leq p \\ \sum_{k=1}^{n-1}\left(\frac{k}{n}\right)c(k)a_{n-k} & n>p \end{cases} \qquad (۴-۳)$$

در رابطه‌ی (۳-۴)، G نشان دهنده‌ی بهره بوده و a_k مجموعه ضرایب پیشگویی را بیان می کند[۸][۹][۱۰].

[1] Voice
[2] Unvoiced

۳.۳) ضرایب مل-کپستروم:

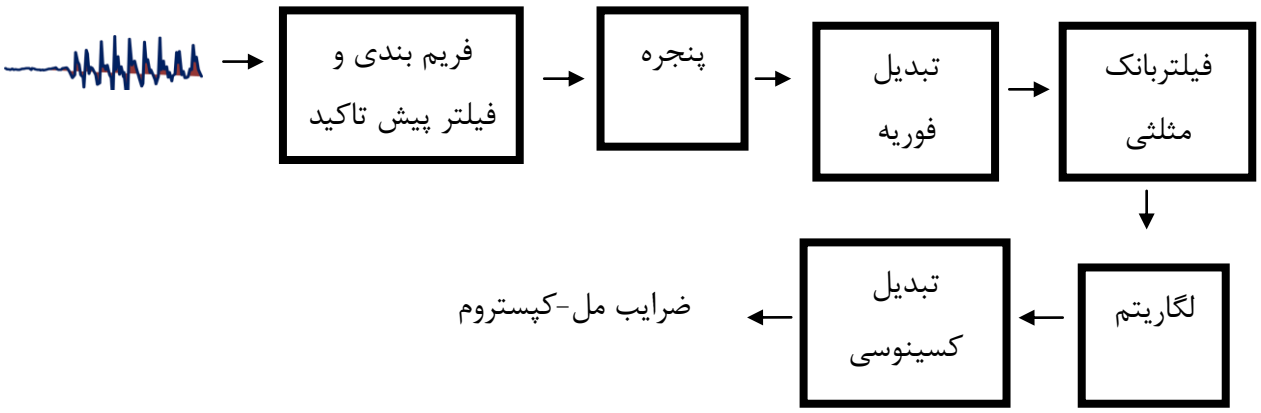
ضرایب کپسترال فرکانس مل، ابتدا در سال ۱۹۸۰ توسط دیویس[1] و مرملستین[2] به عنوان ویژگی، برای استفاده در تشخیص گفتار پیشنهاد شد و امروزه به عنوان پرکاربردترین روش استخراج ویژگی در سیستم های تشخیص گفتار مورد استفاده قرار می گیرد. روش MFCC مقیاس مل را به طیف توان سیگنال اعمال می کند.[۱۰].

ایده‌ی اصلی در استفاده از ضرایب مل-کپستروم الهام گرفتن از خواص شنیداری گوش انسان در دریافت و فهم گفتار انسان است. عملکرد گوش انسان به گونه ای است که فرکانس دریافتی را به همان اندازه‌ی فیزیکی آن درک نمی‌کند و در دریافت فرکانس عملکرد خطی ندارد، بلکه آن را به صورت لگاریتمی و طبق معادله ی (۳-۵) درک می نماید :

$$f_{mel}=2595 \log (1+\frac{f}{700}) \qquad (۵-۳)$$

در این رابطه f فرکانس بر حسب هرتز است و نیز f_{mel} فرکانس در حوزه ی مل است، این رابطه بیان می-دارد که گوش انسان دقت فرکانسی بالایی در درک فرکانس های پایین دارد و به عکس، دقت فرکانسی پایینی در درک فرکانس های بالا دارد بلوک دیاگرام الگوریتم پایه‌ی محاسبه‌ی ضرایب مل-کپستروم در شکل (۳-۱) نشان داده شده است:

[1] Davis
[2] Mermelstein



شکل (۳-۱) بلوک دیاگرام پایه‌ی محاسبه ضرایب مل-کپستروم[۱]

در ادامه به طور مختصر، مراحل مختلف روش استخراج ضرایب MFCC بررسی می شود.

۳.۳.۱) فریم بندی[۱] و فیلتر پیش تاکید[۲]:

برای استخراج ویژگی‌های مورد نیاز به طور معمول ابتدا سیگنال گفتار به پنجره های کوتاهی با طول های مناسب که باهم هم پوشانی دارند تقسیم می گردد، مشخصات سیگنال گفتار در حین بیان یک گویش تغییر می کند بنابراین سیگنال گفتار سیگنالی غیر ایستان است و خواص آماری آن در طول زمان تغییر می کند. ولی از آن جا که اندام گفتاری در طول زمان به کندی تغییر حالت می دهند در بازه های زمانی کوچکی می توان آن ها را ایستان فرض کرد این قطعات گفتاری را فریم گویند. گاه پیش از این مرحله، مرحله‌ی پیش تاکید انجام می گیرد که سبب کاهش اثر تشعشع لب‌ها و نیز بالا رفتن نرخ اندازه‌ی سیگنال به نویز و نیز یکنواختی سیگنال می گردد و معمولا بوسیله‌ی اعمال یک فیلتر دیجیتال مرتبه اول با پارامتر پیش تاکید α با یک قطب استفاده می شود، انجام می گیرد. فیلتر پیش تاکید در معادله‌ی(۳-۶) بیان شده است:

$$P(z)=1-\alpha z^{-1} \qquad (۶-۳)$$

[1] Frame Blocking
[2] Pre-emphasis

معمولا α را مقداری ثابت بین ۰/۹ و ۱ انتخاب می کنند.

۳.۳.۲) پنجره گذاری[1] :

سیگنال حاصل از مرحله‌ی پیش تاکید و فریم بندی وارد بلوک پنجره می گردد. برای از بین بردن ناپیوستگی موجود در مرز بین فریم ها، باید هر فریم را در یک تابع پنجره ضرب کنیم. برای این منظور، پنجره‌های گوناگونی وجود دارد که هر یک کاربرد خاص خود را دارند. پنجره ممکن است پنجره‌ی مستطیلی یا پنجره ای در طرفین کاهنده مانند همینگ باشد. در پنجره‌ی مستطیلی، پنجره هیچ تغییری در دامنه‌ی نمونه ها ایجاد نمی کند اما در پنجره‌ی همینگ که پنجره‌ی رایج مورد استفاده در الگوریتم ضرایب مل-کپستروم نیز می باشد، دامنه‌ی سیگنال در وجه‌ها دچار تغییر می گردد. معادله‌ی یک پنجره ی همینگ ساده در زیر آورده شده است. در این رابطه N طول پنجره می باشد:

$$W(n)=\begin{cases} 0.54-0.46*\cos(\frac{2k\pi}{N-1}) & K=0,1,\ldots,N-1 \\ 0 & Else \end{cases} \qquad (۷-۳)$$

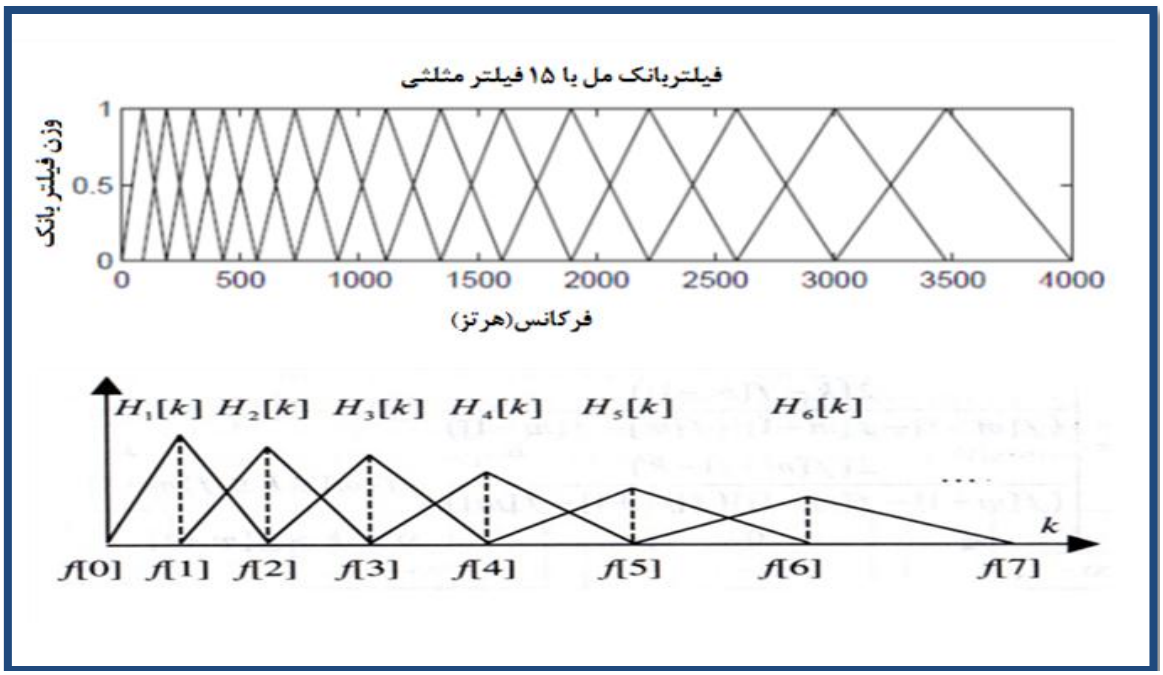
پس از فریم بندی و عبور از پنجره ی مناسب از سیگنال حاصل تبدیل فوریه گرفته شده و اندازه‌ی آن محاسبه، به عنوان طیف توان سیگنال وارد مرحله ی فیلتربانک می گردد.

۳.۳.۳)فیلتربانک:

بر اساس اصول اولیه‌ی تشکیل ضرایب مل در این مرحله سیگنال از یک فیلتربانک که متشکل از تعدادی مثلث است عبور می نماید، در فرکانس‌های بالاتر پهنای باند فیلتر ها زیادتر است و این امر بدان معناست که حساسیت گوش انسان نسبت به تغییر فرکانس در فرکانس‌های بالاتر کمتر از حساسیت آن در

[1] Windowing

فرکانس‌های پایین تر است. در واقع فیلتر بانک مجموعه‌ای از فیلترهای میان گذر با فرکانس های متفاوت است که محدوده ی مورد نظر سیگنال به لحاظ طیفی را می پوشاند. معمولا فرکانس‌ها به گونه ای انتخاب می شوند که حساسیت گوش انسان نسبت به حوزه های مختلف فرکانسی، مدل شود. یکی از تخمین هایی که برای این منظور استفاده می شود، مقیاس مل است. این فیلتربانک مثلثی در شکل(۳-۲) تمایش داده شده است. مقیاس مل[1] به صورت پنجره‌های مثلثی که با یکدیگر دارای همپوشانی هستند، اعمال می‌شود. در نوع و تعداد فیلتر بانک‌های بکار رفته، تنوع و تغییرات زیادی وجود دارد.



شکل (۳-۲) : دو نوع پیاده سازی فیلتربانک مل. در حالت اول، وزن برای تمام فیلترها یکسان است اما در حالت دوم مقدار وزن از ۱ تا ۰.۵ متغیر است[۴۶]

برای بدست آوردن فیلتربانک فوق در ابتدا نیاز به یافتن f_{max} است که از تقسیم ماکسیمم فرکانس نمونه برداری بر دو بدست می آید و سپس انتقال آن به حوزه‌ی مل، مل واحد ارزیابی صدای درک شده است. فرکانس مل طبق الگوریتم پایه در معادله‌ی (۳-۵) بیان شده بود.

[1] Mel-Scale

در این رابطه fفرکانس واقعی بوده و بر حسب هرتز می باشد و f_{max} ماکسیمم فرکانس نمونه برداری تقسیم بر دو می باشد.

سپس Δ_{mel} محاسبه خواهد گردید که در آن i شماره‌ی فیلتربانک (مثلث ها) می باشد:

$$\Delta_{mel}=\frac{f_{max(mel)}}{i+1} \quad (۸-۳)$$

و kb ها که نقاط مرزی در مثلث ها را مشخص می‌کنند، از رابطه ی زیر بدست می آیند:

$$kb_i=(i+1)\ .\ \Delta_{mel} \quad (۹-۳)$$

که در آن i شماره‌ی فیلتربانک است و در نهایت فیلتربانک ها، a_i طبق ضوابط معادله‌ی (۳-۱۰)ساخته می شود:

$$a_i=\begin{cases} 0 & K<kb_{i-1} \\ \frac{k-kb_{i-1}}{kb_i-kb_{i-1}} & kb_{i-1}\le K\le kb_i \\ \frac{kb_{i+1}-k}{kb_{i+1}-kb_i} & kb_i\le K\le kb_{i+1} \\ 0 & kb_{i+1}\le k \end{cases} \quad (۱۰-۳)$$

۳.۳.۴) لگاریتم و تبدیل کسینوسی:

پس از قرار دادن فیلترها روی طیف سیگنال، لگاریتم انرژی فیلتربانک محاسبه می شود در نهایت با اعمال تبدیل فوریه کسینوسی گسسته به لگاریتم انرژی های فیلتر بانک، ضرایب MFCC بدست می آیند. از آنجا که ضرایب پایین تر حاوی اطلاعات مربوط به مجرای صوتی هستند ۱۳ ضریب ابتدایی را نگه

داشته و باقی ضرایب را حذف می کنند[۱۰]. با استفاده از این مقادیر ضرایب مل طبق رابطه ی زیر محاسبه می گردد :

$$C(i)=\sum_j \log(X_j)\ \cos(\frac{\pi i(j-0.55)}{f}) \qquad 1\leq i \leq f \qquad (۱۲-۳)$$

که در آن f تعداد فیلتربانک ها (به طور معمول ۱۳ و گاه ۱۴ فیلتربانک مورد استفاده می باشد) X_j خروجی حاصل از فیلتر j ام و c(i) ضرایب مل-کپستروم حاصل است .

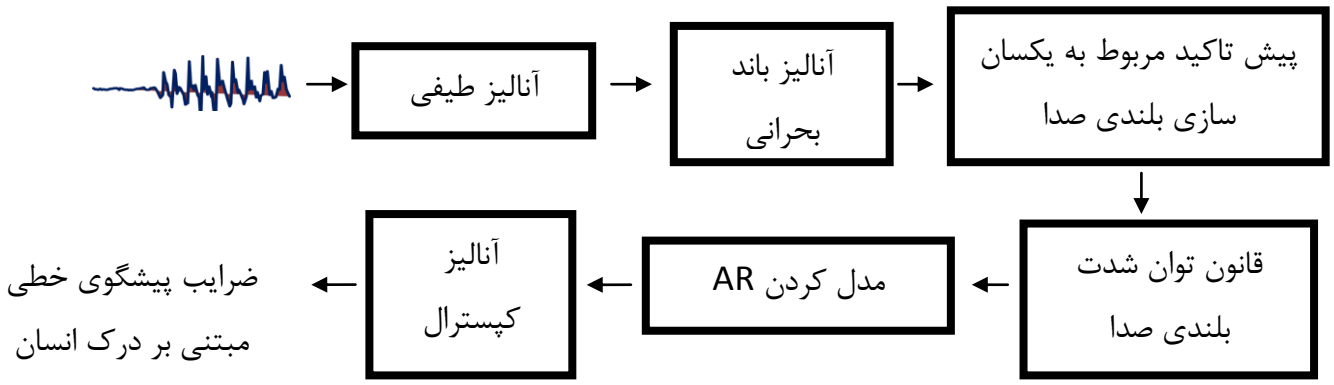
معمولا لگاریتم انرژی هر فریم، به عنوان ضریب صفرم به مجموعه ضرایب کپسترال فرکانس مل افزوده می شود [۲] [۱۲][۱۳].

تا کنون این ضرایب به عنوان متداول ترین روش استخراج ویژگی در بازشناسی گفتار به کار رفته است. همانطور که قبلا نیز اشاره شد مهمترین مشکل این ضرایب آسیب پذیری آن ها در برابر انواع نویز می باشد که به همین دلیل روش های بسیاری در مصون سازی این الگوریتم توسط محققین علم پردازش سیگنال ارائه شده است که موجب شکل‌گیری انواع ضرایب مل گردیده است که به تفصیل در فصل چهارم، مروری بر کارهای انجام شده، پیرامون روند شکل گیری آن ها توضیح داده شده است.

۳.۴) پیشگوی خطی مبتنی بر درک انسان (PLP)

روش پیشگوی خطی برای بدست آوردن ضرایب کپسترال، در قسمت قبل بیان شد یکی از معایب LPCC آن است که مکانیزم شنوایی انسان را در محاسبه ی ویژگی ها منظور نمی کند. به عبارت دیگر پیشگوی خطی در تمام فرکانس ها، سیگنال گفتار را به یک صورت تخمین می زند که این مطابق با سیستم شنوایی انسان نیست. برای هماهنگ کردن روش پیشگوی خطی با سیستم شنوایی انسان، ابتدا

هرمانسکی در سال ۱۹۸۹ روش پیشگوی خطی مبتنی بر در ک انسان را پیشنهاد کرد[۱۴]. هرمانسکی برای هماهنگی ویژگی ها با سیستم شنوایی انسان، آنالیز طیفی را به گونه ای انجام داد که بعضی از نواحی حساس تر از بقیه ی قسمت ها شوند. برای این منظور از مقیاسی شبیه مقیاس مل با نام مقیاس بارک استفاده نمود شکل زیر بلوک دیاگرام محاسبه آن می باشد.



شکل(۳-۳) بلوک دیاگرام پایه‌ی محاسبه ضرایب پیش گوی خطی مبتنی بر درک انسان[۱۴]

۳.۴.۱) آنالیز طیفی:

بعد از فریم بندی و اعمال پنجره، طیف توان زمان کوتاه مربوط به هر فریم با استفاده از تبدیل فوریه محاسبه می شود.

