پایه‌گذاری بستری نو و کارآمد در حوزهٔ بازشناسی گفتار فارسی

باقر باباعلی

دانشکدهٔ ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران، تهران، ایران

چکیده

برخلاف پیشینهٔ سی‌سالهٔ پژوهش در حوزهٔ بازشناسی گفتار فارسی در ایران و دست‌یافتن به پیشرفت‌های درخور توجه، نتایج عمده کارهای انجام‌شده به‌دلیل عدم وجود بستر یکسان، قابل مقایسه و ارزیابی دقیق نیستند. بستر بیش‌تر شامل سامانهٔ بازشناسی و دادگان با تعریف مشخص مجموعه‌های آموزش، توسعه و ارزیابی است. سامانهٔ متن‌باز کلدی با وجود نوظهوربودن آن ویژگی‌های منحصربه‌فردی دارد که در سال‌های اخیر مورد توجه اکثر آزمایشگاه‌های تراز نخست پردازش گفتار دنیا قرار گرفته است و با لحاظ همه جوانب، بهترین انتخاب موجود در راستای پایه‌گذاری این بستر برای تمامی زبان‌ها از جمله زبان فارسی است. در این مقاله پس از بررسی خصوصیات، توانمندی‌ها و اجزای مختلف نرم‌افزار کلدی؛ دادگان فارس‌دات را به‌دلیل ثبت رسمی و قابل دسترس‌بودن آن برای همگان از سراسر دنیا به‌عنوان بخش دیگر این بستر انتخاب کرده و به تأسی از انتخاب انجام‌شده بر روی دادگان TIMIT به تعریف مجموعه‌های آموزش، توسعه و ارزیابی می‌پردازیم. درنهایت بیش‌تر قریب به اتفاق تکنیک‌ها و روش‌های موجود در کلدی بر روی دادگان فارس‌دات، مطابق تعریف صورت گرفته، مورد آزمایش قرار گرفته‌اند. بهترین میزان خطای حاصل در بازشناسی واج برای مجموعه توسعه ۲۰/۳ درصد و برای مجموعه آزمون ۱۹/۸ بوده است. دسترسی به کدهای نوشته در جهت فراهم‌سازی این بستر، در نرم‌افزار کلدی موجود است که با توجه به متن‌بازبودن آن، دسترسی به آنها به‌منظور بازسازی نتایج آمده در این مقاله درصورت دراختیارداشتن دادگان فارس‌دات به‌راحتی قابل انجام است.

واژگان کلیدی : بازشناسی گفتار پیوسته فارسی، دادگان فارس دات، نرم‌افزار متن‌باز کلدی.

A state-of-the-art and efficient framework for Persian Speech Recognition

Bagher BabaAli

School of Mathematics, Statistics and Computer Sciences, College of Science, University of Tehran, Tehran, Iran

Abstract

Although researches in the field of Persian speech recognition claim a thirty-year-old history in Iran which has achieved considerable progresses, due to the lack of well-defined experimental framework, outcomes from many of these researches are not comparable to each other and their accurate assessment won't be possible. The experimental framework includes ASR toolkit and speech database which consists of training, development and test datasets. In recent years, as a state-of-the-art open-source ASR toolkit; Kaldi has been very well-received and welcomed in the community of the world-ranked speech researchers

* نویسندهٔ عهده‌دار مکاتبات

[DOI: 10.18869/acadpub.jsdp.13.3.51]

around the world. considering all aspects, Kaldi is the best option among all of the other ASR toolkits to establish a framework to do research in all languages, including Persian.
In this paper, we chose Fardat as the speech database which is the counterpart of TIMIT for Persian language because not only it has got a standard form but it's also accessible for all researchers around the world. Similar to the recipe on TIMIT database, we defined these three sets on the Farsdat: Training, Development and Test sets. After a survey on Kaldi's components and features, we applied most of state-of-the-art ASR techniques in the Kaldi on the Farsdat based on three sets definition. The best phone error rate on development and test set have been 20.3% and 19.8%. All of the codes and the recipe that was written by author have been submitted to Kaldi repository and they are accessible for free, so all the reported results will be easily replicable if you have access to Farsdat database.

Keyword: **Persian Continuous Speech Recognition, FarsDat Database, Kaldi Toolkit**

۱- مقدّمه

به‌منظور تسهیل ارتباط انسان با ماشین، پژوهش در حوزهٔ بازشناسی گفتار از حدود نیم‌قرن، همواره مورد توجه پژوهش‌گران علوم برق و رایانه بوده است. با توجه وابستگی این مقوله به زبان؛ لذا با تأسی از این پژوهش‌ها در دو دههٔ اخیر در حوزهٔ بازشناسی گفتار فارسی پژوهش‌های متعددی توسط پژوهش‌گران داخلی این عرصه صورت گرفته است [7]-[1].

با وجود دست‌یابی این پژوهش‌گران به نتایج قابل قبول و ارائهٔ یک‌سری محصولات تجاری بر مبنای آن؛ ولی از دیدگاه پژوهشی هم‌چنان مشکلات متعددی پابرجاست که در ادامه به آن می‌پردازیم. پژوهش‌های انجام‌شده بیش‌تر بر مبنای سامانه‌های بازشناسی گفتار و دادگان‌های متنوعی گزارش شده است که ارزیابی و مقایسه نتایج دست‌یافته را ناممکن می‌کند. برخی از این سامانه‌ها در آزمایشگاه خاصی توسعه داده شده و دسترسی به آن برای دیگران مقدور نیست. برخی از این پژوهش‌ها بر مبنای دادگان‌هایی بوده که چندان در دسترس همگان نبوده و یا استاندارد نیست. به این ترتیب مقایسه نتایج و ارزیابی تکنیک‌های ارائه شده ناممکن است.

بنابراین شکی نیست که در مقوله بازشناسی گفتار، فراهم‌آوردن شرایطی که قضاوت در مورد روش‌ها و تکنیک‌های ارائه شده توسط پژوهش‌گران را تسهیل کند یک نیاز اساسی است. برای تحقق این امر مهم می‌بایست یک سامانه بازشناسی گفتار پایه که به‌روز هم است، در دسترس همگان باشد. همچنین نتایج بر مبنای دادگان‌های استاندارد با شرایط مشابه ارائه شود که قضاوت را به‌مراتب روشن و آسان‌تر می‌کند. در راستای تحقق این امر، در کارگاه تابستانی سال ۲۰۰۹ در مرکز پژوهش زبان و گفتار دانشگاه جانز هاپکینز؛ هستهٔ اصلی یک سامانه بازشناسی گفتاری پیوسته‌ای شکل گرفت که کلدی[1] [8] نام گرفت. برخلاف این‌که از زمان پابه‌عرصه‌گذاشتن کلدی زمان زیادی نمی‌گذرد، ولی در بیت پژوهش‌گران علوم گفتار دنیا به‌دلیل ویژگی‌های منحصربه‌فرد آن به‌شدت مورد توجه قرار گرفته است. تمامی تکنیک‌های کارآمد و به‌روز بازشناسی گفتار یا در این سامانه پیاده‌سازی شده و یا در حال پیاده‌سازی است. تا این حد که شبکه‌های عصبی ژرف[2] [9]،[10] که داغ‌ترین مبحث پژوهشی در سال‌های اخیر در این عرصه است، برای مدل‌سازی آکوستیکی [11]،[12] و همچنین امتیازدهی مجدد[3] امتیاز مدل زبانی [13] پیاده‌سازی شده است و پیوسته در حال توسعه و بهبود است. از حدود دو دهه قبل در عرصهٔ بازشناسی گفتار، سامانه‌های متن‌باز متعددی نظیر HTK [14]، SONIC [15]، [16]، SPHINX [17]، [18]، RASR [20]،[19]، JULIUS [21]،SPRAAK [22]، KALDI و درهمین‌اواخر [23] عرضه شده است. با لحاظ همه جوانب و معیارها؛ تنها رقیب سامانه کلدی، سیستم RASR محسوب می‌شود که توسط آزمایشگاه RWTH دانشگاه آخن آلمان توسعه داده شده است. کلدی در مقایسه با RASR از معماری و ساختار نرم‌افزاری به‌مراتب بهتری برخوردار است که کار فهم کد، اعمال تغییرات و توسعه را برای پژوهش‌گران تسهیل می‌کند. و علاوه‌بر موارد ذکرشده، محدودیت قانونی به‌کارگیری در کاربردهای تجاری ندارد و عزم و ارائه توسعه‌دهندگان آن بر پیاده‌سازی آخرین تکنیک‌های کارآمد بازشناسی گفتار و به‌روز نگه‌داشتن آن راسخ‌تر بوده است.

