مقایسهٔ روش‌های طیفی برای بازشناسی زبان گفتاری

شقایق رضا* [1] و سید جهانشاه کبودیان [2]

[1] پژوهشکدهٔ پردازش داده، پژوهشگاه توسعهٔ فناوری‌های پیشرفته خواجه نصیرالدین طوسی، تهران، ایران
[2] گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه رازی، کرمانشاه، ایران

چکیده

شناسایی خودکار زبان گفتاری به تشخیص زبان از روی سیگنال گفتار گفته می‌شود. شناسایی زبان به‌طورمعمول به یکی از دو دسته روش آوایی و طیفی انجام می‌شود. در این مقاله، انواع روش‌های مختلف طیفی برای بازشناسی زبان گفتاری معرفی شده و نتایج به‌کارگیری آنها بر روی یک مجموعه دادگان گفتاری تلفنی محاوره‌ای مقایسه شده است. روش طیفی پایهٔ شناسایی زبان، مدل مخلوط گوسی-مدل جهانی (GMM-UBM) است. برای بهبود مدل گوسی هر زبان از روش تمایزی MMI و برای مدل‌کردن دینامیک زبان از مدل پنهان مارکوف ارگودیک (EHMM) استفاده می‌شود. روش‌های GSV-SVM و روش نشانه‌گذار مبتنی بر GMM (GMM Tokenizer) نیز دو روش طیفی دیگر است که مورد بررسی قرار گرفته است. در این مقاله همچنین روش‌های جدیدِ مدل‌سازی تنوعات کانال و گوینده (تحلیل توأم عامل‌ها (JFA) و بردار شناسایی (i-Vector)) به‌کار رفته و برای بهبود نتایج آن از چند روش جبران‌سازی تنوعات استفاده شده است. علاوه‌براین برای سهولت تصمیم‌گیری و کاهش خطای سامانهٔ شناسایی زبان، از پس‌پردازش امتیاز استفاده شده است. این مقاله بخشی از هفت سال پژوهش در زمینه شناسایی زبان گفتاری در پژوهشگاه توسعهٔ فناوری‌های پیشرفته خواجه نصیرالدین طوسی است و تنها خلاصه‌ای از روش‌ها و نتایج به‌دست‌آمده در این مقاله آورده شده است.

واژگان کلیدی: شناسایی خودکار زبان گفتاری، روش‌های طیفی، آموزش تمایزی، جبران‌سازی تنوعات کانال، بردار شناسایی.

A Comparison of Spectral Approaches to Spoken Language Identification

Shaghayegh Reza[1] and Seyyed Jahanshah Kabudian[2]

[1] Research Center for Development of Advanced Technologies (RCDAT), Tehran, Iran
[2] Department of Computer Engineering and Information Technology, Razi University, Iran

Abstract

Identifying spoken language automatically is to identify a language from the speech signal. Language identification systems can be divided into two categories, spectral-based methods and phonetic-based methods. In the former, short-time characteristics of speech spectrum are extracted as a multi-dimensional vector. The statistical model of these features is then obtained for each language. The Gaussian mixture model is the most common statistical model in spectral-based language identification systems. On the other hand, in phonetic-based methods, speech signals are divided into a sequence of tokens using the hidden Markov model (HMM) and a language model is trained using the obtained sequence. Approaches like PRLM, PPRLM, and PR-SVM are some examples of phonetic-based methods. In research papers, usually a combination of phonetic-based and spectral-based systems are used to achieve a high quality language identification system. Spectral-based methods have been the focus of researchers, since they have no need

* نویسندهٔ عهده‌دار مکاتبات

for labeled data and usually achieve better results than phonetic approaches. Therefore, in this paper, these methods used for language identification and different spectral methods, are introduced, implemented, and compared with spoken language recognition.

The basic spectral language identification method is Gaussian Mixture Model-Universal Background Model (GMM-UBM). In this paper, the MMI discrimination method is used to improve the Gaussian model of each language. Moreover, in order to model the language dynamically, GMM is replaced with the ergodic hidden Markov model (EHMM). GSV-SVM and GMM tokenizer methods are also implemented as two popular spectral approaches. In this paper, novel speaker and channel variation modeling methods are used as language identification approaches, including joint factor analysis (JFA), identity vector (i-Vector) and several variations compensation methods exploited to improve the results of i-Vector.

Furthermore, in order to boost the performance of language recognition systems, different post-processing methods are applied. For post-processing, each element of raw score vector indicates the degree by which the spoken signal belongs to a language. Post-processing methods are applied to this vector as a classifier and allows making better language detection decisions by mapping the raw score vector to a space of desired languages. Different studies have employed different post-processing methods, including GMM, NN, SVM, and LLR. This study exploits several score post-processing methods to improve the quality of language recognition.

The goal of the experiments in this article is to detect and distinguish Farsi, English, and Arabic, individually and simultaneously from other languages. The latter is also called open-set language identification. The signals considered in this paper include two-sided conversations, whose quality is usually not desirable due to strong noise signals, background noises of individuals or music, accents, etc.

Gaussian mixture-universal model (GMM-UBM) was implemented as the basic method. In this approach, mean EER of the three target languages (Farsi, English, and Arabic) was 13.58. Experimental results indicated that training the GMM language identification system with the MMI discrimination training algorithm is more efficient than systems only trained by the ML algorithm. More specifically, the mean EER of the three target languages was reduced about 8 percent in comparison to GMM-UBM. The GMM tokenizer method was also tested as a novel spectral approach. Using this method, the mean EER of the three target languages was also about 5 percent better than GMM-UBM.

In this study, the GSV-SVM discrimination method was also used for language recognition. The results of this method were considerably better than those of common spectral approaches, such that the mean EER of the three target languages was reduced by 11 percent in comparison to GMM-UBM. This study improves the low speed of this method using a model pushing method.

This study also implemented two novel methods, JFA and i-Vector. According to the results, both of these methods provide better results than GMM-UBM, such that the mean EER values of the three target languages in JFA and i-Vector are respectively reduced by 1% and 12%. Generally, experimental results showed that i-Vector provides better results than other spectral language identification systems.

This study is a result of a seven-year research in spoken language identification in the advanced technology development center of Khajeh Nasiredin Tousi. The ongoing research includes studying and implementing novel spectral language identification algorithms like PLDA and state-of-the-art phonetic language identification methods to combine the two spectral and phonetic systems and eventually, achieving a high quality language identification system.

Keywords: Automatic Spoken Language Recognition, Acoustic Approaches, Discriminative training, Channel compensation, Identity Vector.

۱-مقدمه

شناسایی خودکار زبان گفتاری، تشخیص زبان از روی سیگنال گفتار است. شناسایی خودکار زبان می‌تواند شامل تشخیص یک زبان خاص از مجموعه‌ای از زبان‌های دیگر (آشکارسازی زبان[1]) و یا شناسایی و تعیین زبان از یک مجموعه زبان شناخته شده باشد (تعیین زبان[2]). تنوعات ذاتی سیگنال گفتار (مثل مفهوم و گوینده)، نوفه‌ها و اثرات مخرب محیطی، موجب پیچیدگی استخراج اطلاعات مربوط به زبان شده و رسیدن به سامانهٔ شناسایی زبان را با کیفیت مناسب مشکل کرده است [30].

سامانه‌های شناسایی زبان را به دو دسته می‌توان تقسیم‌بندی کرد: روش‌های مبتنی بر طیف[3] و روش‌های مبتنی بر نشانه‌ها یا واحدهای آوایی[4]. در روش نخست، ویژگی‌های زمان-کوتاهِ طیف گفتار به‌صورت یک بردار چندبعدی استخراج می‌شود؛ سپس مدل آماری این ویژگی‌ها برای هر زبان به‌دست می‌آید. مدل مخلوط گوسی

[1] Language Detection
[2] Language Identification
[3] Spectral-Based Methods
[4] Token-Based Methods

معمول‌ترین مدل آماری در سامانه‌های شناسایی زبانِ مبتنی بر طیف است. در مقابل، در روش مبتنی بر نشانه‌ها یا واحدهای آوایی، گفتار با استفاده از مدل پنهان مارکوف (HMM[1]) یا هر بازشناس دیگر، به دنباله‌ای از نشانه‌ها یا واحدهای آوایی تقسیم می‌شود؛ سپس یک مدل زبانی روی دنبالهٔ به‌دست‌آمده تعلیم داده می‌شود [38]. روش‌هایی چون PRLM[2]، PPRLM[3] و PR-SVM[4] از جمله روش‌های مبتنی بر واحدهای آوایی هستند [17]، [49]، [41]، [30]. به‌طورمعمول در مقالات از تلفیق سامانه های مبتنی بر روش های آوایی و طیفی برای رسیدن به سامانه شناسایی زبان با کیفیت بالا استفاده می‌شود [7]، [8].

از آنجا که روش‌های مبتنی بر طیف، نیازی به دادگان برچسب‌دار ندارد و به‌طورمعمول نتایجی بهتر از روش‌های آوایی دارد، بیشتر مورد توجه پژوش‌گران قرار گرفته است؛ از این رو در این مقاله از این دسته از روش‌ها برای شناسایی زبان استفاده شده و انواع روش‌های مختلف طیفی برای بازشناسی زبان گفتاری، معرفی، پیاده‌سازی و مقایسه شده است.

مدل مخلوط گوسی (GMM[5]) معمول‌ترین مدل آماری در سامانه‌های شناسایی زبان مبتنی بر طیف است. در این روش برای هر زبان یک مدل مخلوط گوسی با استفاده از الگوریتم EM[6] و معیار ML[7] تعلیم داده می‌شود [39]. بورگت[8] در سال ۲۰۰۶ برای متمایزترشدن مدل‌های زبانی، از معیار تمایزی MMI[9] استفاده کرد. با استفاده از این روش، روش، بهینه‌سازی در جهتی پیش می‌رود که مرزهای تصمیم‌گیری و جداسازی به‌نحو دقیق‌تری مدل شود. در نتیجه پارامترهای مدل برای مدل‌سازی بخش‌هایی از توزیع ویژگی که در تمامی مدل‌ها یکسان است، هدر نمی‌رود [24]، [8].

در سال ۲۰۰۶، کمپبل[10] و همکارانش، روش تمایزی جدیدی برای شناسایی زبان ارائه کردند. در این روش از هر داده‌ی تعلیم یک اَبَر بردار گوسی (GSV[11]) استخراج شده و این اَبَر بردارها توسط ماشین بردار مرزی (SVM[12]) طبقه‌بندی می‌شود. این روش GSV-SVM نامیده می‌شود [39]، [11]، [10]. در این روش، به‌دلیل زمان‌بَربودنِ مرحلهٔ استخراج اَبَربردارگوسی و محاسبه تابع تصمیم‌گیری SVM، مدت زمان آزمایش زیاد است. برای کاهش این زمان، کمپبل در سال ۲۰۰۸ یک روش برای برگشت به حوزه مدل (MP[13]) پیشنهاد کرده است [9].

یک روش شناسایی زبان طیفی به‌نسبت جدید، جایگزینی بازشناس زبان (در سامانه‌های مبتنی بر واحدهای آوایی) با یک مدل مخلوط گوسی (GMM Tokenizer) است. این روش در سال ۲۰۰۷ توسط یانگ[14] مورد استفاده قرار گرفت [46]. نشانه‌گذار (Tokenizer)، هر فریم (قاب) گفتاری را مورد پردازش قرار داده و رشته‌ای از اندیس‌های مؤلفه‌های گوسی را برای هر قاب تولید می‌کند. دنبالهٔ بردارهای به‌دست‌آمده برای هر داده تعلیم (یا داده آزمایش)، با استفاده از SVM طبقه‌بندی می‌شود [21]، [38].

همان‌طور که می‌دانیم، دادگان مورد استفاده در سامانه شناسایی زبان تنها شامل اطلاعات مفید نیست؛ بلکه این دادگان اطلاعات دیگری نیز دارد که کمکی به شناسایی زبان نمی‌کند. این اطلاعات ناخواسته شامل ویژگی‌های مربوط به گویندهٔ گفتار (شکل مجرای صوتی)، لهجه، احساسات، وضعیت سلامتی فرد و غیره است؛ علاوه‌براین، شرایطِ ضبط صدا همچون نوفه پس‌زمینه، ویژگی‌های میکروفن، کانال انتقال و شیوه کدگذاری صوتی نیز اطلاعات ناخواسته‌ای را به سیگنال گفتار اضافه می‌کند. به تمامی این اطلاعات ناخواسته، اطلاعات وابسته به شرایط ضبط[15] (تنوعات مزاحم) گفته می‌شود. برای حذف این تنوعات مزاحم، روش‌های جدیدی از نوع مبتنی بر طیف پیشنهاد شده است. یکی از این روش‌ها، روش تحلیل توأم عامل‌ها (JFA[16]) است که در سال ۲۰۰۵ توسط کِنی[17] و همکارانش معرفی شد [29]، [26]. در این روش اطلاعات وابسته به زبان و تنوعات جداسازی می‌شود. برای جداسازی اطلاعات وابسته به زبان، فرض می‌شود که اطلاعات زبانی، بخشی است که در تمامی دادگان یکسان است و بخش باقی‌مانده، مربوط به تنوعات است. با وجود کیفیت مناسب این روش، نشان داده شد که بخش مربوط به تنوعات همچنان دارای اطلاعات

[1] Hidden Markov Model
[2] Phone Recognition - Language Modeling
[3] Parallel Phone Recognition and Language Modeling
[4] Phone recognition-SVM
[5] Gaussian Mixture Model
[6] Expectation Maximization
[7] Maximum Likelihood
[8] Burget
[9] Maximum Mutual Information
[10] Campbell
[11] Gaussian Super Vector
[12] Support Vector Machines
[13] Model Pushing
[14] Yang
[15] Session Dependent Information
[16] Joint Factor Analysis
[17] Kenny

زبانی است [13]. برای رفع این مشکل، روش بردار شناسایی (i-Vector[1]) پیشنهاد شد. این روش نیز روشی مشابه روش JFA است با این تفاوت که یک زیرفضای تنوعات کلی[2] در نظر گرفته شده و تنوعات مختلف، جداسازی نمی‌شوند. این روش که جدیدترین روش به‌کاررفته در سامانه‌های شناسایی زبان تا سال ۲۰۱۱ است، ابتدا توسط دهک[3] و همکارانش ارائه شده [16]، [15] و سپس توسط ماتروف[4] و همکارانش ساده‌سازی شد [43]، [31]؛ پس از آن، برای بهینه‌سازی استخراج i-Vector ها و مدل‌کردن و طبقه‌بندی آنها، روش‌هایی توسط گلمبرگ[5] و مارتینز[6] ارائه شد [41]، [19].

