به‌کارگیریِ اطلاعاتِ زبانی در یک سیستمِ بازشناسیِ گفتار پیوستهٔ فارسی

محمد بحرانی
دانشگاه صنعتی شریف(آزمایشگاه پردازش گفتار)
حسین صامتی
دانشگاه صنعتی شریف(آزمایشگاه پردازش گفتار)

چکیده
در این مقاله یک سامانهٔ بازشناسی گفتار پیوسته برای زبان فارسی معرفی می‌شود و نقش مدل آوایی و مدل زبانی در آن مورد بررسی قرار می‌گیرد. مدل‌های آوایی با روش‌های مستقل‌ازبافت و وابسته‌به‌بافت در این سامانه به‌کار رفته و نتایجِ به‌کارگیریِ آن‌ها ارائه شده است. مدل زبانیِ سه‌کلمه‌ای نیز با روش‌های مبتنی‌بر کلمه، مبتنی‌بر مقولهٔ نحوی و مبتنی‌بر طبقه، با استفاده از پیکرهٔ متنیِ زبانِ فارسی استخراج و در سامانهٔ بازشناسی به‌کار گرفته شده است. همچنین مدل زبانیِ دستوری مبتنی‌بر دستور ساخت-گروهیِ تعمیم‌یافته در این سامانه پیاده‌سازی شده و نیز در ترکیب با مدل زبانیِ آماری به‌کار رفته است. نتایج حاصل نشان می‌دهد که مدل آواییِ وابسته به بافت، مطابق انتظار، بهترین عملکرد را دارد. همچنین مدلِ زبانیِ سه‌کلمه‌ایِ مبتنی بر کلمه، نسبت به سایر روش‌های استخراجِ مدلِ زبانیِ آماری برتری دارد. درضمن ترکیبِ مدلِ زبانیِ دستوری با مدلِ زبانیِ آماری منجر به بهبودِ نتایجِ بازشناسی می‌شود. سامانهٔ بازشناسیِ گفتارِ معرفی‌شده در این مقاله، اولین سامانهٔ بازشناسی برای گفتارِ پیوستهٔ فارسی بوده و با پشتوانهٔ فعالیت‌های تحقیقاتیِ متعددی که برای پیاده‌سازیِ آن انجام شده است، قابلیت استفاده به‌صورت کاربردی را یافته است.

کلیدواژه‌ها: بازشناسیِ گفتارِ پیوسته، مدل‌سازیِ آوایی، مدل‌سازیِ زبانی، مدلِ زبانیِ آماری، مدلِ زبانیِ دستوری.

۱. مقدمه

مسئلهٔ بازشناسی گفتار و یا تبدیل گفتار به متن از مسائل مهم تحقیقاتی و کاربردیِ مرتبط با زبان‌شناسی است. یک موتورِ بازشناسیِ گفتار توسط پژوهشگرانی با تخصص‌های متفاوت، شامل مهندسی کامپیوتر، پردازش علائم دیجیتال، هوش مصنوعی، آواشناسی و زبان‌شناسی، طراحی و پیاده‌سازی می‌شود. تخصص‌های درگیر برای کار روی یک سامانهٔ بازشناسیِ گفتار نیاز به ارتباطِ نزدیک با یک‌دیگر برای مدل‌سازی و پیاده‌سازیِ اطلاعاتِ گفتاری و زبانی دارند. گفتارِ مورد بازشناسی می‌تواند به‌صورت یک فایلِ صوتی باشد و یا ازطریق میکروفون، ازطریق خط تلفن و یا فرمان از راه دور باشد. درواقع هدفِ نهاییِ بازشناسیِ گفتار، ساختِ ماشین‌هایی است که بتوانند مانند انسان بشنوند و عکس‌العمل مناسب نشان دهند.

اولین تلاش‌ها برای ساخت سیستم بازشناسیِ گفتار از اوایلِ دههٔ چهل میلادی آغاز شد وتا اکنون این تلاش‌ها ادامه دارند. اگرچه رویکردهای مختلفی برای بازشناسی گفتار وجود دارند اما موفق‌ترینِ آن‌ها رویکردِ تشخیصِ الگو[1] است، که تقریباً همهٔ سیستم‌های موفقِ امروزی براساسِ آن عمل می‌کنند. در این رویکرد، گفتار به‌کمک تعدادی واحدِ آوایی[2] (کلمه[3]، هجا[4]، سه واجی[5] یا واج[6]) مدل می‌شود و برای بازشناسی نیز از تشخیص این واحدها و کنار هم قرار دادنِ آن‌ها، متنِ متناسب با گفتار، تشخیص داده می‌شود. این مدل به‌نوعی متناسب با نظریه‌های زبان‌شناسیِ چامسکی است که براساس آن، هر انسانی در برخورد با گفتاری از زبانی آشنا، آن‌را به رشته‌ای از کوچک‌ترین واحدهای زیرکلمهٔ[7] ممکن تبدیل می‌کند و سپس در مغزِ خود به پردازش این رشته می‌پردازد تا هجاها، کلمات، جملات و نهایتاً گفتار را بازشناسی کند. اگرچه امروزه سیستم‌های موفقِ مختلفی براساس این رویکرد ارائه شده‌اند و در عمل نیز از آن‌ها استفاده می‌شود ولی همگیِ آن‌ها به‌گونه‌ای برخی از محدودیت‌های ساده‌کننده را یدک می‌کشند. حذفِ این محدودیت‌ها می‌تواند به‌طورِ قابل ملاحظه‌ای بر پیچیدگی آن‌ها بیفزاید. درواقع، یک سیستمِ بازشناسیِ کامل، که بتواند مشابه انسان عمل کند باید بتواند:

- قادر به بازشناسیِ گفتارِ پیوسته[8] و محاوره‌ای[9] باشد.
- گفتارِ افرادِ مختلف، حتی با لهجه‌های متفاوت را بازشناسی نماید.
- در محیط‌های شلوغ و نوفه‌ای هم جواب‌گو باشد.

1. pattern recognition
2. acoustic unit
3. word
4. syllable
5.triphone
6. phoneme
7. subword
8. continuous
9. spontaneous

- به‌صورتِ بی‌درنگ[1] عمل کند.
- قادر به فراگیریِ اطلاعاتِ جدید، نظیر کلمات و قواعدِ زبانی باشد.

سامانه‌های موجودِ بازشناسیِ گفتار معمولاً دارای محدودیت‌هایی هستند و مشخصاتِ فوق را، به‌طورکامل برآورده نمی‌کنند. این سامانه‌ها ممکن است گفتار محاوره‌ای را به‌خوبیِ گفتارِ رسمی و کتابی تشخیص ندهند. همچنین عملکرد این سامانه‌ها در محیط‌های واقعی به‌مراتب ضعیف‌تر از محیطِ آکوستیک است. علاوه‌براین، نوعاً محدود به واژگانِ تعریف‌شده و قوانینِ زبانیِ آموزش‌داده‌شده هستند و در مواجهه با واژگانِ جدید و قوانینِ زبانیِ جدید دچار خطا می‌شوند. میزانِ پیچیدگیِ یک سیستمِ بازشناسیِ گفتار به عواملِ متعددی بستگی دارد. در این سیستم‌ها، برای کاربردی کردن، سعی می‌شود با ایجادِ محدودیت‌هایی میزان پیچیدگی را کاهش دهند. این کار باعث تعریفِ دقیقِ سیستم برای یک کاربردِ مشخص و کاهشِ دامنهٔ کاربرد آن (با کاراییِ بهتر) می‌شود. مهم‌ترینِ این محدودیت‌ها که توانمندیِ یک سیستمِ بازشناسیِ گفتار را مشخص می‌کنند موارد زیر هستند.

میزان وابستگی یا استقلال از گوینده: سیستم‌هایی که تنها به یک و یا چند گویندهٔ خاص پاسخ می‌گویند، وابسته‌به‌گوینده[2]، و آن‌هایی که به تمام گویندگانِ یک زبان پاسخ می‌گویند مستقل‌ازگوینده[3] نامیده می‌شوند.

پیوسته یا گسسته‌بودنِ گفتار: سیستم‌های بازشناسیِ گفتار ممکن است محدودیت‌هایی بر نحوهٔ تلفظِ کلمات توسط گوینده اعمال کنند. میزان پیوستگی یا گسستگیِ کلمات به سه دستهٔ سیستم‌های بازشناسیِ کلماتِ مجزا[4]، سیستم‌های بازشناسیِ گفتارِ متصل[5] و سیستم‌های بازشناسیِ گفتارِ پیوسته[6] پیوسته[6] طبقه‌بندی می‌شوند. در حالتِ مجزا، گفتارِ گوینده به‌صورت کلمه‌به‌کلمه و کاملاً مجزا از یک‌دیگر است. در سیستم‌های کلماتِ متصل، گوینده دنباله‌ای از کلماتِ محدود را تولید می‌کند. سیستم‌های بازشناسیِ پیوسته، گوینده را چندان به پیروی از قوانین خاصی در بیان گفتار مجبور نمی‌کنند بلکه گوینده، جملات را به‌طور پیوسته و طبیعی بیان می‌کند، هرچند در این حالت نیز فرضِ کتابی‌بودنِ گفتار وجود دارد. نوعی دیگر، که پیوسته اما کاملاً غیرماشینی است، گفتار محاوره‌ای[7] است است که در آن گوینده می‌تواند گفتار را به‌صورتِ کاملاً طبیعی بیان کند. در چنین گفتاری، جملات ناقص، سرفه، تپق، مکث‌های طولانی و امثال آن وجود دارد.

1. real time
2. Speaker Dependent (SD)
3. Speaker Independent (SI)
4. Isolated Word Recognition (IWR)
5. connected word recognition
6. Continuous Speech Recognition (CSR)
7. spontaneous speech

حجم واژگان: حجم واژگان[1] و تعداد کلماتِ مورد استفاده در یک سیستم بازشناسی، از عوامل مؤثر در دقت و سرعتِ سیستم است. بعضی از سیستم‌های بازشناسی فقط برای تشخیصِ تعدادِ محدودی کلمه طراحی شده‌اند درحالی‌که بعضی دیگر از سیستم‌ها قادرند مجموعهٔ بزرگی از کلمات را تشخیص دهند.

