تاریخ دریافت مقاله: ۹۸/۰۶/۱۱
تاریخ پذیرش مقاله: ۹۸/۱۲/۲۰

شناسایی گوینده در شرایط نوفه‌ای با استفاده از ویژگی‌های فیلتربانک گاماتون و تبدیل کسینوسی گسسته و قطبی

فرنوش عارفی*
کارشناس ارشد فناوری اطلاعات، پژوهشکدهٔ فضای مجازی - دانشگاه شهید بهشتی - تهران - ایران
پست الکترونیکی: farnoosh.arefi@gmail.com

بهزاد سعیدی
کارشناس ارشد مهندسی برق، دانشکده مهندسی برق - دانشگاه شهید بهشتی - تهران - ایران
پست الکترونیکی: b.saeedi@mail.sbu.ac.ir

چکیده

احراز هویت افراد بر مبنای صوت یکی از موارد مهم پژوهشی در دههٔ اخیر بوده است. این موضوع در حوزه‌هایی مانند ردیابی گویندگان، ورود امن گویندگان به دستگاه‌های مختلف و سایر سامانه‌های مربوط به پردازش گفتار، کاربرد فراوانی دارد. هدف از یک سیستم احراز هویت گوینده، تشخیص هویت افراد با استفاده از سیگنال صدای انسان می‌باشد. یکی از چالش‌های مهمِ موجود در حوزهٔ احراز هویت گوینده، افزایش کارایی این سیستم در شرایط نوفه‌ای شدید می‌باشد. در این پژوهش با استفاده از فیلتربانک گاماتون و ارائه ویژگی جدیدی از تبدیل‌های کسینوسی گسسته و قطبی، روشی برای افزایش کارایی سیستم احراز هویت گوینده در شرایط نوفه‌ای شدید طراحی شده است. مقایسه نتایج روش پیشنهادی با روش‌های موجود نشان می‌دهد، روش پیشنهادی توانسته است با دقت بیشتری، هویت افراد را در شرایط نوفه‌ای مختلف شناسایی کند. روش پیشنهادی به‌صورت کمی در محیط نوفهٔ سفید با نسبت سیگنال به نوفهٔ ۲۰، ۱۰ و ۵ به ترتیب ۸۹، ۷۷ و ۶۱ درصد دقت و در محیط نوفهٔ توأمان خیابانی و سفید با نسبت سیگنال به نوفهٔ ۲۰، ۱۰ و ۵ به ترتیب ۸۷، ۷۱ و ۵۱ درصد دقت داشته است. همچنین، دقت روش پیشنهادی در مقایسه با روش جدید ضرایب کپسترالِ فرکانسِ گاماتونِ بهبود یافته، به‌طور متوسط ٤درصد، افزایش یافته است.

واژه‌های کلیدی: احراز هویت گوینده، شرایط نوفه‌ای، فیلتربانک گاماتون، تبدیل کسینوسی گسسته و قطبی.

۱- مقدمه

سیستم احراز هویت گوینده، سیستمی است که از روی ویژگی‌های منحصربه‌فرد در سیگنال صدای انسان می‌تواند هویت آن فرد را شناسایی کند. احراز هویت گوینده به‌صورت کلی به دو حوزهٔ وارسی گوینده[1] و شناسایی گوینده[2] تقسیم‌بندی می‌شود. در حوزهٔ وارسی

1- Speaker verification
2- Speaker identification

* نویسندهٔ مسئول

گوینـــده، فرد یک‌بار ابتدا در سیســـتم احراز هویت گوینده ثبت‌نام کرده و مدل مشـــخصهٔ هویتی خـــود را در پایگاه داده‌ای ذخیره می‌کند. در هنـــگام ورود و یا آزمایش، فرد ادعا می‌کند شخص خاصی است که در سیستم قبلاً ثبت‌نام کرده اســـت، در ادامه سیستم وارسی هویت گوینده وظیفه دارد با تطبیق گفتار شخص گوینده با مدل مشخصهٔ هویتی ادعاشده، تشـــخیص دهد که آیا آن فرد همان کسی هست که ادعا می‌کند یا خیر. در سیســـتم شناسایی گوینده، فرد یک‌بار در سیســـتم احراز هویت گوینده، مدل مشـــخصهٔ هویتی خـــود را ثبت می‌کند ولی در هنـــگام آزمایش، فرد ادعایی مبنی بر این‌که شـــخص خاصی هست نمی‌کند، در ادامه سیستم شناسایی گوینده از روی مدل‌های مشخصهٔ هویتی گویندگان مختلف، باید تشـــخیص دهد این فرد چه کسی میباشد. از کاربردهای سیســـتم وارسی گوینده، به ورود امـــن[3] و از کاربردهای سیســـتم شناســـایی گوینده می‌توان به ردیابی گویندگان[4] اشاره کرد.

از منظری دیگر سیستم احراز هویت گوینده را می‌توان به‌صورت مستقل از متن و وابسته به متن تقسیم‌بندی کرد. در سیســـتم مستقل از متن، فرد در هنگام ثبت‌نام و ورود، کلمات یا جملاتی دلخواه و بـــدون کنترل می‌گوید ولی در سیستم وابسته به متن، کلمات و یا جملات در هنگام ثبت‌نام و ورود باید یکسان باشند. سیستم‌هایی که مستقل از متن هستند غالباً پیچیده‌تر و پرچالش‌تر از سیستم‌های وابسته به متن هســـتند زیرا کنترلی در نوع صحبـــت کردن افراد وجود ندارد.

با وجود چالش‌های زیاد در یک سیســـتم احراز هویت گوینده که بیشتر مربوط به تغییر صدای گوینده در هنگام ثبت‌نام و ورود می‌شـــود، مسئلهٔ وجود نوفه‌های[5] محیطی نیز بســـیار مهم اســـت. نوفه‌های محیطی سیگنال صدا را تخریب می‌کنند و باعث می‌شوند کارایی سیستم تشخیص گوینده با اشـــکال روبرو شود؛ بنابراین ایجاد یک سیستم تشـــخیص گویندهٔ کارا و مقاوم در برابر نوفه‌های محیطی می‌تواند در این محیط‌ها مفید واقع شود.

به‌صورت کلی می‌توان به دو صورت کارایی سیســـتم تشـــخیص گوینده را در صورت وجـــود نوفه افزایش داد. به عنوان اولیـــن راه‌حل می‌توان به اســـتفاده از فیلترهای مختلـــف بهبود کیفیت صدا [۱] مانند فیلترهای وفقی [٤-۲] و یا فیلترهای ثابت و سایر سیستم‌های بهبود و بازسازی صدا اشـــاره کرد [۹-٥]. دومین راه‌حل، عدم بهبود کیفیت سیگنال صدا و ارائه یک سیستم استخراج ویژگی مقاوم به نوفه، برای ثبت شناسهٔ هویتی سیگنال صدا می‌باشد.

راه‌حل دوم این مزیت را دارد که مستقل از این‌که عامل تخریب‌کنندهٔ سیگنال صدا دارای چه ماهیتی باشد، می‌تواند کارایی تشـــخیص گوینده را بالا ببـــرد، زیرا در این حالت، تمرکز طراحی ویژگی‌ها، به صورتی اســـت که نســـبت به سیگنال‌های نوفه‌ای مقاوم و نسبت به سیگنال گفتار انسان حساس خواهدبود.

در ایـــن پژوهش، هدف، ارائه یک سیســـتم شناســـایی گویندهٔ مستقل از متن، بدون اعمال فیلترهای بهبود کیفیت سیگنال می‌باشد. تمرکز این پژوهش استخراج ویژگی‌های مقاوم به نوفهٔ برای ثبت شناســـه هویتی منحصربه‌فرد از گویندگان مختلف می‌باشد، به صورتی که کارایی شناسایی گوینده در محیط‌های نوفه‌ای افزایش یابد.

