نیروی XXI زمین جنگجو :

رویکرد سیستم به تشخیص گفتار

خلاصه

تشخیص "گفتار" به عنوان خط اتصال کاربر در برنامه های جدید به طور پیوسته دیده می شود. با پیشرفت این تکنولوژی، روشهای جدیدی برای انسان فراهم می شود تا با ماشین ها و اطلاعات حاصل از آنها، ارتباط برقرار کنند. انجام این کار در بسیاری از حوزه‌ها، کاربران را به انتظارات خود نزدیکتر کرده است. اگر چه هنوز چالش های زیادی در مورد فناوری وجود دارد، ولی تشخیص گفتار به یک سطح تکامل یافته نیاز به در نظر گرفتن استقرار آن در سیستم ها و محیط های پیچیده دارد. به این ترتیب ما درباره سیستمهایی بحث می کنیم که ما را به اجرای موفقیت آمیز تشخیص گفتار برنامه نیروی XXI زمین جنگجو نزدیک می کند.

ما در مورد اجزای کنترل صدای سیستم در این برنامه بحث می کنیم. نیازمندی های مرتبط با قابلیت اطمینان، تشخیص و مسائل پیچیده محاسباتی به طور کامل در این بخش مورد توجه قرار گرفته است. ما به طور صریح جنبه های مختلف سیستم و نحوه تاثیر آنها بر روی خط اتصال کاربر و آشکار شدن پارامترهای کاربرد واقعی را پوشش می دهیم. در نهایت، اجرای یک طبقه بندی مبتنی بر چند جمله ای را برای تشخیص گفتار در نظر می گیریم و ارزیابی عملکرد نهایی سیستم را در حوزه پایگاه اطلاعاتی خاص، ارائه می دهیم.

1. مقدمه

برنامه توسعه ساختار مهندسی زمین جنگجو (EMD)، برنامه انقلابی ارتش برای توسعه و ایجاد یک سیستم مبارزه با سربازان کاملاً متحد است. این سیستم از فناوری های پیشرفته برای ارائه اثربخشی بی نظیر با ارائه قابلیت بهبود شناسایی، به دست آوردن، قرار دادن و درگیر کردن اهداف در محدوده های بیشتر در طول روز یا شب استفاده می کند. این سیستم یک سرباز منفرد را به میدان جنگ دیجیتالی می فرستد تا باعث بهبود ارتباطات بهتر و آگاهی موقعیتی شود.

هدف از برنامه نیروی جنگجویان زمینی XXI، این است که باعث سرعت ارتقای میدان فناوری پیشرفته به برنامه EMD مبارزه زمینی شود. این برنامه، تضمین کننده مزیت تکنولوژی جهانی برای سیستم های مبارزه جنگجویان پیاده است.

سیستم اجزای کنترل صدای سیستم (SVC)، نیروی جنگنده زمینی XXI یک رابط سخنران را برای سرباز کامپیوتری موجود از برنامه EMD Land Warrior فراهم می کند. هدف این است که سرباز پیاده را با یک روش کارآمد از کنترل های دست بسته و چشم بسته در سیستم، هدایت کند. شکل 1 مفهوم SVC را نشان می دهد.

استفاده از تشخیص گفتار در محیط مبارزه، منجر به استخراج نیازمندی های چالش برانگیز عملکرد می شود. شناخت و واژگان خروجی (OOV) باید باعث حفظ سطوح عملکرد در شرایط ناخوشایندِ صدای مخالف شود. علاوه بر این، سیستم باید به طیف گسترده ای از سطوح صدای صوتی (سطوح کم برای عملیات مخفی و سطوح بالا برای موقعیت های پر سر و صدا) پاسخ دهد. استرس صدا، نگرانی دیگر موجود برای اجرای برنامه است، آنچنان که اغلب با اثر لومپارد (1) مواجه می باشد. در نهایت، الگوریتمها باید از لحاظ محاسباتی، دارای کارایی مناسبی باشند تا باتری سیستم را تخلیه نکند، و مدل های کلمه ای، باید به اندازه کافی کوچک باشند تا بتوانند در حافظه موجود قرار گیرند.

