علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۸، شماره ۲، ۱۳۹۸
صفحات ۷۱- ۶۲
مقاله عادی

بهبود سامانه‌های تصدیق هویت گوینده برای گفتارهای آلوده به نویز با استفاده از بردارهای هویت موزون

محسن محمدی[۱]، حمیدرضا صادق محمدی[۲*]

*نویسنده مسئول، دریافت: ۹۸/۱۱/۲۰، بازنگری: ۹۹/۰۲/۱۵، پذیرش: ۹۹/۰۲/۲۹
[۱] دانشجوی دکتری، مهندسی برق-مخابرات، پژوهشکده برق جهاد دانشگاهی، تهران، ایران
[۲] دانشیار، مهندسی برق، پژوهشکده برق جهاد دانشگاهی، تهران، ایران

چکیده

دسترسی ایمن به سامانه‌های کاربردی متفاوت از فواصل دور و نزدیک، کاربرپسند بودن، پیچیدگی محاسباتی کم و هزینه پیاده‌سازی پایین از ویژگی‌های برجسته روش تصدیق هویت مبتنی بر گفتار است. اما کارایی این شیوه در محیط‌های واقعی به دلیل وجود نویزهای متفاوت صوتی و عوارض کانال به‌شدت افت می‌کند. روش i-vector PLDA ازجمله شیوه‌های موفق در بهبود عملکرد سامانه‌های تصدیق هویت گوینده است. در این مقاله بهره‌مندی از ویژگی‌های آماری بردارهای ثبت‌نام گویندگان هدف برای وزن‌دهی به بردارهای مدل و تست، جهت بهبود دقت امتیازدهی و درنتیجه عملکرد سامانه تصدیق هویت در شرایط آزمون گفتار نویزی پیشنهاد گردیده است. تأثیر استفاده از این بردارهای وزن داده شده، که آن را بردارهای موزون نامیده‌ایم، بر عملکرد سامانه در محیط‌های نویزی مورد ارزیابی قرار گرفته است. آموزش‌ها و آزمون‌ها با استفاده از دادگان گفتار TIMIT، بردارهای ویژگی MFCC و PNCC و روش امتیازدهی PLDA انجام شده است. همچنین برای بهبود عملکرد سامانه در شرایط عدم تطابق نویز، بین گفتار ثبت‌نام و آزمون، از آموزش چند-شرطی برای LDA و PLDA استفاده شده است. همچنین ترکیب امتیازات این آزمون‌ها نیز مورد ارزیابی قرار گرفت. نتایج آزمون‌ها مبین آن است که بهره‌گیری از بردارهای موزون دقت سامانه تصدیق هویت گوینده را برای گفتارهای نویزی نیز افزایش می‌دهد، علاوه بر آن در اکثر قریب به اتفاق موارد ترکیب امتیازات آزمون‌ها نیز عملکرد سامانه را بهبود می‌بخشد.

کلمات کلیدی: تصدیق هویت گوینده، وزن‌دهی بردار، نویز، بردار هویت، PLDA، multi-condition، I-vector.

۱- مقدمه

ازجمله محبوب‌ترین و کارآمدترین شیوه‌های شناسایی خودکار افراد استفاده از ویژگی‌های حیاتی آن‌هاست. از میان معروف‌ترین ویژگی‌ها، می‌توان از اثرانگشت، چهره و صدای افراد یاد کرد. هیچ‌یک از سامانه‌های تشخیص هویتی در تمامی شرایط به‌طور مطلق بهترین راهکار نیست و هر یک با توجه به‌دقت و کاربرد موردنظر نقاط قوت و ضعف خاص خود را دارند. در این میان پدیده‌هایی وجود دارند که سیگنال گفتار افراد را از دیگر ویژگی‌ها متمایز می‌سازد. گفتار یک سیگنال طبیعی است و معمولاً تولید گفتار یک فرد برای فردی دیگر به‌صورت طبیعی ممکن نیست. در بسیاری از کاربردها همانند دسترسی یا ارتباط از راه دور با پهنای باند کم‌نظیر تلفن، گفتار ساده‌ترین سیگنال قابل‌دسترسی از افراد است. همچنین تشخیص از طریق گفتار کاربرپسند بوده و نیاز به تجهیزات و حسگرهای ویژه و گران‌قیمتی ندارد [۱].

تشخیص گوینده را معمولاً به دو شاخه اصلی شناسایی هویت و تصدیق هویت گوینده تقسیم می‌کنند. در شناسایی گوینده، هویت گوینده یک قطعه گفتار از بین چند گوینده از قبل تعیین‌شده مشخص می‌گردد، در مقابل در تصدیق گوینده بایستی ادعای مطرح‌شده درباره هویت گوینده یک نمونه صوتی رد شده یا تأیید گردد. بررسی هویت افراد در سامانه‌های کنترل دسترسی، برچسب‌گذاری صوت و تشخیص هویت قانونی ازجمله مهم‌ترین کاربردهای تصدیق هویت گوینده هستند.

از بعد علمی تصدیق هویت گوینده خود به دو دسته وابسته به متن و مستقل از متن تقسیم می‌شود که سامانه‌های مستقل از متن باوجود دقت کمتر آن‌ها به دلیل دارا بودن سازگاری بیشتر با شرایط محیطی واقعی کاربرد گسترده‌تری دارند. سامانه‌های تصدیق هویت گوینده از چند بخش اصلی شامل استخراج ویژگی، مدل‌سازی گوینده، امتیازدهی، مقایسه و تصمیم‌گیری تشکیل شده‌اند [۲].

هرچند در دهه اول قرن حاضر روش‌های مبتنی بر مدل آمیزه‌های گوسی[۱] نظیر GMM-UBM روش پایه برای اغلب سیستم‌های تأیید گوینده محسوب می‌گردید [۳]، اما در طی دهه اخیر پیشرفت‌های چشمگیری در حوزه تصدیق گوینده به وقوع پیوسته است و شیوه‌های جدیدی در این حوزه معرفی شده‌اند. استفاده از روش تحلیل تفکیک‌کننده خطی احتمالاتی (PLDA)[۲] در سال ۲۰۱۰ برای جبران اثرات منفی کانال با کاهش تغییرات درون‌کلاسی و افزایش تغییرات بین‌کلاسی [۴]، معرفی بردار هویت[۳] (i-vector) در سال ۲۰۱۱ به‌عنوان نمایش برداری با طول ثابت از قطعات با طول متفاوت صدا در فضای گوینده و کانال [۵] و گزارش‌های موفق از کاربرد روش‌های مبتنی بر شبکه‌های عصبی و یادگیری عمیق[۴](DNN) از سال ۲۰۱۴ [۶، ۷] ازجمله این پیشرفت‌ها هستند. علاوه بر آن، تاکنون راهبردهای گوناگونی برای استفاده از شبکه‌های عمیق در این خصوص پیشنهاد شده است. ازجمله شبکه‌هایی که برای استخراج آماره‌های پسین جهت استفاده در فرایند استخراج i-vector به کار می‌روند (DNN i-vector) [۷]، شبکه‌هایی که برای استخراج ویژگی‌های گلوگاهی آموزش داده شده و برخی از آن‌ها با استخراج تنها یک بردار به ازای هر نمونه گفتار جایگزینی برای i-vector بوده و عملکرد موفقی در ترکیب با PLDA از خود نشان داده‌اند (x-vector) [۸] و شبکه‌هایی که به‌صورت انتها به انتها فرایند تشخیص گوینده را انجام می‌دهند [۹] ازجمله این راهبردها هستند. البته بار محاسباتی بالای این روش‌ها و نیاز آن‌ها به داده‌های زیاد و متنوع ازجمله حجم قابل‌ملاحظه نمونه‌های ثبت‌نام از هر گوینده که دسترسی به آن در شرایط دنیای واقعی ازجمله در کاربردهای قانونی تشخیص گوینده دشوار است و ضعف این روش‌ها در شرایط عدم تطابق[۵] کانال و/یا نویز موجب گردیده است که با توجه به کاربردها و شرایط کاری متفاوت همچنان نتایج موفقی از روش‌هایی مبتنی بر i-vector گزارش شود [۱۰].

در میان روش‌های آماری، روش‌های مبتنی بر تحلیل عامل عملکرد برتری نسبت به سایر روش‌ها داشته‌اند. ازجمله مهم‌ترین این روش‌ها تحلیل عامل توأم[۶] (JFA) و بردار هویت قابل ذکر هستند. ایده استفاده از تحلیل عامل در فضای ابر بردارهای GMM[۷] اولین بار در سال ۲۰۰۴ مطرح شد [۲]. از آن زمان تاکنون روش‌های مختلفی بر مبنای تحلیل عامل ارائه شده‌اند که درنهایت منجر به پیشنهاد روش i-vector در سال ۲۰۱۱ گردید [۵]. مهم‌ترین ویژگی بردار هویت تبدیل قطعات گفتار با طول زمانی متفاوت به برداری با طول یکسان و نسبتاً کوتاه است که بهره‌گیری از آن کاهش بار محاسباتی، سازگاری بهتر با روش‌های عمومی یادگیری ماشین و جبران سازی اثرات منفی کانال را در پی دارد. همین امر علاقه‌مندی به استفاده از آن را به‌سرعت افزایش داده و محققان بسیاری را بر آن داشته تا سامانه‌های جدید موردنظر را در این حوزه پیشنهاد و پیاده‌سازی نمایند [۱۱].

برخلاف JFA، روش i-vector تفکیکی بین اطلاعات گوینده و کانال قائل نیست. بنابراین ذاتاً در آن جبران سازی خاصی انجام نگرفته و تاب‌آوری سامانه تشخیص گوینده را بهبود نمی‌بخشد، بلکه با کاهش معنادار ابعاد ابر بردار GMM (معمولاً به ۴۰۰ تا ۸۰۰) علاوه بر دارا بودن اغلب مزایای ابر بردارها، امکان پیاده‌سازی بهینه روش‌های جبران سازی چون تحلیل تفکیک‌کننده خطی[۸] (LDA) و هنجار سازی کوواریانس درون‌گروهی[۹] (WCCN) را فراهم می‌سازد که پیش از آن به دلیل ابعاد بزرگ ابر بردارها در عمل کاربرد چندانی نداشتند [۵].