نخستین نوآوری این پژوهش، ارائه یک ویژگی ترکیبی از تبدیل‌های کسینوســـی قطبی و گســـسته می‌باشـــد که می‌تواند مؤلفه‌های اصلی فرکانســـی گفتـــار گوینده را با مقاومت بالایـــی در محیط‌های مختلف نوفه‌ای اســـتخراج کرده و در فرآیند سیستم شناســـایی گوینده تأثیر بالایی ازنظر کارایی داشـــته باشـــد. نـــوآوری بعدی کـــه از این پژوهـــش حاصل شـــد، ارائه یک سیســـتم خالص‌ســـازی ویژگـــی[6] در کاربـــرد شناســـایی گوینـــده در محیط‌های نوفه‌ای اســـت. ازآنجاکه در سیستم‌های تشخیص گوینده ویژگی‌های زیادی انتخاب می‌شـــوند، نیاز بـــود که تأثیر مقادیـــر ویژگی‌های دورافتـــاده (پرت) و یا پرنوســـان به نحوی تقلیل یابد که اولاً سرعت بخش آموزش را بالا ببرد

3- Secure login
4- Speaker tracking
5- noises
6- Purification feature

و ثانیاً با ارائه یک ویژگی خالص‌شـــده از مقادیر پرنوسان در طول زمان، کارایی دقت تشخیص گوینده افزایش یابد. در این پژوهش از دو روش برای خالص‌ســـازی ویژگی‌ها استفاده شده اســـت. روش اول، استانداردسازی و پیچش ویژگی‌ها[7] می‌باشـــد که در آن مقادیـــر هر بعد از ماتریس بردار ویژگی به یک توزیع نرمال در طول بازه زمانی کوتاه تبدیل می‌شـــود و روش دوم استفاده از فیلترهای گاوسی دوبعدی بـــر روی ماتریس بردار ویژگی می‌باشـــد که در طی این فرآیند، نوســـانات بین مقادیرِ ویژگی‌های مختلف تا حد زیادی کاهـــش می‌یابد. مجموعـــه اقدامات صورت گرفته در خصوص طراحی ویژگی‌های مناســـب ترکیبی و خالص‌ســـازی ویژگی‌ها در این پژوهش باعث شده، روش پیشنهادی نسبت به سایر روش‌های مورد مقایسه، کارایی بالاتری از خود در محیط‌های نوفه‌ای شـــدید نشـــان دهد. روش پیشـــنهادی به‌صورت کمی در محیط نوفهٔ سفید با نســـبت سیگنال به نوفهٔ ۲۰، ۱۰ و ۵ به ترتیب ۸۹، ۷۷ و ۶۱ درصـــد دقت و در محیط نوفهٔ توأمان خیابانی و ســـفید با نســـبت سیگنال به نوفه ۲۰، ۱۰ و ۵ به ترتیب ۸۷، ۷۱ و ۵۱ درصد دقت داشته است. همچنین، دقت روش پیشنهادی در مقایسه با روش جدید ضرایب کپسترالِ فرکانسِ گاماتونِ بهبود یافته، به‌طور متوسط ٤درصد، افزایش یافته است.

در بخـــش دو به مـــرور کارهـــای پیشـــین در حوزهٔ شناسایی گوینده پرداخته خواهد شد، در بخش سه روش پیشنهادی ارائه و در بخش چهار نتایج آزمایشگاهی بیان خواهند شد. در بخش پنج نیز نتیجه‌گیری کلی از پژوهش و کارهای آینده شرح داده خواهدشد.

۲- مرور کارهای پیشین

در دهـــهٔ اخیر بـــا معرفی مدل‌های پنهـــان مارکوف و مدل‌هـــای مخلوط گاوســـی [۱۱و۱۰] و مدل‌هـــای آماری زبـــان [۱٤-۱۲]، عملکـــرد سیســـتم پـــردازش گفتـــار در محیط‌های عاری از نوفه، رشـــد زیادی پیدا کرده اســـت. قدیمی‌ترین روش‌هـــای احراز هویت گوینـــده، روش‌های ضرایب کپســـترال فرکانس مِل[8] (MFCC) [۱۵] و پیش‌بینی ادراکی خطـــی[9] (PLP) [۱٦] بوده‌اند. این دو روش با این‌که متداول‌ترین و پایه‌ای‌ترین روش‌های موجود در تشخیص گوینده هســـتند، ولـــی تحقیقاتی مثل [۱۷] نشـــان داد، این روش‌ها در محیط‌هایی با نوفهٔ شدید، کارایی بسیار پایینی خواهند داشـــت. سیستم‌های تشـــخیص هویت گوینده در تحقیقات پیشـــین، مانند روش‌های MFCC [۱۵]، PLP [۱٦]، ضرایب کپســـترال توان نرمالیزه شـــده[10] [۱۸](PNCC) و ضرایب کپســـترال فرکانس گاماتون[11] [۱۹] (GFCC)، برای اســـتخراج ویژگی‌های فرکانسی منحصربه‌فرد از سیگنال صدای انسان از مؤلفه‌ای به نام فیلتربانک استفاده می‌کنند.

فیلتربانک‌ها بر اســـاس سیســـتم درک شنوایی انسان، مدل‌ســـازی شـــده‌اند. محققان این حـــوزه تأکیـــد دارند، همان‌گونه که انســـان با یک‌بار شنیدن صدای فرد می‌تواند در تکرارهای آینده آن فرد را تشـــخیص دهد، با ســـاختن یک سیســـتم شبیه‌سازی‌شده مشـــابه با ساختار سیستم شـــنوایی انســـان که فیلتربانک نام دارد، می‌تـــوان فرآیند تشخیص گوینده را با کارایی بیشـــتری انجام داد. در یک فیلتربانک، مشـــابه با ساختار شـــنوایی انسان، فیلترهایی مناســـب با فرکانس‌های شـــنوایی انســـان وجود دارد که هرکـــدام از این فیلترهـــا جنبه‌ای خـــاص از فرکانس‌های موجود در سیگنال صدا را بررسی می‌کنند. فیلتربانک مِل در MFCC [۱۵]، فیلتربانک بارک[12] در PLP [۱٦] و فیلتربانک گاماتـــون [۲۱و۲۰] در روش‌هـــای PNCC [۱۸] و GFCC [۱۹]، مرسوم‌ترین فیلترهای تشخیص گوینده می‌باشند. در فیلتربانک‌ها به‌صورت متداول هرچه به سمت فرکانس‌های بالاتر پیمایش می‌شـــود، عرض فیلترها بزرگ‌تر و ضریب آن‌ها کم می‌شـــود. زیرا در سیســـتم شـــنوایی انسان نیز، اهمیت فرکانس‌های بالا از فرکانس‌های پایین کمتر اســـت. سیگنال صدا با هم‌آمیزی[13] در فیلترهای مختلف، وزن‌هایی

7- Feature wrapping

8- Mel frequency cepstral coefficients
9- Perceptual linear prediction
10- Power normalized cepstral coefficients
11- Gammatone frequency cepstral coefficients
12- Bark filterbank
13- convolution

از آن فیلتر دریافت می‌کنـد که با کنار هم قرار دادن آن‌ها می‌توان به یک ویژگی منحصربه‌فرد از سـیگنال صحبت گوینده رسید.

