بازشناسی مقاوم گفتار با استفاده از شبکه‌های عصبی حافظه کوتاه‌مدت ماندگار و ویژگی‌های گلوگاه

امین معاون‌جولا[۱]، دانشجوی دکتری؛ احمد اکبری[۲]، دانشیار؛ بابک ناصرشریف[۳]، استادیار

۱- دانشکده مهندسی کامپیوتر - دانشگاه علم و صنعت - تهران - ایران- joula@comp.iust.ac.ir

۲- دانشکده مهندسی کامپیوتر - دانشگاه علم و صنعت - تهران - ایران- akbari@iust.ac.ir

۳- دانشکده مهندسی کامپیوتر - دانشگاه صنعتی خواجه‌نصیرالدین طوسی - تهران - ایران- bnasersharif@kntu.ac.ir

چکیده: شبکه‌های عصبی عمیق در سال‌های اخیر به طرز گسترده‌ای در سیستم‌های بازشناسی گفتار مورداستفاده قرارگرفته‌اند. بااین‌وجود، مقاوم‌سازی این مدل‌ها در حضور نویز محیط کمتر موردبررسی قرارگرفته است. در این مقاله دو راهکار برای مقاوم‌سازی مدل‌های شبکه حافظه کوتاه‌مدت ماندگار نسبت به نویز جمع‌پذیر محیطی موردبررسی قرارگرفته است. راهکار اول افزایش مقاومت مدل‌های شبکه حافظه کوتاه‌مدت ماندگار نسبت به حضور نویز است که با توجه به‌خصوصیت این شبکه‌ها در یادگیری رفتار بلندمدت نویز ارائه می‌شود. بدین منظور پیشنهاد می‌شود از گفتار نویزی برای آموزش مدل‌ها استفاده شود تا به‌صورت آگاه به نویز آموزش ببینند. نتایج روی مجموعه داده نویزی شده TIMIT نشان می‌دهد که اگر مدل‌ها به‌جای گفتار تمیز با گفتار نویزی آموزش ببینند، دقت بازشناسی تا ۱۸ درصد بهبود خواهد یافت. راهکار دوم کاهش تأثیر نویز بر ویژگی‌های استخراج‌شده با استفاده از شبکه خود رمزگذار کاهنده نویز و استفاده از ویژگی‌های گلوگاه به‌منظور فشرده‌سازی بردار ویژگی و بازنمایی سطح بالاتر ویژگی‌های ورودی است. این راهکار باعث می‌شود مقاومت ویژگی‌ها نسبت به نویز بیشتر شده و درنتیجه دقت سیستم بازشناسی پیشنهادشده در راهکار اول، در حضور نویز ۴ درصد افزایش یابد.

واژه‌های کلیدی: بازشناسی گفتار، مقاومت نسبت به نویز، داده‌های چند شرطی، شبکه خود رمزگذار، شبکه حافظه کوتاه‌مدت ماندگار.

Robust Speech Recognition using Long Short Term Memory Networks and Bottleneck Features

Amin Moaven Joula[1], PhD candidate, Ahmad Akbari, Associate professor [2], Babak Naser Sharif, Assistant professor[3]

1- Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran, Email: joula@comp.iust.ac.ir

2- Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran, Email: akbari@iust.ac.ir

3- Faculty of Computer Engineering, K.N Toosi University of Technology, Tehran, Iran, Email: bnasersharif@kntu.ac.ir

Abstract: Deep neural networks have been widely used in speech recognition systems in recent years. However, the robustness of these models in the presence of environmental noise has been less discussed. In this paper, we propose two approaches for the robustness of deep neural networks models against environmental additive noise. In the first approach, we propose to increase the robustness of long short-term memory (LSTM) networks in the presence of noise based on their abilities in learning long-term noise behavior. For this purpose, we propose to use noisy speech for training models. In this way, LSTMs are trained in a noise-aware manner. The results on the noisy TIMIT dataset show that if the models are trained with noisy speech rather than clean speech, recognition accuracy will be improved up to 18%. In the second approach, we propose to reduce noise effects on the extracted features using a denoised autoencoder network and to use the bottleneck features to compress the feature vector and represent the higher level of input features. This method increases the accuracy of the proposed recognition system in the first approach by 4% in the presence of noise.

Keywords: Speech recognition, Noise robustness, Multicondition data, Autoencoder network, Long short term memory network.

تاریخ ارسال مقاله: ۱۳۹۶/۹/۲۴
تاریخ اصلاح مقاله: ۱۳۹۶/۱۲/۱۶ و ۱۳۹۷/۴/۳۰
تاریخ پذیرش مقاله: ۱۳۹۷/۷/۱۵
نام نویسنده مسئول: دکتر احمد اکبری
نشانی نویسنده مسئول: ایران – تهران – نارمک – خیابان هنگام – خیابان دانشگاه علم و صنعت – دانشگاه علم و صنعت ایران - دانشکده مهندسی کامپیوتر.

۱ -مقدمه

استفاده از یادگیری عمیق و به‌خصوص شبکه‌های عصبی عمیق، در سیستم‌های بازشناسی گفتار به‌صورت مرکب با مدل مخفی مارکف و یا به‌صورت مستقل، در سال‌های اخیر بسیار موردتوجه قرارگرفته است. در ابتدا استفاده از شبکه‌های عمیق پیش‌رو و به‌طور خاص استفاده از آن در کنار مدل مخفی مارکف (DNN-HMM)[۱] بهبود قابل‌ملاحظه‌ای در مدل‌سازی آکوستیکی ایجاد نمود. با توجه به ماهیت دینامیک گفتار، یکی دیگر از گزینه‌ها برای مدل‌سازی آکوستیکی، شبکه‌های عصبی بازگشتی(RNN)[۲] هستند. با آزمایش‌هایی که دیگر پژوهش‌ها انجام دادند، مشاهده شد که RNN-HMM کارایی خوبی به اندازه DNN-HMM ندارد.

به‌جای استفاده از شبکه‌های عصبی بازگشتی در کنار مدل‌های مخفی مارکف، می‌توان آن‌ها را به‌صورت یکجا برای کاربرد بازشناسی گفتار آموزش داد [۱, ۲, ۳]. دلیل عدم اقبال به شبکه‌های عصبی بازگشتی با وجود سابقه زیاد آن‌ها، سخت بودن آموزش این شبکه‌ها در گذشته بود. اما حالا با وجود الگوریتم‌های آموزش هوشمندانه‌تر مثل آموزش لایه به لایه حریصانه [۴, ۵] و سخت‌افزار مناسب، این امکان وجود دارد تا بتوان آن‌ها را به‌خوبی آموزش داد. آموزش یکجای آن‌ها باعث می‌شود تا نسبت به مدل‌های مخفی مارکف، فضای حالت بزرگ‌تری مورداستفاده قرار گیرد و از خاصیت دینامیک شبکه‌های عصبی بازگشتی به‌خوبی بهره گرفته شود. همچنین این نوع آموزش این امکان را می‌دهد تا تأثیر تراز بندی اشتباه در میان داده‌های آموزشی را حذف کرد.

یکی از خواصی که گفتار دارد آن است که خروجی سیستم بازشناسی گفتار به دنباله‌ای بستگی دارد که گفتار در آن واقع است. به‌عبارت‌دیگر اگر زمینه گذشته و آینده در بازشناسی واج مدنظر قرار گیرد، دقت خروجی بهبود خواهد یافت. بدین منظور محققان شبکه‌های عصبی بازگشتی دوطرفه را مطرح کردند که در آن هر دنباله آموزشی را به‌طرف جلو و به‌طرف عقب به دولایه مخفی بازگشتی ارائه کند که هردوی آن‌ها به یک‌لایه خروجی یکسان متصل هستند.

در کنار مزایایی که شبکه‌های عصبی بازگشتی برای ما فراهم می‌کنند، چندین ضعف نیز دارند که استفاده از آن‌ها را در حل مسائل بازشناسی گفتار در دنیای واقعی محدود می‌کنند. متأسفانه بازه زمینه‌ای که معماری‌های RNN استاندارد می‌توانند به آن دسترسی داشته باشند، کاملاً محدود است. درواقع شاید بتوان گفت مهم‌ترین مشکل شبکه‌های عصبی بازگشتی استاندارد آن است که بسیار سخت می‌توان از آن‌ها برای ذخیره اطلاعات در یک بازه زمانی طولانی استفاده کرد. این اشکال شبکه‌های عصبی بازگشتی، دایره زمینه‌ای که این شبکه‌ها می‌توانند بدان دسترسی داشته باشند، را محدود می‌کند. درواقع مشکل این شبکه‌ها، تأثیر یک ورودی داده‌شده به لایه مخفی و درنتیجه تأثیر آن در خروجی شبکه است که یا به‌مرور از بین می‌رود و یا همان‌طور که در اتصالات بازگشتی شبکه گردش می‌کند، به‌صورت نمایی تشدید می‌شود. از این اثر معمولاً در ادبیات موضوع با نام مسئله به صفر رسیدن گرادیان[۳] یاد می‌شود [۶].

