بهسازی گفتار دو مرحله‌ای توسط خودرمزگذار عمیق کاهنده نویز

امیرحسین حاج‌احمدی[۱]، دانشجوی دکتری؛ محمدمهدی همایون‌پور[۲]، دانشیار

۱- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی امیرکبیر - تهران - ایران –a.hadjahmadi@aut.ac.ir
۲- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی امیرکبیر - تهران - ایران – homayoun@aut.ac.ir

چکیده: برای حذف نویز از سیگنال گفتار، هم اطلاعات زمان‌کوتاه و هم اطلاعات زمان‌بلند سیگنال می‌توانند مفید باشند. خصوصا اگر نویز دارای ویژگی‌های غیرایستان باشد. لذا در این مقاله سعی شده است تا با استفاده از کاهش تعداد زیرباندهای فرکانسی در فواصل زمانی بلند امکان اعمال ورودی‌های زمان‌بلند را برای شبکه عصبی خودرمزگذار عمیق کاهنده نویز فراهم سازد. همچنین یک روش دو مرحله‌ای بهسازی گفتار ارائه می‌شود که در مرحله نخست بهسازی زمان‌کوتاه و در مرحله دوم بهسازی زمان‌بلند را انجام دهد. آزمایش‌های این مقاله بر روی مجموعه دادگان Aurora-2 انجام شده است. نتایج نشان داده است که روش پیشنهادی می‌تواند از نظر بهسازی گفتار و معیار PESQ نسبت به فیلتر وینر در شرایط آغشتگی به نویز بالا به میزان ۰/۳ بهبود ایجاد کند. همچنین روش پیشنهادی می‌تواند از نظر دقت بازشناسی خودکار گفتار نسبت به ویژگی‌های مبنا یعنی MFCC حدود ۴٪ بهبود ایجاد کند.

واژه‌های کلیدی: بهسازی گفتار، خودرمزگذار عمیق کاهنده نویز ، رمزگذار عمیق، کاهش نویز.

A Two Phase Speech Enhancement Based on Deep Denoising Autoencoder

A. Hadjahmadi[1], PhD student; M. M. Homayounpour[2], Associate professor

1- Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Iran, Email:a.hadjahmadi@aut.ac.ir
2- Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Iran, Email:homayoun@aut.ac.ir

Abstract: The short-and the long-term information in speech signal are useful for speech enhancement, especially if the speech signal is corrupted by both stationary and non-stationary noises. This paper proposes a new approach to provide long-term speech input for a deep denoising autoencoder by reducing the number of frequency sub-bands of the input data. This paper also proposes a two phase speech enhancement approach. The first phase performs short-term speech enhancement by using a deep denoising autoencoder. In the second phase, long-term speech enhancement denoising autoencoder is applied on the output of short-term enhanced speech data. The proposed models were evaluated on the Aurora-2 Speech recognition corpus and our results show significant improvements of 0.3 in PESQ score at lower SNR values. The proposed models were evaluated on the recognition task where the proposed method results in 4% reduction in word error rate for the multi-condition training when compared to the baseline MFCC front-end.

Keywords: Speech enhancement, denoising autoencoder, deep autoencoder, noise removal.

تاریخ ارسال مقاله: ۱۳۹۶/۰۶/۰۶
تاریخ اصلاح مقاله: ۱۳۹۶/۰۸/۰۸و ۱۳۹۶/۱۰/۱۶
تاریخ پذیرش مقاله: ۱۳۹۶/۱۱/۲۶
نام نویسنده مسئول: محمد مهدی همایون‌پور
نشانی نویسنده مسئول: دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی امیرکبیر - تهران – ایران.

۱- مقدمه

سیگنال گفتار معمولا در مواجهه با انواع نویزهای مخرب محیطی قرار دارد که می‌توانند کیفیت و قابلیت ادراک آن را برای شنونده کاهش دهند. روش‌های بهسازی گفتار به‌همین منظور ایجاد شده‌اند. طبق تعریف، بهسازی گفتار به مجموعه فرایندهایی گفته می‌شود که برای حذف تاثیر نویزهای محیطی از سیگنال گفتار و افزایش ادارک و بهبود کیفیت آن به‌کار می‌روند.

از جمله کاربردهای بهسازی گفتار می‌توان به افزایش کیفیت گفتار در مکالمات تلفنی و موبایلی اشاره کرد. همچنین بهسازی گفتار به‌عنوان یک مرحله اولیه در سیستم‌های خودکار بازشناسی گفتار شناخته می‌شود که از اهمیت بالایی در مقاوم‌سازی آنها در مقابل نویزهای محیطی برخوردار است.

روش های بهسازی گفتار را می‌توان به دو دسته روش های چندکاناله و تک کاناله تقسیم‌بندی نمود که روش ارائه شده در این پروژه در دسته روش‌های بهسازی گفتار تک کاناله قرار می‌گیرد ولی قابلیت تعمیم به حالت چندکاناله را نیز دارا می‌باشند.