از فیلتربانـک گاماتون پیش‌تر در پـردازش گفتار در محیط‌های نوفه‌ای استفاده شده اسـت. به‌طور مثال [۲۲] با اسـتخراج ویژگی از فیلتربانک گاماتون و اسـتفاده از الگوریتم جداسـاز خطی، نشان داد اسـتفاده از فیلتربانک گاماتون نسـبت به فیلتربانک بارک و مِل می‌تواند کارایی مناسـبی داشـته باشـد، منتها در این پژوهش تأثیر این فیلتربانک در شرایط نوفه‌ای به‌صورت جامع بررسی نشد، چـاو در سـال ۲۰۰۸ [۲۳]و ۲۰۰۹ [۲٤] در طی تحقیقاتی جامع، با ادامه دادن همان رهیافت روش MFCC [۱۵] و تنها با تغییر فیلتربانک گاماتون به‌جای فیلتربانک مِل از ویژگی فیلتربانک گاماتون جهت تشـخیص گوینده در محیط‌های نوفه‌ای استفاده کرد. در این پژوهش، آزمایش‌های متعددی از کارایـی این فیلتربانک در محیط نوفه‌ای با شـدت‌های مختلف صورت گرفته است، در اکثر آزمایش‌های صورت گرفته فیلتربانـک گاماتون در محیط‌های نوفه‌ای کاراتر از فیلتربانـک مل بوده اسـت. علت این امر این اسـت که این فیلتربانک با بررسی روان‌شناختی و فیزیولوژیکی سیستم شـنوایی انسان طراحی و توسعه داده شده است و نسبت به فیلتربانک مل، محققان بر جنبه‌های بیشتری از سیستم شـنوایی انسان تمرکز کرده‌اند. در [۲۵] روش پیاده‌سازی جدیدی برای فیلتربانک گاماتـون در حوزهٔ زمان در نظر گرفته شـد که سـریع‌تر از اعمال فیلتر گاماتون در حوزهٔ فرکانس عمـل می‌کند، در این پیاده‌سـازی عملیات تبدیل به حوزهٔ فرکانس وجود ندارد و به همین دلیل به سـرعت محاسـبات افزوده می‌شود و ویژگی‌های جدیدی از حوزهٔ زمان هم استخراج می‌شود که می‌تواند طبق ادعای نویسنده کارایـی را تا حدودی افزایش دهد. در بخش ۳-۱، جزئیات بیشتری از فیلتربانک گاماتون بررسی شده است.

با وجود کارایی بهتر فیلتربانک گاماتون نسبت به سایر فیلتربانک‌هـا در محیط‌های نوفه‌ای، تنها اسـتفاده از این فیلتربانک نمی‌تواند کارایی شناسایی گوینده در محیط‌های نوفه‌ای بسـیار شـدید را افزایش دهد، بلکه تغییر فضای ویژگی‌های حاصل از فیلتربانک‌هـا و انتخاب ویژگی‌های جدیـد در این فضـا مهم‌ترین عامل در کارایی تشـخیص گوینده می‌باشد[۱۸]

مؤلفهٔ دوم پراهمیت در سیستم‌های تشخیص گوینده، بازنمایـی ضرایب پراهمیت حاصل از هم‌آمیزی سـیگنال صدا با فیلتربانک‌ها، در مقیاسـی فشـرده می‌باشد. هدف از ایـن مؤلفه، اسـتخراج ضرایب پراهمیـت از فیلترهای موجود و حذف ضرایب غیراساسـی می‌باشد. با این کار ویژگی‌های به‌دست‌آمده ناهمبسته‌تر شده فرآیند یادگیری باقدرت بیشـتری صورت می‌گیرد. یکی از پرکاربردترین روش‌هـا برای اسـتخراج ضرایب اساسـی در سـیگنال فیلترشده، استفاده از ضرایب حاصل از تبدیل کسینوسی گسسته[14] (DCT) می‌باشـد، تبدیل DCT، ضرایب پراهمیت سـیگنال فیلترشده را به ترتیب مرتب کرده و مقدار اهمیت هر یک را مشخص می‌کند.

در سـال‌های اخیر، توجه به کارایی سیستم تشخیص گوینده در محیط‌های نوفه‌ای بیشتر شده است. روش‌هایی ماننـد PNCC [۱۸]، GFCC [۱۹] و GFCC بهبودیافتـه [۲٦] (GFCC-IMP) نمونـه‌ای از ایـن روش‌ها هسـتند. روش PNCC [۱۸] در سـال ۲۰۰۹، توسط [۲۷] ارائه شد. نسخهٔ بهبودیافتهٔ این روش کـه دارای ویژگی‌های بهتر و کاراتر می‌باشـد در سال ۲۰۱٦ ارائه شـد [۱۸]. مهم‌ترین ویژگی PNCC [۱۸]، جایگزینی توان غیرخطی به‌جای لگاریتم خطی می‌باشـد که پیش‌تر در روش MFCC [۱۵] اسـتفاده شده است. هم‌چنین این روش با ایجاد فیلترهای نامتقارن باعث کاهش نوفه‌های زمینه می‌شـود. نتایج منتشرشده در این پژوهش نشـان می‌دهد، اسـتفاده از این روش در مقایسه بـا MFCC [۱۵] و PLP [۱٦]، بسـیار کاراتر در محیط‌های نوفه‌ای می‌باشد.

روش GFCC [۱۹] نخسـتین بـار در سـال ۲۰۱۲ و نسـخهٔ بهبودیافتهٔ آن توسـط گروهی دیگـر یعنی روش

14- Discrete cosine transform

GFCC-IMP در سال ۲۰۱۶ بهبود پیدا کرده است [۲۶]. این دو روش مانند روش PNCC [۱۸] از فیلتربانک گاماتون برای تحلیل فرکانسی موجود در سیگنال صدا استفاده می‌کنند. استفاده از این فیلتربانک در مقایسه با فیلتربانک مِل نشان می‌دهد، این روش، ویژگی‌های متمایزکنندهٔ بیشتری در فرکانس‌های پایین سیگنال صدا بازنمایی می‌کند که در تشخیص گوینده بسیار کارا خواهدبود. روش GFCC [۱۹] همانند روش MFCC [۱۵] پایه‌گذاری شده است و تنها در بخش فیلتربانک، تغییراتی مناسبی در جهت بهبود دقت ایجاد شده است. روش GFCC-IMP [۲۶] در امتداد روش GFCC [۱۹] بوده ولی بعد از اعمال تبدیل کسینوسی گسسته، عملیات‌هایی مانند استانداردسازی داده‌ها و همچنین پیچش ویژگی‌ها، برای کاهش اثر نوفهٔ انجام می‌دهد. از پیچش ویژگی‌ها برای نرمالیزه‌کردن داده‌ها در یک بازهٔ زمانی خاص استفاده می‌شود، با استفاده از پیچش ویژگی‌ها اثر نوفه‌های موجود در سیگنال صدا، کاهش پیدا کرده و عملیات احراز هویت با قدرت بیشتری صورت می‌گیرد.

مدل مخلوط گاوسی[15] و مدل مخفی مارکوف[16] دو تا از مرسوم‌ترین روش‌ها برای آموزش ویژگی‌های سیستم تشخیص گوینده در تحقیقات پیشین بوده‌اند[۳۱-۲۸]. روش مدل مخلوط گاوسی با مدل‌سازی هر بعد از ویژگی‌ها با یک تابع گاوسی، سعی می‌کند مقادیر اساسی در اجزای تشکیل‌دهندهٔ آن بعد را مدل‌سازی کند و به نوفه‌های موجود در ویژگی‌ها مقاومت بالایی از خود نشان می‌دهد. در تحقیقاتی مانند [۱۷] نشان داده شده است روش آموزش مدل مخلوط گاوسی در سیگنال‌های نوفه‌ای نسبت به مدل مخفی مارکوف کارایی بالاتری در دسته‌بندی گویندگان مختلف داشته است.

روش‌های ارائه‌شده، مانند روش PNCC [۱۸]، دارای محاسبات پیچیدهٔ بالایی هستند و روش‌هایی مثل GFCC [۱۹] و GFCC-IMP [۲۶] با این‌که محاسبات کمتری دارند و در محیط‌های نوفه‌ای کاراتر هستند، اما قابلیت بهبود بیشتری در آن‌ها برای افزایش کارایی وجود دارد. هدف از این پژوهش، طراحی ویژگی‌های مقاوم به نوفه از سیگنال صدا می‌باشد که بتواند کارایی تشخیص گوینده را در محیط‌های نوفه‌ای افزایش دهد. روش پیشنهادی، با در نظرگرفتن توأمان ویژگی‌های تبدیل کسینوسی گسسته و تبدیل کسینوسی قطبی[17] (PCT) و طراحی یک سیستم خالص‌سازی ویژگی مناسب با حوزهٔ شناسایی گوینده، توانسته دقت بهتری در شناسایی گوینده در مقایسه با سایر روش‌های دیگر رقم بزند.

۳- روش پیشنهادی

مراحل کلی سیستم شناسایی گوینده در این پژوهش در شکل ۱ نشان داده شده است. فرآیند این پژوهش به دو قسمت آموزش و آزمایش تقسیم‌بندی می‌شود.

