
بسم الله الرحمن الرحيم

دانشکده مهندسی برق و رباتیک

رشتهٔ مهندسی برق مخابرات سیستم

پایان‌نامه کارشناسی ارشد

بهبود کارایی سیستم تشخیص گفتار با استفاده از ویژگی‌های مبتنی بر اندازه و فاز

نگارنده:

محمد دولتی

استاد راهنما:

دکتر حسین مروی

بهمن ۱۴۰۰

تقدیم به

به‌پاس تعبیر عظیم و انسانی‌شان از کلمه ایثار و ازخودگذشتگان

به‌پاس عاطفه سرشار و گرمای امیدبخش وجودشان که در این سردترین روزگاران بهترین پشتیبان است

به‌پاس قلب‌های بزرگشان که فریادرس است و سرگردانی و ترس در پناهشان به شجاعت می‌گراید

و به‌پاس محبت‌های بی‌دریغشان که هرگز فروکش نمی‌کند

این مجموعه را به **پدر و مادر** عزیزم تقدیم می‌کنم.

سپاسگزاری

بر خود واجب می‌دانم از استاد فرزانه جناب آقای دکتر حسین مروی که به‌عنوان استاد راهنما در مراحل مختلف این پایان‌نامه همواره با سعه‌صدر و گشاده‌رویی در کنار من بودند و در طول مدت تحصیل از راهنمایی‌های اخلاقی و علمی ایشان بهره جسته‌ام تشکر و قدردانی نمایم. همچنین از سایر اساتید محترم دانشکده برق که در طول این دوره از آن‌ها بسیار آموخته‌ام، کمال تشکر را دارم.

از سایر دوستانی که هرکدام به نحوی در تهیه این مجموعه با این‌جانب همکاری داشته‌اند تشکر نموده و موفقیت همه آنها را از خداوند متعال خواهانم.

تعهدنامه

اینجانب **محمد دولتی** دانشجوی دوره کارشناسی ارشد رشته مهندسی مخابرات سیستم دانشگاه صنعتی شاهرود نویسنده پایان‌نامه با موضوع **بهبود کارایی سیستم تشخیص گفتار با استفاده از ویژگی‌های مبتنی بر اندازه و فاز** تحت راهنمایی دکتر حسین مروی متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورداستفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام **دانشگاه صنعتی شاهرود** و یا **Shahrood University of Technology** به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می‌گردد.
- 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 :

امضاء دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن(مقالات، مستخرج، کتاب، برنامه‌های رایانه ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.

- استفاده از اطلاعات نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

با توجه به کاربردهای گسترده سیستم تشخیص گفتار در شرایط محیطی مختلف، مقاوم‌سازی و تقویت عملکرد این سیستم‌ها در برابر انواع نویزها امری ضروری است. یکی از راه‌های بهبود عملکرد این سیستم‌ها، تقویت روش‌های استخراج ویژگی از سیگنال‌های گفتار است. اغلب روش‌های موجود جهت استخراج ویژگی در پردازش گفتار، ویژگی‌های مبتنی بر اندازه را در نظر می‌گیرند و به دلایل مختلفی از به‌کارگیری ویژگی مبتنی بر فاز سیگنال گفتار چشم‌پوشی می‌کنند.

هدف اصلی در این پایان‌نامه استفاده از یک روش جدید استخراج ویژگی مبتنی بر فاز جهت بهبود عملکرد سیستم تشخیص گفتار در شرایط نویزی است. در همین راستا ما در این تحقیق با تغییر در ساختار ضرایب کپسترال نرمالیزه شده توان بهبودیافته (MPNCC) و اضافه‌کردن تأخیر گروهی بهبودیافته (MGD) به آن، یک روش استخراج ویژگی جدید ارائه کرده‌ایم.

عملکرد روش پیشنهادی به کمک شبکه عصبی پرسپترون چندلایه (MLP) در نرم‌افزار متلب و با استفاده از پایگاه‌داده TIIMT جهت تشخیص کلمه در شرایط بدون نویز و نویزی، سنجیده می‌شود. به‌منظور مقایسه عملکرد روش پیشنهادی از دو ویژگی متداول در حوزه تشخیص گفتار یعنی ضرایب کپسترال فرکانس مل (MFCC) و MPNCC در شرایط یکسان استفاده شده است. نتایج شبیه‌سازی بیانگر صحت شناسایی بهتر روش پیشنهادی نسبت به روش MPNCC و قابل‌مقایسه با روش MFCC در شرایط بدون نویز است. در شرایط نویزی در SNRهای مختلف صحت روش پیشنهادی بهتر از دو روش MFCC و MPNCC بوده است. به‌عنوان‌مثال صحت روش پیشنهادی در نویز سفید گوسی در SNR ۵- دسی‌بل ۱۴/۸ و ۹/۷ درصد به ترتیب نسبت به روش‌های MFCC و MPNCC بهبود داشته است.

کلمات کلیدی: بازشناسی گفتار، استخراج ویژگی، مقاوم‌سازی، فاز، تأخیر گروهی، ضرایب کپسترال

لیست مقالات مستخرج از پایان‌نامه:

...

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل اول

کلیات

۱-۱- مقدمه

پیشرفت سریع تکنولوژی مخصوصاً در دهه‌های اخیر سبب شده است تا ماشین‌های گوناگونی جهت سهولت زندگی انسان‌ها اختراع شوند. در سیر تحول تاریخی ارتباطات ارتباط کلامی یکی از مهم‌ترین روش‌های تبادل اطلاعات میان انسان‌ها بوده است که در حال حاضر به تعامل میان انسان و ماشین منجر شده است. لازمهٔ برقراری ارتباط موفق بین انسان و ماشین این است که بازشناسی گفتار توسط ماشین به‌درستی انجام گردد.

در سال‌های اخیر پژوهش‌های زیادی برای طراحی و ایجاد سیستم‌های تعاملی بر پایه گفتار انجام شده است. ازاین‌رو پیشرفت‌های بسیاری در زمینه پردازش گفتار و تحلیل محتوای گفتار انجام شده است. فناوری تشخیص گفتار نوعی فناوری است که به یک کامپیوتر این امکان را می‌دهد که گفتار و کلمات گوینده را از طریق دستگاه‌های ورودی، بازشناسی نماید و به داده‌های موردنیاز و خروجی مطلوب تبدیل کند. در نتیجه سیستم‌های تشخیص خودکار گفتار (ASR[1]) را وسیله ارتباطی آینده بین انسان و ماشین در نظر می‌گیرند [۱].

شرایط ایده‌آل و بدون نویزی که در کاربردهای آزمایشگاهی در نظر گرفته می‌شود در کاربردهای عملی وجود ندارد. به‌عنوان‌مثال وجود نویز در محیط‌هایی مانند قطار، رستوران، خیابان‌های پرسروصدای مرکز شهر، صدای فن کامپیوتر و... سبب برهم‌خوردن شرایط آزمایشگاهی می‌شود. به همین دلیل اگر از سیستمی که در شرایط آزمایشگاهی آموزش‌داده‌شده است جهت بازشناسی گفتار در کاربردهای عملی استفاده گردد، صحت سیستم بازشناسی گفتار به دلیل عدم تطابق داده‌های آموزش و داده‌های آزمایش کاهش می‌یابد. همین امر سبب شده است تا موضوع مقاوم‌سازی و بهبود عملکرد سیستم‌های تشخیص گفتار در کاربردهای عملی به یکی از زمینه‌های مهم پژوهشی در سال‌های اخیر تبدیل شود.

آنالیز سیگنال صوتی ورودی، جهت استخراج ویژگی موردنیاز، از مهم‌ترین قسمت‌های یک سیستم تشخیص گفتار است. روش‌های به‌کارگیری جهت استخراج ویژگی از سیگنال گفتار ورودی از عوامل

[1] Automatic Speech Recognition

مهم و تأثیرگذار در عملکرد سیستم‌های تشخیص گفتار هستند. با تقویت این روش‌ها می‌توان عملکرد این سیستم‌ها را بهبود بخشید. آنالیز فوریه نقش کلیدی در پردازش سیگنال گفتار ایفا می‌کند. به‌عنوان یک کمیت مختلط، می‌توان آن را در شکل قطبی با استفاده از دامنه و فاز طیف سیگنال، نشان داد. مشخصه اندازه طیف تقریباً در هر زمینه‌ای از گفتار به کار می‌رود. اما، مشخصه فازی طیف سیگنال را نقطه شروع مناسبی برای پردازش سیگنال گفتار نمی‌دانند. هدف اصلی در این پایان‌نامه استفاده از ویژگی مبتنی بر فاز استخراج شده از سیگنال گفتار در کنار ویژگی اندازه و نشان‌دادن اهمیت آن در بهبود عملکرد سیستم‌های تشخیص گفتار است.

۱-۲- اهمیت اندازه و فاز طیف سیگنال گفتار

برخلاف اندازه طیف سیگنال که ساختار آن رابطه واضحی با ادراک گفتار دارند، تفسیر و تحلیل مشخصه فاز دشوار است. اکثر الگوریتم‌های موجود در حوزه گفتار، اطلاعات طیف اندازه سیگنال گفتار را در مرکز توجه قرار داده و از اطلاعات طیف فازی بدون انجام هیچ‌گونه پردازشی، صرف‌نظر می‌کردند.

در سال‌های قبل سه دلیل اصلی در عدم به‌کارگیری طیف فازی وجود داشت که عبارت‌اند از: ۱- باور تاریخی که فاز را عاری از اطلاعات می‌دانستند. این باور تاریخی به اواسط قرن نوزدهم برمی‌گردد و منشأ آن قانون آکوستیک اُهم بود که بیان می‌کند گوش انسان‌مانند تحلیل‌گر طیفی عمل می‌کند و طیف فازی هیچ تأثیری بر اینکه گوش چگونه صدا را درک می‌کند، ندارد. ۲- پیچیدگی ناشی از پیچش فاز[1] ، دشواری تفسیر فیزیکی و یا مدل‌سازی ریاضی. باتوجه‌به پدیده پیچش فاز، طیف فازی شکل نویز به خود می‌گیرد که مانع تفسیر فیزیکی و مدل‌سازی ریاضی می‌شود. ۳- نشان‌داده‌شده است که اطلاعات فاز فقط در تجزیه‌وتحلیل طولانی‌مدت (طول پنجره بلند) مفید است درحالی‌که به دلیل غیر ایستان[2] بودن سیگنال گفتار، باید در کوتاه‌مدت (طول پنجره ۲۰-۴۰ میلی‌ثانیه) پردازش شود [۲].

برای بازیابی سیگنال به‌صورت یکتا در حوزه زمان، هر دو مشخصه دامنه و فاز طیف، موردنیاز است.

[1] Phase Wrapping
[2] Non-Stationary

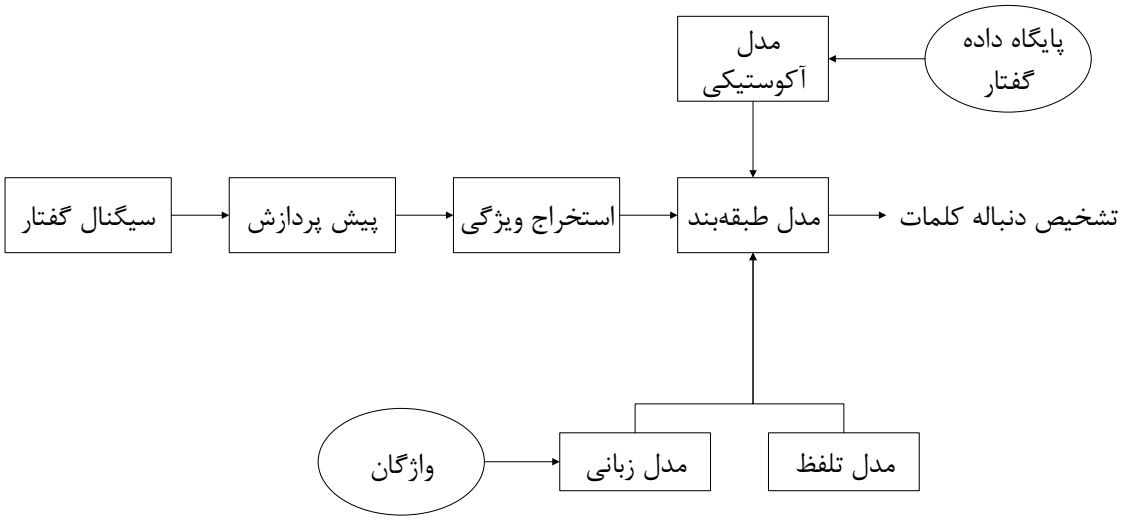
این نشان می‌دهد که مشخصه فازی طیف حاوی اطلاعاتی است. اخیراً فاز طیف گفتار مجدداً توجه‌های زیادی را به خود جلب کرده است. کارهای انجام‌شده در زمینه اهمیت اطلاعات فاز سیگنال گفتار نشان می‌دهد که اطلاعات فاز گفتار می‌تواند در بسیاری از زمینه‌های پردازش گفتار به کار گرفته شود [۳].

۱-۳- گفتار و بازشناسی گفتار

گفتار طبیعی‌ترین، کارآمدترین و ارجح‌ترین شیوه ارتباط بین انسان‌ها است؛ بنابراین می‌توان فرض کرد که ارتباط افراد با ماشین‌های مختلف نیز ازطریق گفتار بسیار راحت‌تر از روش‌های دیگر مانند صفحه‌کلید است. سیستم تشخیص خودکار گفتار (ASR) به ما در رسیدن به این هدف کمک می‌کند. تشخیص گفتار که با عنوان تشخیص خودکار گفتار نیز شناخته می‌شود به کامپیوتر اجازه می‌دهد تا فایل صوتی یا گفتار مستقیم را ازطریق میکروفون به‌عنوان ورودی گرفته و آن را به متن تبدیل کند (ترجیحاً به خط زبان گفتاری) [۴].

۱-۳-۱- ساختار سیستم بازشناسی گفتار

بازشناسی گفتار، به‌عنوان دانش بازیابی کلمات از یک سیگنال صوتی باهدف انتقال محتوای گفتار گوینده به مخاطب تعریف می‌شود. شکل (۱-۱) ساختار کلی سیستم بازشناسی گفتار را نشان می‌دهد. همان‌طور که در شکل (۱-۱) مشاهده می‌شود، یک سیستم بازشناسی گفتار مراحل پیش‌پردازش، استخراج ویژگی، مدل‌سازی آکوستیکی، مدل‌سازی تلفظ، مدل‌سازی زبانی و کدگشایی (طبقه‌بند) را شامل می‌گردد.



شکل ۱-۱ ساختار کلی سیستم بازشناسی گفتار [۵]

۱-۳-۱-۱- پیش‌پردازش

مطابق با شکل ۱-۱ اولین بلوک از سیستم بازشناسی گفتار، بلوک پیش‌پردازش است. سیگنال صوتی ضبط شده یک سیگنال آنالوگ است. یک سیگنال آنالوگ نمی‌تواند مستقیماً به سیستم‌های ASR منتقل شود؛ بنابراین این سیگنال‌های گفتاری باید به شکل سیگنال‌های دیجیتال تبدیل شوند و تنها پس از آن می‌توان آنها را پردازش کرد. همچنین معمولاً ورودی داده شده به ASR با استفاده از میکروفون گرفته می‌شود. این بدان معناست که نویز نیز ممکن است در کنار صدا حمل شود. یکی از اهداف پیش‌پردازش صدا کاهش نویز است. فیلترها و روش‌های مختلفی وجود دارد که می‌توان برای کاهش نویز مربوطه به سیگنال صوتی اعمال کرد. فریم‌بندی[1]، نرمال‌سازی[2] و پیش تاکید برخی از روش‌های در پیش‌پردازش سیگنال هستند [۱]،[۶]. روش‌های پیش‌پردازش بر اساس الگوریتم مورداستفاده برای استخراج ویژگی، متفاوت هستند. برخی از الگوریتم‌های استخراج ویژگی نیاز به نوع خاصی از روش پیش‌پردازش برای سیگنال ورودی آن دارند.

۱-۳-۱-۲- استخراج ویژگی

پس از پیش‌پردازش، سیگنال گفتار وارد بلوک استخراج ویژگی می‌شود. این بلوک سیگنال گفتار

[1] Frame Blocking
[2] Normalization

ورودی را به بردار ویژگی‌ها که برای بازشناسی گفتار مناسب است، تبدیل می‌کند.

مرحله استخراج ویژگی، مجموعه‌ای از پارامترهای گفته‌ها را پیدا می‌کند که با سیگنال‌های گفتاری همبستگی صوتی دارند و این پارامترها از طریق پردازش شکل موج صوتی محاسبه می‌شوند. این پارامترها به‌عنوان ویژگی شناخته می‌شوند. هدف اصلی از استخراج ویژگی حفظ اطلاعات مفید و دور انداختن اطلاعات نامرتبط است [۷]. عملکرد و کارایی طبقه‌بندها به‌شدت به ویژگی‌های استخراج شده وابسته است. روش‌های مختلفی برای استخراج ویژگی‌ها از سیگنال‌های گفتاری وجود دارد. ویژگی‌ها معمولاً تعداد از پیش تعریف شده ضرایب یا مقادیری هستند که با اعمال روش‌های مختلف بر روی سیگنال گفتار ورودی به دست می‌آیند. مدل استخراج ویژگی باید در برابر عوامل مختلف مانند نویز و اثر اکو مقاوم باشد [۸].

۱-۳-۱-۳- مدل آکوستیکی

مدل‌سازی آکوستیک، بخش اساسی سیستم ASR را تشکیل می‌دهد. در مدل‌سازی آکوستیک، ارتباط بین اطلاعات صوتی و آوایی برقرار می‌شود. مدل آکوستیک نقش مهمی در عملکرد سیستم دارد. آموزش، بین واحدهای اصلی گفتار و مشاهدات آکوستیک ارتباط برقرار می‌کند. آموزش سیستم مستلزم ایجاد یک نماینده الگو برای ویژگی‌های کلاس با استفاده از یک یا چند الگوی است که با صداهای گفتاری همان کلاس مطابقت دارد. مدل‌های زیادی برای مدل‌سازی آکوستیک در دسترس هستند. یکی از این مدل‌های بسیار رایج، مدل مخفی مارکوف (HMM[1]) است زیرا الگوریتمی کارآمد برای آموزش و تشخیص است [۴].

۱-۳-۱-٤- مدل زبان و تلفظ

یک مدل زبان شامل محدودیت‌های ساختاری موجود در زبان برای ایجاد احتمال وقوع است. هر زبانی محدودیت‌های خاص خود را دارد. مدل زبان از انواع قواعد و معناشناسی یک زبان تشکیل شده است. مدل‌های زبان برای تشخیص واج پیش‌بینی‌شده توسط طبقه‌بند ضروری هستند و همچنین برای

[1] Hidden Markov Model

تشکیل کلمات یا جملات با استفاده از تمام واج‌های پیش‌بینی‌شده یک ورودی داده شده استفاده می‌شود. اکثر ASRهای مدرن برای کار بدون مدل‌های زبانی نیز طراحی شده‌اند. چنین ASRها می‌توانند کلمات و جملات گفته شده در ورودی داده شده را پیش‌بینی کنند، اما کارایی آنها را می‌توان با استفاده از یک مدل زبان به طور قابل‌توجهی افزایش داد [۹]. در تشخیص گفتار، سیستم کامپیوتری صدا را با توالی کلمات تطبیق می‌دهد. مدل زبان، کلمه و عبارتی را که صدایی مشابه دارند، متمایز می‌کند. به‌عنوان‌مثال، در انگلیسی آمریکایی، عباراتی مانند "recognize speech" و " wreck a nice beach" تلفظ یکسانی دارند اما به معنای آنها کاملاً متفاوت است. وقتی شواهدی از مدل زبان با مدل تلفظ و مدل آکوستیک ترکیب شود، این ابهامات راحت‌تر حل می‌شوند [۱].

۱-۳-۱-۵- مدل طبقه‌بند

این مدل برای پیش‌بینی متن مربوط به سیگنال گفتار ورودی استفاده می‌شود. در بلوک طبقه‌بند، با استفاده از مدل آکوستیکی و مدل تلفظ، بردار ویژگی حاصل از بلوک استخراج ویژگی به یکی از واحدهای گفتاری مشخص مانند واج یا کلمه نگاشت می‌گردد و سپس با استفاده از مدل زبانی و بر اساس یک سری از الگوریتم‌های جستجو، کلمات یا جملات ازروی واحدهای گفتاری استخراج می‌شود.

به‌منظور آموزش مدل مربوط به بازشناسی گفتار نیاز به پایگاه‌داده[1] شامل دادگان گفتار و واژگان است. پایگاه‌دادهٔ موردنیاز باید متناسب با کاربردی باشد که مدل برای آن منظور آموزش می‌بینند.

مدل‌های طبقه‌بند، ویژگی‌های استخراج شده از مرحله استخراج ویژگی را به‌عنوان ورودی جهت پیش‌بینی متن دریافت می‌کنند. مانند مدل استخراج ویژگی، برای مدل طبقه‌بند نیز انواع مختلفی از رویکردها وجود دارد که می‌توان برای تشخیص گفتار استفاده کرد. رویکرد نوع اول از توزیع احتمال مشترک[2] که با استفاده از مجموعه‌داده آموزشی تشکیل شده است، استفاده می‌کند و از توزیع احتمال مشترک برای پیش‌بینی خروجی آینده استفاده می‌شود. این رویکرد، رویکرد مولد نامیده می‌شود.

[1] Database
[2] Joint Probability Distribution

مدل‌های HMM و مدل مخلوط گاوسی (GMM[1])، متداول‌ترین مدل‌های مورداستفاده بر اساس این رویکرد هستند. رویکرد دوم یک مدل پارامتری را با استفاده از مجموعه آموزشی از بردارهای ورودی و بردارهای خروجی مربوطه آنها محاسبه می‌کند. این رویکرد را رویکرد تبعیض‌آمیز می‌نامند. ماشین‌های بردار پشتیبانی (SVM[2]) و شبکه‌های عصبی مصنوعی ANN[3] رایج‌ترین نمونه‌های آن هستند. همچنین رویکردهای ترکیبی نیز می‌توانند برای اهداف طبقه‌بندی استفاده شوند. یکی از نمونه‌های این مدل ترکیبی مدل HMM و ANN است [۱۰].

۱-۳-۲- انواع سیستم‌های تشخیص گفتار

همان‌طور که در شکل (۱-۲) نشان‌داده‌شده است، یک سیستم ASR را می‌توان بر اساس مدل‌های گوینده، واژگان مورداستفاده، انواع کانال و سبک گفتار طبقه‌بندی کرد.

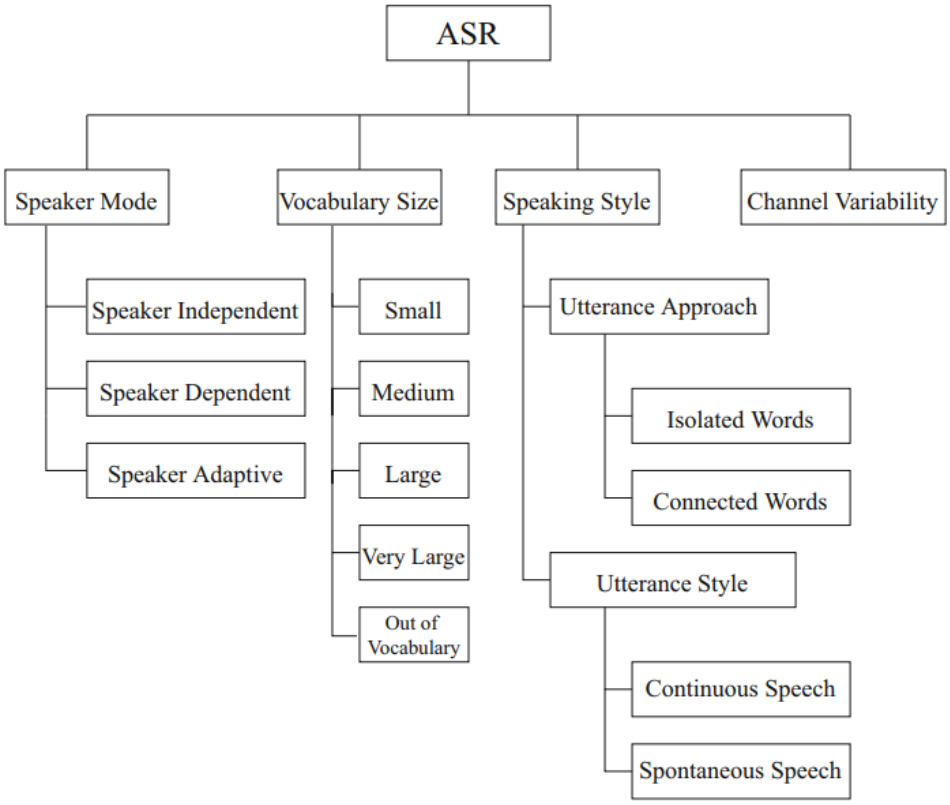
۱-۳-۲-۱- حالت گوینده

هدف از ایجاد سیستم ASR این است که بتواند هر زبانی را برای هر گوینده‌ای ترجمه کند. زبان‌ها از نظر آوایی، مجموعه کاراکترها و قواعد دستور زبان متفاوت هستند. گوینده‌ها نیز از نظر زیروبم صدا، لهجه و شخصیت متفاوت هستند. هر گوینده دارای صدای منحصربه‌فرد و سبک صحبت‌کردن خاص خود را دارد. براین‌اساس، سیستم‌های ASR را می‌توان به سه نوع زیر طبقه‌بندی کرد.

- مدل‌های مستقل از گوینده (SI[4])

ASRهای مستقل از گوینده برای تشخیص گوینده‌های مختلف طراحی می‌شوند. چنین سیستم‌هایی برای کاربر خاصی آموزش داده نمی‌شوند و یکی از پیچیده‌ترین انواع سیستم‌ها برای طراحی هستند. این سیستم‌ها ممکن است دقت کمتری نسبت به روش‌های دیگر ارائه دهند، اما انعطاف‌پذیرتر هستند و می‌توانند کاربرد گسترده‌ای در دنیای واقعی داشته باشند.

[1] Gaussian Mixture Model
[2] Support Vector Machines
[3] Artificial Neural Network
[4] Speaker-Independent



شکل ۱-۲ انواع سیستم‌های ASR [۱]

- مدل‌های وابسته به گوینده (SD[1])

ASRهای وابسته به گوینده برای شناسایی یک یا چندین کاربر از پیش آموزش‌دیده، توسعه‌یافته‌اند. چنین سیستم‌هایی به‌راحتی آموزش داده می‌شوند و همچنین دقت بهتری نسبت به ASRهای مستقل از گوینده ارائه می‌دهند. اما آنها نمی‌توانند همان سطح از نتایج دقیق را برای صداهای خارج از مجموعه که در آن آموزش‌دیده‌اند ایجاد کنند.

- مدل‌های تطبیقی با گوینده (SA[2])

ASRهای تطبیقی با گوینده تا حدودی بین ASRهای مستقل از گوینده و ASRهای وابسته به گوینده قرار دارند. این سیستم‌ها به‌گونه‌ای آموزش‌داده‌شده‌اند که هر زمان که یک گوینده جدید خود را معرفی می‌کند، می‌توانند الگوهای گفتاری جدیدی را بیاموزند [۱].

[1] Speaker-Dependent
[2] Speaker Adaptive

۱-۳-۲-۲- اندازه واژگان

واژگان یک سیستم تشخیص گفتار بسیار مهم است زیرا می‌تواند بر پیچیدگی، زمان پردازش و دقت سیستم تأثیر بگذارد. هر چه اندازه واژگان بزرگ‌تر باشد، سیستم پیچیده‌تر خواهد بود. زمان بیشتری نیز برای آموزش سیستم موردنیاز خواهد بود و متقابلاً دقت بازشناسی سیستم نیز کاهش می‌یابد. برخی از ASRها ممکن است به دایره واژگانی متشکل از ده‌ها کلمه نیاز داشته باشند، به‌عنوان‌مثال، یک سیستم تشخیص گفتار اعداد یا یک سیستم تشخیص کاراکتر. درحالی‌که برای برخی کاربردها حتی ده‌ها هزار کلمه ممکن است کافی نباشد. برای مثال، برای یک ASR که زبان انگلیسی را تشخیص می‌دهد، به واژگان بزرگ‌تری نسبت به سیستم تشخیص اعداد لازم است. در جدول (۱-۱) اندازه واژگان آورده شده است.

جدول ۱-۱ انواع واژگان سیستم تشخیص گفتار

اندازه واژگان	توضیحات
واژگان کوچک	یک واژگان کوچک می‌تواند شامل ده‌ها کلمه باشد.
واژگان متوسط	واژگانی که شامل صدها کلمه باشد، واژگانی با اندازه متوسط در نظر گرفته می‌شود.
واژگان بزرگ	یک واژگان بزرگ می‌تواند از هزاران کلمه تشکیل شده باشد.
واژگان خیلی بزرگ	یک واژگان بسیار بزرگ معمولاً ده‌ها هزار کلمه دارد.
کلمات ناشناخته	تمام کلماتی که بخشی از واژگان نیستند به‌عنوان کلمات ناشناخته ترسیم می‌شوند.

۱-۳-۲-۳- سبک گفتار

در حوزه تشخیص گفتار، یک گفته، یک کلمه ادا شده است. همچنین یک کلمه، چند کلمه، یک جمله و چند جمله را نیز می‌توان به‌عنوان گفته در نظر گرفت. بر اساس نوع گفته‌ها، چندین رویکرد می‌تواند برای توسعه ASR استفاده شود.

- رویکرد بیان

یک گفته به دو نوع تقسیم می‌شود: کلمات مجزا و کلمات متصل.

۱. کلمات مجزا[1]

[1] Isolated Words

سیستمی که مبتنی بر نوع بیان کلمات مجزا است، از کاربرانش می‌خواهد که بین هر کلمه گفتاری مکث کاملاً مشخصی داشته باشند. این لزوماً به این معنا نیست که سیستم هر بار فقط ورودی یک‌کلمه‌ای را دریافت می‌کند و خروجی یک‌کلمه‌ای تولید می‌کند. چنین سیستم‌هایی می‌توانند چندین کلمه را به‌عنوان ورودی دریافت کنند، اما فقط یکی از آنها را در یک‌زمان پردازش می‌کنند [۱۱].

۲. کلمات متصل[1]

کلمات متصل شامل سیستمی است که با جملات متصل کار می‌کند و بین دو یا چند کلمه مکث جزئی یا بدون مکث خواهد داشت. چنین سیستم‌هایی می‌توانند چند کلمه به‌صورت هم‌زمان به‌عنوان ورودی دریافت کنند و آنها را در همان حالت پردازش کند.

- سبک بیان

ازآنجایی‌که هر فرد سبک صحبت‌کردن خاص خود را دارد، گفته‌ها نیز براین‌اساس به دو نوع تقسیم می‌شوند. این دو نوع، گفتار پیوسته و گفتار محاوره‌ای هستند [۱].

۱. گفتار پیوسته[2]

در سیستم‌های تشخیص‌دهندهٔ گفتار پیوسته کاربران اجازه دارند تقریباً با حالت طبیعی صحبت کنند، درحالی‌که رایانه محتوا را تعیین می‌کند. (در اصل، دیکته کامپیوتری است). این نوع سیستم برای پیاده‌سازی بسیار دشوار هستند. زیرا از روش‌های خاصی برای تعیین مرزهای گفتار استفاده می‌کنند.

۲. گفتار محاوره‌ای[3]

گفتار محاوره‌ای کاملاً طبیعی است. چنین گفتاری ممکن است شامل شروع‌های ساختگی، سرفه، خنده، جملات ناقص، تپق و ... باشد.

۱-۳-۲-٤- تغییرپذیری کانال

روش دیگر طبقه‌بندی ASRها بر اساس کیفیت کانال ورودی است. برخی از ASRها به سیگنال‌های

[1] Connected Words
[2] Continuous Speech
[3] Spontaneous Speech

ورودی نیاز دارند که در یک محیط ساکت، یعنی بدون هیچ‌گونه نویز پس‌زمینه، ثبت شوند. نویز اطلاعات غیرضروری یا ناخواسته در سیگنال گفتار ورودی است. نویز می‌تواند هر چیزی باشد، از هرگونه صدایی در پس‌زمینه گرفته تا اعوجاج ناشی از عدم ضبط صحیح صدا. گاهی اوقات وقتی با استفاده از نرم‌افزارهای مختلف کانال موج صوتی ورودی را عوض می‌کنیم، این موج دچار اعوجاج می‌شود. علاوه بر نویز، تفاوت در سن، جنسیت، لهجه، محیط و سرعت صحبت نیز به‌عنوان تغییرات در سیگنال ورودی در نظر گرفته می‌شود. یک سیستم ASR باید بتواند با انواع مختلف نویزهای پس‌زمینه یا تغییرات در سیگنال گفتار ورودی مقابله کند [۱۲].