در این مقاله روشی دو مرحله‌ای برای بهسازی طیف پیشنهاد شده است که در مرحله اول نگاشت حذف نویز را با اطلاعات زمانی کوتاه مدت را با استفاده از یک شبکه خودرمزگذار عمیق کاهنده نویز مدل می‌نماید و سپس با استفاده از شبکه خودرمزگذار عمیق کاهنده نویز دوم نگاشت بین طیف بهسازی شده زمان‌کوتاه در مرحله اول به طیف تمیز را مدل می نماید.

همچنین در این مقاله با استفاده از کاهش تعداد زیرباندهای فرکانسی در فواصل زمانی بلند برای یادگیری زمان‌بلند سیگنال گفتار توسط شبکه عصبی خودرمزگذار عمیق کاهنده نویز پیشنهاد شده است و هم برای بهسازی گفتار با مدل کردن تغییرات زمان‌کوتاه و زمان‌بلند به صورت همزمان مدلی پیشنهاد شده است. آزمایش‌های انجام شده روی مجموعه دادگان Aurora-2 نشان دهنده توانایی بالاتر این روش در مقایسه با خودرمزگذار عمیق کاهنده نویز می‌باشد.

در ادامه این مقاله، ابتدا در بخش دوم کارهای مرتبط مرور شده‌اند. سپس در بخش سوم سیستم بهسازی گفتار دو مرحله‌ای پیشنهادی تشریح شده است. در بخش چهارم چارچوب آزمایش‌های انجام شده تشریح و سپس نتایج آن در بخش پنجم نشان داده شده است. در نهایت در بخش ششم بحث و نتیجه‌گیری کلی مقاله بیان شده است.

۲- مرور کارهای مرتبط

۲-۱- بهسازی گفتار با کمک شبکه عصبی عمیق

بهسازی گفتار مسئله پیچیده و مشکلی است. به‌خصوص زمانی که گفتار تک کاناله باشد. اولین روش حل این مسئله، یعنی روش تفاضل طیف (SS)[1] در سال ۱۹۷۹ معرفی گردید [۱] و از آن زمان تا سال ۲۰۱۲ رایجترین روش‌های بهسازی گفتار روش‌های آماری مانند، فیلتر وینر[2]، MMSE[3] می باشند [۸-۲]. در این روش‌ها، فرض بر ایستان‌بودن نویز، موجب اثر مخربی به نام نویز موزیکال در سیگنال بهبود یافته می شود. لذا روش‌های مبتنی بر ماسک ابداع شدند تا بتوانند فرض بر ایستان‌بودن نویز را جبران نمایند. در این روش‌ها ابتدا برای سیگنال نویزی‌شده ماسک تخمین زده می‌شود تا بخش‌های تخریب‌شده از بخش‌های سالم تفکیک داده شوند. سپس تنها بخش‌های تخریب شده بهسازی می‌شوند [۹، ۱۰].

شبکه‌های عصبی توانایی بالایی در مدل سازی سیگنال‌های زمانی دارند. به‌عنوان مثال در [۱۱] از شبکه عصبی برای پیش‌بینی قیمت روزانه برق و در [۱۲] از شبکه عصبی برای تخمین زمان بحرانی رفع خطا در شبکه استفاده شده است. در چندسال اخیر با توسعه پردازش موازی به کمک GPUها و ارائه روش های پیش‌آموزش، شبکه های عصبی عمیق در بهسازی گفتار بسیار مورد توجه قرار گرفته‌اند [۱۶-۱۳]. توجه زیاد به شبکه‌های عصبی عمیق برگرفته از توانایی بالای آنها در مدل‌سازی نگاشت‌های غیرخطی است [۱۷].

به‌عنوان مثال در [۱۳] استفاده از شبکه عصبی عمیق و در [۱۴، ۱۸] استفاده از شبکه عصبی LSTM برای بهسازی طیف گفتار مورد استفاده قرار گرفته است. در [۱۹] از خود رمزگذار تغییراتی[4] و همچنین در [۲۰، ۲۱] از شبکه خودرمزگذار عمیق کاهنده نویز استفاده شده است.

شبکه‌های عصبی عمیق توانایی محدودی در مدل‌سازی زمانی دادگان ورودی دارند. هرچند شبکه‌های عصبی بازگشتی مانند LSTM قادر هستند تا حدودی این مشکل را برطرف سازند اما به‌دلیل وجود ابر-پارامترهای زیاد و نیز مسئله تضعیف شدن گرادیان، آموزش آنها معمولا سخت و زمان‌بر است. لذا در این مقاله روشی دو مرحله‌ای ارائه شده است که با کمک اعمال دو خود رمزگذار عمیق کاهنده نویز متوالی قادر است هم تغییرات زمان کوتاه و هم تغییرات زمان بلند سیگنال گفتار را در نظر بگیرد. این روش بهسازی گفتار را با دقت بهتری نسبت به روش‌هایی که تنها از یک مرحله بهسازی استفاده می‌نمایند، انجام می‌دهد. البته این روش قابلیت تعمیم به سایر مدل های بهسازی گفتار با کمک شبکه عصبی مانند خود رمزگذار تغییراتی را نیز داراست.