شبکه‌های حافظه کوتاه‌مدت ماندگار (LSTM)[۴] طراحی جدیدتری از معماری شبکه‌های عصبی بازگشتی است که از سلول‌های حافظه استفاده می‌کند. در بسیاری از آزمایش‌ها نشان داده‌شده که LSTM می‌تواند اطلاعات موجود در یک بازه زمانی طولانی و بلند را ذخیره و مورد دسترسی قرار دهد. ترکیب خاصیت استفاده از زمینه گذشته و آینده و سلول‌های حافظه LSTM منجر به ایجاد شبکه‌های LSTM دوطرفه می‌شود [۷].

از سوی دیگر، مطالعه مقاوم‌سازی نسبت به نویز در بازشناسی گفتار پیشینه زیادی دارد که قسمت عمده آن قبل از پیشرفت‌های اخیر یادگیری عمیق بوده است [۸, ۹]. عدم تطبیق میان محیط آزمایش و محیط آموزش که به دلیل وجود عواملی چون نویز، تنوع گوینده و لهجه اتفاق می‌افتد، سبب افت کارایی سیستم‌های بازشناسی گفتار در این شرایط می‌گردد. برای غلبه بر این عدم تطابق معمولاً روش‌های مقاوم‌سازی در سه سطح سیگنال، ویژگی، مدل و یا ترکیبی از آن‌ها کار می‌کنند که روش‌های مبتنی بر مدل معمولاً با توجه به ماهیت سیستم‌های GMM_HMM [۵] توسعه‌یافته‌اند. ازاین‌رو روش‌های مبتنی بر مدل به‌صورت مستقیم قابل‌اعمال بر روی مدل‌های یادگیری عمیق جدید بازشناسی گفتار نیستند. اما تکنیک‌های در سطح سیگنال یا ویژگی را می‌توان به‌صورت مستقیم برای سیستم DNN به کار برد. در [۱۰] یک بررسی همراه با جزییات از استفاده شبکه‌های DNN برای بازشناسی گفتار مقاوم به نویز در دامنه ویژگی توسط سلتزر[۶] و همکاران گزارش‌شده است. محققین در این بررسی، الگوریتم بهبود ویژگی C-MMSE [۷] [۱۱] را روی ویژگی ورودی در DNN اعمال کردند. با پردازش داده‌های آموزش و آزمایش توسط یک روش یکسان، هرگونه خطای سازگاری یا اثری که با روش بهبود رخداده است، توسط بازشناس DNN-HMM قابل یادگیری است. این تحقیق همچنین به‌طور موفق نشان داد الگوی آموزش آگاه به نویز را می‌توان برای آموزش DNN به کار برد که در آن به هر مشاهده، تخمینی از نویز اضافه شد و نتایج خوبی روی داده‌های Aurora4 به دست آمد. همچنین می‌توان هرگونه عدم سازگاری در زمان آموزش و آزمایش را، به‌صورت یک مسئله تطبیق دامنه مدل کرد [۱۲].

در سال‌های اخیر شبکه‌های عصبی و به‌ویژه نوع عمیق آن در حوزه استخراج ویژگی‌های مقاوم در برابر نویز موردتوجه قرارگرفته‌اند [۱۳, ۱۴]. این شبکه‌ها یک تبدیل غیرخطی بر روی ویژگی‌ها اعمال کرده و یک بازنمایی جدید از ویژگی‌های حذف نویز شده ارائه می‌دهند. شبکه‌های خود رمزگذار کاهنده نویز[۸] و شبکه‌های گلوگاه[۹] جزو این نوع شبکه‌ها هستند. ویژگی بارز شبکه‌های گلوگاه، لایه گلوگاه آن‌ها است که ابعاد کمتری نسبت به سایر لایه‌ها دارد و ویژگی‌های استخراج‌شده از این لایه می‌توانند اطلاعات سطوح بالاتری از ویژگی‌های اولیه را در قالبی فشرده‌تر نسبت به ویژگی‌های اولیه ارائه دهند. این شبکه‌ها در بازشناسی گفتار مقاوم در برابر نویز موفق ظاهرشده‌اند.

در [۱۵] از یک شبکه‌ی خود رمزگذار به‌منظور استخراج ویژگی‌های گلوگاهی استفاده‌شده است، اما برخلاف تحقیقات دیگر، ویژگی‌های اولیه‌ی گفتاری نظیر مل کپستروم به‌عنوان ورودی‌های شبکه‌ی خود رمزگذار نیستند بلکه از خروجی‌های یک شبکه‌ی باور عمیق به‌عنوان ورودی برای خود رمزگذار استفاده‌شده است. همین روش در [۱۶] به ساختار عمیق تعمیم داده‌شده است و از یک شبکه‌ی خود رمزگذار عمیق برای نگاشت داده‌ی نویزی به داده‌ی تمیز استفاده‌شده است، اما ورودی‌های این شبکه‌ی عمیق برچسب‌های واجی هستند که از آموزش یک شبکه‌ی عصبی عمیق حاصل‌شده است و استدلال آن‌ها این بوده که وقتی نگاشت با در نظر گرفتن اطلاعات صوتی انجام می‌شود پس می‌تواند وابسته به اطلاعات واجی هم باشد، و بر همین اساس روش جدیدی ارائه کرده‌اند. در سال ۲۰۱۳ متز[۱۰] از تعدادی شبکه‌ی خود رمزگذار که به‌صورت ساختار پشته‌ای به دنبال هم قرارگرفته بودند، استفاده کرد که ویژگی‌های گلوگاه استخراج‌شده از این ساختار نسبت به ویژگی‌های اولیه موجب بهبود عملکرد شدند [۱۷].

در این مقاله دو راهکار برای افزایش مقاومت سیستم بازشناسی گفتار مبتنی بر شبکه‌های حافظه کوتاه‌مدت ماندگار نسبت به نویز جمع‌پذیر ارائه می‌شود. در درجه اول با توجه به‌خصوصیات شبکه LSTM و قابلیت آن در یادگیری رفتار بلندمدت نویز، پیشنهاد می‌شود که این شبکه با دادگان گفتار نویزی آموزش ببیند. از سوی دیگر برای کاهش تأثیرات نویز بر ویژگی‌ها پیشنهاد می‌شود که از یک شبکه خود رمزگذار حذف‌کننده نویز استفاده شود تا ضمن فشرده سازی و کاهش بعد ویژگی‌ها، تأثیرات نویز بر ویژگی‌ها را نیز کاهش دهد. برای آموزش و عملکرد بهتر این شبکه رمزگذار از روش آموزش لایه به لایه استفاده‌شده است.

در ادامه مقاله، در بخش ۲ شبکه‌های حافظه کوتاه‌مدت ماندگار معرفی می‌شوند. در بخش ۳ روش‌های پیشنهادی مبتنی بر آموزش چند شرطی داده‌ها و شبکه خود رمزگذار کاهنده نویز، ارائه‌شده است. در بخش ۴ نیز نتایج روش‌های پیشنهادی آورده شده است. درنهایت در بخش ۵، جمع‌بندی و نتیجه‌گیری انجام خواهد شد.

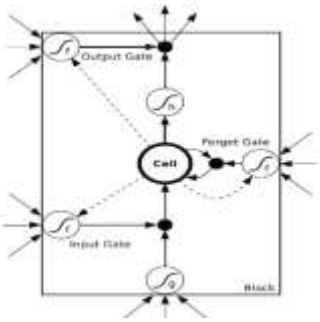
۲ -شبکه حافظه کوتاه‌مدت ماندگار

شبکه LSTM طراحی جدیدتری از معماری شبکه‌های RNN است که از سلول‌های حافظه استفاده می‌کند. در بسیاری از آزمایش‌ها نشان داده‌شده که LSTM می‌تواند اطلاعات موجود در یک بازه زمانی طولانی و بلند را ذخیره و مورد دسترسی قرار دهد. همچنین مزایای استفاده از LSTM در حوزه‌های دنیای واقعی ازجمله پردازش گفتار به اثبات رسیده است [۱۸, ۱۹, ۲۰].

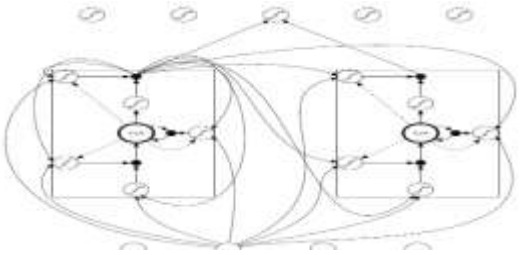
معماری LSTM شامل مجموعه‌ای از زیرشبکه‌های به‌صورت بازگشتی متصل است که بلاک‌های حافظه نامیده می‌شوند. این بلاک‌ها را می‌توان نسخه مشتق‌پذیر چیپ‌های حافظه در کامپیوتر دیجیتال در نظر گرفت. هر بلاک شامل یک یا چند سلول حافظه خود متصل و سه واحد ضرب‌کننده (گیت‌های ورودی، خروجی و فراموشی) است که حالت پیوسته عملیات نوشتن، خواندن و بازنشانی را فراهم می‌کنند.

