دانشکده : برق و رباتیک

گروه : الکترونیک

یک روش استخراج ویژگی برای شناسایی گوینده بر مبنای توزیع ویگنر

دانشجو : جلیل قاسمی

استاد راهنما

دکتر حسین مروی

استاد مشاور

دکتر امید رضا معروضی

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

بهمن ۱۳۹۰

شماره :

بسمه تعالی

تاریخ :

مدیریت تحصیلات تکمیلی

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 جلیل قاسمی رشته برق گرایش الکترونیک تحت عنوان

یک روش استخراج ویژگی برای شناسایی گوینده بر مبنای توزیع ویگنر

که در تاریخ ۹۰/۱۱/۱۲ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح زیر است :

قبول (با درجه : ..... امتیاز .....) ☐ 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

a عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنما			
۲ـ استاد مشاور			
۳ـ نماینده شورای تحصیلات تکمیلی			
۴ـ استاد ممتحن			
۵ ـ استاد ممتحن			

تأیید رئیس دانشکده :

تقدیم به

پدر و مادر بزرگوارم

و

همسرمهربانم

سپاس بی کران خداوند متعال که مرا فرصت اندیشیدن داد.

بر خود لازم می‌دانم که از استاد گرانقدرم جناب آقای دکتر حسین مروی

و تمامی اساتید گرامی که مرا در انجام این تحقیق یاری نمودند

کمال تشکر و قدردانی را داشته باشم

تعهد نامه

اینجانب جلیل قاسمی دانشجوی دوره کارشناسی ارشد رشته برق- الکترونیک دانشکده برق و رباتیک دانشگاه صنعتی شاهرود نویسنده پایان نامه یک روش استخراج ویژگی برای شناسایی گوینده بر مبنای توزیع ویگنرتحت راهنمائی 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:

با پیشرفت روز افزون علم و تکنولوژی نیاز بشر به سیستم‌های هوشمند در تمامی زمینه‌ها مشهود می‌باشد. با توجه به اینکه صدای هر فرد مانند شناسه‌های دیگری چون اثر انگشت، خصوصیات چهره، عنبیه چشم و ... منحصر به فرد می‌باشد به همین دلیل از آن در سیستم‌های هوشمند بازشناسی گوینده استفاده می‌شود. اما یکی از مهمترین مسائل مطرح شده در زمینه سیستم‌های بازشناسی گوینده تأثیر نویز بر سیگنال گوینده و کاهش دقت بازشناسی است. لذا بازشناسی مقاوم گوینده در مقابل نویز یکی از موارد مورد مطالعه محققان می‌باشد که در سالهای اخیر تحقیقات زیادی در این زمینه انجام شده و روشهای مختلفی ارایه شده است.

هدف از انجام این تحقیق استفاده از توزیع زمان- فرکانس ویگنر- ویل برای استخراج ویژگی از سیگنال گفتار می‌باشد. در این تحقیق از ترکیب تابع توزیع ویگنر- ویل و تبدیل هیلبرت و ضرایب MFCC برای استخراج ویژگی استفاده شده است. در این روش از سیگنال گفتار تبدیل هیلبرت گرفته شده و پس از آن، تبدیل ویگنر- ویل گرفته می‌شود. سیگنال خروجی از تبدیل ویگنر- ویل از فیلتر بانک مل عبور داده می‌شود و پس از لگاریتم گیری، از آن تبدیل فوریه کسینوسی گرفته می‌شود. خروجی بدست آمده از این سیستم پیشنهادی بعنوان یک ویژگی استخراج شده از سیگنال گفتار جهت تشخیص گوینده بکار می‌رود. به کمک این ویژگی‌ها برای هر گوینده یک مدل مخلوط گوسی(GMM) بدست آورده شده است. نتایج حاصله در این روش با ضرایب MFCC و PLP مقایسه شده است که جهت پیاده سازی روش‌های پیشنهادی از دیتا بیس TIMIT استفاده شده است. نتایج گویای این مطلب است که روش پیشنهادی در سیگنال به نویزهای پایین جواب بهتری را از ضرایب MFCC و PLP ارایه می‌دهد.

کلید واژه : بازشناسی گوینده و گفتار، توزیع ویگنر- ویل، استخراج ویژگی

فهرست مطالب **صفحه**

فصل اول: سیستم‌های بازشناسی گوینده و گفتار

فصل دوم : روشهای متداول استخراج ویژگی

فصل سوم : توزیع ویگنر- ویل

فصل چهارم : مدل مخلوط گوسی

فصل پنجم : روش پیشنهادی و پیاده سازی الگوریتم

فهرست اشکال **صفحه**

فهرست جداول صفحه

فهرست نمودار صفحه

فصـــــــل اول

سیـــستم های

بازشناسی گوینـــده و گفتـــار

۱-۱- مقدمه

دیر زمانی است که استفاده از شناسه‌های کاربری مختلف برای تعیین هویت افراد مورد استفاده قرار می‌گیرد. از مرسوم‌ترین این شناسه‌ها می‌توان به شماره شناسنامه، کد ملی، نام و نام خانوادگی و ... اشاره کرد. یکی از اشکالات عمده این شناسه‌ها ضعفی است که در مقابل جعل آنها قرار دارد که این موضوع خود امنیت این شناسه‌ها را زیر سؤال برده است و دانشمندان را به یافتن شناسه‌های قابل اطمینانی مانند شناسه‌های بیومتریک[1] سوق داده است. شناسه‌های بیومتریک مانند اثر انگشت، خصوصیات چهره، عنبیه چشم، ویژگیهای صدا و ... کمتر قابل جعل هستند و قابلیت اطمینان بیشتری دارند.

در اغلب کاربردها، مانند سیستم‌های امنیتی، سیستم‌های الکترونیکی، جرم شناسی و ... شناسایی فرد گوینده حائز اهمیت می‌باشد. حال با توجه به اینکه از لحاظ علمی ویژگی‌های گفتار، منحصر بفرد می‌باشد، در سیستم‌های تشخیص گوینده[2] از ویژگی‌های صوتی صدای افراد جهت بازشناسی آنها استفاده می‌شود. الگوهای صوتی افراد به دو دسته تقسیم بندی می‌شود. دسته اول شامل ساختار اندامهای صوتی، اندازه، شکل گلو، دهان و ویژگی‌های تارهای صوتی است و دیگری الگوهای رفتاری آموخته شامل تحصیلات، موقعیت اجتماعی و سبک سخن گفتن است که غالباً در سیستم‌های بازشناسی گوینده امروزی از ویژگی‌های دسته اول استفاده می‌شود[۱۰].

قابلیت بالای این سیستم‌ها و نیز دسترسی ساده از طریق میکروفن، احساس نیاز به طراحی اینگونه سیستم‌ها را جهت جانشینی امضاء، اثر انگشت، کارت شناسایی، رمز عبور و مانند اینها پدید آورده است.

[1] Biometrik
[2] Speaker Recognition

۱-۲- تاریخچه

اولین تلاشها برای ایجاد و ساخت سیستم‌های پردازش گفتار در اوایل دهه ۱۹۵۰ شروع شد. تقریبا ۱۰ سال پس از آن در اوایل دهه ۱۹۶۰ اولین سیستم بازشناسی گوینده و سیستم تأیید هویت گوینده[1] اختراع شد. آقای Pruzanky اولین کسی بود که در آزمایشگاه بل با استفاده از فیلتر بانک و ارتباط بین دو اسپکتوگرام برای اندازه گیری شباهت بین آنها تحقیقات خود را شروع کرد. آقای Li و همکارانش از جداسازهای خطی برای جدا سازی این طیف‌ها استفاده کردند و همچنین در تگزاس آقای Doddington از آنالیز طیف به جای روش فیلتر بانک استفاده کرد. تحقیقات در مورد مقاوم سازی[2] سیستم‌های بازشناسی گوینده و گفتار در مقابل انواع نویزها و شرایط محیطی در دهه ۱۹۹۰ به اوج خود رسید و هم اکنون محققان زیادی بر روی مقاوم‌سازی سیستم‌های بازشناسی گوینده و گفتار فعالیت دارند[۴].

۱-۳- سیستم های تشخیص گفتار یا بازشناسی گفتار[3]

به سیستم‌هایی که اطلاعات گفتاری را دریافت کرده و متن بیانی گوینده را استخراج می‌کنند یا به بیان دیگر گفتار شخص گوینده را تشخیص می‌دهند، سیستم‌های تشخیص گفتار یا بازشناسی گفتار می‌گویند[۴]. کاربردهای این نوع سیستم‌ها بسیار متنوع بوده، که ارتباط با بانک اطلاعاتی از راه دور، ثبت نام دانشجویان از طریق تلفن، اعلام شماره شناسایی شخصی، شماره حساب بانکی، شماره عضویت برای کاربران یک سیستم خدمات رسانی و در کمک به افراد ناشنوا و ... نمونه‌هایی از کاربردهای این نوع سیستم‌ها به حساب می‌آیند.

[1] Speaker Identification
[2] Robustness
[3] Speech Recognition

۱-۴- سیستم های بازشناسی گوینده

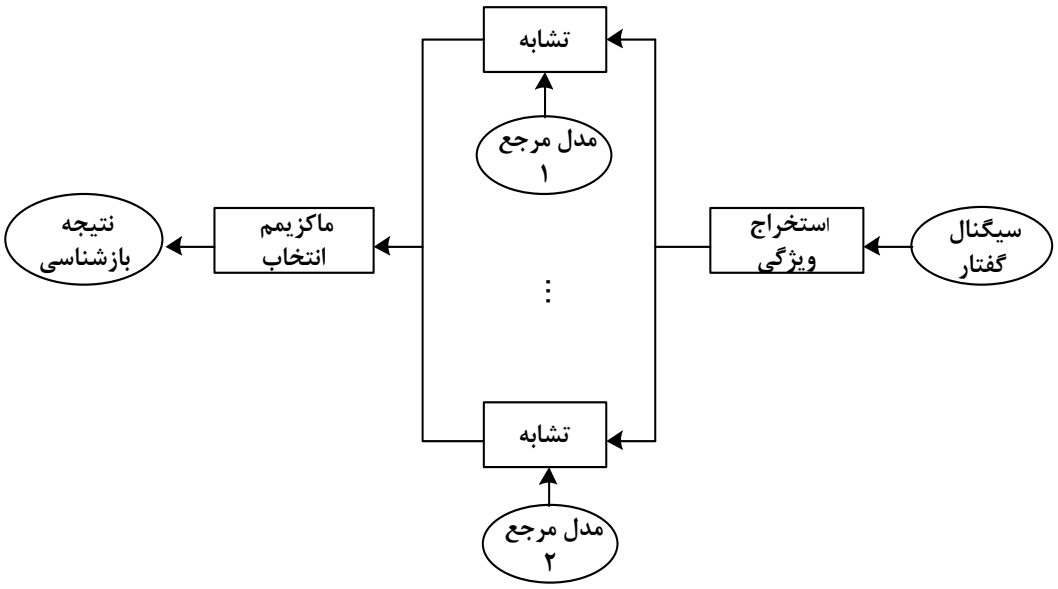
یک سیستم تشخیص گوینده سیستمی است که شخص را از صدای آن تشخیص می‌دهد. سیستم‌های بازشناسی گوینده از لحاظ روش استفاده به دو دسته سیستم‌های تایید هویت گوینده و سیستم های بازشناسی هویت گوینده تقسیم‌بندی می‌شوند. در یک تقسیم‌بندی دیگر سیستم‌های تشخیص گوینده به سیستم‌های وابسته به متن و مستقل از متن تقسم‌بندی می‌شوند که در ادامه توضیح داده می‌شوند[۴].

۱-۴-۱- سیستم های تأیید هویت گوینده

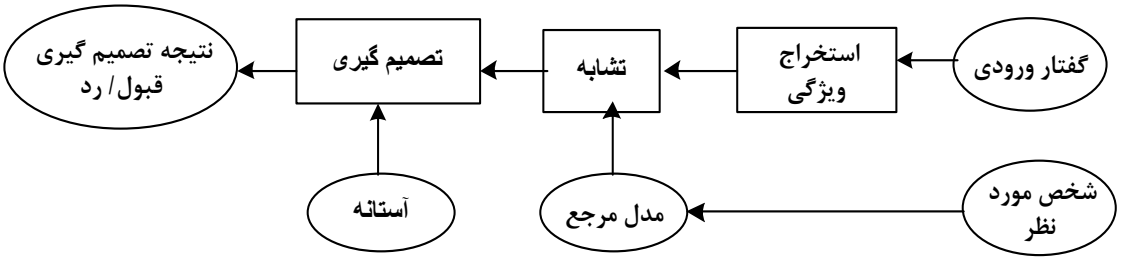
در یک سیستم تایید هویت گوینده شخص عموماً با انتخاب یا وارد کردن نام یکی از کاربران خاص سیستم ادعا می‌کند که او همان کاربر ثبت شده‌ی سیستم است. در این حالت سیستم وظیفه دارد ویژگی‌های صوتی شخص مدعی را با ویژگی‌های صوتی ذخیره شده‌ی کاربر ثبت شده‌ی مورد ادعا مقایسه نموده و با استفاده از نتیجه‌ی بدست آمده ادعای شخص را بپذیرد یا رد کند[۴].

۱-۴-۲- سیستم های بازشناسی هویت گوینده

در یک سیستم بازشناسی هویت گوینده، شخص صحبت کننده ادعای هویت یک کاربر خاص ثبت شده را نمی‌نماید و این سیستم است که وظیفه دارد او را در میان کاربران ثبت شده‌ی سیستم بازشناسی نماید و یا تشخیص دهد که ویژگی‌های صوتی او با هیچ یک از کاربران ثبت شده‌ی سیستم همخوانی ندارد[۴]. در شکل(۱-۱) شمای کلی این نوع سیستم‌ها به نمایش گذاشته شده است.



(الف) : تشخیص گوینده



(ب) : تأیید گوینده

شکل(۱-۱) : ساختار پایه از سیستم های بازشناسی گوینده[۱۱].

۱-۴-۳- سیستم های وابسته به متن

در سیستم‌های وابسته به متن، در مرحله تست، گوینده باید از همان جمله‌ای استفاده کند که در مرحله آموزش استفاده کرده است. به طور کلی این سیستم‌ها کاربرد زیادی ندارند[۴] و [۱۱].

۱-۴-۴- سیستم های مستقل از متن

در سیستم‌های مستقل از متن، در مرحله تست، لزومی ندارد که شخص از همان جمله‌ای استفاده کند که در مرحله آموزش استفاده کرده است. بنابراین همانطور که مشخص است کاربرد سیستم‌های مستقل از متن گسترده تر از سیستم‌های وابسته به متن است[۴] و [۱۱].

دقت تمامی سیستم‌های خودکار بازشناسی گوینده و گفتار در هنگام استفاده در محیط نویزی کاهش می‌یابد. دلیل این افت دقت تفاوت بین داده‌های آموزش و داده‌های تست این سیستم‌ها می‌باشد[۱]. با توجه به وجود نویز در اکثر محیط‌های واقعی، مقاوم‌سازی سیستم‌های خودکار بازشناسی گفتار و گوینده در مقابل نویزها را می‌توان واسطه بین بازشناسی خودکار و کاربرد واقعی آن دانست[۱] و[۴].

مجموعه تلاش‌های انجام شده برای مقاوم‌سازی سیستم‌های باز شناسی گفتار و گوینده را می‌توان به سه دسته تقسیم کرد. در دسته اول که استخراج ویژگی‌های مقاوم[1] نامیده می‌شود، سعی شده است تا مستقیما از سیگنال گفتار ویژگی‌های مقاوم در مقابل نویزها استخراج شود. در این مرحله باید از سیگنال گفتار ویژگی‌هایی انتخاب شود که تأثیر پذیری آنها در مقابل نویز کم باشد. در دسته دوم که نرمال سازی ویژگی‌ها[2] نامیده می‌شوند، سعی شده است تاثیر نویز تا حد امکان از ویژگی‌های استخراج شده حذف گردد. در دسته‌ای دیگر که تطبیق مدل[3] نامیده می‌شود، مقاوم‌سازی در مرحله

[1] Robust Feature Extraction
[2] Feature Normalization
[3] Model Adaptation

ساخت مدلها صورت می‌گیرد. در این مرحله از ترکیب کلاسیفایرها و رسیدن به کلاسیفایر مطلوب جهت مقاوم‌سازی سیستم‌ها استفاده می‌شود[۱] و [۶] .

علی‌رغم بهبود بسیار ایجاد شده در کارایی سیستم‌های خودکار بازشناسی، استخراج ویژگی‌های مقاوم در مقابل نویز بنیادی‌ترین مسئله حل نشده در سیستم‌های بازشناسی گفتار و گوینده محسوب می‌شود[۲]و[۷].

سه معیار خوب بودن برای ویژگی‌های مقاوم عبارتند از[۲] و [۸]:

۱. میانگین[1] ویژگی‌ها تحت تأثیر نویز تغییر نکند.
۲. واریانس مربوط به فضای ویژگی تحت تأثیر نویز افزایش نیابد.
۳. ویژگی‌ها از خاصیت جدایی پذیری برای الگوهای مختلف برخوردار باشند.

با توجه به موارد گفته شده، استخراج ویژگی در محیط‌های نویزی یکی از مهمترین مراحل در سیستم‌های بازشناسی گوینده و گفتار می‌باشد. هدف از انجام این تحقیق این است که در سیستم‌های بازشناسی گوینده از توزیع زمان- فرکانس ویگنر- ویل برای استخراج ویژگی استفاده شود.

۱-۵- ساختار پایان نامه

این پایان‌نامه از پنج فصل تشکیل شده است که به اختصار در زیر توضیح داده شده است. فصل اول مقدمه‌ای از سیستم‌های بازشناسی گوینده و گفتار می‌باشد و انواع سیستم‌های گوینده نیز توضیح داده شده است.

در فصل دوم روشهای متداول استخراج ویژگی آورده شده است و به دو دسته آنالیز زمانی و طیفی تقسیم بندی می‌شوند که در اکثر موارد از روشهای آنالیز طیفی برای استخراج ویژگی استفاده

[1] Mean

می‌کنند. دو روش ضرائب MFCC و PLP در آنالیز طیفی بیشترین کاربرد را در بین روشهای دیگر دارند.

در فصل سوم ابتدا مشکلات تبدیل فوریه در آنالیز سیگنال‌های غیر ایستان گفته شده است و سپس تئوری توزیع ویگنر- ویل و ویژگی‌های آن و در نهایت انواع توزیع ویگنر- ویل توضیح داده شده است.

در فصل چهارم کلاسیفایر مدل مخلوط گوسی و الگوریتم EM و کاربردهای آنها در سیستم های شناسایی گوینده توضیح داده شده است.

در فصل پنجم ابتدا روش پیشنهادی توضیح داده شده است که این روش ترکیبی از تابع توزیع ویگنر- ویل و تبدیل هیلبرت و ضرایب MFCC می‌باشد و سپس مشخصات دیتا بیس TIMIT مورد استفاده در آزمایشات گفته شده است و در نهایت نتایج آورده شده است. نتایج گویای این مطلب است که در سیگنال به نویزهای پایین روش پیشنهادی جواب‌های بهتری را از ضرایب PLP و MFCC ارایه می‌دهد.

فـــــصل دوم

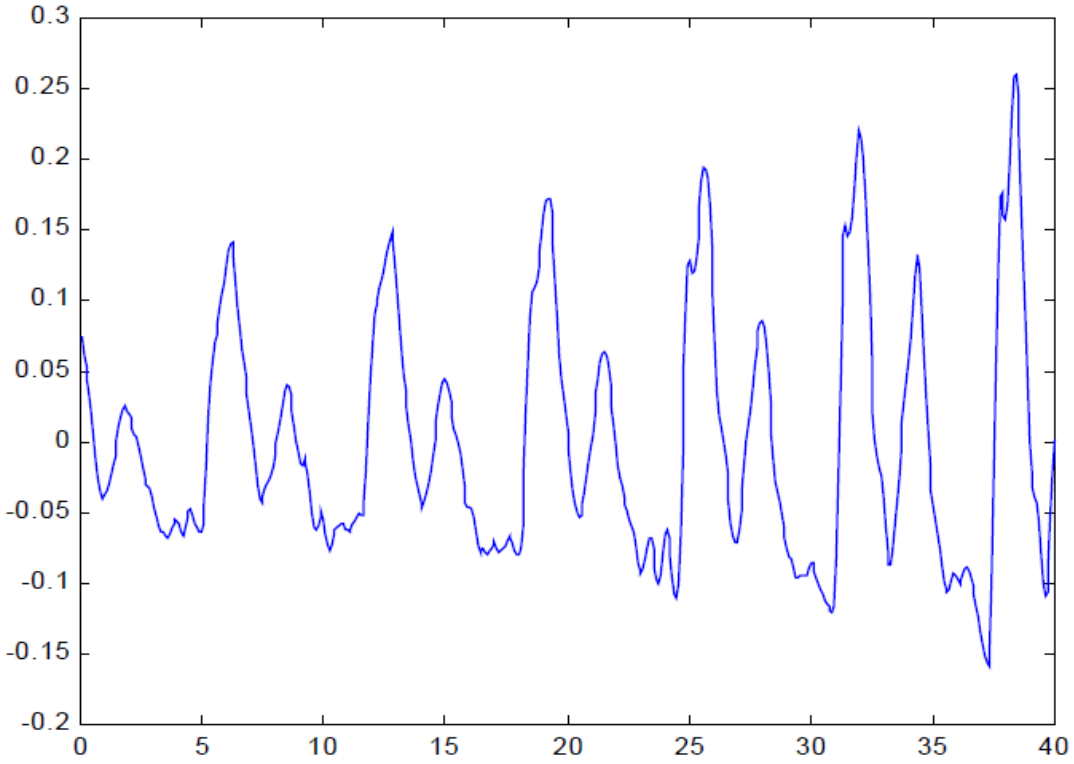
روشـهای متداول استخراج ویـــژگی

۲-۱- مقدمه

یکی از بخش‌های اساسی هر مسئله تشخیص الگو فرآیند استخراج ویژگی‌های مناسب می‌باشد. هدف از استخراج ویژگی، تبدیل سیگنال گفتار به فرم پارامتری(نرخ اطلاعات کمتر) برای آنالیز و پردازش‌های بعدی می‌باشد. سیگنال گفتار سیگنالی است که به آرامی با زمان تغییر می‌کند به همین دلیل سیگنال گفتار، یک سیگنال شبه ایستان نامیده می‌شود.

برای سیگنال گفتار یک ویژگی منحصر به فرد وجود ندارد زیرا سیگنال گفتار حامل سطوح متفاوتی از اطلاعات می‌باشد که به دو دسته اطلاعات سطح پایین و اطلاعات سطح بالا دسته‌بندی می‌شود. در اطلاعات سطح پایین، اطلاعاتی مانند ویژگی‌های فیزیکی دستگاه تولید گفتار در دسترس می‌باشد اما در اطلاعات سطح بالا طرز بیان یک کلمه، سطح تحصیلات فرد، لهجه و ... از ویژگی‌های صدای فرد می‌باشد. ویژگی‌هایی که در اطلاعات سطح پایین گفتار می‌باشد، همانند آنالیز طیف زمان کوتاه، در سیستم‌های تشخیص گوینده غالب می‌باشند زیرا استخراج این ویژگی‌ها از لحاظ محاسباتی آسان می‌باشد. در سالهای اخیر انگیزه برای استخراج ویژگی‌های سطح بالا برای سیستم‌های تشخیص گوینده زیاد شده است. تحقیقات انجام شده نشان می‌دهد اگر چه اطلاعات سطح بالای گفتار برای تشخیص گوینده امید بخش می‌باشد اما هزینه محاسباتی بالایی دارند که باید مصالحه‌ای بین هزینه محاسباتی و دقت این سیستم‌ها در نظر گرفت[۱۰].

ویژگی‌های استخراج شده از سیگنال گفتار برای تشخیص گوینده باید حامل خصوصیاتی باشد که برای هر گوینده تغییراتی اندک و نیز فاصله زیادی با خصوصیات گویندگان دیگر داشته باشد. از ویژگی‌هایی که عموما برای این منظور استفاده می‌شود می‌توان به تغییرات فرکانس گام، فرکانس فرمنت‌ها، ضرایب سری فوریه، ضرایب خود همبستگی، ضرایب انعکاسی، ضرایب LPCC، ضرایب MFCC، و بعضی از مشتقات آنها اشاره کرد. در شکل زیر بازه‌ای از سیگنال گفتار را می‌بینیم که ایستان در نظر گرفته شده است.



شکل (۲-۱): بازه‌ای ایستان از یک سیگنال گفتار[۱۱].

اگر یک بازه زمانی ۵ تا ۱۰۰ میلی‌ثانیه از سیگنال گفتار را انتخاب کنیم، ویژگی‌های آن به طور نسبتا خوبی ایستان خواهند بود با این حال در یک بازه طولانی تری از زمان(۰,۲ ثانیه یا بیشتر)، ویژگی‌های سیگنال گفتار برای انعکاس صداهای متفاوتی که بیان شده است، تغییر خواهد کرد[۱۱]. بنابراین آنالیز طیفی زمان کوتاه یک روش معمول برای توصیف کردن سیگنال گفتار می‌باشد.

۲-۲- *ویژگی‌های مورد استفاده در تحلیل گفتار*

در تجزیه و تحلیل سیگنال گفتار ویژگی‌های متفاوتی استفاده می‌شود که انتخاب ویژگی مورد نظر بسته به کاربرد آن دارد، زیرا شرایط مناسب کاربرد هر یک با دیگری متفاوت است. برای مثال، ممکن است برخی ویژگی‌ها در محیط‌های نویزی جواب بهتری دهند و یا اینکه حجم محاسبات کمتری را طلب کنند. این ویژگی‌ها گاه در کل سیگنال و گاه در یک پنجره از آن محاسبه می‌شوند، ولی در تحلیل سیگنال گفتار به طور معمول ویژگی‌های سیگنال پنجره بندی شده تحلیل می‌شوند.

این ویژگی‌ها گاه یک عددند، نظیر انرژی یا میانگین پنجره سیگنال گفتار و گاه یک بردارند. برای مثال، خود اندازه سیگنال گفتار در حوزه زمان را می‌توان به صورت برداری یک بعدی با طول تعداد نقاط پنجره گفتار در نظر گرفت. از این رو، گاه از این ویژگی‌ها با نام ضرایب ویژگی یا بردارهای ویژگی یاد می‌شود[۲۹].

