سال دوم، شماره‌های ۵ و ٦ ، پاییز و زمستان ۱۳۸۹
صص: ۸- ۱

فصلنامه علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

بهبود دقت مدل GMM در قالب سیستم GMM-VSM در کاربرد تشخیص زبان گفتاری

فهیمه قاسمیان ▪* **محمدمهدی همایون‌پور** **

* کارشناس ارشد، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر

** دانشیار، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر

تاریخ دریافت: ۱۳۸۹/۰۱/۲۰ تاریخ پذیرش: ۱۳۸۹/۰۵/۲۵

چکیده

مدل GMM یکی از پرکاربردترین و موفق‌ترین مدل‌ها در زمینه تشخیص خودکار زبان است. در این مقاله مدلی جدید با نام Adapted Weight-GMM(AW-GMM) ارائه شده است. این مدل مشابه GMM است با این تفاوت که وزن مولفه‌های آن در قالب سیستم GMM-VSM بر اساس قدرت مولفه‌ها در تمایز یک زبان از سایر زبان‌ها تعیین می‌گردد. همچنین با توجه به پیچیدگی محاسباتی که در سیستم GMM-VSM در حالتی که توالی ۲تایی مولفه‌ها در نظر گرفته شود، وجود دارد، تکنیکی برای ساخت توالی ۲تایی مولفه‌ها ارائه شده است که می‌توان از آن برای ساخت توالی‌های از مرتبه بالاتر نیز استفاده نمود. ارزیابی‌های صورت گرفته بر روی ٤ زبان انگلیسی، فارسی، فرانسوی و آلمانی از دادگان OGI کارایی تکنیک‌های ارائه شده را نشان می‌دهد.

کلید واژگان: مدل مخلوط گاوسی (GMM)، بردار BOS، ماشین بردار پشتیبان (SVM)، تشخیص زبان.

۱- مقدمه

به دلیل نیاز روزافزون به برقراری ارتباط انسان و رایانه و گرایش به سمت برقراری ارتباط‌های طبیعی‌تر با ماشین، تحقیقات زیادی در زمینه طراحی و پیاده‌سازی سیستم‌هایی با قابلیت پردازش گفتار طبیعی صورت گرفته است. تشخیص خودکار زبان (LID[1]) جزء این دسته از سیستم‌هاست که با استفاده از آن، رایانه زبان مربوط به گفتار دیجیتال شده را تشخیص می‌دهد. از جمله کاربردهای سیستم‌های تشخیص زبان می‌توان به هدایت تماس‌های ضروری، سرویس‌های چند زبانه، سرویس‌های نظامی، کاربردهای امنیتی، اندیس‌گذاری فایل‌های صوتی و غیره اشاره کرد [1].

محققان با الهام از معیارهایی که شنوندگان انسانی جهت تمایز میان زبان‌ها مورد استفاده قرار می‌دهند، توانسته‌اند به موفقیت‌های قابل توجهی در زمینه تشخیص خودکار زبان گفتاری دست پیدا کنند. در سال ۱۹۹٦ آقای زیسمن[2]، مقاله‌ای را منتشر کرد که در آن ٤ روش پایه برای تشخیص زبان را به طور کامل شرح داد و از نظر کارایی با یکدیگر مقایسه نمود. این روش‌ها شامل روش‌های PRLM[3]، PPRLM[4]، PPR و GMM بود[2]. کارهای بعدی که پس از انتشار این مقاله صورت گرفت، معمولا کارایی خود را با سیستم‌های تشخیص زبانی که در این مقاله ارائه شده، مقایسه نمودند و سعی کردند که این روش‌ها را بهبود بخشند، همچنین مسابقاتی که جهت ارزیابی سیستم‌های تشخیص زبان از سال ۱۹۹٦ تا کنون هر ۲ سال یکبار توسط NIST تحت عنوان NIST-LRE برگزار می‌شود، بستری برای بهبود سیستم‌های تشخیص زبان بود. طبق نتایج ارائه شده توسط این موسسه، کارایی سیستم‌های تشخیص زبان در هر دوره بهبود پیدا می‌کند.

▪ نویسنده عهده‌دار مکاتبات (f_ghasemian@yahoo.com)

[1] Language Identification

[2] Zissman

[3] Phone Recognizer Followed by Language Model

[4] Parallel Phone Recognizers Followed by Language Model

بهبود سیستم‌های تشخیص زبان، از جنبه‌های مختلفی صورت گرفته است. این بهبودها را می‌توان از نقطه‌نظر ویژگی-های اکوستیکی و نوایی مورد استفاده برای تشخیص زبان، شناساگرهای آوایی و مدل‌های زبانی مورد استفاده، نحوه‌ی ترکیب نتایج حاصل از مدل‌های زبانی، استفاده از طبقه‌بندی-کننده‌های تمایزی و ترکیب سیستم‌های تشخیص زبان مختلف مورد بررسی قرار داد. اکثر سیستم‌های تشخیص زبان موفق نیازی به دانش سطح بالا برای تمایز زبان‌ها ندارند، بلکه از اطلاعات اکوستیکی، نوایی و واج‌آرایی (قوانین حاکم بر توالی مجاز واج‌ها) برای تمایز میان زبان‌ها استفاده می‌کنند. گرچه واج‌ها به طور قابل ملاحظه‌ای میان زبان‌ها مشترکند اما فرکانس تکرار این واج‌ها و توالی چندتایی آن‌ها می‌تواند به طور قابل ملاحظه‌ای از یک زبان به زبان دیگر متفاوت باشد [1].