۲-۲- خودرمزگذار عمیق کاهنده نویز

در این بخش ابتدا خود رمزگذار عمیق و خودرمزگذار عمیق کاهنده نویز تشریح و ویژگی‌های آنها مرور خواهند شد. به شبکه عصبی در صورتی که ورودی و خروجی آن یکسان باشد خودرمزگذار[5] (AE) گفته می‌شود. در صورتی که تعداد لایه‌های چنین شبکه‌ای بیش از سه لایه باشند به آن شبکه خودرمزگذار عمیق[6] (DAE) گفته می‌شود. در صورت یادگیری مناسب وزن‌ها در DAE در هر لایه میانی یک بازنمایی غیرخطی از داده ورودی تولید می شود که قادر به بازتولید مجدد همان داده است [۲۲].

هر DAE از دوبخش رمزگذار[۷] و رمزگشا[۸] تشکیل می شود. بخش رمزگذار از لایه ورودی شروع و تا یکی از لایه های مخفی میانی ادامه می یابد. خروجی آخرین لایه بخش خودرمزگذار به عنوان یک مانیفولد[۹] غیرخطی شناخته می‌شود. سپس در بخش کدگشا خروجی بخش رمزگذار به داده اولیه نگاشت می‌شود [۲۲].

از مهمترین مزایای DAE می توان به بدون ناظر بودن آن اشاره کرد. در واقع DAE این امکان را فراهم می سازد تا از مزایای یادگیری عمیق در مسائلی که داده برچسب گذاری شده به اندازه کافی در دسترس نیست، نیز استفاده شود. به‌همین دلیل از DAE در یادگیری ویژگی[۱۰]، کاهش ابعاد و خوشه‌بندی استفاده می‌شود.

نوعی از DAE به خودرمزگذار عمیق کاهنده نویز[۱۱] (DDA) معروف است که در آن داده ورودی، نمونه نویزی شده‌ای از داده خروجی است. در واقع در این نوع شبکه عصبی یک نگاشت غیرخطی به‌دست می‌آید که قادر است فرایند حذف نویز را از داده ورودی مدل نماید [۲۳].

استفاده از DDA در بهسازی گفتار در چند سال اخیر بسیار مورد توجه بوده است. به عنوان مثال در [۲۰] برای بهسازی طیف با استفاده از DDA روشی ارائه شده است که توسط Ch. H. Lee و همکاران در [۱۳، ۲۱، ۲۴] تکمیل شده است. در این روش‌ها طیف گفتار نویزی شده توسط یک DDA با ورودی زمان‌کوتاه بهسازی می‌گردد.

در این مقاله استفاده از دو خودرمزگذار عمیق کاهنده نویز به صورت متوالی برای بهسازی گفتار پیشنهاد شده است. با این روش امکان بهسازی گفتار هم با کمک مدل سازی تغییرات زمان‌کوتاه و هم با کمک مدلسازی تغییرات زمان‌بلند امکان‌پذیر خواهد بود.

۳- سیستم بهسازی گفتار دو مرحله‌ای پیشنهادی

تاثیر نویز بر سیگنال گفتار گاهی دارای وابستگی زمانی طولانی‌مدت مانند اثر نویزهای غیرایستان و گاهی دارای وابستگی زمانی کوتاه‌مدت مانند اثر نویزهای ایستان مانند نویز سفید است. همچنین استفاده از دانش طولانی‌مدت سیگنال گفتار مانند اطلاعات مربوط به توالی واجی و یا اطلاعات زبانی موجود در سیگنال گفتار، می تواند در بازسازی بخش‌های تخریب‌شده طیف گفتار مفید واقع شود.

لذا در این مقاله روشی پیشنهاد شده است که هم قادر است بهسازی زمان‌کوتاه و هم بهسازی زمان‌بلند سیگنال گفتار را به صورت همزمان انجام دهد. دیاگرام کلی مرحله آموزش روش پیشنهادی در شکل ۱ نشان داده شده است. دادگان گفتاری آموزشی شامل دادگان نویزی و داده تمیز متناظر آنها می‌باشند.