۲-۳- مراحل استخراج ویژگی

برای استخراج ویژگی از سیگنال گفتار، از مراحل و روشهای مختلفی استفاده می‌شود که به طور کلی شامل پیش تأکید، پنجره‌گذاری، فیلتر کردن، و ... می‌باشد. در ادامه برخی از اصلی‌ترین این مراحل را بررسی می‌کنیم[۲۹].

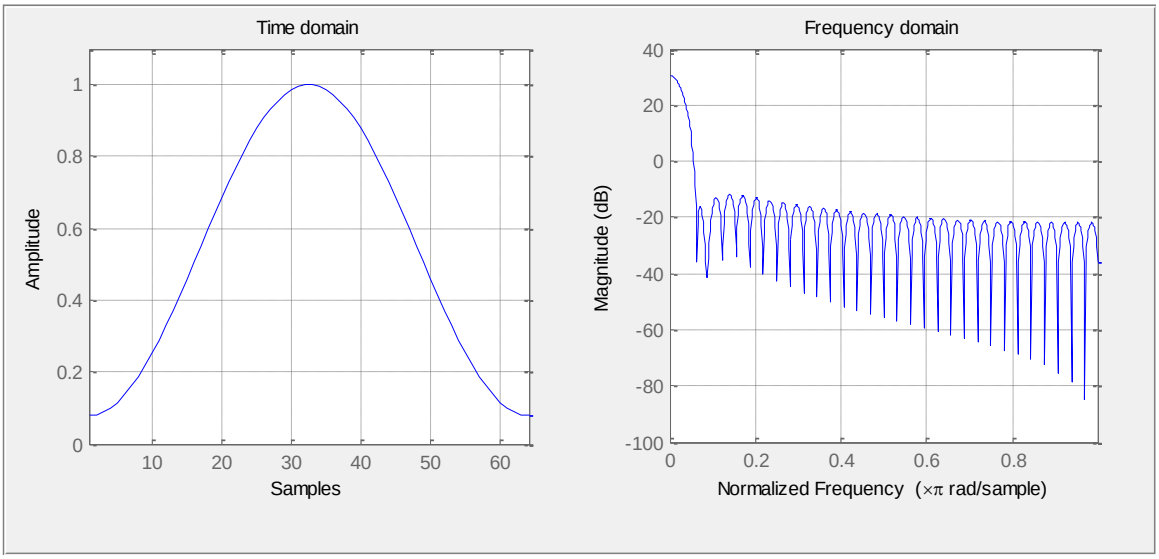
۲-۳-۱- پیش تأکید

سیگنال گفتار با فرکانس خاصی نمونه برداری می‌شود که این فرکانس به طور معمول بین ۸تا۱۶کیلو هرتز می‌باشد. همچنین این مرحله شامل جداسازی حروف بی‌صدا از حروف صدادار و اطلاعات غیر ضروری یا اضافی نظیر نویزهای موجود می‌باشد.

۲-۳-۲- پنجره بندی

پس از پیش تأکید، یک پنجره به سیگنال اعمال می‌شود، که طول این پنجره با توجه به فرض ناایستان بودن سیگنال گفتار اغلب بین ۲۰ تا ۳۰ میلی ثانیه انتخاب می‌شود. از معروفترین پنجره‌ها، پنجره همینگ می‌باشد. فرمول (۲-۱) بیانگر پنجره همینگ می‌باشد و همچنین می‌توان مشخصات آن در حوزه زمان و فرکانس را در شکل(۲-۲) مشاهده کرد.

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), \qquad 0 \le n \le N \tag{۲-۱}$$



شکل(۲-۲): حوزه فرکانس و زمان پنجره همینگ[۲۴]

۲-۳-۳- *فیلتر کردن*

این مرحله در سیگنال‌هایی نظیر سیگنال‌های تلفنی یا سیگنال‌هایی که نویز یا اطلاعات اضافه دارند استفاده می‌شود. برای مثال، ممکن است درکاربردی نظیر بازشناسی گفتار تلفنی‌خاص، اطلاعات موجود در فرکانس‌های بالای گفتار اهمیت چندانی نداشته باشد. از طرفی تمامی دستگاه‌هایی که با برق کار می‌کنند، متأثر از نویز ناشی از فرکانس برق شهر (۵۰Hz) هستند. برای آنکه تا حدی آثار ناشی از نویز خط تلفن رفع شود، ممکن است از فیلتر استفاده کنند تا فرکانس های‌پایین تر از Hz ۲۰۰ و بالاتر از ۳/۳KHz را حذف نماید. برای این منظور از هر پنجره، تبدیل فوریه گرفته می‌شود، سپس، این فرکانس‌ها از طیف حذف می‌گردند و بر طیف حاصل، تبدیل فوریه معکوس یا IFFT اعمال می‌گردد[۲۹].

استخراج ویژگی را می‌توان از آنالیز زمانی و آنالیز طیفی بدست آورد. در آنالیز طیفی از پردازش طیف سیگنال گفتار استفاده می‌شود که مزایایی از قبیل یافتن پارامترهایی مانند وکال‌ترکت[1] و فرمنت‌ها و ... را دارند اما دارای پیچیدگی محاسباتی می‌باشند.

[1] Vocal tract

اما در آنالیز زمانی از پردازش شکل موج زمانی سیگنال گفتار استفاده می‌شود که مزایایی از قبیل محاسبات کمتر و پیاده سازی آسان را دارا می‌باشد اما به پارامترهایی از قبیل انرژی و عبور از صفر و دوره تناوب محدود می‌شود[۲۹].

الگوریتم های استخراج ویژگی به دو روش

۱. آنالیز زمانی

۲. آنالیز طیفی

دسته بندی می‌شوند که در ادامه هر دو روش طیفی و زمانی توضیح داده می‌شوند.

۲-۴- روشهای آنالیز زمانی

در روشهای آنالیز زمانی از شکل موج زمانی سیگنال گفتار برای استخراج ویژگی استفاده می-شود[۲۹].

۲-۴-۱- انرژی

انرژی(E) کل هر سیگنالی می‌تواند به عنوان یک ویژگی از آن سیگنال در نظر گرفته شود که برای سیگنالی به طول N از رابطه (۲-۲) محاسبه می‌شود.

$$E_n = \sum_{m=n-N+1}^{n} x^2[m] \qquad (۲-۲)$$

۲-۴-۲- تعداد عبور از صفر[1]

تعداد عبور از صفر(ZCR) نشان دهنده تعداد دفعاتی است که سیگنال تغییر علامت می‌دهد و از صفر عبور می‌کند و به صورت رابطه (۲-۳) به دست می‌آید.

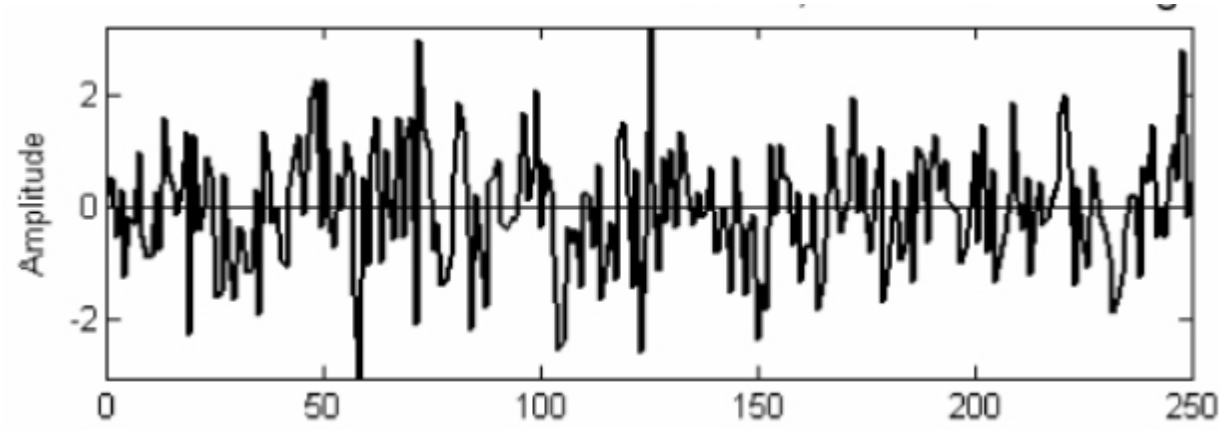
$$z_n = Z_1 = \sum_{m=-\infty}^{\infty} |\mathrm{sgn}(x[m]) - \mathrm{sgn}(x[m-1])|\, \omega[n-m] \qquad (۲-۳)$$

[1] Zero crossing (ZCR)

در این روابط $x[n]$ سیگنال گفتار، $\omega[n]$ پنجره مورد نظر و Nطول پنجره است. تابع Sgn به صورت زیر تعریف می‌شود.

$$\mathrm{sgn}(x[n]) = \begin{cases} 1 & x[n] \geq 0 \\ -1 & x[n] < 0 \end{cases} \qquad (۴-۲)$$

از این پارامتر در مواردی نظیر تشخیص نواحی گفتار از سکوت یا نویز استفاده می‌شود. واضح است که سیگنال در سکوت مطلق تعداد عبور از صفر برابر صفر و در نواحی نویزی تعداد عبور از صفر بالاتری نسبت به نواحی حاوی گفتار دارد. شکل(۲-۳) متوسط تعداد عبور از صفر یک جمله را نشان می‌دهد.



شکل (۲-۳): متوسط عبور از صفر یک جمله

۲-۴-۳- سیگنال به نویز[1]

پارامتر سیگنال به نویز یا SNR نسبت توان سیگنال به نویز را به ما نشان می‌دهد. از این پارامتر برای مقدار تحت تأثیر قرار گرفتن سیگنال گفتار تمیز در برابر نویز استفاده می‌شود. واضح است که هر چه این پارامتر مقدار بزرگتری داشته باشد، نشان دهنده آن است که سیگنال تمیزتر است و کیفیت بهتری دارد. در صورتی که توان سیگنال به نویز با توان سیگنال گفتار برابر باشد، نسبت سیگنال به نویز برابر صفر خواهد بود. در صورتی که این مقدار منفی باشد، یعنی اینکه توان نویز از توان سیگنال گفتار بیشتر است.

[1] Signal noise ratio (SNR)

۲-۵- روشهای آنالیز طیفی

روشهای آنالیز طیفی از روشهای غالب استخراج ویژگی در سیگنال گفتار می باشند[۱۳]. در تعدادی از این روشها توضیح داده شده است. در ادامه تعدادی از این روشها توضیح داده شده است.

- کپستروم[۱]
  - ✓ کپستروم حقیقی[۲]
  - ✓ کپستروم مختلط[۳]
  - ✓ کپستروم ریشه[۴]
- کدینگ پیشگوئی خطی[۵]
- ضرائب مل- کپستروم[۶]
  - ✓ روش خود همبستگی
- پیشگوئی خطی ادراکی[۷]
- پیشگوئی خطی ادراکی - رستا[۸]

[1] Cepstrum
[2] Real Cepstrum
[3] Cpmplex Cepstrum
[4] Root Cepstrum
[5] Linear Predictive Coding(LPC)
[6] Mel frequency Cepstrum Coeffcient(MFCC)
[7] Perceptual Linear Prediction(PLP)
[8] PLP-Rasta

۲-۵-۱- روش کپستروم

در بسیاری از سیستم‌های پردازش گفتار مطلوب است که اطلاعات منبع تحریک را از اطلاعات مجرای گفتار جدا کنیم و سیگنال گفتار را به دو جزء تشکیل دهنده آن تقسیم نماییم. یکی از راه‌های انجام این کار آنالیز کپسترال است. سیگنال گفتار $s(t)$ حاصل تاثیرگذاری دو سیگنال مختلف می-باشد، که یکی سیگنال منبع تحریک e(t) و دیگری سیگنال کانال صوتی $v(t)$ است. سیگنال $s(t)$ در حوزه زمان از کانولوشن این دو سیگنال بدست می‌آید که در رابطه (۲-۵) نشان داده شده است.

$$s(t) = e(t) * v(t) \qquad (۵-۲)$$

اما در حوزه فرکانس، پاسخ فرکانسی سیگنال گفتار از حاصل ضرب پاسخ فرکانسی منبع تحریک در پاسخ فرکانسی مجرای گفتار بدست می‌آید که در رابطه (۲-۶) نشان داده شده است.

$$S(\omega) = E(\omega) \times V(\omega) \qquad (۶-۲)$$

بنابراین برای بدست آوردن بردار ویژگی از سیگنال گفتار دو مرحله اساسی انجام می‌گیرد که سیگنال‌ها ابتدا به حوزه فرکانس منتقل می‌شوند وسپس حاصل‌ضرب سیستم‌ها به جمع سیستم‌ها تبدیل می‌شود. در مرحله اول به طور معمول از تبدیل فوریه استفاده می‌شود و برای مرحله دوم از توابع غیر خطی استفاده می‌شود که این موضوع باعث بوجود آمدن روشهای مختلف کپستروم می‌گردد. و در نهایت بعد از اعمال تابع غیرخطی، از آن تبدیل فوریه معکوس گرفته می‌شود[۱۳].

۲-۵-۱-۱- کپستروم حقیقی

در این نوع کپستروم، از لگاریتم به عنوان تابع غیرخطی استفاده می‌شود و اگر درالگوریتم استخراج ویژگی کپستروم، از اندازه تبدیل فوریه لگاریتم گرفته شود، ضرایب کپستروم حقیقی حاصل می‌شود[۱۳] که در رابطه (۲-۷) بیان شده و مراحل آن در شکل (۲-۴) نشان داده شده است.

$$\log\{|S(\omega)|\} = \log\{|E(\omega)|\} + \log\{|V(\omega)|\} \quad (۷-۲)$$

شکل(۲-۴): مراحل کپستروم حقیقی

۲-۵-۱-۲- کپستروم مختلط

این نوع ضرایب ، زمانی حاصل می شود که از خود تبدیل فوریه سیگنال گفتار لگاریتم گرفته شود[۱۳] که در ابطه (۲-۸) نشان داده شده است.

$$\log\{S(\omega)\} = \log\{E(\omega)\} + \log\{V(\omega)\} \quad (۸-۲)$$

۲-۵-۱-۳- کپستروم ریشه

برخلاف حالات قبلی در این روش از یک تابع نمایی مانند x^{γ} ، به عنوان یک تابع غیرخطی استفاده شده است. $\gamma = 1/3$ برای کپستروم ریشه انتخاب شده است[۱۳].

از جمله کاربردهای ضرایب کپسترال استفاده در تخمین پوش طیف سیگنال می‌باشد که برای تشخیص فرمنت‌ها و فرکانس پیچ استفاده می‌شود اما دارای پیچیدگی محاسباتی است.

۲-۵-۲- ضرایب پیشگویی خطی یا LPC

ایده اصلی در این روش این است که یک نمونه از سیگنال گفتار را می‌توان از نمونه‌های قبلی آن بدست آورد. پیشگویی خطی یکی از روشهای قدرتمند برای محاسبه و تخمین پارامترهای سیگنال گفتار مانند فرکانس‌های پیچ، فرمنت و توابع مقطع کانال صوتی می‌باشد. اهمیت این روش در دقت بالای تخمین و سرعت بالای محاسبات می‌باشد. این مدل از دو بخش تشکیل شده است که یک بخش فقط شامل صفر است و دیگری تمام قطب است[۱۳] و [۲۹].

در عمل، چون همه واکه‌ها تنها دارای قطب‌اند، تنها از بخش تمام قطب استفاده می‌شود که با فرمول (۲-۹) بیان شده است. روش LPC یک روش دقیق، مقاوم، و قابل اعتماد برای تخمین پارامترهای یک سیستم متغیر با زمان خطی می‌باشد.

$$H(z)=\frac{Y(z)}{X(z)}=\frac{1}{1-\sum_{j=1}^{p}a_j z^{-1}} \qquad (۹-۲)$$

ضرایب a_j را ضرایب پیشگویی خطی یا ضرائب LPC می‌نامند. در این رابطه مرتبه پیشگویی برابرP است. دلیل استفاده از این نام در این فیلتر این است که اگر بخواهیم نمونه زمانی لحظه n سیگنال y یعنی $y(n)$را بدست آوریم، آن را با استفاده از ترکیب خطیP نمونه قبلی y پیشگویی می‌کنیم و آن را مطابق رابطه (۲-۱۰) با $\tilde{y}(n)$ نشان می‌دهیم.

$$\tilde{y}(n)=\sum_{j=1}^{p}a_j y(n-j) \qquad (۱۰-۲)$$

اگرمطابق رابطه (۲-۱۱) به سیگنال گفتار یک پنجره اعمال کنیم داریم:

$$y(n)=\omega(n)s(n) \qquad (۱۱-۲)$$

که تفاوت بین نمونه های تخمینی و نمونه های واقعی را خطای پیشگویی می‌گویند و در رابطه (۲-۱۲) بیان شده است.

$$e(n)=y(n)-\sum_{j=1}^{p}a_j\, y(n-j) \qquad (۱۲-۲)$$

با داشتن خطا، میانگین مجذور خطای پیشگویی می‌تواند به صورت رابطه (۲-۱۳) تعریف شود.

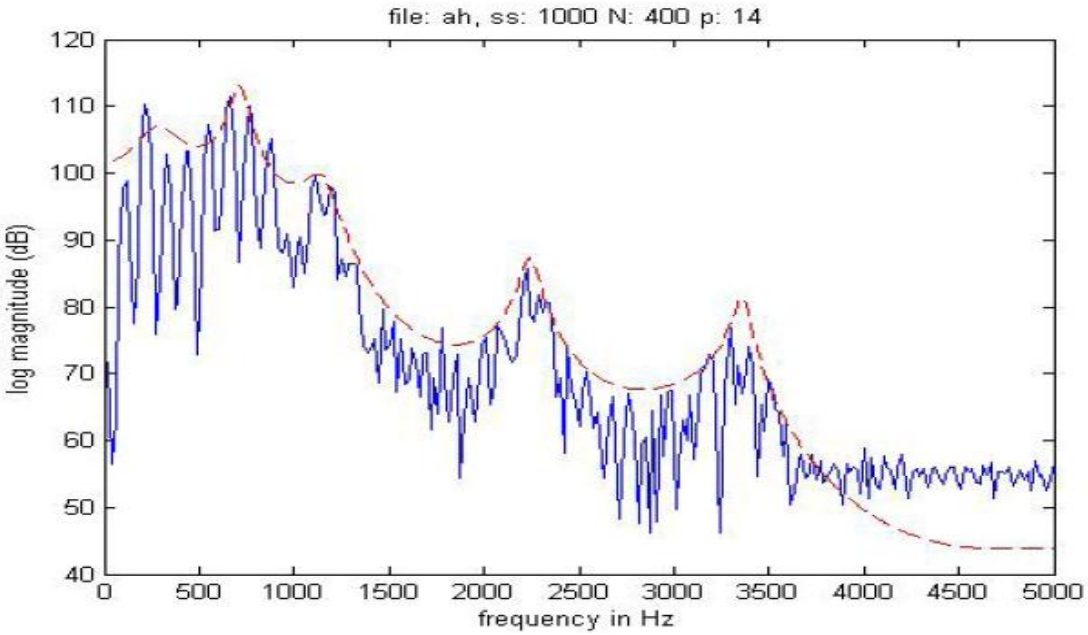
$$E=\sum_{n=0}^{N-1}e^2(n)=\sum_{n=0}^{N-1}\left(y(n)-\sum_{j=1}^{p}a_j y(n-j)\right)^2 \qquad (۱۳-۲)$$

حال ضرایب را نیز می‌توان با مینیمم کردن میانگین مجذور خطا بدست آورد. که در این صورت می‌توان از آن مشتق گرفت. با این عمل برای هر ضریب a، به تعداد p معادله خواهیم داشت که در رابطه (۲-۱۴) نشان داده شده است.

$$\sum_{j=1}^{p} a_j \phi(i,j) = \phi(i,0) \quad for \;\; i \in \{1,2,3,...,p\} \qquad (۱۴-۲)$$

در رابطه (۲-۱۴)، $\phi(i,j)$ تابع همبستگی سیگنال‌های تاخیر یافته توسط نقاط i و j می‌باشد

در شکل (۲-۵) نمونه‌ای از یک سیگنال گفتار به همراه طیف حاصل ازDFT و پوش طیف حاصل ازLPC نشان داده شده است. از دیگر کاربردهای مطرح LPC، استفاده در کد کردن سیگنال است. این قبیل کاربردها باعث شده که سعی شود روش‌های بهینه‌ای در محاسبه ضرایب پیش‌گویی خطی به دست آید. برای مثال، برخی روشهای محاسبه این ضرایب عبارت‌اند از روش خود همبستگی[۱]، روش کواریانس[۲]، روش لتیس[۳]، روش تخمین طیف[۴]، روش فیلتر معکوس[۵] و روش ماکزیمم پیشامد[۶] که تئوری این روشها در[۱۳] توضیح داده شده است.



شکل (۲-۵): پوش طیف سیگنال گفتار با استفاده از LPC

[1] Auto correlation method
[2] Covariance method
[3] Lattice method
[4] Spectral estimation
[5] Inverse filter
[6] Maximum likelihood

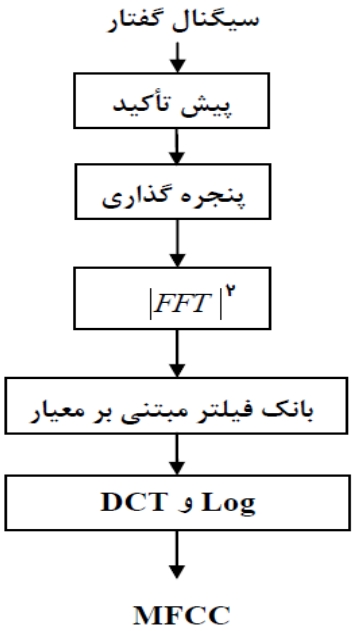
۲-۵-۳- ضرائب *LPC-CEPSTRUM*

اگر ضرایب کپستروم را از ضرایب LPC بدست آوریم به این ضرایب lpc-cepstrum می‌گویند. البته ضرایب اصلی کپستروم را fft-cepstrum می‌گویند که برای بدست آوردن آن از مدل صفر و قطب استفاده می‌کنند در حالیکه lpc-cepstrum از مدل تمام قطب استفاده می‌کند. از لحاظ محاسباتی ضرایب lpc-cepstrum سریعتر از fft-cepstrum می‌باشند[۱۳].

۲-۵-۴- ضرائب کپسترال مبتنی بر معیار مل *(MFCC)*

ایده اصلی در استفاده از ضرایب MFCC، الهام گرفته از خواص شنیداری گوش انسان در دریافت و فهم گفتار است. عملکرد گوش انسان به گونه ای است که فرکانس ادراکی آن با فرکانس واقعی صدا متفاوت است[۱۳] و [۲۹]. مقیاس فرکانس مل برای فرکانس‌های زیر ۱۰۰۰Hz خطی و برای فرکانس‌های بالای ۱۰۰۰Hz به صورت لگاریتمی تغییر می‌کند که این دو فرکانس مطابق رابطه (۲-۱۵) زیر با هم مرتبطاند و مراحل آن در شکل (۲-۶) نشان داده شده است.

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \qquad (۱۵-۲)$$



شکل(۲-۶): بلوک دیاگرام روش MFCC [۲]

مراحل این بلوک دیاگرام به طور مختصر در ادامه توضیح داده شده است[۲۹].

۱- پیش تاکید

سیگنال گفتار با فرکانس خاصی نمونه برداری می‌شود که این فرکانس به طور معمول بین ۸ تا ۱۶کیلو هرتز می‌باشد. همچنین این مرحله شامل جداسازی حروف بی‌صدا از حروف صدادار و اطلاعات غیر ضروری یا اضافی نظیر نویزهای موجود می‌باشد.

۲- پنجره بندی

برای استخراج ویژگی سیگنال گفتار فریم بندی می‌شود که برای این عمل از پنجره همینگ استفاده می‌شود. طول هر فریم حدود۲۰ تا۳۰ میلی ثانیه می‌باشد. هر فریم باید با فریم مجاور خود مقداری همپوشانی داشته باشد که به طور متداول این همپوشانی حدود ۱۰میلی ثانیه می‌باشد.

۳- طیف توان

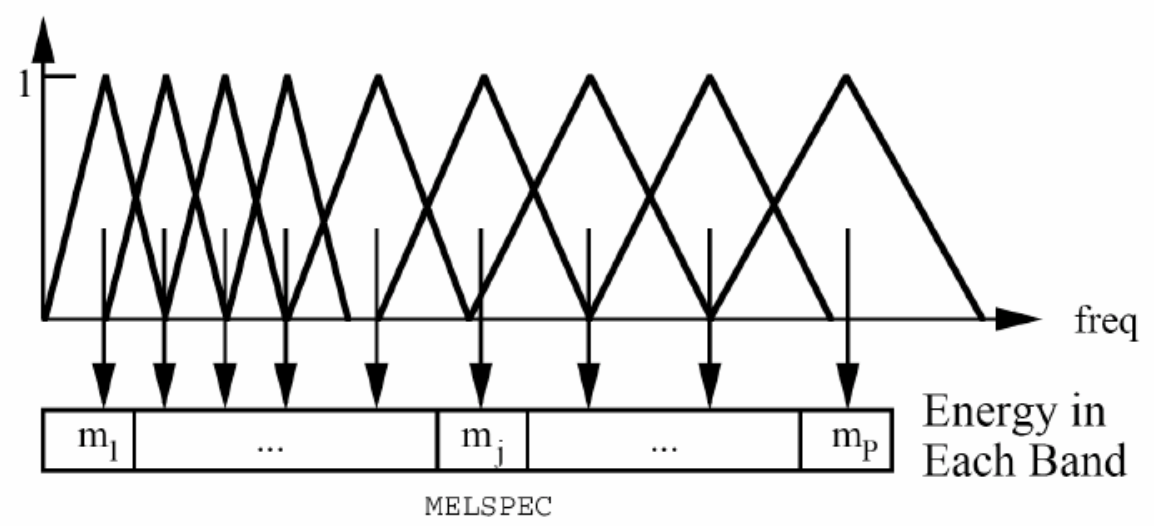
در این مرحله از سیگنال فریم بندی شده، تبدیل فوریه گرفته می‌شود. و سپس مجذور دامنه آن محاسبه می‌گردد و به نوعی طیف توان اسپکترال آن محاسبه می‌گردد.