پس از ارائه سیستم تشخیص زبان PPRLM[2] و موفقیت این سیستم در تشخیص زبان، تحقیقات بیشتری در زمینه اطلاعات واج‌آرایی صورت گرفت. آزمایشات صورت گرفته بر روی شنوندگان انسانی نشان داده است که شنوندگان چندزبانه قدرت بالاتری در تشخیص زبان‌ها نسبت به شنوندگان تک زبانه دارند. PPR که در قسمت ابتدایی این سیستم قرار دارد، از مجموعه‌ی موازی از شناساگرهای آوایی تشکیل شده است و روشی موثر در تبدیل قطعات گفتار ورودی به توالی‌های آوایی است.

سیستم‌های تشخیص زبان نظیر PPRLM و PPR-VSM[1][3] که جز موفق‌ترین سیستم‌های تشخیص زبان هستند، برای آموزش نیاز به داده‌هایی دارند که در سطح واج برچسب-گذاری شده باشند. برچسب‌گذاری عملی وقت‌گیر است و تعمیم این سیستم‌ها به تعداد زبان‌های بالاتر را دشوار می‌سازد. برای حل این مشکل روشی ارائه شده است که در آن از مجموعه‌ای از شناساگرهای GMM جهت تشخیص زبان استفاده می‌شود. در این روش نیازی به داده‌های برچسب‌خورده وجود ندارد اما دقت پایین‌تری در تبدیل قطعه گفتاری به توالی آوایی دارد [4]. به این ترتیب می‌توان از مدل GMM به عنوان قسمت ابتدایی سیستم‌های PPRLM و PPR-VSM استفاده نمود که به ترتیب سیستم‌های GMM-LM و GMM-VSM نامگذاری شده‌اند.

با بررسی سیستم‌های تشخیص زبان ارائه شده، مشاهده می‌شود که در اکثر این سیستم‌ها از مدل GMM استفاده می‌شود و این مدل در کاربرد تشخیص زبان خوب عمل کرده است. بنابراین افزایش دقت مدل GMM می‌تواند در افزایش دقت سیستم‌های تشخیص زبانی که از این مدل استفاده می‌کنند، تاثیر بسزایی داشته باشد. هر مدل GMM با استفاده از روش بیشینه-سازی امید ریاضی (EM[2])، به طور مستقل از سایر داده‌ها و با استفاده از داده‌های آموزشی مربوط به آن زبان، آموزش داده می‌شود و وزن هر مولفه در مدل، متناسب با فرکانس تکرار آن در مجموعه داده‌های آموزشی تعیین می‌شود. در این مقاله مدلی جدید با نام AW-GMM ارائه شده است که مشابه GMM با استفاده از الگوریتم EM آموزش داده می‌شود اما پس از آموزش در قالب سیستم GMM-VSM عمل تطبیق وزن صورت می‌گیرد. در فاز تطبیق وزن، وزن هر مولفه از GMM بر اساس قدرت آن در تمایز زبان مربوطه از سایر زبان‌ها تعیین می‌شود.

در این مقاله همچنین تکنیکی جهت کاهش پیچیدگی محاسباتی سیستم GMM-VSM ارائه شده است که علاوه بر کاهش پیچیدگی سبب افزایش دقت این سیستم شده است.

در ادامه در بخش ۲، سیستم‌های تشخیص زبان گفتاری بر مبنای GMMو سیستم GMM-VSM شرح داده شده است. در بخش ۳ به بیان تکنیک‌های ارائه شده جهت بهبود دقت GMM و سیستم GMM-VSM پرداخته شده است. نتایج آزمایش‌ها و ارزیابی‌های صورت گرفته نیز در بخش ٤ بیان گردیده و در خاتمه در بخش ۵ به جمع‌بندی و نتیجه‌گیری پرداخته شده است.

۲- معرفی سیستم‌های تشخیص زبان گفتاری

در تشخیص خودکار زبان‌ها، یافتن مشخصه‌های موثر برای جداسازی زبان‌ها از اهمیت بالایی برخوردار است [5]. انسان و ماشین می‌توانند معیارهای مختلفی از جمله نوع آواها، فرکانس تکرار آن‌ها، توالی‌های آوایی، اطلاعات نوایی و غیره را جهت تشخیص زبان مورد استفاده قرار دهند. در سیستم‌هایی که از اطلاعات آوایی جهت تشخیص زبان استفاده می‌شود، برای استخراج دنباله آوایی متناظر با یک قطعه گفتاری از یک سری ویژگی‌های اکوستیکی استفاده می‌شود. این ویژگی‌ها معمولا ویژگی MFCC[3] یا SDC[4] در نظر گرفته می‌شود.

[1] Parallel Phone Recognizer-Vector Space Modeling

[2] Expectation Maximization

[3] Mel Frequency Cepstral Coefficient

[4] Shifted Delta Cepstral

۲-۱-استخراج ویژگی

فاز استخراج ویژگی در همه سیستم‌های تشخیص زبان مشترک است. برای استخراج ویژگی ابتدا سیگنال ورودی به فریم‌های با طول ثابت تقسیم شده و سپس از هر فریم ویژگی اکوستیکی (معمولا ضرایب MFCC) یا پروزودی مورد نظراستخراج می‌شود. به این ترتیب قطعات گفتاری به توالی از بردارهای ویژگی با طول ثابت تبدیل می‌شوند. همچنین از روش CMS[1] برای حذف اثر کانال نیز استفاده می‌شود.

