EasyChair Preprint

№ 1641

Design and collecting the speaker independent data set of voice commands for controlling the smart home appliances based on Persian speech

Leila Safarpoor Kalkhoran, Shima Tabibian and
Elaheh Homayounvala

October 12, 2019

طراحی و جمع‌آوری دادگان مستقل از گویشور حاوی فرامین صوتی برای کنترل لوازم خانگی هوشمند مبتنی برگفتار فارسی

برای بیست و پنجمین کنفرانس بین‌المللی انجمن کامپیوتر ایران

لیلا صفرپور کلخوران[۱]، شیما طبیبیان[۲] و الهه همایون‌والا[۳]

[۱] دانشجوی کارشناسی ارشد، پژوهشکده فضای مجازی، دانشگاه شهید بهشتی ، تهران
l.safarpour@mail.sbu.ac.ir
[۲] استادیار، پژوهشکده فضای مجازی، دانشگاه شهید بهشتی ، تهران
sh_tabibian@sbu.ac.ir
[۳] استادیار، دانشکده کامپیوتر، گلد اسمیت، دانشگاه لندن، لندن
e.homayounvala@gold.ac.uk

چکیده

امروزه با پیشرفت روزافزون تکنولوژی و سیستم‌های پردازش گفتار، استفاده از گفتار علاوه بر تعاملات انسان‌ها با یکدیگر در تعاملات میان انسان و ماشین نیز ورود پیدا کرده است. یکی از جنبه‌های به‌کارگیری گفتار در زمینه‌هایی مانند خانه‌های هوشمند، جایگزین کردن دستورات صوتی، با فرمان‌هایی است که از طریق لمس به دستگاه داده می‌شود. به این منظور یک سیستم تشخیص فرامین صوتی می‌تواند طراحی و پیاده‌سازی شود. سیستم مذکور قابلیت آن را دارد که هرکجا کاربر درخواستی از سیستم داشت، به راحتی با استفاده از دستورات سادهٔ گفتاری، نیاز خود را برآورده کند. یکی از نیازهای اولیه برای داشتن هر سیستم بازشناس گفتار، وجود دادگان غنی برای آموزش و ارزیابی آن می‌باشد. با توجه به در دسترس نبودن دادگان حاوی فرامین صوتی برای کنترل لوازم خانگی هوشمند به زبان فارسی، در این مقاله به بررسی روند طراحی، جمع‌آوری و ارزیابی یک مجموعه دادگان فرامین صوتی مستقل از گویشور برای کنترل لوازم خانگی هوشمند (تلویزیون، ضبط صوت، لامپ) مبتنی بر گفتار فارسی پرداخته شده است. این دادگان به دو بخش آموزش و آزمون تقسیم شده است. برای ارزیابی دادگان از روش مدل مخفی مارکف و روش حافظهٔ کوتاه مدت ماندگار استفاده شده است. نتایج ارزیابی دادگان مبتنی بر مدل مخفی مارکف نشان می‌دهد که بازشناس کلمات آموزش یافته بر روی دادگان مذکور در بهترین حالت از صحت ۹۳٪ و دقت ۹۲٪ برخوردار بوده و میزان خطا در سطح کلمه تنها ۸٪ است. همچنین نتایج ارزیابی مبتنی برحافظهٔ کوتاه مدت ماندگارحاکی از آن است که دقت بازشناسی در سطح کلمه با بهترین تنظیمات از میان تنظیمات مختلف انجام شده، حدودا برابر با ۹۸٪ است.

کلمات کلیدی

فرامین صوتی مستقل از گویشور، لوازم خانگی هوشمند، دادگان گفتاری فارسی، مدل مخفی مارکف، حافظهٔ کوتاه مدت ماندگار

۱- مقدمه

گفتار یکی از طبیعی‌ترین راه‌های برقراری ارتباط برای انسان‌هاست. ارتباط انسان با ماشین نیز از این امر مستثنی نیست. یکی از حوزه‌های ارتباط افراد با وسایلشان، خانه‌های هوشمند می‌باشد. مقصود از هوشمند شدن خانه، خودکارسازی و تعاملی کردن فعالیت‌هایی است که در خانه انجام می‌شود؛ مانند تهویه مطبوع هوا، سیستم حرارتی، کنترل میزان روشنایی و سیستم امنیتی. تعاملی نمودن این فعالیت‌ها مبتنی بر بازشناسی گفتار و درک زبان طبیعی توسط لوازم خانهٔ هوشمند می‌باشد[۱]. فرمان‌های صوتی، نقش مهمی را در زندگی امروزه انسان‌ها بازی می‌کنند. سیستم تشخیص فرامین صوتی، در خانه‌های هوشمند، دستیارهای صوتی، کنترل و مدیریت نرم افزارها و جستجوی صوتی مورد استفاده قرار می‌گیرد. در اغلب کاربردهای مذکور، استفاده از گفتار به جای لمس، کلیک کردن یا فشردن دکمه، علاوه بر محبوبیت و سادگی بیشتر، این امکان را فراهم می‌کند که کاربران بدون از دست دادن تمرکزشان، هم‌زمان به کاردیگری هم مشغول باشند[2].