یکی از بخش‌های مهم سامانه تصدیق هویت گوینده، امتیازدهی به آزمون‌ها است. در فضای i-vector روش‌های مختلفی برای امتیازدهی معرفی شده‌اند که برخی مانند فاصله اقلیدسی و فاصله کسینوسی [۵] تنها بر مبنای فاصله بین بردار ثبت‌نام (مدل) و آزمون (تست) می‌باشند و برخی دیگر مانند فاصله ماهالانوبیس[۱۰] [۱۲] و PLDA [۴] از اطلاعات آماری بردارهای هویت حاصل از دادگان توسعه نیز برای افزایش دقت سامانه استفاده می‌کنند. شیوه‌های گوناگونی برای بهبود عملکرد مرحله امتیازدهی توسط محققان پیشنهاد گردیده که اغلب با اعمال تغییراتی در بردارها به این مهم دست می‌یابند. کاهش ابعاد بردارها و درعین‌حال افزایش تفاوت‌های بین‌گروهی و کاهش تفاوت‌های درون‌گروهی با روش‌هایی چون LDA و WCCN، به‌منظور جبران اثرات مخرب کانال، ازجمله این شیوه‌ها محسوب می‌شوند [۵]. هنجار سازی[۱۱] بردارها از دیگر روش‌های مؤثر در کاهش خطای سامانه‌های تصدیق هویت گوینده مبتنی بر i-vector است. همچنین، سفیدسازی[۱۲] و تغییر اندازه بردارها به یک[۱۳] از پرکاربردترین و مؤثرترین این روش‌ها به شمار می‌آیند. تبدیل سفیدسازی یا بیضی کردن، یک تبدیل خطی است که برای نا همبسته کردن رشته‌ای از داده‌ها به کار می‌رود. داده‌های سفید شده دارای میانگین صفر، واریانس یک و همبستگی صفر می‌باشند. از سوی دیگر اندازه بردارهای هویت تا حدود زیادی متأثر از جلسه، کانال و مدت‌زمان قطعه گفتاری است که بردار از آن استخراج شده و بنابراین هنجار سازی اندازه بردارها عملکرد سامانه را بهبود می‌بخشد [۱۳]. LDA با منبع هنجار شده و موزون [۱۴] و وزن‌دهی به بردارهای هویت بر اساس مدت‌زمان قطعه گفتاری که این بردارها از آن استخراج شده‌اند [۱۵] از دیگر روش‌های کاهش خطا هستند. ازجمله مؤثرترین شیوه‌ها برای افزایش دقت سامانه، اعمال وزن‌های متفاوت به هر یک از درایه‌های بردار هویت است. فاصله ماهالانوبیس یکی از این روش‌هاست که فاصله اقلیدسی وزن‌دار نیز نامیده می‌شود. وزن‌های به‌کاررفته در این روش بر مبنای عکس واریانس دادگان توسعه به دست می‌آید و روش‌های مختلفی برای محاسبه این وزن‌ها در مراجع متفاوت پیشنهاد شده است [۱۲، ۱۴، ۱۶].

نتایج عملکرد سامانه‌های تصدیق گوینده در محیط‌های آزمایشگاهی با گفتار تمیز خوب است، اما در محیط‌های زندگی واقعی که عوامل مزاحمی همانند انواع نویزهای جمع‌شونده، کانولوشنی، اعوجاج‌های کانال و یا پژواک محیطی وجود دارند، دقت آن‌ها با افت شدیدی همراه است [۲]. بنابراین یکی از عوامل مهم برای هر سامانه تصدیق گوینده مطلوب، تاب‌آوری آن در برابر نویزهای محیطی است. همین امر باعث می‌گردد که روش‌های مقاوم‌سازی سامانه در مقابل تأثیر نویز در سامانه‌های تصدیق گوینده از اهمیت شایان توجهی برخوردار باشند. بااینکه تطبیق شرایط آموزش و آزمون ازنظر نوع نویز و سطح سیگنال به نویز بهترین نتیجه را در پی دارد، اما به علت بار محاسباتی بالای این روش و مهم‌تر از آن عدم سازگاری آن با شرایط دنیای واقعی، که در آن پیش‌بینی اعوجاجات زمان آزمون و یا تخمین دقیق آن‌ها در عملاً امکان‌پذیر نیست، بسیاری از محققان این حوزه به ارائه روش‌هایی برای غلبه بر شرایط عدم تطبیق پرداخته‌اند. یکی از روش‌های مرسوم مقاوم‌سازی در این سامانه‌ها، بهبود کیفیت گفتار به‌عنوان پیش‌پردازش برای سیستم تصدیق هویت است. روش‌هایی چون تفریق طیفی و وینر[۱۴] ازجمله این روش‌ها به شمار می‌آیند. استفاده از شبکه‌های عصبی برای بهبود گفتار نیز نتایج موفقی را در پی داشته است [۱۷].

روش I-Vector PLDA نیز علی‌رغم عملکرد بسیار موفق در شرایط تمیز، در شرایط عدم تطبیق و در جبران اثرات نامطلوب اعوجاج‌هایی چون تفاوت سطح نویز و نوع نویز دچار افت کارایی می‌شود. راهکارهای محدودی برای غلبه بر این مشکل ارائه شده‌اند که ازجمله آن‌ها می‌توان به تخمین نویز و حذف آن در فضای i-vector و محاسبه پارامترهای PLDA مستقل از سطح سیگنال به نویز اشاره کرد [۱۸،۱۹]. از دیگر روش‌های کارآمد برای غلبه بر شرایط عدم تطبیق آموزش و آزمون استفاده از آموزش چند-شرطی[۱۵] است. در این روش داده‌های آموزش را با چند نوع نویز در سطوح متفاوت نسبت سیگنال به نویز آغشته کرده و در کنار دادگان تمیز برای آموزش سامانه به کار می‌گیرند. بااینکه اعمال این روش در آموزش بخش‌های مختلف سامانه امکان‌پذیر است اما نتایج تحقیقات نشان داده است اعمال این روش در محاسبه پارامترهای LDA و آموزش PLDA علاوه بر بار محاسباتی کمتر، نتیجه عملکردی بهتری را در پی دارد [۲۰، ۲۱].

در این مقاله شیوه‌ای برای وزن‌دار کردن بردارهای هویت به‌منظور بهبود عملکرد سامانه تصدیق گوینده و افزایش تاب‌آوری آن در شرایط نویزی و عدم تطابق نویز پیشنهاد گردیده است. ساختار ادامه این مقاله به‌صورت زیر است. ابتدا در بخش ۲ کارهای مرتبط و روش‌های پیش‌زمینه استفاده شده در این پژوهش ازجمله روش استخراج ویژگی، فضای i-vector و روش‌های امتیازدهی معرفی شده و بررسی گردیده‌اند. بخش ۳ به معرفی روش پیشنهادی این مقاله اختصاص دارد. در بخش ۴ ملاحظات اجرایی این تحقیق اعم از دادگان مورد استفاده، پیکربندی سامانه آزمون، اجزا مختلف آن و روش‌های ارزیابی عملکرد شرح داده شده‌اند و نتایج حاصل از آزمون‌ها مورد بحث و ارزیابی قرار می‌گیرد. درنهایت بخش پایانی به نتیجه‌گیری از این مقاله اختصاص دارد.

۲- کارهای مرتبط و پیش‌زمینه پژوهش

در این بخش ابتدا به مرور کارهای مرتبط پرداخته شده و سپس روش‌های پایه مورد استفاده در این پژوهش شرح داده شده‌اند.

۲-۱- بردارهای ویژگی گفتار

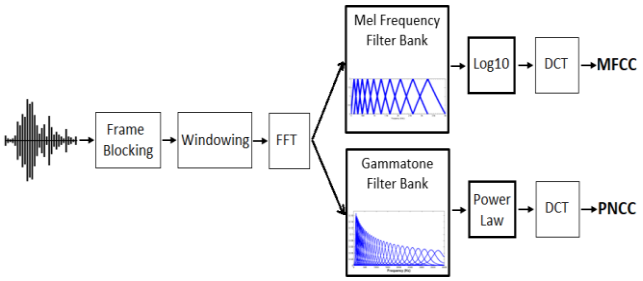
صدای هر گوینده مشخصه‌های مختص به خود را دارد. هرچند مشخصه‌های صدای گویندگان متفاوت به‌راحتی تفکیک‌پذیر نیستند، اما به دلیل تفاوت‌های فیزیولوژیکی مسیر صوتی و عادات گفتاری هر گوینده، مشخصه‌های گفتار گویندگان وابسته به شخص و منحصربه‌فرد هستند [۲۲]. بخش اعظمی از ویژگی ایده‌آل مورد نیاز برای تصدیق هویت گوینده را انرژی‌های بانک فیلتر مقیاس مل و نمایش کپسترال[۱۶] آن‌ها در بر دارد. باوجوداینکه این خصوصیات به‌صورت ذاتی وابسته به‌سلامت شخص و کیفیت کانال انتقال هستند، اما با استفاده از روش‌های ساده می‌توان این اثرات نامطلوب را کاهش داد.

در این مقاله از دو بردار ویژگی سیگنال گفتار، یعنی ضرایب کپسترال فرکانس مل[۱۷] (MFCC) و ضرایب کپسترال با توان هنجارشده[۱۸] (PNCC) استفاده شده‌اندکه از متداول‌ترین بردارهای ویژگی در سامانه‌های شناسایی گوینده به شمار می‌آیند. ایده اصلی این بردارها از خواص شنیداری انسان الهام گرفته شده و به تغییرات در فرکانس‌های پایین حساسیت بیشتری دارند. بر این اساس برای پیاده‌سازی آن‌ها بانک فیلتری طراحی می‌شود که تأکید بیشتری بر فرکانس‌های پایین دارد.

خروجی این بانک فیلتر پس از گذر از فشرده‌ساز لگاریتمی و تبدیل گسسته کسینوسی DCT ضرایب MFCC، c_n را نتیجه می‌دهد [۲۲]. اگر خروجی‌های بانک فیلتر M کاناله را به‌صورت $Y(m), m = 1, \ldots, M$ در نظر بگیریم ضرایب MFCC به‌صورت زیر محاسبه می‌شوند:

$$c_n = \sum_{m=1}^{M} [\log(Y(m))] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right]. \qquad (۱)$$

برای استخراج ضرایب PNCC از یک بانک فیلتر گاماتون ۳۰ کاناله در بازه ۱۰۰ تا ۴۰۰۰ هرتز استفاده می‌شود. پهنای باند فیلترها نیز به‌گونه‌ای در نظر گرفته می‌شوند که بر فرکانس‌های پایین تأکید بیشتری داشته باشند. جهت بهبود عملکرد سامانه ، سطح زیر نمودار هر کانال به یک هنجار سازی می‌شود. اصلی‌ترین ویژگی‌های PNCC عبارتند از: به‌کارگیری تابع غیرخطی توان (که جایگزین تبدیل لگاریتمی در MFCC می‌شود،) استفاده از الگوریتم غلبه بر نویز بر مبنای فیلتر نامتقارن و برخورداری از روشی برای پوشش زمانی[۱۹]. مجموعه این ویژگی‌ها عملکرد آن را در شرایط نویزی بهبود می‌بخشد [۲۳]. در شکل ۱ بخش‌های اصلی ساختار استخراج بردارهای ویژگی MFCC و PNCC نمایش داده شده و قسمت‌های متفاوت آن‌ها مشخص گردیده‌اند.



