تعیین و تصدیق هویت گوینده بر روی خط تلفن به کمک یک سیستم هیبرید مقاوم در برابر نویز و اثر انتقال کانال همراه با نرمالیزاسیون امتیازات

محمدمهدی همایون‌پور[1*]، جهانشاه کبودیان[2]

۱- استادیار دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر

۲- دانشجوی دکترای مهندسی کامپیوتر، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر

*تهران، صندوق پستی ۴۴۱۳- ۱۵۸۷۵

kabudian, homayoun@ce.aut.ac.ir

(دریافت مقاله: مرداد ۱۳۸۱، پذیرش مقاله: مهر ۱۳۸۲)

چکیده – در این مقاله یک سیستم کارامد برای تعیین و تصدیق هویت گوینده معرفی می‌شود که در برابر نویز و اثر کانال انتقال مقاوم است. با استفاده از تکنیک آمیختن داده ها، دو سیستم HMM و GMM ، موازی شده و یک سیستم ترکیبی را برای کاربرد در تعیین و تصدیق هویت گوینده بر روی خط تلفن تشکیل داده‌اند. آزمایشها نشان می‌دهد که مدل ترکیبی HMM⊕GMM در بازشناسی گوینده از هر یک از سیستمهای HMM و GMM بهتر عمل می‌کند. برای مقابله با نویز جمع‌شونده، از روش تفاضل طیفی و نیز از معیارِ تصویر وزن‌دهی شده، و برای جبران‌سازی اثرِ کانالِ تلفن از روش تفاضل میانگین در حوزه کپسترال استفاده شده است که هر سه روش موجب بهبود سیستم بازشناسی گوینده شده‌اند. برای جمعیت ۱۰۰ نفری گویندگان (۶۱ مرد و ۳۹ زن) و بر روی پایگاه دادهٔ تلفنی FARSDIGITS1 با $SNR=8.8^{dB}$ به ازای داده‌های آزمایشی، نرخ صحت تعیین هویت گوینده برای جمعیت ۱۰۰ نفری برابر ۹۵/۵۱٪ و نرخ خطا در تصدیق هویت گوینده ۰/۳۷٪ به‌دست آمده است. چند روش نرمالیزاسیون امتیازات در سطح گویش و در سطح فریم و نیز روش وزن‌دهی امتیازات مدل برای افزایش کارایی سیستمهای تصدیق و تعیین هویت گوینده به کار رفته و نشان داده شده که این روشها به افزایش تمایز بین گویندگان و در نتیجه کاهش خطا در سیستمهای تصدیق و تعیین هویت گوینده منجر می‌شوند. بر روی پایگاه دادهٔ تلفنی FARSDIGITS1 استفاده از روشهای نرمالیزاسیون امتیازات، کارایی سیستم شناسایی گوینده مبتنی بر مدل مخلوط گاوسی در حالت نرمالیزه نشده را در تعیین و تصدیق هویت گوینده به ترتیب از ۹۵/۳۴٪ (صحت) و ۳/۲۵٪ (خطا)، به ۹۵/۵۱٪ (صحت) و ۰/۳۳٪ (خطا) در حالت نرمالیزه شده بهبود می‌بخشد که بویژه برای تصدیق هویت گوینده این بهبود بسیار چشمگیر است.

کلید واژگان: شناسایی گوینده، سیستم ترکیبی، مدل پنهان مارکُف، مدل مخلوط گاوسی، نرمالیزاسیون امتیازات.

۱- مقدمه

تصدیق و تعیین خودکار هویت از طرق مختلف، کاربردهای بسیاری از جمله در کنترل دسترسی افراد به اطلاعات، خدمات، منابع و مکانهای مهم دارد و در چند دهه اخیر به‌شدت مورد توجه قرار گرفته است. یکی از ساده ترین راههای تصدیق و تعیین هویت افراد استفاده از صدای شخص است، زیرا صدا مانند کارت شناسایی و

همواره با فرد همراه است. اهمیت تصدیق و تعیین هویت هویت افراد را از راه دور به‌عنوان مثال از طریق تلفن یا اینترنت - انجام دهیم. استفاده عمومی از تلفن موجب شده است که شناسایی اتوماتیک و هوشمند گویندگان از راه دور توسط تلفن از اهمیت و کاربرد وسیعی برخوردار باشد.

تعیین هویت گوینده بدین معنا است که گوینده، گفتاری را ادا می‌کند. سیستم باید تعیین کند که این گوینده کدام یک از گویندگان است، محمد ، علی، حسن ، ... . در تعیین هویت گوینده، گوینده هویت خود را بیان نمی‌کند. اما در تصدیق یا تأیید هویت گوینده، گوینده ابتدا هویت خود را اعلام می‌کند، مثلاً می‌گوید که علی است، سپس سیستم هویت او را تکذیب یا تأیید، و مشخص می‌کند که آیا او علی است یا خیر.

خط تلفن و مکالمات تلفنی خصوصیاتی دارند که بازشناسی بر روی خط تلفن را کاملاً از بازشناسی سیگنال میکروفنی (احیاناً فقط با یک میکروفن،) متفاوت می‌سازد. از جمله عوامل مزاحم بر روی خط تلفن می‌توان به نویزِ جمع‌شونده بر روی خط، پژواک، نویزِ برق شهر، نویزِ همشنوایی، نویزِ حاصل از ارتباطات مایکروویو، نویز آکوستیکی زمینه، محدود بودن پهنای باند خط تلفن و از بین رفتن فرکانسهای بالا - که در تمایز گویندگان نقش مهمی دارند - هموار نبودن مشخصهٔ فرکانسی خط تلفن، متفاوت بودن کانال تلفنی در تماسهای مختلف، استفادهٔ گویندگان از دهنی‌های مختلف در مکالمات مختلف که مشخصه‌های فرکانسی بسیار ناهموار و بعضاً بسیار متفاوت با یکدیگر دارند، اشاره کرد. با توجه به موارد ذکر شده می‌توان دریافت که اعوجاج و عوامل مزاحم بر روی مکالمات تلفنی بسیار زیاد بوده و این کار بازشناسی گوینده و گفتار را مشکل می‌سازد.

به طور کلی مدل پنهان مارکُف و مدل مخلوط گاوسی توانایی خود را در بازشناسی گفتار و گوینده بخوبی نشان داده‌اند. یکی از راههای ارتقای کارایی چنین سیستمهایی افراد توسط صدا هنگامی مشخص می‌شود که بخواهیم ساختن مدل ترکیبی موازی متشکل از دو سیستم با استفاده از روش آمیختن داده‌ها است که از آن در طراحی سیستم تعیین و تصدیق هویت گوینده مورد نظر استفاده شده است.

یکی از مسائل مهم در سیستمهای بازشناسی گوینده، جنبهٔ تصمیم‌گیری است. در سیستم های تصدیق و تعیین هویت گوینده اگر امتیازات خام[1] نرمالیزه شوند، تمایز گوینده از گویندگان دیگر افزایش یافته و کارایی سیستم افزایش می‌یابد. روشهای نرمالیزاسیون را می‌توان به روشهای نرمالیزاسیون امتیازات در سطح گویش، روشهای نرمالیزاسیون امتیازات در سطح فریم و روشهای نرمالیزاسیون امتیازات در سطح گویش و فریم تقسیم کرد. در بعضی سیستمها امتیازات در سطح فریم را برای بالابردن کارایی سیستم وزن‌دهی می‌کنند. بررسی و مقایسه روشهای نرمالیزاسیون امتیازات در بهبود کارایی سیستم مورد نظر در این مقاله مورد نظر است. یکی از جنبه های دیگر در سیستم های بازشناسی گوینده، تعداد گویندگان است. در [۱،۲] نشان داده شده که با افزایش جمعیت گویندگان، نرخ خطای تعیین هویت گوینده افزایش می‌یابد، اما نرخ خطای تصدیق هویت با افزایش جمعیت گویندگان تقریباً ثابت می‌ماند و چندان اضافه نمی‌شود. از طرفی نشان داده شده است که نرخ خطای تصدیق و تعیین هویت برای زنان بیش از مردان است [۱].

در این کار تحقیقاتی سیستم تصدیق و تعیین هویت افراد از طریق تلفن پیاده سازی شده که در مقابل نویزهای جمعی، اثر کانال انتقال و اثر دهنی های مختلف تلفن مقاوم بوده و در سطوح مختلف سعی می‌کند با این پدیده‌های مزاحم مقابله نماید. در این سیستم برای افزایش کارایی از مدل ترکیبی موازی متشکل از مدل پنهان مارکف و مدل مخلوط گوسی و برای تمایز بیشتر

1. Raw Score

گویندگان از یکدیگر از روشهای نرمالیزاسیون امتیازات استفاده شده است.