۳.۴.۲) آنالیز باند بحرانی:

بعد از محاسبه‌ی طیف توان، محور فرکانس با توجه به رابطه (۳-۱۳) به فرکانس بارک نگاشته می شود.

$$\Omega(w)=6\ln\left\{\frac{w}{1200\pi}+\left[\left(\frac{w}{1200\pi}\right)^2+1\right]^{0.5}\right\} \qquad (۱۳-۳)$$

که در این رابطه w فرکانس زاویه ای بر حسب رادیان بر ثانیه است. سپس طیف توان نگاشت داده شده، با نمودار باند بحرانی پوشاننده،$\psi(\Omega)$ کانوال می شود.این مرحله تقریبا شبیه به آنالیز کپسترال فرکانس مل است که در قسمت قبلی بیان شد رابطه‌ی مربوط به نمودار باند بحرانی به صورت رابطه‌ی (۳-۱۴) نشان داده شده است.

$$\psi(\Omega) = \begin{cases} 0 & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5} & -1.3 \le \Omega < -0.5 \\ 1 & -0.5 \le \Omega < 0.5 \\ 10^{-0.1(\Omega-0.5)} & 0.5 \le \Omega < 2.5 \\ 0 & \Omega \ge 2.5 \end{cases} \qquad (۱۴-۳)$$

۳.۴.۳) پیش تاکید مربوط به یکسان سازی بلندی صدا:

در این مرحله نمودار مربوط به یکسان سازی بلندی صدا، به رشته بدست آمده از مرحله‌ی قبلی اعمال می شود تابع استفاده شده در این مرحله، تخمینی از حساسیت غیر یکنواخت شنوایی انسان در فرکانس - های مختلف می باشد و حساسیت شنوایی انسان را در حدود ۴۰ دسی بل شبیه سازی می کند.

۳.۴.۴) قانون توان شدت بلندی صدا:

آخرین مرحله قبل از مدل کردن AR ، اعمال ریشه‌ی سوم به طیف برای تخمین قانون توان مربوط به سیستم شنوایی انسان می باشد. به عبارت دیگر در گوش انسان، میزان احساس بلندی صدا، با ریشه سوم انرژی آن متناسب است.

۳.۴.۵) مدل کردن AR:

در این مرحله برای بدست آوردن ضرایب کپسترال حاصل از پیشگویی خطی مبتنی بر درک انسان، ابتدا تبدیل فوریه‌ی معکوس گرفته می شود تا رشته‌ی خودهمبستگی بدست آید. سپس توسط معادله های یول-واکر، که یکی از روش های محاسبه‌ی ضرایب پیشگوی خطی هستند، فیلتر تمام قطب مدل می شود و ضرایب AR بدست می آید[۱۴].

۳.۴.۶) آنالیز کپسترال:

بعد از بدست آمدن ضرایب AR، ضرایب کپسترال مانند حالت LPCC با استفاده از معادله ی(۲-۶) به دست می آید. آنالیز باند بحرانی انجام شده در PLP ، شباهت زیادی به آنالیز فیلتر بانک مل در محاسبه- ی ضرایب MFCC دارد. برخی آزمایش های انجام شده نشان داده اند که روش MFCC نسبت به PLP

دارای نتایج بهتری در تشخیص گفتار است[۱۵]. اما روش PLP نسبت به تغییرات تعداد ضرایب و تعداد فیلترهای مورد استفاده در محاسبه‌ی ضرایب، دارای نتایج پایدارتری نسبت به MFCC می باشد [۱۶]. علاوه بر این PLP مقاومت بیشتری نسبت به نویز دارد.

۳.۵) ضرایب موجک

تبدیل موجک یک سیگنال را به سطوح فرکانسی پایین به بالا، در درجات تفکیک مختلف تجزیه می کند که به عنوان تبدیل چندتفکیکه[1] شناخته شده است. این خاصیت سبب تمایز روشنی بین مولفه های فرکانس پایین گفتار، در برگیرنده‌ی ویژگی های صوتی نظیر ضرب آهنگ است.

فیلتر موجک نیاز به تقسیم نمونه سیگنال در مقیاس کوچک ندارد و به عبارتی یک فیلتر بالاگذر می-باشد. برای فیلتر در هر مرحله می توان قسمت های تخمین (خروجی فیلتر پایین گذر) و جزئیات (خروجی فیلتر بالاگذر) را مشخص نمود.

شایستگی فیلتر موجک در فشرده سازی گفتار، به دلیل قابلیتش در متمرکز نمودن اطلاعات می باشد هنگامیکه گفتار، صدادار[2] و یا ترکیبی[3] است.یک فیلتر موجک مناسب باید بیش از ۹۰٪ انرژی گفتار صدادار را در نخستین N/2 ضرایب و حدود ۹۰٪ را در نخستین N/4 ضرایب متمرکز نماید. اما برای گفتار بی صدا تنها حدود ۳۵٪ ضرایب در نخستین N/2 و حدود ۸٪ در نخستین N/4 ضرایب و برای گفتار ترکیبی ۲۴٪ ضرایب در نخستین N/2 و ۱۹٪ در نخستین N/4 ضرایب موجود است.

به دلیل اینکه تبدیل‌های موجک محلی هستند، تفکیک پذیری در حوزه های زمان و فرکانس قابلیت جابه جایی دارند. این ویژگی ها آن ها را برای مشاهده ی سریع و موثر سیگنال در بازه ی زمانی خاص، موجه می سازد.

[1] Multi Resolution
[2] Voiced
[3] Mixed

در واقع تبدیل موجک از توابع پایه ی محلی استفاده می کند، بنابراین قابلیت تخمین سیگنال به نحوی مطلوب با تعداد کمی از جملات را دارا می باشد. به ویژه تبدیل موجک دابیچی[1]، فرم فشرده ای در حوزه ی زمان دارند،یعنی فقط تعداد محدودی از جملات برای ایجاد موجک مورد نیاز است. تبدیل موجک میر[2] دارای فرم فشرده در حوزه ی فرکانس بوده و موجک باتل-لماری[3]، هم در حوزه‌ی زمان و هم در حوزه‌ی فرکانس فرم فشرده را دارا می باشد [۱۷][۱۸][۱۹].

[1] Daubechies
[2]Meyer
[3]Battle-Lemarie

فصل چهارم

مروری بر کارهای انجام شده

فصل چهارم: مروری بر کارهای انجام شده

۴.۱) مقدمه:

مشکل عمده در به کارگیری ضرایب مل کپستروم آسیب پذیری و حساسیت بالای آن ها در محیط های نویزی می باشد. در بازشناسی گفتار مصون در برابر نویز نیاز به استخراج ویژگی هایی است که در برابر نویز حداکثر مصونیت را داشته باشند و ضرایب مل نیز به علت حساسیت بالا در محیط های نویزی نیازمند این مصون سازی هستند[۲۰]. بدین منظور محققین علم پردازش سیگنال با ایجاد تغییراتی در الگوریتم پایه ی استخراج آن ها، سعی در رسیدن به هدف فوق داشته و تا به امروز نیز این روند بهبود به سرعت در حال پیشرفت بوده است.

در این فصل برخی روش های مورد استفاده برای مصون سازی الگوریتم استخراج ضرایب مل-کپستروم را بررسی می کنیم.

۴.۲) مرور و طبقه بندی روش‌های استفاده شده در بهبود الگوریتم ضرایب مل-کپستروم:

ضرایب مل-کپستروم یکی از فراگیرترین ویژگی‌های مورد استفاده در سیستم های تشخیص گفتار است. اساسی ترین مشکل این ضرایب، آسیب پذیری و حساسیت زیاد آن ها در برابر نویز می باشد. برای دست -یابی به یک عملکرد رضایت بخش تحت شرایط نویزی در سیستم های تشخیص گفتار، نیاز به طراحی الگوریتم مصون در برابر نویز وجود دارد که موضوع کار محققین در این زمینه بوده است. در ابتدا بارز-ترین تحقیقات صورت گرفته در این زمینه ،کوتاه و تیتروار معرفی شده در ادامه به تفصیل پیرامون هریک بحث خواهد شد.

در سال ۱۹۹۹ کاربرد ضرایب خودهمبستگی در بهبود این الگوریتم مطرح شد. اساس این بود که تابع خود همبستگی نویز در بسیاری از موارد می توانست در طول زمان ثابت فرض شود بنابراین اثر نویز سیگنال ورودی پس از استخراج ضرایب خود همبستگی با عبور از یک فیلتر بالاگذر مناسب تا حدودی خنثی می شد،در ادامه این نوع ضریب را با نماد AMFCC (Autocorrelation MFCC) به کار خواهیم برد. در همان سال طرحی بیان شد که در آن روش فوق با اعمال فیلتری به نام RAS، به عنوان یک فیلتر بالاگذر پیاده سازی و نام RAS-MFCC ، (Relative Autocorrelation Sequences) را گرفت[۲۱][۲۲].
در سال ۲۰۰۱ توجه به آسیب پذیری بلوک لگاریتم اساس یکی از تغییرات در الگوریتم پایه‌ی استخراج ضرایب مل گشت. محققان در این سال با حذف این بلوک، بلوک ریشه گیری را جای گزین نمودند که علاوه بر مصون بودن در برابر نویز از پراکندگی و حذف اطلاعات جلوگیری می نمود، ROOT-MFCC [۲۳] [۲۴] .

در سال۲۰۰۴ نوعی دیگر از تغییرات شامل افزودن بلوکی مکمل به الگوریتم پایه معرفی گشت تکنیک به

کار رفته در این الگوریتم شامل تفریق میانگین طیفی بود که الگوریتم استاندارد را در برابر نویز مصون می ساخت SMN-CMN MFCC (Spectral Mean Normalization-Cepstral Mean Normalization) [۲۵] [۲۶].

در سال ۲۰۰۵ توجه به بلوک لگاریتم به عنوان یکی از بلوک های اصلی در الگوریتم ضرایب مل و آسیب پذیری آن در برابر انواع نویز به خصوص نویز کانولوشن مجددا در دستور کار محققین قرار گرفت و به جهت مصون سازی، یک تابع تکه‌ای- خطی که دارای همان خواص فشرده سازی بلوک لگاریتم بود و آسیب پذیری کمتری نسبت به نویز نیز داشت پیشنهاد گردید.و حذف بلوک لگاریتم صورت پذیرفت [۲۷] [۲۸] [۲۹].

در سال ۲۰۰۶ ضرایب خود همبستگی نویز، مجددا مورد توجه قرار گرفت و با فرض ناهمبسته بودن نویز و سیگنال اصلی، نشان داده‌شد که تخریب سیگنال توسط نویز در ضرایب مرتبه پایین تابع خودهمبستگی، بیشتر از ضرایب مرتبه بالا است بنابراین با حذف آن‌ها ، که آستانه ی آن به صورت سعی و خطا مشخص می شد، اثر نویز تاحدود زیادی کاسته شد[۲۱][۲۲][۳۰].

در سال ۲۰۰۹ به جای فیلتر بانک مثلثی از توابع گوسی استفاده شد که به علت ایجاد همبستگی بیشتر بین فریم ها موجب بهبود در نرخ تشخیص گشت. از این ضریب با علامت GMFCC (Gaussian MFCC) استفاده خواهیم کرد. در همان سال طرحی بیان شد که در آن به جای اینکه مثلث های به کار رفته در فیلتربانک از عرض کم به عرض زیاد تشکیل شوند به عکس، از عرض زیاد به عرض کم تشکیل می شدند، که در ترکیب با فیلتر بانک اصلی به علت توجه به اطلاعات مکمل نادیده گرفته شده در فرکانس های بالا عملکرد بهتری در کاربرهای تشخیص گفتار از خود نشان می داد[۳۱] [۳۲].

در سال ۲۰۱۰ ایده‌ی استفاده از ضرایب مرتبه بالای تابع خود همبستگی با یک تفریق فرکانسی تکمیل گشت و نوع دیگری ازضرایب مل-کپسترم مصون شده در برابر نویز را به نام DRHOASS-MFCC

(Differential Relative Higher Order Autocorrelation Sequence Spectrum) بوجود آورد [۲۲][۳۰][۳۳] .

در سال ۲۰۱۲ علاوه بر بخش نرم افزاری به بخش سخت افزاری در پیاده سازی ، این الگوریتم نیز توجه گردید به نحوی که با ایجاد تغییراتی در الگوریتم پایه و کم نمودن محاسبات ضرب و جمع در آن، گیت های منطقی برای پیاده سازی این الگوریتم کاهش چشمگیری یافت(MFCCE) [۳۴][۳۵].

در جدول (۴-۱) روند بهبود در این الگوریتم طی سال های ۱۹۹۹ تا ۲۰۱۲ بیان شده است. که مسیر تغییرات طی این سال ها به مسیر آتی کار در این تحقیق کمک شایانی می نماید.

جدول (۴-۱): سیر روش های ایجاد بهبود در الگوریتم ضرایب مل-کپستروم

انواع روش های به کار رفته در بهبود الگوریتم ضرایب مل-کپستروم	
RAS-MFCC & AMFCC	1999
ROOT-MFCC	2001
CMN-SMN-MFCC	2004
Piece wise LOG-MFCC	2005
GMFCC &IMFCC &IGMFCC	2009
DRHOASS-MFCC	2010
MFCC-E	2012
NEW WINDOWED MFCC	2012

روش های به کار رفته برای ایجاد بهبود در این الگوریتم را می توان به سه دسته کلی زیر تقسیم نمود :

- ❖ مدل های بهبود یافته شامل تغییر در بلوک های پایه ی این الگوریتم
- ❖ مدل های بهبود یافته شامل یک بلوک تکمیل کننده که به الگوریتم پایه اضافه گردیده است
- ❖ مدل های بهبود یافته در پیاده سازی سخت افزاری این الگوریتم توسط کاستن محاسبات ضرب و جمع در الگوریتم پایه

در ادامه با استفاده از تقسیم بندی پیشنهادی به بررسی هر یک از ضرایب مل-کپستروم بهبود یافته خواهیم پرداخت و تعلق روش های فوق را به این گروه بندی بررسی می نماییم.

۴.۳) ضرایب مل-کپستروم بهبود یافته بر اساس افزودن بلوک مکمل به الگوریتم پایه:

پیش از این نویز اضافی شونده به سیگنال معرفی گردید و گفتیم نویز اضافی شونده به سیگنال را می توان با معادله ی (۲-۱) نمایش داد که در آن s(n) سیگنال ورودی و d(n) نویز اضافی شونده به سیگنال است. منابع تولید نویز اضافی شونده به سیگنال که عملکرد سیستم تشخیص گفتار را تحت شعاع قرار می دهد بسیار است، از آن جمله می توان به :

- نویز ناشی از ترافیک
- نویز ناشی از وسایل گرمایشی سرمایشی
- نویز ناشی از جابجایی وسایل اتاق و... اشاره نمود.

ویژگی این نویز با فرض ناهمبسته بودن[1] 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.

معادله (۴-۱) روابط بین تابع خودهمبستگی را نشان می دهد که حاصل اعمال تابع خودهمبستگی به معادله ی (۲-۱) می باشد :

$$R_{xx(m,k)}=R_{ss(m,k)}+R_{dd(m,k)} \qquad (۱-۴)$$

که در آن k شماره ی فریم، m شماره ی نمونه وRتابع خودهمبستگی می باشد با توجه به نکته یاد شده

[1] Uncorrelated

که تغییرات تابع خودهمبستگی سیگنال نویز نسبت به زمان بسیار ناچیز و گاه نزدیک به صفر است، می-توان معادله (۴-۱) را به صورت معادله (۴-۲) نمایش داد.

$$R_{xx(m,k)}=R_{ss(m,k)}+R_{dd(m)} \qquad (۴-۲)$$

معادله‌ی (۴-۲) این نکته را خاطر نشان می‌کند که تابع خودهمبستگی نویز ناهمبسته به سیگنال، مستقل از فریم می باشد و طبق این فرض معادله (۴-۱) به معادله ی (۴-۲) تبدیل خواهد شد.

با توجه به ناچیز بودن دامنه ی سیگنال $R_{dd(m)}$ ، می توان با این روش اثر نویز را تا حدودی کاهش داد . با این روش نوع خاصی از ضرایب مل به نام AMFCC [1] بدست خواهد آمد که ضرایب مل به جای استخراج از سیگنال اصلی از تابع یک طرفه ی خود همبستگی آن بدست می آید [۲۱][۲۲].