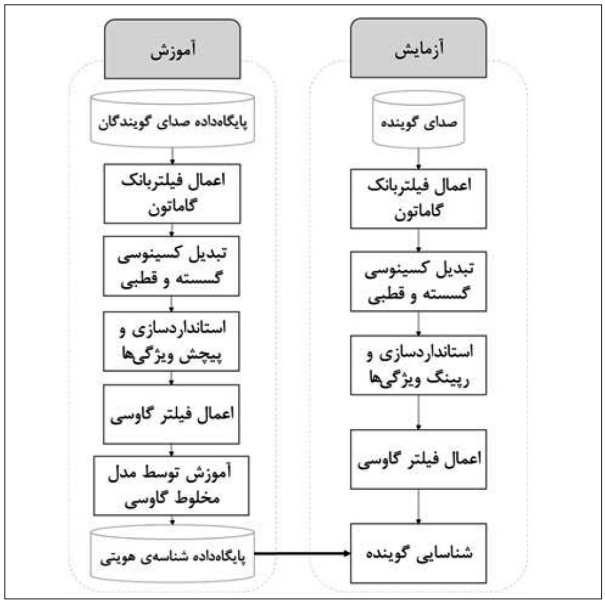
در قسمت آموزش ابتدا پایگاه داده‌ای از گویندگان مختلف تهیه می‌شود. سپس سیگنال صدای هر گوینده در فیلتربانک گاماتون هم‌آمیزی می‌شود. در ادامه از ماتریس سیگنال فیلترشده جهت استخراج ویژگی، تبدیل‌های کسینوسی گسسته و قطبی گرفته می‌شود. ماتریس بردار ویژگی حاصل شده، بعد از فرآیند استانداردسازی و پیچش توسط مدل مخلوط گاوسی آموزش می‌بیند و در پایگاه دادهٔ شناسهٔ هویتی ذخیره می‌شود.

در مرحلهٔ آزمایش، از سیگنال صدای گوینده، مشابه آن چیزی که در مرحلهٔ آموزش بیان شد، ماتریس بردار ویژگی ساخته می‌شود و با مقایسه و جستجو در پایگاه دادهٔ شناسهٔ هویتی، هویت فرد شناسایی می‌شود. در ادامه هر یک از فازهای مرحلهٔ استخراج ویژگی و آموزش و آزمایش بیان می‌شوند.

۳-۱- اعمال فیلتربانک گاماتون بر روی سیگنال صدا

در این مرحله، فیلترهای گاماتون F با فرکانس‌های مشخص در سیگنال صدای S (طول سیگنال N، S فرض

15- Gaussian mixture model
16- Hidden markov model
17- Polar cosine transform



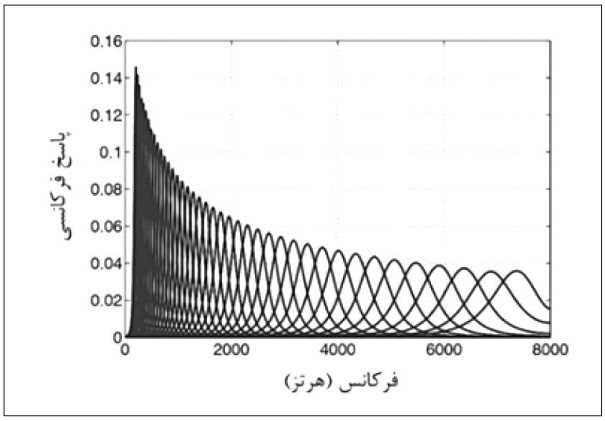
شکل ۱: فرآیند سیستم شناسایی گوینده در این پژوهش

می‌شود) هم‌آمیزی می‌شوند. پیش‌تر در مورد فیلتربانک‌ها و دلیل استفاده از آن‌ها توضیح داده شد. در [۳۲] نشان داده شده است فیلتربانک گاماتون در برابر نوفه‌ها نسبت به فیلتربانک مثلثی مِل مقاوم‌تر است. علت کارایی بالاتر فیلتربانک گاماتون، استفاده از فیلترهای نرمال‌شده با هم‌پوشانی بالا می‌باشد که وقتی در مقادیر پر نوسان سیگنال صدا هم‌آمیزی می‌شود، این مقادیر را به متوسط نمونه‌ها نزدیک می‌کند و در حقیقت اثر این مقادیر را کاهش می‌دهد، بنابراین استفاده از این فیلتربانک در وجود نوفهٔ بسیار مفید خواهدبود. در شکل ۲ برخی از فیلترهای این فیلتربانک قابل‌مشاهده است. محدودهٔ فرکانسی این فیلتربانک از ۲۰۰ هرتز تا ۸۰۰۰ هرتز تعریف می‌شود [۳۳]. در این پژوهش از ۳۲ فیلتربانک جهت استخراج ویژگی‌های فرکانسی مختلف استفاده شده است.

برای اعمال این فیلتر بر روی سیگنال صدا، ابتدا سیگنال صدا به حوزهٔ فرکانس تبدیل می‌شود و سپس با هر فیلتر F در این حوزه ضرب می‌شود. در رابطهٔ ۱ نحوهٔ ساخت فیلتربانک گاماتون در حوزهٔ فرکانس و در رابطهٔ ۲ نحوهٔ اعمال این فیلتربانک به سیگنال S مشخص شده است.

$$\cos(2\pi f_c t + h)\, e^{j\omega t}\partial t \qquad (1)$$

$$Gammaton_F = \int_{-\infty}^{+\infty} At^{(F-1)} e^{-2\pi Bt}$$

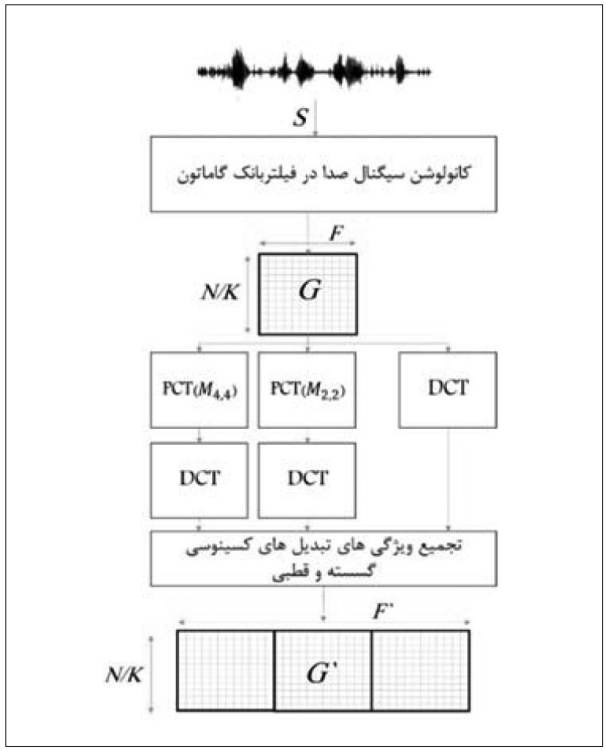


شکل ۲: فیلتربانک گاماتون در حوزهٔ فرکانس

$$G_{[N\times F]} = \{S \otimes Gammaton_F, F = 1, \ldots, 32\} \qquad (2)$$

در رابطهٔ ۱، A متغیر دامنه، fc فرکانس مرکزی، h جابه‌جایی فاز و B نمایانگر طول مدت‌زمان پاسخ ضربه می‌باشد. هم‌آمیزی حاصل از سیگنال صدا در فیلتربانک گاماتون یک ماتریس F×N می‌باشد که هر سطر آن، میزان وجود مؤلفه‌های فرکانسی فیلترهای گاماتون را در سیگنال صدا نشان می‌دهد. با توجه به این‌که که تعداد نمونه‌های سیگنال صدا زیاد می‌باشد، فرآیند آموزش با کندی پیش خواهدرفت. برای افزایش سرعت محاسبات و حذف نمونه‌های تکراری یا غیرتاثیرگذار از ماتریس F×N عملیات نمونه‌برداری با گام مشخص K صورت می‌گیرد. به این صورت اندازهٔ ماتریس فیلترشدهٔ صدا به ماتریسی به ابعاد F×(N/K) کاهش می‌یابد. در این پژوهش بر اساس آزمایش‌های مختلف مقدار K=100 در نظر گرفته شده است. پارامتر دیگر اعمال‌شده در این بخش، اعمال نرخ غیرخطی سطحی[18] به ماتریس بردار ویژگی می‌باشد. با توجه به این‌که انسان فرکانس‌های صدا را به‌صورت غیرخطی درک می‌کند، محققان مدل‌های مختلفی برای وزن‌دهی فرکانس‌ها جهت نزدیک شدن به درک شنوایی انسان، مستقل از فیلتربانک‌ها طراحی کرده‌اند [۱۸]. در این پژوهش مانند پژوهش [۲۶] از توان ۱/۳ جهت اعمال نرخ غیرخطی سطحی استفاده شده است. به‌صورت واضح‌تر ماتریس G به توان ۱/۳ می‌رسد.