ساختار مقاله بدین صورت است که ابتدا در بخش ۲، تاریخچه‌ای از کارهای انجام شده در زمینه بازشناسی گوینده در ایران را ارائه خواهیم کرد. در بخش۳ به تشریح ساختار سیستم تعیین و تصدیق هویت می‌پردازیم. بخش ٤ خصوصیات خط تلفن و مکالمات تلفنی را ارائه می‌نماید. در بخش ٥ تکنیک تفاضل میانگین در حوزه کپسترال توضیح داده خواهد شد. در بخش ٦ جزئیات پیش پردازش و استخراج ویژگی به اختصار بررسی می‌شود. در بخش ۷ آموزش مدلهای گویندگان به روش مدل مخفی مارکف و مدل مخلوط گاوسی را توضیح می‌دهیم. در بخش ۸ نحوهٔ ساختن سیستم ترکیبی و در بخش ۹ روش معیارِ تصویرِ وزن‌دهی شده شرح داده خواهد شد. روشهای نرمالیزاسیون در سطح گویش و در سطح فریم و وزن‌دهی امتیازاتِ مدل به ترتیب در بخش های ۱۰، ۱۱ و ۱۲ ارائه خواهند شد. بخش ۱۳ به چگونگی انجام آزمایشها و نتایج به‌دست آمده می‌پردازد. بخش آخر به نتیجه گیری اختصاص دارد.

۲- تاریخچهٔ بازشناسی گوینده در ایران

در زمینه بازشناسی گوینده در زبان فارسی کارهایی انجام شده که مختصراً به آن خواهیم پرداخت. در سال ۱۳۷۳ آقایان ذهابی و سپهری [۳]، با استفاده از مدل پنهان مارکف، برای جمعیت ٥ نفری گویندگان، به ازای ٤٠ رقم برای آموزش و ۲۰ رقم برای بازشناسی، سیستمی‌را پیاده سازی کردند. در سال ۱۳۷۳، آقایان مندولکانی و لطفی زاد [٤]، با استفاده از تکنیک DTW برای جمعیتی ۱۰ نفری و به ازای ۱۰ جمله برای آموزش و ۱۰ جمله برای آزمایش، به کارایی ۹۸٪ برای تعیین هویت گوینده دست یافته‌اند. در همین سال آقایان اصغری و عارف [٥]، با استفاده از کوانتیزاسیون برداری و بر روی جمعیتی ۳۰ نفری از مردان، به ازای ۳۰ عبارت کوتاه برای آموزش و ۳۰ عبارت کوتاه برای آزمایش به کارایی ۹۹٪ برای تعیین هویت گوینده رسیده‌اند. باز هم در سال ۱۳۷۳ آقایان حدائق و لطفی زاد [٦]، با استفاده از تکنیک DTW و بر روی جمعیت ۱۰ نفری و به ازای ۱۰ تکرار جملهٔ خاص برای آموزش و ۱۰ تکرار همان جمله برای آزمایش، به کارایی ۱۰۰٪ برای تصدیق هویت گوینده دست یافته‌اند. در سال ۱۳۷٤، آقایان صیادیان و غفوری فرد [۷]، با استفاده از کوانتیزاسیون برداری و بر روی جمعیت ٥۰ نفری گویندگان، به ازای ۱۰ جمله برای آموزش و یک جمله برای آزمایش به کارایی متوسط ۹۸/۰۳ درصد برای تعیین هویت گوینده رسیده‌اند. باز هم در سال ۱۳۷٤، آقایان مقصودلو، نخعی و تبیانی [۸]، با استفاده از کوانتیزاسیون برداری و بر روی جمعیت ۱۰ نفری مردان، به ازای ۸ کد پنج رقمی برای آموزش و کدهای سه رقمی برای آزمایش برای تصدیق هویت گوینده به کارایی ۹۹/۸۳٪ رسیده‌اند. در سال ۱۳۷۷، آقایان فیض آبادی و صدوقی [۹]، با استفاده از کوانتیزاسیون برداری و بر روی جمعیت ۳۰ نفری گویندگان و ازای ۲۰ جمله و ۲۰ رقم برای آموزش و یک جمله برای آزمایش، به کارایی ۱۰۰٪ برای تصدیق هویت گوینده رسیده‌اند. در سال ۱۳۷۹ آقایان صیادیان، بدیع، حکاک و بیک زاده [۱۰]، با استفاده از مدل مخلوط گاوسی (GMM) در سطح واج و یک مدل به ازای هر واج برای هر گوینده، بر روی جمعیت ٦۰ نفری (٤۰ مرد و ۲۰ زن) و به ازای ۱۰۰۰ جمله در دورهٔ آموزش - که به‌صورت دستی واج نگاری می‌شود – و به ازای ۳ ثانیه گویش در دورهٔ آزمایش به کارایی ۱۰۰٪ برای تعیین هویت گوینده رسیده‌اند. تمام کارهای انجام شده که در بالا ذکر شد در محیط میکروفنی انجام شده و تنها کاری که در محیط تلفنی در زبان فارسی برای شناسایی گوینده انجام شده، کاری است که در سال ۷۸ توسط آقایان نجاری و همایون پور [۱۱]، بر روی جمعیت ٥۸ نفری (۳٦ مرد و ۲۲ زن) و با استفاده از دو روش شبکه های عصبی- الگوریتم های ژنتیک، و کوانتیزاسیون

برداری به ازای ۵۰ رقم برای آموزش و ۷ رقم برای آزمایش انجام شده است و در محیط تلفنی به کارایی ۹۷/۸٪ برای تصدیق هویت گوینده رسیده‌اند. همانطور که مشاهده می‌شود برای زبان فارسی در محیط تلفنی، کار زیادی انجام نشده است.

۳- ساختار کلی سیستم تعیین و تصدیق هویت گوینده

ساختار کلی سیستم پیاده سازی شده برای تعیین و تصدیق هویت گوینده به شرح زیر است. در این سیستم دو مرحله آموزش و بازشناسی وجود دارد. در مرحله آموزش سیگنال گفتار نخست - بدلیل محدودیت پهنای باند تلفن - از ۲۰۰ تا ۳۴۰۰ هرتز فیلتر می‌شود. این کار موجب می‌شود که نویزهای احتمالی موجود در خارج از این محدوده فیلتر شوند. سپس برای حذف نویز جمعی، از الگوریتم تفاضل طیفی [۱۲] استفاده می‌شود. آشکارسازی نواحی غیر گفتار (سکوت) برای الگوریتم تفاضل طیفی ضروری است که این کار توسط بخش تشخیص گفتار از سکوت [۱۳] انجام می‌شوند. پس از این مرحله، عملیات قاب بندی ، پیش‌تاکید و اعمال پنجره انجام و سپس ضرایب کپسترال MFCC به عنوان ویژگی استخراج و به روش تفاضل میانگین ضرایب کپسترال، اثر کانال انتقال و دهنی تلفن - که نوعی نویز کانولوشنال محسوب می‌شوند - از ویژگیهای به‌دست آمده حذف می‌شوند. بعد از این مرحله، مشتقات اول و دوم ضرایب کپسترال MFCC نیز به‌عنوان سایر ویژگیهای مورد استفاده در این سیستم به‌دست می‌آیند. ویژگیهای به‌دست آمده تا این مرحله از لحاظ نویزهای جمعی و کانولوشنال تا حد زیادی پاکسازی شده‌اند. این ویژگیها از داده های آموزشی هر یک از گویندگان استخراج شده و سپس برای آموزش مدلهای HMM و GMM آنها به‌کار می‌روند. در مرحله بعد اطلاعات وابسته به متن موجود در مدلهای HMM و اطلاعات مستقل از متن موجود در مدلهای GMM با یکدیگر آمیخته شده و سیستم ترکیبی موازی ساخته می‌شود. پس از این مرحله امتیازات حاصل از مدل ترکیبی هر گوینده نرمالیزه و برای مرحله تست سیستم، سطوح آستانه تصمیم گیری به‌طور بهینه تعیین می‌شوند.

در مرحله آزمایش ازای گویش تست، مراحل استخراج ویژگی به‌طور مشابه با مرحلهٔ آموزش انجام و پس از انجام تفاضل میانگین ضرایب کپسترال، احتمال رشته بردارهای ویژگی گویش تست روی مدل، به‌دست آمده و پس از نرمالیزه کردن، برای تصدیق هویت گوینده، امتیاز نرمالیزه شده، با سطح آستانه مقایسه و گوینده قبول یا رد می‌شود یا برای تعیین هویت گوینده، مدلی که بیشترین امتیاز نرمالیزه شده را دارد هویت گوینده را تعیین می‌نماید. برای حذف بیشتر اثر نویز جمعی بر روی ضرایب کپسترال، در مرحله تست سیستم از روش دیگری به‌نام روش WPM نیز استفاده می‌شود.