بدیهی است که این قبیل ارتباطات مبتنی بر گفتار نیازمند استفاده از سیستم‌های کارای بازشناسی و سنتز گفتار می‌باشد. آموزش یک سیستم کارای بازشناس گفتار در یک کاربرد خاص علاوه بر استفاده از روش‌های نوین پردازش گفتار، نیازمند وجود دادگان غنی متناسب با آن کاربرد می‌باشد.

درحیطهٔ فرمان‌های صوتی، دادگان گفتاری بسیار معدودی در کشور تهیه شده است، که اغلب این دادگان نیز، توسط آزمایشگاه‌های پردازش صوت و گفتار مقیم در دانشگاه‌ها و برای استفاده‌های شخصی تهیه شده‌اند که معمولا در معرض استفاده عموم قرار نگرفته و درجایی ثبت نشده‌اند.لازم به ذکر است که دادگان موجود در مقالات [3,4]، حاوی فرامین صوتی است که به ترتیب در زمینهٔ ربات اسکات و فضاپیما، در داخل کشور جمع آوری و ارائه شده است. برخلاف کارهای معدودی که در داخل کشور درحیطهٔ فرمان‌های صوتی مخصوصا در حوزهٔ خانه‌های هوشمند انجام شده‌اند، در خارج از کشور دادگان بسیاری با هدف‌های مختلف بر روی فرامین صوتی مورد استفاده در خانه‌های هوشمند جمع‌آوری شده‌اند[8-5]. از میان کارهای انجام شده می‌توان به مقاله‌های [9,10] اشاره نمود.

در مقالهٔ [9]، به منظور کنترل بهتر لوازم خانگی هوشمند، از فاصله دور و همچنین تشخیص صدای گفتار انسان از صداهای موجود در طبیعت، یک سری فرمان‌های مربوط به کار با لوازم خانگی هوشمند از افراد مختلف در شرایط مختلف با استفاده از میکروفن‌هایی در مسافت‌های مختلف از لوازم خانگی هوشمند، جمع‌آوری شده‌اند. فرامین ضبط شده مطابق جدول (۱) می‌باشد.

جدول (۱): دادگان تهیه شده در مقالهٔ [9]

فرمان‌ها	روشن کن، خاموش کن، روشن شو، رد شو، افزایش دادن، کاهش دادن، بازی کردن، شروع کن، متوقف کن، مکث کن، ادامه دادن، بررسی کردن، خواندن، تغییردادن، حرکت بعدی، حرکت قبلی
اشیاء	لامپ، لامپ روشنایی، تلویزیون، ضبط صوت، ماشین لباس‌شویی، ساعت، هوا، تهویه کنندهٔ هوا، یخچال، یخچال فریزر، ماشین لباسشویی، ساعت، زمان‌سنج
موقعیت	آشپزخانه، اتاق خواب، اتاق پذیرایی، اتاق نشیمن، اتاق غذاخوری، حمام

در مقالهٔ [10]، یک سیستم برای کنترل لوازم خانهٔ هوشمند[۱] ارائه شده است که در آن کاربران می‌توانند از دستورات صوتی برای کنترل لوازم خانگی خود استفاده کنند. سیستم پیشنهادی می‌تواند دستورات صوتی را مستقل از ویژگی‌های گوینده مانند لهجه تشخیص دهد. بنابراین یک سیستم بسیار انعطاف پذیر است. فرمان‌هایی که برای ارزیابی این سیستم استفاده شده‌اند، شامل موارد موجود در جدول (۲) می‌باشد که از افراد مختلف در رده‌های سنی مختلف جمع‌آوری شده است.

جدول (۲): دادگان تهیه شده در مقالهٔ [10]

فرمان‌های انواع گروه‌ها				
گروه دسترسی	گروه روشنایی	گروه تهیه هوا	گروه سودمندی	گروه امنیت
باز کن بسته شو سکوت کن قرار بگیر	آشپزخاله آشپزخانه–خاموش پذیرایی پذیرایی–خاموش اتاق اتاق–خاموش	روشن خاموش	پرده کشیدن–پرده آبپاش خاموش شدن آبپاش	کمک سکوت