18- Rate level nonlinearity



شکل ۳: استخراج مقادیر پراهمیت از ماتریس صدای فیلترشده G

۳-۲ تبدیل کسینوسی گسسته و قطبی

در این پژوهش از الگویی جدید برای استخراج مقادیر پراهمیت از ماتریس صدای فیلترشده G استفاده شده است. در شکل ۳، این فرآیند با جزئیات کامل قابل‌مشاهده است. با تبدیل DCT مؤلفه‌های پراهمیت فرکانسی به ترتیب، مرتب می‌شوند. بنابراین می‌توان مؤلفه‌هایی که در ساختن ماتریس فیلترشده، نقش بیشتری داشتند را تنها انتخاب کرد. در این پژوهش از تصویرپایهٔ ۳۲*۳۲ برای اعمال DCT استفاده شده است. در این بخش، علاوه بر تبدیل مستقیم DCT از تبدیل PCT نیز استفاده می‌شود. تبدیل PCT مانند سایر تبدیل‌های ریاضی دارای یک سری تصویرپایه می‌باشد که هر تصویری با هم‌آمیزی شدن در این تصاویرپایه می‌تواند به این فضا تبدیل شود. PCT یک استخراج‌کنندهٔ بسیار قوی بوده که پیش‌تر درزمینهٔ تشخیص اشیا و پردازش تصویر استفادهٔ زیادی از آن شده است [۳۶-۳۴]. وجه تمایز این تبدیل‌کننده نسبت به سایر تبدیل‌کننده‌ها این است که می‌تواند داده‌ها را با مقاومت بالایی در برابر نوفه و مستقل از جابه‌جایی و چرخش توصیف کند. در داده‌های صوتی همواره تغییر نوع صحبت گوینده وجود دارد، این توصیف‌گر به دلیل مقاوم بودن در برابر جابه‌جایی و چرخش، تغییرات جزئی در جابه‌جایی مقادیر فرکانسی صدا را نادیده می‌گیرد و توصیف دقیقی از ویژگی‌های اصلی دادهٔ صوتی ارائه می‌کند.

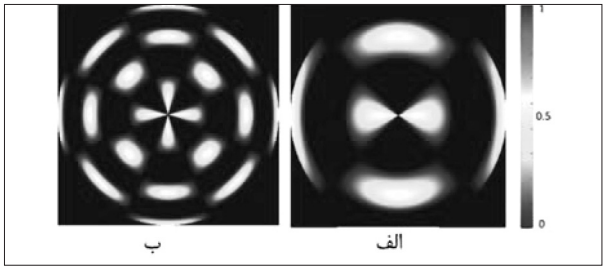
برای اعمال PCT بر روی ماتریس G، دو تصویر پایه به شکل ٤ در نظر گرفته شد. در این شکل هر چه به نواحی پررنگ‌تر پیمایش می‌شود، ضریب متناظر در ماتریس بردار ویژگی در فضای جدید قطبی مقدار بیشتری خواهد گرفت و هرچه از به نواحی کم‌رنگ‌تر پیمایش شود، این ضریب کاهش خواهدیافت. درواقع نواحی کم‌رنگ‌تر، مقادیر پر نوسان از ماتریس بردار ویژگی در فضای قطبی هستند که ضریبی کمتری از مقادیر اساسی در این فضای جدید خواهند گرفت. در تبدیل PCT می‌توان تصویرپایه‌های متعددی تولید کرد، بعد از آزمایش‌های مختلف دو تا از بهترین تصویرپایه‌ها به شکل ٤ انتخاب شدند. هر یک از این تصویرپایه‌ها که اندازهٔ آن‌ها ۳۲×۳۲ می‌باشد در ماتریس G هم‌آمیزی شده و سپس از نتایج آن مجدداً DCT گرفته می‌شود. در رابطهٔ ۳، نحوهٔ تولید تصویرپایه PCT و در رابطهٔ ٤، نحوهٔ اعمال تبدیل کسینوسی گسسته و قطبی به ماتریس فیلترشدهٔ G قابل مشاهده است. در رابطهٔ ۳، g(r,θ) تصویرپایه در مختصات قطبی می‌باشد، نحوهٔ تولید یک تصویرپایه، از درجهٔ n و تکرار l در این رابطه مشخص شده است. در این پژوهش از تصویرپایهٔ M2,2 و M4,4 استفاده شده است.

$$M_{n,l} = \int_{0}^{2\pi} \int_{0}^{1} [H_{n,l}(r,\theta)]^{*} g(r,\theta) r dr d\theta$$

$$H_{n,l}(r,\theta) = \Omega_n \cos(\pi n r^2) e^{jl\theta}, \Omega_n = \begin{cases} 1/\pi & n = 0 \\ 2/\pi & n \neq 0 \end{cases} \quad (۳) \quad (۴)$$

$$G_{[N\times F']}^{`} = \{DCT_{(G,T)}, DCT_{(G\otimes M_{2,2},T)}, DCT_{(G\otimes M_{4,4},T)}\}$$

در رابطهٔ ٤، $DCT_{(x,T)}$ بدان معنی است که از ماتریس x، تبدیل کسینوسی گرفته شده و تنها T ضریب یا ویژگی



شکل ۴: تصویر پایهٔ PCT، الف) M2,2، ب) M4,4

اولیه حفظ می‌شود، بنابراین خروجی این مرحله، ماتریس ویژگـی `G با ابعـاد (N/K)*3T خواهد بـود. همان‌طور کـه پیش‌تر گفته شـد، با تبدیـل DCT مؤلفه‌های پراهمیت فرکانسی مشخص می‌شوند. از این‌رو می‌توان تنها بخشی از ویژگی‌های پراهمیت را نگه داشـت و مابقی آن‌ها را از فرآیند استخراج ویژگی حذف کرد. در این پژوهش بعد از عملیات DCT تنها 22 (T=22) ویژگی پراهمیت از ۳۲ ویژگی حفظ می‌شوند. چون در این طرح از سه تبدیل DCT استفاده شده است، بنابراین درمجموع 66 (F`=66) ویژگی به ازای هر نمونه از سیگنال صدا وجود خواهد داشت.

۳-۳ استانداردسازی و پیچش ویژگی‌ها

منظـور از استانداردسـازی ویژگی‌هـا، فرآینـد بهنجارکـردن داده‌ها می‌باشـد. طی این فرآینـد، هر بعد ویژگی، دارای میانگین صفر و واریانس یک می‌شـود. این فرآیند در رابطهٔ ۵ مشـخص اسـت، در این رابطه هر دادهٔ ماتریـس بردار ویژگی `G از میانگین هر بعد ویژگی Pj کم شده و بر انحراف استاندارد Vj آن بعد، تقسیم می‌شود.

$$P_j = \frac{1}{N}\sum_i^N G`(i,j), V_j = \sqrt{\frac{1}{N-1}\sum_i^N (G`(i,j) - P_j)^2}$$
$$G`(i,j) = \frac{G`(i,j) - P_j}{V_j} \qquad (۵)$$

در فرآینـد پیچـش ویژگی‌ها [۳۹-۳۷] که پیشـتر در کاربرد تشـخیص گوینده استفاده شـده بود [۲۶]، هدف، برازشِ توزیعِ استانداردِ نرمال، در هر بعد ویژگی، در یک بازه زمانی کوتاه می‌باشـد. برای این کار یک پنجرهٔ لغزان روی هر بعد ویژگی پیمایش می‌شـود و داده‌ها را در یک بازهٔ زمانی کوتاه به توزیع استاندارد نرمال تبدیل می‌کند. با این کار اثرات نوفه‌ها کاهش می‌یابد، زیرا در فرآیند نگاشت به تابع نرمال، به تغییرات اساسی داده‌ها توجه می‌شود و مقاومت زیادی بر روی تغییرات ناگهانی که بیشتر نوفه‌ها می‌باشد صورت می‌گیرد و آن‌ها را حذف می‌کند.