شکل ۱- الگوریتم استخراج بردارهای ویژگی MFCC و PNCC

۲-۲- بردار هویت I-Vector

i-vector برداری است با بعد کم که از یک قطعه گفتار استخراج شده و اطلاعات گوینده، کانال و نویز را دربر می‌گیرد. بردارهای هویت توسط تحلیل عامل از نمونه‌های صحبت استخراج می‌گردند. تحلیل عامل روابط بین متغیرهای آشکار را با استفاده از تعداد کمی متغیر پنهان بیان می‌کند. هدف اصلی در محاسبه بردار هویت کاهش ابعاد ابر بردار GMM است. بنابراین ایده استفاده از این بردارها همان ایده استفاده از ابر بردارهاست، یعنی نمایش قطعات صدا به‌صورت بردارهایی با ابعاد یکسان. ابر بردار به برداری با بعد بالا گفته می‌شود که از ترکیب تعداد زیادی بردار با بعد کوچک ساخته شود. ابر بردار GMM یک قطعه گفتار از ترکیب بردارهای میانگین آمیزه‌های گاوسی، که برای آن قطعه تطبیق داده شده‌اند، به دست می‌آید [۲].

طول بردار هویت مستقل از مدت‌زمان قطعه صوتی بوده و سیگنال گفتار با هر اندازه را با یک طول ثابت مدل می‌کند. این بردار با استفاده از مدل پس‌زمینه جهانی (UBM)[۲۰]، آمارگان بام-ولش[۲۱] و ماتریس فضای تغییرپذیری کل (T)[۲۲]، که از تحلیل عامل به‌دست‌آمده، محاسبه می‌شود [۵]. ابر بردار GMM وابسته به گوینده s و جلسه h توسط رابطه زیر نشان داده می‌شود:

$$m_{s,h} = m_0 + T\, w_{s,h} \qquad (۲)$$

که در آن m_0 ابر بردار مستقل از گوینده و کانال است که از UBM به دست می‌آید و $w_{s,h}$ عامل‌های پنهان با توزیع نرمال استاندارد هستند. در این روش با در نظر گرفتن این واقعیت که اطلاعات تفکیک‌شده برای کانال شامل اطلاعات گوینده نیز هست، عامل‌های گوینده و کانال در یک فضای واحد یعنی ماتریس تغییرپذیری کل ترکیب می‌شوند. پس از استخراج آمارگان بام-ولش مرتبه اول و دوم برای هر نمونه صوتی و به کمک UBM، با به‌کارگیری از این آمارگان، UBM و ماتریس T، بردار هویت از رابطه زیر به دست می‌آید

$$w = \left(I + T^t \Sigma^{-1} N(s) T\right)^{-1} T^t \Sigma^{-1} F(s). \qquad (۳)$$

که در آن N و F به ترتیب آمارگان مرتبه اول و دوم هستند. Σ ماتریس کوواریانس قطری است که در فرایند آموزش تحلیل عامل تخمین زده شده و تغییرات اضافی را مدل می‌کند که ماتریس T دربر نمی‌گیرد.

پس از استخراج بردارها، اثرات منفی کانال معمولاً با استفاده از LDA و WCCN جبران شده و بردارها با تبدیل سفیدسازی و تغییر اندازه‌شان به یک، هنجار سازی می‌شوند که اثر مثبت آن‌ها بر بهبود عملکرد چشمگیر است [۵، ۱۳]. پارامترهای مورد نیاز برای این روش‌ها با استفاده از بردارهای استخراج شده از گفتار تعداد زیادی گوینده، که برای هر کدام چندین فایل از جلسات مختلف وجود دارد و اشتراکی با دادگان آموزش و آزمون ندارند، محاسبه می‌شوند.

مدل هر گوینده در فضای i-vector، بردار به‌دست‌آمده از قطعه گفتار متعلق به همان گوینده است. ازآنجایی‌که در فضای i-vector استخراج بردار مدل گوینده هدف و بردار آزمون دقیقاً به یک شیوه صورت می‌گیرد، می‌توان i-vectorها را به‌عنوان بردارهای ویژگی برای تشخیص گوینده در نظر گرفت. با این نگرش تحلیل

عامل نقش استخراج‌کننده ویژگی را ایفاء می‌کند [۱۲]. در مرحله آزمون پس از استخراج بردار مربوط به نمونه گفتار تحت آزمون، امتیاز آزمایش (مقایسه بردار آزمون و بردار مدل گوینده ادعا شده) توسط معیارهایی همانند فاصله کسینوسی، فاصله ماهالانوبیس و PLDA محاسبه می‌شود.

۲-۳- PLDA

PLDA شامل دو بخش مدل‌سازی و امتیازدهی است. پس از استخراج i-vector از یک نمونه گفتار، PLDA این بردار را از منظر یک مدل مولد احتمالات مورد ملاحظه قرار می‌دهد. در PLDA گاوسی، که حالت ساده برگرفته از PLDA می‌باشد و در این مقاله مورد استفاده قرار گرفته است، دو فرض معمول در نظر گرفته می‌شود. اول اینکه اجزای گوینده و کانال از نظر آماری مستقل از یکدیگرند و دوم اینکه هر از دو توزیع گاوسی برخوردارند [۱۳].

با فرض اینکه $\{\eta_r : r = 1,...,R\}$ مجموعه بردارهای هویت متعلق به یک گوینده مشخص است، بردار هویت به‌صورت زیر قابل تجزیه است

$$\eta_r = m + \Phi\beta + \Gamma\alpha_r + \varepsilon_r . \quad (۴)$$

این مدل متشکل از دو بخش است: بخش مختص گوینده یعنی $s = m + \Phi\beta$ و بخش کانال $c_r = \Gamma\alpha_r + \varepsilon_r$. در بخش مربوط به گوینده m آفست کلی است، ستون‌های Φ و Γ به ترتیب بردارهای پایه فضای گوینده (eigenvoices) و کانال (eigenchannels) هستند، α_r و β بردارهای مشخصه پنهان با توزیع گاوسی استاندارد هستند و ε_r بخش باقیمانده با توزیع گاوسی (میانگین صفر و کوواریانس قطری Σ) است [۱۳]. علاوه بر این استقلال آماری برای متغیرهای پنهان مفروض است. به دلیل طول نسبتاً کم بردارهای هویت، یعنی ۴۰۰، Σ به‌صورت کامل در نظر گرفته شده و بخش اختصاصی کانال از رابطه (۳) حذف گردیده و درنتیجه مدل تغییریافته G-PLDA به این صورت بازنویسی می‌شود:

$$\eta_r = m + \Phi\beta + \varepsilon_r . \quad (۵)$$

پارامترهای مدل یعنی $\{m, \Phi, \Sigma\}$ با استفاده از الگوریتم بیشینه شباهت[۲۳] (EM) و از مجموعه بزرگی از داده‌ها تحت عنوان دادگان توسعه برآورد می‌شوند.

برای امتیازدهی به دو بردار مدل η_1 و آزمون η_2 در فرایند تصدیق هویت دو فرض در نظر گرفته می‌شود، نخست H_s که در آن هر دو بردار متعلق به یک گوینده با بردار مدل β هستند و دیگری H_d یعنی دو بردار مورد ارزیابی متعلق به گوینده‌های متفاوت با بردارهای مدل β_1 و β_2 هستند. درنهایت امتیاز آزمایش تصدیق گوینده به‌صورت زیر محاسبه می‌گردد:

$$score = \log \frac{p(\eta_1, \eta_2 \mid H_s)}{p(\eta_1 \mid H_d)\, p(\eta_2 \mid H_d)} . \quad (۶)$$

برای G-PLDA که فرض گاوسی بودن توزیع‌ها را دربر دارد، نرخ شباهت به این صورت بیان می‌شود:

$$score = \log N\left\{\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}; \begin{bmatrix}m\\ m\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & \Sigma_{ac}\\ \Sigma_{ac} & \Sigma_{tot}\end{bmatrix}\right\} - \log N\left\{\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}; \begin{bmatrix}m\\ m\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & 0\\ 0 & \Sigma_{tot}\end{bmatrix}\right\} \quad (۷)$$

که در آن $\Sigma_{tot} = \Phi\Phi^T + \Sigma$ و $\Sigma_{ac} = \Phi\Phi^T$ می‌باشد. جزییات بیشتر محاسبه و پیاده‌سازی در [۱۳] ذکر شده است.

۲-۴- آموزش چند-شرطی

آموزش چند-شرطی[۲۴] سامانه یک روش معروف و موفق برای افزایش تاب‌آوری تصدیق گوینده در برابر نویز است. در این نوع آموزش، نسخه‌های نویزی مختلفی از دادگان برای آموزش ایجاد می‌شوند. این نسخه‌ها شامل نویزهای مختلف در سطوح متفاوت نسبت سیگنال به نویز هستند. معمولاً در فرایند آزمون و برای بررسی شرایط عدم تطبیق، انواع مختلف نویز اعم از اینکه در فرایند آموزش چند-حالته باشد یا نباشد را مورد بررسی قرار می‌دهند. البته ازآنجایی‌که داده ثبت‌نام در حالت تمیز باقی می‌ماند شرایط عدم تطبیق مابین داده ثبت‌نام و آزمون وجود دارد.