۴- فیلتر بانک مل[1]

همانطور که گفته شد مقیاس مل بر مبنای ساختار شنیداری انسان طرح ریزی شده است. کلمه مل از لغت ملودی[2] گرفته شده است. مقدار مل به طور تجربی در آزمایشاتی مربوط به ساختار شنوایی انسان به دست آمده است که یک مل بیانگر یک‌هزارم یک پیچ با فرکانس یک کیلو هرتز می-باشد. این مقیاس در یک فیلتر بانک به همین نام استفاده شده است.

[1] - Mel Filter - Bank
[2] - Melody

در شکل (۲-۷) شمای فیلتر بانک مل نشان داده شده است.



شکل(۲-۷): فیلتر بانک مل

تعداد این فیلترها متغیر است. همانگونه که در شکل (۲-۷) مشاهده می‌شود، در فرکانس‌های بالا، پهنای باند فیلترها زیادتر است و این امر بدان معناست که حساسیت گوش انسان نسبت به تغییر فرکانس‌های پایین بیشتر است. پس از قرار دادن فیلتر روی طیف سیگنال، خروجی فیلترها محاسبه می‌شود.

برای این فیلتر بانک، از فیلترهای مقیاس مل استفاده شده است که دارای پاسخ فرکانسی مثلثی میان گذر با پهنای باند معین می‌باشد. همچنین این فیلترها دارای پنجاه درصد همپوشانی هستند که ترکیبی از لگاریتمی و خطی می‌باشند و تعداد معمول آن بین ۲۰ تا۴۰ عدد می‌باشد. با وجود ۴۰ فیلتر، این فیلتر بانک شامل ۱۳عدد فیلتر خطی و۲۷ عدد فیلتر لگاریتمی می‌باشد که به طیف توان سیگنال پنجره بندی اعمال می‌شود.

۵- لگاریتم و تحلیل کپستروم

در این مرحله از خروجی‌های فیلتر بانک لگاریتم گرفته می‌شود که به نوعی عمل فشرده سازی انجام می‌شود. از آنجاییکه این فیلترها دارای همپوشانی هستند، خروجی‌های فیلتر بانک دارای همبستگی هستند که برای کاهش این همبستگی از تبدیل فوریه کسینوسی استفاده می‌شود. البته این تبدیل تا حدودی باعث نرم شدن طیف اسپکترال[1] می‌گردد و به نوعی ضرایب را کاهش می‌دهد.

همچنین این تبدیل، هارمونیک‌ها را حذف نمی‌کند و خروجی آن مستقل از سطح سیگنال می‌باشد. خروجی‌های فیلتر بانک به نوعی توان سیگنال را در باندهای فرکانسی مختلف بیان می‌کنند. برای تبدیل فوریه کسینوسی داریم:

$$C_m = \frac{1}{N}\sum_{i=1}^{N} E_i \, Cos\left[\frac{m\pi}{N}(i-\frac{1}{2})\right] \qquad (۱۶-۲)$$

N: تعداد فیلترها است که از ۲۰ تا۴۰ متغیر است.

E_i: خروجی i امین فیلتر $1 < i < N$

C_m : m امین ضریب کپسترال $0 \le m \le M$ (C_0 متوسط log-energy است)

m : مرتبه کپستروم یا همان تعداد ضرایب کپستروم می باشد که از ۸ تا ۱۸ متغییر است .

بعد از بدست آوردن ضرایب کپستروم، باید یک دنباله‌ای از این ضرایب را انتخاب کنیم که معمولاً از همه ضرایب در دنباله ویژگی استفاده نمی‌شود. به طور متداول حدود ۱۳ ضریب را انتخاب می‌کنند. معمولاً اولین ضریب حذف می‌گردد زیرا مقدار متوسط انرژی است که حاوی اطلاعات مفیدی نمی‌باشد.

[1] - Spectral smoothing

این دنباله فقط از نوع ویژگی‌های استاتیک است و برای اضافه نمودن ویژگی‌های دینامیک باید مقدار مشتق اول و دوم را نیز به آن اضافه نمود که به ترتیب ضرایب دلتا و دلتا - دلتا نامیده می‌شوند. ضرایب ویژگی، شکل و ویژگی مجرای گفتار در هنگام بیان گفتار را نشان می‌دهند. در برخی موارد نظیر بازشناسی گفتار می‌توان از سرعت تغییر این ویژگی‌ها و نیز آهنگ تغییر آنها استفاده کرد، که این امر مزایای خود را دارد. به همین دلیل، مشتقات زمانی بردارهای ویژگی حاصل نیز در عمل استفاده می‌شود(مثلاً مشتقات اول و دوم). روشی که در محاسبه مشتق بردارها به کار می‌رود در رابطه (۲-۱۷) بیان شده است.

$$\Delta c_t^{'} = \frac{\sum_{\tau=1}^{2} \tau (c_{t+\tau} - c_{t-\tau})}{\sum_{\tau=1}^{2} \tau^2} \qquad (۲-۱۷)$$

در این رابطه Δc_t مشتق حاصل و c بردارهای ویژگی‌اند. در مرزها نیز تفاوت بردار مرزی با آخرین بردار نزدیک آن محاسبه می‌شود که در رابطه (۲-۱۸) بیان شده است.

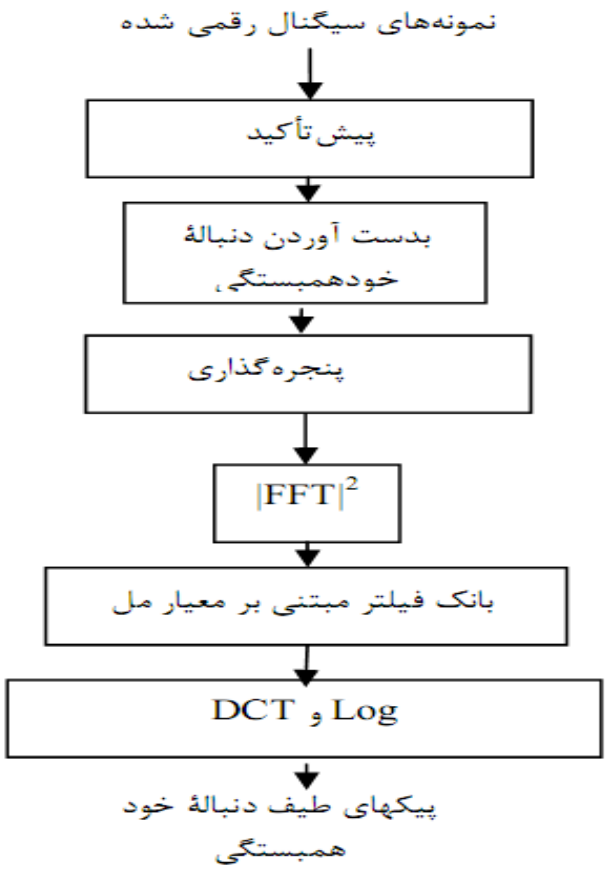
$$\Delta c_t^{'} = c_T - c_{T-1} \qquad (۲-۱۸)$$

همانطور که گفته شد هدف از استخراج ویژگی استفاده از آنها در آنالیز یا کاربردی خاص نظیر باز شناسی گوینده یا گفتار است. در مواردی نظیر بازشناسی، هدف این است که سیستم تا حد امکان راندمان بیشتری داشته باشد لذا نتایج به دست آمده تحلیل می‌شود و ایده‌های متفاوتی جهت افزایش راندمان سیستم بررسی می‌گردد.

برای مثال دیده شده است که در کاربردی نظیربازشناسی اگر ضرایب کپسترال استفاده شوند، به طور معمول ضرایب پایین به شیب کلی طیف سیگنال حساس‌اند و ضرایب بالا به نویز، که این نویز ممکن است ناشی از مواردی نظیر خطای چندی سازی یا محاسبات باشد. برای کاهش این حساسیت-ها، وزن دهی عناصر بردار کپسترال روش معمولی است و بردار کپسترال در یک بردار وزن دیگر به منظور کاهش این حساسیت‌ها ضرب می‌گردد.

۲-۵-۴-۱- استخراج ویژگی با استفاده از طیف‌های دنباله خود همبستگی

این روش برگرفته از ضرایب MFCC است با این تفاوت که در این روش از دنباله خود همبستگی به جای خود سیگنال استفاده شده است. تاثیر انواع نویزهای محیطی بر سیگنال طیف دنباله خود همبستگی کمتر از تاثیر آن بر خود سیگنال است. بلوک دیاگرام شکل (۲-۸) مراحل این روش را نشان داده است. همانطور که در این شکل نشان داده شده، ابتدا سیگنال گفتار پیش تأکید می‌شود و پس از آن دنباله خود همبستگی سیگنال بدست می‌آید. سپس سیگنال حاصل پنجره گذاری می‌شود و تبدیل فوریه آن محاسبه و طیف توان آن بدست می‌آید. به دنباله حاصل در انتها یک بانک فیلتر مبتنی بر معیار مل اعمال می‌شود و از خروجی‌های آن لگاریتم و تبدیل فوریه کسینوسی گرفته می‌شود و بدین ترتیب ویژگی‌های نهایی بدست می‌آیند[۱] و [۶].

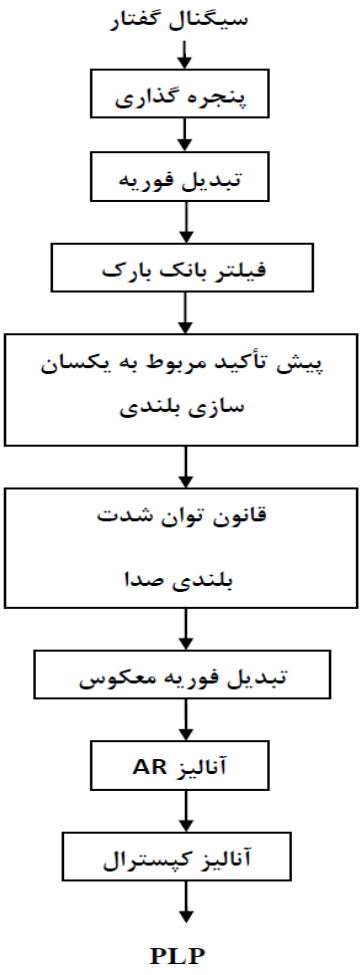


شکل(۲-۸): بلوک دیاگرام روش استخراج گفتار در حوزه خود همبستگی[۱]

در [۱] بیان شده است که در مقایسه با ضرائب MFCC، پیک‌های طیف دنباله خود همبستگی در مقابل نویز مقاومتر هستند.

۲-۵-۵- پیشگویی خطی ادراکی

الگوریتم پیشگویی خطی ادراکی(PLP) به همراه روش ضرایب کپسترال مل(MFCC) از پر کاربردترین الگوریتم‌های استخراج ویژگی در بازشناسی گفتار و گوینده می‌باشند و اینکه کدامیک دارای عملکرد بهتری می‌باشد بستگی به نوع کاربرد دارد. در مواردی که یک عدم تطابق بین دیتای تست و آموزش وجود دارد روش پیشگویی خطی ادراکی بهتر عمل می‌کند به عبارت دیگر در مقابل نویز مقاوم‌تر است. در شکل زیر بلوک دیاگرام این روش نشان داده شده است[۱۳] و [۱۴].



شکل(۲-۹) بلوک دیاگرام PLP [۱۴]

این الگوریتم شامل چندین مرحله اساسی است [۱۴] که به ترتیب عبارتند از:

۱-پیش پردازش

سیگنال گفتار با فرکانس خاصی نمونه برداری می‌شود که این فرکانس به طور معمول بین ۸ تا ۱۶کیلو هرتز می‌باشد. همچنین این مرحله شامل جداسازی حروف بی‌صدا از حروف صدادار و اطلاعات غیر ضروری یا اضافی نظیر نویزهای موجود می‌باشد.

۲-پنجره گذاری

سیگنال فیلتر شده توسط یک پنجره، فریم‌بندی می‌شود که عموماً از پنجره همینگ برای فریم بندی استفاده می‌شود. این فریم ها می‌بایست بین ۳۰ تا ۷۵درصد همپوشانی داشته باشند.

۳- طیف توان

سیگنال فریم‌بندی شده باید به حوزه فرکانس برده شود که برای این منظور از تبدیل فوریه استفاده می‌شود و سپس طیف توان آن محاسبه می‌گردد.

۴- فیلتر بانک بارک

این فیلتر بانک بر مبنای مقیاس بارک می‌باشد. مقیاس بارک بیانی بیولوژیکی از حلزونی گوش انسان است. حلزونی گوش انسان، طیف توان سیگنال‌های گفتار را در محدوده و سطوح معینی از فرکانس‌ها بیان می‌کند. برای نحوه ارتباط و تبدیل مقیاس هرتز به مقیاس بارک داریم:

$$Bark(f) = 6\log_e\left(\left(f_{Hz}/600\right)+\sqrt{\left(f_{Hz}/600\right)^2+1}\right) \qquad (۱۹-۲)$$

و همچنین برای تبدیل مقیاس بارک به هرتز داریم :

$$f_{Hz} = 13arctan\left(0.00076 f_{bark}\right) + 3.5\arctan\left[\left(f_{bark}/7500\right)^2\right] \quad (۲۰-۲)$$

فیلتر بانک بارک از یک سری باندهای فرکانسی خاص استفاده می‌کند که بر مبنای عملکرد حلزونی گوش طراحی شده است. این باندها در فرکانس‌های پایین، باریک ترند و در فرکانس‌های بالا، این وسیعتر می‌شوند. این فیلتر بانک دارای فیلترهای ذوزنقه‌ای می‌باشد که دارای پنجاه درصد همپوشانی هستند.

۵- اکولایزر[1]

حساسیت گوش انسان نسبت به شدت و انرژی صوت در فرکانس‌های مختلف، متفاوت است. جهت شبیه سازی یکسان نبودن حساسیت گوش و احساس یکسان نبودن بلندی صدا در فرکانس‌های مختلف، عمل وزن‌گذاری و اکولایز توسط یک تابع پیش تاکید بر روی خروجی‌های فیلتر بانک انجام می‌گیرد.برای این منظور از رابطه (۲-۲۱) جهت وزن گذاری خروجی‌های فیلتربانک استفاده می شود.

$$E_1(f) = \frac{\left(f^2 + 1.44*10^6\right) f^4}{\left(f^2 + 1.6*10^5\right)^2 \left(f^2 + 9.61*10^6\right)} \quad (۲۱-۲)$$

۶- قانون توان شدت بلندی صدا[2]

در این مرحله رابطه‌ای غیر خطی بین شدت صدا و بلندی صدای دریافتی برقرار می‌گردد و از آنجاییکه در گوش انسان میزان احساس بلندی صدا با ریشه سوم آن متناسب است، مقادیر اکولایز شده طبق قانون استیون به توان ۰/۳۳ رسانده می‌شود.

[1] Equlizer
[2] Intensity – Loudness Power Law

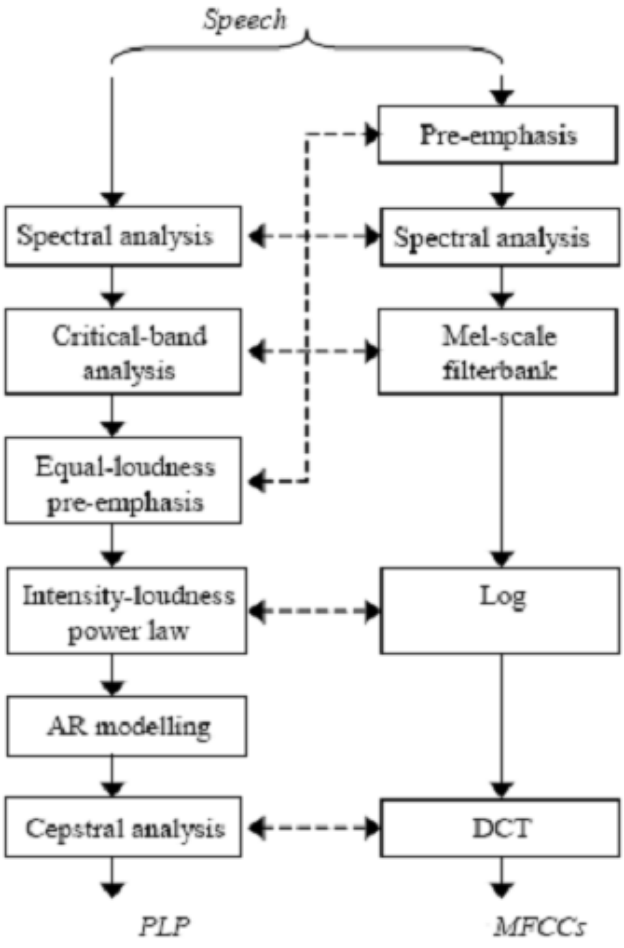
۷- پیشگویی خطی

در این مرحله ضرایب پیشگویی با اعمال روش پیشگویی خطی (LP) ، محاسبه می‌گردد. که باعث نرم شدن طیف فرکانسی می‌گردد و همچنین باعث کاهش ضرایب می‌شود.

۸- آنالیز کپسترال

در نهایت ضرایب کپسترال آن با عکس تبدیل فوریه گرفتن از لگاریتم ضرایب پیشگویی بدست می‌آید که به این ضرایب، ضرایب PLP می‌گویند

در شکل (۲-۱۰) مقایسه‌ای بین روش ضرایب PLP و MFCC انجام شده است



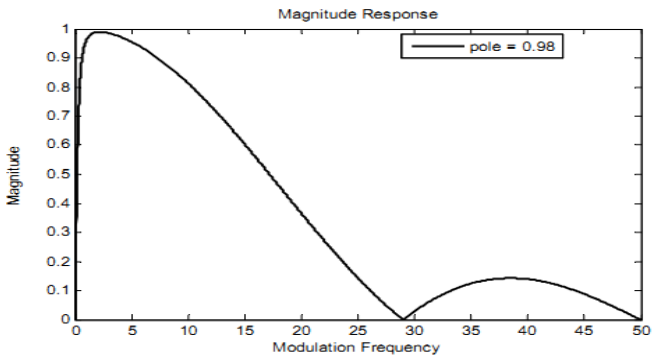
شکل(۲-۱۰) مقایسه MFCC و PLP [۱۴].

همانطور که در شکل (۲-۱۰) مشخص است فشرده سازی در ضرایب MFCC توسط تابع لگاریتم انجام می شود اما در PLP با گرفتن ریشه سوم این عمل انجام می شود. کاهش همبستگی و نرم کردن طیف اسپکترال در MFCC توسط استخراج ضرایب کپستروم با تبدیل فوریه کسینوسی انجام می شود اما در PLP با اعمال پیشگویی خطی و تحلیل کپسترال ضرایب پیشگویی انجام می-شود[۱۴].

۲-۵-۶- تحلیل *PLP-RASTA*

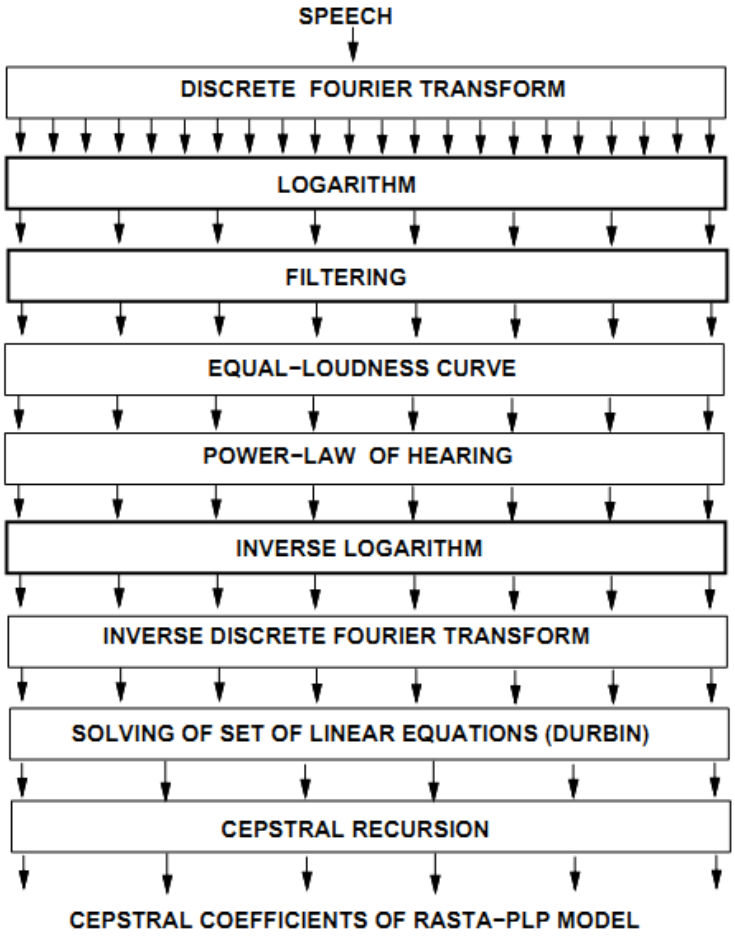
اثرات مخرب محیطی بر سیگنال گفتار، موجب کاهش کارایی سیستم‌های بازشناسی گوینده و گفتار می‌شود. روش‌هایی چون اعمال فیلتر رستا به کاهش این اعوجاجات و استخراج ویژگی‌های مقاوم در برابر تغییرات کمک می‌کند. فیلتر رستا یک فیلتر میان گذر است که در حوزه فرکانس بر دنباله زمانی طیف سیگنال اعمال شده و اعوجاجات غیر مفید فرکانس- پایین و فرکانس- بالای گفتار را کاهش می‌دهد. شرایط اولیه و مقدار قطب این فیلتر تأثیر اساسی برکارایی آن دارد و هر چه قطب فیلتر به یک نزدیکتر شود زمان پایداری فیلتر طولانی‌تر می‌شود[۳]. این فیلتر از درجه چهار است که در رابطه (۲-۲۲) نشان داده شده است. پاسخ فرکانسی این فیلتر در شکل (۲-۱۱) نشان داده شده است.

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - 0.98z^{-1})} \qquad (۲۲-۲)$$



شکل (۲-۱۱): پاسخ فرکانسی فیلتر رستا

بلوک دیاگرام استخراج ضرائب PLP- RASTA در شکل (۲-۱۲) نشان داده شده است. همانطور که در شکل (۲-۱۲) مشخص است عمل فیلترکردن بعد از لگاریتم گیری انجام شده است.



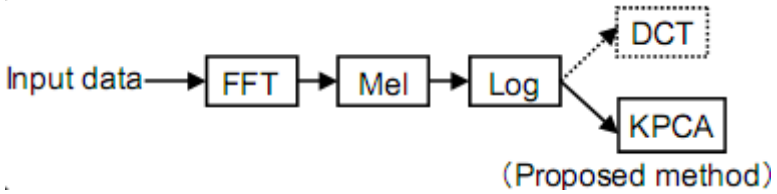
شکل (۲-۱۲): مراحل استخراج ضرائب PLP- RASTA

روشهای استخراج ویژگی زیادی وجود دارند که ضرایب PLP و MFCC بیشترین کاربرد را در بین روشهای دیگر دارند. در سال‌های اخیر تکنیک‌های استخراج ویژگی زیادی مورد بررسی قرار گرفته‌اند که در ادامه به چند روش اشاره می‌کنیم.

۲-۵-۷- ترکیب MFCC و PCA برای استخراج ویژگی

در مرجع [۱۸] بیان شده است که برای استخراج ویژگی از سیگنال گفتار به جای DCT در روش MFCC از آنالیز اجزای اصلی[1] استفاده شده است.

بلوک دیاگرام این روش در شکل (۲-۱۳) نشان داده شده است. استفاده از PCA در این روش باعث می‌شود که المان‌های اصلی گفتار به فضای مرتبه پایین و المان‌های نویز به فضای مرتبه بالا تصویر شوند. نتایج نشان داده شده در [۱۸] گویای این مطلب است که استفاده از PCA به جای DCT کارایی بهتری برای استخراج ویژگی از سیگنال گفتار نویزی دارد.



شکل (۲-۱۳): مراحل استخراج ضرایب با استفاده از MFCCوPCA [۱۸]

۲-۵-۸- استخراج ویژگی با استفاده از MFCC وزن دهی شده

در مرجع [۱۹] بیان شده است که به جای استفاده از ضرایب MFCC از ضرایب MFCC وزن دهی شده استفاده شده است. ویژگی استخراج شده جدید در رابطه (۲-۲۳) بیان شده است.

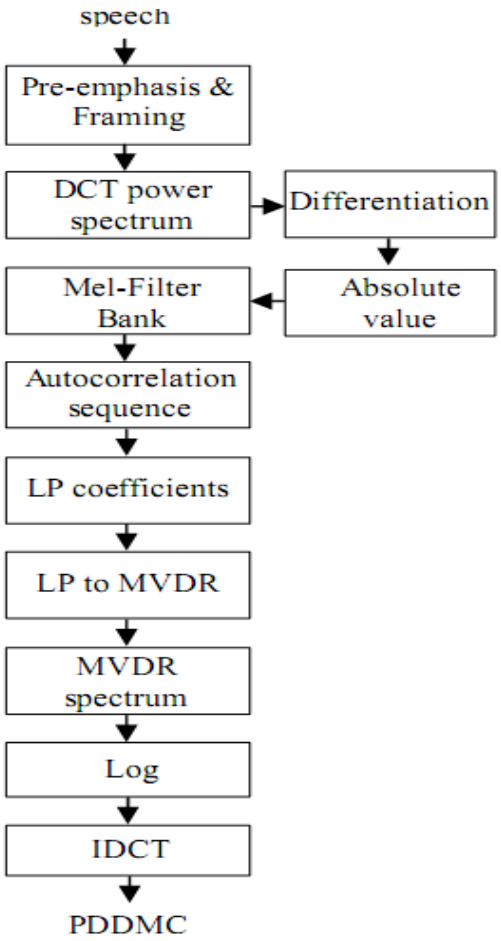
$$newMFCC = MFCC + a \cdot \Delta MFCC + b \cdot \Delta^2 MFCC \qquad (۲۳-۲)$$

در این رابطه $\Delta MFCC$ مشتق اول و $\Delta^2 MFCC$ مشتق دوم ضرایب MFCC هستند و همچنین شرط $b \langle a \langle 1$ باید برقرار باشد. نتایج گویای این مطلب است که با این ضرایب نرخ بازشناسی بهتر و همچنین پیچیدگی محاسباتی این روش از $MFCC + \Delta MFCC$ کمتر می‌باشد.