همان طور که گفته شد تابع خود همبستگی نویز دارای دو ویژگی است اول اینکه دامنه‌ی طیف آن بسیار کوچک بوده و دوم اینکه تغییرات آن نسبت به زمان کم می باشد. بنابراین کافیست با استفاده از ویژگی دوم، حاصل اتوکورلیشن را از یک فیلتر بالاگذر مناسب مانند فیلتر معادله ی (۴-۳) نیز عبور دهیم که به عبور نکردن سیگنال خودهمبستگی نویز به علت داشتن مولفه های فرکانسی پایین منتهی می‌گردد. نحوه ی محاسبه ی فیلتر بالاگذر به کاررفته در معادلات زیر آمده است : [۲۱][۲۲].

$$T_l=\sum_{-L}^{L} t^2 \qquad t=-L,-L+1,\ldots,L \qquad (۴-۳)$$

$$H(z)=\frac{1}{T_l}\sum_{-L}^{L} tz^t \qquad (۴-۴)$$

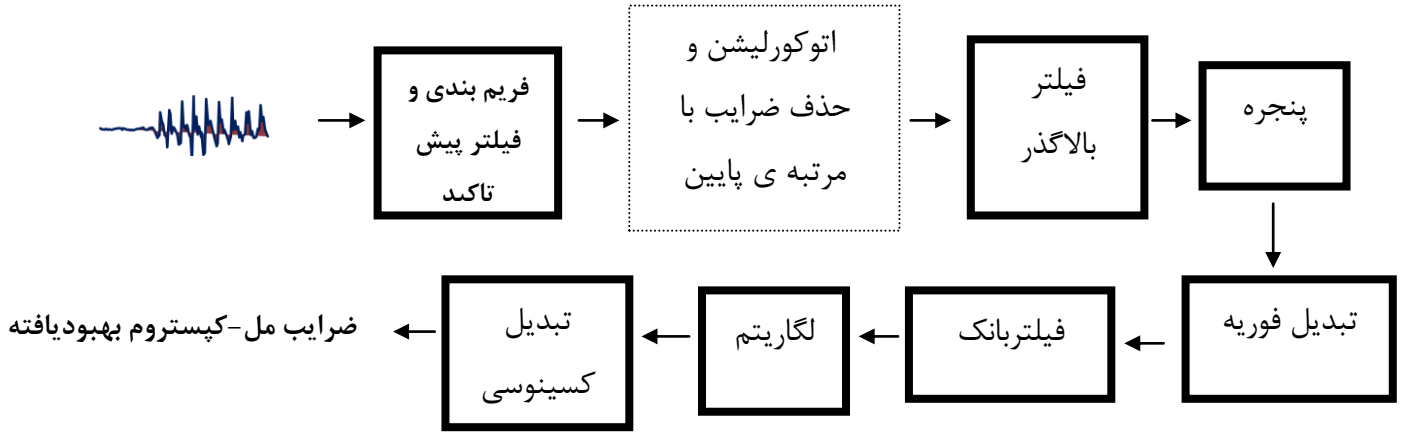
در این فیلتر L بیان‌گر طول فیلتر و نیز t بیان‌گر شماره ی نمونه هاست. معادله ی (۴-۳) نیز تابع تبدیل فیلتر در حوزه ی فرکانس می باشد.

[1] Autocorrelation Mel Frequency Cepstral Coefficients

همچنین از فیلتر بالاگذر دیگری با نام RAS-Filtre به عنوان فیلتر بالاگذر در این الگوریتم استفاده می-شود که ضرایب ایجاد شده توسط آن RAS-MFCC نام گرفت. معادله ی تابع تبدیل این فیلتر در ادامه آورده شده است :

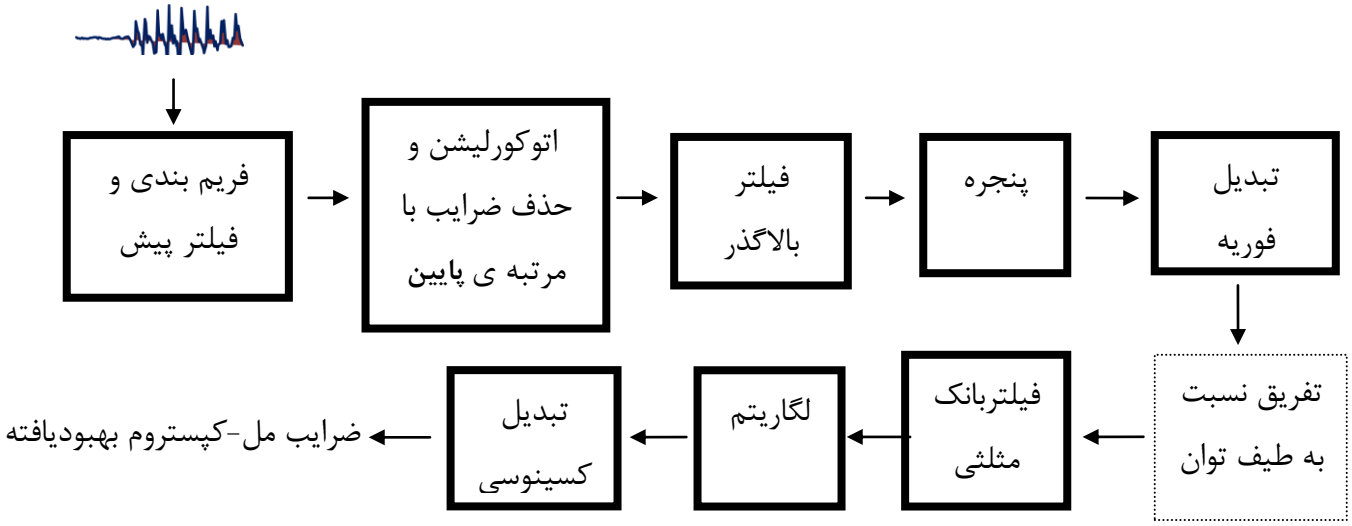
$$H_{RAS}=0.1(2z^{2}+z-z^{-1}-z^{-2}) \qquad (۵-۴)$$

در سال ۲۰۰۶ ضرایب خود همبستگی نویز، مجددا مورد توجه قرار گرفت و با فرض ناهمبسته بودن نویز و سیگنال اصلی، نشان داده شد که تخریب سیگنال توسط نویز در ضرایب با مرتبه پایین‌تر تابع خودهمبستگی، بیشتر از ضرایب با مرتبه بالاتر است، بنابراین با حذف آن‌ها، که آستانه‌ی آن به صورت سعی و خطا مشخص می شد، اثر نویز تا حدود زیادی کاسته شد. در این روش پس از فریم بندی و عبور از بلوک اتوکورلیشن ضرایب با مرتبه‌ی پایین که حامل بخش زیادی از اطلاعات نویزی هستند حذف می-گردند. بلوک دیاگرام این الگوریتم نیز در شکل (۴-۱) آمده است :



شکل (۴-۱) الگوریتم حاصل از حذف ضرایب مرتبه پایین[۲۱] [۲۲]

ایده‌ی توجه به ضرایب خودهمبستگی که در سال ۱۹۹۹ مطرح و مجددا در سال ۲۰۰۶ دنبال شده بود. در سال ۲۰۱۰ با افزوده شدن بلوک تفریق فرکانسی پیگیری شد و نوع دیگری ازضرایب مل-کپستروم مصون شده در برابر نویز را بوجود آورد. در این روش پس از بلوک تبدیل فوریه طیف توان سیگنال محاسبه می گردد، سپس نسبت به فرکانس عملیات تفریق انجام می دهیم، بدین معنا که از طیف توان سیگنال در این مرحله میانگین گرفته و میانگین طیف توان را از آن کم می نماییم، در سایر بخش های الگوریتم بلوک‌ها مانند بلوک‌های به کار رفته در الگوریتم سال ۲۰۰۶ می باشد این تغییرات منجر به ایجاد ضرایب جدیدی از مل گشت به نام DRHOASS-MFCC [1] بلوک دیاگرام این الگوریتم در شکل (۳-۲) آمده است . [۲۱][۲۲] [۳۰][۳۳].



شکل (۴-۲)بلوک دیاگرام الگوریتم DRHOASS-MFCC [۲۲] [۳۰] [۳۳]

[1] Differentiated Higher Order Relative Autocorrelation

در سال۲۰۰۴ نوعی دیگر از تغییرات شامل افزودن بلوکی مکمل به الگوریتم پایه معرفی گشت تکنیک به کار رفته در این الگوریتم شامل تفریق میانگین طیفی بود که الگوریتم استاندارد را در برابر نویز مصون می ساخت. MFCC SMN-CMN (Spectral Mean Normalization-Cepstral Mean Normalization)

اساس کار در این روش بر ناهمبسته بودن سیگنال گفتار ورودی و نیز سیگنال نویز می باشد.

در این صورت چنانچه سیگنال بدون نویز را s(m,k) در نظر گرفته و نویز اضافی شونده را d(m,k) در نظر بگیریم که در آن ها m شماره‌ی نمونه‌ی سیگنال و نیز k شماره ی فریم مربوطه باشد در این صورت داریم:

$$s(m,k)+ d(m,k)= x(m,k) \qquad (۶-۴)$$

از طرفی می توانیم سیگنال نویز را در طول زمان به صورت ثابت[1] فرض و مدل سازی نماییم بنابراین هر نمونه، مستقل از شماره‌ی فریم خواهد بود به عبارتی دیگر نویز مستقل از فریم خواهد بود یعنی معادله‌ی (۴-۶) را می توان به صورت معادله ی (۴-۷) بازنویسی نمود که در آن عبارت مربوط به نویز مستقل از فریم (m) بیان شده است:

$$s(m,k)+ d(k)= x(m,k) \qquad (۷-۴)$$

چنانچه به دو طرف معادله‌ی (۴-۷) تابع امید ریاضی را اعمال نماییم و از تابع مجموع سیگنال و نویز، کم کنیم به علت اینکه امید ریاضی یک عبارت ثابت با خودش برابر است خواهیم داشت:

$$\begin{aligned}\hat{x}(m,k) &= x(m,k)-E[x(m,k)] \\ &= \{s(m,k)+d(k)\}-\{E[s(m,k)+d(k)]\} \\ &= s(m,k)-E[s(m,k)] \\ &= \hat{s}(m,k)\end{aligned} \qquad (۸-۴)$$

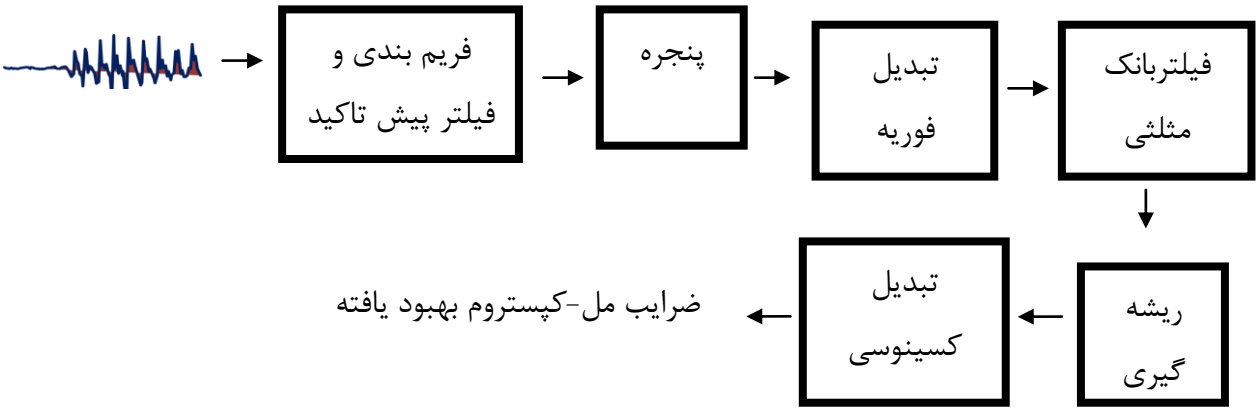
[1] Invariant

چنانچه از معادله ی فوق پیداست سیگنال حاصل از این فرایند بدون نویز خواهد بود و عامل نویز d حذف شده است. [۲۵][۲۶].

۴.۴) ضرایب مل-کپستروم بهبود یافته بر اساس تغییر در بلوک‌های پایه‌ی الگوریتم:

در سال ۲۰۰۱ توجه به آسیب پذیری بلوک لگاریتم اساس یکی از تغییرات در الگوریتم پایه‌ی استخراج ضرایب مل گشت. محققان در این سال با حذف این بلوک، بلوک ریشه گیری را جای گزین نمودند که علاوه بر مصون بودن در برابر نویز از پراکندگی و حذف اطلاعات جلوگیری می نمود، ROOT-MFCC .

در شکل (۴-۳) بلوک دیاگرام محاسبه ی این الگوریتم نمایش داده شده است.



شکل (۴-۳) بلوک دیاگرام الگوریتم ROOT-MFCC [۲۳] [۲۴]

در بلوک ریشه گیری عملگر $(.)^{\alpha}$ به کار می رود و داریم $0 <\alpha< 1$ که در آن مقدار مناسب برای α در آزمایش‌های تشخیص گفتار به صورت سعی و خطا مشخص شده است[۲۳][۲۴].

در سال ۲۰۰۵ به مشکلات ایجاد شده در بحث آسیب پذیری الگوریتم استخراج ضرایب مل-کپستروم ناشی از بلوک لگاریتم پرداخته شد. تابع لگاریتم نسبت به نویز از حساسیت بالایی برخوردار است و پس از محاسبه‌ی انرژی خروجی فیلتربانک ها، در انرژی های کم این آسیب پذیری بسیار بیشتر دیده می شود.با توجه به همین مطلب یک تابع جای‌گزین معرفی شد تا فشرده سازی را در این مناطق اسیب پذیر، بیشتر انجام دهد در این تابع جای‌گزین که یک تابه تکه‌ای خطی می باشد که در عین حال پیوسته نیز هست تابع لگاریتم با یک عبارت توانی جای گزین شده است.

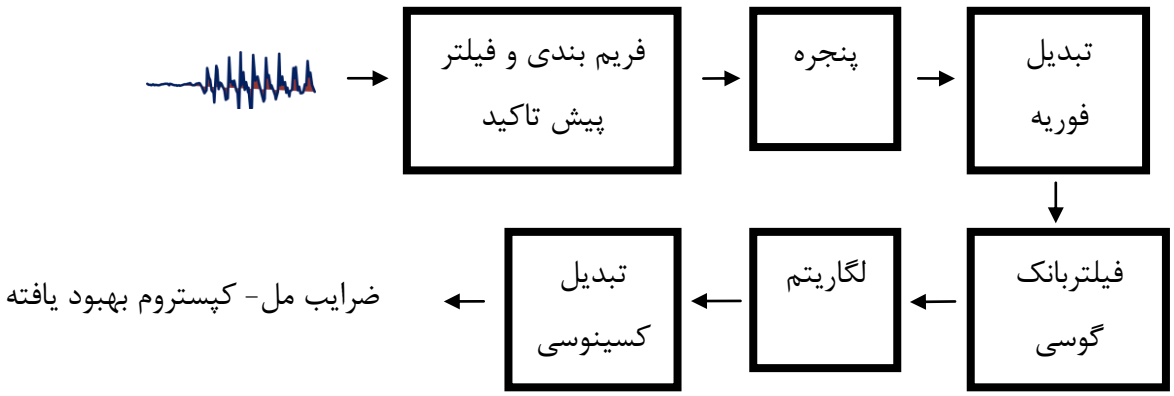
در این تابع که معادله‌ی آن در صفحه‌ی بعد نمایش داده شده است λ ضریب فشردگی و c نیز ضریب پوشش نویز نام دارد معمولا λ را برابر ۲ و نیز c را برابر ۵×۱۰^{۶} در نظر می گیریم[۲۷][۲۸][۲۹].

$$\begin{cases} (\lambda\, x^{\frac{1}{\lambda}})/(c^{\frac{1}{\lambda}}) & x\leq c \\ \text{Log } x + \lambda - \log c & x>c \end{cases} \qquad (۹-۴)$$

همان‌طور که اشاره شد، در سال ۲۰۰۹ به جای فیلتربانک مثلثی در الگوریتم پایه از توابع گوسی استفاده شد. در فیلتربانک مثلثی اطلاعات بخش هایی از فریم که در نقاط ابتدایی و انتهایی و خارج از زیر بخش-ها، قرار می گیرند از دست می روند زیرا مثلث ها در خارج از زیر باند ها وزنی ندارند اما در فیلتربانک به

کار رفته که یک فیلتربانک گوسی متقارن می باشد در خارج از زیر باند ها وزن وجود دارد و مانع از دست رفتن اطلاعات در این بخش ها می گردد. مزیت دیگر آن نسبت به فیلتربانک مثلثی شروع و اتمام آن با شیب کم تر و ملایم تر می باشد. این فیلتربانک با ایجاد همبستگی بیشتر بین زیرباندها اطلاعات از دست رفته در مرزها را کاهش داده و در بهبود الگوریتم و متعاقبا بالا بردن نرخ بازشناسی سیستم اتوماتیک تشخیص گفتار موثر بوده است. روش ساخت فیلتربانک گوسی از بعد پارامتر ها تاحدود زیادی مانند فیلتربانک مثلثی است با این تفاوت که kb ها که نقاط مرزی را در فیلتربانک مثلثی مشخص می کردند، در این فیلتربانک در تعیین پارامتر سیگما کاربرد دارند و به بیان دیگر در تعیین پراکندگی هر گوسی مشخص کننده هستند. و نهایت امر این فیلتربانک با استفاده از دو معادله که در معادله ی (۴-۱۰) بیان شده است، ساخته می شوند و بلوک دیاگرام مربوط به این الگوریتم نیز در شکل(۴-۴) آمده است. با این روش نوع دیگری از ضرایب مل به نام GMFCC[1] بدست آمد[۳۱][۳۲] :

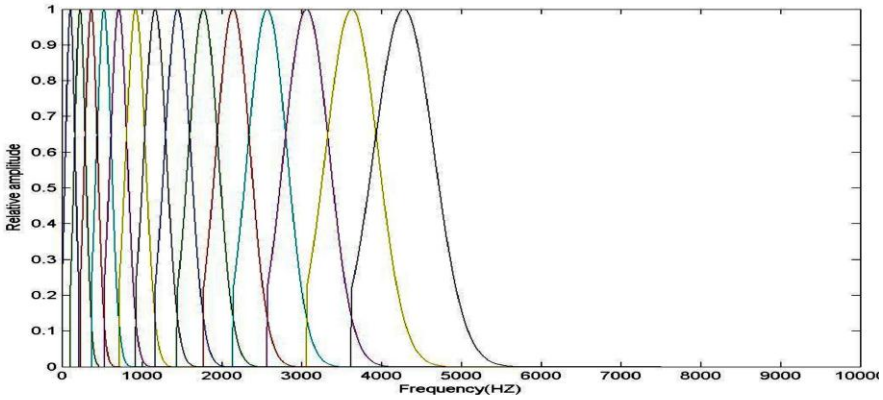
$$\sigma_i=\frac{kb_i-kb_{i-1}}{2} \qquad \Psi_i=e^{\frac{-(k-kb_i)}{2\sigma_i^2}} \qquad (۴-۱۰)$$