۱-٤- کاربردهای تشخیص گفتار

فناوری تشخیص گفتار و استفاده از دستیارهای دیجیتال به‌سرعت از تلفن‌های همراه به منازل و ادارات منتقل شده است و کاربرد آن در صنایع مختلف مانند بانکداری، بازاریابی، آژانس مسافرتی و مراقبت‌های بهداشتی در حال آشکار شدن است.

- در اداره

۱. درخواست جستجو اسناد در رایانه

۲. درخواست چاپ

۳. درخواست شروع کنفرانس‌های ویدئویی

٤. برنامه‌ریزی جلسات

معمولاً از سیستم بازشناسی گفتار جهت تایپ خودکار گفتار در امور اداری استفاده می‌گردد. ازآنجایی‌که در یک محیط اداری فن کامپیوتر، تهویه اتاق و نیز گفتگوی افراد ایجاد نویز صوتی می‌کند، در این شرایط اشتباهات بسیاری ممکن است در حین تایپ خودکار گفتار توسط سیستم رخ دهد.

- بانکداری

۱. درخواست اطلاعات مربوط به مانده‌حساب و تاریخچه تراکنش‌ها را

۲. پرداخت‌ها

۳. پاسخ به سؤالات مربوط به وام

٤. تجزیه‌وتحلیل خواسته‌های مشتری

- پزشکی

۱. بازیابی فوری اطلاعات بدون دخالت دست انسان

۲. پرداخت‌ها

۳. دسترسی سریع و هدفمند به اطلاعات موردنیاز

٤. تکرار فرایندها و مراحل دستورالعمل‌های خاص

با استفاده از تقویت سیستم‌های تشخیص گفتار در شرایط نویزی مختلف می‌توان پتانسیل بالایی را در تشخیص گفتار برای افزایش کارایی همه این حوزه‌ها مشاهده کرد و در بسیاری از صنایع دیگر نیز می‌تواند بسیار مؤثر باشد [۱۳].

۱-۵- نویز در سیستم بازشناسی گفتار

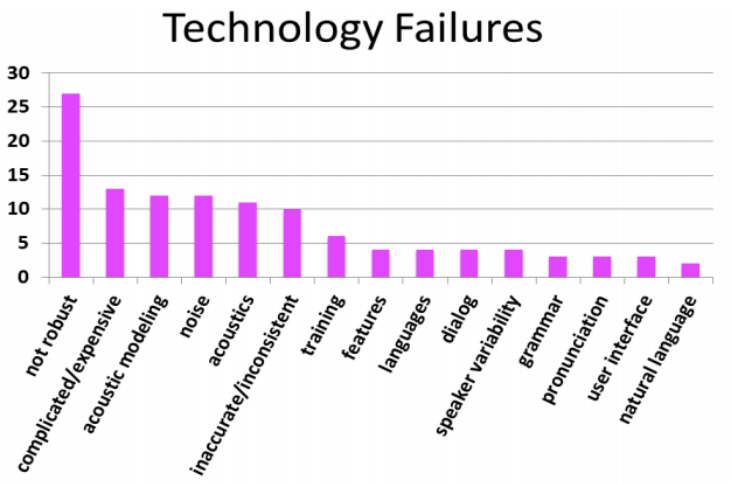
نویز در یک سیستم ارتباطی اساساً سیگنال‌های نامطلوب یا ناخواسته‌ای است که به طور تصادفی به سیگنال حامل اطلاعات واقعی، اضافه می‌شود. نویزها به دودسته کلی ایستا[1] و غیرایستا تقسیم‌بندی می‌گردد. نویزی که دارای چگالی طیف توان ثابت در طول زمان باشد، نویز ایستان نامیده می‌شود. نویز فن کامپیوتر مثالی از نویز ایستان است و نویزی که دارای خواص آماری متغیر در طول زمان باشد نویز غیرایستا نامیده می‌شود. صدای بوق اتومبیل مثالی از نویز غیرایستا است.

نویزها را از لحاظ تأثیرگذاری بر سیگنال گفتار می‌توان به دودسته جمع‌شونده[2] و ضرب شونده[3] دسته‌بندی کرد. نویز جمع‌شونده در حوزه زمان با سیگنال گفتار جمع می‌گردد درحالی‌که اثر نویز ضرب شونده در حوزه فرکانس به‌صورت عمل کانولوشن با سیگنال گفتار است.

به‌منظور جلوگیری از کاهش دقت بازشناسی گفتار ناشی از نویزهای مختلف باید از الگوریتم‌هایی

[1] Stationary
[2] Additive
[3] Convolution

استفاده گردد که سیستم بازشناسی گفتار را نسبت به تغییر شرایط محیطی مقاوم سازد. طبق نظرسنجی انجام شده (شکل ۱-۳) از شرکت‌های فعال در زمینهٔ بازشناسی گفتار، عدم مقاوم‌سازی را به‌عنوان بزرگ‌ترین نقطه‌ضعف این فناوری معرفی کرده‌اند [۱۴].



شکل ۱-۳ ضعف‌های سیستم بازشناسی گفتار مطابق با نظرسنجی متخصّصان [۱۴]

۱-۶- ضرورت و اهمیت تحقیق

باتوجه‌به کاربردهای فراوان سیستم بازشناسی مقاوم گفتار در محیط‌های نویزی گوناگون، نیاز به طراحی سیستم بازشناسی مقاوم گفتار نسبت به انواع نویزها وجود دارد. در حال حاضر تحقیقات زیادی در زمینهٔ مقاوم‌سازی بازشناسی گفتار در حال انجام است. برخی از این تحقیقات مربوط به بازشناسی مقاوم گفتار نسبت به نویز جمعی و برخی هم مربوط به بازشناسی مقاوم گفتار در برابر نویز کانولوشنال است. امّا بررسی تحقیقات انجام شده نشان می‌دهد که سیستم‌های بازشناسی گفتار طراحی شده به این منظور در مرحلهٔ آزمون به علت عدم تطابق محیطی دچار مشکل می‌گردد و این در حالی است که سیستم شنوایی انسان نسبت به این شرایط نویزی محیطی مقاوم عمل می‌کند؛ بنابراین در چنین شرایطی باید از الگوریتم‌هایی در سیستم بازشناسی گفتار استفاده نمود که مانند ساختار شنوایی انسان نسبت به تغییرات شرایط محیطی مقاوم باشد. به همین علت در پژوهش حاضر سعی شده است از الگوریتم استخراج ویژگی استفاده گردد که نسبت به سایر الگوریتم‌ها بیشترین شباهت را به ساختار شنوایی انسان دارد.

۱-۷- ساختار پایان‌نامه

در فصل اول به تعریف و توضیح سیستم خودکار بازشناسی گفتار و مقاوم‌سازی آن پرداختیم و در نهایت ضرورت‌ها و کاربردهای آن را بیان کردیم. در فصل دوم پژوهش‌های صورت‌گرفته در زمینه به‌کارگیری ویژگی فاز را مورد بررسی قرار می‌دهیم. در فصل سوم به مبانی نظری پردازش سیگنال مبتنی بر فاز و استخراج ویژگی پرداخته شده است. در فصل چهارم ابتدا روش پیشنهادی جهت استخراج ویژگی به‌منظور بازشناسی گفتار بیان می‌شود سپس در بخش بعدی پارامترهای بهینه را به دست می‌آوریم و در بخش نهایی پایگاه‌داده استفاده شده در این تحقیق را تشریح می‌کنیم. فصل پنجم به مدل‌سازی سیستم تشخیص گفتار و ارزیابی نتایج به‌دست‌آمده اختصاص دارد و در فصل ششم نتیجه‌گیری‌ها و پیشنهادها مطرح می‌شود.

فصل دوم

ادبیات ومروری بر

پژوهش‌های پیشین

۲-۱- مقدمه

در سال‌های دور و حتی امروزه ویژگی‌های مبتنی بر فاز سیگنال، یکی از بحث‌برانگیزترین موضوعات در زمینه پردازش گفتار بوده است و اکثر ویژگی‌هایی که از سیگنال گفتار استخراج می‌شود مبتنی بر اندازه سیگنال است. در این فصل ابتدا مرور و بررسی اجمالی بر تحقیق‌های انجام شده و بحث‌های صورت‌گرفته از سال‌های دور در خصوص ویژگی فاز و اهمیت آن انجام می‌شود. سپس در بخش بعدی به برخی از کارها و تحقیقات برجسته اخیر انجام شده در خصوص اهمیت به‌کارگیری ویژگی فاز سیگنال در پردازش گفتار، با جزئیات مبسوط‌تر پرداخته می‌شود.

۲-۲- ادبیات تاریخی فاز سیگنال

درگذشته پردازش و به‌کارگیری فاز طیف سیگنال در کاربردهای پردازش سیگنال گفتار تا حد زیادی نادیده گرفته شده است. اولین بینش دررابطه‌با درک فاز به سال ۱۸۴۳ بازمی‌گردد، زمانی که جورج سیمون اهم و هرمان فون هلمهولتز از آزمایش‌های خود به این نتیجه رسیدند که گوش‌های انسان نسبت به فاز حساس نیستند. هلمهولتز مفهوم تحلیل سری فوریه سیگنال‌های متناوب را مطالعه کرد. او حلزون گوش انسان را به‌عنوان یک تحلیلگر طیفی تجسم کرد. هلمهولتز ادعا کرد که بزرگی مؤلفه‌های فرکانس تنها عامل در درک آهنگ‌های موسیقی است. او پیشنهاد کرد که انسان صداهای موسیقی را در قالب آهنگ‌های هارمونیک درک می‌کند و با این جمله نتیجه گرفت: "کیفیت درک شده صداها به فازها بستگی ندارد، بلکه به وجود فرکانس‌های خاص در سیگنال بستگی دارد [۱۵]."

در سال ۱۸۴۴ سیبک از آژیرها برای تولید محرک‌های متناوب با چندین پالس به طور نامنظم در یک دوره استفاده کرد. او تشخیص داد که درک یک تون[1] تنها به قدرت نسبی هارمونیک‌های اصلی در سیگنال بستگی ندارد که با مشاهدات هلمهولتز در تناقض بود. او دریافت که سرعت تکانه‌های درگیر در تولید صدا (فرکانس پیچ) به کیفیت ادراکی نهایی صدا کمک می‌کند. این مشاهدات مهم برای

[1] Tone

برجسته‌کردن اهمیت اطلاعات زیروبم[1] در یک سیگنال بود و منجر به مجادله اهم - زیبک شد [۱۵].

بعدها، چندین پژوهش برای مخالفت با اعتقاد قبلی مبنی بر اینکه گوش انسان فاز ناشنوا است، انجام شد. این سری از مطالعات شامل بیلسن (۱۹۷۳) و کارلسون و همکاران بود. (۱۹۷۹). برخلاف نتایج قبلی هلمهولتز، آنها تغییرات محسوسی را در وضوح پیچ یا طنین[2] به دلیل تغییرات فاز در سیگنال نشان دادند. بعدها، شرودر و استروب (۱۹۸۶) وابستگی دامنه طیفی زمان کوتاه به تغییرات فاز را نشان دادند. آنها این فرضیه را با انجام آزمایش‌هایی که امکان ساخت سیگنال‌های گفتاری صوتی قابل فهمی را که دامنه طیفی مسطح دارند، توجیه کردند. به‌عنوان‌مثال دیگر، نی و هیو در سال ۲۰۰۷ یک تفسیر آماری از اطلاعات فاز در بازسازی سیگنال و تصویر ارائه کردند. آنها در طول آزمایش‌های خود نشان دادند که ویژگی‌های مهم یک سیگنال در طیف فاز آن حفظ می‌شود [۱۵]. همه این پژوهش‌ها و مشاهدات موجود در آنها معمولاً نشان می‌دهند که مدل‌های ساده شنیداری که قبلاً توسط اهم پیشنهاد و توسط هلمهولتز تأیید شده‌اند، هم برای لحن‌های موسیقی و هم برای مصوت‌ها نادرست هستند.

در دهه ۱۹۸۰، چندین محقق بررسی کردند که چگونه می‌توان سیگنال حوزه زمان را تنها ازروی اطلاعات اندازه تبدیل فوریه زمان کوتاه ($STFT^3$) آن تخمین زد. در بسیاری از کاربردها، از جمله اصلاح مقیاس زمانی، رمزگشایی گفتار، سنتز گفتار، تقویت گفتار و جداسازی منبع که در آن هیچ دسترسی به طیف فاز اصلی امکان‌پذیر نیست، یا فاز سیگنال دریافتی تحریف شده یا به دلیل نویز خراب است، به مشکل خوردند [۱۵].

اپنهایم و همکاران نشان دادند که فاز حاوی مقدار زیادی اطلاعات در مورد یک سیگنال است و این واقعیت را از طریق مثال‌هایی در سیگنال‌های تصویر و گفتار نشان دادند که در آن بازسازی فقط فاز می‌تواند درک (برای گفتار) یا تجسم (برای تصاویر) سیگنال اصلی را حفظ کند. در نهایت، آنها الگوریتم‌هایی را برای بازسازی سیگنال از اطلاعات فاز پیشنهاد کردند [۱۶].

[1] Pitch
[2] Timber
[3] Short-Time Fourier Transform

گووژن و موریس نقش فاز در طیف سیگنال دیجیتال را بررسی کردند و یک تکنیک تکرارشونده[1] برای بازسازی تصاویر از اطلاعات فاز آنها پیشنهاد کردند. این آثار مشاهدات قبلی هلمهولتز و اهم را رد کردند و اهمیت بالقوه پردازش سیگنال اطلاعات فاز طیفی را به‌طورکلی و به طور خاص در پردازش صوت نشان دادند [۱۷].

در سال‌های اخیر نیز پژوهش‌های زیادی در خصوص اهمیت ویژگی‌های مبتنی بر فاز سیگنال گفتار در حوزه‌های مختلف پردازش گفتار از جمله گوش‌دادن توسط انسان[2]، بهبود گفتار[3]، تشخیص خودکار گفتار، درک گفتار[4]، و تحلیل گفتار[5] انجام شده است. پالیوال و همکاران در مجموعه‌ای از کارها به اهمیت طیف فاز پرداختند. آلستریس و پالیوال در سال ۲۰۰۷ یک مرور کلی از اهمیت فاز در پردازش گفتار ارائه کرده‌اند. آنها مروری بر فاز STFT در کاربردهای پردازش گفتار ارائه کردند و نتایج تجربی را برای استفاده از طیف فاز کوتاه‌مدت در پردازش گفتار بررسی کردند. این کار سودمندی طیف فاز را برای تشخیص خودکار گفتار، تشخیص گفتار توسط انسان و درک گفتار نشان داد. مشاهده می‌شود که اطلاعات فاز طیفی به بهبود دقت تشخیص‌دهنده‌های خودکار گفتار کمک می‌کند [۱۸].

در طی تحقیقی نشان‌داده‌شده که استفاده از تابع تأخیر گروهی که یک ویژگی مبتنی بر فاز است، باعث بهبود عملکرد سیستم در بازشناسی آواها نسبت به روش PLP که یک مبتنی بر اندازه است، می‌شود [۱۹].

اکثر سیستم‌های ASR یا سیستم‌های تشخیص گوینده بر اساس نمایش‌های کوتاه‌مدت ویژگی‌های طیفی - زمانی ساخته شده‌اند که معمولاً از طیف دامنه محاسبه می‌شوند بااین‌حال، ویژگی‌های جدید به‌دست‌آمده از فاز طیفی به یک رویکرد محبوب تبدیل شده‌اند و گزارش شده است که برای بهبود دقت تشخیص کلی، مفید هستند. به‌عنوان‌مثال، تابع تأخیر گروه اصلاح شده(MGDF[6]) را می‌توان به‌عنوان

[1] Iterative
[2] Human Listening
[3] Speech Enhancement
[4] Speech Intelligibility
[5] Speech Analysis
[6] Modified Group Delay Function

یکی از ویژگی‌های مفید فاز نام برد [۲۰], [۲۱].

در قسمت قبل یک مروری اجمالی بر ادبیات فاز و در مورد ظهور، زوال، و احیای پردازش فاز برای کاربردهای مختلف گفتار و دیدگاه‌های مختلف در مورد بی‌اهمیت بودن یا اهمیت فاز ارائه شد. نتایج، اهمیت و پتانسیل طیف فاز را در پردازش گفتار نشان می‌دهند.

۲-۳- کارهای انجام شده

ژو و پالیوال از ویژگی‌های مبتنی بر فاز برای تشخیص گفتار استفاده کردند [۲۲]. ازآنجایی‌که تبدیل فوریه سیگنال گفتار از طیف اندازه و طیف فاز تشکیل شده است، ویژگی‌های مشتق شده از طیف توان یا مشخصه فاز دارای محدودیت‌هایی در نمایش سیگنال هستند. در این مقاله، طیف حاصل‌ضرب (PS) به‌عنوان حاصل‌ضرب طیف توان در تابع تاخیر گروه (GDF) تعریف شده است که این کار طیف اندازه و طیف فاز را ترکیب می‌کند. در این مقاله آن‌ها ضرایب کپسترال فرکانس مل (MFCC[1]) را از طیف حاصل‌ضرب استخراج می‌کنند و آنها را ضرایب کپسترال طیف حاصلضرب فرکانس مل (MFPSCC[2]) می‌نامند. همچنین MFCC استخراج شده از MGDF را ضرایب کپسترال تأخیر گروهی اصلاح شده فرکانس مل (MFMGDCCs[3]) نام‌گذاری می‌کنند. آنها عملکرد ویژگی پیشنهادی خود یعنی MFPSCCs را در تشخیص گفتار با سه ویژگی زیر مقایسه می‌کنند:

ضرایب کپسترال فرکانس مل (MFCCs)

ضرایب کپسترال تأخیر گروهی اصلاح شده (MGDCCs)

ضرایب کپسترال فرکانس مل تأخیر گروه اصلاح شده (MFMGDFCCs[4])

پایگاه‌داده Aurora2 برای ارزیابی عملکرد استفاده شده است. از این پایگاه‌داده می‌توان برای ارزیابی عملکرد الگوریتم‌های تشخیص گفتار در شرایط نویز استفاده کرد. گفتار منبع برای این پایگاه‌داده، TIDigits است که شامل تلفظ ارقام متصل به زبان انگلیسی آمریکایی است که با فرکانس ۸ کیلوهرتز

[1] Mel-frequency Cepstral coefficients
[2] Mel-frequency Product Spectrum Cepstral Coefficients
[3] Mel-frequency Modified-Group-Delay Cepstral Coefficients
[4] Mel-frequency Modified-Group-Delay Cepstral Coefficients

نمونه‌برداری شده است. دو مجموعه آموزش و سه مجموعه آزمایش وجود دارد. مجموعه‌های آزمایشی اول (A) و دوم (B) دارای گفتار آغشته به نویز هستند که توسط نویزهای مختلف جمع‌شونده در دنیای واقعی با سیگنال به نویزهای (SNR) از ۵- دسی‌بل تا ۲۰ دسی‌بل با گام ۵ دسی‌بل آغشته می‌شوند. مجموعه آزمایشی سوم (C) هم تحت‌تأثیر نویز جمع‌شونده و هم نویز کانولوشنی است.

از مجموعه آموزش بدون نویز برای آموزش مدل مخفی مارکوف (HMM) استفاده شده است. در محاسبه تمامی ویژگی‌ها، سیگنال گفتار هر ۱۰ میلی‌ثانیه با عرض فریم ۳۰ میلی‌ثانیه (با پنجره همینگ و پیش تأکید) بررسی شده است. فیلتر بانک مل با ۲۳ باند فرکانسی در محدوده ۶۴ هرتز تا ۴ کیلوهرتز طراحی شده است. در نهایت ۱۲ ضریب کپسترال به‌دست‌آمده است. در محاسبه MGDF، طیف هموار شده کپسترالی از ۱۳ ضریب کپسترال مرتبه پایین (شامل ضریب مرتبه ۰) به‌دست‌آمده است. در محاسبه MGDCCs، دو پارامتر α و γ (پارامترهای کنترل محدوده تغییرات که به‌صورت تجربی به دست می‌آیند) به صورت $\alpha = 0.4$ و $\gamma = 0.9$ تنظیم شده‌اند.

جدول‌های (۲-۱) و (۲-۲) صحت ویژگی‌ها را در سه مجموعه تست نشان می‌دهند. در جدول (۲-۱)، ویژگی‌ها فقط شامل ۱۲ ضریب کپسترال است. در جدول (۲-۲) ویژگی‌ها شامل ۱۲ ضریب کپسترال، انرژی، دلتا و ضرایب شتاب‌دهنده[1] (دلتا دلتا) است که در مجموع ۳۹ ضریب هستند.

جدول ۲-۱ مقایسه ویژگی‌های متشکل از ضرایب کپسترال [۲۱]

Test set	Feature set	SNR(dB)							
		Clean	20	15	10	5	0	-5	Ave
A	MFCC	96.64	88.71	79.00	60.14	34.88	18.69	12.86	56.28
	MGDCC	81.32	70.22	60.53	47.00	28.49	10.85	-0.64	43.42
	MFMGDCC	55.32	50.69	48.43	42.02	32.47	19.90	12.09	38.70
	MFPSCC	96.41	89.11	80.08	63.17	37.62	19.98	13.38	57.99
B	MFCC	96.64	90.14	82.98	66.87	41.94	22.54	13.75	60.90
	MGDCC	81.32	69.46	60.76	47.09	28.60	11.61	-0.16	43.50
	MFMGDCC	55.32	48.07	45.56	38.14	28.23	16.99	9.71	35.40
	MFPSCC	96.41	90.40	83.30	68.61	44.82	23.94	14.20	62.22
C	MFCC	96.65	89.82	80.67	62.65	38.80	20.97	14.11	58.58
	MGDCC	81.46	68.66	54.76	36.95	17.74	1.38	-5.16	35.90
	MFMGDCC	52.47	51.86	49.23	43.35	34.46	22.56	12.86	40.29
	MFPSCC	96.32	90.41	81.59	65.33	42.18	22.66	14.49	60.43

[1] Accelerator Coefficients

برای هر مجموعه تست، صحت در نویزهای مختلف به طور میانگین محاسبه شده است. ستون آخر میانگین صحت در سیگنال به نویزهای ۲۰ تا صفر دسی‌بل است.

با تحلیل نتایج به‌دست‌آمده داریم:

۱. MFCCها عملکرد بهتری نسبت به MGDCC و MFMGDCC ارائه می‌دهند. این نشان می‌دهد که طیف توان عملکرد بهتری نسبت به طیف فاز دارد.

جدول ۲-۲ مقایسه ویژگی‌های متشکل از ضرایب کپسترال، انرژی، دلتا و ضرایب شتاب دهنده (دلتا دلتا)[۲۲]

Test set	Feature set	SNR(dB)							
		Clean	20	15	10	5	0	-5	Ave
A	MFCC+Δ+ΔΔ	99/34	97/04	92/24	76/79	44/70	22/36	13/04	66/63
	MGDCC+Δ+ΔΔ	86/50	75/27	65/60	50/71	27/69	4/17	-10/26	44/69
	MFMGDCC+Δ+ΔΔ	81/06	7176	68/96	61/09	45/39	2533	13/20	54/98
	MFPSCC+Δ+ΔΔ	99/31	96/94	92/36	78/68	48/60	2137	13/43	67/99
B	MFCC+Δ+ΔΔ	99/34	97/81	94/28	82/73	54/48	26/93	14/26	7115
	MGDCC+Δ+ΔΔ	86/50	7158	64/37	48/84	26/07	5/06	-8/94	4158
	MFMGDCC+Δ+ΔΔ	81/06	7101	67/47	57/70	41/40	20/64	10/57	52/04
	MFPSCC+Δ+ΔΔ	99/31	97/77	94/22	84/08	57/86	28/45	14/66	72/48
C	MFCC+Δ+ΔΔ	99/27	96/65	90/82	74/05	43/44	21/96	13/59	65/38
	MGDCC+Δ+ΔΔ	86/81	71/78	56/00	34/59	8/34	-13/91	-26/44	31/36
	MFMGDCC+Δ+ΔΔ	80/65	72/55	67/76	59/11	43/73	24/17	12/20	53/46
	MFPSCC+Δ+ΔΔ	99/28	96/73	91/43	76/31	46/82	23/29	13/99	66/91

۲. MFMGDCCها در SNRهای کم عملکرد بهتری نسبت به MGDCCها دارند، اما در SNRهای بالا یا در شرایط بدون نویز بدتر هستند. این نشان می‌دهد که هموارسازی طیفی با فیلتر بانک فرکانس مل برای به‌دست‌آوردن یک نمایش قوی در شرایط ناسازگار مفید است.

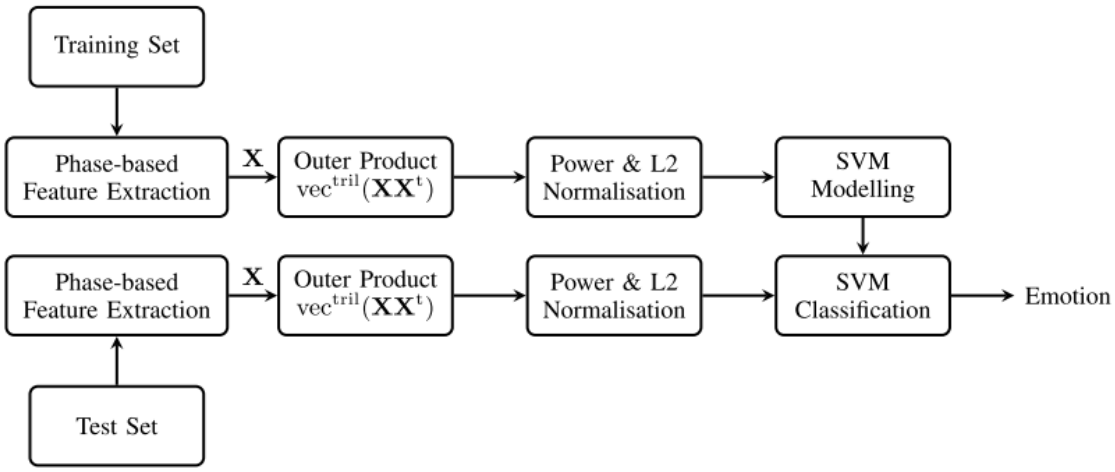
۳. MFPSCCها بهترین عملکرد را دارند. این نشان می‌دهد که طیف حاصل‌ضرب بهتر از طیف توان و طیف فاز برای تشخیص گفتار است.

در[۲۳] تحقیقی در زمینه تشخیص احساسات از گفتار زمزمه شده با بهره‌گیری از ویژگی‌های مبتنی بر فاز انجام شده است. آنها دو نوع ویژگی مبتنی بر فاز را انتخاب کرده‌اند که عبارت‌اند از: ویژگی‌های تأخیر گروهی اصلاح‌شده (MGD)، و ویژگی‌های تأخیر گروه تمام قطبی (APGD[1]) که هر دو کاربرد

[1] All-Pole Group Delay

گسترده‌ای در انواع تحلیل‌های گفتاری مختلف نشان داده‌اند و اکنون در تشخیص احساسات گفتار زمزمه‌شده موردمطالعه قرار می‌گیرند. در این مقاله همچنین یک چارچوب جدید تشخیص احساسات گفتار، پیشنهاد شده است که از ضرب خارجی و نرمال‌سازی توان و L2 تشکیل شده است.

آنها از تکنیک ضرب خارجی[1] برای نگاشت دنباله های بردارهای ویژگی مبتنی بر فاز با طول متغیر به یک بردار ویژگی با طول ثابت استفاده کرده اند که از الزامات کلاسه بند ماشین بردار پشتیبان(SVM) است. علاوه بر این، نرمال‌سازی توان و L2 برای تصحیح مقادیر حاصل از بردار ضرب خارجی انجام می‌شود تا در محدوده مناسب برای مدل‌سازی SVM قرار بگیرند. سیستم پیشنهاد شده دراین تحقیق در شکل (۲-۱) آورده شده است.



شکل ۲-۱ روند نمای سیستم تشخیص احساسات گفتار پیشنهاد شده در این تحقیق، با استفاده از استخراج ویژگی مبتنی بر فاز، ضرب خارجی، توان و نرمال‌سازی L2 و SVM ها [۲۳]

این سیستم به‌طورکلی از چهاربخش اصلی شامل بخش استخراج ویژگی مبتنی بر فاز، بخش ضرب خارجی، یک بخش نرمال‌سازی و یک بخش SVM تشکیل شده است.

در این مقاله از دیتاست مجموعه احساسات زمزمه شده ژنو[2] (GeWEC) برای ارزیابی اثربخشی سیستم پیشنهادی استفاده شده است. این مجموعه شامل جفت نطق عادی و زمزمه شده است. در این پایگاه‌داده دو بازیگر مرد و دو بازیگر زن حرفه‌ای فرانسوی‌زبان در ژنو سوئیس استخدام شدند تا هشت

[1] Outer Product
[2] Geneva Whispered Emotion Corpus

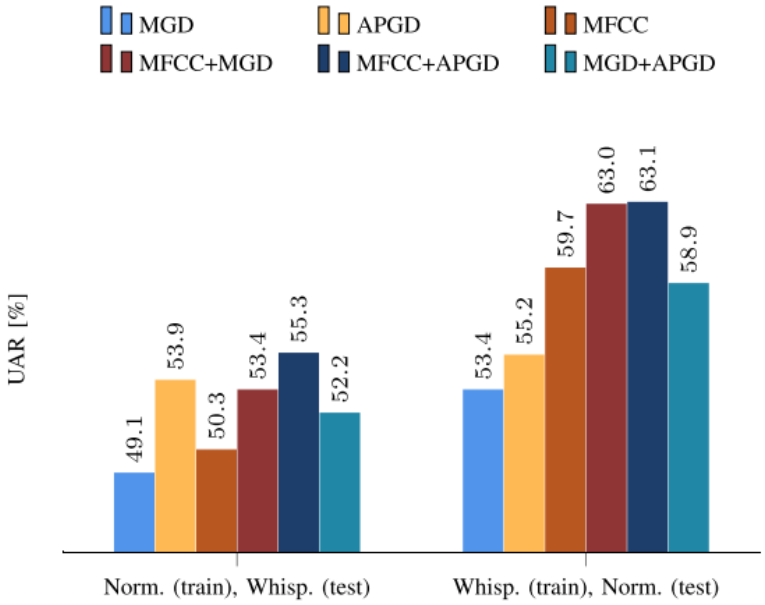
شبه کلمه (کلماتی که وجود خارجی ندارند ولی ساختار کلمه را دارند) فرانسوی از پیش تعریف شده (مانند «belam» و «molen») را با یک حالت احساسی مشخص، در هر دو حالت گفتاری عادی و زمزمه‌ای صحبت کنند. از بازیگران خواسته شده است که هر کلمه را در هر چهار حالت احساسی (عصبانیت، ترس، شادی و حالت عادی) پنج بار بیان کنند. در نتیجه GeWEC در مجموع از ۱۲۸۰ نمونه تشکیل شده است. یک حالت گفتاری از داده‌های GeWEC برای آموزش استفاده می‌شود درحالی‌که داده‌های حالت گفتار دیگر برای آزمایش استفاده می‌شود. در این تحقیق یک حالت گفتاری از داده‌های GeWEC برای آموزش و بقیه حالت‌ها برای آزمایش استفاده می‌شود.

فریم‌بندی با استفاده از پنجره همینگ[1] با طول فریم ۲۵ میلی‌ثانیه و تغییر فریم ۱۰ میلی‌ثانیه انجام می‌شود. هنگام محاسبه ویژگی‌های MGD دو پارامتر تنظیم α و γ به ترتیب روی ۰.۱ و ۰.۲ تنظیم می‌کنند که در نهایت به ویژگی‌های ۳۶ بعدی از جمله ضرایب دلتا و دلتا، دلتا ختم می‌شوند. در مورد ویژگی‌های APGD،مرتبه حالت تمام قطبی30 (p) است و ۱۸ ضریب APGD حفظ می‌شود. ضرایب دلتا و شتاب به ضرایب APGD اضافه می‌شوند تا بردارهای ویژگی ۵۴ بعدی را تشکیل دهند. علاوه بر این دو ویژگی مبتنی بر فاز، متداول‌ترین ویژگی‌های مبتنی بر اندازه یعنی MFCC با ۳۶ بعد نیز استخراج می‌شوند. تشخیص میانگین بدون وزن[2] (UAR) به‌عنوان معیار عملکرد استفاده می‌شود.

