علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۳، شماره ۱ (الف)، بهار ۱۳۸۴
صفحات ۳۳-۴۳
مقاله عادی

بهبود سرعت و دقت در تعیین هویت گوینده مجموعه باز

محمد مهدی همایون‌پور　　هدیه رزازان

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر،تهران، ایران

چکیده

به منظور دست یابی به دقت بیشتر در تعیین هویت گوینده مجموعه باز تاثیر نوع ویژگی و امکان استفاده از مشتقات اول و دوم ویژگی ها و نیز تعداد مناسب مخلوط ها در مدل مخلوط گوسی GMM[1] مورد بررسی قرار گرفت. دریک روش ترکیبی پیشنهادی از توانایی GMM در مدل کردن گویندگان و نیز از توان تمایز دهندگی بالای SVM در کنار هم بهره برده و از این ویژگی که مدل های مخلوط گوسی GMM[2]و ماشین بردار پشتیبان SVM[3] با وجود داشتن کارایی نسبتا مشابه، دارای خطاهای نا همبسته نیز می باشند به منظور افزایش دقت تعیین هویت گوینده، استفاده شده است. در روش پیشنهادی در مرحله اول تعیین هویت گوینده توسط GMM صورت می گیرد و در مواردیکه به دلیل تشابه گویندگان احتمال اشتباه وجود داشته دارد، از SVM برای کاهش خطای تعیین هویت استفاده میگردد. تعیین گویندگان مشابه هر گوینده توسط GMM ایجاد میشود و ساخت مدل های SVM جهت تمایز بین گویندگان مشابه در مرحله آموزش صورت می گیرد. روش پیشنهادی موجب گردید که خطای تعیین هویت از ۴/۱۵٪ که مربوط به روش GMM به تنهایی است به ۱/۷٪ مربوط به سیستم هیبرید کاهش یابد. به منظور افزایش سرعت از ایده گروه بندی گویندگان استفاده نموده و گویندگان را به گروه هایی با بیشترین شباهت در درون هر گروه تقسیم و در هنگام تعیین هویت ابتدا گروه گوینده را تعیین و سپس در درون گروه مربوطه به جستجوی گوینده مورد نظر می پردازیم. همچنین در این مقاله از هنجار سازی گروهی در جهت کاهش خطاهای قبول و رد اشتباه در تعیین هویت مجموعه باز استفاده نمودیم که این کار منجر به کاهش زیادی در میزان خطاهای فوق گردید.

کلمات کلیدی: بازشناسی تعیین هویت گوینده مجموعه باز، مدل مخلوط گوسی، ماشین بردار پشتیبان، هنجار سازی گروهی

۱- مقدمه

مطالعه و تحقیق درباره شناسایی گوینده تقریبا از سال ۱۹۶۰ شروع شد [۱]. در سال ۱۹۷۴ آقای آتال یک سیستم تعیین هویت با استفاده از ویژگی های کپستروم و روش تطبیق الگو طراحی کرد. تعداد گوینده های مرجع ده گوینده و سیستم به صورت وابسته به متن طراحی شده بود. میزان خطای گزارش شده ۲٪ است[۲]. سیستم دیگری در سال ۱۹۷۹ توسط آقایان مارکل و همکاران با استفاده از ویژگی ضرائب پیشگویی خطی حاصل از آنالیز بلند مدت و در شرایط آزمایشگاهی طراحی شد. این سیستم به صورت مستقل از متن و با هفده گوینده مرجع آموزش و مورد ارزیابی قرار گرفت و کارایی ۲٪ حاصل شد[۳].

سیستم تعیین هویت دیگری در سال ۱۹۸۱ توسط آقای فوروی در شرکت AT&T با استفاده از ویژگی های کپستروم، به صورت مستقل از متن و به روش تطبیق الگو طراحی شد. پایگاه داده های گفتار شامل داده های تلفنی مربوط به بیست و یک گوینده مرجع بود. میزان خطای سیستم ۲٪ گزارش شده است[۴]. سیستم دیگری در سال ۱۹۸۶ توسط آقای هیگینز در شرکت ITT با استفاده از ویژگی های کپستروم، روش در هم پیچش زمانی(DTW) در شرایط آزمایشگاهی و به صورت مستقل از متن طراحی شد[۱]. در بین سال های ۱۹۹۱ و ۱۹۹۵ آقای بنانی مطالعات و آزمایشاتی بر روی سیستم های تعیین هویت با استفاده از شبکه های عصبی تاخیر زمانی و سیستم های غیر یکپارچه انجام داده است[۵,۶,۷,۸]. آقای بنانی در طراحی سیستم های تعیین هویت خود، از ایده گروه بندی گویندگان بر اساس لهجه یا جنسیت بهره گرفته است. در سال های اخیر نیز تلاش های بسیاری برای افزایش دقت سیستم های تعیین هویت و تصدیق هویت گوینده به ویژه مستقل از متن انجام شده است. برای نمونه می توان به سیستم تعیین هویت طراحی شده توسط آقایان وانگ و همکاران در سال ۲۰۰۱ و با استفاده از مدل های مخلوط گوسی و شبکه های عصبی اشاره نمود. این سیستم مستقل از متن بوده و تعداد گویندگان مرجع برابر ۴۹ گوینده است. نرخ تشخیص گوینده در حدود ۹۲/۰۸٪ گزارش شده است[۹]. معین و همکاران در مقاله خود [۱۰] روشهای HMM، GMM و SVM را بطور جداگانه برای منظور تصدیق هویت بکار برده اند. نتایج ارزیابی آنها حاکی از برتری SVM از نقطه نظر دسته بندی و GMM از نقطه نظر سرعت محاسبات می باشد.آقایان گوپینات و فاین در سال

۲۰۰۱ سیستم هیبریدی را طراحی کردند که در آن مدل GMM و مدل SVM با یکدیگر ترکیب شده بودند. هدف آنها از این ترکیب افزایش دقت تعیین هویت بوده است [۱۲,۱۱].

مدل GMM دارای کارایی بالایی در مدل کردن گویندگان برای تعیین هویت مستقل از متن به ویژه در حالتی که شرایط آموزش و تست متفاوت است، میباشد [۱۳]. SVM از روش های کلاس بندی است که در سال های اخیر جایگاه ویژه ای در مسائل شناسایی الگو و طبقه بندی پیدا کرده است [۱۵,۱۴]. SVM از توان تمایز دهندگی بالایی برخوردار می باشد. از معایب بزرگ آن، محاسبات پیچیده و سرعت پایین آن است که همین امر استفاده از SVM را محدود می کند. در این مقاله روشی برای تعیین هویت گوینده بر اساس ترکیب توانایی های مدل و SVM مطرح می شود که علاوه بر افزایش دقت تعیین هویت، مشکل سرعت را که در سیستم های هیبرید قبلی مانند [۱۲,۱۱]. مطرح بود را تا حد زیادی بهبود می بخشد. در این روش سعی شده است از این ویژگی که نواحی خطای دو نوع کلاس بند GMM و SVM لزوما هم پوشانی کامل ندارند، برای افزایش کارایی سیستم هیبرید، استفاده شود.

سیستم های تعیین هویت گوینده به دو دسته مجموعه بسته[۴] و مجموعه باز[۵] تقسیم میشوند. در سیستم های مجموعه بسته گوینده مجهول عضو مجموعه گویندگانی است که سیستم با گفتار آنها آموزش دیده است. اما در سیستم های مجموعه باز این امکان وجود دارد که گفتار ورودی متعلق به هیچ یک از اعضای مجموعه گوینده گان مرجع نباشد. در تعیین هویت مجموعه باز ابتدا باید تعلق گوینده مجهول به مجموعه گویندگان مرجع تشخیص داده شده و در صورتیکه وی متعلق به مجموعه گویندگان مرجع باشد، تعیین هویت او انجام می شود. سیستم های تعیین هویت مجموعه باز دارای پیچیدگی بیشتری نسبت به سیستم های تعیین هویت مجموعه بسته هستند، در واقع این نوع سیستم ترکیبی از سیستم تعیین هویت مجموعه بسته و تصدیق هویت می باشد.

سیستم های تعیین هویت از نقطه نظر متن گویشی که برای تعیین هویت شدن نیاز دارند میتوانند از نوع مستقل از متن و یا وابسته به متن باشند. سیستم های مستقل از متن برخلاف سیستم های وابسته به متن به کاربران این امکان را می دهند که لزوما متن از پیش تعیین شده ای را استفاده نکنند.

بعد از بیان تاریخچه و مقدمات ارائه شده، در بخش های ۲ و۳ این مقاله ابتدا مروری کوتاه بر مدل های GMM و SVM خواهیم داشت. بخش ۴ به بیان انواع خطاها در تعیین هویت مجموعه باز اختصاص دارد. بخش ۵ به چگونگی تشخیص گفتار از سکوت می پردازد. بخش ۶ دادگان گفتاری مورد استفاده در این تحقیق و نیز چگونگی استخراج ویژگیها را بیان می نماید. بخش ۷ به چگونگی پیاده سازی، ارزیابی و تحلیل نتایج حاصل از انجام آزمایشات برای انجام تعیین هویت گوینده بروش پیشنهادی،گروه بندی گویندگان، تعیین هویت مجموعه باز، تعیین سطح آستانه و هنجار سازی گروهی اختصاص دارد. بخش ۸ نیز به نتیجه گیری و جمع بندی بررسی های صورت گرفته در این مقاله می پردازد.

۲- مدل مخلوط گوسی GMM

مدل GMM کارایی مناسبی در سیستم های تعیین هویت به ویژه در حالت مستقل از متن دارند. به همین دلیل از مدل های GMM به عنوان سیستم مبنا در بسیاری از مدل های هیبرید استفاده می شود و نیز در بسیاری از آزمایشات به عنوان مدل مرجع جهت مقایسه استفاده شده است. مدل مخلوط گوسی l_l مجموع وزنی M جزء چگالی مخلوط گوسی است. احتمال بردار ویژگی تصادفی $\vec{x}$ در مدل l_l توسط رابطه زیر محاسبه می شود:

$$P(\vec{x}/\, l_l)=\sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (۱)$$

$b_i(\vec{x})$ احتمال مشاهده بردار $\vec{x}$ است که این احتمال از مخلوط i ام ناشی شده است و p_i وزن مخلوط i ام می باشد. $b_i(\vec{x})$ از مخلوط گوسی با میانگین $\vec{m}_i$ و ماتریس کوواریانس Σ_i به صورت زیر بدست می آید:

$$b_i(\vec{x})=\frac{1}{(2p)^{\frac{D}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\vec{x}-\vec{m}_i)'\Sigma^{-1}(\vec{x}-\vec{m}_i)\right) \qquad (۲)$$

$\vec{m}_i$ بردار میانگین مخلوط i ام و Σ_i ماتریس کوواریانس این مخلوط می باشد. وزن مخلوط ها باید شرط $\sum_{i=1}^{M} p_i = 1$ را برآورده کنند. این شرط تضمین می کند که احتمال گوسی هر مخلوط، یک تابع چگالی احتمال صحیح می باشد. مدل GMM را می توان توسط بردار میانگین و ماتریس کوواریانس و وزن های هر مخلوط به صورت زیر تعریف نمود:

$$\lambda_{\ l}=\{p_i,\mu_i,\Sigma_i\} \qquad (۳)$$

مدل ترکیب گوسی بسته به نوع ماتریس کوواریانس چندین نوع می تواند داشته باشد. اگر ماتریس کوواریانس، کامل باشد، مدل گوسی مدل کامل گفته می شود. می توان ماتریس کوواریانس را به صورت یک ماتریس قطری در نظر گرفت، همچنین میتوان از یک ماتریس مشترک برای تمام مخلوط ها استفاده کرد. برای توضیحات بیشتر می توان به مرجع [۱۳] مراجعه نمود.