شکل (۴-۴)بلوک دیاگرام الگوریتم GMFCC [۳۱] [۳۲]

[1] Gaussian Mel-Frequency Cepstral Coefficient

فیلتربانک گوسی پس از کدنویسی و شبیه سازی در نرم افزار متلب به شکل (۴-۵) خواهد بود :

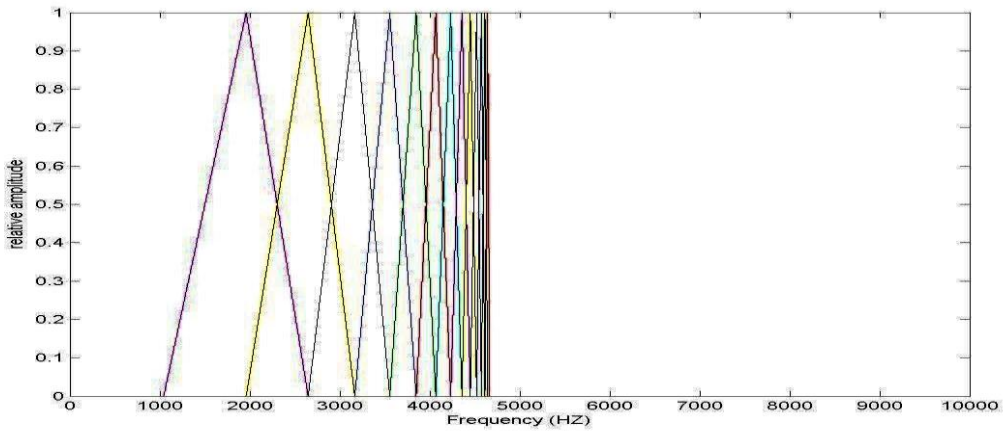


شکل(۴-۵) فیلتربانک گوسی

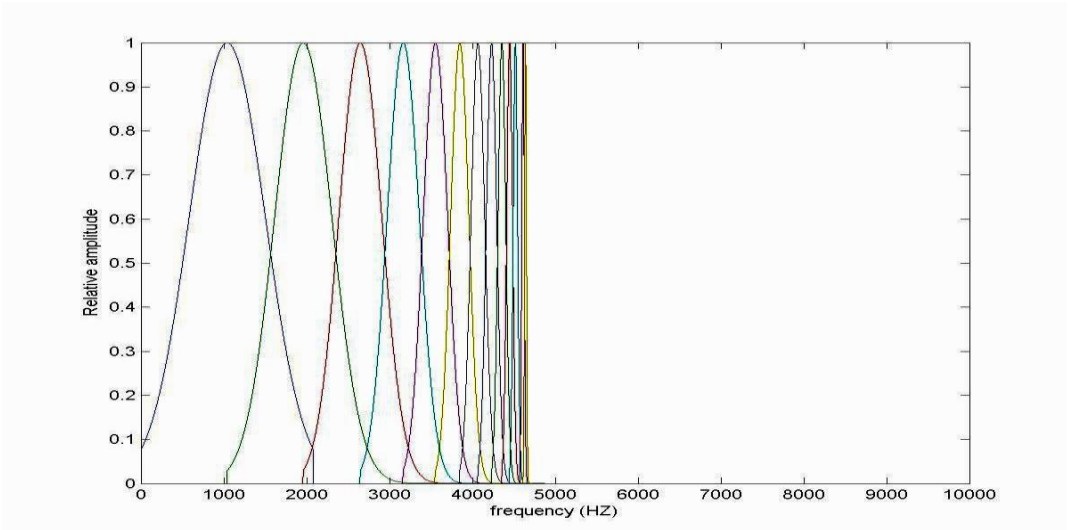
مجددا در همان سال، (۲۰۰۹)، طرحی بیان شد که در آن به جای این که مثلث های به کار رفته در فیلتربانک از عرض کم به عرض زیاد تشکیل شوند به عکس، از عرض زیاد به عرض کم تشکیل می شدند که در ترکیب با فیلتربانک اصلی به علت توجه به اطلاعات مکمل نادیده گرفته شده در فرکانس های بالا عملکرد بهتری در کاربردهای تشخیص گفتار از خود نشان می دادند[۳۱][۳۲].

الگوریتم های حاصل از به کار رفتن فیلتربانک های معکوس به ترتیب" IMFCC"[۱] و نیز "IGMFCC"[۲] می باشد.فیلتربانک های معکوس شبیه سازی شده در نرم افزار متلب در شکل های (۴-۶) و (۴-۷) نمایش داده شده است:

[1] Inverse Mel Frequency Cepstral Coefficient
[2] Inverse Gaussian Mel Frequency Cepstral Coefficient



شکل (۴-۶) فیلتربانک با توزیع معکوس مثلثی



شکل(۴-۷) فیلتربانک با توزیع معکوس گوسی

در سال ۲۰۱۲ استفاده از یک پنجره ی بهبود یافته در زمره‌ی پیشرفت های انجام شده پیرامون این الگوریتم بود.

در مرحله‌ی اعمال پنجره در الگوریتم پایه تغییری صورت پذیرفت که بواسطه‌ی آن به جای استفاده از پنجره‌ی رایج همینگ از یک پنجره‌ی جدید استفاده شد که موجب ایجاد بهبود در نرخ بازشناسی گردید.

در پنجره‌ی به کار رفته در این تحقیق سه پارامتر مهم در نظر گرفته شده و بررسی شده است:

۱. پارامتر پراکندگی

۲. همگرایی بخش جانبی[1]

۳. عرض بخش اصلی پنجره[2]

در این پنجره که معادله‌ی آن در ادامه بیان خواهد شد، نسبت به یک پنجره‌ی همینگ ساده فاکتور پراکندگی طیفی و نیز عرض وجه اصلی افزایش و فاکتور همگرایی وجه های جانبی کاهش می یابد که دو مورد اول تغییراتی مطلوب و مورد آخر نامطلوب می باشد.

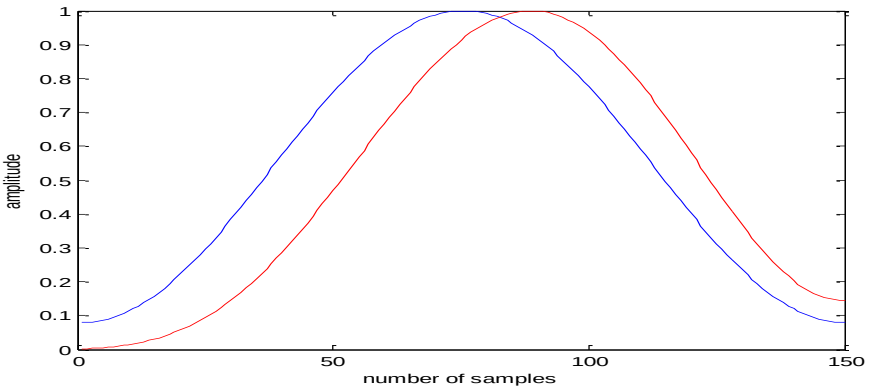
تغییرات انجام شده، حاکی از ایجاد بهبود در الگوریتم و متعاقبا بهبود در نرخ تشخیص گفتار می باشد و بنابراین از نقص ایجاد شده نسبت به دو مزیت فوق می توان چشم پوشی کرد.

موارد ذکر شده در شکل(۴-۸) در قیاس با یک پنجره ی همینگ ساده قابل رویت می باشد.

معادله‌ی پنجره‌ی بهبود یافته نیز در معادله ی(۴-۱۱) بیان شده است.

$$w_{new}(n)=\mathrm{n}\ w_{hamming}(\mathrm{n}) \qquad (۱۱-۴)$$

[1] Side Lobe
[2] Main Lobe



شکل (۴-۸) پنجره‌ی همینگ و همینگ بهبود یافته[۳۶]

همانطورکه در تصویر مشخص است پنجره‌ی تغییر یافته دارای بخش جانبی بیشتری می باشد در حالیکه پنجره‌ی همینگ ساده دارای تقارن در دو بخش جانبی می باشد و عرض وجه اصلی آن نسبت به پنجره ی تغییر یافته کمتر شده است.

نتایج یک نمونه از این پنجره که در نرم افزار متلب شبیه سازی شده است در حالت مقایسه در جدول نشان داده شده است[۳۶].

جدول (۴-۲) مقایسه‌ی پارامترهای پنجره‌ی بهبود یافته با پنجره‌ی همینگ ساده در یک نمونه شبیه سازی [۳۶]

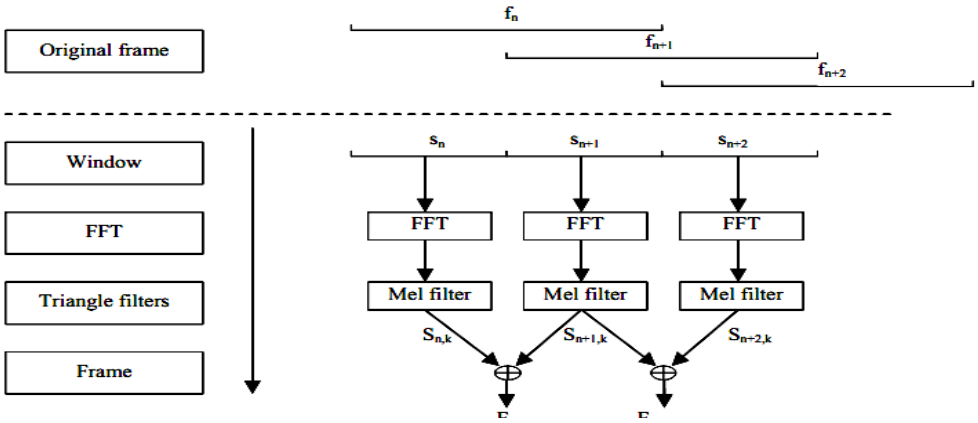
پنجره	پارامتر پراکندگی	همگرایی بخش جانبی	عرض بخش اصلی پنجره
همینگ	۰/۰۴٪	۴۲.۶ *dB*	۰/۰۱۵*cm*
پنجره‌ی بهبود یافته	۰/۱۷٪	۳۷.۹ *dB*	۰/۰۱۸*cm*

۵.۴) ضرایب مل-کپستروم بهبودیافته بر اساس تغییر در پیاده سازی سخت افزاری

این الگوریتم :

همان طور که بیان شد در سال ۲۰۱۲ مقوله‌ی کاهش در حجم سخت افزاری این الگوریتم مطرح شد. در این روش با کاهش محاسبات به کار رفته در الگوریتم پایه بوسیله‌ی انجام فریم بندی در دو مرحله، موفق به کاهش حجم محاسبات و در نتیجه کاهش گیت های منطقی و سهولت در پیاده سازی گردیدند.

الگوریتم به کاررفته در این طرح در شکل(۳-۱۰) نمایش داده شده است :



شکل (۴-۹) بلوک دیاگرام الگوریتم MFCCE [۳۴]

چنانچه مشاهده می گردد هر دو خروجی از فیلتربانک در این الگوریتم تبدیل به یک خروجی شده است که این ایده منجر به کاهش پنجاه‌درصدی در گیت‌های منطقی در روند پیاده سازی خواهد شد[۳۴][۳۵].

فصل پنجم

روش پیشنهادی

فصل پنجم: روش پیشنهادی

۵.۱)مقدمه:

در این فصل به معرفی روش پیشنهادی پرداخته و عملکرد سیستم را بیان می نماییم.

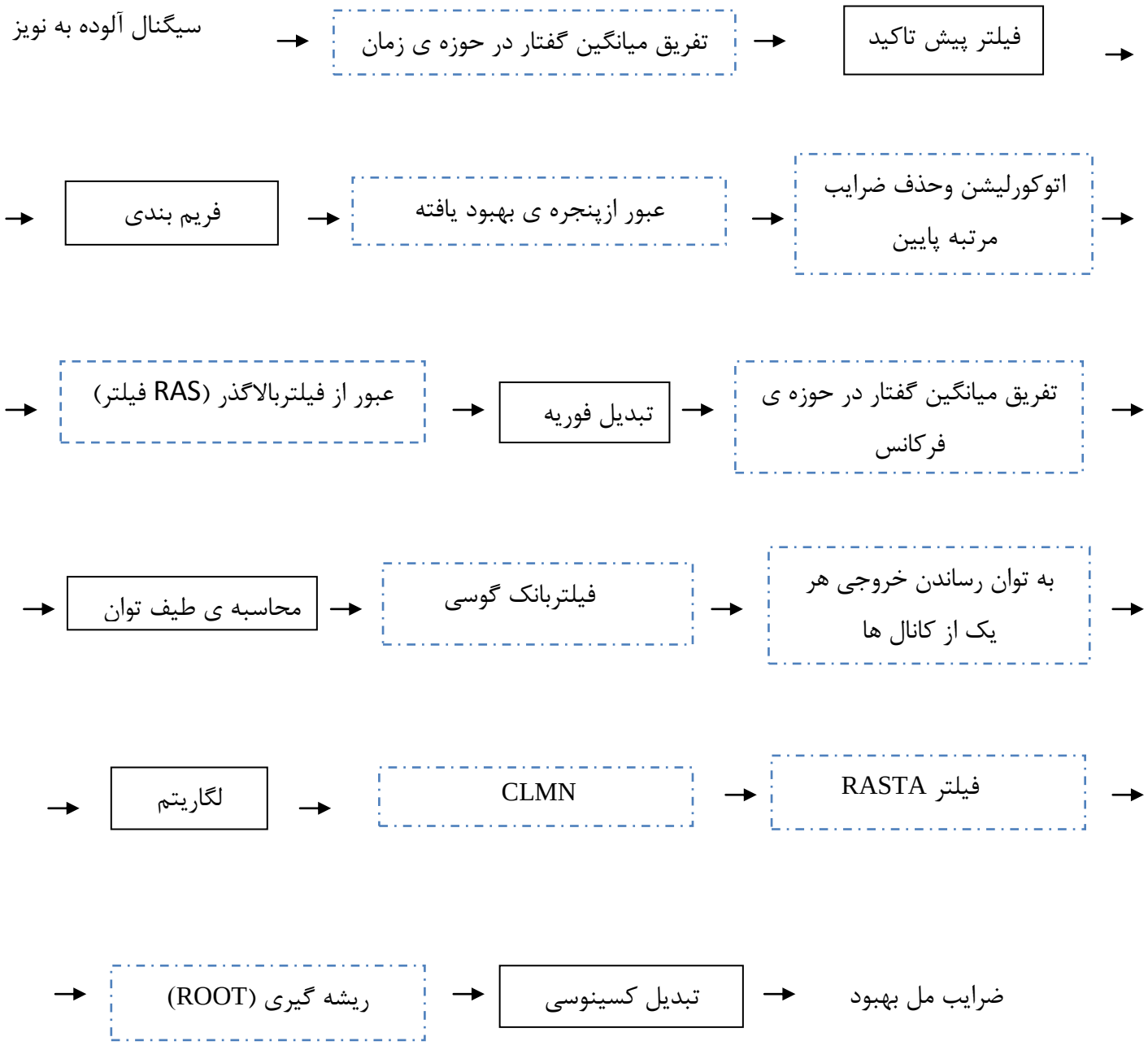
پس از آن که در فصول دوم و سوم روش های مطالعه پیرامون بازشناخت گفتار در محیط نویزی و نیز استخراج ویژگی های مقاوم در برابر نویز که مورد نیاز در کاربرد های بازشناخت مقاوم گفتار فارسی می-باشد، بیان گردید و به صورت خاص ضرایب مل-کپستروم و سیر پیشرفت های حاصله در مقاوم سازی این الگوریتم به عنوان یکی از پرکاربردترین ویژگی های مورد استفاده در کاربردهای بازشناخت گفتار بررسی گردید، به جمع بندی و ارائه ی روش پیشنهادی خواهیم پرداخت.

همان‌طور که در طبقه بندی روش های مقاوم سازی الگوریتم ضرایب مل-کپستروم اشاره گردید، پیشرفت های حاصل در سه گروه طبقه بندی شد، آنچه در این پایان نامه مد نظر بوده است بحث های شبیه سازی در نرم افزار متلب و نیز کدنویسی های مربوطه است در نتیجه در الگوریتم پیشنهادی به مورد پیاده سازی سخت افزاری و کاهش محاسبات ضرب و جمع توجه نگردیده است وتنها به دو مورد اول یعنی ایجاد بهبود شامل تغییر در بلوک های پایه ی این الگوریتم و نیز افزودن بلوک مکمل به الگوریتم پایه اشاره شده است.

چنانچه از مطالعات پیشین منتج گردید جهت مصون سازی ضرایب مل-کپستروم نیاز به ایجاد تغییرات مطلوب در بلوک های پایه ی این الگوریتم و نیز ایجاد تغییرات مطلوب در افزودن بلوک های مکمل به الگوریتم پایه است.

۵.۲)بررسی روش پیشنهادی:

پس از مطرح نمودن توضیحات ابتدایی بلوک دیاگرام الگوریتم پیشنهادی در صفحه ی بعد بیان و سپس به بررسی بخش های مختلف آن خواهیم پرداخت.



شکل(۵-۱) : بلوک دیاگرام الگوریتم پیشنهادی

در ابتدا سیگنال های ورودی که به صورت ۸۰۰ لغت[1] شامل ۲۰ کلاس و در هر کلاس ۴۰ تکرار، پس از طبقه بندی و نام گذاری جهت استفاده در نرم افزار متلب در یک فایل مناسب قرار داده شده و توسط دستورهای مربوطه خوانده خواهد شده و سپس آماده ی پردازش می شوند.