محدودیت‌های زبانی: یکی از مهم‌ترین بخش‌های یک سیستمِ بازشناسیِ گفتار، مدل زبانی[2] است، که درواقع بیانگر محدودیت‌های زبانی است. مدل زبانیِ یک زبانِ طبیعی مرکب از چهار جزءِ نمادها، دستور[3]، معنی‌شناسی[4] و کاربردشناسی[5] است. نمادهای زبان، واحدهایی هستند که پیام‌ها را تشکیل تشکیل می‌دهند و درواقع کلمات یا واحدهایی کوچک‌تر از کلمه، نظیر هجاها یا واج‌ها هستند. دستور زبان مرکب از محدودیت‌های واژگانی[6] و نحوی[7] است که بیانگر نحوهٔ شکل گرفتنِ کلمات از واحدهایی واحدهایی کوچک‌تر از کلمه و نیز شکل گرفتنِ جملات از کلمات است. جنبهٔ معناییِ مرتبط با نحوهٔ ترکیبِ کلمات، برای شکل‌دادنِ پیغام‌های با معنا است. به‌عنوان‌مثال: جملهٔ **صندلی غذا می‌خورد** از لحاظ نحوی درست ولی از لحاظ معنایی نادرست است. جنبهٔ کاربردیِ یک زبان، در بالاترین سطح جای دارد و بیانگر وابستگیِ گفتار به گوینده‌ها و محیط است. محدودیت‌های معنایی و جنبهٔ کاربردی به‌ندرت در سیستم‌های بازشناسی گفتار استفاده می‌شوند، زیرا که این محدودیت‌ها را به‌دشواری می‌توان به‌شکل فرمول بیان کرد. ولی محدودیت‌های دستوری تقریباً در تمامیِ سیستم‌های بازشناسیِ گفتارِ پیوسته، به‌صورتِ محدودیت‌های واژگانی و نحوی، مورد استفاده قرار می‌گیرند و تعداد جملات مُجاز برای بازشناسی را کاهش می‌دهند و به‌عبارت‌دیگر، فضای مورد جستجو را کوچک‌تر می‌کنند. میزان محدودیتی که توسط مدل زبانی، درون یک سیستمِ بازشناسی ایجاد می‌شود، سرگشتگیِ[8] حاصل از آن مدل زبانی نامیده می‌شود و سعیِ محققین، کاهشِ هرچه‌بیشترِ سرگشتگی با استفاده از اطلاعات زبانی است.

کارایی در حضور نوفه و در محیط‌های کاربردیِ مختلف: بعضی از سیستم‌های بازشناسی تنها در شرایط محیطیِ مساعد و با نوفهٔ کم قادر به تشخیصِ گفتار با دقتِ مطلوب هستند. کارایی این سیستم‌ها با تغییر محیط به‌شدت کاهش می‌یابد. این‌که سیستمی بتواند با تغییر محیط و حضور نوفه،

1. vocabulary size
2. language model
3. grammar
4. semantics
5. pragmatics
6. lexical
7. syntactic
8. perplexity

دقتِ بازشناسی خود را حفظ کند، یکی از معیارهای مهم، به‌ویژه در کاربردی بودنِ آن، است. درنظرگرفتنِ این مهم بر پیچیدگی مسئله می‌افزاید.

ابهام آکوستیکی[1] و میزان اشتباه[2] بین کلمات: کلماتی که شکل نوشتاریِ متمایزی دارند ولی از لحاظ گفتاری مانند هم تلفظ می‌شوند، ابهام آکوستیکی ایجاد می‌کنند. کلماتی نظیر «سمر» و «ثمر». همچنین کلماتی که تلفظ آن‌ها به یک‌دیگر شباهت دارد، مانند «دو» و «نه» که ممکن است به‌جای یک‌دیگر بازشناسی شوند و دقت را پایین بیاورند. هرچه تعداد این‌گونه کلمات در واژگان بیشتر شود، دقت سیستم بازشناسی پایین‌تر می‌آید. لذا برای جبران این مسئله، سیستم بازشناسی باید از مدل زبانی در سطوح دستوری و معنایی کمک بگیرد.

تحقیقاتِ بی‌شماری برای طراحی و پیاده‌سازیِ موتورها و سامانه‌های بازشناسیِ گفتار انجام شده است. چون مسائلِ زبانی و زبان‌شناسی از بخش‌های اساسیِ بازشناسیِ گفتار هستند، زبانِ یک سامانهٔ بازشناسیِ گفتار، آن‌را به‌صورتِ کلی، از سایر سامانه‌ها مجزا می‌کند. برای بازشناسیِ گفتار در زبان فارسی نیز فعالیت‌های زیادی انجام شده و مقالات و پایان‌نامه‌های زیادی منتشر شده است. از مهم‌ترینِ آن‌ها می‌توان به (احدی، ۱۹۹۹)، (غلامپور، ۱۳۷۹)، (الماس‌گنج و دیگران، ۱۳۸۰)، (الماس‌گنج و دیگران، ۱۳۸۳)، (همایون‌پور، ۱۳۸۳)، (ولی، ۱۳۸۵)، (صامتی و دیگران، ۱۳۸۳)، (باباعلی و صامتی، ۲۰۰۴)، (صامتی و دیگران، ۲۰۰۸)، و (صامتی و دیگران، ۲۰۰۹) اشاره کرد. در این مقاله موتورِ بازشناسیِ گفتار به‌صورتِ اجمالی معرفی می‌شود و بعضی پژوهش‌های انجام یافته برای به‌کارگیریِ مدل زبانیِ زبان فارسی شرح داده می‌شود. در بخش ۲، ساختارِ کلیِ سیستمِ بازشناسیِ گفتارِ موردبحث در این مقاله، ارائه و واحدهای آن به‌طور مختصر تشریح می‌شود. بخش ۳ به تشریحِ نحوهٔ مدل‌سازیِ آوایی، اعم از مستقل‌ازبافت[3] (CI) و یا وابسته‌به‌بافت[4] (CD) اختصاص دارد. در بخش ۴ نحوهٔ ساخت مدل زبانی برای موتور بازشناسی در زبان فارسی، و مسائلِ مربوط به آن، مورد بحث قرار می‌گیرد. بخش ۵ به توضیح در مورد آزمایش‌ها و ارائهٔ نتایجِ حاصل از به‌کارگیریِ مدل‌های مختلفِ زبانی اختصاص دارد. نهایتاً در بخش ۶، خلاصه و نتیجه‌گیری ارائه می‌شود.

۲. معرفی کلی سیستمِ بازشناسیِ گفتار

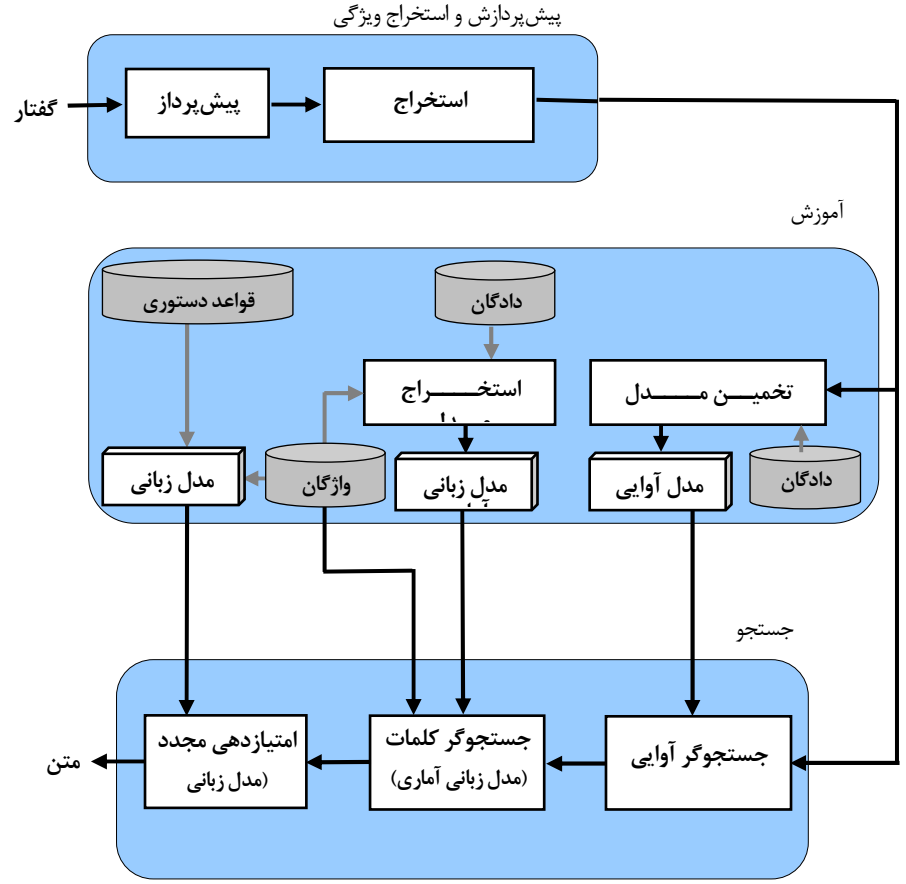
موتورِ بازشناسیِ گفتارِ پیوستهٔ فارسی، مورد بحث در این مقاله، حاصلِ به کارگیریِ آخرین روش‌های شناخته‌شده برای پیاده‌سازیِ واحدهای مختلفِ یک سامانهٔ بازشناسیِ گفتار و انطباقِ آن با اطلاعاتِ واج‌شناسی و خواص دستوری و واژگانیِ زبانِ فارسی است. شکل ۱، نمودارِ جعبه‌ایِ این موتور

1. acoustic ambiguity
2. confusability
3. context independent
4. context dependent

را نشان می‌دهد. موتور به‌صورتی طراحی شده است که واحدهای مختلفِ آن به‌راحتی قابل تعویض و یا اصلاح باشند. بنابراین با ادامهٔ کارِ گروه تحقیقاتی و به‌دست‌آمدنِ نتایجِ جدید، در مورد هریک از واحدهای سامانهٔ بازشناسیِ گفتار، به‌راحتی می‌توان اصلاحاتِ موردنظر را در این سامانه وارد کرد. این سامانه شامل واحدهای مقاوم‌سازی نسبت‌به نوفه و شرایط محیطی است که وظیفهٔ پیش‌پردازش، جداسازیِ سکوت از گفتار و حذفِ اولیهٔ نوفه و استخراج مشخصه‌های اولیه برای بازشناسی را برعهده دارد. سپس این مشخصه‌ها با روش‌های متعدد و مؤثر، مقاوم‌سازی می‌شوند.

این سامانه هر دونوع مدلِ وابسته‌به‌بافت و مستقل‌ازبافت را به‌کار می‌گیرد. این مدل‌ها از نوع مدلِ مخفیِ مارکوف با مشاهداتِ پیوسته[1] (CDHMM) هستند و شامل حالت‌هایی با تلفیقِ توزیع‌های گوسی در حوزهٔ کپسترال هستند. در این مدل‌ها، پَرِش به جلو، پَرِش از روی حالت‌ها، و پَرِش به خودِ حالت، مجاز است و ماتریس‌های کوواریانس[2] از نوع قطری هستند (باباعلی و صامتی، ۲۰۰۴)، (صامتی و دیگران، ۲۰۰۸)، (صامتی و دیگران، ۲۰۰۹). پارامترهای مربوط به احتمالِ مشاهدات، با روشِ شباهتِ بیشینه[3] (ML) آموزش داده شده، و مدلِ اولیه با قطعه‌بندیِ یک‌نواخت به‌دست می‌آید. هر دور از فرایندِ آموزش، شامل قطعه‌بندی با برنامه‌ریزیِ پویا (الگوریتم ویتربی)[4] و تخمینِ مجددِ پارامترهای مدل است که منجر به روشِ کا- مینزِ قطعه‌ای[5] می‌شود (رابینر[6] و یوانگ[7]، ۱۹۹۳). در واحدِ جستجوگرِ آوایی[8]، یک جستجوی ویتربی با هرس از نوع شعاعی و هیستوگرام اجرا می‌شود. این بخش بخش واحدهای صوتیِ شناسایی‌شده را، به‌عنوانِ فرضیاتِ اولیه، در اختیار واحد استخراجِ کلمات قرار می‌دهد. واحد استخراج، وظیفهٔ جستجو در درختِ واژگان را برعهده دارد. مدل‌های زبانیِ آماری و دستوری می‌توانند در هر دو واحدِ استخراجِ کلمات و امتیازدهیِ مجدد به‌کار روند. در این سامانه، از مدل زبانیِ آماری در واحد استخراجِ کلمات استفاده می‌شود و مدل‌های زبانیِ دستوری در بخشِ امتیازدهیِ مجدد به‌کار گرفته می‌شوند.