[1] Principal component analysis(PCA)

۲-۵-۹- استخراج ویژگی با استفاده از MVDR و DCT

در مرجع [۲۰] یک روش استخراج ویژگی بر اساس تبدیل فوریه کسینوسی(DCT) و تخمین طیف MVDR[1] بیان شده است که بلوک دیاگرام این روش در شکل (۲-۱۴) نشان داده شده است. روش MVDR با کاهش واریانس تخمین طیف توان، مقاوم بودن ویژگی‌ها را افزایش می‌دهد. نتایج نشان داده شده در [۲۰] گویای این مطلب است که این روش نسبت به روش متداول MFCC بار محاسباتی بیشتر و نتایج بازشناسی بهتری را ارایه می‌دهد.



شکل (۲-۱۴): مراحل استخراج ضرایب با استفاده از DCT و MVDR [۲۰]

در مرجع [۲۱] برای بازشناسایی گفتار از فیلتر گابور برای استخراج ویژگی استفاده شده است. همچنین در مرجع [۲۲] برای شناسایی گوینده مقاوم از ویژگی‌های وکال‌ترکت و وکال‌سورس[2] بدون نویز استفاده شده است.

[1] Minimum variance distortionless response
[2] Vocal source

فصـــــــل ســـــوم

توزيـــــــع ويگنـــر – ويل

۳-۱- مقدمه

تقریباً از زمانی که کار بر روی سیگنال گفتار شروع شد تحقیقات بر روی سیستم‌های تشخیص گوینده نیز جای خود را باز کرد. موضوع اصلی در سیستم‌های تشخیص گوینده استخراج ویژگی از سیگنال گفتار می‌باشد . با توجه به اینکه سیگنال گفتار یک سیگنال غیر ایستان[۱] می‌باشد و همچنین تبدیل فوریه توانایی آنالیز سیگنال‌های غیر ایستان را ندارد، باید از تبدیلات زمان- فرکانس برای آنالیز سیگنال‌های غیر ایستان استفاده کرد. تبدیلات زمان- فرکانس زیادی تاکنون معرفی شده‌اند که از معروفترین آنها تبدیل فوریه زمان کوتاه، تبدیل گابور[۲]، تبدیل ویگنر- ویل و ... می‌باشند. در این تحقیق از تبدیل ویگنر- ویل که یکی از تبدیل‌های مهم در آنالیز زمان- فرکانس سیگنال‌های غیر ایستان می‌باشد برای استخراج ویژگی استفاده شده است. توزیع ویگنر- ویل[۳] در حقیقت تبدیل فوریه[۴] تابع خود همبستگی[۵] سیگنال ورودی می‌باشد که فقط دارای مقادیر حقیقی می‌باشد. توزیع ویگنر- ویل یک توزیع زمان- فرکانس درجه دو[۶] (دو خطی) می‌باشد که بیان کننده توزیع انرژی زمان- فرکانس به جای طیف توان لحظه‌ای سیگنال می‌باشد. در مقایسه با توزیع‌های دیگر، توزیع ویگنر- ویل دارای وضوح[۷] بیشتری می‌باشد.

در این فصل ابتدا تعریفی از سیگنال‌های ایستان و غیر ایستان و ضعف تبدیل فوریه در آنالیز سیگنال‌های غیرایستان بیان شده است. سپس خلاصه‌ای از توابع توزیع زمان- فرکانس و در نهایت تئوری توزیع ویگنر- ویل و انواع آن بیان می‌شوند.

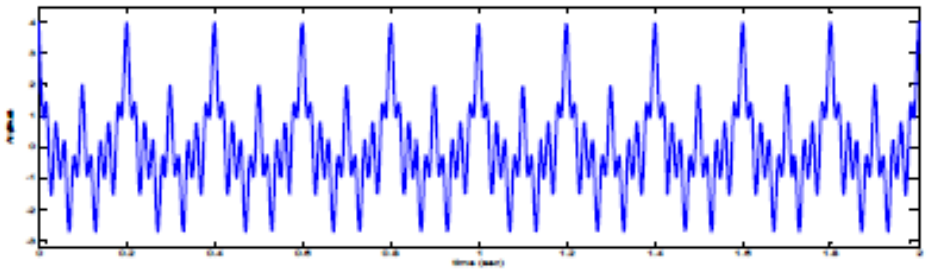
[1] Nonstationary
[2] Gabor
[3] Wigner-Ville Distribution(WVD)
[4] Fourier Transform
[5] Auto correlation
[6] Bilinear
[7] Clearity

۳-۲- سیگنال‌های ایستان[1] و غیر ایستان[2]

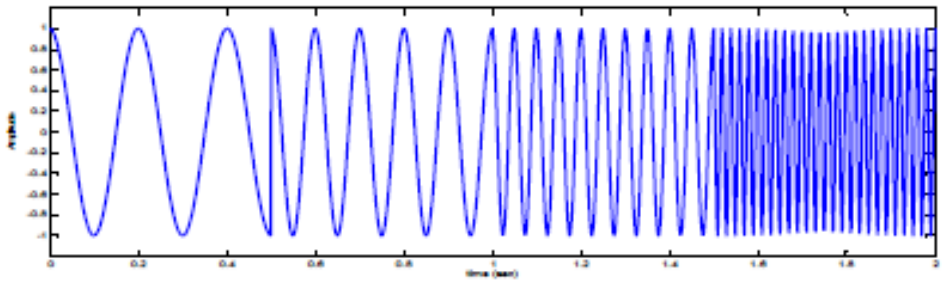
سیگنالی که محتوای فرکانسی آن در زمان تغییر نکند، سیگنال ایستان نامیده می‌شود. در این حالت نیاز نیست بدانیم که مؤلفه‌های فرکانسی در چه زمانی اتفاق می‌افتند زیرا تمام مؤلفه‌های فرکانسی در همه زمانها موجود می‌باشند. اما سیگنال غیر ایستان سیگنالی است که محتوای فرکانسی آن در زمان تغییر می‌کند. در یک سیگنال غیر ایستان تمام مؤلفه‌های فرکانسی در همه زمانها موجود نمی‌باشند[۱۵].

سیگنال نشان داده شده در رابطه (۳-۱) یک سیگنال ایستان می‌باشد که نمودار آن در شکل (۳-۱) نشان داده شده است. سیگنال نشان داده شده در شکل (۳-۲) یک سیگنال غیر ایستان می‌باشد که به سیگنال chrip معروف است.

$$x(t)=\cos(2\pi\times 10t)+\cos(2\pi\times 25t)+\cos(2\pi\times 50t)+\cos(2\pi\times 100t) \qquad (۳-۱)$$



شکل (۳-۱): سیگنال ایستان



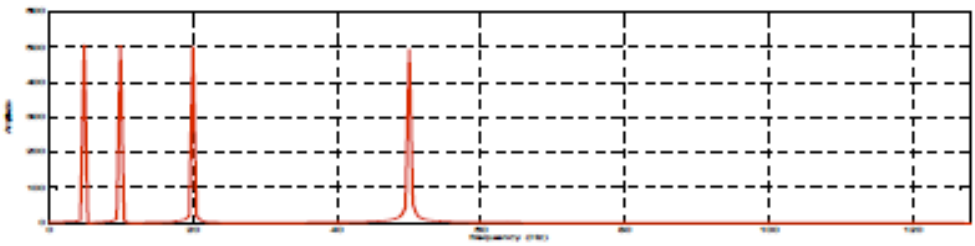
شکل (۳-۲): سیگنال غیر ایستان

[1] Stationary signal
[2] Nonstationary signal

۳-۳- مشکل تبدیل فوریه در آنالیز سیگنال‌های غیر ایستان

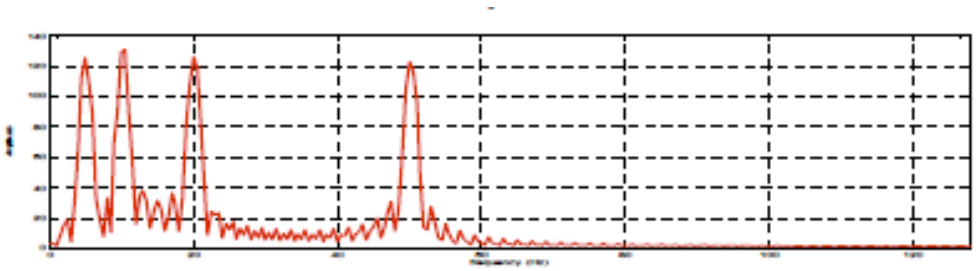
طیف فرکانسی یک سیگنال، عناصر فرکانسی آن سیگنال می‌باشد که نشان می‌دهد در آن سیگنال چه فرکانس‌هایی موجود می‌باشد. اگر متغیرهای یک سیگنال به سرعت تغییر کنند آن سیگنال فرکانس بالا می‌باشد و اگر تغییرات آنها نرم باشد فرکانس پایین می‌باشد. محتوای فرکانسی یک سیگنال با استفاده از تبدیل فوریه (FT) انجام می‌شود. اگر FT در حوزه زمان گرفته شود نمایش دامنه - فرکانس به دست می‌آید. نمودار دامنه - فرکانس نشان می‌دهد که هر فرکانس به چه مقدار در سیگنال موجود می‌باشد که به خاطر تقارن هرمیتی[1] در سیگنال‌های حقیقی معمولاً طیف مثبت را درنظر می‌گیریم.

در تبدیل فوریه فقط اطلاعات حوزه فرکانس موجود می‌باشد و اطلاعاتی از حوزه زمان نداریم و همینطور برای معکوس تبدیل فوریه فقط اطلاعات حوزه زمان را داریم و اطلاعاتی از حوزه فرکانس نداریم. تبدیل فوریه نشان می‌دهد که هر فرکانسی به چه مقداری در سیگنال موجود می‌باشد، اما این آگاهی را نمی‌دهد که در چه زمان‌هایی این مؤلفه‌های فرکانسی وجود دارند. در شکل (۳-۳) تبدیل فوریه شکل (۳-۱) ودر شکل (۳-۴) تبدیل فوریه شکل (۳-۲) نشان داده شده است.



شکل(۳-۳): تبدیل فوریه شکل(۳-۱)

[1] Hermisian Symitric



شکل(۳-۴): تبدیل فوریه شکل(۳-۲)

در یک نگاه کلی، هر دو طیف نشان دهنده وجود ۴ فرکانس غالب در سیگنال می‌باشند اما هیچ گونه اطلاعات زمانی(مکانی) در اختیار نمی‌گذارند. لذا تبدیل فوریه ابزار مناسبی برای تمایز بین این دو سیگنال نیست. به همین دلیل از تبدیلات زمان- فرکانس برای آنالیز سیگنال‌های غیر ایستان استفاده می‌شود.

۳-۴- *نمایش زمان – فرکانس*[1]

نمایش زمان – فرکانس (TFR) نمایشی از سیگنال (که تابعی از زمان می‌باشد) در هر دو حوزه زمان و فرکانس می‌باشد. اغلب TFR در حوزه زمان و فرکانس، مقادیر مختلط دارند که قدر مطلق یا دامنه آنها نمایانگر چگالی انرژی می‌باشد و زاویه آنها نمایانگر فاز می‌باشد.

یک سیگنال به عنوان تابعی از زمان، دارای نمایشی با رزولوشن زمانی کامل می‌باشد اما هیچ اطلاعاتی از فرکانس ندارد. در مقابل، دامنه تبدیل فوریه یک سیگنال، نمایشی از رزولوشن فرکانسی کامل است اما هیچ اطلاعاتی از زمان ندارد زیرا تبدیل فوریه تنها شامل محتوای فرکانسی بوده و زمان اتفاق افتادن آن فرکانس را نشان نمی‌دهد[۱۵].

نمایش زمان- فرکانس پلی بین این دو نمایش می‌باشد زیرا آنها به طور همزمان هم دارای اطلاعات زمانی و هم اطلاعات فرکانسی می‌باشند. بنابراین TFR ها برای تحلیل سیگنال‌هایی که شامل چندین فرکانس متغیر در زمان می‌باشند، مفید هستند[۱۵].

[1] Time-frequency Representation (TFR)

در پردازش بسیاری از سیگنال‌ها، تحلیل زمان- فرکانس بدنه روشهای توصیف و پردازش سیگنال‌های غیرایستان مانند سیگنال‌های گذرا[1] در زمان می‌باشد. برای سیگنال‌هایی که فرکانس‌های آن در زمان ثابت هستند، تبدیل فوریه کافی است ولی برای سیگنال‌هایی مانند گفتار[2] و موزیک[3] وتصویر[4] و سیگنال‌های پزشکی[5] تحلیل زمان- فرکانس مناسب تر است.

از معروفترین تابع‌های توزیع زمان- فرکانس عبارتند از:

1- short-time Fourier transform (including the Gabor transform)

2- Cohen's class distribution function (Wigner distribution function)

3- modified Wigner distribution function

4- Gabor-Wigner distribution function

[1] Transient
[2] Speech
[3] Music
[4] Image
[5] Medical

یک تابع توزیع زمان- فرکانس ایده آل باید دارای ۴ ویژگی باشد[۱۵]:

۱- وضوح بالا[1]: هر چه تابع زمان- فرکانس دارای وضوح بالاتری باشد تحلیل آن سیگنال راحت‌تر می-شود.

۲- ترم تداخلی: هر چه تابع زمان- فرکانس دارای ترم تداخلی کمتری باشد تشخیص المان‌های نویز از المان‌های غیر نویز آسانتر است.

۳- خصوصیات ریاضی مناسب[2] : هر چه تابع زمان- فرکانس دارای خصوصیات ریاضی مناسب‌تری باشد آن را کاراتر می‌کند.

۴- پیچیدگی محاسباتی[3]: هر چه تابع زمان- فرکانس دارای پیچیدگی محاسباتی کمتری باشد زمان کمتری برای محاسبه نیاز دارد.

در جدول(۳-۱) سه تابع توزیع با هم مقایسه شده‌اند.

جدول(۳-۱): مقایسه چند تابع توزیع

تابع زمان-فرکانس	وضوح	ترم تداخلی	خصوصیات ریاضی	پیچیدگی محاسباتی
تبدیل گابور	بدترین	ندارد	بدترین	کم
تبدیل ویگنر ویل	بهترین	دارد	بهترین	زیاد
تبدیل گابور-ویگنر	خوب	بسیار کم	خوب	زیاد

[1] High clarity
[2] Good mathematical properties
[3] Computational complexity

برای یک آنالیز بهتر، انتخاب یک تابع توزیع مناسب مهم است. وضوح بالای تابع توزیع ویگنر (WDF) وابسته به تابع خود همبستگی می‌باشد که این موضوع باعث به وجود آمدن ترم تداخلی می‌شود. بنابراین اگر بخواهیم یک سیگنال یک ترمه[1] را آنالیز کنیم، توزیع ویگنر- ویل مناسب ترین تابع توزیع می‌باشد. بعضی از کاربردهای تحلیل زمان- فرکانس عبارتند از[۱۵]:

۱- یافتن مقدار لحظه‌ای فرکانس

۲- طراحی فیلتر

۳- تجزیه سیگنال

۴- نمونه برداری[2]

۵- مدولاسیون و مالتی پلکس کردن

۶- اپتیک[3]

۷- تشخیص سیگنال[4]

۸- صوت شناسی

۹- انتشار موج الکترومغناطیس

بر خلاف ارایه زمان- فرکانس خطی که سیگنال را به اجزای خطی تفکیک می‌کند هدف توزیع انرژی، تعمیم دادن انرژی سیگنال به دو متغییر زمان و فرکانس می‌باشد. نمایش زمان- فرکانس سیگنال‌ها، به فرم دو جمله‌ای و خطی تقسیم می‌شود که در ادامه توضیح داده شده است.

[1] single-term
[2] Sampling
[3] Optic
[4] Signal detection

۳-۴-۱- فرم دو جمله‌ای[1]

یک نوع از TFR مقایسه یک سیگنال با خودش می‌باشد که در جهت‌های مختلف نسبت به یک نقطه جابه‌جا می‌شود. این فرمول بندی به عنوان quadratic TFRs(QTFRs) شناخته می‌شود زیرا نمایشی از توان دو سیگنال می‌باشد[۱۵].

این فرمول بندی اولین بار توسط Eugene Wigner در سال ۱۹۳۲ در زمینه مکانیک کوانتوم ارایه شد و بعداً در سال ۱۹۴۸ توسط Ville به عنوان یک TFR عمومی، تغییر یافت که با عنوان Wigner–Ville distribution شناخته شد.

امروزه QTFR های مختلف وجود دارد که تنها به اسپکتوگرام[2] (توان دوم دامنه تبدیل فوریه زمان کوتاه[3]) و اسکلوگرام[4] (توان دوم تبدیل ویولت[5]) و smoothed pseudo-Wigner distribution (SPWV) ختم نمی‌شود. هرچند QTFR ها، به طور همزمان رزولوشن زمانی و فرکانسی خوبی دارند، اما طبیعت درجه دو تبدیل‌ها، باعث ترم تداخلی[6] می شود که در ادامه توضیح داده می‌شود.

۳-۴-۲- فرم خطی[7]

ترم تداخلی که یک مشکل اساسی TFR های درجه دو می‌باشد ممکن است توسط مقایسه سیگنال با یک تابع دیگر برطرف شود. چنین نمایش‌هایی linear TFR نامیده می‌شوند زیرا نسبت به سیگنال خطی است[۱۵]. برای مثال‌هایی از این نمونه می‌توان به تبدیل فوریه پنجره بندی شده[8](که به عنوان تبدیل فوریه زمان کوتاه شناخته می‌شود) و تبدیل ویولت اشاره کرد.

[1] Quadratic
[2] Spectogram
[3] Short-time Fourier Transform
[4] Scaleogram
[5] Wavelet Transform
[6] Cross-term
[7] Linear
[8] Windowed Fourier transform

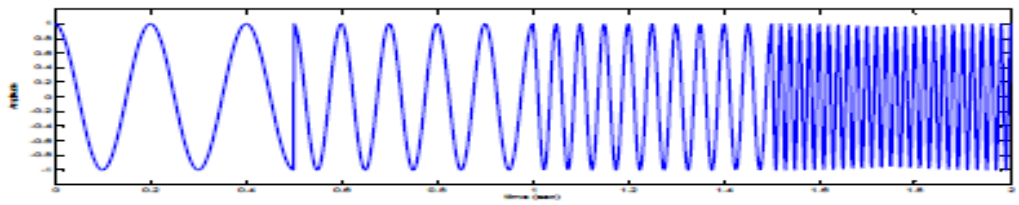
۳-۵- توزیع ویگنر - ویل

تابع توزیع ویگنر- ویل (WVD) اولین بار در سال ۱۹۳۲ توسط Eugene Wigner، برای تصحیح مکانیک آماری کلاسیک مطرح شد. تابع توزیع ویگنر- ویل یک تبدیل مناسب برای آنالیز زمان- فرکانس می‌باشد. در مقایسه با تبدیل فوریه زمان کوتاه، مانند تبدیل گابور، توزیع WDF در موارد مشابه وضوح بیشتری دارد. بنابراین WDF بیشتر ویژگی‌هایی که سایر تبدیلات آنالیز زمان- فرکانس دارا می‌باشند را دارد. هرچند تعاریف مختلفی برای WDF وجود دارد، تعریفی که مخصوص آنالیز زمان- فرکانس است به صورت روابط (۳-۲) و (۳-۳) بیان می‌شود که $W_x(t,v)$ ویگنر- ویل سیگنال $\mathbf{x}$ می‌باشد[۱۶].

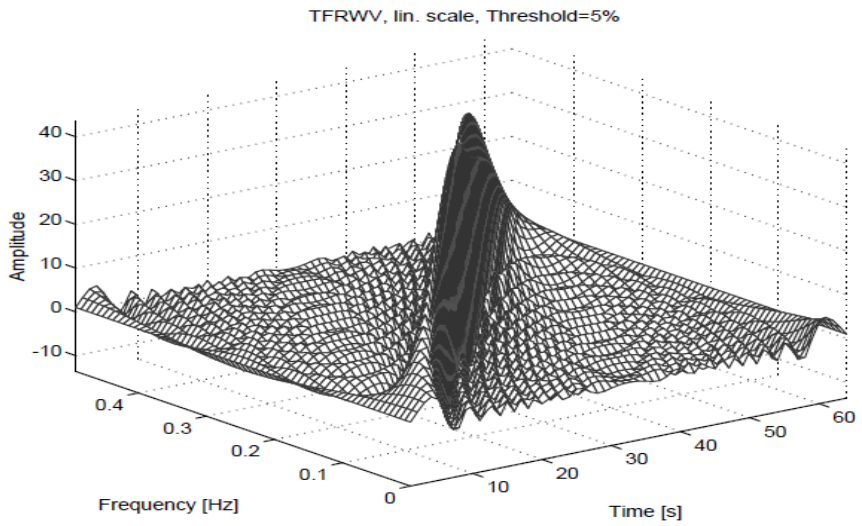
$$W_x(t,v)=\int_{-\infty}^{+\infty} x(t+\tau/2)x^*(t-\tau/2)e^{-j2\pi v\tau}d\tau \qquad (۳-۲)$$

$$W_x(t,v)=\int_{-\infty}^{+\infty} x(v+\xi/2)x^*(v-\xi/2)e^{j2\pi\xi t}d\xi \qquad (۳-۳)$$

به طور ویژه تابع توزیع ویگنر- ویل همیشه دارای مقادیر حقیقی است و شیفت زمان فرکانس را نگه می‌دارد و ویژگی‌های حاشیه‌ای را هم برآورده می‌کند[۱۶]. توزیع ویگنر ویل شکل (۳-۵) که به سیگنال chirp معروف است در شکل (۳-۶) نشان داده شده است. همانطور که در شکل مشخص است با افزایش زمان، فرکانس سیگنال هم افزایش می یابد.



شکل(۳-۵): سیگنال chirp



شکل (۳-۶): توزیع ویگنر- ویل سیگنال chirp [۱۶]

۳-۵-۱- ویژگی های توزیع ویگنر ویل

توزیع ویگنر ویل دارای خواص و ویژگی‌هایی می‌باشد که منجر شده است این توزیع در بعضی کاربردها مفید واقع شود. در ادامه این خواص مختصراً شرح داده می‌شود[۱۶].

۱. نگه داشتن انرژی[1]

اگراز توزیع ویگنر- ویل تابع x روی تمام صفحه زمان- فرکانس انتگرال گرفته شود انرژی کل سیگنال x بدست می‌آید که در رابطه (۳-۴) نشان داده شده است.

$$E_x = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} W_x(t,v)\,dtdv \qquad (۳-۴)$$

این پارامتر می‌تواند یک ویژگی از هر سیگنالی باشد.

[1] Energy conservation

۲. ویژگی‌های حاشیه‌ای[1]

چگالی انرژی طیفی و قدرت لحظه‌ای سیگنال x می‌توانند از توابع حاشیه‌ای توزیع ویگنر- ویل تابع x بدست آیند. یعنی انتگرال‌گیری در یک زمان خاص از تابع توزیع روی تمام فرکانس‌ها مقدار انرژی لحظه‌ای را در آن زمان و انتگرال‌گیری در یک فرکانس خاص از تابع توزیع روی تمام زمان‌ها مقدار چگالی طیف توان در آن فرکانس را می‌دهد.

$$\int_{-\infty}^{+\infty} w_x(t,v)dt = |X(v)|^2 \qquad (۵-۳)$$

$$\int_{-\infty}^{+\infty} w_x(t,v)dv = |X(t)|^2 \qquad (۶-۳)$$

۳. حقیقی بودن[2]

تابع توزیع ویگنر- ویل به ازای هر مقدار از t و v دارای مقادیر حقیقی است

$$w_x(t,v) \in R, \forall t,v \qquad (۷-۳)$$

۴. انتقال

یکی از خواص WVD خاصیت انتقال می‌باشد یعنی اگر سیگنال مورد نظر در زمان یا فرکانس انتقال یابد توزیع ویگنر- ویل آن هم طبق روابط (۳-۱۴) و (۳-۱۵) در زمان یا فرکانس انتقال میابد.

$$y(t) = x(t-t_0) \Rightarrow w_y(t,v) = w_x(t-t_0,v) \qquad (۸-۳)$$

$$y(t) = x(t)e^{j2\pi v_0 t} \Rightarrow w_y(t,v) = w_x(t,v-v_0) \qquad (۹-۳)$$

[1] Marginal properties
[2] Real- valued

۵. خاصیت ضرب

اگر سیگنال y حاصل‌ضرب سیگنال‌های m و x باشد توزیع ویگنر- ویل سیگنال y برابر کانولوشن فرکانسی توزیع ویگنر- ویل سیگنال‌های x و m می‌باشد. این خاصیت در رابطه (۳-۱۰) نشان داده شده است.

$$y(t)=m(t)x(t)\Rightarrow W_y(t,v)=\int_{-\infty}^{+\infty}W_m(t,v-\xi)W_x(t,\xi)d\xi \qquad (۳-۱۰)$$

۶. فیلتر کردن

یعنی اگر سیگنال y برابر کانولوشن سیگنال‌های x و h باشد آنگاه توزیع ویگنر- ویل سیگنال y برابر کانولوشن زمانی ویگنر- ویل سیگنال‌های x و h می‌باشد که در رابطه (۳-۱۱) نشان داده شده است.

$$y(t)=\int_{-\infty}^{+\infty}h(t-s)x(s)ds\Rightarrow W_y(t,v)=\int_{-\infty}^{+\infty}W_h(t-s,v)W_x(s,v)ds \qquad (۳-۱۱)$$

۷. فرکانس لحظه‌ای[1]

این ویژگی مقدار فرکانس در یک لحظه خاص از سیگنال را نشان می‌دهد. فرکانس لحظه‌ای سیگنال x را می‌توان از گشتاور مرتبه اول در فرکانس ویگنر- ویل تابع x_a بدست آورد که x_a سیگنال تحلیلی وابسته به x می‌باشد. این خاصیت در رابطه (۳-۱۲) نشان داده شده است.

$$f_x(t)=\frac{\int_{-\infty}^{+\infty}vW_{x_a}(t,v)dv}{\int_{-\infty}^{+\infty}W_{x_a}(t,v)dv} \qquad (۳-۱۲)$$

[1] Instantaneous frequency

۸. تاخیر گروهی[1]

این کمیت که میانگین زمانی ورود فرکانس v را اندازه گیری می‌کند، می‌تواند از گشتاور مرتبه اول در زمان ویگنر- ویل تابع x_a بدست آید که x_a سیگنال تحلیلی وابسته به سیگنال x می‌باشد و در رابطه (۳-۱۳) نشان داده شده است.

$$t_x(v) = \frac{\int_{-\infty}^{+\infty} tW_{x_a}(t,v)dt}{\int_{-\infty}^{+\infty} W_{x_a}(t,v)dt} \qquad (۱۳-۳)$$

۹. نگه داشتن شرایط بازه‌ای

این خاصیت بیان می‌دارد که اگر سیگنال در حوزه زمان یا فرکانس خارج از یک محدوده خاصی صفر باشد، توزیع ویگنر- ویل آن هم خارج از آن نواحی صفر می‌باشد روابط (۳-۱۴) و (۳-۱۵) بیانگر این خاصیت می‌باشند.

$$x(t) = 0, |t| > T \Rightarrow W_x(t,v) = 0, |t| > T \qquad (۱۴-۳)$$

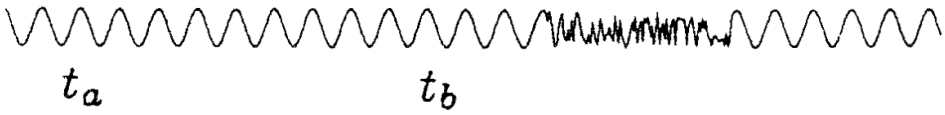
$$X(v) = 0, |v| > B \Rightarrow W_x(t,v) = 0, |v| > B \qquad (۱۵-۳)$$

۳-۵-۲- *ترم تداخلی*

اگر جنبه هندسی توزیع ویگنر- ویل را مورد بررسی قرار دهیم مقدار ویگنر- ویل در یک نقطه را می‌توان با تا زدن قسمت چپ سیگنال در آن نقطه روی قسمت راست سیگنال در آن نقطه و سپس تبدیل فوریه گرفتن از آن بدست آورد[۱۶].