٤- خصوصیات خط تلفن و مکالمات تلفنی

پردازش سیگنالهای صوتی و گفتاری عبور داده شده از خط تلفن و نیز مکالمات تلفنی، بسیار متفاوت از پردازش سیگنالهای میکروفنی و بدون نویز است. پهنای باند خطوط تلفن محدود است و به‌عنوان مثال محدودهٔ ۲۰۰ Hz تا ۳٤۰۰ Hz و حتی محدودتر از این درنظر گرفته می‌شود که این بسیاری از اطلاعات مفید سیگنال گفتار را از بین می‌برد. این پدیده در سیستمهای بازشناسی گوینده - که اطلاعات فرکانسهای بالا از اهمیت خاصی برای تمایز گویندگان برخوردار است - بیشتر اثر خود را نشان می‌دهد. بر روی خط تلفن پژواک وجود دارد. مشخصهٔ کانال تلفنی در باند عبور، مشخصه‌ای هموار نیست و در فرکانسهای مختلف، تضعیف یا تقویت متفاوت است که این نیز کار بازشناسی را مشکل‌تر می‌سازد. نکتهٔ بسیار مهمی که دربارهٔ مکالمات تلفنی وجود دارد این است که گویندگان مختلف از دهنی‌های متفاوتی در دستگاه تلفن خود استفاده می‌نمایند که پاسخ

فرکانسی آنها ممکن است بسیار متفاوت و بسیار ناهموار باشد. در مرجع [۱٤] نشان داده شده که تضعیف یا تقویت در مشخصهٔ فرکانسی دهنی در باند تلفنی ممکن است تا 25 dB تغییر داشته باشد و به‌عنوان مثال دو دهنی یکی از نوع خازنی و دیگری از نوع کربنی با یکدیگر مقایسه شده‌اند. علاوه بر مسائل فوق، اگر گوینده فقط از یک دهنی استفاده کند، در زمانهای متفاوت هیچ تضمینی وجود ندارد که مشخصهٔ کانال ارتباطی در تماسهای مختلف یکسان باشد. بر روی خط تلفن نویز نیز وجود دارد که لزوماً نویز جمع شونده نیست. در مکالمات تلفنی وضعیت قرار گرفتن دهان گوینده نسبت به دهنی - در مقایسه با ضبط میکروفنی کنترل شده- تغییرات بیشتری دارد. در بعضی از دهنی ها، مانند نوع کربنی، اعوجاج هارمونیکی ایجاد شده و حتی پاسخ فرکانسی دهنی متغیر با زمان است. نویزهای دیگری نیز بر روی خط تلفن وجود دارد که از آن جمله می‌توان به نویز آکوستیکی زمینه، نویز برق شهر، نویز همشنوایی، نویز حاصل از ارتباطات مایکروویو و ... اشاره کرد. به علت وجود پدیده‌های فوق، برای سیستم های بازشناسی بر روی خط تلفن باید تمهیداتی را بیندیشیم. در این مقاله برای کاهش اثرِ نویزهای جمع شونده و نیز برای جبران سازی[1] مشخصهٔ کانالِ تلفنی راه حلی درنظر گرفته و نتایج آن ارائه شده است.

۵- تفاضل میانگین در حوزه کپسترال [۱۵]

یکی از روشهایی که برای جبران سازیِ اثرِ کانالِ انتقال پیشنهاد شده، روش تفاضل میانگین در حوزهٔ کپسترال یا CMS است. اگر فرض کنیم S(z) متناظر با سیگنال گفتار، G(z) متناظر با مشخصهٔ کانال و T(z) مربوط به سیگنالِ عبور داده شده از کانالِ انتقال باشد، آنگاه:

$$T(z)=S(z).G(z) \qquad (۱)$$

$$\log T(z)= \log S(z)+\log G(z) \qquad (۲)$$

با اعمالِ تبدیل فوریهٔ معکوس به طرفین رابطهٔ فوق داریم:

$$c_t(n)=c_s(n)+c_g(n) \qquad (۳)$$

و برای فریم m اُم از سیگنال گفتار داریم:

$$c_t(n,m)=c_s(n,m)+c_g(n,m) \qquad (٤)$$

$$c_t(n,m)=c_s(n,m)+c_g(n) \qquad (٥)$$

در این رابطه‌ها $c_t(n)$ ، $c_s(n)$ و $c_g(n)$، ضرایب کپسترال متناظر با سیگنال عبور داده شده از کانال، سیگنال گفتار و مشخصهٔ کانال هستند و نیز فرض شده که مشخصهٔ کانال انتقال یعنی $c_g(n)$ در طول زمان (با تغییر m) ثابت است. رابطهٔ بالا نشان می‌دهد که بردارهای کپسترال به‌صورت جمع شونده[2] تحت تأثیر بردارهای کپسترالِ مربوط به مشخصهٔ کانالِ انتقال یعنی $c_g(n)$ قرار می‌گیرند. از طرفی اگر سیگنالِ گفتارِ (عبارت) بیان شده به‌اندازهٔ کافی طولانی و از لحاظ فونتیکی متعادل باشد[3]، آنگاه می‌توان نوشت:

$$E\{c_s(n,m)\}=\sum_m c_s(n,m)\cong \underline{0} \qquad (٦)$$

علاوه بر این اگر عبارت بیان شده توسط گوینده عبارتی ثابت باشد (به عنوان مثال در سیستم‌های شناسایی گوینده وابسته به متن)، آنگاه:

$$E\{c_s(n,m)\}=\sum_m c_s(n,m)\cong \underline{C} \qquad (۷)$$

که $\underline{C}$ برداری ثابت است. یعنی با فرض بالا، میانگین بردارهای کپسترال در طول زمان در عبارت بیان شده برابر صفر یا مقداری ثابت خواهد بود. واضح است که در این صورت میانگین بردارهای کپسترال حاوی اطلاعات مفیدی نبوده و تفریق بردار میانگین از بردارهای اولیه، هیچگونه مشکلی ایجاد نخواهد کرد. اگر بردار جدید را چنین تعریف کنیم:

$$\tilde{c}_s(n,m)=c_s(n,m) - E\{c_s(n,m)\} \qquad (۸)$$

$$\tilde{c}_s(n,m)=c_s(n,m) - \underline{C} \qquad (۹)$$

1. Channel Compensation

2. Additive
3. Phonetically Balanced

آنگاه دو بردار c_s و $\tilde{c}_s$ از لحاظ اطلاعات مفید، یکسان بوده و معادل هستند. با توجه به رابطهٔ بالا داریم:

$$E\{c_t(n,m)\}=E\{c_s(n,m)\} + c_g(n) \qquad (۱۰)$$

$$E\{c_t(n,m)\}=\underline{C} + c_g(n)$$
$$\tilde{c}_t(n,m)=c_t(n,m) - E\{c_t(n,m)\}=$$
$$[c_s(n,m)+c_g(n)]-[\underline{C} + c_g(n)] =$$
$$c_s(n,m) - \underline{C}=\tilde{c}_s(n,m) \qquad (۱۱)$$

و به‌عبارت دیگر:

$$\tilde{c}_t(n,m)= \tilde{c}_s(n,m) \qquad (۱۲)$$

رابطهٔ فوق بدین معنی است که کم کردنِ میانگینِ بردارهای کپسترالِ مربوط به سیگنالِ عبور داده شده از کانال انتقال در طول زمان و در نواحیِ گفتار از خود بردارهای کپسترال، اثر مشخصهٔ کانال انتقال را از بین برده و معادلِ بردار کپسترال اولیه یعنی $\tilde{c}_s$ را به ما می‌دهد. با توجه به روابط فوق، روش CMS یکی از روشهایی است که برای جبران اثر کانال تلفن به‌کار می‌رود.

٦- پیش پردازش و استخراج ویژگی

برای مقابله با نویزِ جمعی موجود در مکالمات از روش مشهور تفاضلِ طیفی [۱۲] استفاده شده است. در این روش برای تخمین طیفِ نویز به الگوریتمی مقاوم برای تشخیصِ گفتار از سکوت نیاز داشتیم که از الگوریتم [۱۳] استفاده شد. ویژگیهای استفاده شده، ویژگیهای مبتنی بر بانک فیلتر هستند که به طریق زیر به‌دست می‌ایند: فریم‌بندی سیگنالِ صحبت به فریم‌های ^{ms} 35 که فاصلهٔ شروع هر دو فریمِ مجاور ^{ms} 10 است؛ اعمال پیش تأکید (α=0.975)؛ اعمال پنجرهٔ همینگ؛ به‌کار بردن ۱۸ فیلتر مثلثی که بر اساس معیار *Mel* بر روی طیف فوریهٔ سیگنال توزیع شده‌اند؛ استخراج ۱۲ ضریب کپسترال با اعمال لیفتر کاهنده در طرفین (لیفتر جوانگ)؛ به‌دست آوردن مشتقات اول و دوم ضرایب کپسترال؛ و در نهایت اعمال روش تفاضل میانگین در حوزهٔ کپسترال که در بخش قبلی توضیح داده شد.