آموزش چند-شرطی را می‌توان در مراحل مختلف سامانه به کار برد، برای مثال در بخش‌های ساخت UBM، ایجاد ماتریس T، در LDA، در آموزش PLDA و همچنین در توسعه دادگان ثبت‌نام گوینده هدف. به خدمت گرفتن آموزش چند-شرطی در آموزش UBM و ماتریس T سبب افزایش شدید محاسبات و بهبود ناچیز دقت می‌شود. اعمال آن به دادگان ثبت‌نام در فرایند مدل‌سازی نیز مطابق با شرایط کاربری واقعی نیست. اما در مورد آموزش LDA و PLDA به‌صورت چند-حالته نتایج بسیار خوبی گزارش شده است. ازجمله اینکه افزایش بار محاسباتی کمی داشته و در مقابل به کاهش چشمگیری در خطا منجر می‌شود [۱۴].

۲-۵- ثبت‌نام چندگانه بردارهای هویت[۲۵]

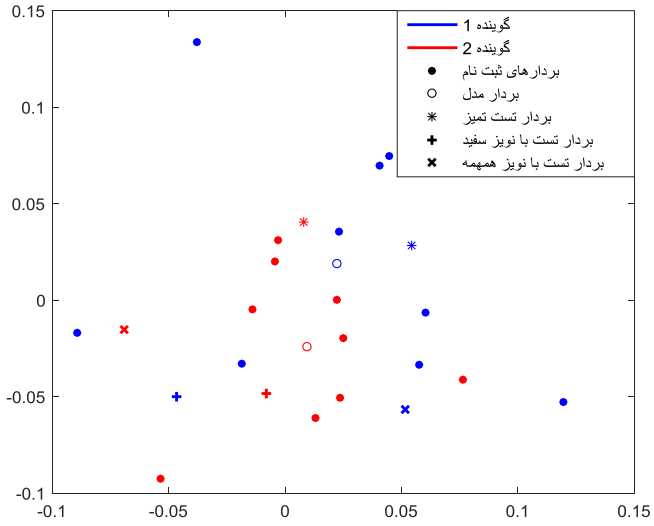
در صورت وجود چند نمونه گفتار از جلسات مختلف برای گوینده هدف، تکنیک‌های مختلفی برای امتیازدهی به او وجود خواهد داشت. با فرض استقلال بردارهای ثبت‌نام یک گوینده می‌توان معادلات امتیازدهی در PLDA را برای چند بردار ثبت‌نام بازنویسی کرده و محاسبه کرد. اما از آنجائی که نمونه‌های مختلف گفتار یک شخص اشتراکات فراوانی با یکدیگر دارند، فرض استقلال بردارهای حاصل از آن‌ها با واقعیت سازگار نیست. بنابراین روش‌های محاسبه امتیاز کاربردی و موفق در این حالت عبارتند از:

- محاسبه امتیاز آزمون برای هر بردار ثبت‌نام به‌صورت جداگانه و محاسبه میانگین آن‌ها به‌عنوان امتیاز نهایی؛
- الحاق نمونه‌های گفتار ثبت‌نام به یکدیگر و ساخت تنها یک بردار برای گوینده هدف؛
- میانگین‌گیری از آمارگان بام-ولش چند نمونه گفتار و ساخت یک بردار؛
- میانگین‌گیری از بردارهای حاصل از چند نمونه گفتار و استفاده از آن به‌عنوان بردار ثبت‌نام.

نتایج تحقیقات گزارش‌شده در مقالات حاکی از عملکرد موفق‌تر روش میانگین‌گیری از بردارهای ثبت‌نام گوینده هدف در مقایسه با دیگر روش‌هاست [۲۰].

۳- روش بردارهای هویت موزون

در این مقاله روشی جدید، که عنوان بردارهای موزون را برای آن برگزیده‌ایم، برای بهبود عملکرد سامانه‌های تصدیق هویت گوینده مبتنی بر i-vector پیشنهاد می‌شود. ایده اصلی این مقاله که به‌صورت مقدماتی برای بهبود عملکرد سامانه‌های مبتنی بر بردارهای ویژگی MFCC و در شرایط گفتار تمیز در [۲۴، ۲۵] معرفی شده بود در این مقاله توسعه داده شده و برای بردارهای ویژگی PNCC و گفتار آلوده به نویز نیز تعمیم یافته است. همان‌گونه که پیش‌ازاین ذکر شد در استخراج بردار هویت از گفتار هر گوینده ویژگی‌های آماری سیگنال گفتار وی اعم از اطلاعات گوینده، نویز و کانال دخیل هستند. زمانی که این بردار از چند قطعه گفتار ثبت‌نام آن گوینده، که هر یک در جلسات متفاوتی ضبط شده‌اند، استخراج می‌شود آنگاه میانگین آمیزه‌های گاوسی بردارهای هویت مرتبط با این جلسات به‌صورت مستقیم تأثیرگذار هستند، اما پراکندگی درایه‌های بردارهای ثبت‌نام گوینده هدف که قاعدتاً بیانگر



شکل ۲- توزیع دو درایه اول بردارهای ثبت‌نام دو گوینده متفاوت به همراه بردار مدل و تست متناظرشان.

برخی از مشخصات گفتاری گوینده است، در نظر گرفته نمی‌شود. نظریه اصلی روش پیشنهادی بر این اصل استوار است که با بهره‌گیری مناسب از اطلاعات پراکندگی درایه‌های بردارهای ثبت‌نام گویندگان هدف مختلف می‌توان عملکرد سامانه‌های تشخیص گوینده را برای گفتارهای تمیز و نویزی بهبود بخشید.

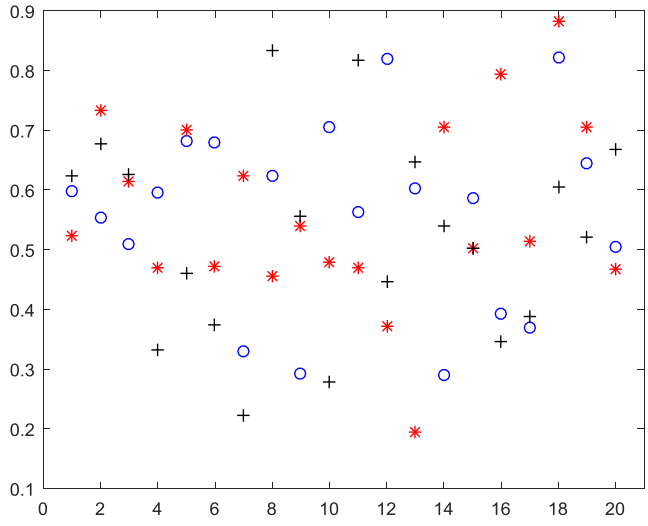
روش پیشنهادی برای استفاده از این اطلاعات پراکندگی، وزن‌دار کردن بردارهای هویت مدل و آزمون، بر مبنای پراکندگی درایه‌های بردارهای ثبت‌نام گویندگان هدف است. در این روش با توجه به تفاوت پراکندگی درایه‌های بردارهای ثبت‌نام، درایه‌های بردار مدل و آزمون وزن‌دهی می‌شوند.

همان‌طور که در بخش ۲-۲ اشاره شد، شیوه‌های رایج محاسبه امتیاز در حوزه i-vector تفاوتی بین تأثیرگذاری درایه‌های متفاوت بردارهای ثبت‌نام قائل نمی‌شوند. مبنای روش پیشنهادی بردارهای موزون بر آن است که با عنایت به تفاوت پراکندگی درایه‌های بردارهای ثبت‌نام می‌توان به‌صورت مناسبی تفاوت از این تفاوت پراکندگی‌ها استفاده کرد. مقدار پراکندگی درایه‌های بردارهای ثبت‌نام برخی گویندگان، به‌صورت ذاتی بیشتر از دیگران است و این امر نه‌تنها نباید در هنگام آزمایش تأثیر نامناسب در نتایج بگذارد بلکه از آن می‌توان در جهت افزایش دقت آزمون‌ها سود برد. بر همین اساس محاسبه بردار وزنی بر پایه تفاوت‌های موجود در پراکندگی درایه‌ها پیشنهاد می‌گردد. جهت درک بهتر این تفاوت‌ها، در شکل ۲ توزیع متفاوت دو درایه اول بردارهای ثبت‌نام، مدل و آزمون دو گوینده مختلف نشان داده شده است که مبین پراکندگی متفاوت در درایه‌های متناظر برای دو گوینده است.

اساس این روش به این صورت است که بردار مدل یک گوینده مشخص و بردار تستی که ادعا شده متعلق به آن گوینده است توسط بردار وزنی که از بردارهای ثبت‌نام همان گوینده به‌دست‌آمده، وزن‌دهی شوند. در اعمال وزن به بردارها باید دقت شود که بر روی بردار مدل و تست یک نوع وزن اعمال شود و این وزن از بردارهای ثبت‌نام مدل به‌دست‌آمده باشد. فرضیه اصلی آن است که این وزن‌دهی، امتیاز آزمونی را که بردارهای تست و مدل آن متعلق به یک گوینده باشند افزایش خواهد داد. برای دستیابی به این هدف دو شیوه برای محاسبه بردار وزن پیشنهاد می‌شود. در هر یک از این روش‌ها، بردار وزن به نوعی پراکندگی درایه‌های متناظر در بردارهای ثبت‌نام یک گوینده را نمایندگی می‌کنند.

روش اول:

در این روش ، عکس انحراف استاندارد درایه‌های متناظر از بردارهای ثبت‌نام هر گوینده، یعنی فاصله اقلیدسی آن‌ها از میانگین مقادیر درایه‌ها، به‌عنوان وزن آن درایه و برای همان گوینده در نظر گرفته می‌شود. محاسبه بردار وزن هر گوینده به‌صورت است:



شکل ۳- بردار وزن بر مبنای فاصله کسینوسی برای سه گوینده متفاوت؛ برای نمایش بهتر تمایز، تنها ۲۰ درایه اول رسم شده است.

$$m(l)=\frac{1}{K}\sum_{k=1}^{K}s_k(l) \qquad l=1,\dots,L \qquad \text{(۸- الف)}$$

$$std(l)=\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left[s_k(l)-m(l)\right]^2} \qquad \text{(۸- ب)}$$

$$ws(l)=1/std(l) \qquad \text{(۸- پ)}$$

که در آن‌ها همه بردارها L بعدی بوده و s_kبردار ثبت‌نام k-ام گوینده موردنظر، m میانگین بردارهای ثبت‌نام (مدل) گوینده، std انحراف استاندارد بردارهای ثبت‌نام گوینده و $ws(l)$ بردار وزن گوینده بر مبنای انحراف استاندارد می‌باشند.