در این مقاله، کلیه روش‌های طیفی معرفی، پیاده‌سازی و کیفیت این روش‌ها با یکدیگر مقایسه شده است. هدف ما در آزمایش‌های انجام‌شده، شناسایی و جداسازی هر یک از زبان‌های فارسی، انگلیسی و عربی از زبان‌های غیر فارسی، غیر انگلیسی و غیر عربی (STLD[7]) و نیز شناسایی و جداسازیِ همزمانِ مجموعه سه زبان فارسی، انگلیسی و عربی از زبان‌هایی غیر از این سه زبان است (MTLD[8]). حالت دوم، تعیین زبان به‌صورت مجموعه-باز[9] نیز نامیده می‌شود. سیگنال‌های مورد نظر در این مقاله، سیگنال‌های گفتار تلفنی محاوره‌ای از مکالمات دوطرفه است که به‌طورعمومی کیفیت مناسبی ندارند. به‌عنوان مثال می‌توان به مواردی چون سیگنال‌های به‌شدت نوفه‌ای، وجود صدای افراد و موسیقی در پس‌زمینه، لهجه‌های غلیظ و مواردی از این قبیل اشاره کرد.

یکی از روش‌های بهبود کیفیت سامانه‌های شناسایی زبان، استفاده از روش‌های پس‌پردازش بردار امتیازهای خام است. مقصود از بردار امتیازهای خام، برداری است که هر یک از مؤلفه‌های آن، میزان تعلق سیگنال گفتار به یکی از زبان‌ها است. روش‌های پس‌پردازش به‌صورت یک طبقه‌بندی‌کننده روی این بردار امتیازها عمل‌کرده و با نگاشت بردار امتیازهای خام به فضای زبان‌های مطلوب، موجب تصمیم‌گیری بهتر در زمینه تشخیص زبان می‌شود. در مقالات مختلف از روش‌های متفاوتی از جمله GMM، NN، SVM و LLR [34]، [30]، [4]، [3] برای پس‌پردازش استفاده شده است. در این مقاله نیز برای بهبود کیفیت بازشناسی زبان از چندین روش پس‌پردازش امتیازات استفاده شده است.

در ادامه ابتدا در بخش دوم، مراحل استخراج ویژگیِ طیفی از سیگنال گفتار معرفی می‌شود؛ سپس در بخش سوم روش‌های مختلف بازشناسی زبان همچون GMM-UBM، GMM-MMI، GSV-SVM، GMM Tokenizer-SVM، JFA و i-Vector معرفی و در بخش چهارم، روش‌های پس‌پردازش امتیازات توضیح داده شده است؛ سپس در بخش پنجم، دادگان مورد استفاده و در بخش ششم، آزمایش‌ها و نتایج حاصله ارائه و در بخش هفتم رویکردها و روش‌های جدید از مقالات به‌روز معرفی شده و درنهایت در بخش هشتم، یک جمع‌بندی کلی از مقاله و از نتایج آزمایش‌ها بیان شده است.

۲- استخراج ویژگی طیفی

ویژگی‌های طیفی معمول برای سامانه شناسایی زبان، ضرایب کپسترال هفت‌بعدیِ مبتنی بر بانک فیلتر با توزیع فرکانسی مل (MFCC[10]) به‌علاوه ضریب c_0، تلفیق‌شده با بردارهای شیفت‌یافته کپسترال (SDC[11]) هستند. ضرایب SDC برای مدل‌کردن پویای زمان-کوتاهِ گفتار به بردارهای MFCC معمول اضافه شده و به میزان قابل ملاحظه‌ای موجب بهبود نتایج می‌شود [42]. فرمول استخراج SDC از ضرایب MFCC به‌صورت زیر است:

$$c(t) = [c_0(t), c_1(t), ..., c_{N-1}(t)] \quad (۱)$$

$$\Delta c(t,i) = c(t + iP + h) - c(t + iP - h) \quad (۲)$$

$$c^{tot}(t) = [c(t), \Delta c(t,0), \Delta c(t,1), ..., \Delta c(t,k-1)] \quad (۳)$$

در این رابطه $c^{tot}(t)$ بردار ویژگی نهایی در زمان t، $c(t)$ بردار کپسترال در زمان t، P شیفت زمانی بین دو بلوک متوالی و h میزان تقدم و تأخر زمانی برای محاسبه ضرایب دلتا است. پارامتر N نیز تعداد ضرایب بردار کپسترال معمولی است که در هر قاب استفاده می‌شوند و k تعداد بلوک‌هایی است که ضرایب دلتای آن‌ها باید محاسبه شود [2]. در این مقاله تنظیماتِ استخراج پارامترهای SDC به‌صورت $(N,P,h,k) = (8,1,3,8)$ در نظر گرفته شده است. بنابراین در مجموع از هر قاب گفتاری، یک بردار

[1] Identity Vector
[2] Total Variability
[3] Dehak
[4] Matrouf
[5] Glemberg
[6] Martinez
[7] Single-Target Language Detection
[8] Multi-Target Language Detection
[9] Open-Set Language Identification

[10] Mel-Frequency Cepstral Coefficient
[11] Shifted Delta Cepstral

ویژگی ۷۲ بُعدی استخراج می‌شود.

در این مقاله، برای گفتار تلفنی با فرکانس نمونه-برداری ۸ کیلو هرتز، تعداد فیلترها در بانک فیلتر برابر ۱۴و طول پنجره تقطیع سیگنال ۳۲ میلی‌ثانیه و مقدار پیشروی زمانی (فاصله‌ی زمانِ شروع دو قاب متوالی) ۱۰ میلی‌ثانیه است. برای آشکارسازی نواحی گفتاری (VAD[1])، از یک روش مقاوم مبتنی بر انرژی با سطح آستانهٔ تطبیقی استفاده شده است.

۲-۱-بهبود ویژگی‌ها با فیلتر رستا[2]

فیلتر رستا یک روش مهم و پر کاربرد در زمینه پردازش گفتار است که درصورت وجود تغییرات کانال انتقال یا ابزار ضبط سیگنال، موجب بهبود ویژگی‌های استخراجی و درنهایت بهبود کارایی سامانه‌های شناسایی گفتار، گوینده و زبان می‌شود. این فیلتر با حذف فرکانس‌های نزدیک به صفر (در حوزهٔ فرکانس مدولاسیون)، موجب حذف برخی فرکانس‌های کوچک مخرب (نوفه‌های ایستان و اعوجاجات ثابت کانال) از سیگنال می‌شود. از طرف دیگر این فیلتر فرکانس‌های بالا و غیر مفید (در حوزهٔ فرکانس مدولاسیون) را نیز حذف می‌کند. فرکانس قطع بالای فیلتر به‌نحوی تنظیم شده است که نوفه‌های گذرا و تغییرات سریع محیطی از سیگنال گفتار حذف می‌شوند [25].

با وجود فواید مهم فیلتر رستا در حذف مؤلفه‌های فرکانس-پایین و مخرب از سیگنال گفتار، این فیلتر ممکن است، برخی مؤلفه‌های فرکانس-پایین و مفید گفتار را نیز حذف کند. بنابراین هر چه مقدار قطب این فیلتر به عدد یک نزدیک‌تر باشد، فیلتر در فرکانس صفر به حذف مقدار DC سیگنال (در حوزهٔ فرکانس مدولاسیون) نزدیک‌تر شده و مشخصاً تأثیرات منفی فیلتر بر مؤلفه‌های فرکانس-پایین و مفید گفتار کاهش می‌یابد. علاوه‌براین، مشخصه مرحلهٔ فیلتر، به مرحلهٔ خطی نزدیک‌تر خواهد شد. فیلتر IIR[3] رستا نیاز به پنج مقدار اولیه $[x(-1), x(-2), x(-3), x(-4), y(-1)]$ دارد. در صورت صفر قراردادن این مقادیر، فیلتر به زمانی برای پایدارشدن نیاز دارد. بنابراین درصورت اعمال این فیلتر بر روی ضرایب ویژگی، بخش‌های ابتدایی بردارهای ویژگیِ فیلترشده، به‌دلیل عدم پایداری فیلتر، نامعتبر خواهند بود. برای استفاده از مزایای افزایش مقدار قطب (نزدیک به عدد یک)، باید به جای مقادیر اولیه صفر، مقادیر اولیه معتبر و مناسبی قرار دهیم تا فیلتر به‌سرعت پایدار شده و نقش تخریبی بر روی ضرایب ویژگی گفتار نداشته باشد. در مرجع [1] روش تعیین مقادیر اولیه مناسب برای فیلتر رستا شرح داده شده است. در این مقاله از مقادیر معتبر اولیه تعیین‌شده برای قطب ۰٫۹۹۹ در فیلتر رستا استفاده شده است.

۲-۲-کاهش بُعد به روش LDA[4]

روش LDA به دنبال یافتن یک ماتریس انتقال A برای تبدیل بردار ویژگی x به x' است؛ به‌گونه‌ای که این انتقال خطی بتواند تفکیک‌پذیری بین طبقه‌های مختلف را افزایش و در مقابل پراکندگی درون‌طبقه‌ای را کاهش دهد. در این روش، بُعد فضای جدید حداکثر برابر $c-1$ است که c تعداد طبقه‌های موجود است [5].

$$x' = Ax \quad (۴)$$

از آنجا که تعداد زبان‌های تعلیم در این مقاله پانزده زبان است، با استفاده از روش LDA بُعد بردار ویژگی را می‌توان از ۷۲ به حداکثر۱۴ تبدیل کرد. از آنجا که کاهش بُعد بردارهای ویژگی تا این حد مطلوب نیست و باعث ازبین‌رفتن اطلاعات می‌شود؛ لذا باید به طریقی تعداد طبقه‌ها را به‌صورت مجازی افزایش داد؛ بدین منظور ابتدا یک مدل مخلوط گوسی (با ۱۰۲۴ مؤلفه گوسی) با استفاده از الگوریتم EM روی دادگان کلیه زبان‌ها تعلیم داده شد؛ سپس لگاریتم احتمال تعلق هر یک از بردارهای ویژگی استخراج‌شده از دادگان تعلیم به هر یک از ۱۰۲۴ مؤلفه گوسی محاسبه می‌شود. برای هر یک از بردارهای ویژگی، شمارهٔ مؤلفهٔ گوسی که به‌ازای آن مقدار لگاریتم احتمال بیشینه است، به‌عنوان طبقهٔ مجازی بردارهای ویژگی در نظر گرفته می شود. بدین ترتیب کلیه بردارهای ویژگی، در ۱۰۲۴ طبقهٔ مجازی مختلف قرار می‌گیرند.

در این مقاله با استفاده از LDA و به روش بالا، بُعد بردار ویژگی از ۷۲ به چهل کاهش داده شده است. کاهش بعد بردار ویژگی علاوه‌بر افزایش سرعت برنامه، موجب حذف اطلاعات اضافه (و فاقد اطلاعات زبانی) از بردار ویژگی شده و اغلب موجب بهبود کیفیت سامانهٔ شناسایی زبان می‌شود.

۳-سامانهٔ شناسایی زبان

در این بخش روش‌ها و اجزای هریک از روش‌های طیفی

[1] Voice Activity Detection
[2] RASTA
[3] Infinite Impulse Response
[4] Linear Discriminant Analysis

برای شناسایی زبان معرفی شده است.

۳-۱-روش مدل مخلوط گوسی

شناسایی زبان یک مسألهٔ طبقه‌بندی الگو است که هر زبان در آن نشان‌دهندهٔ یک طبقه است. یکی از روش‌های معمول برای مدل‌کردن توزیع دادگان هر طبقه، استفاده از مدل مخلوط گوسی (GMM) است. در این مقاله توزیع ویژگی-های MFCC+SDC در هر زبان، با یک توزیع GMM مدل می‌شود:

$$p(x;\lambda)=\sum_{i=1}^{M} w_i \mathcal{N}_i(x) \quad (۵)$$

x در این رابطه بردار ویژگی d بعدی، $\mathcal{N}_i(x)$ تابع چگالی گوسی i اُم و w_i وزن هر تابع گوسی است. هر تابع چگالی گوسی d متغیره با بردار میانگین μ_i و ماتریس کوواریانس Σ_i به‌صورت زیر است:

$$\mathcal{N}_i(x)=(2\pi)^{-\frac{d}{2}}\left|\Sigma_i\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\} \quad (۶)$$

$$\sum_{i=1}^{M} w_i = 1 \quad (۷)$$

۳-۱-۱-روش GMM-UBM

در این روش، مدل آکوستیکی جهانی UBM[1] (مدل تعلیم‌یافته روی دادگان کلیه زبان‌ها)، یک مدل مخلوط گوسی (GMM) است که با استفاده از معیار ML تعلیم داده می‌شود. در الگوریتم تعلیم استاندارد ML، مجموع لگاریتم احتمال کل دادگان تعلیم، به‌عنوان تابع هدف در نظر گرفته می‌شود و هدف بیشینه‌کردن این تابع است:

$$F_{ML}(\lambda)=\sum_{r=1}^{R}\log\, p(x_r \mid l_r) \quad (۸)$$

در این رابطه λ پارامترهای مدل، x_r نشانگر r امین گفتار تعلیم، R تعداد گفتارهای تعلیم و l_r برچسب صحیح داده تعلیم r ام است. برای بیشینه‌کردن تابع هدف، پارامترهای GMM به‌صورت تکراری و با استفاده از روابط تخمین الگوریتم EM محاسبه می‌شود.

مدل آکوستیکی هر یک از زبان‌ها نیز یک مدل مخلوط گوسی (GMM) است که از طریق تطبیق دادن مدل جهانی UBM بر روی تمام دادگان آن زبان خاص (با استفاده از روش تطبیق Relevance MAP[2]) آموزش می‌بیند (به‌صورت معمول فقط مقادیر بردارهای میانگین توابع گوسی تطبیق داده می‌شود و وزن‌ها و ماتریس‌های کوواریانس تطبیق داده نمی‌شوند). استفاده از روش UBM-GMM نسبت به GMM موجب افزایش سرعت محاسبات و بهترشدن دقت نتایج نهایی می‌شود [38]، [8]، [42]، [34]. در این مقاله ماتریس کوواریانس، قطری فرض شده و تعداد توابع گوسی (M در رابطه (۵))، برابر با ۱۰۲۴ در نظر گرفته شده است. مقدار اولیه عناصر قطری ماتریس کوواریانس (یا همان واریانس‌ها) در اولین تکرار از الگوریتم تعلیم UBM (الگوریتم EM) برابر با واریانس کل دادگان است و در طی فرایند تخمین، مقدار تخمین‌زده‌شدهٔ واریانس نباید از مقدار کف واریانس[3] کمتر شود. مقدار کف واریانس در هر بُعد به‌طور نسبی تعیین می‌شود و برابر است با ضریبی از واریانس کل دادگان آموزشی. هر چه میزان کف واریانس کمتر در نظر گرفته شود، مدل GMM روی دادگان تعلیم بیشتر آموزش یافته و قابلیت تعمیم آن روی دادگان آزمایش کمتر می‌شود. در این مقاله مقدار نهایی کف واریانس برای آموزش UBM برابر ۰/۰۱ در نظر گرفته شده است (به‌صورت نسبی و نسبت به واریانس کل دادگان آموزشی).