با عنایت به برتری سامانه متن‌باز کلدی از جمیع جوانب و نیاز به وجود یک سامانه بازشناسی گفتار متن‌باز و

[1] Kaldi
[2] Deep Neural Network
[3] Rescoring

[DOI: 10.18869/acadpub.jsdp.13.3.51]

به‌روز برای انجام پژوهش‌ها در حوزه بازشناسی گفتار فارسی، بر آن شدیم که در سامانه کلدی، بستری برای این منظور بنا کنیم. از بین دادگان‌های گفتاری موجود برای زبان فارسی دادگان فارس‌دات [24]، استاندارد، ثبت شده و در سطح بین‌المللی عرصه شده است که تهیه آن را برای پژوهش‌گران در سراسر دنیا مقدور می‌کند. این دادگان توسط پژوهشگاه توسعه فناوری‌های پیشرفتهٔ خواجه نصیرالدین طوسی برای اهداف متعددی در حوزه زبان‌شناسی و پردازش گفتار زبان فارسی طراحی و گردآوری شده است و شباهت زیادی به دادگان انگلیسی TIMIT [25] دارد. به‌منظور تعادل فنوتیکی، جملات آن طوری انتخاب شده که پوشش خوبی بر روی واج‌ها و دوواجی‌های زبان فارسی داشته باشد. بنابراین متن برخی جملات ساختگی، یعنی چندان متداول نیست. با توجه به ساختگی‌بودن جملات و کوچک‌بودن حجم؛ این دادگان مشابه دادگان TIMIT برای بازشناسی واج مناسب است نه کلمه. بنابراین دادگان فارس‌دات جهت بنای بستر بازشناسی گفتار فارسی در سامانه کلدی انتخاب شد.

به‌طورمعمول در ارزیابی‌های متداول در حوزهٔ بازشناسی گفتار سه مجموعه آموزش[1]، توسعه[2] و ارزیابی[3] یا آزمون در نظر گرفته می‌شود. بر مبنای مجموعه آموزش؛ مدل‌های آکوستیکی واحدهای آوایی آموزش داده می‌شود. مجموعه توسعه برای تنظیم و بهینه‌سازی پارامترها از جمله وزن مدل زبانی استفاده می‌شود و در آخر براساس مدل‌های آموزش داده شده و پارامترهای تنظیم‌شده، مجموعه آزمون مورد آزمایش قرار می‌گیرد. با توجه به چندمنظوره‌بودن دادگان فارس‌دات، این سه مجموعه برای این دادگان تعریف نشده است. در انتخاب این سه مجموعه برای دادگان TIMIT سعی شده که هیچ‌گونه گوینده مشترکی بین آنها نباشد و همچنین بین متن جملات ادا شده توسط گویندگان مجموعه آموزش از یک طرف و مجموعه توسعه و آزمون از طرف دیگر جملهٔ یکسانی وجود ندارد [26]. در این مقاله نیز با لحاظ این دو محدودیت ذکرشده؛ بر روی دادگان‌فارس دات سه مجموعه ذکر شده تعریف شده و در آزمایش‌های انجام‌شده، بر مبنای آن عمل شده است. در این مقاله هدف تعریف یک بستر استاندارد و به‌روز در سامانه کلدی برای پژوهش‌گران بازشناسی گفتار فارسی بر روی دادگان فارس‌دات و ارائه نتایج حاصله به‌منظور انجام مقایسات بعدی است.

در ادامه، در بخش بعد به بررسی سامانه کلدی شامل روش‌ها، تکنیک‌های و قابلیت‌های موجود در آن می‌پردازیم. بخش سوم به بررسی دادگان فارس‌دات و شرایط آزمایش تعریف‌شده بر روی آن اختصاص داده شده است. نتایج و بررسی و تحلیل آنها در بخش چهارم آمده و درنهایت در بخش پنجم جمع‌بندی ارائه شده است.

۲- سامانه بازشناسی گفتارکلدی

کلدی یک سامانه متن باز برای بازشناسی گفتار است که در محیط در لینوکس به زبان C++ توسعه داده شده است هرچند که قابلیت به‌کارگیری در محیط ویندوز را هم دارد. خلق کلدی با هدف ارائه یک کد مدرن، انعطاف‌پذیر و ساده جهت فهم، تغییر و توسعه صورت گرفت. از ویژگی‌های مهم کلدی می‌توان این موارد را برشمرد : برای پیاده‌سازی بخش کدگشایی[4] به روش ایستای مبنی بر FST[5] [27] که به‌روزترین و کارآمدترین روش در این حوزه است، عمل شده است و برای پیاده‌سازی FST از کتابخانه متن باز OpenFST [28] بهره گرفته شده است. در روش‌های ایستا کدگشایی برخلاف روش‌های پویا؛ کل فضای جستجو به صورت یک گراف بزرگ، برون‌خط ساخته شده و در هنگام بازشناسی در آن فقط جستجو صورت می‌گیرد. اندازهٔ گراف حاصل به اندازهٔ مدل زبانی و مدل آکوستیکی و تعداد مدخل واژگان بستگی دارد. در کلدی برای انجام محاسبات ماتریسی؛ از یک طریق یک واسط نرم‌افزاری، توابع کتابخانه‌های استاندارد LAPACK و BLAS فراخوانی می‌شوند. الگوریتم‌ها در کلدی به عمومی‌ترین شکل ممکن طراحی شده‌اند، به‌نحوی که توسعه به‌راحتی قابل انجام باشد. برای مثال کدگشا در کلدی از طریق یک واسط نرم‌افزاری امتیاز هر قاب گفتار را به‌ازای هر حالت HMM دریافت می‌کند. این امتیاز می‌تواند هم بر مبنای SGMM[6] [29] یا GMM محاسبه شده باشد و یا از خروجی شبکه عصبی باشد. کلدی تحت لیسانس Apache v2,0 عرضه شده که یکی از کم محدودیت ترین‌هاست. در کلدی دستورالعمل کامل برای ساخت یک سامانه بازشناسی گفتار بر مبنای اکثر قریب به اتفاق دادگان‌های گفتاری استاندارد در قالب یک فایل bash ارائه شده است که کمک بزرگی به پژوهش‌گران محسوب می‌شود. برای بیش‌تر کدهای کلدی تولبعی جهت تست صحت عملکرد آنها نوشته شده است. هر چند در حال حاضر