در مرحله آموزش، ابتدا ویژگی های طیفی از دادگان آموزشی تمیز/نویزی استخراج می‌شوند و سپس شبکه عصبی خودکدگذار عمیق کاهنده نویز زمان‌کوتاه آموزش داده می‌شود تا به‌صورت زیر نگاشت بین طیف داده های نویزی ورودی (Y) و طیف داده تمیز خروجی (X) را مدل نماید:

$$\widehat{X_{ST}} = G_{DDA-ST}(Y), \qquad (۱)$$

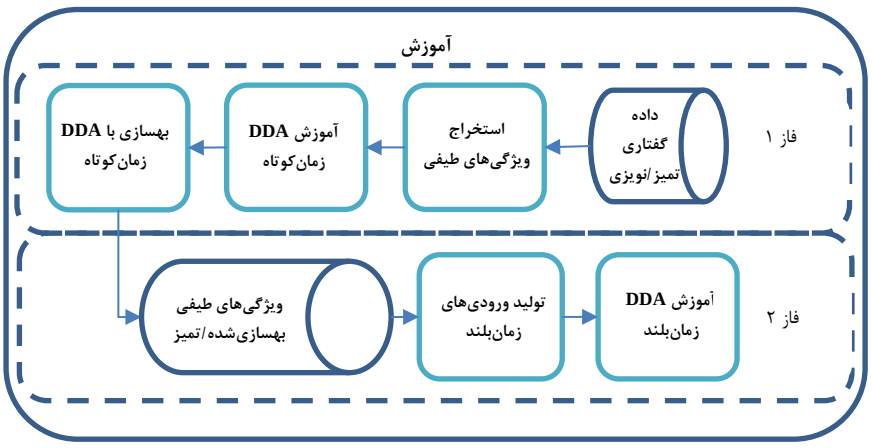
که در آن G_{DDA-ST} نگاشت بهسازی زمان کوتاه و $\widehat{X_{ST}}$ ویژگی های طیفی بهسازی‌شده زمان‌کوتاه هستند.

پس از آن ورودی زمان‌بلند از طیف بهسازی شده زمان‌کوتاه تولید شده و شبکه عصبی خودرمزگذار عمیق کاهنده نویز زمان‌بلند آموزش داده می شود تا توسط رابطه زیر نگاشت بین طیف بهسازی زمان‌کوتاه شده $(\widehat{X_{ST}})$ و طیف داده تمیز را مدل نماید:

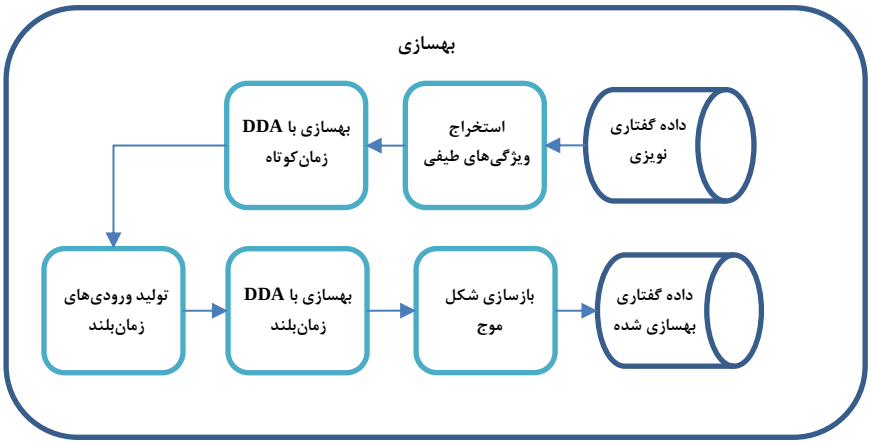
$$\widehat{X_{LT}} = G_{DDA-LT}(\widehat{X_{ST}}), \qquad (۲)$$

که در آن G_{DDA-LT} نگاشت بهسازی زمان‌بلند و $\widehat{X_{LT}}$ طیف بهسازی‌شده زمان‌بلند هستند.

مراحل مختلف بهسازی گفتار نیز در شکل ۲ نشان داده شده است. به‌طور مشابه ویژگی‌های طیفی از سیگنال گفتار نویزی شده استخراج شده و بهسازی زمان‌کوتاه توسط مدل آموزش داده شده در مرحله اول آموزش انجام می شود و ورودی های زمان بلند تولید و توسط مدل آموزش داده شده زمان‌بلند بهسازی می شوند. در نهایت بازسازی سیگنال از طیف بهسازی‌شده صورت می‌پذیرد.



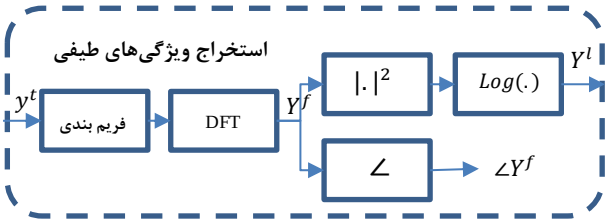
شکل ۱: دیاگرام بلوکی مراحل مختلف آموزش مدل‌های بهسازی.