۳-۱-۲-استفاده از الگوریتم تمایزی MMI

برای متمایزترشدن مدل‌های گوسی زبان‌ها از الگوریتم‌های تعلیم تمایزی استفاده می‌شود. در تعلیم تمایزی، تابع هدف به نحوی طراحی می‌شود که دقت بازشناسی بهبود یابد. یکی از الگوریتم‌های تعلیم تمایزی مشهور، الگوریتم تعلیم MMI است که در آن تابع هدف، احتمال پسین تشخیص صحیح گفتارهای تعلیم است [46]:

$$F_{MMI}(\lambda)=\sum_{r=1}^{R}\log\frac{p_\lambda(x_r \mid l_r)p(l_r)}{\sum_{\forall r} p_\lambda(x_r \mid l_r)p(l_r)} \quad (۹)$$

غالباً احتمال پیشین کلیه طبقه‌ها یکسان در نظر گرفته می‌شود. بنابراین جمله‌های $p(l_r)$ و $p(l)$ از رابطهٔ بالا حذف می‌شوند. مخرج رابطه ($\sum_{\forall r} p_\lambda(x_r \mid l_r)p(l_r)$) نیز، احتمال گفتار x_r با درنظرگرفتن کلیه مدل‌های رقیب (کلیه زبان‌ها) است [24]. در ادامه روش‌هایی برای بهبود کیفیت سامانهٔ شناسایی زبان آکوستیکی با روش MMI شرح داده شده است.

*** تعلیم مدل‌های وابسته به جنسیت گوینده**

یک روش برای بهبود نتایج مدل‌های گوسی تعلیم‌یافته با

[1] Universal Background Model
[2] Relevance Maximum A Posteriori Adaptation
[3] Final Variance Floor

الگوریتم MMI، استفاده از مدل‌های وابسته به جنسیت گوینده[1] است. بدین منظور دادگان تعلیم بر حسب جنسیت به دو دسته تقسیم‌بندی شده و برای هر دسته به‌طور مجزا مدل‌های گوسی با الگوریتم MMI تعلیم داده شده است.

*** اعمال وزن در رابطه MMI**

برای جلوگیری از متمایل‌شدن مدل‌ها به زبان‌هایی که تعداد دادگان تعلیمِ زیاد دارند، در رابطه بهینه‌سازی تابع هدف، به هر داده مقدار وزن زیر اعمال می‌شود:

$$w_l^i = \frac{1}{N_l * M_l^i} \qquad (۱۰)$$

در این رابطه N_l تعداد دادگان زبان l، M_l^i تعداد قاب‌های دادهٔ iام از زبان l و w_l^i وزن اعمال‌شده بر این داده است.

*** مدل‌کردن پویای زبان با HMM**

یک روش برای بهبود کیفیت مدل‌های تعلیم‌یافته به روش MMI، جایگزین‌کردن GMM با مدل پنهان مارکوف ارگودیک (EHMM[2]) است. از آنجا که مدل گوسی نمی‌تواند به‌خوبی رفتار پویای گفتار را مدل کند، تبدیل مدل هر زبان به یک مدل چندحالته در HMM می‌تواند موجب بهبود کیفیت سامانه شناسایی زبان شود. در این مقاله هر مدل GMM تعلیم‌یافته با معیار MMI به یک مدل EHMM با اتصال کامل[3] تبدیل شده است. برای این کار کافی است، مقادیر هر یک از مؤلفه‌های گوسی از GMM به‌عنوان یک حالت جدا در EHMM فرض شود؛ سپس توسط الگوریتم تعلیمِ ML، تنها مقادیر احتمالات گذر[4] بین حالات مختلف HMM را تعلیم داده و سایر پارامترهای توابع گوسی (میانگین‌ها و واریانس‌ها) ثابت نگه داشته می‌شود [48]، [24].

۳-۲-روش GSV-SVM

روش GSV-SVM نیز یک روش طیفی برای حل مسئلهٔ بازشناسی زبان است. در این روش، از هر داده تعلیم (یک قطعه ۳۰ ثانیه‌ای از گفتار) یک آبَر بُردار گوسی استخراج شده و این آبَر بُردارها توسط SVM طبقه‌بندی می‌شوند. مفهوم آبَربردار به‌زودی توضیح داده خواهد شد. در ادامه ابتدا روش طبقه‌بندی SVM و سپس نحوه استخراج آبَر بردارها شرح داده شده است.

[1] Gender Dependent Models
[2] Ergodic HMM
[3] Fully Connected
[4] Transition Probabilities

۳-۲-۱-روش SVM

روش SVM یک روش تمایزی برای حل مسائل شناسایی الگو است. ایدهٔ اصلی SVM تنظیم توابع تمایز به‌صورتی است که به‌طور بهینه از اطلاعات الگوهای مرزی (نمونه‌های واقع‌شده در مرز جداکنندهٔ دو طبقه) استفاده کند. به بیان دیگر SVM یک آبَر مرز خطی یا غیرخطی را پیدا می‌کند که اول‌این‌که طبقه‌ها را به بهترین صورت جدا کند (برای کاهش خطای دسته‌بندی) و دوم‌این‌که پهنای ناحیه مرزی جداکنندهٔ طبقه‌ها نیز بیشینه باشد (برای افزایش میزان تعمیم طبقه‌بندی‌کننده). به الگوهای انتخاب شده و سهیم در تعیین مرز طبقه‌ها بردار مرزی[5] گفته می‌شود [30]. درواقع، SVM استاندارد یک طبقه‌بندی‌کننده دودویی است که از ترکیب خطی وزن‌دار توابع هسته[6] $k(.,.)$ تشکیل می‌شود:

$$f(x) = \sum_{i=1}^{N_{sv}} \alpha_i y_i K(x, x_i) + b \qquad (۱۱)$$

$$\sum_{i=1}^{N_{sv}} \alpha_i y_i = 0, \quad \alpha_i > 0, \qquad (۱۲)$$

در این رابطه N_{sv} تعداد بردارهای مرزی، y_i خروجی مطلوب (یک از مقادیر $\{-1, +1\}$) و α_i وزن بردار مرزی و درواقع ضریب لاگرانژ متناظر با نمونه x_i است [34]، [30]. برای تعلیم SVM باید مسأله بهینه‌سازی مقید زیر حل شود:

$$\min_{w} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \qquad (۱۳)$$

$$w.r.t. \quad y_i[w^T \phi(x_i) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0 \qquad (۱۴)$$

در این رابطه بردار x_i توسط تابع غیرخطی ϕ به فضایی با بُعد بیشتر نگاشت شده و به $\phi(x_i)$ تبدیل می‌شود [23]. پارامتر C هزینه خطا است که هر قدر مقدر آن بیشتر باشد، اهمیت خطای طبقه‌بندی بیشتر شده و در نتیجه میزان تعمیم‌پذیری SVM آموزش‌یافته کاهش می‌یابد.

تابع هسته به‌صورت ضرب داخلی $\phi(x)$ و $\phi(x_i)$ تعریف می‌گردد (ϕ یک تابع نگاشت است) [14].

$$K(x, x_i) = \phi(x)^t \phi(x_i) \qquad (۱۵)$$

یکی از بهترین روش‌های ساختن طبقه‌بندی‌کنندهٔ

[5] Support Vectors
[6] Kernel Functions

چندطبقه[1] مبتنی بر SVM، استفاده از مجموعه‌ای از طبقه‌بندی‌کننده‌های SVM دودویی است. در این مقاله از روش یکی در مقابل بقیه[2] استفاده شده است [45].

۳-۲-۲-آبَر بُردار گوسی

گام نخست در روش GSV-SVM، تعلیم یک مدل مخلوط گوسی روی دادگانی از کلیه زبان‌های تعلیم و درواقع ساختن یک مدل مخلوط گوسی جهانی از نوع GMM به نام UBM مطابق با رابطه (۵) است.

گام بعدی در روش GSV-SVM استخراج یک مدل مخلوط گوسی برای هر پوشه گفتاری از طریق تطبیق UBM به آن پوشه گفتاری و به‌دست آوردن یک GMM تطبیق‌داده‌شده به‌ازای هر پوشه گفتاری است. این مدل مخلوط گوسی، با تطبیق مدل UBM به هر پوشه گفتاری به‌طور جداگانه (با استفاده از الگوریتم Relevance MAP) به‌دست می‌آید. به‌طور معمول، تنها بردارهای میانگین (μ_i) توابع گوسی تطبیق داده می‌شود؛ بنابراین، GMMهای به‌دست‌آمده از کلیه پوشه‌های گفتاری، ماتریس کوواریانس (Σ_i) و وزن یکسان دارند و تنها تفاوت آنها در مقدار میانگین است. از کنار هم قراردادن بردارهای میانگین GMM تطبیق‌داده‌شده به هر پوشه گفتاری، یک آبَر بُردار گوسی به‌ازای آن پوشه گفتاری تولید می‌شود؛ سپس این ابربردارهای گوسی توسط SVM طبقه‌بندی می‌شوند [16]، [11]، [10].

۳-۲-۳-فاصله بین GMMها

اگر p^a و p^b مدل‌های احتمالاتی متناظر با گفتارهای s_a و s_b باشند، فاصلهٔ بین این دو مدل را می‌توان با معیار واگرایی KL[3] به‌صورت زیر تعریف کرد:

$$KL(p^a \parallel p^b) = \int_{R^n} p^a(x) \log\left(\frac{p^a(x)}{p^b(x)}\right) dx \qquad (۱۶)$$

معیار فاصله KL2، شکل متقارنی از معیار واگرایی KL است که به‌صورت زیر تعریف می‌شود:

$$KL2(p^a \parallel p^b) = KL(p^a \parallel p^b) + KL(p^b \parallel p^a) \qquad (۱۷)$$

حد بالای معیار واگرایی KL بین دو مدل GMM در رابطهٔ (۱۸) نوشته شده است. در این رابطه، ${\mu^a}_i$ و ${\mu^b}_i$ به‌ترتیب نشان‌گر بردار میانگین مربوط به گوسین i ام، در مدل GMM تطبیق‌داده‌شده به گفتارهای s_a و s_b است.

$$KL(p^a \parallel p^b) \le KL(w^a \parallel w^b) + \sum_{i=1}^{N} {w_i}^a KL(\mathcal{N}(.;{\mu_i}^a,{\Sigma_i}^a) \parallel \mathcal{N}(.;{\mu_i}^b,{\Sigma_i}^b)) \qquad (۱۸)$$

اگر تطبیق پارامترهای مدل گوسی هر گفتار با استفاده از الگوریتم MAP و تنها بر روی میانگین‌های مدل‌های گوسی صورت گرفته باشد (در این صورت $w^a = w^b$ و $\Sigma_i^a = \Sigma_i^b, i = 1..N$ خواهد بود)، حد بالای معیار فاصلهٔ KL2 می‌تواند به‌صورت زیر باشد:

$$KL2(p^a \parallel p^b) \le \sum_{i=1}^{N} w_i ({\mu_i}^a - {\mu_i}^b)^t {\Sigma_i}^{-1} ({\mu_i}^a - {\mu_i}^b) \le {D_e}^2(\mu^a, \mu^b) \qquad (۱۹)$$

$${D_e}^2(\mu^a,\mu^b) = \sum_{i=1}^{N} w_i ({\mu_i}^a - {\mu_i}^b)^t {\Sigma_i}^{-1} ({\mu_i}^a - {\mu_i}^b) \qquad (۲۰)$$

درصورتی‌که ماتریس کوواریانس قطری باشد، فاصله ${D_e}^2$ بین ابر بردارهای GMM μ^a و μ^b نشان‌گر فاصله اقلیدسی وزن‌دارِ بین ابر بردارهای GMM مقیاس شده است. فاصلهٔ ${D_e}^2$ حد بالای فاصلهٔ KL است. بنابراین اگر فاصلهٔ ابر بردارهای GMM μ^a و μ^b کم باشد، معیار فاصله KL2 متناظر آنها نیز کم خواهد بود [14]، [11]، [10].

۳-۲-۴-هسته خطی و غیر خطی

برای تعیین فاصله بین دادگان هدف از سایر دادگان‌ها (غیر هدف)، نخستین مرحله تعیین تابع هسته مناسب است. معیار فاصلهٔ بین دو مدل گوسی به‌دست آمده از تطبیق در رابطه (۲۰) را می‌توان به ضرب داخلی (تابع هسته خطی) تبدیل کرد. با اعمال هستهٔ خطی، فاصله ${D_e}^2$ به‌دست خواهد آمد.

$$K_{Lin}(s_a, s_b) = \sum_{i=1}^{N} w_i ({\mu_i}^a)^t {\Sigma_i}^{-1} {\mu_i}^b = \sum_{i=1}^{M} \left(\sqrt{w_i}{\Sigma_i}^{-1/2}{\mu_i}^a\right)^t \left(\sqrt{w_i}{\Sigma_i}^{-1/2}{\mu_i}^b\right) \qquad (۲۱)$$

تابع هسته K_{Lin} در فضای ابر بردارهای GMM، خطی است و فضای ویژگی جدید با مقیاس‌کردن آبَربردارهای GMM (ضرب ابر بردارهای GMM در $\sqrt{w_i}\Sigma_i^{-1/2}$) به‌دست می‌آید. در رابطه (۲۲)، F_i مقدار میانگین مقیاس‌شده تابع گوسی i ام و SV ابر بردار

[1] Multi-Class
[2] One Versus the Rest
[3] Kullback-Leibler

GMM مقیاس شده است. این رابطه بدین معناست که در روش GSV-SVM بردارهای میانگین به‌دست آمده از هر گوینده را طبق رابطهٔ (۲۲) مقیاس‌دهی می‌کنیم و ابر بردار تشکیل‌شدهٔ جدید را به یک SVM استاندارد می‌دهیم.

$$F_i = \sqrt{w_i}\Sigma_i^{-1/2}\mu_i \qquad i = 1,2,...,N$$
$$SV = [F_1, F_2, \cdots, F_N] \tag{۲۲}$$

آنگاه:

$$K_{linear}(SV^a, SV^b) = (SV^a)^t SV^b \tag{۲۳}$$

یک روش برای به‌دست‌آوردن هسته غیرخطی استفاده از تابع نمایی است:

$$K_{Non-linear}(s_a, s_b) = e^{-\gamma . D_c^2(\mu^a, \mu^b)} \tag{۲۴}$$

با قرار دادن رابطه (۲۰) در رابطهٔ بالا و ساده‌سازی داریم:

$$K_{Non-linear}(SV^a, SV^b) = e^{-\gamma\left\|SV^a - SV^b\right\|^2} \tag{۲۵}$$

۳-۲-۵-امتیازدهی به روش برگشت به حوزه مدل

در روش SVM استاندارد با استفاده از رابطهٔ (۱۱) مقدار امتیاز دادهٔ آزمون محاسبه می‌شود. روش برگشت به حوزهٔ مدل (MP) یک روش جدید و بسیار جالب است که با تغییر شیوهٔ امتیازدهی، موجب بهبود کیفیت (برای فایل‌های کوتاه) و افزایش سرعت می‌شود. در این روش با استفاده از مجموعهٔ بردارهای پشتیبان به‌دست‌آمده از طبقه‌بندی کنندهٔ SVM، یک GMM به‌ازای هر طبقه ساخته می‌شود و سپس مرز جداسازی توزیع دو طبقه با استفاده از این دو مدل GMM ساخته می‌شود. ابتدا تابع امتیازدهی معمول SVM به دو بخش تقسیم می‌شود:

$$f(x) = \sum_{\{i|\alpha_i>0\}} \alpha_i K(x, x_i) - \sum_{\{i|\alpha_i<0\}} \alpha_i K(x, x_i) + d \tag{۲۶}$$