۳- ماشین بردار پشتیبان SVM

فرض کنید تعدادی نمونه آموزشی داریم که به صورت جدایی پذیر خطی بوده و به صورت زیر باشند:

$$(x_i,y_i)\quad x_i\in R^n \qquad y_i=\pm 1 \qquad i=1,\ldots,N \qquad (۴)$$

هدف یافتن تابع تصمیم گیری خطی f(x) و ابرصفحه بهینه جدا کننده دو کلاس می باشد.

$$H: w.x+b=0 \qquad (۵)$$

$$F(x)=\mathrm{sgn}(w.x+b) \qquad (۶)$$

بردار وزن w، بردار عمود بر ابرصفحه جدا کننده و b مقدار بایاس است. $w.x$ حاصل ضرب داخلی بردار ویژگی x و بردار وزن w است. فاصله عمودی ابرصفحه تا مبدا مختصات برابر $\frac{|b|}{\|w\|}$ است. فرض کنید d_+ (d_-) کوتاه ترین فاصله ابرصفحه تا نزدیکترین نمونه مثبت (منفی) باشد. $d_+ + d_-$ را حاشیه ابرصفحه می نامیم. ابرصفحه بهینه ابرصفحه ای است که دارای بزرگترین حاشیه باشد. الگوریتم بردار پشتیبان به دنبال یافتن چنین ابرصفحه ای است. بنابراین باید $\|w\|^2$ مینیمم شود. می توان روش کار را به صورت زیر فرموله کرد:

نقاط آموزشی باید شرایط زیر را ارضاء کنند :

$$w.x_i + b \geq 1 \qquad for\ y_i = 1 \qquad (۷)$$

$$w.x_i + b \leq 1 \qquad for\ y_i = 1 \qquad (۸)$$

شرایط فوق در نامعادله زیر خلاصه می شوند:

$$y_i\ (w.x_i + b) - 1 \geq 0 \qquad \forall i \qquad (۹)$$

حال نقاطی را که به ازاء آنها، تساوی نامعادله (۷) برقرار باشد، در نظر می گیریم. این نقاط روی ابرصفحه H_1 قرار می گیرند.

$$H_1 : w.x_i + b = 1 \qquad (۱۰)$$

فاصله عمودی این ابرصفحه تا مبدا برابر $\frac{|b|}{\|w\|}$ می باشد. به طور مشابه، نقاطی که به ازاء آنها حالت تساوی نامعادله (۸) برقرار باشد، روی ابرصفحه H_2 قرار می گیرند.

$$H_2 : w.x_i + b = -1 \qquad (۱۱)$$

فاصله عمودی این ابرصفحه تا مبدا برابر $\frac{|1-b|}{\|w\|}$ است.

$d_+ = d_- = \frac{1}{\|w\|}$ و حاشیه ابرصفحه برابر است با $\frac{2}{\|w\|}$. ابرصفحه های H_2 و H_1 موازی هستند و هیچ یک از نمونه های آموزشی بین این دو ابرصفحه قرار نمی گیرند. همانطور که قبلا نیز اشاره شد، هدف یافتن ابرصفحه ای با بزرگترین حاشیه می باشد. بنابراین برای رسیدن به این هدف باید $\|w\|^2$ را با در نظر گرفتن شرط ۹ مینیمم کرد. روال فوق را می توان به صورت زیر خلاصه کرد:

$$\begin{aligned} &Minimize \quad \frac{1}{2}\|w\|^2 \\ &subject\ \ to\ \ y_i(w.x_i + b) \geq 1 \\ &for\ \ i = 1,2,...,N \end{aligned} \qquad (۱۲)$$

زیر مجموعه ای از نقاط آموزشی که روی ابرصفحه های H_2 و H_1 قرار می گیرند و حذف هر یک از آنها راه حل پیدا شده را تغییر خواهد داد، بردار پشتیبان نامیده می شود. مساله (۱۲) یک مساله بهینه سازی مقید درجه دوم محدب[i] است. برای حل مساله فوق ابتدا آن را به فرم لاگرانژ تبدیل می کنیم. در فرم لاگرانژ نمونه های آموزشی به صورت ضرب داخلی بردارها دیده می شوند. فرم لاگرانژ مساله به صورت زیر است:

$$L_P = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} a_i\ y_i\ (x_i.w + b) + \sum_{i=1}^{N} a_i \qquad a_i \geq 0 \qquad (۱۳)$$

L_P باید نسبت به w و b مینیمم و نسبت به ضرائب a_i ماکزیمم شود. برای اینکه (w,b,a) جواب مساله باشد، باید در شرایط KKT صدق کرده و نیز در نقطه جواب مشتق L_P نسبت به a, b, w صفر شود. اگر گرادیان تابع L_P را نسبت به w و b برابر صفر قرار دهیم ، خواهیم داشت :

$$w = \sum_{i=1}^{N} a_i\ y_i\ x_i \qquad (۱۴)$$

$$\sum_{i=1}^{N} a_i\ y_i = 0$$

با جایگذاری معادلات فوق در تابع هدف، به فرم دوگان آن خواهیم رسید:

$$\begin{aligned} &Maximize\ L_D = \sum a_i - \frac{1}{2}\sum_{i,j} a_i\ a_j\ y_i\ y_j\ x_i\ x_j \\ &Subject\ to \qquad a_i \geq 0 \qquad for\ i = 1,...,N \\ &\sum_{i=1}^{N} a_i\ y_i = 0 \end{aligned} \qquad (۱۵)$$

L_P و L_D هر دو از یک تابع هدف ولی با شرایط متفاوت مشتق شده اند. حل مورد نظر با مینیم کردن L_P و یا ماکزیمم کردن L_D بدست می آید.

آموزش بردار پشتیبان شامل ماکزیمم کردن L_D نسبت به a_i و با در نظر گرفتن شرط (۱۴) و مثبت بودن a_i می باشد. پس از حل این مساله، ضرائب لاگرانژ a_i بدست می آیند. هر یک از این ضرائب متناظر با یکی از نمونه های آموزشی می باشد. آن دسته از نمونه های آموزشی که ضریب لاگرانژ آن بزرگتر از صفر باشد، بردار پشتیبان نامیده می شود و بر روی یکی از ابرصفحه های H_2 و H_1 قرار می گیرد.

بردارهای پشتیبان عناصر اصلی مجموعه داده های آموزشی هستند و در نزدیکترین فاصله مرز جدا کننده قرار می گیرند. چنانچه به جز بردارهای پشتیبان، سایر نقاط آموزشی حذف شوند و یا در پیرامون خود حرکت کنند ولی از H_2 و H_1 عبور نکنند، با تکرار آموزش، ابرصفحه جدا کننده مشابهی بدست خواهد آمد. پس از بدست آوردن ضرائب لاگرانژ، بردار وزن و مقدار بایاس b از روابط زیر بدست می آیند:

$$\begin{aligned} &w = \sum_{i=1}^{Ns} a_i\ y_i\ Sv_i \\ &b_i = y_i - \sum_{j=1}^{Ns} a_j\ y_j\ Sv_j\ Sv_i \\ &b = \frac{1}{N}\sum_{j=1}^{Ns} b_j \end{aligned} \qquad (۱۶)$$

N_s تعداد بردارهای پشتیبان و Sv_i بردار پشتیبان می باشد. تابع تصمیم گیری نمونه ورودی x به صورت زیر خواهد بود:

$$f(x) = sgn\sum_{i=1}^{Ns} a_i\ \ y_i\ x.Sv_i + b \qquad (۱۷)$$

چنانچه الگوریتم فوق درباره داده های جدایی پذیر خطی، به داده های جدایی ناپذیر اعمال شود، راه حلی عملی پیدا نخواهد شد. در مورد داده های جدایی ناپذیر باید هزینه ای اضافی برای خطاهای آموزشی احتمالی تعریف کرد. بدین منظور متغیرهای z را تعریف می کنیم طوری که شرایط زیر برقرار باشند:

$$\begin{aligned} &w.x_i + b \geq 1 - z_i \qquad for\ \ y_i = +1 \\ &w.x_i + b \leq -1 + z_i \qquad for\ \ y_i = -1 \end{aligned} \qquad (۱۸)$$

$$z_i \geq 0 \qquad \forall i \qquad (۱۹)$$

با رخ دادن هر خطا، متغیر z_i متناظر آن یک واحد افزایش خواهد یافت. $\sum_i z_i$ حد بالای خطای آموزشی را نشان می دهد. برای اعمال هزینه اضافی به خطاهای آموزشی باید تابع هدف به جای مینیمم کردن $\frac{\|w\|^2}{2}$، $\frac{\|w\|^2}{2}+C(\sum_i z_i i)$ را مینیمم کند که C پارامتری است که توسط کاربر انتخاب می شود. انتخاب مقداری بزرگ برای C معادل با اعمال هزینه بالا برای خطای آموزشی می باشد. با توجه به توضیحات عنوان شده، مساله ماشین بردار پشتیبان در مورد داده های جدایی ناپذیر به صورت زیر قابل ارائه می باشد.

$$Maximize\ \ L_D=\sum a_i-\frac{1}{2}\sum_{i,j} a_i\ a_j\ y_i\ y_j\ x_i\ x_j \qquad (۲۰)$$

$$Subject\ \text{to}\quad 0\le a_i\le C\qquad \text{for } i=1,...,N$$

$$\sum_{i=1}^{N} a_i\ y_i=0 \qquad (۲۱)$$

$$w=\sum_{i=1}^{Ns} a_i\ y_i x_i \qquad (۲۲)$$

N_s تعداد بردارهای پشتیبان می باشد. همان طور که مشاهده می شود، تنها تفاوت مساله در حالت داده های جدایی ناپذیر خطی با داده های جدایی پذیر، محدود بودن a_i به C می باشد. در اکثر کاربردهای عملی، داده ها جدایی پذیر خطی نیستند، لذا باید بتوان به نحوی ایده ماشین پشتیبان خطی را به داده های غیر خطی تعمیم داد. راه حلی که به نظر می رسد، نگاشت داده ها به فضایی جدید است که در آن فضا تحت تبدیل خاص، داده ها به صورت جدایی پذیر خطی تبدیل شوند. فضای جدید را H و تابع نگاشت را Φ می نامیم.