۷- آموزش مدلهای گویندگان به روش مدل مخفی مارکف و مدل مخلوط گاوسی

یکی از موفق‌ترین روشهای مدلسازی دنباله‌های تصادفی و از جمله سلسله بردارهای ویژگی استخراج شده از سیگنال صحبت، مدل پنهان مارکف است. مدل پنهان مارکف پیوسته با توابع چگالی احتمال مشاهدات از نوع مخلوط گاوسی، پر استفاده ترین نوع مدل مارکف است که پارامترهایی مانند احتمالات حالات اولیه، احتمال انتقال بین حالات، وزن‌های هر یک از توابع گوسی، بردارهای میانگین و ماتریس های کوواریانس هر یک از این توابع را در بر دارد. مدل پنهان مارکف پیوسته‌ای با توابع چگالی احتمال مخلوط گاوسی - که فقط دارای یک حالت باشد - مدل مخلوط گاوسی (GMM) نامیده می‌شود. در واقع در اینجا از مدل پنهان مارکف برای مدلسازی اطلاعات وابسته به متن گوینده و از مدل مخلوط گاوسی برای مدلسازی اطلاعات مستقل از متن گوینده استفاده شده و با ساختن سیستم ترکیبی - که در بخش بعدی به تشریح آن خواهیم پرداخت - اطلاعات وابسته به متن و اطلاعات مستقل از متن گویندگان با یکدیگر آمیخته می‌شوند. آموزش مدل پنهان مارکُف در طی مراحل زیر انجام می‌شود.

- تقسیم یکسان بردارهای ویژگی بین حالات مدل و به دست آوردن تخمین اولیه پارامترهای مدل با استفاده از خوشه‌بندی.
- تقسیم بهینهٔ بردارهای ویژگی بین حالات مدل با استفاده از الگوریتم ویتربی و خوشه‌بندی بردارها توسط الگوریتم *k-means* و به دنبال آن تخمین پارامترهای مدل و تکرار این کار تا حصول همگرایی.
- تصحیح پارامترهای مدل با استفاده از فرمولهای تخمین بام- ولش و تکرار این کار تا رسیدن به همگرایی.

آموزش مدلهای مخلوط گاوسی نیز به‌صورت زیر انجام می‌شود:

- تخمین اولیه برای میانگین‌ها، واریانس‌ها و وزنهای توابع گاوسی با استفاده از خوشه‌بندی.
- تصحیح پارامترهای مدل با استفاده از فرمولهای تخمین EM و تکرار این کار تا رسیدن به همگرایی.

۸- نحوه ساختن سیستم ترکیبی

اگر فرض کنیم O رشتهٔ مشاهدات (بردارهای ویژگی) باشد، آنگاه احتمال تولید مشاهدات توسط مدل HMM را با احتمال تولید مشاهدات توسط مدل GMM به نحو زیر آمیخته می‌کنیم:

$$P_{hyb}(O|\lambda_i)=\alpha.P_n(O|\lambda_{i,GMM})+(1-\alpha).P_n(O|\lambda_{i,HMM}) \quad (۱۳)$$

که P_n ، احتمال نرمالیزه شده است که چنین به دست می‌آید:

$$P_n(O|\lambda_i)=P(O|\lambda_i)-\underset{j\neq i}{Max}\,P(O|\lambda_j) \quad (۱۴)$$

λ_i گویندهٔ ادعا شده و λ_j گوینده‌ای غیر از گویندهٔ ادعا شده است.

۹- معیار تصویر وزن‌دهی شده

یکی از روشهایی که برای مقابله با نویز جمع‌شونده باند پهن[1] پیشنهاد شده، استفاده از معیار تصویر وزن‌دهی شده یا *WPM* است [۱۶-۱۸]. آزمایشها نشان داده است که نویز سفید (یا نویزِ با باند پهن) به صورت جمع‌شونده، بر اندازه یا طول بردارهای کپسترال تأثیر می‌گذارد، اما جهتِ بردارها نسبت به نویزِ جمعی مقاوم‌تر است. برای محاسبهٔ فاصله بین دو بردار کپسترال نیز پیشنهاد شده است که به جای محاسبهٔ فاصلهٔ بین دو بردار، از معیار فاصلهٔ مقاوم‌تری - که زاویهٔ بین دو بردار را درنظر می‌گیرد - استفاده شود. تابع گاوسی برای محاسبهٔ احتمال در مدل پنهان مارکُف یا مدل مخلوط گاوسی چنین است:

$$N(o_t,\mu_i,\Sigma_i)=(2\pi)^{-\frac{n}{2}}.|\Sigma_i|^{\frac{1}{2}}.\exp(-\frac{1}{2}(o_t-\mu_i)^T\Sigma_i^{-1}(o_t-\mu_i)) \quad (۱۵)$$

در این تابع گاوسی، فاصلهٔ اقلیدسی وزن‌دار را به عنوان معیاری برای فاصله به شکل زیر می‌توان در نظر گرفت:

$$d_{WED(o_t,\mu_i)}=(o_t-\mu_i)^T\Sigma_i^{-1}(o_t-\mu_i) \quad (۱۶)$$

معیار فاصلهٔ اقلیدسی وزن‌دار و مبتنی بر تصویر یا همان WPM به صورت زیر تعریف می‌شود:

$$d_{WPM(o_t,\mu_i)}=(o_t-\lambda\mu_i)^T\Sigma_i^{-1}(o_t-\lambda\mu_i) \quad (۱۷)$$

مقدار بهینهٔ λ باید طوری تعیین شود که بدون تغییر جهت در بردارهای o_t و μ_i ، فاصلهٔ وزن‌دار بین بردار o_t و $\lambda\mu_i$ می‌نیمم شود. مقدار بهینهٔ λ از رابطهٔ زیر به دست می‌آید:

$$\lambda=\frac{o_t^T\Sigma_i^{-1}\mu_i}{\mu_i^T\Sigma_i^{-1}\mu_i} \quad (۱۸)$$

آزمایشها نشان داده است که با حضور نویز سفید جمع‌شونده باند پهن[2] و حتی نویزهای رنگیِ جمع‌شونده با باند پهن[3]، معیار WPM موجب ارتقای کارایی سیستم‌های بازشناسی می‌شود. لازم است ذکر شود که معیار WPM فقط در بازشناسی و تست سیستم اعمال می‌شود و در هنگام آموزش به هیچ تمهیدی نیاز نیست. یکی از روشهایی که می‌تواند برای مقابله با نویزهای جمع‌شوندهٔ موجود در خط تلفن به‌کار رود، همین معیار WPM است که در این مقاله برای سیستمهای بازشناسی گوینده به کار برده شده است.

۱۰- روشهای نرمالیزاسیون امتیازات در سطح گویش [۱۹]

معیار بیز برای طبقه بندی مسألهٔ دو کلاسه را می‌توان به‌صورت زیر به‌کار برد:

$$if \quad P(q_1|X)\geq P(q_2|X) \quad then \quad X\in q_1 \quad else \quad X\in q_2 \quad (۱۹)$$

اگر فرض کنیم X رشتهٔ مشاهدات یا رشتهٔ بردارهای ویژگی به‌صورت $X=\{x_1,x_2,\ldots,x_t,\ldots,x_T\}$ و q_1

1. Additive Broadband Noise

2. Additive Broadband White Noise
3. Additive Broadband Colored Noise

کلاس خودگوینده و q_2 کلاس گویندگان دیگر باشد آنگاه:

$$\frac{P(q_1).P(X|q_1)}{P(X)} \geq \frac{P(q_2).P(X|q_2)}{P(X)} \qquad (۲۰)$$

$$if \quad \frac{P(X|q_1)}{P(X|q_2)} \geq \left(\frac{P(q_2)}{P(q_1)} = Thr\right) \quad then \quad X \in q_1 \quad else \quad X \in q_2 \qquad (۲۱)$$

Thr سطح آستانهٔ تصمیم‌گیری است. اگر λ_i و λ_c مدل گویندهٔ مدعی[1] یا ادعا شده[2] و λ_j به طوری که $j \neq i$، مدل سایر گویندگان باشد و نیز $\lambda_{\bar{c}}$ مدل گویندگان غیر از گویندهٔ ادعا شده و S تعداد کل گویندگان باشد، آنگاه:

$$if \quad \frac{P(X|\lambda_c)}{P(X|\lambda_{\bar{c}})} \geq Thr \quad then \quad X \in c \quad else \quad X \in \bar{c} \qquad (۲۲)$$

این احتمال را احتمال نرمالیزه شده یا نسبت احتمالات می‌نامند. اگر مبنای تصمیم گیری به‌صورت زیر باشد، احتمال نرمالیزه نشده یا خام مورد نظر است:

$$if \quad P(X|\lambda_c) \geq Thr \quad then \quad X \in c \quad else \quad X \in \bar{c} \qquad (۲۳)$$

گاهی از لگاریتم نسبت احتمالات[3] استفاده می‌شود:

$$Log\left(\frac{P(X|\lambda_c)}{P(X|\lambda_{\bar{c}})}\right) = LogP(X|\lambda_c) - LogP(X|\lambda_{\bar{c}}) \qquad (۲۴)$$