دو جملهٔ سمت راست رابطهٔ بالا بیان‌گر نسبت لگاریتم شباهت بین مدل‌های درون‌طبقه‌ای و برون‌طبقه‌ای است. در واقع می‌توان بردارهای پشتیبان با مقدار $\alpha_i > 0$ را به‌عنوان دادگان درون‌طبقه‌ای و بردارهای پشتبان با مقدار $\alpha_i < 0$ را به‌عنوان دادگان برون‌طبقه‌ای در نظر گرفت. برای به‌دست‌آوردن دو مدل گوسی جدید، یعنی مدل گوسی مثبت و مدل گوسی منفی، ابتدا روابط زیر محاسبه می‌شود [9]:

سپس آَبربردارهای جدیدِ x_p و x_n با انجام عملیات عکسِ استخراج آَبر بردار، به مدل‌های گوسی g_p و g_n تبدیل می‌شوند. به‌طور شهودی به‌نظر می‌رسد که آَبربردارهای استخراج‌شده بر مرز جداسازی دادگان قرار دارند. بنابراین

$$x_p = \frac{1}{\sum_{\{i|\alpha_i>0\}}\alpha_i}\sum_{\{i|\alpha_i>0\}}\alpha_i x_i$$
$$x_n = \frac{1}{\sum_{\{i|\alpha_i<0\}}\alpha_i}\sum_{\{i|\alpha_i<0\}}\alpha_i x_i \tag{۲۷}$$

مدل‌های گوسی مثبت g_p و گوسی منفی g_n، محل مرزهای مثبت و منفی را مدل می‌کنند. روش امتیازدهی برای یک دادهٔ آزمایش به‌صورت زیر است:

$$score = \sum_t \log(g_p(y_t)) - \sum_t \log(g_m(y_t)) \tag{۲۸}$$

در این رابطه y_t بردار ویژگی قاب t ام از داده‌ی آزمایش است [9]. استفاده از این روش دو فایده مهم دارد:

الف-در دادگان با طول زمانی کم که مرحلهٔ تطبیق مدل گوسی با دقت کم صورت می‌گیرد، آَبربردار گوسی استخراج‌شده دقیق نخواهد بود و درنتیجه مقدار احتمال به‌دست‌آمده از طبقه‌بندی‌کنندهٔ SVM معتبر نخواهد بود؛ ولی با استفاده از روش MP تنها با استفاده از محاسبهٔ احتمال گوسی‌هایِ بیانگرِ مرزِ بین طبقه مورد نظر و طبقه-های دیگر (گوسی‌های کلاس ۱+ و ۱-)، می‌توان احتمال تعلق دادهٔ آزمایش را به طبقه مورد نظر به‌دست آورد.

ب-از آنجا که مرحلهٔ تطبیق مدل گوسی و استخراج آَبر بردار گوسی حذف می‌شود و نیازی به محاسبهٔ احتمال تعلق آَبر بردارها به طبقه‌های ۱+ و ۱- در طبقه‌بندی‌کننده‌های SVM نیست، لذا سرعت اجرای برنامه افزایش می‌یابد.

۳-۳-روش GMM Tokenizer

یکی از روش‌های معمول در بازشناسی زبان، استفاده از سامانه‌های مبتنی بر واحدهای آوایی مانند PPRLM است. در این روش‌ها از یک بازشناس آوای موازی و یک مدل زبانی برای هر زبان استفاده می‌شود. بازشناس آوا برای تخمین‌زدن رشتهٔ آوای گفتار و مدل زبانی برای تخمین احتمال تعلق این رشته آوا به زبان مورد نظر استفاده می‌شود. با وجود این‌که این روش نتایج خوبی دارد، ولی تعلیم بازشناس زبان بسیار زمان‌بر است و از نظر محاسباتی نیز گران است؛ زیرا در این روش باید تعداد بسیار زیادی دادگان آوا نویسی‌شده

از هر زبان در دسترس باشد.

در روش GMM Tokenizer، بازشناس آوا با یک مدل گوسی جایگزین می‌شود. این مدل گوسی به‌عنوان تجزیه‌کننده گفتار (Tokenizer) عمل می‌کند. برای این کار هر دادهٔ گفتاری مورد پردازش قرار گرفته و احتمال تعلق هر قاب از آن به هر یک از مؤلفه‌های مدل گوسی محاسبه می‌شود. برای هر قاب، اندیس مؤلفهٔ گوسیِ با احتمال بیشتر ذخیره شده و درنهایت برای هر داده گفتاری، رشته‌ای از اندیس‌های مؤلفه‌های گوسیِ با احتمال بیشتر محاسبه می‌شود. بدین ترتیب در این روش مدل گوسی با تخصیص یک اندیس به جای هر قاب گفتاری نقش بازشناس آوا را اجرا می‌کند و گفتار را بدون نیاز به دادگان برچسب‌دار تجزیه می‌کند [21].

اندیس‌های تولیدشده از یک قطعه یا یک فایل از سیگنال گفتار را می‌توان به یک بردار M بُعدی با نام p تبدیل کرد. M تعداد مؤلفه‌های مدل گوسی است. مقدار هریک از مؤلفه‌های این بردار به‌صورت زیر محاسبه می‌شود:

$$p_j = \frac{\text{count}(C_j)}{\sum_i \text{count}(C_i)} \qquad (۲۹)$$

در این رابطه منظور از C_j، j امین تابع گوسی است و p_j با شمارش تعداد نسبی رخداد اندیس مورد نظر در کل قاب‌های آن پوشه گفتاری محاسبه می‌شود. مقدار سیگما در رابطه بالا روی تمامی اندیس‌ها محاسبه می‌شود. پارامتر $Count(C_j)$ نیز تعداد دفعاتی است که اندیس تابع گوسی شماره C_j ام در رشته اندیس‌ها اتفاق افتاده است. بدین ترتیب برای هر داده گفتاری یک بردار M بُعدی تولید می‌شود.

فرض کنیم p^{tar} و p^{bkg} به‌ترتیب بردارهای احتمال مربوط به زبان‌های هدف و غیر هدف باشد، در این صورت یک طبقه‌بندی‌کنندهٔ N-gram معمول یک ابر صفحه h خواهد داشت که بردارهای هدف و غیر هدف را از هم جدا می‌کند:

$$h_i = \log(p_i^{tar} \;/\; p_i^{bkg}) \qquad (۳۰)$$

طبقه‌بندی‌کنندهٔ N-gram خطی است. به جای آن می‌توان با استفاده از روش طبقه‌بندی SVM با انواع مختلفی از توابع هسته، ابرصفحهٔ h را به‌نحو مناسب‌تری تخمین زد [21].

*** مدل ترکیب گوسی چند زبانی (MLM[1])**

در روش GMM Tokenizer، افزایش تعداد مؤلفه‌های گوسی باعث می‌شود که اطلاعات زبانی در مؤلفه‌های جداگانه‌ای بیان شوند. درصورتی‌که بخواهیم تعداد مؤلفه‌های گوسی را افزایش دهیم، به تعداد بسیار بیشتری داده برای تعلیم مدل نیازمندیم و از طرف دیگر زمان محاسبات نیز به‌شدت افزایش پیدا می‌کند. برای رفع این مشکل از MLM استفاده می‌شود. MLM ترکیب مدل‌های گوسی مستقل از زبان‌های مختلف است:

$$p(\mathrm{x}_t \mid \lambda_l) = \sum_{\forall i} \omega_l^i \mathcal{N}(\mathrm{x}_t; \mu_l^i; \Sigma_l^i) \qquad (۳۱)$$

در رابطهٔ بالا، $\lambda_l = \{\omega_l, \mu_l, \Sigma_l\}$ پارامترهای مدل گوسی زبان lام است. ω_l^i، μ_l^i و Σ_l^i به‌ترتیب وزن، میانگین و کوواریانس i امین مؤلفه گوسی مربوط به مدل GMM زبان l ام است و x_t بردار ویژگی فریم tام است. درصورتی‌که فاصلهٔ KL بین برخی مؤلفه‌های گوسی از مقدار خاصی کمتر باشد، باید مؤلفه‌های مشابه با هم ترکیب و به یک مؤلفه تبدیل شود.

برای تشکیل MLM، پارامترهای مدل‌های مستقل به‌سادگی و تنها با تقسیم مقدار وزن مدل‌های گوسی بر تعداد مدل‌های زبانی (L) با هم ترکیب می‌شوند.

$$\lambda_{MLM} = \{\frac{[\omega_1,...,\omega_L]}{L}, [\mu_1;...\mu_L], [\Sigma_1;...\Sigma_L]\} \qquad (۳۲)$$

در این رابطه λ_{MLM} پارامترهای مدل MLM تمام زبان‌ها است [21].

۳-۴-روش JFA

دادگان مورد استفاده در سامانه شناسایی زبان تنها شامل اطلاعات مفید نیست؛ بلکه این دادگان اطلاعات دیگری نیز دارد که کمکی به شناسایی زبان نمی‌کند. از جمله این اطلاعات می‌توان به اطلاعات مربوط به هویت گوینده و اطلاعات کانال اشاره کرد که آن را تنوعات ناخوسته یا اطلاعات غیرمفید می‌نامیم. در روش تحلیلِ توأمِ عامل‌ها (JFA) برای رفع این مشکل، تنوعات ناخواسته و اطلاعات وابسته به زبان از همدیگر جدا می‌شود. در این روش، فرض می‌شود که اطلاعات زبانی بخشی است که در تمامی دادگان یکسان و بخش باقی‌مانده مربوط به تنوعات است و اثر اطلاعات مفید و غیر مفید به‌صورت جمع قابل بیان است.

در مرحله آزمایش تنوعات موجود در هر دادهٔ آزمایش تخمین زده می‌شود و با توجه به تنوعات دادگان تعلیم،

[1] Multi Language Model

حذف می‌شود. با حذف تنوعات ناخواسته، بخش باقی‌مانده حاوی اطلاعات مفید است و درنتیجه طبقه‌بندی به‌نحو بهتری صورت می‌گیرد [48]، [29].

فرض‌کنیم اَبَربردارگوسی (m_l) برای تمامی زبان‌های l، به‌صورت آماری مستقل بوده (نسبت به l) و دارای توزیع پیشین به‌صورت نرمال با میانگین m و کوواریانس $DD^T = (\Sigma / \tau)$ باشد، و m و Σ پارامترهای GMM-UBM و τ عامل رابطه[1] در الگوریتم استاندارد Relevance-MAP باشد. متغیر تصادفی m_l را می‌توان به‌صورت زیر نوشت:

$$m_l = m + Dy_l \quad (۳۳)$$

در این رابطه y_l یک بردار تصادفی پنهان با توزی نرمال استاندارد چندبُعدی $\mathcal{N}(0,I)$ است. با استفاده از دادگانی از زبان l، الگوریتم MAP توزیع پسین m_l را تخمین می‌زند.

حال مجموعه‌ای از ابربردارهای دادگان زبان l، یعنی $m_{(h,l)}$، با شرایط ضبط $h\,(h=1,2,...)$ را در نظر می‌گیریم و فرض می‌کنیم برای یک زبان l، ابر بردارهای گوسی $m_{(h,l)}$ به‌صورت آماری مستقل باشند (نسبت به h) و احتمال پیشین آنها نرمال باشد. در این صورت ماتریسی با رتبهٔ پایین با نام U برای هر ضبط h وجود خواهد داشت:

$$m_{(h,l)} = m_l + Ux_{(h,l)} \quad (۳۴)$$

در این رابطه $x_{(h,l)}$ بردار متغیرهای تصادفی پنهان با توزیع نرمال $\mathcal{N}(0,I)$ است. با داشتن دادگان تطبیق از زبانِ l و ضبط شده از کانالِ h، الگوریتم تطبیق MAP میانگین توزیع را برای $x_{(h,l)}$ محاسبه می‌کند. برای درنظرگرفتن ترکیب تنوعات زبان و تنوعات ناخواسته روابط (۳۴) و (۳۵) با هم ترکیب می‌شوند. مدل نهایی عبارت است از:

$$m_{(h,l)} = m + Dy_l + Ux_{(h,l)} \quad (۳۵)$$

درصورتی‌که مدل UBM دارای M مؤلفه‌ی گوسی باشد و بُعد بردارهای ویژگی d باشد، در این رابطه، $m_{(h,l)}$ ابر بردار وابسته به زبان و تنوعات ناخواسته، D ماتریس قطری ($Md \times Md$)، y_l بردار فاکتور زبان (بردار Md بُعدی)، U ماتریس تنوعات با رتبهٔ n_x ($Md \times n_x$) و $x_{(h,l)}$ عوامل تنوعات ناخواسته (بردار n_x بُعدی) است. در این روابط به‌صورت نظری فرض می‌شود که $x_{(h,l)}$ مستقل از زبان است و سایر تنوعات غیر زبان را در بر می‌گیرد.

هر دو متغیر $x_{(h,l)}$ و y_l توزیعی نرمال دارند. ماتریس $DD^T = (\Sigma / \tau)$ تنوعات ابر بردار میانگین زبان را نشان می‌دهد و ماتریس کوواریانس بین‌طبقه‌ای[2] است. ماتریس UU^T تنوعات ناخواسته را نشان می‌دهد و ماتریس کوواریانس درون‌طبقه‌ای[3] است. از آنجا که فرض شده تنوعات گوینده و کانال مستقل هستند، تنوعات کلی برابر با $DD^T + UU^T$ است. بنابراین، توزیع پیشین تمامی ابر بردارها با میانگین m و کوواریانس $DD^T + UU^T$ بیان می‌شود [2].

۳-۴-۱-مرحله تعلیم روش JFA

برای تعلیم روش JFA، ابتدا یک مدل مخلوط گوسی روی دادگانی متشکل از کلیه دادگان تمامی زبان‌های تعلیم با استفاده از الگوریتم EM تعلیم داده می‌شود که همان مدل UBM است. از کنار هم قرارگرفتن بردارهای میانگین این مدل مخلوط گوسی، مدل مستقل از تنوعات ناخواسته و تنوعات زبان به‌دست می‌آید (m در رابطهٔ (۳۶)). سپس با استفاده از قطعاتی از زبان‌های تعلیم که توسط گویندگان مختلف و در شرایط مختلف ضبط شده است، پارامترهای مدل با استفاده از الگوریتم امیدریاضی-بیشینه‌سازی (الگوریتم EM) قابل تخمین است (هر فایل گفتاری بزرگ به تعدادی پوشه‌های کوچک (قطعات ۳۰ ثانیه‌ای) تقسیم می‌شود):

* مرحلهٔ E: تخمین y_l و $x_{(h,l)}$ با استفاده از الگوریتم MAP انجام می‌شود. y_l برای سگمنت‌های هر زبان ثابت نگاه داشته می‌شود.
* مرحلهٔ M: پارامترهای مدل به‌نحوی به روز می‌شود که احتمال روی دادگان تعلیم بیشتر شود. برای این کار تابع کمکی EM بیشینه می‌شود.