1. Continuous Density Hidden Markov Models
2. covariance
3. Maximum Likelihood
4. Viterbi
5. segmental K-means
6. L. Rabiner
7. B. H. Juang
8. acoustic decoder



شکل ۱. ساختارسامانهٔ بازشناسی گفتار

۳. مدل‌سازیِ آوایی

برای‌این‌که گفتار به‌طور دقیق و مناسب مدل‌سازی شود، لازم است گفتار به بخش‌های کوچک‌تری به نام «واحد آوایی» تقسیم و هربخش جداگانه مدل‌سازی شود. بسته به نوع و کاربردِ سیستمِ بازشناسی (پیوسته یا منفصل بودنِ بازشناسی،حجم واژگان و زبانِ موردِ استفاده)، واحدهای آواییِ

مختلفی ممکن است به‌کار روند. واحدهای آواییِ رایج عبارت‌اند از: واج[1]، هجا[2]، دیاد (نیم‌هجا)[3]، کلمه و و واحدهای آوایی وابسته به بافت (مانند دوواج[4]، سه‌واج[5] و واج‌گونه[6]).

متداول‌ترین ساختارهایی که برای مدل‌سازیِ گفتار در بازشناسی مورد استفاده قرار می‌گیرند، عبارت‌اند از: مدل مخفیِ مارکوف (HMM)، مدل انطباق زمانی پویا[7] (DTW)، شبکهٔ عصبی مصنوعی[8] مصنوعی[8] (ANN) و مدل‌های ترکیبی[9] (ترکیب HMM با ANN). از بین مدل‌های مذکور، مدل مخفی مارکوف به‌دلیلِ توانایی بالای آن در مدل‌سازیِ تغییراتِ زمانیِ سیگنالِ گفتار، بیش از سایر مدل‌ها در بازشناسیِ گفتارِ پیوسته با مجموعهٔ واژگان بزرگ به‌کار می‌رود (رابینر و یوانگ، ۱۹۹۳). در این مقاله نیز از مدل مخفی مارکوف برای مدل‌سازیِ آوایی استفاده شده است.

برای آموزش مدل‌های آوایی نیاز به دادگانِ گفتاری داریم. زیرا برای مدل‌سازی لازم است از هر واحد آواییِ زبان، به تعدادِ کافی نمونهٔ ضبط شده دردست باشد. درعمل تعداد این واحدها بسیار زیاد است و معمولاً به مهم‌ترینِ آن‌ها اکتفا می‌شود. با توجه به کاربردهای مختلفِ بازشناسیِ گفتار، دادگان‌های گفتاریِ مختلفی طراحی شده‌اند. ازآن‌جاکه در بسیاری از سیستم‌های بازشناسیِ امروزی، واج‌ها به‌عنوانِ کوچک‌ترین واحدهای بازشناسی انتخاب می‌شوند، در طراحیِ دادگان‌های گفتاری به واج‌ها توجه خاصی می‌شود. در این مقاله از دو دادگانِ گفتاریِ فارس‌داتِ کوچک (بی‌جن‌خان و دیگران، ۱۹۹۴)، و فارس‌داتِ بزرگ (شیخ‌زادگان و بی‌جن‌خان، ۱۳۸۵) برای مدل‌سازی آوایی استفاده کرده‌ایم. برای مدل‌سازیِ آوایی، دو رویکردِ عمده را به‌کار برده‌ایم: مدل‌سازیِ مستقل‌ازبافت و مدل‌سازیِ وابسته‌به‌بافت. در مدل‌سازیِ مستقل‌ازبافت، هر واحدِ آوایی به‌طور مستقل و بدونِ درنظرگرفتنِ واحدهای آواییِ مجاور، مدل‌سازی می‌شود. در این مقاله از واج به‌عنوانِ واحدِ آواییِ پایه استفاده شده است و به‌ازای هرکدام از ۲۹ واجِ زبانِ فارسی، یک مدلِ مخفیِ مارکوف آموزش داده شده است. یک مدلِ مخفیِ مارکوف نیز برای مدل‌سازیِ سکوت و نوفه درنظر گرفته شده است. بنابراین در مدل‌سازیِ مستقل‌ازبافت جمعاً ۳۰ مدل داریم.

در مدل‌سازیِ وابسته‌به‌بافت، هر واحدِ آوایی (مثلاً هر واج)، با توجه به واحدهای آواییِ مجاورِ آن، مدل‌سازی می‌شود. واحدهای آوایی در هنگام مجاورت با هم، اثر متقابل بر یک‌دیگر می‌گذارند و تغییراتی در طرز تلفظ آن‌ها ایجاد می‌شود، که این تغییرات بسته‌به نوعِ واحدهای مجاور متفاوت است.

1. phoneme
2. syllable
3. dyad
4. diphone
5. triphone
6. allophone
7. Dynamic Time Warping
8. Artificial Neural Network
9. hybrid

مجاورت واج‌ها در گفتار پیوسته، باعث ایجاد «هم‌تولیدی»[1] و «همگونی»[2] می‌شود. «هم‌تولیدی» به‌این‌معناست که مجاورت واج‌ها با یک‌دیگر، در ناحیهٔ مرزی بین واج‌ها تأثیر می‌گذارد. به‌عبارت‌دیگر، حالتِ شروع و پایانِ یک واج کاملاً وابسته به واج‌های اطرافِ آن در گفتار است. پدیدهٔ «همگونی» نیز در اثر وضعیت واحدهای تکلّمیِ دخیل درتلفظ واج‌ها پیش می‌آید، زیرا وضعیتِ این واحدها هنگام گذار از یک واج به واج بعدی، به‌یک‌باره تغییر نمی‌کند و این باعث می‌شود که نحوهٔ تلفظِ یک واج، به واج یا واج‌های قبلی و یا واج‌های بعدی وابسته شود (به تلفظ‌های گوناگونِ یک واج در گفتار، «واج‌گونه» می‌گویند). به‌عنوان مثال واج [η]، که یک واج بی‌واک است، در کلمهٔ ناهار [ν/η/ρ] در مجاورتِ واکه‌ها به‌صورتِ واک‌دار تلفظ می‌شود، یا در کلمهٔ موم [μυμ]، واج [υ] درمجاورتِ واجِ خیشومی [μ] به‌حالتِ خیشومی درمی‌آید. همچنین در کلمهٔ حدس[ηαδσ]، واجِ [δ] در مجاورتِ واجِ بی‌واک [σ] «واک‌رفته» شده و شبیه [τ] تلفظ می‌شود. اثر واج‌های مجاور گاهی نیز باعث می‌شود که یک واج کاملاً به واج دیگری تبدیل گردد، مانند واج [ν] در کلمهٔ شنبه [.ανβε] که در تلفظ به واج [μ] تبدیل می‌شود.

بنابر آنچه گفته شد، خصوصیاتِ طیفیِ یک واج در گفتار، کاملاً به بافتِ اطرافِ آن وابسته است. برای مدل‌سازیِ دقیق و مناسبِ گفتار، به‌منظورِ بازشناسی، باید دگرگونی‌های حاصل از بافت را نیز درنظر گرفت و درواقع بافتِ آوایی را به‌نحوی در مدل‌سازی دخالت داد. تجربه نشان داده است که اطلاعاتِ وابسته‌به‌بافت، نقش مهمی در بازشناسی دارند و با به‌کارگیریِ آن‌ها میزانِ خطای بازشناسی به‌مقدار قابل‌توجهی کاهش می‌یابد. برای این منظور ساختارهای وابسته به بافتِ متنوعی برای مدل‌سازیِ اثراتِ بافتِ آوایی به‌کار رفته‌اند. در این مقاله از دو روش مدل‌سازیِ وابسته‌به‌بافت استفاده شده است: مدل‌سازیِ مبتنی بر سه‌واج و مدل‌سازیِ مبتنی بر دسته‌بندیِ واج‌ها.

۳. ۱. مدل‌سازیِ مبتنی بر سه‌واج

واحد آواییِ سه‌واج یکی از زیرکلمه‌های وابسته‌به‌بافتِ بسیار پرکاربرد و مؤثر در بازشناسیِ گفتار است. در سه‌واج‌ها، هر واج با درنظرگرفتنِ واجِ قبلی و واجِ بعدی‌اش در گفتار به‌صورتِ یک واحدِ آواییِ مستقل درنظر گرفته شده و مدل می‌شود. برای یک واج مانند P، مجموعهٔ سه‌واج‌های مرتبط‌به‌آن را به‌صورتِ P_L-P-P_R نمایش می‌دهیم.

مزیت استفاده از مدل‌سازیِ مبتنی بر سه‌واج در این است که برای هر واج، اثرات هر دو واجِ مجاور در مدل‌سازی دخالت داده می‌شود. درعوض، مشکلِ کمبودِ دادهٔ آموزشی در این‌جا بیشتر مشهود است، زیرا تعداد سه‌واج‌ها، به‌لحاظِ نظری، برابر با تعداد واج‌ها به‌توانِ سه می‌شود، که تعداد بسیار زیادی

1. coarticulation
2. assimilation

است (برای مثال در زبان فارسی که حدود ۳۰ واج دارد، تعداد سه‌واج‌ها ۳۰[۳] می‌شود). اگرچه در عمل تعداد زیادی از سه‌واج‌ها در زبان گفتاری به‌کار نمی‌روند ولی بازهم تعداد باقی‌مانده بسیار زیادند. زیادبودنِ تعدادِ سه‌واج‌ها باعث می‌شود که در یک دادگانِ گفتاریِ محدود، بسیاری از آن‌ها خیلی به‌ندرت تلفظ شوند و به‌همین‌دلیل، نمی‌توان مدلِ دقیقی برای آن‌ها آموزش داد. همچنین ممکن است بعضی از سه‌واج‌های موجود در زبان، در دادگانِ آموزشی اصلاً وجود نداشته باشند. بنابراین نمی‌توان مدلی برای این سه‌واج‌ها تشکیل داد (به این سه‌واج‌ها، «سه‌واج مشاهده نشده[۱]» می‌گویند). در این حالت اگر در دادگان آزمون با این سه‌واج‌ها مواجه شویم، برای بازشناسیِ آن‌ها هیچ مدلی در اختیار نداریم، درنتیجه، سیستم بازشناسی در این حالت نمی‌تواند کاملاً مستقل از مجموعهٔ واژگان باشد.