۲-۲-سیستم تشخیص زبان بر مبنای مدل مخلوط گاوسی (GMM[2])

این روش یک روش ساده و آماری برای تشخیص زبان است و سعی بر جداسازی زبان‌ها بر مبنای تفاوت‌هایی که میان آواها و فرکانس تکرار آن‌هاست، دارد. GMM فرض می‌کند که هر بردار ویژگی V_t که مربوط به فریم زمانی t است، توسط یک تابع چگالی احتمالاتی تولید می‌شود. به عبارت دیگر بردارهای ویژگی (ویژگی MFCC) استخراج شده از فریم‌ها، دارای یک تابع توزیع احتمالاتی است که این تابع توزیع را می‌توان به صورت مجموع توابع توزیع نرمال چند متغیره[3] به صورت تساوی ۱ نوشت [2].

$$p(V_t|\lambda_t) = \sum_{k=1}^{K} p_k b_k(V_t) \qquad (۱)$$

که در آن λ مجموعه پارامترهای مدل است.

$$\lambda = \{p_k, \mu_k, \sum_k\} \qquad (۲)$$

k نشان‌دهنده شماره مخلوط، p_k وزن مخلوط با محدودیت $\sum_{k=1}^{N} p_k = 1$ و b_k تابع توزیع گاوسی است که با استفاده از دو پارامتر μ_k و $\sum_k$ مشخص می‌شود [2].

به ازای هر زبان هدف یک مدل GMM با استفاده از الگوریتم EM و داده‌های آموزشی مربوط به آن زبان، آموزش داده می‌شود. در طول فاز تشخیص نیز ابتدا بردارهای ویژگی از سیگنال ورودی استخراج شده و سپس لگاریتم احتمال تعلق قطعه گفتاری به تک تک مدل‌های زبانی با استفاده از رابطه‌ی ۳ محاسبه می‌شود.

$$L(\{x_t\}|\lambda_l)) = \sum_{t=1}^{T} [\log p(x_t|\lambda_l)] \qquad (۳)$$

که در آن λ_l مدل GMM زبان l، $\{x_t\}$ توالی بردارهای ویژگی قطعه گفتاری ورودی و T طول زمانی آن است. در نهایت زبان مربوط به قطعه‌ی گفتاری با استفاده از تساوی ٤ تعیین می‌شود [2].

$$\hat{l} = \arg\max_{l} L(\{x_t\}|\lambda_l)) \qquad (٤)$$

۲-۳-مدل GMM به عنوان شناساگر آوا

یکی از مشکلاتی که در سیستم‌های PPRLM و PPR-VSM وجود دارد، نیاز به داده‌های آموزشی است که در سطح آوا برچسب خورده باشند. برای رفع این مشکل از مدل‌های GMM به عنوان شناساگرهای آوایی استفاده می‌شود[4]. این شناساگرها دقت پایین‌تری نسبت به شناساگرهایی که در PPRLM و PPR-VSM استفاده می‌شوند دارند، اما مزیت آن‌ها این است که به راحتی می‌توان تعداد شناساگرها را افزایش داد، همچنین از نظر محاسباتی سریع‌تر هستند.

روش کار به این صورت است که به تعداد زبان‌های مورد نظر مدل GMM آموزش داده می‌شود. هر مدل GMM فضای آوایی زبان مربوطه را مدل می‌کند و در صورتی که هر مولفه یا کرنل گاوسی از GMM، به عنوان مدل‌کننده یک یا بخشی از یک آوا در نظر گرفته شود، شماره مولفه‌ای که به ازای بردار ویژگی ورودی، بالاترین میزان احتمال را از میان مولفه‌های دیگر GMM برگرداند، می‌تواند به عنوان آوای معادل بردار ویژگی ورودی در نظر گرفته شود. به این ترتیب به ازای هر قطعه گفتار ورودی، دنباله‌ای از شماره‌ها (از ۱ تا تعداد مولفه-های GMM) به عنوان خروجی هر شناساگر برگردانده می‌شود. برای مثال در صورتی که یک قطعه گفتاری که پس از استخراج ویژگی از ۱۰ بردار ویژگی متوالی ($F_1, F_2, \ldots, F_{10}$) تشکیل شده باشد را به یک شناساگر GMM با ۲۰ مولفه بدهیم، دنباله-ی آوایی با طول ۱۰ را به عنوان خروجی برمی‌گرداند، به عنوان نمونه ۱۲، ۱، ۵، ٦، ١٤، ۱۳، ۱۹، ۱۹، ۱۱ می‌تواند این دنباله‌ی خروجی باشد.

۲-٤-سیستم تشخیص زبان GMM-VSM

سیستم GMM-VSM از F شناساگر آوایی تشکیل شده است که هر کدام مجموعه‌ی آوای v_f را در بردارد. هر شناساگر به ازای قطعه گفتار ورودی، توالی از آواها را به عنوان خروجی

[1] Cepstral Mean Subtraction
[2] Gaussian Mixture Model
[3] Multivariate Gaussian Densities