شکل (۲-۲) نتایج تجربی ترکیبات مختلف ویژگی را در دو حالت نشان می‌دهد. باتوجه‌به شکل (۲-۲) مشاهده می‌شود که زمانی که ویژگی‌های MGD یا APGD با ویژگی‌های MFCC ترکیب می‌شوند، هر ترکیب افزایش قابل‌توجهی در عملکرد نسبت به حالتی که هرکدام به‌صورت جداگانه باشند، ایجاد می‌کند. علاوه بر این، ترکیب ویژگی‌های MFCC و APGD حتی از نظر آماری به طور قابل‌توجهی بهتر از ویژگی‌های MFCC برای آزمایش تشخیص احساسات زمزمه‌شده عمل می‌کند. همچنین ترکیب ویژگی‌های MGD و APGD در مقایسه با ویژگی‌های اصلی MGD یا APGD باعث بهبود عملکرد

[1] Hamming
[2] Unweighted Average Recall

متوسطی می‌شود.



شکل ۲-۲ مقایسه عملکرد ویژگی‌ها با معیار UAR برای ترکیب‌های مختلف دو ویژگی مبتنی بر فاز (یعنی MGD و APGD) و ویژگی‌های مبتنی بر اندازه (یعنی MFCCs) [۲۳]

در طی تحقیقی از ویژگی‌های مبتنی بر فاز، جهت بهبود تشخیص گفتار دیزارتریک (DYSARTHRIC) استفاده شده است [۲۴]. دیزارتری یک اختلال گفتاری عصبی است که معمولاً به ازدست‌دادن کنترل گفتار به دلیل ضعیف بودن عضلات و ناهماهنگی مفصل‌ها منجر می‌شود. در نتیجه، به دلیل ناسازگاری در سیگنال صوتی و پراکندگی داده، گفتار کمتر قابل‌درک و مدل‌سازی با الگوریتم‌های یادگیری ماشینی دشوار می‌شود.

ویژگی فازی استفاده شده، طیف تأخیر گروه است. چنین نمایش‌هایی برای توصیف طنین‌های مجرای گفتار مناسب‌تر هستند و همچنین قابلیت‌های تشخیص واج بهتری در سیگنال‌های دیزارتریک نشان می‌دهند و در نتیجه عملکرد سیستم تشخیص خودکار گفتار (ASR) را بهبود می‌بخشند. ویژگی‌های فازی که استفاده کردند MODGDF و طیف حاصل‌ضرب (PS[1]) هستند. پارامترهای هموارسازی MODGDF به‌صورت تجربی با مقادیر بهینه تنظیم شده در $\alpha = 0.95$ و $\gamma = 0.2$ برای آزمایش‌ها تعیین شدند. طیف‌های مبتنی بر MODGDF و PS با استفاده از پنجره Hanning-Poisson

[1] Power Spectrum

که بهترین وضوح را ارائه می‌دهد، پردازش شدند. ضرایب کپسترال مبتنی بر MODGDF و PS به‌عنوان MODGDFCC و PSCC نامیده می‌شوند.

در این تحقیق از سه پایگاه‌داده برای همه تجزیه‌وتحلیل‌ها و آزمایش‌ها استفاده شده است. مجموعه وال‌استریت ژورنال[1] (WSJ SI-84) و معادل بریتیش آن (WSJCAM0) به‌عنوان داده‌های گفتار معمولی استفاده شده‌اند. علاوه بر این، مجموعه UASPEECH داده‌های دیزآرتریک را تشکیل می‌دهند. گوینده‌ها شامل ۱۵ گوینده مبتلا به فلج مغزی و ۱۳ فرد سالم همسن هستند. برای هر گوینده ۷۶۵ کلمه جدا شده (ایزوله) وجود دارد که در سه بلوک جداگانه (B1,B2,B3) جمع‌آوری‌شده‌اند. هر بلوک شامل ۱۰ رقم، ۲۶ کلمه رادیویی بین‌المللی، ۱۹ فرمان کامپیوتری، ۱۰۰ کلمه رایج و ۱۰۰ کلمه غیرمعمول متمایز است که در همه بلوک‌ها تکرار نشده‌اند. تمام داده‌های B1+B3 برای آموزش و B2 برای آزمایش استفاده می‌شود. عملکرد ویژگی‌های پیشنهادی با ویژگی معروف MFCC مقایسه شده است. هر یک از ویژگی‌های کپسترال به‌عنوان یک بردار ۳۹ بعدی با ۱۲ ضریب استاتیک، c0، مشتقات زمانی مرتبه اول و دوم نشان‌داده‌شده است. از مدل مخفی مارکوف (HMM) برای مدل‌سازی استفاده شده است. در جدول (۲-۳) انواع سیستم‌ها و داده‌های آموزشی که برای هرکدام از آنها استفاده شده است را مشاهده می‌کنید.

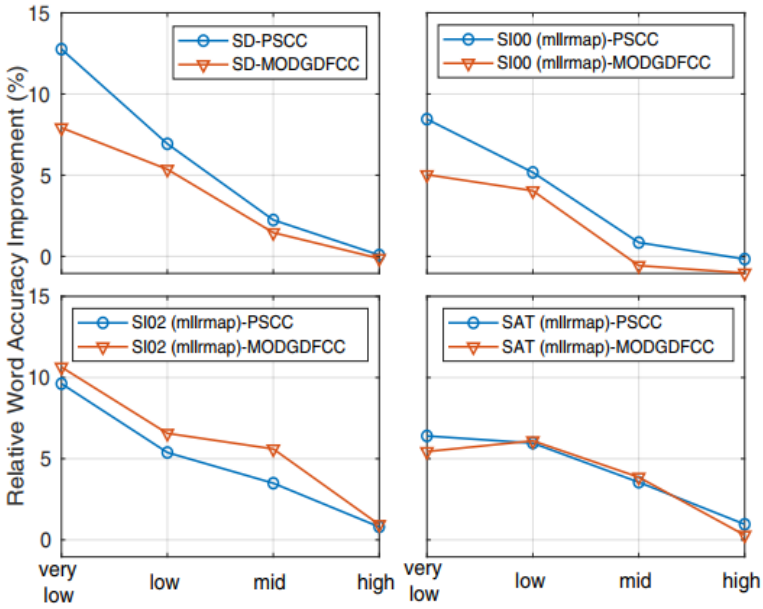
- وابسته به گوینده (SD)
- مستقل از گوینده (SI)
- سازگار با گوینده (SA)

جدول ۲-۳ سیستم‌های ASR و پایگاه‌داده‌های استفاده شده [۲۴]

System	Training Dataset Used
SD	UASPEECH-Dysarthria
SI-00	WSJ SI-84 + WSJCAM0
SI-02	UASPEECH-Dysarthria
SAT	UASPEECH-Dysarthria

[1] Wall Street Journal

در شکل (۲-۳) درصد بهبود دو روش پیشنهادی نسبت به روش MFCC در چهار سیستم مختلف و در ۴ حالت قابل‌فهم بودن کلمات آورده شده است.



شکل ۲-۳ بهبود نسبی صحت تشخیص کلمه برای ضرایب کپسترال مبتنی بر PSCC و MODGDFCC [۲۴]

همان‌طور که مشاهده می‌کنید هر دو ویژگی PSCC و MODGDFCC برای مدل‌سازی گفتار دیزارتریک با درجه بیشتری از اختلال پاتولوژیک در مقایسه با MFCC مبتنی بر اندازه بسیار مؤثر بودند. همچنین به‌منظور بررسی آماری بهبود عملکرد سیستم تشخیص خودکار گفتار (ASR) مبتنی بر ویژگی‌های فاز، یک آزمون کوکران Q (Cochran's Q test) به‌صورت دوبه‌دو برای ویژگی‌های ویژگی MODGDFCC/Standard-MFCC و PSCC/Standard-MFCC انجام شده است. جدول (۲-۴) امتیازات صحت تشخیص کلمه را برای دو ویژگی نشان می‌دهد.

جدول ۲-۴ امتیازات صحت تشخیص کلمه در حالت های مختلف [۲۴]

Intelligibility	**PSCC Features**				**MODGDFCC Features**			
	SD	SI-02	SAT	SI-00	SD	SI-02	SAT	SI-00
very-low	23/52	27/36	28/71	20/61	23/52	27/36	28/71	20/61
	26/52	30/00	30/55	22/36	25/33	30/27	30/28	21/65
low	62/48	62/92	62/98	57/89	62/48	62/92	62/98	57/89
	66/81	66/30	66/72	60/89	65/82	67/05	66/83	60/23
mid	64/08	68/51	69/54	66/12	64/08	68/51	69/54	66/12
	65/52	70/90	72/02	66/69	65/02	72/34	72/23	66/23
high	83/07	86/17	86/87	87/08	83/07	86/17	86/87	87/08
	83/14	86/86	87/71	86/93	82/96	86/98	87/12	86/19

از ۱۶ ترکیب ممکن بین چهار سیستم و حالت‌های قابل‌فهم بودن کلمات، ویژگی‌های PSCC در

۱۳ سیستم و ویژگی‌های MODGDFCC در ۱۲ سیستم عملکرد قابل‌توجهی نشان می‌دهند. همچنین برای هر دو نمایش ویژگی، همه سیستم‌ها عملکرد بسیار قابل‌توجهی را برای گروه‌ها با قابلیت فهم بسیار کم، نشان داده‌اند. این یک نتیجه دلگرم‌کننده است، زیرا هدف اکثر سیستم‌های تشخیص گفتاردیزآرتریک در درجه اول تشخیص گفتار کاربرانی با درجه بالایی از اختلال (قابلیت فهم بسیار کم) است. ازاین‌رو، نمایش ویژگی بر اساس طیف‌های تأخیر گروهی (ویژگی مبتنی بر فاز) در گفتار نارسا می‌تواند به طور قابل‌توجهی برای مدل‌سازی صوتی قوی، مفید باشد.

همچنین شایان‌ذکر است که ویژگی PSCC برای مدل‌سازی وابسته به گوینده نسبت به MODGDFCC، برای گوینده‌هایی با قابلیت فهم بسیار پایین، عملکرد بسیار بهتری دارد. به نظر می‌رسد انتخاب بین PSCC و MODGDFCC یک موضوع اختیاری است و می‌تواند به کاربردهای خاص وابسته باشد. همچنین MODGDFCC با یک محدودیت اضافی برای یافتن مقادیر بهینه (α, γ) همراه است که می‌تواند به پایگاه‌داده وابسته باشد، درحالی‌که PSCC از چنین محدودیت‌هایی عاری است و می‌تواند از اطلاعات در هر دو طیف بزرگی و فازی بهره ببرد.

در[۲۵] تشخیص گفتار بازپخش شده (PSD[1]) را به کمک ویژگی‌های مبتنی بر اندازه و فاز سیگنال گفتار موردمطالعه قرار داده‌اند. ازآنجایی‌که در مطالعات پیشین PSD اغلب مرسوم بود که فقط ویژگی اندازه (مانند MFCC) را از سیگنال استخراج کنند، در این تحقیق ابتدا به‌منظور استخراج اطلاعات بیشتر از گفتار، جهت تشخیص بهتر در سیستم PSD، ویژگی طیف اندازه - فاز (MPS[2]) پیشنهاد شده است. سپس یک ویژگی جدید، یعنی ضرایب اکتاو فاز - اندازه ثابتQ [۳] (CMPOC) ، از MPS استخراج می‌شود. برای این منظور تبدیل ثابت Q[۴] با MPS و تقسیم‌بندی اکتاو و همچنین تبدیل کسینوسی گسسته (DCT[5]) ترکیب می‌شوند. ازآنجاکه ویژگی جدید به‌دست‌آمده از CQT و MPS استخراج

[1] Playback Speech Detection
[2] Magnitude–Phase Spectrum
[3] Constant-Q Magnitude–Phase Octave Coefficients
[4] Constant-Q Transform
[5] Discrete Cosine Transform

می‌شود، به‌عنوان CMPOC نام‌گذاری شده است.

پایگاه‌داده‌ای که در آزمایش مورداستفاده قرار می‌گیرد، مجموعه Asvspoof 2017 است که با استفاده از ۱۵ دستگاه پخش و ۱۶ دستگاه ضبط در چهار مکان به دست می‌آید. این پایگاه‌داده از سه قسمت تشکیل شده است: آموزش، توسعه[1] و ارزیابی[2] مجموعه. جدول (۲-۵) جزئیات مربوط به Asvspoof 2017 را نشان می‌دهد.

باتوجه‌به قوانین چالش Asvspoof 2017، دو نوع مدل آموزش وجود دارد. مدل اول برای ارزیابی الگوریتم پیشنهاد شده توسط عملکرد مجموعه ارزیابی، استفاده می‌شود. در این مدل، ۴۷۲۶ نطق از مجموعه آموزش و مجموعه توسعه استفاده می‌شود. مدل دوم برای ارزیابی الگوریتم پیشنهاد شده توسط عملکرد مجموعه توسعه، استفاده می‌شود. در این مدل، ۳۰۱۶ گفتار از مجموعه آموزش استفاده می‌شود. در این مجموعه نرخ خطای معادل (EER[3]) به‌عنوان معیار ارزیابی استفاده می‌شود.

در CQT، چندین پارامتر مهم برای عملکرد نهایی اهمیت دارد. این پارامترها شامل تعداد سطوح در هر اکتاو، محدوده اکتاو، دوره نمونه‌برداری و گاما است. تعداد سطوح در هر اکتاو ۹۶ تنظیم شده است و همچنین تعداد اکتاوها ۹، دوره نمونه‌برداری ۱۶ و گاما ۳.۳۰۲۶ تنظیم شده است. در این کار از ویژگی‌های پویا استفاده شده است.D و A به ترتیب بیانگر دلتا و شتاب (دلتا دلتا) هستند.

جدول ۲-۵ جزئیات مجموعه ASVspoof 2017 [۲۵]

Subset	**Number**			
	Speakers	Utterances	Genuine	Spoofed
training	10	3016	1508	1,508
development	8	1710	760	950
evaluation	24	13,306	1298	12,008

در این تحقیق از Toolkit شبکه محاسباتی (CNTK[4]) برای آموزش شبکه عصبی عمیق استفاده شده است که به‌عنوان یک طبقه‌بند در آزمایش‌ها استفاده می‌شود. در آزمایش‌ها مجموعه‌ای از

[1] Development
[2] Evaluation
[3] Equal Error Rate
[4] Computational Network Toolkit

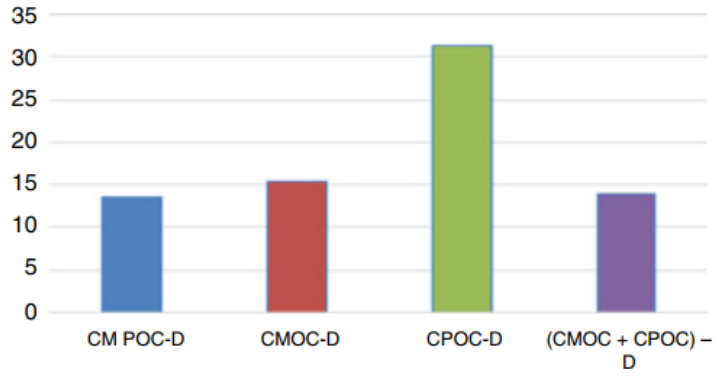
طبقه‌بندهای چهار لایه DNN آموزش داده می‌شوند، هرکدام دارای دولایه پنهان با ۵۱۲ گره در هر لایه و یک‌لایه خروجی با دو گره و یک‌لایه ورودی تشکیل شده از یک پنجره دارای ۱۱ فریم از بردار ویژگی ورودی است. از جدول (۲-۶) مشاهده شده است که CMPOC-D عملکرد بهتری نسبت به CMPOC-A و CMPOC-DA دارد. بنابراین ویژگی CMPOC-D را می‌توان به‌عنوان یک ویژگی برای ارزیابی عملکرد CMPOC در مجموعه ارزیابی ASVSpoof2017 استفاده کرد.

شکل (۲-۴) نتیجه آزمایش را با استفاده از CMPOC-D در مجموعه ارزیابی ASVSpoof2017 را برحسب درصد EER نشان می‌دهد. علاوه بر این، دو ویژگی دیگر برای مقایسه با عملکرد CMPOC استفاده می‌شود. اولین ویژگی ضرایب اکتاو اندازه ثابت Q (CMOC) است که با استفاده از طیف اندازه به دست می‌آید. ویژگی دوم ضرایب اکتاو فاز ثابتQ (CPOC) است که با استفاده از طیف فاز به دست می‌آید.

جدول ۲-۶ نتایج تجربی برحسب (EER٪) در مجموعه ASVspoof 2017 با استفاده از ترکیب ویژگی‌های دینامیکی مختلف CMPOC [۲۵]

Feature	Feature combinations	EER
CMPOC	D	21.22
	A	22.92
	DA	21.56

از شکل (۲-۴) می‌توان چندین نتیجه گرفت. ۱. عملکرد CMPOC-D بسیار بهتر از CMOC-D است. علت این امر می‌تواند این باشد که در MPS اطلاعات بیشتری نسبت به طیف فاز وجود دارد. ۲. عملکرد CMPOC-D بسیار بهتر از CPOC-D است که بدین معنی است که اطلاعات خیلی بیشتری در MPS نسبت به طیف فاز وجود دارد. ۳. CMPOC-D بهتر از ترکیب دو ویژگی (CMOC + CPOC) D- عمل می‌کند، دلیل این امر می‌تواند ناسازگاری اطلاعات در این ترکیب باشد.



شکل ۲-۴ مقایسه نتایج تجربی در مجموعه ارزیابی ASVspoof 2017 با استفاده از CMPOC-D، CMOC-D، CPOC-D و (CMOC + CPOC)-D [۲۵]

همچنین روش پیشنهادی با برخی از ویژگی‌های رایج مانند MFCC و CQCC[1] در ارزیابی ASVspoof2017 مقایسه شده است. آموزش شبکه عمیق طبقه‌بندها برای همه یکسان و به همراه ویژگی دلتا بوده است. همان‌طور که در جدول (۲-۷) مشاهده می‌کنید، عملکرد CMPOC-D بهتر از CQCC-D و خیلی بهتر از MFCC-D است. دلیل این امر می‌توان این باشد که اطلاعات باقدرت تشخیص بالای بیشتری از MPS برای CMPOC استخراج می‌شود.

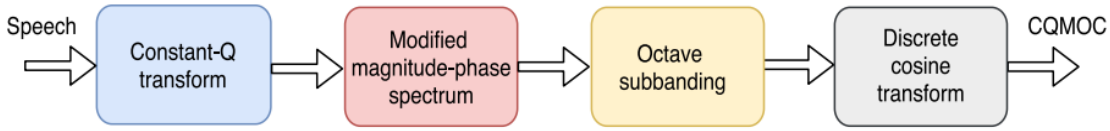
جدول ۲-۷ مقایسه روش پیشنهاد شده با سایر سیستم‌های شناخته شده در ارزیابی ASVspoof 2017 با معیار(%EER) [۲۵]

Features	Classifier	EER
MFCC-D	DNN	27/09
CQCC-D	DNN	14/61
CMPOC-D	DNN	13/55

در طی تحقیقی در [۲۶] به اهمیت استفاده از ویژگی‌های اندازه و فاز به‌صورت هم‌زمان در تشخیص گفتار جعل[2]، پرداخته‌اند. آنها از ویژگی طیف اندازه - فاز اصلاح شده[3] (MMPS) که به‌صورت هم‌زمان از ویژگی اندازه و فاز سیگنال گفتار بهره می‌برد، استفاده می‌کند. ازآنجایی‌که ابتدا از سیگنال گفتار تبدیل ثابت Q (CQT) گرفته می‌شود و سپس از خروجی این تبدیل که در حوزه فرکانس است، ویژگی MMPS را به دست می‌آورند، این ویژگی را CQT-MMPS می‌نامند. ویژگی بعدی که با تبدیل زیر

[1] Constant-Q Cepstral Coefficient
[2] Spoofing Detection
[3] Modified Magnitude-Phase Spectrum

باند اکتاو OST[1] (زیر باند مبتنی بر مقیاس اکتاو) بر روی CQT-MMPS حاصل می‌شود را SQMOC[2] می‌نامند. به‌عبارت‌دیگر، هر سطح فرکانسی پهنای باند متفاوتی دارد و پهنای باند اکتاو اولی نصف پهنای باند اکتاو دومی است. بلوک دیاگرام این دو ویژگی در شکل (۲-۵) آورده شده است.



شکل ۲-۵ روند نمای استخراج ضرایب اکتاو اصلاح شده با ثابت Q (CQMOC) [۲۶]

برای ارزیابی روش پیشنهادی، سه مدل کلاسیک ضد جعل به کار گرفته شده است. مدل اول مدل مخلوط گوسی در سطح فریم (GMM) است که به‌صورت گسترده در تشخیص جعل به کار گرفته می‌شود. در سطح فریم بودن به‌این‌علت گفته می‌شود که با استفاده از فریم‌های گفتار آموزش داده می‌شود. مدل دوم شبکه عصبی کانولوشنال سبک (LCNN[3]) لایه است. این مدل بر خلاف GMM در سطح فریم نیست بلکه در سطح نطق (در اینجا کمتر از ۱۶ فریم) است. مدل سوم مدل شبکه عصبی باقیمانده (ResNet[4]) است که این نیز مانند LCNN در سطح نطق است. در این تحقیق از مدل ResNet18 متشکل از ۸ بلوک باقیمانده {۲، ۲، ۲، ۲} استفاده شده است.

در این تحقیق از پایگاه‌داده ASVspoof 2019 استفاده شده است. این پایگاه‌داده از دو بخش تشکیل شده است. بخش ASVspoof 2019 LA که برای تشخیص گفتار سنتز شده استفاده می‌شود و بخش ASVspoof 2019 PA که برای تشخیص گفتار بازپخش استفاده می‌شود.

پایگاه‌داده ASVspoof 2019 برای چالش ASVspoof 2019 منتشر شده است که به‌اختصار در جدول (۲-۸) آورده شده است. هر دو بخش ASVspoof 2019 LA و ASVspoof 2019 PA شامل سه زیرمجموعه هستند: مجموعه آموزش (Train)، توسعه (Dev) و ارزیابی (Eval). هر دو بخش تعداد گوینده‌های یکسانی دارند. بااین‌حال، مشاهده می‌شود که مقدار داده بخش ASVspoof 2019 PA

[1] Octave Subband Transform
[2] Constant-Q Modified Octave Coefficients
[3] Light Convolutional Neural Network
[4] Residual Neural Network

بزرگ‌تر از بخش ASVspoof 2019 LA است. در این تحقیق طبق طرح ارزیابی ASVspoof 2019، تابع هزینه تشخیص پیاپی (t-DCF[1]) و نرخ خطای برابر (EER) به ترتیب به‌عنوان معیار ارزیابی اولیه و ثانویه استفاده می‌شوند.

برای مدل GMM، پارامترها در CQT بر اساس تحقیق [۲۷] تنظیم می‌شوند. به‌عنوان‌مثال، عدد اکتاو V برابر ۹ و عدد bin فرکانس در هر اکتاو B به‌عنوان ۹۶ تنظیم شده است. در نتیجه، ابعاد استاتیک CQT-MMPS، ۸۶۳ است.

جدول ۲-۸ مشخصات پایگاه‌داده ASVspoof 2019 [۲۶]

Portion	Subset	# Speakers		# Utterances	
		Male	Female	Bonafide	Spoofed
ASVspoof 2019 LA	Train	8	12	2/580	22/800
	Dev	8	12	2/548	22/296
	Eval	30	37	7/355	63/882
ASVspoof 2019 PA	Train	8	12	5/400	48/600
	Dev	8	12	5/400	24/300
	Eval	30	37	18/089	134/630

در مورد OST برای استخراج CQMOC، از همان پارامترهای کار قبلی خودشان [۲۸] استفاده کرده‌اند که در آن P به‌عنوان ۱۲ در نظر گرفته می‌شود؛ بنابراین بعد استاتیک CQMOC، $9\times12=108$ است. آنها همچنین ضرایب دلتا و دلتادلتا را نیز در نظر گرفته‌اند، بنابراین ابعاد ویژگی نهایی $3\times108=324$ است. در آزمایش‌ها، دو مدل GMM از ۵۱۲ جزء مخلوط را با استفاده از نمونه‌های آموزشی واقعی و جعلی، با مشخصات سیستم پایه ASVspoof 2019 آموزش داده می‌شوند.

برای مدل‌های مبتنی بر شبکه عصبی (LCNN و ResNet)، پارامترها در CQT را مطابق کار قبلی خودشان در [۲۹] تنظیم می‌کنند. به‌عنوان‌مثال، V و B به ترتیب ۷ و ۱۲ تنظیم می‌شوند. در نتیجه، بعد استاتیک CQT-MMPS، ۸۴ است. همچنین برای ویژگی CQMOC، P را برابر ۸ در نظر می‌گیرند؛ بنابراین بعد نهایی CQMOC، $7\times8=56$ است.

ازآنجایی‌که طول گفته‌ها متفاوت است، طول تمام گفته‌ها از طریق تکرار ویژگی‌های آنها در هر دسته

[1] Tandem Detection Cost Function

با طول حداکثر، یکسان می‌کنند که به آنها این امکان را می‌دهد که در طول فرایند آموزش به‌صورت موازی پردازش شوند. اما ازآنجایی‌که در مرحله آزمایش، گفته‌های ورودی را یکی‌یکی به شبکه می‌دهند در نتیجه در این مرحله نیازی به یکسان‌سازی طول نیست و می‌توان مرحله آزمایش را با گفته‌ها با طول‌های مختلف اجرا کرد. جدول (۲-۹) نتایج تجربی را بر روی مجموعه ASVspoof 2019 LA نشان می‌دهد.

جدول ۲-۹ عملکرد ویژگی‌های CQT-MMPS و CQMOC با معیار T-DCF و (EER٪) در پایگاه‌داده ASVSPOOF 2019 LA CORPUS [۲۶]

Model	Feature	Development		Evaluation	
		t-DCF	EER	t-DCF	EER
GMM	CQMOC	0/161	4/87	0/199	7/10
LCNN	CQT-MMPS	0/064	2/00	0/176	5/99
	CQMOC	0/118	3/53	0/260	8/59
Res Net	CQT-MMPS	0/050	1/52	0/119	3/72
	CQMOC	0/084	2/51	0/197	6/49

شایان‌ذکر است که GMM برای CQT-MMPS استفاده نمی‌شود زیرا ابعاد بسیار بالایی دارد (۸۶۳). در این کار، GMM عمدتاً برای ارزیابی عملکرد ویژگی‌های CQMOC استفاده می‌شود. از جدول (۲-۹) مشاهده می‌شود که سیستم‌های مبتنی بر (CQT-MMPS) در هر دو مدل (LCNN و ResNet)، به طور قابل‌توجهی بهتر از سیستم‌های مبتنی بر CQMOC عمل می‌کنند، دلیل این امر این است که ویژگی CQMOC از CQT-MMPS استخراج شده است که می‌تواند منجر به ازدست‌رفتن برخی از اطلاعات شود و باعث کاهش عملکرد در تشخیص جعل (SD) شود.

در جدول (۲-۱۰) مقایسه عملکرد روش‌های پیشنهادی با ویژگی‌های مبتنی بر اندازه صورت‌گرفته است. ویژگی‌های مبتنی بر اندازه که در این بخش استفاده شده از لگاریتم طیف توان (LPS[1]) به دست می‌آیند. ویژگی CQT-LPS که از تبدیل Q لگاریتم طیف توان به دست می‌آید و همچنین ویژگی CQ-OST که از ویژگی CQT-LPS مشتق می‌شود.

[1] Log Power Spectrum

جدول ۲-۱۰ مقایسه عملکرد ویژگی‌های مبتنی بر LPS، MPS، و MMPS با معیار T-DCF و (EER٪) در مجموعه ارزیابی ASVSPOOF 2019 LA [۲۶]

Type	Model	Feature	t-DCF	EER
Frame-level	GMM	CQ-OST	**0.196**	**6.81**
		CQMOC	0.199	7.10
Utterance - level	LCNN	CQT-LPS	0.215	6.73
		CQT-MPS	0.215	7.18
		CQT-MMPS	**0.176**	**5.99**
		CQ-OST	0.270	8.78
		CQMOC	0.260	8.59
	ResNet	CQT-LPS	0.122	4.04
		CQT-MPS	0.123	4.03
		CQT-MMPS	**0.119**	**3.72**
		CQ-OST	0.198	6.64
		CQMOC	0.197	6.49

همان‌طور که در جدول (۲-۱۰) مشاهده می‌کنید، CQT-MPS می‌تواند عملکرد قابل‌مقایسه با CQT-LPS به دست آورد، درحالی‌که CQT-MMPS از هر دوی آنها به طور مداوم از نظر t-DCF و EER بهتر عمل می‌کند.

جدول (۲-۱۱) مقایسه بین ویژگی‌های پیشنهادی و برخی دیگر از ویژگی‌های رایج مورداستفاده را نشان می‌دهد. این ویژگی‌ها عبارت‌اند از: MFCC، ضرایب کپسترال فرکانس خطی (LFCC[1])، ضرایب کپسترال فرکانس لحظه‌ای (IFCC[2])، CQCC، (CQSPIC[3])، CosPhase، APGDF، و MODGDF. برای ارزیابی عملکرد این ویژگی‌ها از مدل GMM استفاده شده است.

جدول ۲-۱۱ مقایسه عملکرد ویژگی پیشنهادی با معیار T-DCF و (EER٪) با برخی از ویژگی‌های رایج در مجموعه ارزیابی ASVSPOOF 2019 LA [۲۶]

Feature	t-DCF	EER	Feature	t-DCF	EER
MFCC	0.238	8.71	CQSPIC	0.207	7.35
LFCC	0.227	8.32	CosPhase	0.541	23.98
IFCC	0.298	10.40	APGDF	0.164	6.09
CQCC	0.237	9.57	MODGDF	0.259	11.49
CQMOC	0.199	7.10			

از جدول (۲-۱۱)، می‌توان دریافت که ویژگی پیشنهادی CQMOC از بیشتر ویژگی‌های رایج به جز APGDF بهتر عمل می‌کند. این کارایی ویژگی CQMOC مبتنی بر MMPS را برای گرفتن اطلاعات

[1] Linear Frequency Cepstral Coefficient
[2] Instantaneous Frequency Cepstral Coefficients
[3] Constant-Q Statistics-Plus-Principal Information Coefficients

اندازه و فاز برای تشخیص گفتار سنتز شده را نشان می‌دهد.

جدول (۲-۱۲) سیستم‌های پیشنهادی را با برخی از سیستم‌های شناخته شده در مجموعه ارزیابی ASVspoof 2019 LA مقایسه می‌کند. ZTWCC[1]-GMM مخفف ضرایب کپسترال پنجره زمانی صفر با یک طبقه‌بند GMM است، FFT-LCNN سیستم شبکه عصبی مبتنی بر LCNN را با تبدیل فوریه سریع (FFT) به‌عنوان ورودی نشان می‌دهد. براین‌اساس، LFCC-LCNN و LCNN-(LFCC-CMVN) سیستم‌های شبکه عصبی مبتنی بر LCNN را با استفاده از LFCC و LFCC به ترتیب با میانگین کپسترال و نرمال‌سازی واریانس (CMVN[2]) به‌عنوان ورودی نشان می‌دهند. به طور مشابه، MFCC-ResNet، Spec-ResNet و CQCC-ResNet سیستم‌های شبکه عصبی مبتنی بر ResNet را به ترتیب با MFCC، LPS مبتنی بر DFT و CQCC به‌عنوان ورودی نشان می‌دهند.

مشاهده می‌شود که سیستم‌های پیشنهادی عملکرد قابل‌مقایسه یا بهتری با کارهای قبلی دارند. همچنین اطلاعات فاز - اندازه که توسط ویژگی‌های MMPS و CQMOC به کار گرفته می‌شوند، می‌تواند به سیستم‌های ضد جعل کمک کند تا به نتایج بهتری دست یابند. لازم به ذکر است که FFT-LCNN و LFCCLCNN در زیرچالش ASVSpoof 2019 LA بهترین عملکردها را در بین تمام سیستم‌های منفرد از خود نشان می‌دهند. اگرچه (CQT-MMPS) –ResNet از نظر t-DCF کمی بدتر از FFT-LCNN و LFCC-LCNN عمل می‌کند، اما از نظر EER عملکرد بهتری دارد.

جدول (۲-۱۳) مقایسه عملکرد بین ویژگی‌های مبتنی بر LPS، MPS و MMPS را که از CQT مشتق شده‌اند در مجموعه ارزیابی ASVspoof 2019 PA نشان می‌دهد. مشاهده می‌کنیم که سیستم‌های مبتنی بر MMPS می‌توانند بهتر از سیستم‌های مبتنی بر MPS مربوطه در مجموعه ارزیابی ASVspoof 2019 PA عمل کنند، و هر دوی آنها از نظر t-DCF و EER از سیستم‌های مبتنی بر LPS بهتر عمل می‌کنند.