$$\Phi: R^d \rightarrow H \qquad (۲۳)$$

از آنجا که در دوگان تابع هدف، داده های آموزشی فقط به صورت ضرب داخلی بردارها دیده شده اند، الگوریتم آموزشی در این مورد نیز فقط به ضرب داخلی داده های نگاشت یافته در فضای H یعنی $\Phi(x_i).\Phi(x_j)$ وابستگی دارد. حال اگر تابع هسته، k را به صورت زیر تعریف کنیم.

$$k(x_i,x_j)=\Phi(x_i).\Phi(x_j) \qquad (۲۴)$$

در الگوریتم آموزش فقط نیاز به دانستن تابع $k(x_i,x_j)$ داریم و نیازی به دانستن تابع نگاشت نخواهیم داشت. با اعمال تابع $k(x_i,x_j)$ به داده های آموزشی، مساله به یک مساله جدایی پذیر خطی ولی در فضایی متفاوت تبدیل می شود. مساله بهینه سازی دوگان در حالت جدایی ناپذیر و غیر خطی به صورت زیرخواهد بود:

$$Maximize \qquad \sum_{i=1}^{N} a_i-\frac{1}{2}\sum\sum a_i a_j y_i y_j\ K(x_i,x_j) \qquad (۲۵)$$

$$\text{Subject to}\quad 0\le a_i\le C$$

$$\sum_{i=0}^{N} a_i y_i=0 \qquad (۲۶)$$

و تابع تصمیم گیری به صورت زیر تبدیل می شود :

$$f(x)=\sum_{i=1}^{Ns} a_i y_i \Phi(Sv_i).\Phi(x)+b = \sum_{i=1}^{Ns} a_i y_i\ K(Sv_i,x)+b \qquad (۲۷)$$

برای توضیحات بیشتر می توان به مرجع [۱۴] مراجعه نمود.

۴- انواع خطا در تعیین هویت مجموعه باز

در سیستم های تعیین هویت مجموعه باز سه نوع خطا وجود دارد: پذیرش اشتباه، رد اشتباه و کلاس بندی اشتباه[۷]. منظور از پذیرش اشتباه در سیستم های تعیین هویت مجموعه باز این است که گوینده مجهول به عنوان عضوی از مجموعه گویندگان مرجع قبول شود، در حالیکه عضو این مجموعه نیست. منظور از رد اشتباه، عدم پذیرش گوینده مجهول به عنوان عضوی از مجموعه گویندگان مرجع است در حالیکه وی عضو این مجموعه می باشد. نوع سوم خطا که کلاس بندی اشتباه می باشد و بین سیستم های تعیین هویت مجموعه باز و مجموعه بسته مشترک است و به خطایی گفته می شود که گفتار مربوط به یک گوینده به گوینده دیگری نسبت داده شود.

۵- تشخیص گفتار از سکوت

به منظور تشخیص گفتار از سکوت و بدنبال آن حذف سکوت از فایل های صوتی از روش ساده حذف سکوت مبتنی بر انرژی استفاده نمودیم. تحقیق صورت گرفته در [۱۶] گویای مناسب بودن این روش در شرایطی که دادگان گفتاری از SNR خوبی برخوردار باشد می باشد. در این روش حذف سکوت از تفاوت انرژی بخش های سکوت و گفتار استفاده شده است. ابتدا گفتار به سگمنتهای متوالی ۰/۰۵ ثانیه تبدیل شده و انرژی آن محاسبه می گردد. سپس انرژی سگمنت جاری با یک سطح آستانه که به صورت تجربی به دست آمده، مقایسه می شود. چنانچه انرژی سگمنت جاری کمتر از سطح آستانه باشد، سگمنت جاری، سکوت تشخیص داده شده، از فایل گفتار حذف می شود. سعی شده است سطح آستانه به گونه ای در نظر گرفته شود که سگمنت های حاوی اصوات بی واک که دارای سطح انرژی پایین تری نسبت به اصوات واکدار هستند، به عنوان سکوت تشخیص داده نشوند.

۶- دادگان گفتاری و استخراج ویژگی

در این مقاله از بانک اطلاعاتی فارس دات بزرگ استفاده شده است. فارس دات شامل ۱۰۰ گوینده با لهجه های مختلف می باشد. هر گوینده ۴۰۰۰ کلمه را در قالب جملاتی از روزنامه بیان کرده است. از چهار نوع میکروفن برای ضبط صدا استفاده شده و گفتار با فرکانس ۲۲۰۵۰ هرتز و ۱۶بیت به ازای هر نمونه و به صورت مونو ضبط شده است. مجموعه گویندگان در نظر گرفته شده برای این تحقیق، شامل ۵۰ گوینده با لهجه تهرانی از این بانک اطلاعاتی می باشد که نمونه های صوتی استفاده شده به ازای هر گوینده شامل ۵ فایل مجموعا ۱۰ دقیقه ای گفتار ضبط شده با میکروفون پایه ای می باشد. به دلیل آنکه بخش های سکوت برای تعیین هویت گوینده مفید نمی باشند و با توجه به یکسان بودن آن درگفتار گویندگان مختلف، سکوت می تواند بر تعیین هویت گوینده تاثیر منفی داشته باشد، لذا ابتدا بخش های سکوت از سیگنال گفتار حذف گردید و پس از اعمال فیلتر دیجیتال پایین گذر، با انتخاب یک در میان نمونه های صوتی، فرکانس نمونه برداری از ۲۲۰۵۰ به ۱۱۰۲۵ کاهش داده شد. برای استخراج ویژگی از فریم های ۳۰ میلی ثانیه ای با همپوشانی ۲۰ میلی ثانیه استفاده شده است. پنجره اعمال شده به فریم ها، پنجره همینگ می باشد. با توجه به مطالعات انجام شده، ویژگی

های MFCC[۸] و LPCC[۹] و مشتقات اول و دوم آنها حاوی اطلاعات مفیدی برای تمایز بین گویندگان می باشند و لذا برای سیستم تعیین هویت گوینده انتخاب مناسبی هستند. برای انتخاب ویژگی بهینه و نیز برای درک تاثیر مشتقات اول و دوم بر دقت تعیین هویت گوینده، ویژگی های MFCC و LPCC و مشتق اول و دوم آنها از گفتار استخراج و آزمایش گردیدند.

۷- پیاده سازی و آزمایشات

قبل از هر چیز لازم به ذکر است که آزمایشات ارائه شده در این مقاله بر روی کامپیوتر Pentium III با ۱۲۸ مگا بایت حافظه RAM صورت گرفته اند.

۷-۱ تعیین نوع ویژگی

تجربات قبلی در زمینه تعیین هویت گوینده گویای کارایی بهتر ضرایب LPCC برای تعیین هویت در محیط های تمیز و غیر تلفنی و کارایی بهتر ضرایب MFCC برای محیط های نویزی و تلفنی است[۱۷]. با توجه به نسبت بالای نسبت سیگنال به نویز دادگان مورد استفاده در این تحقیق و غیر تلفنی بودن آن، از ضرایب LPCC به عنوان ویژگی استفاده گردید. به منظور بررسی تاثیر استفاده از مشتق اول و دوم این ویژگیها بر راندمان تعیین هویت گوینده آزمایشی صورت گرفت. در این آزمایش تعداد ضرائب LPCC ۱۲ ضریب در نظر گرفته شد. مرتبه تحلیل LPC برای تعیین ضرایب LPCC برابر 12 و طول پنجره برابر ۳۰ میلی ثانیه با همپوشانی ۲۰ میلی ثانیه انتخاب شد. در این آزمایشات از مدل GMM بعنوان کلاسبند استفاده گردید. لازم بذکر است که به منظور نرمالیزه کردن ضرائب کپسترال و جبران اثر نوع میکروفون بکار رفته در ضبط صدا، میانگین بردارهای ویژگی از هر یک از آنها کم گردید. از بردارهای ویژگی حاصل از 25 ثانیه داده گفتاری برای آموزش مدل GMM هر گوینده و نیز از ۴۰ نمونه گفتار ۲ ثانیه ای گفتار بازاء هر گوینده برای تست های تعیین هویت استفاده گردید. تعداد مخلوط های گوسی هر مدل ۳۲ انتخاب شد. نتایج آزمایشات در جداول ۱ آمده است. در این جدول اضافه کردن حروف D و A در ادامه نام ویژگی به ترتیب به معنای اضافه کردن مشتق اول و مشتق دوم به بردار ویژگی می باشد. اضافه کردن حرف Z به نام ویژگی نیز به معنای آن است که میانگین ضرائب کپسترال از آنها کاسته شده است[۱۰].

جدول ۱- بررسی تاثیر مشتق اول و دوم ضرائب LPCC در کارایی تعیین هویت گوینده

نوع ویژگی	خطا (٪)
LPCC_D_A_Z	٤/١٥
LPCC_D_Z	٤/٩٥
LPCC_Z	٩/٥

با بررسی نتایج به دست آمده از آزمایشات مشاهده می شود که اضافه کردن مشتق اول و دوم به ویژگی های LPCC منجر به افزایش کارایی و کاهش خطای تعیین هویت می گردد. بنابر نتایج بدست آمده ویژگی LPCC همراه با مشتقات اول و دوم آن برای تعیین هویت گوینده در آزمایشات بخشهای بعد در این مقاله مورد استفاده قرار خواهند گرفت.

۷-۲ تعیین تعداد مخلوط های گوسی

از جمله عواملی که تاثیر بسیار در کارایی مدل های GMM دارند، تعداد مخلوط های گوسی هر مدل می باشد. تعداد مخلوط ها مرتبط با میزان داده آموزشی و میزان کارایی مورد نظر در تعیین هویت مرتبط است. با افزایش تعداد مخلوط ها زمان آموزش مدل نیز افزایش می یابد. لذا هدف اصلی انتخاب حداقل تعداد مخلوط های گوسی لازم برای مدل کردن گوینده با حفظ کارایی مناسب می باشد. انتخاب تعداد کم مخلوط ها باعث می شود که نتوان به خوبی خصوصیات و ویژگی های گوینده را مدل کرد. انتخاب تعداد زیاد مخلوط ها به نسبت میزان داده موجود برای آموزش نیز علاوه بر کاهش سرعت آموزش، باعث کاهش کارایی نیز میگردد. در آزمایشات زیر کارایی مدل های GMM به ازاء تعداد مخلوط های ۱۶، ۳۲ و ۶۴ برای حجم داده یادگیری به میزان۲۵ و ۶۰ ثانیه، بررسی شده اند.