به‌منظور فائق آمدن بر مشکلِ کمبودِ دادهٔ آموزشی برای مدل‌های وابسته به متن، معمولاً سه‌واج‌های مشابه (یا حالت‌های مربوط به مدل آن‌ها) به‌هم گره[۲] زده می‌شوند. به‌این‌معنا که حالت‌های مشابه در سه‌واج‌های مختلف در یک دسته قرار می‌گیرند و با هم یکی درنظر گرفته می‌شوند. دو روشِ مشهور برای گره‌زدنِ حالت‌ها در مدل‌های مبتنی بر سه‌واج، دسته‌بندیِ مشتق‌شده از داده[۳] (یانگ[۴] و وودلند[۵]، ۱۹۹۳) و دسته بندیِ مبتنی بر درختِ تصمیم‌گیری[۶] (یانگ و دیگران، ۱۹۹۴) است. در این مقاله، از روش دوم، که رایج‌تر است، استفاده کرده‌ایم. در هردوی این روش‌ها رویهٔ کلیِ آموزش به‌این‌صورت است که همهٔ سه‌واج‌های موجود در دادگانِ آموزش، ابتدا با استفاده از همان تعداد دادهٔ آموزشیِ موجود، آموزش داده می‌شوند، و سپس حالت‌های مربوط به سه‌واج‌های با بافت مشابه، دسته‌بندی می‌شوند و حالت‌های شبیه‌تر در یک دسته قرار می‌گیرند. درنهایت، حالت‌های موجود در یک دسته به‌عنوانِ یک حالتِ واحد درنظر گرفته، و به اصطلاح به هم گره زده می‌شوند. در مرحلهٔ بعد، حالت‌های گره‌زده‌شده، مجدداً تحت آموزش قرار می‌گیرند.

در روشِ مبتنی بر درختِ تصمیم‌گیری، که در این مقاله استفاده شده است، رویهٔ دسته‌بندیِ حالت‌ها با استفاده از درخت تصمیم‌گیری صورت می‌گیرد. در این روش، تمامِ حالت‌های مربوط به بافت‌های مشابه (واجِ میانیِ یک‌سان در سه‌واجی‌ها) و همچنین مکان مشابه در مدل، در ریشهٔ درخت قرار می‌گیرند. سپس تعدادی پرسش در مورد بافتِ چپ و راستِ حالت‌ها مطرح می‌شود. با توجه به پرسش مطرح شده، حالت‌ها به دو دسته تقسیم می‌شوند، و دو گرهٔ جدید در درخت ایجاد می‌کنند. از بین پرسش‌های مطرح‌شده، پرسشی در نهایت انتخاب می‌شود که دو دستهٔ حاصل از آن بیشترین

1. unseen triphone
2. tie
3. data driven clustering
4. S. J. Young
5. P. C. Woodland
6. tree-based state tying decision

افزایش در میزانِ شباهت[1] را نسبت به دستهٔ اولیه، داشته باشند. این روند ادامه پیدا می‌کند و درخت به‌تدریج رشد می‌کند تا درنهایت به‌جایی برسد که افزایش در میزانِ شباهت از حد آستانهٔ خاصی کمتر شود و یا دادهٔ آموزش، برای حالت‌های موجود در گره‌های درخت، به‌حدکافی نباشد. پس از تشکیلِ درخت، هربرگ از درخت نشان‌دهندهٔ یک طبقه از حالت‌هاست. حالت‌های موجود در هر کلاس به‌هم گره زده می‌شوند و یک حالت را تشکیل می‌دهند.

در این مقاله برای آموزش مدل‌های مبتنی بر سه‌واج از دادگان‌های فارس‌داتِ کوچک و فارس‌داتِ بزرگ استفاده کرده‌ایم. فرهنگِ لغتِ به‌کاربرده‌شده برای دادگانِ فارس‌داتِ کوچک شامل ۱۱۴۷ کلمه و برای فارس‌داتِ بزرگ شامل ۴۷۸۰۲ کلمه است. با استفاده از فرهنگ لغات می‌توان تمام انواع سه‌واج‌های دیده‌شده در دادگان آموزش را مشخص کرد. با توجه به این فرهنگ لغات، در مجموع، ۴۳۶۲ نوع سه‌واجِ درون‌کلمه‌ای[2] در دادگان فارس‌داتِ کوچک و ۱۶۳۶۴ نوع سه‌واجِ درون‌کلمه‌ای در دادگان فارس‌داتِ بزرگ وجود دارد. در رَوَندِ آموزش، تعداد حالت‌ها برای هر مدل، ۵ مورد و تعداد گوسی‌ها در هر حالت، ۱ مورد درنظرگرفته شد. درنتیجه تعداد کل حالت‌ها در مجموعهٔ دادگان فارس‌داتِ کوچک، ۲۱۸۱۰=۵×۴۳۶۲ و در مجموعهٔ دادگان فارس‌داتِ بزرگ، ۸۱۸۲۰=۵×۱۶۳۶۴ است.

در مجموع، ۱۳۰ پرسش خودکار با استفاده از روش مذکور، در مقالهٔ سینگ[3] (سینگ و دیگران، ۱۹۹۹)، برای ساخته شدنِ درخت تصمیم‌گیری به‌کار برده شده است، که ۶۵ عدد از این پرسش‌ها درخصوصِ بافتِ چپ و ۶۵ عددِ بقیه درخصوصِ بافتِ راست مطرح می‌شوند. با توجه به تعداد واج‌ها (۳۰ واج)، و تعداد حالت‌ها در هرمدل (۵ حالت)، تعداد ۱۵۰=۵*۳۰ درختِ تصمیم‌گیری تشکیل می‌شود. بعد از انجام رویهٔ دسته‌بندی و پیداشدنِ حالت‌های مشابه، داده‌های آموزش مربوط به همهٔ این حالت‌ها برای آموزشِ یک حالتِ عمومی (گره‌زده‌شده) به‌کار برده می‌شود. به این حالتِ عمومی، زنون[4] گفته می‌شود. تعداد زنون‌ها را می‌توان با تغییر مقادیر آستانه در الگوریتمِ دسته‌بندی، کنترل کرد. در مرحلهٔ آموزش زنون‌ها می‌توان تعداد گوسی‌ها را در هر حالت افزایش داد. در این مقاله دسته‌بندی و آموزش سه‌واج‌ها را با تعداد زنون‌های مختلف و تعداد گوسی‌های مختلف در هر زنون انجام دادیم. آزمایش‌ها نشان می‌دهند که تعداد ۵۰۰ زنون در فارس‌داتِ کوچک و ۴۰۰۰ زنون در فارس‌داتِ بزرگ و همچنین ۸ گوسی در هر زنون، بهترین دقت بازشناسی را نتیجه می‌دهند. نتایج بازشناسی در بخش ۵ آمده است.

1. likelihood
2. within-word triphone
3. R. Singh
4. senone

۳. ۲. مدل‌سازیِ وابسته‌به‌بافت با استفاده از دسته‌بندیِ واج‌ها

در این روش، از ایدهٔ واحدهای آواییِ چندگانه[1] (مدل‌سازیِ واج‌گونه‌ها[2])، برای مدل‌سازیِ وابسته‌به‌بافت استفاده کرده‌ایم. مزیتِ استفاده از این روش این‌است که می‌توان با توجه به حجمِ دادهٔ آموزشِ در دسترس، تعداد واحدهای آواییِ وابسته‌به‌بافت را انتخاب کرد.

اندیشهٔ کلّیِ واحدهای آواییِ چندگانه این‌است که یک واج در کلمه‌ها و متن‌های مختلف، به حالت‌های گوناگونی تلفظ می‌شود. بنابراین می‌توان گفت که خصوصیاتِ طیفیِ یک واج (واحدِ آواییِ مستقل‌ازبافت) در بافت‌های مختلف دچار تغییرات و دگرگونی‌هایی می‌شود. این تغییرات را می‌توان در چند دستهٔ کلی جای داد. به‌عبارت‌دیگر، هر واحد آوایی مانند P، بسته‌به‌این‌که در چه بافتی قرار دارد، چند حالت گوناگون خواهد داشت (مانند P_1، P_2، P_3). پس می‌توان به‌جای یک واحد آواییِ مستقل‌ازبافت، انواعِ گوناگونِ وابستگی‌به‌بافتِ آن‌را به‌عنوان واحدهای بازشناسی به‌کار برد.

روش‌های مختلفی برای پیدا کردنِ خودکارِ حالاتِ گوناگونِ یک واج وجود دارد. یک روش مناسب برای دسته‌بندیِ[3] حالاتِ گوناگونِ یک واج، استفاده از تعدادی قواعد آوایی است. با استفاده از این قواعد می‌توان نوع تأثیرپذیریِ یک واج را از واج‌های اطرافش تعیین کرده و براساسِ آن، واج را در یک دستهٔ خاص جای داد (غلامپور، ۱۳۷۹). برای به‌کارگیریِ این روش نیاز به قواعد آواشناسی داریم. در این مقاله به‌دلیل عدم دسترسی به قواعدِ آواییِ زبانِ فارسی، دسته‌بندیِ حالاتِ گوناگونِ واج‌ها را به‌صورتِ بی‌نظارت و با استفاده از الگوریتم دسته‌بندیِ کا- مینز[4] انجام داده‌ایم (بحرانی و صامتی، ۱۳۸۴). دلیل انتخاب این الگوریتم سادگی و سهولتِ پیاده‌سازی آن است.

برای دسته‌بندی واج‌ها باید هر نمونه واج از گفتار را با یک دنبالهٔ ویژگی[5] نمایش داد. برای این منظور لازم است دادگانِ گفتاری، تقطیعِ واجی شده باشد، به‌این‌معنا که واج‌های تلفظ‌شده در گفتار و مرز آن‌ها در سیگنالِ گفتار دقیقاً مشخص باشد.

ابتدا سیگنال گفتار، پیش‌پردازش می‌شود. به‌این‌صورت که سیگنالِ گفتار به‌تعدادی قاب[6] با طول مساوی (و با مقداری همپوشانی بین قاب‌ها) تقسیم شده و از هر قاب، پس از طی مراحلی، تعدادی ضریب بازنمایی (در این سیستم، ضرایب مل-کپستروم) به‌عنوان بردار ویژگی استخراج می‌گردد. چون دادگانِ گفتاری، تقطیع شده هستند، قاب‌های ابتدایی و انتهاییِ هر واج را در سیگنال گفتار می‌توان مشخص کرد. بنابراین هر واج از تعدادی قاب تشکیل شده که هر قاب نیز با یک بردارِ ویژگیِ n بُعدی مشخص می‌گردد. پس می‌توان برای هر واج یک دنبالهٔ ویژگی به‌صورت دنباله‌ای از l بردار ویژگی قاب

1. multiple phone units
2. allophones
3. clustering
4. K-means
5. feature sequence
6. frame

تعریف کرد که l مشخص‌کنندهٔ تعداد قاب‌های تشکیل دهندهٔ واج است. در این مقاله، تعدادِ قابِ تشکیل‌دهندهٔ دنبالهٔ ویژگیِ یک واج را به‌عنوان «طول دنبالهٔ ویژگی» می‌شناسیم. تعداد قابِ تشکیل‌دهندهٔ هر واج بسته‌به میزانِ کشیدگیِ زمانیِ آن واج در گفتار متفاوت است، بنابراین دنبالهٔ ویژگی واج‌ها دارای طول‌های متفاوتی خواهند بود.