مدل $\lambda_{\bar{c}}$ را مدل ضد گوینده[4] نیز می‌نامند. ($P(X|\lambda_{\bar{c}})$ را به‌صورت زیر می‌توان به‌دست آورد:

$$P(X|\lambda_{\bar{c}}) = \sum_{j \neq i} P(\lambda_j).P(X|\lambda_j) = \frac{1}{S-1}\sum_{j \neq i} P(X|\lambda_j) \qquad (۲۵)$$

فرمول بالا با این فرض نوشته شده که احتمال پسین مدلها یعنی $P(\lambda_j)$ ها مساوی و برابر با $\frac{1}{S}$ باشند، بنابراین می‌توان نوشت:

$$Log\left(\frac{P(X|\lambda_c)}{P(X|\lambda_{\bar{c}})}\right) = LogP(X|\lambda_i) - Log\left(\frac{1}{S-1}\sum_{j \neq i} P(X|\lambda_j)\right) \qquad (۲۶)$$

در این رابطه برای محاسبهٔ احتمال بر روی سایر گویندگان غیر از گوینده ادعا شده از میانگین استفاده شده است. به‌طور کلی و به‌صورت تقریبی برای محاسبهٔ احتمال نرمالیزه شده در سطح گویش از فرمول کلی زیر می‌توان استفاده کرد:

$$Log\left(\frac{P(X|\lambda_c)}{P(X|\lambda_{\bar{c}})}\right) = LogP(X|\lambda_i) - Log(\underset{j \neq i}{Stat}\{P(X|\lambda_j)\}) \qquad (۲۷)$$

منظور از Stat عملیات آماری مانند میانگین بر روی امتیازات حاصل از مدل‌های گوینده‌های دیگر است. برای محاسبهٔ احتمال نرمالیزه‌شده به‌جای Stat می‌توان از آماره‌های[5] زیر استفاده کرده:

۱۰-۱- احتمال پسین[6] [۱۹]

در محاسبه احتمال نرمالیزه شده به این روش، در واقع از رابطهٔ محاسبهٔ احتمال پسین برای گویندهٔ i استفاده می‌کنیم:

$$P_n(X|\lambda_i) = LogP(X|\lambda_i) - Log\left(\frac{1}{S}\sum_{j=1}^{S} P(X|\lambda_j)\right) \qquad (۲۸)$$

همانطور که مشاهده می‌شود، برای محاسبهٔ $P(X|\lambda_{\bar{c}})$ در طرف راست معادله، احتمال $P(X|\lambda_i)$ نیز در $\sum$ در نظر گرفته می‌شود. استفاده از این رابطهٔ برای محاسبهٔ احتمال نرمالیزه شده موجب می‌شود که اگر گویندهٔ مدعی، دروغگو باشد، به ازای یکی از λ_j ها در قسمت $\sum$ - که در حقیقت مدل خود گویندهٔ دروغگو است - $P(X|\lambda_j)$ زیاد شده و در کل $P_n(X|\lambda_i)$ کاهش یابد و احتمال قبول شدن او کمتر شود.

1. Claimant Speaker
2. Claimed Speaker
3. Log-Likelihood Ratio (LLR)
4. Anti-Speaker
5. Statistics
6. Posteriori Probability

۱۰-۲- میانگین [۲۰، ۲۱]

فرمول محاسبه احتمال نرمالیزه شده به این روش به شکل زیر است:

$$P_n(X|\lambda_i) = LogP(X|\lambda_i) - Log(\frac{1}{S-1}\sum_{j=i}^{S} P(X|\lambda_j)) \quad (۲۹)$$

تفاوت این رابطه با روش احتمال پسین این است که $P(X|\lambda_i)$ یعنی احتمال به ازای مدل گوینده ادعا شده در قسمت $\sum$ منظور نمی‌شود. محاسبهٔ احتمال نرمالیزه‌شده به این روش نیز به دلیل مشابه با احتمال پسین موجب کاهش خطا می‌شود.

۱۰-۳- ماکزیمم [۲۱]

احتمال نرمالیزه شده در این روش به روش زیر محاسبه می‌شود:

$$P_n(X|\lambda_i) = LogP(X|\lambda_i) - Log\,(\underset{j\neq i}{Max}\, P(X|\lambda_j)) \quad (۳۰)$$

استفاده از این آماره نیز هم احتمال قبول شدن فرد دروغگو را کم می‌کند، زیرا ماکزیمم امتیاز فرد مدعی دروغگو بر روی مدلهای غیر از مدل ادعا شده - که اتفاقاً مدل حقیقی فرد دروغگو نیز یکی از آنها است - زیاد بوده و امتیاز نرمالیزه شده او کم است. این نحوهٔ محاسبهٔ احتمال نرمالیزه شده، همچنین به گوینده مدعی راستگو - که ماکزیمم امتیاز او بر روی مدلهای دیگر معمولاً کم است - کمک می‌کند و احتمال نرمالیزه شده شخص مدعی راستگو، در سطح بالایی باقی می‌ماند.

۱۰-۴- می‌نیمم [۲۰]

احتمال نرمالیزه شده را در این روش چنین محاسبه می‌کنیم:

$$P_n(X|\lambda_i) = LogP(X|\lambda_i) - Log\,(\underset{j\neq i}{Min}\, P(X|\lambda_j)) \quad (۳۱)$$

این روش از این ایده بهره می‌برد که می نیمم امتیاز فرد مدعی دروغگو بر روی مدلهای غیر ادعا شده - که مدل خود او نیز در میان آنها است - زیاد است، اما می‌نیمم امتیاز فرد مدعی راستگو بر روی مدلهای غیر ادعا شده معمولاً کم است و بنابراین، این روش مدعی دروغگو را سرکوب کرده و به مدعی راستگو کمک می‌کند.

۱۰-۵- شبیه‌ترین M گوینده[1] [۲۲]

اگر امتیازات بردارهای ویژگی روی مدل‌های غیر ادعا شده را به ترتیب نزولی مرتب کنیم و M عدد از این امتیازات را که بیش از دیگران هستند برداریم، آنگاه احتمال نرمالیزه شده چنین است:

$$P_n(X|\lambda_i) = LogP(X|\lambda_i) - Log(\frac{1}{M}\sum_{j=1, j\neq i}^{M} P(X|\lambda_j)) \quad (۳۲)$$

که در آن:

$$P(X|\lambda_j) \geq P(X|\lambda_{j+1}) \geq \ldots \quad (۳۳)$$

۱۰-۶- نرمالیزاسیون گروهی[2] [۲۱]

در این روش که نرمالیزاسیون گروهی نام دارد، گروهی از گویندگان از میان گویندگان غیر از گوینده ادعا شده - که به گوینده ادعا شده بیشتر شبیه هستند - در دوره آموزش تعیین می‌شوند. تعداد گویندگان این گروه، C و احتمال نرمالیزه شده چنین محاسبه می‌شود:

$$P_n(X|\lambda_i) = LogP(X|\lambda_i) - \frac{1}{C}\sum_{j\in Cohort(i),\ j\neq i} P(X|\lambda_j) \quad (۳۴)$$

شباهت گویندهٔ i و گوینده j در دورهٔ آموزش به این طریق به‌دست می‌آید که امتیازات نمونه‌های آموزشی گویندهٔ i ام بر روی مدل گویندهٔ j ام و نیز امتیازات نمونه‌های آموزشی گویندهٔ j ام بر روی مدل i ام، محاسبه شده و میانگین این دو عدد، شباهت دو گویندهٔ i و j را نشان می‌دهد. بدیهی است که در روشهایی مانند

1. Top M Speakers
2. Cohort Normalization

کوانتیزاسیون برداری، مقدار فاصلهٔ به‌دست آمده یا اعوجاج به‌دست آمده، درجهٔ عدم شباهت دو گوینده را نشان می‌دهد. مزیت بزرگی که نرمالیزاسیون گروهی نسبت به پنج روش قبلی دارد، این است که امتیاز مشاهدات به ازای تمام مدلهای گویندگان غیر مدعی محاسبه نمی‌شود و فقط به ازای گروهی از آنها محاسبه می‌شود. قابل ذکر است که گویندگانی که می‌توانند در یک گروه، شبیه به گوینده i ام قرار گیرند، لزومی ندارد که از همان جنسیت باشند.

۱۰-۷- نرمالیزاسیون گروهی ترکیبی[1] [۲۰]

اگر می‌نیمم امتیازی (احتمال) را که گویش‌های گویندهٔ i ام در دورهٔ آموزش بر روی مدل خودِ گویندهٔ i ام یعنی λ_i کسب می‌کنند، $S^i_{\min}$ بنامیم، آنگاه احتمال نرمالیزه‌شده به این طریق به‌دست می‌آید:

$$\begin{aligned} &if \quad P(X \mid \lambda_i) > kS^i_{\min} \quad then \\ &P_n(X \mid \lambda_i) = LogP(X \mid \lambda_i) - Log\left(\frac{1}{C}\sum_{j \in Cohort(i),\ j \neq i} P(X \mid \lambda_j)\right) \\ &else \quad P_n(X \mid \lambda_i) = -\infty \end{aligned} \qquad (۳۵)$$

که در این فرمول k می‌تواند مقداری به‌عنوان مثال در حدود 0.9 داشته باشد. حجم محاسبات این روش نیز مانند روش نرمالیزاسیون گروهی است.