با توجه به مشاهدات و تحقیقات، یکی از چالش‌های اصلی در زمینهٔ فرامین صوتی آموزش سیستم‌های تشخیص فرامین صوتی بر اساس ساختار کلمات و بدون درکی از مفهوم کلمات (متضاد بودن کلمات، هم معنی بودن کلمات و ...) انجام می‌شود. یکی از راهکارهای پیشنهادی نویسندگان این مقاله برای حل این چالش، ترکیب بحث بازشناسی گفتار با بحث هستان‌شناسی می‌باشد. در این صورت سیستم تشخیص فرمان صوتی، کلمه را تشخیص داده و کلمه تشخیص داده شده با مفاهیم موجود در پایگاه دانش که به وسیلهٔ هستان شناسی ایجاد شده‌اند، مقایسه می‌شود و به دستگاه مربوطه ارسال می‌شود. در واقع با انجام این‌کار، سیستم مفهوم درخواست کاربر را درک می‌کند و متناسب با آن عمل می‌کند. چالش دیگری که در حوزهٔ جمع آوری دادگان می‌توان به آن اشاره کرد، وابسته بودن دادگان به لهجه یا گویش خاص می‌باشد، که برای رفع این چالش، از افراد با گویش‌ها، جنسیت‌ها و رده‌های سنی متنوع استفاده شده است. باتوجه به دسترس‌ناپذیری دادگان حاوی فرامین صوتی مبتنی بر گفتار فارسی در حوزهٔ کنترل لوازم خانگی هوشمند که علاوه بر بحث بازشناسی گفتار در بحث هستان‌شناسی هم مورد استفاده قرار بگیرد، بر آن شدیم که به طراحی و جمع‌آوری این دادگان بپردازیم. دراین مقاله به معرفی دادگان صوتی مذکور خواهیم پرداخت. ساختار این مقاله به صورت ذیل می‌باشد. در بخش دوم مقاله، روند جمع‌آوری دادگان و شرایط ضبط آن مورد بررسی قرار خواهد گرفت. سپس دربخش سوم، دادگان ارائه شده مورد ارزیابی قرار خواهد گرفت. جمع‌بندی مقاله در بخش چهارم ارائه خواهد شد.

۲- روند جمع‌آوری دادگان

در این بخش به طراحی و روند ضبط دادگان حاوی فرمان‌های صوتی مبتنی بر زبان فارسی برای کنترل لوازم خانگی هوشمند(PVC_SHA)[۲] خواهیم پرداخت.

۱-۲- شرایط ضبط دادگان

ضبط دادگان (PVC_SHA) تحت گوشی تلفن همراه با کمک نرم افـزاری به نام Voice Recorder (یک نرم‌افزار مبتنی بـر اندرویـد) صـورت گرفتـه است. برای هماهنگی بین داده‌های جمع‌آوری شده، صـدای افـراد بـا فرمـت Wav، نرخ بیت ۱۶ کیلوهرتز و به صورت تک بانده ضبط شده است.
دادگان (PVC_SHA) شامل فرامین صوتی مطـابق جـدول (۳) مـی‌باشـد. همانطور که در جدول مشخص است، ایـن دادگـان از ۳۲ کلمـه کلیـدی و ۶ کلمه غیر کلیدی و ۵۸ فرمان تشکیل شده است.

جدول (۳) : مشخصات مربوط به دادگان

فرمان‌ها	کلمات کلیدی	کلمات غیرکلیدی
تلویزیون! خاموش شو	یک	است
تلویزیون! قطع شو	دو	ایست
تلویزیون! کافیست	سه	شو
تلویزیون! بسته شو	چهار	کن
تلویزیون! شروع شو	پنج	رو
تلویزیون! باز شو	وصل	بده
تلویزیون! نمایش بده	بخون	
تلویزیون! پخش کن	صفحه	
تلویزیون! روشن شو	صدا	
تلویزیون! صدا رو کم کن	روشن	
تلویزیون! صدا رو معمولی کن	پخش	
تلویزیون! نور صفحه رو کم کن.	نمایش	
تلویزیون! نور صفحه رو زیاد کن	باز	
تلویزیون! نور صفحه رو معمولی کن	شروع	
تلویزیون! نور صفحه رو متوسط کن	بسته	
تلویزیون! صدا رو زیاد کن	متوقف	
تلویزیون! صدا رو متوسط کن	کافی	
تلویزیون! صدا ضعیف است.	قطع	
تلویزیون! صدا خیلی ضعیف است.	خاموش	
تلویزیون! صدا بلند است.	زیاد	
تلویزیون! صدا خیلی بلند است.	کم	
تلویزیون! متوقف شو ،	متوسط	
ضبط صوت! خاموش شو	معمولی	
ضبط صوت! قطع شو	نور	
ضبط صوت! کافی است	سیستم روشنایی	
ضبط صوت! متوقف شو	لامپ	
ضبط صوت! بسته شو	ضبط صوت	
ضبط صوت! شروع شو	تلویزیون	
ضبط صوت! بازشو	بلند	
ضبط صوت! پخش کن	خیلی	
ضبط صوت! روشن شو	تاریک	
ضبط صوت! بخون	ضعیف	
ضبط صوت! صدارو کم کن		
ضبط صوت! صدا رو معمولی کن		
ضبط صوت! صدا رو متوسط کن		
ضبط صوت! صدا رو زیاد کن		
ضبط صوت! صدا ضعیف است		
ضبط صوت! صدا خیلی ضعیف است		
ضبط صوت! صدا بلند است		
ضبط صوت! صدا خیلی بلند است		
سیستم روشنایی! روشن شو		
سیستم روشنایی! وصل شو		
سیستم روشنایی! قطع شو		
سیستم روشنایی! خاموش شو		
سیستم روشنایی! نور رو زیاد کن		
سیستم روشنایی! نور رو معمولی کن		
سیستم روشنایی! نور رو متوسط کن		
سیستم روشنایی! نور رو کم کن		
لامپ! روشن شو		
لامپ! وصل شو		
لامپ! قطع شو		
لامپ! خاموش شو		
لامپ! نور رو کم کن		
لامپ! نور رو زیاد کن		
لامپ! نور رو معمولی کن		
لامپ! نور رو متوسط کن		
لامپ! تاریک است		
لامپ! خیلی تاریک است		