روش دوم:

در این روش فاصله کسینوسی به‌عنوان مبنای محاسبه بردار وزن در نظر گرفته می‌شود. برای محاسبه آن ابتدا برای هر درایه دو بردار جدید تشکیل شده و سپس فاصله کسینوسی آن‌ها معیار وزن‌دهی قرار می‌گیرد. یک بردار متشکل از درایه‌های متناظر از بردارهای ثبت‌نام گوینده‌ای مشخص است که ابعاد آن برابر تعداد بردارهای ثبت‌نام خواهد بود. بردار دیگر با همین ابعاد بوده و درایه‌های آن یکسان و برابر میانگین درایه‌های متناظر از بردارهای ثبت‌نام (بردار اول) خواهد بود.

محاسبه بردار وزن گوینده موردنظر بر اساس فاصله کسینوسی:

$$x_l(k)=s_k(l) \quad k=1,\dots,K;\ l=1,\dots,L \qquad \text{(۹- الف)}$$

$$y_l(k)=m(l) \qquad \text{(۹- ب)}$$

$$wc(l)=\frac{\langle x_l,y_l\rangle}{\|x_l\|\|y_l\|} \qquad \text{(۹- پ)}$$

که در آن‌ها S نماینده بردارهای ثبت‌نام گوینده موردنظر، s_k بردار ثبت‌نام k-ام از گوینده، x_l برداری K بعدی متشکل از درایه‌های l-ام K بردار ثبت‌نام گوینده، m میانگین بردارهای ثبت‌نام (مدل) گوینده، y_l برداری K بعدی با درایه‌های یکسان و برابر با درایه l-ام بردار مدل گوینده و wc بردار وزن گوینده بر مبنای فاصله کسینوسی هستند.

در شکل ۳ بیست درایه اول بردار وزن بر مبنای فاصله کسینوسی برای سه گوینده مختلف نشان داده شده است. همان‌طور که مشاهده می‌شود به دلیل تفاوت

در توزیع درایه‌های متناظر در بردارهای ثبت‌نام گویندگان مختلف، بردار وزن آن‌ها با یکدیگر متفاوت است.

البته حالت‌های متفاوتی برای محاسبه بردار وزن نهایی و نحوه امتیازدهی در هر دو روش تعریف گردیده و مورد ارزیابی قرار گرفت و درنهایت مناسب‌ترین حالت با در نظر گرفتن عملکرد مناسب سامانه برای کار در شرایط نویزی، به‌صورت روابط (۸- پ) و (۹- پ) استنتاج شد. برای محاسبه امتیاز حاصل از مقایسه بردار مدل یک گوینده (m_s) و بردار تست (t) به روش PLDA، ابتدا بردار تست مذکور با ضرب شدن در بردار وزن آن گوینده وزن‌دار شده و سپس در فرایند امتیازدهی مورد استفاده قرار می‌گیرد.

۴- ارزیابی کارایی سامانه تصدیق هویت گوینده

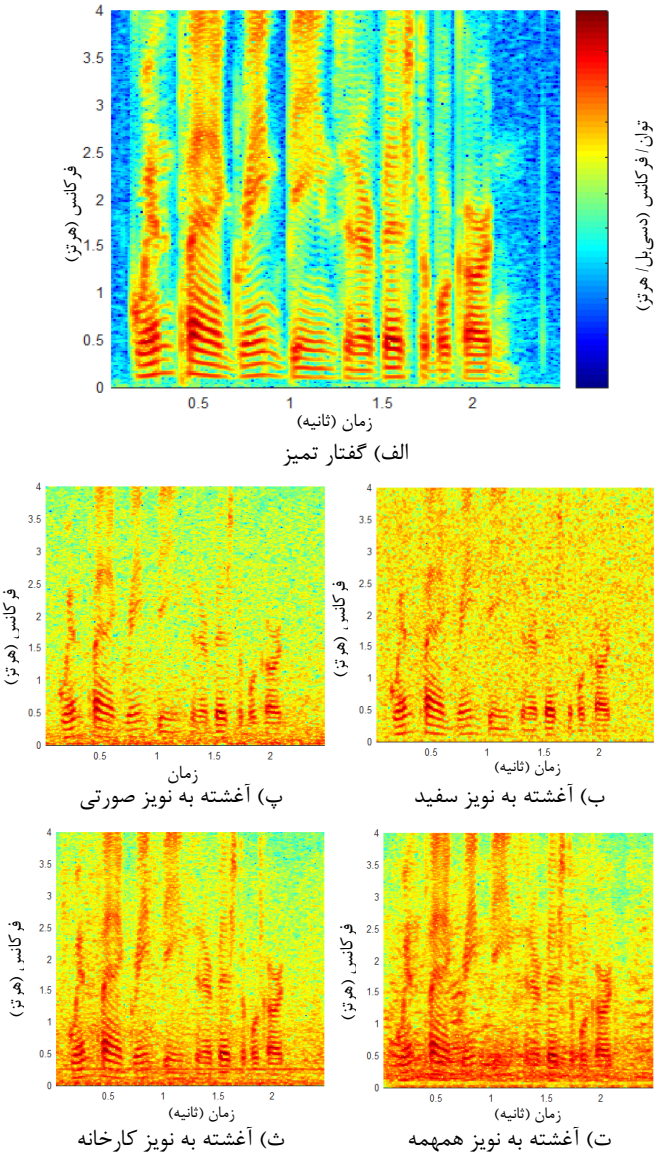
در این مقاله عملکرد بردارهای موزون پیشنهادی برای گفتار تمیز و آلوده به نویز در شرایط عدم تطبیق نویزی بین مرحله ثبت‌نام و آزمون و در کنار آموزش چند-شرطی LDA و PLDA مورد بررسی قرار گرفته است. برای بردارهای ویژگی نیز عملکرد برای بردارهای ویژگی MFCC و PNCC بررسی شده و مورد مقایسه قرار گرفته‌اند. شایان ذکر است که بردارهای ویژگی مبتنی بر PNCC به دلیل استفاده از تدابیر کاهش نویز در استخراج آن‌ها معمولاً عملکرد بهتری نسبت به بردارهای ویژگی بر پایه MFCC در شرایط نویزی دارند. ادامه این بخش به بیان چارچوب مورد استفاده در آزمون‌ها و ارزیابی عملکرد سامانه می‌پردازد.

۴-۱- دادگان سیگنال گفتار

در این مقاله از دادگان سیگنال گفتار TIMIT بهره گرفته شده است. این دادگان مشتمل بر نمونه‌های گفتار ۶۳۰ گوینده با هشت لهجه مختلف انگلیسی است که متشکل از ۴۳۸ مرد و ۱۹۲ زن می‌باشد که با فرکانس ۱۶ کیلوهرتز و دقت ۱۶ بیت نمونه‌برداری شده‌اند. در این دادگان برای هر گوینده ۱۰ جمله کوتاه وجود دارد که در شرایط تمیز بیان شده‌اند. هر جمله در حدود ۳ ثانیه و از نظر آوایی متنوع است [۲۶]. در آزمایش‌های ابتدا فرکانس صداها به ۸ کیلوهرتز کاهش داده شد. در آزمون‌ها تنها از نمونه‌های گفتاری گویندگان مرد استفاده شده است. ۳۶۸ گوینده مرد و از هرکدام ده جمله به‌عنوان دادگان توسعه برای ایجاد مدل پس‌زمینه UBM، ماتریس T، محاسبات مورد نیاز برای LDA و آموزش PLDA در نظر گرفته شد. ۷۰ گوینده مرد دیگر و از هرکدام ۹ جمله برای ثبت‌نام و یک جمله حدود ۳ ثانیه‌ای برای آزمون استفاده شده است. تعداد کل آزمون‌های تصدیق هویت گوینده برابر ۴۹۰۰ می‌باشد.

برای ارزیابی عملکرد سامانه در شرایط ورودی گفتار نویزی داده‌های نویزی NOISEX-92 با نسبت‌های سیگنال به نویز قطعه‌ای صفر، ۵ و ۱۰ دسی‌بل به دادگان تمیز TIMIT افزوده شده‌اند [۲۷]. عملکرد سامانه برای گفتارهای همراه با چهار نوع نویز، در شرایط عدم تطبیق بین مرحله ثبت‌نام و آزمون، مورد بررسی قرار گرفته است. نویزهای سفید، همهمه و صورتی در آموزش چند-شرطی PLDA و LDA استفاده شده و علاوه بر این سه نوع نویز، نویز کارخانه به‌عنوان نویز دیده نشده فقط در مرحله آزمون به کار گرفته شده است.

چهار نوع نویز انتخاب‌شده از معمول‌ترین نویزهای حاضر در محیط‌های دنیای واقعی هستند. نویزهای همهمه، صورتی و کارخانه فرکانس پایین بوده ولی نویز سفید همه فرکانس‌ها را پوشش می‌دهد. نویز همهمه به این دلیل که به‌صورت ذاتی از جنس گفتار است ازجمله چالش‌برانگیزترین نویزها در ارزیابی تاب‌آوری سامانه‌های تصدیق هویت گوینده برای گفتارهای آلوده به نویز محسوب می‌شود. شکل ۴ طیف یک نمونه سیگنال گفتار تمیز را در کنار نمونه‌های آلوده شده آن به چهار نوع نویز مورد استفاده در آزمون‌ها در حال نسبت سیگنال به نویز ۵ دسی‌بل نمایش می‌دهد.



شکل ۴- مقایسه تأثیر نویزهای سفید، صورتی، همهمه و کارخانه بر طیف سیگنال گفتار؛ نسبت انرژی سیگنال به نویز (SNR) در همه حالت‌ها ۵ دسی‌بل است.

۴-۲- پیکربندی سامانه

برای قطعه‌بندی سیگنال گفتار از پنجره همینگ با عرض ۲۵ میلی‌ثانیه و همپوشانی ۱۵ میلی‌ثانیه استفاده شد. فرکانس نمونه‌برداری دادگان TIMIT از ۱۶ کیلوهرتز به ۸ کیلوهرتز کاهش داده شده و فیلتر بانک ۲۶ تایی مل برای استخراج ضرایب ویژگی MFCC به دادگان TIMIT اعمال شد. به‌منظور استخراج ضرایب ویژگی PNCC از یک بانک فیلتر گاماتون ۳۰ کاناله استفاده شد. ابعاد بردار ویژگی استفاده شده متشکل از ۱۹ ضریب اصلی به همراه لگاریتم انرژی و مشتق اول و دوم آن‌هاست که یک بردار ویژگی ۶۰ بعدی را تشکیل می‌دهند. به‌منظور حذف اطلاعات زائد از دادگان گفتار از روشی مبتنی بر انرژی برای تشخیص صحبت از سکوت و حذف فریم‌های سکوت استفاده شد [۲۸].