۵.۲.۱)حذف سکوت از گفتار:

سیگنال گفتار معمولا از دو ناحیه ی سکوت و غیر سکوت تشکیل می شود که در سیگنال گفتار حاوی نویز این ناحیه‌ی سکوت نویزی است. تشخیص ناحیه‌ی سکوت در سیگنال گفتار اهمیت بسزایی دارد و در کاربردهای نظیر تشخیص گفتار، فشرده سازی اطلاعات گفتار، تخمین و حذف نویز استفاده می شود. برای مثال، در سیستم بهسازی گفتار در صورتی که بتوانیم نواحی سکوت را از گفتار تشخیص دهیم و جداسازی کنیم، می توانیم اطلاعات خوبی از نویز ترکیب شونده با گفتار بدست آوریم.

در سیستم های تشخیص گفتار یا تشخیص گوینده می‌توان با مشخص نمودن نواحی سکوت، انجام عملیات تشخیص را فقط به نواحی غیر سکوت محدود کرد و از هزینه های سربار اضافی کاست.

در سیستم های تشخیص گفتار گاهی سکوت نیز مانند گفتار مدل می شود بدین معنا که سیگنال سکوت نیز مانند سیگنال گفتار مراحل استخراج ویژگی و عملیات دیگر را طی می نماید. در این گونه سیستم ها معمولا به بخشی برای تشخیص گفتار از سکوت نیار نداریم، مگر اینکه برای مثال بهسازی گفتار در مرحله ی اول انجام شود و برای بدست آوردن طیف سکوت به الگوریتمی برای تشخیص گفتار از سکوت احتیاج داشته باشیم.

در بعضی از سیستم ها نیز برای سکوت مدلی در نظر گرفته نمی شود، بلکه ابتدا با الگوریتم قسمت هایی از سیگنال صوتی حذف می شوند و پردازش های بعدی مانند استخراج ویژگی و دسته بندی بردار ویژگی فقط روی سیگنال گفتاری جدا شده انجام می گیرد در این دسته از سیستم ها، که در روش پیشنهادی از

[1] Isolated Word

آن ها بهره برده‌ایم، میزان محاسبات نسبت به دسته ی قبلی کاهش می یابد. ولی البته به دلیل وجود خطا در الگوریتم تشخیص گفتار از سکوت، ممکن است بر راندمان سیستم تاثیر بگذارد.

۵.۲.۲) تفریق میانگین گفتار در حوزه ی زمان:

هدف از پیاده سازی الگوریتم پیشنهادی کاربرد آن در محیط‌های نویزی و به خصوص نویز اضافی شونده می باشد و تشخیص گفتار در محیط نویزی باید صورت بپذیرد. بنابراین پس از افزودن نویز سفید گوسی[1] به صورت مصنوعی به سیگنال بدون نویز ورودی، در نرخ سیگنال به نویز دلخواه (که مقدار آن در ابتدا از کاربر خواسته شده) آن را از بلوک تفریق میانگین گفتار عبور می دهیم اساس کار این بلوک بر ناهمبسته بودن[2] نویزاضافی و سیگنال ورودی است. در این صورت چنانچه سیگنال بدون نویز را s(m,k) در نظر گرفته و نویز اضافی شونده را d(m,k) در نظر بگیریم که در آن ها m شماره‌ی نمونه‌ی سیگنال و نیز k شماره‌ی فریم مربوطه باشد نحوه‌ی ترکیب را در معادله‌ی (۲-۱) بیان نمودیم.

از طرفی می‌توانیم سیگنال نویز را در طول زمان به صورت ثابت فرض و مدل سازی نماییم بنابراین هر نمونه، مستقل از شماره‌ی فریم خواهد بود که معادله ی (۴-۷) مطلب فوق را یادآوری می نمود.

و در اثر اعمال تابع امید ریاضی به دو طرف معادله ی (۴-۷) و سپس کم نمودن از تابع مجموع سیگنال و نویز، به علت اینکه امید ریاضی یک عبارت ثابت با خودش برابر است خواهیم داشت:

$$\begin{aligned}\hat{x}(m,k) &= x(m,k) - E[x(m,k)] \\ &= \{s(m,k)+d(k)\} - \{E[s(m,k)+d(k)]\} \\ &= s(m,k) - E[s(m,k)] \\ &= \hat{s}(m,k)\end{aligned} \qquad (۱-۵)$$

[1] Artificial White Gaussian Noise (WGN)
[2] Uncorrelated

چنانچه از معادله‌ی (۵-۱) مشخص است در اثر این فرایند که در ابتدا در حوزه‌ی زمان انجام می شود می توان تا حدود زیادی نویز اضافی را که با حرف d نمایش میدادیم حذف نمودیم.

۵.۲.۳) فیلتر پیش تاکید:

برای استخراج ویژگی های مورد نیاز در هر فرایند پردازش گفتار معمولا ابتدا سیگنال گفتار به پنجره های کوتاهی با طول مثلا ۲۰ ms که با هم همپوشانی دارند تقسیم می گردد اما گاه پیش از این مرحله هر پنجره مرحله‌ی پیش تاکید را می گذراند سپس پنجره بندی و پس از آن فیلتر می شود تا بسته به کاربرد، اطلاعات غیر ضروری یا اضافی نظیر نویزهای موجود را رفع نماید. فیلتر کردن می تواند نسبت سیگنال به نویز را بهبود بخشد سپس بسته به روش استخراج ویژگی، اعمال لازم انجام شود.

دلیل انجام مرحله‌ی پیش تاکید این است که طیف سیگنال تحریک واکدار خروجی از حنجره شیب ۱۲ dB/oct- دارد. تابع تبدیل اثر تشعشع لب ها، فیلتر بالاگذری با مشخصه ی فرکانسی ۶ dB/oct استفاده می شود. این فیلتر فیلتر پیش تاکید نام دارد. پاسخ ضربه ی این فیلتر از رابطه ی زیر بدست می آید.

$$P(z)= 1-\alpha z^{-1} \qquad (۲-۵)$$

مقدار α در این تابع عددی بین صفر تا یک انتخاب می گردد این مقدار در واقع صفر تابع است .

۵.۲.۴) فریم بندی:

سپس در قدم بعد گفتار پیش تاکید شده وارد مرحله‌ی فریم بندی می گردد که طول آن ها در این پایان نامه برابر ۴۴ میلی ثانیه می باشد و پنجاه درصد نیز روی هم افتادگی دارند.

۵.۲.۵)عبور از پنجره ی بهبود یافته:

آنچه در الگوریتم پایه محاسبه‌ی ضرایب مل انجام می شود عبور سیگنال پیش تاکید از یک پنجره‌ی همینگ ساده است اما در این تحقیق ما از یک پنجره‌ی همینگ تغییر یافته استفاده می کنیم که از معادله‌ی (۵-۳) حاصل می گردد.

$$\widehat{w}=nw(n) \qquad (۳-۵)$$

پیش از این در بخش مروری بر روش ها نیز این پنجره ی همینگ تغییر یافته معرفی گردیده بود و بیان شد که در پنجره‌ی به کار رفته در این الگوریتم، سه پارامتر مهم در نظر گرفته شده است:

-پارامتر پراکندگی

-همگرایی بخش ها جانبی (Side lobe)

-عرض بخش اصلی پنجره (Main lobe)

دانستیم که در این پنجره نسبت به یک پنجره‌ی همینگ ساده فاکتور پراکندگی طیفی و نیز عرض وجه اصلی افزایش و فاکتور همگرایی وجه های جانبی کاهش می یابد که دو مورد اول تغییراتی مطلوب و مورد آخر نامطلوب می باشد.

پس از به کاربردن این پنجره و نیز پنجره‌ی همینگ ساده و انجام آزمایش های تشخیص گفتار با ثابت در نظرگرفتن سایر پارامترها و متغیر در نظر گرفتن پنجره‌ی به کار رفته، این نتیجه حاصل شد که تغییرات، حاکی از بهبود نتایج می باشد و بنابراین از نقص ایجاد شده در مقابل دو مزیت فوق چشم پوشی کردیم و در الگوریتم پیشنهادی از پنجره ی فوق در مرحله‌ی پنجره بندی استفاده نمودیم.

۵.۲.۶) اتوکورلیشن و حذف ضرایب مرتبه پایین:

تابع خودهمبستگی[1] عبارت است از ضرب سیگنال در انتقال همان سیگنال در حالت کلی خودهمبستگی تاخیر k، سیگنال y با رابطه ی زیر محاسبه می شود (N تعداد کل طول سیگنال می باشد) :

$$R_{yy}(k)=1/N \sum_{m=0}^{N-1} y(m) * y(m-k) \qquad (۴-۵)$$

چنانچه پیشتر در فصل مروری بر روش ها بیان گردید تابع خودهمبستگی دارای این خاصیت است که نویز ضرایب خودهمبستگی سیگنال را بیشتر در تاخیرهایی با مرتبه پایین تخریب می نماید در حالی که تاخیرهایی با مرتبه‌های بالاتر یعنی ضرایب اتوکورلیشن با مرتبه های بالاتر کمتر تحت تخریب نویز اضافی شونده قرار می گیرند.

بنابراین با حذف ضرایب مرتبه پایین تابع خود همبستگی نویز تا حدود زیادی از بین خواهد رفت و جهت تخمین طیف تابع از سیگنال اتوکورلیشن با ضرایب مرتبه بالا استفاده خواهیم نمود. بنابراین خواهیم داشت:

$$R_{xx}(m,k)=R_{xx}(m,k) \quad \text{if} \quad D\le k\le N-1 \quad \text{and} \quad R_{xx}(m,k)=0 \quad \text{if } 0\le k\le D \qquad (۵-۵)$$

[1] Autocorrelation

که در آن k شماره ی فریم، m شماره ی نمونه، R تابع خودهمبستگی و N طول سیگنال خودهمبستگی می باشد.

همچنین بالاترین مرتبه تابع خودهمبستگی (D) جهت حذف با استفاده از آزمایش های تشخیص گفتار بدست آمده‌است که در ادامه و در فصل ششم به تفصیل به همراه جداول مربوط به آزمایش‌های تشخیص گفتار بیان خواهد گردید.

۵.۲.۷) عبور از فیلتربالاگذر (RAS فیلتر):

تحقیقات محققین پردازش سیگنال پیرامون تابع خودهمبستگی حاکی از این اصل بود که می توان تابع خودهمبستگی نویز را در بسیاری از موارد تقریبا در طول زمان ثابت در نظر گرفت و با استفاده از یک فیلتر بالاگذر که تغییرات زیاد (فرکانس بالا) را عبور می دهد و تغییرات کم (فرکانس پایین) از آن عبور نمی کند، می توان تا حدود زیادی از اثر نویز کاست و فیلتر بالاگذر مانع عبور سیگنال نویز به علت داشتن فرکانس پایین، می گردد.

یکی از بهترین فیلترهای بالاگذر از نظر عملکرد، که اولین بار در الگوریتم RAS-MFCC پیشنهاد و به کار رفته است، به فیلتر بالاگذر RAS شهرت یافته است[۲۱] که تابع تبدیل آن در رابطه ی (۵-۶) بیان شده است :

$$H_{RAS=}0.1(2z^{2}+z-z^{-1}-z^{-2}) \qquad (۶-۵)$$

۵.۲.۸)تبدیل فوریه و تفریق میانگین گفتار در حوزه ی فرکانس:

پس از فیلتر نمودن خروجی‌های تابع خودهمبستگی، تبدیل فوریه گرفته شده و سپس ایده‌ی تفریق طیفی در حوزه‌ی فرکانس انجام می گردد. تبدیل فوریه‌ی سیگنال های x,s,d را که پیش‌تر به عنوان سیگنال‌های نویز، بدون نویز و سیگنال مجموع نویز و نویز اضافی شونده معرفی شدند، به ترتیب با X,S,D نمایش می دهیم بنابراین خواهیم داشت:

$$X(m,k)= S(m,k)+ D(k) \qquad (۷-۵)$$

$$\begin{aligned}\widehat{X}(m,k) &= X(m,k)-E[X(m,k)] \\ &= \{S(m,k)+D(k)\}-\{E[S(m,k)+D(k)]\} \\ &= S(m,k)-E[S(m,k)] \\ &= \hat{S}(m,k)\end{aligned} \qquad (۸-۵)$$

در اثر این عملیات که ایده‌ی آن همان ایده‌ی مورد استفاده در تفریق میانگین در حوزه‌ی زمان می باشد و با تکیه بر این اصل که امید ریاضی یک تابع ثابت برابر خود آن تابع می باشد، (در اینجا تابع ثابت همان نویز اضافی شونده (D(k)) می باشد) بخشی از مولفه‌های نویز در حوزه‌ی فرکانس حذف و جهت مقاوم نمودن الگوریتم پایه‌ی ضرایب مل قدم دیگری برداشته خواهد شد.

پس از انجام این عملیات طیف توان خروجی محاسبه و سپس وارد فیلتر بانک گوسی می گردد.

۵.۲.۹)فیلتربانک گوسی:

همانطورکه در فصل چهارم، مروری بر روش ها ، بیان شد در الگوریتم پایه محاسبه ضرایب مل-کپستروم معمولا از یک فیلتر بانک مثلثی استفاده می گردد. در این نوع فیلتربانک اطلاعات بخش هایی از فریم که

در نقاط ابتدایی و انتهایی و خارج از زیر بخش‌ها، قرار می گیرند، از دست می روند، زیرا مثلث‌ها در خارج از زیر باندها وزنی ندارند. اما اگر به جای این فیلتربانک از یک فیلتربانک با گوسی های متقارن استفاده کنیم مزیت های فوق را خواهیم داشت :

- ✓ به دلیل وجود وزن، در خارج از زیرباندهای آن ، مانع از دست رفتن اطلاعات در این بخش ها می گردد.
- ✓ مزیت دیگر آن نسبت به فیلتر بانک مثلثی شروع و اتمام آن با شیب کمتر و ملایم‌تر می باشد.
- ✓ این فیلتر بانک با ایجاد همبستگی بیشتر بین زیرباندها اطلاعات از دست رفته در مرزها را کاهش داده و در بهبود الگوریتم و متعاقبا بالا بردن نرخ بازشناسی سیستم اتوماتیک تشخیص گفتار موثر خواهد بود.

مراحل بدست آوردن و ترسیم این فیلتر بانک نیز در معادلات زیر بیان شده است :

برای ایجاد یک فیلتربانک گوسی ابتدا باید فرکانس ها را که با f نشان داده شده است، به حوزه مل انتقال دهیم (f_{mel})، معادله ی (۵-۹) ، و سپس Δ_{mel} را با استفاده از رابطه ی (۵-۱۰) محاسبه نماییم که در آن $f_{max(mel)}$ از قراردادن بزرگترین فرکانس نمونه برداری در رابطه ی (۵-۱۱) بدست آمده و i نیز تعداد فیلتربانک مورداستفاده در الگوریتم است. سپس با استفاده از رابطه ی (۵-۱۲) k_b ها که در فیلتر بانک مثلثی نقاط مرزی رامشخص می کردند و در این فیلتر بانک در تعیین پارامتر سیگما کاربرد دارند و به بیان دیگر در تعیین پراکندگی هر گوسی مشخص کننده هستند به دست می آیند. بعد از آن با استفاده از رابطه (۵-۱۳)، σ_i را که واریانس هر زیربخش از فیلتر بانک است محاسبه کرده و نهایتا با استفاده از رابطه (۵-۱۴) معادله پایانی گوسی های فیلتر بانک را به دست می آوریم.

$$f_{mel}=2595 \log\left(1+\frac{f}{700}\right) \qquad (۹-۵)$$

$$\Delta_{mel} = \frac{f_{max(mel)}}{i+1} \qquad (۱۰-۵)$$

$$kb_i = (i+1) \cdot \Delta_{mel} \qquad (۱۱-۵)$$

$$\sigma_i = \frac{kb_i - kb_{i-1}}{2} \qquad (۱۲-۵)$$

$$\Psi_i = e^{\frac{-(k-kb_i)}{2\sigma_i^2}} \qquad (۱۳-۵)$$

۵.۲.۱۰) به توان رساندن خروجی هر یک از کانال:

پس از عبور سیگنال از فیلتر بانک گوسی که در روش پیشنهادی تعداد آن برابر ۱۴ گوسی در نظر گرفته شده است، ۱۴خروجی بدست آمده و سپس جهت ورود به بلوک لگاریتم و جبران گرهای مربوط به آن که عملکردشان در ادامه به تفصیل بیان خواهد گردید، نیاز است، داده ها به توان مناسبی برسند. علت ایجاد شدن این نیاز، استفاده از بلوک CLMN پس از لگاریتم می باشد که گاه داده هایی نزدیک به صفر تولید نموده و سبب از بین رفتن اطلاعات می‌گردد بلوک ریشه گیری نیز تا حدودی این مشکل را حل می‌نماید،

اما پیش از استفاده از بلوک ریشه گیری به توان مناسب رساندن ورودی‌ها به بخش لگاریتم، عدم از دست رفتن اطلاعات را تضمین می نماید که توان مناسب به صورت سعی و خطا در آزمایش های تشخیص گفتار مشخص می‌گردد. محدوده‌ی انتخاب عدد مناسب جهت توان، بین یک تا نه می باشد که از مبنای لگاریتم مورد استفاده (که در اینجا ده در نظر گرفته شده است) کمتر است.

سپس خروجی این بلوک وارد بلوک لگاریتم می شود.

