مقایسه تأثیر استفاده از بردارهای متفاوت ویژگی بر کارایی سیستم تایید هویت گوینده مبتنی بر مدل‌سازی گوینده توسط بردارهای هویت i-Vectors

محسن محمدی[1]، حمیدرضا صادق محمدی[2]

[1] پژوهشکده برق جهاد دانشگاهی، تهران، mohammadi.mohsen@gmail.com

[2] پژوهشکده برق جهاد دانشگاهی، تهران، mohammadis@iust.ac.ir

چکیده

گفتار یک پارامتر مناسب برای تشخیص هویت است که علاوه بر کاربرپسند بودن و پیچیدگی پیاده‌سازی کم، هزینه پایینی هم دارد. با وجود مزایای زیاد، این روش محدودیت‌هایی نیز دارد که از آن جمله می‌توان به کاهش دقت این روش در محیط‌های واقعی به دلیل حضور نویزهای مختلف اشاره کرد. تا کنون روش‌های گوناگونی برای حل این مشکل در مراحل مختلف سیستم تایید هویت گوینده یعنی استخراج ویژگی، مدل‌سازی و مقایسه و امتیازدهی ارائه شده است که البته هیچکدام کامل نیست. در این مقاله تأثیر استفاده از چهار بردار ویژگی MFCC، IMFCC، LFCC و PNCC بر کارایی سیستم تایید هویت گوینده مبتنی بر مدل‌سازی گوینده توسط بردارهای هویت i-vector در شرایط گفتار تمیز و نویزی مورد ارزیابی و مقایسه قرار گرفته است. در پیاده‌سازی آزمون‌ها برای سیگنال گفتار و نویز به ترتیب از دادگان‌های TIMIT و NOISEX92 استفاده گردید. نتایج آزمون‌ها نشان می‌دهد که در حضور نویز ویژگی‌های IMFCC و PNCC عملکرد بهتری دارند. به منظور افزایش کارایی سیستم در شرایط نویزی استفاده از الگوریتم بهبود گفتار تفریق طیفی مورد بررسی واقع شد که تنها برای سیگنال گفتار همراه با نویز سفید در برخی حالات عملکرد سیستم را تا حد قابل ملاحظه‌ای بهبود بخشید.

واژه‌های کلیدی

تایید هویت، بردارهای ویژگی گفتار، مدل آمیزه‌های گاوسی، گفتار نویزی.

۱ - مقدمه

تا به امروز روش‌های مختلفی برای شناسایی افراد از روی ویژگی‌های حیاتی آن ها مورد مطالعه قرار گرفته‌اند. از میان معروف‌ترین ویژگی‌ها، می‌توان اثر انگشت، چهره و صدای افراد را نام برد. هر کدام از این ویژگی‌ها مزایا و معایب خاص خود را با توجه به دقت و کاربرد مورد نظر دارند و هیچ سیستم تایید هویتی در تمامی شرایط به طور مطلق بهترین راهکار نیست. عواملی وجود دارد که صدای افراد در هنگام صحبت را از دیگر ویژگی‌ها متمایز می‌سازد. گفتار یک سیگنال طبیعی است و تولید گفتار یک فرد برای فرد دیگر ممکن نیست. در بسیاری از کاربردها، گفتار تنها راه دسترسی به افراد می‌باشد مانند ارتباط از راه دور با پهنای باند کم نظیر تلفن. همچنین شناسایی از طریق گفتار کاربر پسندتر بوده و نیاز به تجهیزات و حسگرهای ویژه و گران قیمتی ندارد [1].

تاکنون سیستم‌های شناسایی و تایید گوینده متنوعی مطرح شده‌اند و در محیط‌های آزمایشگاهی به نتایج خوب و قابل قبولی نیز دست یافته‌اند ولی این سیستم‌ها در محیط‌های طبیعی که عوامل مزاحم همانند انواع نویزهای کانولوشن، جمع شونده، اعوجاجات کانال و یا انعکاس‌های محیطی موجود هستند، افت شدیدی در کارایی خود دارند [2]. بنابراین برای سیستم شناسایی گوینده مطلوب، یکی از فاکتورهای مهم این است که سیستم بتواند در برابر نویزهای محیطی مقاوم باشد. از این‌رو روش‌های مقاوم‌سازی سیستم نسبت به نویز در سیستم‌های شناسایی و تایید گوینده از اهمیت به‌سزایی برخوردارند.

سیستم‌های تشخیص گوینده از چند بخش اصلی تشکیل شده‌اند که عبارتند از: استخراج ویژگی، مدل‌سازی گوینده، مقایسه و تصمیم‌گیری. یکی از روش‌های مقاوم‌سازی سیستم تشخیص گوینده استخراج و بهره‌گیری از ویژگی‌هایی است که نسبت به شرایط نویزی مقاوم باشند و یا بتوان با روش‌هایی اثرات نامطلوب آنها را کاهش داد [3]. بنابراین در کنار استفاده ضرایبی همچون MFCC[1] [4]، به‌کارگیری روش‌های مقاوم برای تخمین طیف مانند پیشگویی خطی وزن‌دار[2] (WLP) [5]، روش‌های نوین استخراج ویژگی مانند PNCC[3] [6] و روش‌های پس‌پردازش بردارهای

ویژگی چون feature warping [7] ارائه شده‌اند. از آنجایی که تحلیل و تخمین طیف از روش‌هایی همانند تبدیل فوریه و پیشگویی خطی نتوانسته‌اند به تنهایی بر اثرات مخرب نویز جمع شونده غلبه کنند، بهره‌مندی از روش‌هایی چون تفریق طیفی، فیلتر وینر و فیلتر کالمن برای بهبود سیگنال گفتار آلوده به نویز جمع شونده و در نتیجه بهبود عملکرد سیستم‌های تشخیص هویت در دهه‌های اخیر پیشنهاد گردیده‌اند [8].