جهت محاسبه بردارهای هویت در این تحقیق، ابتدا از ۲۵۶ آمیزه گاوسی برای ایجاد مدل پس‌زمینه جهانی استفاده شد. پیش‌آزمون‌های انجام‌شده با تعداد متفاوت آمیزه‌های گاوسی نشان داده بود که انتخاب ۲۵۶ گزینه بهینه برای این ارزیابی است. با استفاده از UBM ایجادشده و تحلیل عامل، ماتریس فضای تغییرپذیری کل توسط داده‌های توسعه آموزش داده شده و i-vector مربوط به هر قطعه گفتار با طول

جدول ۱- نتایج آزمایش‌ها با آموزش چند-شرطی LDA و PLDA برای بردارهای ویژگی MFCC و PNCC و دادگان TIMIT با و بدون بردارهای موزون بر مبنای نرخ خطای برابر (٪) و minDCF×100

بردار ویژگی		MFCC						PNCC					
امتیازدهی / شرایط		no weight		std weight		cds weight		no weight		std weight		cds weight	
		EER %	minDCF	EER %	minDCF	EER %	minDCF	EER %	minDCF	EER %	minDCF	EER %	minDCF
تمیز		۱/۰۶	۰/۷۱۶۱	۱/۱۲	۰/۵۵۲۲	**۰/۵۴**	۰/۴۷۰۲	۰/۵۶	۰/۵۵۲۲	۱/۰۸	۰/۳۸۸۸	**۰/۴۳**	۰/۳۸۸۸
نویز سفید	۱۰ dB	۹/۶۷	۴/۸۸۸۲	۸/۵۷	۴/۶۰۰۶	**۷/۵۸**	۴/۲۳۴۸	**۴/۹۹**	۳/۱۸۸۸	۵/۲۸	۳/۴۷۶۴	۵/۲۳	۲/۸۴۱۶
	۵ dB	۱۴/۲۹	۶/۸۸۷۶	**۱۲/۸۶**	۷/۰۱۱۸	۱۴/۲۹	۶/۷۴۰۷	۷/۳۵	۴/۹۸۹۴	۷/۱۴	۴/۹۹۱۳	**۵/۷۱**	۴/۵۰۲۵
	۰dB	۲۴/۷۸	۸/۶۳۶۰	۲۱/۴۳	۸/۴۳۵۴	**۲۰/۲۷**	۷/۹۴۳۷	۱۴/۲۹	۷/۲۵۴۷	**۱۴/۲۹**	۷/۱۰۸۱	۱۳/۷۳	۶/۳۵۷۱
نویز همهمه	۱۰ dB	۶/۱۷	۴/۱۶۸۹	۵/۷۱	۳/۹۶۲۱	**۴/۸۴**	۳/۸۰۳۱	۴/۱۶	۳/۰۶۷۱	۴/۴۷	۳/۱۶۹۶	**۴/۰۴**	۳/۰۶۵۸
	۵ dB	۱۱/۴۳	۵/۴۷۵۸	۱۰/۰۰	۵/۴۳۷۳	**۱۰/۰۰**	۴/۸۰۱۹	۷/۱۴	۵/۲۵۴۷	۷/۲۵	۵/۲۹۵۷	**۷/۱۲**	۵/۰۴۷۸
	۰dB	۱۸/۷۰	۷/۴۳۳۵	۱۷/۱۴	۷/۱۱۰۶	**۱۶/۷۱**	۷/۵۵۹۰	۲۲/۸۶	۸/۶۰۲۵	۲۱/۴۳	۸/۰۴۹۱	**۲۰/۰۰**	۸/۱۶۸۳
نویز صورتی	۱۰ dB	۸/۵۷	۴/۴۷۲۰	۷/۱۴	۳/۸۲۴۲	**۷/۱۴**	۳/۱۵۱۶	۵/۴۵	۳/۱۹۰۷	۴/۲۹	۳/۵۱۸۶	**۴/۲۹**	۳/۰۲۸۶
	۵ dB	۱۵/۷۱	۶/۸۶۳۴	۱۴/۳۱	۶/۲۷۳۹	**۱۲/۸۶**	۶/۱۰۷۵	۱۲/۷۱	۶/۱۶۷۱	۱۲/۸۶	۶/۱۶۵۸	**۱۱/۷۲**	۵/۸۸۵۷
	۰dB	۲۹/۰۷	۹/۲۶۸۳	۲۵/۷۱	۹/۰۲۸۶	**۲۵/۷۱**	۸/۷۴۰۴	۲۴/۲۹	۸/۷۲۱۷	۲۲/۸۶	۸/۴۱۱۸	**۲۱/۴۳**	۸/۵۱۸۶
نویز کارخانه	۱۰ dB	۹/۰۳	۴/۲۵۴۰	۸/۷۰	۴/۱۶۷۱	**۷/۳۳**	۳/۵۵۷۱	۴/۲۹	۳/۳۱۴۳	۴/۹۹	۳/۲۵۲۸	**۴/۲۹**	۳/۲۷۲۰
	۵ dB	۱۷/۱۴	۶/۹۸۶۳	۱۵/۷۱	۶/۷۸۳۲	**۱۵/۷۱**	۶/۴۵۷۸	۱۳/۹۵	۶/۳۳۲۳	**۱۰/۱۷**	۶/۳۱۹۹	۱۲/۸۶	۵/۶۱۹۳
	۰dB	۲۸/۳۲	۹/۴۰۹۳	۲۷/۸۷	۹/۴۰۹۳	**۲۵/۷۱**	۹/۰۸۲۳	۲۷/۱۴	۸/۷۷۶۴	۲۶/۷۹	۸/۶۷۳۹	**۲۵/۷۴**	۸/۳۰۷۵

۴۰۰ استخراج گردید. سپس با استفاده از تحلیل تفکیک‌کننده خطی (LDA) ابعاد بردارها به ۲۰۰ کاهش یافت. این اقدام علاوه بر کاهش بار محاسباتی سامانه، با افزایش تغییرات بین‌گروهی و کاهش تغییرات درون‌گروهی، میزان تفکیک‌پذیری بین گویندگان را نیز افزایش می‌دهد.

توزیع بردارهای به‌دست‌آمده از دادگان توسعه نماینده خوبی از توزیع بردارها است. بنابراین بردارهای ثبت‌نام و آزمون توسط پارامترهای آماری بردارهای توسعه سفیدسازی شدند. این بردارها توسط ماتریسی که از بردارها و مقادیر ویژه ماتریس کوواریانس داده‌های توسعه به دست آمد ناهمبسته شده و با تبدیل واریانس آن‌ها به یک هنجار سازی گردیدند. سپس مدل هر گوینده با میانگین‌گیری بین بردارهای ثبت‌نام آن گوینده ایجاد شد.

برای آموزش PLDA ابتدا دادگان توسعه به‌صورت تصادفی و به نسبت مساوی به نوع سه نویز سفید، همهمه و صورتی با سطوح سیگنال به نویز ۱۰، ۵ و صفر دسی‌بل آغشته شدند. این دادگان به همراه دادگان تمیز برای آموزش چند-شرطی PLDA به کار گرفته شدند. علاوه بر این ماتریس تبدیل LDA نیز در شرایط چند-شرطی محاسبه شد. امتیازدهی به آزمایش‌ها با و بدون وزن‌دهی به بردارهای آزمون انجام شده و پارامترهای ارزیابی سامانه محاسبه گردید.

۴-۳- معیار ارزیابی

ارزیابی سامانه تصدیق گوینده بر مبنای میزان خطا در آزمون‌ها و ترسیم منحنی مصالحه خطای آشکارسازی[۲۶](DET) انجام شد که عملکرد سامانه را با مقادیر آستانه متفاوت نشان می‌دهد. این منحنی مصالحه‌ای است بین خطای پذیرش اشتباه[۲۷](FA) و رد اشتباه[۲۸](FR). خطاهای رد اشتباه و پذیرش اشتباه را می‌توان با تنظیم حد آستانه تصمیم‌گیری سبک-سنگین کرد. از دید قابلیت جداسازی و بدون توجه به کاربرد، نقطه بهینه جایی است که دو نرخ خطای FA و FR باهم برابر گردند که آن نقطه نرخ خطای برابر[۲۹](EER) نامیده می‌شود. علاوه بر آن در مرحله ارزیابی سامانه، از تابع هزینه تصمیم[۳۰](DCF) نیز استفاده شد تا ارزیابی کامل‌تری از کارایی سامانه حاصل گردد. این تابع توسط سازمان ملی استاندارد و فناوری (NIST) معرفی شده و به‌صورت رابطه (۱۰) تعریف می‌گردد

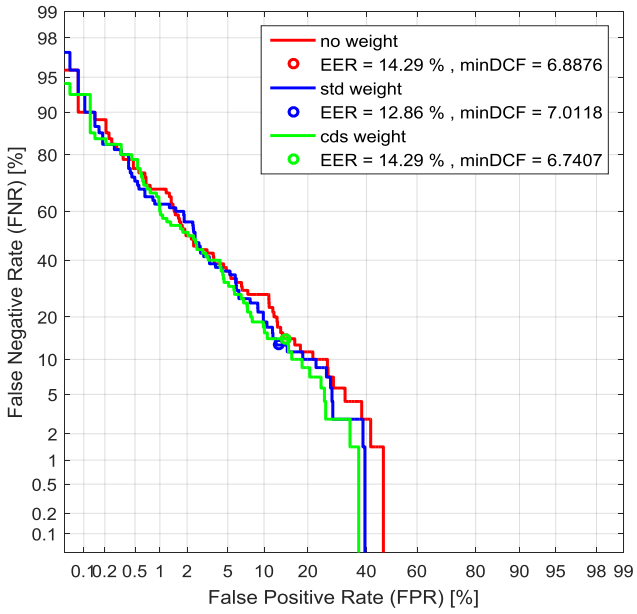
$$DCF = C_{fr}\ E_{fr}\ P_{target} + C_{fa}\ E_{fa}\ (1-P_{target}), \qquad (۱۰)$$