[1] Zero Time Windowing Cepstral Coefficients
[2] Cepstral Mean and Variance Normalization

ویژگی‌های پیشنهادی با استخراج اطلاعات اندازه و فاز، می‌توانند از LPS که فقط از ویژگی اندازه بهره می‌برد، در تشخیص گفتار بازپخش بهتر عمل کنند.

در جدول (۲-۱۴) ویژگی پیشنهادی CQMOC با برخی از ویژگی‌های رایج دیگر مقایسه شده است. به طور مشابه، این ویژگی‌ها عبارت‌اند از: MFCC، LFCC، IFCC، CQCC CQSPIC، CosPhase، APGDF، و MODGDF. همانند مقایسه در بخش قبلی در این قسمت نیز از مدل GMM استفاده شده است.

همان‌طور که در جدول (۲-۱۴) مشاهده می‌شود، ویژگی پیشنهادی CQMOC می‌تواند بسیار بهتر از سایر ویژگی‌ها عمل کند. در نهایت جدول (۲-۱۵) مقایسه بین عملکرد سیستم‌های پیشنهادی در این تحقیق و برخی از سیستم‌های شناخته شده در مجموعه ارزیابی PA را در ASVspoof 2019 نشان می‌دهد.

جدول ۲-۱۲ مقایسه عملکرد ویژگی پیشنهادی با معیار T-DCF و (EER٪) با برخی از ویژگی‌های رایج در مجموعه ارزیابی ASVSPOOF 2019 LA [۲۶]

Type	System	t-DCF	EER
GMM-based	CQCC-GMM	0.237	9.57
	LFCC-GMM	0.212	8.09
	ZTWCC-GMM	0.141	6.13
LCNN-based	LFCC-LCNN	0.100	5.06
	(LFCC-CMVN)-LCNN	0.183	7.86
	FFT-LCNN	0.103	4.53
ResNet-based	MFCC-ResNet	0.204	9.33
	Spec-ResNet	0.274	7.69
	CQCC-ResNet	0.217	7.69
Proposed	CQMOC-GMM	0.199	7.10
	(CQT-MMPS)-LCNN	0.176	5.99
	CQMOC-LCNN	0.260	8.59
	(CQT-MMPS)-ResNet	0.119	3.72
	CQMOC-ResNet	0.197	6.49

باتوجه‌به سیستم‌های مبتنی بر GMM، مشاهده می‌کنیم که ویژگی CQMOC می‌تواند بسیار بهتر از ویژگی‌های CQCC، LFCC و ZTWCC عمل کند. همچنین در صورت استفاده از مدل‌های مبتنی بر بیان، مانند LCNN و ResNet، بهبودهای قابل‌توجهی حاصل می‌شود. همان‌طور که در جدول (۲-۱۵) نشان‌داده‌شده است، سیستم ResNet (CQT-MMPS) به طور مداوم از سایر سیستم‌های منتشر

شده مبتنی بر ResNet بدون افزایش داده (DA[1]) بهتر عمل می‌کند. علاوه بر این، می‌تواند نتیجه‌ای قابل‌مقایسه با سیستم ResNetDA - (GD gram) به دست آورد.

جدول ۲-۱۳ مقایسه عملکرد ویژگی‌های مبتنی بر LPS، MPS، و MMPS در مجموعه ارزیابی PA ASVSPOOF 2019 [۲۶]

Type	Model	Feature	t-DCF	EER
Frame-level	GMM	CQ-OST	0.149	7.28
		CQMOC	0.135	6.95
Utterance - level	LCNN	CQT-LPS	0.040	1.39
		CQT-MPS	0.033	1.22
		CQT-MMPS	0.024	0.90
		CQ-OST	0.037	1.31
		CQMOC	0.036	1.32
	ResNet	CQT-LPS	0.038	1.30
		CQT-MPS	0.035	1.19
		CQT-MMPS	0.031	1.08
		CQ-OST	0.040	1.38
		CQMOC	0.038	1.36

جدول ۲-۱۴ مقایسه عملکرد ویژگی‌های پیشنهادی با برخی از ویژگی‌های رایج مورد استفاده در مجموعه ارزیابی PA ASVSPOOF 2019 [۲۶]

Feature	t-DCF	EER	Feature	t-DCF	EER
MFCC	0/360	14/21	CQSPIC	0/164	7/73
LFCC	0/302	13/54	CosPhase	0/484	19/71
IFCC	0/357	15/85	APGDF	0/328	13/51
CQCC	0/245	11/04	MODGDF	0/242	11/24
CQMOC	0/135	6/95			

در بین تمام سیستم‌های مبتنی بر LCNN، سیستم LCNN-(CQTMMPS) از نظر معیارهای t-DCF (۰.۰۲۴) و EER (۰.۹۰) بهترین عملکرد را دارد. این نتایج نشان می‌دهند که ویژگی‌های CQT-MMPS و CQMOC پیشنهاد شده در این تحقیق که مبتنی بر اندازه و فاز هستند، برای تشخیص گفتار بازپخش در GMM و همچنین مدل‌های LCNN و ResNet مبتنی بر شبکه عصبی، بسیار مؤثر هستند.

[1] Data Augmentation

جدول ۲-۱۵ مقایسه عملکرد در میان سیستم‌های پیشنهادی و برخی از سیستم‌های شناخته شده در مجموعه ارزیابی ASVSPOOF 2019 PA [۲۶]

Type	System	t-DCF	EER
	CQCC-GMM	0/245	11/04
GMM-based	LFCC-GMM	0/302	13/54
	ZTWCC-GMM	0/281	1220
LCNN-based	DCT-LCNN	0/56	206
	LFCC-LCNN	0/105	4/60
	CQT-LCNN	0/030	1/23
ResNet-based	CQCC-ResNet	0/107	4/43
	Spec-ResNet	0/099	3/81
	(GD gram)-ResNet	0/044	1/79
	(GD gram)-ResNet-DA	0/028	1/08
Proposed	CQMOC-GMM	0/135	6/95
	(CQT-MMPS)-LCNN	0/024	0/90
	CQMOC-LCNN	0/036	1/32
	(CQT-MMPS)-ResNet	0/031	1/08
	CQMOC-ResNet	0/038	1/36

در [۳۰] از ترکیب ویژگی‌های مبتنی بر اندازه و فاز در تأیید گوینده (SV[1]) استفاده کرده‌اند. ویژگی مبتنی بر اندازه، ضرایب کپسترال فرکانس مل (MFCC) و ویژگی مبتنی بر فاز، تابع تأخیر گروهی اصلاح شده (MOGDF) است. برای ارزیابی عملکرد روش‌های پیشنهادی از پایگاه‌داده TIMIT استفاده شده است. این پایگاه‌داده شامل گفتار بدون نویز ۶۳۰ گوینده با فرکانس ۱۶ کیلوهرتز است. هر فایل گفتاری در ۲۰ میلی‌ثانیه با ۵۰ درصد هم‌پوشانی فریم‌بندی می‌شود.

۲۰ گوینده از پایگاه‌داده انتخاب شده است. آزمایش ابتدا با استخراج ویژگی‌ها با استفاده از MFCC و MODGDF انجام می‌شود.MFCC بردار ویژگی ۱۳ بعدی را ارائه می‌دهد. همچنین با بهره‌گیری از دلتا و دلتا - دلتا، یک بردار ویژگی ۳۹ بعدی ایجاد می‌شود. تابع تأخیر گروهی اصلاح شده بردار ویژگی را بر اساس مقادیر FFT و ضرایب لیفتر [۲]به دست می‌آورد. برای مدل‌سازی از یک شبکه عصبی با دولایه که هر لایه دارای ۳۰ نرون است، استفاده شده است.

برای آموزش این شبکه چندلایه از الگوریتم پیش‌خور پس انتشار[۳] استفاده می‌شود. ۷۰ درصد از داده‌ها برای آموزش شبکه استفاده می‌شود درحالی‌که ۱۵ درصد برای آزمایش شبکه و ۱۵ درصد برای

[1] Speaker Verification
[2] Lifter
[3] Feed Forward Back Propagation

اعتبارسنجی شبکه استفاده می‌شود. خروجی‌های شبکه عصبی برای هر دو ویژگی با استفاده از چند استراتژی ترکیب می‌شوند.

استراتژی اول برای ترکیب ویژگی‌ها استفاده از عملگر OR است که در رابطه (۲-۱) نشان‌داده‌شده است. در این استراتژی، بردار ترکیب دو ویژگی، با گرد کردن مقادیر و اعمال عملیات OR بر روی دو بردار ویژگی به دست می‌آید.

$$Y_{OUT} = X_{MFCC} + X_{MODGDF} \qquad ۱-۲$$

استراتژی دوم برای ترکیب ویژگی‌ها در رابطه (۲-۲) نشان‌داده‌شده است. در این استراتژی بردارهای ویژگی قبل از اعمال عملگر OR، وزن‌دهی می‌شوند. ویژگی که به‌صورت جداگانه عملکرد بهتری دارد با وزن بیشتر و ویژگی دیگر با وزن کمتر ضرب شده است. در این تحقیق ویژگی MFCC در ۰.۸ و بردار ویژگی MODGDF در ۰.۲ ضرب شده است. شایان‌ذکر است که پارامتر β بین ۰ و ۱ قرار دارد.

$$Y_{OUT} = \beta^* X_{MFCC} + (1-\beta) * X_{MODGDF} \qquad ۲-۲$$

استراتژی سوم برای ترکیب ویژگی‌ها به‌صورت زیر است.

۱. خروجی‌های شبکه عصبی را برای هر دو بردار ویژگی به دست آورده می‌شود.

۲. مقادیر متناظر هر دو بردار ویژگی با هم مقایسه می‌شوند.

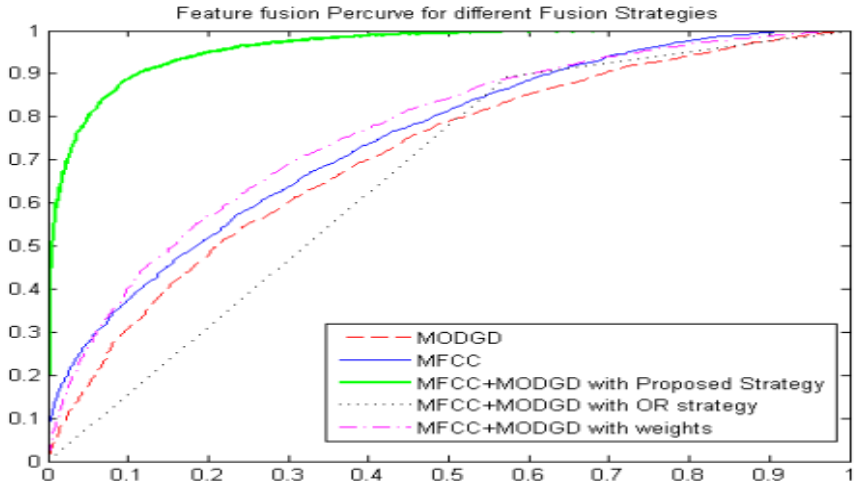
۳. مقدار بزرگ‌تر از این دو انتخاب شده و یک بردار جدید تشکیل می‌شود.

در شکل (۲-۶) منحنی‌های عملکرد در استراتژی‌های مختلف ترکیب به همراه عملکرد فردی MFCC و MODGDF رسم شده است. ناحیه زیر منحنی (ROC[1]) به‌عنوان معیار سنجش عملکرد، انتخاب شده است که با عملکرد ویژگی رابطه مستقیم دارد. به‌عبارت‌دیگر هر چه ناحیه زیر منحنی بزرگ‌تر باشد عملکرد آن ویژگی بهتر است. ROC یک منحنی عملکرد است که نمودار نرخ مثبت صحیح[2] در مقابل نرخ مثبت کاذب[3] است. اگر سطح زیر ROC به‌دست‌آمده تقریباً واحد باشد، عملکرد

[1] Receiver Operating Characteristic
[2] True Positive
[3] False Positive

سیستم تقریباً بهینه است. همان‌طور که در جدول (۲-۱۶) مشاهده می‌کنید استراتژی OR از ادغام بردارهای ویژگی منجر به عملکردی کمتر از ویژگی‌های تکی می‌شود.

اختصاص وزن به بردارهای ویژگی عملکرد را تا حدودی بهبود می‌بخشد. استراتژی سوم برای ترکیب ویژگی‌ها، بهترین عملکرد را در بین استراتژی‌های اتخاذ شده و همچنین از عملکرد تکی ویژگی‌ها، ارائه می‌دهد. همان‌طور که مشاهده کردیم ترکیب ویژگی مبتنی بر اندازه با ویژگی مبتنی بر فاز با اتخاذ استراتژی مناسب، عملکرد سیستم را بهبود می‌بخشد که اهمیت به‌کارگیری ویژگی فاز را نشان می‌دهد.



شکل ۲-۶ ROC برای استراتژی های مختلف ترکیب ویژگی‌ها [۳۰]

جدول ۲-۱۶ ارزیابی عملکرد استراتژی های مختلف [۳۰]

Feature	Area under curve
MFCC	0.7470
MODGDF	0.7089
MFCC+MODGDF WITH PROPOSED FUSION STRATEGY	0.9611
MFCC+MODGDF WITH OR STRATEGY	0.7360
MFCC+MODGDF WITH WEIGHT STRATEGY	0.7631

فصل سوم

مبانی نظری پردازش

سیگنال مبتنی بر فاز

۳-۱- مقدمه

در این فصل جنبه‌های اساسی پردازش سیگنال مبتنی بر فاز و برخی از ویژگی متداول جهت استخراج ویژگی بررسی می‌شوند. ابتدا تعریف طیف فاز و مشکلات مربوط به استفاده از طیف فاز در پردازش سیگنال گفتار بررسی می‌شود. سپس تأخیر گروهی به‌عنوان ویژگی اصلی طیف فاز، با جزئیات مورد بررسی قرار می‌گیرد. تعریف، خواص آن، رابطه آن با طیف اندازه، مسئله تیزی[1] و راه‌حل‌های پیشنهاد شده توسط محققین جهت برطرف‌کردن این موضوع پوشش داده شده است. در نهایت، استخراج ویژگی و برخی از روش‌های متداول در استخراج ویژگی بررسی می‌شود.

۳-۲- نقش کم‌رنگ طیف فاز در پردازش گفتار

تحلیل فوریه نقش کلیدی در پردازش سیگنال گفتار ایفا می‌کند. به‌عنوان یک کمیت مختلط، می‌توان آن را به شکل قطبی با استفاده از طیف اندازه و فاز بیان کرد. طیف اندازه تقریباً در هر گوشه‌ای از پردازش گفتار به طور گسترده استفاده می‌شود. درصورتی‌که طیف فاز نقطه شروع جذابی برای پردازش سیگنال گفتار نیست. اکثر الگوریتم‌های موجود در این زمینه، طیف اندازه را در مرکز توجه قرار می‌دهند و طیف فاز را بدون هیچ پردازشی مستقیماً به خروجی منتقل می‌کنند (به‌عنوان‌مثال در تقویت گفتار) یا کاملاً آن را کنار می‌گذارند. به‌عنوان‌مثال در سیستم‌های ASR، از ویژگی MFCC که رایج‌ترین روش استخراج ویژگی در پردازش گفتار است، استفاده می‌کنند. سه دلیل اصلی وجود داشت که تمایل محققان را در استفاده از طیف فاز کاهش می‌داد:

- جبهه‌گیری‌های تاریخی علیه فاز که فاز را فاقد اطلاعات مهم ادراکی می‌دانند
- ویژگی‌های مبتنی بر فاز فقط در تجزیه‌وتحلیل سیگنال گفتار زمان - بلند اهمیت پیدا می‌کند، درحالی‌که به دلیل غیرایستا بودن سیگنال گفتار، باید در کوتاه‌مدت پردازش شود.
- شکل پیچیده به دلیل پیچش فاز که مانع از تفسیر فیزیکی و مدل‌سازی ریاضی می‌شود

[1] Spikiness

۳-۲-۱- جبهه‌گیری تاریخی علیه فاز

جبهه‌گیری تاریخی به اواسط قرن نوزدهم برمی‌گردد و از قانون آکوستیک اهم نشئت می‌گیرد که بیان می‌کند گوش انسان به‌عنوان یک تحلیلگر طیفی عمل می‌کند و طیف فاز تأثیری بر نحوه درک صدا توسط گوش، ندارد. به‌عبارت‌دیگر، سیستم شنوایی انسان فاز ناشنوا است. این قانون در ابتدا مورد استقبال قرار نگرفت (۱۸۴۳) و بحث‌های تندی بین سیبک و اهم در گرفت. سیبک قانون فاز اهم را بی‌اعتبار دانست و او را مجبور کرد که آن را پس بگیرد. بااین‌حال، در سال ۱۸۶۳ هلمهولتز با اهم هم‌عقیده شد و جامعه را متقاعد کرد که قانون آکوستیک اهم درست است. سیبک در سال ۱۸۴۹ درگذشت، اما مطالعات بعدی در قرن ۲۰ توسط شوتن عقیده سیبک را تأیید کرد. بااین‌حال، کار اهم و هلمهولتز منجر به جبهه‌گیری در برابر سودمندی فاز از لحاظ حمل اطلاعات مهم ادراکی شد [۳۱].

۳-۲-۲- اهمیت فاز در پردازش بلندمدت گفتار

در[۲] مشاهده کردند که وقتی سیگنال گفتار به فریم‌های طولانی تجزیه می‌شود (مثلاً به مدت ۱ ثانیه)، طیف فاز گفتار اهمیت پیدا می‌کند، به این دلیل که گفتار بازسازی‌شده توسط فاز به‌تنهایی، قابل‌درک می‌شود. این عقیده در[۳۲] مورد تأیید قرار گرفت و نشان‌داده‌شد که با افزایش طول فریم، اهمیت طیف فاز افزایش می‌یابد. اگرچه، بر اساس کارهای فعلی، برای تجزیه‌وتحلیل سیگنال‌های غیرایستا (مانند گفتار انسان)، قبل از تجزیه‌وتحلیل، آنها باید به فریم‌های کوتاهی که در آن ایستایی وجود دارد، تجزیه شوند. برای سیگنال گفتار، طول فریم معمولی که در آن ویژگی‌های گفتار ثابت است، در محدوده ۲۰-۴۰ میلی‌ثانیه است. این بسیار کمتر از فریم‌های بلندی است که در آنها طیف فاز اهمیت خاصی پیدا می‌کند.

۳-۲-۳- تفسیر فیزیکی و مدل‌سازی ریاضی

خواص سیستم تولید گفتار انسان از نظر فیزیولوژیکی و فیزیکی تا حد زیادی واضح است. به‌عنوان‌مثال، برای صداهای واک‌دار، تارهای صوتی به ارتعاش درمی‌آیند و این منجر به یک رفتار شبه

تناوبی در حوزه زمان می‌شود که با یک دوره تناوب اصلی مشخص می‌شود که مفهوم معینی در حوزه فرکانس دارد. همچنین، مجرای گفتار را می‌توان به‌عنوان مجموعه‌ای از لوله‌ها با مقطع‌های مختلف مدل کرد. بر اساس قوانین فیزیک، چنین سیستمی دارای برخی فرکانس‌های تشدید است. این خواص را می‌توان به‌راحتی در طیف اندازه مشاهده کرد. اما به دلیل پدیده پیچش فاز، طیف فاز شکلی شبیه نویز به خود می‌گیرد که مانع از تفسیر فیزیکی و مدل‌سازی ریاضی می‌شود.

۳-۲-۴- عملکرد طیف اندازه

طیف اندازه از هیچ یک از مشکلات ذکر شده را ندارد. ازنقطه‌نظر ادراکی، اتفاق‌نظر وجود دارد که حامل اطلاعاتی است که بسیاری از جنبه‌های سیگنال گفتار را مشخص می‌کند. ازنقطه‌نظر فیزیکی، به‌خوبی با درک ما از سیستم تولید گفتار مطابقت دارد. علاوه بر این، با استفاده از این بخش از تبدیل فوریه به‌راحتی می‌توان فرمنت‌ها و فرکانس اصلی (فرکانس پیچ) را شناسایی و تخمین زد. از نظر ریاضی، اکسترمم[1]ها و روند آنها تفسیر روشنی دارد و می‌توان به طور مؤثر مدل‌سازی کرد.

۳-۳- تجزیه‌وتحلیل سیگنال با استفاده از تبدیل فوریه

تحلیل فوریه یکی از مهم‌ترین ابزارهای ریاضی است که به طور گسترده در پردازش سیگنال به کار گرفته شده است. این تبدیل سیگنال را از حوزه زمان به حوزه فرکانس می‌برد و با استفاده از نمایی مختلط به‌عنوان توابع پایه، نمایش جدیدی را ارائه می‌دهد. این عملیات خطی به طور یکتا معکوس‌پذیر است، بنابراین هیچ اطلاعاتی پس از اعمال آن از بین نمی‌رود. در بسیاری از کاربردها، این نمایش، اطلاعات مطلوب و مفیدی که در سیگنال اصلی پنهان شده است را بهتر برجسته و متمایز می‌کند و روشی مؤثر برای تحلیل سیگنال ارائه می‌دهد. تبدیل فوریه زمان گسسته، F، به‌صورت زیر تعریف می‌شود [۳۳].

$$F\{x[n]\} = X(\omega) = \sum_{n=0}^{N-1} x[n] e^{-j\omega n} \qquad ۱-۳$$

[1] Extremum

که در آن n، N و ω به ترتیب متغیرهای زمان، تعداد نمونه‌ها و فرکانس شعاعی را نشان می‌دهند و X (ω) تبدیل فوریه سیگنال (زمان گسسته) x[n] است. نمایی‌های مختلط ($e^{-j\omega n}$) توابع پایه این تبدیل را تشکیل می دهند و X (ω) را به یک کمیت مختلط تبدیل می‌کنند. اعداد مختلط را می‌توان در مختصات دکارتی بصورت زیر بیان کرد.

$$X(\omega) = X_{Re}(\omega) + jX_{Im}(\omega) \qquad ۲-۳$$

که در آن $X_{\mathrm{Re}(\omega)}$ و $X_{\mathrm{Im}(\omega)}$ به ترتیب قسمت حقیقی[1] و موهومی[2] را نشان می دهند. اعداد مختلط نیز به شکل قطبی به صورت زیر قابل نمایش هستند.

$$X(\omega) = |X(\omega)|\, e^{\phi_X(\omega)} \qquad ۳-۳$$

که در آن $|X(\omega)|$ و $\phi_{X(\omega)}$ به ترتیب اندازه و طیف فاز هستند و به صورت زیر تعریف می‌شوند.

$$|X(\omega)| = \sqrt{X_{Re}^2(\omega) + X_{Im}^2(\omega)} \qquad ۴-۳$$

$$\phi_X(\omega) = \arctan\left(\frac{X_{Im}(\omega)}{X_{Re}(\omega)}\right) \qquad ۵-۳$$

به طور دقیق‌تر دو نوع معکوس تابع تانژانت (arctan)، وجود دارد، یعنی ۲ ربعی و ۴ ربعی. اولی به‌طورکلی به‌صورت atan (.) شناخته می‌شود که یک ورودی می‌گیرد و خروجی آن محدود به $(-\frac{\pi}{2}, \frac{\pi}{2})$ است، در حالی که دومی با atan2 (.،.) نشان داده می‌شود که دارای دو ورودی (قسمت های حقیقی و موهومی) است که خروجی در محدوده $(-\pi, \pi]$ پیچش می‌کند. خروجی atan2 طیف فاز اصلی (ARG{ X(ω) }) نامیده می‌شود. به این ترتیب، فرم دقیقتر رابطه (۳-۵) بصورت زیر خواهد بود.

$$ARG\{X(\omega)\} = \operatorname{atan2}\left(X_{Im}(\omega), X_{Re}(\omega)\right) = \operatorname{atan2}\left(\frac{X_{Im}(\omega)}{X_{Re}(\omega)}\right) \qquad ۶-۳$$

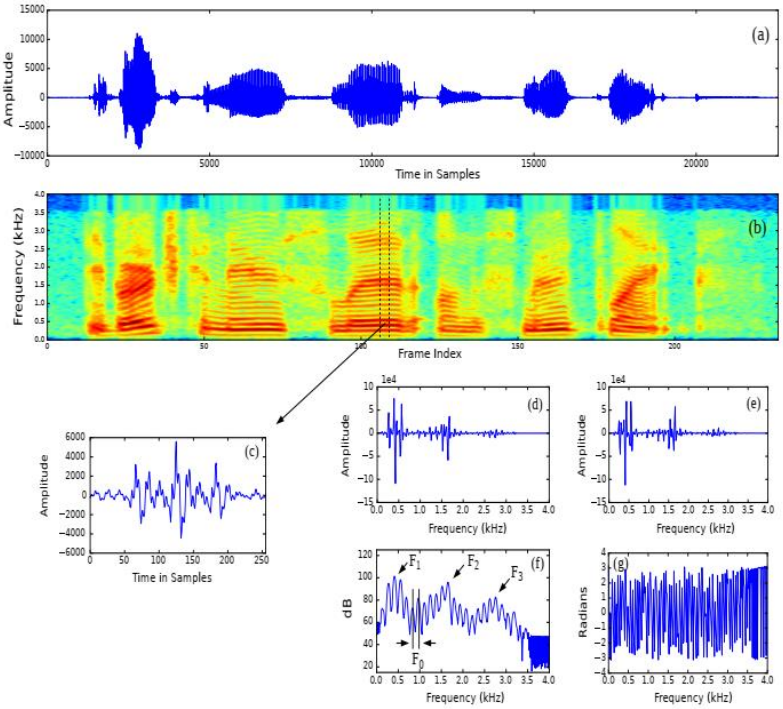
در محاسبه atan (.)، فقط تقسیم $\frac{X_{\mathrm{Im}(\omega)}}{X_{\mathrm{Re}(\omega)}}$ در نظر گرفته می‌شود و اگر این مقدار منفی شود،

[1] Real
[2] Imaginary

مشخص نیست که قسمت حقیقی یا موهومی منفی بوده است. اما از آنجایی که (.,.) atan2 دارای دو ورودی است، ربع مربوطه را می‌توان بدون ابهام تعیین کرد.

۳-۳-۱- مزیت طیف اندازه

شکل (۳-۱) شکل موج و اسپکتروگرام (طیف‌نگار) یک سیگنال گفتار را به همراه بخش‌های حقیقی و موهومی و همچنین طیف‌های اندازه و فاز نشان می‌دهد.



شکل ۳-۱ مقایسه نمایش سیگنال با استفاده از بخش های مختلف تبدیل فوریه (a) سیگنال گفتار جمله (" We find joy in the simplest things "، fs = 8000 هرتز)، (b) اسپکتوگرام، (c) شکل موج قطعه ای به طول ۳۲ میلی ثانیه، (d) بخش حقیقی از STFT، (e) بخش موهومی STFT،(f) طیف اندازه، (g) طیف فاز[۳۱]

همان‌طور که مشاهده می‌شود، در میان بخش‌های تبدیل فوریه، طیف اندازه، سیگنال گفتار را بهتر مشخص می‌کند و با درک ما از سیستم تولید گفتار انسانی مطابقت دارد. این سیستم را می‌توان به‌عنوان اتصال تعدادی لوله با سطح مقطع‌های مختلف در نظر گرفت. بر اساس قانون فیزیک، چنین سیستمی دارای فرکانس‌های رزونانسی است. این فرکانس‌ها در پردازش گفتار، فرمنت نامیده می‌شوند که با F1، F2، F3 و... نشان داده می‌شوند. علاوه بر این، برای صداهای واک‌دار، یک حرکت شبه تناوبی در تارهای

صوتی وجود دارد که منجر به شبه تناوبی در حوزه زمانی می‌شود. این تناوب مربوط به فرکانس پایه (پیچ یا گام) است که اغلب با F0 نشان داده می‌شود. فرکانس پایه و فرمنت‌ها حاوی اطلاعات مفیدی در مورد اجزای تحریک[1] و مجرای صوتی[2] (گفتار) سیگنال گفتار هستند. همان‌طور که در شکل (۳-۱) نشان‌داده‌شده است، آنها در حوزه طیف اندازه به‌خوبی برجسته و متمایز شده‌اند؛ بنابراین، به نظر می‌رسد که طیف اندازه یک انتخاب طبیعی برای پردازش اطلاعات رمزگشایی شده در سیگنال گفتار باشد [۳۱].

یکی دیگر از مزایای طیف اندازه این است که نقاط مینیمم و ماکزیمم محلی آن تفسیر دارند. اگر به تبدیل z یک سیگنال به‌صورت تحلیلی بنگریم، مینیمم‌های محلی به صفرها و ماکزیمم‌های محلی به قطب‌ها مربوط می‌شوند. قطب‌ها و صفرها در توصیف سیگنال (سیستم) مربوطه و مطالعه ویژگی‌های آن مفید هستند که این مورد، پردازش سیگنال مبتنی بر اندازه را تسهیل می‌کند.

به طور خلاصه، رفتار قابل‌درک و بیان طیف اندازه، امکان استفاده از آن را در طیف گسترده‌ای از کاربردها در پردازش گفتار و سیگنال فراهم می‌کند.

۳-۳-۲- مزیت طیف فاز

همان‌طور که در شکل (۳-۱) نشان‌داده‌شده است، طیف فاز به شفافیت طیف اندازه رفتار نمی‌کند. ظاهراً شبیه نویز، بدون اطلاعات و بدون نقاط اکسترممی است که تفسیر فیزیکی یا مدل‌سازی ریاضی را تسهیل کند. چنین ساختار مبهم، سؤال اساسی را مطرح می‌کند: آیا اصلاً اطلاعاتی در این فاز وجود دارد؟ اگر اطلاعاتی در طیف فاز وجود نداشته باشد، کار بر روی آن ارزشی ندارد [۳۱].

لازم به ذکر است که برای بازیابی منحصربه‌فرد سیگنال در حوزه زمان، هر دو طیف اندازه و فاز موردنیاز است. این بدان معناست که طیف فاز حاوی برخی اطلاعات است. بااین‌حال، مقدار، نوع و اهمیت اطلاعات کدگذاری شده در طیف فاز نامشخص است.

[1] Excitation
[2] Vocal Tract

۳-۳-۳- پیچش فاز

دلیل اصلی شکل نامنظم طیف فاز، پدیده پیچش فاز است که از تابع تانژانت معکوس چهار ربعی ناشی می‌شود که در محاسبه فاز اصلی استفاده می‌شود. پیچش، امکان نمایش یک مقدار را زمانی که از محدوده مجاز فراتر رود، فراهم می‌کند. راه‌حل دیگر برای حفظ مقادیر در یک محدوده مشخص، برش دادن مقادیر زیر یا بالاتر از محدوده مجاز است. ضعف برش این است که برگشت‌ناپذیر است و مقادیر اصلی دیگر قابل بازیابی نیستند. پیچش، مکانیسمی است که مقادیر را در محدوده‌های داده شده نگه می‌دارد و درعین‌حال امکان بازیابی مقادیر اصلی را از طریق واپیچش[1] می‌دهد. با فرض اینکه محدوده معتبر $(L_{\min}, L_{\max})$و $\Delta L = L_{\max} - L_{\min}$ باشد، فرآیند پیچش به صورت زیر عمل می‌کند: اگر نمونه خارج از محدوده داده شده بود $m\Delta L$ را اضافه می کنیم تا نمونه $+ m\Delta L$ در محدوده $(L_{\min}, L_{\max})$قرار گیرد (mیک عدد صحیح است).

طیف فاز اصلی بین $(-\pi, \pi]$پیچیده شده است و این باعث ایجاد شکل آشفته آن با ناپیوستگی‌های فراوان می‌شود. از عواقب این ناپیوستگی‌ها و همچنین شکل آشفته آن، این است که مانع محاسبه مشتق فاز می‌شود. فرآیند معکوس، واپیچش فاز نامیده می‌شود که در آن هدف یافتن یک نمایش پیوسته برای طیف فاز است. طیف فاز واپیچی شده یا پیوسته با (arg{X(ω}) نشان داده می‌شود. مشکل اصلی در واپیچی فاز این است که افزودن هر مضرب صحیح 2π به فاز اصلی، عدد مختلط $X(\omega)$ را تغییر نمی‌دهد.

$$X(\omega) = |X(\omega)| e^{j\phi_X(\omega)]} = |X(\omega)| e^{j(\phi_X(\omega)+2m\pi)} \qquad ۷-۳$$

بنابراین، گزینه‌های ممکن زیادی برای m وجود دارد، اما فقط یک m صحیح و فاز پیوسته وجود دارد. از نظر ریاضی، $\arg\{X(\omega)\} = ARG\{X(\omega)\} + 2\, m(\omega)\pi$ و فرایند واپیچش یافتن مقدار مناسب برای $m(\omega)$ است .