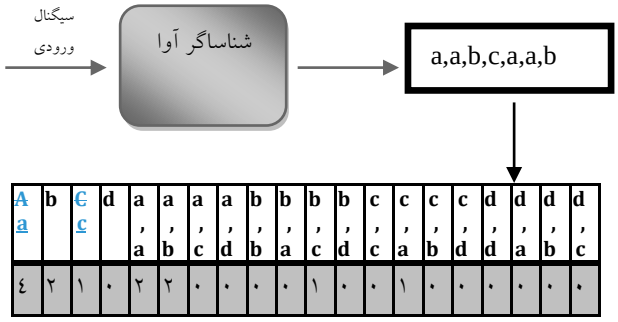
برمی‌گرداند. این توالی آواها به یک بردار با اندازه ثابت تبدیل شده که به این بردار(Bag-of-Sound (BOS گفته می‌شود که بر اساس فرکانس تکرار مجموعه آواها و توالی‌های آن‌ها ساخته می‌شود. بعد این بردار به ۲ پارامتر وابسته است: تعداد آواهای شناساگر و مرتبه (n) در نظر گرفته شده. به عنوان مثال در صورتی که تعداد آواهای یک شناساگر ۱۰ و مرتبه در نظر گرفته شده ۲ باشد، بعد بردار BOS برابر خواهد بود با:

$$10 + (10 \times 10) \quad (۵)$$

به همین ترتیب در صورتی که با همین تعداد آوا مرتبه برابر با ۳ در نظر گرفته شود بعد برابر است با

$$10 + (10 \times 10) + (10 \times 10 \times 10) \quad (۶)$$

به عبارت دیگر در حالتی که مرتبه ۲ در نظر گرفته شود فرکانس تکرار هر آوا و همچنین فرکانس تکرار توالی‌های ۲ تایی آن‌ها در توالی آوایی تولید شده توسط شناساگر برای ساخت بردار BOS در نظر گرفته می‌شود و در حالتی که مرتبه ۳ در نظر گرفته شود، فرکانس تکرار هر آوا، توالی ۲تایی و ۳تایی آواها در ساخت بردار در نظر گرفته می‌شود و به این ترتیب می‌توان بردار BOS را برای مرتبه‌های بالاتر نیز تولید کرد.در شکل ۱ نحوه ساخت بردار BOS برای توالی آوایی تولید شده توسط یک شناساگر که مجموعه آوای آن از ۳ آوای a، b، c، d تشکیل شده و مرتبه BOS برابر با ۲ در نظر گرفته شده، نشان داده شده است.



Aa	b	Cc	d	a,a	a,b	a,c	a,d	b,b	b,a	b,c	b,d	c,c	c,a	c,b	c,d	d,d	d,a	d,b	d,c
٤	٢	١	٠	٢	٢	٠	٠	٠	٠	١	٠	٠	١	٠	٠	٠	٠	٠	٠

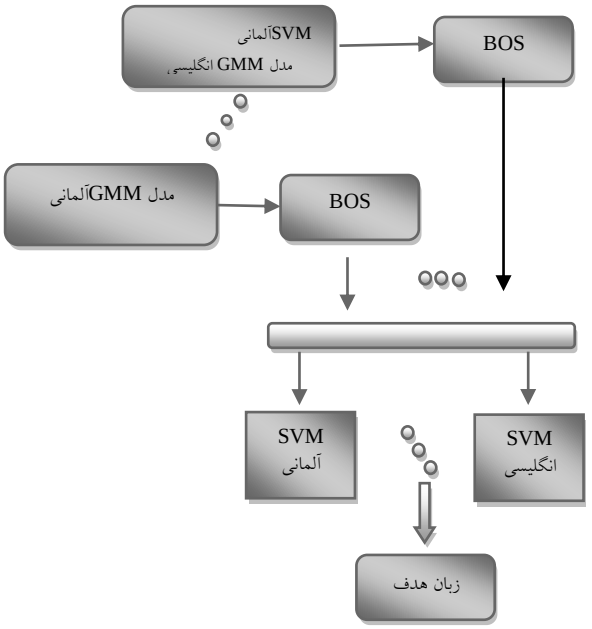
شکل ۱–نحوه تبدیل توالی آوایی به بردار BOS از مرتبه ۲

بردار BOS حاصل از هر شناساگر آوایی در کنار یکدیگر قرار گرفته و یک بردار واحد را تشکیل می‌دهد. حال که سیگنال به یک بردار واحد تبدیل شد، می‌توان این مسئله را به دید یک مسئله طبقه‌بندی نگاه کرد و از یک طبقه‌بندی کننده مانند SVM برای پیدا کردن ابرصفحه جداکننده زبان‌ها استفاده نمود [3]. این سیستم در شکل ۲ نشان داده شده است.

۳–مدل‌های ارائه شده

۳–۱– مدل AW-GMM

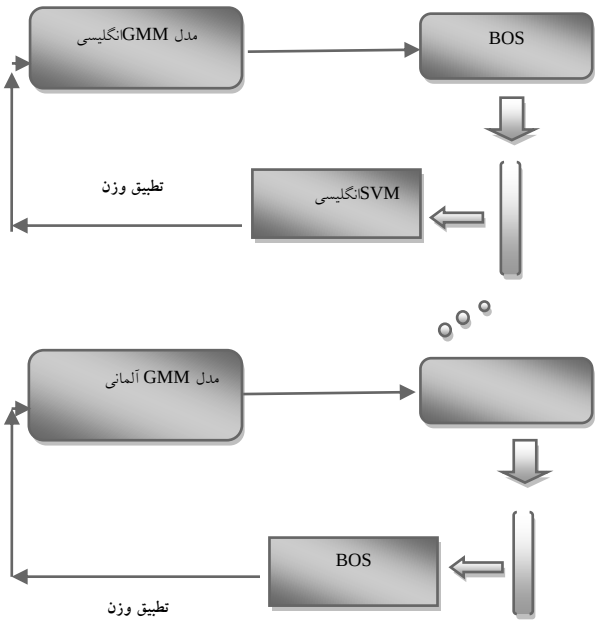
در آموزش GMM، الگوریتم EM به ازای هر مولفه، یک بردار میانگین، واریانس و وزن را محاسبه می‌کند. وزن هر مولفه بر اساس احتمال رخداد آن مولفه به ازای داده‌های آموزشی تعیین می‌شود، به عبارت دیگر وزن هر مولفه، متناسب با فرکانس تکرار آن مولفه در فضای آواهای آن زبان است و آموزش درست این وزن‌ها تاثیر زیادی در دقت تشخیص زبان دارد.ایرادی که در روش مدل کردن GMM وجود دارد این است که هر زبان مستقل از زبان دیگر مدل می‌شود. به عنوان نمونه ممکن است یک یا بخشی از یک آوا فرکانس تکرار کمی در یک زبان داشته باشد و در مقابل آوای دیگر فرکانس تکرار زیادتری داشته باشد اما تاثیر آوای اول در تشخیص این زبان از سایر زبان‌ها بیشتر از آوای دوم باشد،بنابراین یک راه برای افزایش دقت مدل‌های GMM می‌تواند تغییر مکانیزم وزن‌دهی مولفه‌ها بر اساس اطلاعات تمایزدهندگی آن‌ها باشد. به این منظور مدل AW-GMM ارائه شده است.مدل AW-GMM مشابه GMM است و با استفاده از الگوریتم EM آموزش داده می‌شود اما پس از آموزش، وزن مولفه‌های هر یک از مدل‌های GMM در قالب سیستم GMM-VSMدر حالت مدل‌کننده شنونده تک زبانه به طوری که در شکل ۳نشان داده شده است، تطبیق داده می‌شود.