طراحی سیستم کلی و خط اتصال کاربر در بخش 2 مورد بحث قرار گرفته است. ساختار الگوریتم تشخیص میدان واقعی گفتار، که بر اساس یک طبقه بندی چندجمله ای است، در بخش 3 توضیح داده شده است. در بخش 4 و 5،

راه حل مشکلات فنی استحکام سر و صدا، سخنرانی استرس زا و غیرقابل انکار واژگان مورد بحث قرار گرفته است. در نهایت، اعتبارسنجی و نتایج سیستم نهایی در بخش 6 ارائه شده است.

2. طراحی سیستم و رابط کاربر

نمودار بلوکی اجزای پایه SVC در شکل 2 نشان داده شده است. سرباز، یک دکمه را روی سلاح خود فشار می دهد تا سیستم شروع به شناسایی کند. یک میکروفون لغو سر و صدای نزدیک به صحبت کردن، فرمان سخن گفتن توسط A / D دیجیتالی را ضبط می کند. فرایندهای شناسایی کننده گفتار کاربر و پاسخ مناسب او، توسط سرباز کامپیوتری اجرا می شود.

استقرار موفق در تشخیص گفتار، به رویکرد سیستم ها به طراحی و اجرای SVC نسبت داده می شود. هر جزء، از رابط کاربر تا طبقه بندی پایانی برای همکاری با یکدیگر برای به حداکثر رساندن عملکرد این برنامه و برای کاربران هدف، طراحی شده است.

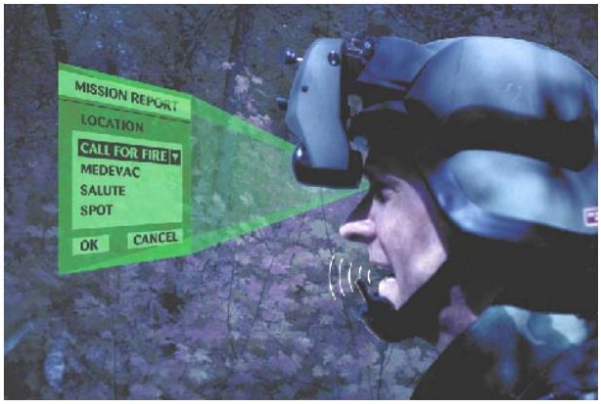
درک محیط کاربر و صحت استفاده واقعی، از طریق مطالعات گسترده تجربی کاربر و مشارکت واقعی در تمرینات زنده (و دیگران) به دست می آید. بر اساس داده های جمع آوری شده، می توان جزئیاتی از رابط کاربر و نیز شرایط محیطی مورد انتظار که در آن الگوریتم های تشخیص گفتار باید عمل کنند، استخراج نمود.

یکی از مهمترین مسائل مربوط به رابط کاربر، تمرکز بر روش شروع موتور تشخیص گفتار است. دو معیار وجود دارد. سرباز باید محاصره نشده باشد و دقت سیستم باید انتظارات او را برآورده کند. یک سیستم مبتنی بر کلمات کلیدی، باعث عملیات کاملاً آزاد می شود، اما دارای دو نقص مشخص می باشد. اول اینکه الگوریتم کلمه کلیدی باید به طور مداوم اجرا شود، که باعث تخلیه باتری می شود. دوم اینکه، چنین سیستمی به خطاهای درج و یا تشخیص فرمان دروغ حساس است. برای حل این مسائل، یک خط اتصال "فشار دادن برای صحبت کردن" مورد نیاز است.