ازنظر کاربرد، واپیچی فاز به‌ویژه در پردازش تصویر اهمیت دارد و بخشی جدایی‌ناپذیر از روش‌هایی

[1] Unwrapping

مانند تصویربرداری تشدید مغناطیسی (MRI[1])، رادار دیافراگم مصنوعی (SAR[2]) و پردازش الگوی حاشیه‌ای است. به همین دلیل بیشتر تحقیقات انجام شده در مورد موضوع پیچش فاز در زمینه پردازش تصویر است[۳۴].

۳-٤- واپیچش فاز

برای واپیچی فاز می‌توان از مشتق طیف فاز استفاده کرد که بدون واپیچی طیف فاز، بر اساس فرمول زیر قابل محاسبه است:

$$\arg'\{X(\Omega)\} = \frac{d\arg\{X(\omega)\}}{d\omega} = \frac{X_{Re}(\omega)X'_{Im}(\omega) - X'_{Re}(\omega)X_{Im}(\omega)}{|X(\omega)|^2}$$
$$\Rightarrow \begin{cases} \arg\{X(\omega)\} = \int_0^{\omega} \arg'\{X(\Omega)\}d\Omega \\ \arg\{X(\omega)\}\big|_{\omega=0} = 0 \end{cases} \qquad ۳-۸$$

که در آن ′ نشان‌دهنده مشتق نسبت به ω است. در عمل، ازآنجایی‌که متغیرها گسسته هستند، مشتق و انتگرال به‌صورت تقریبی به ترتیب با استفاده از تفاضل و جمع نمونه انجام می‌شود. این تقریب‌ها خطای مرتبط با این رویکرد را افزایش می‌دهد. به همین جهت محققان روش‌های مختلفی برای واپیچی فاز ارائه کرده‌اند. یک روش محبوب برای واپیچی، بر اساس تفاوت بین دو نمونه متوالی است [۳۳].دستور unwrap در کتابخانه Matlab این الگوریتم را پیاده‌سازی می‌کند. در این رویکرد، تفاوت بین نمونه‌های فاز اصلی مجاور محاسبه می‌شود. اگر قدر مطلق اختلاف بیش از یک آستانه (معمولاً π) شود، 2π اضافه یا کم می‌شود به‌طوری‌که اختلاف بین نمونه‌های مجاور کمتر از π شود. این تکنیک به‌عنوان تشخیص ناپیوستگی (DD[3]) شناخته می‌شود. به‌طورکلی، برای مسئله واپیچی فاز، هیچ راه‌حل جهانی و نهایی وجود ندارد و تمام روش‌های پیشنهادی ابتکاری هستند.

[1] Magnetic Resonance Imaging
[2] Synthetic Aperture Radar
[3] Discontinuity Detection

۳-۵- تأخیر گروهی[1]

همان‌طور که در شکل (۳-۱) نشان‌داده‌شده است، بر خلاف طیف اندازه که ساختارهای آن رابطه واضحی با درک گفتار دارند، تفسیر و تحلیل طیف فاز دشوار است. به این دلیل که ظاهراً روند و ساختار معناداری وجود ندارد که فرایند ساخت مدل را تسهیل کند. به همین علت، محققان به کار با نمایش‌های دیگری از طیف فازی روی آوردند که اطلاعات فاز را حمل می‌کنند اما درعین‌حال قابل‌تحلیل و کنترل هستند. تابع تأخیر گروهی نمایش اصلی طیف فاز است که به‌صورت منفی مشتق طیفی طیف فاز پیوسته (بدون پیچش) تعریف می‌شود.

$$\tau(t,\omega) = -\frac{\partial \arg\{X(t,\omega)\}}{\partial \omega} \qquad ۹\text{-}۳$$

که $\tau(t,\omega)$، تأخیر گروهی (زمان کوتاه) فریم t ام است. همچنین می‌توان تابع تأخیر گروهی (زمان کوتاه) را به روش زیر محاسبه کرد.

$$\tau(t,\omega) = -\frac{\partial \arg\{X(t,\omega)\}}{\partial \omega} = -\frac{X_{Re}(t,\omega)X'_{Im}(t,\omega) - X'_{Re}(t,\omega)X_{Im}(t,\omega)}{|X(t,\omega)|^2} \qquad ۱۰\text{-}۳$$

مزیت اصلی این فرمول این است که نیازی به واپیچش فاز قبل از محاسبه تأخیر گروهی نیست.

برای محاسبه، تأخیر گروهی باید مشتق $X_{\mathrm{Re}(\omega)}$ و $X_{\mathrm{Im}(\omega)}$ را محاسبه کرد. محاسبه مشتق نمونه های گسسته سرراست نیست و با مقداری خطا همراه است. برای حل این مشکل، می‌توان از ویژگی تبدیل فوریه که در زیر آورده شده است، استفاده کرد.

$$F[nx(n)] = j\frac{dX(\omega)}{d\omega} = -X'_{Im}(\omega) + jX'_{Re}(\omega) \qquad ۱۱\text{-}۳$$

بدین ترتیب به‌جای استفاده از (۳-۱۰) برای محاسبه تأخیر گروه، می‌توان از فرمول زیر استفاده کرد.

[1] Group Delay

$$\tau(t,\omega)=\frac{X_{Re}(p,\omega)Y_{Im}(t,\omega)+Y_{Re}(t,\omega)X_{Im}(t,\omega)}{|X(t,\omega)|^2} \qquad ۱۲-۳$$

که در آن $Y(\omega)$ تبدیل فوریه $y[n] = nx(n)$ است. اثبات باتوجه‌به روابط زیر ساده است.

$$\begin{cases} Y_{Re}(\omega) &= -X'_{Im}(\omega) \\ Y_{Im}(\omega) &= X'_{Re}(\omega) \end{cases} \qquad ۱۳-۳$$

بنابراین با این ترفند می‌توان تأخیر گروهی را بدون پرداختن به پیچش فاز و مسائل مشتق محاسبه کرد.

همچنین روش دیگری برای محاسبه تأخیر گروه وجود دارد. ازآنجایی‌که $LnX(\omega) = Ln|X(\omega)| + j\arg\{X(\omega)\}$ که در آن Ln لگاریتم مختلط است، می‌توان نوشت.

$$\tau(\omega)=-\frac{d}{d\omega}\arg\{X(\omega)\}=-\frac{d}{d\omega}\mathrm{Im}\{\mathrm{Ln}\,X(\omega)\}=-\mathrm{Im}\left\{\frac{d}{d\omega}\mathrm{Ln}\,X(\omega)\right\} \qquad ۱۴-۳$$

زیرا $\mathrm{Im}\{.\}$ و مشتق (d) عملیات خطی هستند؛ بنابراین، می‌توان ترتیب آنها را عوض کرد. با استفاده از تبدیل فوریه و خواص عدد مختلط تأخیر گروه را می‌توان به‌صورت زیر محاسبه کرد.

$$\tau(\omega)=-\mathrm{Im}\left\{\frac{X'(\omega)}{X(\omega)}\right\}=\mathrm{Re}\left\{\frac{jX'(\omega)}{X(\omega)}\right\}=\mathrm{Re}\left\{\frac{F[nx[x]]}{F[x[n]]}\right\} \qquad ۱۵-۳$$

که $\mathrm{Re}\{.\}$ قسمت حقیقی را نشان می‌دهد.

۳-۵-۱- مزیت‌های تأخیر گروهی

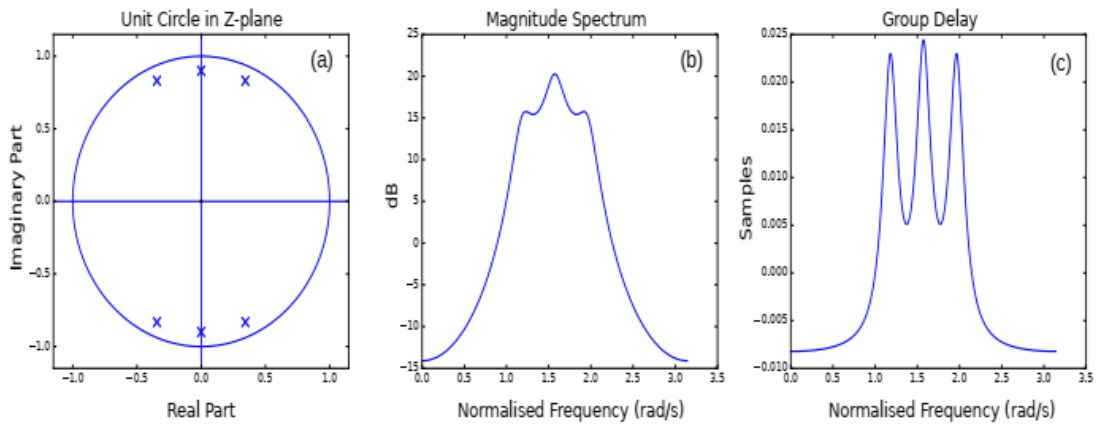
تابع تأخیر گروهی چهار مزیت اصلی دارد:

- وضوح طیفی بالا
- نشت طیفی[1] کم
- خاصیت جمع‌شوندگی
- شباهت به طیف اندازه

[1] Spectral Leakage

تأخیر گروهی اطلاعات طیف فاز را نشان می‌دهد اما شکل کلی آن قابل‌درک‌تر است. برای سیگنال‌های مینیمم فاز، مشابه طیف اندازه، در قطب‌ها جهش دارد و در صفرها به حداقل می‌رسد. اما، برای سیگنال‌های ماکزیمم فاز، برخلاف طیف اندازه، در قطب مقدار حداقل و در صفر مقدار حداکثر دارد.

شکل (۳-۲) طیف اندازه و تأخیر گروهی یک سیگنال (سیستم) ساده را نشان می‌دهد که با شش قطب تخمین زده شده با استفاده از تحلیل ضرایب پیشگویی خطی (LPC) مشخص می‌شود [۳۱]. همانطور که مشاهده می‌شود، تأخیر گروهی وضوح فرکانس بهتر و نشت فرکانس کمتری دارد. همچنین از نظر اوج گرفتن(پیک) در قطب ها و دره ها در صفرها به طیف اندازه شباهت دارد.



شکل ۳-۲ LPC و GD یک سیگنال که با داشتن شش قطب مشخص می‌شود. (a) قطب ها در صفحه z، (b) تخمین طیف توان مبتنی بر LPC (پارامتری)، (c) تابع تأخیر گروهی [۳۱]

در نهایت، اگر دو سیگنال در حوزه زمان در هم کانوال شوند، به‌عنوان‌مثال یک سیگنال (x[n]) و یک پاسخ ضربه (h[n]) یک فیلتر، تأخیرهای گروهی آنها با هم جمع خواهند شد؛ بنابراین، کانولوشن در حوزه زمان معادل جمع در حوزه‌های فاز واپیچی شده و تأخیر گروهی است.

$$F\{x[n]*h[n]\} = X(\omega)H(\omega) = |X(\omega)||H(\omega)|\, e^{j(\arg\{X(\omega)\}+\arg\{H(\omega)\})}$$
$$\Rightarrow \tau_{x*h}(\omega) = \tau_x(\omega) + \tau_h(\omega) \qquad (۳-۱۶)$$

۳-۵-۲- مشکل اصلی تأخیر گروهی

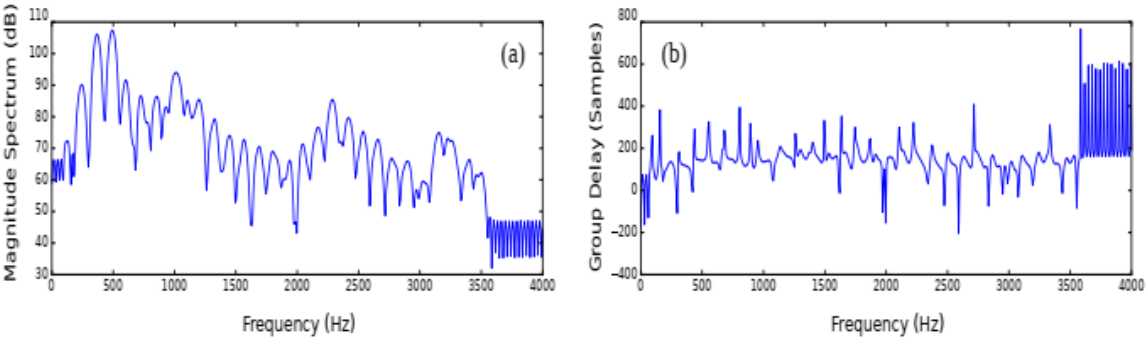
در استفاده از تأخیر گروهی یک مشکل اساسی وجود دارد که سودمندی آن را در عمل محدود

می‌کند. شکل (۳-۳) طیف اندازه و تأخیر گروهی یک سیگنال گفتار معمولی را نشان می‌دهد. همان‌طور که مشاهده می‌شود، در این مورد، تأخیر گروهی بسیار نوک‌تیز است و نه فرکانس پایه (پیچ) و نه فرمنت‌ها را نمی‌توان تشخیص داد. این قله‌های تیز به دلیل قطب‌ها و (یا) صفرهایی هستند که در نزدیکی دایره واحد قرار دارند. برای یک مدل صفر و قطب (ARMA[1]) داریم:

$$X(z) = \frac{B(z)}{A(z)} = \frac{\prod_{q=1}^{Q}\left(z - z_q\right)}{\prod_{p=P}\left(z - z_p\right)} \Rightarrow \tau_X(\omega) = \sum_{q=1}^{Q}\tau_q(\omega) - \sum_{p=1}^{P}\tau_p(\omega) \qquad ۱۷\text{-}۳$$

که در آن τ_q, P, Q, Z_p, Z_q و τ_p به ترتیب صفرها، قطب‌ها، تعداد صفرها، تعداد قطب‌ها، تأخیر گروهی Z_q و تأخیر گروهی Z_P را نشان می‌دهند. با استفاده از (۳-۱۴)، تأخیر گروهی چنین سیستمی را می‌توان به صورت زیر محاسبه کرد.

$$\tau(\omega) = \mathrm{Im}\frac{d\log\{X(\omega)\}}{d\omega} = -\mathrm{Im}\left\{\frac{\frac{d}{d\omega}\left[\frac{B(\omega)}{A(\omega)}\right]}{\frac{B(\omega)}{A(\omega)}}\right\} = \mathrm{Im}\left\{\frac{A'(\omega)B(\omega) - A(\omega)B'(\omega)}{A(\omega)B(\omega)}\right\} \qquad ۱۸\text{-}۳$$



شکل ۳-۳ (a) لگاریتم طیف اندازه و (b) تأخیر گروهی برای یک فریم گفتاری واکدار [۳۱]

همان‌طور که در (۳-۱۸) مشاهده می‌شود، هنگامی که یک قطب یا صفر به دایره واحد نزدیک می‌شود، مخرج به سمت صفر میل می‌کند و تأخیر گروهی در فرکانس مربوطه تبدیل به قله (تیزی) می‌شود. ازآنجایی‌که تأخیر گروهی قطب‌ها و (یا) صفرها با هم جمع می‌شوند، اگر تنها یکی از تأخیرهای

[1] Auto-Regressive Moving Average

گروه به‌صورت قله باشد باعث می‌شود کل تأخیر گروه قله باشد. این قله‌ها، فرمنت‌ها و دیگر خواص مفید طیف را می‌پوشانند و در نهایت باعث می‌شوند که تأخیر گروهی یک نمایش غیر مفیدی داشته باشد. برای سیگنال‌های گفتار، قطب‌ها که عمدتاً با مجرای گفتار مرتبط هستند، در فاصله تقریباً مطمئنی از دایره واحد قرار دارند. به همین دلیل است که تأخیر گروهی یک مدل تمام قطب گفتار، هموار است. اما، صفرهای گفتار که عمدتاً مربوط به مؤلفه تحریک است، نزدیک به دایره واحد قرار دارند و منجر به تأخیر گروهی تیز می‌شوند.

برای بهره‌مندی از مزایای تأخیر گروهی، باید مشکل تیزی آن حل شود. چهار راه‌حل اصلی برای این موضوع وجود دارد که در قسمت بعدی به بررسی دو راه‌حل اصلی می‌پردازیم.

- تأخیر گروهی اصلاح شده (MODGD)
- طیف حاصل‌ضرب (PS)
- تابع تأخیر گروه چیرپ (CGDF[1])
- تأخیر گروهی (پارامتری) مبتنی بر مدل

۳-۵-۲-۱- تأخیر گروه اصلاح شده

در سال ۱۹۹۲، مورتی و یگنارا تأخیر گروهی اصلاح شده (MODGD) را پیشنهاد کردند که تا حد زیادی مشکل تیزی تأخیر گروه را حل می‌کند [۳۵]. آنها پیشنهاد کردند که جزء تحریک را از مخرج تأخیر گروه در (۳-۱۲)، از طریق هموارسازی کپسترال طیف مجذور اندازه، فیلتر کنند.

$$\tau(\omega) = \frac{X_{Re}(\omega)Y_{Re}(\omega) + X_{Im}(\omega)Y_{Im}(\omega)}{S(\omega)} \qquad ۱۹\text{-}۳$$

که در آن $S(\omega)$ هموار شدهٔ کپسترالی طیف توان است.

هموارسازی کپسترال در واقع لیفتینگ زمان کوتاه کپستروم است. در واقع، لگاریتم طیف اندازه گفتار را می‌توان به‌عنوان برهم‌نهی دو عنصر در نظر گرفت: یک مؤلفه به‌سرعت در حال نوسان که توسط

[1] Chirp Group Delay Function

یک پوش به‌آرامی تغییر می‌کند. اولی مربوط به جزء تحریک و دومی به مجرای گفتار مرتبط است. در حوزه کپستروم، مؤلفه‌های فرکانس پایین به جزء کم تغییر، یعنی مجرای صوتی، و اجزای فرکانس بالا به بخش تحریک مرتبط هستند. حال اگر یک لیفتر[1] (حذف‌کننده) با زمان کوتاه اعمال شود، می‌توان جزء تحریک را فیلتر کرد و یک طیف توان هموار به دست آورد. شایان‌ذکر است که هموار کردن نام دیگری برای فیلتر پایین گذر است و هموارسازی کپسترال اساساً حذف اثرات زمان کوتاه در حوزه شبه فازی[2] است.

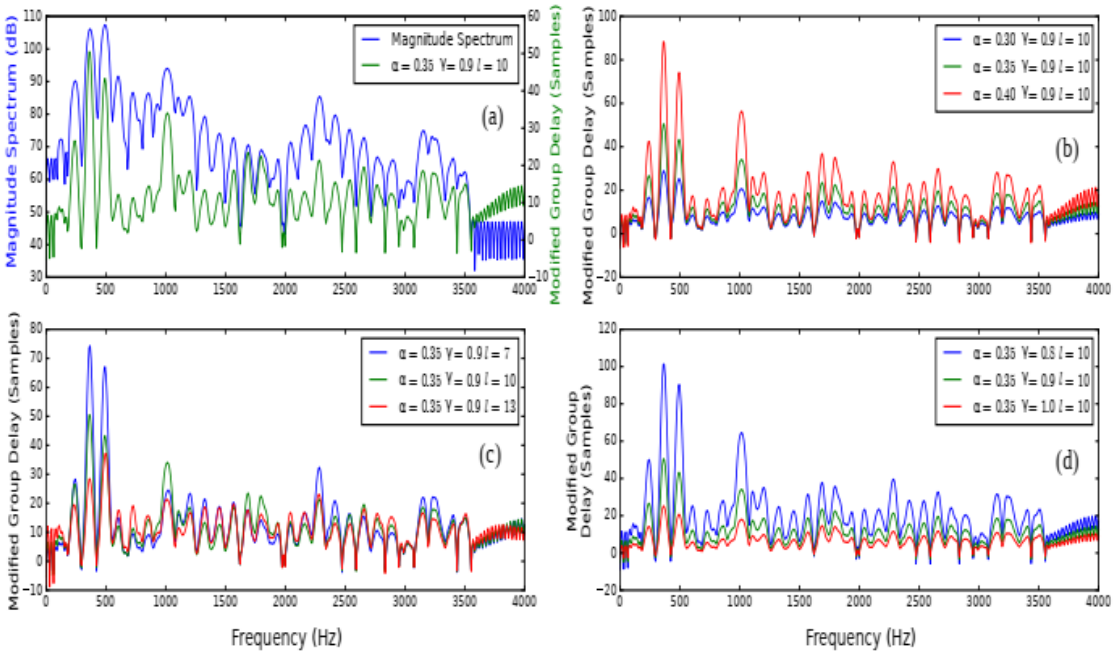
تأخیر گروه اصلاح شده هنوز مشکل دیگری دارد که باید برطرف شود. محدوده دینامیکی همچنان بالاست و پهنای باند فرمت‌ها به دلیل تیزی پیک‌ها، بسیار کم است. دراین‌خصوص مورتی و گاد اصلاح اضافی زیر را پیشنهاد کردند [۳۶].

$$\tau(\omega) = \frac{X_{Re}(\omega)Y_{Re}(\omega) + X_{Im}(\omega)Y_{Im}(\omega)}{S^{\gamma}(\omega)} \qquad \text{۲۰-۳}$$
$$\tau(\omega) \leftarrow \text{sign}\{\tau(\omega)\}\,|\,\tau(\omega)\,|^{\alpha}$$

که α و γ پارامترهای جدیدی هستند که برای محدوده دینامیکی و تنظیم پهنای باند فرمنت معرفی شده‌اند. مقدار مناسب برای α در محدوده ۰.۳ تا ۰.۴ و مقدار بهینه برای γ حدود ۰.۹ پیشنهاد شده است [۳۶], [۳۷]؛ بنابراین، MODGD دارای سه پارامتر به نام‌های α، γ و L است که در آن L مرتبه فیلترمدین[3] است. شکل (۳-۴) تأثیر این پارامترها را بر روی این تابع نشان می‌دهد. با گرفتن تبدیل کسینوسی گسسته (DCT) از تأخیر گروهی اصلاح شده می‌توان مجموعه‌ای از ویژگی‌ها را استخراج کرد.

[1] لیفتر همان عمل فیلتر کردن در حوزه شبه فرکانسی است

[2] Qufrency
[3] Median Filter



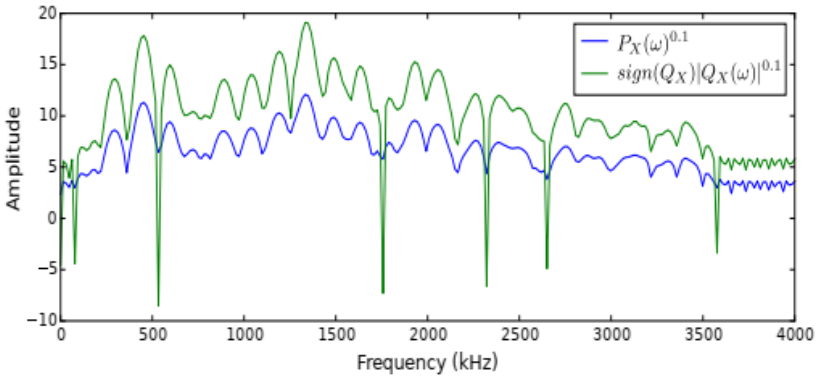
شکل ۳-۴ مقایسه طیف توان با تأخیر گروه اصلاح شده (a) طیف توان رسم شده همراه با تابع تأخیر گروه اصلاح شده (۳-۲۰)، (b) اثر α بر تأخیر گروه اصلاح شده، (c) اثر L بر تأخیر گروه اصلاح شده، (d) اثر γ بر تأخیر گروه اصلاح شده[۳۱]

۳-۵-۲-۲- طیف حاصل‌ضرب

این روش در سال ۲۰۰۴ توسط ژو و پالیوال [۲۲] پیشنهاد شد و همان‌طور که از نام آن پیداست تأخیر گروه در چیزی ضرب می‌شود. همان‌طور که ذکر شد، موضوع تیزی از مؤلفه تحریک گفتار ناشی می‌شود که مخرج را به سمت صفر می‌برد. برای غلبه بر این مشکل، در این کار، مخرج از طریق ضرب تأخیر گروه در تخمین پریودوگرام طیف توان حذف می‌شود.

$$Q(\omega) = \frac{X_{Re}(\omega)Y_{Re}(\omega) + X_{Im}(\omega)Y_{Im}(\omega)}{|X(\omega)|^2} |X(\omega)|^2 \qquad ۳-۲۱$$

که در آن Q(ω) طیف حاصل‌ضرب تأخیر گروهی در توان است که به‌صورت خلاصه طیف حاصل‌ضرب (PS) گفته می‌شود. شکل (۳-۵) طیف اندازه را به همراه طیف حاصل‌ضرب نشان می‌دهد. همان‌طور که مشاهده می‌شود، طیف حاصل‌ضرب شباهت زیادی به طیف اندازه دارد. تفاوت اصلی بین آنها در دره‌های تیز در طیف حاصل‌ضرب است که این امر، از حساسیت تأخیر گروه به صفرهای واقع در مجاورت دایره واحد نشات می‌گیرد [۳۱].



شکل ۳-۵ طیف حاصلضرب ($Q_X(\omega)$) همراه با تخمین طیف توان پریودوگرام ($P_X(\omega) = |X|^2$) [۳۱]

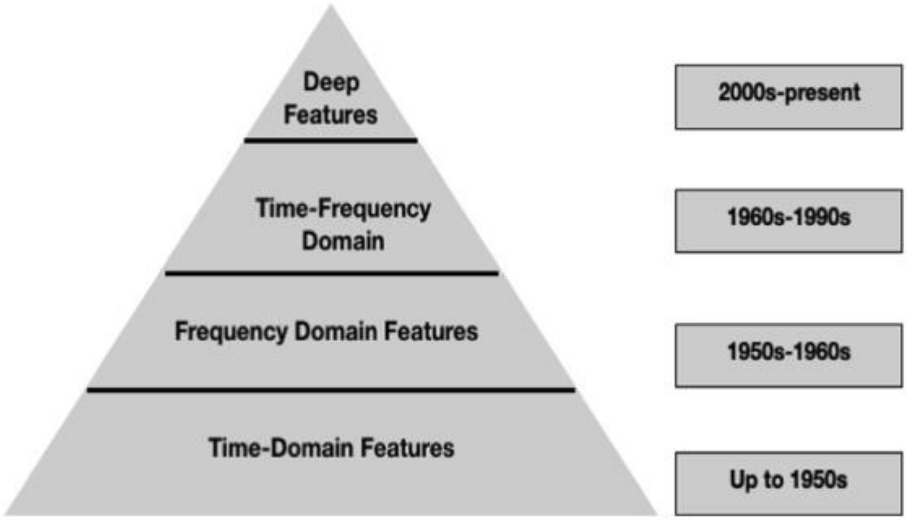
۳-۶- کاربردهای طیف فاز در پردازش گفتار

طیف فاز در زمینه‌های مختلف پردازش گفتار به کار گرفته می‌شود که مهم‌ترین آنها عبارت است از:

- تقویت (بهسازی) گفتار
- تشخیص گفتار مصنوعی و جعل شده
- تشخیص احساسات
- تشخیص گوینده
- استخراج ویژگی برای تشخیص خودکار گفتار (ASR)
- سنتز گفتار
- کدگذاری گفتار
- تجزیه‌وتحلیل گفتار

۳-۷- استخراج ویژگی

هدف اصلی از استخراج ویژگی حفظ اطلاعات مفید و حذف اطلاعات نامرتبط است. ویژگی‌هایی که از سیگنال صوتی در سیستم‌های بازشناسی گفتار استخراج می‌شوند در ۳ حوزه اصلی زمان، فرکانس و زمان - فرکانس هستند. سیر تکامل استخراج ویژگی از سیگنال صوتی در شکل (۳-۶) نشان‌داده‌شده است.



شکل ۳-۶ سیر تکاملی استخراج ویژگی [۳۸]

قدیمی‌ترین و ساده‌ترین ویژگی‌ها از حوزه زمان استخراج می‌شوند. ویژگی‌های حوزه زمانی تا اواخر دهه ۱۹۵۰ تکامل‌یافته‌اند. تا به امروز ویژگی‌های حوزه زمانی نقش مهمی در تجزیه‌وتحلیل و طبقه‌بندی صدا ایفا کرده‌اند. نرخ عبور از صفر (ZCR[1])، ویژگی‌های مبتنی بر انرژی مانند انرژی زمان کوتاه و ویژگی‌های مبتنی بر خود همبستگی[۲] از جمله ویژگی‌های رایج در این حوزه هستند [۳۸].

برای تجزیه‌وتحلیل طیف سیگنال‌های صوتی، چندین ویژگی مانند زیروبم[۳]، فرمنت[۴]ها و غیره از حوزه فرکانس استخراج شده‌اند و تا به امروز در کاربردهای مختلفی مورداستفاده قرار گرفته‌اند. تکامل ویژگی‌های حوزه فرکانس در حدود دهه ۱۹۵۰ تا ۱۹۶۰ بوده است. متداول‌ترین روش‌های استخراج ویژگی در این حوزه عبارت‌اند از ضرایب کپسترال فرکانس مل (MFCCs[5])، کدگذاری پیش‌بینی خطی (LPC[6])، و تبدیل موجک گسسته (DWT[7]) [۱], [۳۸].

در اواخر دهه ۱۹۶۰، الگوریتم‌های استخراج ویژگی مشترک زمان - فرکانس توسعه یافت. از آن زمان این ویژگی‌ها در الگوریتم‌های پردازش سیگنال صوتی استفاده می‌شوند. یکی از رایج‌ترین ویژگی‌ها

[1] Zero-crossing rate
[2] Autocorrelation
[3] Pitch
[4] Formant
[5] Mel-frequency cepstral coefficients
[6] Linear Predictive Coding
[7] Discrete Wavelet Transform

در این حوزه ضرایب کپسترال نرمال‌سازی شده توان (PNCC[1]) است که اخیراً نظر بسیاری از محققان را در حوزه پردازش گفتار به خود جلب کرده است.

یکی از قوی‌ترین روش‌هایی که اخیراً در استخراج ویژگی به کار می‌برند روش استخراج ویژگی عمیق است. ویژگی‌های استخراج شده از لایه‌های مخفی مدل‌های مختلف یادگیری عمیق به‌عنوان ویژگی‌های عمیق شناخته می‌شود. ویژگی‌های عمیق را می‌توان از هر مدل یادگیری عمیق مانند شبکه‌های عصبی کانولوشن (CNN[2])، شبکه‌های عصبی عمیق (DNN[3]) استخراج کرد. از زمان تکامل یادگیری عمیق، ویژگی‌های عمیق به طور گسترده در کاربردهای مختلف مورداستفاده قرار می‌گیرند، در پردازش سیگنال صوتی ویژگی‌های عمیق از سال ۲۰۱۰ در حوزه طبقه‌بندی صحنه آکوستیک، تشخیص گوینده و تجزیه‌وتحلیل صوتی تصویری استفاده شده است [۳۸].

۳-۷-۱- روش‌های متداول استخراج ویژگی در بازشناسی گفتار

ویژگی MFCC یکی از قوی‌ترین و متداول‌ترین ویژگی‌های استفاده شده در حوزه بازشناسی گفتار است. در این پژوهش همچنین از ویژگی MPNCC که عملکرد بهتری نسبت به روش متداول PNCC از خود نشان داده است، برای مقایسه با عملکرد روش پیشنهادی، استفاده شده است [۳۹].