ذکر این نکته لازم است که در تمامی روشهای فوق به‌طور تقریبی می‌توان به جای $Log(\frac{\sum P_i}{S-1})$ از $\frac{\sum Log P_i}{S-1}$ استفاده کرد که به‌جای میانگین حسابی، میانگین هندسی را محاسبه می‌کند. در مرجع [۱۹] نشان داده شده که این دو تقریبا یکسانند و در برخی موارد، واسطهٔ هندسی جواب بهتری را داده است. همچنین می‌توان نشان داد که روشهای نرمالیزاسیون در سطح گویش، نرخ خطای تصدیق هویت گوینده را تحت تأثیر قرار می‌دهند اما بر نرخ خطای تعیین هویت گوینده تأثیری ندارند.

۱۱- روشهای نرمالیزاسیون امتیازات در سطح فریم [۲۲]

اگر $P_n(x_t|\lambda_i)$ احتمال نرمالیزه شدهٔ بردار x_t بر روی مدل گوینده i ام باشد، این احتمال را به روش زیر حساب می‌کنیم:

$$P_n(x_t \mid \lambda_i) = P(x_t \mid \lambda_i) - Log(Stat_j\{P(x_t \mid \lambda_j)\}) \qquad (۳۶)$$

آمارهٔ $Stat$ می‌تواند هر یک از آماره‌های مطرح شده در قسمت قبل باشد. احتمال رشتهٔ بردارها در سطح گویش چنین به‌دست می‌آید:

$$P(X \mid \lambda_i) = \frac{1}{T}\sum_{t=1}^{T} P_n(x_t \mid \lambda_i) \qquad (۳۷)$$

اگر بخواهیم $P(X|\lambda_i)$ را باز هم در سطح گویش نرمالیزه کنیم و به‌عنوان مثال آمارهٔ مورد نظر ما در سطح گویش آمارهٔ ماکزیمم باشد، چنین عمل می‌کنیم:

$$P_n(X \mid \lambda_i) = P(X \mid \lambda_i) - \underset{j \neq i}{Max}\, P(X \mid \lambda_j) \qquad (۳۸)$$

فرمول فوق با این فرض است که $P(X|\lambda_i)$ خود لگاریتم احتمال است نه احتمال. به این احتمال به‌دست آمده، احتمال نرمالیزه شده هم در سطح فریم و هم در سطح گویش می‌گویند.

۱۲- وزن‌دهی امتیازاتِ مدل [۲۲]

اگر فرض کنیم $P(x_t|\lambda_i)$ احتمال تولید بردار x_t توسط مدل گوینده i اُم باشد، آنگاه احتمالات $P(x_t|\lambda_1)$، $P(x_t|\lambda_2)$ ... و $P(x_t|\lambda_S)$ را به‌صورت نزولی مرتب کرده و به هر مدل یک رتبه اختصاص می‌دهیم و برای مدل λ_i، رتبه را r_i می‌نامیم. مدلی که بیشترین احتمال را تولید کند، دارای رتبهٔ ۱ و مدلی که کمترین احتمال را تولید کند، دارای رتبهٔ S است (S تعداد گویندگان جمعیت است). امتیاز وزن‌دهی شدهٔ مدل به روش زیر محاسبه می‌شود:

$$P_w(x_t|\lambda_i) = w(r_i).P(x_t|\lambda_i) \qquad (۳۹)$$

1. Hybrid Threshold Cohort Normalization

که در آن $w(r)$، نوعی تابع وزنِ کاهنده و به‌عنوان مثال به‌صورت زیر است:

$$w(r) = \frac{S}{\alpha . r} \qquad (٤٠)$$

αبه عنوان مثال می‌تواند برابر یک باشد. در این مقاله پس از وزن‌دهی امتیازات مدل در سطح فریم، عمل نرمالیزاسیون در سطح گویش نیز انجام می‌شود، بدین معنا که ابتدا وزن‌دهی امتیازات در سطح فریم انجام می‌شود، احتمالات وزن‌دهی شدهٔ بردارها برای محاسبهٔ احتمال رشتهٔ بردار بر روی هم انباشته شده و سپس نرمالیزاسیون بر روی احتمال در سطح گویش (رشتهٔ بردار) انجام می‌شود.

۱۳- آزمایشها

دادگان مورد استفاده در این کار، پایگاه دادهٔ تلفنیِ FARSDIGITS1 متشکل از ۱۰۰ گویندهٔ زن و مرد است که گفتارهای ٦١ مرد و ۳۹ زن با کیفیت SNR= 8.8^{dB} از مکالمات تلفنیِ شهری و تعدادی مکالمهٔ راه دور ضبط شده است. هر گوینده ارقام صفر تا نه را در یک تا سه جلسه و از ۱۰ تا ١٦ بار تکرار کرده است. ٪۷۰ گویندگان گفتار خود را در دو یا بیش از دو جلسه ضبط کرده‌اند. نیمی از داده های این دادگان برای آموزش و نیم دیگر برای آزمایش سیستم استفاده شده است. محدودهٔ سنی مردان از ۱۲ تا ٦١ سال و محدودهٔ سنی زنان از ١٤ تا ٥٢ سال است. به ازای هر گوینده، یک مدل مخلوط گاوسی با ٦٤ تابع گاوسی در مدل و ۱۰ مدل پنهان مارکُف به ازای هر یک از ارقام صفر تا نُه آموزش داده شد. لذا به طور کلی برای ۱۰۰ گوینده، ۱۰۰ مدل مخلوط گاوسی و ۱۰۰۰ مدل پنهان مارکُف درنظر گرفته شده است. هر یک از مدلهای پنهان مارکُف دارای ٦ حالت و ٥ تابع گاوسی در هر حالت است. لازم است توجه شود که ساخت مدل HMM بازای هر رقم بدان معنا است که HMM در تعیین و تصدیق هویت گوینده به‌صورت وابسته به متن عمل کند. در حالی که ساخت تنها یک مدل GMM برای کلیه ارقام، به معنای آن است که GMM برای تعیین و تصدیق هویت به‌صورت مستقل از متن استفاده شده است. در مرحلهٔ آزمایش هر یک از گویندگان، ارقامِ صفر تا نُه را یک بار برای شناخته شدن خود آدا می‌کنند و کارایی سیستم اندازه‌گیری می‌شود. خطای تصدیق هویت گوینده به‌صورت $Error(\%) = \frac{FA + FR}{2} \times 100$ درنظر گرفته شده است. تعیین سطح آستانه برای تصمیم‌گیری به روش EER[1] انجام شد. اگر سطح آستانهٔ به‌دست آمده به این روش Thr باشد، سطح آستانهٔ نهایی برابر 0.9*Thr قرار داده شده است.

آزمایش الف: در این آزمایش کارایی سیستم ترکیبی نسبت به هر یک از سیستمهای HMM و GMM و به ازای ویژگیهای مختلف مقایسه می‌شود. نتایج این آزمایش در جدول ۱ آمده است. کارایی به صورت درصد صحت در تعیین هویت و درصد خطا در تصدیق هویت در نظر گرفته شده است. یادآوری می‌شود که نتایج سیستم ترکیبی بدون اعمال CMS و WPM است. α_V و α_I مقادیر بهینهٔ α برای تصدیق و تعیین هویت گوینده است. ملاحظه می‌شود که ازای سیستمهای منفرد (غیر ترکیبی)، بهترین کارایی برای تصدیق هویت متعلق به HMM به ازای پارامترهای MFCC+ΔMFCC است؛ و بهترین کارایی برای تعیین هویت متعلق به GMM به ازای پارامترهای MFCC+ΔMFCC+ΔΔMFCC است. اضافه کردن مشتق اول ضرایب یعنی ΔMFCC کارایی HMM را بالا می‌برد (به علت مدلسازی دینامیک محلی مجرای گفتار - که مدل مارکف با تعداد حالات محدود برابر شش، احتمالاً قادر به مدلسازی آن نیست) - اما اضافه کردن ضرایب ΔΔMFCC (مشتق دوم)، کارایی HMM را پایین می‌آورد. این پدیده شاید به این دلیل

1. Equal-Error Rate

باشد که با اضافه کردن مشتق دوم ضرایب به مدلسازی سراسری دینامیک گفتار نزدیک می‌شویم که خود مدل پنهان مارکف با ماتریس گذر بین حالات، آن را بهتر مدل می‌کند و اضافه کردن مشتق دوم ضرایب سودی ندارد اما اضافه کردن مشتق دوم ضرایب در مدل مخلوط گاوسی - که فاقد احتمالات گذر بین حالات است - کارایی GMM را افزایش می‌دهد.