برای خط اتصال "فشار دادن برای صحبت کردن"، محل دکمه به توانایی سربازان بستگی دارد تا به راحتی تشخیص داده و شروع به کار کند. چندین گزینه پیشنهادی شامل افزودن یک دکمه به یک سلاح پیشرفته وجود دارد. به وضوح

نشان داده شده است که با استفاده از ترکیب قرارگیری دکمه و استفاده از تشخیص گفتار به عنوان رابط اولیه، سرباز می تواند کارآیی خود را افزایش دهد. شکل 3 یک سرباز با استفاده از رابط گفتار (سمت چپ) را با یک سرباز با استفاده از یک دستگاه اشاره گر سنتی (راست) مقایسه می کند. سرباز با استفاده از SVC، هر دو دست خود را بر روی سلاح نگه می دارد که باعث کاهش تمرکز سرباز می شود، در صورتی که در روش دیگر، با دستکاری موس، سرباز با سلاح خود به زمین اشاره می کند.

علاوه بر اهمیت رابط کاربر، محدودیت های مصرف انرژی و استحکام نیز بر روی انتخاب الگوریتم های تشخیص گفتار مزبور تاثیر می گذارد (بخش 3). به منظور ارائه کارایی مورد نیاز، نرم افزار به طور کلی با سیستم کلی معیارهای طراحی خاص یکپارچه می شود تا به محیط های شلوغ ناسازگار و سخنرانی استرس زا (بخش 4)، و همچنین رد صحیح سر و صدا OOV (بخش 5) رسیدگی کند.



شکل 1) مفهوم کنترل صدای سیستم

3. ساختار طبقه بندی شده

در حال حاضر روش های طبقه بندی زیادی برای حل مشکل تشخیص گفتار مورد استفاده قرار می گیرد. به طور سنتی، از روشهای آماری برای مدل سازی نطق سخنرانان استفاده می شود؛ محبوب ترین رویکرد، مدل مخفی مارکوف (HMM) است. اخیراً، روشهای طبقه بندی مختلفی برای این مشکل اعمال شده است. به منظور ارائه بهترین عملکرد

برای سیستم های تشخیص گفتار، این عملکردها شامل داده های خارج از کلاس در مرحله آموزش هستند. برای SVC، یک رویکرد بر اساس طبقه بندی چند جمله ای اجرا می شود.

ساختار اصلی طبقه بندی ما در شکل 4 نشان داده شده است. بردارهای ویژگی xM ... x1، به یک تابع مجزا وارد می شوند ((wtp (x)، و سپس خروجی به صورت میانگین همه Mها برای تولید یک امتیاز، محاسبه می گردد. این استراتژی شبیه به رویکردهای روابط ترکیبی است که در آن شبکه های عصبی مصنوعی به عنوان ژنراتورهای احتمالی، مورد استفاده قرار می گیرند (2).

طبقه بندی الگوی ما از یک تابع تشخیص چندجمله ای استفاده می کند:

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^t \mathbf{p}(\mathbf{x}) . \qquad (1)$$

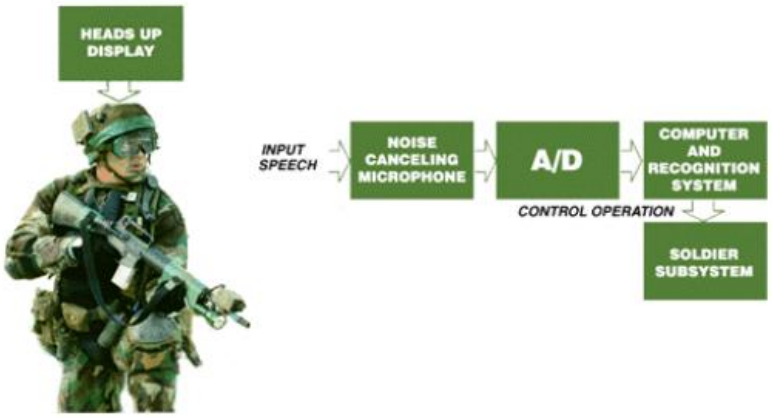
تابع تشخیصی، متشکل از دو بخش است. بخش اول مدل (w)، برای رده های مخصوص است. بخش دوم ((p(x)، یک بردار مبتنی بر چند جمله ای است که از ویژگی ورودی بردار x ساخته شده است. این بردار، مبتنی بر شرایط تک جمله ای تا درجه K از ویژگی های ورودی است. بنابراین، خروجی تابع جدا شده، یک ترکیب خطی از عناصر پایه چند جمله ای است.