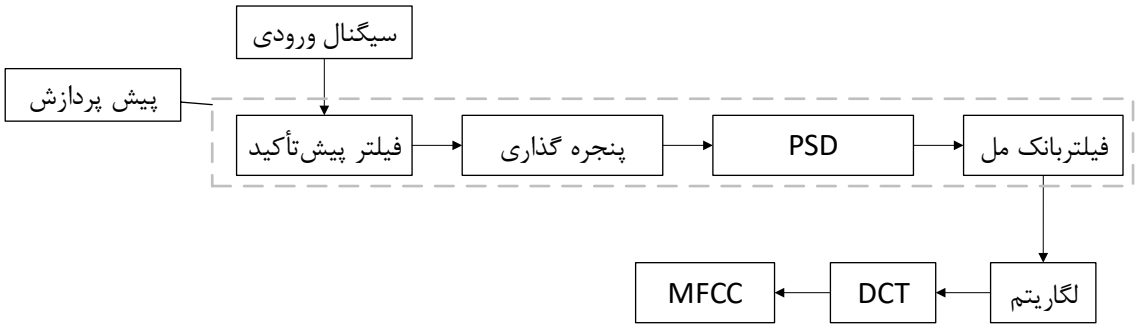
۳-۷-۱-۱- ضرایب کپسترال فرکانس مل

امروزه MFCC را می‌توان یکی از پرکاربردترین روش استخراج ویژگی در پردازش گفتار دانست. حساسیت سیستم شنوایی انسان به زیروبمی صدا در همه فرکانس‌ها یکسان نیست (در فرکانس‌های بالاتر حساسیت کمتری دارد) و به‌صورت غیرخطی عمل می‌کند. برای مدل‌سازی سیستم شنوایی گوش انسان در دهه ۱۹۴۰ مقیاسی به نام مقیاس مل معرفی شد. از فرمول زیر می‌توان برای تبدیل کردن فرکانس به مقیاس Mel استفاده کرد [۱]:

[1] Power-Normalized Cepstral Coefficients
[2] Convolutional Neural Network
[3] Deep Neural Network

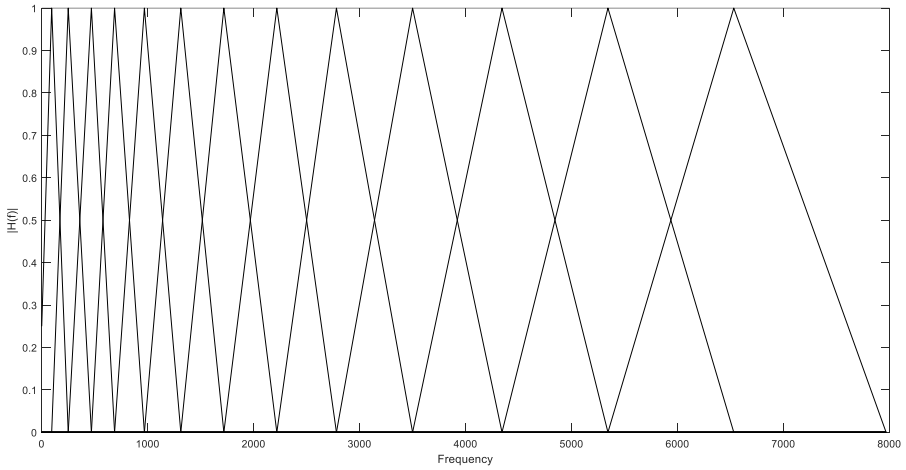
$$F_{mel} = \frac{1000}{\log(2)} \log\left[1 + \frac{F_{Hz}}{1000}\right] \qquad ۲۲-۳$$

اساس کار MFCC متناسب با مدل‌سازی سیستم شنوایی انسان است [۱].در شکل (۳-۷) مراحل استخراج ضرایب ویژگی MFCC نشان‌داده‌شده است.



شکل ۳-۷ مراحل استخراج ویژگی MFCC [۱]

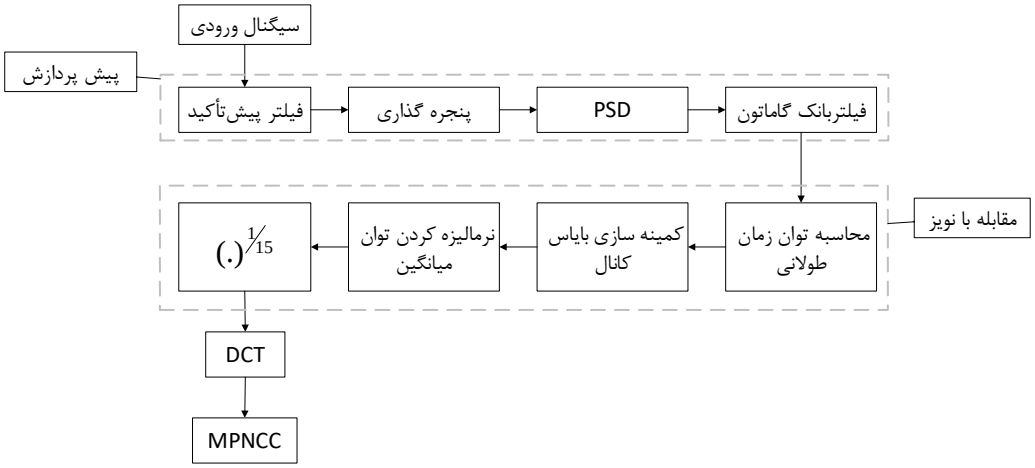
ابتدا سیگنال صوتی به فریم‌های کوچک بین ۲۰-۴۰ میلی‌ثانیه با هم‌پوشانی تقریباً ۵۰ درصد تقسیم می‌شوند. بعد از تقسیم‌بندی سیگنال صوتی به فریم‌های مختلف، جهت ازبین‌بردن اثر ناپیوستگی موجود در مرز بین فریم‌ها هر فریم در تابع پنجره hamming ضرب می‌شود، سپس تبدیل فوریه گسسته (DFT) به نتیجه اعمال می‌شود. در مرحله بعد بر روی طیف به‌دست‌آمده فیلتر بانک با مقیاس مل اعمال می‌شود. در شکل (۳-۸) نحوه توزیع این فیلتر بانک‌ها نشان‌داده‌شده است که متناسب با سیستم شنوایی انسان است. بدین معنی که در فرکانس‌های پایین به‌صورت خطی و در فرکانس‌های بالا به‌صورت لگاریتمی است. سپس لگاریتم انرژی خروجی فیلترها محاسبه می‌شود. در مرحله آخر، تبدیل کسینوس گسسته (DCT) بر روی خروجی‌های لگاریتم انرژی اعمال می‌شود و بدین صورت ویژگی‌های ایستای MFCC به دست می‌آید. همچنین می‌توان از مشتقات اول و دوم MFCC جهت دریافت اطلاعات مربوط به تغییرات سیگنال از یک فریم به فریم دیگر، در کنار ویژگی‌های ایستای MFCC استفاده کرد.



شکل ۳-۸ پاسخ فرکانسی فیلتر بانک مل با ۱۳ فیلتر

۳-۷-۱-۲- روش PNCC اصلاح شده

اولین‌بار روش استخراج ویژگی PNCC در سال ۲۰۰۹ ارائه شد[۴۰] . پس از آن، در سال ۲۰۱۹، تمازین و همکاران [۳۹] توانستند با تغییر آن، به یک روش جدید مبتنی بر PNCC که PNCC بهبود یافته نام دارد (MPNCC)، دست پیدا کنند که می توان روندنمای آن را در شکل (۳-۹) مشاهده کرد.



شکل ۳-۹ روندنمای روش استخراج ویژگی MPNCC [۳۹]

همان‌طور که در شکل (۳-۹) مشاهده می‌شود، فرایند استخراج ویژگی pncc را می‌توان به سه بخش کلی پیش‌پردازش، مقابله با نویز و پردازش نهایی تقسیم کرد که در ادامه، هر قسمت توضیح داده می‌شود.

- پیش‌پردازش:

در این قسمت ابتدا سیگنال گفتار ورودی از یک فیلتر بالاگذر پیش تأکید ((H(z)) عبور داده می‌شود. سپس خروجی فیلتر به قاب‌ها یا فریم‌هایی با طول و هم‌پوشانی مشخصی تقسیم شده و با ضرب هر قسمت در یک پنجره همینگ، مجموعه فریم‌های پنجره‌گذاری شده به دست می‌آید. پس از پنجره‌گذاری، با اعمال یک فیلتر بانک گاماتون روی چگالی طیف توان (PSD[1]) هر فریم و محاسبه توان خروجی هر کانال فیلتربانک، توان طیفی زمان کوتاه به‌صورت رابطه (۳-۲۳) محاسبه می‌شود.

$$P[m,l] = \sum_{k=1}^{M_s/2} |Y[m,k]|^2 \, G_l(k) \qquad ۲۳\text{-}۳$$

که m ، l و M_s به ترتیب اندیس فریم، شماره کانال و رزولوشن طیفی هستند و منظور از $G_l(k)$ تبدیل فوریه گسسته فیلتر کانال lام از مجموعه فیلترهای فیلتربانک گاماتون است.

- مقابله با نویز:

برای مقابله با نویز در mpncc از یک فیلتر توان زمان طولانی به‌عنوان یک فیلتر پایین گذر استفاده شده است که خروجی آن را می‌توان از رابطه (۳-۲۴) به دست آورد:

$$Q[m,l] = \frac{1}{2M+1} \sum_{m'=m-M}^{m+M} P\left[m', l \right] \qquad ۲۴\text{-}۳$$

طبق توضیحاتی که در مرجع [۳۹] آمده است، برای مقدار M=5 خروجی فیلتر منجر به جواب نهایی بهتری می‌شود.

پس از محاسبه زمان طولانی، مشکلی که فرایند استخراج ویژگی ایجاد می‌شود، این است که اثر هموارسازی باعث پخش شدن انرژی‌های نویز بیشتر از انرژی‌های گفتاری در طول هر کانال گاماتون می‌شود؛ بنابراین، یک بایاس کانال ایجاد می‌کند. این بایاس بیشتر به توزیع نویز طیفی تا توزیع طیفی گفتار بستگی دارد. ازآنجایی‌که PSD نویز معمولاً یکنواخت نیست، به‌ویژه در مورد نویز محیطی، مقدار بایاس از هر کانال گاماتونی به کانال دیگر متفاوت است. طبق رابطه (۳-۲۵)، اثر بایاس کانال با کم‌کردن

[1] Power Spectral Density

مقدار کمینه هر کانال که در ضریب ثابت (0<d<1) d از مقادیر کانال، میزان بایاس کانال به حداقل می‌رسد.

$$\tilde{Q}[l] = Q[l] - d \times \min(Q[m,l]) \qquad ۲۵-۳$$

که برای d در [۳۹] مقدار ۰.۶ در نظر گرفته شده است.

راهکار دیگری که در [۳۹] برای مقابله با نویز در نظر گرفته شده است. نرمالیزه کردن توان میانگین نام دارد که برای اینکار ابتدا میانگین توان هر فریم محاسبه میشود که در رابطه (۳-۲۶) آمده است.

$$\mu[m] = \lambda_\mu \mu[m-1] + \frac{(1-\lambda_\mu)}{L} \sum_{l=0}^{L-1} \tilde{Q}[m,l] \qquad ۲۶-۳$$

که λ_μ ضریب فراموشی است و مقدار آن برابر با ۰.۹۹۹ می باشد. در نهایت با استفاده از رابطه (۳-۲۷) توان نرمال شده محاسبه می شود.

$$U[m,l] = k \frac{\tilde{Q}[m,l]}{\mu[m]} \qquad ۲۷-۳$$

در رابطه (۳-۲۷) مقدار k به‌صورت اختیاری تعیین می‌شود.

به طور تجربی ثابت شده است که یک تابع قانون توان با فرجه $\frac{1}{15}$ تطبیقی مناسب و منطقی را با داده های فیزیولوژی در هنگام بهبود دقت بازشناسی در حضور نویز فراهم می کند. به همین منظور در آخرین گام از مرحله مقابله با نویز از رابطه (۳-۲۸) استفاده می شود.

$$V[m,l] = U[m,l]^{1/15} \qquad ۲۸-۳$$

- پردازش نهایی:

در این مرحله از mpncc، یک تبدیل کسینوسی گسسته DCT به‌منظور کاهش بیشتر ابعاد بردار طیفی به ضرایب به‌دست‌آمده از مراحل قبل، به رابطه (۳-۲۸) اعمال می‌شود.

طبق نتایج ارائه شده در [۳۹] روش استخراج ویژگی MPNCC عملکرد بهتری در بازشناسی کلمات نسبت به سایر روش های متداول (MFCC، RASTA-PLP و PNCC) دارد.

فصل چهارم

روش پیشنهادی

٤-١- مقدمه

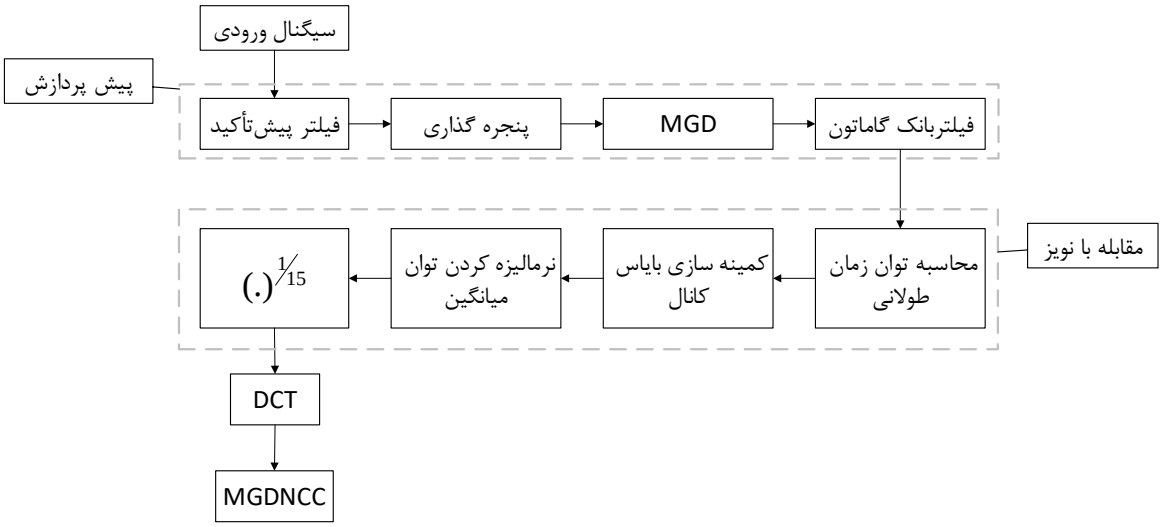
باتوجه‌به مطالب ارائه شده در فصل قبل می‌توان گفت که روش MPNCC نسبت به سایر روش‌ها هم جدیدتر است و هم کارایی بهتری در بازشناسی کلمات دارد. به همین منظور ما روش پیشنهادی خود را بر مبنی MPNCC ارائه می‌کنیم. به این صورت که در روش پیشنهادی ضرایب کپسترال نرمالیزه شده MGD را به‌عنوان ویژگی از سیگنال صوت استخراج می‌کنیم. از نظر ساختاری نیز در این فصل ابتدا روش استخراج ویژگی و سیستم بازشناسی گفتار پیشنهادی را ارائه خواهیم کرد. سپس به چگونگی به‌دست‌آمدن مقادیر پارامترهای اساسی روش استخراج ویژگی و سیستم بازشناسی گفتار پیشنهادی، می‌پردازیم و در نهایت به توضیح پایگاه‌داده به‌خصوص پایگاه‌داده استفاده شده در این پایان‌نامه (TIMIT) پرداخته خواهد شد.

٤-٢- روش پیشنهادی:

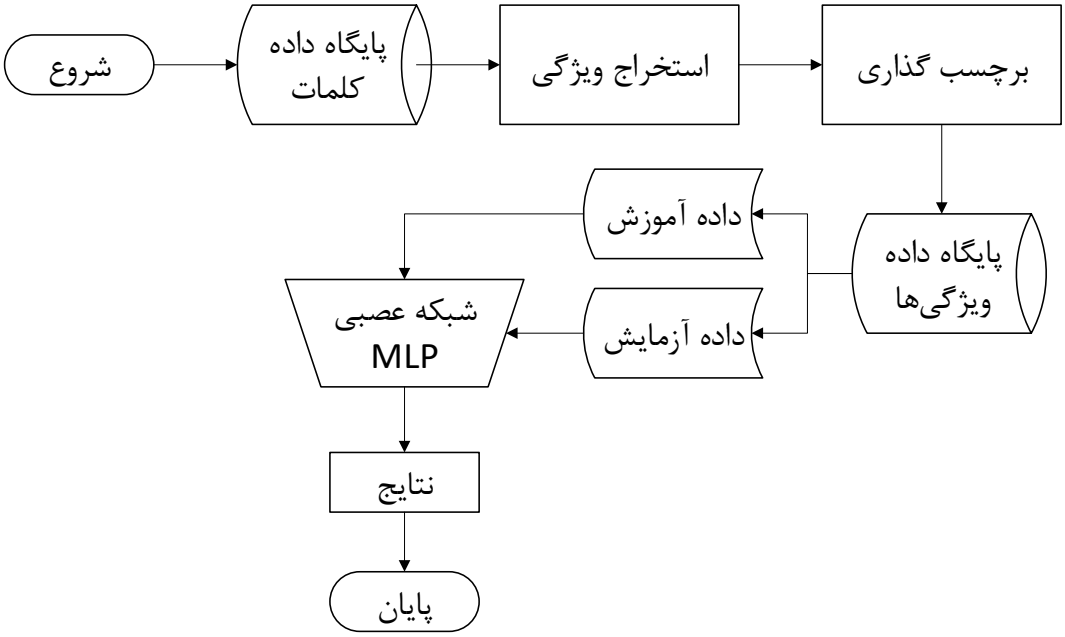
روشی که در این تحقیق جهت استخراج ویژگی پیشنهاد می‌شود، یک روش نوین است که در آن با استفاده از MGD به‌جای PSD در فرآیند استخراج ویژگی MPNCC یک روش جدید مبتنی بر فاز جهت استخراج ویژگی ایجاد می‌شود. روندنمای این روش که می‌توان آن را ضرایب کپسترال نرمالیزه شده تأخیر گروهی اصلاح شده (MGDNCC[1]) نامید، در شکل (۴-۱) نمایش‌داده‌شده است. مانند MPNCC شامل سه بخش پیش‌پردازش، مقابله با نویز و پردازش نهایی است که به دلیل این که دو بخش نهایی دقیقاً مانند MPNCC است، در بخش بعدی از شرح آنها چشم‌پوشی خواهیم کرد.

همچنین برای بازشناسی گفتار و مشخص‌کردن میزان کارایی روش استخراج ویژگی از یک شبکه عصبی MLP دولایه استفاده کرده‌ایم. در نتیجه سیستم بازشناسی گفتار کلی را به‌صورت روند نمای شکل (۴-۲) نشان داد.

[1] Modified Group Delay Normalized Cepstral Coefficient’s



شکل ۴-۱ روندنمای روش استخراج ویژگی پیشنهادی



شکل ۴-۲ روندنمای سیستم بازشناسی پیشنهادی

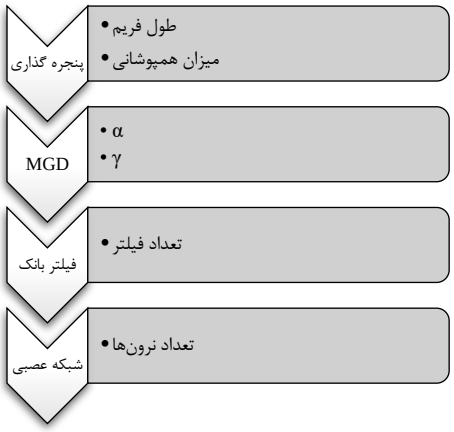
یکی از اصلی‌ترین چالش‌های موجود در سیستم بازشناسی شکل (۴-۲)، این است که طول کلمات باهم برابر نیستند در نتیجه تعداد فریم‌های به‌دست‌آمده برای هر کلمه نسبت به سایر کلمات حتی کلمات هم کلاس با آن متفاوت می‌شود. این تفاوت در تعداد فریم باعث می‌شود که تعداد ویژگی‌ها برای هر نمونه کلمه نیز نسبت به سایر کلمات متفاوت شود که این عدم برابری در تعداد ویژگی‌ها با ملزومات استفاده از شبکه عصبی مغایرت دارد.

برای حل این مشکل در این تحقیق از هر ویژگی به‌دست‌آمده از هر کانال مشخص در تمامی فریم‌ها میانگین گرفته‌ایم که این کار منتج به یک بردار ویژگی که تعداد درایه‌های آن با تعداد فیلترهای گاماتون برابر است، می‌شود. در نتیجه ما برای هر کلمه در پایگاه‌داده ویژگی‌ها یک بردار ویژگی با ۲۱ درایه خواهیم داشت. دلیل اینکه تعداد فیلترهای گاماتون در نتیجه تعداد ویژگی‌ها، برابر با ۲۱ است در بخش بعد توضیح داده خواهد شد.

۴-۳- پیاده‌سازی

برای پیاده‌سازی روش پیشنهادی نیاز است که ما مقدار بهینه پارامترهای اصلی را داشته باشیم. پارامترهایی که ما برای بخش‌های مختلف در نظر گرفته‌ایم را می‌توان در شکل (۴-۳) مشاهده کرد.

برای به‌دست‌آوردن مقادیر پارامترهای شکل (۴-۳) نیاز است که یک مقداردهی اولیه برای آنها داشته باشیم که مقادیر اول بخش MGD را از مرجع [۲۲] برابر با $\alpha = 0.4$ و $\gamma = 0.9$ و همچنین تعداد کانال فیلتر بانک را از مرجع [۳۹]آورده‌ایم که برابر ۲۵ است. برای تعداد نرون‌های هر دو لایه مخفی مقدار ۱۰۰ را به عنوان مقدار اولیه در نظر گرفته ایم.



شکل ۴-۳ پارامتر های اساسی در روش استخراج ویژگی و سیستم بازشناسی گفتار پیشنهادی

پس از مقداردهی در اولین گام مقدار طول فریم و هم‌پوشانی بهینه را به دست می‌آوریم. برای این کار با طول فریم و هم‌پوشانی‌هایی که در اکثر مراجع استفاده شده است، میزان موفقیت بازشناسی سیستم پیشنهادی را به دست آورده و بهترین جواب را انتخاب می‌کنیم. در جدول (۴-۱) می‌توان نتیجه

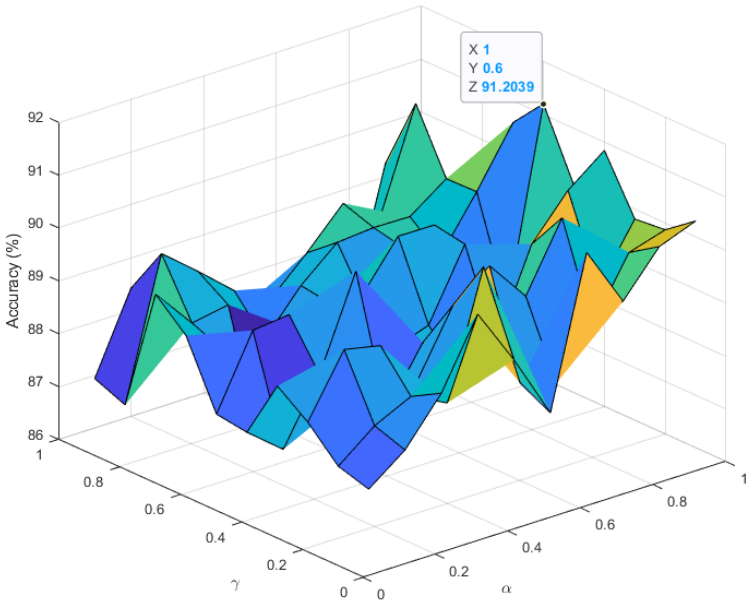
را مشاهده کرد.

جدول ۴-۱نتایج بدست آمده برای طول فریم و میزان هم پوشانی

طول فریم (ms)	۲۰	۲۵	۳۲
هم‌پوشانی (ms)	۱۰	۱۰	۱۶
صحت (٪)	۹۱	۸۹	۸۸/۳

طبق نتایج به‌دست‌آمده، بهترین جواب با طول فریم ۲۰ میلی‌ثانیه و هم‌پوشانی ۱۰ میلی‌ثانیه حاصل می‌شود.

برای مقادیر α و γ در MGD نیز آنها را در دامنه تعریفشان که بین ۰ تا ۱ است با گام ۰/۱ تغییر داده و میزان صحت بدست آمده در هر حالت را محاسبه کرده ایم که نتیجه آن به صورت شکل (۴-۴) بدست آمده است.

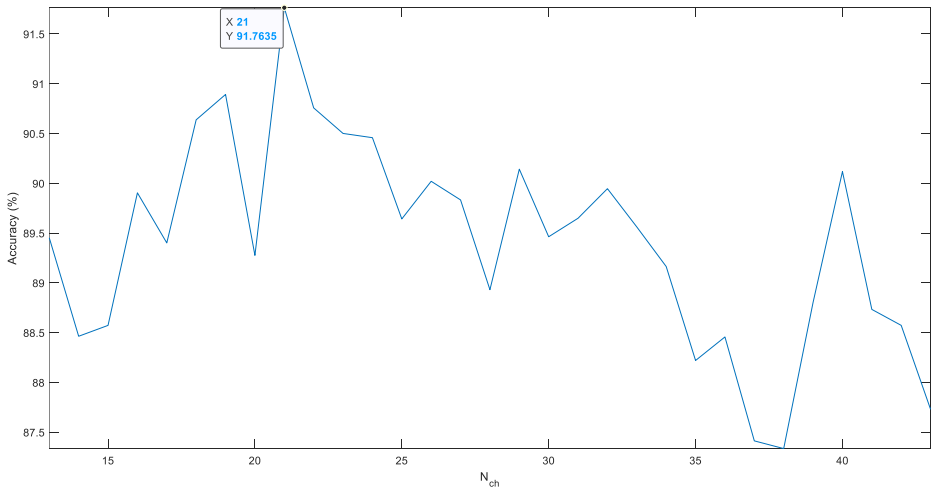


شکل ۴-۴ نتایج به دست آمده برای پارامترهای MGD

که بنا بر نتایج شکل (۴-۴) هنگامی که α و γ به ترتیب برابر با ۱و ۰/۶ باشند صحت بدست آمده

بیشینه می شود.

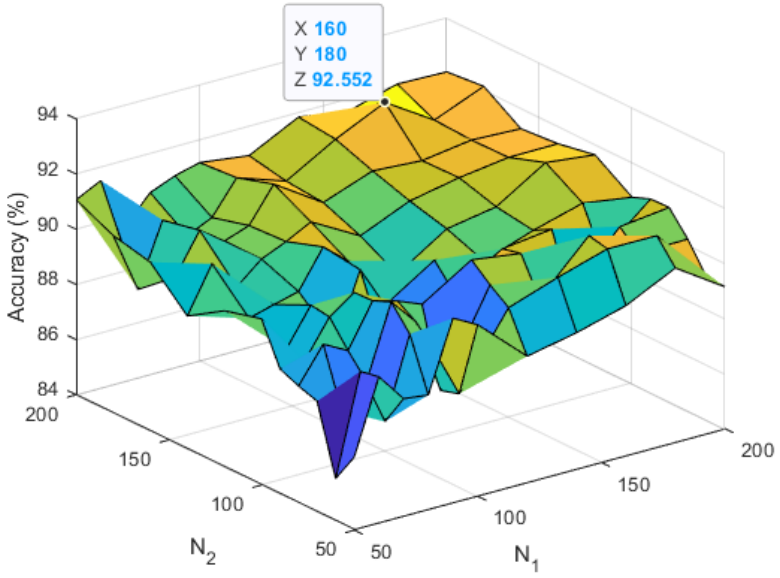
در گام سوم از مراحل به‌دست‌آوردن مقادیر پارامترهای بهینه، نوبت به تعداد کانال می‌رسد که مقادیر صحت خروجی برای تعداد کانال‌های ۱۳ تا ۴۳ در شکل (۴-۵) آمده است.



شکل ۴-۵ نتایج به دست آمده برای تعداد کانال ها

باتوجه‌به شکل (۴-۵) مقدار بهینه برای پارامتر تعداد کانال برای با ۲۱ است.

در نهایت برای به‌دست‌آوردن تعداد نرون بهینه نیز می‌توان از نمودار شکل (۴-۶) استفاده کنیم که از کنار هم قراردادن نتایج شبیه‌سازی در تعداد نرون‌های مختلف به‌دست‌آمده است.



شکل ۴-۶ نتایج به دست آمده برای تعداد نرون های لایه های مخفی شبکه عصبی

در نهایت می‌توان نتایج به‌دست‌آمده در این قسمت را به‌صورت خلاصه در شکل (۴-۷) مشاهده کرد.

شکل ۴-۷ مقادیر بهینه به دست آمده برای پارامترهای اساسی

٤-٤- پایگاه‌داده

یکی از مهم‌ترین بخش‌های سیستم‌های بازشناسی گفتار، پایگاه‌داده است. یک پایگاه‌داده گفتار شامل فایل‌های صوتی گفتار و متن متناظر با این فایل‌ها است که برای آموزش و آزمایش سیستم‌های بازشناسی گفتار ضروری است. به‌صورت کلی پایگاه‌داده گفتار شامل دو بخش اساسی یعنی گفتار بازخوانی شده و گفتار طبیعی است.

- گفتار بازخوانی شده

در این پایگاه‌داده از گوینده خواسته می‌شود که کلمات یا جملات از پیش تعیین شده مانند دنباله‌ای از اعداد، گزیده‌ای از کتاب و ... را بازخوانی کنند و در محیط‌های کنترل شده‌ای ضبط می‌شوند.

- گفتار طبیعی

در این پایگاه‌داده از گفتار روزمره افراد مانند مکالمه بین دو یا چند نفر، شخصی که یک داستانی را تعریف می‌کند و... استفاده می‌شود.

ازآنجایی‌که داده‌برداری در پایگاه‌داده گفتار بازخوانی شده در محیط‌های کنترل شده، انجام می‌شود در مقایسه با پایگاه‌داده گفتار طبیعی، درباز شناسی گفتار عملکرد بهتری را از خود نشان می‌دهند. هرچند امروزه هدف اصلی از بهبود سیستم‌های تشخیص گفتار، افزایش عملکرد آنها در شرایط طبیعی است. اما دستیابی به چنین پایگاه‌های داده طبیعی به دلیل محدودیت‌های متعدد دشوار است. بااین‌وجود می‌توان با افزودن نویز به پایگاه‌داده بازخوانی شده در شرایط آزمایشگاهی، پایگاه‌داده‌ای مشابه گفتار طبیعی ایجاد کرد و عملکرد سیستم‌های بازشناسی گفتار را به کمک آن‌ها سنجید.

٤-٤-١- پایگاه‌داده TIMIT

مجموعه گفتار بازخوانی شده TIMIT برای ارائه داده‌های گفتاری در زمینه افزایش دانش آکوستیک - آوایی و برای توسعه و ارزیابی سیستم‌های تشخیص خودکار گفتار طراحی شده است. TIMIT حاصل تلاش‌های مشترک چندین سازمان و تحت حمایت و اسپانسری آژانس پروژه‌های تحقیقاتی پیشرفته دفاعی (DARPA[1]) آمریکا است.

طراحی این مجموعه یک تلاش مشترک بین مؤسسه فناوری ماساچوست (MIT[2])، مؤسسه تحقیقاتی استنفورد (SRI[3]) و تگزاس اینسترومنتز (TI[4]) بود. عملیات ضبط این پایگاه‌داده در تگزاس اینسترومنتز انجام گرفت و سپس در دانشگاه MIT متن گذاری انجام گرفت، و توسط مؤسسه ملی استاندارد و فناوری (NIST[5]) نگهداری، تأیید و برای تولید CD-ROM آماده شد.

TIMIT از گفتارهای ضبط شده به زبان انگلیسی که از نظر آوایی متعادل هستند، تشکیل شده است. این گفتارها با استفاده هم‌زمان از دو میکروفون حساس به صدا و سنهایزر با نرخ ۲۰ کیلوهرتز در یک محیط ساکت با نرخ سیگنال به نویز ۲۹ دسی‌بل در TI به‌صورت دیجیتال ضبط شده است. سپس نوارهای دیجیتال به NIST ارسال شدند و در آنجا با نرخ ۱۶ کیلوهرتز نمونه‌برداری شدند تا برای

[1] Defense Advanced Research Projects Agency
[2] Massachusetts Institute of Technology
[3] Stanford Research Institute
[4] Texas Instruments
[5] National Institute of Standards and Technology

برچسب‌گذاری در MIT آماده شوند.TIMIT از ۲۳۴۲ جمله مختلف از سه مجموعه تشکیل شده است که شامل ۲ جمله گویشی (SA) ارائه شده توسط SRI که توسط تمامی گوینده‌ها بیان می‌شود، ۴۵۰ جمله فشرده آوایی (SX) طراحی شده توسط MIT که هر جمله توسط ۷ گوینده بیان می‌شود و ۱۸۹۰ جمله تصادفی با آواهای متنوع (SI) انتخاب شده توسط TI است که هر جمله توسط یک گوینده بیان شده است. تعداد گوینده‌ها ۶۳۰ نفر شامل ۴۳۸ نفر (۷۰٪) مرد و ۱۹۲ نفر زن (۳۰٪) از ۸ منطقه (نیوانگلند شمالی، میدلند شمالی، میدلند جنوبی، جنوبی، نیویورک، غربی و ناحیه چرخشی) گویش اصلی ایالات متحده است. هر گوینده ۱۰ جمله (تکراری یا جدید) را بیان می‌کند که در مجموع از ۶۳۰۰ جمله بیان شده توسط گوینده‌ها (۵.۴ ساعت) تشکیل شده است. همه جملات به‌صورت دستی در سطح واج و کلمه برچسب‌گذاری شده‌اند [۴۱].