شکل ۲: سیستم تشخیص زبان GMM-VSM

در سیستم GMM-VSM در حالت مدل‌کننده شنونده تک-زبانه، تنها از بردار BOS حاصل از هر مدل GMM برای آموزش طبقه‌بندی کننده SVM استفاده می‌شود. به عبارت دیگر هر SVM تنها از فضای آوای زبان مربوط به خود اطلاع دارد.

تطبیق وزن به این صورت انجام می‌شود که در فاز آموزش سیستم GMM-VSM، توالی بردارهای استخراج شده از داده-های آموزشی، به بردارهایی با طول ثابت (بر اساس فرکانس تکرار مولفه‌ها)، تبدیل می‌شوند. سپس این بردارهای ویژگی برای آموزش مدل‌های SVM با کرنل خطی، مورد استفاده قرار می‌گیرند که به ازای هر شناساگر GMM یک طبقه‌بندی کننده SVM با استفاده از بردارهای ویژگی مربوط به آن شناساگر و با در نظر گرفتن بردارهای متعلق به زبان آن شناساگر به عنوان نمونه‌های مثبت (نمونه‌های با برچسب ۱+) و بردارهای ویژگی متعلق به سایر زبان‌ها به عنوان نمونه‌های منفی (نمونه‌های با برچسب ۱–)، آموزش داده می‌شود.



شکل ۳: تکنیک تطبیق وزن ارائه شده

پس از آموزش، از هر SVM یک بردار وزن بدست می‌آید که از رابطه‌ی زیر محاسبه می‌شود.

$$w = \sum_{i=1}^{N} \alpha_i b(x_i) + d \qquad (۸)$$

که در آن $b(x_i)$ بردارهای پشتیبان، α_i ضرایب مربوط به بردارهای پشتیبان و N تعداد بردارهای پشتیبان است. هر وزن $|wi|$ نشان‌دهنده سهم هر بعد از بردارهای ویژگی در ساخت ابرصفحه جداکننده حاصل از طبقه بندی کننده SVM است. به عبارت دیگر هر چقدر $|wi|$ بیشتر باشد، بعد iام از بردارهای ویژگی تاثیر بیشتری در متمایز کردن زبان هدف از سایر زبان‌ها دارد [6]. به این ترتیب می‌توان از بردارهای وزن حاصل از مدل‌های SVM به عنوان وزن‌های جدید مولفه‌های GMM استفاده نمود. از آن‌جایی که این وزن‌ها بر اساس قدرت تمایزکنندگی هر مولفه در تشخیص یک زبان از سایر زبان‌ها تعیین شده‌اند، انتظار می‌رود که دقت سیستم‌های تشخیص زبانی که از مدل GMM با وزن‌های تطبیق یافته (AW-GMM) استفاده می‌کنند، افزایش یابد.

۳-۲-انتخاب ویژگی

استفاده از اطلاعات واج‌آرایی در بالا بردن دقت تشخیص زبان موثر است. این اطلاعات را می‌توان با استفاده از توالی‌های آوایی مدل کرد که در سیستم. در نظر گرفتن توالی‌های آوایی که در سیستم GMM-VSMمعادل مرتبه در نظر گرفته شده (n)، در ساخت بردارهای BOS است. هر چه مقدار n بیشتر در نظر گرفته شود، بردارهای BOS اطلاعات واج آرایی بیشتری را شامل خواهند شد اما پیچیدگی محاسباتی سیستم به میزان قابل توجهی بالا می‌رود. به عنوان مثال در حالتی که از مدل‌های GMM با تعداد مولفه‌های برابر با ۲٥٦ برای مدل کردن فضای آوایی استفاده شود، افزایش مرتبه از ۱ به ۲، سبب افزایش بعد بردارهای BOS از ۲٥٦ به ۲٥٦*۲٥٦ می‌شود و عملا در نظر گرفتن توالی‌های با مرتبه‌ی بالاتر را غیرممکن می‌سازد.

برای رفع این مشکل در این مقاله روشی ارائه شده که در ادامه شرح داده شده است. در این روش از تکنیکریچاردسون برای انتخاب کلماتی که سبب ایجاد تمایز میان لهجه‌های مختلف می‌شوند [7]الهام گرفته شده است.