ابتدا داده آموزشی شامل ۲۴۹۸ بردار ویژگی معادل ۲۵ ثانیه گفتار می باشد. برای تست مدل ها از ۲۰۰۰ نمونه تست ۲ ثانیه ای (۴۰ نمونه به ازای هر گوینده) استفاده شد. مدل ها با تعداد اجزائ ۱۶، ۳۲و ۶۴ آموزش داده و تست شدند. نتیجه این آزمایش بر حسب در صد خطای تعیین هویت، در جدول شماره ۲ آورده شده است.

جدول۲- خطای تعیین هویت گوینده به ازاء تعداد مخلوط های گوسی در مدل GMM برای ۲۵ ثانیه داده آموزشی

تعداد مخلوط گوسی	تعداد خطا	خطا (٪)
۶۴	۸۷	۴/۳۵
۳۲	۸۳	۴/۱۵
۱۶	۱۱۰	۵/۵

با توجه به نتایج آزمایش معلوم می شود که برای ۲۵ ثانیه داده آموزشی،۳۲ مخلوط گوسی برای هر مدل، مناسب ترین انتخاب می باشد. علت آنکه بازاء ۶۴ مخلوط گوسی خطا افزایش یافته است ناکافی بودن داده های آموزشی می باشد. آزمایش بعدی نیز مؤید این نکته میباشد. زمان آموزش هر مدل با ۳۲ مخلوط گوسی ۱۹۲ ثانیه می باشد.

در آزمایش دیگری، آزمایش قبل با داده های آموزشی بیشتری شامل ۵۹۹۸ بردار ویژگی معادل ۶۰ ثانیه گفتار و تعداد مخلوط های گوسی ۱۶، ۳۲، ۶۴ و ۱۲۸ تکرار گردید. نتایج آزمایشات در جدول 3 آورده شده است. کمترین خطا به ازاء ۶۴ مخلوط گوسی بدست آمده است.

با مقایسه جداول ۲ و ۳ می توان دریافت که با افزایش داده های آموزشی می توان تعداد مخلوط های گوسی را افزایش داده و به راندمان بهتری دست یافت. لیکن باید توجه داشت که با حجم محدود داده های آموزشی، افزایش بیش از حد تعداد مخلوط ها نه تنها راندمان را بهبود نمی بخشد بلکه منجر به کاهش آن می شود.

جدول ۳- خطای تعیین هویت گوینده با توجه به تعداد مخلوط گوسی به ازاء ۶۰ ثانیه داده آموزشی

تعداد مخلوط گوسی	تعداد خطا	خطا (٪)
۱۲۸	۷۸	۳/۹
۶۴	۷۱	۳/۵
۳۲	۹۸	۴/۹
۱۶	۱۴۵	۷/۲

۷-۳ روش پیشنهادی برای تعیین هویت

سیستم تعیین هویت پیاده سازی شده دارای دو سطح کلاس بندی می باشد. ابتدا نمونه تست (یک مجموعه از داده های آزمایشی) با کلاس بندی کننده های سطح اول تعیین هویت می شود. کلاس بندی کننده های سطح اول مدل های GMM هستند که توان توصیفی بالایی داشته و کارایی خوبی در مدل کردن گویندگان به ویژه در حالتی که شرایط آموزش و تست یکسان نباشند، دارند. در کلاسبند سطح اول هدف آن است که با ایجاد ماتریس اشتباهات[۱۱]، مجموعه گویندگانی را که از

لحاظ گفتاری به یکدیگر شبیه هستند پیدا نمائیم. پس از یافتن مجموعه های گویندگان مشابه، برای متمایز نمودن گویندگان در هر یک از این مجموعه ها و برای رفع عدم قطعیت و اشتباهات کلاس بندی کننده های سطح اول از کلاسبند سطح دوم یعنی SVM استفاده می نمائیم.

در هنگام تعیین هویت، گفتار گوینده تست ابتدا به مدل های GMM که کلاس بندی کننده های سطح اول هستند، اعمال می شود. نتیجه تعیین هویت توسط کلاس بندی کننده سطح اول را گوینده i مینامیم، چنانچه مجموعه گویندگان مشابه گوینده i تهی باشد، یعنی هیچ گوینده دیگری اشتباها به جای این گوینده تعیین هویت نشده باشد، این گوینده به عنوان نتیجه نهایی تعیین هویت پذیرفته می شود. در غیر این صورت کلاس بندی کننده های سطح دوم به عنوان تعیین کننده گوینده واقعی وارد عمل می شوند.

۷-۳-۱ کلاسبند سطح اول و تولید ماتریس اشتباهات

برای تولید ماتریس اشتباهات، مدل‌های مخلوط گوسی با ۱۵۰ نمونه تست ۲ ثانیه ای به ازای هر گوینده تست شدند. برای این منظور از بخش دیگری از دادگان که در مرحله آموزش مدل های مخلوط گوسی استفاده نشده بود استفاده گردید. به ازاء هر نمونه تست، تعیین هویت صورت گرفت، یعنی اینکه گوینده مربوط به مدل با بیشترین مقدار احتمال مشخص گردید. به کمک نتایج تعیین هویت حاصل از این آزمایشاتط ماتریس اشتباهات تولید شد. قسمتی از ماتریس اشتباهات حاصل از نتایج تعیین هویت مدل های مخلوط گوسی در جدول شماره ۴ آورده شده است. سطر اول ماتریس اشتباهات حاوی شناسه گویندگان واقعی و ستون اول این ماتریس شامل شناسه گویندگان شناسایی شده است. هر درایه ماتریس نشاندهنده تعداد نمونه های تست متعلق به گوینده ستون متناظر است که به عنوان گوینده سطر متناظر تعیین هویت شده اند. سطح آستانه ۲ در نظر گرفته شده است. به بیان دیگر آن گویندگانی که تعداد دفعاتی که به اشتباه به عنوان گوینده مرجع i تعیین هویت شده اند، از ۲ بیشتر باشد در مجموعه $F(i)$ قرار می گیرند. برای نمونه در سطر اول این ماتریس که متناظر با گوینده مرجع ۱ می باشد، ۵ بار گوینده ۱۲، ۵ بار گوینده ۵۴ و ۲ بار گوینده ۵۶ به عنوان این گوینده به اشتباه تعیین هویت شده اند و در مجموعه $F(1)$ قرار می گیرند. این ماتریس نشان می دهد که گویندگان متفاوت، درجات کلاس بندی متفاوتی دارند.

جدول ۴- قسمتی از ماتریس اشتباهات حاصل از نتایج کلاس بندی کننده سطح اول

	۱	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸	۱۹	۲۳	۲۶
۱	۱۵	۰	۰	۰	۰	۰	۰	۰	۰	۰	۵	۰	۰	۰	۰	۰	۰	۰	۰	۰
۳	۰	۱۵	۰	۰	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲	۰	۰	۰
۴	۰	۰	۱۵	۰	۰	۰	۰	۰	۲	۰	۰	۰	۰	۸	۰	۰	۰	۰	۰	۰
۵	۰	۰	۰	۱۴	۰	۱	۰	۰	۰	۰	۱	۰	۰	۱	۰	۰	۰	۰	۰	۰
۶	۰	۰	۰	۰	۱۴	۰	۰	۰	۰	۰	۰	۰	۰	۱	۰	۰	۰	۰	۰	۰
۷	۰	۰	۰	۰	۰	۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۸	۰	۰	۰	۰	۰	۰	۱۴	۴	۰	۰	۰	۰	۰	۱	۰	۰	۰	۰	۲	۰
۹	۰	۰	۰	۰	۰	۰	۰	۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۰	۰	۰	۰	۱	۰	۱	۰	۰	۱۴	۰	۰	۰	۰	۲	۰	۰	۰	۰	۰	۰
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۵	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۸۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۱۳	۰	۰	۰	۰	۱	۰	۰	۰	۰	۰	۰	۱۴	۱	۳	۰	۰	۴	۰	۰	۰
۱۴	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲	۱۴	۰	۰	۴	۰	۰	۰	۰
۱۵	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۳	۰	۰	۰	۰	۰	۰
۱۶	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۴	۰	۰	۰	۰	۰
۱۷	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۰	۰	۰	۱۴	۰	۰	۰	۰
۱۸	۰	۰	۰	۰	۱	۰	۰	۰	۰	۰	۰	۰	۰	۱	۰	۰	۱۳	۰	۰	۰
۱۹	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۴	۰	۰
۲۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۳	۰
۲۶	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۵

۷-۳-۲ کلاسبند سطح دوم

در کلاسبند سطح دوم که در این بخش به بیان جزئیات طراحی آن می پردازیم، برای هر یک از مجموعه های غیر تهی $F(i)$ یعنی گویندگان مشابه گوینده i، از SVM به عنوان یک کلاس بندی کننده چند کلاسی برای تشخیص میان گویندگان در مجموعه $i \vee F(i)$ استفاده می نمائیم. برای تعمیم SVM از حالت باینری به حالت چند کلاسی از روش یکی در مقابل همه[۱۲] استفاده نموده ایم. در این روش برای هر گوینده یک مدل آموزش داده می شود. بنابراین تعداد مدل ها برابر با تعداد گویندگان در هر گروه می باشد. داده های آموزشی هر مدل SVM شامل ۲۵ ثانیه گفتار گوینده اصلی با برچسب 1+ و ۱۰ ثانیه گفتار هر یک از گویندگان دیگر با برچسب 1- می باشد. برای آموزش ماشین بردار های پشتیبان همانند مدل های GMM از بردار های ویژگی ضرائب کپستروم مبتنی بر آنالیز پیشگویی خطی همراه با مشتقات اول و دوم استفاده شده است. پارامترهای مهم و تاثیر گذار در دقت مدل های SVM، فاکتور پنالتی، ضریب پنالتی هر کلاس و نیز پارامتر مربوط به هسته مورد استفاده می باشند. تابع مورد استفاده بعنوان هسته مدل SVM نیز هسته گوسی انتخاب گردید. برای یافتن مقادیر بهینه مدل، گفتار یک مجموعه از گویندگان به عنوان مجموعه ارزیابی در نظر گرفته شد و پارامترهای مدل بر اساس تست مدل با گویندگان این مجموعه بدست آمد. معمولا پارامتر ها برای اکثر گروه ها یکسان بود. لذا نتایج حاصل از تعیین پارامترها برای دو یا سه گروه به سایر گروهها تعمیم داده شد. برای مثال برای بیشترگروهها فاکتور پنالتی ۴ یا ۱۰ انتخاب گردید. ضریب پنالتی برای گروه با برچسب 1+، مقدار ۱ و برای گروه با برچسب 1-، متناظر با نسبت داده های آموزشی گروه با برچسب 1- به داده های آموزشی گروه با برچسب 1+، یکی از مقادیر ۳ یا ۲ یا ۱ انتخاب گردید و پارامتر هسته گوسی (انحراف معیار) نیز به صورت تجربی ۰/۱۲۵ انتخاب شد.