امتیاز دادن به کارایی محاسبات برای تشخیص سخنران، مهم می باشد. بار سیستم سربازان، توسط پیچیدگی ارزیابی تابع مجزا تعیین می شود. از آنجایی که w وابسته به شاخص قاب نمی باشد، امتیاز دهی را می توان به صورت ساده به صورت زیر نوشت:

$$s = \mathbf{w}^t \frac{1}{M} \sum_{k=1}^{M} \mathbf{p}(\mathbf{x}_k) = \mathbf{w}^t \bar{\mathbf{p}} . \qquad (2)$$

تنها یک بردار منفرد، نشان دهنده گفتار ورودی محاسبه شده است، و هر ارزیابی نمره، با محاسبه یک محصول درونی برابر می باشد. تعداد عملیات نقطه شناور (FLOPS) شامل 2Nmodel-1 است که در آن، Nmodel، طول w می باشد.

بنابراین، برای 15 ویژگی و گسترش چند جمله ای (k=3)، W، طول معادل 816 است که تنها منجر به 1631 FLOPS در هر امتیاز کلمه، و یک اندازه مدل 3264 بایت برای نمایش نقطه شناور می شود. شرح مفصل این روابط در (3) توضیح داده شده است.



شکل 2) دیاگرام بلوکی کنترل صدای سیستم

شکل 3) آزمایش خط اتصال کاربر

4. سخنرانی پر صدا و استرس زا

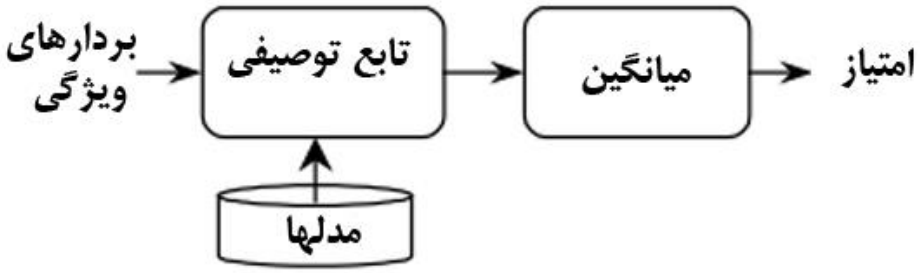
برای اطمینان از اینکه SVC به طور قابل اعتماد برای سرباز پیاده انجام می شود، داده های مربوط به سر و صدای خاص از تمرینات آتش سوزی طبیعی مورد ارزیابی قرار می گیرد. منابع قابل توجه سر و صدا عبارتند از آتش وسیله نقلیه و اسلحه. تجزیه و تحلیل این داده ها (و اطلاعات سر و صدا مشابه از پایگاه داده NOISEX-92)، باعث درک

ویژگی های طیفی خاصی قابل استفاده در سناریوهای واقعی می شود. صدای خودرو معمولی به صورت نیمه ثابت با طیف عبور کم مشخص می شود. آتش سلاح ها، عمدتاً به صورت پرتاب کننده است.

چندین آزمایش صوتی با چندین میکروفن انجام می شود تا بتوانند پاسخ کلی سیستم را مشخص کنند. در این روش از یک مبدل واکنش، به عنوان پایه استفاده می شود و دو میکروفون، سر و صدا را مورد آزمایش قرار می دهد. اولین آن "آندریا" است که ما در آن مورد تجربه داریم؛ یکی دیگر، عنصر "Electret Gentex" است که میکروفون ارتباطی مورد استفاده در سیستم Land Drive EMD می باشد.

نمونه هایی که از سر و صدای وسیله نقلیه و تفنگ ساچمه ای با استفاده از بلندگو پخش و ضبط می شود، هر دو با خود سرباز و با دستورات فردی با او صحبت می کنند. تجزیه و تحلیل سخنرانی های ضبط شده باعث ایجاد ویژگی های طیفی پاسخ های نزدیک به میدان و دور از میکروفون می شود. از این داده ها، یک فیلتر انفجار محدود (IFIR) یکپارچه شده برای از بین بردن فرکانس های پایین غالب در منابع صوتی طراحی شده است.