در بخش مدل‌سازی گوینده تاکنون روش‌های مختلفی ارائه شده است که مدل آمیزه گاوسی GMM، پیچش زمانی پویا DTW و ماشین بردار پشتیبان SVM از این جمله‌اند [4]. در سال‌های اخیر روش‌هایی بر مبنای آنالیز عامل FA پیشنهاد گردیده‌اند که عملکرد بسیار بهتری نسبت به روش‌های پیشین داشته‌اند. از جمله مهمترین این روش‌ها آنالیز عامل مشترک JFA و بردار هویت i-vector قابل ذکر هستند. هدف آنالیز عامل آشکارسازی روابط پنهان بین داده‌هاست. در این روش روابط بین تعداد زیادی متغیر آشکار توسط تعداد کمی متغیر پنهان بازنویسی می‌شود. با این روش می‌توان تعداد زیادی متغیر را در قالب یک یا چند متغیر بیان کرد. متغیرهای وابسته درون یک گروه قرار می‌گیرند. ایده استفاده از FA در فضای ابربردارهای GMM اولین بار در سال ۲۰۰۴ مطرح شد [2]. از آن زمان تاکنون روش‌های مختلفی برمبنای FA ارائه شدند که در نهایت منجر به پیشنهاد روش i-vector در سال ۲۰۱۱ شد [9].

این مقاله به بررسی انواع رایج بردارهای ویژگی سیگنال گفتار و ارزیابی و مقایسه عملکرد آنها در حضور چند نمونه مهم از نویز جمع‌شونده با سطوح سیگنال به نویز متفاوت می‌پردازد. همچنین این عملکرد در شرایط بهره‌گیری از یک الگوریتم بهبود گفتار همراه با نویز نیز بررسی می‌گردد. ساختار کلی مقاله به صورت زیر است. ابتدا در بخش ۲، چهار بردار ویژگی رایج سیگنال گفتار مرور می‌شود. بخش ۳ به بیان مدل‌سازی تایید هویت گوینده به روش i-vector اختصاص یافته است. در بخش ۴ ملاحظات اجرایی این تحقیق اعم از دادگان مورد استفاده، پیکربندی سیستم آزمون و اجزاء مختلف آن، روش‌های ارزیابی عملکرد شرح داده شده و نتایج حاصل از آزمون مورد بحث و ارزیابی قرار می‌گیرد. در نهایت بخش پایانی به نتیجه‌گیری از این تحقیق می‌پردازد.

۲ – بردارهای ویژگی گفتار

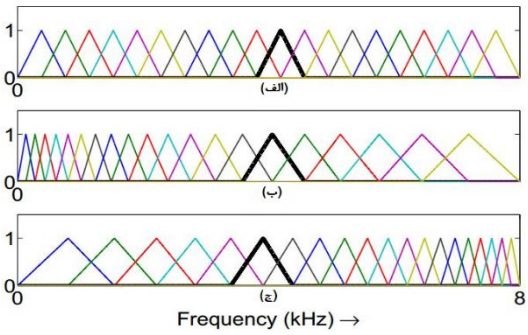
هر گوینده مشخصه‌هایی در صدای خود دارد که مختص به اوست. با وجود اینکه مشخصه‌های صدای گویندگان مختلف به راحتی تفکیک‌پذیر نیستند اما به دلیل فیزیولوژی متفاوت مسیر صوتی و عادات رفتاری هر فرد، وابسته به شخص و واحد هستند [4]. انرژی‌های بانک فیلتر مقیاس مل و نمایش کپسترال آنها بسیاری از خصوصیاتی را شامل می‌شوند که یک ویژگی ایده‌آل برای تشخیص گوینده نیاز دارد. با وجود اینکه این خصوصیات به صورت ذاتی وابسته به سلامت شخص و کیفیت کانال انتقال هستند، اما با استفاده از روش‌های ساده‌ای می‌توان این اثرات را کمینه کرد.

در این مقاله چهار بردار ویژگی رایج سیگنال گفتار، یعنی MFCC، IMFCC[۴]، LFCC[۵] و PNCC مورد مطالعه قرار گرفته است. ضرایب کپسترال فرکانس مل یکی از متداول‌ترین بردارهای ویژگی در سیستم‌های شناسایی گوینده است. ایده اصلی آن از خواص شنیداری گوش انسان الهام گرفته است که به تغییرات فرکانسی در فرکانس‌های پایین حساس‌تر هستند. بر همین اساس بانک فیلتری طراحی می‌شود که تاکید بیشتری بر فرکانس‌های پایین دارد.