شکل ۱ یک بلاک حافظه LSTM با یک سلول را نمایش می‌دهد. سه گیت موجود در آن، واحدهای جمعی غیرخطی هستند که فعالیت‌های درون و بیرون بلاک را جمع‌آوری کرده و فعالیت سلول را از طریق ضرب (دایره‌های سیاه کوچک) کنترل می‌کنند. گیت‌های ورودی و خروجی، ورودی و خروجی سلول را ضرب می‌کنند درحالی‌که گیت فراموشی، حالت قبلی سلول را ضرب می‌کند. هیچ تابع فعالیتی درون سلول اعمال نمی‌شود. تابع فعالیت گیت معمولاً سیگموید لجستیک[۱۱] است به‌گونه‌ای که فعالیت‌های گیت بین صفر (گیت بسته) و یک (گیت باز) هستند. توابع فعالیت ورودی و خروجی سلول (g و h) معمولاً tanh یا سیگموید لجستیک است. گاهی اوقات نیز h تابع همانی است. اتصالات وزن‌دار از سلول به گیت‌ها با خطوط فاصله نشان داده‌شده‌اند. همه اتصالات دیگر در بلاک بدون وزن هستند (یا وزن ثابت ۱ دارند). تنها خروجی‌های بلاک به بقیه شبکه از عملیات ضرب گیت خروجی می‌آیند.

یک شبکه LSTM همانند RNN استاندارد است با این تفاوت که واحدهای جمعی در لایه مخفی با بلاک‌های حافظه جایگزین می‌شوند به صورتی که در شکل ۲ نشان داده‌شده است. در این شکل، شبکه شامل چهار واحد ورودی، یک‌لایه مخفی از دو بلاک حافظه LSTM تک‌سلولی و ۵ واحد خروجی است. همه اتصالات نشان داده نشده‌اند. توجه کنید که هر بلاک چهار ورودی و تنها یک خروجی دارد. بلاک‌های LSTM را می‌توان با واحدهای جمعی معمولی ترکیب کرد بااین‌وجود این کار غیرضروری است. همان لایه‌های خروجی که برای شبکه‌های RNN استاندارد مورداستفاده قرار می‌گیرند، را می‌توان برای شبکه‌های LSTM به کار گرفت.



شکل ۱: بلاک حافظه شبکه کوتاه‌مدت ماندگار با یک سلول [۶]



شکل ۲: یک نمونه از شبکه حافظه کوتاه‌مدت ماندگار [۶]

گیت‌های ضربی به سلول‌های حافظه LSTM این امکان را می‌دهند تا اطلاعات در بازه‌های زمانی طولانی را ذخیره کنند و به آن‌ها دسترسی یابند و درنتیجه اثر مسئله به صفر رسیدن گرادیان را کاهش دهند. برای مثال تا زمانی که گیت ورودی بسته می‌ماند(یعنی فعالیتی نزدیک به صفر دارد)، فعالیت سلول توسط ورودی‌های جدیدی که به شبکه می‌رسند، تغییر نمی‌کند و با باز کردن گیت، خروجی در دنباله بسیار دیرتر استفاده می‌شود.

۳ -روش پیشنهادی

۳ ۱ -آموزش با داده‌های چند شرطی

تکنیک‌های زیادی برای پاک‌سازی سیگنال‌های گفتار، طیف و یا ویژگی‌ها از تخریب‌ها در ادبیات موضوع طراحی‌شده‌اند. اکثر این روش‌ها به نویز جمع‌شونده و فیلتر کردن خطی ارتباط دارند. در هنگام استفاده، یکی از این روش‌ها مورداستفاده قرار می‌گیرد تا گفتار ورودی را بهبود دهد و پس‌ازآن به سیستم بازشناسی داده می‌شود که روی گفتار تمیز آموزش‌دیده است. فرض ضمنی در این راهکار آن است که روش‌های به‌کاررفته، کاملاً عملیات بهبود را انجام می‌دهند و هیچ عدم تطابقی میان داده‌های آزمایش بهبودیافته و مدل‌های آکوستیکی تمیز باقی نمی‌ماند. البته کاملاً مشخص است که این فرض درست نیست. همه این الگوریتم‌ها مقداری تخریب و اعوجاج باقی‌مانده، در خروجی دارند. اصل پایه‌ای آموزش آگاه به نویز (NAT)[۱۲] آن است که همان الگوریتم تطبیق که در طول آموزش مورداستفاده قرار می‌گیرد، همان الگوریتم در زمان آزمایش اعمال می‌شود تا منابع تغییرپذیری یکسانی را از داده‌های آموزشی و آزمایشی حذف کنند. برای بازشناسی گفتار نویزی، جبران نویز معادل تطبیق نویز است. منابع نویز مختلف و تغییراتی که موجب می‌شوند با محیط‌های مختلف و اعوجاج‌هایی که آن‌ها سبب می‌گردند، جایگزین می‌شوند. بنابراین الگوریتم بهبود در زمان آزمایش ، باید زمان آموزش نیز اعمال شود. در آموزش آگاه به نویز (NAT)، گفتار تخریب‌شده توسط نویز ، شامل انواع نویزها و سطوح مختلف آن است که در مرحله آزمایش دیده خواهند شد. پس اگر بدانیم که سیستم تنها برای محیط‌های خاص به کار گرفته خواهد شد، برای مثال در یک ماشین، در این صورت نوع داده‌هایی که برای آموزش مورداستفاده قرار می‌گیرند، می‌توانند این واقعیت را انعکاس دهند.

هدف ما در این مقاله آن است تا برای آموزش مدل مقاوم به نویز، از تکنیک‌هایی استفاده کنیم که عدم تطابق میان شرایط آموزش و آزمایش را از بین می‌برند و یا کاهش می‌دهند. به‌عنوان‌مثال یکی از این تکنیک‌ها استفاده از NAT در فضای مدل است. دستورالعمل پایه برای آموزش تطبیقی نویز در فضای مدل آن است که در زمان آموزش مدل ترکیبی از داده‌ها به‌گونه‌ای برای مجموعه آموزش انتخاب شوند که این مجموعه شامل انواع نویزها و انواع SNR [۱۳] باشد که در زمان آزمایش با آن مواجه خواهد شد. به این صورت چون مدل در زمان آزمایش، با تنوعی از انواع نویز و سطوح نویز، مواجه خواهد شد لذا در زمان آموزش سعی بر آن است که همین تنوع در نمونه‌های آموزشی نیز موجود باشد. به مجموعه داده‌های آموزشی که شامل انواع نویز و سطوح نویز است، مجموعه داده چند شرطی[۱۴] می‌گویند. یکی دیگر از نکاتی که انتخاب مدل LSTM را بر دیگر گزینه‌ها برتری می‌دهد، توان یادگیری و مدل‌کردن شبکه‌های LSTM در رفتار بلندمدت نویز است. بنابراین اگر این مدل‌ها با استفاده از داده‌های چند شرطی آموزش ببینند، قادر هستند تا با یادگیری رفتار بلندمدت نویز، دقت بهتری در مواجهه با داده‌های نویزی مرحله آزمایش داشته باشند.

لازم به ذکر است که الزامی به تناظر میان انواع نویز موجود در داده‌های آموزشی و آزمایشی نیست، بلکه هدف آن است که شبکه LSTM با مشاهده انواع نویز در داده‌های آموزشی، رفتار تخریبی بلندمدت نویز را یاد بگیرد و بتواند در زمان آزمایش که با داده‌های نویزی مواجه می‌شود، تطبیق لازم را برای برچسب‌گذاری درست انجام دهد. بدیهی است که داده‌های آزمایش توسط نویزهایی تخریب‌شده‌اند که یا مشابه با نویزهای داده‌های آموزشی هستند و یا با آن‌ها تفاوت دارند.