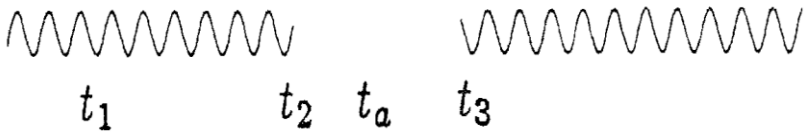
[1] Group delay

با این تعریف در شکل (۳-۷) مقدار ویگنر- ویل در نقطه t_b شامل نویزهای زمان‌های بعدی آن هم می‌شود



شکل (۳-۷): نمایش ترم تداخلی در سیگنال

حال با توجه به این تفسیر در شکل (۳-۸) در نقطه t_a مقدار واقعی سیگنال صفر می‌باشد اما مقدار ویگنر- ویل در این نقطه صفر نمی‌باشد که این موضوع ترم‌های تداخلی[1] را توجیه می‌کند. یعنی در جاهایی که مقدار خود سیگنال صفر است مقدار ویگنر- ویل آن صفر نمی‌باشد.



شکل (۳-۸): نمایش ترم تداخلی در سیگنال

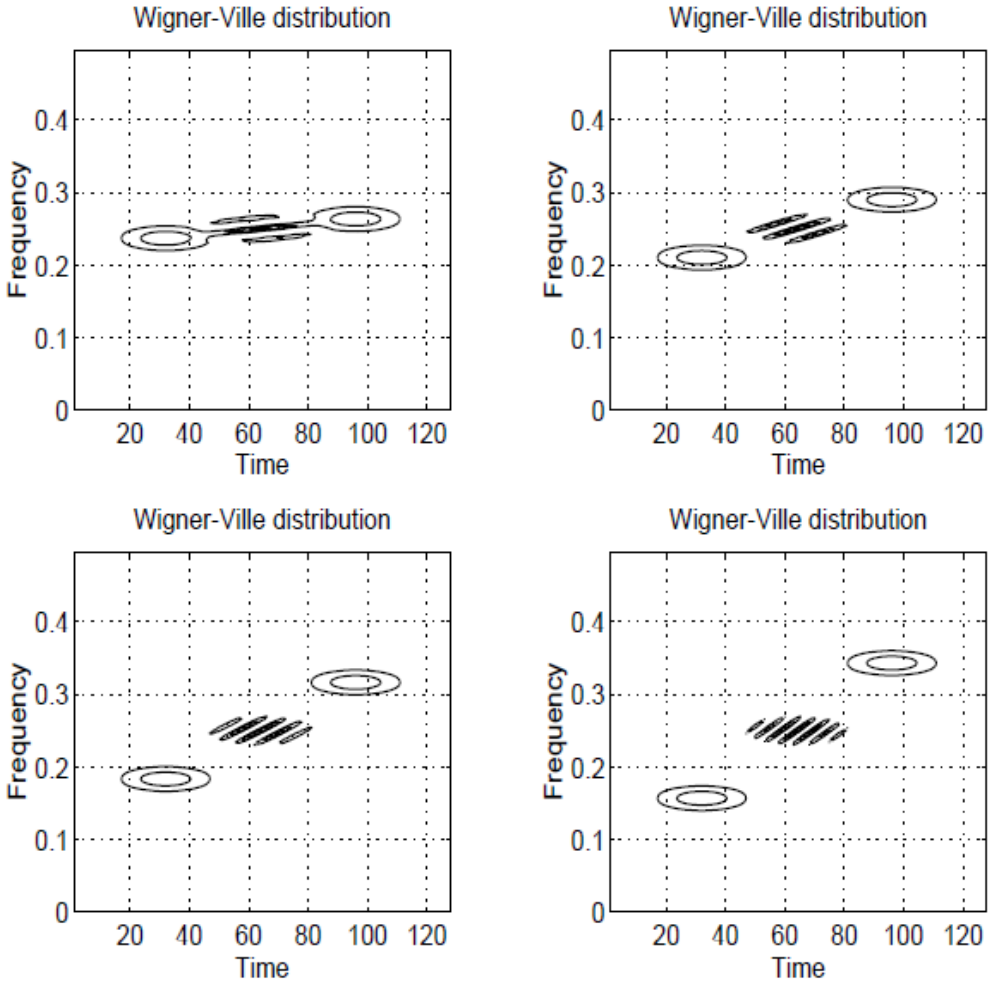
بنابراین ویگنر- ویل مجموع دو سیگنال برابر مجموع ویگنر- ویل دو سیگنال نیست که این موضوع در روابط (۳-۱۶) و (۳-۱۷) نشان داده شده است و $W_{x,y}(t,v)$ ترم تداخلی می‌باشد.

$$W_{x+y}(t,v) = W_x(t,v) + W_y(t,v) + 2R\left\{W_{x,y}(t,v)\right\} \qquad (۳-۱۶)$$

$$W_{x,y}(t,v) = \int_{-\infty}^{+\infty} x(t+\tau/2)y^*(t-\tau/2)e^{-j2\pi v\tau}d\tau \qquad (۳-۱۷)$$

در شکل(۳-۹) ساختار ترم‌های تداخلی بین دو جزء گوسی با موقعیت‌های متفاوت را نشان می‌دهد. یعنی اگر یک سیگنال با دو جزء گوسی داشته باشیم توزیع ویگنر- ویل آن شامل دو گوسی نیست و ترمهای اضافی نیز تولید می‌کند.

[1] Cross term



شکل (۳-۹): ساختار ترم تداخلی بین دو جزء گوسی با موقعیت‌های متفاوت [۱۶]

در این شکل دوایر ترم‌های اصلی و خطوط بین این دوایر ترم‌های تداخلی را نشان می‌دهند. این ترم‌های تداخلی تفسیر توزیع ویگنر- ویل را از لحاظ دیداری مشکل می‌کند زیرا ممکن است ترم-های تداخلی با ترم‌های اصلی تداخل داشته باشند. بنابراین باید مصالحه‌ای[1] بین ترم‌های تداخلی و ویژگی‌های خوب توزیع ویگنر- ویل وجود داشته باشد. همچنین این شکل‌ها نشان می‌دهند که هر چه فاصله بین این دوایر بیشتر شود حضور ترم‌های تداخلی نیز بیشتر می‌شود.

[1] Trade off

۳-۵-۳- نمونه برداری از تابع توزیع ویگنر- ویل

به دلیل طبیعت دو جمله‌ای توزیع ویگنر- ویل نمونه‌برداری باید با دقت کافی انجام شود[۱۶]. با توجه به فرمول توزیع ویگنر- ویل که در رابطه (۳-۱۸) بیان شده است

$$w_x(t,v) = 2\int_{-\infty}^{+\infty} x(t+\tau)x^*(t-\tau)e^{-j\,4\pi v\tau}\,d\tau \qquad (۱۸-۳)$$

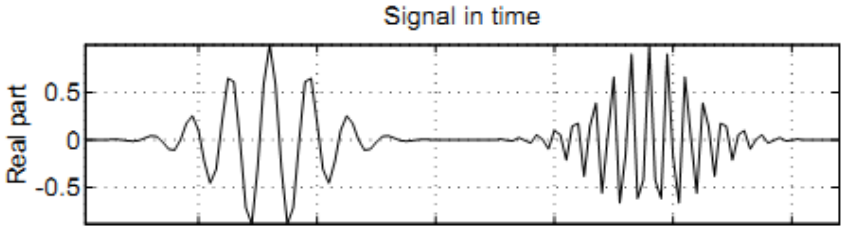
اگر از سیگنال x با دوره تناوب Te نمونه‌برداری کنیم یعنی $x[n]=x(nT_e)$ ، و تابع توزیع ویگنر- ویل را در نقاط nT_e در زمان ارزیابی[1] کنیم، عبارت زمان گسسته- فرکانس پیوسته، مانند رابطه(۳-۱۹) بدست می‌آید.

$$w_x[n,v) = 2T_e\sum_k x[n+k]x^*[n-k]\,e^{j\,4\pi vk} \qquad (۱۹-۳)$$

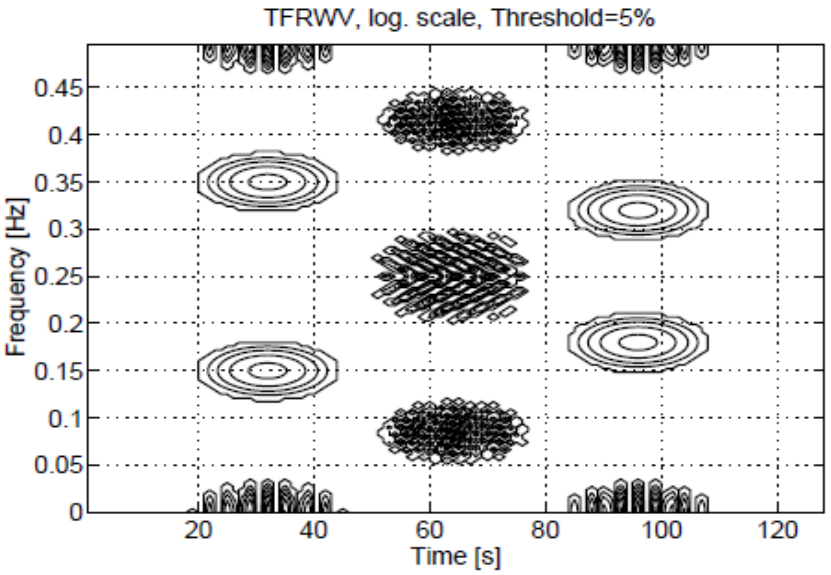
با توجه به این که این عبارت با پریود ۱/۲Te در فرکانس متناوب می باشد (بر خلاف پریود $1/T_e$ بدست آمده برای تبدیل فوریه سیگنال نمونه‌برداری شده در نرخ نایکوئیست) اگر سیگنال حقیقی باشد و با نرخ نایکوئیست نمونه‌برداری شود ممکن است نمونه گسسته WVD تحت تاثیر تداخل فرکانسی قرار گیرد. برای حل این مشکل باید از سیگنال تحلیلی وابسته به سیگنال x استفاده کرد.

اگر سیگنال x دارای دو ترم گوسی اصلی باشد که در شکل (۳-۱۰) نشان داده شده است آنگاه شکل (۳-۱۱) توزیع ویگنر ویل سیگنال x و شکل (۳-۱۲) توزیع ویگنر ویل سیگنال x_a می‌باشد که x_a سیگنال تحلیلی x می‌باشد.

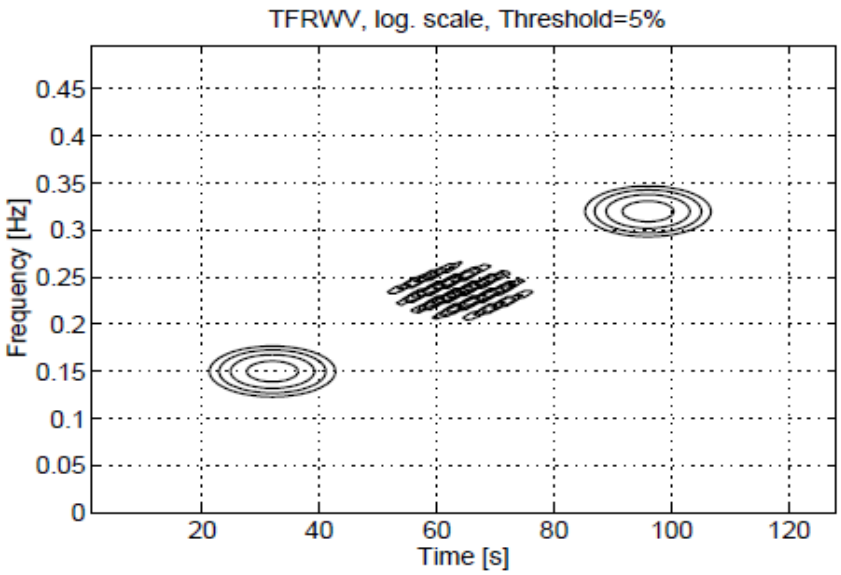
[1] Evaluate



شکل (۳-۱۰): سیگنالی با دو جزء گوسی[۱۶]



شکل (۳-۱۱): توزیع ویگنر- ویل سیگنال غیر تحلیلی متشکل از دو گوسی[۱۶]



شکل (۳-۱۲): توزیع ویگنر- ویل سیگنال تحلیلی متشکل از دو گوسی[۱۶]

همانطور که در شکلهای (۳-۱۱) و (۳-۱۲) نشان داده شده است اگر سیگنال تحلیلی باشد ترم‌های اصلی به تعداد واقعی و ترم‌های تداخلی کمتر می‌باشند. پس برای ویگنر- ویل گرفتن از یک سیگنال باید سیگنال تحلیلی باشد.

۳-۵-۴- *توزیع ویگنر ویل هموار شده در فرکانس*[1]

چون در تابع توزیع ویگنر- ویل انتگرال از منفی بی‌نهایت تا مثبت بی‌نهایت تغییر می‌کند، در عمل می‌تواند ایجاد مشکل کند بنابراین باید عبارت زیر انتگرال را به یک نمونه پنجره شده تبدیل کرد[۱۶]. که در فرمول (۳-۲۰) نشان داده شده است.

$$PW_x(t,v) = \int_{-\infty}^{+\infty} h(\tau) x(t+\tau/2) x^*(t-\tau/2) e^{-j2\pi v\tau} d\tau \qquad (۲۰-۳)$$

که در این عبارت h(t) یک پنجره می‌باشد. این پنجره‌گذاری معادل هموار سازی فرکانسی می-باشد که دررابطه (۳-۲۱) نشان داده شده است.

$$PW_x(t,v) = \int_{-\infty}^{+\infty} H(v-\xi) W_x(t,\xi)\, d\xi \qquad (۲۱-۳)$$

در رابطه (۳-۲۱)، H(v) تبدیل فوریه پنجره h(t) می‌باشد.

این عمل باعث می‌شود که ترم‌های تداخلی در مقایسه با تابع توزیع ویگنر- ویل تضعیف شود اما نتیجه این بهبود خوانایی این است که بسیاری از ویژگی‌های تابع توزیع ویگنر- ویل از قبیل ویژگی‌های حاشیه‌ای و ... از دست می‌رود. همچنین عرض فرکانسی ترم‌های اصلی توسط این عمل افزایش می‌یابد.

[1] Pseudo-WVD(PWVD)

۳-۵-۵- توزیع ویگنر- ویل هموار شده در زمان و فرکانس[1]

اگر علاوه بر هموارسازی روی فرکانس، هموارسازی روی زمان نیز داشته باشیم می‌توانیم ترم-های تداخلی را هم در راستای محور زمان و هم در راستای محور فرکانس حذف کنیم که به آن Smoothed pseudo WVD یا توزیع ویگنر- ویل هموار شده در زمان و فرکانس می‌گویند[۱۶]. یعنی می‌توانیم کنترل هموار سازی را به طور همزمان روی زمان و فرکانس داشته باشیم این عبارت در معادله (۳-۲۲) نشان داده شده است.

$$SPW_x(t,\nu) = \int_{-\infty}^{+\infty} h(\tau) \int_{-\infty}^{+\infty} g(s-t)\, x(s+\tau/2)\, x^*(s-\tau/2)\, ds\; e^{-j2\pi\nu\tau}\, d\tau \qquad (۲۲-۳)$$

این درجه آزادی هموار سازی در محور زمان و فرکانس را با رابطه (۳-۲۲) انجام می‌دهیم

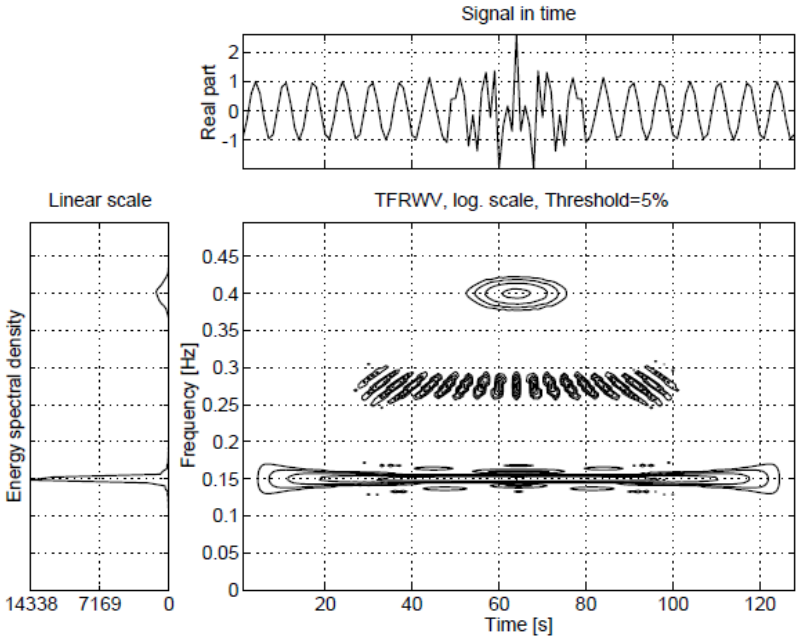
$$\Pi(t,\nu) = g(t)\, H(-\nu) \qquad (۲۳-۳)$$

که H(v) تبدیل فوریه تابع هموار سازی پنجره h(t) می باشد.

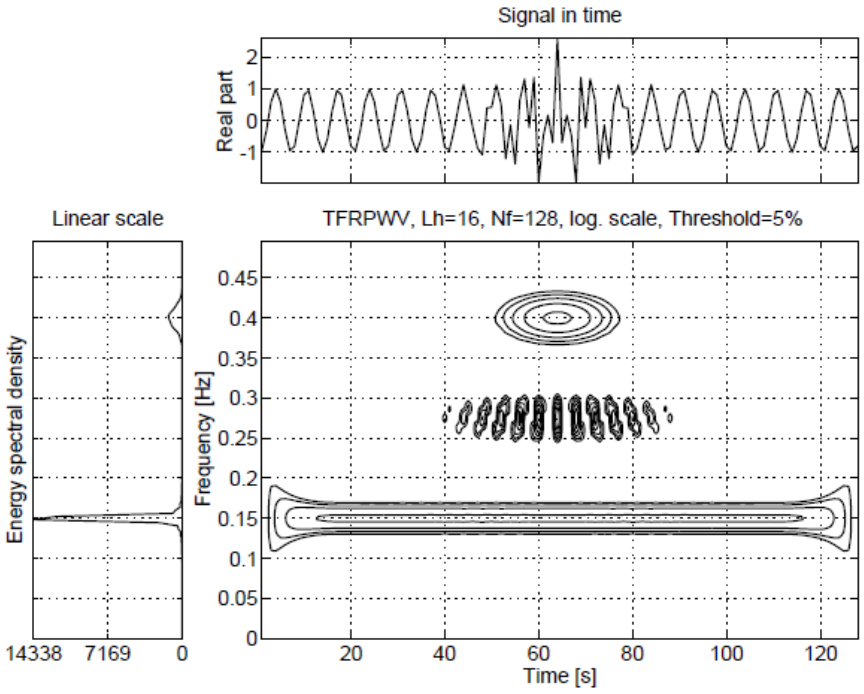
در اسپکتروگرام، مصالحه بین رزولوشن زمانی و فرکانسی است اما در SPWVD مصالحه بین رزولوشن زمان- فرکانس و سطح ترم تداخلی است یعنی هموارسازی بیشتر در زمان و یا فرکانس، رزولوشن کمتر را در زمان و یا فرکانس می‌دهد. توجه کنید که اگر هموار سازی فقط در فرکانس می-بود SPWVD به PWVD تبدیل می‌شد.

در اشکال (۳-۱۳) و (۳-۱۴) و (۳-۱۵) مقایسه‌ای بین WVD و PWVD و SPWVD انجام شده است. سیگنال پردازش شده مجموع یک فرکانس ثابت و تابع گوسی است

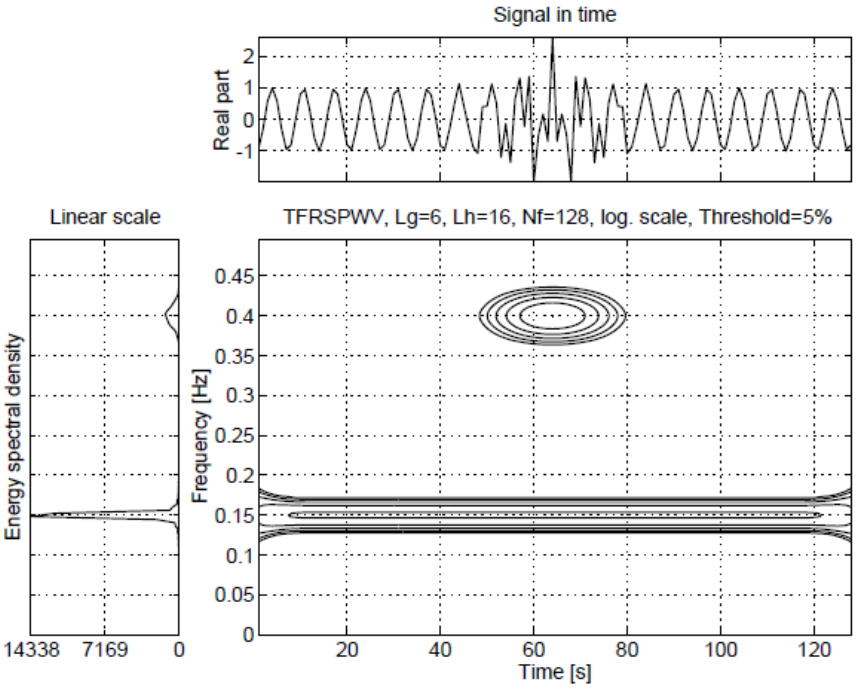
[1] Smoothed pseudo WVD



شکل (۳-۱۳): توزیع WVD سیگنالی متشکل از یک گوسی و یک سینوسی[۱۶].



شکل (۳-۱۴): توزیع PWVD سیگنالی متشکل از یک گوسی و یک سینوسی[۱۶].



شکل (۳-۱۵): توزیع SPWVD سیگنالی متشکل از یک گوسی و یک سینوسی[۱۶].

همانطور که در این مثال نشان داده شده است در گذر از WVD به SPWVD ترم‌های تداخلی کاهش می‌یابند اما رزولوشن را در فرکانس و زمان از دست می‌دهیم.

بنابراین درPWVD کنترل فقط روی هموارسازی فرکانس است ولی در SPWVD هموارسازی روی محور فرکانس و زمان با هم می‌باشد.

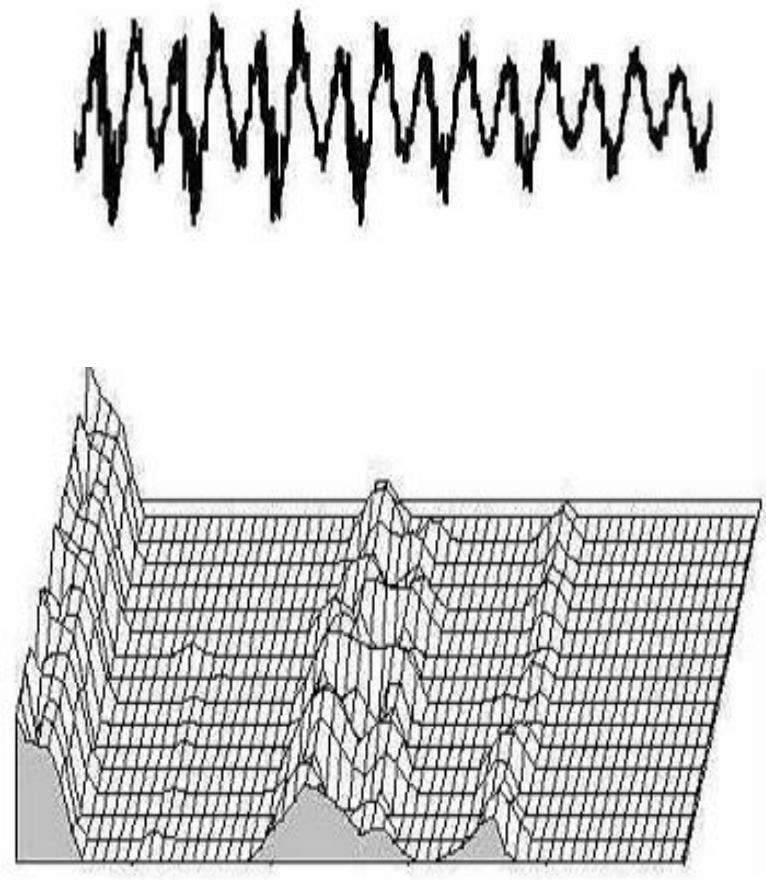
یک ویژگی جالب SPWVD این است که یک گذر پیوسته از اسپکتروگرام به WVD تحت شرایط گوسین بودن تابع هموار سازی g وh ایجاد می کند.