برای به‌دست‌آوردن نتیجه مطلوب از روش JFA باید تخمین خوبی از ماتریس U با استفاده از تعداد مناسبی از دادگان تعلیم (ضبط‌های مختلفی از هر زبان) محاسبه شود. در ادامه جزئیات پیاده‌سازی JFA توضیح داده شده است.

۳-۴-۱-۱-آمارگان کافی[4]

برای تخمین متغیرهای پنهان و ماتریس U از آمارگان کافی

[1] Relevance Factor

[2] Across-Class Covariance Matrix
[3] Within-Class Covariance Matrix
[4] Sufficient Statistics

استفاده می‌شود. این آمارگان کافی، مُمان‌های مرتبهٔ نخست و صفرم (بر پایه مدل UBM) است. اگر $N(l)$ و $N(h,l)$ به‌ترتیب شامل آمارگان وابسته به زبان و تنوعات مزاحم باشد داریم:

$$N_g(l) = \sum_{f\in l} \gamma_g(f); \quad N_g(h,l) = \sum_{f\in(h,l)} \gamma_g(f), \qquad (۳۶)$$

در این رابطه $\gamma_g(f)$ احتمال پسین گوسی g برای بردار کپسترال f است. در این رابطه، $\Sigma_{f\in l}$ مجموع را روی تمامی قاب‌های زبان l و $\Sigma_{f\in(h,l)}$ مجموع را روی تمامی قاب‌های متعلق به ضبط hام از زبان l حساب می‌کند. اگر $X(l)$ و $X(h,l)$ به‌ترتیب بردارهای شامل آمارگان مرتبهٔ نخست وابسته به زبان و تنوعات مزاحم باشد، بُعد این ابر بردارها برابر با Md است و به‌صورت زیر محاسبه می‌شود [31]:

$$X_g(l) = \sum_{f\in l} \gamma_g(f).f; \quad X_g(h,l) = \sum_{f\in(h,l)} \gamma_g(f).f, \qquad (۳۷)$$

۳-۴-۱-۲-تخمین متغیرهای پنهان

متغیرهای $x(h,l)$ و $y(l)$ در این زیر بخش، بیان‌گر تخمین MAP نقطه‌ای[1] از عوامل تنوعات $x_{(h,l)}$ و زبان y_l است. اگر $\overline{X}(l)$ و $\overline{X}(h,l)$ به‌ترتیب آمارگان وابسته به زبان و وابسته به تنوعات باشد، می‌توان تعریف کرد:

$$\begin{aligned} \overline{X}_g(l) &= X_g(l) - \sum_{h\in l} N_g(h,l).\{m + Ux(h,l)\}_g \\ \overline{X}_g(h,l) &= X_g(h,l) - N_g(h,l).\{m + Dy(l)\}_g \end{aligned} \qquad (۳۸)$$

در این رابطه، $\overline{X}(l)$ برای تخمین بردار زبان (تنوعات حذف شده است) و $\overline{X}(h,l)$ برای تخمین عوامل کانال (تنوعات ناخواسته غیر زبانی) استفاده می‌شود.

اگر $L(h,l)$ یک ماتریس $n_x \times n_x$ بُعدی و $B(h,l)$ یک بردار n_x بُعدی باشد:

$$\begin{aligned} L(h,l) &= I + \sum_{g\in UBM} N_g(h,l).U_g^T.\Sigma_g^{-1}.U_g \\ B(h,l) &= \sum_{g\in UBM} U_g^T.\Sigma_g^{-1}.\overline{X}_g(h,l) \end{aligned} \qquad (۳۹)$$

در این رابطه، Σ_g ماتریس کوواریانس مربوط به مؤلفهٔ g ام UBM است. با استفاده از $B(h,l)$ و $L(h,l)$، می‌توان $x(h,l)$ و $y(l)$ (یا همان تخمین MAP نقطه‌ای) را به‌دست آورد [31].

۳-۴-۱-۳- تخمین ماتریس U

$$\begin{aligned} x(h,l) &= L(h,l)^{-1}.B(h,l) \\ y_g(l) &= \frac{\tau}{\tau + N_g(l)}.D_g.\Sigma_g^{-1}.\overline{X}_g(l) \\ D_g &= \frac{1}{\sqrt{\tau}}\Sigma_g^{1/2} \end{aligned} \qquad (۴۰)$$

اگر U_g^i، i امین ردیف U_g باشد، برای تخمین U_g داریم:

$$\begin{aligned} U_g^i &= L(g)^{-1}.R^i(g), \\ L(g) &= \sum_l \sum_{h\in l} N_g(h,l).(L(h,l)^{-1} + x(h,l)x(h,l)^T) \\ R^i(g) &= \sum_l \sum_{h\in l} \overline{X}_g(h,l)[i].x(h,l) \end{aligned} \qquad (۴۱)$$

الگوریتم ۱، مراحل تخمین ماتریس U را با استفاده از دادگان مستقل از زبان‌های مختلف با تنوعات مختلف بیان می‌کند. برای تخمین x و y نیز از همین الگوریتم (بدون تکرار) و با ثابت در نظر گرفتن U استفاده می‌شود [19].

۳-۴-۲- مرحله آزمایش روش JFA

با اعمال JFA بر داده آزمایش می‌توانیم بنویسیم:

$$m(h_{test},l) = m + Dy_l + Ux_{h_{test}} \qquad (۴۲)$$

برای دنباله‌ای از ویژگی‌های گفتاری $Y = \{y_1,\ldots,y_T\}$، لگاریتم نسبت احتمال (LLR[2]) با استفاده از رابطهٔ زیر محاسبه می‌شود. در این رابطه $P(.\mid m)$ احتمال قاب‌های داده آزمایش بر روی مدل مخلوط گوسی با ابر بردار میانگینِ m است [2].

$$score(Y \mid l) = \log\left(\frac{P(Y \mid m + Dy_l + Ux_{l_{test}})}{P(Y \mid m + Ux_{l_{test}})}\right) \qquad (۴۳)$$

در این زیر بخش تلاش شد که کلیات روش JFA بیان شود. از آنجا که توضیح بیشتر روابط در حوصله این مقاله نیست، برای فهم بیشتر روابط به مرجع بسیارخوب [31] رجوع شود.

[1] MAP Point Estimate

[2] Log Likelihood Ratio

(الگوریتم-۱): الگوریتم تخمین ماتریس زیر فضای تنوعات در روش JFA

(Algorithm-1): Estimation of total variability matrix in JFA method

For each language l and session h :
  $y(l) \leftarrow 0, x(h,l) \leftarrow 0,$
  $U \leftarrow \text{random}(U \text{ is initialized randomly})$
Estimate statistics : $N(l), N(h,l), X(l), X(h,l)$
for $i = 1$ to $nb\_iterations$ do
  for all h and l do
    Center statistics : $\overline{X}(h,l)$
    Estimate $L(h,l)^{-1}$ and $B(h,l)$
    Estimate $x(h,l)$
    Center statistics : $\overline{X}(l)$
    Estimate $y(l)$
  end
  Estimate matrix U
end

۳-۵-شناسایی زبان به روش i-Vector

روش i-Vector جدیدترین و به‌روزترین روش در سامانه‌های شناسایی گوینده و زبان است. در این روش، بردارِ شناسایی (i-Vector)، برداری با بُعد پایین و طول ثابت است که با روشی مشابه JFA استخراج می‌شود. این فضای برداری با بُعد کم زیر فضای تنوعات کلی (شامل تنوعات زبان و شرایط ضبط) را نشان می‌دهد. در این روش تلاشی برای جداسازی تنوعات مربوط به شرایط ضبط و زبان صورت نمی‌گیرد، بلکه تنها زیر فضای تنوعات کلی مربوط به دو منبعِ ایجاد تنوع محاسبه شده و به‌عنوان ویژگی به طبقه‌بندی‌کننده‌ها داده می‌شود. مزیت استفاده از این روش این است که مدل به‌کاررفته برای استخراج i-Vector ها را می‌توان به‌صورت بدون سرپرستی (بدون برچسب زبان) تعلیم داد.

به‌طور خلاصه در این روش دادگان با بُعد زیاد به ویژگی‌هایی با طول ثابت به‌نحوی نگاشت می‌شود که اطلاعات مفید تا حد ممکن باقی بماند. ابر بردار وابسته به تنوعات شرایط ضبط و زبان (m_l) را می‌توان به‌صورت زیر بیان کرد:

$$m_l = m + Tw \qquad (۴۴)$$

در این رابطه m مؤلفهٔ مستقل از تنوعات و زبان است و مشابه روش JFA برابر با اَبر بردار میانگین مدل UBM قرار داده می‌شود. T یک ماتریس است که فضای تنوعات کلی را نشان می‌دهد و w هم متغیرهای پنهان با توزیع نرمال استاندارد است که i-Vector نامیده می‌شود [16]، [33].

درواقع ماتریس T به‌عنوان استخراج کنندهٔ i-Vectorها محسوب شده و ماتریسی با رتبهٔ پایین است که زیرفضای n_w بُعدی مربوط به تنوعات ابر بردار m_l را نشان می‌دهد. n_w بُعد بردار w است.

روش تعلیم ماتریس استخراج i-Vector (T) مشابه تعلیم زیر فضای D در JFA است. در این صورت تنها کافی است در روابط JFA، بخش مربوط به مدل‌سازی تنوعات کانال ($Ux_{(h,l)}$) حذف شود. در این صورت برای تعلیم ماتریس نگاشت T نیازی به دادگان با برچسب زبان نیست و فرض می‌شود که هر داده از منبع تنوع مختلفی ایجاد شده است. به بیان دیگر یکسان‌بودن زبان دادگان تعلیم یک زبان در تعلیم T درنظر گرفته نمی‌شود و این ماتریس زیر فضایی از تنوعات کلی شامل تنوعات زبان، کانال، گوینده و ... را نشان می‌دهد. بنابراین استخراج i-Vector (w) را می‌توان به‌عنوان یک مرحله استخراج ویژگی در نظر گرفت. توضیح بیشتر روابط استخراج ماتریس نگاشت T و بردار شناسایی w به‌دلیل مشابهت با روابط JFA در حوصله این مقاله نیست. توضیحات دقیق‌تر در مراجع [32]، [16] موجود است.

در سال‌های اخیر روش i-Vector به‌دلیل کیفیت مناسب در تمامی مقالات و سامانه‌های به‌روز شناسایی زبان و گوینده به‌عنوان سامانه پایه در نظر گرفته شده است. در سال ۲۰۱۴ مؤسسه NIST مسابقه‌ای را در این زمینه ترتیب داد و گروه‌های مختلف پژوهشی در این زمینه سامانه‌های شناسایی گوینده و زبان بر پایه i-Vector خود را در این مسابقه ارائه کردند [20]. هم اکنون نیز سامانه i-Vector تلفیق‌شده با شبکهٔ عصبی بهترین نتایج را در سامانه‌های شناسایی گوینده و زبان دارد [35].

۳-۵-۱-سامانه شناسایی زبان بر پایه i-Vector

در سامانه شناسایی زبان بر پایه i-Vector ابتدا در مرحلهٔ تعلیم، ماتریس نگاشت T تعلیم یافته و سپس از کلیهٔ دادگان تعلیم زبان‌های موجود، i-Vector استخراج می‌شود. در این مقاله برای طبقه‌بندی i-Vector ها از

طبقه‌بندی‌کنندهٔ تمایزی SVM استفاده شده است. درواقع برای هر زبان یک طبقه‌بندی‌کنندهٔ SVM به‌صورت یکی در مقابل بقیه تعلیم می‌یابد. دادگان تعلیم طبقه ۱+، i-Vectorهای استخراج‌شده از دادگان یک زبان تعلیم و دادگان طبقه ۱-، i-Vectorهای استخراج‌شده از دادگان سایر زبان‌های تعلیم است. بنابراین به تعداد زبان‌های تعلیم، طبقه‌بندی کننده SVM داریم.

در مرحلهٔ آزمایش ابتدا از دادهٔ آزمایش i-Vector استخراج شده و به طبقه‌بندی کننده‌های SVM داده می‌شود. امتیاز به‌دست آمده از این طبقه‌بندی کننده‌ها به‌طورمعمول با آستانه‌ای از پیش تنظیم‌شده مقایسه و در زمینه زبان آن داده تصمیم‌گیری می‌شود.

***هنجارسازی‌کردن i-Vector ها**: برای بهبود نتایج سامانه شناسایی زبان به روش i-Vector به‌طورمعمول i-Vector ها را با تقسیم بر مقدار طولشان هنجارسازی می‌کنند [16].

$$w' = \frac{w}{\|w\|} \qquad (۴۵)$$

۳-۵-۲-جبران‌سازی تنوعات در روش i-Vector

همان‌طور که در زیربخش‌های قبل شرح داده شد، در روش ivector زیرفضای تنوعات کلی محاسبه شده و در مرحله تعلیم ماتریس نگاشت ivector از اطلاعات برچسب زبان استفاده نمی‌شود. برای بهبود کارایی سامانه شناسایی زبان به روش i-Vector می‌توان تنوعات درون هر طبقه را با استفاده از روش‌های جبران‌سازی مختلف با استفاده از اطلاعات برچسب زبان کاهش داده و کارایی سامانه شناسایی زبان را بهبود داد. در ادامه سه روش جبران‌سازی مختلف برای حذف تنوعات درون هر طبقه (هر زبان) شرح داده شده است:

***روش LDA:** این روش فاصلهٔ بین‌طبقه‌ای را افزایش و فاصله درون‌طبقه‌ای را کاهش می‌دهد. جهت LDA با یک ماتریس انتقال A تعریف می‌شود. ماتریس A شامل بردارهای ویژه‌ی یک مسألهٔ مقدار ویژه به‌صورت زیر است:

$$S_b v = \lambda S_w v \qquad (۴۶)$$

در این رابطه، λ ماتریس قطری مقادیر ویژه است. ماتریس‌های S_b و S_w نیز به‌ترتیب کوواریانس بین‌طبقه‌ای و درون‌طبقه‌ای است که از روابط زیر محاسبه می‌شود:

$$S_b = \sum_{l=1}^{L} (\overline{w}_l - \overline{w})(\overline{w}_l - \overline{w})^t \qquad (۴۷)$$

$$\overline{w} = \frac{\sum n_l \times \overline{w}_l}{\sum n_l}$$

$$S_w = \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i=1}^{n_l} (w_i^l - \overline{w}_l)(w_i^l - \overline{w}_l)^t \qquad (۴۸)$$

در این رابطه $\overline{w}_l = \frac{1}{n_l}\sum_{i=1} n_l \times w_i^l$ میانگین i-Vectorهای مربوط به هر زبان و L تعداد کل زبان‌ها و n_l تعداد دادگان هر زبان l است [25]، [33]، [16].

*** روش NAP**[1]: هدف از اعمال این روش پیداکردن ماتریس انتقالی است که به‌نحو مناسب مؤلفه‌های تنوعات را حذف کند. جزئیات این روش در مرجع [11]، [10] شرح داده شده است. ماتریس نگاشت به‌صورت زیر تعریف می‌شود:

$$P = I - RR^t \qquad (۴۹)$$

در این رابطه، R ماتریس مستطیلی[2] با رتبهٔ پایین است که ستون‌های آن بهترین k بردار ویژه از ماتریس کوواریانس درون‌طبقه‌ای است. این بردارهای ویژه فضای کانال را مشخص می‌کند. ماتریس P برای نگاشت i-Vector ها مورد استفاده قرار می‌گیرد [16, 33].