۷-۳-۳ کلاسبندی گویندگان

در مرحله آزمایش ابتدا نمونه های تست گویندگان با کلاس بندی کننده های سطح اول (مدل های GMM) کلاس بندی می شوند. در صورتیکه گوینده تشخیص داده شده (i) دارای مجموعه غیر تهی F(i) باشد، کلاس بندی کننده های سطح دوم (ماشین های بردار پشتیبان) مربوط به این گوینده به کار گرفته می شود و نتیجه حاصل از تعیین هویت گوینده توسط آن به عنوان نتیجه نهایی پذیرفته می شود. با توجه به آزمایشات صورت گرفته، در حالتی که تنها از مدل های GMM برای تعیین هویت استفاده گردد میزان خطای تعیین هویت ۴/۱۵٪ می باشد. استفاده از ماتریس اشتباهات و نیز بکارگیری SVM برای تعیین گوینده اصلی از بین گویندگان در مجموعه گویندگان مشابه، موجب گردید که میزان خطا به ۱/۷٪ تقلیل یابد (جدول ۵). منظور از خطا* در این جدول خطاهایی است که امکان تصحیح توسط مدل های ماشین بردار پشتیبان را ندارند.

جدول ۵- میزان خطای تعیین هویت مدل های مخلوط گوسی, ماشین بردار پشتیبان و سیستم هیبرید بر حسب درصد

خطای کل	خطا	خطای SVM	خطای GMM
۱/۷	۰/۷۵	۰/۹۵	۴/۱۵

با دقت در نتایج آزمایشات در گروه های مختلف مجموعه گویندگان مشابه، نتایج زیر حاصل گردید:

الف- در بیشتر گروه ها استفاده از کلاسبند سطح دوم منجر به کاهش و یا صفر شدن خطای تعیین هویت می گردد. نکته مهم قابل توجه عدم هم پوشانی کامل نواحی خطای دو کلاس بند سطح اول و سطح دوم است که عامل بسیار موثر در بالا رفتن کارایی سیستم هیبرید می باشد. باید توجه داشت که با وجود شباهت گفتاری دو یا چند گوینده با یکدیگر لزوما نواحی خطای گویندگان در مدل های GMM و مدل های SVM یکسان نیست و لذا آن نمونه هایی که با مدل های GMM اشتباه کلاس بندی شده اند می توانند توسط مدل های SVM تصحیح

شوند، حتی اگر مدل های SVM نتوانند این گویندگان را بهتر از مدل های GMM مدل نمایند.

ب- در برخی گروه ها استفاده از کلاسبند سطح دم منجر به کاهش خطا نمیشود. در حالتی که نواحی خطای هر دو کلاس بندی کننده با یکدیگر تطابق داشته باشند، در این صورت باز هم خطا نسبت به قبل افزایشی نخواهد داشت. فقط در مواردی خطا پیش خواهد آمد که ناحیه خطای SVM در مورد سرگروه ها در ناحیه ای رخ دهد که مدل های GMM خطایی نداشته اند. در تعداد محدودی از گروه ها این مورد رخ داده است ولی به دلیل تصحیح خطا در سایر نمونه های گروه، خطای کل کاهش یافته یا تغییری نکرده است.

ج- در این سیستم تعیین هویت، سعی شده است که سرعت تعیین هویت نسبت به روش های مشابه دیگر بهینه باشد. در سیستم تعیین هویت پیاده سازی شده توسط آقایان فاین و گوپیناچ که در مقدمه نیز بدان اشاره شد[۱۱]، ابتدا نمونه های تست به مدل های GMM اعمال میشوند. بر اساس نتایج تعیین هویت بدست آمده، N گوینده (N=10) که بالاترین احتمال برنده شدن را دارند انتخاب گردیده سپس مدل های SVM به منظورکلاس بندی میان این گویندگان به کار گرفته میشوند. در روش پیشنهادی ما در بسیاری از موارد تعیین هویت با استفاده از کلاسبند سطح اول و بدون نیاز به کلاسبند سطح دوم صورت میگیرد. در مواردی هم که نیاز به استفاده از کلاسبند سطح دوم یعنی مدلهای SVM می باشد تعداد مدل های SVM ارزیابی شونده، برای تمامی نمونه های تست ثابت نبوده و در بازه کوچکی (۲ الی ۶ در آزمایشات صورت گرفته) قرار دارد. نکات فوق حاکی از سرعت بالاتر روش پیشنهادی در این مقاله (ضمن دارا بودن دقت بالا) در مقایسه با روش ارائه شده در مرجع [۱۱] می باشد. در سیستم تعیین هویت دیگری که توسط آقایان فاین و گوپیناچ به منظور بهبود کارایی سیستم قبلی پیاده سازی شد[۱۲]، در صورت بروز عدم قطعیت در نتیجه تعیین هویت توسط مدل های GMM، در مرحله بعد مدل های SVM مورد استفاده قرار می گیرند و نمونه تست با این مدلها مقایسه می شود. لیکن چنانچه این نمونه با مدل های SVM مربوط به تمامی گویندگان تست شود، زمان تعیین هویت بسیار بالا خواهد رفت. حتی اگر در این روش نیز مشابه روش N-Best، N مدل SVM مربوط به گوینده انتخابی بر اساس نتایج مدل های GMM استفاده و در ارزیابی دخالت داده شوند، زمان تعیین هویت همچنان طولانی خواهد بود. زیرا برای کاهش احتمال خطا نمیتوان اندازه مجموعه (N) را چندان کوچک در نظر گرفت. همچنین از آنجا که مدل های SVM مورد نیاز در زمان تعیین هویت مشخص می شوند، باید مدل های SVM مربوط به کلیه گویندگان از پیش ایجاد شوند. در مجموع در روش تعیین هویت پیاده سازی در این تحقیق، همراه با افزایش دقت تعیین هویت، سرعت نیز تا حد زیادی در مقایسه با روش های مشابه بالا رفته است.

۷-۴ گروه بندی گویندگان

عامل اصلی در افزایش زمان تعیین هویت، تعداد مقایسات بالا و استفاده از کلاس بندی کننده‌های پیچیده با سرعت پایین می‌باشد. روش پیشنهادی برای افزایش سرعت تعیین هویت، گروه بندی گویندگان به منظور کاهش تعداد مقایسات و استفاده از کلاس بندی کننده‌هایی با سرعت بالا می‌باشد. از این ایده در برخی از سیستم‌های تعیین هویت استفاده شده است. تفاوت روش‌ها در نحوه گروه بندی گویندگان است. در برخی از سیستم‌های تعیین هویت، جنسیت گویندگان ملاک گروه بندی قرار گرفته است. ایرادی که بر روش گروه بندی گویندگان بر اساس جنسیت وارد است، شباهت صدای برخی زنان به مردها و یا بالعکس می‌باشد، که می‌تواند باعث بروز خطا شود. در برخی سیستم‌های تعیین هویت معیار گروه بندی لهجه گویندگان، در نظر گرفته شده است. این روش نیز فقط در مواردی که بانک اطلاعاتی از تنوع لهجه برخوردار باشد، امکان پذیر است. در برخی از سیستم‌های تعیین هویت نیز از روش‌های خوشه بندی برای گروه بندی گویندگان استفاده شده است.

روشی که در این مقاله برای گروه بندی گویندگان پیاده سازی شده است، بر اساس شباهت‌های گفتاری گویندگان می‌باشد. معیار سنجش شباهت دو گوینده i و j، متوسط میزان درستنمایی حاصل از مدل مخلوط گوسی i به ازاء نمونه‌های گفتار گوینده j و میزان درستنمایی حاصل از مدل مخلوط گوسی گوینده j به ازای نمونه‌های گفتار گوینده i می‌باشد. الگوریتم گروه بندی گویندگان به صورت زیر است:

۱- ابتدا شباهت میان تمامی جفت گویندگان به روش فوق محاسبه می شود.

۲- دو گوینده که کمترین میزان شباهت را به یکدیگر دارند، به عنوان سر گروه های سطح اول انتخاب می‌شوند.

۳- سایر گویندگان بر اساس میزان شباهت، به هر یک از دو گروه تعلق پیدا می‌کنند. معیار سنجش میزان شباهت به هر یک از گروه‌ها، متوسط شباهت گوینده به اعضای گروه می‌باشد.

۴- پس از اینکه تمامی گویندگان گروه بندی شدند، گروه بندی مجدد انجام می شود. بدین معنی که میزان شباهت گویندگان به هر یک از دو گروه با اعضای جدید محاسبه شده و بر اساس میزان شباهت مجددا گروه بندی می شوند. برای محاسبه میزان شباهت گوینده به گروهی که قبلاً به آن تخصیص یافته، شباهت گوینده به خودش لحاظ نمی‌شود.

۵- مرحله قبل تا زمانی که وضعیت گروه‌ها تثبیت شود، یعنی گویندگان از گروهی به گروه دیگر جابجا نشوند، ادامه می‌یابد.

گروه بندی گویندگان بروش فوق انجام شد. گویندگان ۱۲ و ۵۳ به عنوان سر گروه‌های سطح اول انتخاب شده و سپس سایر گویندگان به دو گروه با سر گروهی گویندگان ۱۲ و ۵۳ تقسیم گردیدند. گروه ۱۲ شامل ۲۴ گوینده و گروه ۵۳ شامل ۲۶ گوینده است. هر یک از گروه‌های ۱۲ و ۵۳ توسط یک مدل مخلوط گوسی، با ۲۵۶ مخلوط مدل شده و با ۲۵ ثانیه گفتار به ازای هر یک از گویندگان عضو گروه آموزش داده شدند. برای ارزیابی کارایی گروه بندی، مدل‌های فوق با ۴۰ نمونه گفتار ۲ ثانیه‌ای به ازای هر یک از گویندگان تست شدند. در بین ۲۰۰۰ نمونه تست، ۱۸ مورد خطا رخ داد و سرعت متوسط تست هر مدل برابر۰/۲ ثانیه بود. یعنی به طور متوسط برای تعیین گروه در سطح اول بین گروه‌های ۱۲ و ۵۳ حدود ۰/۴ ثانیه زمان لازم است.

در مرحله بعد، گروه ۵۳ که شامل ۲۶ گوینده است، خود به دو گروه کوچکتر با روشی مشابه تقسیم گردید. سرگروه‌های انتخاب شده زیر گروه‌های گروه ۵۳، گویندگان ۱۴ و ۱۵ هستند.