[1] Training Set
[2] Development Set
[3] Evaluation Set

[4] Decoding
[5] Finite State Transducer
[6] Sub-space Gaussian Mixture model

[DOI: 10.18869/acadpub.jsdp.13.3.51]

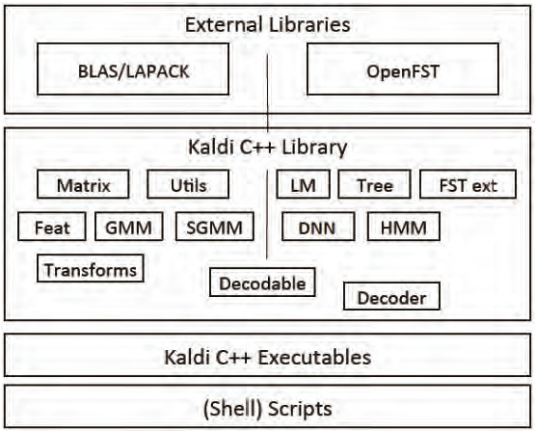
کلدی از روش‌های کارآمد و به‌روزی برای مدل‌سازی زبانی و کدگشایی بهره برده است و قابلیت کدگشایی برخط را هم دارد، ولی در ابتدا بیشتر با هدف بستری برای انجام پژوهش‌ها در حوزه مدل‌سازی آکوستیکی عرضه شد. در ادامه در این بخش به تشریح ساختار کد کلدی، توصیف مؤلفه‌های اصلی بازشناسی گفتار در کلدی نظیر بخش استخراج ویژگی، مدل‌سازی آکوستیکی، درخت تصمیم فونتیکی، مدل‌سازی زبانی و در آخر کدگشا خواهیم پرداخت.

۲-۱- ساختار نرم‌افزاری

در شکل (۱) ساختار نرم‌افزاری کلدی آمده است. همان‌طوری‌که در این شکل مشاهده می‌شود در آن از دو کتابخانه خارجی متن‌باز استفاده است : یکی OpenFST و دیگری BLAS[1]/LAPACK[2] که به‌منظور انجام محاسبات عددی به کار گرفته شده است. دیگر کتابخانه‌های کلدی که در شکل نمایش داده به‌صورت داخلی هستند. این کتابخانه‌ها به‌صورت طبقه در زبان C++ نوشته شده‌اند و فراخوانی توابع آنها از طریق برنامه‌های اجرای که به‌صورت خط فرمان اجرا می‌شود، صورت می‌گیرد. برای ایجاد یک سامانه بازشناسی، این توابع اجرایی در یک زبان اسکریپتی مانند bash فراخوانی می‌شوند. هر یک از این توابع اجرایی یک وظیفه مشخص برای آن تعریف شده است که از طریق آرگومان‌های خط فرمان به‌نحو دقیق‌تری مشخص می‌شود. همه این توابع اجرایی می‌توانند عمل خواندن از و یا نوشتن به خروجی را از طریق لوله[3] انجام دهند که این امکان به‌رشته درآوردن تعدادی از آنها را فراهم می‌کند.

۲-۲- استخراج ویژگی

سه روش استخراج ویژگی MFCC، PLP [30] و BNF[4] [31] موجود است و پارامترهای هر یک از این روش‌ها دارای مقدار پیش فرضی می‌باشد، هرچند که امکان تغییر و تنظیم آن توسط کاربر در قالب فایل Config وجود دارد. روش استخراج ویژگی BNF در سال‌های اخیر ارائه و مورد توجه قرار گرفته است. در این روش، بردار ویژگی از لایهٔ میانی یک شبکهٔ عصبی ژرف که ورودی آن بردار ویژگی فعلی به همراه تعداد محدودی بردارهای قبل و بعد (سه یا پنج) و خروجی آن حالت‌های تمام HMMهاست به‌دست می‌آید که الحاق آن با بردارهای ویژگی متداول، بهبود درخور توجهی را در دقت بازشناسی به همراه دارد. در نسخه برخط کلدی که دررهمین‌اواخر پیاده‌سازی شده‌است، به‌دلیل آن امکان به‌کارگیری مؤثر fMLLR[5] [32] و CMVN[6] [33] وجود ندارد، از بردار ویژگی ivector [34] بهره گرفته شده که به‌طورکامل افت دقت ناشی از عدم به‌کارگیری دو تکنیک مؤثر ذکر شده را جبران می‌کند، به‌نحوی که هیچ‌گونه تفاوتی در دقت بین نسخه برخط و برون‌خط کلدی وجود ندارد [12].



(شکل- ۱): نمایی ساده‌شده از اجزای نرم‌افزاری سامانه کلدی
(Figure-1): A simplified diagram on the software modules from the Kaldi toolkit

در روش استخراج ویژگی ivector که در سالیان اخیر بیش‌تر برای بازشناسی گوینده ارائه شده است، از گفتار با طول متغیر یک بردار ویژگی با طول ثابت استخراج می‌کند که گویای خصوصیات گوینده و مستقل از متن گفتار است. با الحاق این بردار ویژگی به بردار ویژگی MFCC یا PLP و به‌کارگیری آن در بازشناسی گفتار، اثر عدم اعمال fMLLR و CVN در حالت به‌کارگیری برخط جبران می‌شود. در بخش استخراج ویژگی کلدی، امکان به‌کارگیری VTLN [35]، [36]، به‌منظور هنجارسازی اثر طول کانال صوتی گوینده و CMVN به‌منظور حذف اثر کانال و اعمال تبدیلات خطی LDA، HLDA[7] [37] و MLLT[8] [38] وجود دارد. از آنجا که به‌کارگیری فرکاس گام در بهبود دقت بازشناسی زبان‌های آهنگین نظیر چینی و ویتنامی دارد [39]، در کلدی یک روش استخراج فرکانس گام، اختصاصی‌شده به‌منظور

[1] Basic Linear Algebra Subroutines
[2] Linear Algebra PACKage
[3] Pipe
[4] Bottleneck Feature
[5] Feature-space Maximum Likelihood Linear Regression
[6] Cepstral Mean and Variance Normalization
[7] Heteroscedastic Linear Discriminant Analysis
[8] Maximum Likelihood Linear Transform

[DOI: 10.18869/acadpub.jsdp.13.3.51]

بازشناسی گفتار ارائه شده است. در روش ارائه‌شده [40] به‌ازای هر قاب گفتار دو مؤلفه POV[1] و فرکانس گام که در بخش‌های بی‌صدا مقدار آن درون‌یابی شده؛ استخراج و به سایر ویژگی‌های استخراج‌شده از آن قاب گفتار الحاق می‌شود.

روند مرسوم در کلدی برای استخراج ویژگی به این صورت است: در آموزش نخستین سامانه بر مبنای تک‌واج[2]، بردار ویژگی MFCC یا PLP به همرا مشتقات نخست و دوم آن به‌صورت یک بردار ۳۹ مؤلفه‌ای استفاده می‌شود. در آموزش و ارزیابی سامانه بر مبنای سه‌واجی[3] بر روی هر بردار ویژگی ۱۳ مؤلفه‌ای به همراه سه بردار ویژگی قبل و بعد تبدیل LDA و سپس تبدیل MLLT اعمال شده و درنهایت به یک بردار ویژگی چهل مؤلفه‌ای تبدیل می‌شود. با توجه به خطی‌بودن این تبدیلات، ماتریس آنها در مرحلهٔ آموزش و بر مبنای داده‌های آموزشی محاسبه می‌شود.