در این تحقیق با استفاده برچسب‌گذاری‌های موجود در پایگاه‌داده TIMIT ابتدا تمام کلمات را جدا می‌کنیم. سپس از ۱۸ کلمه که نسبت به باقی کلمات تعداد دفعات تکرار آنها بیشتر است، استفاده کرده‌ایم که می‌توان آنها در جدول (۴-۲) مشاهده کرد.

جدول ۴-۲ کلمات استفاده شده در روش پیشنهادی

شماره کلاس	**۱**	**۲**	**۳**	**۴**	**۵**	**۶**	**۷**	**۸**	**۹**
کلمه (کلاس)	That	She	Had	Your	Dark	Suit	Greasy	Wash	Water
شماره کلاس	**۱۰**	**۱۱**	**۱۲**	**۱۳**	**۱۴**	**۱۵**	**۱۶**	**۱۷**	**۱۸**
کلمه (کلاس)	All	Year	Don’t	Ask	Me	Carry	Oily	Rag	Like

تعداد نمونه‌های هر کلاس نیز برابر با ۶۳۰ کلمه است که این مقدار را با درنظرگرفتن حداقل تعداد نمونه کلاس‌ها به دست آورده‌ایم. از این ۶۳۰ نمونه برای هر کلاس ۷۰ درصد آن را که ۴۴۱ نمونه است، برای داده‌های آموزش و مابقی را که ۱۸۹ نمونه است، به‌عنوان داده‌های آزمایش در نظر می‌گیریم.

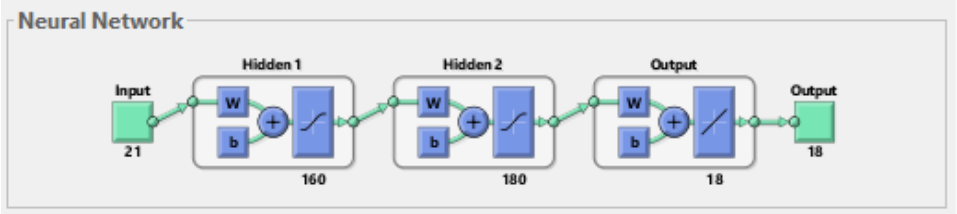
فصل پنجم

شبیه سازی و نتایج

۵-۱- مقدمه

در این پایان‌نامه جهت ارزیابی عملکرد روش پیشنهادی در بازشناسی گفتار در مقایسه با دو روش بسیار متداول در این حوزه یعنی MFCC و MPNCC، و همچنین ویژگی مبتنی بر فاز MGD از پایگاه‌داده TIMIT جهت تشخیص کلمات استفاده شده است. برای این منظور از یک شبکه عصبی MLP با دولایه مخفی و با تعداد نرون‌های به ترتیب ۱۶۰ و ۱۸۰ نرون در هر لایه، در نرم‌افزار متلب (۲۰۲۰b) و با سیستمی با مشخصات زیر استفاده شده است. جزئیات این شبکه عصبی در شکل (۵-۱) آورده شده است.

Windows 10 Pro.Intel(R) Core(TM) i5-4200M CPU @ 2.50GHz 64-bit RAM-8GB



شکل ۵-۱ ساختار شبکه عصبی استفاده شده

مقایسه عملکرد روش‌ها در دو حالت بدون نویز و نویزی انجام شده است. برای این منظور از نویزهای سفید گوسی، کانال HF، تداخل سیگنال، کارخانه و نویز VOLVO در SNRهای ۵-، ۰، ۵ و ۱۰ دسی‌بل استفاده شده است. معیار استفاده شده جهت ارزیابی عملکرد روش‌ها، معیار "صحت " است.

- **صحت**

درجه‌ای از اندازه‌گیری که منجر به مقدار صحیح می‌شود را صحت می‌نامند.

صحت یکی از محبوب‌ترین معیارها در طبقه‌بندی داده‌ها در چند گروه مختلف است و مستقیماً از ماتریس درهم‌ریختگی محاسبه می‌شود که از رابطه زیر به دست می‌آید.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (۱-۵)$$

پارامترهای این رابطه به‌صورت زیر تعریف می‌شوند:

موارد مثبت صحیح (TP[1]): مواردی که برچسب واقعی آنها مثبت است و کلاس آنها به‌درستی مثبت پیش‌بینی‌شده است.

موارد مثبت کاذب (FP[2]): مواردی که برچسب واقعی آنها منفی است و کلاس آنها به‌اشتباه مثبت پیش‌بینی‌شده است.

موارد منفی صحیح (TN[3]): مواردی که برچسب واقعی منفی است و کلاس آنها به‌درستی منفی پیش‌بینی‌شده است.

موارد منفی کاذب (FN[4]): مواردی که برچسب واقعی آنها مثبت است و کلاس آنها به‌اشتباه منفی پیش‌بینی‌شده است.

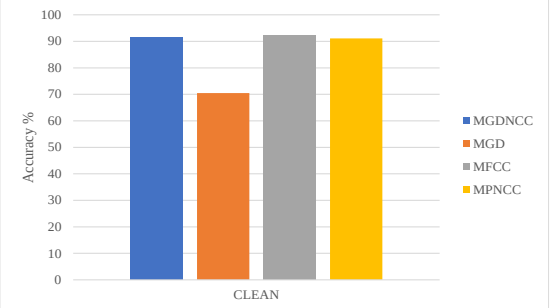
در ادامه نتایج آزمایش‌ها در حالت‌های بدون نویز و نویزی در SNRهای مختلف آورده شده است.

۵-۲- نتایج بازشناسی گفتار با استخراج ویژگی‌های مختلف در شرایط بدون نویز

جدول (۵-۱) و شکل (۵-۲) نتایج عملکرد سیستم بازشناسی گفتار با روش‌های مختلف استخراج ویژگی را بر حسب معیار صحت بازشناسی، نشان می‌دهد.

جدول ۵-۱ صحت بازشناسی گفتار در حالت بدون نویز

ویژگی	MGDNCC (پیشنهادی)	MGD	MFCC	MPNCC
صحت (٪)	۹۱.۳۴	۷۰.۴۷	۹۲.۱۷	۹۱.۱۳



شکل ۵-۲ صحت بازشناسی گفتار در حالت بدون نویز

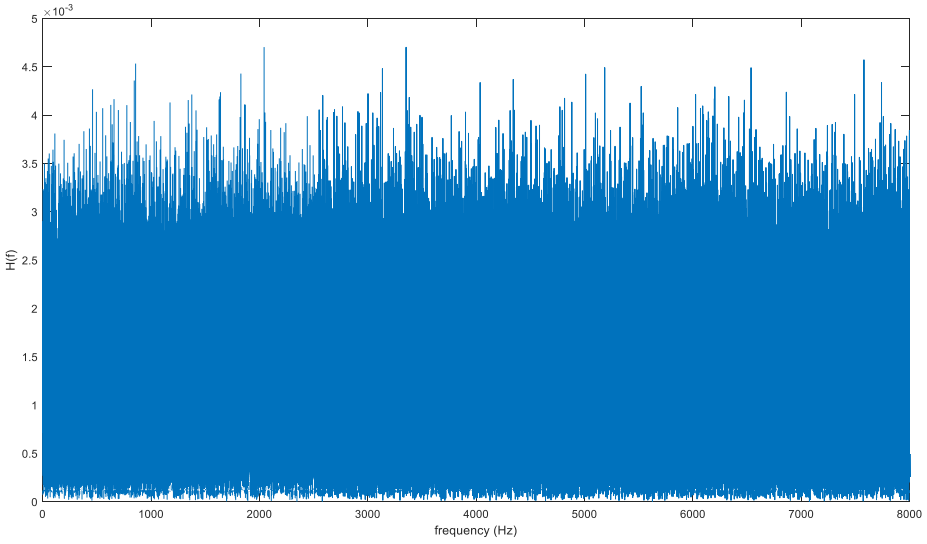
[1] True Positive
[2] False Positive
[3] True Negative
[4] False Negative

باتوجه‌به جدول (۵-۱) مشاهده می‌گردد که:

- روش MFCC در مقایسه با سایر روش‌های ارائه شده بهترین عملکرد را داشته است درحالی‌که روش MGD بدترین عملکرد را در مقایسه با سایر روش‌ها از خود نشان داده است.
- روش پیشنهادی عملکرد بسیار خوبی را داشته است و اختلاف اندکی با روش MFCC دارد. همچنین عملکرد روش پیشنهادی نسبت به روش MPNCC بهتر بوده است.
- روش MGD ازآنجایی‌که ویژگی فاز خالص است عملکرد نسبتاً خوبی را از خود نشان داده است، هرچند در مقایسه با سه روش دیگر عملکرد پایینی دارد.
- نکته قابل‌تأمل این است استخراج ویژگی از طریق روش پیشنهادی منجر به بهبود عملکرد بازشناسی گفتار نسبت به هر دو روش MGD و MPNCC به‌تنهایی، شده است.

۵-۳- نتایج بازشناسی در نویز سفید گوسی

جدول (۵-۲) نتایج عملکرد سیستم بازشناسی گفتار با روش‌های مختلف استخراج ویژگی را بر حسب معیار صحت بازشناسی، در نویز سفید گوسی در SNRهای مختلف، نشان می‌دهد. طیف فرکانسی این نویز در شکل (۵-۳) آمده است.



شکل ۵-۳ طیف فرکانسی نویز سفید گوسی

باتوجه‌به جدول (۵-۲) مشاهده می‌گردد که:

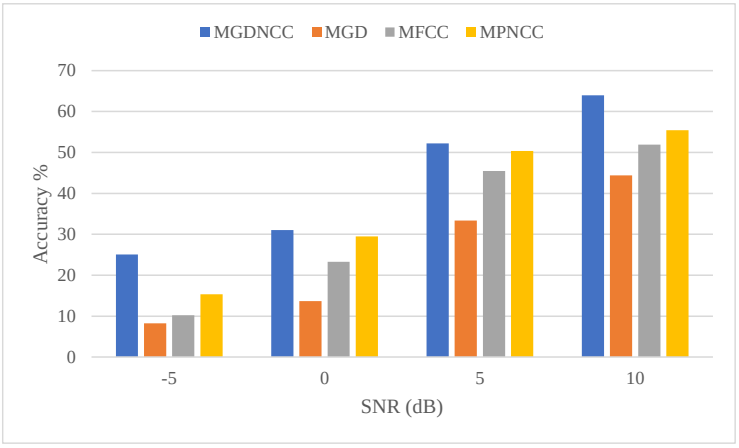
- در شرایط نویزی شدید عملکرد روش‌های متداول بسیار ضعیف می‌شود درحالی‌که روش پیشنهادی عملکرد بهتری را از خود نشان می‌دهد. همان‌طور که مشاهده می‌شود، در SNR ۵- دسی‌بل روش پیشنهادی بسیار بهتر از دو روش متداول MFCC و MPNCC عمل کرده است.
- روش MGD به‌تنهایی عملکرد ضعیفی را در مقایسه با روش‌های دیگر از خود نشان داده است که می‌توان تأثیرگذاری روش پیشنهادی را که بر پایه MGD است را به‌وضوح مشاهده کرد.
- همان‌طور که در ستون آخر جدول مشاهده می‌گردد، روش پیشنهادی به‌صورت میانگین بهترین صحت را در مقایسه با سه روش دیگر داشته است.

جدول ۵-۲ صحت بازشناسی گفتار در نویز سفید گوسی

ویژگی‌ها	SNR (dB)				
	۵-	۰	۵	۱۰	میانگین
MGDNCC (پیشنهادی)	۲۵.۰۵	۳۱.۰۵	۵۲.۲۳	۶۳.۹۳	۴۳.۰۷
MGD	۸.۲۳۰	۱۳.۶۴	۳۳.۳۵	۴۴.۴۱	۲۴.۹۱
MFCC	۱۰.۲۳	۲۳.۲۹	۴۵.۴۶	۵۱.۸۸	۳۲.۷۱
MPNCC	۱۵.۳۵	۲۹.۴۷	۵۰.۳۵	۵۵.۴۱	۳۷.۶۴

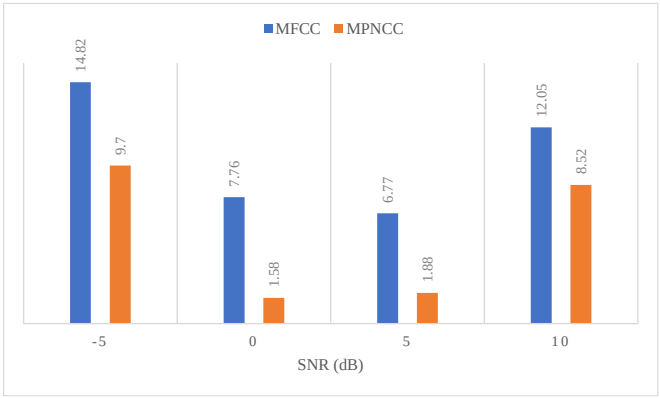
شکل (۵-۴) مقایسه بصری بهتری از عملکرد این چهار روش را ارائه می‌دهد. همان‌طور که در این شکل نیز مشاهده می‌گردد، در شرایط نویزی بیشتر، روش پیشنهادی صحت بیشتری را نسبت به روش‌های دیگر داشته است. وبا کم‌شدن مقدار توان نویز، عملکرد روش‌های MFCC و MPNCC با شیب بیشتری افزایش می‌یابد (به دلیل افت شدید عملکرد آن‌ها در شرایط نویزی زیاد) اما همچنان عملکرد آنها کمتر از روش پیشنهادی بوده است.

شکل (۵-۵) میزان درصد بهبود روش پیشنهادی در تشخیص کلمه را در مقایسه با دو روش متداول MFCC و MPNCC، در SNRهای مختلف نشان می‌دهد.



شکل ۵-۴ صحت بازشناسی گفتار در نویز سفید گوسی

همان‌طور که در این شکل مشاهده می‌شود، بیشترین درصد بهبود را در ۵- دسی‌بل و کمترین درصد بهبود را در ۰ دسی‌بل به ترتیب نسبت به MFCC و MPNCC داشته است.



شکل ۵-۵ میزان بهبود عملکرد روش پیشنهادی نسبت به ویژگی های MFCC و MPNCC در نویز سفید گوسی

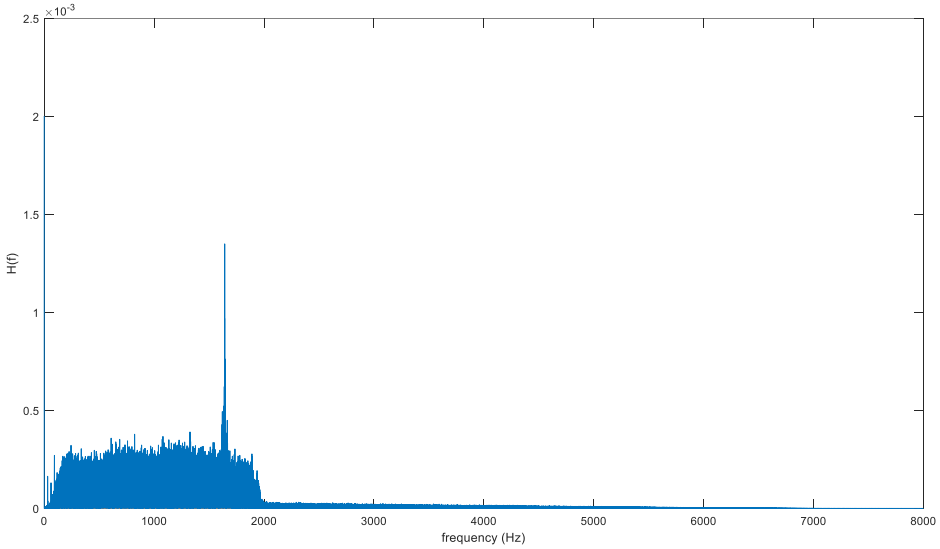
۵-٤- نتایج بازشناسی در نویز کانال HF

جدول (۵-۳) نتایج عملکرد سیستم بازشناسی گفتار با روش‌های مختلف استخراج ویژگی را بر حسب معیار صحت بازشناسی، در نویز کانال HF در SNRهای مختلف، نشان می‌دهد. طیف فرکانسی این نویز در شکل (۵-۶) آمده است.

باتوجه‌به جدول (۵-۳) مشاهده می‌گردد که:

- روش پیشنهادی در SNRهای پایین بهترین عملکرد را در مقایسه با سه روش دیگر دارد که حاکی از مقاوم بودن روش پیشنهادی در برابر نویز در مقایسه با روش‌های ارائه شده دیگر است.

- روش MGD در مقایسه با سایر روش‌ها بدترین عملکرد را داشته است.



شکل ۵-۶ طیف فرکانسی نویز کانال HF

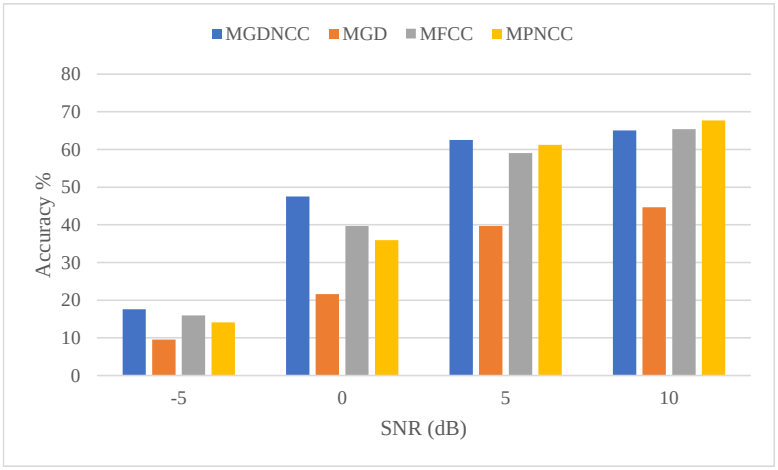
- در سیگنال به نویز ۱۰ دسی‌بل روش MPNCC نسبت به سایر روش‌ها عملکرد بهتری را از خود نشان داده است. همچنین در این مقدار نویز، عملکرد روش MFCC نیز از روش پیشنهادی تقریباً به‌اندازه ۰/۴درصد بهتر بوده است.
- روش پیشنهادی به‌صورت میانگین در SNRهای مختلف بهترین صحت را در مقایسه با سه روش دیگر داشته است.

جدول ۵-۳ صحت بازشناسی گفتار در نویز کانال HF

ویژگی‌ها	SNR (dB)				
	-۵	۰	۵	۱۰	میانگین
MGDNCC (پیشنهادی)	۱۷.۶۲	۴۷.۵۵	۶۲.۵۵	۶۵.۰۲	۴۸.۱۹
MGD	۹.۵۷۰	۲۱.۶۴	۳۹.۷۰	۴۴.۶۴	۳۰.۴۸
MFCC	۱۵.۹۴	۳۹.۶۷	۵۹.۰۸	۶۵.۳۷	۴۱.۹۲
MPNCC	۱۴.۰۹	۳۵.۹۷	۶۱.۲۶	۶۷.۷۲	۴۴.۷۶

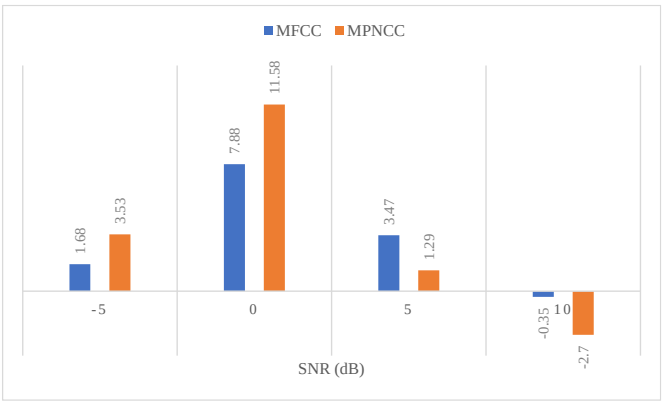
شکل (۵-۷) مقایسه بصری بهتری از عملکرد این چهار روش را ارائه می‌دهد. همان‌طور که در این شکل مشاهده می‌گردد، در شرایط نویزی بیشتر، روش پیشنهادی عملکرد بهتری را نسبت به روش‌های دیگر دارد. وبا کم‌شدن مقدار توان نویز، عملکرد روش‌های MFCC و MPNCC با شیب بیشتری

افزایش می‌یابد.



شکل ۵-۷ صحت بازشناسی گفتار در نویز کانال HF

شکل (۵-۸) میزان درصد بهبود روش پیشنهادی در تشخیص کلمه را در مقایسه با دو روش متداول MFCC و MPNCC، در SNRهای مختلف نشان می‌دهد. همان‌طور که در این شکل مشاهده می‌شود، بیشترین درصد بهبود را در ۰ دسی‌بل و کمترین درصد بهبود را در ۵ دسی‌بل، هر دو نسبت به MPNCC داشته است. همچنین در ۱۰ دسی‌بل عملکرد روش‌های MFCC و MPNCC نسبت به روش پیشنهادی بهتر بوده است.

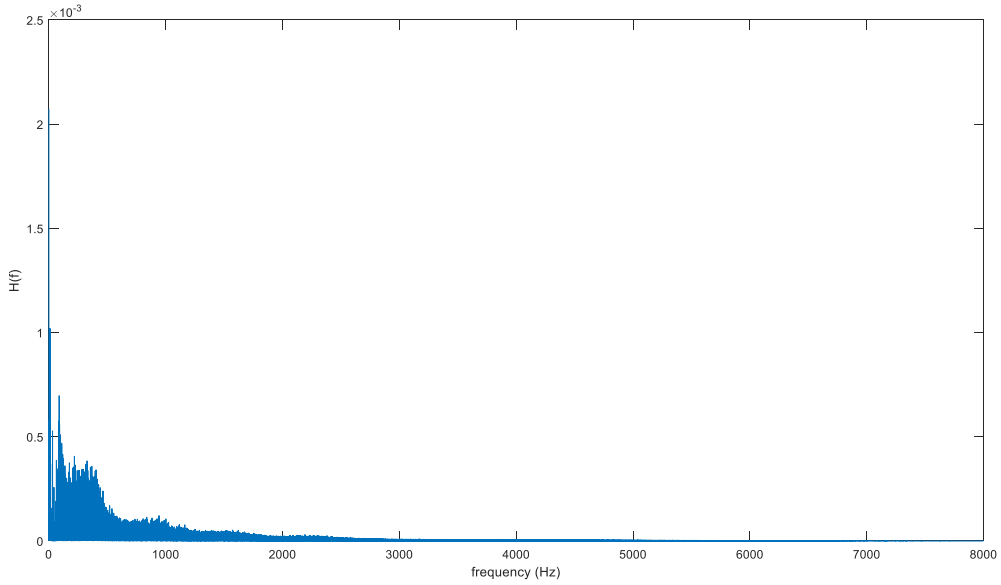


شکل ۵-۸ میزان بهبود عملکرد روش پیشنهادی نسبت به ویژگی های MFCC و MPNCC در نویز کانال HF

۵-۵- نتایج بازشناسی در نویز تداخل سیگنال

جدول (۵-۴) نتایج عملکرد سیستم بازشناسی گفتار با روش‌های مختلف استخراج ویژگی را بر حسب معیار صحت بازشناسی، در نویز تداخل سیگنال در SNRهای مختلف، نشان می‌دهد. طیف

فرکانسی این نویز در شکل (۵-۹) آمده است.



شکل ۵-۹ طیف فرکانسی نویز تداخل سیگنال

باتوجه‌به جدول (۵-۴) مشاهده می‌گردد که:

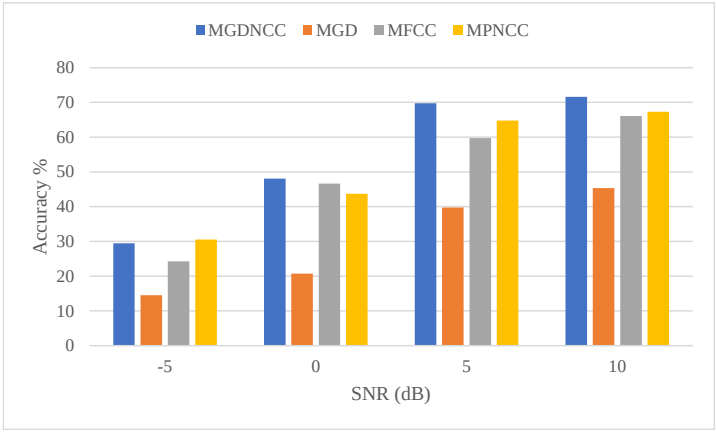
- در SNR ۵- دسی‌بل روش پیشنهادی نسبت به دو روش MCC و MGD عملکرد بهتری داشته است اما عملکرد آن نسبت به روش MPNCC به‌اندازه ۱ درصد کمتر بوده است.
- در تمامی SNRها به‌غیر از -۵ دسی‌بل، روش پیشنهادی بهترین عملکرد را داشته است.
- در تمامی SNRها روش MGD بدترین عملکرد را در مقایسه با روش‌های دیگر از خود نشان داده است.
- در حضور این نویز، روش پیشنهادی به‌صورت میانگین در مقایسه با سه روش دیگر بهترین عملکرد را داشته است.

شکل (۵-۱۰) مقایسه بصری بهتری از عملکرد این چهار روش را ارائه می‌دهد. همان‌طور که در این شکل مشاهده می‌گردد، در شرایط نویزی بیشتر، روش پیشنهادی عملکرد بهتری را نسبت به روش‌های دیگر به جز روش MPNCC دارد.

جدول ۵-۴ صحت بازشناسی گفتار در نویز تداخل سیگنال

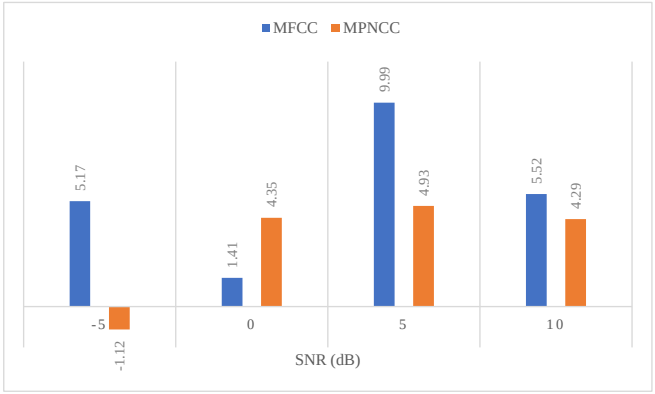
ویژگی‌ها	SNR (dB)				
	-۵	۰	۵	۱۰	میانگین
MGDNCC (پیشنهادی)	۲۹.۴۳	۴۸.۰۸	۶۹.۷۲	۷۱.۶۰	۵۴.۷۱
MGD	۱۴.۵۳	۲۰.۷۶	۳۹.۷۶	۴۵.۳۵	۳۰.۱۰
MFCC	۲۴.۲۶	۴۶.۶۷	۵۹.۷۳	۶۶.۰۸	۴۹.۱۹
MPNCC	۳۰.۵۵	۴۳.۷۳	۶۴.۷۹	۶۷.۳۱	۵۱.۶۰

شکل (۵-۱۱) میزان درصد بهبود روش پیشنهادی در تشخیص کلمه را در مقایسه با دو روش متداول MFCC و MPNCC، در SNRهای مختلف نشان می‌دهد.



شکل ۵-۱۰ صحت بازشناسی گفتار در نویز تداخل سیگنال

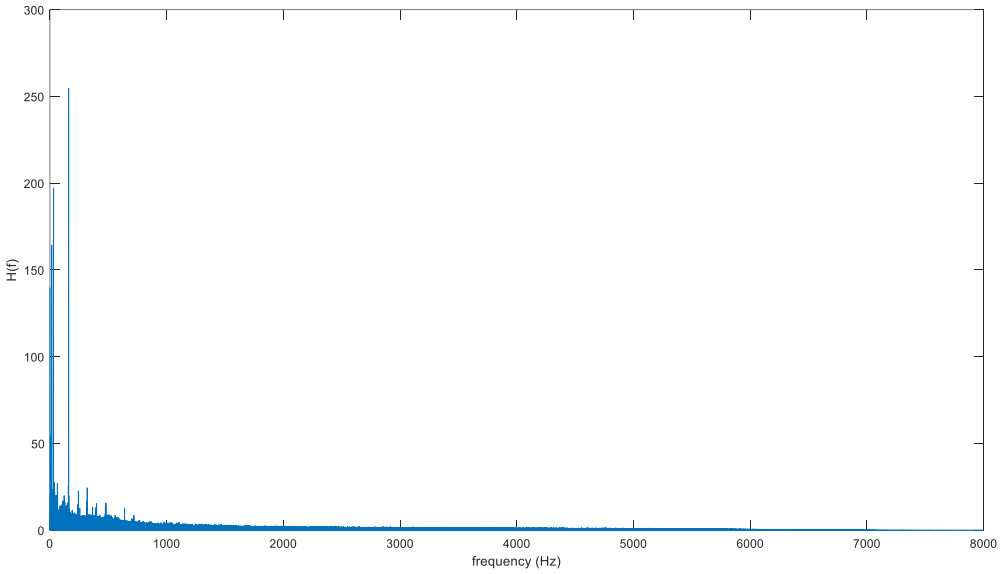
همان‌طور که در این شکل مشاهده می‌شود، بیشترین درصد بهبود را در ۵ دسی‌بل و کمترین درصد بهبود را در ۰ دسی‌بل، هر دو نسبت به MFCC داشته است. همچنین در ۵- دسی‌بل عملکرد روش MPNCC به میزان تقریباً ۱ درصد نسبت به روش پیشنهادی بهتر بوده است.



شکل ۵-۱۱ میزان بهبود عملکرد روش پیشنهادی نسبت به ویژگی های MFCC و MPNCC در نویز تداخل سیگنال

۵-۶- نتایج بازشناسی در نویز کارخانه

جدول (۵-۵) نتایج عملکرد سیستم بازشناسی گفتار با روش‌های مختلف استخراج ویژگی را بر حسب معیار صحت بازشناسی، در نویز کارخانه در SNRهای مختلف، نشان می‌دهد. طیف فرکانسی این نویز در شکل (۵-۱۲) آمده است.



شکل ۵-۱۲ طیف فرکانسی نویز کارخانه

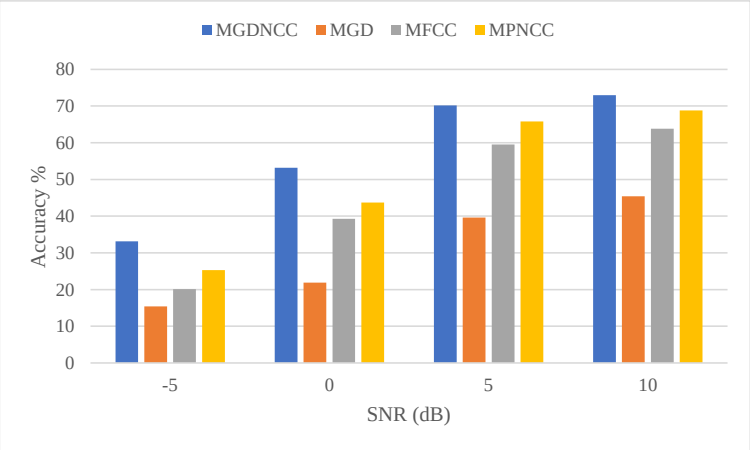
باتوجه‌به جدول (۵-۵) مشاهده می‌گردد که:

- روش پیشنهادی در تمامی حالت‌های تراز نویز، از روش‌های دیگر بهتر عمل کرده است.
- بعد از روش پیشنهادی، روش MPNCC بهترین عملکرد را در تمامی حالت‌های تراز نویز از خود نشان داده است.
- در SNR ۵- دسی‌بل روش متداول MFCC نسبت به دو روش MPNCC و روش پیشنهادی، عملکرد بهتری نداشته است.
- همان‌طور که انتظار می‌رود روش MGD در مقایسه با سه روش دیگر عملکرد ضعیفی را از خود نشان داده است. در این نویز، روش پیشنهادی هم به‌صورت تکی و هم به‌صورت میانگین بهترین عملکرد را در مقایسه با روش‌های دیگر داشته است.