توزیع ویگنر ویل (WVD) بهترین رزولوشن را در زمان و فرکانس دارد و بدترین تداخل، اما اسپکتروگرام بدترین رزولوشن را در زمان و فرکانس می دهد اما کمترین تداخل را دارد. اما SPWVD به ما اجازه می دهد تا مصالحه‌ای بین رزولوشن و تداخل را داشته باشتم.

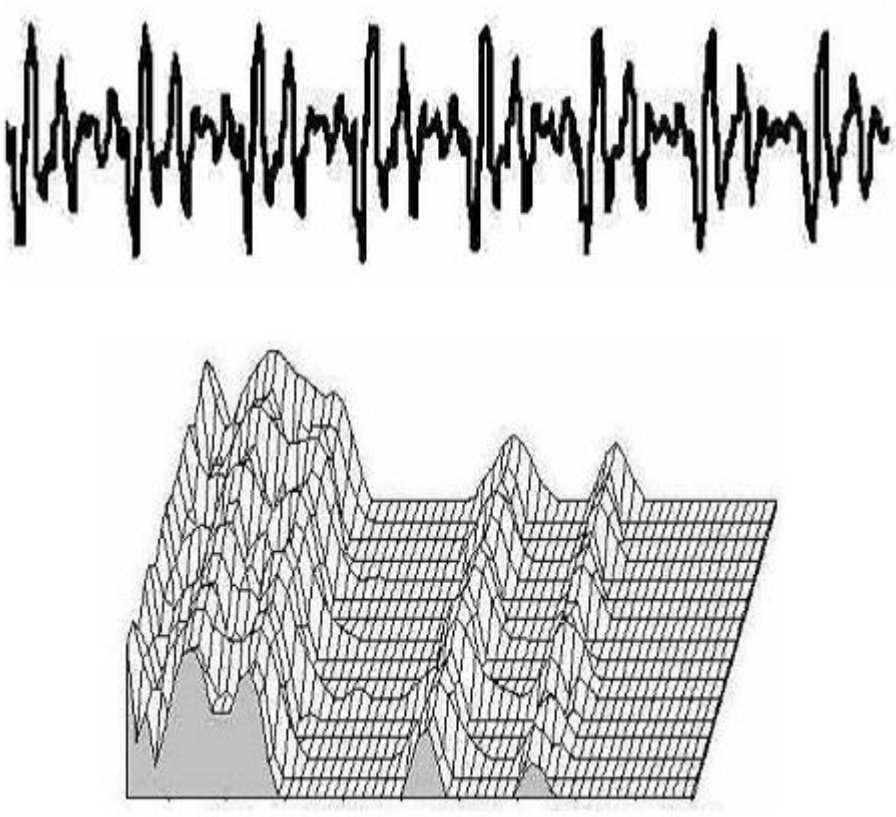
۳-۶- استفاده از WVD در تحلیل سیگنال گفتار

کارهای انجام گرفته با استفاده از تابع توزیع ویگنر- ویل برای استخراج ویژگی از سیگنال گفتار در سالهای اخیر بر روی استخراج فرکانس‌های فرمنت و فرکانس پیچ بوده است که در ادامه به چند روش اشاره می‌شود.

در مرجع[۲۳] برای استخراج فرکانس‌های اصلی از هر واج از تابع توزیع ویگنر- ویل هموار شده در زمان و فرکانس استفاده شده است. در این روش در بازه‌های زمانی مشخص شده فرکانس‌های فرمنت استخراج شده‌اند ولی به دلیل هموار سازی در زمان و فرکانس دارای رزولوشن نمی‌باشند. در شکل-های (۳-۱۶) و (۳-۱۷) به ترتیب سیگنال زمانی واج و نمودار SPWVD مربوط به واج /i/ و /a/ نشان داده شده است.



شکل (۳-۱۶): سیگنال زمانی واج /i/ و نمودار SPWVD



شکل (۳-۱۷): سیگنال زمانی واج /a/ و نمودار SPWVD

همچنین در مرجع [۲۴] مقایسه‌ای بین تابع توزیع هموار شده در زمان و فرکانس (SPWVD) و اسپکتوگرام انجام شده است. در این مقاله بیان شده است که می‌توان از بعد زمان تابع SPWVD فرکانس‌های پیچ و از بعد فرکانس تابع SPWVD فرکانس‌های فرمنت را استخراج کرد.

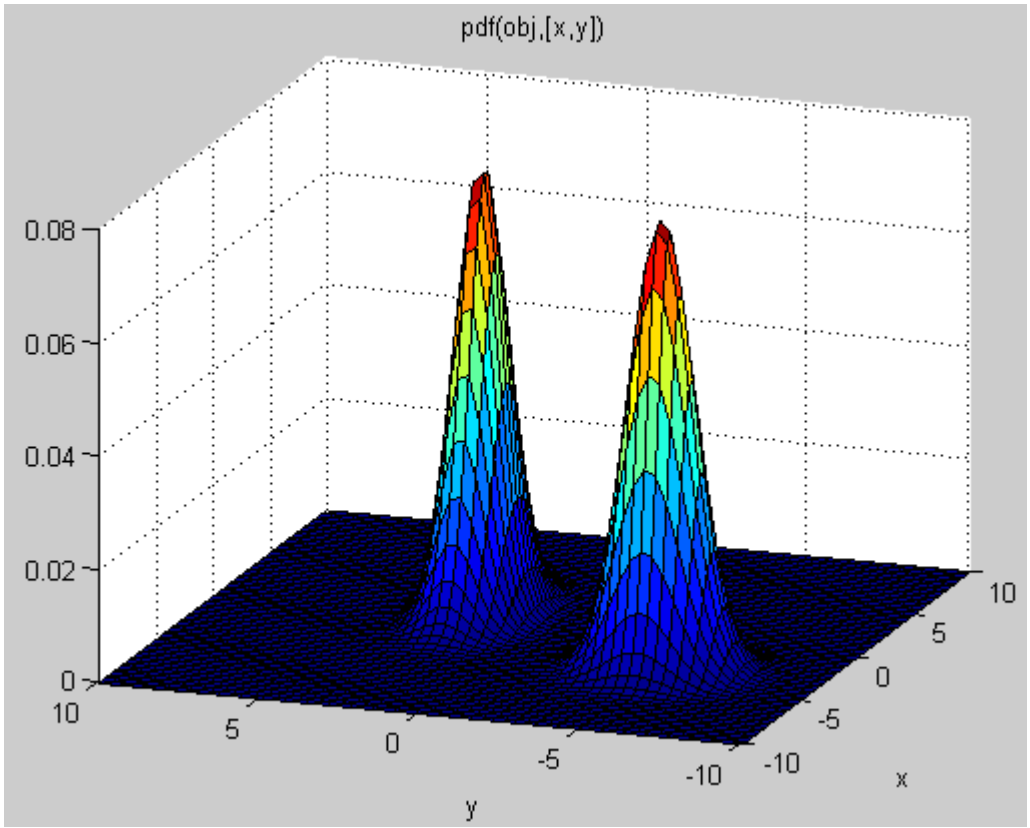
در مرجع[۲۵] تحلیل گذرای سیگنال گفتار توسط تابع توزیع ویگنر – ویل و نمونه هموار شده آن انجام شده است.

فـــــــــصل چهارم

مـــــدل مخـــلوط گوســـــی

۴-۱- مقدمه

مخلوط گوسی برای مدل نمودن بعضی از فرآیندهای اتفاقی بکار می‌رود که برای نشان دادن یک زیر مجموعه از یک مجموعه کلی مورد استفاده قرار می‌گیرد. مدل مخلوط گوسی[1] با ترکیب اجزای چگالی نرمال چندگانه تشکیل می‌شود که از آن برای تخمین تابع چگالی احتمال[2](PDF) دنباله‌ای از بردارهای ویژگی استفاده می‌شود[۵]. در شکل زیر یک مدل مخلوط گوسی دو بعدی با دو جزء گوسی نشان داده شده است.



شکل(۴-۱): نمایشی از مدل مخلوط گوسی با دو بعد و دو گوسی[۲۷]

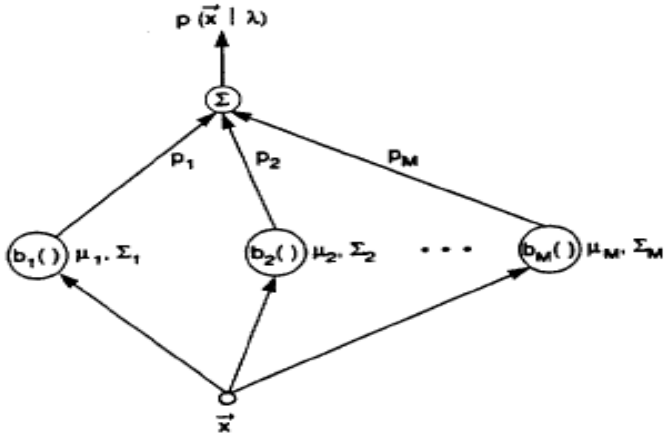
[1] Gaussian mixture model (GMM)
[2] Probability Density Function(PDF)

۴-۲- انگیزه استفاده از مدل مخلوط گوسی در سیستم‌های تشخیص گوینده

در استفاده از چگالی مخلوط گوسی برای شناسایی گوینده دو انگیزه وجود دارد. انگیزه اول در درک شهودی از مدل است که هر جزء گوسی از یک چگالی چند متغیره مانند GMM، مجموعه‌ای از کلاس‌های صوتی اصلی را مدل می‌کند که این مجموعه از کلاس‌های صوتی می‌توانند واولها[1]، نزلها[2] و صداهای فرسایشی[3] باشد. این کلاس‌های صوتی ترکیب‌بندی وکال‌ترکت وابسته به هر گوینده را انعکاس می‌دهد که می‌تواند برای شناسایی تشخیص هویت گوینده مفید باشد. انگیزه دوم این است که مدل مخلوط گوسی تخمین هموواری از نمونه‌های زمان طولانی گفته شده توسط گوینده را فراهم می‌کند[۵].

۴-۳- مدل مخلوط گوسی و فرمول بندی آن

مدل مخلوط گوسی از M جزء چگالی وزن دهی شده تشکیل شده است که در نهایت این اجزاء با هم جمع می‌شوند و مدل نهایی را تشکیل می‌دهند. شمای کلی این مدل در شکل(۴-۲) نشان داده شده است. در ادامه فرمول‌بندی مدل مخلوط گوسی برای استفاده در سیستم‌های تشخیص گوینده آورده شده است[۵].



شکل(۴-۲): ساختار مدل مخلوط گوسی[۵].

[1] Vowel
[2] Nasal
[3] Fricative sound

ساختار مدل مخلوط گوسی در شکل(۴-۲) با رابطه (۴-۱) بیان می‌شود.

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (۱-۴)$$

که $\vec{x}$ یک بردار تصادفی D بعدی، $b_i(\vec{x})$ برای i=1…m اجزای چگالی گوسی و p_i برای i=1…m وزن‌های مخلوط گوسی می‌باشند[۵]. در فرمول (۴-۲) هر جزء چگالی تابع گوسی D متغیره نشان داده شده است.

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \mid \sum_i \mid^{1/2}} \exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)' \sum_i^{-1} (\vec{x} - \vec{\mu}_i)\} \qquad (۲-۴)$$

که $\vec{\mu}_i$ بردار میانگین و $\sum_i$ ماتریس کوواریانس تابع گوسی می‌باشند. در فرمول (۴-۱) وزن-های مخلوط گوسی (p_i) باید شرط بیان شده در رابطه (۴-۳) را برآورده نماید تا اینکه

$$\sum_{i=1}^{M} p_i = 1 \qquad (۳-۴)$$

احتمال گوسی بدست آمدهٔ مخلوط $p(\vec{X} \mid \lambda)$، یک تابع چگالیِ احتمال صحیح باشد[۵].

یک چگالی مخلوط گوسی به طور کامل با بردار میانگین، ماتریس کوواریانس و وزن‌های مخلوط گوسی برای تمام اجزای گوسی مشخص می‌شوند. این پارامترها به طور کامل توسط رابطه (۴-۴) برای یک گوینده ارایه می‌شوند[۵].

$$\lambda = \{p_i, \vec{\mu}_i, \sum_i\} \quad i = 1, \ldots, M \qquad (۴-۴)$$

در یک سیستم بازشناسی گوینده، هر گوینده توسط یک مدل مخلوط گوسی (GMM) که شامل بردار میانگین، ماتریس کوواریانس و وزن‌های مخلوط گوسی برای هر جزء گوسی می‌باشند مشخص می‌شود که λ یک مدل مرجع برای آن شخص می‌باشد.

مدل مخلوط گوسی دارای چندین ماتریس کوواریانس می‌باشد. در یک نوع از GMM برای هر جزء گوسی یک ماتریس کوواریانس موجود می‌باشد که به این ماتریس، ماتریس کوواریانس گره‌ای[1] گفته می‌شود. در نوعی دیگر از GMM یک ماتریس کوواریانس برای تمام اجزای گوسی موجود می‌باشد که به این ماتریس، ماتریس کوواریانس بزرگ[2] گفته می‌شود. و در نوعی دیگر از GMM برای تمام گوینده‌ها تنها یک ماتریس کوواریانس وجود دارد که به این ماتریس، ماتریس کوواریانس عمومی[3] گفته می‌شود که هر یک از این ماتریس‌های کوواریانس می‌توانند کامل یا قطری باشند[۵].

۴-۴- الگوریتم EM[4]

هدف از مدل آموزش گوینده برای یک مجموعه آموزش، تخمین پارامترهای GMM می‌باشد. برای تخمین پارامترهای GMM تکنیک‌های متفاوتی وجود دارد که یکی از بهترین روش‌ها تخمین بیشترین پیشامد[5] (ML) می‌باشد.

هدف از تخمین ML پیدا کردن پارامترهای مدل مخلوط گوسی می‌باشد که برای داده‌های آموزشی، پیشامد GMM را ماکزیمم‌کند. برای دنباله‌ای از T بردار آموزش پیشامد GMM مانند رابطه(۴-۵) می‌باشد.

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda) \qquad (۵-۴)$$

[1] Nodal
[2] Grand
[3] Global
[4] Expectation Maximization(EM)
[5] Maximum Likelihood(ML)

متاسفانه این عبارت یک تابع غیر خطی از پارامترهای λ است و ماکزیمم کردن مستقیم آن امکان پذیر نمی‌باشد. با این حال با استفاده از الگوریتم EM به صورت پی در پی پارامترهای ML تخمین زده می‌شوند.

به طور کلی الگوریتم EM در مسائلی به کار می‌رود که می‌خواهیم مجموعه‌ای از پارامترهای λ را تخمین بزنیم که مبتنی بر یک توزیع احتمالی‌اند. الگوریتم EM یک روش تکراری[1] است که ایدهٔ اصلی آن این است که با شروع از یک مدل اولیهٔ λ، یک مدل جدید $\hat{\lambda}$ تخمین زده شود بگونه‌ای که عبارت $p(X \mid \hat{\lambda}) \geq p(X \mid \lambda)$ برقرار باشد. مدل جدید، مدل اولیه برای تکرار بعدی می‌باشد واین فرآیند تکرار می‌شود تا زمانی که به آستانه‌ایی که مد نظر است همگرا شود[۵].

الگوریتم EM شامل ۳ مرحله جهت تخمین پارامترهای مدل مخلوط گوسی می‌باشد[۱۷]:

- مدل اولیه: شامل مقداردهی اولیه برای پارامترهای $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \;\; i = 1, ..., M$ می‌باشد، که می‌توان هم بصورت تصادفی و هم از روش k-means، جهت مقداردهی اولیه استفاده نمود.

- مرحلهٔ E: شامل محاسبهٔ احتمال پسین[2] برای جزءِ iاُم می‌باشد که مبنای روش بدست آوردن آن قانون بیزین[3] است، و طبق معادلهٔ (۴-۶) قابل محاسبه می‌باشد.

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \qquad (۶-۴)$$

[1] Iteratively
[2] posteriori probability
[3] Bayesian Rule

- مرحلهٔ M: شامل به روزرسانی[1] مقادیر پارامترهای وزن، میانگین و کوواریانس می‌باشد که در رابطه (۴-۷) بیان شده‌اند.

$$\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)$$

$$\vec{\bar{\mu}}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)\vec{x}_i}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} \qquad (۴-۷)$$

$$\bar{\sigma}_i^{\,2} = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)x_t^{\,2}}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} - \bar{\mu}_i^{\,2}$$

مراحل E و M تا زمانی تکرار می‌شوند که تابع لگاریتمِ درستنمایی[2]، مطابق رابطه (۴-۸) ماکزیمم گردد.

$$\log\, p(X \mid \lambda) = \sum_{t=1}^{T} \log\left(\sum_{i=1}^{M} p_i b_i(\vec{x})\right) \qquad (۴-۸)$$

$$\hat{\lambda} = \arg\max \log p(X \mid \lambda)$$

لازم به ذکر است که مدل اولیه در آموزش مدل‌های مخلوط گوسی بسیار مهم است. در حالت کلی و مخصوصاً برای آموزش تعداد زیادی از مخلوط‌های گوسی، مقداردهی اولیهٔ صحیح می‌تواند نرخ همگرایی را بالاتر برده و همچنین باعث شود تا الگوریتم EM در ماکزیمم‌های محلی به تله نیفتاده و به ماکزیمم عمومی نزدیکتر گردد. یکی از روشهای مناسب بدین منظور، استفاده از الگوریتم خوشه بندی K-means می باشد[۱۷].

[1] Update
[2] Log- Likelihood Function

دو فاکتور بحرانی در استفاده از مدل مخلوط گوسی، انتخاب مرتبه M و مقدار دهی اولیه به پارامترهای مدل مخلوط گوسی در الگوریتم EM است که باید از طریق آزمایش هر یک از این پارامترها را تعیین گردد[۵].

۴-۵- کاربرد GMM در تشخیص گوینده

با توجه به رابطه (۴-۴) در یک سیستم شناسایی گوینده متشکل از S گوینده S={1,2,…,S} که با GMM هایی از λ_1 و λ_2 و ... و λ_s ارایه می‌شوند، هدف پیدا کردن مدل گوینده‌ای است که برای دنباله‌ای از مشاهدات بیشترین احتمال پسین را داشته باشد[۵].

در معادله (۴-۹) این موضوع نشان داده شده است که این معادله همان قانون بیز می‌باشد.

$$\hat{S} = \arg \max_{1 \le k \le S} \Pr(\lambda_k \mid X) = \arg \max_{1 \le k \le S} \frac{p(X \mid \lambda_k) \Pr(\lambda_k)}{p(X)} \qquad (۹-۴)$$

با فرض احتمال یکسان برای گوینده‌ها و اینکه P(x) برای تمام گوینده‌ها یکسان می‌باشد قانون کلاسه‌بندی به معادله(۴-۱۰) ساده می‌شود.

$$\hat{S} = \arg \max_{1 \le k \le S} \mathrm{P}(X \mid \lambda_k) \qquad (۱۰-۴)$$

با استفاده از لگاریتم و استقلال بین مشاهدات، از رابطه (۴-۱۱) در سیستم بازشناسی گوینده استفاده می‌شود.

$$\hat{S} = \arg \max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_k) \qquad (۱۱-۴)$$

که $p(\vec{x}_t \mid \lambda_k)$ از رابطه (۴-۱) محاسبه می‌شود. این معادله بیانگر این موضوع است که اگر این معادله برای گوینده مورد نظر ماکزیمم شود آن فرد درست تشخیص داده شده است.

فـــــصل پنـــجم

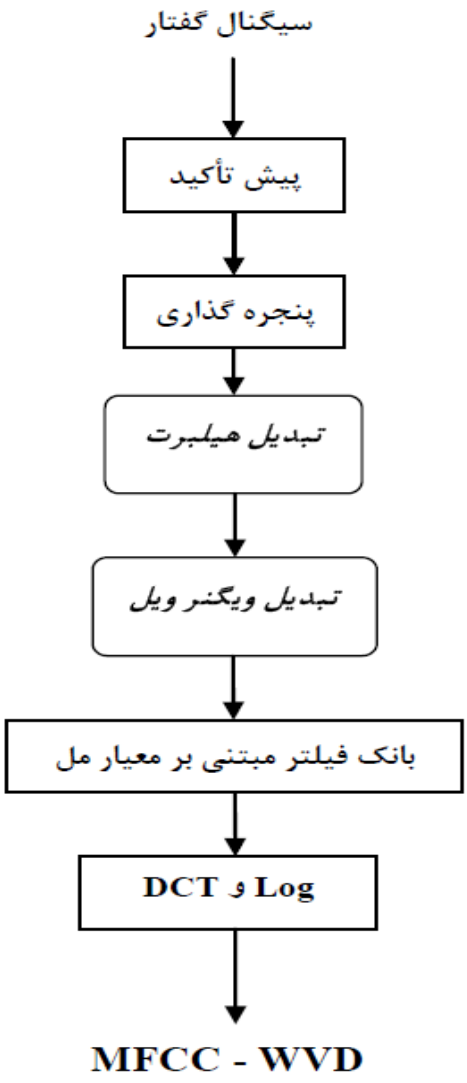
روش پیشنهادی و پیـــاده سازی الگوریتـــم

۵-۱- مقدمه

با توجه به ضعف تبدیل فوریه در آنالیز سیگنال‌های غیر ایستان، از تبدیلات زمان- فرکانس برای آنالیز این سیگنال‌ها استفاده می‌شود. از آنجا که تابع توزیع ویگنر- ویل یکی از تبدیلات مهم زمان- فرکانس در آنالیز سیگنال‌های غیر ایستان می‌باشد و همچنین دارای خواص مفیدی از قبیل رزولوشن فرکانسی و خواص حاشیه‌ای و ... می‌باشد، در این پایان‌نامه از آن برای استخراج ویژگی از سیگنال گفتار جهت استفاده در سیستم‌های تشخیص گوینده استفاده شده است. در روش پیشنهادی از ترکیب تابع توزیع ویگنر- ویل و تبدیل هیلبرت و ضرایب MFCC استفاده شده است که نتایج این روش با نتایج ضرایب PLP و MFCC مقایسه شده است. برای ارزشیابی کارایی از دو روش متفاوت استفاده شده است که این نتایج گویای جواب بهتر روش پیشنهادی در سیگنال به نویزهای پایین نسبت به دو روش متداول PLP و MFCC می‌باشد.

۵-۲- روش پیشنهادی

بلوک دیاگرام روش پیشنهادی در شکل (۵-۱) نشان داده شده است. همانطور که در این شکل مشخص است، پس از پیش پردازش و پنجره گذاری، از سیگنال گفتار تبدیل هیلبرت و تبدیل ویگنر-ویل گرفته می‌شود و سپس از فیلتر بانک مل عبور داده می‌شود و در نهایت از آن لگاریتم و تبدیل فوریه کسینوسی گرفته می‌شود و ضرایب بدست می‌آید. مراحل این روش به تفکیک در ادامه توضیح داده شده است.



شکل(۵-۱): بلوک دیاگرام روش پیشنهادی

۵-۲-۱- پیش پردازش

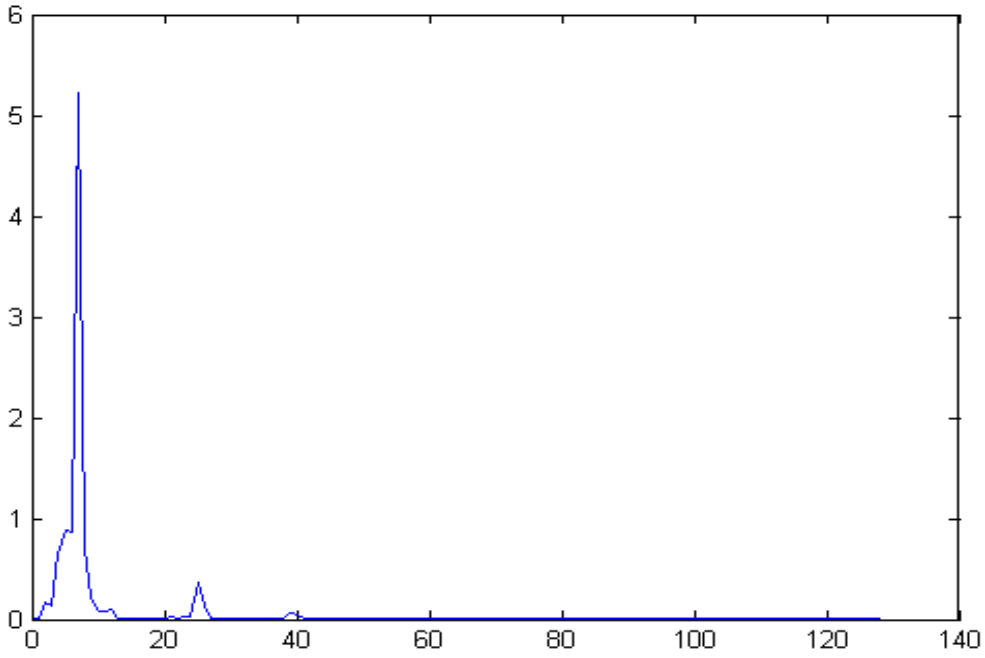
این مرحله شامل جداسازی حروف بی‌صدا از حروف صدادار و اطلاعات غیر ضروری و اضافی نظیر نویزهای موجود می‌باشد.