پس از پیش‌پردازش می‌توان به‌ازای هریک از واج‌های زبان، دنبالهٔ ویژگی تمام نمونه‌های آن را در دادگان گفتاری جمع‌آوری کرد. اگر زبان دارای N واج باشد، N مجموعه دنبالهٔ ویژگی داریم که هرمجموعه مربوط به یکی از واج‌های زبان خواهد بود. حال می‌توان هرکدام از این مجموعه‌ها را به چنددسته تقسیم کرد به‌طوری‌که هر دسته بیان‌گر یکی از گوناگونی‌های آن واج باشد. به‌دلیل غیرهم‌طول‌بودنِ دنباله‌های ویژگی واج‌ها، برای دسته‌بندی آن‌ها با استفاده از الگوریتم *کا- مینز* با دو مشکل مواجه هستیم. مشکل اول این‌که نمی‌توان برای محاسبهٔ مرکز دسته‌ها از میانگین‌گیریِ معمولیِ دنباله‌ها استفاده کرد و مشکل دوم این‌که برای محاسبهٔ فاصلهٔ بین دنباله‌ها، توابع فاصلهٔ معمولی (مانند فاصلهٔ اقلیدسی) را نمی‌توان به‌کار برد. برای حل این مشکلات از روش انطباق زمانی پویا[1] (DTW) استفاده کرده‌ایم. پس از تعیین تعداد دستهٔ مناسب برای هر واج، که معمولاً بین ۱ تا ۵ دسته است، الگوریتم *کا- مینز* را برای دسته‌بندیِ دنباله‌های ویژگی مربوط به نمونه‌های مختلف آن واج به‌کار می‌بریم. به‌جای تابعِ فاصله از روشِ انطباقِ زمانیِ پویا استفاده می‌کنیم و برای محاسبهٔ مرکز دسته‌ها، روشِ «انطباق و میانگین‌گیری» (بحرانی و صامتی، ۱۳۸۴) را به‌کار می‌بریم.

پس از اعمال الگوریتم دسته‌بندی، واجِ موردنظر به K دسته تقسیم می‌شود که هر دسته را می‌توان به‌عنوان یکی از حالاتِ گوناگونِ آن واج درنظر گرفت. درواقع هردسته بیان‌گر یک واحدِ آواییِ وابسته‌به‌بافت است و دنباله‌های ویژگی‌ای که در یک دسته قرار دارند داده‌های آموزش را برای مدل‌سازیِ آن واحد آوایی فراهم می‌سازند. به‌عنوانِ‌مثال اگر واج [a] به سه دسته تقسیم شود، سه واحدِ آوایی [a1]، [a2] و [a3] تولید می‌شود که هریک بیان‌گر یکی از گوناگونی‌های واج [a] در متن‌های مختلف خواهد بود. حال به‌جای‌آن‌که یک مدل برای واج [a] ساخته شود، برای هرکدام از سه نوعِ گوناگونِ واجِ [a]، مدلِ جداگانه‌ای ساخته می‌شود. مدل‌سازی می‌تواند به‌طورمستقیم با استفاده از دنباله‌های ویژگیِ موجود در هر دسته صورت گیرد. تعداد کل مدل‌هایی که باید آموزش داده شوند برابر با تعداد کل واحدهای آواییِ وابسته‌به‌بافت و به‌عبارت‌دیگر برابر با تعدادِ کلِ دسته‌های مربوط به همهٔ واج‌ها است. آزمایش‌ها نشان می‌دهند که استفاده از مدلِ وابسته‌به‌بافت در سامانهٔ بازشناسیِ گفتارِ پیوستهٔ فارسی، نرخ خطای کلمات را به‌میزانِ ۲٪ کاهش می‌دهد.

1. Dynamic Time Warping

۴. مدل‌سازیِ زبانی

یکی از مؤثرترین راه‌های افزایشِ دقتِ سیستمِ بازشناسیِ گفتارِ پیوسته، به‌کارگیریِ اطلاعاتِ زبانی است (به‌صورتِ آماری، نحوی و معنایی). انسان‌ها به‌هنگامِ شنیدنِ گفتار، علاوه‌بر اطلاعاتِ آوایی، از اطلاعاتِ زبانی هم استفاده می‌کنند. آن‌ها از احتمالاتِ رخدادِ یک کلمه، که به‌نوعی در ذهنِ آن‌ها نقش بسته است، استفاده می‌کنند. اگر بخواهیم اطلاعاتِ زبانی را، که انسان‌ها به‌طور ناخودآگاه به‌کار می‌برند، سطح‌بندی کنیم، در اولین سطح، اطلاعاتِ واژگانی و در سطوحِ بعد اطلاعاتِ نحوی و معنایی قرار دارند (آلن[1]، ۱۹۹۵). در کاربردهای بازشناسیِ گفتار، سعی می‌شود که این‌سطوح از اطلاعات، مدل‌سازی شوند. هدف از مدل‌سازیِ زبان این است که سیستمِ بازشناسی، به‌سمتِ جمله‌هایی متناسب با کاربردِ موردنظر سوق داده شود و از جمله‌های بی‌معنا اجتناب گردد. به‌این‌منظور مدل‌های زبانیِ متنوعی پیشنهاد شده‌اند که هریک سعی می‌کنند تمام یا بخشی از سطوح اطلاعاتیِ مذکور را پوشش دهند.

برای استخراج مدل‌های زبانی در هر زبانِ خاص، نیاز به حجمِ عظیمی از داده‌های متنیِ آن زبان داریم. هرچه حجم داده‌های مورداستفاده بیشتر باشد، تخمینِ بهتری از این مدل‌ها به‌دست می‌آید. داده‌های متنیِ مورداستفاده برای استخراج مدل‌های زبانی، به‌صورتِ دادگان‌هایی به‌نامِ پیکرهٔ متنی[2] در دسترس هستند. پیکره‌های متنی برای هرزبان، شامل متن‌های مختلفِ آن زبان در زمینه‌های مختلف و در حجمِ بالاست (درحدود چند میلیون کلمه)، که معمولاً با برچسب‌های مختلفِ زبانی و گرامری مشخص شده‌اند. در این مقاله از نسخهٔ اولیهٔ "پیکرهٔ متنیِ زبانِ فارسی" (بی‌جن‌خان و دیگران، ۲۰۰۹)، برای استخراج مدل‌های زبانیِ آماری استفاده کرده‌ایم.

۴.۱. مدلِ زبانیِ چندکلمه‌ای

مدل زبانیِ چندکلمه‌ای[3] از متداول‌ترین مدل‌های زبانیِ مورد استفاده در سیستم‌های بازشناسی گفتار پیوسته است. مدل‌های چندکلمه‌ای احتمالِ شرطیِ رخدادِ یک کلمه در زبان، پس از هر رشته n-1 کلمه‌ای از آن زبان را مشخص می‌سازند. در این روش احتمال دنبالهٔ کلمات $W = w_1 w_2 \cdots w_T$ از رابطهٔ زیر محاسبه می‌شود:

$$p(W) = p(w_1 w_2 \cdots w_T) \cong \prod_{i=1}^{T} p(w_i \mid w_{i-n+1} \cdots w_{i-1}) \qquad (۱)$$

1. J. Allen
2. text corpus
3. n-gram

چون همیشه میزان داده‌های متنیِ در دسترس برای آموزش مدل‌های چندکلمه ای محدود است، معمولاً n، برابر با ۱، ۲ و یا ۳ انتخاب می‌شود، و مدل‌های آماریِ استخراج‌شده به‌ترتیب یک‌کلمه‌ای[1]، دوکلمه‌ای[2] و سه‌کلمه‌ای[3] نامیده می‌شوند. همان‌طور که گفتیم در این مقاله از "پیکرۀ متنیِ زبان فارسی" برای آموزش مدل‌های زبانی استفاده شده است. متونِ "پیکرۀ متنیِ زبان فارسی" از منابع مختلف شامل روزنامه‌ها، مجلات و کتاب‌های مختلف فراهم شده است. این پیکره شامل حدود ۱۰۰ میلیون کلمه است ولی نسخۀ اولیۀ آن، که در این مقاله استفاده شده، شامل حدود ۱۰ میلیون کلمه است که به‌ازای هر کلمۀ آن یک برچسب مشخص شده است. این برچسب‌ها شامل مقولۀ نحوی یا اجزای کلام (POS)[4]، و درصورت‌لزوم شامل زیرمقوله‌های نحوی- معناییِ هر کلمه هستند. در کل ۸۸۲ نوع برچسب در پیکرۀ متنی موجود است. با استفاده از این پیکرۀ متنی، مدل‌های چندکلمه‌ایِ مختلفی شامل مدل‌های چندکلمه‌ایِ مبتنی‌بر کلمه، مبتنی‌بر مقولۀ نحوی و مبتنی‌بر طبقه را استخراج کرده‌ایم. مدل‌های چندکلمه‌ایِ مبتنی‌بر مقولۀ نحوی و مبتنی‌بر طبقه، از آن جهت استخراج شدند که در وهلۀ اول به‌نظر می‌رسید میزانِ داده‌های متنی برای آموزشِ مدل‌های مبتنی‌بر کلمه کافی نباشد.

برای استخراج مدل‌های چندکلمه‌ایِ مختلف از پیکرۀ متنی، با دو مشکل مواجه بودیم. مشکلِ اول یک‌دست‌نبودنِ پیکره از لحاظِ املایی، و مشکلِ دوم زیادبودنِ تعدادِ برچسب‌ها در پیکرۀ متنی بود. مشکلِ اول از سیستمِ نگارشِ فارسی و استفاده از الفبای عربی در نگارشِ فارسی ناشی می‌شود. در نگارشِ فارسی بعضی از واژک‌ها[5] (مانند پسوندها و پیشوندها) اجازه دارند که به‌صورتِ پیوسته یا جدا از کلمۀ اصلی نوشته شوند. در حالتِ جدانویسی نیز دو گزینه داریم: استفاده از فاصله به‌عنوانِ جداکننده و استفاده از نویسۀ ZWNJ[6] یا نیم‌فاصله. به‌عنوان‌مثال سه شکلِ ممکن برای اتصالِ پسوندِ جمعِ "ها" و پیشوندِ استمراریِ "می" به کلمات، در جدولِ ۱ نشان داده شده است. همۀ این اَشکالِ املایی در پیکرۀ متنی استفاده شده‌اند.

جدول ۱. صورت‌های نگارشی مختلف برای اتصال پیشوندِ "می" و پسوندِ "ها" به کلمۀ اصلی

جدا با فاصله	جدا بدون فاصله	متصل
کتاب ها	کتاب‌ها	کتابها
می روند	می‌روند	میروند

1. monogram
2. bigram
3. trigram
4. parts of speech
5. morphemes
6. Zero-Width Non-Joiner

مشکلِ دیگر، عدمِ وجودِ استاندارد خاص برای نگارشِ کلماتِ فارسی است که باعث می‌شود یک کلمهٔ خاص چند شکلِ املاییِ مختلف داشته باشد. به‌عنوان‌مثال کلمهٔ **مسئولیت** می‌تواند به شکل‌های **مسوولیت** و **مسؤولیت** نیز نوشته شود. برای حل این مشکلات با بررسی‌های فراوانی که برروی متون پیکره انجام گرفت، همهٔ شکل‌های مختلفِ نگارشیِ یک کلمه به‌شکلِ واحدی تبدیل شدند. مثلاً همهٔ پسوندها و پیشوندها با استفاده از نویسهٔ نیم‌فاصله به کلمهٔ اصلی متصل شدند و کلماتِ دارای چند شکلِ املایی، همگی به‌شکلِ استانداردِ فرهنگستان زبان و ادب فارسی (صادقی و زندی‌مقدم، ۱۳۸۵) تبدیل شدند.