خروجی این بانک فیلتر پس از گذر از فشرده‌ساز لگاریتمی و تبدیل گسسته کسینوسی DCT ضرایب MFCC را نتیجه می‌دهد [4]. اگر خروجی‌های یک بانک فیلتر Mکاناله را به صورت $Y(m), m = 1,...,M$ در نظر بگیریم ضرایب MFCC به‌دست می‌آیند:

$$c_n = \sum_{m=1}^{M}\left[\log\left(Y(m)\right)\right]\cos\left[\frac{\pi n}{M}\left(m-\frac{1}{2}\right)\right] \qquad (1)$$

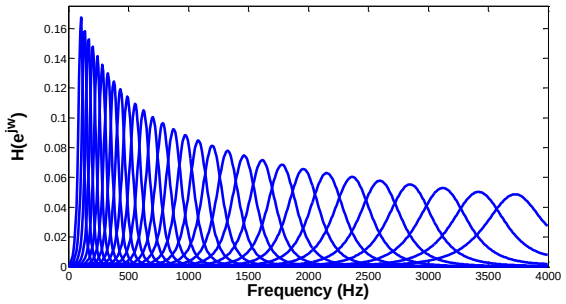
استخراج ضرایب IMFCC تا حدود زیادی مشابه MFCC است با این تفاوت که تاکید بیشتری بر فرکانس‌های بالای سیگنال وجود دارد، بنابراین تنها تفاوت آن در استخراج با ضرایب MFCC در بانک فیلتر به کار گرفته شده است. IMFCC اطلاعات نهفته در فرکانس‌های بالا را که در MFCC دارای اهمیت کمتری هستند، بیشتر مورد توجه قرار می‌دهد [10]. استخراج ضرایب LFCC نیز با MFCC مشابهت زیادی دارد و تفاوت آن در بانک فیلتر به کار رفته است به این صورت که در استخراج ضرایب LFCC، همه محدوده‌های فرکانسی سیگنال گفتار دارای اهمیت یکسانی تلقی شده و به یک میزان مورد توجه قرار می‌گیرند. شکل 1 تفاوت‌های بانک‌های فیلتر مورد استفاده در محاسبه ضرایب MFCC، IMFCC و LFCC را به روشنی نشان می‌دهد [11].



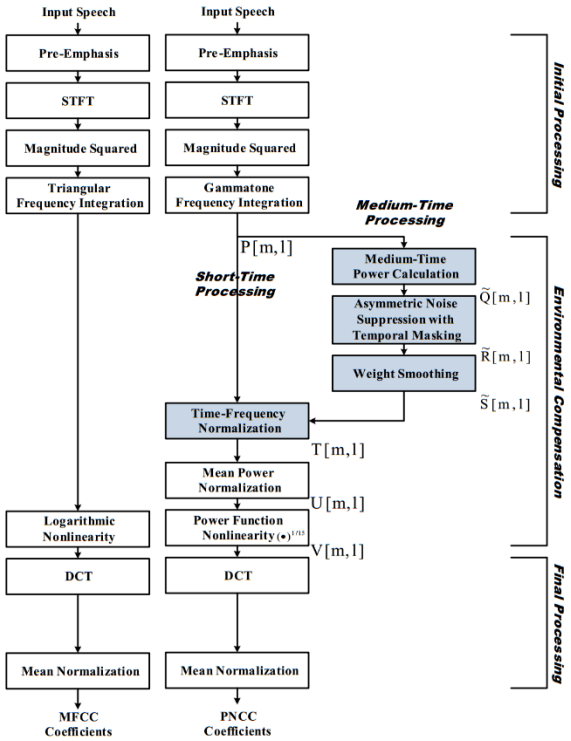
شکل 1: فیلتر بانک های به کار رفته در محاسبه (الف) LFCC، (ب) MFCC و (ج) IMFCC [11]

استخراج بردار ویژگی PNCC نیز مانند MFCC از سیستم شنوایی انسان الهام گرفته و سعی در شبیه‌سازی آن دارد. برای استخراج ضرایب PNCC از یک بانک فیلتر گاماتون ۳۰ کاناله در بازه ۱۰۰ تا ۴۰۰۰ هرتز استفاده می‌شود. پهنای باند فیلترها را به‌گونه‌ای در نظر می‌گیرند که همانند MFCC تاکید بیشتر بر فرکانس‌های پایین باشد. شکل 2 پاسخ فرکانسی این فیلتر را نمایش می‌دهد. به منظور افزایش کارایی سیستم سطح زیر نمودار هر کانال به یک نرمالیزه شده است. اصلی‌ترین ویژگی PNCC عبارت است از استفاده از تابع غیرخطی توان که جایگزین تبدیل لگاریتمی در MFCC می‌شود، الگوریتم غلبه بر نویز بر مبنای فیلتر نامتقارن و روشی برای پوشش زمانی است [6]. در شکل 3 ساختار

استخراج ویژگی به دو روش PNCC و MFCC نشان داده شده‌اند و بخش‌های متفاوت مشخص گردیده است.