۵.۲.۱۱) بلوک CLMN :

تحقیقات محققین درزمینه‌ی مقاوم سازی ویژگی ضرایب مل بیان گر این موضوع بود که یکی از بخش های آسیب‌پذیر الگوریتم پایه، بلوک لگاریتم می‌باشد. جهت رفع مشکل، تلاش های زیادی صورت پذیرفت که از جمله‌ی آن‌ها می‌توان به استفاده از توابع تکه‌ای- خطی و یا ریشه گیری، به جای بلوک لگاریتم و در واقع حذف این بلوک اشاره کرد در اثر حذف این بلوک هرچند که مشکل آسیب پذیری تا حدودی بر طرف می شود اما از خاصیت فشرده سازی بسیار خوب لگاریتم بی بهره می شویم، بنابراین در الگوریتم پیشنهادی سه بلوک پس از بلوک لگاریتم به عنوان جبرانگرهای مناسب به کار رفته است تا ضمن برطرف نمودن مشکل آسیب پذیری آن، هیچ حذفی نیز صورت نپذیرد و از خواص مثبت بلوک لگاریتم بهره مند شویم.

از میان انواع نویز، نویز کانولوشن بیشترین تاثیر **را** در تخریب این ضرایب در بخش لگاریتم خواهد گذاشت که این نویز ناشی از اعوجاج کانال‌ها و نیز نقصان در دریافت صدا در مسیری که گفتار از دهان خارج، تا زمانی که به منطق دیجیتال تبدیل می گردد، می باشد.

نحوه‌ی ترکیب نیز کانولوشن با سیگنال بدون نویز و نیز عملکرد بلوک تفریق میانگین لگاریتم در ادامه بیان شده است.

$$x(m,k)=s(m,k) * h(k) \quad (۱۴-۵)$$

در این معادله h(k) تابع تبدیل نویز کانولوشن در حوزه‌ی زمان و نیز s سیگنال بدون نویز و x سیگنال حاصل از تخریب توسط نویز کانولوشن می باشد.

پس از عبور از بلوک تبدیل فوریه معادله (۵-۱۴) به صورت معادله ی (۵-۱۵)، وارد تابع لگاریتم می-گردد:

$$X(m,k)= S(m,k) . H(k) \quad (۱۵-۵)$$

در ادامه و با استفاده از خواص تابع لگاریتم و اعمال نمودن این تابع به دو طرف معادله‌ی (۵-۱۵) خواهیم داشت:

$$\text{Log } X(m,k) =\log(S(m,k) . H(k))= \log S(m,k)+\log H(k) \quad (۱۶-۵)$$

سپس عملگر تفریق میانگین سیگنال را در معادله‌ی(۵-۱۶) انجام می دهیم که حاصل را CLMN می نامیم(معادله ی(۵-۱۷))، بنابراین داریم :

$$\begin{aligned}&\text{Log } X(m,k)-E(\text{Log } X(m,k)) \\ &= \log S(m,k)+\log H(k) –E(\log S(m,k))-\log H(k)= \\ &\log S(m,k) –E(\log S(m,k))\end{aligned} \quad (۱۷-۵)$$

چنانچه مشاهده می گردد جمله‌ی حامل نویز کانوالوشن یعنی log H(k) دراثر این فرایند حذف خواهد شد.

۵.۲.۱۲) فیلتر RASTA:

این فیلتر یکی از فیلتر های مشهور است که کاربرد آن در الگوریتم ضرایب مل موجب بهبود در مقاوم سازی این الگوریتم به وسیله‌ی اثر گذاری بر نویز کانولوشن می باشد. بهترین مکان جهت اعمال آن در جبران سازی تابع لگاریتم و پس از جبران ساز CLMN می باشد. تابع تبدیل آن در ادامه در معادله‌ی (۵-۱۸) بیان شده است [۴۷][۴۸].

$$H_{RASTA}=0.1z^{4}\left[\frac{2z+z^{-1}-z^{-3}-2z^{-4}}{1-0.98z^{-1}}\right] \qquad (۵-۱۸)$$

۵.۲.۱۳) بلوک ریشه گیری (ROOT):

همان طور که بیان گردید، پس از بلوک CLMN که در جبرانسازی لگاریتم به کار رفت، داده هایی کوچک و نزدیک به صفر تولید می شود که در واقع دلالت بر حذف اطلاعات مورد نیاز و ضروری را دارد. پیش از ورود به این تابع بخشی از این مشکل را توسط بلوک توان رسانی، پیش بینی و جبران نمودیم اما همچنان نیاز به یک جبرانساز با خواص فشرده سازی به نحوی که داده هایی یک دست تر به ما بدهد و از، از دست رفتن اطلاعات جلوگیری نماید خواهیم داشت زیرا داده های پراکنده، تبدیل کسینوسی را نیز در فشرده سازی دچار مشکل می نماید.

بلوک پیشنهادی بلوک ریشه گیری است که با $(.)^{\alpha}$ نمایش داده شده و داریم $0 <\alpha< 1$ که در آن مقدار مناسب برای α در آزمایش های تشخیص گفتار به صورت سعی و خطا مشخص شده است که در فصل بعد جداول مربوط به محاسبه‌ی آن آورده شده است.

در نهایت پس از اعمال این بلوک از خروجی تبدیل کسینوسی گرفته شده و ۱۴ ضریب بهبود یافته‌ی مل به نام AGDCR-MFCC تولید می گردد که در آن A حرف ابتدای واژه ی کلیدی (Autocorrelation) ، G (Gaussian) ، D(Differential) ، C (CLMN) و R حرف ابتدای واژه های کلیدی (RAS-RASTA-ROOT) می باشد.

۵.۳) ارزیابی الگوریتم پیشنهادی:

ویژگی‌های استخراج شده از سیگنال گفتار را به وسیله‌ی یک شبکه عصبی طبقه بندی می کنیم. در این بخش از یک شبکه پرسپترون با یک لایه ورودی، دو لایه میانی یا مخفی و یک لایه خروجی استفاده می-کنیم.تعداد نرون‌های لایه ورودی به اندازه بردار ویژگی ها و تعداد نرون های دو لایه میانی به دلخواه توسط کاربر تعیین شده است ،تعداد ۴۰ لغت[1] متفاوت، در هر لغت ۲۰ تکرار توسط ۲۰ گوینده‌ی متفاوت شامل زن و مرد بزرگسال، جهت ارزیابی الگوریتم پیشنهادی به کار رفته است. طول فریم‌ها و پنجره‌ی پیشنهادی، برابر ۴۴ میلی ثانیه انتخاب شده است فرکانس نمونه برداری برابر ۲۲۰۰۰ در نظر گرفته شده است.

هفتاد در صد کل داده ها، که در مجموع شامل ۸۰۰ لغت می باشد، به عنوان داده های آموزش و سی-درصد داده ها به عنوان داده های تست در نظر گرفته شده است. در هر یک از الگوریتم های به کار رفته

[1] Isolated Word

قسمت سکوت حذف و سپس استخراج ویژگی صورت گرفته است. نویز مورد استفاده نویز سفید گوسی است که به صورت مصنوعی و با استفاده از کد متلب به داده های صوتی اضافه شده است.

۵.۴) دیتا بیس:

دیتا بیس مورد استفاده در مرحله‌ی آزمایش، دیتا بیس مشهور فارسی، فارس دات می باشد. دادگان فارس دات، دارای ۶۰۰۰ جمله است که توسط حدود ۳۰۰ گویشور فارسی زبان تولید شده اند.گویشوران بر حسب سن لهجه و میزان تحصیلات انتخاب گردیده‌اند و هر گویشور ۲۰ جمله در اطاقک ضد صدای آزمایشگاه زبان شناسی دانشگاه تهران بیان نموده است. فرکانس نمونه برداری این دادگان ۴۴.۱کیلو هرتز بوده است که به ۲۲۰۰۰هرتز کاهش داده شده است. این دیتا بیس حاوی کلمات مجزا، دادگان تلفنی و.. نیز می باشد که برگرفته از همان ۶۰۰۰ جمله‌ی ابتدایی است.

فصل ششم

نتایج آزمایش های انجام شده

فصل ششم : نتایج آزمایش های انجام شده

۶.۱) مقدمه:

جهت انجام آزمایش‌های بازشناخت مقاوم گفتار در حضور نویز و به طور خاص نویز اضافی شونده از داده های صوتی پایگاه مشهور فارسی، فارس دات استفاده شده است بدین صورت که بر روی ۴۰ لغت،(۴۰ کلاس)، هر لغت ۲۰ تکرار توسط ۲۰ گوینده‌ی متفاوت فرایند استخراج ویژگی صورت گرفته است.

ویژگی مورد ارزیابی ضرایب مل-کپستروم بهبود یافته می باشد که در اثر اصلاحات و تغییرات صورت گرفته بر روی الگوریتم پایه‌ی استخراج ویژگی ضرایب مل-کپستروم صورت گرفته است.

در ارزیابی الگوریتم پیشنهادی، AGDCR-MFCC ، در ابتدا نیاز به بررسی و تعیین چند پارامتر به کار رفته در بلوک‌های الگوریتم است که تعیین آن ها از طریق آزمایش‌های مناسب تشخیص گفتار و به صورت سعی و خطا در محدوده‌ی مشخص شده، صورت می پذیرد. این پارامترها عبارتند از: توان مورد نیاز در بلوک توان رسانی پیش از ورود به مرحله‌ی لگاریتم، مرتبه‌ی مورد نیاز در حذف ضرایب اتوکورلیشن، و نیز انتخاب و تعیین ریشه‌ی مناسب در مرحله‌ی بلوک ریشه‌ی گیری.

جهت تعیین هر یک از پارامترهای فوق نیاز به ثابت فرض نمودن دو پارامتر دیگر است در واقع نتایج کاملا نسبی است و با سعی و خطا می توان هر سه متغیر را در بهترین حالت با یکدیگر هماهنگ کرد طوری که

بهترین نتایج بازشناخت مقاوم گفتار فارسی را داشته باشیم بنابراین در این بخش تنها نتایج و ارزیابی حاکی از آزمایش‌های نهایی، را گردآوری و ارائه نمودیم و تحلیلی از جداول بیان شده است. سپس آزمایش‌های تشخیص گفتار در ارزیابی الگوریتم پیشنهادی صورت می پذیرد. آنچنانکه بیان گردید نویز به کار رفته در طول آزمایش‌ها، نویز سفید گوسی مصنوعی، در نرخ سیگنال به نویزهای متفاوت شامل مقادیر ۴۰دسی بل(بدون نویز)،۱۰دسی بل،۵ دسی بل،۰ دسی بل و نیز ۵- دسی بل، خواهد بود. همچنان که بیان شد شرایط بدون نویز را نرخ سیگنال به نویز ۴۰ دسی بل در نظر گرفته ایم.

روش‌پیشنهادی،الگوریتم پایه‌ی استخراج ضرایب مل،AMFCC،GMFCC،ROOT-MFCC،CMN-SMN-MFCC و نیز RAS-MFCC ،پیاده سازی و نتایج نرخ تشخیص در هر یک از سیگنال به نویزهای ذکر شده بیان و با یکدیگر مقایسه شده اند.

۶.۲) نتایج آزمایش‌های انجام شده برای تعیین توان مورد نیاز در بلوک توان رسانی:

همان طور که بیان گردید، پیش از ارزیابی و مقایسه‌ی روش پیشنهادی نیاز به تعیین بعضی پارامترهای نامشخص در برنامه می باشد که تعیین آن ها از طریق آزمایش‌های مناسب تشخیص گفتار و به صورت سعی و خطا در محدوده‌ی مشخص شده، صورت می پذیرد. از جمله‌ی آن ها تعیین پارامتر توان مورد نیاز در مرحله‌ی توان رسانی می باشد.

پیش از بررسی جدول (۶-۱) و نتایج مربوط به بلوک توان رسانی یادآوری می کنیم منظور از میانگین نرخ تشخیص در محیط نویزی میانگین درصد تشخیص در حضور نویز با نرخ سیگنال به نویز ۱۰،۵،۰و ۵- دسی بل می باشد در ادامه جدول مربوط به نتایج آورده شده است :

جدول(۶-۱) : مقایسه وتعیین مقدار عددی توان، در مرحله ی توان رسانی در نرخ سیگنال به نویز های متفاوت

سیگنال به نویز/مقدار توان	۱۰dB	۵dB	۰dB	-۵dB	میانگین نرخ تشخیص
p=۲	۵۵	۳۵/۲	۱۲	۱۰/۱۸	۲۸/۰۹۵
p=۳	۵۹	۴۲	۱۷/۸	۹/۹	۳۲/۱۷۵
p=۴	۷۳/۵	۷۳	۷۲/۵	۴۰	۶۴/۷۵
p=۵	۷۷/۳	۷۷	۶۱/۶	۵۵/۵	۶۷/۸۵
p=۶	۸۷/۵	۸۶/۸	۷۷/۷	۶۴/۴	۷۹/۱
p=۷	۸۷	۸۵/۸	۵۵	۳۰/۳	۶۴/۵۲۵
p=۸	۷۷	۷۵	۵۸/۵	۱۲	۵۵/۶۲۵
p=۹	۵۵/۵	۳۰/۳	۱۵	۱۲/۱۲	۲۸/۲۳

محدوده‌ی انتخاب عدد مناسب جهت توان، بین یک تا نه می باشد که از مبنای لگاریتم مورد استفاده (که در اینجا ده در نظر گرفته شده است) کمتر است.

همان طور که از جدول (۶-۱) پیداست الگوریتم پیشنهادی در حالتی‌که توان مورد نیاز برابر ۶ در نظر گرفته شده است دارای بالاترین نرخ میانگین تشخیص در محیط نویزی می باشد و نیز در حالتی که توان مورد نیاز به یک نزدیک است میانگین درصدهای تشخیص در محیط های نویزی پایین می باشد که این امر به علت از دست رفتن اطلاعات پس از عبور از بلوک های جبرانگر لگاریتم می باشد. در توان های مرکزی (توان ۴، توان ۵ و توان ۶) بهترین نرخ های تشخیص را داریم که از میان آن‌ها، توان ۶ بهترین انتخاب خواهد بود و عملکرد تشخیصی در بهترین حالت صورت پذیرفته است.

۶.۳) مرتبه ی مورد نیاز در حذف ضرایب اتوکورلیشن:

آنچنان‌که بیان گردید بالاترین مرتبه تابع خودهمبستگی جهت حذف با استفاده از آزمایش های تشخیص گفتار بدست خواهد آمد این مرتبه را با حرف T (Threshold) نمایش می دهیم و جدول مربوط به نتایج ،جدول (۶-۲)، که نحوه‌ی انتخاب را نشان می‌دهد در ادامه بیان و بررسی می گردد.

جدول(۶-۲):مقایسه و تعیین مقدار عددی آستانه حذف، در مرحله ی حذف ضرایب مرتبه پایین خودهمبستگی، در نرخ سیگنال به نویز های متفاوت

میانگین نرخ تشخیص	-۵dB	۰dB	۵dB	۱۰dB	سیگنال به نویز/مرتبه(T)
۶۱/۰۷۵	۵۵	۶۸/۸	۵۵/۵	۶۵	۱۰
۶۰/۴۱۵	۳۰	۶۶/۶۶	۷۰	۷۵	۳۰
۵۷/۹۸	۳۵	۴۸/۵	۷۱	۷۷/۴۵	۵۰
۶۶/۳۷	۵۵	۵۸/۲	۷۵	۷۷/۳	۷۰
۷۹/۱	۶۴/۴	۷۷/۷	۸۶/۸	۸۷/۵	۱۰۰
۳۱/۱۲۵	۲۲.۵	۲۷	۳۵	۴۰	۱۱۰

پس از حذف صد ضریب مرتبه پایین تابع خود همبستگی مشاهده می گردد بالاترین نرخ میانگین تشخیص در محیط نویزی را خواهیم داشت چنانچه مشاهده می شود پس از افزایش نرخ حذف به بالاتر از ۱۰۰ ضریب، درصد میانگین نرخ تشخیص در محیط نویزی به شدت نزول می کند بنابراین به علت واگرایی درصدهای تشخیص، آزمایش ها را ادامه نمی دهیم و بهترین مرتبه جهت حذف برابر ۱۰۰ انتخاب می گردد.

۶.۴) مرتبه ی مورد نیاز در بلوک ریشه گیری (ROOT):

همان طور که بیان گردید بلوک ریشه گیری که با $(.)^{\alpha}$ نمایش داده شده و داریم $0 <\alpha< 1$ ، که در آن مقدار مناسب برای α در آزمایش های تشخیص گفتار به صورت سعی و خطا مشخص خواهد گردید جدول مربوط به نتایج ،جدول (۶-۳) ، و نحوه‌ی انتخاب در ادامه بیان و بررسی می‌گردد.

جدول(۶-۳):مقایسه درصد تشخیص گفتار در مقادیر متفاوت ریشه (α)، در نرخ های متفاوت سیگنال به نویز

میانگین نرخ تشخیص در حضور نویز	-۵dB	۰dB	۵dB	۱۰dB	سیگنال به نویز/مقدار α
۷۹/۱	۶۴/۴	۷۷/۷	۸۶/۸	۸۷/۵	۰/۹
۷۲/۳۲۵	۵۵/۶	۶۹	۷۸/۷	۸۶	۰/۸
۶۳/۴۷۵	۳۵	۶۶.۲	۶۵	۸۷/۷	۰/۷
۴۶/۲۲۵	۳۰	۳۹/۴	۴۵	۷۰/۵	۰/۵
۳۲/۷۲۵	۱۲/۶	۲۵	۳۸/۳	۵۵	۰/۳

چنانچه از نتایج پیداست بالاترین نرخ میانگین تشخیص در محیط نویزی در α=۰/۹ اتفاق افتاده است و نیز پایین‌ترین نرخ میانگین تشخیص در α=۰/۳ اتفاق افتاده است در واقع با نزدیک شدن α به صفر یعنی بی‌اثر شدن این بلوک (هر عدد به توان صفر برابر یک بوده و در واقع این بلوک بی‌اثر می گردد) نتایج درصد تشخیص کاهش یافته که این امر دلیلی دیگر بر اثر بخشی بالای این بلوک در الگوریتم پیشنهادی است.