مشکل دیگر زیاد بودنِ تعدادِ برچسب‌ها در پیکرهٔ متنی است. همان‌طور که گفته شد، ۸۸۲ نوع برچسب در پیکرهٔ متنی به‌کار رفته است. علاوه‌براین‌که این تعداد برچسب بسیار زیاد و بعضی از آن‌ها بیش‌ازحد جزئی هستند، تعداد قابل‌ملاحظه‌ای از آن‌ها نیز بسیار کم به‌کار رفته‌اند. بنابراین با بررسیِ برچسب‌ها و استخراجِ آمارِ هر برچسب، آن‌ها را براساسِ تشابهشان (از لحاظ نحوی)، به ۱۶۴ دستهٔ کلی تقسیم کردیم و به هر دسته، یک برچسبِ کلی اختصاص دادیم. با این کار هم تعدادِ کلِ برچسب‌ها کاهش می‌یابد و هم برچسب‌های کم‌کاربرد در دسته‌های کلی‌تر جای می‌گیرند. به تعدادی از برچسب‌ها نیز که بسیار کم به‌کار رفته بودند، و درضمن در هیچ دسته‌ای جای نمی‌گرفتند، برچسبِ IGNORE نسبت داده شد. یک برچسبِ NULL نیز به‌عنوانِ نشانهٔ پایانِ جمله و شروع جملهٔ جدید درنظر گرفته شد. بنابراین در کل، ۱۶۶ برچسب به‌دست آمد که در استخراجِ مدل‌های مبتنی‌بر مقولهٔ نحوی مورد استفاده قرار گرفتند (بحرینی و دیگران، ۲۰۰۶). پس‌از یک‌دست‌سازیِ املایی و کاهشِ تعدادِ برچسب‌ها، آمارهای زیر از پیکرهٔ متنی استخراج شدند (بحرانی و دیگران، ۱۳۸۵):

آمارِ یک‌کلمه‌ای‌ها یا تعدادِ رخدادِ هرکلمه در پیکرهٔ متنی (با استفاده از این آمار ۲۰.۰۰۰ کلمهٔ پرکاربرد از پیکرهٔ متنی استخراج و به‌عنوانِ مجموعهٔ واژگان درنظر گرفته شد)، آمار دوکلمه‌ای‌ها، آمار سه‌کلمه‌ای‌ها، آمار هر برچسب، آمار دو برچسبِ متوالی، آمار سه برچسبِ متوالی و تعدادِ رخدادِ هر برچسب به‌ازای هرکلمه (آمار برچسب-کلمه).

لازم‌به‌ذکر است که آمارهای دوکلمه‌ای، سه‌کلمه‌ای و برچسب- کلمه برای ۲۰۰۰۰ کلمهٔ موجود در مجموعهٔ واژگان، استخراج شدند. سایر کلمات، همگی به‌عنوانِ یک کلمهٔ خارج از واژگان درنظر گرفته شدند.

با استفاده از آمارِ استخراج‌شده، مدل‌های چندکلمه‌ایِ مبتنی‌بر کلمه و مبتنی‌بر مقولهٔ نحوی ساخته شدند. از آن‌جا که در مدل‌های دوکلمه‌ای و سه‌کلمه‌ای، تعدادِ زیادی پارامتر با مقدار صفر وجود

دارند، روشِ هموارسازی ویتن[1]- بل[2] (ویتن و بِل، ۱۹۹۱) برای ازبین‌بردنِ مقادیر صفر برروی مدل‌ها اعمال گردید.

علاوه‌بر مدل‌های چندکلمه‌ایِ مبتنی‌بر کلمه و مبتنی‌بر مقولهٔ نحوی، مدل‌های چندکلمه‌ایِ مبتنی‌بر طبقه نیز از پیکرهٔ متنی استخراج شد. در این روش، کلمات دسته‌بندی شدند و هرکلمه در یک یا چند طبقه جای گرفت. سپس مدل‌های چندکلمه‌ای، براساسِ دسته‌های کلمات (به‌جای خودِ کلمات) استخراج گردیدند. مدل‌های مبتنی‌بر طبقه، مانند مدل‌های مبتنی‌بر مقولهٔ نحوی، مشکلِ کمبودِ داده‌های آموزشی را تاحدزیادی برطرف می‌سازند. برای دسته‌بندیِ کلمات، روش‌های خودکارِ زیادی پیشنهاد شده‌اند. دو روشِ مشهور در این زمینه، الگوریتمِ مارتین[3] (مارتین و دیگران، ۱۹۹۸)، و الگوریتمِ براون[4] (براون و دیگران، ۱۹۹۲) هستند. در این مقاله از الگوریتمِ مارتین برای دسته‌بندیِ کلمات استفاده شده است. در این الگوریتم ابتدا یک دسته‌بندیِ اولیه از کلمات به‌وجود می‌آید و سپس کلمات، بینِ طبقات، آن‌قدر جابه‌جا می‌شوند تا معیار سرگشتگی[5] به حداقل برسد. پس‌از دسته‌بندیِ کلمات، آمارِ دوکلمه‌ای و سه‌کلمه‌ای، بینِ دسته‌های کلمات استخراج می‌گردد، و مدل‌های دوکلمه‌ای و سه‌کلمه‌ایِ مبتنی‌بر طبقه با استفاده از این آمار ساخته می‌شوند.

مدل‌های چندکلمه‌ایِ مذکور، در بخشِ جستجوی کلمات[6]، در سیستمِ بازشناسی به‌کار می‌روند (شکل ۱). در این روش امتیازِ مدلِ زبانی و امتیازِ مدلِ آکوستیک "در حین جستجو" با هم ترکیب می‌شوند (هارپر[7] و دیگران، ۱۹۹۴). به‌کارگیریِ مدلِ چندکلمه‌ای، درحینِ جستجو به‌این‌صورت است که هنگامی‌که الگوریتمِ جستجو، فرضیه‌های مختلف را برای بازشناسیِ کلمات به‌پیش می‌برد، با شناساییِ یک کلمهٔ جدید، احتمالِ چندکلمه‌ایِ آن‌را نیز، همراه با امتیازِ آواییِ آن، در امتیازِ فرضیه ضرب می‌کند. به‌این معنا که اگر امتیازِ کنونیِ یک فرضیه پس‌از شناساییِ کلمهٔ w_n، S_n باشد، و این فرضیه پس‌از بسط‌داده‌شدن، کلمهٔ w_{n+1} را به‌عنوانِ کلمهٔ بعدی شناسایی کند، امتیازِ جدیدِ فرضیه برابر است با:

$$S_{n+1} = S_n . S_{AM}(w_{n+1}) . S_{LM}(w_{n+1})^{LMW} \qquad (۲)$$

1.I. Witten
2. T. Bell
3. S. Martin
4.P. Brown
5. perplexity
6. word decoding
7. M.P. Harper

$S_{AM}(w_{n+1})$ امتیاز مدل آوایی کلمهٔ w_{n+1}، و $S_{LM}(w_{n+1})$ امتیازِ مدلِ زبانیِ آن است. معمولاً به‌دلیلِ تفاوت در مقیاس‌های $S_{AM}(w_{n+1})$ و $S_{LM}(w_{n+1})$، یک پارامترِ وزن (LMW)[1] برای امتیازِ مدلِ زبانی درنظر گرفته می‌شود. معمولاً برای پرهیز از به‌کاربردنِ اعدادِ خیلی کوچک، به‌جای خودِ امتیازها از لگاریتمِ آن‌ها استفاده می‌گردد:

$$\log S_{n+1} = \log S_n + \log S_{AM}(w_{n+1}) + LMW.\log S_{LM}(w_{n+1}) \qquad (۳)$$

بنابراین لگاریتمِ امتیازِ زبانیِ کلمهٔ جدید به‌صورتِ وزن‌دار، با لگاریتمِ امتیازِ آوایی کلمه و همچنین با لگاریتمِ امتیازِ فرضیه جمع زده می‌شود و امتیازِ جدیدِ فرضیه را می‌سازد. امتیازِ فرضیه به‌همین‌روش، با استفاده از ترکیبِ امتیازهای آوایی و زبانی به‌دست می‌آید، و درنهایت پس‌از کامل‌شدنِ فرضیه‌ها، فرضیه با بالاترین امتیاز (در بین N دنبالهٔ خروجیِ الگوریتم جستجو)، خروجیِ بخشِ بازشناسیِ سیستم خواهد بود. در این روش، مدلِ چندکلمه‌ای، درواقع، رویهٔ جستجو را برای یافتنِ دنبالهٔ کلماتِ محتمل‌تر هدایت می‌کند.

$S_{LM}(w_{n+1})$ برای مدل‌های یک‌کلمه‌ای، دوکلمه‌ای و سه‌کلمه‌ایِ مبتنی‌بر کلمه، به‌ترتیب با روابطِ زیر محاسبه می‌شود:

$$S_{monogram}(w_{n+1}) = P(w_{n+1}) = \frac{N_{monogram}(w_{n+1})}{N_{total}} \qquad (۴)$$

$$S_{bigram}(w_{n+1}) = P(w_{n+1} \mid w_n) = \frac{N_{bigram}(w_n w_{n+1})}{N_{monogram}(w_n)} \qquad (۵)$$

$$S_{trigram}(w_{n+1}) = P(w_{n+1} \mid w_{n-1}w_n) = \frac{N_{trigram}(w_{n-1}w_n w_{n+1})}{N_{bigram}(w_{n-1}w_n)} \qquad (۶)$$

در روابطِ بالا، $N_{monogram}(w_{n+1})$، $N_{bigram}(w_n w_{n+1})$ و $N_{trigram}(w_{n-1}w_n w_{n+1})$، به‌ترتیب آمارهای یک‌کلمه‌ای، دوکلمه‌ای و سه‌کلمه‌ای و N_{total}، تعدادِ کلِ کلمات، در پیکرهٔ متنی است.

1. Language Model Weight

۴. ۲. مدلِ زبانیِ دستوری

خروجیِ سیستم‌های بازشناسیِ گفتارِ پیوسته، معمولاً به‌صورتِ فهرستی از بهترین دنباله‌های کلمات (N-best) معادل با گفتارِ ورودی هستند. در این سیستم‌ها، بهترین دنباله‌های کلمات، معمولاً با درنظرگرفتنِ مدل‌های آوایی و مدل‌های زبانیِ آماری به‌دست می‌آیند، و امتیازِ آن‌ها براساسِ این مدل‌ها محاسبه می‌شود و از میانِ آن‌ها، دنبالهٔ با بالاترین امتیاز به‌عنوانِ خروجیِ نهایی درنظر گرفته می‌شود. چون مدلِ آوایی و مدلِ زبانیِ آماری هیچ محدودیتِ دستوری بر روی خروجی‌ها اعمال نمی‌کنند، پس دنباله‌های کلماتِ حاصل ممکن است ازلحاظِ دستوری، درست یا نادرست باشند. بنابراین هیچ تضمینی برای درست بودنِ بهترین فرضیه ازلحاظِ دستوری نیست. درحالی‌که ما انتظار داریم که اگر جملهٔ گفتاریِ ورودی ازنظرِ دستوری درست باشد، سیستمِ بازشناسی، یک جملهٔ درستِ دستوری به‌دست دهد. بنابراین لازم است خروجی‌های سیستمِ بازشناسی ازلحاظِ درست‌بودنِ دستوری نیز بررسی گردند. برای این کار احتیاج به یک مدلِ دستوری داریم. مدلِ دستوری، شاملِ الگوهای نحویِ جملات و همچنین مشخصه[1]های دستوریِ کلماتِ مورد استفاده است. با استفاده از این مدل، یک تحلیل‌گرِ نحوی، خروجیِ سیستمِ بازشناسی را پردازش می‌کند و دنباله‌های نادرست (ازنظر دستوری) را رد کرده و درنهایت دنباله‌ای را به‌عنوانِ خروجیِ نهایی تولید می‌کند که هم ازلحاظِ دستوری درست باشد، و هم دارای بیشترین امتیاز (ازلحاظِ آوایی و آماری) در بینِ N دنبالهٔ خروجی باشد.