شکل ۲: دیاگرام بلوکی مراحل مختلف بهسازی گفتار در مدل پیشنهادی.

۳-۱- استخراج طیف

مراحل استخراج طیف سیگنال گفتار در شکل ۳ نشان داده شده است. ابتدا بر روی سیگنال گفتار y^t پنجره‌گذاری و سپس تبدیل فوریه اعمال می گردد و طیف گفتار Y^f بدست می آید. لگاریتم اندازه طیف به‌عنوان ویژگی طیفی در نظر گرفته می‌شود و فاز آن برای بازسازی شکل موج استفاده می‌شود.

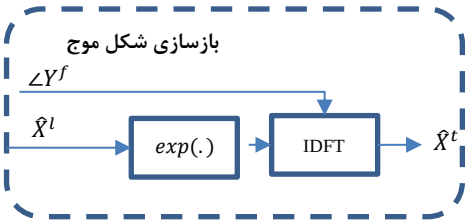


شکل ۳: شکل مراحل استخراج ویژگی‌های طیفی.

در تمام آزمایش‌های انجام‌شده فرکانس نمونه‌برداری ۸ کیلو هرتز است و از اندازه پنجره ۲۵ میلی‌ثانیه و نرخ شیفت ۱۰ میلی‌ثانیه استفاده شده است. تعداد باندهای فرکانسی در مدل سازی زمان‌کوتاه ۱۲۸ عدد در نظر گرفته شده است.

۳-۲- بازسازی شکل موج

در ماجول باز سازی شکل موج که جزئیات آن در شکل ۴ نشان داده شده است، ابتدا تاثیر لگاریتم از اندازه طیف بهسازی‌شده توسط اعمال تابع "$exp(.)$" خنثی و سپس با استفاده از فاز سیگنال نویزی اولیه توسط اعمال یک تابع فوریه معکوس سیگنال بهسازی‌شده $\hat{X}^t$ حاصل خواهد شد.



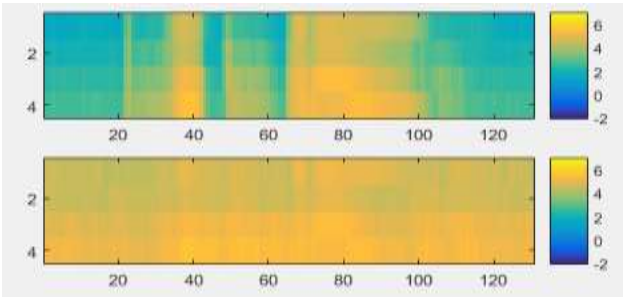
شکل ۴: مراحل بازسازی شکل موج.

۳-۳- مدل‌سازی طیف زمان‌بلند

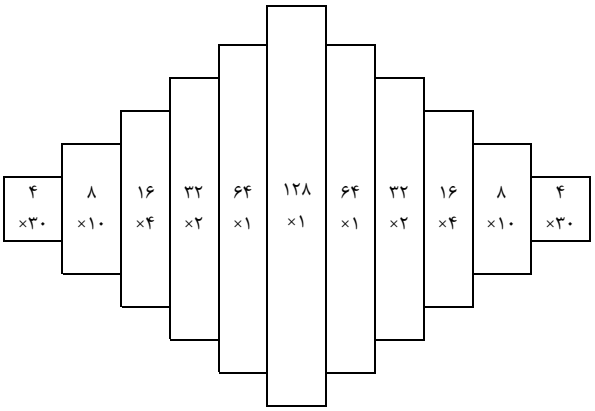
با توجه به اینکه تعداد زیرباندهای فرکانسی ۱۲۸ زیرباند در نظر گرفته شده است، اگر قرار باشد شبکه عصبی تغییرات زمان‌بلند را مدل نماید، بایستی تعداد زیادی از فریم‌های زمانی به‌صورت همزمان به‌عنوان ورودی به شبکه عصبی اعمال شوند که این امر موجب می شود تعداد ورودی‌های شبکه عصبی خیلی زیاد شوند.

برای حل این مسئله پیشنهاد شده است که با دور شدن از فریم جاری و مد نظر در ورودی شبکه، از طیف با تعداد زیرباند کمتر استفاده شود و برای تبدیل طیف ۱۲۸ باندی به طیف‌های با زیرباند کمتر از بانک فیلتر مل استفاده شود. در شکل ۵ نمونه‌ای از یک طیف گفتاری با تنها چهار زیرباند و نیز طیف گفتار نویزی شده با نسبت آغشتگی به نویز (SNR) صفر دسیبل نشان داده شده است. هرچند در چنین بازنمایی بخش زیادی از جزئیات سیگنال گفتار نادیده گرفته شده است اما همچنان اطلاعات زیادی نیز قابل مشاهده است. به‌عنوان مثال انرژی گفتار و اینکه تغییرات فرکانسی اصلی گفتار که در کدام بخش ۴گانه از محدوده فرکانسی است، قابل مشاهده می‌باشد.