۶.۵) ارزیابی های نهایی روش پیشنهادی:

پس از مشخص شدن سومین پارامتر متغیر از الگوریتم پیشنهادی، برنامه‌ی مورد نظر تکمیل و آماده‌ی قیاس با سایر الگوریتم های بهبود یافته، مطرح در خانواده‌ی ضرایب مل-کپستروم می باشد.

این الگوریتم‌ها شامل الگوریتم‌های مطرح در خانواده‌ی ضرایب مل-کپستروم می باشد که در فصل مروری بر روش‌ها چگونگی استخراج هریک و بلوک دیاگرام استخراج این ضرایب توضیح داده شد. بررسی صورت گرفته بر روی این الگوریتم‌ها که شامل الگوریتم پایه ی استخراج ضرایب مل، AMFCC ، GMFCC ، ROOT-MFCC، CMN-SMN-MFCC و نیز RAS-MFCC می باشد، انجام گرفته است.

نتایج و تحلیل‌ها در جدول (۶-۴) آورده شده است.

جدول(۶-۴): مقایسه‌ی درصد تشخیص گفتار در روش پیشنهادی با نرخ تشخیص گفتار در روش های متداول، از خانواده‌ی ضرایب مل

SNR/feature	MFCC	AMFCC	GMFCC	ROOT-MFCC	CMN-SMN MFCC	RAS-MFCC	AGDCR-MFCC (روش پیشنهادی)
بدون نویز(۴۰dB)	۹۷/۳	۹۰/۲	۹۱/۵	۸۹/۹	۹۱/۲	۸۴/۴	۹۴/۶
۱۰dB	۸۲/۶	۸۴/۴	۸۳	۸۲/۱	۸۴	۸۴/۳	۸۶/۵
۵dB	۶۸/۵	۷۷/۱	۷۰/۳	۶۸/۵	۷۵/۷	۷۳/۵	۸۳/۵
۰dB	۶۰/۱	۷۱/۶	۷۴/۶	۶۵/۵۵	۶۴/۴	۷۵/۰۹	۸۱/۷
-۵dB	۳۹	۶۵/۲	۶۲/۲	۵۵/۶	۶۵/۶	۶۵/۸	۷۲/۷
میانگین نرخ تشخیص در حضور نویز	۶۲/۵۵	۷۴/۵۷۵	۷۲/۵۲۵	۶۷/۹۳	۷۲/۴۲	۷۴/۶۷	۸۱/۱

چنانچه از نتایج نهایی در جدول (۶-۴) پیداست ضرایب پایه‌ی مل-کپستروم همان‌طور که انتظار می رفت بالاترین نرخ تشخیص را در محیط بدون نویز نسبت به سایر الگوریتم ها داراست (۹۷٪).

اما این الگوریتم در محیط نویزی نسبت به الگوریتم های بهبود یافته مطرح در این خانواده، بسیار آسیب پذیر است و پس از افزودن نویز به الگوریتم پایه، در نسبت سیگنال به نویز ۱۰ دسی بل، نرخ تشخیص ۱۴/۷ ٪ نزول پیدا می کند در حالیکه الگوریتم AMFCC در سیگنال به نویز ۱۰ دسی بل ۵/۸ ٪ نسبت به حالت بدون نویز نزول داشته، الگوریتم GMFCC ۸/۵٪ نزول، الگوریتم ROOT-MFCC ۷/۸ ٪ نزول ،الگوریتم CMN-SMN MFCC ۷/۲ ٪نزول و نیز الگوریتم RAS-MFCC ۰/۱ ٪ نزول، داشته است.

در مورد الگوریتم پیشنهادی (AGDCR-MFCC) در نسبت سیگنال به نویز ۱۰ دسی بل، هر چند که در ابتدا ۸٪ نزول می یابد که نسبت به اکثر الگوریتم های بهبود یافته نتیجه مطلوبی نمی باشد، اما از دو جنبه قابل بررسی است، اول اینکه شیب این نزول با افزوده شدن نویز در مراحل بعدی تقریبا ثابت و الگوریتم پیشنهادی در درصد تشخیص دچار افت بسیار ناچیزی نسبت به سایر الگوریتم ها می‌گردد در حالی که سایر الگوریتم ها دچار افت شدیدتری با افزایش نویز می‌گردند مثلا در نرخ سیگنال به نویز ۵ دسی بل نسبت به ۱۰ دسی بل الگوریتم پیشنهادی ۳٪ دچار نزول می‌گردد در حالی که الگوریتم AMFCC ۷/۳ ٪، GMFCC ۱۲/۷٪ نزول ، الگوریتم ROOT-MFCC ۱۳/۶٪ نزول ،الگوریتم CMN-SMN MFCC ۸/۳ ٪نزول و نیز الگوریتم RAS-MFCC ۱۰/۸٪ نزول داشته است و الگوریتم پیشنهادی با افزایش نویز کمتر دچار افت در نرخ تشخیص می‌گردد.

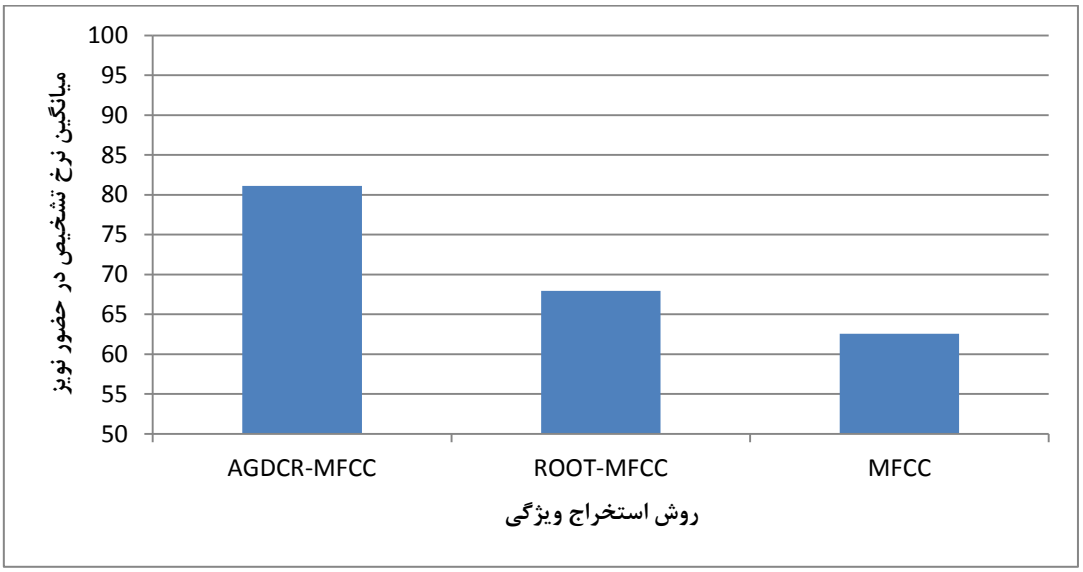
مزیت دیگر الگوریتم پیشنهادی نسبت به سایر الگوریتم‌های بهبود یافته این است که الگوریتم پیشنهادی در حالت محیط بدون نویز نسبت به سایر الگوریتم های بهبود یافته درصد تشخیص بالاتری دارد. همان-طور که اشاره شد درمیان الگوریتم‌های بهبودیافته در شرایط نویزی، الگوریتم پیشنهادی کندترین نزول در نرخ تشخیص را داراست که این نزول از ۹ درصد در شرایطی همراه با نویز اضافی شونده تجاوز نمی - نماید.

همچنین درصد میانگین تشخیص گفتار در محیط نویزی، در الگوریتم پیشنهادی، از سایر الگوریتم های مقاوم و نیز الگوریتم پایه بالاتر می باشد. یادآوری می کنیم منظور از درصد میانگین نرخ تشخیص در محیط نویزی، میانگین درصد تشخیص در حضور نویز با نرخ سیگنال به نویز ۱۰،۵،۰و ۵- دسی بل می - باشد که این میانگین در الگوریتم پیشنهادی نسبت به سایر الگوریتم ها دارای مزیتی به ترتیب ۲۲/۸۷، درصد بهبود در نرخ تشخیص گفتار نسبت به الگوریتم پایه،۸/۴ درصد بهبود نسبت به الگوریتم AMFCC ۱۰/۵۷، درصد بهبود نسبت به الگوریتم GMFCC ،۱۶/۲۳درصد بهبود نسبت به الگوریتم ROOT-

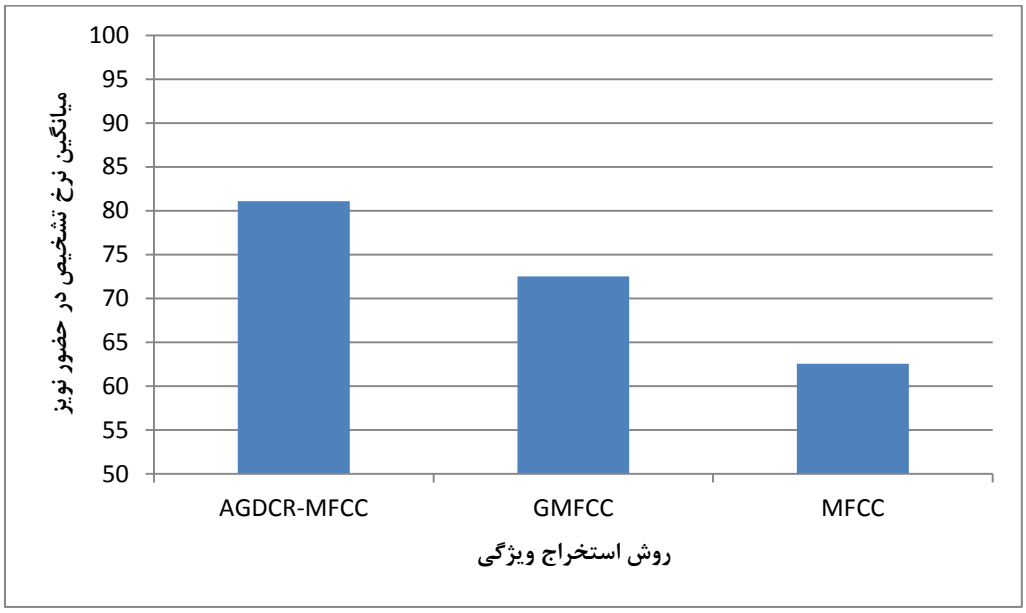
MFCC، ۱۰/۷ درصد بهبود نسبت به الگوریتم CMN-SMN MFCC و نیز۹/۴۶ درصد بهبود نسبت به الگوریتم RAS-MFCC می باشد. چنانچه از این اعداد پیداست، بیشترین بهبود نسبت به، الگوریتم پایه یعنی۲۲/۸۷ درصد، صورت پذیرفته است.

پس از الگوریتم پیشنهادی الگوریتم‌های RAS-MFCC و نیز AMFCC به ترتیب دارای عملکرد تشخیصی بهتری نسبت به سایر الگوریتم‌ها دارند از نظر زمانی نیز این دو الگوریتم در سال هایی پس از سایر الگوریتم ها پیشنهاد شده اند بنابراین انتظار نیز می رفت که از سایر الگوریتم ها که قدیمی تر می-باشند بهتر عمل نمایند وهمین‌طور نیز بوده است.

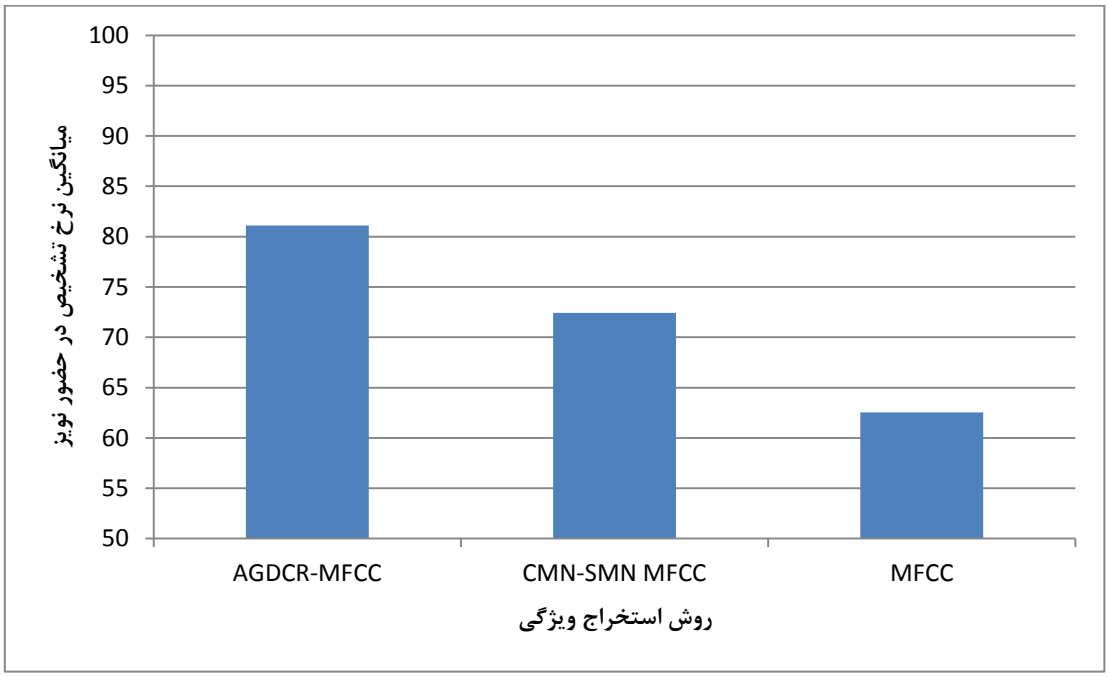
در ادامه، الگوریتم پیشنهادی، الگوریتم پایه و هر یک از الگوریتم‌های بهبود یافته در قالب نمودارهای تصویری در اشکال (۶-۱)،(۶-۲)،(۶-۳)،(۶-۴) و (۶-۵) با یکدیگر مقایسه خواهند شد.



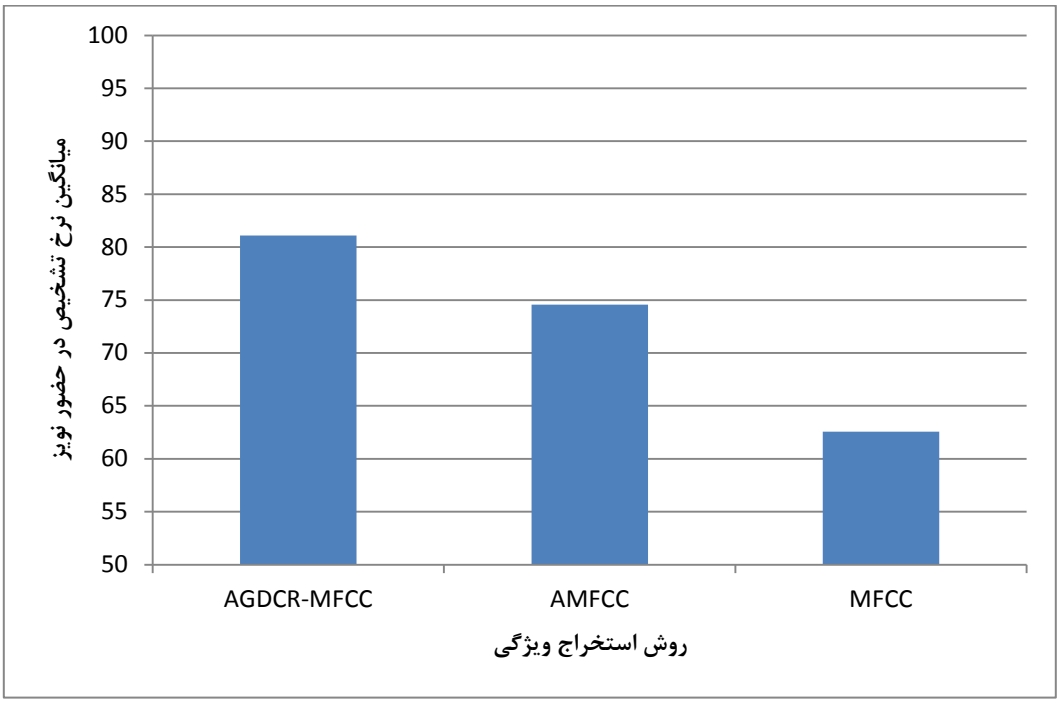
شکل (۶-۱) : مقایسه ی عملکرد روش پیشنهادی با الگوریتم پایه و نیزROOT-MFCC



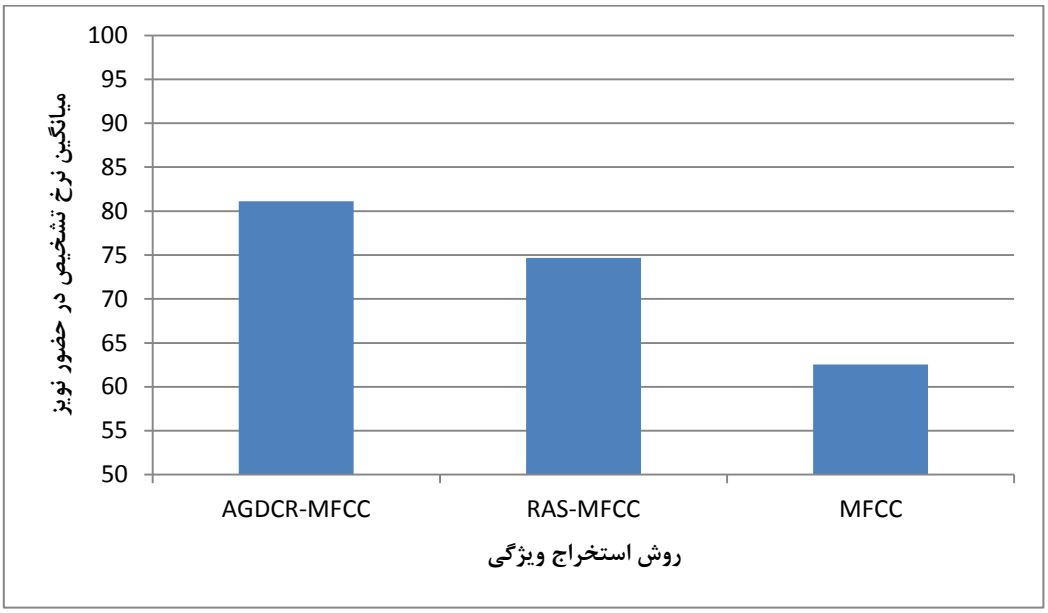
شکل (۶-۲) : مقایسه ی عملکرد روش پیشنهادی با الگوریتم پایه و نیزGMFCC