زیرگروه با سر گروه ۱۵ شامل ۷ گوینده و زیرگروه با سر گروه ۱۴ شامل ۱۹ گوینده است. زیر گروه اول با یک مدل مخلوط گوسی با ۱۲۸ مخلوط و زیر گروه دوم با یک مدل مخلوط گوسی با ۲۵۶ مخلوط مدل شدند. برای آموزش مدل‌ها از ۲۵ ثانیه گفتار به ازای هر یک از گویندگان عضو گروه استفاده شد. مدل‌ها با ۴۰ نمونه گفتار ۲ ثانیه‌ای به ازای هر یک از گویندگان، تست شدند. خطای تعیین هویت ۷/۷٪ بدست آمد. سرعت متوسط تعیین گروه بین زیرگروه‌های فوق برای نمونه های تستی که در سطح قبل به گروه ۵۳، تعلق پیدا کرده‌اند، ۰/۲ ثانیه بود. همان طور که از نتایج آزمایشات مشاهده می‌شود، با افزایش تعداد گروه‌ها، میزان خطا نیز افزایش می‌یابد. در حالیکه به دلیل کم شدن تعداد مقایسات، زمان کل تعیین هویت کاهش می یابد. گروه ۱۲ را نیز به دو گروه کوچکتر تقسیم کردیم.

گوینده ۱۲ نسبت به سایر گویندگان، میزان خطای بالاتری را موجب می شود. از ۸۳ خطای مدل‌های مخلوط گوسی، ۱۴ مورد خطا مربوط به گوینده ۱۲ می‌باشد. پس از گوینده ۱۲، گوینده ۹ بالاترین میزان خطا را دارا می‌باشد. از بین ۸۳ نمونه خطای مدل‌های مخلوط گوسی، ۹ نمونه خطا مربوط به گوینده ۹ است که این گوینده نیز در گروه ۱۲ وجود دارد. زیرگروهی که شامل گویندگان ۹ و ۱۲

می‌باشد، نسبت به زیرگروه دیگر بسیار پر جمعیت‌تر است و دارای ۲۲ عضو است در حالی که زیرگروه دیگر فقط شامل دو عضو می‌باشد.

به دلیل تفاوت جمعیت بسیار زیاد میان این دو زیر گروه پیش‌بینی می‌شود خطای کلاس بندی میان این دو زیر گروه بسیار بالا باشد. زیرا مدل مربوط به زیر گروهی که دارای ۲۲ گوینده است نسبت به مدل زیر گروه دیگر خاصیت تعمیم بسیار بالاتری به دلیل تنوع بیشتر گویندگان خواهد داشت که همین امر باعث می‌شود، نمونه‌های گفتار مربوط به گویندگان زیر گروه دوم به مدل مربوط به زیر گروه اول که دارای جمعیت بیشتری است تمایل پیدا کنند. نتایج آزمایشات، این پیش‌بینی را تایید می‌نمایند. تقریباً نیمی از نمونه‌های تست مربوط به گویندگان زیر گروه با جمعیت کمتر، به اشتباه به زیر گروه با جمعیت بیشتر کلاس بندی شده‌اند.

در مرحله ارزیابی سیستم تعیین هویت، پس از تعیین گروه نمونه تست، قدم بعد تعیین هویت گوینده درون گروه می‌باشد. کلاس بندی میان دو گروه ۱۲ و ۵۳ باعث بروز ۰/۹٪ خطا می شود. بنابراین حد بالای راندمان سیستم تعیین هویت ۹۹/۱٪ خواهد بود. چنانچه نمونه تست مربوط به گروه ۵۳ باشد، از آنجا که این گروه خود به دو زیرگروه کوچکتر تجزیه شده است، کلاس بندی دیگری میان زیرگروه‌های ۱۵-۵۳ و ۱۴-۵۳ خواهیم داشت که خطای آن بنابر نتایج بدست آمده ۰/۷۵٪ می‌باشد. در جداول ۶ و۷ متوسط سرعت تعیین هویت گروه ها و نیز میزان خطای سیستم پیشنهادی بر اساس گروه بندی گویندگان آورده شده است.

زمان تعیین هویت در سیستم مبنا (مدل‌های مخلوط گوسی) برابر ۱/۴۷ ثانیه است. با دقت در نتایج مشاهده می‌شود که سرعت تعیین هویت در این روش نسبت به سیستم مبنا افزایش یافته (در بهترین حالت ۰/۷۵ ثانیه. در بدترین حالت ۱/۱۵ ثانیه) ولی دقت آن کاهش داشته است که موید این نکته است که پارامترهای دقت و سرعت تعیین هویت، در تقابل با یکدیگر هستند.

۷-۵ تعیین هویت گوینده مجموعه باز

سیستم تعیین هویت مطرح شده در بخشهای قبل را به سیستم تعیین هویت مجموعه باز تعمیم دادیم. در یک سیستم تعیین هویت مجموعه باز دو مرحله زیر صورت میگیرد:

الف) انتخاب گوینده ای در بین مجموعه گویندگان مجاز که گفتار تست با بیشترین احتمال، به مدل آن گوینده تعلق دارد.

ب) مقایسه میزان درستنمایی[۱۳] (لگاریتم شباهت) حاصل از مقایسه گفتار تست با مدل گوینده انتخاب شده با یک سطح آستانه به منظور تصمیم گیری درباره اینکه آیا گفتار ورودی متعلق به گوینده مدل انتخابی است یا مربوط به گوینده ای غیر از مجموعه گویندگان مرجع است.

جدول ۶- متوسط سرعت تعیین هویت گروه های گویندگان

	گروه با سرگروه ۱۴	گروه با سرگروه ۱۵	گروه با سرگروه ۱۲
تعداد مقایسات برای تعیین گروه	۴	۴	۲
تعداد مقایسات جهت تعیین هویت درون گروه	۱۹	۷	۲۴
متوسط زمان لازم برای تعیین گروه برحسب ثانیه	۰/۶	۰/۶	۰/۴
متوسط زمان لازم برای تعیین هویت بر حسب ثانیه	۰/۵۵	۰/۱۵	۰/۷
متوسط زمان کل جهت تعیین هویت بر حسب ثانیه	۱/۱۵	۰/۷۵	۱/۱

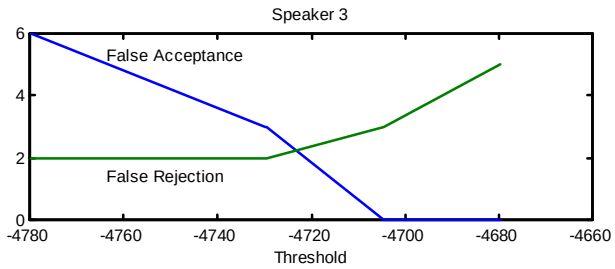
جدول ۷- میزان خطای سیستم و سیستم مبنا

میزان خطای سیستم مبنا (مدل های مخلوط گوسی)	٪ ۴/۱۵
میزان خطای سیستم پیشنهادی بر اساس گروه بندی گویندگان	٪ ۵/۴۵

همانطور که در مقدمه این مقاله گفته شد، در تعیین هویت مجموعه باز علاوه بر خطای کلاس بندی، دو نوع خطای پذیرش اشتباه[۱۴] و رد اشتباه[۱۵] نیز مطرح می باشند. خطای کلاس بندی مربوط به مرحله اول یعنی مرحله انتخاب گوینده و خطای پذیرش اشتباه و رد اشتباه مربوط به مرحله دوم می باشند. خطای رد اشتباه زمانی رخ می دهد که گفتار متعلق به گوینده انتخابی باشد ولی چون مقدار درستنمایی کمتر از سطح آستانه است، این گوینده جزء مجموعه گویندگان مجاز در نظر گرفته نشده و پذیرفته نشود. خطای پذیرش اشتباه زمانی رخ میدهد که گفتار ورودی متعلق به گوینده انتخابی نباشد ولی میزان درستنمایی بزرگتر یا مساوی سطح آستانه بوده و این گوینده بعنوان گوینده مجاز مورد پذیرش قرار گیرد. مقدار این دو خطا بستگی به سطح آستانه انتخاب شده دارد. معمولا سطح آستانه را به گونه ای در نظر می گیرند که این دو خطا با یکدیگر برابر شوند.

برای بررسی تاثیر روش استفاده برای تعیین سطح آستانه از گفتار ۲۵ گوینده از بانک اطلاعاتی فارس دات بزرگ خارج از مجموعه گویندگان مرجع استفاده گردید. مشابه با سایر ارزیابی های انجام شده، طول گفتار تست ۲ ثانیه در نظر گرفته شد. ابتدا گفتار تست به مدل های مخلوط گوسی گویندگان اعمال می شود. مدلی که بیشترین مقدار لگاریتم شباهت را داشته باشد، انتخاب می شود. در مرحله بعد مقدار درستنمایی با سطح آستانه مقایسه می شود.

برای انتخاب سطح آستانه دو روش در نظر گرفته شده است. در روش اول برای هر گوینده سطح آستانه به صورت مجزا تعیین می شود. در روش دوم تمامی گویندگان سطح آستانه مشترکی دارند. در روش اول برای تعیین مقدار مناسب سطح آستانه، کمترین مقدار لگاریتم شباهت مدل های مخلوط گوسی برای هر گوینده، به ازای داده های آموزشی تعیین می شود. سطح آستانه گویندگان با این مقادیر مینیمم، مقدار دهی اولیه می گردد. سپس میزان خطای پذیرش اشتباه و رد اشتباه به ازاء مقادیر مختلف سطح آستانه برای هر گوینده به صورت مجزا محاسبه می شود. برای هر گوینده سطح آستانه ای که به ازای آن، خطای پذیرش اشتباه و خطای رد اشتباه به تساوی برسند (نرخ خطای EER[۱۶])، به عنوان حد آستانه نهایی آن گوینده پذیرفته می شود. در شکل ۱ نمودار مربوط به خطای پذیرش اشتباه و رد اشتباه برای یکی از گویندگان به عنوان نمونه آورده شده است.



شکل ۱- نمودار خطای پذیرش اشتباه و رد اشتباه به ازای مقادیر مختلف سطح آستانه ای برای گوینده ۳

در نقطه تلاقی دو منحنی، مقادیر پذیرش اشتباه و رد اشتباه تقریبا برابر با ۲۳/۲٪ می باشد. در مورد گویندگانی که دارای خطای قبول اشتباه بالایی هستند، با دقت در ماتریس اشتباهات مشاهده گردید که این گویندگان عمدتا گویندگانی هستند که مجموعه گویندگان مشابه بزرگتری به نسبت سایر گویندگان دارند. این نکته باعث میشود که احتمال اینکه گویندگانی غیر از مجموعه مرجع به این گوینده شباهت گفتاری پیدا کنند، بالا باشد. بـرای اجتناب از خطای قبول اشتباه بالا ناچار به افزایش سطح آستانه می باشیم و بالا رفتن سطح آستانه برای دستیابی به نرخ خطای برابر EER خود باعث بالا رفتن خطای رد اشتباه می شود. در مقابل گویندگانی که خطای کلاس بندی بالایی دارند، یعنی مدل این گویندگان لگاریتم

شباهت نه چندان بالایی نسبت به سایر گویندگان تولید می کند، موجب پایین آمدن سطح آستانه می شود که این امر باعث بالا رفتن خطای پذیرش اشتباه می گردد.