۲-۳- مدل‌سازی آکوستیکی

در کلدی سه روش مدل‌سازی اکوستیکی GMMs، SGMMs و DNN موجود است و با توجه به تدابیر اندیشیده شده در معماری کلدی، به‌راحتی می‌توان سایر روش‌های مدل‌سازی را هم پیاده‌سازی و به‌کار گرفت.

۲-۳-۱- مدل‌های مخلوط گوسی (GMM)

مدل‌های مخلوط گوسی با ماتریس کواریانس کامل و قطری در کلدی موجود است و پیاده‌سازی آنها به‌نحوی صورت گرفته که محاسبه میزان شباهت به‌ازای هر GMM تنها با ضرب نقطه‌ای قابل انجام است که افزایش سرعت زیادی را به همراه دارد. در کلدی امکان توصیف مجزای توپولوژی مدل مخفی مارکوف مربوط به هر واج وجود دارد. توپولوژی هر مدل مخفی مارکوف شامل تعداد حالت‌ها و نحوهٔ پرش بین آنها است. در کلدی مقدار پیش‌فرض تعداد حالت‌ها برای واج‌ها سه حالت است و برای سکوت و نوفه و سایر اصوات غیرکلامی پنج حالت در نظر گرفته شده است. بطور واضح مدل HMM واج‌ها چپ به راست و مدل HMM سکوت و نوفه و اصوات غیرکلامی ارگودیگ[4] است؛ یعنی از هر حالت به تمام حالات دیگر امکان پرش وجود دارد.

[1] Probability Of Voicing
[2] Monophone
[3] Triphone
[4] Ergodic

۲-۳-۲- مدل‌های مخلوط گوسی زیرفضا (SGMM)

این روش مدل‌سازی آکوستیکی با الهام از بحث تحلیل مؤلفه‌های مشترک[5] که یک رویکرد جدید در مبحث بازشناسی گوینده محسوب می‌شود، در کارگاه تابستانی سال ۲۰۰۹ مرکز زبان و گفتار دانشگاه جانزهاپکینز ارائه شد. به‌طورکلی در این روش یک مدل مخلوط جهانی (UBM) بر مبنای کل گفتار آموزش داده می‌شود و آنگاه با تطبیق آن به مدل‌های مربوط به هر حالت می‌رسیم. در مواقعی که داده آموزشی در محدوده[6] کمی در اختیار است، به‌عنوان‌مثال بازشناسی گفتار برای زبان‌ها با منابع محدود[7]، این روش امکان بهرمندی از داده خارج از محدوده[8] و یا خارج از زبان[9] را جهت جبران کمبود داده در این موارد فراهم می‌کند [29].

۲-۳-۳- شبکه‌های عصبی ژرف (DNN)

به‌کارگیری شبکهٔ عصبی به‌صورت ترکیب با HMM (به جای GMM) برای مدل‌سازی آکوستیکی به‌دلیل قدرت تمایز بالای آن در مقایسه با GMM در دهه نود میلادی مورد توجه پژوهش‌گران بازشناسی قرار گرفت [41]، [42]. در روش ترکیبی، یک شبکهٔ عصبی جهت تخمین احتمال پسین[10] هر بردار ویژگی به‌ازای حالات HMMها آموزش داده می‌شود که درنهایت با تقسیم آنها بر احتمال پیشین[11] هر حالت؛ به میزان شباهت[12] حالات می‌رسیم. در تلاش‌های نخستین که در دهه نود صورت گرفت، به‌دلیل محدودبودن توان محاسباتی، تعداد لایه‌های مخفی شبکهٔ عصبی یک و یا حداکثر دو لایه بود و هیچ‌گونه تلاشی در جهت مقداردهی نخستین وزن‌های شبکهٔ عصبی به‌صورت هوشمندانه صورت نمی‌گرفت. موفقیت رویکرد ترکیبی[43]،[45]،[10]،[12]، در سال‌های اخیر بیش‌تر مدیون بکارگیری شبکه‌های عصبی ژرف (تعداد لایه‌های مخفی بیشتر) به لطف افزایش توان محاسباتی رایانه‌ها و به‌کارگیری ماشین بولتزمن محدودیت‌دار (RBM[13]) [46] برای پیش‌آموزش شبکهٔ عصبی بوده است. در کلدی دو کد برای پیاده‌سازی سامانه ترکیبی به‌صورت مجزا توسط Daniel Povey و Karel Vesely توسعه داده شده است. عمده تمایز این دو سامانه در

[5] Joint Factor Analysis
[6] In-domain
[7] Low Resource Languages
[8] Out-of-domain
[9] Out-of-language
[10] Posterior Probability
[11] Prior Probability
[12] Likelihood
[13] Restricted Boltzmann Machine

[DOI: 10.18869/acadpub.jsdp.13.3.51]

آن است که Karel مقداردهی نخستین شبکه را با روش پیش‌آموزش RBM و Daniel به‌صورت تصادفی انجام می‌دهد. این دو سامانه به‌لحاظ دقت بازشناسی، بسیار مشابه هم بوده و در برخی مواقع سامانه Karel به میزان کمی بهتر عمل می‌کند. با توجه به این‌که به‌کارگیری شبکهٔ عصبی ژرف در بازشناسی گفتار، داغ‌ترین مبحث پژوهشی این حوزه در سالیان اخیر بوده است، در سامانه کلدی پیوسته روش‌های جدیدی در این حوزه توسط پژوهش‌گران سراسر دنیا توسعه داده شده و آزمایش می‌شود که پرداختن به آن خارج از حوصله این مقاله است.

۲-۳-۴- تطبیق گوینده

تطبیق گوینده[1] در کلدی به دو صورت، یکی در فضای مدل (MLLR) و دیگری در فضای ویژگی (fMLLR) قابل انجام است. برای هر دو روش می‌بایست تبدیلات لازم با استفاده از درخت رگرسیون تخمین زده شود [47]. همچنین امکان هنجارسازی گوینده با استفاده از VTLN و یا به‌صورت کلی‌تر هنجارسازی جنسیت[2] با استفاده از تبدیل‌نمایی[3] [48] وجود دارد. برای آموزش تطبیق به گوینده (SAT[4]) مدل‌های آکوستیکی، امکان استفاده از هر دو روش VTLN و fMLLR وجود دارد که استفاده از fMLLR متداول‌تر است.

۲-۳-۵- درخت‌های تصمیم فونتیکی

در کلدی کد ساخت درخت تصمیم فونتیکی به‌نحوی است که امکان مدل‌سازی بافت[5] با هر طولی را مقدور می‌سازد و همچنین به‌اندازهٔ کافی عمومی است که امکان به‌کارگیری طیف وسیعی از روش‌ها را دارد. روش کلاسیک به این صورت است که به ازای هر حالت HMM مربوط به هر تک‌واج یک درخت تصمیم وجود دارد، که بر مبنای سؤالاتی در مورد واج‌های سمت چپ و راست تصمیم‌گیری می‌کند. در کلدی، ریشه‌های درخت‌های تصمیم بین واج‌ها و حالت‌های یک واج می‌تواند به اشتراک گذاشته شود. سؤالات فونتیکی بر مبنای دانش زبان‌شناسی می‌تواند تهیه شود؛ ولی در کلدی سؤالات بر مبنای خوشه‌بندی درختی[6] واج‌ها به‌صورت خودکار تولید می‌شوند [49]. سؤالات در مورد چیزهایی نظیر استرس فونتیکی – درصورتی‌که در واژگان آمده باشد- و اطلاعات شروع و پایان کلمات منجر به یک مجموعه واج گسترش‌یافته می‌شود. در این موارد ریشه‌های درخت‌های تصمیم بین صورت‌های مختلف یک واج مشترک است.