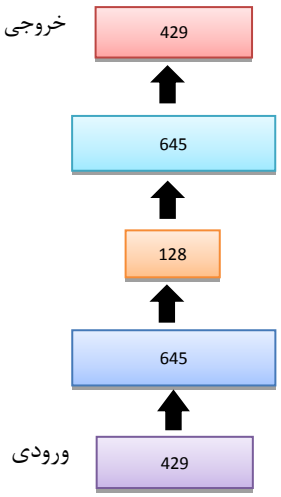
۳ ۲ -تبدیل ویژگی با شبکه گلوگاه

یکی از روش‌های تبدیل ویژگی، استفاده از یک شبکه‌ی خود رمزگذار کاهنده‌ی نویز عمیق جهت استخراج ویژگی‌های حذف نویز شده است. در شکل ۳ ساختار شبکه خود رمزگذار کاهنده نویز پیشنهادی در کار حاضر را مشاهده می‌کنید. در این شکل، ورودی یک سیگنال نویزی و خروجی آن سیگنال تمیز معادل آن است. به همین دلیل به این شبکه کاهنده نویز گفته می‌شود چون شبکه وزن‌هایی را یاد می‌گیرد که بتواند یک نمونه نویزی را به نمونه تمیز آن نگاشت کند. درواقع این شکل، یک شبکه خود رمزگذار ۵ لایه را نشان می‌دهد که تعداد نورون‌های لایه‌های آن به ترتیب ۴۲۹، ۶۴۵، ۱۲۸، ۶۴۵ و ۴۲۹ هستند. همچنین لایه سوم این شبکه نیز لایه گلوگاه نامیده می‌شود

که ابعاد آن نسبت به دیگر لایه‌ها کمتر است. خروجی این لایه، اطلاعات غنی‌تری است که شبکه از ویژگی‌های ورودی یاد گرفته است. در اینجا بردار ویژگی گلوگاه ۱۲۸تایی حاوی اطلاعات غنی مستخرج از بردار ویژگی ۴۲۹تایی ورودی است.

ورودی ۴۲۹تایی بیانگر یک بردار ویژگی ۴۲۹تایی از ویژگی‌های MFCC است که حاصل الحاق بردار ویژگی ۳۹تایی (۱۳ ضریب MFCC و مشتقات اول و دوم آن) استخراج‌شده از ۱۱ پنجره متوالی است. به‌عبارت‌دیگر در هرلحظه برای بازشناسی هر پنجره، در کنار ویژگی‌های MFCC مستخرج از خود آن پنجره، ویژگی‌های استخراج‌شده از ۵ پنجره قبل و ۵ پنجره بعد نیز به شبکه داده می‌شود تا ورودی، اطلاعات زمینه را نیز در برداشته باشد.

آموزش یک شبکه‌ی خود رمزگذار کاهنده‌ی نویز عمیق شامل دو مرحله است: مرحله‌ی اول پیش آموزش است که با یادگیری لایه به لایه‌ی نورون‌ها در هر لایه در این مرحله، یک وزن دهی اولیه‌ی مناسب برای شبکه حاصل می‌شود. بدین ترتیب که پس از یادگیری اولین لایه نورون‌ها، خروجی این لایه (یک توصیف و نمایش جدید از داده‌های خام ورودی) به‌عنوان ورودی برای لایه دوم استفاده می‌شود و این روند ادامه می‌یابد تا تمام لایه‌ها آموزش یابند.



شکل ۳: شبکه خود رمزگذار کاهنده نویز پیشنهادی

هر لایه نورون‌ها با الگوریتم آموزش CE [۱۵] می‌یابد. پس از یادگیری لایه‌ها به‌صورت بدون ناظر، در مرحله دوم کل شبکه عصبی با یک الگوریتم با ناظر که معمولاً انتشار رو به عقب است آموزش داده‌شده و وفق پیدا می‌کند. در یک شبکه‌ی خود رمزگذار کاهنده‌ی نویز عمیق انتشار رو به عقب با هدف مینیمم‌سازی خطای بین داده‌ی خروجی شبکه و داده‌ی تمیز انجام می‌شود. نتایج دیگر پژوهش‌ها حاکی از آن است که پیش آموزش شبکه عصبی خود رمزگذار کاهنده‌ی نویز با داده‌های نویزی و آموزش آن با داده‌های تمیز موجب نتایج بهتری در بازشناسی خواهد شد.

نکته قابل‌توجه دیگر، استفاده از یک‌لایه گلوگاه در یک شبکه‌ی خود رمزگذار عمیق، به‌منظور استخراج ویژگی‌های گلوگاهی است تا تأثیر کاهش بعد مؤثر در لایه‌ی گلوگاه مشاهده گردد. این ویژگی‌ها حاوی مهم‌ترین اطلاعات ویژگی‌های اولیه هستند. ویژگی‌های ورودی به شبکه خود رمزگذار عمیق دارای لایه گلوگاه، ویژگی‌های MFCC استخراج‌شده از چندین قاب متوالی گفتار هستند. این کار سبب می‌شود که شبکه‌ی خود رمزگذار کاهنده‌ی نویز، رفتار درازمدت نویز را یاد بگیرد و بر روی حذف نویز تأثیر بهتری بگذارد. نتیجه آن، استخراج ویژگی‌های گلوگاهی مطلوب‌تری در شبکه‌ی گلوگاه است. با توجه به اینکه ویژگی‌های گلوگاهی بازنمایشی از ویژگی‌های ورودی هستند، الحاق آن‌ها به ویژگی‌های MFCC ورودی، تأثیر چندانی در افزایش مقاومت آن‌ها به نویز نخواهد داشت. لذا بردار ویژگی گلوگاه استخراج‌شده از لایه گلوگاه به‌تنهایی مورد ارزیابی قرار خواهد گرفت.

۴ -نتایج آزمایش‌ها

به‌منظور ارزیابی روش‌های پیشنهادی برای یک سیستم بازشناسی واج و انجام آزمایش‌ها از مجموعه داده TIMIT استفاده‌شده است. این مجموعه داده یکی از رایج‌ترین مجموعه‌های داده تمیز برای ارزیابی بازشناسی گفتار پیوسته و واجها است که با نرخ ۱۶ کیلوهرتز نمونه‌برداری شده است. برای تشکیل داده‌های نویزی، به مجموعه داده TIMIT نویزهای مجموعه NOISEX-92 افزوده‌شده است. به عبارت دقیق‌تر، از میان نویزهای مجموعه NOISEX-92، نویزهای pink، factory و babble با سطوح 0db، 10db و 20db به مجموعه داده TIMIT اضافه‌شده است.

برای پیاده‌سازی آزمایش‌ها در این مقاله از ابزار CNTK [۲۱] استفاده‌شده است. ویژگی‌های استخراج‌شده از نوع MFCC با پنجره زمانی ۲۵ میلی‌ثانیه و شیفت ۱۰ میلی‌ثانیه است. ویژگی‌های MFCC به همراه مشتق اول و دوم تشکیل بردار ویژگی ۳۹تایی را می‌دهند. همچنین مدلی که در اینجا برای بازشناسی واج استفاده می‌شود، از نوع شبکه‌های LSTM است دارای ۱۲۰ واحد خروجی متشکل از سه بسته ۴۰تایی. بسته ۴۰تایی متناظر با ۴۰واجی است که مورد بازشناسی قرار می‌گیرد و سه بسته، متناظر با مفهوم سه‌واج[۱۶] در شبکه‌های عصبی عمیق است. ورودی این شبکه نیز ویژگی‌های استخراج‌شده از یک پنجره ورودی است.

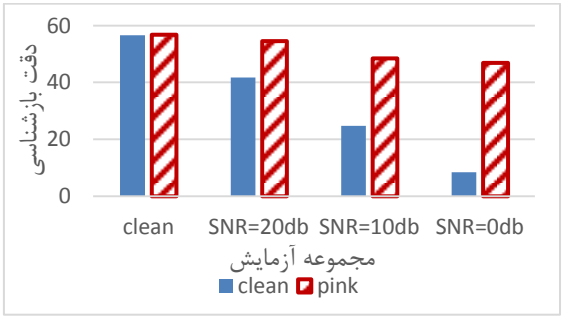
۴ ۱ -نتایج آموزش چند شرطی

در این مقاله برای انجام آموزش به‌صورت NAT از روش چند شرطی (multicondition) استفاده می‌شود. آموزش آگاه به نویز (NAT) به‌نوعی از آموزش اطلاق می‌شود که سیستم بازشناسی در مرحله آموزش نمونه‌هایی از گفتار نویزی را ببیند تا بتواند در مرحله آزمایش، عملکرد بهتری داشته باشد.