در آزمایش‌های انجام‌شده در این مقاله از دنباله طیفی به‌صورت نشان داده‌شده در شکل ۶ به‌عنوان ورودی شبکه عصبی زمان‌بلند استفاده شده است که با توجه به آن اندازه ورودی شبکه عصبی زمان‌بلند ۹۱۲ نرون خواهد بود.



شکل ۵: نمایش طیف سیگنال با تنها چهار باند فرکانسی برای گفتار تمیز (قسمت بالا) و گفتار نویزی‌شده با SNR برابر با صفر dB در پایین.



شکل ۶: نحوه تولید یک ورودی زمان‌بلند برای شبکه عصبی از ترکیب ویژگی‌های طیفی با اندازه ۹۱۲ نرون. (منظور از $m \times n$ در شکل بالا یعنی استفاده از طیف m فریم متوالی با اندازه n زیر باند فرکانسی).

مطابق شکل ۶ که به مرکزیت فریم لحظه lام است (l همان لحظه‌ای است که در خروجی بازسازی خواهد شد)، ۹۵ فریم متوالی از ویژگی‌های طیفی در مدل‌سازی زمان‌بلند به‌کار می‌روند، که حدود یک ثانیه گفتار را مدل خواهند کرد.

۴- چارچوب آزمایش‌ها

چارچوب کلی مجموعه آزمایش‌های انجام‌شده در این مقاله در این بخش تشریح شده است. ابتدا مجموعه دادگان استفاده شده، چگونگی استخراج ویژگی‌های طیفی، معماری به‌کار رفته شبکه خودرمزگذار عمیق کاهنده نویز و ساختار مدل مخفی مارکوف بکار رفته برای رمزگشایی نهایی تشریح شده است.

۴-۱- دادگان

مجموعه دادگان Aurora-2 برای آزمایش‌های این مقاله مورد استفاده قرار گرفته است [۲۵]. این دادگان شامل ۸۴۴۰ داده آموزشی تمیز و ۸۴۴۰ داده آموزشی نویزی است که از آغشته‌شدن همان دادگان آموزشی تمیز به نویزهای مترو، همهمه، خودرو و نمایشگاه با نسبت سیگنال به نویزهای تمیز، ۲۰، ۱۵، ۱۰ و ۵ دسیبل ایجاد شده‌اند. وجود این دادگان استریو[۱۲] برای آموزش DDA لازم است. دادگان تست Aurora-2 شامل ۳ مجموعه است که عبارتند از:

- Set A شامل همان نویزهای مجموعه آموزشی نویزی ولی با نسبت‌های سیگنال به نویز تمیز، ۲۰، ۱۵، ۱۰، ۵، ۰ و ۵- دسیبل.
- Set B شامل نویزهای جمع‌شونده متفاوت از نویزهای آموزشی شامل نویزهای رستوران، خیابان، فرودگاه و ایستگاه قطار می‌باشد که با نسبت‌های سیگنال به نویز تمیز، ۲۰، ۱۵، ۱۰ و ۵ دسیبل ایجاد شده‌اند.
- Set C که علاوه بر نویزهای جمع‌شونده در آن نویزهای کانال نیز اضافه شده است.

۴-۲- معماری خودرمزگذار عمیق کاهنده نویز

معماری کلی بکاررفته در شبکه خودرمزگذار عمیق کاهنده نویز در این آزمایش‌ها یک شبکه عصبی عمیق با ۵ لایه مخفی سیگموئیدی است. در لایه ورودی و خروجی از اعمال همزمان، تعدادی از فریم که هرکدام بین ۴ تا ۱۲۸ ویژگی طیفی را شامل می‌شوند، استفاده شده است. در آزمایش‌های انجام‌شده بجز لایه آخر و لایه ورودی در تمام لایه‌های مخفی از ۳۰۰۰ نرون استفاده شده است. این معماری هم برای شبکه عصبی بهسازی گفتار زمان‌کوتاه و هم شبکه عصبی بهسازی زمان‌بلند استفاده شده است.

جهت آموزش شبکه عصبی، مقدار پارامتر ضریب یادگیری و میزان مومنتوم نیز به‌ترتیب ۰/۰۱ و ۰/۹در نظر گرفته شده‌اند. اندازه مینی بچ[۱۳] نیز در شروع آموزش ۲۵۶ در نظر گرفته شده است که به مرور کاهش یافته تا به ۶۴ رسیده است و از ۲۰۰ مرحله تکرار (epoch) برای آموزش شبکه استفاده شده است.