۲-۴- مدل‌سازی زبانی

از آنجا که کدگشای کلدی بر مبنای روش استاتیک FST عمل می‌کند، بنابراین قادر به استفاده از هر گونه مدل زبانی است که به یک FST قابل تبدیل باشد. ابزاری جهت تبدیل مدل زبانی از فرمت ARPA به شکل FST وجود دارد. در مواقعی که نیاز به هرس مدل زبانی[7] باشد از نرم‌افزار IRSTLM [50] استفاده می‌شود. برای ساخت مدل زبانی از روی متن خام نیز همین نرم‌افزار IRSTLM قابل استفاده است؛ هرچند که می‌توان از نرم‌افزار قدرتمند SRILM [51] هم استفاده کرد.

به‌تازگی استفاده از شبکه‌های عصبی بازگشتی[8] برای مدل‌سازی زبانی مورد توجه پژوهش‌گران قرار گرفته است [13]. در کلدی این ابزار مدل‌سازی زبانی که به‌طور مجزا توسعه داده شده، موجود است و امکان به‌کارگیری آن فراهم شده است. یک روش متداول برای به‌کارگیری آن، امتیازدهی مجدد امتیاز مدل زبانی خروجی‌های بازشناسی (N-best) است که به‌طورمعمول بهبود دقت بازشناسی را به همراه دارد.

۲-۵- کدگشا

در کلدی مرحلهٔ بازشناسی گفتار از WFST استفاده می‌شود. فرایند ساخت گراف کدگشایی بر مبنای روش آمده در [27] است هر چند که تغییراتی به‌منظور بهبود انجام شده است. به‌عنوان‌مثال این تغییر صورت گرفته که FST تصادفی[9] باشد؛ به این معنی که مجموع اوزان خروجی‌ها در هر حالت برابر یک باشد. مدل‌های HMM واج‌ها (H)، وابستگی به متن (C)، واژگان (L) و مدل زبانی (G) به‌صورت FST های مجزا با هم ترکیب شده و گراف بزرگی تحت عنوان HCLG را می‌سازند که فضای جستجو را در هنگام کدگشایی شکل می‌دهد.

در کلدی، بسته به برخط یا برون‌خط‌بودن فرایند بازشناسی، نوع خروجی (شبکه[10] یا n-تا بهترین خروجی)، میزان بهینه‌بودن و سرعت اجرا، کدگشاهای متعددی وجود دارد. همگی کدگشاها این ویژگی را دارند که به‌راحتی

[1] Speaker Adaptation
[2] Gender Normalization
[3] Exponential Transform
[4] Speaker Adaptive Training
[5] Context
[6] Tree-Clustering
[7] Language Model Pruning
[8] Recurrent Neural Network
[9] Stochastic
[10] Lattice

[DOI: 10.18869/acadpub.jsdp.13.3.51]

می‌توان نوع مدل آکوستیکی آنها را تغییر داد. کدگشاهای موجود در کلدی در سطح کد C++ همگی تک‌گذره هستند و چندگذره‌بودن در سطح اسکریپت پیاده‌سازی شده است.

۳-دادگان فارس‌دات

فارس‌دات اولین دادگان استاندارد گفتاری زبان فارسی است که با هدف مطالعه مبانی و مدل‌سازی آکوستیکی زبان فارسی به‌منظور به‌کارگیری در سامانه‌های بازشناسی گفتار فارسی توسط پژوهشگاه توسعه فناوری‌های پیشرفتهٔ خواجه نصیرالدین طوسی تولید شده است [24]. این دادگان شامل ۳۰۴ گوینده فارسی‌زبان با یکی از ده لهجه رایج تهرانی، ترکی، اصفهانی، شمالی، جنوبی، لری، کردی، یزدی، خراسانی و بلوچی با سن، جنسیت و میزان تحصیلات متنوع است. نسبت گویندگان مرد به زن دو به یک است. هر گوینده بیست جمله را در دو جلسه مجزا در یک اتاق آکوستیک به سبک رسمی بیان کرده است و بنابراین تعداد کل جملات ۶۰۸۰ است. جملات به صورت ۱۶بیتی با فرکانس نمونه‌برداری ۲۲۰۵۰ هرتز و با نسبت سیگنال به نویز[1] متوسط ۳۲ دسی‌بل ضبط شده است. تعداد کل جملات فارس‌دات ۴۰۵ جمله است که دو جمله آن شامل کل واج‌های زبان فارسی (به جز /f/) است و این دو جمله توسط همه گویندگان بیان شده است. جملات به‌گونه‌ای انتخاب شده‌اند که شامل تمامی دوواجی‌های رایج زبان فارسی باشد. فارس‌دات استاندارد مؤسسه اروپایی داده‌های زبانی[2] را کسب کرده و در کاتالوگ این مؤسسه ثبت شده است.

در ارزیابی‌های متداول در حوزهٔ بازشناسی گفتار در دنیا، دادگان گفتاری به سه زیرمجموعهٔ آموزش، توسعه و آزمون یا ارزیابی تقسیم می‌شود. بر روی مجموعهٔ آموزش مدل آکوستیکی آموزش داده می‌شود. بر روی مجموعه توسعه، بهینه‌سازی و تنظیم پارامترها انجام می‌شود و درنهایت دقت بر روی مجموعهٔ آزمون با فرض مشخص و ثابت‌شدن پارامترها بر مبنای مجموعهٔ توسعه گزارش می‌شود. با عنایت به این که فارس‌دات برای اهداف متنوعی گردآوری شده، بر روی آن این سه مجموعه به‌طور رسمی تعریف نشده است؛ و در نتایج گزارش‌شده، مجموعه توسعه در نظر گرفته نشده است و یا اشتراکی بین این سه مجموعه از نظر گوینده و یا متن جمله وجود دارد. در این پژوهش با بررسی دقیق این دادگان و الهام از آنچه برای دادگان TIMIT در [26] صورت گرفته است، سعی شد این سه مجموعه را انتخاب کنیم. با لحاظ این محدودیت که بین سه مجموعه، گویندهٔ مشترکی نباشد و متن جملات مجموعهٔ آموزش با مجموعه توسعه و آزمون اشتراکی نداشته باشد، این انتخاب صورت گرفت. همچنین سعی شده تا حد امکان توزیع نوع گویش، نسبت مرد به زن, میزان تحصیلات و سن بین این سه مجموعه مطابق توزیع کل دادگان باشد که چندان کار راحتی نبود و نیاز به بررسی و استخراج آمارهای متعدد و لحاظ آنها داشت. مشخصات سه مجموعه انتخاب شده، مطابق جدول (۱) است. همان‌طوری‌که که درقبل ذکر شد، تعداد کل جملات بیان‌شده در فارس‌دات ۶۰۸۰ است؛ درحالی‌که مجموع جملات سه مجموعه تعریف شده (۳۹۹۴+۴۷۵+۲۸۷) ۴۷۵۶ جمله است. شماره گویندگان هریک از این سه مجموعه نیز در جدول (۲) آمده است. برخی از جملات بیان‌شده توسط بعضی از گویندگان کنار گذاشته شده است؛ به این دلیل که این محدودیت را رعایت کرده باشیم : هیچ گونه جمله با متن مشابه بین مجموعه آموزش از یک طرف و مجموعه توسعه و آزمون وجود نداشته باشد. در تعریف این سه مجموعه بر روی دادگان TIMIT هم به خاطر رعایت این محدودیت، تعدادی از جملات بیان شده (در همین مقیاس) کنار گذاشته شده است.