۵-۲-۲- پنجره گذاری

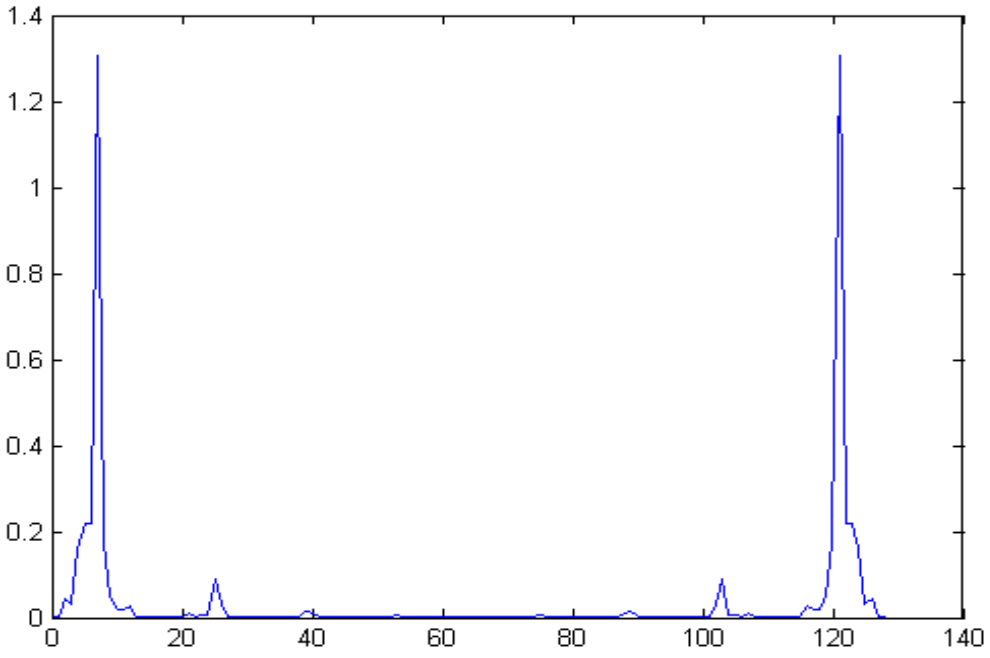
پس از پیش پردازش عمل پنجره گذاری روی سیگنال گفتار اعمال می‌شود که از پنجره همینگ برای این منظور استفاده شده است. طول هر پنجره با توجه به ناایستان بودن سیگنال گفتار بین ۲۰ تا ۳۰ میلی ثانیه می‌باشد که در این پایان‌نامه طول هر پنجره (۲۳,۲۲) میلی ثانیه انتخاب شده است و هر فریم دارای ۲۵۶ نمونه می‌باشد.

۵-۲-۳- تبدیل هیلبرت

پس از عمل پنجره‌گذاری روی هر فریم از تبدیل هیلبرت جهت تحلیلی نمودن سیگنال خروجی استفاده شده است تا تبدیل ویگنر- ویل آن به سادگی انجام شود و از تداخل ترم‌های اضافی جلوگیری گردد. دلیل استفاده از تبدیل هیلبرت در شکل‌های (۵-۲) و (۵-۳) نشان داده شده است. در شکل (۵-۲) که مشخصه فرکانسی یک پنجره از سیگنال گفتار می‌باشد از تبدیل هیلبرت استفاده شده است در صورتی که در شکل (۵-۳) از تبدیل هیلبرت استفاده نشده است. همانطور که مشخص است شکل (۵-۳) ترم‌های اصلی بیشتری نسبت به شکل(۵-۲) دارد که این امر می‌تواند ترم‌های تداخلی بیشتری را تولید کند و تحلیل نمودار زمان فرکانس مشکل شود.



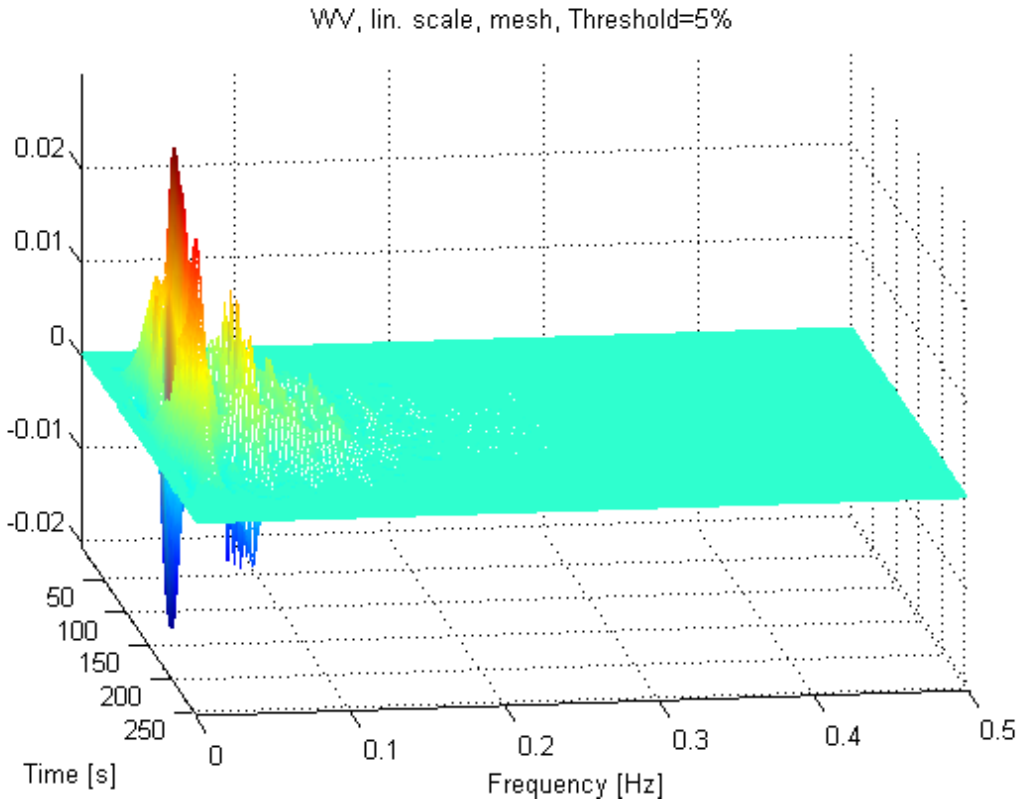
شکل (۵-۲): مشخصه فرکانسی تابع توزیع ویگنر- ویل با استفاده از تبدیل هیلبرت



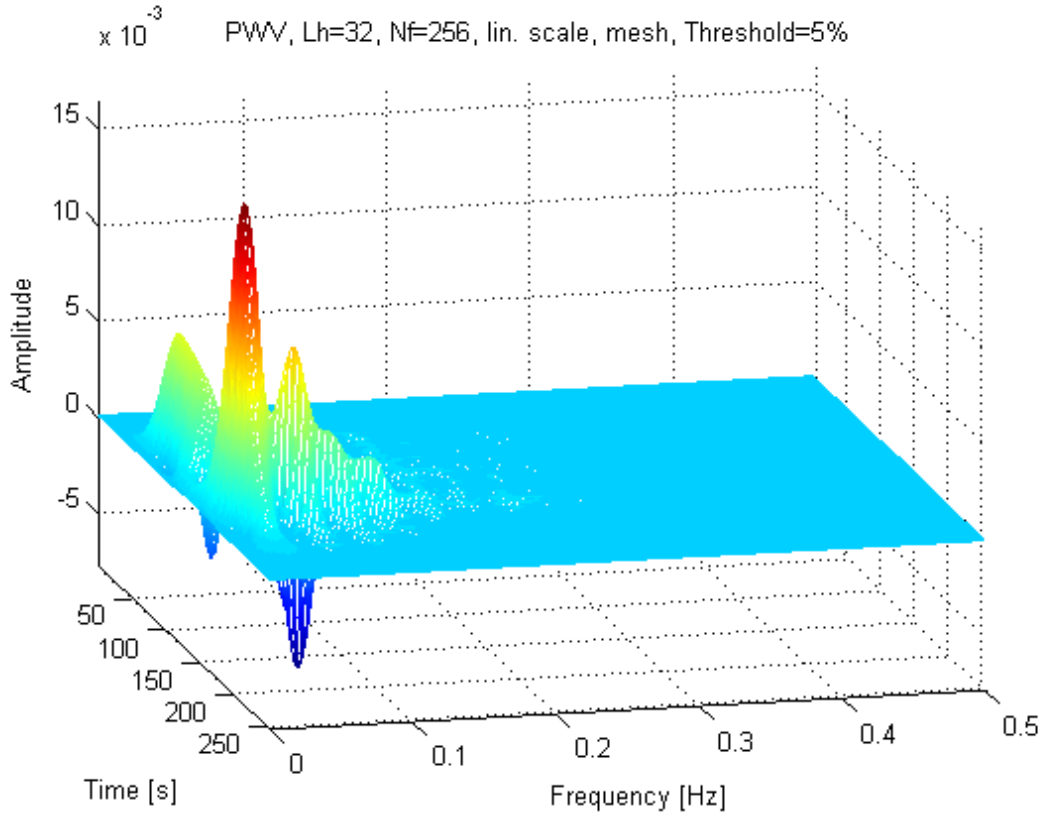
شکل (۵-۳): مشخصه فرکانسی تابع توزیع ویگنر- ویل بدون استفاده از تبدیل هیلبرت

۵-۲-۴- تبدیل ویگنر- ویل

تبدیل ویگنر ویل روی هر فریم n نمونه‌ای، ماتریسی با ابعاد $n \times n$ می باشد که یک بعد آن زمان و بعد دیگر آن فرکانس می‌باشد. یک نمونه از تبدیل ویگنر- ویل روی یک فریم ۲۵۶ نمونه‌ای در شکل زیر نشان داده شده است. با توجه به شکل (۵-۴) کاملاً مشخص است که هر مؤلفه فرکانسی در چه زمانی رخ داده است. اما در پنجره‌ای به طول (۲۳,۲۲) میلی ثانیه اطلاعات زمانی مفیدی قابل استخراج نیست و اگر طول هر فریم را آنقدر زیاد کنیم که اطلاعات زمانی مفیدی قابل استخراج باشد آنگاه فرض ایستان بودن سیگنال گفتار را از دست داده‌ایم. به همین دلیل در این پایان نامه فقط از اطلاعات فرکانسی توزیع ویگنر- ویل استفاده شده است.



شکل (۵-۴): نمایشی سه بعدی از تبدیل ویگنر- ویل یک پنجره از سیگنال گفتار

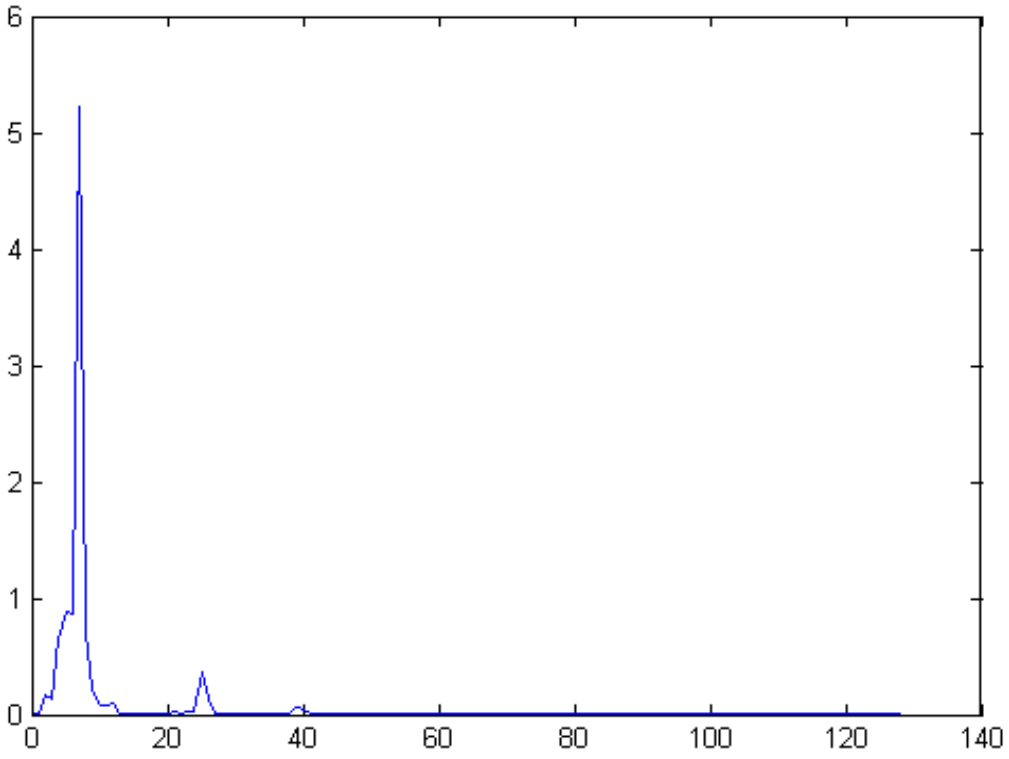


شکل (۵-۵): نمایشی سه بعدی از تبدیل ویگنر- ویل هموار شده یک پنجره از سیگنال گفتار

شکل (۵-۴) و شکل (۵-۵) به ترتیب تابع توزیع ویگنر- ویل و نمونه هموار شده آن از یک فریم مشخص از سیگنال گفتار می‌باشند. در این پایان نامه برای استخراج ویژگی از هر فریم $n \times n$ توزیع ویگنر- ویل از خاصیت حاشیه‌ای توزیع ویگنر-ویل (WVD) و توزیع ویگنر- ویل هموار شده در فرکانس (PWVD) استفاده کرده‌ایم که دو روش پیشنهادی به تفکیک در زیر توضیح داده شده است.

- ***روش اول***

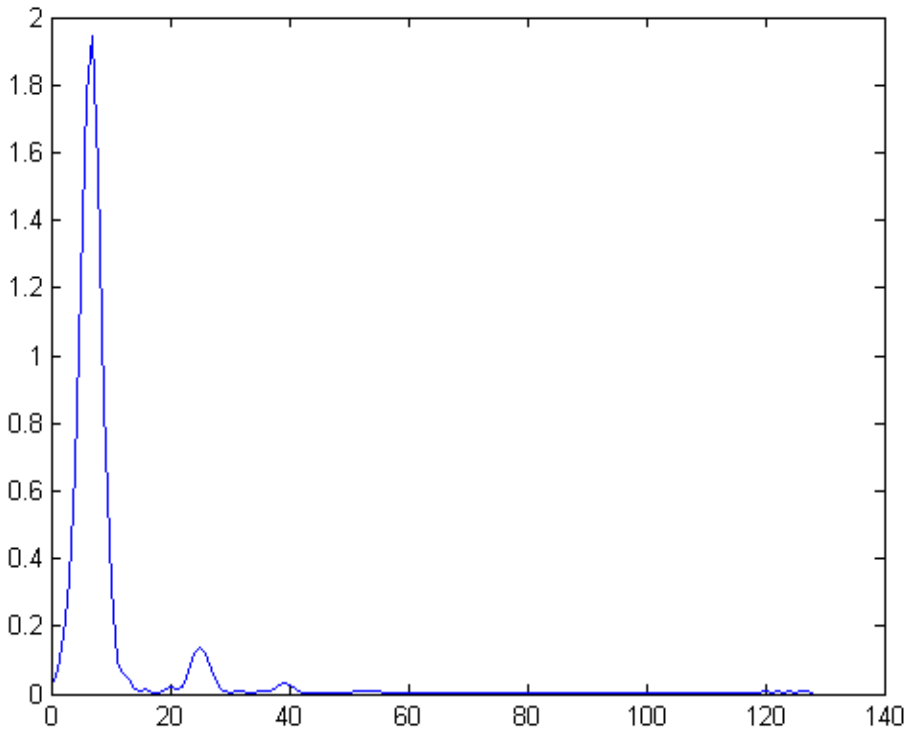
در روش اول از خاصیت حاشیه‌ای روی هر فریم از تابع توزیع ویگنر- ویل (WVD) نشان داده شده در شکل (۵-۴) استفاده کرده‌ایم که نتیجه حاصل در شکل (۵-۶) نشان داده شده است. مشخصه فرکانسی استخراج شده در این روش را با WVD –S نشان می‌دهیم. همانطور که در شکل (۵-۶) نشان داده شده است فرکانس‌های فرمنت از رزولوشن بسیار خوبی برخوردارند و دامنه فرکانس‌های فرمنت هم از اندازه خوبی برخوردارند. در شکل (۵-۶) سه فرمنت موجود می‌باشد که دامنه اولین فرمنت ۵,۵ و دامنه سومین فرمنت نزدیک به ۰,۱ می‌باشد.



شکل (۵-۶): مشخصه فرکانسی حاشیه‌ای روی یک فریم WVD

- ***روش دوم***

در روش دوم از خاصیت حاشیه‌ای روی هر فریم تابع توزیع ویگنر- ویل هموار شده در فرکانس (PWVD) در شکل (۵-۵) استفاده کرده‌ایم که در شکل (۵-۷) نشان داده شده است. مشخصه فرکانسی استخراج شده در این روش را با PWVD – S نشان می‌دهیم. همانطور که در شکل (۵-۷) نشان داده شده است فرکانس‌های فرمنت به دلیل هموار سازی از رزولوشن خوبی برخوردار نیستند. در شکل (۵-۷) سه فرمنت موجود می‌باشد که دامنه اولین فرمنت ۲ و دامنه سومین فرمنت نزدیک به ۰,۰۲ می باشد. در شکل (۵-۷) هم دامنه و هم رزولوشن نسبت به شکل (۵-۶) کم شده است که در هر دوی این اشکال، از خاصیت حاشیه‌ای روی هر فریم تابع توزیع ویگنر- ویل و نمونه هموار شده آن استفاده شده است.



شکل (۵-۷): مشخصه فرکانسی حاشیه‌ای روی یک فریم PWVD

۵-۲-۵- فیلتر بانک مل

در این مرحله هر یک از مشخصه‌های فرکانسی بدست آمده در مرحله قبل از این فیلتر عبور داده می‌شود. همانطور که گفته شد مقیاس مل بر مبنای ساختار شنیداری انسان می‌باشد. در این پایان نامه از ۲۷ فیلتر در فیلتر بانک مل استفاده شده است. در فرکانس‌های بالا، پهنای باند این فیلترها زیادتر از فرکانس‌های پایین است و این امر بدان معناست که حساسیت گوش انسان نسبت به تغییر فرکانس‌های پایین بیشتر است. پس از قرار دادن فیلتر روی مشخصه‌های فرکانسی در شکل‌های (۵-۶) و (۵-۷)، خروجی این فیلترها محاسبه می‌شود.

۵-۲-۶- لگاریتم گیری

در این مرحله از خروجی های فیلتر بانک مل لگاریتم گرفته می‌شود که این لگاریتم‌گیری به نوعی عمل فشرده سازی را انجام می‌دهد.

۵-۲-۷- تبدیل فوریه کسینوسی

از آنجائیکه فیلترهای بانک مل دارای همپوشانی می‌باشند بنابراین خروجی‌های این فیلتر بانک دارای همبستگی هستند که برای کاهش این همبستگی از تبدیل فوریه کسینوسی استفاده شده است. البته این تبدیل تا حد زیادی باعث نرم شدن طیف فرکانسی می‌گردد که به نوعی ضرایب را کاهش می‌دهد و باعث فشرده‌سازی می‌گردد. پس از بدست آوردن ضرایب کپستروم اولین ضریب را که مقدار متوسط انرژی است حذف می‌کنیم زیرا حاوی اطلاعات مفیدی نمی‌باشد و در نهایت دنباله‌ای ۱۳ تایی از ضرایب را به عنوان دنباله ضرایب ویژگی انتخاب می‌کنیم.

در مرحله آخر بردار میانگین تمام بردارها محاسبه می‌شود و از تک تک آنها کسر می‌گردد که به این روش نرمال سازی میانگین کپسترال (CMN) می‌گویند که با این عمل به ویژگی‌های مقاوم-تری دست یافته می‌شود.

۵-۳- دیتا بیس

دیتا بیس مورد استفاده در پیاده سازی روش‌های پیشنهادی، دیتا بیس انگلیسی TIMIT می باشد. این دیتا بیس یکی از معروفترین و پر کاربردترین دیتا بیس زبان انگلیسی مورد استفاده در زمینه پردازش گفتار می باشد. این دیتا بیس شامل ۲۳۴۲ جمله متنی و ۶۳۰ گوینده و ۶۳۰۰ جمله گفته شده می‌باشد. مجموعه تمام گوینده‌ها از ۸ ایالت آمریکا می‌باشند که به ۸ زیر گروه dr[1-8] تقسیم‌بندی شده‌اند و هر گوینده ۱۰ جمله از تمام جملات را بیان می‌کند. در جدول (۵-۱) توزیع گوینده‌ها برای هر ایالت مشخص شده است و در جدول (۵-۲) ایالت گوینده‌ها نشان داده شده است.

جدول (۵-۱): توزیع گوینده ها برای هر ایالت

Region(dr)	#Male	#Female	Total
1	31 (63%)	18 (27%)	49 (8%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
8	438 (70%)	192 (30%)	630 (100%)

جدول (۵-۲): هشت ایالت گوینده ها

dr1: New England
dr2: Northern
dr3: North Midland
dr4: South Midland
dr5: Southern
dr6: New York City
dr7: Western
dr8: Army Brat (moved around)

جملات به سه دسته Dialect (SA) و Compact (SX) و Diverse (SI) تقسیم بندی می-شوند که جملات مجموعه (SA) مربوط به لهجه می‌باشد و جملات مجموعه (SX) از لحاظ واج متراکم می‌باشد و جملات مجموعه (SI) از لحاظ واج متنوع و تقریبا دارای همه واج‌ها می‌باشد.

کل جملات ۲۳۴۲ جمله می‌باشد که از این مجموعه ۲ جمله آن از نوع (SA) و ۴۵۰ جمله از نوع (SX) و ۱۸۹۰ جمله از نوع (SI) می‌باشد که از ۱۰ جمله‌ای که هر فرد بیان می‌کند ۲ جمله از (SA) و ۵ جمله از (SX) و ۳ جمله از (SI) می‌باشد که به طور خلاصه در جدول (۵-۳) نشان داده شده است.

جدول (۵-۳): تقسیم بندی جملات بر حسب SX و SA و SI

Sentence Type	#Sentences	Total	#Sentences/Speaker
Dialect (SA)	2	1260	2
Compact (SX)	450	3150	5
Diverse (SI)	1890	1890	3
Total	2342	6300	10

مجموعه متن و گویندگان این دیتا بیس به دو زیر مجموعه تست و آموزش تقسیم بندی می-شوند که حدود ۲۰ تا ۳۰ درصد برای تست و حدود ۷۰ تا ۸۰ درصد برای آموزش استفاده می‌شود. هیچکدام از گویندگان در هر دو مجموعه تست و آموزش عضویت ندارند و برای هر مجموعه منحصر به فرد می‌باشند. در مجموعه تست این دیتا بیس از هر لهجه تعداد مختلف زن و مرد انتخاب شده است که توزیع آن را می‌توان در جدول (۵-۴) مشاهده کرد.

جدول (۵-۴): تعداد زن و مرد در هر لهجه

Dialect	#Male	#Female	Total
1	7	4	11
2	18	8	26
3	23	3	26
4	16	16	32
5	17	11	28
6	8	3	11
7	15	8	23
8	8	3	11
Total	112	56	168

هر جمله بیان شده دارای سه نوع فایل مختلف با پسوند های ".txt" و ".wrd" و ".phn" می‌باشد که در فایل اول کل جمله انگلیسی همراه با تعداد نمونه‌ها از شروع تا پایان جمله وجود دارد. فایل دوم شامل کلمات جمله است که در آن شماره نمونه از آغاز کلمه تا پایان آن، بعد از هر کلمه ذکر شده است و در فایل سوم هر کلمه به فونم‌های خاص خود شکسته شده است و شماره نمونه شروع و پایان هر فونم، به طور جداگانه بعد از هر فونم در آن ذکر گردیده است.

در این دیتا بیس برای شناسایی مشخصات گوینده، برچسپ هر گوینده از سه قسمت تشکیل شده است که یکی جنس گوینده که با M یا F در ابتدا مشخص می‌شود. بخش دوم آن دارای سه حرف است که در برگیرنده نام گوینده می‌باشد و قسمت آخر دارای مشخصات دیگر گوینده است. اما هر جمله در این دیتا بیس شامل سه نوع مختلف است که با sa و si و sx بیان می‌شوند و عدد بعد از این دو حرف بیانگر شماره جمله است و هر جمله دارای یک فایل صوتی است که نیاز به خواندن این فایل صوتی می‌باشد.

هر فایل صوتی به طور معمول در نرم افزار Matlab با دستور Waveread قابل خواندن است اما این فایل صوتی دارای فرمت خاصی است و نمی‌توان برای آن از این دستور استفاده کرد. بنابراین از دستور خاصی به نام readsph استفاده می‌شود که دارای پارامترهای خاصی است تا بتوان کلمه یا فونم مورد نظر را خواند. این تابع شماره شروع نمونه برای فونم یا کلمه را به همراه طول فونم یا کلمه مورد نظر را به عنوان ورودی دارد.

در این پایان نامه از ۳۲ گوینده مربوط به مجموعه تست dr[4] استفاده شده است که تعداد آقایان و خانم‌ها هر کدام شانزده می‌باشد. در ابتدا هر جمله با پسوند wav به نام خود آن جمله آدرس دهی می‌شود و سپس توسط دستور readsph فراخوانی می‌شود و در یک آرایه از نوع cell به نام خود آن جمله و با فرمت (mat.) ذخیره می‌شود.

۵-۴- مشخصات سیستم

میانگین داده‌های آموزش برای ۳۲ گوینده ۳۶۰۰۰۰ نمونه و میانگین داده تست ۸۰۰۰۰ هزار نمونه می‌باشد که نسبت داده‌های آموزش به داده‌های تست ۴,۵۳ می‌باشد. به ازای تعداد نمونه‌های مشخص برای داده‌های تست هر چه تعداد نمونه‌های داده‌های آموزش بیشتر باشد درصد تشخیص گوینده بیشتر خواهد بود[۵].

بنابراین میانگین داده‌های آموزش برای ۳۲ گوینده ۳۲,۶۵ ثانیه و میانگین داده تست ۷,۲۶ ثانیه می‌باشد. یعنی از کل جملات بیان شده توسط یک فرد حدوداً ۸۲ درصد برای آموزش و ۱۸درصد برای تست استفاده شده است.

طول هر فریم ۲۵۶ نمونه و مقدار پیش‌روندگی ۱۲۸ نمونه در نظر گرفته شده است. بنابراین طول هر فریم ۲۳,۲۲ میلی ثانیه و مقدار پیش روندگی ۱۱,۶۱ میلی ثانیه می‌باشد. در فیلتر بانک مل از ۲۷ فیلتر استفاده شده است که تعداد این فیلترها از ۲۰ تا ۴۰ متغییر می‌باشد و با احتساب حذف اولین ضریب، دنباله ضرایب ۱۳ تایی می‌باشد[۱۳].