که در آن E_{fr} و E_{fa} به ترتیب خطای رد اشتباه و پذیرش اشتباه هستند. P_{target} احتمال پیشین گویندگان واقعی، C_{fr} هزینه از دست دادن اشتباه و C_{fa} هزینه پذیرش اشتباه است که مقادیر پیشنهادی NIST برای این سه پارامتر به ترتیب ۱۰، ۰/۰۱ و ۱ می‌باشد [۲۹]. 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

۴-۴- پیاده‌سازی آزمون

سامانه تصدیق هویت گوینده در حوزه i-vector و بر اساس جعبه‌ابزارهای VoiceBox [۳۰] و MSR Identity [۳۱] پیاده‌سازی شده است. در این سامانه عملکرد روش i-vector PLDA در آموزش چند-شرطی با و بدون روش وزن‌دهی بردارها و با استفاده از دو بردار ویژگی MFCC و PNCC با یکدیگر مقایسه شده‌اند. به دلیل آنکه آزمایش‌ها در حالت عدم استفاده از آموزش چند-شرطی LDA و PLDA نتیجه بسیار ضعیفی داشته و اغلب خطاها مقادیر بالایی بودند از ارائه نتایج آن‌ها در این مقاله اجتناب شده است و نتایج موجود در جداول همه حاصل آزمایش‌های مبتنی بر آموزش چند-شرطی می‌باشد. آزمایش‌ها در شرایط نویزی و با دادگان TIMIT و NOISEX-92 انجام شده است. دادگان TIMIT در شرایط بسیار تمیز و آزمایشگاهی جمع‌آوری شده و از این نظر شرایط ایده‌آلی را برای تشخیص گوینده فراهم کرده است؛ اما همین شرایط سبب می‌شود نتایج آزمایش‌های حاصل از آغشته کردن این دادگان به نویز تنها ناشی از نویز اضافه شده به آن‌ها باشد و نه هیچ اعوجاج‌های دیگر.

نتایج آزمایش‌ها در جدول ۱ ارائه شده است که در آن عملکرد امتیازدهی برای بردارهای هویت موزون و غیر موزون با استفاده از بردارهای ویژگی MFCC و PNCC و دادگان TIMITبه ترتیب به‌وسیله دو معیار نرخ خطای برابر و minDCF با یکدیگر مقایسه گردیده‌اند. برای هر حالت (سطر) و ویژگی بهترین نتیجه (روش امتیازدهی) از نظر نرخ کمینه خطا مشخص شده است.



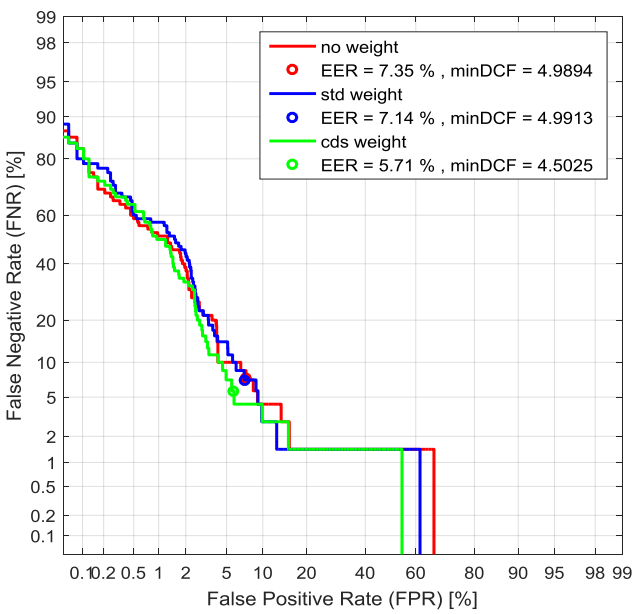
شکل ۵- مقایسه عملکرد سیستم مبتنی بر بردار ویژگی MFCC و LDA و PLDA چند-شرطی با و بدون بردارهای موزون برای دادگان TIMIT در حضور نویز سفید ۵ دسی‌بل – مقادیر minDCF با مقیاس ۱۰۰ نمایش داده شده‌اند.

مقادیر جدول ۱ نشان می‌دهد که استفاده از بردارهای موزون پیشنهادی ثبت‌نام و آزمون، عملکرد سامانه تصدیق هویت گوینده را در شرایط نویزی بهبود بخشیده است. علاوه بر آن محاسبه بردارهای وزن با روش شباهت کسینوسی عملکرد بهتری نسبت به انحراف استاندارد داشته است. همچنین عملکرد بردارهای ویژگی PNCC در مقایسه با بردارهای ویژگی MFCC بهتر بوده است.

در شکل‌های ۵ و ۶ عملکرد سامانه تصدیق هویت گوینده بر مبنای بردارهای موزون در حضور نویز سفید با سیگنال به نویز ۵ دسی‌بل به ترتیب با استفاده از بردار ویژگی MFCC و PNCC و به‌صورت منحنی‌های DET نمایش داده شده‌اند. بااینکه مقایسه منحنی‌ها نشان می‌دهد که استفاده از بردارهای موزون منجر به پذیرش اشتباه کمتری شده است اما با توجه به تعداد بسیار کمتر آزمایش‌هایی که بردار مدل و تست متعلق به یک گوینده هستند، مقایسه دقیق خطای پذیرش و رد اشتباه و اظهارنظر در این خصوص تا حدودی دشوار است.

۴-۵- ترکیب نتایج

روش‌های مختلف به‌تنهایی بخشی از اطلاعات را مدل کرده و تصمیم‌گیری می‌کنند و ترکیب آن‌ها می‌تواند اطلاعات کامل‌تری را در اختیار بگذارد. ترکیب روش‌ها در مراحل مختلفی چون ترکیب بردارهای ویژگی، شیوه‌های امتیازدهی و تصمیم‌گیری قابل پیاده‌سازی است. ترکیب در مرحله ویژگی با به‌هم‌پیوستن دو یا چند بردار و تشکیل بردار جدید صورت می‌گیرد. ترکیب امتیازهای حاصل از دو یا چند سامانه نیز با روش‌های گوناگونی چون میانگین (جمع)، بیشینه و ماشین بردار پشتیبان (SVM) قابل انجام است. البته در برخی موارد قبل از ترکیب امتیازها نیاز به هنجار سازی آن‌ها می‌باشد. ترکیب سامانه‌ها در مرحله تصمیم‌گیری به این صورت است که ابتدا هر سامانه تصمیم خود را مبنی بر رد یا قبول آزمون اعلام می‌کند و تصمیم نهایی بر پایه رأی‌گیری و مطابق با گزینه‌ای است که بیشترین تعداد آراء را داشته باشد. استفاده از این راهکار برای تصدیق هویت مرسوم‌تر از شناسایی گوینده است، زیرا برای تصدیق گوینده این شیوه حداقل با سه سامانه متفاوت قابل پیاده‌سازی است. در همه حالت‌های مختلف ترکیب، اعمال وزن‌های متفاوت به نتایج حاصل از روش‌های مختلف می‌تواند نتیجه بهتری به دست دهد، البته برای دستیابی به وزن‌های مطلوب دسترسی به داده‌های آموزشی بیشتری مورد نیاز



شکل ۶- مقایسه عملکرد سیستم مبتنی بر بردار ویژگی PNCC و LDA و PLDA چند-شرطی با و بدون بردارهای موزون برای دادگان TIMIT در حضور نویز سفید ۵ دسی‌بل – مقادیر minDCF با مقیاس ۱۰۰ نمایش داده شده‌اند.

می‌باشد. در این میان برای ترکیب در مرحله امتیازدهی نتایج بهتری گزارش شده است [۳۲، ۳۳].

در این مقاله ترکیب امتیازهای دو سامانه حاصل از بردارهای ویژگی MFCC و PNCC با استفاده از روش میانگین و SVM، که عملکرد موفق آن‌ها در نتایج تحقیقات محققان متفاوت گزارش شده است، مورد آزمایش و بررسی قرار گرفته‌اند [۳۳، ۳۴]. نتایج این ارزیابی در جدول ۲ نشان داده شده است. بررسی این نتایج نشان می‌دهد ترکیب امتیازها در اغلب شرایط نویزی عملکرد خطای سامانه را کاهش می‌دهد. همچنین میزان بهبود عملکرد سامانه در شرایط نویزی با سیگنال به نویز پایین بیشتر بوده و ترکیب با استفاده از SVM عملکرد بهتری نسبت به بهره‌گیری از میانگین در ترکیب امتیازها داشته است.

جدول ۲- نتایج ترکیب امتیازات حاصل از دو سامانه مبتنی بر دو بردار متفاوت ویژگی MFCC و PNCC به دو روش میانگین و SVM بر مبنای نرخ خطای برابر (EER ٪)

شرایط / بردار ویژگی		MFCC	PNCC	ترکیب بر پایه میانگین	ترکیب بر پایه SVM
تمیز		۰/۵۴	۰/۴۳	**۰/۳۵**	**۰/۳۵**
نویز سفید	۱۰ dB	۷/۵۸	۵/۲۳	**۳/۲۴**	۳/۲۸
	۵ dB	۱۴/۲۹	۵/۷۱	۶/۶۷	**۵/۵۳**
	۰ dB	۲۰/۲۷	۱۳/۷۳	۱۰/۱۱	**۹/۹۸**
نویز همهمه	۱۰ dB	۴/۸۴	**۴/۰۴**	۴/۳۲	۴/۰۶
	۵ dB	۱۰/۰۰	۷/۱۲	۵/۸۸	**۵/۳۹**
	۰ dB	۱۶/۷۱	۲۰/۰۰	۱۳/۲۶	**۱۳/۱۶**
نویز صورتی	۱۰ dB	۷/۱۴	**۴/۲۹**	**۴/۲۹**	**۴/۲۹**
	۵ dB	۱۲/۸۶	۱۱/۷۲	۹/۹۴	**۹/۸۰**
	۰ dB	۲۵/۷۱	۲۱/۴۳	۱۷/۶۶	**۱۵/۷۹**
نویز کارخانه	۱۰ dB	۷/۳۳	۴/۲۹	۳/۷۳	**۳/۲۶**
	۵ dB	۱۵/۷۱	۱۲/۸۶	**۹/۷۷**	۱۰/۰۰
	۰ dB	۲۵/۷۱	۲۵/۷۴	**۱۸/۶۸**	**۱۸/۶۸**

۵- جمع‌بندی

در این مقاله تأثیر استفاده از روش پیشنهادی بردارهای هویت موزون بر عملکرد سامانه تصدیق هویت گوینده پیاده‌سازی شده در فضایi-vector با استفاده از بردارهای ویژگی MFCC و PNCC و در شرایط نویزی مورد ارزیابی قرار گرفت. برای بهبود عملکرد سامانه از آموزش چند-شرطی برای LDA و PLDA استفاده شد. نتایج حاصل از آزمایش‌ها نشان داد که استفاده از بردارهای موزون استخراج شده از بردارهای ثبت‌نام گویندگان هدف در کنار آموزش چند-شرطی، سبب کاهش خطای سامانه در گفتارهای نویزی می‌شود. همچنین ترکیب نتایج حاصل از دو بردار متفاوت ویژگی در مرحله امتیازدهی عملکرد سامانه را بهبود می‌بخشد.