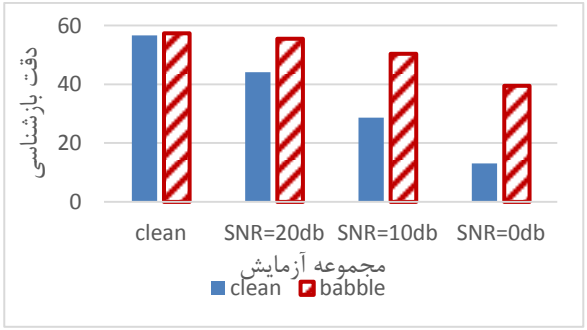
برای انجام آموزش چند شرطی، به دو صورت می‌توان عمل کرد. در روش اول فرض بر آن است که محیطی که سیستم بازشناسی در زمان آزمایش در آن قرار می‌گیرد، کاملاً مشخص است. بنابراین در زمان آموزش، می‌توان از نمونه‌هایی در مجموعه داده چند شرطی استفاده کرد که مختص همان نوع و سطح نویز باشند. در این سناریو

سیستم بازشناسی گفتار، داده‌های نویزی یک نوع نویز را با سطوح مختلف نویز در داده‌های آموزشی می‌بیند و در زمان آزمایش، دقت آن روی همان نوع نویز سنجیده می‌شود. عملاً با این کار، یک مدل اختصاصی برای هر نوع نویز آموزش داده می‌شود تا در زمان آزمایش با فرض دانستن نوع نویز، از مدل اختصاصی همان نویز برای بازشناسی استفاده شود. در شکل ۴ نتایج بازشناسی واج با استفاده از سناریوی اول در آموزش را مشاهده می‌کنید. همان‌طور که مشخص است، درصورتی‌که یک مدل اختصاصی برای هر یک از نویزها آموزش داده شود و در زمان آزمایش، نوع نویز را تشخیص داده و از مدل اختصاصی همان نوع نویز برای بازشناسی استفاده شود، دقت بهبود می‌یابد. استفاده از مدل اختصاصی باعث بهبود در بازشناسی مجموعه آزمایش تمیز، 20db، 10db و 0db به‌طور میانگین به میزان ۰/۴، ۱۲، ۲۲/۶ و ۳۰/۵۳ درصد شده است.

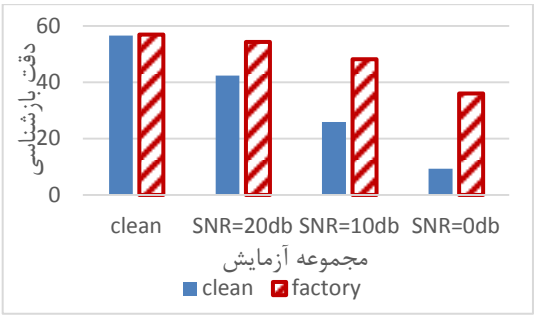
بهبود نتایج روی مجموعه‌های آزمایشی نویزی بدین دلیل حاصل‌شده است که سیستم بازشناسی گفتار از نوع شبکه حافظه کوتاه‌مدت ماندگار، نمونه‌هایی از همان نوع نویز و همان سطح نویز در زمان آموزش‌دیده است و قادر است، رفتار نویز را که در مجموعه آموزش یاد گرفته، به خاطر بسپارد و در زمان آزمایش عملکرد بهتری داشته باشد.



(الف) آموزش با داده‌های تمیز و نویزی pink

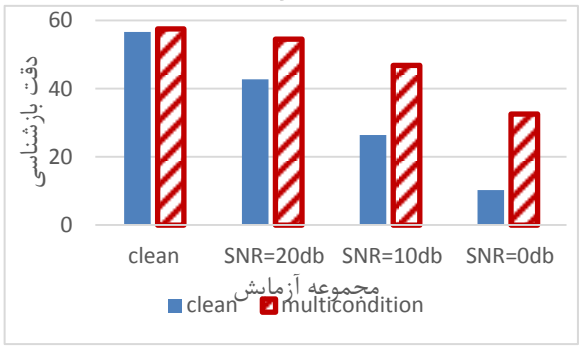


(ب) آموزش با داده‌های تمیز و نویزی babble



(پ) آموزش با داده‌های تمیز و نویزی factory

شکل ۴: دقت بازشناسی واج برای آموزش با یک نویز و آزمایش روی همان نوع نویز



شکل ۵: نتایج سیستم بازشناسی واج آموزش‌دیده با داده‌های ترکیبی همه نویزها

لازم به ذکر است که هدف سناریوی اول، ارائه یک سیستم کاربردی بازشناسی گفتار نیست بلکه ارائه سناریویی است که توان یادگیری شبکه‌های LSTM را در آموزش و مواجهه با حضور یک نوع نویز نشان دهد. در شرایط واقعی، با توجه به اینکه نوع نویز از قبل مشخص نیست، در یک محیط واقعی نمی‌توان از این سناریو بهره گرفت. مگر آنکه سیستم در محیطی بکار گرفته شود که فقط یک نوع نویز مشخص در آن موجود باشد.در روش دوم آموزش چند شرطی، فرض بر آن است که اطلاعاتی از نویز موجود در شرایط آزمایش در دسترس نیست که معمولاً این فرض برقرار است. در این سناریو در زمان آموزش، سیستم داده‌های مربوط به انواع نویزها با سطوح مختلف را می‌بیند و مدل خود را به‌صورت عمومی روی تمام این داده‌ها آموزش می‌دهد. در زمان آزمایش نیز این مدل عمومی، بازشناسی گفتار را انجام می‌دهد. در این روش، ممکن است نویزی که سیستم بازشناسی واج در زمان آزمایش با آن مواجه می‌شود، هیچ‌گونه نمونه‌ای در مجموعه داده چند شرطی مرحله آموزش نداشته باشد. در شکل ۵ نتایج بازشناسی گفتار با استفاده از سناریوی دوم در آموزش مشاهده می‌کنید.

همان‌طور که در شکل ۵ مشخص است، اگر همه نویزها با سطوح مختلف را در مجموعه آموزش قرار دهیم، سیستم بازشناسی گفتار نسبت به حالتی که اصلاً نویز را ندیده است، بسیار بهتر عمل می‌کند تا جایی که روی مجموعه آزمایش pink0db حدود ۳۰۰ درصد افزایش دقت نسبی داشته‌ایم. دلیل این بهبود، دیدن نمونه‌های نویزی در زمان

آموزش است که باعث می‌شود در زمان آزمایش عملکرد سیستم بهبود یابد.

در اینجا نیز بدین دلیل روی مجموعه آزمایش نویزی، شاهد بهبود زیاد نتایج هستیم که شبکه حافظه کوتاه‌مدت ماندگار با دیدن نمونه‌های نویزی مختلف از انواع نویز و سطوح مختلف آن، توانسته است رفتار نویز را یاد بگیرد و در بازشناسی نمونه‌های نویزی در زمان آزمایش، دقت بیشتری داشته باشد.

در ارزیابی مقاومت سیستم‌های بازشناسی گفتار، با افزایش نویز، افت دقت سیستم بازشناسی طبیعی است چون سطح نویز بیشتر شده

جدول ۱: مقایسه انواع روش‌های ترکیب داده‌های نویزی برای آموزش مدل (دقت بازشناسی واج برحسب ٪)

مدل	SNR=20db	SNR=10db	SNR=0db
آموزش‌یافته روی Babble	۵۵/۵	۵۰/۴	۳۹/۵
آموزش‌یافته روی Pink	۵۴/۶	۴۸/۵	۴۶/۹
آموزش‌یافته روی Factory	۵۴/۳	۴۸/۲	۳۶
چند شرطی	۵۴/۵	۴۶/۸۳	۳۲/۶

و تخریب گفتار افزایش‌یافته و درنتیجه دقت بازشناسی کاهش می‌یابد. در میان روش‌های افزایش مقاومت به نویز، آن روشی عملکرد بهتری دارد که دقت بازشناسی با استفاده از آن، همراه با افزایش سطح نویز، افت با شیب کمتری داشته باشد. با بررسی دقیق نتایج شکل ۴ و شکل ۵ مشاهده می‌کنید که هر دو سناریو منجر به کاهش افت دقت بازشناسی گفتار با شیب کمتر، همگام با افزایش سطح نویز شده‌اند.

حال وقت آن است که دو سناریوی آموزش با داده‌های چند شرطی را نسبت به یکدیگر بسنجیم. در جدول ۱ دو سناریوی ترکیب داده‌های نویزی در مجموعه آموزش با یکدیگر مقایسه شده‌اند. همان‌طور که در این شکل می‌بینید، سناریوی اول که یک مدل خاص برای هر نویز آموزش می‌دهد بهتر از سناریوی دوم عمل می‌کند که از مدل عمومی برای بازشناسی استفاده می‌کند. این نتیجه بدین دلیل به دست آماده است که چون سیستم در سناریوی اول تنها داده‌های نویزی یک نوع نویز خاص را دیده قاعدتاً در بازشناسی همان نوع نویز بهتر عمل می‌کند چون اطلاعات آن نوع نویز را به‌طور خاص خوب یاد گرفته است. در مقابل در مدل عمومی، سعی شده اطلاعات همه نویزها یاد گرفته شود که در این سناریو ممکن است دیدن نمونه‌های چندین نویز در آموزش، باعث گمراهی آن در برخی موارد شود. البته بدیهی است هر یک از این سناریوها نسبت به حالتی که سیستم اصلاً نمونه نویزی ندیده است، دقت بالاتری دارند. بنابراین می‌توان نتیجه گرفت زمانی که فقط یک نوع نویز در محیط موجود باشد، شبکه LSTM یادگیری بهتری از این نوع نویز دارد. طبعاً حالت عملیاتی سیستم زمانی است که شبکه LSTM با چند نوع نویز آموزش ببیند. لذا مقایسه فقط ازنظر توان آموزش شبکه LSTM است و هدف، آموزش دو سیستم مختلف بازشناسی گفتار نیست. با مقایسه دو سناریوی آموزش چند شرطی، می‌توان گفت تنها سناریوی دوم در عمل، قابل به‌کارگیری است. چون نوع نویز موجود در شرایط آزمایش، در حالت کلی از قبل مشخص نیست.