این روش بر این اساس است که لزومی ندارد که همه‌ی توالی‌های ۲تایی مولفه‌ها برای ساخت بردارهای ویژگی در نظر گرفته شود. می‌توان تنها توالی‌های ۲تایی را درنظر گرفت که حداقل یک عضو تشکیل دهنده‌ی آن مولفه‌ای باشد که از وزن بالایی در تمایز زبان هدف از سایر زبان‌ها برخوردار باشد.

برای آزمایش موثر بودن این روش ابتدا ٤ مدل GMM با تعداد مولفه‌های برابر با ۲٥٦ آموزش داده شد و از آن‌ها برای آموزش دو سیستم GMM-VSM استفاده شد. در سیستم اول از فرکانس تکرار مولفه‌های GMM و در سیستم دوم از فرکانس

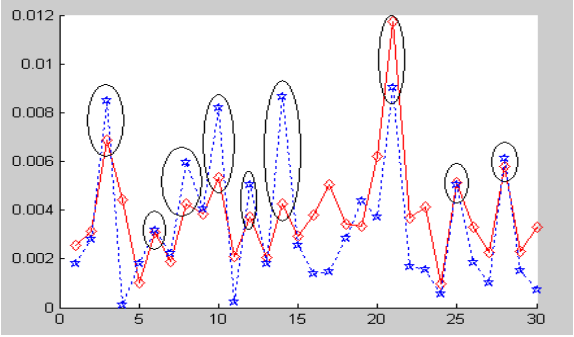
تکرار توالی ۲تایی مولفه‌ها برای ساخت بردارهای BOS استفاده گردید. در سیستم اول بر اساس بردارهای وزنی حاصل از طبقه‌بندی کننده‌های SVM، ۱۰۰ مولفه گاوسی که دارای وزن بالاتری نسبت به بقیه بودند، انتخاب شد، همچنین بر اساس وزن‌های طبقه‌بندی کننده‌های SVM در سیستم دوم نیز از میان توالی‌های ۲ تایی مولفه‌ها، ۲ تایی‌هایی که وزن بالاتری نسبت به سایر ۲ تایی‌ها داشتند، انتخاب گردید. درصد شرکت مولفه‌های گاوسی با وزن بالا در توالی ۲ تایی مولفه‌های با وزن بالا در جدول ۱، برای ٤ زبان انگلیسی، فارسی، فرانسوی و آلمانی نشان داده شده است.

جدول ۱- درصد حضور مولفه‌های گاوسی با وزن بالا در تشکیل توالی ۲ تایی مولفه‌ها با وزن بالا

انگلیسی	فارسی	فرانسوی	آلمانی
٪٦٨	٪٦٩	٪٦٣	٪٧٠

همان‌طور که مشاهده می‌شود درصد بالایی از توالی‌های ۲تایی که وزن بالایی در تشخیص زبان داشته‌اند را مولفه‌هایی تشکیل می‌دهند که آن‌ها نیز از وزن بالایی برخوردارند.

در شکل ٤، وزن مولفه‌ها برای ۳۰ مولفه‌ی اول و توالی ۲ تایی آن‌ها برای زبان آلمانی نشان داده شده است. همان‌طور که در این شکل مشخص شده است، اکثر نقاط پیک منحنی که با خط‌چین مشخص شده است و مربوط به وزن توالی ۲تایی مولفه‌هاست، معادل نقاط پیک منحنی که با خطوط مستقیم مشخص شده است و مربوط به وزن مولفه‌هاست، می‌باشد.



شکل ٤-مقادیر وزن مولفه‌های گاوسی (خطوط توپر) و توالی ۲ تایی مولفه‌های گاوسی (خطوط نقطه‌چین) برای زبان آلمانی

با توجه به این شهود می‌توان برای اضافه کردن احتمال رخداد توالی‌های ۲ تایی، تنها توالی‌هایی را در نظر گرفت که مولفه-های با وزن بالا در آن‌ها حضور دارند. به همین ترتیب می‌توان از این تکنیک برای ساخت توالی‌های ۳تایی و بالاتر استفاده نمود. از نظر پیچیدگی محاسباتی هنگامی که از تکنیک انتخاب ویژگی استفاده شود تنها در زمان آموزش توالی‌های ۲تایی که در تشخیص زبان موثرترند انتخاب می‌شوند و در زمان ارزیابی نه تنها محاسبات اضافه‌تری صورت نمی‌گیرد بلکه به دلیل کاهش بعد از پیچیدگی محاسباتی کاسته می‌شود.

٤- آزمایشات

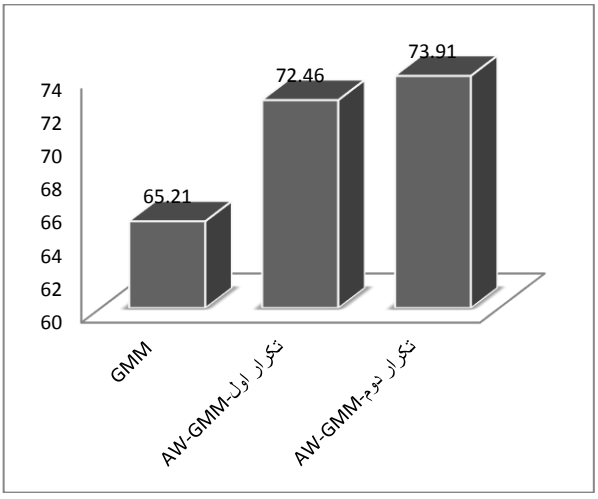
تمامی آزمایشات در این مقاله با استفاده از دادگان تلفنی چند زبانه OGI [8] صورت گرفته است. این دادگان شامل ۱۰ زبان انگلیسی، فارسی، فرانسوی، آلمانی، ژاپنی، کره‌ای، اسپانیایی، ماندارین، تامیل و ویتنامی است. از بین این زبان‌ها، ٤ زبان انگلیسی، فرانسه، فارسی و آلمانی جهت آموزش و ارزیابی سیستم‌ها انتخاب شده و آموزش سیستم‌ها با استفاده از قسمت آموزش و ارزیابی با استفاده از قطعات ٤٥ ثانیه‌ای قسمت ارزیابی دادگان، صورت گرفته است.