ادامه این تحقیق با هدف بهره‌گیری از دادگان بزرگ شامل چندین نمونه گفتار ثبت‌نام که در شرایط واقعی جمع‌آوری شده باشند و همچنین استفاده از شبکه‌های عصبی عمیق برای بهبود کیفیت گفتار نویزی به‌عنوان پیش‌پردازش و آموزش یکپارچه آن با سامانه تأیید هویت گوینده پی گرفته خواهد شد.

۶- 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] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, pp. 19-41, 2000.

[4] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, pp. 1-10, Czech Republic, 2010.

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

[6] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, pp. 1671-1675, 2015.

[7] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 4814-4818, 2015.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 5329-5333, 2018.

[9] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *Proc. 2018 IEEE Spoken Language Technology Workshop* (SLT), pp. 1021-1028, 2018.

[10] E. Lleida and L. J. Rodriguez-Fuentes, "Speaker and language recognition and characterization: Introduction to the CSL special issue," ed: Elsevier, 2018.

[11] C. S. Greenberg, D. Bansé, G.R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A.F. Martin, A. McCree, M. Przybocki, and D.A. Reynolds, "The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge," in *Proceedings ofOdyssey: The Speaker and Language Recognition Workshop*, Joensuu, Finland, 2014.

[12] P.-M. Bousquet, D. Matrouf,and J.-F. Bonastre, "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition," in *Proc. Interspeech*, pp. 485-488, 2011.

[13] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Proc. Interspeech*, pp. 249-252, 2011.

[14] M. McLaren and D. Van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2011*, pp. 5456-5459, 2011.

[15] B. Vesnicer, J. Zganec-Gros, S. Dobrisek, and V. Struc, "Incorporating duration information into i-vector-based speaker-recognitionsystems," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, pp. 241-248, 2014.

[16] Z. Lei, Y. Wan, J. Luo, and Y. Yang, "Mahalanobis Metric Scoring Learned from Weighted Pairwise Constraints in I-Vector Speaker Recognition System," in *Proc. INTERSPEECH*, pp. 1815-1819, 2016.

[17] O. Novotný, O. Plchot, O. Glembek, and L. Burget, "Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition," *Computer Speech & Language,* 2019.

[18] W. B. Kheder, D. Matrouf, J.-F. Bonastre, M. Ajili, and P.-M .Bousquet, "Additive noise compensation in the I-vector space for speaker recognition," in *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4190-4194, 2015.

[19] N. Li and M. Mak, "SNR-invariant PLDA with multiple speaker subspaces," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5565-5569, 2016.

[20] P. Rajan, A. Afanasyev, V. Hautamäki, and T. Kinnunen, "From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification," *Digital Signal Processing,* vol. 31, pp. 93-101, 2014.

[21] A. Sholokhov, T. Kinnunen, and S. Cumani, "Discriminative multi-domain PLDA for speaker verification," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5030-5034, 2016.

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

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

[۲۴] م. محمدی، و ح. ر. صادق محمدی، "بهبود عملکرد سامانه‌های تصدیق هویت گوینده در فضای I-Vector با استفاده از بردارهای هویت موزون ،" *بیست و سومین کنفرانس ملی سالانه انجمن کامپیوتر ایران*، تهران، دانشگاه صنعتی شریف، ۱۳۹۶.

[25] M. Mohammadi and H. R. S. Mohammadi, "Weighted I-Vector Based Text-Independent Speaker Verification System," in *Proc. 27th Iranian Conference on Electrical Engineering (ICEE)*, pp. 1647-1653, 2019.

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

[27] 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, 1993.

[28] R. Saeidi, H. R. Sadegh Mohammadi, T. Ganchev, and R. D. Rodman, "Particle swarm optimization for sorted adapted gaussian mixture models," *IEEE Transactions on Audio, Speech, and Language processing,* vol. 17, pp. 344-353, 2009.

[29] "The NIST Year 2008 Speaker Recognition Evaluation Plan," *Available: https://www.nist.gov/itl/iad/mig/speaker-recognition,* 2008.

[30] M. Brookes, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html,* vol. 47, 1997.

[31] 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.

[32] F. Răstoceanu and M. Lazăr, "Score fusion methods for text-independent speaker verification applications," in *Proc. 2011 6th Conference on Speech Technology and Human-Computer Dialogue* (SpeD), pp. 1-6, 2011.

[33] S. Garcia-Salicetti, M. A. Mellakh, L. Allano, and B. Dorizzi, "Multimodal biometric score fusion: The Mean Rule vs. support vector classifiers*," in Proc. 2005 13th European Signal Processing Conference*, pp. 1-4, 2005.

[۳۴] م. محمدی، و ح. ر. صادق محمدی، "بهبود عملکرد سامانه مستقل از متن تصدیق هویت گوینده برای گفتار آلوده به نویز با ترکیب دو روش GMM-UBM و I-Vector PLDA،" *چهارمین کنفرانس پردازش سیگنال و سامانه‌های هوشمند*، تهران، دانشگاه صنعتی امیرکبیر، ۱۳۹۷.

محسن محمدی مدرک کارشناسی خود را در رشته مهندسی برق با گرایش الکترونیک در سال ۱۳۸۸ از دانشگاه صنعتی امیرکبیر و کارشناسی ارشد رشته مهندسی برق با گرایش الکترونیک را در سال ۱۳۹۰ از دانشگاه صنعتی سهند تبریز اخذ نموده است. وی در حال حاضر در دوره دکتری مهندسی برق با گرایش مخابرات سیستم در پژوهشکده برق جهاد دانشگاهی و در زمینه تصدیق مستقل از متن هویت گوینده به تحصیل و پژوهش اشتغال دارد. زمینه‌های تحقیقاتی موردعلاقه ایشان پردازش صحبت و شناسایی گوینده می‌باشد. آدرس پست الکترونیکی ایشان عبارت است از:

mohammadi.mohsen@gmail.com

حمیدرضا صادق محمدی تحصیلات خود را در مقطع کارشناسی مهندسی برق با گرایش مخابرات و کارشناسی ارشد مهندسی الکترونیک به ترتیب در سال‌های ۱۳۶۲ و ۱۳۶۶ در دانشکده مهندسی برق دانشگاه علم و صنعت و دوره دکتری در رشته مهندسی برق با گرایش مخابرات را در سال ۱۳۷۵ در دانشگاه نیوساوت‌ولز استرالیا به اتمام رسانده است. وی در سال ۱۳۶۰ به جهاد دانشگاهی پیوست و مسئولیت‌های مختلفی در این مجموعه بر عهده داشته است و در حال حاضر به‌عنوان عضو هیئت‌علمی با درجه دانشیاری به فعالیت اشتغال دارد. همچنین ایشان از بدو تأسیس عهده‌دار مسئولیت سردبیری نشریه مهندسی برق و مهندسی کامپیوتر ایران از انتشارات پژوهشکده برق جهاد دانشگاهی بوده است. زمینه‌های تحقیقاتی موردعلاقه ایشان پردازش سیگنال، پردازش صحبت و تصویر، شناسایی گوینده و سیستم‌های بهینه‌سازی است.

آدرس پست الکترونیکی ایشان عبارت است از:

mohammadis@acecr.ac.ir

---

[1] Guassian Mixture Model (GMM)
[2] Probabilistic Linear Discriminant Analysis (PLDA)
[3] Identity Vector
[4] Deep Neural Networks (DNN)
[5] Mismatch
[6] Joint Factor Analysis
[7] Supervectors
[8] Linear Discriminant Analysis (LDA)
[9] Within-Class Covariance Normalization
[10] Mahalanobis
[11] Normalization
[12] Whitening
[13] Length Normalization
[14] Weiner
[15] Multi-Condition
[16] Cepstral
[17] Mel Frequency Cepstral Coefficients (MFCC)
[18] Power-Normalized Cepstral Coefficients (PNCC)
[19] Temporal Masking
[20] Universal Background Model (UBM)
[21] Baum-Welch
[22] Total Variability Space
[23] Expectation Maximization (EM)
[24] Multi-Condition Training
[25] Multiple Enrollment I-Vectors
[26] Detection Error Trade-off (DET)
[27] False Acceptance (FA)
[28] False Rejection (FR)
[29] Equal Error Rate (EER)
[30] Decision Cost Function (DCT)

The CSI Journal on Computing Science and Information Technology

The CSI Journal on
Computing Science and
Information Technology
Vol. 17, No.2, 2020

Enhancement of Speaker Verification Systems for Noise Contaminated Speech Using Weighted Identity-Vectors

Mohsen Mohammadi, Hamid Reza Sadegh Mohammadi

Department of Communications, Iranian Research Institute for Electrical Engineering, ACECR, Tehran, Iran.

Abstract

Secure access to different application systems from far and near distances, user-friendliness, low computational complexity, and low implementation costs are prominent features of the speech-based authentication method. However, the performance of this method in real environments is greatly degraded due to the presence of different noise and channel effects. The i-vector PLDA technique is one of the successful approaches to improve the performance of speaker authentication systems. In this paper, the use of statistical characteristics of target speakers' enrolment vectors for weighting model and test i-vectors have been proposed which improves the scoring accuracy of the verification system under both clean and noisy conditions. The performance of speaker verification system based on the proposed weighted i-vectors is evaluated. Experiments were implemented using TIMIT dataset, MFCC and PNCC feature vectors, and PLDA scoring method. Multi-condition training for LDA and PLDA has also been used to improve system performance under noise-mismatched conditions. In addition, the fusion of trial scores were also evaluated. The results confirm that the use of proposed weighted i-vectors improve the accuracy of the speaker verification system for clean and noise contaminated speech in both seen and unseen noisy conditions. Moreover, in the vast majority of cases the fusion of trial scores improves the system performance.

Keywords: speaker verification, weighting, noise, i-vector, multi-condition PLDA.