آزمایش ب: در این آزمایش اثر معیار تصویر وزن‌دهی شده بر روی سیستم بازشناسی گوینده مبتنی بر GMM بررسی می‌شود. با استفاده از پارامترهای MFCC و مشتق اول و دوم آنها و با ٦٤ تابع گاوسی، نتایج برای تصدیق و تعیین هویت گوینده بدون اعمال WPM و با اعمال WPM، در جدول ۲ آمده است. ملاحظه می‌شود که در هر دو حالت تصدیق و تعیین هویت گوینده اعمال WPM موجب بهبود کارایی سیستم شده است.

آزمایش ج: در این آزمایش اثر تفاضل میانگین در حوزهٔ کپسترال یا CMS بر روی سیستم تصدیق هویت گوینده بررسی می‌شود. ملاحظه می‌شود که CMS موجب کاهش اثر کانال انتقال تلفنی بر روی پارامترهای کپسترال و در نتیجه کاهش خطای تصدیق هویت گوینده می‌شود که در جدول ۳ نشان داده شده است. ذکر این نکته لازم است که از WPM در این آزمایش استفاده نشده است.

آزمایش د: در این آزمایش یک بررسی بر روی افراد جمعیت انجام و کارایی سیستم برای تصدیق و تعیین هویت گوینده به ازای جمعیت‌های ۲۰، ٤٠، ۷۰ و ۱۰۰ نفری اندازه‌گیری می‌شود. سیستم پایه برای این آزمایش، مدل مخلوط گاوسی با ٦٤ تابع گاوسی وبا استفاده از معیار تصویر وزن‌دهی شده است. نتایج این آزمایش در جدول ٤ درج شده و مشاهده می‌شود که با افزایش تعداد گویندگان، نرخ خطا در تعیین هویت گوینده بسیار سریعتر از نرخ خطا در تصدیق هویت گوینده رشد می‌کند.

آزمایش ه: در این آزمایش هدف آن است که تأثیر روشهای نرمالیزاسیون در سطح گویش بررسی شود. جدول ۵ نتایج به‌دست آمده را به ازای روشهای مختلف نشان می‌دهد. ستون اول کارایی سیستم را به ازای احتمالات خام و نرمالیزه نشده و ستون‌های بعدی، کارایی سیستم را به ازای امتیازات نرمالیزه شده و با هر یک از ۸ روشِ ذکر شده در بخش ۱۰ نشان می‌دهد.

جدول ۱ نتایج بازشناسی گوینده ازای سیستم ترکیبی

	نرخ صحت در تعیین هویت (٪)				نرخ خطا در تصدیق هویت (٪)			
نوع ویژگی استفاده شده	HMM	GMM	HMM⊕GMM	α_I	HMM	GMM	HMM⊕GMM	α_V
MFCC	۹۳/۳٤	۹۵/۰۱	۹۵/۰۳	۰/۹۲	۰/٤٤	۰/٤٥	۰/٤۲	۰/٤۳
MFCC+ΔMFCC	۹۳/۵۱	۹۵/۱۷	۹۵/۳٤	۰/٤۱	۰/۳۰	۰/٤۱	۰/۳۰	۰/۰۰
MFCC +ΔMFCC + ΔΔMFCC	۹۳/۳٤	۹۵/۳٤	۹۵/۵۱	۰/٦٦	۰/٤۱	۰/٤۰	۰/۳۹	۰/۸۷

جدول ۲ نتایج بازشناسی گوینده پس از اعمال WPM

	نرخ صحت در تعیین هویت (٪)	نرخ خطا در تصدیق هویت (٪)
بدون اعمال WPM	۹۵/۳٤	۰/٤۰
با اعمال WPM	۹۵/۵۱	۰/۳۷

جدول ۵ نشان می‌دهد که در بهترین حالت و به ازای یک بار بیانِ ارقام صفر تا نُه توسط گوینده، خطای تصدیقِ هویتِ گوینده از ۳/۲۵٪ به ۰/٤٪ رسیده که کاهش بسیار چشمگیری را نشان می‌دهد. همچنین می‌توان مشاهده کردکه حتی آمارهٔ می‌نیمم برای نرمالیزه کردنِ امتیازات نیز تا حدّی خطای تصدیق هویت را کاهش می‌دهد. آزمایشها نشان دهندهٔ این موضوع است که آمارهٔ ماکزیمم، کمترین خطا را داشته و نیز روشهای نرمالیزاسیون در سطح گویش تأثیری بر نرخ صحتِ تعیین هویت گوینده ندارند که این با توجه به فرمولهای ارائه شده، منطقی است.

آزمایش و: در این آزمایش، هدف آن است که اثر نرمالیزاسیون امتیازات – هم در سطحِ فریم و هم در سطحِ گویش – بر روی نرخ صحت بازشناسی گوینده بررسی شود. امتیازات مربوط به هر بردار ویژگی (در سطح فریم) به پنج روش از روشهای مذکور در بخش ۱۰ نرمالیزه و سپس احتمالِ انباشته شده (در سطح گویش) با استفاده از آمارهٔ ماکزیمم نرمالیزه می‌شود. نتایج حاصل از این آزمایش در جدول ٦ درج شده است.

جدول ۳ نتایج بازشناسی گوینده پس از اعمال CMS

	نرخ خطا در تصدیق هویت (٪)
بدون اعمال CMS	۰/٤۰
با اعمال CMS	۰/۱٦

جدول ٤ نتایج بازشناسی گوینده به ازای جمعیت با تعداد متغیر از گویندگان

	۲۰ نفر	٤۰ نفر	۷۰ نفر	۱۰۰ نفر
نرخ صحت در تعیین هویت (٪)	۱۰۰	۹۹/۵۲	۹۸/۵۱	۹۵/۵۱
نرخ خطا در تصدیق هویت (٪)	۰/۰۰	۰/۰۰	۰/۰۳	۰/۳۷۳

جدول ۵ نتایجِ به‌دست آمده پس از نرمالیزه کردن امتیازات در سطح گویش

	نرمالیزه نشده	احتمال نرمالیزه شده						
	امتیاز خام	میانگین	ماکزیمم	می‌نیمم	شبیه ترین M گوینده	نرمالیزاسیون گروهی	نرمالیزاسیون گروهی ترکیبی	احتمال پسین
تعیین هویت (درصد صحت)	۹۵/۳٤	۹۵/۳٤	۹۵/۳٤	۹۵/۳٤	۹۵/۳٤	۹۵/۳٤	۹۵/۳٤	۹۵/۳٤
تصدیق هویت (درصد خطا)	۳/۲۵	۰/۹۷	۰/٤۰	۲/۱٤	۰/۹۲	۰/۹۹	۰/۹۹	۰/۹۷

جدول ٦ نتایج به‌دست آمده با نرمالیزه کردن امتیازات در دو سطحِ گویش و فریم

	روش نرمالیزاسیون احتمال در سطح فریم				
	میانگین	ماکزیمم	می‌نیمم	شبیه ترین M گوینده	نرمالیزاسیون گروهی
نرخ صحت تعیین هویت(٪)	۹۲/۸۵	۹۵/۵۱	۹۵/۳٤	۹۵/۵۱	۹۲/۱۸
نرخ خطا در تصدیق هویت(٪)	۰/٦۸	۰/۳۳	۰/۳۸	۰/۳٦	۰/٤۱

ملاحظه می‌شود که در بهترین حالت- یعنی استفاده از آمارهٔ ماکزیمم- نرمالیزه کردن امتیازات در دو سطحِ فریم و گویش، کارایی را نسبت به بهترین نتیجهٔ حاصل شده از نرمالیزاسیون امتیازات، فقط در سطحِ گویش ارتقا می‌دهد. همچنین ملاحظه می‌شود که نرمالیزاسیون امتیازات در سطح فریم، نرخ صحت در تعیین هویت گوینده را بر خلاف روشهای نرمالیزاسیون در سطح گویش، تحتِ تأثیر قرار می‌دهد.

آزمایش ز: هدف از این آزمایش آن است که اثر وزن‌دهیِ امتیازاتِ مدل بر روی نرخ صحت تعیین هویت گوینده و نرخ خطای تصدیق هویت گوینده بررسی شود. همانطور که در بخش ۱۲ ذکر شد، ابتدا امتیازاتِ مدلها در سطح فریم وزن‌دهی شده و سپس احتمال انباشته شده را در سطحِ گویش با استفاده از آمارهٔ ماکزیمم، نرمالیزه می‌کنیم. نتایج به‌دست آمده در جدول ۷ درج شده است. ملاحظه می‌شود که وزن‌دهیِ امتیازات در سطح فریم، خطای تصدیق هویت را اندکی کاهش داده اما بر نرخِ صحت در تعیینِ هویت گوینده تأثیری نداشته است.