۴-۳- ساختار مدل مخفی مارکوف

جهت بازشناسی کلمات از مدل‌های مخفی مارکوف تمام-کلمه استفاده شده است. برای هر کلمه ۱۶ حالت در مدل مخفی مارکوف در نظر گرفته شده است (البته با لحاظ کردن دو حالت شروع و خاتمه برای هر کلمه ۱۸ حالت می‌شود). مدل مخفی مارکوف از نوع چپ به راست بوده و در هر حالت، از یک مدل مخلوط گوسی با سه گوسی استفاده شده است. هچنین مدل قطری برای ماتریس کواریانس ضرائب ویژگی در نظر گرفته شده است [۲۵].

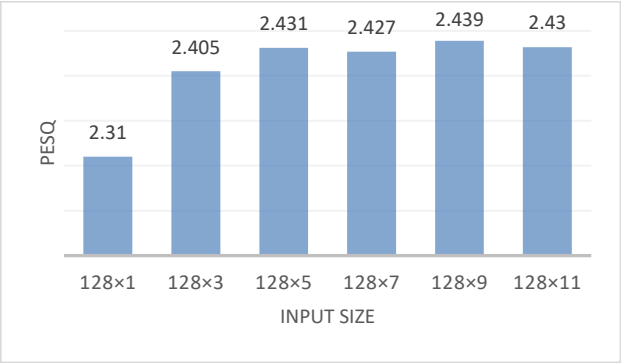
۵- آزمایش‌های انجام‌شده

در این بخش مجموعه آزمایش‌های انجام شده شامل، بررسی بهسازی زمان‌کوتاه، بررسی بهسازی زمان‌بلند و بهسازی با روش ترکیبی تشریح خواهد شد.

۵-۱- بهسازی گفتار زمان‌کوتاه

با استفاده دادگان آموزش چند-شرطی Aurora-2، شش مدل DDA آموزش داده شده است که اندازه ورودی در آنها بین ۱ فریم تا ۱۱ فریم تغییر کرده است. سپس بهسازی با استفاده از هر کدام از این مدل‌ها

برای دادگان تست A از دادگان Aurora-2 انجام شده است. سپس معیار PESQ برای همه آنها محاسبه شده است که میانگین آن در نمودار شکل ۷ نشان داده شده است.



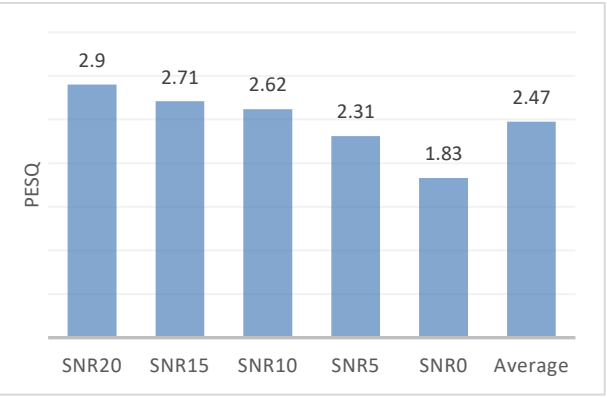
شکل ۷: مقدار معیار PESQ برای مجموعه testA از Aurora-2 برای انواع DDA با اندازه ورودی‌های مختلف.

با توجه به نتایج نشان داده شده در شکل ۷ مشاهده می‌شود که با افزایش اندازه ورودی از ۱ فریم تا ۵ فریم همواره دقت به‌سازی افزایش یافته است اما از ۵ تا ۱۱ فریم تغییرات چندان محسوس نیست. به‌عبارت دیگر یک شبکه عصبی خودرمزگذار عمیق کاهنده نویز به تنهایی قادر به کاهش تاثیرات زمان‌بلند و زمان‌کوتاه به صورت همزمان نیست.

۵-۲- بهسازی گفتار زمان‌بلند

برای DDA که ورودی آن مشابه بخش ۳-۳ به صورت زمان‌بلند در نظر گرفته شده باشد مقادیر معیار PESQ برای مجموعه تست A از Aurora-2 محاسبه شده است و در شکل ۸ نشان داده شده است.

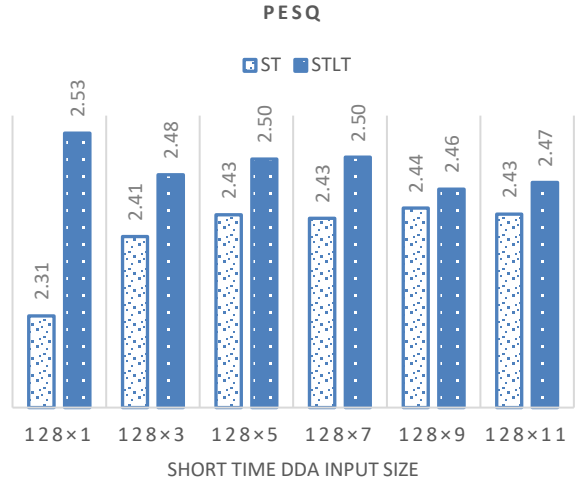
با توجه به این نتایج مشاهده می‌شود که دقت بدست‌آمده برای میانگین داده‌های مجموعه تست A برابر با مقدار ۲/۴۷ است که از بهترین دقت بدست‌آمده در حالت زمان‌کوتاه مبتنی بر نمودار شکل ۷ یعنی ۲/۴۳۹ بهتر است. به‌عبارتی استفاده از اطلاعات زمان‌بلند گفتاری در بهسازی گفتار اثری مثبت داشته است.