*** روش WCCN**[3]: در این روش ابتدا کوواریانس بین کلاسی محاسبه می‌شود [22]:

$$W = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{n_l}\sum_{i=1}^{n_l}(w_i^l - \overline{w}_l)(w_i^l - \overline{w}_l)^t \qquad (۵۰)$$

در این رابطه $\overline{w}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} w_i^l$، میانگین i-Vector های هر زبان، L تعداد زبان‌ها و n_l تعداد دادگان هر زبان است. ابتدا معکوس این ماتریس با روش تجزیهٔ چولسکی[4] تجزیه شده ($W^{-1} = BB^t$) و سپس از ماتریس B برای نگاشت i-Vector ها به‌صورت زیر استفاده می‌شود:

[1] Nuisance Attribute Projection
[2] Rectangular
[3] within-class Covariance Normalization
[4] Cholesky Decomposition

$$\varphi(w) = B^t w \quad (۵۱)$$

در این رابطه φ تابع نگاشت به‌دست‌آمده از روش WCCN است [33]، [16].

۴-پس‌پردازش امتیازات

برای تصمیم‌گیری مناسب‌تر از روی بردار امتیازهای به‌دست‌آمده از روش‌های طیفی معرفی‌شده در بخش سوم، از روش‌های مختلفی چون شبکه عصبی (NN)، ماشین بردار مرزی (SVM) و مدل مخلوط گوسی (GMM) به‌عنوان پس‌پردازش امتیازهای خام برای طبقه‌بندی امتیازها به طبقه‌های مطلوب استفاده می‌شود. همچنین اعمال روش LLR به‌عنوان یک روش پس‌پردازش ساده نیز موجب بهبود تصمیم‌گیری می‌شود. این روش‌ها در ادامه توضیح داده شده‌اند. در شکل (۱) اعمال روش‌های پس‌پردازش به یک سامانه شناسایی زبان گوسی نشان داده شده است.

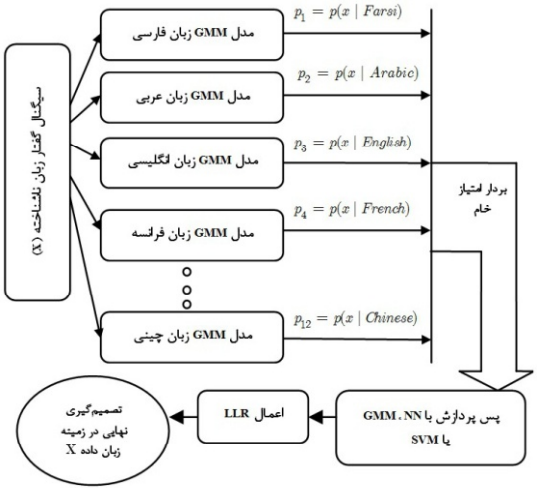
۴-۱-روش شبکهٔ عصبی

شبکه‌های عصبی مصنوعی که از مدل سیستم عصبی بیولوژیکی الهام گرفته شده است، از عناصر ساده‌ای شامل وزن‌ها و نورون‌ها تشکیل‌شده و کاربردهای زیادی از جمله شناسایی الگو، خوشه‌بندی، مدل‌سازی و ... دارند. شبکه‌های MLP[1] از جمله ساده‌ترین انواع شبکه‌های عصبی است که با پس‌انتشار خطای خروجی مطلوب و خروجی هر مرحله و تنظیم وزن‌های آن تعلیم می‌بینند. اگر در ورودی این شبکه دادگان یک طبقه و در خروجی آن کد طبقه قرار بگیرد، این شبکه به‌عنوان طبقه‌بندی‌کننده عمل می‌کند.

در این مقاله از شبکه MLP به‌عنوان طبقه‌بندی کننده امتیازهای خروجی استفاده می‌شود. تعداد مدل‌های زبانی پانزده عدد و تعداد طبقه‌ها (تعداد گره‌های خروجی) در شبکه عصبی برابر چهار و متشکل از سه زبان هدف عربی، فارسی و انگلیسی باضافه سایر زبان‌ها است، بنابراین می‌توان بردار امتیاز پانزده بعدی دادهٔ آزمایش به‌ازای مدل زبان‌های مختلف را به‌عنوان ورودی شبکه و کد چهار بیتی زبان را (فارسی، عربی، انگلیسی، غیره) به‌عنوان خروجی به شبکه MLP داده و تصمیم‌گیری روی امتیازات را با تنظیم وزن‌ها به شبکه تعلیم دهیم. درنهایت از شبکه تعلیم‌یافته برای تصمیم‌گیری روی دادگان آزمایش استفاده می‌کنیم.

شبکهٔ استفاده‌شده در این مقاله دو لایهٔ پنهان و با تعداد نورون‌های لایهٔ پنهان نخست پنجاه و لایهٔ پنهان دوم بیست انتخاب شده است. تعداد مطلوب لایه‌ها و نورون‌ها با آزمایش به‌دست آمده است. تابع غیرخطی لایه‌های پنهان log-sigmoid و لایه خروجی تابع tangent-sigmoid قرار داده شده است. خروجی مطلوب شبکه با چهار بیت برای طبقه‌های "عربی، فارسی، انگلیسی و سایر زبان‌ها" نشان داده می‌شود. برای نشان‌دادن یک طبقه خاص بیت مربوط به آن طبقه یک و سایر طبقه‌ها صفر قرار داده می‌شود. شبکه با الگوریتم تعلیم پس‌انتشار خطای گرادیان مزدوج[2] و با معیار خطای کمینه مجموع مربعات خطا (MSE[3]) تعلیم می‌بینند. برای تعلیم شبکه از نرم‌افزار MATLAB استفاده می‌شود.

از دادگان اعتبارسنجی[4] برای کنترل تعلیم شبکه و جلوگیری از آموزش بیش از اندازه[5] شبکه روی دادگان تعلیم، استفاده می‌شود، بدین صورت که اگر خطای MSE[6] روی دادگان اعتبارسنجی در بیش از بیست دوره تعلیم کاهش نیابد، تعلیم متوقف می‌شود.



(شکل-۱): استفاده از روش‌های NN، SVM و GMM به عنوان روش طبقه‌بندی برای پس‌پردازش امتیازات در سیستم شناسایی زبان مبتنی بر مدل مخلوط گوسی

(Figure-1): Using NN, SVM, and GMM as a classification approach for score post-processing in the GMM based language identification system

[1] Multi Layer Perceptron

[2] Conjugate-Gradient Back-Propagation
[3] Mean Squared Error
[4] Validation
[5] Overtrain
[6] Mean Square Error

۴-۲-روش GMM

همان‌طور که در بخش قبل نیز ذکر شد، روش GMM یک روش مدل‌کردن توزیع دادگان تصادفی با استفاده از ترکیب خطیِ توابع چگالی گوسی است. تا کنون GMM به‌طور گسترده به‌عنوان روش پس‌پردازش برای کالیبره‌کردن و ترکیب نتایج استفاده شده و نتایج را به‌طور قابل ملاحظه‌ای بهبود داده است [23]، [4]، [3].

در این مقاله، برای استفاده از مدل مخلوط گوسی به‌عنوان روش پس‌پردازش، برای هر یک از زبان‌های هدف (عربی، فارسی و انگلیسی) یک مدل با استفاده از امتیازات فایل‌های این زبان تعلیم داده می‌شود. به همین ترتیب به‌ازای کل زبان‌های غیر هدف نیز یک مدل تعلیم داده می‌شود. بنابراین چهار مدل GMM با استفاده از امتیازهای پانزده بعدی تعلیم می‌بینند (سه مدل برای هر یک زبان‌های هدف و یک مدل برای سایر زبان‌ها). برای تسریع محاسبات، ماتریس کوواریانس توزیع‌های گوسی قطری فرض شده است.

در این مقاله، در روش پس‌پردازش GMM، تعداد توابع گوسی هر یک از سه زبان هدف (M در رابطهٔ (۵))، تعداد توابع گوسی زبان‌های غیر هدف ($q \times M$)، از آنجا که تنوعات مجموعه زبان‌های غیر هدف از یک زبان هدف بیشتر است، تعداد توابع گوسی غیر هدف ضریبی از تعداد توابع هدف در نظر گرفته شده است. q عددی صحیح و مثبت است) و کف واریانسِ توابع گوسی ($\frac{1}{M^p}$) به‌صورت متغیر در نظر گرفته شده است. با افزایش مقدار عدد صحیح و مثبت p مقدار کف واریانس کمتر شده و مدل GMM روی دادگان تعلیم بیشتر آموزش می‌یابد. این پارامترها (p ، q و M) در هنگام تعلیم مدل‌ها با استفاده از دادگان اعتبارسنجی محاسبه می‌شوند. اعتبارسنجی به‌صورت N-fold روی دادگان تعلیم انجام می‌شود. یعنی هر بار $1-\frac{1}{N}$ از دادگان تعلیم برای آموزش و $\frac{1}{N}$ به‌عنوان دادگان اعتبارسنجی در نظر گرفته می‌شود. خطای میانگین روی N دسته دادگان اعتبارسنجی محاسبه می‌شود (در این مقاله N برابر ۵ قرار داده شده است). درنهایت پارامترهایی که به‌ازای آنها این خطا کمینه است، به‌عنوان پارامترهای بهینه برای دادگان آزمایش انتخاب و استفاده می‌شود.

۴-۳-روش SVM

روش SVM نیز می‌تواند به‌عنوان یک ابزار پس‌پردازش امتیازها استفاده شود. در این روش امتیازهای دادگان تعلیم و طبقه‌های این دادگان به‌عنوان ورودی به برنامه SVM داده می‌شود. بنابراین ورودی برنامه تعلیم SVM، پانزده بعدی و خروجی آن چهار طبقه عربی، انگلیسی، فارسی و سایر زبان‌ها است. در مرحله تعلیم، مرز مناسب برای جداسازی این چهار طبقه تعیین شده و از این مرزها در مرحله آزمایش استفاده می‌شود. تعیین مرز در مرحله تعلیم می‌تواند توسط توابع هسته مختلف از جمله خطی، چندجمله‌ای، شعاعی و غیره صورت گیرد. با توجه به نوع هسته انتخاب شده پارمترهایی باید تنظیم شوند. هسته غیرخطی RBF که در این مقاله استفاده شده است، به‌صورت زیر تعریف می‌شود:

$$K(x_i,x_j)=\exp(-\gamma \mid\mid x_i-x_j \mid\mid^2), \quad \gamma>0 \qquad (۵۲)$$

پارامتر γ در رابطه (۵۴) و پارامتر هزینه C در رابطه (۱۳) با استفاده از دادگان اعتبارسنجی تخمین زده می‌شود. روش تنظیم این پارامترها نیز همان روش N-fold توضیح داده شده در بخش GMM است. برای پیاده‌سازی SVM با هسته RBF و به‌صورت احتمالاتی از جعبه‌ابزار LIBSVM استفاده شده است [12].

۴-۴-روش LLR

روش لگاریتم نسبت احتمال (LLR) یک روش پس‌پردازش ساده برای کالیبره‌کردن امتیازات است که می‌توان آن را بر بردار خروجی روش‌های پس‌پردازش ذکرشده در بخش‌های قبل اعمال کرد. در این روش از خروجی احتمالاتی پسین[1] لگاریتم گرفته می‌شود. مقصود از خروجی احتمالاتی پسین این است که مجموع خروجی‌های طبقه‌بندی‌کننده به‌ازای یک دادهٔ آزمایش و روی طبقه‌های مختلف برابر یک باشد. استفاده از روش LLR در مسائل آشکارسازی، موجب گسترده‌ترشدن محدوده امتیازات و در بیش‌تر موارد موجب افزایش دقت و کارایی تصمیم‌گیری می‌شود. اگر $\vec{S}=[S_1,S_2,...,S_L]^t$ بردار لگاریتم احتمال خروجی طبقه‌بندی‌کنندهٔ پس‌پردازشی باشد، مقادیر خروجی پس از اعمال LLR (S') به‌صورت زیر محاسبه می‌شود [34]، [30]:

$$S'_i=S_i-\log\left(\frac{1}{L-1}\sum_{j\#i}e^{S_j}\right) \qquad (۵۳)$$

[1] Posterior Probability

همچنین، LLR به‌صورتی دیگر نیز اعمال می‌شود:

$$S_i' = S_i - \log\left(\sum_{j=1}^{L} e^{S_j}\right) \quad (۵۴)$$

۵-دادگان

در این بخش دادگان لازم برای تعلیم و آزمایش سامانه معرفی شده است.

۵-۱-دادگان تعلیم

دادگان تعلیم از پانزده زبان تهیه شده است. این دادگان ترکیبی از پوشه‌های سی ثانیه‌ای از دادگان NIST2003 و NIST96، دو دقیقه از فایل‌های دادگان فارس‌دات تلفنی بزرگ و مجموعه‌ای از دادگان محاوره‌ای از زبان‌های عربی، فارسی، انگلیسی، چینی، فرانسه، آلمانی، کردی و روسی و ترکی است. طول زمانی فایل‌های استفاده‌شده برای تعلیم بر حسب ساعت در جدول (۱) درج شده است.

(جدول-۱): طول زمانی دادگان تعلیم سیستم شناسایی زبان

(Table-1): the duration of training data

زبان	عربی	انگلیسی	فارسی	فرانسه	چینی	آلمانی	هندی	ژاپنی	اسپانیایی	کره‌ای	تامیل	ویتنامی	ترکی	کردی	روسی
طول زمانی (ساعت)	۱۱/۹۳	۱۶/۷۴	۱۲/۳۸	۸/۸۲	۱۰/۸۹	۸/۸۳	۴/۲۰	۵/۴۲	۳/۴۰	۴/۰۵	۱/۹۸	۱/۹۵	۸/۶۴	۶/۲۷	۱۱/۱۶

۵-۲-دادگان آزمایش

دادگان مورد استفاده برای آزمایش سامانه بازشناسی زبان، دادگان گفتار تلفنی محاوره‌ای هستند.

دادگان آزمایش بین سی ثانیه تا پنج دقیقه و با فرکانس هشت‌هزار هرتز هستند و در چهار دسته عربی، فارسی، انگلیسی و سایر زبان‌ها دسته‌بندی شده است. هدف در این مقاله جداسازی زبان‌های هدف عربی، فارسی و انگلیسی از سایر زبان‌ها است. تعداد دادگان آزمایش هر زبان در جدول (۲) درج شده است.

(جدول-۲): تعداد دادگان آزمایش زبان‌های هدف و سایر زبانها

Table-2): The number of test records of target and other (languages

عربی	فارسی	انگلیسی	سایر زبان‌ها
۳۰۲	۲۰۷	۲۰۶	۳۶۹

۶-آزمایش

در این بخش با استفاده از سامانه‌های بازشناسی شرح‌داده‌شده در بخش ۲، آزمایش‌هایی روی دادگان آزمایش انجام شده است. نتایج ارائه‌شده به‌صورت خطای EER[1] ارائه شده است. خطای EER محلی است که مقادیر خطای FA[2] (درصد تعداد دادگان غیر هدفی که به‌اشتباه هدف تشخیص داده شده) و خطای FR[3] (درصد تعداد دادگان هدفی که به‌اشتباه غیر هدف تشخیص داده شده) یکسان است.