با استانداردسـازی و پیچش ویژگی‌ها، ماتریس بردار ویژگی تا حدود زیـادی از نوفه‌ها و تغییرات ناگهانی پاک می‌شود. در این پژوهش بعد از آزمایش‌های مختلف، پنجرهٔ لغزانی به طول ۳۰۰ نمونه در نظر گرفته شد.

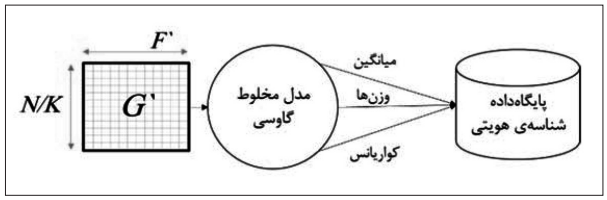
۳-۴ اعمال فیلتر گاوسی

بـا توجه به آزمایـش مختلف صـورت گرفته در این پژوهش، اگر بتوان بین ویژگی‌های تعیین‌شده نیز تغییرات ناگهانی را حذف نمود، می‌توان دقت سیستم را در شرایط نوفه‌ای شدید افزایش داد.

به‌همین دلیل در آخرین مرحله یک فیلتر گاوسی با اندازهٔ ۲×۳ بر روی ماتریس `G لغزانده می‌شود. با این کار تغییرات ناگهانی در بین ویژگی‌های همسایه و نمونه‌های مختلف از سـیگنال صدا به‌صورت توأمان حذف‌شـده و به‌اصطلاح ماتریس بردار ویژگی از دو جهت ویژگی‌ها و نمونه‌ها دارای تغییرات نرم می‌شـود. فرآیند خالص‌سازی ویژگی‌ها که با دو عمل استانداردسـازی و پیچش ویژگی‌ها و اعمال فیلتر گاوسی صورت گرفت، باعث افزایش سرعت آموزش شده و اثر مقادیر پرنوسان از ویژگی‌های مختلف را کاهش می‌دهد.

۳-۵ آموزش توسط مدل مخلوط گاوسی

مـدل مخلوط گاوسـی یکـی از رده‌بندهای مرسـوم جهت آموزش فرآیند تشـخیص گوینده می‌باشـد. در این پژوهش نیز بعد آزمایش‌های مختلف، این رده‌بند به‌عنوان آموزش‌دهنـدهٔ ویژگی‌ها انتخاب شـد. در این مرحله طبق شکل ۵، ماتریس بردار ویژگی تحت این الگوریتم یادگیری، آموزش دیده و سـرانجام مقادیر میانگین، کوواریانس و وزن حاصـل از ویژگی‌هایی که در طی فرآیند آموزش به ازای هر گوینده حاصل می‌شـود، در پایگاه دادهٔ شناسـهٔ هویتی ذخیره می‌شود.



شکل ۵: فرآیند آموزش توسط مدل مخلوط گاوسی

طبق روش بیان‌شـــده، ٦٦ ویژگـــی در ماتریس بردار ویژگـــی وجـــود دارد کـــه مدل مخلـــوط گاوســـی باید با پـــردازش بر روی این ماتریـــس، مؤلفه‌های میانگین، وزن و کوواریانس ویژگی‌ها را برای هر گوینده به‌دســـت آورد. در رابطـــهٔ ٦، مدل مخلوط گاوســـی به‌صـــورت یک مدل احتمالاتی پارامتریک نشـــان داده شده‌است. این رابطه به ازای هر گویندهٔ موجود در پایگاه‌داده به‌دســـت می‌آید. در این رابطه هدف به دست آوردن ω (وزن)، μ (میانگین) و Σ (ماتریس کوواریانس) یک تابع گاوســـی N به ازای هر مخلوطِ j تعریف شـــده می‌باشد، به صورتی که با مشاهدهٔ این پارامترها، احتمال P(G`) بیشینه شود.

$$P(G`) = \sum_j^J \omega_j N(G`|\mu_j, \Sigma_j) \qquad (۶)$$

۳-۶ شناسایی گوینده

در مرحلهٔ آزمایش، همهٔ مراحل استخراج ویژگی برای ســـیگنال گوینده کـــه در مرحلهٔ آموزش بیان شـــد، تکرار می‌شود. بعد از استخراج ماتریس بردار ویژگی `G، میزان شـــباهت ویژگی‌های گویندهٔ مورد آزمایـــش با هر یک از گوینده‌های موجود در پایگاه‌دادهٔ شناسه هویتی، محاسبه می‌شـــود. بیشترین امتیاز حاصل‌شـــده در بین گویندگان، هویت فرد را مشخص می‌کند.

۴- نتایج آزمایشگاهی

برای آزمایش کارایی روش پیشـــنهادی، یک مجموعه داده با ۱۰۰ گوینده از مجموعه‌دادهٔ بزرگ Voxforge [٤٠] تهیـــه گردید. در این مجموعه داده بـــه ازای هر گوینده ۱۰ فایل صوتـــی، با مدت‌زمـــان ٦ ثانیه وجـــود دارد. در این پژوهـــش از ۷ فایل صوتی جهت آموزش و ۳ فایل صوتی جهت آزمایش برای هر گوینده استفاده شده است.

تمـــام روش‌های موجود و روش پیشـــنهادی تحت یک مدل مخلوط گاوســـی با تعداد مخلوط‌های برابر ١٦، تعداد خوشه‌های (به‌دست آمده توسط الگوریتم k-means) برابر ۱۰ و متغیر تنظیم سازی 10^{-6} آموزش دیدند. برای آزمایش میزان کارایی این الگوریتم‌ها در هنگام وجود نوفه، در این پژوهش سه شرایط محیطی مختلف در نظر گرفته شد.

محیط عاری از نوفهٔ، محیط همراه با نوفهٔ سفید باقدرت سیگنال به نوفهٔ[19] (SNR) ۲۰،۱۰ و ٥ دسی‌بل و محیط همراه با نوفهٔ خیابانی و ســـفید باقدرت SNR ۲۰،۱۰ و ٥ دسی‌بل، شرایط ذکرشـــده می‌باشـــند. لازم به ذکر است در هنگام آموزش، صدای گویندگان با شرایط عادی و عاری از نوفه آموزش می‌بینند ولی در مرحلهٔ آزمایش، صدای گویندگان در شرایط محیطی نوفه‌ای مورد آزمایش قرار می‌گیرند.

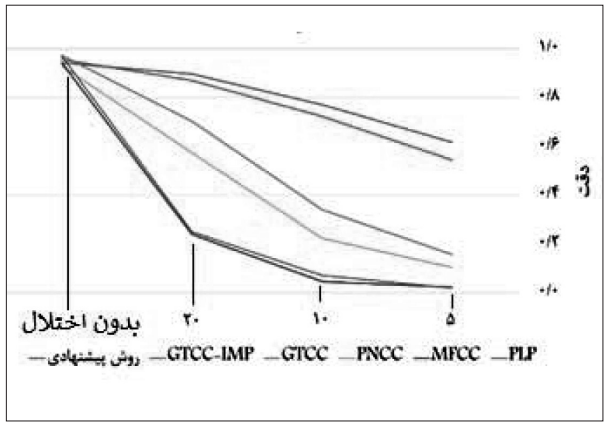
منظور از نوفهٔ خیابانی، نوفهٔ موجود در داخل شـــهر، همراه با صدای اتومبیل‌ها و هیاهوی مردم می‌باشـــد که با قدرت‌های سیگنال به نوفهٔ مختلف بر روی صدای گویندگان اعمال می‌شـــود. معیار کارایی در نمودارها و جداول پیش رو، بر اســـاس تعداد تشـــخیص درســـت گوینـــده در هر روش، می‌باشـــد. در جدول ۱، کارایی روش پیشنهادی در مقایســـه با روش‌های موجود در شرایط وجود نوفهٔ سفید قابل‌مشاهده است.