در این قسمت رفتار شبکه‌های LSTM در قبال آموزش چند شرطی داده‌ها در دو سناریو مورد ارزیابی قرار گرفت. می‌توان گفت به‌طورکلی ترکیب چند شرطی داده‌ها دقت را افزایش داده است و سبب افزایش مقاومت مدل بازشناس گفتار شده است. همچنین تشخیص نوع نویز می‌تواند به بازشناسی مقاوم گفتار کمک شایانی کند چون مدل مختص آن نوع نویز را می‌توان برای بازشناسی به کار گرفت.

یکی از نکات مهمی که در بازشناسی گفتار مقاوم به نویز موردتوجه قرار می‌گیرد، ارزیابی سیستم بازشناسی گفتار در مواجهه با نویزهای دیده نشده[۱۷] در داده‌های آموزشی است. در این قسمت از ارزیابی، دقت بازشناسی واج سیستم بازشناسی گفتار روی داده‌های نویزی شده با نویزهای car و factory2 در جدول ۲ گزارش می‌شود. نتایج جدول ۲ نشان می‌دهد که عملکرد سیستم بازشناسی گفتار در مواجهه با نویزهای دیده نشده نیز افت زیادی نداشته و شبکه توانسته است که رفتار نویز در داده‌های آموزشی را یاد بگیرد.

جدول ۲: دقت بازشناسی واج سیستم آموزش‌دیده با داده‌های چند شرطی روی نویزهای دیده نشده

داده‌های آزمایش	SNR=20db	SNR=10db	SNR=0db
نویزهای حاضر در آموزش	۵۴/۵	۴۶/۸۳	۳۲/۶
نویز غایب در آموزش car	۵۳/۹	۴۶/۱	۳۳/۲
نویز غایب در آموزش factory2	۵۲/۷	۴۴/۴	۳۱/۵

۴ ۲ -نتایج استفاده از ویژگی‌های گلوگاه

در این بخش با شبکه خود رمزگذار کاهنده نویز یک تبدیل برای ویژگی‌ها یاد گرفته می‌شود تا مقاومت آن‌ها نسبت به نویز افزایش یابد. در این مقاله از یک شبکه خود رمزگذار با سه لایه مخفی با اندازه ۶۴۵-۱۲۸-۶۴۵ استفاده می‌شود. همان‌طور که در شکل ۳ مشاهده می‌کنید، ورودی این شبکه ۴۲۹تایی است که متشکل از ۱۱ پنجره متوالی MFCC است. خروجی آن نیز ۴۲۹تاست. ورودی آن ۱۱ پنجره ویژگی‌های استخراج‌شده از سیگنال نویزی است و خروجی آن ۱۱ پنجره ویژگی‌های استخراج‌شده از سیگنال تمیز متناظر آن است. وظیفه این شبکه آن است که تبدیلی از ویژگی‌های نویزی به ویژگی‌های تمیز یاد بگیرد. لایه ۱۲۸تایی وسط یک‌لایه گلوگاه است که از خروجی‌های این لایه به‌عنوان ویژگی‌های گلوگاه یاد می‌شود.

در این مقاله دو سناریو برای آموزش شبکه خود رمزگذار کاهنده نویز موردبررسی قرار می‌گیرد. سناریوی اول آموزش کل این شبکه به‌صورت هم‌زمان است. سناریوی دوم آموزش لایه به لایه شبکه است. با اجرای سناریوی اول مشاهده می‌شود که استفاده از ویژگی‌های گلوگاه یاد گرفته‌شده چه به‌تنهایی و چه در کنار ویژگی‌های MFCC تأثیر خاصی روی دقت بازشناسی ندارد. بنابراین مانند پژوهش‌های

دیگر روی شبکه خود رمزگذار این نتیجه مشاهده شد که نباید شبکه خود رمزگذار را به‌صورت یکجا آموزش داد. در سناریوی دوم پیش‌آموزش لایه‌ای حریصانه انجام‌شده و شبکه خود رمزگذار به‌صورت لایه‌ای آموزش‌دیده است. بعد از آموزش شبکه، وزن‌های ورودی شبکه تا لایه گلوگاه ذخیره‌شده تا در فرایند بازشناسی گفتار مورداستفاده قرار گیرد.

در فرایند بازشناسی ورودی شبکه ابتدا توسط لایه‌های شبکه خود رمزگذار تبدیل به ویژگی‌های گلوگاه می‌شود و سپس این ویژگی‌ها وارد شبکه حافظه کوتاه‌مدت ماندگار که با داده‌های چند شرطی آموزش‌دیده، برای بازشناسی می‌شوند. نتایج بازشناسی حاکی از آن بود که استفاده از ویژگی‌های گلوگاه در حالتی که شبکه خود رمزگذار به‌صورت لایه‌ای آموزش‌دیده دارای دقت بازشناسی بهتری است.

برای نشان دادن نتایج مقاله و مقایسه آن‌ها بهتر است برای هر یک از سیستم‌های بازشناسی با توجه به سناریوی آموزش آن‌ها، یک علامت اختصاری تعریف شود. ازاین‌رو در جدول ۳ علامت‌های اختصاری سیستم‌های مختلف آورده شده است.

در جدول ۴، نتایج مقایسه دو سناریوی آموزش شبکه خود رمزگذار برای استخراج ویژگی‌های گلوگاه را مشاهده می‌کنید. همان‌طور که در این جدول مشخص است، نتایج استفاده از ویژگی‌های گلوگاه مستخرج از شبکه خود رمزگذار آموزش‌دیده به‌صورت یکجا، تفاوت چندانی با شبکه اصلی حافظه کوتاه‌مدت ماندگار که با داده‌های چند شرطی آموزش‌دیده بود، نمی‌کند. در قسمت (الف) جدول نتایج بازشناسی روی داده‌های نویزی با نویز babble با سطوح SNR مختلف نشان داده‌شده است در قسمت‌های (ب) و (پ) نیز به ترتیب همین نتایج را برای نویزهای pink و factory نمایش داده‌شده است.

با مقایسه نتایج جدول۴، مشاهده می‌شود که در حالتی که شبکه خود رمزگذار به‌صورت لایه‌ای آموزش‌دیده است، نتایج به‌صورت چشم‌گیری بهتر است. بدین ترتیب می‌توان گفت آموزش یک شبکه خود رمزگذار باید به‌صورت لایه‌ای انجام شود و ویژگی‌هایی که از آموزش یکجای آن به دست می‌آیند، اطلاعات بامعنایی از ویژگی‌های ورودی یاد نمی‌گیرند. بدین دلیل می‌بایست شبکه خود رمزگذار کاهنده نویز به‌صورت لایه‌ای آموزش داده شود که این نوع آموزش باعث مقداردهی اولیه بهتری برای وزن‌های شبکه می‌شود.

جدول ۴: مقایسه دقت بازشناسی واج دو سناریوی آموزش شبکه خود رمزگذار برای استخراج ویژگی‌های گلوگاه

Babble	20db	10db	0db
MC_AE_LW	۵۷/۶	۵۲/۵	۴۱/۳
MC_AE_NLW	۵۵/۳	۴۹/۴	۳۸/۵

(الف) نویز babble

pink	20db	10db	0db
MC_AE_LW	۵۷/۶	۵۲/۵	۴۱/۳
MC_AE_NLW	۵۵/۳	۴۹/۴	۳۸/۵

(ب) نویز pink

factory	20db	10db	0db
MC_AE_LW	۵۷/۶	۴۹/۵	۳۳/۱
MC_AE_NLW	۵۴/۱	۴۳/۴	۲۳/۷

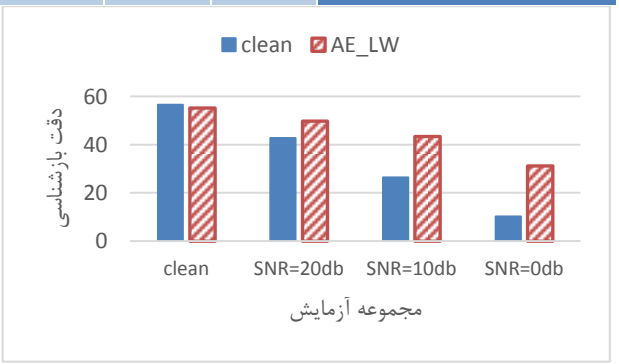
(پ) نویز factory

جدول ۵: ارزیابی معماری‌های مختلف شبکه خود رمزگذار کاهنده نویز برای دقت بازشناسی واج