خطای EER برای هر زبان، خطای جداسازی آن زبان از چهارده زبان دیگر است. در جدوال متوسط EER سه زبان عربی، انگلیسی و فارسی برای مقایسه روش‌های مختلف گزارش شده است.

۶-۱-نتایج روش مدل مخلوط گوسی

نتایج روش مدل مخلوط گوسی در جدول (۳) درج شده است. در روش GMM-UBM تعداد مؤلفه‌های گوسی برابر ۱۰۲۴ قرار داده شده است. در روش GMM-ML مدل هر زبان به‌طور جداگانه با استفاده از معیار ML و با ۱۲۸ گوسی تعلیم داده می‌شود. در روش MMI نیز برای هر زبان یک مدل گوسی با ۱۲۸ مؤلفه در نظر گرفته شده است و الگوریتم تعلیم MMI تا ۱۵ تکرار اجرا شده است. در این جدول مقصود از اعمال وزن، رابطه (۱۰) و مقصود از GD، تعلیم مدل‌های گوسی وابسته به جنسیت گوینده است.

(جدول-۳): نتایج خطاهای EER زبان‌های هدف با روش شناسایی زبان مدل مخلوط گوسی

(Table-3): EER results of target languages of the GMM based language identification system

	عربی	فارسی	انگلیسی	میانگین خطای سه زبان
GMM-UBM	۱۸/۰۴	۱۵/۲۴	۷/۴۵	۱۳/۵۸
GMM-ML	۲۰/۵۲	۲۰/۰۵	۱۲/۴۸	۱۷/۶۸
GMM+MMI	۷/۱۰	۸/۶۷	۴/۰۰	۶/۵۹
EHMM + MMI	۷/۱۰	۸/۴۴	۴/۱۱	۶/۵۵
MMI + اعمال وزن	۷/۴۹	۷/۵۱	۶/۵۹	۷/۲۰
MMI + اعمال وزن + GD	۵/۳۴	۸/۵۴	۴/۰۱	۵/۹۶

۶-۲-نتایج روش GSV-SVM

نتایج روش GSV-SVM با تعداد مؤلفه‌های گوسی مختلف

[1] Equal Error Rate
[2] False Accept
[3] False Reject

(M) در جدول (۴) درج شده است. بُعد بردارهای ویژگی با استفاده از روش LDA از ۷۲ به ۴۰ کاهش یافته است. نتایج این آزمایش کیفیت مناسب این روش را در بازشناسی زبان نشان می‌دهد.

استفاده از روش برگشت به حوزهٔ مدل (MP) گرچه موجب بهبود نتایج نشده ولی زمان آزمایش برای دادگان طولانی را تا $\frac{1}{4}$ کاهش داده است. در جدول (۵) زمان اجرای برنامه برای دو داده‌ی آزمایش با طول زمانی مختلف با و بدون اعمال روش MP مقایسه شده است.

۶-۳-نتایج روش GMM Tokenizer-SVM

نتایج این روش در جدول (۶) درج شده است. به‌عنوان tokenizerدر آزمایش نخست از مدل گوسی جهانی (UBM) با ۱۰۲۴ مؤلفه و در آزمایش دوم از مدل تلفیق گوسی چند زبانی (MLM) استفاده شده است. در آزمایش دوم برای هر زبان ابتدا یک مدل گوسی ۱۲۸ مؤلفه‌ای با الگوریتم تعلیم MMI تعلیم‌یافته و سپس با ترکیب مدل پانزده زبان تعلیم، مدل گوسی Tokenizer به‌دست آمده است. اشکال اصلی این روش طولانی‌بودن مراحل تعلیم و آزمایش و زیاد بودن بُعد بردار ورودی به طبقه‌بندی‌کننده‌های SVM است.

(جدول-۴): نتایج خطاهای EER با روش GSV-SVM

(Table-4): EER results of target languages of the GSV-SVM approach

	عربی	فارسی	انگلیسی	میانگین خطای سه زبان
GSV-SVM M=32	۵/۸۰	۴/۸۱	۳/۶۵	۴/۷۵
GSV-SVM M=128	۲/۴۸	۳/۶۴	۲/۷۱	۲/۹۴
GSV-SVM M=256	۲/۸۶	۲/۷۰	۲/۷۱	۲/۷۶
GSV-SVM + MP M=128	۳/۹۱	۲/۷۰	۲/۹۴	۳/۱۸

(جدول-۵): زمان اجرای برنامه با اعمال و بدون اعمال MP

(Table-5): The program speed with and without model pushing method

طول زمانی داده آزمایش (ثانیه)	بدون MP (ثانیه)	با MP (ثانیه)
۹۰	۱۶۵	۱۵
۴۵۰	۲۵۵	۶۰

(جدول-۶): نتایج خطاهای EER با روش GMM Tokenizer

(Table-6): EER results of target languages of GMM Tokenizer approach

Tokenizer	عربی	فارسی	انگلیسی	میانگین خطای سه زبان
UBM	۱۱/۷۳	۹/۸۴	۷/۵۴	۹/۷۰
MLM	۷/۸۲	۸/۵۶	۸/۸۲	۸/۴۰

۶-۴-نتایج روش JFA و i-Vector

نتایج به‌دست‌آمده از دو روش شناسایی زبان (JFA و i-Vector) در جدول (۷) درج شده است. در روش JFA تعداد مؤلفه‌های گوسی ۲۵۶ و ۱۰۲۴ و بُعد زیرفضای تنوعات (n_x) برابر بیست در نظر گرفته شده است. در روش i-Vector نیز تعداد مؤلفه‌های مدل گوسی ۲۵۶، ۵۱۲ و ۲۰۴۸ در نظر گرفته شده و بُعد زیر فضای تنوعات کلی (n_w) برابر دویست در نظر گرفته شده است. با توجه به نتایج، روش i-Vector نسبت به روش JFA نتایج بسیار بهتری دارد. در هر دو روش افزیش تعداد مؤلفه‌های گوسی موجب بهبود نتایج می‌شود.

بُعد بهینه زیر فضای i-Vector به‌صورت تجربی و با انجام آزمایش با ابعاد مختلف به دست آمده است. نتایج این آزمایش در جدول (۸) نوشته شده است. با توجه به نتایج بعد ivector برای سایر آزمایش‌ها برابر دویست قراردادهشد.

(جدول-۷): نتایج خطاهای EER زبان‌های هدف در سامانهٔ شناسایی زبان به روش JFA و i-Vector با تعداد مؤلفه‌های گوسی (M) مختلف

(Table-7): EER results of target languages of JFA and i-Vector methods with different Gaussian components.

	عربی	فارسی	انگلیسی	میانگین خطای سه زبان
JFA M=256	۱۱/۷۳	۱۸/۸۸	۲۸/۷۷	۱۹/۷۹
JFA M=1024	۷/۹۵	۱۰/۳۰	۱۸/۶۷	۱۲/۳۱
i-Vector M=256	۳/۳۲	۲/۷۰	۲/۰۰	۲/۶۷
i-Vector M=512	۳/۰۷	۱/۷۶	۲/۰۰	۲/۲۸
i-Vector M=2048	۲/۴۸	۱/۷۶	۰/۹۴	۱/۷۱

برای بهبود نتایج روش i-Vector سه روش جبران تنوعات LDA، NAP و WCCN اعمال شده و نتایج آنها در جدول (۹) نوشته شده است. با توجه به نتایج آزمایش‌ها اعمال هنجارسازی طولی (Norm در جدول زیر) و دو روش NAP و WCCN هر سه تا حدودی نتایج سامانه i-Vector را بهبود می‌دهد.

(جدول-۸): تاثیر بُعد زیر فضای تنوعات بر کارایی روش i-Vector (در این آزمایش M برابر ۲۵۶ است)

(Table-8): The influence of i-Vector dimention on language identification results (the number of guassian components is 256)

n_w	عربی	فارسی	انگلیسی	میانگین خطای سه زبان
۱۰۰	۴/۱۷	۴/۵۸	۳/۶۵	۴/۱۳
۲۰۰	۳/۳۲	۲/۷۰	۲/۰۰	۲/۶۷
۴۰۰	۳/۷۸	۲/۷۰	۱/۷۷	۲/۷۵
۶۰۰	۳/۷۸	۳/۰۴	۱/۷۷	۲/۸۶

(جدول-۹): نتایج جبران‌سازی تنوعات در روش i-Vector (در این آزمایش M برابر ۲۵۶ است)

(Table-9): Applyingf variability compensation methods in i-Vector approach (the number of guassian components is 256).

	عربی	فارسی	انگلیسی	میانگین خطای سه زبان
بدون Norm و جبران‌سازی	۳/۳۲	۲/۷۰	۲/۰۰	۲/۶۷
Norm	۳/۰۷	۲/۷۰	۱/۸۸	۲/۵۵
Norm +LDA	۴/۱۷	۲/۸۲	۲/۷۱	۳/۲۳
Norm +NAP	۲/۸۶	۱/۷۶	۲/۷۱	۲/۴۴
Norm +WCCN	۲/۸۶	۲/۷۰	۱/۸۸	۲/۴۸

۶-۵-نتایج اعمال پس‌پردازش امتیازات

روش‌های مختلف پس‌پردازش امتیازات (روش شبکه عصبی (NN)، روش مدل گوسی (GMM)، روش ماشین بردار مرزی (SVM) و روش LLR بر سامانه شناسایی زبان GMM-UBM با ۱۰۲۴ مؤلفه گوسی اعمال شد. با توجه به نتایج این آزمایش‌ها (جدول (۱۰)) روش شبکه عصبی به همراه LLR بهترین نتایج را دارد.

(جدول-۱۰): مقایسه نتایج اعمال روش پس پردازش درسامانه‌های مختلف بازشناسی زبان

(Table-10): Comparison of different score post-processing methods.

	LLR	عربی	فارسی	انگلیسی	میانگین خطای سه زبان
NN	×	۵/۰۸	۳/۶۴	۳/۸۸	۴/۲۰
	√	۴/۵۰	۲/۹۳	۳/۶۵	۳/۶۹
GMM	×	۱۰/۴۳	۸/۴۵	۵/۰۵	۷/۹۸
	√	۱۰/	۹/۲۳	۵/۴۵	۸/۳۰
SVM	×	۶/۳۹	۳/۶۴	۴/۸۲	۴/۹۵
	√	۶/۵۲	۳/۷۵	۳/۶۵	۴/۶۴

در جدول (۱۱) پس‌پردازش شبکهٔ عصبی به همراه LLR بر سه روش شناسایی زبان اعمال شده است. مقصود از (BB) در این جدول، اعمال LLR قبل از اعمال پس‌پردازش امتیازات و مقصود از (AB)، اعمال LLR پس از اعمال پس‌پردازش امتیازات است. همان‌طورکه نتایج نشان می‌دهد، استفاده از روش پس‌پردازش امتیاز در سامانه GMM-UBM که کیفیت پایینی دارد به‌نحو موثری نتایج را بهبود می‌دهد، ولی بهبود مناسبی در دو سامانه مبتنی بر ابربردار گوسی (GSV) و i-Vector دیده نمی‌شود. به نظر می‌رسد امتیازدهی با روش SVM و اعمال هنجارسازی LLR، امتیازات مناسبی در خروجی تولید می‌کند، به‌نحوی که پس‌پردازش در این دو سامانه مفید واقع نمی‌شود.

(جدول-۱۱): مقایسه نتایج اعمال روش پس پردازش شبکه عصبی به همراه LLR در سه سامانه شناسایی زبان مختلف

(Table-11): Applying neural network post-processing method on different language identification approaches

		عربی	فارسی	انگلیسی	میانگین خطای سه زبان
GMM-UBM	BB	۱۸/۰۴	۱۵/۲۴	۷/۴۵	۱۳/۵۸
	AB	۴/۵۰	۲/۹۳	۳/۶۵	۳/۶۹
GSV-SVM	BB	۲/۴۸	۳/۶۴	۲/۷۱	۲/۹۴
	AB	۳/۷۸	۲/۷۰	۲/۸۳	۳/۱۰
i-Vector	BB	۲/۴۸	۱/۷۶	۰/۹۴	۱/۷۱
	AB	۲/۴۸	۱/۷۶	۰/۹۴	۱/۷۱

در جدول (۱۲) نتایج بهترین جواب‌های به‌دست‌آمده از هر کدام از روش‌های طیفی مختلف برای مقایسه آورده شده است. نتایج به‌دست‌آمده نشان می‌دهد که سامانه شناسایی زبان به روش i-Vector بهترین نتایج را دارد.

(جدول-۱۲): مقایسه خطای بازشناسی زبان در روش‌های طیفی مختلف

(Table-12): Performance comparison of different language identification methods.

	عربی	فارسی	انگلیسی	میانگین خطای سه زبان
GMM-UBM	۱۸,۰۴	۱۵,۲۴	۷,۴۵	۱۳,۵۸
GMM-UBM + Backend	۴/۵۰	۲/۹۳	۳/۶۵	۳/۶۹
GMM-ML	۲۰/۵۲	۲۰/۰۵	۱۲/۴۸	۱۷/۶۸
GMM-MMI	۵/۳۴	۸/۵۴	۴/۰۱	۵/۹۶
GMM tokenizer	۷/۸۲	۸/۵۶	۸/۸۲	۸/۴۰
GSV-SVM	۲/۸۶	۲/۷۰	۲/۷۱	۲/۷۶
JFA	۷/۹۵	۱۰/۳۰	۱۸/۶۷	۱۲/۳۱
i-Vector	۲/۴۸	۱/۷۶	۰/۹۴	۱/۷۱

۷-رویکردهای اخیر در بازشناسی زبان

در این مقاله مروری بر روش‌های معمول و به‌روز شناسایی زبان گفتاری طیفی (روش‌های GMM-UBM، GMM-MMI، GSV-SVM، JFA و iVector) انجام شد. علاوه‌بر این روش‌ها رویکردهای جدیدی در مقالات برای بهبود کارایی سامانه‌های شناسایی زبان طیفی پیشنهاد شده است. چند نمونه از روش‌های جدید برای بهبود کارایی سامانه شناسایی زبان در ادامه معرفی شده است:

***روش مبتنی بر شبکهٔ عصبی:** یکی از رویکردهای جدید در سامانه‌های شناسایی خودکار زبان گفتاری طیفی استفاده از روش‌های مبتنی بر شبکه‌های عصبی است. استفاده از ساختارهای شبکه عصبی عمیق[1] برای دسته‌بندی ویژگی‌های آکوستیکی به زبان‌های هدف [36]، استفاده از ویژگی‌های استخراج‌شده از لایه گلوگاه[2] یک شبکه عصبی برای شناسایی زبان [44]، [35] و استفاده از شبکه‌های عصبی کانولوشنی[3] [18] نمونه‌هایی از این مقالات است.