همان‌طور کـــه در جدول ۱ قابل‌ملاحظه اســـت، روش پیشـــنهادی در شـــرایط عاری از نوفه، اختلاف دقت کمی با ســـایر روش‌هـــا دارد ولی هرچه به نوفه‌های شـــدیدتر نزدیک می‌شویم (مقدار SNR کاهش می‌یابد)، کارایی روش پیشـــنهادی نمایان‌تر می‌شود. همچنین در شکل ٦، نمودار روند تغییرات کارایی به‌صورت شـــهودی مشـــخص شده است. کارایی روش پیشـــنهادی با مقایسه با روش GFCC [۱۹] وضوح بیشـــتری از خود نشان می‌دهد، روش GFCC [۱۹] بـــا این‌که بیشـــترین دقت (۹۷ درصد) را در شـــرایط عاری از نوفه داشته است، اما در شرایط نوفهٔ سفید با SNR ٥ دسی‌بل، تنها ۱٥ درصد کارایی داشته است، این در حالی است که روش پیشـــنهادی در این شرایط ٦۱ درصد دقت

19- Signal to Noise Ratio

جدول ۱: کارایی تشخیص گوینده در محیط نوفهٔ سفید با SNRهای مختلف

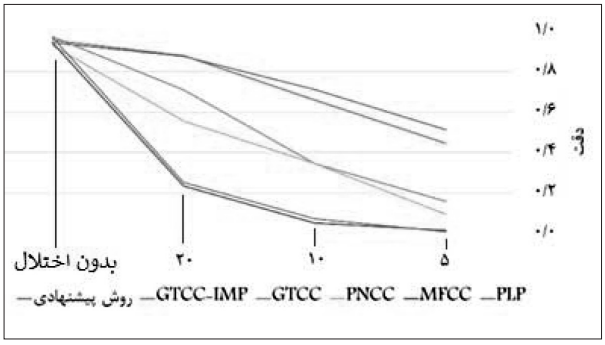
	بدون اختلال	۲۰	۱۰	۵
روش پیشنهادی	۰/۹۴۳۳	۰/۸۹۶۶	۰/۷۷	۰/۶۱۶۶
GFCC-IMP [26]	۰/۹۵۶۶	۰/۸۷	۰/۷۲۳۳	۰/۵۴۳۳
GFCC [19]	۰/۹۷۳۳	۰/۷۰۳۳	۰/۳۴	۰/۱۵۶۶
PNCC [18]	۰/۹۳	۰/۵۷۳۳	۰/۲۲۳۳	۰/۱۰۳۳
MFCC [15]	۰/۹۴۳۳	۰/۲۴	۰/۰۴۶۶	۰/۰۲۳۳
PLP [16]	۰/۹۷	۰/۲۵	۰/۰۷	۰/۰۲



شکل ۶: نمودار کارایی روش‌های مختلف در شرایط نوفه‌ای سفید با SNR های مختلف

جدول ۲: کارایی تشخیص گوینده در محیط نوفهٔ توأمان سفید و خیابانی با SNRهای مختلف

	بدون اختلال	۲۰	۱۰	۵
روش پیشنهادی	۰/۹۴۳۳	۰/۸۷۶۶	۰/۷۱۳۳	۰/۵۱۳۳
GFCC-IMP [26]	۰/۹۵۶۶	۰/۸۸	۰/۶۶	۰/۴۴۶۶
GFCC [19]	۰/۹۷۳۳	۰/۷۱	۰/۳۴۶۶	۰/۱۶
PNCC [18]	۰/۹۳	۰/۵۵۶۶	۰/۱۹۳۳	۰/۰۹۶۶
MFCC [15]	۰/۹۴۳۳	۰/۲۳۶۶	۰/۰۵۳۳	۰/۰۱۶۶
PLP [16]	۰/۹۷	۰/۲۵۳۳	۰/۰۷۳۳	۰/۰۱



شکل ۷: نمودار کارایی روش‌های مختلف در شرایط نوفه‌ای سفید و خیابانی با SNR های مختلف

دارد. نزدیک‌ترین روش نسبت به روش پیشنهادی، روش GFCC-IMP [۲۶] می‌باشد که در مقایسه با این روش نیز، روش پیشنهادی به‌صورت میانگین، در سه شرایط نوفه‌ای در حدود ۵ درصد، دقت بهتری نسبت به آن داشته است. مهم‌ترین دلیل کارایی روش پیشنهادی، اعمال ویژگی‌های توأمان تبدیل کسینوسی قطبی و گسسته می‌باشد که با استخراج ویژگی‌های اساسی، مقاومت بیشتری در قبال نوفه، نسبت به روش GFCC-IMP [۲۶] رقم زده است.

آزمایش بعدی، کارایی روش پیشنهادی در مقایسه با روش‌های موجود در شرایط نوفه‌های توأمان خیابانی و نوفهٔ سفید می‌باشد. این آزمایش شرایط سختی برای تشخیص گوینده هست، زیرا وجود دو نوفهٔ، سیگنال صدا را بیشتر از حالت قبلی تخریب می‌کند و استخراج ویژگی‌های اساسی را با مشکلات زیادی روبرو خواهدکرد. در جدول ۲، کارایی روش پیشنهادی در مقایسه با سایر روش‌ها، تحت شرایط ذکر شده نمایان است. همچنین در شکل ۷، نمودار روند تغییرات کارایی به‌صورت شهودی مشخص شده است.

همان‌طور که از نمودار شکل ۷ مشخص است در تمامی روش‌ها با افزایش قدرت نوفهٔ (کاهش SNR)، کارایی کاهش می‌یابد، ولی در روش پیشنهادی این کاهش کارایی از سایر روش کمتر است. همان‌طور که در اعداد جدول ۲ مشخص است، روش پیشنهادی نسبت به روش IMP-GFCC [۲۶] در این حالت در حدود ٤ درصد افزایش دقت داشته است.

همان‌طور که پیش‌تر توضیح داده شد، در این پژوهش از فیلترهای پیش‌پردازش سیگنال برای افزایش کارایی استفاده نشده است و تنها هدف این پژوهش ارائهٔ یک روش پیشنهادی برای استخراج ویژگی‌های اساسی و مقاوم به نوفه برای تشخیص گوینده بوده است. مطمئناً با اعمال فیلترهای پیش‌پردازش سیگنال، کارایی تشخیص گوینده بالاتر از این مقدار می‌رود. با توجه به آزمایش‌های

صورت گرفته مشخص می‌شود، روش پیشنهادی توانسته در حضور نوفه‌های شـــدید کارایی مناسبی از خود نشان دهد، بنابراین می‌توان از این سیســـتم در محیط‌های آلوده به نوفهٔ استفادهٔ مناسبی برد.

۵- نتیجه‌گیری

در این پژوهش یک سیســـتم تشـــخیص گویندهٔ مقاوم به نوفه ارائه شـــد که می‌تواند در شـــرایط نوفه‌ای مختلف کارایی خوبی از خود نشان دهد. شالودهٔ روش پیشنهادی، طراحی ویژگی‌های مقاوم به نوفه از سیگنال صدای انسان بوده است.

هم‌آمیزی سیگنال صدا در فیلتربانک گاماتون، تبدیل‌های توأمان کسینوسی قطبی و گسسته و همچنین طراحی یک سیستم اختصاصی خالص‌سازی ویژگی متناسب با حوزهٔ تشخیص گوینده، موجب ساخت یک ویژگی مقاوم در این کاربرد شده است. طبق آزمایش‌های صورت گرفته، روش پیشـــنهادی در محیط‌های آلوده به نوفهٔ سفید و نوفه‌های خیابانی، گوینده را با دقت بهتری نسبت به روش‌های دیگر شناسایی کرده است.