نگرانی اصلی برای سخنرانی استرس زا، اثر لومپارد است (1) که مشخصه غالب آن، یک جابجایی در طیف (مشابه با عدم انطباق کانال) است. برای جبران این ناهنجاری، از میانگین تفریق cepstral (CMS) و نرمال سازی انحراف استفاده می شود. نتایج پایگاه داده SUSAS (4) در جدول 1 نشان داده شده است.



شکل 4) ساختار طبقه بندی

جدول 1) عملکرد تشخیص استرس زا برای پایگاه داده SUSAS (Lom: لومبارد، Ques: پرسش)

شرایط	حوزه سخنرانی استرس زا (مستقل از سخنران)									
	آرام	سریع	نرم	با صدای بلند	خشمگین	شفاف	مسابقه	C50	C70	Lom
بدون جبران خسارت	٪77	٪79	٪77	٪61	٪56	٪80	٪78	٪81	٪78	٪69
CMS و Var. Norm	٪81	٪81	٪83	٪74	٪70	٪86	٪82	٪87	٪87	٪81

5. رد کردن خارج از واژه

رویکرد رد OOV باید بدون اینکه به طور قابل توجهی باعث کاهش عملکرد تشخیص شود، به خوبی در شرایط پر سر و صدا انجام شود. علاوه بر این، "حافظه" پیچیدگی محاسباتی و مسائل مربوط به زمان تاخیر، محل نگرانی می باشد. الگوریتم رد کردن خارج از رده در جدول 2 ارائه شده است. ایده این است که فضای ویژگی با تعداد زیادی از مدل های خارج از رده (باطله) پر شود. این عملکرد تمایل دارد که توزیع اطراف مدل های رده را تسریع کرده و مناطق خارج از رده را بین مناطق درون رده مشخص کند. این موضوع مشابه با معیارهای رد ابهام است، اما به توزیع آستانه ها وابسته نیست.

از یک "چشم پوشی از رتبه" برای اعطای تغییرات ویژگی های ورودی در شرایط ناسازگار استفاده می شود. این اغماض رتبه به آمارهای متغیر بردارهای ویژگی درون رده، بدون رد کردن تولید اشتباه، اجازه می دهد تا در سیستم وارد شود. برای SVC، این معیار رد کردن، نیاز به حافظه اضافی برای ذخیره سازی مدل های باطله ندارد، و تنها هزینه محاسباتی حداقل، برای تولید نمرات باطله لازم است.این برنامه به زیر مجموعه ها (منوها) تقسیم شده است. برای هر زیرمجموعه، واژگان غیرفعال به عنوان رده های باطله استفاده می شود. این ساختار، نیاز به حافظه اضافی برای مدل های باطله مشخص را از بین می برد. از آنجائی که امتیاز گیری تابع تشخیصی به سادگی یک محصول درونی است، هزینه محاسبات اضافه شده امتیازدهی تمام مدل ها برای هر کلمه ورودی در مقایسه با استخراج ویژگی اولیه، ناچیز است. این تکنیک، به طور گسترده مورد آزمایش قرار گرفته و عملکرد آنها نشان می دهد که اساساً قوی تر از روش های آستانه استاندارد می باشد. برای توصیف کامل الگوریتم و مشخصات عملکرد آن (5) را ببینید.