***استفاده از ویژگی‌های جدید:** در مقالات شناسایی زبان به‌طورمعمول از ویژگی‌های طیفی چون MFCC استفاده می‌شود. یکی از رویکردهای جدید برای بهبود کارایی سامانه شناسایی زبان استفاده از ویژگی‌های دیگری چون ویژگی‌های نوایی[4] [32]، ویژگی‌های مبتنی بر مدل مخلوط گوسی زیر فضایی (SGMM[5]) [37] و ... است. به‌طور معمول این ویژگی‌ها با ویژگی‌های معمول MFCC در سطح ویژگی و یا مدل تلفیق شده و موجب بهبود کیفیت شناسایی زبان می‌شود.

***روش‌های جبران‌سازی تنوعات:** از آنجا که یکی از چالش‌های مهم در سامانه‌های شناسایی زبان، تنوعات کانال و شرایط ضبط صدا است، در مقالات جدید با استفاده از روش‌هایی چون JFA [24] و NAP [40] به حذف و جبران‌سازی این تنوعات در حوزه ویژگی و مدل پرداخته شده است.

۸-نتیجه‌گیری

در این مقاله بازشناسی زبان گفتاری برای مکالمات محاوره‌ای تلفنی با استفاده از چندین روش طیفی انجام شد. روش پایه شناسایی زبان، مدل مخلوط گوسی-مدل جهانی (GMM-UBM) به عنوان روش پایه پیاده‌سازی شد. میانگین خطای EER سه زبان هدف (فارسی، انگلیسی و عربی) در این روش حدود ۱۳/۵۸ درصد شد. نتایج آزمایش‌ها نشان داد که تعلیم سامانه شناسایی زبان GMM با الگوریتم تعلیم تمایزی MMI نسبت به سامانه‌هایی که تنها با الگویتم ML تعلیم یافته‌اند، کارایی بسیار بهتری دارد. به‌نحوی‌که میانگین خطای EER سه زبان هدف نسبت به روش GMM-UBM حدود هشت درصد (به‌صورت مطلق) کاهش یافته است. روش GMM tokenizer نیز به‌عنوان یک روش طیفی جدید مورد آزمایش قرار گرفت. میانگین خطای EER سه زبان هدف این روش نیز نسبت به روش GMM-UBM حدود پنج درصد بهتر است.

روش تمایزی GSV-SVM نیز در این مقاله برای بازشناسی زبان مورد استفاده قرار گرفت. نتایج به‌دست‌آمده از این روش به‌طور قابل ملاحظه‌ای از روش‌های طیفی معمول بهتر است؛ به‌طوری‌که میانگین خطای EER سه زبان هدف نسبت به الگویتم GMM-UBM حدود یازده درصد (به‌صورت مطلق) کاهش یافته است. در این مقاله سرعت کم این روش، با استفاده از یک روش برگشت به حوزهٔ مدل (MP) بهبود داده شد.

دو روش جدید JFA و i-Vector نیز در این مقاله مورد پیاده‌سازی قرار گرفت. با توجه به نتایج به‌دست‌آمده هریک از این دو روش، نسبت به روش GMM-UBM نتایج بسیار بهتری دارد؛ به‌نحوی‌که میانگین خطای EER سه زبان هدف در روش JFA حدود یک درصد و در روش i-Vector حدود دوازده درصد (به‌صورت مطلق) کاهش یافته است. به‌طورکلی نتایج آزمایش‌ها نشان داد که روش i-Vector نسبت به سایر روش‌های طیفی شناسایی زبان نتایج بسیار بهتری دارد.

لازم به ذکر است که این مقاله حاصل پژوهش هفت ساله در زمینه شناسایی زبان گفتاری در پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی است. در این پژوهش مطالعه و پیاده‌سازی الگوریتم‌های جدید شناسایی زبان طیفی چون PLDA و روش‌های شناسایی زبان آوایی بسیار جدید برای تلفیق دو سامانه آوایی و طیفی و رسیدن به سامانه شناسایی زبان با کیفیت بسیار خوب همچنان در حال انجام است.

---

[1] Deep Neural Networks
[2] Bottle-neck Features
[3] Convolutional Neural Networks
[4] Prosodic
[5] Subspace Gaussian Mixture Models

۹-مراجع 9-Refrences

[۱]ش.، رضا، ز.، زینل‌خانی، ج.، کبودیان، "تأثیر شرایط اولیه مناسب بر کارآیی فیلتر رستا در سیستم‌های شناسایی زبان،" پانزدهمین کنفرانس انجمن کامپیوتر ایران، شرکت متن، تهران، ۱۳۸۸.

[۲]ش. رضا، ز. زینل‌خانی، ج. کبودیان، "بهبود تصمیم‌گیری در سیستم‌های شناسایی زبان با استفاده از پس‌پردازش امتیازات،" پانزدهمین کنفرانس انجمن کامپیوتر ایران، شرکت متن، تهران، ۱۳۸۸.

[1] Sh. Reza, Z. Zeynalkhani, J. Kabudian, "On the Impact of Initial Conditions in RASTA-Based Modulation Spectrum Filtering for Spoken Language Recognition Systems," Proc. 15thInternational Computer-Society-of-Iran Computer Conference (CSICC), Tehran, Iran, 2010. (in Persian).

[2] Sh. Reza, Z. Zeynalkhani, J. Kabudian, "Improved Decision-Making in Spoken Language Recognition Systems Back-End," Proc. 15th International Computer-Society-of-Iran Computer Conference (CSICC), Tehran, Iran, 2010. (in Persian).

[3] M. F., BenZeghiba, J. L., Gauvain and L., Lamel, "Gaussian Backend Design for Language Detection," Proc. IEEE ICASSP, 2009.

[4] M. F., BenZeghiba, J. L., Gauvain and L., Lamel, "Language Scores Calibration Using Adapted Gaussian Backend," Proc. INTERSPEECH, Brighton, UK, 2009.

[5] C. Bishop. "Pattern Recognition and Machine Learning", Springer, 2006.

[6] N., Brummer, L., Burget, P., Kenny, P., Matejka, E. D., Villers, M., Karafiat, "ABC System Description for NIST SRE 2012," in Processing NIST Speaker Recognition Evaluation, 2012.

[7] N. Brummer, S. Cumani, O. Glembek, M. Karafiat, P. Matejka, J. Pesan, O. Plchot, M. Soufifar, E. Villiers, and H. Cernocky, "Description and analysis of the brno276 system for LRE2011," in Proceedings of Odyssey 2012.

[8] L., Burget, P., Matejka and J., Cernocky, "Discriminative Training Techniques for Acoustic Language Recognition," IEEE International Conference on Acoustics, Speech and Signal Processing, 2006.

[9] W.M. Campbell, "A Covariance Kernel for Language Recogniton," Proc. IEEE ICASSP, 2008.

[10] W.M. Campbell, D.E. Sturim, D.A. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," IEEE Signal Processing Letters, vol. 13, no. 5, 2006.

[11] W., Campbell, D., Sturim, D., Reynolds and A., Solomonoff, "SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP Variability Compensation," IEEE International Conference on Acoustic Speech and Signal Processing, 2006.

[12] C. C., Chang and C.-J., Lin., "LIBSVM: A Library for Support Vector Machines," Software Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[13] N., Dehak, "Discriminative and Generative Approches for Long - and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification," Ph.D. thesis, Ecole de Technologie Superieure, Montreal, 2009.

[14] R. Dehak, N. Dehak, P. Kenny, P. Dumouchel, "Linear and Non Linear Kernel GMM SuperVector Machines for Speaker Verification," Proc. INTERSPEECH, Belgium, 2007.

[15] N. Dehak, P., Kenny, R., Dehak, P., Dumouchel, and P., Ouellet, "Front-end Factor Analysis for Speaker Verification,", IEEE Transaction on Audio Speech and Language Processing, 2010.

[16] N. Dehak, P.A. Torres-Carrasquillo, D. Reynolds, R. Dehak, "Language Recognition via ivectors and Dimensionality Reduction," Proc. INTERSPEECH, 2011.

[17] L. F. Dharo, O. Glembek, O. Plchot, P. Matejka, M. Soufifar, R. Cordoba, and J. Cernocky, "Phonotactic language recognition using i-vectors and phoneme posteriogram counts," in Interspeech, 2012.

[18] S., Ganapathy, K., Han, S., Thomas, M., Omar, M. V., Segbroeck, and Sh. S., Narayanan, "Robust Language Identification Using Convolutional Neural Networks Features," Singapore, INTERSPEECH, 2014.

[19] O., Glembek, L., Burget, P., Matejka, M., Karafiat and P., Kenny, "Simplification and Optimization of i-vector Extraction," IEEE, 2011.

[20] C., Greenberg et all, "The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge", Odyssey, 2014.

[21] A., Hanani, M., Carey and M., Russell, "Improved Language Recognition using Mixture Components Statistics," INTERSPEECH, 2010.

[22] A., Hatch, S., Kajarekar and A., Stolcke, "Withen-Class Covariance Normalization for SVM-Based Speaker Recognition," International Conference on Spoken Language Processing, USA, September 2006.

[23] C.W., Hsu, C.C., Chang and C. J., Lin, "A Practical Guide to Support Vector Classification," Available Online at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2009.

[24] V., Hubeika, L., Burget, P., Matejka and P., Schwarz, "Discriminative Training and Channel Compensation for Acoustic Language Recognition, " Brisbane, 2008.

[25] Kanagasundaram, A., et al. "I-vector Based Speaker Recognition on Short Utterances.", International Speech Communication Association (ISCA), 2011.

[26] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithm," Technical Report, CRIM Research Center, Canada, 2005.

[27] P., Kenny, G., Boulianne and P., Ouellet, "Factor Analysis Simplified," In: Proceedings of International Conterence on Acoustic Speech and Signal Processing, ICASSP 2005.

[28] P., Kenny, G., Boulianne, P., Dumouchel, "Eigenvoice Modeling With Sparse Training Data," IEEE Transaction on Speech and Audio Processing, May 2005.

[29] P., Kenny, P., Ouellet, N., Dehak, V., Gupta and P., Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," IEEE Trans. Audio, Speech and Language Procesing, July 2008.

[30] H., Li, M., Suo, Y., Yan, "Using SVM as Backend Classifier for Language Identification," EURASIP Journal on Audio, Speech and Music Processing, vol., pp. 2008.

[31] D., Matrouf, F., Verdet, M., Rouvier, J., Bonastre and G., Linares, "Modeling Nuisance Variability with Factor Analysis for GMM-based Audio Pattern Classification," Computer science and language, 2010.

[32] D., Martinez, L., Burget, L., Ferrer and N., Scheffer, "Ivector-based Prosodic System for Language Identification," ICASSP, 2012.

[33] D. Martinez, O. Plchot, L. Burget, O. Glemberk, P. Matejka, "Language Recognition in i-Vector Space," Proc. INTERSPEECH, 2011.

[34] P., Matejka, L., Burget, O., Glembek, P., Schwarz, V., Hubeika, M., Fpso, T., Mikolov, O., Plchon and J., Honza Cernohy, "BUT Language Recognition for NIST 2007 Evaluation," Proc. INTERSPEECH, 2008.

[35] P., Matejka, L., Zhang, T., Ng, S. H., Mallidi, O., Glembek, J., Ma, and B., Zhang, "Neural Networks Bottleneck Features for Language Identification," Odyssey, 2014.

[36] I. L., Moreno, J. G., Dominguez, O., Plchot, D., Martinez, J. G., Rodriguez, and P., Moreno, "Automatic Language Identification Using Deep Neural Networks," IEEE Conference on Acoustic, Speech and Signal Processing ICASSP, 2014.

[37] O., Plchot, M., Karafiat, N., Brummer, O., Gelembek, P., Matejka, E. D., Villiers, and J., Cernocky, "Speaker vectors from Subspace Gaussian Mixture Model as complementary features for Language Identification," Odyssey, 2012.

[38] D.A., Reynolds, W. M., Campbell, W., Shen and E., Singer, "Automatic Language Recognition via Spectral and Token Based Approaches," in Springer Handbook of Speech Processing, 2008.

[39] D., Reynolds, T., Quatieri and R., Dunn, "Speaker Verification Using Adapted Gausiian Mixture Models", Digital Signal Processing, 2000.

[40] F. S., Richardson, and W. M., Campbell, "NAP for High Level Language Identification," ICASSP, 2011.

[41] W. Shen, et al. "Experiments with lattice-based PPRLM language identification, " in Proc. Odyssey, Puerto Rico, 2006.

[42] P.A., Torres-Carrasquillo, E., Singer, M. A., Kohler, R. J., Greene, D. A., Reynoldsand J. R., Deller, Jr., "Approaches to Language

Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features," International Conference on Spoken Language Processing, 2002.

[43] F., Verdet, D., Matrouf, J.-F., Bonastre, J., Hennebert, "Factor Analysis and SVM for Language Recognition," ISCA, 2009.

[44] K., Vesely, M., Karafiat, F., Grezl, M., Janda, and E., Egorova, "The Language-Independent Bottleneck Fetaures," IEEE, Brno University of Technology , SLT 2012.

[45] J., Weston and C., Watkins, "Support vector machines for multi-class pattern recognition", In ESANN, 1999.

[46] X., Yang, M., Siu, "N-Best Tokenization in a GMM-SVM Language Identification System," IEEE, 2007.

[47] S. Young, *et al.*, HTK Book, http://htk.eng .cam.ac.uk/.

[48] C., Yu, G., Liu, S., Hahm, and J.H.L., Hansen, "Uncertainty Propagation in Front End Factor Analysis For Noise Robust Speaker Recognition", ICASSP, 2014.

[49] Q. Zhang, H. Boril, and J.H.L. Hansen, "Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification, " in Proc. ICASSP, Vancouver, Canada, May 2013.

شقایق رضا تحصیلات خود را در مقطع کارشناسی در رشته مهندسی پزشکی (بیوالکتریک) در دانشگاه صنعتی امیرکبیر (۱۳۸۵) و کارشناسی ارشد را در همان رشته و دانشگاه (۱۳۸۷) به پایان رساند. وی هم اکنون دانشجوی مقطع دکترای بیوالکتریک در دانشگاه صنعتی امیرکبیر است. موضوعات مورد علاقهٔ ایشان پردازش گفتار، پردازش سیگنال و تصویر است.
نشانی رایانامهٔ ایشان عبارت است از:
shaghayegh.reza@gmail.com

سیدجهانشاه کبودیان تحصیلات خود را در مقاطع کارشناسی، کارشناسی ارشد و دکترا در دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران) گذراند و دکترای خود را در سال ۱۳۸۹ از دانشگاه یادشده دریافت کرد. وی هم‌اکنون استادیار گروه مهندسی کامپیوتر و فناوری اطلاعات دانشگاه رازی کرمانشاه و همچنین مشاور ارشد پژوهشگاه توسعه فناوریهای پیشرفته خواجه نصیرالدین طوسی است. زمینه‌های پژوهشی مورد علاقهٔ ایشان عبارتند از پردازش سیگنال، پردازش گفتار و صوت، پردازش زبان طبیعی، شناسایی الگو، یادگیری ماشین و الگوریتم‌های فراابتکاری است.
نشانی رایانامهٔ ایشان عبارت است از:
kabudian@{razi,rcisp,aut}.ac.ir