این فرامین به دو بخش تقسیم شده است: فرامین مستقیم و فرامین غیرمستقیم. فرامین مستقیم، فرامینی هستند که کاربر، درخواست خود را به صورت واضح بیان می‌کند؛ مانند تلویزیون! خاموش شو. این بخش از فرمان‌ها در مدت زمان ۱ ساعت و ۴۲ دقیقه و ۳۰ ثانیه از ۳۰ گویشور، شامل ۱۵ نفر مرد و ۱۵ نفر زن با محدودهٔ سنی بین ۱۳ تا ۵۰ سال و با میزان تحصیلات مختلف تهیه شده است. تعداد گویشوران در محدودهٔ سنی ۱۰ تا ۲۰ سال، دو خانم و چهارآقا، در محدودهٔ سنی ۲۱ تا ۳۰، نه خانم و ۱۳ آقا، در محدودهٔ سنی ۳۱ تا ۴۰، یک خانم و در محدودهٔ سنی ۴۱ تا ۵۰، یک خانم بوده است. فرامین غیرمستقیم، فرامینی است که سیستم تشخیص فرمان، به کمک استدلالی که توسط هستان‌شناسی[۳] انجام می‌دهد، به نیاز کاربر پاسخ می‌دهد. مانند لامپ! اتاق تاریک است. این فرامین، در مدت زمان ۱۶ دقیقه و ۲۰ ثانیه از ۲۰ گویشور شامل ۱۰ نفر مرد و ۱۰ نفر زن با محدودهٔ سنی بین ۱۳ تا ۵۵ سال و با میزان تحصیلات مختلف تهیه شده است. تعداد گویشوران در محدودهٔ سنی ۱۰ تا ۲۰ سال، یک خانم، ۲۱ تا ۳۰ سال، هشت آقا و هفت خانم، ۳۱ تا ۴۰، یک خانم و در نهایت در محدودهٔ سنی ۴۱ تا ۵۵، یک آقا و یک خانم بوده است.

همانطورکه در جدول (۳) مشاهده می‌کنید، برای یک عمل مانند: خاموش شدن تلویزیون، از فرامین متنوعی استفاده شده است که هدف از این کار، استفاده از دادگان در بحث هستان شناسی برای رفع چالش بیان شده در قسمت مقدمه می‌باشد.

اطلاعات دقیق گویشوران بر حسب سن، جنسیت، میزان تحصیلات و لهجه در جدول (۴) و (۵) قرار داده شده است. اطلاعات جدول (۴) و(۵) بـر اسـاس جنسیت مرتب شده‌اند. شماره گویشوران عدد منحصر به فردی بـین ۱ تـا ۵۰ می‌باشد.

جدول (٤) : مشخصات مربوط به گویشوران فرمان‌های مستقیم

شمارهٔ گویشور	سن	جنسیت	لهجه	میزان تحصیلات
۱	۲۵	مرد	کاشانی	فوق دیپلم
۲	۲۴	مرد	اراکی	لیسانس
۳	۲۸	مرد	مازندرانی	فوق لیسانس
۴	۲۴	مرد	مازندرانی	لیسانس

۵	۲۵	مرد	سمنانی	لیسانس
۶	۲۶	مرد	تهرانی	فوق لیسانس
۷	۱۵	مرد	عربی	دانش‌آموز
۸	۲۶	مرد	تهرانی	لیسانس
۹	۲۴	مرد	تهرانی	لیسانس
۱۰	۲۰	مرد	کردی	فوق دیپلم
۱۱	۲۵	مرد	کردی	لیسانس
۱۲	۲۶	مرد	قمی	لیسانس
۱۳	۲۷	مرد	آذری	فوق لیسانس
۱۴	۲۵	مرد	آذری	لیسانس
۱۵	۲۲	مرد	کردی	فوق دیپلم
۱۶	۲۵	زن	آذری	لیسانس
۱۷	۳۱	زن	اراکی	لیسانس
۱۸	۲۹	زن	شیرازی	فوق لیسانس
۱۹	۲۴	زن	سمنانی	لیسانس
۲۰	۲۴	زن	عربی	لیسانس
۲۱	۳۰	زن	گلستانی	فوق لیسانس
۲۲	۱۹	زن	عربی	دیپلم
۲۳	۲۰	زن	قمی	دیپلم
۲۴	۲۱	زن	سمنانی	دیپلم
۲۵	۲۶	زن	خراسانی	لیسانس
۲۶	۲۶	زن	آذری	لیسانس
۲۷	۱۹	زن	آذری	دیپلم
۲۸	۱۳	زن	آذری	دانش‌آموز
۲۹	۲۲	زن	کردی	فوق دیپلم
۳۰	۵۰	زن	آذری	لیسانس