معماری	20db	10db	0db
(۶۴۵ ،۱۲۸ ،۶۴۵)	۵۷/۶	۴۹/۵	۳۳/۱
(۱۰۲۸ ،۱۲۸ ،۱۰۲۸)	۵۶/۱	۴۷/۶	۳۰/۴
(۱۰۲۸ ،۶۴۵ ،۱۲۸ ،۶۴۵ ،۱۰۲۸)	۵۲/۳	۴۳/۷	۲۵/۲

جدول ۳: علامت اختصاری سیستم‌های بازشناسی

علامت اختصاری	توضیحات
clean	سیستم پایه که با داده‌های تمیز آموزش‌دیده
MC	سیستمی که با داده‌های چند شرطی متشکل از نویزها با سطوح نویز متنوع، آموزش‌دیده
AE_LW	سیستمی که از ویژگی‌های گلوگاه شبکه خود رمزگذار آموزش‌دیده به‌صورت لایه‌ای استفاده می‌کند.
MC_AE_NLW	سیستمی که با داده‌های چند شرطی آموزش‌دیده و از ویژگی‌های گلوگاه شبکه خود رمزگذار آموزش‌دیده به‌صورت یکجا استفاده می‌کند
MC_AE_LW	سیستمی که با داده‌های چند شرطی آموزش‌دیده و از ویژگی‌های گلوگاه شبکه خود رمزگذار آموزش‌دیده به‌صورت لایه‌ای استفاده می‌کند



شکل ۶: مقایسه دقت بازشناسی واج سیستم پایه با سیستم مبتنی بر ویژگی‌های گلوگاه

برای در نظر گرفتن معماری شبکه خود رمزگذار، آزمایش‌های مختلفی صورت گرفت تا معماری مناسب با مجموعه داده TIMIT و

تبدیل موردنظر برای نگاشت ویژگی‌های نویزی به ویژگی‌های تمیز مشخص شود. بدین منظور سه معماری مختلف شبکه خود رمزگذار همراه با آموزش چند شرطی و لایه‌ای مورد ارزیابی قرار گرفت که نتایج آن در جدول ۵، آمده است. در این جدول، تعداد نورون‌های هر یک از لایه‌های مخفی شبکه خود رمزگذار، به ترتیب ذکرشده‌اند.

نتایج جدول ۵، نشان می‌دهد که پیچیده‌تر کردن معماری شبکه خود رمزگذار، کمکی به یادگیری تبدیل میان ویژگی‌های نویزی و تمیز نمی‌کند. با افزایش پیچیدگی معماری، تعداد پارامترهایی که شبکه باید آن‌ها را یاد بگیرد، بیشتر شده و با توجه به اندازه مجموعه داده آموزشی، این امر موجب کاهش دقت خواهد شد. لذا معماری به‌کاررفته در این مقاله همان (۶۴۵، ۱۲۸، ۶۴۵) در نظر گرفته‌شده است.

یکی از جنبه‌هایی که باید بدان توجه داشت، میزان تأثیر استفاده از ویژگی‌های گلوگاه نسبت به سیستم پایه است. در این قسمت سیستم پایه را که با داده‌های تمیز آموزش‌دیده است (clean)، با سیستمی مورد مقایسه قرار می‌دهیم که با داده‌های تمیز آموزش‌دیده و از ویژگی‌های گلوگاه مستخرج از شبکه خود رمزگذار با معماری (۶۴۵، ۱۲۸، ۶۴۵) استفاده می‌کند (AE_LW). در شکل ۶ نتایج این مقایسه آورده شده است.

با توجه به نتایج به‌دست‌آمده، می‌توان به کارآمدی ویژگی‌های گلوگاه در کم کردن تأثیر مخرب نویز در ویژگی‌ها پی برد. همان‌طور که در کاربردهای دیگر مثل بازشناسی گفتار نجوا گونه [۲۲] و بازشناسی زبان گفتار [۲۳]، استفاده از ویژگی‌های گلوگاه باعث بهبود نتایج شده است، در اینجا هم شبکه خود رمزگذار چگونگی کم کردن تأثیر نویز بر ویژگی‌های ورودی را فرامی‌گیرد.

۴ ۳ -مقایسه نتایج دو روش پیشنهادی

بعد از استفاده از دو ایده مطرح‌شده در بخش روش‌های پیشنهادی، مشخص شد که استفاده از داده‌های چند شرطی باعث شد تا مقاومت مدل‌های بازشناسی گفتار نسبت به نویز افزایش یابد. دلیل این بهبود را می‌توان بدین گونه عنوان کرد که وقتی مدل در داده‌های آموزشی، نمونه‌هایی از انواع نویز و انواع SNR را ببیند، می‌تواند با به خاطر سپردن رفتار نویز، آن نویزها و سطوح SNR را با دقت بیشتری بازشناسی کند.

همچنین استفاده از ویژگی‌های گلوگاه مستخرج از شبکه خود رمزگذار کاهنده نویز که به‌صورت لایه‌ای آموزش‌دیده، سبب افزایش مقاومت ویژگی‌ها نسبت به نویز می‌گردد. لذا در جدول ۶، نتایج ۴ سیستم روی داده‌های نویزی با یکدیگر مقایسه می‌شود.

نتایج جدول ۶، نشان می‌دهد استفاده از داده‌های چند شرطی (MC) دقت بازشناسی بیشتری نسبت به مدل Clean آموزش‌دیده با داده‌های تمیز به دست می‌دهد. همچنین شبکه حافظه کوتاه‌مدت ماندگاری که با داده‌های چند شرطی آموزش‌دیده و از ویژگی‌های گلوگاه مستخرج از شبکه خود رمزگذار آموزش‌دیده به‌صورت لایه‌ای، استفاده می‌کند (MC_AE_LW)، نسبت به مدل آموزش‌دیده با داده‌های چند شرطی (MC) و مدل تمیز نسبت به نویز مقاوم‌تر است.

۴ ۴ -مقایسه با روش‌های دیگر

در این قسمت کارایی ترکیب دو روش پیشنهادی در مقایسه با دو راهکار دیگر سنجیده می‌شود. اولین راهکار، تفاضل طیفی[۱۸] [۲۴] است که در آن، برای تخمین نویز از راهکار مارتین[۱۹] استفاده‌شده که با تخمین طیف نویز، اثر آن را در فضای فرکانسی حذف می‌کند.

برای مقایسه با روش تفاضل طیفی در این مقاله، مدل بازشناسی واج که روی داده‌های تمیز آموزش‌دیده، با داده‌های نویزی ارزیابی می شود. قبل از ارائه هر نمونه آزمایشی به این مدل، ابتدا با استفاده از روش مارتین، نویز تخمین زده می‌شود و از طریق روش تفاضل طیفی، حذف نویز صورت می‌گیرد و سپس نمونه آزمایشی بهبودیافته برای برچسب‌زنی به مدل ارائه می‌شود.

دومین راهکار، استفاده از ویژگی‌های استخراج‌شده از شبکه خود رمزگذار است که قبلاً برای بازشناسی گفتار نجوا گونه از این ایده استفاده‌شده است [۲۲]. برای مقایسه با این روش، ابتدا یک شبکه خود رمزگذار کاهنده نویز برای فراگیری نگاشت ویژگی‌های نویزی به ویژگی‌های تمیز، آموزش داده می‌شود. سپس در زمان بازشناسی، ویژگی‌ها با استفاده از همین شبکه خود رمزگذار استخراج‌شده و سپس به‌عنوان ورودی به شبکه LSTM داده می‌شوند تا بازشناسی واج انجام شود.

در جدول ۷، نتایج دقت بازشناسی واج برای مدل تمیز، تفاضل طیفی، ویژگی‌های شبکه خود رمزگذار[۲۰] (AEF) و ترکیب دو روش پیشنهادی آورده شده است. همان‌طور که در جدول ۷ مشاهده می‌کنید، روش تفاضل طیفی به دلیل حذف نویز از نمونه آزمایشی توانسته دقت بالاتری کسب کند. همچنین با افزایش سطح نویز (کاهش SNR)، تفاوت دقت میان مدل تمیز و روش مارتین بیشتر می‌شود چون مدل تمیز هیچ‌گونه حذف نویزی روی نمونه آزمایشی اعمال نکرده و همین امر موجب بالا رفتن خطا می‌گردد.