جدول 2) الگوریتم رد واژگان خارج شده

1	تولید نمرات برای همه مدل های در رده و باطله
2	مرتب سازی نمرات به طوری که بالاترین امتیاز، یک است (رتبه = 1)
3	پیدا کردن بالاترین نمره مدل در کلاس، * w
4	از همگرایی این مدل (* w) از فهرست نمره مرتب شده حذف شود
5	رتبه کلی، r برای این مدل (* w) را در لیست نمره مرتب شده بدست آورید
6	اگر این رتبه، r، کمتر یا برابر تحمل رتبه R است، پس خروجی * w به عنوان رده شناخته شده است؛ در غیر این صورت ورودی را رد کنید

6. اعتبارسنجی

ما آزمایش های خود را بر روی پایگاه کلمات خاص جداگانه دامنه انجام می دهیم. این موضوع، شامل عبارات امری کوتاه است که برای کنترل سرباز کامپیوتری مورد استفاده قرار می گیرند. در مجموع 131 عبارت واژه منحصر به فرد وجود دارد که به 27 زیر مجموعه (محدوده) تقسیم می شوند که در حدود 2 تا 35 قرار دارند. داده های اطلاعاتی پایگاه، با استفاده از میکروفون Gentex با نرخ نمونه برداری تک جمله ای 11.025 کیلو هرتز و 16 بیتی جمع آوری می شود. مجموعه آموزشی شامل 100 سرباز مرد است که هر کدام فقط یک بار با یک دستور، کلمات را تکرار می کنند. مجموعه آزمایشی شامل 66 سرباز مرد است که دستورات مشابه را یک بار تکرار می کنند. سخنرانان در مجموعه های آموزشی و آزمایشی متمایز بوده، و دارای طیف گسترده ای از لهجه ها هستند.

آزمایش ها در چهار محیط انجام می شود. محیط آرام، که دارای شرایط سازگار است (یعنی ضبط گفتار اصلی به هیچ وجه اصلاح نشده است). سه محیط دیگر، شرایط معمولی پر سر و صدا را نشان می دهند. آنها با مخلوط کردن سر و صدا از پایگاه داده NOISEX-92 تنها با نطق سخنران آزمایش ساخته می شوند؛ نطق سخنران آموزشی بدون تغییر باقی می ماند. برای ایجاد سر و صدا از پلنگ، m109 و منابع تفنگ استفاده می شود. نسبت های سیگنال به صدا در عبارات آزمون، در جدول 3 آمده است.

سیستم چند جمله ای سه بعدی، اجرا شده است. ما این سخنرانی را با استفاده از فریم های 30 میلی ثانیه مورد تجزیه و تحلیل قرار می دهیم. حذف میانگین به دنبال فیلتر IFIR ادامه می یابد و یک پنجره Hamming با 100 فریم

در ثانیه انجام می شود. ما از تجزیه و تحلیل دوازدهمین دستور LP استفاده می کنیم و سپس، 14 پارامتر cepstral را تولید می کنیم. Endpointing با استفاده از انرژی فریم انجام می شود؛ بردارهای ویژگی مربوط به فریم های غیر سخنران، کنار گذاشته می شود. نرمال سازی حذف میانگین cepstral و واریانس، برای جبران ویژگی های سخنرانی استرس زا اعمال می شود. علاوه بر این، یک ویژگی زمان نرمال شده، i / Nframes برای مجموع 15 ویژگی اضافه شده است. در نتیجه، بردارهای حاصل، به سیستم وارد می شوند.

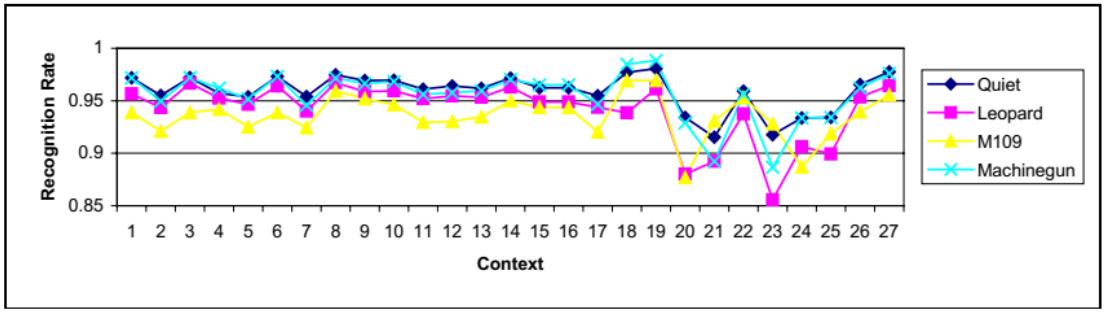
عملکرد هر یک از 27 موضوع، مورد ارزیابی قرار می گیرد. دقت تشخیص (با رد OOV فعال شده است) در شکل 5 نشان داده شده است. موضوع جالب این حقیقت است که این عملکرد به طور قابل توجهی باعث کاهش سر و صدای حاضر می شود.