جدول (۵) : مشخصات مربوط به گویشوران فرمان‌های غیر مستقیم

شمارهٔ گویشور	سن	جنسیت	لهجه	میزان تحصیلات
۱	۲۲	مرد	کردی	فوق دیپلم
۲	۵۵	مرد	آذری	لیسانس
۳	۲۶	مرد	یزدی	لیسانس
۴	۲۴	مرد	آذری	لیسانس
۵	۲۱	مرد	تهرانی	فوق دیپلم
۶	۲۵	مرد	کردی	فوق لیسانس
۷	۳۶	مرد	کرمانی	فوق لیسانس
۸	۲۵	مرد	مازندرانی	لیسانس
۹	۲۸	مرد	مازندرانی	فوق لیسانس
۱۰	۲۷	مرد	گلستانی	لیسانس
۱۱	۲۴	زن	کرمانی	لیسانس
۱۲	۲۸	زن	شیرازی	فوق لیسانس
۱۳	۳۰	زن	یزدی	دکترا
۱۴	۳۱	زن	کردی	دکترا
۱۵	۱۳	زن	آذری	دانش‌آموز
۱۶	۵۰	زن	آذری	لیسانس
۱۷	۲۴	زن	مازندرانی	فوق لیسانس
۱۸	۲۳	زن	خراسانی	لیسانس
۱۹	۲۶	زن	لری	لیسانس
۲۰	۲۵	زن	آذری	لیسانس

هرفایل wav به صورت منحصر به فرد نامگذاری می‌شود. این نام مخفف عبارت ضبط شده در هر فایل (به عنوان مثال برای فرمان (Mediaplayer Mto (turnon) و Sp (مخفف Speaker) و شمارهٔ گویشور که عددی بین ۱ تا ۵۰ است، می‌باشد. به عنوان مثال Mtosp50 نام یکی از فایل‌های حاوی دستور (Mediaplayer turn on) می‌باشد.

۲-۲- ویرایش و برچسب‌گذاری

به منظور ویرایش دادگان از نرم افزار CoolEdit استفاده شده است. هر فایل ضبط شده در محیط نرم افزار، باز و نمایش داده می‌شود. چنانچه صداهایی از قبیل کلیک، سرفه، باز و بسته شدن در و... در بخش‌های سکوت فایل وجود داشته باشد، توسط نرم افزار حذف می‌شود. لازم به ذکر است که هنگام ضبط دادگان، از گویشوران خواسته شده بود در محیط تمیز و عاری از هرگونه نویزی، اقدام به ضبط صدا کنند. بنابراین می‌توان ادعا کرد که تمام فایل‌ها بدون هیچگونه نویز قابل اهمیتی جمع‌آوری شده‌اند.

برچسب گذاری دادگان در سطح کلمه و به صورت دستی با استفاده از نرم افزار CoolEdit انجام شده است. برای قسمت‌های سکوت از برچسب sil و برای کلمات غیر کلیدی از برچسب filler استفاده شده است. سایر کلمات برچسب کلمه کلیدی متناظر را خورده‌اند. بنابراین دادگان (PVC_SHA) حاوی ۲۷۳۷ فایل wav با طول متوسط ۴ ثانیه و ۲۷۳۷ فایل lab (برچسب متناظر با هر فایل) می‌باشد.

۳- ارزیابی دادگان

به منظور ارزیابی دادگان (PVC_SHA) از دو روش مبتنی بر مدل مخفی مارکف و حافظه کوتاه‌مدت ماندگار (LSTM)[۴] استفاده شده است. دادگان به دو بخش آزمون و آموزش تقسیم شده است. مجموعهٔ آموزش حاوی ۱۷۷۷ فایل و مجموعهٔ آزمون حاوی ۹۶۰ فایل می‌باشد، که این فایل‌ها به صورت تصادفی از بین گویشوران انتخاب شده است. بدیهی است که گویشوران در مجموعهٔ آموزش و آزمون مستقل از هم هستند. در ادامه ارزیابی دادگان مبتنی بر هر یک از دو رویکرد مذکور توضیح داده می‌شود.

۳-۱- ارزیابی دادگان مبتنی بر مدل مخفی مارکف

برای آموزش بازشناسی فرامین از جعبه ابزار (HTK) استفاده شده است. تعداد مدل‌های مخفی مارکف، ۳۴ مدل به ازای ۳۲ کلمهٔ کلیدی، یک مدل سکوت (sil) و یک مدل کلمهٔ غیرکلیدی (Filler) است. برای هر مدل وضعیت‌های مختلفی (۸، ۱۰، ۱۲، ۱۴، ۱۶، ۱۸) در نظر گرفته شده است. تعداد مخلوط‌های گوسی در هر وضعیت (از ۴ تا ۶۴ مخلوط گوسی) لحاظ شده و حالت بهینه بر روی این دادگان پس از انجام ارزیابی‌های مختلف مشخص شده است. ویژگی‌های استخراج شده از هر فایل، ضرایب مل کپسترام بوده و تعداد مطلوب این ویژگی‌ها پس از انجام ارزیابی‌های مختلف حاصل شده است.