(جدول-۱): مشخصات سه مجموعه انتخاب شده بر روی دادگان فارس‌دات

(Table-1): Specification of three datasets defined on the FarsDat database

مجموعه	تعداد گویندگان (نفر)	تعداد جملات
آموزش	224	3994
توسعه	50	475
آزمون	30	287

۴-نتایج

سامانه کلدی در محیط لینوکس توسعه داده شده است و کدنویسی آن در سه سطح صورت گرفته است. به‌طورمعمول توابع پایه‌ای به زبان برنامه نویسی C++ نوشته می‌شود و این توابع در محیط شل اسکریپت برای انجام کاری خاص نظیر استخراج ویژگی، آموزش، اعمال تبدیلات به‌منظور مقاوم‌سازی و کدگشایی فراخوانی می‌شوند. برای دادگان‌های استاندارد متعددی، در قالب یک فایل شل، کارهای

[1] Signal to Noise Ratio
[2] European Linguistic Resource Association (ELRA)

[DOI: 10.18869/acadpub.jsdp.13.3.51]

آماده‌سازی و آمایش دادگان، واژگان و مدل زبانی، و فراخوانی فایل‌های شل به‌منظور استخراج ویژگی، آموزش و کدگشایی صورت گرفته است. برای دادگان فارس‌دات هم این کار توسط نویسنده این مقاله انجام گرفته است.

کدهای لازم برای آماده‌سازی سه مجموعهٔ آموزش، توسعه و آزمون تعریف‌شده بر روی دادگان فارس و انجام آزمایش‌های زیر، در سامانه کلدی توسط نگارنده مقاله، نوشته شده و در اختیار همگان است. با دراختیارداشتن دادگان فارس‌دات و نصب سامانه کلدی، کلیه نتایج آمده در ادامه به‌راحتی و بدون هیچگونه زحمت برنامه‌نویسی قابل تولید مجدد است. به این ترتیب پژوهش‌گران حوزه بازشناسی گفتار زبان فارسی به‌راحتی می‌توانند به پیاده‌سازی ایده‌های خود و مقایسه آن در یک بستر بروز بپردازند.

(جدول-۲): شماره گویندگان سه مجموعه انتخاب شده بر روی دادگان فارس‌دات

(Table-2): Speaker numbers of three datasets defined on the FarsDat database

مجموعه	شماره گویندگان
آموزش	001 002 003 004 005 006 008 011 012 013 014 015 016 017 019 020 021 022 023 024 025 026 027 028 029 030 031 033 034 035 038 039 040 041 042 043 045 047 048 050 051 052 053 054 055 056 057 058 059 060 061 062 064 065 066 067 068 069 070 071 072 073 074 075 076 077 078 079 081 082 083 084 085 086 087 088 089 091 092 093 094 095 096 099 100 101 102 104 105 106 107 108 109 110 111 112 114 116 117 119 120 121 122 123 124 126 127 128 129 130 131 132 133 134 136 137 138 140 141 142 145 146 147 148 151 153 154 156 157 158 160 161 163 164 165 166 170 171 173 174 175 176 177 178 179 180 182 183 184 185 186 187 188 189 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 207 208 210 211 215 216 216 217 218 220 221 223 224 227 228 229 230 230 231 232 234 235 237 238 240 241 243 244 245 246 247 248 249 251 252 253 256 257 258 260 261 262 264 265 266 267 268 269 279 293 297 298 300 301 302 303 304
توسعه	009 018 036 046 049 063 080 098 103 113 115 118 135 144 149 150 159 162 168 169 190 206 209 214 222 226 239 254 255 259 263 270 271 272 276 277 278 281 282 284 285 286 287 289 290 291 292 294 296 299
آزمون	007 010 032 037 044 090 097 125 139 143 152 155 167 172 181 212 213 219 225 233 236 242 250 273 274 275 280 283 288 295

(جدول-۳): درصد خطای بازشناسی واج برای مجموعه توسعه و مجموعه آزمون یا ارزیابی تعریف شده بر روی دادگان فارس‌دات به ازای روش‌های مختلف استخراج ویژگی‌های و مدل‌سازی آکوستیکی در نرم افراز کلدی

(Table-3): Percent of Phoneme error rate (PER) on the development & evaluation datasets defined on the FarsDat database for applying various methods of feature extraction and acoustic modeling in the Kaldi Toolkit

شماره	ویژگی	مدل	خطای بازشناسی واج (٪)	
			مجموعه توسعه	مجموعه آزمون
1	MFCC + delta + delta-delta	Mono-phone	29.8	30.5
2	MFCC + delta + delta-delta	Tri-phone	25.8	25.1
3	MFCC + LDA + MLLT	Tri-phone	24.4	23.3
4	MFCC + LDA + MLLT + SAT	Tri-phone	21.2	21.8
5	MFCC + LDA + MLLT + SAT	SGMM	20.2	19.9
6	MFCC + LDA + MLLT + SAT	SGMM + MMI	20.3	19.8
7	MFCC + LDA + MLLT + SAT	Karel-DNN	20.3	19.8

[DOI: 10.18869/acadpub.jsdp.13.3.51]

در جدول (۳) نتایج حاصل از انجام بازشناسی مجموعه توسعه و مجموعه آزمون یا ارزیابی به‌ازای ویژگی‌های مختلف استخراج‌شده و روش‌های مختلف مدل‌سازی آکوستیکی آمده است. سطر نخست نتایج را به‌ازای ساده‌ترین حالت یعنی روش استخراج ویژگی MFCC و مشتقات نخست و دوم آن و مدل‌سازی آکوستیکی مبتنی بر تک‌واج نشان می‌دهد. در سطر دوم، فقط مدل‌سازی از تک‌واجی به سه‌واجی تغییر کرده است که طبق انتظار، بهبود درخور توجهی در کاهش خطای بازشناسی واج بر روی هر دو مجموعه توسعه و آزمون را به همراه داشته است. در سطر سوم جدول، در مرحله استخراج ویژگی به جای بردار ویژگی MFCC و مشتقات نخست و دوم آن، بر روی هر بردار ویژگی MFCC سیزده‌مؤلفه‌ای به همراه سه بردار ویژگی قبل و بعد که تشکیل یک بردار ۹۱ مؤلفه‌ای می‌دهد دو تبدیل خطی LDA و MLLT به‌ترتیب اعمال شده و درنهایت به یک بردار چهل‌مؤلفه‌ای می‌رسیم که مدل‌سازی سه‌واجی بر مبنای آن صورت گرفته است. با مقایسهٔ نتایج این مرحله و مرحلهٔ قبل، به این نتیجه می‌رسیم که ویژگی MFCC + LDA + MLLT نسبت به MFCC و مشتقات نخست و دوم آن خطای بازشناسی واج را به‌طور مطلق حدود ۱/۵ درصد کاهش داده است و در مراحل بعدی از این ویژگی استفاده خواهیم کرد. در مرحلهٔ چهارم هم در مرحلهٔ آموزش آن، به کمک تکنیک fMLLR آموزش تطبیقی به گوینده (SAT) صورت گرفته است و هم در مرحلهٔ بازشناسی به روش fMLLR ، ویژگی‌های هر گوینده تطبیق داده شده است که سرجمع بر روی مجموعهٔ توسعه ۳/۲ درصد و بر روی مجموعه آزمون ۱/۵درصد به‌طور مطلق بهبود دقت به همراه داشته است. در مرحلهٔ پنجم به‌کارگیری مدل‌سازی SGMM موجب بهبود یک‌درصدی بر روی مجموعهٔ توسعه و بهبود حدود دودرصدی بر روی مجموعه آزمون شده است. و درنهایت اعمال آموزش تمایزی[1] MMI [26] حدود یک‌صدم درصد بهبود به همراه داشته است.