در این تحقیق از نظریهٔ ساخت- گروهیِ تعمیم‌یافته، یا GPSG[2] (گزدار[3] و دیگران، ۱۹۸۵)، برای ارائهٔ مدلِ دستوریِ موردنظر استفاده شده است. ازآن‌جاکه GPSG یک نوع دستور ساخت- گروهی است، تلاش می‌کند جملاتِ زبان را به‌صورتِ ترکیبی از تعدادی گروهِ دستوری[4] درنظر بگیرد و سپس این گروه‌های دستوری را نیز ترکیبی از گروه‌های دستوریِ کوچک‌تر، و همین‌طور الی‌آخر درنظر گرفته تا نهایتاً به مرزِ کلمات برسد. در انتخابِ گروه‌های زبانِ فارسی، با اندکی انحراف، از نظریهٔ X-تیره[5] (ردفورد[6]، ۱۹۸۸) پیروی شده و اسم (N)، فعل (V)، صفت (ADJ)، قید (ADV) و حرف اضافه (P) به‌عنوانِ مقوله‌های نحوی پایه‌ای در نظر گرفته شده‌اند، که می‌توانند به‌عنوانِ هستهٔ گروه‌های اسمی ($\overline{N}, \overline{\overline{N}}$)، فعلی($\overline{V}, \overline{\overline{V}}$)، صفتی ($\overline{ADJ}, \overline{\overline{ADJ}}$)، قیدی ($\overline{ADV}, \overline{\overline{ADV}}$) و حرف اضافه‌ای ($\overline{P}, \overline{\overline{P}}$) قرار گیرند. سپس تلاش شده است تا ساختِ نحویِ هریک از این گروه‌ها براساسِ مقوله‌های نحویِ کوچک‌تر صورت‌بندی شود. به‌عنوان‌مثال، $\overline{N}$ ترکیبِ اسم با همهٔ وابسته‌های پسین آن و $\overline{\overline{N}}$ ترکیبِ

1. feature
2. Generalized Phrase Structured Grammar
3. G. Gazdar
4. phrase
5. x-bar Theory
6. A. Radford

$\overline{N}$ با همهٔ وابسته‌های پیشین اسم درنظر گرفته شده‌اند. البته در برخی موارد، درنظرگرفتنِ دو لایه برای تعریف قواعدی که بتوانند تمامیِ ساختارهای نحویِ ممکنِ یک گروه را دربر بگیرند، کافی نبود. در این‌گونه موارد بینِ لایه‌های ۱ و ۲، لایه‌های میانی تعریف شدند. به‌عنوان‌مثال، درموردِ گروه‌های اسمی، به‌دلیلِ وجودِ چهار نوع وابستهٔ پسین برای اسم (وابستهٔ صفتی، وابستهٔ حرف اضافه‌ای، وابستهٔ گروه‌اسمی و وابستهٔ جمله)، بالاخص با توجه به امکانِ تکرارِ وابستهٔ صفتی، یک لایهٔ $\overline{N}^{+}$ بینِ لایه‌های $\overline{N}$ و $\overline{\overline{N}}$ تعریف شد، و $\overline{N}$ ترکیبِ اسم با همهٔ وابسته‌های صفتیِ آن، و $\overline{N}^{+}$ نیز ترکیبِ $\overline{N}$ با سایر وابسته‌های پسینِ اسم درنظر گرفته شدند.

یکی از قواعدِ ساختِ گروهی $\overline{N}^{+}$ در زیر نشان داده شده است:

$$\overline{N}^{+} \rightarrow *\overline{N}\,[GEN+, PRO-]\,\overline{\overline{N}}\,(\overline{\overline{P}})\,(S[COMP+, GAP])$$

این قاعده، ساختارِ گروهِ اسمی را با وابسته‌های پسینِ آن، هنگامی‌که اسم دارای وابستهٔ گروه اسمی (مضاف الیه) است، نشان می‌دهد. همان‌گونه که گفته شد، $\overline{N}$ در سمتِ راستِ قاعدهٔ فوق، ترکیبِ اسم با وابسته‌های صفتیِ احتمالیِ آن را نشان می‌دهد. مشخصهٔ دوارزشیِ +GEN، بیان می‌کند که ترکیبِ $\overline{N}$ باید حتماً دارای کسرهٔ اضافه باشد و مشخصهٔ دوارزشیِ - PRO ، تأکید می‌کند که $\overline{N}$ نمی‌تواند دارای هستهٔ ضمیر باشد.

درحدود ۱۷۰ قاعده، مشابه با قاعدهٔ فوق استخراج شدند، که بسیاری از ساختارهای رایجِ گروه‌های اسمی، فعلی، حرف اضافه‌ای، صفتی، قیدی و بالاخره جمله را پوشش می‌دهند. جزئیاتِ بیشتر در موردِ دامنهٔ پوششِ قواعدِ دستوریِ استخراج شده، در مقالهٔ حافظی و دیگران (۱۳۸۵) آمده است.

۵. آزمایش‌ها و نتایج

در این بخش، نتایجِ به‌کارگیریِ مدل‌های زبانی در سامانهٔ بازشناسیِ گفتارِ پیوستهٔ زبانِ فارسی ارائه می‌شود. سیستم برای ۲۹ واجِ زبانِ فارسی، و سکوت به‌عنوانِ واجِ سی‌ام آموزش داده شد. مدل‌های مخفیِ مارکوف، که به‌کار گرفته شدند، ازنوعِ چپ‌به‌راست و شامل ۶ حالت و ۱۶ تلفیقِ گوسی در هرحالت هستند. حجمِ واژگان، ۱۰۰۰ کلمه است و از زیرمجموعه‌های آموزش و آزمایشِ دادگانِ فارس‌داتِ کوچک، به‌ترتیب برای آموزش و بازشناسی استفاده شد. برای ارزیابیِ مدل‌های زبانیِ به‌کاررفته با پارامترهای مختلف، از دو معیارِ سرگشتگی و نرخِ خطای کلمه[1] (WER) استفاده شد. جدول ۲ نتایجِ بازشناسی را روی زیرمجموعهٔ آزمایشِ فارس‌داتِ کوچک نشان می‌دهد، که شامل ۱۴۰

1. Word Error Rate

جمله از ۷ گوینده است. برای همهٔ مدل‌های زبانی، روشِ هموارسازیِ[1] ویتن- بل استفاده شد. تعداد دسته‌های مورد استفاده برای آزمایش‌های مربوط‌به مدلِ زبانیِ مبتنی‌بر طبقه، ۲۰۰ دسته است. اولین نتیجه‌ای که از نتایجِ این جدول حاصل می‌شود آن است که سیستمِ پایه[2]، بدونِ مدلِ زبانی، دارای نرخِ خطای کلماتِ بالایی است و اطلاعاتِ موجود در مدلِ زبانیِ عملکردِ سیستم را بهبود می‌بخشد. ازآن‌جاکه مدلِ زبانیِ آماریِ مبتنی‌بر کلمه، بهبودِ بیشتری را در عملکردِ سیستم، به‌نسبتِ سایر مدل‌های زبانیِ آماری ایجاد می‌کند، از این مدلِ زبانی برای آزمایش‌های بعد استفاده شد. همچنین اثرِ استفاده از اطلاعاتِ دستوری در کاهشِ نرخِ خطای کلمات در جدول ۲ مشهود است.

جدول ۲. عمل‌کردِ موتورِ بازشناسی در سطحِ کلمه در شرایطِ بدونِ نوفه.

مدل زبانی	نرخ خطای کلمات (WER%)
بدون مدل زبانی	۳۸.۱۴
سه‌کلمه‌ای مبتنی بر مقوله نحوی	۲۴.۶۸
سه‌کلمه‌ای مبتنی بر طبقه	۲۳.۴۰
سه‌کلمه‌ای مبتنی بر کلمه	۲۱.۷۶
سه‌کلمه‌ای مبتنی بر مقوله نحوی به همراه قواعد دستوری	۱۸.۲۰

جدول ۳ میزانِ سرگشتگیِ محاسبه‌شده روی ۷۵۰ جمله (حدود ۱۰.۰۰۰ کلمه) از دادگانِ فارس‌داتِ بزرگ را براساسِ مدلِ چندکلمه‌ایِ مبتنی‌بر کلمه (براساسِ واژگانِ ۲۰.۰۰۰ کلمه‌ای) نشان می‌دهد. به‌منظور کاهشِ اندازهٔ حافظهٔ مورد نیاز برای مدلِ زبانی، چندکلمه‌ای‌های با فراوانیِ کم، از مدل خارج شدند و تعداد آستانهٔ مورد نظر برای تعیینِ حذف یا عدمِ حذفِ چندکلمه‌ای‌ها، به‌عنوانِ پارامترِ میان‌بُر[3] (کلارکسون[4] و روزنفلد[5]، ۱۹۹۷) در نتایج آمد. نتایجِ مندرج در جدول ۳ نشان می‌دهد که چگونه میان‌بُرهای مربوط به دوکلمه‌ای و سه‌کلمه‌ای، روی اندازهٔ حافظهٔ مورد نیاز و سرگشتگیِ مدلِ زبانیِ سه‌کلمه‌ای تأثیر می‌گذارند. نتیجهٔ خلاصه‌ای که از این جدول می‌توان گرفت این است که مقادیرِ میان‌بُر، اندازهٔ حافظهٔ موردنیاز برای مدلِ زبانی را کاهش می‌دهد، ولی افزایشِ معناداری در سرگشتگی ایجاد نمی‌کند. براساسِ نتایجِ این جدول برای تعدادِ دوکلمه‌ای، مقدارِ میان‌بُر برابر با صفر و

1. smoothing
2. baseline
3. cutoffs
4. P. Clarkson
5. R. Rosenfeld

برای تعداد سه‌کلمه‌ای، مقدارِ آن برابر با ۱ درنظر گرفته شد. همچنین در جدولِ ۴، نرخِ خطای کلماتِ موتورِ بازشناسی روی فارس‌داتِ کوچک و بزرگ، با استفاده از مدلِ واجیِ مستقل‌ازبافت و وابسته‌به‌بافت آمده است. حجمِ واژگان در این آزمایش ۲۰.۰۰۰ کلمه است. نتایجِ مندرج در این جدول نشان می‌دهد که استفاده از مدلِ وابسته‌به‌بافت در همهٔ حالات منجر به کاهشِ نرخِ خطای کلمات می‌شود. همچنین بهترین نتایج، در حالتِ استفاده از دادگانِ فارس‌داتِ بزرگ، برای آموزش و آزمایش و مدل‌سازیِ وابسته به بافت، حاصل می‌شود.