به منظور انتخاب بهینهٔ تعداد ویژگی‌های مل کپسترام، به ازای هر مجموعه مدل مخفی مارکف، دو دسته ویژگی (۱۲ ضریب مل کپسترام و یک ضریب انرژی به همراه مشتقات اول و دومشان (در مجموع ۳۹ ضریب) و ۱۲ ضریب مل کپسترام و یک ضریب انرژی به همراه مشتقات اول و دوم و

سومشان(در مجموع ۵۲ ضریب)) استخراج شده اند. نتایج ارزیـابی در جـدول (۶) ارائه شده است.

لازم به ذکر است که ارزیابی ها بـر اسـاس سـه معیـار ارزیـابی دقـت[۵]، صحت[۶] و Word Error Rate (WER) انجام شده‌انـد کـه بـا اسـتفاده از روابط زیر محاسبه می‌شوند[4]:

$$Accuracy=(H-I)/N \quad (۱)$$

$$Correctness=H/N \quad (۲)$$

$$WER=D+S+I/N \quad (۳)$$

که H، معرف تعداد تشخیص‌های درست، N، تعداد کـل کلمـات قابـل شناسایی و I، S و D به ترتیب تعداد خطاهای درج، جایگزینی و حذف کلمـه می‌باشند.

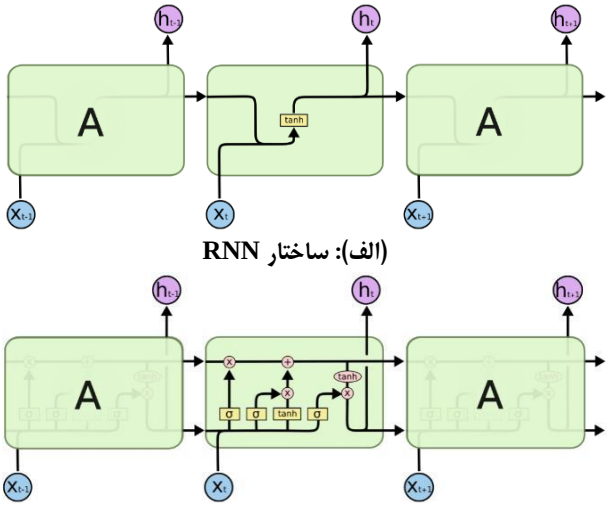
جدول (٦) : نتایج مربوط به ارزیابی

تعداد مخلوط گوسی	تعداد وضعیت	تعداد ضرایب ویژگی	دقت	صحت	٪WER
۱۶	۸	۳۹	٪۹۱.۲۳	٪۹۲.۵۴	٪۱۰.۱۳
۱۶	۱۰	۳۹	٪۸۹.۷۵	٪۹۱.۲۵	٪۱۰.۲۴
۱۶	۱۲	۳۹	٪۹۰.۸۸	٪۹۲.۰۷	٪۹.۱۲
۱۶	۱۴	۳۹	٪۹۰.۲۴	٪۹۱.۲۳	٪۹.۷۵
۱۶	۱۶	۳۹	٪۸۹.۸۷	٪۹۰.۸۴	٪۱۰.۱۲
۱۶	۱۸	۳۹	٪۸۹.۰۳	٪۹۰.۰۰	٪۱۰.۹۶
۴	۸	۳۹	٪۸۹.۶۹	٪۹۱.۶۶	٪۱۰.۳۰
۸	**۸**	**۳۹**	**٪۹۱.۴۱**	**٪۹۲.۸۳**	**٪۸.۵۸**
۳۲	۸	۳۹	٪۸۷.۷۰	٪۸۹.۲۲	٪۱۲.۲۹
۶۴	۸	۳۹	٪۸۱.۶۶	٪۸۳.۴۴	٪۱۸.۳۴
۸	۸	۵۲	٪۸۱.۲۷	٪۸۳.۶۲	٪۱۸.۷۳

با توجه به نتایج حاصل شده در جدول (۶)، بهترین دسـته مـدل و ویژگـی، ۸ مخلوط گوسی در هر وضعیت و ۳۹ ویژگی می‌باشد. این تنظـیم دارای دقـت حدود ۹۲٪ و صحت ۹۳٪ و WER حدود ۸ ٪ برروی دادگان آزمـون مـی-باشد. تاکید می‌شود که میـان دادگـان آزمـون هـیچ‌گونـه اشـتراکی از لحـاظ گویشوران وجود نداشته و بازشناسی فرامین صوتی کاملا مسـتقل از گویشـور عمل می‌کند.