در این پایان نامه از کلاسیفایر مدل مخلوط گوسی (GMM) استفاده شده است. در کلاسیفایر مدل مخلوط گوسی، هر گوینده با یک ماتریس کوواریانس، یک بردار میانگین و یک بردار ضرایب مشخص می‌شود که مدل هر گوینده شامل ۱۶ مخلوط گوسی و ۱۳ بعد می‌باشد. علاوه بر این از ماتریس کوواریانس گره‌ای و قطری برای مدل کردن گوینده‌ها استفاده شده است. این انتخاب بر اساس نتایج آزمایشاتی است که در [۵] آمده است و نشان از کارایی بهتر این نوع ماتریس‌ها برای تشخیص گوینده می‌باشد.

۵-۵- سنجش کارایی

در این پایان‌نامه برای سنجش کارایی سیستم از دو روش متفاوت استفاده شده است. در روش اول که روش مورد استفاده در اکثر مقالات می‌باشد[۱] و [۵]، در ابتدا باید از سیگنال گفتار ویژگی-های مورد نظر استخراج شود و بردار ویژگی تشکیل داده شود. سپس برای داده‌های تست با طول متفاوت، دنباله بردارهای ویژگی به بخش‌هایی با همپوشانی دارای T بردار ویژگی تقسیم بندی می-شود که دو بخش اول این دنباله در روابط (۵-۱) و (۵-۲) نشان داده شده است.

$$\overbrace{\vec{x}_1,\vec{x}_2\ldots,\vec{x}_T}^{segment\ 1},\vec{x}_{T+1},\vec{x}_{T+2},\ldots \quad (۱-۵)$$

$$\vec{x}_1,\overbrace{\vec{x}_2\ldots,\vec{x}_T,\vec{x}_{T+1}}^{segment\ 2},\vec{x}_{T+2},\ldots \quad (۲-۵)$$

سپس شخص (گوینده) شناخته شده در هر قسمت T برداری با شخص (گوینده) اصلی مقایسه می‌شود و نتیجه آن در یک جدول ذخیره می‌شود. همین مراحل برای داده‌های تست تمام گویندگان تکرار می‌شود. سنجش کارایی نهایی مانند رابطه (۵-۳) محاسبه می‌شود.

$$\text{درصد شناسایی صحیح} = \frac{\text{بخش هایی از } T \text{ که صحیح تشخیص داده شده اند}}{\text{مجموعه تمام بخش ها}} \times 100 \quad (۳-۵)$$

ارزیابی کارایی برای مقادیر متفاوتی از T می‌تواند تکرار شود. برای این که سنجش کارایی نهایی تحت تأثیر گوینده ویژه‌ای نباشد باید تقریباً طول گفتارهای تست تمام گویندگان مساوی باشد. در نهایت درصد تشخیص در یک سیگنال به نویز خاص، از میان‌گیری درصد تشخیص صحیح تمام گوینده‌گان در آن سیگنال به نویز خاص بدست می‌آید.

در روش دوم بردار ویژگی تست گوینده به طور کامل به سیستم داده می‌شود و این روش برای تمام گوینده‌ها تکرار می‌شود و در نهایت درصد شناسایی صحیح طبق رابطه (۵-۴) محاسبه می‌شود

$$\text{درصد شناسایی صحیح} = \frac{\text{افرادی که صحیح تشخیص داده شده اند}}{\text{مجموعه تمام افراد}} \times 100 \quad (۴-۵)$$

در نمودارهایی که در ادامه آمده‌اند ستون سمت چپ نتایج حاصل از روش اول در سنجش کارایی و ستون سمت راست نتایج حاصل از روش دوم در سنجش کارایی می‌باشد.

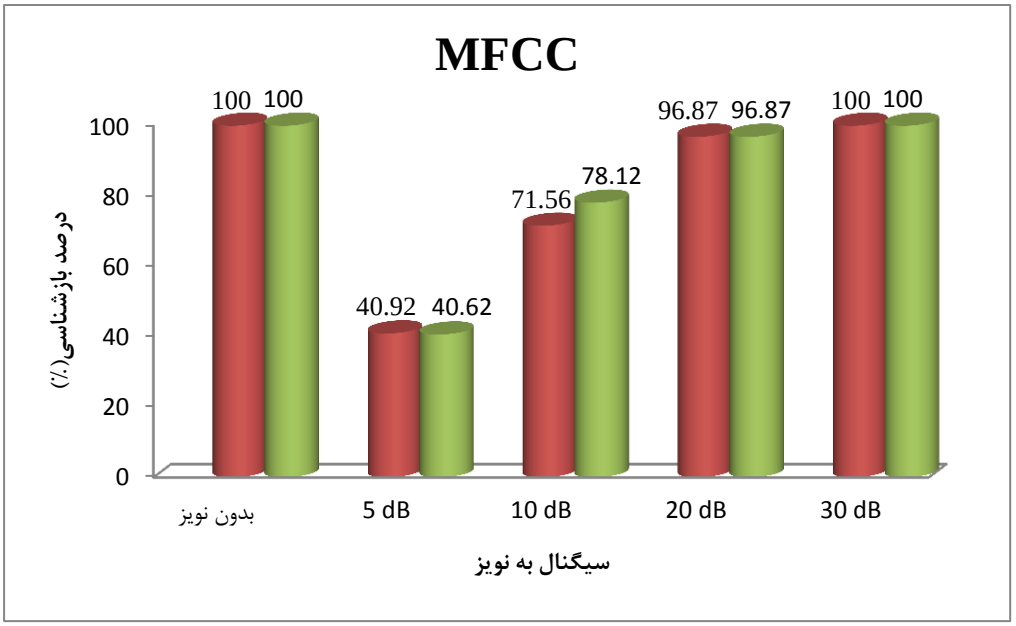
۵-۵- نتایج

در این پایان‌نامه ابتدا روش ضرایب MFCC و PLP پیاده سازی شده‌اند و سپس نتایج حاصل از این روشها با نتایج حاصل از روشهای پیشنهادی مقایسه شده‌اند. برای هر گوینده پنج آزمایش متفاوت انجام شده است که در آزمایش اول از سیگنال بدون نویز و در آزمایش‌های بعدی از سیگنال به نویزهای ۵ dB و ۱۰ dB و ۲۰ dB و ۳۰ dB استفاده شده است. در این آزمایشات از نویز همهمه[1] برای سنجش مقاومت سیستم‌ها استفاده شده است[۲۶]. در ادامه ابتدا نتایج حاصل از پیاده‌سازی هر روش آورده شده است و در نهایت این جواب‌ها مقایسه و تحلیل شده‌اند.

۵-۵-۱- نتایج حاصل از ضرایب MFCC

با توجه به نمودار (۵-۱)، در سیگنال بدون نویز و سیگنال به نویز ۳۰ dB نتایج حاصل از کلاسه بندی برای هر دو روش ۱۰۰ ٪ می باشد و در سیگنال به نویز های دیگر در نمودار (۵-۱) نشان داده شده است.

نمودار (۵-۱): نتایج حاصل از ضرایب MFCC

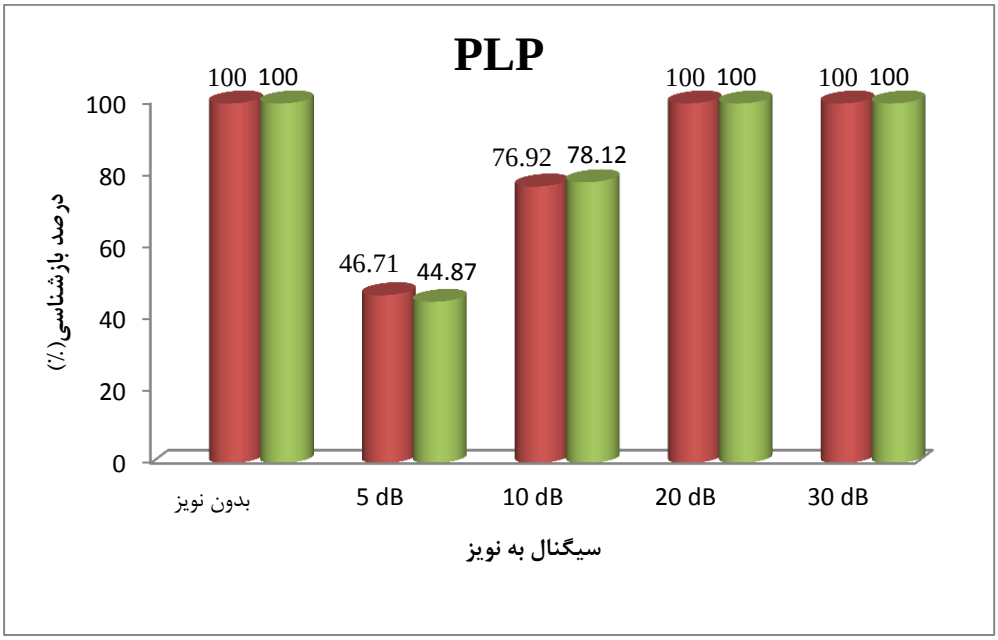


[1] Babble

۵-۵-۲- نتایج حاصل از ضرایب PLP

با توجه به نمودار (۵-۲)، در سیگنال بدون نویز و سیگنال به نویز dB ۲۰ و سیگنال به نویز dB ۳۰ نتایج حاصل از کلاسه بندی برای هر دو روش ۱۰۰ ٪ و در سیگنال به نویزهای دیگر در نمودار (۵-۲) نشان داده شده است. با توجه به نمودارهای (۵-۱) و (۵-۲) نتایج حاصل از ضرایب PLP از نتایج حاصل از ضرایب MFCC بهتر می‌باشد.

نمودار (۵-۲): نتایج حاصل از ضرایب PLP

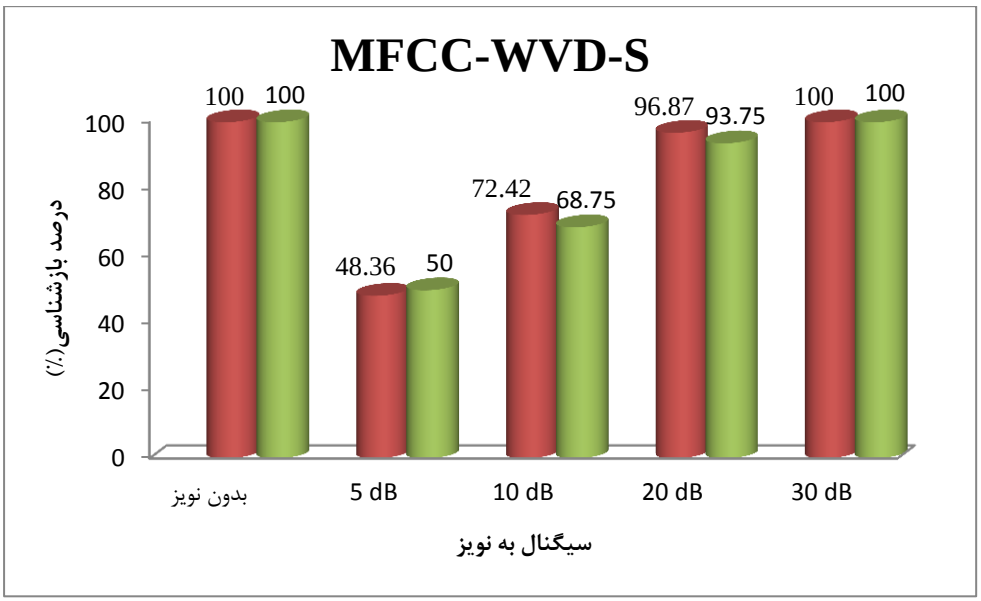


ضرایب بدست آمده از WVD-S را MFCC-WVD-S و ضرایب بدست آمده از PWVD-S را MFCC-PWVD-S را نامگذاری می‌کنیم.

۵-۵-۳- نتایج حاصل از ضرایب *MFCC – WVD - S*

با توجه به نمودار (۵-۳)، در سیگنال بدون نویز و سیگنال به نویز ۳۰ dB نتایج حاصل از کلاسه بندی در هر دو روش ۱۰۰ ٪ می‌باشد. همچنین این روش پیشنهادی در سیگنال به نویز ۵ dB نسبت به روشهای دیگر بهترین نتایج را ارایه می‌دهد. نتایج حاصل از کلاسه‌بندی در سیگنال به نویز های دیگر در نمودار (۵-۳) نشان داده شده است.

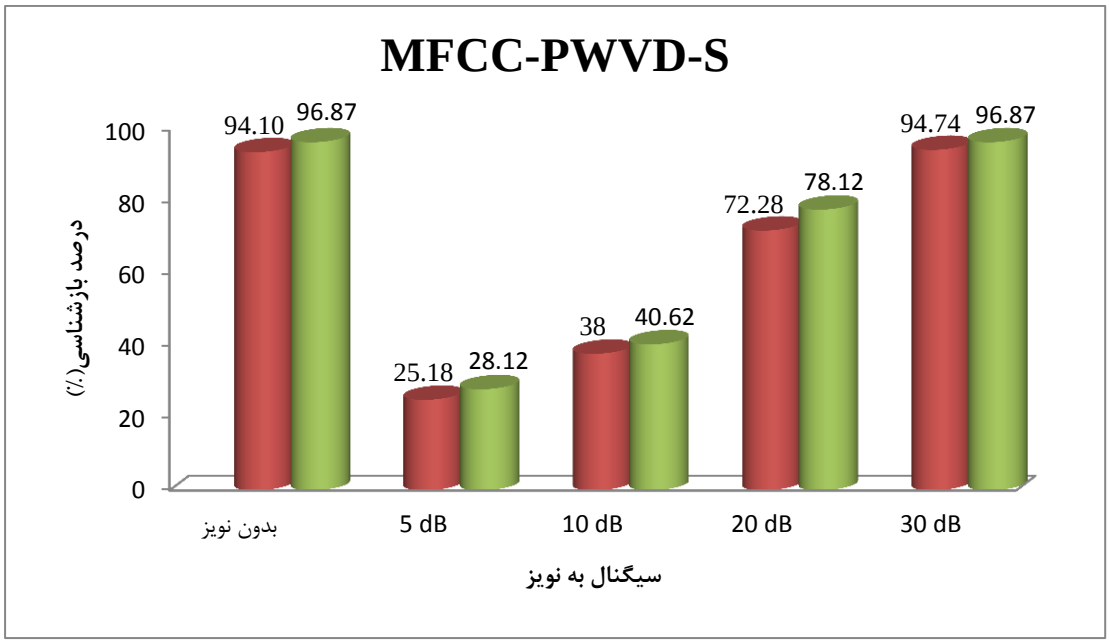
نمودار (۵-۳): نتایج حاصل از ضرایب MFCC – WVD - S



۵-۵-۴- نتایج حاصل از ضرایب *MFCC –PWVD – S*

با توجه به نمودار (۵-۴)، در سیگنال بدون نویز و سیگنال به نویزهای دیگر نتایج حاصل از کلاسه بندی هیچ یک از روشهای ارزشیابی اول و دوم ۱۰۰ ٪ نمی‌باشد. همچنین درصد بازشناسی در سیگنال به نویز ۳۰ dB در روش ارزشیابی اول بیشتر از درصد بازشناسی در سیگنال بدون نویز می-باشد که یک جواب غیر معقول می‌باشد و یکی از دلایل این نتیجه می‌تواند از دست رفتن رزولوشن در نمودار زمان - فرکانس نمونه هموار شده توزیع ویگنر- ویل باشد. همچنین PWVD خاصیت حاشیه-ای را برآورده نمی‌کند

نمودار (۵-۴): نتایج حاصل از ضرایب MFCC – PWVD - S



۵-۶- نتیجه گیری

در جداول (۵-۵) و (۵-۶) نتایج حاصل از همه ضرایب در روش‌ارزشیابی اول و دوم، در سیگنال به نویزهای مختلف نشان داده شده است. همانطور که بیان شد رزولوشن مشخصه فرکانسی تأثیر بسزایی در نتایج تشخیص گوینده‌ها دارد. نتایج نشان داده شده در جدول (۵-۵) گویای این مطلب است که نتایج حاصل از ضرایب MFCC – WVD – S در روش ارزشیابی اول از نتایج حاصل از ضرایب MFCC بهتر می‌باشد. همچنین ضرایب MFCC – WVD – S در سیگنال به نویز ۵ dB جواب‌های بهتری را از ضرایب PLP و MFCC در روشهای ارزشیابی اول و دوم ارایه داده است. به طور کلی در سیگنال به نویزهای پایین روش پیشنهادی حاصل از ضرایب MFCC – WVD – S بهترین جواب و در سیگنال به نویز های بالا ضرایب حاصل از PLP بهترین جواب را دارند. در تابع توزیع PWVD علیرغم اینکه ترم‌های تداخلی کاهش یافته است، عرض فرکانسی ترم‌های اصلی نسبت به WVD بیشتر شده است که این موضوع باعث می‌شود رزولوشن فرکانسی در PWVD کاهش یابد. به همین دلیل ضرایب MFCC – WVD – S جواب‌های بهتری را نسبت به ضرایب MFCC – PWVD – S ارایه می‌دهد.

جدول(۵-۵): نتایج کلاسه بندی حاصل از چهار روش متفاوت در روش ارزشیابی اول

روش / سیگنال به نویز	MFCC	PLP	MFCC-WVD-S روش پیشنهادی۱	MFCC-PWVD-S روش پیشنهادی۲
سیگنال تمیز	۱۰۰	۱۰۰	۱۰۰	۹٤,۱
۵ dB	٤۰,۹۲	٤٦,۷۱	٤۸,۳٦	۲۵,۱۸
۱۰ dB	۷۱,۵٦	۷٦,۹۲	۷۲,٤۲	۳۸
۲۰ dB	۹٦,۸۷	۱۰۰	۹٦,۸۷	۷۷,۲۸
۳۰ dB	۱۰۰	۱۰۰	۱۰۰	۹٤,۷٤

جدول(۵-۶): نتایج کلاسه بندی حاصل از چهار روش متفاوت در روش ارزشیابی دوم

روش / سیگنال به نویز	MFCC	PLP	MFCC-WVD-S روش پیشنهادی۱	MFCC-PWVD-S روش پیشنهادی۲
سیگنال تمیز	۱۰۰	۱۰۰	۱۰۰	۹۶,۸۷
۵ dB	٤۰,٦۲	٤٦,۸۷	۵۰	۲۸,۱۲
۱۰ dB	۷۸,۱۲	۷۸,۱۲	٦۸,۷۵	٤۰,٦۲
۲۰ dB	۹٦,۸۷	۱۰۰	۹۳,۷۵	۷۸,۱۲
۳۰ dB	۱۰۰	۱۰۰	۱۰۰	۹٦,۸۷

برای مقایسه این روشها از لحاظ زمان اجرا، برای یک شخص، آزمایشی انجام شده که نتایج آن در جدول(۵-۷)نشان داده شده است که PLP کمترین و MFCC-PWVD-S بیشترین زمان اجرا را دارند.

جدول(۵-۷): مقایسه زمان برای روش‌های مختلف

ضرایب	MFCC	PLP	MFCC-WVD-S	MFCC-PWVD-S
زمان	۲۱,۱۴ ثانیه	۲۰,۲۵ ثانیه	۲۴,۷۲ ثانیه	۲۵,۴۷ ثانیه

۵-۷- کارهای آینده

با توجه به مباحث بحث شده در فصل سوم در مورد تئوری توزیع ویگنر- ویل، اگر بتوان توزیع ویگنر- ویل (WVD) و نمونه هموار شده آن در فرکانس (PWVD) را با هم ترکیب کرد تا هم از خاصیت رزولوشن بالای توزیع ویگنر- ویل و هم از کم شدن ترم های تداخلی در نمونه هموار شده توزیع ویگنر- ویل (PWVD) استفاده کرد قطعاً به جواب‌های بهتری خواهیم رسید. همچنین برای بهبود نتایج حاصل از بازشناسی می‌توان از ترکیب کلاسیفایرهای مناسب مانند SVM ، HMM و شبکه های عصبی و ... استفاده کرد.

[۱] امیر حسین حاج احمدی، محمد مهدی همایون پور، سید محمد احدی، غلامرضا فراهانی، (۱۳۸۷) "بررسی تأثیر تکنیک‌های مقاوم سازی پیک‌های دنباله خود همبستگی بر مقاومت سیستم‌های بازشناسی گوینده" شانزدهمین کنفرانس مهندسی برق ایران، دانشگاه تربیت مدرس، صفحه ۳۸۳-۳۷۸.

[۲] محمد مهدی همایون پور، امیر حسین حاج احمدی، سید محمد احدی (۱۳۸۶) "استخراج ویژگی ترکیبی مقاوم در مقابل نویز برای سیستم‌های بازشناسی گوینده" سیزدهمین انجمن ملی کامپیوتر ایران، دانشگاه صنعتی شریف، صفحه ۶ .

[۳]رضا شقایق، زینب زینل خانی، جهانشاه کبودیان، "تأثیر شرایط اولیه مناسب بر کارائی فیلتر رستا در سیستم‌های شناسائی زبان" پژوهشکده پردازش هوشمند علائم تهران، صفحه ۸

[۴] Furui, S. (2010) “50 years of progress in speech and speaker recognition” Department of computer science tokyo institute of technology, pp. 1-9.

[۵] Douglas A. Reynolds and Richard C. Rose (1995) “Robust text-independent speaker identification using Gaussian mixture speaker model ” *IEEE Transactions on Speech and audio processing*, Vol. 3, No. 1, pp. 72-83.

[۶] Frahani GH. and Ahadi S. M. and Homayounpoor M. M. (2006) “Feature based on filtering and spectral peaks in autocorrelation domain for robust speech recognition ” *computer speech and language* Vol.21, issue 1, pp. 187-205.

[۷] Huang K. C. and Juang Y. T. and Charg W. CH. (2006) “Robust integration for speech feature ” *signal processing*

[٨] Hermansky H. and Morgan N. (1994) "Rasta processing of speech " *IEEE Transactions on Speech and audio processing*, Vol. 2, No. 4.

[٩] Huang X. and Acero A and Hon H. W. (2001) "spoken language processing " prentice Hall

[١٠] Reynolds D. A. and Campbell W. M. (2008) "text - independent speaker recognition " *Springer Handbook of Speech Processing*, pp. 763-779.

[١١] Minh N. Do "An Automatic speaker Recognition System " *Audio Visual Communication Laboratory , Swiss Federal Institute of Technology*, page. 14.

[١٢] Zufeng Weng, Lin Li, Donghui Guo, "Speaker Recognition using weighted Dynamic MFCC on Gmm," *Proc. Int. Conf. on Signal and Image Processing Application* , pp. 285-288, 2010.

[١٣] Rabiner L. (2007) "Digital speech processing " *Dept. of electrical and computer engineering.*

[١٤] Hachkar Z. Mounir B. Farchi A. Elabbadi J. (2011) "Comparsion of mfcc and plp parameterization in pattern recognition of Arabic Alphabet speech " *Canadian Journal on Artificial Intelligence* , Machine Learning & Pattern Recognition Vol. 2, No. 3, pp. 56-60.

[١٥] Papandreou A. (2003) "Application in Time-Frequency Signal processing" Electrical engineering and applied signal processing series, CRC Press LLC *Arizona State University.*

[١٦] Auger F. and Flandrin P. and Lemoine O. (1995-1996)"Time – Frequency Toolbox For Use with MATLAB" papers. 150.

[١٧] Christopher M. (2006) "Mixture Model and the EM Agorithm Microsoft Research, Cambridge, page. 69.

[۱۸] Takiguchi T. Ariki Y. (2006) “Robust Feature Extraction Using Kernel PCA ” IEEE, Department of Computer and system engineering , pp. 509-512.

[۱۹] Weng Z. and Li L. and Guo D. (2010) “Speaker Recognition Using Weighted Dynamic MFCC Based on GMM ” School of information science and technology, Xiamen University, pp. 285-288.

[۲۰] Seyedin S. and Ahadi M. (2008) “Feature extraction based on DCT and MVDR spectral estimation for robust speech recognition” IEEE Electrical engineering department Amirkabir university of technology,Tehran, iran, pp. 605-608.

[۲۱] Wu Q. and Zahang L. and Shi G. (2010) “Robust Multifactor speech feature extraction based on gabor analysis ” *IEEE*, paper. 10.

[۲۲] Wang N. and Ching P. C. and Zheng N. and Lee T. (2010) “Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features” *IEEE*, paper. 11.

[۲۳] Marir F. “Time frequency analysis of speech signals smoothed pseudo-wigner-ville distribution ” makkan college of thechnology,paper. 7.

[۲۴] Wokurek W. and Hlawatsch F. and Kubin G. “wigner distribution analysis of speech signals ” paper. 5.

[۲۵] Edgar F. Velez and Richard G. Absher “،Transient analysis of speech signal using the wigner time-frequency representation ” department of computer science and electrical engineering, paper. 4.

[۲۶] http://spib.rice.edu/spib/select_noise.html

[۲۷] http://www.mathworks.com/

[۲۸] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[۲۹]. مبانی پردازش سیگنال گفتار، دکتر سعید آیت، انتشارات دانشگاه پیام نور، ۱۳۸۷

Abstract:

With the increasing advance in science and technology, the human needs in all areas of intelligent systems have become more obvious. As a person's voice like every other identifiers such as fingerprint, facial features, iris, etc is unique, it can be used in speaker recognition smart systems. But one of the most important issues in the field of speaker recognition systems is the effect of noise on the speech signal that may lead to a decrease in the recognition accuracy. Due to the importance of robust speaker recognition in noise (or noisy conditions), many studies recently investigated the issue and various methods have been proposed.

The purpose of this thesis is to use the time-frequency distribution of Wigner-Ville for feature extraction. The study employed the combination of Wigner-Ville distribution function and Hillbert transform and MFCC coefficient for feature extraction. In this method, after pre-emphasis and using window, Hilbert transform of speech signal is obtained and then the signal is analysed by Wigner-Ville transformation .The output signal obtained from Wigner-Ville transform is passed through the Mel Filter Bank and after taking the logarithm, the cosine Fourier transform is used. The output of the proposed system is used as a feature vector for speaker recognition. A GMM model also is used to model any speaker. The results of proposed method are compared with MFCC and PLP coefficients. The experiment indicates that the proposed method provides a better results than MFCC and PLP coefficients in low signal to noise ratio.

Keyword: speaker and speech recognition, wigner-ville distribution, feature extraction

Shahrood university of technology

Faculty of Electrical and Robatic Engineering

A **feature extraction method for speaker recognition based on wigner distribution**

Jalil ghasemi

Supervisor:

Dr. Hossain Marvi

Assistance:

Dr. Omidreza Marouzi

February 2012