جدول ۶: مقایسه دقت بازشناسی واج برای انواع روش‌های استفاده از ویژگی های گلوگاه و داده‌های چند شرطی

SNR	20db	10db	0db
Clean	۴۲/۷۳	۲۶/۴	۱۰/۲۷
MC	۵۴/۵	۴۶/۸۳	۳۲/۶
MC_AE_LW	**۵۷/۵۳**	**۵۱/۳۳**	**۳۸/۰۳**

جدول ۷: مقایسه دقت بازشناسی واج ترکیب دو روش پیشنهادی با روش تفاضل طیفی

SNR	20db	10db	0db
Clean	۴۲/۷۳	۲۶/۴	۱۰/۲۷

Spectral subtraction [24]	۴۷/۶۲	۳۷/۵۷	۲۳/۴۶
AEF	۴۹/۸	۴۳/۴	۳۱/۲۶
MC_AE_LW	**۵۷/۵۳**	**۵۱/۳۳**	**۳۸/۰۳**

از سوی دیگر ترکیب دو روش پیشنهادی استفاده از داده‌های چند شرطی و ویژگی‌های گلوگاه توانسته دقت بالاتری نسبت به روش تفاضل طیفی کسب کند. بدین دلیل که در سیستم ترکیب دو روش پیشنهادی، حذف اثر نویز، هم در حوزه ویژگی و هم در حوزه مدل توسط شبکه یاد گرفته می‌شود. درحالی‌که در روش تفاضل طیفی، با استفاده از یک الگوریتم، تنها تخمینی از نویز محاسبه‌شده و در فضای فرکانسی اثر آن از سیگنال نویزی حذف می‌گردد.

نتایج استفاده از ویژگی‌های استخراج‌شده از شبکه خود رمزگذار کاهنده نویز، نشان می‌دهد که این راهکار برای کم کردن تأثیر نویز از ویژگی‌های استخراج‌شده مؤثر بوده و این شبکه توانسته روند کاهش تخریب نویز از ویژگی‌ها را یاد بگیرد. البته با توجه به اینکه این روش تنها به کاهش تخریب نویز در حوزه ویژگی پرداخته، نتوانسته است به اندازه ترکیب دو روش پیشنهادی مؤثر باشد چون MC_AE_LW هم چگونگی تخریب نویز در حوزه ویژگی را با استفاده از شبکه خود رمزگذار کاهنده نویز، فراگرفته و هم با توجه به استفاده از داده‌های چند شرطی، شبکه LSTM توانسته رفتار نویز در تخریب مدل را فرابگیرد. لذا ترکیب دو روش پیشنهادی می‌توانند تخریب نویز هم در حوزه مدل و هم در حوزه ویژگی را کاهش دهند و به دقت بازشناسی واج بالاتری دست یابند.

۵ -نتیجه‌گیری

در این مقاله دو روش برای افزایش مقاومت شبکه‌های حافظه کوتاه‌مدت ماندگار در حضور نویز جمع‌پذیر پیشنهاد شد. با توجه به انتخاب شبکه حافظه کوتاه‌مدت ماندگار به‌عنوان مدل بازشناسی گفتار، خصوصیات این مدل موردبررسی قرار گرفت تا بتوان راهکاری برای افزایش مقاومت این مدل نسبت به حضور نویز در زمان آزمایش، یافت. با توجه به‌خصوصیت یادگیری رفتار بلندمدت نویز توسط این شبکه، یکی از روش‌های پیشنهادی این مقاله بر همین اصل ارائه شد. چون این مدل این توانایی را دارد که با دیدن رفتار نویز در مجموعه آموزشی، رفتار بلندمدت آن را به خاطر بسپارد، پیشنهاد شد تا از داده‌های چند شرطی در زمان آموزش استفاده شود تا مقاومت مدل نسبت به حضور نویز افزایش یابد.

همچنین پیشنهاد دوم که استفاده از ویژگی‌های گلوگاه مستخرج از شبکه خود رمزگذار آموزش‌دیده به‌صورت لایه‌ای بود، باعث افزایش مقاومت ویژگی‌ها شد. بنابراین سیستمی که هم از داده‌های چند شرطی برای آموزش استفاده می‌کند و هم از ویژگی‌های گلوگاه بهره می‌برد، در شرایط نویزی عملکرد بهتری دارد.

مراجع

[1] A. Graves and N. Jaitly, “Towards End-To-End Speech Recognition with Recurrent Neural Networks,” Proceedings of the 31st International Conference on Machine Learning, 2014.

[2] Y. Miao, M. Gowayyed and F. Metze, “EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding,” ASRU, 2015.

[3] D. Amodei and a. et, “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin,” International Conference on Machine Learning, New York, NY, USA, 2016.

[4] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, “Greedy Layer-Wise Training of Deep Networks,” NIPS, 2006.

[5] H. Larochelle, Y. Bengio, J. Louradour and P. Lamblin, “Exploring Strategies for Training Deep Neural Networks,” JMLR, vol. 10, pp. 1-40, 2009.

[6] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, 2012.

[7] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter and H. Ney, “A Comprehensive Study of Deep Bidirectional LSTM RNNs for Acoustic Modeling in Speech Recognition,” Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, USA, 2017.

[۸] مجتبی حاجی‌آبادی, عباس ابراهیمی مقدم و حسین خوش‌بین, «حذف نویز صوتی مبتنی بر یک الگوریتم وفقی نوین»، مجله مهندسی برق دانشگاه تبریز, جلد ۴۶, شماره ۳, صفحه‌های ۱۳۹-۱۴۷, پائیز ۱۳۹۵.

[۹] مسعود گراوانچی‌زاده و ساناز قائمی سردرودی, «بهبود کیفیت گفتار مبتنی بر بهینه‌سازی ازدحام ذرات با استفاده از ویژگی‌های ماسک‌گذاری سیستم شنوایی انسان»، مجله مهندسی برق دانشگاه تبریز, جلد ۴۶, شماره ۳, صفحه‌های ۲۸۷-۲۹۷, پاییز ۱۳۹۵.

[10] M. Seltzer, D. Yu and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

[11] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong and A. Acero, “Robust speech recognition using cepstral minimum-mean-square-error noise suppressor,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 5, 2008.

[12] S. Sun, B. Zhang, L. Xie and Y. Zhang, “An unsupervised deep domain adaptation approach for robust speech recognition,” Neurocomputing, vol. 257, pp. 79-87, 2017.

[13] V. Mitra, H. Franco, R. M. Stern, J. v. Hout, L. Ferrer, M. Graciarena, W. Wang, D. Vergyri, A. Alwan and J. H. L. Hansen, "Robust features in Deep Learning based Speech Recognition,” New Era for Robust Speech Recognition: Exploiting Deep Learning, Springer, 2017, pp. 187 - 217.

[14] A. M. C. Martinez, S. H. Mallidi and B. T. Meyer, “On the relevance of auditory-based Gabor features for deep learning in robust speech recognition,” Computer Speech and Language, vol. 45, no. C, pp. 21-38, 2017.

[15] D. Yu and M. Seltzer, “Improved Bottleneck Features Using Pretrained Deep Neural Networks,” INTERSPEECH, 2011.

[16] T. N. Sainath, B. Kingsbury and B. Ramabhadran, “Auto-encoder bottleneck features using deep belief networks,” Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012.

[17] J. e. a. Gehring, “Extracting deep bottleneck features using stacked auto-encoders,” Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

[18] A. Senior, H. Sak, F. de Chaumont Quitry, T. N. Sainath and K. Rao, “Acoustic Modelling with CD-CTC-SMBR LSTM RNNS,” ASRU , 2015.

[19] H. Sak, A. W. Senior and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” INTERSPEECH, 2014.

[20] A. L. Maas, Z. Xie, D. Jurafsky and A. Y. Ng., “Lexicon-Free

Conversational Speech Recognition with Neural Networks," NAACL, 2015.

[21] D. Yu, K. Yao and Y. Zhang, "The Computational Network Toolkit," IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 123 - 126, 2015.

[22] D. Dorde, T. Grozdic, S. T. Jovicic and M. Subotic, "Whispered speech recognition using deep denoising autoencoder," Engineering Applications of Artificial Intelligence, vol. 59, pp. 15-22, 2017.

[23] R. Fr, P. Matjka, F. Grzl, O. Plchot, K. Vesel and J. H. ernock, "Multilingually trained bottleneck features in spoken language recognition," Computer Speech and Language, vol. 46, no. C, pp. 252-267, 2017.

[24] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," IEEE transactions on speech and audio processing, vol. 9, no. 5, pp. 504-512, 2001.

زیرنویس‌ها

[1] Deep Neural Network-Hidden Markov Model
[2] Recurrent Neural Networks
[3] Gradient Vanishing
[4] Long Short-Term Memory
[5] Gaussian Mixture Model-Hidden Markov Model
[6] Seltzer
[7] Cepstral Minimum Mean Square Error
[8] Denoising Autoencoder
[9] Bottleneck
[10] Metze
[11] logistic sigmoid
[12] Noise Aware Training
[13] Signal to Noise Ratio
[14] Multicondition
[15] Cross Entropy
[16] triphone
[17] Unseen
[18] Spectral Subtraction
[19] Martin
[20] Autoencoder Features