شکل 2: پاسخ فرکانسی بانک فیلتر گاماتون که در آن مساحت هر ناحیه از پاسخ فرکانسی به یک نرمالیزه شده است

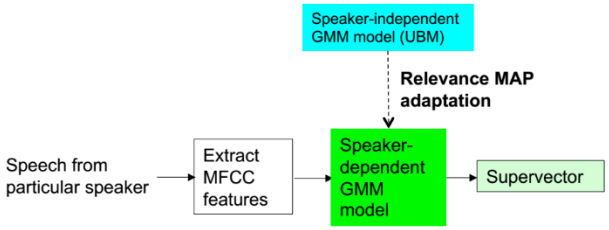


شکل 3 : مقایسه ساختار دو روش استخراج ویژگی MFCC و PNCC [6]

۳- مدل‌سازی گوینده با i-Vector

مدل‌سازی مشخصه‌های آکوستیکی وابسته به گوینده، اثرات ناشی از تفاوت محتویات متنی نمونه‌های صحبت آموزش و آزمون را کمینه می‌کند. در این مقاله مدل‌سازی با استفاده از بردارهای هویت i-vector انجام شده است که توسط FA از نمونه‌های صحبت استخراج می‌گردد.

هدف اصلی در i-vector کاهش ابعاد ابربردار GMM است. بنابراین ایده استفاده از i-vector همان ایده استفاده از ابربردارهاست، یعنی نمایش قطعات صدا به صورت بردارهای هم اندازه. ابربردار به برداری با ابعاد بالا گفته می‌شود که از ترکیب تعداد زیادی بردار با ابعاد کوچک ساخته می‌شود. همانطور که در شکل 4 نشان داده شده است، در تشخیص گوینده برای ساختن ابربردار یک فایل صوتی معمولا بردارهای میانگین مثلاً d-بعدی k آمیزه گاوسی که برای آن قطعه تطبیق داده شده را در یک بردار $k \times d$-بعدی قرار می‌دهند [2].



شکل 4: استخراج ابربردار از GMM تطبیق داده شده برای نمونه گفتار یک گوینده

تطبیق MAP علاوه بر مشخصات مختص به گوینده، اطلاعات اضافی همانند اطلاعات کانال و نویز را هم انتقال می‌دهد. بنابراین ابربردار به دست آمده از این طریق ایده‌آل نیست. JFA فرض می‌کند که هم اطلاعات گوینده و هم اطلاعات کانال در زیرفضاهای با ابعاد کمتری از فضای ابربردار GMM نهفته است و در پی استخراج آنهاست. این زیرفضاها با ماتریس‌های $\mathbf{V}$ و $\mathbf{U}$ مشخص می‌شوند. برای یک نمونه صوتی گوینده s در جلسه h ابربردار میانگین GMM به صورت زیر است

$$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{V}_{y_S} + \mathbf{U}_{x_h} + \mathbf{D}_{z_{s,h}} \qquad (2)$$

که در آن $\mathbf{m}_0$ ابربردار مستقل از گوینده است که از UBM به دست می‌آید، $\mathbf{V}$ ماتریس eigenvoice، y فاکتورهای گوینده با فرض توزیع اولیه نرمال با میانگین صفر و واریانس یک $N(0,1)$ ، $\mathbf{U}$ ماتریس eigenchannel، x فاکتورهای کانال با توزیع اولیه نرمال N(0,1)، D ماتریس اطلاعات باقیمانده‌ها (residual) و $\mathbf{z}$ فاکتورهای باقیمانده خاص گوینده با توزیع اولیه نرمال $N(0,1)$ است. برای محاسبه ماتریس $\mathbf{V}$، ماتریس‌های $\mathbf{U}$ و $\mathbf{D}$ صفر در نظر گرفته می‌شود، محاسبه $\mathbf{V}$ با تخمین زده شده و صفر در نظر گرفتن ماتریس $\mathbf{D}$ به‌دست می‌آید و سپس با تخمین های به دست آمده از $\mathbf{V}$ و $\mathbf{U}$ ماتریس $\mathbf{D}$ محاسبه می‌شود. در نهایت با استفاده از این سه ماتریس، فاکتورهای گوینده y، کانال x و باقیمانده $\mathbf{z}$ به دست می‌آید [9].

روش i-vector با در نظر گرفتن این واقعیت که اطلاعات کانال شامل اطلاعات گوینده نیز هست، فاکتورهای گوینده و کانال را در یک فضای واحد به نام فضای تغییرپذیر کلی[۶] ترکیب می‌کند. در این مدل ابربردار GMM وابسته به گوینده و جلسه توسط رابطه زیر نشان داده می‌شود:

$$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{T}_{w_{s,h}} \qquad (3)$$

که در آن متغیرهای پنهان $w_{s,h} \sim N(0,1)$ فاکتورهای کلی و $\mathbf{T}$ ماتریس تغییرپذیری کلی نامیده می‌شوند. آموزش ماتریس $\mathbf{T}$ دقیقا" به همان روش آموزش ماتریس $\mathbf{V}$ در JFA انجام می‌شود. تخمین‌های به دست آمده از فاکتورهای کلی در رابطه (۳) همان i-vector ها هستند.

برای ایجاد مدل هر گوینده، پس از آن که برای هر فایل صوتی مربوط به آن گوینده یک i-vector محاسبه شد، با میانگین‌گیری بین این بردارها یک i-vector به عنوان مدل گوینده به‌دست می‌آید. در مرحله آزمون نیز

ابتدا i-vector مربوط به فایل صوتی آزمون محاسبه می‌شود و سپس توسط الگوریتم‌های چون محاسبه فاصله کسینوسی و یا آنالیز جداساز خطی احتمالی[7] (PLDA) فرایند مقایسه و امتیازدهی انجام می‌گیرد.