۱۴- نتیجه گیری

در این مقاله مدلی ترکیبی برای تصدیق و تعیین هویت گوینده، متشکل از مدل پنهان مارکف و مدل مخلوط گاوسی ارائه شد. آزمایشها نشان داد که این مدل از هر یک از مدلهای پنهان مارکف و مخلوط گاوسی کارایی بیشتری دارد. همچنین برای مقابله با نویز جمع‌شوندهٔ موجود بر روی مکالمات تلفنی، از روش تفاضل طیفی و برای کاهش اثر نویز جمع‌شونده بر روی بردارهای کپسترال از روش تصویر وزن‌دهی شده یا WPM استفاده شد که کارایی سیستم بازشناسی گوینده را ارتقا بخشید. برای مقابله با اثر کانال انتقال تلفنی و اثر دهنی‌های مختلف بر روی بردارهای کپسترال، از روش تفاضل میانگین در حوزهٔ کپسترال یا CMS استفاده شد که این روش نیز کارایی سیستم را افزایش داد. در نهایت بر روی پایگاه دادهٔ تلفنی متشکل از ۱۰۰ نفر گوینده (۶۱ مرد و ۳۹ زن) با $SNR=8.8^{dB}$ نرخ خطای تصدیق هویت در بهترین حالت ۰/۱۶٪ و نرخ صحت در تعیین هویت گوینده ۹۵/۵۱٪ بوده است. روشهای نرمالیزاسیون امتیازات در سطح گویش و در سطح فریم و نیز روش وزن‌دهی امتیازاتِ مدل برای بهبود کارایی سیستمهای تعیین و تصدیق هویت گوینده به‌کار گرفته شد. آزمایشها نشان داده است که نرمالیزاسیونِ امتیازات در هر دو سطحِ گویش و فریم نتیجهٔ بهتری را نسبت به نرمالیزاسیون فقط در سطح گویش به دست می‌دهد. همچنین ملاحظه شد که روش وزن‌دهیِ امتیازاتِ مدل در سطح فریم و قبل از نرمالیزاسیونِ امتیازات در سطح گویش، کارایی سیستم را نسبت به حالت نرمالیزاسیون فقط در سطح گویش ارتقا می‌دهد. در نهایت، بر روی پایگاه دادهٔ تلفنی FARSDIGIT متشکل از ۱۰۰ گوینده (۶۱ مرد و ۳۹ زن) با $SNR=8.8^{dB}$ نرخِ صحت تعیین هویت گوینده و نرخ خطای تصدیق هویت گوینده در حالت نرمالیزه نشده به ترتیب از ۹۵/۳۴٪ و ۳/۲۵٪، به ۹۵/۵۱٪ و ۰/۳۳٪ در حالت نرمالیزه شده رسید که بویژه به ازای تصدیق هویت گوینده کاهش بسیار چشمگیری را نشان می‌دهد.

جدول ۷ مقایسه نتایج بدون وزن‌دهی و با وزن‌دهیِ امتیازات در سطحِ فریم

	نرمالیزاسیون فقط در سطح گویش	وزن‌دهی در سطح فریم و نرمالیزاسیون در سطح گویش
نرخ صحت در تعیین هویت(٪)	۹۵/۳۴	۹۵/۳۴
نرخ خطا در تصدیق هویت(٪)	۰/۴۰	۰/۳۹

۱۵- قدردانی

این کار تحقیقاتی در راستای طرح ملی تحقیقات به شماره NRCI357 انجام و از طرف شورای پژوهشهای علمی کشور حمایت شده است.

۱۶- منابع

[1] S. Furui; Digital Speech Processing Synthesis and Recognition; Marcel Dekker, New York; 1989.

[2] A. E. Rosenberg; "Automatic Speaker Verification: A Review"; Proc. IEEE; Vol. 64 Apr 1976; pp. 475-487.

[۳] م. ر. ذهابی؛ ا. ا. سپهری؛ "استفاده از تصمیم گیرنده‌های باینری در مدل مخفی مارکُف برای شناسایی گوینده "؛ دومین کنفرانس مهندسی برق ایران؛ دانشگاه تربیت مدرس؛ تهران، ایران؛ ۱۳۷۳. صص. ۳۶۱-۳۶۷.

[٤] م. مندولکانی؛ م. لطفی‌زاد؛ "تشخیص هویت گوینده توسط کامپیوتر"، دومین کنفرانس مهندسی برق ایران؛ دانشگاه تربیت مدرس؛ تهران، ایران؛ ۱۳۷۳؛ صص. ۳۵۳-۳۶۰.

[٥] ح. اصغری؛ م.ر. عارف؛ "بازشناسی گوینده با تحقق چندی‌کننده‌های برداری"؛ دومین کنفرانس مهندسی برق ایران؛ دانشگاه تربیت مدرس؛ تهران، ایران؛ ۱۳۷۳؛ صص. ۳۳۵-۳٤٤.

[٦] م. ص. حدائق؛ م. لطفی‌زاد؛ "دومین کنفرانس مهندسی برق ایران؛ دانشگاه تربیت مدرس؛ تهران، ایران؛ ۱۳۷۳؛ صص. ۲۱۲-۲۲۱.

[۷] ا. صیادیان؛ ح. غفوری‌فرد؛ "استفاده از تغییرات دینامیکی ضرایب LSPF جهت کاهش خطای سیستمهای بازشناسی گوینده"؛ سومین کنفرانس مهندسی برق ایران؛ دانشگاه علم و صنعت ایران؛ تهران، ایران؛ ۱۳۷٤؛ صص. ۲۰۷-۲۱۲.

[۸] ح. مقصودلو؛ م. ر. نخعی؛ م. تبیانی؛ "سیستم تأیید هویت گوینده وابسته به متن با استفاده از روش کوانتیزاسیون برداری"؛ سومین کنفرانس مهندسی برق ایران؛ دانشگاه علم و صنعت ایران؛ تهران، ایران؛ ۱۳۷٤؛ صص. ۱۵۵-۱٦۲.

[۹] س. ذ. فیض‌آبادی؛ س. صدوقی؛ "سیستم تشخیص گوینده"؛ ششمین کنفرانس مهندسی برق ایران؛ دانشگاه صنعتی خواجه نصیرالدین طوسی؛ تهران، ایران؛ ۱۳۷۷؛ صص. ۳٦۹-۳۷۲.

[۱۰] ا. صیادیان؛ ک. بدیع؛ م. حکاک؛ م. ر. بیک زاده؛ "ارائهٔ روش TSD-PGMM در بازشناسی گوینده مستقل از متن"؛ هشتمین کنفرانس مهندسی برق ایران؛ دانشگاه صنعتی اصفهان؛ اصفهان، ایران؛ ۱۳۷۹؛ صص. ۳۷٦-۳۸۲.

[۱۱] م. م. همایون‌پور؛ ا. نجاری؛ "تصدیق هویت گوینده توسط تلفیق شبکه‌های عصبی و الگوریتم‌های ژنتیکی"؛ پنجمین کنفرانس بین المللی سالانهٔ انجمن کامپیوتر ایران؛ دانشگاه شهید بهشتی؛ تهران، ایران؛ ۱۳۷۸؛ صص. ۲۵۷-۲٦٤.

[12] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction", IEEE Trans. on ASSP; Vol. ASSP-27, No. 2; April, 1979. pp. 113-120.

[13] J. Pencak, D. Nelson, 'The NP Speech Activity Detection Algorithm', ICASSP-95; Vol. 1; May 1995. pp. 381-384.

[14] A. K. Hunt; "New Commercial Applications of Telephone-Network-based Speech Recognition and Speaker Verification"; EuroSpeech-91; Genova, Italy, 1991. pp. 431-433.

[15] R. J. Mammone et al.; "Robust Speaker Recognition: A Feature-based Approach", IEEE Signal Processing Magazine; Sept. 1996. pp. 58-71

[16] D. Mansour et al., "A Family of Distortion Measures based upon

Projection Operation for Robust Speech Recognition"; IEEE Trans. on ASSP; Vol. 37, No. 11, Nov. 1989.

[17] B. A. Carlson et al. ; "A Projection-based Likelihood Measure for Speech Recognition in Noise"; IEEE Trans. SAP, Vol. 2, No. 1, Jan. 1994.

[۱۸] م. ر. میرحسینی؛ س. م. احدی؛ "معیار تصویر وزن‌دهی شده برای بازشناسی مقاوم گفتار فارسی"؛ کنفرانس مهندسی برق ایران؛ دانشگاه صنعتی اصفهان؛ اصفهان، ایران؛ ۱۳۷۹.

[19] T. Matsui, S. Furui, "Similarity Normalization Method for Speaker Verification based on a Posteriori Probability"; ESCA Workshop on

[20] Automatic Speaker Recognition, Identification, and Verification; Martigny, Switzerland; 1994; pp. 59-62,

[21] F. Chen, et al.; "Hybrid Threshold Approach in Text-Independent Speaker Verification", ICSLP-94; Yokohama, Japan, 1994; pp. 1855-1858.

[22] A. E. Rosenberg et al.; "The Use of Cohort Normalized Scores for Speaker Verification"; ICSLP-92, Banff, Canada; 1992; pp. 599-602.

[23] K. P. Markov et al.; "Text-Independent Speaker Recognition Using Non-linear Frame Likelihood Transformation"; Speech Communication, Vol. 24 1998; pp. 193-209.