از ایـــن روش تشـــخیص گوینده می‌تـــوان در صنعت نظامی و محیط‌های صنعتی اســـتفاده خوبی برد، زیرا در این محیط‌ها همـــواره نوفه‌های مختلفی وجود دارد. روش پیشـــنهادی باوجود این‌که می‌تواند دقت خوبی در شرایط نوفه‌ای بســـیار شدید نسبت به سایر روش‌ها داشته باشد ولی به علت در نظر نگرفتن یک بخش پیش‌پردازش سیگنال صدا، همچنان ســـیگنال مخرب نوفه بر کارایی شناســـایی گوینـــده تأثیـــر زیـــادی دارد، ازاین‌رو، کار آینـــده در این پژوهش، طراحی یک روش مناســـب جهـــت پیش‌پردازش ســـیگنال صدا در کاربرد تشـــخیص گوینده خواهدبود. با طراحی این ســـامانه، سیستم تشـــخیص گوینده با کارایی بیشتری می‌تواند در محیط‌های نوفه‌ای عمل کند.

مراجع

1. X. Yong, J. Du, L. Dai, “A regression approach to speech enhancement based on deep neural networks”, IEEE/ ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 23, pp. 7-19, 2015.
2. K. Kais, A. Boudraa, M. Turki, “Voiced/unvoiced speech classification-based adaptive filtering of decomposed empirical modes for speech enhancement”, IET Signal Processing, vol. 10, pp. 69-80, .2016.
3. N. Jingen, X. Chen, J. Yang, “Two variants of the sign subband adaptive filter with improved convergence rate”, Signal Processing, vol. 96, pp. 325-331,2014.
4. Y. Yi, H. Zhao, B. Chen, “A new normalized subband adaptive filter algorithm with individual variable step sizes”, Circuits Systems and Signal Processing, vol. 35, pp. 1407-1418, 2016.
5. G. Elizabeth, M. Koutsogiannaki, Y. Stylianou, “Approaching speech intelligibility enhancement with inspiration from Lombard and clear speaking styles”, Computer Speech & Language, vol. 28,pp. 629-647, 2014.
6. H. Cheng, S. Wang, Y. Lai, Y. Tsao, H. Chang, and H. Wang, “Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks”, IEEE Transactions on Emerging Topics in Computational, vol.2, pp. 117-128, 2018.
7. D. Vishwakarma, K. Kapoor, R. Dhiman, A. Goyal and D. Jamil, “De-noising of Audio Signal using Heavy Tailed Distribution and comparison of wavelets and thresholding techniques”, In Computing for Sustainable Global Development, 2nd International Conference on IEEE, pp. 755-760, 2015.
8. O. Plchot, L. Burget, H. Aronowitz and P. Matejka, “Audio enhancing with DNN autoencoder for speaker recognition”, In Acoustics, Speech and Signal Processing, International Conference on IEEE, pp. 5090-5094, 2016.
9. R. Tong, Y. Zhou, L. Zhang, G. Bao and Z. Ye, “A Robust time-frequency decomposition model for suppression of mixed Gaussian-impulse noise in audio signals”, IEEE/ ACM Transactions on Audio, Speech and Language Processing, vol. 23, pp. 69-79, 2015.
10. L. Rabiner and B. Juang, “Fundamentals of speech recognition”, Englewood Cliffs: PTR Prentice Hall, vol. 14, 1993.
11. D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models”, IEEE transactions on speech and audio processing, vol. 3, pp. 72-83, 1995.
12. M. Collins, “Head-driven statistical models for natural language parsing. Computational linguistics”, vol. 29, pp. 589-637, 2003.
13. B. Wang, Z. Ou and Z. Tan, “Learning trans-dimensional random fields with applications to language modeling”, IEEE transactions on pattern analysis and machine intelligence, vol. 40, pp. 876-890. 2018.
14. F. Jelinek, “Statistical Methods for Speech Recognition

(Language, Speech, and Communication)", MIT Press, 1998.
15. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Speech and Signal Processing, vol. 28, pp. 357–366, 1980.
16. H. Hermansky, "Perceptual linear prediction analysis of speech", The Journal of the Acoustical Society of America. Soc. Am, vol. 87, pp. 1738–1752, 1990.
17. N. Scheffer, L. Ferrer and A. Lawson, Y. Lei, M. McLaren, "Recent developments in voice biometrics: Robustness and high accuracy", In Technologies for Homeland Security (HST), International Conference on IEEE, pp. 447-452, 2013.
18. C. Kim, R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition", IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 24, pp. 1315-1329, 2016.
19. X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification", IEEE Transactions on Multimedia, vol. 14, pp.1684-1689, 2012.
20. R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function", a meeting of the IOC Speech Group on Auditory Modelling at RSRE, vol. 2, pp. 1, 1987.
21. X. Valero, A. Francesc, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification", IEEE Transactions on Multimedia, vol. 14, pp, 1684-1689, 2012.
22. R. Schluter, I. Bezrukov, H. Wagner and Hermann Ney, "Gammatone features and feature combination for large vocabulary speech recognition", In Acoustics, Speech and Signal Processing, International Conference IEEE, vol. 4, pp. IV-649, 2007.
23. Y. Shao, D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis. In Acoustics, Speech and Signal Processing", International Conference on IEEE, pp.1589-1592. 2008.
24. Y. Shao, Z. Jin, D. Wang, S. Srinivasan, "An auditory-based feature for robust speech recognition", In Acoustics, Speech and Signal Processing, International Conference on IEEE, pp. 4625-4628, 2009.
25. J. Qi, D. Wang, Y. Jiang and R. Liu, R, "Auditory features based on gammatone filters for robust speech recognition", In Circuits and Systems, International Symposium on IEEE, pp. 305-308, 2013.
26. S. Agarwal and D. Muralidharan, "Speaker Verification in Noisy Environment", UCLA Electrical Engineering, https://github.com/ShubhamAgarwal12/Automatic-Speaker-Recognition/blob/master/report.pdf, April 20, 2018.
27. C. Kim and M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction", 10th Annual Conference of the International Speech Communication Association, pp. 28-31, 2009.
28. J. Ding, Jr and C. Yen, "Enhancing GMM speaker identification by incorporating SVM speaker verification for intelligent web-based speech applications", Multimedia Tools and Applications, vol. 74, pp. 5131-5140, 2015.
29. D. Bone, M. Li, M. Blac and S. Narayanan, "Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functional and GMM supervector", Computer speech & language, vol. 25, pp.375-391, 2014.
30. N. Tomashenko and K. Yuri Khokhlov, "Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing", Fifteenth Annual Conference of the International Speech Communication Association, pp. 2997-3001, 2014.
31. D. Leon, M. Phillip, J. Yamagishi, I. Herneaz and I. Saratxage, "Evaluation of speaker verification security and detection of HMM-based synthetic speech", IEEE Transactions on Audio, Speech and Language Processing, vol. 20, pp. 2280-2290, 2012.
32. C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in INTERSPEECH, pp. 28–31.2009.
33. B. Moore and B. Glasberg, "A revision of Zwicker's loudness model", Acustica - Acta Acustica, vol. 82, pp. 335–345, 1996.
34. Y. Li, "Image copy-move forgery detection based on polar cosine transform and approximate nearest neighbor searching", Forensic Science International, vol. 224, pp. 59-67, 2013.
35. D. Cozzolino, G. Poggi, and L. Verdoliva, "Efficient dense-field copy-move forgery detection", IEEE Transactions on Information Forensics and Security, vol. 10, pp. 2284-2297, 2015.
36. Y. Thian, J. Xudong, and A. C. Kot, "Two-dimensional polar harmonic transforms for invariant image representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pp. 1259-1270, 2010.
37. B. Kim, B. Kang, H. Kim, and S. Baek, "Prognosis prediction for class III malocclusion treatment by feature wrapping method", The Angle orthodontist, vol. 79, pp. 683-691, 2009.
38. H. Huang, C. Hsieh and M. Lu, "Hybrid feature selection by combining filters and wrappers", Expert Systems with Applications, vol. 38, pp. 8144-8150, 2011.
39. X. Liu, Y. Liang, S. Wang, Z. Yang, and H. Ye, "A Hybrid Genetic Algorithm With Wrapper-Embedded Approaches for Feature Selection", IEEE Access, vol. 6, pp.22863-22874, 2018.