در نهایت در سطر آخر جدول نتیجه به‌کارگیری شبکهٔ عصبی ژرف برای مدل‌سازی آکوستیکی سه‌واجی‌ها آمده که نسبت به بهترین دقت حاصله در دو مرحلهٔ قبل هیچ‌گونه بهبودی به همراه ندارد که دلیل آن هم کوچک‌بودن اندازهٔ مجموعه آموزشی است. در کارهای آتی به تبیین اثر به‌کارگیری شبکه‌های عصبی ژرف در بازشناسی گفتار فارسی بر روی دادگان‌های بزرگتر فارسی خواهیم پرداخت.

[1] Maximum Mutual Information

۵- جمع‌بندی

در این مقاله، نگارنده با اعتقاد به این که در حال حاضر سامانه متن‌باز بازشناسی گفتار کلدی، مدرن‌ترین و کارآمدترین سامانه حال حاضر دنیاست، ابتدا به بررسی ویژگی‌ها و توانمندی‌های مختلف آن پرداخته است. یکی از ویژگی‌های منحصر به فرد کلدی، وجود کدهای آماده در آن برای انجام بازشناسی بر روی دادگان‌های گفتاری متداول به زبان‌های مختلف است؛ که کمک بزرگی به پژوهش‌گران در بازسازی نتایج قبلی و ارائه نتایج جدید در یک بستر یکسان و قابل دسترس برای همه محسوب می‌شود. در راستای بنای یک بستر همگانی و استاندارد برای پژوهش‌گران فعال در عرصهٔ بازشناسی گفتار فارسی، در این مقاله بر مبنای دادگان فارس‌دات و با الهام از کار صورت‌گرفته برای دادگان TIMIT، مجموعه‌های آموزش، توسعه و ارزیابی انتخاب شده و دقت بازشناسی با اعمال تکنیک‌های مختلف موجود در کلدی محاسبه شده است. بهترین میزان خطای حاصل در بازشناسی واج برای مجموعه توسعه ۲۰/۳ درصد و برای مجموعه آزمون ۱۹/۸ بوده است. در این کار، بررسی گسترده اثر به‌کارگیری شبکه‌های عصبی ژرف و سامانه‌های ترکیبی، استخراج ویژگی BNF، به‌کارگیری فرکانس گام برای زبان فارسی در بستر ارائه شده در این مقاله مد نظر است.

سپاس‌گزاری

نگارندهٔ مقاله بر خود لازم می‌داند که از همکاری و حمایت بی‌دریغ آقای دکتر دنیل پووی (Daniel Povey) پژوهش‌گر ارشد مرکز پردازش زبان و گفتار (CLSP) دانشگاه جانز هاپکینز، دکتر حسین صامتی عضو هیئت علمی دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف و مدیر آزمایشگاه پردازش گفتار آن دانشکده، مهندس خسرو حسین‌زاده مدیر فنی شرکت عصرگویش‌پرداز، دکتر محمد بحرانی عضو هیئت علمی گروه زبان و زبان‌شناسی دانشگاه صنعتی شریف و مدیر بخش پردازش زبان‌های طبیعی شرکت عصرگویش‌پرداز، دکتر هادی ویسی عضو هیئت علمی دانشکده علوم و فنون دانشگاه تهران و مدیر عامل شرکت عصرگویش‌پرداز، دکتر دیگو جیولیانی (Diego Jiuliani) پژوهش‌گر آزمایشگاه تکنولوژی زبان‌های گفتاری مؤسسه پژهشی برنوکسلر کشور ایتالیا صمیمانه تشکر کند.

[DOI: 10.18869/acadpub.jsdp.13.3.51]

6-Refrence

۶- مراجع

[1] B. BabaAli and H. Sameti, The Sharif speaker-independent large vocabulary speech recognition system, in Proceedings of the 2nd Workshop on Information Technology & Its Disciplines (WITID '04), pp. 24–26, Iran, 2004.

[2] H. Sameti, H. Veisi, M. Bahrani, B. Babaali, K. Hosseinzadeh, A large vocabulary continuous speech recognition system for Persian language, in EURASIP Journal on Audio, Speech, and Music Processing 2011, 2011:6

[3] F. Almasganj, S.A. Seyyed Salehi, M. Bijankhan, H. Razizade, M. Asghari, Shenava 2: a Persian continuous speech recognition software, in The first workshop on Persian language and Computer, pp. 77–82, Tehran, 2004.

[4] M. Sheikhan, M. Tebyani, M. Lotfizad, Continuous speech recognition and syntactic processing in iranian farsi language. Inter J Speech Technol. 1(2), 135 (1997). doi:10.1007/BF02277194

[5] S.M. Ahadi, Recognition of continuous Persian speech using a medium-sized vocabulary speech corpus, in European Conference on Speech communication and technology (Eurospeech'99), pp. 863–866, Switzerland, 1999.

[6] N. Srinivasamurthy, S.S. Narayanan, Language-adaptive Persian speech recognition, in European Conference on Speech Communication and Technology (Eurospeech'03), Switzerland, 2003.

[7] H. Sameti, H. Veisi, M. Bahrani, B. Babaali, K. Hosseinzadeh, Nevisa, a Persian continuous speech recognition system, in Communications in Computer and Information Science (Springer Berlin Heidelberg), pp. 485–492, 2008.

[8] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, The Kaldi Speech Recognition Toolkit, in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011.

[9] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82–97, 2012.

[10] F. Seide, G. Li, and D. Yu, Conversational speech transcription using context-dependent deep neural networks. in Proc. of Interspeech, pp. 437–440, 2011.

[11] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, Sequence discriminative training of deep neural networks, in Proc. of Interspeech, pp. 2345–2349, 2013.

[12] D. Povey, X. Zhang, and S. Khudanpur, Parallel training of Deep Neural Networks with Natural Gradient and Parameter Averaging, in Proc. of 3rd International Conference on Learning Representations (ICLR2015), USA, 2015

[13] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, Recurrent neural network based language model, in Proc. of Interspeech, pp. 1045-1048, Japan, 2010

[14] S.Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for version 3.4). Cambridge Univ. Eng. Dept., 2009.

[15] B. Pellom, SONIC: The University of Colorado Continuous Speech Recognizer, Technical Report TRCSLR-2001-01, Center for Spoken Language Research, University of Colorado, USA, 2001.