جدول ۸- مقادیر خطای رد اشتباه، پذیرش اشتباه و میانگین حسابی دو خطا بر حسب درصد

رد اشتباه	۲۳/۵	۲۶/۵۵	۲۹/۹	۳۶/۶	۴۲/۵۵
پذیرش اشتباه	۳۲/۱	۲۶/۷	۲۱/۶	۱۴/۴	۷/۴
میانگین	۲۷/۸	۲۶/۶	۲۵/۷۵	۲۵/۵	۲۵

در روش دوم، سطح آستانه ای مشترکی برای تمامی گویندگان در نظر گرفته می شود. نحوه تعیین این مقدار مشابه روش اول می باشد. بدین ترتیب که به ازای مقادیر مختلف سطح آستانه ای خطای پذیرش اشتباه و رد اشتباه محاسبه می شود. مقداری که به ازای آن میزان دو خطا با یکدیگر برابر شوند به عنوان سطح آستانه نهایی پذیرفته می شود. انتظار می رود خطا نسبت به روش قبل بالاتر باشد که نتایج بدست آمده نیز این امر را تایید می کنند. در حالت تساوی دو خطا برابر ۲۶/۵٪ می باشد. در جدول ۹ مقادیر خطای پذیرش اشتباه و رد اشتباه و میانگین حسابی دو خطا به ازای برخی مقادیر مختلف سطح آستانه ای آورده شده است.

جدول ۹- مقادیر خطای رد اشتباه، پذیرش اشتباه و میانگین حسابی دو خطا بر حسب درصد در روش حدآستانه ای مشترک

رد اشتباه	۱/۶۵	۵/۹۵	۹/۶	۱۶/۴۵	۲۰/۴۵
پذیرش اشتباه	۵۹/۴	۵۱/۷۵	۴۴/۸۵	۳۴/۶۵	۲۷/۳
میانگین	۳۲/۵	۲۵/۸۵	۲۷/۲۲	۲۵/۵۵	۲۳/۸۷

۶-۷ هنجار سازی گروهی

تغییرات در ویژگی های گفتار نظیر نویز های محیطی، اثرات کانال های ارتباطی و یا هر گونه تفاوت در شرایط آموزش و تست می تواند تاثیر نامطلوب بر کارایی تعیین هویت گوینده داشته باشد. اثر این تغییرات به ویژه در سیستم های تعیین هویت گوینده مجموعه باز و در مرحله دوم تعیین هویت که مربوط به تصمیم گیری درباره پذیرش یا رد گوینده به عنوان یکی از گویندگان مرجع می باشد، بسیار بیشتر است.

علت این امر این است که در مرحله اول تعیین هویت که مدل گوینده با بالاترین لگاریتم شباهت انتخاب می شود، یک نمونه تست به تمامی مدل های گویندگان اعمال می شود و لذا تغییر شرایط تاثیر کم و بیش یکسانی در محاسبه لگاریتم شباهت نمونه تست با کلیه مدل ها می گذارد. ولی در دومین مرحله تعیین هویت که لگاریتم شباهت مدل انتخاب شده با یک سطح آستانه که در شرایطی متفاوت با شرایط تست محاسبه شده است، مقایسه می شود، تغییرات در شرایط می تواند اثر نامطلوبی بر نتیجه مقایسه داشته باشد. برای رفع این مشکل می توان از تکنیک های هنجارسازی استفاده نمود. از بین تکنیک های هنجارسازی، روش هنجارسازی مدل جهانی[۱۷] و روش هنجارسازی ماکزیمم[۱۸] به دلیل کارایی بالاتر نسبت به سایر روش ها انتخاب گردیده اند [۱۹,۱۸]. در ذیل به توضیح این روشهای هنجار سازی می پردازیم.

قانون تصمیم گیری در مرحله دوم تعیین هویت گوینده مجموعه باز به صورت زیر قابل بیان است:

$$P(l^{ML}|O) > P(l^{U}|O) \rightarrow O \in l^{ML} \quad else \quad O \in l^{U} \qquad (۲۸)$$

که $l^{ML} = \arg\max\{p(O|l_n)\}$ ، l^{U} مدل گوینـده مجهــول[۱۹] و O مجموعه بردار های ویژگی گویش تست می باشد. با اعمال تئوری بیز به رابطه فوق، خواهیم داشت:

$$\frac{p(O|l^{ML})}{p(O|l^{U})} > \frac{P(l^{U})}{P(l^{ML})} \rightarrow O \in l^{ML} \qquad else \quad O \in l^{U} \qquad (۲۹)$$

عبارت $l(O) = \frac{p(O|l^{ML})}{p(O|l^{U})}$ در این مرحله محاسبه می شود و سطح $\frac{P(l^{U})}{P(l^{ML})}$ آستانه رد یا قبول گوینده می باشد. در عمل بیشتر از فرم لگاریتمی استفاده می گردد.

$$L(O) = \log p(O|l^{ML}) - \log p(O|l^{U}) \qquad (۳۰)$$

محاسبه $p(O|l^{U})$ در عمل امکان پذیر نمی باشد. بنابراین تخمینی از آن جایگزین می گردد. در روش هنجارسازی مدل جهانی، $p(O|l^{U})$ با $p(O|l^{WM})$ جایگزین می شود که l^{ML} مدلی است که با گفتار تعداد زیادی از گویندگان آموزش داده شده است.

در این مقاله از مدل های مخلوط گوسی که متناظر با گروه های ۱۲ و ۵۳ در بخش گروه بندی گویندگان هستند و عملا کلیه گویندگان را شامل میشود ، به عنوان l^{WM} استفاده گردیده است. توانایی روش فوق وابسته به این است که شباهت حاصل از تعیین هویت نمونه های تستی که مربوط به گویندگان خارج از مجموعه گویندگان مرجع هستند، کوچکتر از شباهت حاصل از تعیین هویت نمونه های تست مربوط به گویندگان مرجع باشد. روال تعیین هویت به صورت زیر می باشد:

ابتدا نمونه تست با مدل های مخلوط گوسی گویندگان مرجع تعیین هویت شده و مدل با بزرگترین لگاریتم شباهت انتخاب می شود. سپس نمونه تست به مدل های مخلوط گوسی متناظر با گروه های ۱۲ و ۵۳ اعمال می شود. ماکزیمم لگاریتم شباهت خروجی این مدل ها از لگاریتم شباهت اولیه کسر می شود. سپس مقدار بدست آمده با سطح آستانه مربوط به گوینده مدل انتخاب شده از میان مدل های گویندگان مرجع، مقایسه می شود. برای دستیابی به کارایی بالاتر برای هر گوینده حدآستانه مجزایی در نظر گرفته شده است. مشابه روش های قبل سطح آستانه که به ازای آن مقادیر خطای پذیرش اشتباه و رد اشتباه برابر شوند، سطح آستانه نهایی پذیرفته می شود. میانگین خطای پذیرش اشتباه و رد اشتباه در حالت تساوی دو خطا از ۲۳/۵٪ بدون هنجارسازی به ۶/۷۵٪ با هنجارسازی به روش مدل جهانی کاهش یافته است. این نتیجه نشاندهنده تاثیر بالای هنجارسازی در کاهش خطای پذیرش و رد اشتباه در تعیین هویت مجموعه باز می باشد.

در روش هنجار سازی ماگزیمم $p(O|l^{U})$ در رابطه (۳۰) با $\underset{i \neq ML}{Max}\, p(O|l^{i})$ جایگزین می شود. به بیان دیگر پس از اعمال نمونه تست به مدل های گویندگان مرجع و انتخاب مدلی که بیشترین لگاریتم شباهت را دارا باشد به عنوان مدل برنده، مدلی که ماکزیمم لگاریتم شباهت را در بین سایر مدل‌ها داشته باشد، انتخاب می شود و مقدار لگاریتم شباهت مدل انتخابی از لگاریتم شباهت مدل برنده کسر شده و حاصل تفریق با حدآستانه مربوط به مدل برنده مقایسه می شود. در جدول ۱۰ مقادیر خطای پذیرش اشتباه و رد اشتباه به ازای ۴ مقدار متفاوت سطح آستانه و میانگین دو خطا آورده شده است.

با توجه به این جدول متوسط خطای پذیرش اشتباه و رد اشتباه در حالت تساوی دو خطا در حدود ۴/۲ ٪ است که نسبت به روش مدل جهانی بهبود بیشتری را در کاهش میزان خطا نشان می دهد.

جدول ۱۰- مقادیر خطای رد اشتباه، پذیرش اشتباه و میانگین حسابی دو خطا در روش هنجارسازی ماکزیمم بر حسب درصد

رد اشتباه	۳	۳/۵۵	۴/۱۵	۴/۱
پذیرش اشتباه	۷/۹۵	۶/۳	۴/۵	۴/۳
میانگین	۵/۴۷۵	۴/۹۲	۴/۳۲۵	۴/۲