۲-۳- مبتنی بر حافظهٔ کوتاه مدت ماندگار(LSTM)

شبکهٔ مبتنی بر حافظهٔ کوتاه مدت ماندگار(LSTM) یکی از انواع شبکه‌های بازگشتی (RNN) می‌باشد که توانایی یادگیری وابستگی‌های بلندمدت را دارد. شبکهٔ LSTM، برای اولین بار توسط هاکریتر و اشمیدر در سال ۱۹۹۷ معرفی شد[10]. درحقیقت هدف از طراحی شبکه LSTM، حل مشکل به یادسپاری وابستگی‌های بلند مدت در شبکه‌های بازگشتیRNN بود. شبکه‌های بازگشتی RNN ساختاری بسیار مشابه با شبکه‌های چندلایهٔ پرسپترون دارند با این تفاوت که نورون‌های لایهٔ مخفی، علاوه بر یال‌های رو به جلو، یک یال هم به صورت بازگشتی و با احتساب یک زمان تاخیر، از خودشان دارند. چنین ساختاری به یاد سپاری وابستگی‌های کوتاه مدت (Short Term) را تضمین می‌کند. لیکن امکان یادگیری وابستگی‌های مربوط به گذشته‌های دور (Long) را ندارد. برای رفع این مشکل نورون‌های مخفی با یک بلاک حافظه با ساختار پیچیده‌تری، جایگزین شده و منجر به ظهور شبکه‌های مبتنی بر حافظهٔ کوتاه مدت ماندگار یا همان LSTM گردید. شکل (۱) تفاوت ساختاری شبکه‌های RNN و LSTM را نشان می‌دهد.



(الف): ساختار RNN

(ب): ساختار LSTM

شکل(۱): تفاوت ساختار شبکه‌ها [9]

چنانچه شکل (۱) نشان می‌دهد، در مورد شبکه بازگشتی RNN (الف)، ماژول‌های تکرار شونده ساختار بسیار ساده‌ای تنها شامل یک لایه تانژانت هایپربولیک دارند. ورودی هر شبکه، شامل خروجی ماژول قبلی و ورودی جدید می‌باشد. خروجی هر ماژول حاصل اعمال تابع تانژانت هایپربولیک (یا هر تابع فعالیت دیگری) بر ترکیب وزن‌دار دو رودی مذکور می‌باشد. درشبکهٔ بازگشتی LSTM (ب)، ساختار ماژول تکرار شونده به منظور به خاطرسپاری وابستگی‌های طولانی مدت، پیچیدگی بیشتری دارد. ورودی هر ماژول شامل ورودی جدید و دو خروجی از ماژول قبلی است. همچنین هر ماژول دارای دو خروجی می‌باشد. برای تولید هر یک از این دو خروجی از اعمال توابع فعالیت سیگموئید و تانژانت هایپربولیک (یا هر تابع فعالیت دیگری) بر خروجی‌های ماژول قبلی و ورودی جدید و ترکیب‌های وزن‌دار آن‌ها استفاده شده است. جزئیات مربوط به این بلاک حافظه، اجزای تشکیل دهنده و روابط مربوطه‌اش در [11] ارائه شده اند.

نتایج ارزیابی مربوط به روش LSTM در جدول (۷) نوشته شده است.

جدول (۷) : نتایج مربوط به ارزیابی دادگان با LSTM

دقت	تعدادواحد درهر لایهٔ مخفی[۷]	بیشترین دوره[۸]	اندازه کوچک‌ترین بسته[۹]
٪۹۶.۴۱	۳۰۰	۳۰۰	۲۷
٪۹۷.۴۵	۵۰۰	۵۰۰	۳۰
٪۹۷.۱۸	۱۰۰۰	۱۰۰۰	۳۰
٪۹۷.۲۹	۱۵۰۰	۱۵۰۰	۲۰
٪۹۸.۰۵	**۲۰۰۰**	**۱۲۰۰**	**۵**

با توجه به نتایج حاصل شده در جدول (۷)، بهترین نتیجه، ۲۰۰۰ واحد در هرلایهٔ مخفی و ۱۲۰۰ دوره می‌باشد. این تنظیمات دارای دقت حدود ۹۸٪ ، بر بروی دادگان آزمون می‌باشد. همانطور که در بالا اشاره شد، میان دادگان آزمون هیچ گونه اشتراکی از لحاظ گویشوران وجود نداشته و بازشناسی فرامین صوتی، کاملا مستقل از گویشور عمل می‌کند.