برخلاف JFA روش i-vector تفکیکی بین گوینده و کانال قائل نیست. بنابراین در ذات خود جبران‌سازی خاصی انجام نمی‌دهد و سیستم تشخیص گوینده را مقاوم نمی‌کند. بلکه با کاهش معنادار ابعاد ابربردار GMM (معمولاً به ۴۰۰ تا ۸۰۰) علاوه بر دارا بودن اغلب مزایای ابربردارها، امکان پیاده‌سازی روش‌های جبران‌سازی چون آنالیز جداساز خطی[8] (LDA) و نرمالیزاسیون کواریانسی درون گروهی[9] (WCCN) را فراهم می‌کند که پیش از آن به دلیل ابعاد بسیار بالای ابربردارها در عمل امکان‌پذیر نبود [2].

۴ – ارزیابی کارایی سیستم تایید گوینده

۴-۱- دادگان سیگنال گفتار

در این مقاله از دادگان سیگنال گفتارTIMIT استفاده شده است [12]. این دادگان شامل نمونه‌های گفتار ۶۳۰ گوینده است که ۴۳۸ مرد و ۱۹۲ زن را در بر می‌گیرد. برای هر گوینده ۱۰ جمله کوتاه وجود دارد. در این مقاله تنها از نمونه‌های گفتاری گویندگان مرد استفاده شده است. ۳۶۸ گوینده مرد و از هر کدام ده جمله برای ایجاد مدل پس زمینه جهانی، ۷۰ گوینده مرد دیگر و از هر کدام ۹ جمله برای آموزش و یک جمله ۳ ثانیه‌ای برای آزمون در نظر گرفته شده است. تعداد کل آزمایش‌های تایید هویت گوینده ۴۹۰۰ می‌باشد.

برای ارزیابی عملکرد سیستم در شرایط ورودی گفتار نویزی داده‌های نویزی NOISEX92 با نسبت‌های سیگنال به نویز صفر، ۵ و ۱۰ دسی‌بل به دادگان تمیز TIMIT افزوده شده‌اند [13]. در این مقاله عملکرد سیستم برای گفتارهای همراه با نویزهای سفید، همهمه و ماشین مورد بررسی قرار گرفته‌اند.

۴-۲- پیکربندی سیستم

برای قطعه‌بندی سیگنال گفتار از پنجره همینگ با عرض ۳۰ میلی‌ثانیه و همپوشانی ۱۵ میلی‌ثانیه استفاده شده است. ابعاد بردار ویژگی استفاده شده برای همه انواع ویژگی‌های به کار رفته یعنی MFCC، IMFCC، LFCC و PNCC یکسان بوده متشکل از ۱۲ ضریب اصلی به همراه مشتق اول و دوم آن‌هااست که یک بردار ویژگی ۳۶ بعدی شکل می‌دهند. از به کارگیری ضریب کپسترال صفر صرف نظر شده است. در ورودی از فیلتر پیش تاکید (فیلتر بالاگذر مرتبه اول)، با ضریب ۰,۹۷ استفاده شده است.

به منظور حذف اطلاعات زاید از داده‌های گفتار از روش آماری مورد استفاده در [14] برای حذف سکوت استفاده شده است. برای بهبود عملکرد سیستم در حضور نویز همراه با گفتار از تفریق طیفی در حوزه دامنه طیف استفاده شده است. به این ترتیب که پس از تخمین توان طیف نویز این مقدار از توان طیف سیگنال کم شده و مقادیر منفی با صفر جایگزین شده است. سپس طیف توان به دست آمده جدید با فاز اصلی بازترکیب شده و سیگنال گفتار بهبود یافته ایجاد شده است [15]. برای کاهش اثر عدم تطبیق‌های احتمالی، تابع feature warping توسط یک پنجره ۳ ثانیه‌ای به بردارهای ویژگی اعمال شده تا توزیع آنها را به یک توزیع نرمال استاندارد یعنی $N(0,1)$ تبدیل کند.

از ۲۵۶ آمیزه گاوسی برای ایجاد مدل پس زمینه جهانی استفاده شد. پیش‌آزمون‌های انجام شده با تعداد متفاوت آمیزه‌های گاوسی نشان داده بود که انتخاب ۲۵۶ گزینه بهینه است. با استفاده از UBM به دست آمده و آنالیز عامل، i-vector مربوط به هر فایل صوتی با طول ۴۰۰ استخراج گردید. سپس توسط آنالیز جداساز خطی ابعاد i-vectorها به ۱۰۰ کاهش یافت. کاهش ابعاد i-vectorها توسط LDA علاوه بر کاستن از بار محاسباتی سیستم، با افزایش تغییرات بین گروهی و کاهش تغییرات درون گروهی، تفکیک‌پذیری بین گویندگان را نیز افزایش می‌دهد. مدل هر گوینده با میانگین‌گیری بین i-vector های به دست آمده برای فایل‌های آموزشی آن گوینده ایجاد شد. در نهایت با محاسبه i-vector برای فایل‌های آزمون و با استفاده از الگوریتم PLDA امتیازدهی به آزمون‌ها انجام گرفت.