جدول ۳. اثر پارامتر میان‌بُر بر اندازه و سرگشتگیِ یک مدلِ زبانیِ سه‌کلمه‌ای.

اندازه (MB)	سرگشتگی	میان‌بُر (سه‌کلمه‌ای)	میان‌بُر (دوکلمه‌ای)
۳۶	۱۳۴.۵۴	۰	۰
۲۰	۱۳۴.۷۶	۱	۰
۱۷	۱۳۵.۸۲	۲	۰
۱۰	۱۴۳.۱۸	۱	۱
۷.۸	۱۴۳.۲۶	۲	۱

جدول ۴. نرخ خطای کلماتِ موتورِ بازشناسی روی فارس‌داتِ کوچک و بزرگ با استفاده از مدلِ واجیِ مستقل‌ازبافت (واج)، و وابسته‌به‌بافت (سه‌واج).

تست		آموزش	
فارس‌داتِ کوچک	فارس‌داتِ بزرگ	وابستگی به بافت	دادگان
۲۵.۷۷	۲۹.۶۰	مستقل	فارس‌داتِ کوچک
۱۶.۷۹	۲۰.۵۱	وابسته	
۳۷.۳۹	۶.۱۰	مستقل	فارس‌داتِ بزرگ
۲۶.۸۵	۵.۲۱	وابسته	

۶. خلاصه و نتیجه‌گیری

در این مقاله، خلاصه‌گزارشِ تحقیقاتِ انجام‌شده برای طراحی و ساختِ یک موتورِ بازشناسیِ گفتارِ پیوستهٔ زبانِ فارسی ارائه شد. این سامانه حاصلِ سال‌ها فعالیتِ تحقیقاتی بوده و اولین موتورِ بازشناسیِ زبانِ فارسی است که درعمل به‌کارگیری شده است. پس‌از تشریحِ ساختارِ کلیِ سامانه، روش‌های به‌کاررفته برای مدل‌سازیِ واجی و طراحیِ مدلِ زبانیِ موتورِ بازشناسی، مطرح شد. در مدل‌سازیِ آوایی، برای به‌دست‌آوردنِ نتایجِ مطلوب، از مدل‌های وابسته‌به‌بافتِ مبتنی‌بر سه‌واج، استفاده شد. درضمن برای کاهشِ تعدادِ مدل‌های لازم، روشِ دسته‌بندیِ واج‌ها به‌کار رفت، و پس‌از طراحیِ اولیه، بر روی دادگانِ مناسب پیاده‌سازی و اجرا شد. تأکیدِ اصلیِ این مقاله برروی مدل‌های زبانیِ به‌کاررفته بود. مدل‌های زبانیِ آماری و دستوری در موتور بازشناسی استفاده شدند. مدل‌های زبانیِ آماری، از سه‌نوعِ مبتنی‌بر کلمه، مبتنی‌بر مقولهٔ نحوی و مبتنی‌بر طبقه، استفاده و نتایجِ عمل‌کردِ آن‌ها با یک‌دیگر مقایسه شد. دراین‌بین، از مدلِ زبانیِ آماریِ مبتنی‌بر کلمه، بهترین نتایج حاصل شد. درضمن اثرِ پارامترِ میان‌بُر، که نقشِ آستانهٔ تعدادِ چندکلمه‌ایِ مشاهده‌شده برای حذف یا لحاظ‌شدن در مدلِ زبانیِ آماری را ایفا می‌کند، بررسی شد و مقدارِ بهینهٔ آن به‌دست آمد. همچنین مدلِ زبانیِ دستوریِ به‌کاررفته، شرح داده شد و درعمل موردِ آزمایش قرار گرفت.

منابع

الماس‌گنج، فرشاد و سیدعلی سیدصالحی و محمود بی‌جن‌خان و حسین صامتی و جواد شیخ‌زادگان (۱۳۸۰). "شنوا ۱ - سیستم بازشناسی گفتار پیوستهٔ فارسی"، *مجموعه مقالاتِ کنفرانسِ بین‌المللیِ مهندسیِ برقِ ایران*، ۶۳-۵۶.

_____ و سیدعلی سیدصالحی و محمود بی‌جن‌خان و حسین رازی‌زاده و محمدرضا اصغری (۱۳۸۳). " نرم‌افزارِ بازشناسیِ گفتارِ پیوستهٔ فارسیِ شنوا ۲". *مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه*، ۸۲- ۷۷.

بحرانی، محمد و حسین صامتی (۱۳۸۴). "استخراج و مدل‌سازی واحدهای آواییِ وابسته‌به‌بافت برای بهبودِ دقتِ بازشناسیِ گفتارِ پیوسته با روشِ دسته‌بندیِ واج‌ها". *نشریهٔ مهندسی برق و مهندسی کامپیوتر ایران*. سال ۳، شماره ۱، ۵۱- ۴۵.

_____ و حسین صامتی و نازیلا حافظی و سعیده ممتازی و حامد موثق (۱۳۸۵). "به‌کارگیریِ پیکرهٔ متنیِ زبانِ فارسی در ساختِ مدل‌های زبانیِ آماری برای سیستم‌های بازشناسیِ گفتارِ پیوستهٔ فارسی". *مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه*، ۱۰۹- ۹۲.

حافظی، محمدمهدی و حسین صامتی و نیلوفر منصوری و نیلوفر منتظری و محمد بحرانی و حامد موثق (۱۳۸۵). "ارائهٔ یک مدل دستوری برای بهبود دقت سیستم‌های بازشناسی گفتار پیوستهٔ فارسی". *مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه*، ۹۱- ۸۰.

شیخ‌زادگان، جواد و محمود بی‌جن‌خان (۱۳۸۵). "دادگان‌های گفتاری زبان فارسی". *مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه*، ۲۶۱- ۲۴۷.

صادقی، علی‌اشرف و زهرا زندی‌مقدم (۱۳۸۵). *فرهنگ املایی خط فارسی*. انتشارات فرهنگستان زبان و ادب فارسی: تهران.

صامتی، حسین و حامد موثق و باقر باباعلی و محمد بحرانی و خسرو حسین‌زاده و امین فاضل دهکردی و حمیدرضا ابوطالبی و هادی ویسی و یاسمین مکری و نیلوفر منتظری و محمد نظامی رنجبر (۱۳۸۳). "سیستم بازشناسی گفتار پیوستهٔ فارسی با واژگان بزرگ". *مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه*، ۷۵- ۶۹.

غلامپور، ایمان (۱۳۷۹). "بازشناسی مستقل از گویندهٔ واج‌های فارسی در صحبت پیوسته". پایان‌نامهٔ دکترا. دانشکدهٔ مهندسی برق، دانشگاه صنعتی شریف.

ولی، منصور (۱۳۸۵). "بازشناسی مقاوم گفتار به‌منظور جبران‌سازیِ تنوعات گفتار میکروفونی- تلفنی توسط شبکه‌های عصبی". پایان‌نامهٔ دکترا. دانشکده مهندسی پزشکی، دانشگاه امیرکبیر.

همایون‌پور، محمدمهدی (۱۳۸۳). "سیستم بازشناسی گفتار پیوسته به کمک هیبرید شبکهٔ عصبی و مدل مخفیِ مارکوف و با استفاده از مدل‌های زبانی و روش‌های جستجو". *مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه*، ۱۸۳.

Ahadi, S. M. (1999). "Recognition of Continuous Persian Speech Using a Medium-sized Vocabulary Speech Corpus". *In Proc. Eurospeech99*. 863-866.

Allen, J. (1995). *Natural Language Understanding*. the Benjamin/Cummings Publishing Company, Inc: Redwood City, CA.

Babaali, B. & H. Sameti (2004). "The Sharif Speaker-Independent Large Vocabulary Speech Recognition System". *The 2nd Workshop on Information Technology & Its Disciplines*. Kish Island: Iran.

Bahrani, M. & H. Sameti & N. Hafezi & H. Movasagh (2006). "Building and Incorporating Language Models for Persian Continuous Speech Recognition Systems". *In Proc. the 5th international conference on Language Resources and Evaluation*. 2590-2593. Genoa: Italy.

Bijankhan, M. & J. Seikhzadeghan & M. Roohani & Y. Samareh & K. Lucas & M. Tebyani (1994). "FARSDAT-The Speech Database of Farsi Spoken Language". *In Proc. the 5th Australian International Conference on Speech Science and Technology*. 826-831.

_____. & J. Seikhzadeghan & M. Bahrani & M. Ghayoomi (2011). "Lessons from Creation of a Persian Written Corpus: Peykare". *Language Resources and Evaluation Journal*. Vol. 45, No. 2. 143-164.

Brown, P. & V. Della Pietra & P. deSouza & J. Lai & R. L. Mercer (1992). "Class-based

n-gram models of natural language". *Computational Linguistics,* 18(4). 467–479.

Clarkson, P. & R. Rosenfeld (1997). "Statistical Language Modeling Using the CMU-Cambridge Toolkit". *In Proc. Eurospeech97*. vol. 5. 2707-2710.

Gazdar, G. & E. H. Klein & G. K. Pullum & I. A. Sag (1985). *Generalized Phrase Structure Grammar*. Harvard University Press: Great Britain.

Harper, M. P. & L. H. Jamieson & C. D. Mitchell & G. Ying & S. Potisuk & P. N. Srinivasan & R. Chen & C. B. Zoltowski & L. L. McPheters & B. Pellom & R. A. Helzerman (1994). ''Integrating Language Models with Speech Recognition''. *AAAI-94 Workshop on the Integration of Natural Language and Speech Processing*. 139-146.

Martin, S. & J. Liermann & H. Ney (1998). "Algorithms for bigram and trigram word clustering". *Speech Communication*. vol. 24. 19-37.

Rabiner, L. & B. H. Juang (1993). *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall.

Radford, A. (1988). *Transformational Grammar*. Cambridge University Press.

Sameti, H. & H. Veisi & M. Bahrani & B. Babaali & K. Hosseinzadeh (2008). "Nevisa, a Persian Continuous Speech Recognition System". *In Communications in Computer and Information Science*, *Advances in Computer Science and Engineering*. Vol. 6. 485-492. Springer Berlin: Heidelberg.

_____. & H. Veisi & M. Bahrani & B. Babaali & K. Hosseinzadeh (2009). "A Large Vocabulary Continuous Speech Recognition System for Persian Language". submitted to *IEICE - Transactions on Information and Systems*.

Singh, R.& B. Raj & R. M. Stern (1999). "Automatic Clustering and Generation of Contextual Questions for Tied States in Hidden Markov Models". In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)*. 117-120.

Witten, I. & T. Bell (1991). "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression". In *IEEE Transactions on Information Theory 37(4)*. 1085-1094.

Young, S.J. & P. C. Woodland (1993). "The Use of State Tying in Continuous Speech Recognition". In *European Conference on Speech Communication and Technology (EUROSPEECH'93)*. Berlin: Germany ISCA. 2203-2206.

_____. &. J. J. Odell & P. C. Woodland (1994). "Tree-based State Tying for High Accuracy Acoustic Modeling". In *Human Language Technology Conference*. Association for Computational Linguistics Morristown.