[16] B. Pellom and K. Hacioglu, Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task, in Proc. of ICASSP, Hong Kong, 2003.

[17] K.F. Lee, H.W. Hon, and R. Reddy, An overview of the SPHINX speech recognition system, in IEEE Transactions on Acoustics, Speech and Signal Processing 38.1, pp.35-45, 1990.

[18] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, Sphinx-4: A flexible Open Source Framework for Speech Recognition, Sun Microsystems Inc., Technical Report SML1 TR2004-0811, 2004.

[19] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, The RWTH Aachen University Open Source Speech Recognition System, in Proc. of Interspeech, pp. 2111–2114, 2009.

[20] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney: RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit, In IEEE Automatic Speech Recognition

[DOI: 10.18869/acadpub.jsdp.13.3.51]

and Understanding Workshop (ASRU), USA, 2011.

[21] A. Lee, T. Kawahara, and K. Shikano, JULIUS - an open source real-time large vocabulary recognition engine, in Proc. of INTERSPEECH, pp. 1691-1694, 2001.

[22] K. Demuynck, J. Roelens, D.V. Compernolle and P. Wambacq, SPRAAK: An Open Source SPeech Recognition and Automatic Annotation Kit, In Proc. Interspeech, pp. 495-498, Australia, 2008.

[23] D. Bolaños, The Bavieca Open-Source Speech Recognition Toolkit, in Proc. of IEEE Workshop on Spoken Language Technology (SLT), Miami, FL, USA, 2012.

[24] M. Bijankhan and M. J. Sheikhzadegan, FARSDAT—the speech database of Farsi spoken language, in Proceedings of the 5th Australian International Conference on Speech Science and Technology (SST '94), pp. 826–829, Perth, Australia, December 1994.

[25] V. Zue, S. Seneff, and J. Glass, Speech database development at MIT: TIMIT and beyond, Speech Communication, vol. 9, no. 4, pp. 351–356, 1990.

[26] K.F. Lee, and H.W. Hon, Speaker-independent phone recognition using hidden Markov models, in IEEE Transactions on Acoustics, Speech and Signal Processing, vol 37, no. 11, pp. 1641-1648, 1989.

[27] M. Mohri, F. Pereira, and M. Riley,Weighted finite-state transducers in speech recognition, in Computer Speech & Language 16, no. 1, pp. 69-88, 2002.

[28] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, OpenFst: a general and efficient weighted finite-state transducer library, in Proc. CIAA, 2007.

[29] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafit, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, The subspace Gaussian mixture model -A structured model for speech recognition, in Computer Speech & Language, vol. 25, no. 2, pp. 404 – 439, 2011.

[30] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, Journal of the Acoustical Society of America, vol. 87, no.4, pp. 1738–1752, 1990.

[31] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, Probabilistic and bottle-neck features for LVCSR of meetings, in Proc. ICASSP, pp. 757–760, 2007.

[32] M. J. F. Gales, Maximum likelihood linear transformations for HMM based speech recognition, in Computer Speech and Language, vol. 12, no. 2, pp. 75–98, 1998.

[33] S. Furui, Cepstral analysis technique for automatic speaker verification, in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 2, pp. 254–272, 1981.

[34] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, Front-end factor analysis for speaker verification, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[35] L. Lee and R. Rose, Speaker Normalization Using Efficient Frequency Warping Procedure, in Proc. of ICASSP, pp. I–353–356, Atlanta, USA, 1996,

[36] D. Kim, S. Umesh, M. Gales, T. Hain, and P. Woodland, Using VTLN for Broadcast News Transcription, In Proc. of the 8th ICSLP, Jeju Island, Korea, 2004.

[37] L. Burget, Combination of Speech Features Using Smoothed Heteroscedastic Linear Discriminant Analysis. In Proc. of the 8th ICSLP, Jeju Island, Korea, pp. 2549–2552, 2004

[38] K. Visweswariah, S. Axelrod, R.A. Gopinath, Acoustic modeling with mixtures of subspace constrained exponential models. In Proc. of the 7th Eurospeech'2003, Geneva, Switzerland, pp. 2613–2616, 2003.

[39] Xin Lei, Modeling Lexical Tones for Mandarin Large Vocabulary Continuous Speech Recognition, Ph.D. thesis, University of Washington, 2006.

[40] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, A pitch extraction algorithm tuned for automatic speech recognition, In Proc. of ICASSP, pp. 2494-2498, Italy, 2014.

[41] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, Connectionist probability estimators in HMM speech recognition, in IEEE Trans. Speech Audio Process., vol. 2, no. 1, pp. 161–174, 1994.

[42] A. J. Robinson, An application of recurrent nets to phone probability estimation," in IEEE Trans. on Neural Networks, vol. 5, no. 2, pp. 298–305, 1994.

[DOI: 10.18869/acadpub.jsdp.13.3.51]

[43] G. E. Dahl, D. Yu, L. Deng, and A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, in IEEE Trans. Audio Speech Lang. Process., vol. 20, no. 1, pp. 30–42, 2012.

[44] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, Application of pretrained deep neural networks to large vocabulary speech recognition, in Proc. INTERSPEECH, September 2012.

[45] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A. Mohamed, Making deep belief networks effective for large vocabulary continuous speech recognition, in Proc. IEEE ASRU, pp. 30–35, December 2011.

[46] G. E. Hinton, S. Osindero, and Y. Teh, A fast learning algorithm for deep belief nets, Neural Computation, vol. 18, pp. 1527–1554, 2006.

[47] M. Gales, The generation and use of regression class trees for MLLR adaptation, University of Cambridge, Department of Engineering, 1996.

[48] D. Povey, G. Zweig, and A. Acero, The Exponential Transform as a generic substitute for VTLN, in Proc. IEEE ASRU, December 2011.

[49] S.J. Young, P.C. Woodland, The use of state tying in continuous speech recognition, in European Conference on Speech Communication and Technology (EUROSPEECH'93), pp. 2203–2206, Germany, September 1993,

[50] M. Federico, N. Bertoldi, and M. Cettolo, IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models, in Proc. of Interspeech, Brisbane, Australia, 2008.

[51] A. Stolcke, SRILM-an extensible language modeling toolkit, in Proc. of INTERSPEECH, 2002.

[52] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, Boosted MMI for model and feature-space discriminative training, in Proc. of ICASSP, pp. 4057–4060, 2008.

باقر باباعلی مدرک کارشناسی خود را در سال ۱۳۷۹ در رشتهٔ مهندسی کامپیوتر (گرایش سخت‌افزار) از دانشگاه شیراز اخذ کرد. همچنین مدرک کارشناسی ارشد و دکترای خود را به‌ترتیب در سال‌های ۱۳۸۲ و ۱۳۸۹ در رشتهٔ مهندسی کامپیوتر (گرایش هوش مصنوعی) از دانشگاه صنعتی شریف دریافت کرد. زمینه‌های پژوهشی مورد علاقهٔ ایشان یادگیری ماشین و بازشناسی الگوی آماری، بازشناسی گفتار، بازشناسی الگوهای دنباله‌ای، یادگیری ژرف و کاربردهای آن بوده و در حال حاضر عضو هیأت علمی در دانشکدهٔ ریاضی، آمار و علوم کامپیوتر دانشگاه تهران است.
نشانی رایانامهٔ ایشان عبارت است از:

babaali@ut.ac.ir

[DOI: 10.18869/acadpub.jsdp.13.3.51]