۴-۳- معیار ارزیابی

ارزیابی سیستم تایید هویت گوینده بر مبنای میزان خطا در آزمون‌ها و ترسیم نمودار مصالحه آشکارسازی خطا[10] (DET) انجام شده است که عملکرد سیستم را با مقادیر آستانه متفاوت نشان می‌دهد. این نمودار بده‌بستانی است از خطای پذیرش اشتباه[11] (FA) و رد اشتباه[12] (FR). همچنین از تابع هزینه تصمیم[13] (DCF) که توسط NIST معرفی شده و به صورت زیر تعریف می‌گردد نیز استفاده شده است:

$$DCF = C_{miss} E_{miss} P_{t\,arget} + C_{fa} E_{fa} (1 - P_{t\,arget}) \qquad (4)$$

که در آن E_{miss} و E_{fa} به ترتیب خطای رد اشتباه و پذیرش اشتباه هستند. P_{target} احتمال پیشین گویندگان واقعی، C_{miss} هزینه از دست دادن اشتباه و C_{fa} هزینه پذیرش اشتباه است که مقادیر پیشنهادی NIST برای این سه پارامتر به ترتیب ۰,۰۱، ۱۰ و ۱ می باشد [16]. نقطه بهینه جایی است که مقدار این تابع هزینه کمینه شود. با توجه به مقادیر پارامترهای ثابت، نقطه بهینه به سمت نرخ خطای پذیرش اشتباه کمتر متمایل می‌شود. خطاهای رد اشتباه و پذیرش اشتباه را می‌توان با تنظیم حدآستانه تصمیم‌گیری سبک- سنگین کرد. از دید قابلیت جداسازی و بدون در نظر گرفتن کاربرد، نقطه بهینه جایی است که دو نرخ خطای FA و FR با هم برابر شوند که آن نقطه نرخ برابری خطا[14] (EER) نامیده می‌شود. در ارزیابی EER هم محاسبه شده که ارزیابی کامل‌تری از کارایی سیستم حاصل گردد.

۴-۴- پیاده‌سازی آزمون

سیستم تایید هویت گوینده بر اساس جعبه ابزار MSR مایکروسافت پیاده‌سازی شده است [17]. در این سیستم عملکرد ویژگی‌های MFCC, IMFCC, LFCC و PNCC با روش مدل‌سازی i-vector-PLDA با یکدیگر مقایسه شده است. آزمایش‌ها با سیگنال‌های گفتار در شرایط تمیز

و همراه با نویزهای سفید، همهمه و ماشین با سطوح سیگنال به نویز صفر، ۵ و ۱۰ دسی‌بل انجام شده است.

در جدول (۱) عملکرد ۴ بردار ویژگی در شرایط تمیز و نویزی از نظر نرخ برابری خطا با یکدیگر مقایسه شده‌اند. همچنین برای بررسی استفاده از الگوریتم کاهش اثر نویز بر عملکرد سیستم تایید هویت، آزمون‌هایی نیز برای حالت استفاده از الگوریتم بهبود کیفیت گفتار تفریق طیفی در ورودی سیستم انجام گردید. تاثیر این الگوریتم بر نتیجه نهایی برای این چهار بردار ویژگی از نظر نرخ برابری خطا در جدول (۲) ارائه شده است.

جدول 1: مقایسه نرخ برابری خطای حاصل از سیستم مبتنی بر بردارهای ویژگی چهارگانه در شرایط تمیز و نویزی

	نویز ماشین (dB)			نویز همهمه (dB)			نویز سفید (dB)			تمیز
	صفر	۵	۱۰	صفر	۵	۱۰	صفر	۵	۱۰	
MFCC	۲/۷	۲/۷	۲/۵	۶/۲	۴/۵	۳/۵	۱۰/۰	۴/۸	۵/۷	۲/۵
IMFCC	**۱/۹**	**۲/۱**	۲/۰	۸/۰	**۲/۹**	۲/۹	۱۱/۴	۸/۳	۵/۷	۱/۵
LFCC	۱/۸	۲/۸	۲/۲	**۵/۲**	۳/۴	۲/۹	۱۰/۰	۵/۲	**۴/۳**	**۱/۱**
PNCC	۲/۱	۳/۲	**۱/۹**	۵/۹	۳/۵	**۱/۶**	**۷/۱**	**۴/۴**	**۴/۳**	۱/۴

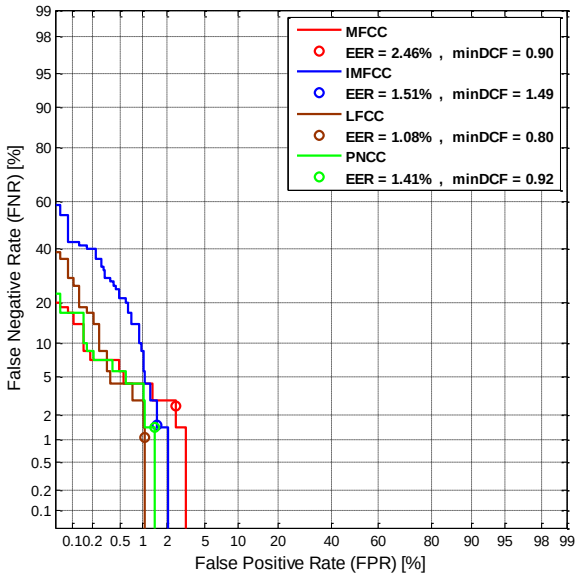
جدول 2: مقایسه نرخ برابری خطای حاصل از سیستم مبتنی بر بردارهای ویژگی چهارگانه در شرایط تمیز و نویزی در صورت استفاده از الگوریتم تفریق طیفی