جدول ۵-۵ صحت بازشناسی گفتار در نویز کارخانه

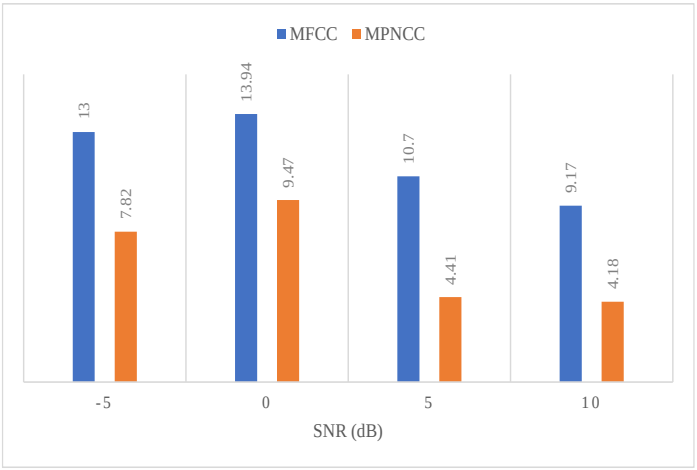
ویژگی‌ها	SNR (dB)				
	-۵	۰	۵	۱۰	میانگین
MGDNCC (پیشنهادی)	۳۳.۱۴	۵۳.۲۰	۷۰.۱۹	۷۲.۹۶	۵۷.۳۷
MGD	۱۵.۴۱	۲۱.۸۸	۳۹.۵۹	۴۵.۴۱	۳۰.۵۷
MFCC	۲۰.۱۴	۳۹.۲۶	۵۹.۴۹	۶۳.۷۹	۴۵.۶۷
MPNCC	۲۵.۳۲	۴۳.۷۳	۶۵.۷۸	۶۸.۷۸	۵۰.۹۰

شکل (۵-۱۳) مقایسه بصری بهتری از عملکرد این چهار روش را ارائه می‌دهد. همان‌طور که در این شکل مشاهده می‌گردد، در تمام حالات تراز نویز، روش پیشنهادی عملکرد بهتری را نسبت به روش‌های دیگر دارد.



شکل ۵-۱۳ صحت بازشناسی گفتار در نویز کارخانه

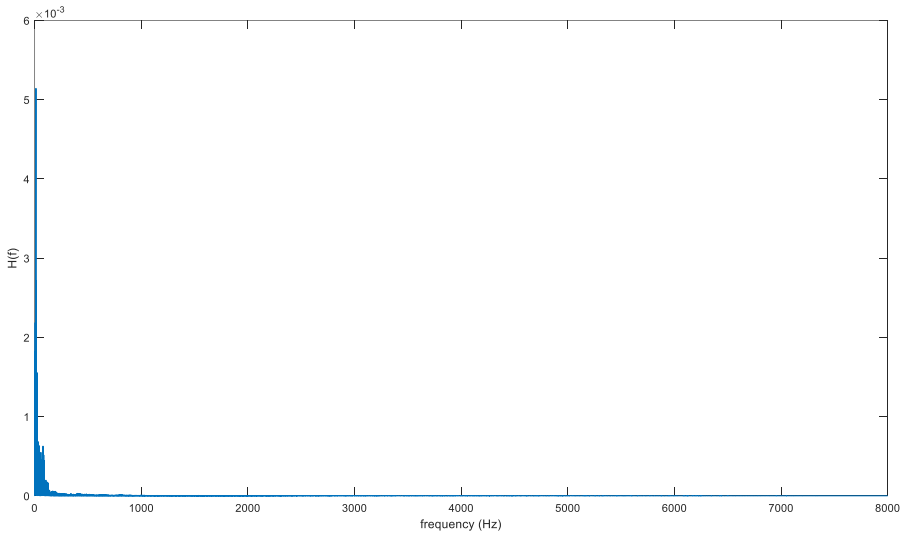
شکل (۵-۱۴) میزان درصد بهبود روش پیشنهادی در تشخیص کلمه را در مقایسه با دو روش متداول MFCC و MPNCC، در SNRهای مختلف نشان می‌دهد. همان‌طور که در این شکل مشاهده می‌شود، بیشترین درصد بهبود را در ۰دسی‌بل و کمترین درصد بهبود را در ۱۰ دسی‌بل، به ترتیب نسبت به روش‌های MFCC و MPNCC داشته است.



شکل ۵-۱۴ میزان بهبود عملکرد روش پیشنهادی نسبت به ویژگی های MFCC و MPNCC در نویز کارخانه

۵-۷- نتایج بازشناسی در نویز VOLVO

جدول (۵-۶) نتایج عملکرد سیستم بازشناسی گفتار با روش‌های مختلف استخراج ویژگی را بر حسب معیار صحت بازشناسی، در نویز VOLVO در SNRهای مختلف، نشان می‌دهد. طیف فرکانسی این نویز در جدول (۵-۱۵) آمده است.



شکل ۵-۱۵ طیف فرکانسی نویز VOLVO

باتوجه‌به جدول (۵-۶) مشاهده می‌گردد که:

- روش پیشنهادی در همه SNRها در مقایسه با روش‌های دیگر و بخصوص دو روش رایج MFCC و MPNCC بهترین عملکرد را از خود نشان داده است.
- دو روش MFCC و MPNCC در SNRهای مختلف عملکرد نزدیک به همی را

داشته‌اند.

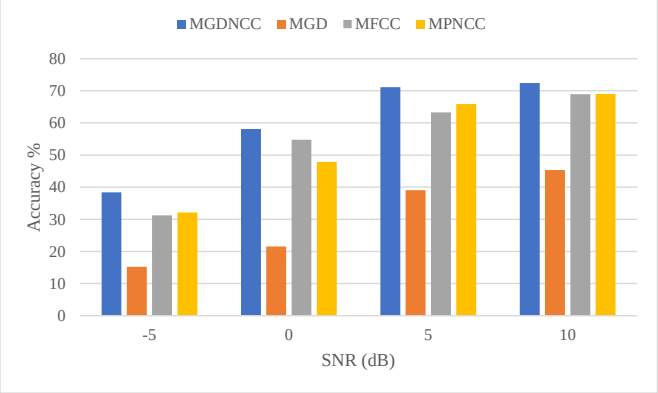
- روش MPNCC بعد از روش پیشنهادی، بهترین عملکرد را در تمامی ترازهای نویز به جز در ۰ دسی‌بل، داشته است.
- روش MGD در این نویز نیز نسبت به روش‌های دیگر عملکرد ضعیف‌تری را داشته است.
- صحت تشخیص سه روش MFCC، MPNCC و روش پیشنهادی در ۱۰ دسی‌بل نزدیک به هم بوده است.
- روش پیشنهادی هم به‌صورت تکی و هم به‌صورت میانگین بهترین عملکرد را در مقایسه با روش‌های دیگر داشته است. همچنین میانگین صحت روش MPNCC از روش MFCC به مقدار تقریباً ۱درصد بیشتر بوده است.

جدول ۵-۶ صحت بازشناسی گفتار در نویز VOLVO

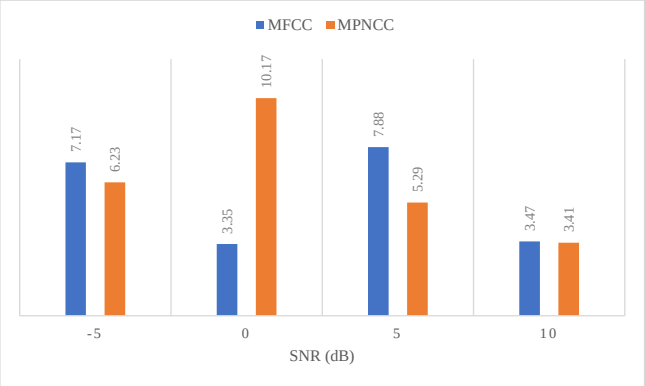
ویژگی‌ها	SNR (dB)				
	-۵	۰	۵	۱۰	میانگین
MGDNCC (پیشنهادی)	۳۸.۳۷	۵۸.۰۷	۷۱.۱۳	۷۲.۴۳	۶۰
MGD	۱۵.۲۳	۲۱.۵۳	۳۹.۰۶	۴۵.۳۵	۳۰.۲۹
MFCC	۳۱.۲۰	۵۴.۷۲	۶۳.۲۵	۶۸.۹۶	۵۴.۵۳
MPNCC	۳۲.۱۴	۴۷.۹۰	۶۵.۸۴	۶۹.۰۲	۵۳.۷۲

شکل (۵-۱۶) مقایسه بصری بهتری از عملکرد این چهار روش را ارائه می‌دهد. همان‌طور که در این شکل مشاهده می‌گردد، در تمام حالات تراز نویز، روش پیشنهادی عملکرد بهتری را نسبت به روش‌های دیگر دارد.

شکل (۵-۱۷) میزان درصد بهبود روش پیشنهادی در تشخیص کلمه را در مقایسه با دو روش متداول MFCC و MPNCC، در SNRهای مختلف نشان می‌دهد. همان‌طور که در این شکل مشاهده می‌شود، بیشترین درصد بهبود را در ۰ دسی‌بل و کمترین درصد بهبود را در ۱۰ دسی‌بل، هر دو نسبت به روش MPNCC داشته است.



شکل ۵-۱۶ صحت بازشناسی گفتار در نویز VOLVO



شکل ۵-۱۷ میزان بهبود عملکرد روش پیشنهادی نسبت به ویژگی های MFCC و MPNCC در نویز VOLVO

۵-۸- زمان استخراج ویژگی

در جدول (۵-۷) زمان استخراج ویژگی روش‌های مختلف آورده شده است. همانطور که مشاهده می‌شود زمان استخراج ویژگی روش پیشنهادی از سه روش دیگر بیشتر است که از پیچیدگی بیشتر آن ناشی می‌شود. علت این امر می‌تواند استفاده از MGD در روش پیشنهادی باشد.

جدول ۵-۷ زمان استخراج ویژگی روش‌ها

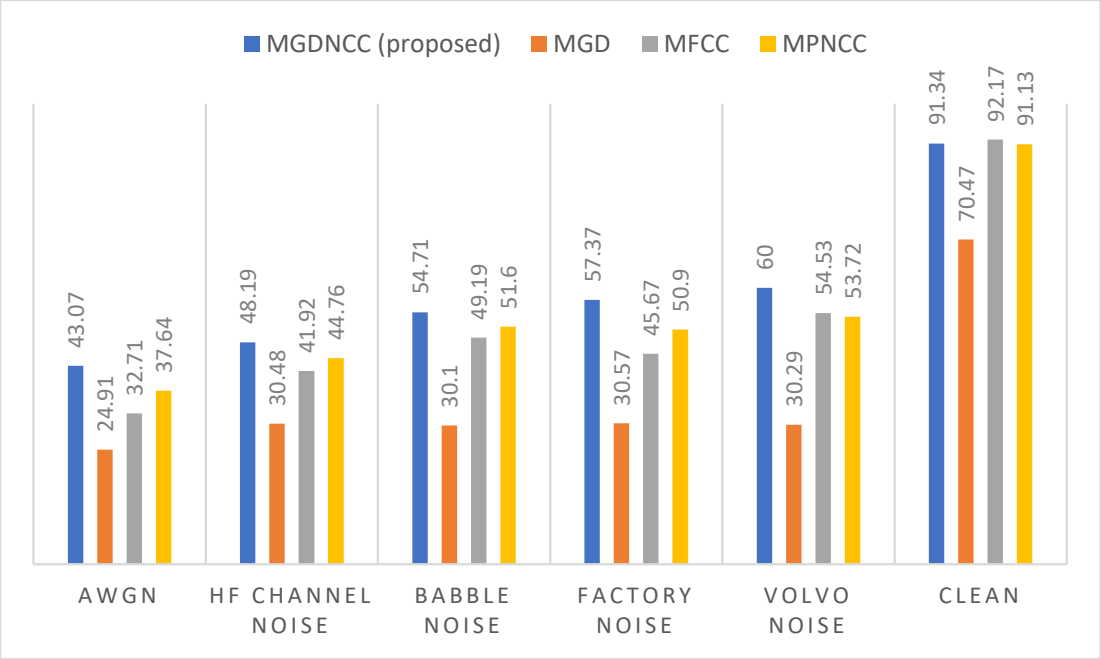
ویژگی‌ها	زمان استخراج ویژگی (میلی‌ثانیه)
MGDNCC (پیشنهادی)	۲.۲
MGD	۱.۵
MFCC	۱.۹
MPNCC	۱.۸

۵-۹- جمع‌بندی نتایج ارزیابی

نتایج عملکرد روش‌ها در شرایط بدون نویز و میانگین صحت شناسایی در SNRهای مختلف برای هرکدام از نویزها در شکل (۵-۱۸) آورده شده است.

همانطور که شکل (۵-۱۸) مشاهده می شود، در شرایط بدون نویز روش پیشنهادی از دو روش MPNCC و MGD عملکرد بهتری داشته است در حالی که روش MFCC بهتر از روش پیشنهادی عمل کرده است. در شرایط نویزی میانگین صحت شناسایی روش پیشنهادی در SNRهای مختلف برای همه نویز ها از سه روش MFCC، MPNCC و MGD بهتر بوده است.

باتوجه‌به نتایج به‌دست‌آمده، روش پیشنهادی جدید ارائه شده برای استخراج ویژگی کارایی بهتری در شرایط نویزی نسبت به روش‌های متداول دیگر از خود نشان می‌دهد.



شکل ۵-۱۸ میانگین صحت بازشناسی روش پیشنهادی و روش های MFCC، MPNCC و MGD در شرایط نویزی و بدون نویز

۵-۱۰- مقایسه با پژوهش‌های پیشین

در این بخش به مقایسه عملکرد روش پیشنهادی با برخی از پژوهش‌های انجام شده در سال‌های اخیر می‌پردازیم.

همان‌طور که در جدول (۵-۸) مشخص است، در شرایط بدون نویز روش پیشنهادی عملکرد ضعیفتری نسبت به روش MFCC در [۴۲] و [۴۳] داشته است. شایان‌ذکر است که عملکرد روش پیشنهادی با روش MFCC در [۴۲] قابل‌مقایسه بوده است. علت عملکرد ضعیف‌تر روش پیشنهادی را می‌توان تعداد کلاس بیشتر (تقریباً دوبرابر) در روش پیشنهای و استفاده از شبکه عصبی MLP معمولی در روش پیشنهادی دانست.

جدول ۵-۸ مقایسه عملکرد روش پیشنهادی با مراجع [۴۲] و [۴۳]

مرجع	ویژگی‌ها	تعداد کلاس‌ها	پایگاه داده	طبقه‌بند	نویز	صحت (٪)
[۴۲]	MFCC	۱۰	TIMIT	SANN	بدون نویز	۹۲/۰۱
[۴۳]	MFCC	۱۰	TIMIT	HMM	بدون نویز	۹۷/۰۲
پیشنهادی	MGDNCC	۱۸	TIMIT	MLP	بدون نویز	**۹۱/۳۴**

در جدول (۵-۹) عملکرد روش پیشنهادی با روش‌های به کار گرفته شده در [۲۲] مقایسه شده است. باتوجه‌به جدول (۵-۹) مشاهده می‌گردد که:

- در مرجع [۲۲] از پایگاه داده Aurora-2 که همان پایگاه‌داده TIDigits در حالت نویزی است، استفاده شده است. همچنین از طبقه‌بند HMM و از نویزهای مختلف واقعی استفاده شده است.
- در حالت بدون نویز روش پیشنهادی از دو روش MGDCC و MFMGDCC بهتر عمل کرده است و از دو روش MFCC و MFPSCC عملکرد ضعیف‌تری داشته است.
- در شرایط نویزی میانگین عملکرد روش پیشنهادی در تمامی نویزهای به کار گرفته در ۵- تا ۱۰ دسی‌بل از تمامی روش‌های استفاده شده در [۲۲] بهتر بوده است و از روش MFPSCC که بهترین عملکرد را در این مرجع داشته است، به میزان تقریباً ۱۵ درصد بهتر عمل کرده است.

جدول ۵-۹ مقایسه عملکرد روش پیشنهادی با مرجع [۲۲]

مرجع	پایگاه‌داده	تعداد کلاس‌ها	طبقه‌بند	ویژگی‌ها	صحت (٪)	
					بدون نویز	میانگین (۵- تا ۱۰ dB)
[۲۲]	Aurora-2	۱۰	HMM	MFCC	۹۶/۶۴	۳۶/۲۷
				MGDCC	۸۱/۳۲	۲۱/۸۲
				MFMGDCC	۵۵/۳۲	۲۳/۲۶
				MFPSCC	۹۶/۴۱	۳۷/۸۹
پیشنهادی	TIMIT	۱۸	MLP	**ویژگی**	**بدون نویز**	**میانگین (۵- تا ۱۰ dB) در ۵ نویز مورد آزمایش**
				MPNCC (پیشنهادی)	**۹۱/۳۴**	**۵۲/۶۶**

در جدول (۵-۱۰) عملکرد روش پیشنهادی با روش‌های به کار گرفته شده در [۳۹] مقایسه شده است. باتوجه‌به جدول (۵-۱۰) مشاهده می‌گردد که:

- پایگاه‌داده استفاده شده در [۳۹] TIDigits و نوع طبقه‌بند به کار گرفته شده، طبقه‌بند ترکیبی GMM-HMM بوده است.
- در حالت بدون نویز تمامی روش‌های استفاده شده در [۳۹] از روش پیشنهادی بهتر عمل کرده است.
- با وجود تقریباً دوبرابر بودن تعداد کلاس‌ها و ضعیف‌تر بودن طبقه‌بند در روش پیشنهادی، در هر سه نویز سفید گوسی، ماشین (VOLVO) و تداخل سیگنال، میانگین صحت روش پیشنهادی از دو روش MFCC و RASTA-PLP بهتر بوده است.
- روش پیشنهادی در دو نویز ماشین و تداخل سیگنال به‌صورت میانگین، از روش GFCC بهتر عمل کرده است و در نویز سفید گوسی عملکرد ضعیف‌تری داشته است.

- روش PNCC بهبودیافته ارائه شده در [۳۹] در هر سه نویز مورد آزمایش میانگین صحت بهتری نسبت به روش پیشنهادی داشته است.

جدول ۵-۱۰ مقایسه عملکرد روش پیشنهادی با مرجع [۳۹]

مرجع	تعداد کلاس‌ها	پایگاه‌داده	طبقه‌بند	ویژگی‌ها	صحت (٪)			
					بدون نویز	میانگین (۵- تا ۱۰ dB)		
						AWGN	CAR	BABBLE
[۳۹]	۱۰	TIDigits	GMM-HMM	MPNCC	۹۶	۷۲	۷۹/۲۵	۶۷/۷۵
				MFCC	۹۸	۳۵	۴۵/۷۵	۴۶/۷۵
				R-PLP	۹۷	۳۶/۷۵	۴۱/۲۵	۳۹/۷۵
				GFCC	۹۸	۵۰/۲۵	۵۳/۷۵	۵۳
پیشنهادی	۱۸	TIMIT	MLP	**MGDNCC (پیشنهادی)**	**۹۱/۳۴**	**۴۳/۰۷**	**۶۰**	**۵۴/۷۴**

فصل ششم

جمع‌بندی و پیشنهادها

۶-۱- جمع‌بندی

هدف اصلی در این پایان‌نامه بهبود عملکرد سیستم بازشناسی گفتار با استفاده از یک روش جدید استخراج ویژگی که از ویژگی‌های فازی سیگنال گفتار بهره می‌برد، است. در این تحقیق ابتدا در فصل اول به بررسی ساختار سیستم تشخیص گفتار و تأثیرگذاری بخش‌های مختلف آن در عملکرد آن پرداخته شد. سپس در فصل دوم یک مروری اجمالی بر ادبیات فاز و در مورد ظهور، زوال، و احیای پردازش فاز برای کاربردهای مختلف گفتار و دیدگاه‌های مختلف در مورد کارایی ویژگی فاز ارائه شد و برخی از کارهای برجسته اخیر انجام شده در این حوزه بررسی شد. در فصل سوم به مبانی نظری فاز سیگنال گفتار پرداخته شد و چالش‌های موجود دراین‌خصوص و راه‌حل‌های آن بررسی شد. سپس ویژگی‌های مبتنی بر فاز استخراج شده از سیگنال گفتار مانند تأخیر گروهی اصلاح شده و طیف حاصل‌ضرب تشریح شد. در نهایت استخراج ویژگی از سیگنال گفتار وبرخی از روش‌های متداول در این حوزه بررسی شد.

در فصل چهارم یک روش جدید استخراج ویژگی از سیگنال گفتار در تشخیص گفتار ارائه شد که در این روش ضرایب کپسترال نرمالیزه شده تابع تأخیر گروه اصلاح شده (MGDNCC) به‌عنوان ویژگی از سیگنال گفتار استخراج می‌شود. همچنین در این فصل یک سیستم بازشناسی گفتار پیشنهاد شد که در آن از یک شبکه عصبی MLP به‌عنوان طبقه‌بند استفاده می‌شود. سیستم بازشناسی و روش استخراج ویژگی پیشنهادی دارای یک سری پارامتر بهینه است که برای عملکرد بهتر سیستم لازم است که برای آنها مقداردهی مناسبی صورت گیرد. به همین منظور در فصل چهارم و در بخش پیاده‌سازی، به چگونگی به‌دست‌آوردن مقدار بهینه این پارامترها پرداخته شد و مقادیر بهینه به‌دست‌آمده، ارائه شدند.

پس از تشریح روش پیشنهادی، نتایج شبیه‌سازی در فصل پنجم برای ارزیابی عملکرد روش پیشنهادی در مقایسه با سایر روش‌های متداول در حوزه تشخیص گفتار ارائه شد. عملکرد روش پیشنهادی در تشخیص کلمه، با دو روش متداول MFCC و MPNCC که هر دو آنها ویژگی‌های

مبتنی بر اندازه هستند، و همچنین روش MGD در شرایط برابر مقایسه شد. آزمایش‌ها در شرایط بدون نویز و نویزی با اضافه‌کردن نویزهای سفید گوسی، کانال HF، تداخل سیگنال، کارخانه و VOLVO به پایگاه‌داده TIMIT با SNRهای ۵-، ۰، ۵ و ۱۰ دسی‌بل انجام شد. کارایی روش‌ها با معیار صحت، مورد ارزیابی قرار گرفته است.

نتایج به‌دست‌آمده مبین آن است که در شرایط بدون نویز روش پیشنهادی از دو روش MPNCC و MGD عملکرد بهتری داشته است درحالی‌که روش MFCC بهتر از روش پیشنهادی عمل کرده است. در شرایط نویزی میانگین صحت شناسایی روش پیشنهادی در SNRهای مختلف برای همه نویزها از سه روش MFCC، MPNCC و MGD بهتر بوده است. باتوجه‌به نتایج به‌دست‌آمده، روش پیشنهادی جدید ارائه شده برای استخراج ویژگی کارایی بهتری در شرایط نویزی نسبت به روش‌های متداول دیگر از خود نشان می‌دهد.

۶-۲- پیشنهادها برای کارهای آتی

جهت بهبود عملکرد سیستم‌های پردازش گفتار در کارهای آتی پیشنهادهای زیر مطرح می‌گردد.

- استفاده از طول گام کوچک‌تر برای محاسبه مقادیر بهینه برای پارامترهای ضروری جهت محاسبه تابع تأخیر گروهی اصلاح شده.
- استفاده از ضرایب کپسترال نرمالیزه شده طیف حاصل‌ضرب (PS) که عاری از محدودیت محاسبه مقادیر بهینه برای پارامترهای α و γ است و بصورت همزمان از هردو مشخصه اندازه و فاز طیف سیگنال گفتار بهره می‌برد.
- ارزیابی عملکرد سیستم تشخیص گفتار با استخراج ویژگی به روش ضرایب کپسترال فرکانس بارک تأخیر گروهی اصلاح شده.
- ترکیب ویژگی‌های استخراج شده به کمک روش پیشنهادی با MPNCC
- استفاده از طبقه‌بندهای دیگر جهت بهبود عملکرد سیستم تشخیص گفتار مانند مدل مخفی مارکوف، ماشین بردار پشتیبان، مدل ترکیبی HMM و ANN و...

- به‌کارگیری دیگر ویژگی‌های مبتنی بر فاز مانند تأخیر گروهی چیرپ جهت استخراج ویژگی به کمک روش ارائه شده در تشخیص گفتار.
- استفاده از روش استخراج ویژگی پیشنهادی در این تحقیق در دیگر زمینه‌های دیگر پردازش گفتار مانند تشخیص سن، تشخیص گفتار جعل شده، تشخیص نارسایی گفتار، تشخیص گوینده و...

منابع

[1] M. Malik, M. K. Malik, K. Mehmood and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications,* vol. 80, p. 9411–9457, November 2020.

[2] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE,* vol. 69, p. 529–541, 1981.

[3] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida and Y. Kawakami, "Spoofing Speech Detection Using Modified Relative Phase Information," *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, p. 660–670, June 2017.

[4] M. A. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *Proceeding IEEE, vol. abs/1001.2267, 2010.*

[5] S. K.Saksamudre, P. P. Shrishrimal and R. R. Deshmukh, "A Review on Different Approaches for Speech Recognition System," *International Journal of Computer Applications,* vol. 115, p. 23–28, April 2015.

[6] I. Mporas, T. Ganchev, M. Siafarikas and N. Fakotakis, "Comparison of Speech Features on the Speech Recognition Task," *Journal of Computer Science,* vol. 3, p. 608–616, August 2007.

[7] T. F. Kleinschmidt, "Robust speech recognition using speech enhancement,"Ph.D. dissertation, School of Queensland University of Technology, 2010.

[8] B. Zamani, A. Akbari, B. Nasersharif and A. Jalalvand, "Optimized discriminative transformations for speech features based on minimum classification error," *Pattern Recognition Letters,* vol. 32, p. 948–955, May 2011.

[9] W. Chan, N. Jaitly, Q. V. Le and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211,* 2015.

[10] X. Tang, "Hybrid Hidden Markov Model and Artificial Neural Network for Automatic Speech Recognition," in *Proceeding Pacific-Asia Conference on Circuits, Communications and Systems*, 2009.

[11] A. V. Haridas, R. Marimuthu and V. G. Sivakumar, "A critical review and analysis on techniques of speech recognition: The road ahead," *International Journal of Knowledge-based and Intelligent Engineering Systems,* vol. 22, pp. 39-57, 2018.

[12] P. Kaur, P. Singh and V. Garg, "Speech recognition system; challenges and techniques," *International Journal of Computer Science and Information Technologies,* vol. 3, p. 3989–3992, 2012.

[13] N. Ahmadi, "A comparative study of state-of-the-art speech recognition

models for english and dutch,"M.S. thesis, School of Groningen, 2020.

[14] N. Morgan, J. Cohen, S. H. Krishnan, S. Changand and S. Wegmann, "Ouch project (outing unfortunate characteristics of hmms)," in *Tech. Rep.*, International Computer Science Institute, 2013.

[15] P. Mowlaee, "Introduction: Phase Processing, History," in *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*, John Wiley & Sons, Ltd, 2016, p. 1–32.

[16] A. V. Oppenheim, J. S. Lim and S. R. Curtis, "Signal synthesis and reconstruction from partial Fourier-domain information," *Journal of the Optical Society of America*, vol. 73, p. 1413, November 1983.

[17] G. Duan and R. A. Morris, "The importance of phase in the spectra of digital type," *Electronic Publishing*, vol. 2, p. 47–60, 1989.

[18] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, p. 578–616, May 2007.

[۱۹] ع. عباسیان، "ترکیب ویژگی‌های دامنه و فاز جهت بهبود تشخیص آواها"، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود، ۱۳۸۷.

[20] R. M. Hegde, H. A. Murthy and G. V. R. Rao, "Application of the modified group delay function to speaker identification and discrimination," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[21] E. Loweimi and S. M. Ahadi, "A new group delay-based feature for robust speech recognition," in *Proceeding IEEE International Conference on Multimedia and Expo*, 2011.

[22] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Proceeding IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[23] J. Deng, X. Xu, Z. Zhang, S. Fruhholz and B. Schuller, "Exploitation of Phase-Based Features for Whispered Speech Emotion Recognition," *IEEE Access*, vol. 4, p. 4299–4309, 2016.

[24] S. Sehgal, S. Cunningham and P. Green, "Phase-based feature representations for improving recognition of dysarthric speech," in *Proceeding IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[25] J. Yang and L. Liu, "Playback speech detection based on magnitude–phase spectrum," *Electronics Letters*, vol. 54, p. 901–903, July 2018.

[26] J. Yang, H. Wang, R. K. Das and Y. Qian, "Modified Magnitude-Phase Spectrum Information for Spoofing Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, p. 1065–1078, 2021.

[27] M. Todisco, H. Delgado and N. Evans, "A New Feature for Automatic

Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients," in *Proceeding The Speaker and Language Recognition Workshop* , 2016.

[28] J. Yang, R. K. Das and H. Li, "Significance of Subband Features for Synthetic Speech Detection," *IEEE Transactions on Information Forensics and Security,* vol. 15, p. 2160–2170, 2020.

[29] Z. Feng, Q. Tong, Y. Long, S. Wei, C. Yang and Q. Zhang, "SHNU Anti-spoofing Systems for ASVspoof 2019 Challenge," in *Proceeding Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019.

[30] K. K. Jain, G. Kirthan M, V. R. Pai and C. Narendra K, "Speaker verification by combining information from magnitude and phase spectrum," in *Proceeding International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2016.

[31] E. Loweimi, "Robust phase-based speech signal processing from source-filter separation to model-based robust ASR,"Ph.D. dissertation, School of University of Sheffield, 2018.

[32] L. Liu, J. He and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication,* vol. 22, p. 403–417, September 1997.

[33] R. S. Alan Oppenheim, Discrete-Time Signal Processing: Pearson New International Edition, Pearson Education Limited, 2013.

[34] S. Chavez, Q.-S. Xiang and L. An, "Understanding phase maps in MRI: a new cutline phase unwrapping method," *IEEE transactions on medical imaging,* vol. 21, p. 966–977, 2002.

[35] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing,* vol. 40, p. 2281–2289, 1992.

[36] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proceeding IEEE International Conference on Acoustics, Speech, and Signal Processing,. Proceedings.(ICASSP'03).*, 2003.

[37] R. M. Hegde, H. A. Murthy and V. R. R. Gadde, "Significance of the Modified Group Delay Feature in Speech Recognition," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 15, p. 190–202, January 2007.

[38] G. Sharma, K. Umapathy and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics,* vol. 158, p. 107020, January 2020.

[39] M. Tamazin, A. Gouda and M. Khedr, "Enhanced Automatic Speech Recognition System Based on Enhancing Power-Normalized Cepstral Coefficients," *Applied Sciences,* vol. 9, p. 2166, May 2019.

[40] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition

using a power-law nonlinearity and power-bias subtraction," in *Proceeding Interspeech*, 2009.

[41] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[42] H.-N. Ting, B.-F. Yong and S. M. Mirhassani, "self-adjustable neural network for speech recognition," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 9, pp. 2022-2027, 2013.

[43] P. Paramonov, "Fast algorithm for isolated words recognition based on Hidden Markov model stationary distribution," in *IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 2017.

Abstract

Considering the extensive applications of speech recognition systems in different environmental conditions, strengthening and enhancing the performance of these systems against different types of noise is essential. One way to improve the performance of these systems is to strengthen the feature extraction methods from the speech signal. Most of the existing feature extraction methods in speech processing consider magnitude-based features and ignore the use of phase-based features of speech for various reasons.

The main purpose of this thesis is to use a new phase-based feature extraction method to improve the performance of a speech recognition system in noisy conditions. In this regard, we present a novel feature extraction method by changing the structure of the Modified Power-Normalized Cepstral Coefficients (MPNCC) and adding the Modified Group Delay (MGD) to it.

The performance of the proposed method is measured by utilizing a Multilayer Perceptron (MLP) neural network in MATLAB software and using the TIIMT database in order to do word recognition tasks in clean and noisy conditions. In order to compare the performance of the proposed method, two commonly used features in speech recognition, namely Mel-Frequency Cepstral Coefficients (MFCC) and MPNCC, are used under the same conditions. The experimental results demonstrate better recognition accuracy of the proposed method compared to the MPNCC method and is comparable to the MFCC method in noise-free conditions. Under noisy conditions with different SNRs, the accuracy of the proposed method was better than MFCC and MPNCC. For instance, the accuracy of the proposed method in white Gaussian noise at -5dB has improved by 14.8% and 9.7% compared to MFCC and MPNCC, respectively.

Keywords: Speech recognition, feature extraction, strengthening, phase, group delay, cepstral coefficients

Shahrood University of Technology

Faculty of Electrical Engineering

M.Sc. Thesis in Communication Systems Engineering

Improving the Efficiency of Speech Recognition system based on Phase and Magnitude features

By:

Mohammad Dolati

Supervisor:

Dr. Hossein Marvi

February 2022