۸- نتیجه گیری

در این مقاله به بررسی و پیاده سازی روش های افزایش دقت و سرعت تعیین هویت گوینده مستقل از متن و مجموعه باز پرداخته شد و روش ها، نحوه پیاده سازی، آزمایشات انجام شده و نیز نتایج حاصل از آنها ارائه گردید. کارایی ضرائب LPCC و تاثیر مشتقات اول و دوم آنها بر کارایی تعیین هویت بررسی شدند. بنابر نتایج بدست آمده ضرائب LPCC به همراه مشتقات اول و دوم این ضرائب در شرایطی که داده های گفتاری از SNR بالایی برخوردار بوده و دارای پهنای باند کافی باشند از کارایی بسیار خوبی در تعیین هویت گوینده برخوردار می باشند. بررسی ارتباط حجم داده آموزشی و تعداد مخلوط های گوسی با کارایی تعیین هویت و زمان آموزش نشان داد که در صورت کافی بودن حجم داده های آموزشی، میتوان با افزایش تعداد مخلوط های گوسی به کارائی بیشتری در تعیین هویت دست یافت. لیکن باید توجه داشت که افزایش دادههای آموزشی و تعداد مخلوط های گوسی بر زمان آموزش و زمان آزمایش می افزایند. روشهایی برای افزایش دقت و سرعت تعیین هویت ارائه گردید که هدف اصلی آنها ایجاد یک سیستم هیبرید متشکل از SVM و مدل های GMM به منظور افزایش دقت و سرعت در تعیین هویت گوینده بوده است. مهم ترین ویژگی روش استفاده شده یعنی ترکیب توانایی ها و مزایای مدل های GMM و SVM و بهره گیری از عدم هم پوشانی کامل نواحی خطای این دو روش به منظور افزایش کارایی در تعیین هویت با توجه به نتایج آزمایشات به خوبی ملاحظه گردید. در سیستم پیاده سازی شده، مدل های SVM فقط در مواردی که مدل‌های GMM در شناسایی گوینده اصلی ناتوان باشند یا ضعیف عمل کنند، استفاده می‌شوند. البته این کار باعث افزایش چندانی در پیچیدگی سیستم نمی شود چرا که مدل های SVM فقط برای گویندگان موجود در گروه های گویندگان مشابه ساخته می شوند. ویژگی مهم روش پیشنهادی این مقاله در ترکیب مدل های GMM و SVM در مقایسه با روش های مطرح شده توسط دیگران آن است که روش پیشنهادی ضمن برخورداری از کارایی بالا، از سرعت بیشتری نیز برخوردار می باشد. نتایج بدست آمده بهبود کارایی سیستم هیبرید را نسبت به سیستم مبنا که در آن فقط از روش GMM استفاده شده باشد، بدون افزایش میزان داده آموزش و تست، نشان می دهد. میزان خطای تعیین هویت سیستم هیبرید برابر ۱/۷٪ بدست آمد، در حالیکه در سیستم مبنا این خطا در حدود ۴/۱۵٪ بود. گروه‌بندی گویندگان به منظور افزایش سرعت تعیین هویت نیز مورد بررسی قرار گرفت. گویندگان بر اساس میزان شباهتی که بین آنها وجود دارد به چند گروه تقسیم میشوند. گروه بندی گویندگان سرعت تعیین هویت را بهبود می بخشد ولیکن موجب افزایش خطای تعیین هویت می‌گردد. به عنوان مثال در آزمایشات انجام شده سرعت تعیین هویت از ۱/۴۷ ثانیه در سیستم مبنا به ۰/۷۵، ۱/۱ و ۱/۱۵ ثانیه در گروه های مختلف گویندگان کاهش یافت، لیکن میزان خطا از ۴/۱۵٪ در سیستم مبنا به ۵/۴۵٪ افزایش پیدا کرد.

دو روش متفاوت برای محاسبه سطح آستانه در تعیین هویت گوینده مجموعه باز مورد بررسی قرار گرفت. در روش اول برای کلیه گویندگان سطح آستانه‌ای مشترکی در نظر گرفته شد. متوسط خطای پذیرش اشتباه و رد اشتباه در حالت تساوی دو خطا ۲۶/۵٪ بدست آمد. در روش دوم برای هر یک از گویندگان مرجع سطح آستانه مجزایی در نظر گرفته شد. متوسط خطای پذیرش اشتباه و رد اشتباه در این روش به ۲۳/۲ ٪ کاهش یافت. هنجار سازی گروهی موجب بهبود قابل توجهی در کاهش متوسط خطای پذیرش اشتباه و رد اشتباه در تعیین هویت مجموعه باز میگردد. تکنیک های هنجارسازی مدل جهانی و ماکزیمم به منظور افزایش کارایی به کار گرفته شدند. متوسط خطای پذیرش اشتباه و رد اشتباه در روش مدل گسترده و ماکزیمم به ترتیب برابر ۶/۷۵٪ و ۴/۲٪ بدست آمد که نشان دهنده تاثیر چشمگیر هنجارسازی مخصوصاً هنجارسازی به روش ماکزیمم بر کاهش متوسط خطای پذیرش اشتباه و رد اشتباه می‌باشد.

قدردانی

این مقاله حاصل از پروژه تحقیقاتی طرح ماده ۱۰۲ به شماره ۱۵۱۷ می باشد که از طرف سازمان مدیریت و برنامه ریزی و وزارت فناوری اطلاعات و ارتباطات حمایت گردیده است.

مراجع

[1] J. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462 1997.

[2] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," *Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[3] J. D. Markel and S. B. Davis, "Text Independent Speaker Recognition from a Large Linguistically Unconstrained Time Spaced Data Base," *IEEE Transaction on ASSP*, vol. 27, no. 1, pp. 74–82, 1979.

[4] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Transaction on ASSP*, vol. 29, pp. 254–272, 1981.

[5] Y. Bennani, "Probabilistic Cooperation of Connectionist Expert Modules: Validation on a Speaker Identification Task," *ICASSP*, pp. 541-544, 1993.

[6] Y. Bennani, "Text-Independent Talker Identification System Combining Connectionist and Conventional Models," IEEE *Workshop on Neural Networks for Signal Processing, IEEE Service Center Press*, pp. 131-138, 1992.

[7] Y. Bennani and P. Gallinari, "On the Use of TDNN-Extracted Features Information In Talker Identification," *ICASSP*, pp. 385-388, 1991.

[8] Y. Bennani and P. Gallinari, "A Connectionist Approach for Speaker Identification," *ICASSP*, pp. 265-268, 990.

[9] L. Wang, K. Chen and H. S. Chi, "Towards Better Capturing Inter-Speaker Information by Active Learning for Speaker Identification," *IJCNN*, vol.4, pp. 2975-2980, 2001.

[10] م. ش. معین و ر. بوستانی، " مقایسه روشهای HMM, GMM و SVM به منظور بررسی هویت گوینده"، **یازدهمین کنفرانس مهندسی برق ایران**،۱۳۸۲، ص ۲۸۶-۲۹۳.

[11] S. Fine, J. Navratil and R. Gopinatth, "A Hybrid GMM/SVM Approach to Speaker Identification," *ICASSP*, 2001.

[12] S. Fine, J. Navratil and R. A. Gopinatth, "Enhancing GMM Scores Using SVM 'Hints," *EUROSPEECH*, 2001.

[13] D.A. Reynolds, *Gaussian Mixture Modeling Approach to Text Independent Speaker Identification, Ph.D.* Thesis, Georgia Institute of Technology, 1992.

[14] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.

[15] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[16] م. م. همایونپور و ا. شریف نبوی، "مقایسه و ارزیابی روشهای تشخیص گفتار از سکوت"، **اولین کنفرانس فناوری اطلاعات و دانش**، ص ۶۲۹-۶۳۹، ۱۳۸۲.

[17] ا.شریف نبوی، **تعیین هویت گوینده در شرایط غیر متعارف، ارزیابی روشها و ارائه دستورالعمل های مناسب**، تز کارشناسی ارشد، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، ۱۳۸۲.

[18] A. E. Rosenberg et al., "The use of Coort Normalization Scores for Speaker Verification," *ICSLP*, pp. 599-602, 1992.

[19] P. Sivakumaran, J. Fortuna, A. M. Ariyaeenia, "Score Normalization Applied to Open-set, Text Independent Speaker Identification," *EUROSPEECH*, pp. 2669-2672, 2003.

[1] Gaussian Mixture Model (GMM)
[2] Gaussian Mixture Model (GMM)
[3] Support Vector Machine (SVM)
[4] Closed Set
[5] Open Set
[6] Convex Quadratic Constrained Optimization
[7] Misclassification
[8] Mel Frequency Cepstral Coefficient (MFCC)
[9] Linear Predictive Cepstral Coefficient (LPCC)
[10] Cepstral Mean Subtraction (CMS)
[11] Confusion matrix
[12] One Against All
[13] Likelihood
[14] False Acceptance
[15] False Rejection
[16] Equal Error Rate (EER)
[17] World Model Normalization (WMN)
[18] Maximum Normalization
[19] Unknown Speaker

محمد مهدی همایونپور در سال ۱۳۳۹ در شهر شیراز متولد شد. تحصیلات تا مقطع دیپلم را در شهر شیراز سپری و دیپلم متوسطه خود را در سال ۱۳۵۸ دریافت کرد. وی تحصیلات خود در مقطع کارشناسی را در رشته مهندسی برق در دانشگاه صنعتی امیرکبیر (سال ۱۳۶۶)، کارشناسی ارشد را در رشته برق (مخابرات)، از دانشگاه خواجه نصیرالدین طوسی (سال ۱۳۶۹)، کارشناسی ارشد دوم خود را در زمینه فونتیک (۱۳۷۴) در دانشگاه سوربون جدید در فرانسه و همزمان دورهٔ دکترای خود را در دانشگاه پاریس ۱۱ در زمینه مهندسی برق (۱۳۷۴) بپایان رسانید. نامبرده از سال ۱۳۷۴ در سمت عضو هیأت علمی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیر کبیر به تدریس و تحقیق مشغول می‌باشد. ایشان علاوه بر تدریس، راهنمایی پروژه های کارشناسی، کارشناسی ارشد و دکتری در زمینه های مهندسی کامپیوتر و فناوری اطلاعات و نیز هدایت تعداد زیادی پروژه های صنعتی و ملی را برعهده داشته است. نامبرده عضو انجمن های علمی کامپیوتر، ارتباطات و فناوری اطلاعات و رمز می باشد و مسئولیت های اجرایی متعدد از جمله ریاست و معاونت های آموزشی و پژوهشی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیر کبیر و شرکت در برگزاری چندین کنفرانس و مسابقه علمی را بر عهده داشته و موفق به انتشار بیش از ۸۰ مقاله علمی- پژوهشی در مجلات و کنفرانس های علمی داخل و خارج از کشور گردیده است.

آدرس پست الکترونیکی ایشان عبارت است از:
homayoun@ce.aut.ac.ir

و آدرس سایت اینترنتی:
www.autice.org

هدیه رزازان در سال ۱۳۵۸ در شهرستان بجنورد متولد شد. تحصیلات خود را تا مقطع دیپلم در بجنورد سپری و دیپلم متوسطه خود را در رشته ریاضی و فیزیک در سال ۱۳۷۶ دریافت نمود. وی سپس مدرک کارشناسی در رشته مهندسی کامپیوتر گرایش نرم افزار از دانشگاه فردوسی مشهد و کارشناسی ارشد در رشته مهندسی کامپیوتر، گرایش هوش مصنوعی از دانشگاه صنعتی امیرکبیر دریافت کرد. وی تجربیات کاری خود را از سال ۱۳۸۱ در مرکز تحقیقات مخابرات ایران و سپس شرکت عصر دانش افزار آغاز نمود. وی در حال حاضر در سمت مشاور IT در شرکت عصر دانش افزار مشغول به کار می‌باشد. زمینه های تحقیقاتی مورد علاقه ایشان عبارتند از: پردازش سیگنال گفتار، تحلیل و طراحی سیستم و تحلیل و طراحی نرم افزار می باشد.

آدرس پست الکترونیکی ایشان عبارت است از:
h_razazan@hotmail.com