	نویز ماشین (dB)			نویز همهمه (dB)			نویز سفید (dB)			تمیز
	صفر	۵	۱۰	صفر	۵	۱۰	صفر	۵	۱۰	
MFCC	**۱/۹**	۲/۷	**۱/۴**	۸/۸	۴/۴	۴/۳	۸/۶	۶/۲	**۳/۱**	۲/۵
IMFCC	۲/۹	۳/۶	۳/۴	**۷/۱**	**۳/۶**	**۳/۴**	۱۰/۰	۷/۱	۷/۴	۱/۵
LFCC	۲/۰	۲/۵	**۲/۳**	۹/۳	۵/۷	۴/۳	۹/۲	۵/۷	۵/۲	**۱/۱**
PNCC	۲/۷	۲/۴	۲/۹	۷/۸	۵/۷	۵/۱	**۶/۵**	**۴/۳**	۳/۸	۱/۴

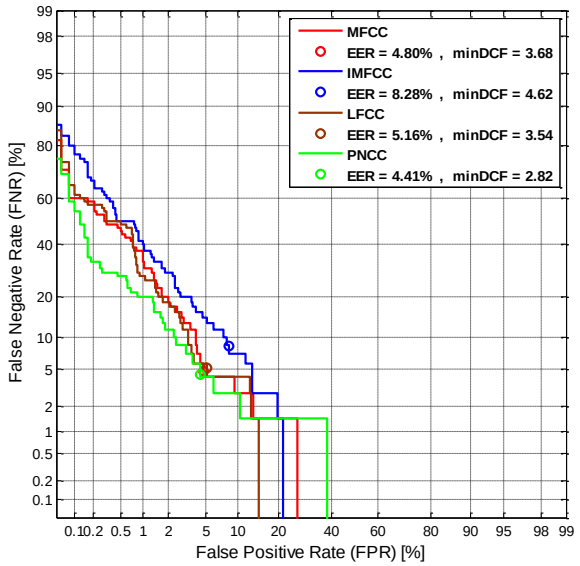
نتایج جدول 1 نشان می‌دهد که در شرایط تمیز استخراج ویژگی با روش LFCC عملکرد بهتری نسبت به بقیه ویژگی‌ها دارد و در حضور نویز سفید عملکرد ضرایب PNCC در مقایسه با دیگر روش‌ها برتر است، در حالی که در حضور نویز همهمه و ماشین ضرایب IMFCC بهبود بیشتری در کارایی سیستم ایجاد می‌کنند. نویز سفید به دلیل حضور نویز در سراسر باند فرکانسی هر چهار روش استخراج ویژگی را تقریباً به یک میزان تحت تاثیر اثر مخرب خود قرار می‌دهد. ضرایب MFCC و PNCC بیشتر تحت تاثیر نویز همهمه قرار می‌گیرند که احتمالاً به دلیل تاکید بیشتر فیلتر بانک‌های به کار رفته در این دو روش بر فرکانس‌های پایین‌تر است. نویز ماشین یک نویز فرکانس پایین است، به همین دلیل تأثیر مخرب آن تا میزان قابل توجهی توسط فیلتر پیش تاکید کاهش می‌یابد و تاثیر چندانی بر عملکرد سیستم تایید هویت ندارد. همچنین فیلتر پیش تاکید در بهبود کارایی سیستم در حضور نویز همهمه هم مفید است.

نتایج جدول 2 مبین آن است که افزودن الگوریتم تفریق طیفی عملکرد سیستم تایید هویت گوینده را برای سیگنال‌ها گفتار همراه با نویز سفید بهبود می‌بخشد ولی بر نویز ماشین تاثیر چندانی نداشته و حتی برای سیگنال گفتار همراه با نویز همهمه دقت سیستم را کاهش می‌دهد. بهبود عملکرد تنها برای نویز سفید، احتمالاً به دلیل تناسب این نوع نویز به لحاظ ویژگی‌های آماری (توزیع تقریباً یکسان در تمام باند فرکانسی) با و فرض مورد استفاده در الگوریتم تفریق طیفی است که سراسر باند فرکانسی را به طور یکسان مورد اصلاح قرار می‌دهد.