شکل (۶-۳) : مقایسه ی عملکرد روش پیشنهادی با الگوریتم پایه و نیز CMN-SMN MFCC

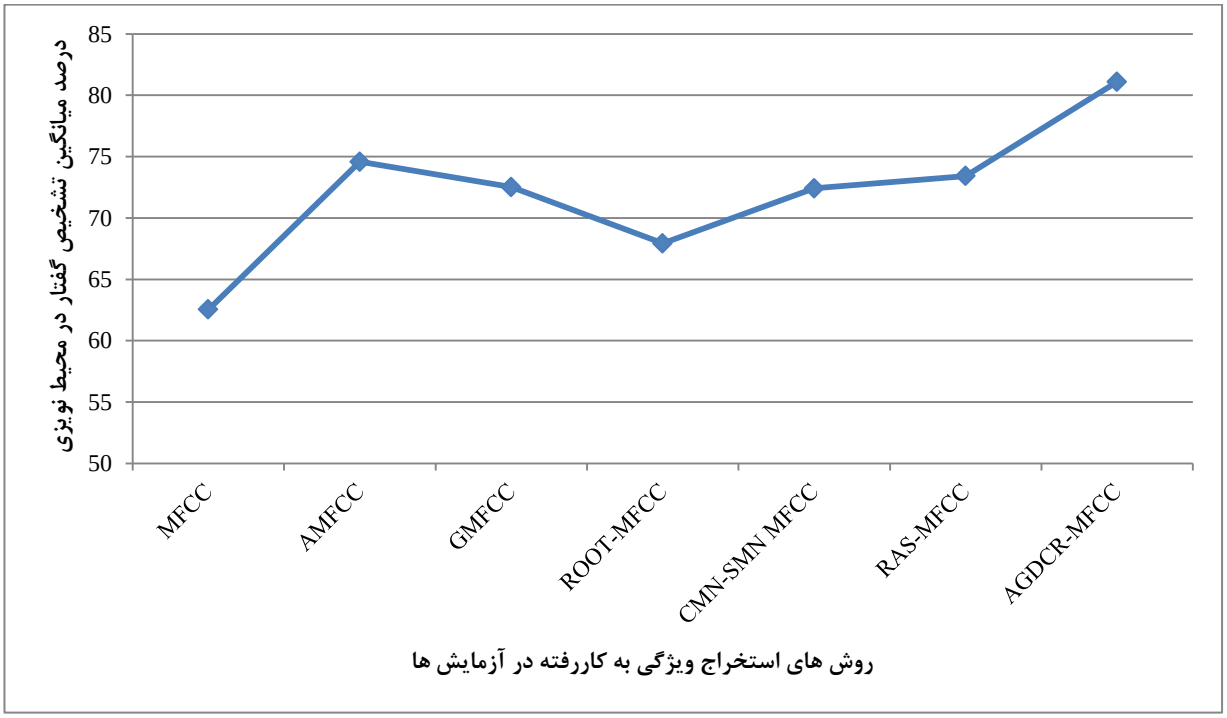


شکل (۶-۴) : مقایسه ی عملکرد روش پیشنهادی با الگوریتم پایه و نیز AMFCC



شکل (۶-۵) : مقایسه ی عملکرد روش پیشنهادی با الگوریتم پایه و نیز RAS-MFCC

و در ادامه نمودار کلی از مقایسه‌ی روش های پیاده سازی شده و نیز روش پیشنهادی در شکل (۶-۶) آورده شده است که نمایانگر عملکرد بهتر الگوریتم پیشنهادی می باشد:



شکل(۶-۶):مقایسه ی عملکرد روش پیشنهادی با سایر روش های مطرح در خانواده ی الگوریتم های ویژگی ضرایب مل با استفاده از، درصد میانگین تشخیص گفتار نویزی

۶.۶)نتیجه‌گیری و پیشنهادات:

در این پایان نامه یک روش استخراج ویژگی ضرایب مل، مقاوم در برابر نویز معرفی گردید، سپس با استفاده از نویز اضافی شونده از نوع نویز سفید گوسی، به علت خواص ویژه ی این نوع نویز، جهت ارزیابی الگوریتم پیشنهادی در مقایسه با روش پایه و نیز روش‌های رایج مقاوم در برابر نویز، AMFCC ، GMFCC ، ROOT-MFC، CMN-SMN-MFCC و نیز RAS-MFCC، به انجام آزمایش های تشخیص

گفتار پرداختیم. نتایج در نرخ سیگنال به نویزهای متفاوت بررسی و مورد مقایسه قرار گرفت. جداول و نمودارها حاکی از عملکرد بهتر الگوریتم پیشنهادی در محیط نویزی در برابر الگوریتم پایه و نیز نمونه های رایج در نوع مقاوم این الگوریتم با افزایش نویز، می‌باشد. همچنین در شرایط بدون نویز همچنان الگوریتم پایه‌ی استخراج این ضرایب بهترین عملکرد را در کاربردهای تشخیص گفتار داراست و بالاترین نرخ تشخیص در محیط بدون نویز (۴۰ دسی بل) در میان الگوریتم های بهبود یافته باز هم متعلق به الگوریتم پیشنهادی(AGDCR-MFCC) می باشد که این مورد نیز مزیت دیگر الگوریتم فوق نسبت به سایر الگوریتم های بهبود یافته می باشد.

در بررسی سیر مقاوم سازی این الگوریتم دو گروه عمده از تغییرات انجام و پیاده سازی گشت که شامل افزودن بلوک مکمل به الگوریتم پایه و نیز تغییر در بلوک های اصلی الگوریتم پایه بود، اما پیرامون بهبود در پیاده‌سازی سخت افزاری از طریق کاهش در محاسبات، و در نتیجه کاهش در گیت‌های منطقی مطالعه‌ای صورت نپذیرفت و نیز همچنان بلوک هایی در الگوریتم پایه مانند بلوک تبدیل فوریه و نیز تبدیل کسینوسی بدون تغییر باقی‌ماند که پیرامون آسیب پذیری شان در برابر نویز هیچ گونه مطالعه‌ای صورت نپذیرفت که هر یک از موارد فوق می تواند زمینه‌ی مطالعات آتی محققین علم پردازش سیگنال باشد.

مراجع

[1] Young s**"A Review of Larg-vocabulary Continues-Speech Recognition"** IEEE signal processing magazine.no13, pp.45-57, 1996

[2] L.Rabinner and B-Hjuang,**" Fundamental of Speech Recognition"** Prentice Hall PTR, USA, 1993

[3] A.Acero, PhD, thesise**,"Acoustical and Environmental Robustness in Automatic Speech Recognition"** Electronice and computer, Mellon University, 1990

[4] K.J.Palonki, G.g.Brwn and J.Barker**"Missing Data Speech Recognition in Reverbration Conditions"** Proc.ICASSP.25. PP.894-900, 2002

[5] Ljubomir Josifovski, PhD, thesise**,"Speech Recognition with Missing and Unreliable Data"** Dparteman of Computer science, University of Sheffield, 2002

[6] Lee ch. **"Adaptive Compensation for Robust Speech Recognition"** 01n:proc.IEEE Workshop on Automatic Speech Recognition.12.pp127

[7] Aik Wing Toh, PhD, thesise**,"Feature Extraction for Robust Speech Recognition in Hostile Environments"**Departeman of Computer science, University of Western Australia

[8] D.O'shaushnessy,**"Linear Predictive coding"** IEEE potentials, vol.7, pp.29-32, 1988

[9] J.Makhoul,**"Linear Prediction: A tutorial review"**IEEE, vol.63, pp561-580, 1975

[10] B.J.Sannon, PhD thesise,**"Speech Recognition and Enhancment Using Autocorralation Domain Processing"**, School of engeenering Griffith University, Brisbone, Australia, 2006

[11] L.Rabinner and R.W.Schafer,**"Digital Processing of Speech Signal**"prentice Hall, USA, 1978

[12] S.B.Davis and P.Mermelstein, **"Fundamentals of Speech Recognition"**, Prentice Hall, USA, 1980

[13] A.Koc, MS thesise,**"Acoustic Feature Analysis for Robust Speech Recognition"**

Bogazici University, Turkey, 2002

[14] H.Hermansky, **"Perceptual Linear Predictive (PLP) Analysis of Speech"**, J.Acoust.Soc.Am, vol.87, No.4, pp.1738-1752, 1989

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

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

[17] M.Gupta and A.Gilbert**,"Robust Speech Recognition Using Wavelet Coeeficients Features"**, in IEEE Automatic Speech Recognition And Understanding Workshop, USA, pp.445-448, 2001

[18] R.Modic, B.Lindberg, B.Petec, **"Comprison Wavelet and MFCC Speech Recognition Experiments on the Slovenian and English SpeechDat2"**in ISCA Tutarial and Research Workshop on Non-Linear Speech processing (NOLISP 03) .Le Croisic, France, 2003

[19] Ching-Tang Hsieh, Eugene Lai, You-Chuang Wang**, "Robust Speaker Identification System based on Wavlet Transform and Gussian Mixture Model"**.J.Inf.Sci.Eng, 19(2):267-282, 2003

[20] XIAO-XInONG, PhD thesise,**"Robust Speech Features and acoustic Models For Speech Recognition"** computer enggenering Departmant, Nanians Technological University, 2009

[21] Kuo-Hwei You and Hsiao-Chuan wang **"Roubust Features for Noisy Speech Recognition based on Temporal Trajectory Filtering of Short time Autocorrelation Sequences"**speech communication 28, pp 13-24, 1999

[22] Benjamin j.Shannon , Kuldip k.Paliwal **" Feature Extraction from Higher Autocorrelation Coefficient for Robust Speech Recognition"** 48, pp.1458-1481, 2006

[23] Wu Zunjing, Cao Zhigang **"Improved MFCC-Based Feature for Robust**

Speaker Identification "Identification Tsinghua Science and Technology ISSN, volume10 Number 2, pp.158-161, 2005

[24] Ruhi Sarikaya and John H.L. Hansen, **"Analysis of the Root-Cepstrum for Aacoustic Modeling and Fast Decoding in Speech Recognition"** in Eurospeech, Aalborg, Denmark, sept 2001

[25] H. Hermansky, N. Morgan. **"RASTA processing of speech signal"** IEEE Trans. On speech and Audio Processing.vol.4, pp. 578–589 ,1994

[26] F.H. Liu, A. Acero, R. Stern. **"Efficient Joint Compensation of Speech For the Effects of Additive Noise and Linear Filtering"** Proc. Of IEEE ICASP, pp. 257–260. 1992

[27] J.S. Lim, **"Spectral Root Homomorphic Deconvolution System,"** IEEETrans.ASS, vol. 27, no. 3, pp. 223–233, 1979

[28] P. Alexandre and P. Lockwood**, "Root cepstral analysis: A unified view"** Speech Communication, vol.3, pp. 277–288, 1993

[29] Ruhi Sarikaya and John H.L. Hansen, **"Analysis of The Root-Cepstrum for Acoustic Modeling and Fast Decoding in Speech Recognition,"** in Eurospeech, Aalborg, Denmark, 2001

[30] AmitaDevandPoonambansal**"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, 2010

[31] Sandipan Chakroborty and Goutam Saha**"Improved text-independent Speaker Identification using Fused MFCC and IMFCC Feature sets Based on Gaussian Filter"**5(2009)

[32] Yegnanarayana B., Prasanna S.R.M., Zachariah J.M. and Gupta C. **features for a fixed-text speaker verification system Combining evidence from source, suprasegmental and spectral",** IEEE Trans Speech and Audio Processing, Vol. 13, No. 4, pp. 575-582, July 2005.

[33] J. Kittler, M. Hatef, R. Duin, J. Mataz, **"On combining classifiers "**

IEEE Trans Pattern Anal. Mach. Intell. 20, PP.226-239, 1998

[34] Chuan Xie and Xiaoi Cao and Lingling **"Algorithm of Abnormal Audio Recognition on Improved MFCC "**International Workshop on Information and Electronic Engineering (IWIEE), pp.731-737

[35] Wang J C, Wang J F, Wang Y S. **"Chip design of MFCC Extraction for Speech Recognition [J] Integration"**, 32 (1/2) PP.111-131, 2002

[36] Md Sahidullah Guatam Saha,**"A novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition"**Arxiv 1206-2437 v1, 2012

[37] F.Phan, M.T.Evangelia.sideman,**"Speaker Identification using Nerula Network and Wavelets"** IEEE Engineering in medicin and Biology Magazine, vol.191, 200, pp.92-101

[38] Rabaou I, Lachir I Z, Ellouze. N**"Using HMM-based Classifier Adapted to Background Noises with Improved Sounds Features for Audio Surveillance Application [J]"**. International Journal of Signal Processing,5(1):pp.46-55.2008

[39] Van Compernal **"D.Noise Adaption in a Hidden Markov Model Speech Recognition system"** Computer Speech Language 3:pp.67-151.1988

[40] Nakagawa,S.,Ueda,Y.,Seino,T.**"Speaker-Independent, Text independent Language Identification by HMM"** proc.ICSLP, 92pp.1011-1014, 1992

[41] D.A Reynols and R.C.Rose **"Robust Text-Independent speaker Identification Using Gaussian Mixture Speaker Models"**IEEE Transactions on information & System Vol.E84-D, PP.847-855

[42] Hu Yi-ping. **"Research and Implementation of Speaker Recognition Based on GMM "**[D]. Xiamen: Xiamen University, 2007.

[43] Godino-Llrente J.I., Gomez-Vilda p.,and Blanco vdascom.,**"Dimensionality of Pathological Voice Quality Assesment System Based on Gaussian Mixture Models and Short Term Cepstral parameters"** 53(10):pp.1943-1953, 2006

[44] R. Auckenthaler, et al., **"Improving a GMM Speaker Verification System by Phonetic Weighting"**, Proc. ICASSP Phoenix, Arizona, Vol. 1, pp. 313-316, May 1999

[۴۵] پ.مهارلویی،(۱۳۸۹)،پایان نامه ارشد،"طراحی و پیاده سازی سیستم شناسایی زبان گفتاری به صورت به صورت خودکار"

[۴۶] س آیت،(۱۳۸۷)،"مبانی پردازش سیگنال گفتار"چاپ اول،انتشارات پیام نور،تهران،ص۱۸۶-۱۴۹

[47] Hermansky, H.Morgan.N,**"RASTA processing of speech"**, IEEE Trans. Speech Audio Processing 2, 578±589, 1994

[48] Bojan kotnik, Damjan vlaj, Zdravk **"Robust MFCC Feature Extraction Algorithm Using Efficient Additive and Convolutional Noise Reduction Precedures"** ICSLP, 02 Proceeding, USA, PP.445-448, 2002

Abstract

Automatic Speech Recognition (ASR) consists in recognizing based on a sample of speech from an unknown speaker.In order to recognize a speech signal. Any ASR systems need the feature extraction block. The Mel Frequency Cepstral Coefficient (MFCC) is one of the most common features which are used in ASR systems.

The problem occurs when ASR systems is in the noisy environment because MFCC performance degrade drastically in noisy condition. In order to achieve a satisfactorily performance under noisy condition we need to improve the standard MFCC feature extraction method. In this thesis we introduce a noise robustness new set of MFCC vector estimated through some basic variation in the standard algorithm.

We use the mean subtraction technique in both time domain and frequency domain, higher order autocorrelation coefficient are extracted and we apply eliminating the lower order of autocorrelation coefficient, using suitable filters to suppress convolution noise, using Gaussian shape filter bank in place of triangular shape filter bank and adding compensator block to enhance robustness of algorithm better.

To evaluate the performance of proposed MFCC method and to classify the results we use MLP neural network with one input layer, two hidden layer and one output layer. We use forty isolated word spoken by twenty different speakers including male and female .Frame length is 44ms and sampling frequency rate is 22000. 70% of the entire data is used for train and 30% is used for testing.

So far lots of approaches is used to improve MFCC algorithm some of them improve it through insisting on some basic block and some of the others improve it by adding complementary block to MFCC basic algorithm .In our proposed method we attend to most of last variations we use them in the best place and find the way to combine them in a best manner and we add some complementary block to standard algorithm too.

Recognition experiments show significant improvement in recognition rate compare with standard MFCC algorithm and some popular algorithm in MFCC's familes including: AMFCC ،GMFCC ،ROOT-MFCC ،CMN-SMN-MFCC ، RAS-MFCC

Key words: MFCC-Speech recognition-Autocorrelation-Gaussian shape filter bank-Mean subtraction-Logarithm compensator

Shahrood University of Technology

Faculty of Electrical and Robotic Engineering

Robust Farsi Speech Recognition Using Modified Mel-Frequency Cepstral Coefficient and Neural Network

Danial Darabian

Supervisor:

Dr.Hossein Marvi

September 2013