۴- نتیجه گیری

تعامل انسان و دستگاه (HDI[10]) شامل ترجمه دستورات انسان به دستورات کنترلی دستگاه است. علاوه براین، تعامل یک ارتباط تک جهته نیست، یعنی علاوه براینکه دستگاه باید منظور انسان را درک کند، انسان هم باید عملکرد دستگاه را متوجه شود. یک مسئله مهم در HDI این است که مردم به جای کارکردن با رابطهای سنتی، با استفاده از دستورات خاصی راحت بتوانند اقدامات خود را انجام دهند. برای مثال از طریق یک دستیارصوتی. با اینکه این‌کار باعث رفاه کاربران می‌شود، اما برای کسانی که با این روش آشنایی ندارند، بسیار آزاردهنده است. موضوعی که از تحقیق‌ها، بررسی‌ها و مشاهدات به دست آمده این است که، افراد برای دستور دادن به دستگاه‌ها از کلمات و عبارات کلیدی خاصی استفاده می‌کنند. به این ترتیب کلمات کلیدی که توسعه دهندگان برای توصیف یک عمل در یک دستگاه استفاده می‌کنند، لزوماً با کلمه‌ای که افراد آشنا به آن وسیله استفاده می‌کنند یکسان نیست و این یکسان نبودن باعث نارضایتی کاربران می‌شود.

برای حل مشکلات مذکور، در این مقاله یک مجموعه دادگان حاوی فرامین صوتی برای کنترل لوازم خانگی هوشمند (لامپ، تلویزیون، ضبط صوت)، ضبط و جمع آوری شده است. این دادگان در کار آتی نویسندگان این مقاله در ترکیب هستان شناسی با تشخیص فرامین صوتی، به منظور استفاده از مفهوم کلمه به جای ساختار کلمه با هدف افزایش دقت بازشناسی فرامین صوتی مورد استفاده قرار خواهد گرفت.

چنانچه نتایج ارزیابی دادگان ارائه شده با مدل مخفی مـارکف در ایـن مقالـه نشان می‌دهند، کارایی کلمه آموزش یافته بر روی دادگان در بهترین حالـت از دقت ۹۲ درصد و صحت ۹۳ درصد و ۸ درصد خطای در سطح کلمه برخوردار است.

همچنین نتیجه ارزیابی دادگان ارائه شده مبتنی بر حافظه کوتاه مدت مانـدگار با بهترین شبکه حدودا دارای دقت ۹۸٪ می‌باشد.

مراجع

[۱] میرحسینی آرانی, سیدمصطفی و مهدی شعبانی، **"طراحی رابط صوتی تشخیص فرامین صوتی جهت کنترل سیستم های صنعتی"**، ششمین کنفرانس ملی علوم و مهندسی کامپیوتر و فناوری اطلاعات، موسسه علمی تحقیقاتی کومه علم آوران دانش، بابل،۱۳۹۸.

[2] Mehrabani, M., Bangalore, S., & Stern, B.. "*Personalized speech recognition for Internet of Things.*" 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT). IEEE, 2015.

[3] Azargoshasb S., Korayem A. H., Tabibian SH. "*A Voice Command Detection system for controlling Movement of SCOUT Robot.*" 2018 6th RSI International Conference on Robotics and Mechatronics (IcRoM). IEEE, 2018.

[4] Tabibian SH.,"*A voice command detection system for aerospace applications.*" International Journal of Speech Technology 20.4 (2017): 1049-1061.

[5] Brenon, A., Portet, F., & Vacher, M. . *"Preliminary study of adaptive decision-making system for vocal command in smart home.*" 2016 12th International Conference on Intelligent Environments (IE). IEEE, 2016.

[6] Rubio-Drosdov, E., Díaz-Sánchez, D., Almenárez, F., Arias-Cabarcos, P., & Marín, A. "*Seamless human-device interaction in the internet of things.*" IEEE Transactions on Consumer Electronics 63.4 (2017): 490-498.

[7] Pakpahan, R. L. H., Sudiharto, D. W., & Satwiko, A. G. P. "*The prototype of automated doors and windows by using voice commands.*" 2016 International Seminar on Application for Technology of Information and Communication (ISemantic). IEEE, 2016.

[8] Mittal, Y., Toshniwal, P., Sharma, S., Singhal, D., Gupta, R., & Mittal, V. K. (2015, December). "*A voice-controlled multi-functional Smart Home Automation System.*" In 2015 Annual IEEE India Conference (INDICON) (pp. 1-6). IEEE.

[9] Han, Y., Hyun, J., Jeong, T., Yoo, J. H., & Hong, J. W. K ."*A smart home control system based on context and human speech.*" 2016 18th International Conference on Advanced Communication Technology (ICACT). IEEE, 2016.Zachman, John A., "A Framework for Information Systems Architecture", IBM Systems Journal, Vol. 26, No. 3, 1987.

[10] Arriany, A. A., & Musbah, M. S. "*Applying voice recognition technology for smart home networks.*" 2016 International Conference on Engineering & MIS (ICEMIS). IEEE, 2016.

[11] Hochreiter, S., & Schmidhuber, J. "*Long short-term memory.*" Neural computation 9.8 (1997): 1735-1780.

پانویس ها

[1] Smart Home Automation System

[2] Persian Voice Commands for Controlling Smart Home Application

[3] Ontology

[4] Long Short Time Memory

[5] Accuracy

[6] Correctnees

[7] Number OF Hidden Unit

[8] MaxEpochs

[9] Mini Bach Size

[10] Human Device intraction