ابتدا قطعات گفتاری به فریم‌های ۳۰ میلی ثانیه با ۱۰ میلی ثانیه همپوشانی تبدیل و بخش‌های سکوت از آن‌ها حذف گردید. سپس از هر فریم بردار ویژگی MFCC شامل ۱۲ ضریب اول و انرژی به همراه مشتق اول و دوم این ضرایب استخراج شد. پس از استخراج بردارهای ویژگی، از CMS جهت نرمالسازی و حذف اثرات کانال استفاده شد. در مرحله‌ی بعد، ٤ مدل GMM (مدل GMM انگلیسی، فارسی، فرانسوی و آلمانی) با ۲۵٦ مولفه با استفاده از الگوریتم EM و ابزارHTK[9]و داده‌های آموزشی زبان هدف، آموزش داده شد. در مورد هر مدل، الگوریتم EM، ۱۰ مرتبه تکرار شد.در مرحله‌ی بعد مدل‌های حاصل برای آموزش سیستم GMM-VSM مورد استفاده قرار گرفت و با استفاده از تکنیک تطبیق وزن ارائه شده، ٤ مدل AW-GMM حاصل شد.

در آزمایش اول تاثیر استفاده از مدل AW-GMM به عنوان جایگزین مدل GMM در قالب سیستم تشخیص زبانی که در قسمت ۲-۲ توصیف شد، مورد بررسی قرار گرفت. بدین منظور ۲ سیستم تشخیص زبان پیاده‌سازی شد. در سیستم اول از ٤ مدل GMM استفاده شد و تصمیم‌گیری با استفاده از ماکزیمم‌گیری بر روی احتمالات خروجی حاصل از مدل‌ها صورت گرفت. سیستم دوم مشابه سیستم اول درنظر گرفته شد با این تفاوت که از AW-GMM به جای GMM استفاده شد. دقت تشخیص زبان هر یک از سیستم‌ها در نمودار شکل ۵نشان داده شده است. همان‌طور که در این شکل مشاهده می-

شود، استفاده از AW-GMM به جای GMM سبب افزایش دقت تشخیص زبان می‌شود (افزایش از ٦۵/۲۱ درصد به ۷۲/٤٦ درصد). در صورتی که عمل تطبیق وزن را یکبار دیگر تکرار نماییم، به این صورت که از مدل‌های AW-GMM در سیستم GMM-VSM استفاده کرده و وزن این مدل‌ها را به‌طوری که در قسمت ۳-۱ توضیح داده شد، تطبیق دهیم، دقت تشخیص زبان بار دیگر افزایش پیدا می‌کند که البته این افزایش نسبت به افزایش مرحله‌ی اول چشمگیر نیست (افزایش از ۷۲/٤٦ به ۷۳/۹۱ درصد) و از این رو ما تعداد تکرارها را در همین مرحله متوقف نمودیم.

در آزمایش دوم تاثیر استفاده از مدل‌های AW-GMM به عنوان شناساگرهای آوایی مورد بررسی قرار گرفته است. به این منظور ۲ سیستم تشخیص زبان GMM-VSMاز مرتبه ۱ (مقدار n برای ساخت بردارهای BOS برابر ۱ در نظر گرفته شد) آموزش داده شده که در سیستم اول از مدل GMM و در سیستم دوم از مدل AW-GMM به عنوان شناساگر آوا استفاده شد.



شکل ۵: تاثیر استفاده از مدل AW-GMM به عنوان جایگزین مدل GMM

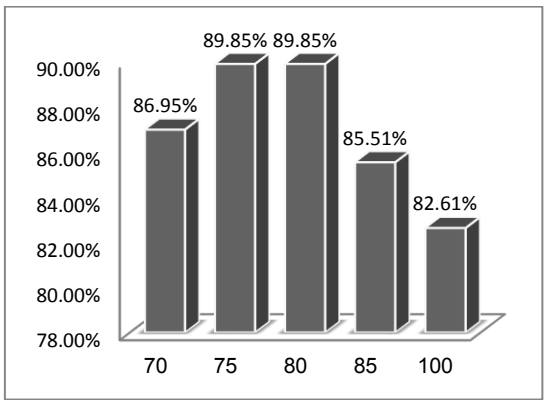
دقت هر یک از این سیستم‌ها در جدول ۲ نشان داده شده است. همان‌طور که مشاهده می‌شود استفاده از مدل AW-GMM به عنوان جایگزین مدل GMM در سیستم GMM-VSM سبب افزایش دقت تشخیص زبان می‌شود.

جدول ۲: تاثیر استفاده از مدل AW-GMM به عنوان شناساگر آوا در سیستم GMM-VSM

AW-GMM-VSM مرتبه ۱	GMM-VSM مرتبه ۱
%۸۸/٤۱	%۸۱/۱۶

در آزمایش سوم تاثیر تکنیک انتخاب ویژگی ارائه شده که در قسمت ۳-۲ توصیف شد، مورد بررسی قرار گرفت. بدین منظور ۲ سیستم GMM-VSMاز مرتبه ۲ پیاده‌سازی شد. در سیستم اول از تمامی توالی‌های ۲تایی مولفه‌ها برای ساختبردارهای BOS استفاده شد و در سیستم دوم تنها توالی-های ۲تایی در نظر گرفته شد که حداقل یک عضو تشکیل دهنده‌ی آن‌ها از بین مولفه‌های با وزن بالا انتخاب شده باشد.