پایگاه داده های خارج از واژگان شامل 118 عبارت (متمایز از عبارات در لغت نامه) و 3 ورودی جعلی (نفس، لب-خستگی و سرفه) است. تمام سخنرانی ها به شیوه ای مشابه آنچه که در بالا شرح داده شد، جمع آوری گردید.

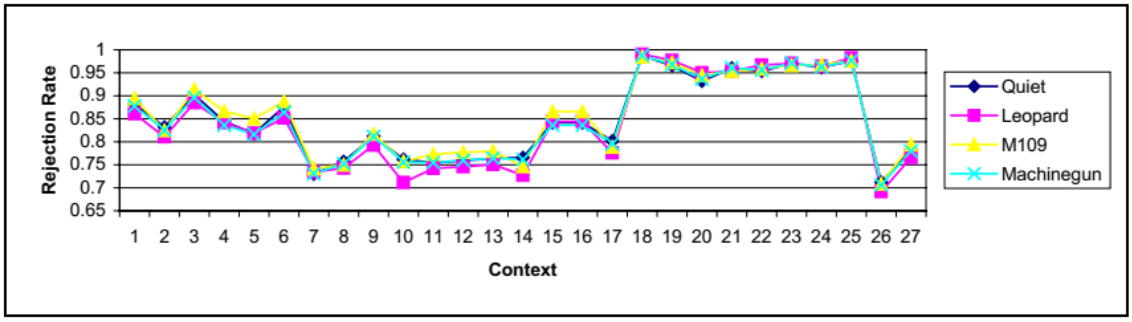
باز هم، عملکرد برای هر یک از 27 زمینه مورد بررسی قرار می گیرد، (شکل 6)، و ما می بینیم که رد OOV در شرایط پر سر و صدا، قوی است.

جدول 3) میانگین SNRs برای عبارت های آزمایشی

آرام	dB 33
پلنگ	dB 14
M109	dB 13
تفنگ	dB 28



شکل 5) عملکرد تشخیص سخنرانی برای همه موضوعات.



شکل 6) عملکرد رد واژگان خارجی برای همه موضوعات

7. نتیجه گیری

اجرای موفق رابط های تشخیص گفتار از طریق رویکرد سیستم ها به دست می آید. مهمتر از همه این است که کاربران برای اولین بار در فرآیند طراحی می توانند انتظارات خود و پارامترهای استفاده از سیستم را به طور کامل ضبط کنند. این موضوع باید به عنوان تمرین استاندارد برای حل کردن خصوصی در حوزه ها و محیط های جدید در نظر گرفته شود.

REFERENCES

[1] J. C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” in J. Acoust. Soc. Am., vol. 1, pp. 510-524, 1993.

[2] H. A. Bourlard and N. Morgan, Connectionist Speech Recognition: A Hybrid Approach. Kluwer Academic Publishers, 1994.

[3] W. M. Campbell, K. T. Assaleh, and C. C. Broun, “LowComplexity Small-Vocabulary Speech Recognition for Portable Devices,” in Proceedings of the Fifth International Conference on Signal Processing and its Applications, pp. 619-622, 1999.

[4] J. H. L. Hansen and S. E. Bou-Ghazale, “Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database,” in Proceedings of Eurospeech, vol. 4, pp. 1743- 1746, 1997.

[5] C. C. Broun and W. M. Campbell, “Robust Out-ofVocabulary Rejection for Low-Complexity Speaker Independent Speech Recognition,” in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, vol. 3, pp. 1811-1814, 2000.