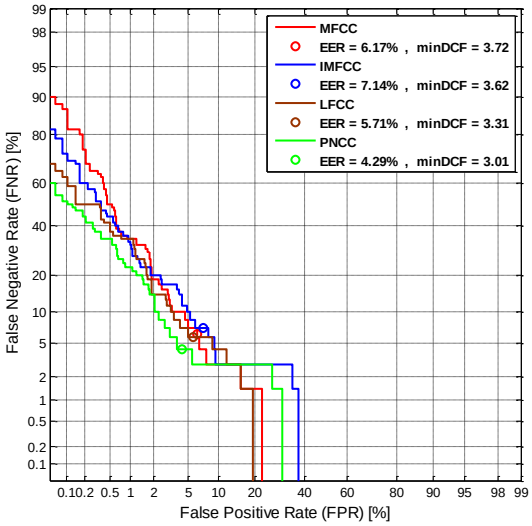
در شکل‌های 5، 6 و 7 عملکرد سیستم تایید هویت گوینده در شرایط مختلف و با به کارگیری ۴ ویژگی MFCC، IMFCC، LFCC و PNCC در شرایط گفتار تمیز، و همچنین گفتار همراه با نویز سفید با نسبت سیگنال به نویز ۵ dB در حالت عادی و استفاده از الگوریتم تفریق طیفی توسط نمودار مصالحه آشکارسازی خطا (DET) و دو معیار ارزیابی نرخ خطای برابر (EER) و حداقل تابع هزینه تصمیم[۱۵] (minDCF) در مقایسه با یکدیگر نمایش داده شده است.



شکل 5: عملکرد سیستم مبتنی بر ویژگی‌های متفاوت در شرایط تمیز



شکل 6: مقایسه عملکرد سیستم مبتنی بر ویژگی‌های متفاوت در حضور نویز سفید با سیگنال به نویز 5 dB



شکل 7: عملکرد سیستم مبتنی بر ویژگی‌های متفاوت در حضور نویز سفید با سیگنال به نویز dB ۵ پس از بهبود گفتار نویزی با تفریق طیفی

۵- نتیجه گیری

در این مقاله یک سیستم تایید هویت گوینده مبتنی بر مدل‌سازی i-vector-PLDA مورد بررسی قرار گرفت و عملکرد آن در حالت بهره‌گیری از چهار بردار ویژگی رایج ارزیابی شد. آزمون‌های انجام شده جهت بررسی تاثیر بردارهای مختلف ویژگی بر عملکرد سیستم تایید هویت در حالت وجود سیگنال گفتاری همراه با نویزهای سفید، همهمه و ماشین نشان داد که استفاده از بردارهای ویژگی IMFCC و PNCC کارایی بهتری را نسبت به بردارهای ویژگی MFCC در پی دارد و استفاده از بهبود گفتار نویزی با تفریق طیفی عملکرد سیستم تایید هویت گوینده را تنها برای سیگنال گفتار همراه با نویز سفید بهبود می‌دهد.

ادامه این تحقیق با هدف بهره‌گیری از سایر رویکردهای رایج در مدل‌سازی گوینده و شیوه‌های کارآمدتر بهبود سیگنال گفتار همراه با نویز به عنوان پیش‌پردازش سیستم تایید هویت گوینده و ترکیب نتایج به دست آمده از ویژگی‌های مختلف با یکدیگر برای دستیابی به نتیجه بهتر پی‌گرفته خواهد شد.

مراجع

[1] R. de Luis-García, C. Alberola-López, O. Aghzout, and J. Ruiz-Alzola, "Biometric identification systems," *Signal Processing*, vol. 83, pp. 2539-2557, 2003.

[2] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, pp. 74-99, 2015.

[3] R. Saeidi, "Advances in Front-end and Back-end for Speaker Recognition," *Nition: A Feature Based Aapproach*, vol. 13, pp. 58-71, 2011.

[4] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12-40, 1// 2010.

[5] C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, pp. 69-81, 1993.

[6] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4101-4104.

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Odyssey*, Crete, Greece, Jun. 2001.

[8] T. Ganchev, I. Potamitis, N. Fakotakis, and G. Kokkinakis, "Text-independent speaker verification for real fast-varying noisy environments," *International Journal of Speech Technology*, vol. 7, pp. 281-292, 2004.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 788-798, 2011.

[10] S. Chakroborty, A. Roy, S. Majumdar, and G. Saha, "Capturing Complementary Information via Reversed Filter Bank and Parallel Implementation with MFCC for Improved Text-Independent Speaker Identification," in *Proc. Int. Conf. on Computing: Theory and Applications, 2007. ICCTA '07.*, 2007, pp. 463-467.

[11] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Sixteenth Annual Conf. of the Int. Speech Communication Association*, 2015.

[12] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, pp. 351-356, 1990.

[13] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247-251, 7// 1993.

[14] S. Jongseo, K. Nam Soo, and S. Wonyong, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, pp. 1-3, 1999.

[15] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP '79.*, 1979, pp. 208-211.

[16] "The NIST Year 2008 Speaker Recognition Evaluation Plan," [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/.

[17] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1. 0: A MATLAB Toolbox for Speaker Recognition Research," *Speech and Language Processing Technical Committee Newsletter*, 2013.

پانویس ها

[1] Mel Frequency Cepstral Coefficients
[2] Weighted Linear Prediction
[3] Power Normalized Cepstral Coefficients
[4] Inverted Mel Frequency Cepstral Coefficients
[5] Linear Frequency Cepstral Coefficients
[6] Total Variability Space
[7] Probabilistic Linear Discriminant Analysis
[8] Linear Discriminant Analysis
[9] Within-Class Covariance Normalization
[10] Detection Error Trade-off
[11] False Acceptance
[12] False Rejection
[13] Decision Cost Function
[14] Equal Error Rate
[15] Minimum Decision Cost Function