در نمودار شکل ٦دقت تشخیص زبان سیستم GMM-VSM در حالتی که از تکنیک انتخاب ویژگی استفاده شود، به ازای تعدادمولفه‌های انتخابی مختلف نشان داده شده است. همان‌طور که مشاهده می‌شود، بهترین کارایی زمانی حاصل می‌شود که عمل انتخاب بر اساس توالی‌های ۲ تایی ۷۵ مولفه اول که دارای وزن بالاتری هستند صورت گیرد.به عبارت دیگر از نظر محاسباتی در حالت اول سیستم بر اساس بردارهای ویژگی با بعد ۲۵٦*۲۵٦ و در حالت دوم بر اساس بردارهای با بعد ۷۵*۷۵ عمل تصمیم‌گیری در مورد زبان هدف را انجام می‌دهد که به طور قابل ملاحظه‌ای پیچیدگی محاسباتی کم می‌شود.



شکل ٦: دقت سیستم GMM-VSM در حالت استفاده از توالی ۲تایی مولفه‌ها به ازای تعداد مولفه‌های انتخابی مختلف

دقت تشخیص زبان هر یک از این سیستم‌ها در جدول ۳ نشان داده شده است. همان‌طور که مشاهده می‌شود علاوه بر اینکه استفاده از تکنیک انتخاب ویژگی سبب کاهش پیچیدگی محاسباتی شده، دقت تشخیص زبان را نیز افزایش داده است.

جدول ۳: تاثیر استفاده از انتخاب ویژگی ارائه شده بر دقت سیستم تشخیص زبان GMM-VSM در حالت استفاده از توالی‌های ۲تایی مولفه‌ها

سیستم GMM-VSM– مرتبه ۲ بدون استفاده از تکنیک انتخاب ویژگی	سیستم GMM-VSM–مرتبه ۲ با استفاده از تکنیک انتخاب ویژگی
٪۸۲/٦۱	**٪۸۹/۸۵**

۵- جمع‌بندی ونتیجه‌گیری

در این مقاله مدلی جدید با نام AW-GMM ارائه گردید. این مدل مشابه GMM است با این تفاوت که وزن مولفه‌ها بر اساس قدرت تمایزکنندگی آن‌ها در قالب سیستم تشخیص زبان GMM-VSM تعیین می‌شود. ارزیابی‌های صورت گرفته با استفاده از ٤ زبان انگلیسی، فارسی، فرانسوی و آلمانی از دادگان OGI نشان داد که استفاده از مدل ارائه شده به عنوان جایگزین GMM در سیستم‌های تشخیص زبانی که از آن استفاده می‌کنند سبب افزایش دقت تشخیص زبان می‌شود. همچنین تکنیکی برای ساخت توالی‌های ۲تایی مولفه‌ها در سیستم GMM-VSM ارائه گردید که علاوه بر کاهش پیچیدگی سبب افزایش دقت تشخیص زبان در این سیستم گردید. نتیجه کلی که از آزمایشات صورت گرفته در این مقاله حاصل می‌شود این است که در نظر گرفتن فضای آوایی زبان‌ها به گونه‌ای که به هر آوا بر اساس قدرت متمایز کردن زبان هدف از سایر زبان‌ها وزن داده شود نه بر اساس تعداد دفعات تکراری که در آن زبان دارد، می‌تواند به میزان قابل توجهی دقت تشخیص زبان را بالا برد و از پیچیدگی محاسباتی کم کند.

قدردانی:
این مقاله مورد حمایت مالی مرکز تحقیقات مخابرات ایران در قالب قرارداد ۵۰۰/۱٤۹۳۹/ت قرار گرفته است.

٦-مراجع

[1] Ziaei A., Ahadi S. M.,Mirrezaie S. M. and Yeganeh H., "Spoken Language Identification Using a New Sequence Kernel-based SVM Back-end Classifier", ISSPIT, 2008, pp.324-329.

[2] Zissman M. A., "Comparision of Four Approaches to Automatic Language Identification of Telephone Speech", IEEE Transactions on Speech and Audio Processing, vol. 4, 1996, pp.31-44.

[3] Li H., Ma B. and Lee C. H., "A Vector space modeling approach to spoken language identification," IEEE Transactions on Audio, Speech and Language Processing, vol. 15, 2007, pp.271-284.

[4] Torres-Carrasquillo P. A., Singer E., Kohler M. A., Greene R. J., Reynolds D. A. and Deller J. A., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features", ICSLP, 2002, pp.89-92.

[5] Tong, R.,Bin, M.,Zhu, D.,Li, H., Chng, E. S., "Integrating acoustic, prosodic and phonotactic features for spoken language identification," ICASSP, 2006, pp. 205-208.

[6] Tong R., Ma B., Li H., and Chng E. S., "Target-Oriented Phone Tokenizers for Spoken Language Recognition", ICASSP 2008, pp. 200-203.

[7] Richardson F. S., Campbell W. M., Torres-Carrasquillo P. A., "Discriminative N-gram selection for dialect recognition", interspeech, 2009, pp. 192-195.

[8] Muthusamy Y. K., Cole R. A., Oshika B. T., "The OGI multi-language telephone speech corpus", ICSLP, 1992.

[9] [1] Available at: http://htk.eng.cam.ac.uk/