شکل ۸: مقدار معیار PESQ برای مجموعه testA از Aurora-2 در نسبت آغشتگی به نویزهای مختلف برای DDA آموزش داده‌شده با ورودی زمان‌بلند.

۵-۳- بهسازی دو مرحله‌ای

مدل دو مرحله‌ای پیشنهادی در بخش ۳ نیز در این قسمت مورد ارزیابی قرار گرفته است. برای این منظور ۶ مدل خودرمزگذار عمیق کاهنده نویز زمان‌کوتاه (DDA_ST) با اندازه ورودی‌های متغیر از ۱ تا ۱۱ فریم آموزش داده شده است و برای هر کدام از آنها خروجی بهسازی دادگان آموزشی محاسبه شده است. سپس مدل خودرمزگذار عمیق کاهنده نویز زمان‌بلند (DDA_LT) بر روی آن اعمال شده است و خروجی آن را به‌صورت مخفف (STLT) نشان داده شده است. با محاسبه مقدار PESQ به عنوان معیار ارزیابی بهسازی نمودارهای نشان داده‌شده در شکل ۹ حاصل شده است.

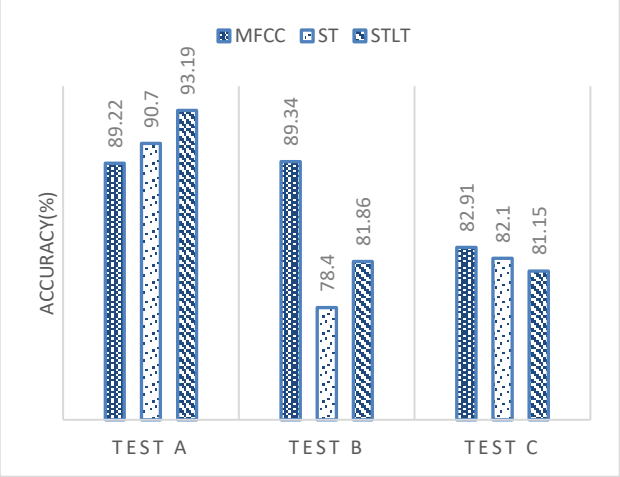


شکل ۹: مقایسه مقدار میانگین معیار PESQ در مدل زمان‌کوتاه (ST) و مدل دو مرحله‌ای پیشنهادی (STLT) برای مجموعه testA از Aurora-2 در انواع اندازه ورودی مدل زمان‌کوتاه مختلف.

با توجه به این نتایج بدست‌آمده در شکل ۹ مشاهده می‌شود که دقت مدل پیشنهادی دو مرحله‌ای بهسازی زمان‌کوتاه با ورودی تک فریم و سپس بهسازی زمان‌بلند بیشترین بهبود را در میانگین معیار PESQ ایجاد نموده است که نشان دهنده موفقیت روش پیشنهادی است.

همچنین میزان بهبود ایجاد شده در راستای بازشناسی گفتار نیز مورد بررسی قرار گرفته است. بدین صورت که ابتدا سیگنال گفتار نویزی توسط مدل زمان‌کوتاه و مدل زمان‌بلند بهسازی‌شده است و سپس از سیگنال بهسازی‌شده ویژگی‌های گفتاری استخراج شده است. ویژگی‌های گفتاری استخراج شده ۱۳ ویژگی کپسترال (MFCC) بدون در نظرگرفتن مشتقات اول و دوم هستند که جهت مقایسه بهتر میزان بهسازی صورت‌گرفته توسط شبکه عصبی، فرایند نرمال سازی میانگین کپسترال (CMN) اعمال نشده است. مقدار ضریب پیش‌تاکید ۰/۹۷، از پنجره همینگ با طول ۲۵ میلی ثانیه و شیفت فریم ۱۰ میلی ثانیه استفاده شده است.

بازشناسی گفتار با آموزش مدل‌های بازشناسی توسط دادگان آموزشی چند حالته که بهسازی شده‌اند صورت پذیرفته است. میانگین دقت بدست‌آمده برای مجموعه‌های تست A، B و C در انواع نویز و میزان آغشتگی به نویز ۰ تا ۲۰ دسیبل محاسبه شده است که در شکل ۱۰ نشان داده شده است.



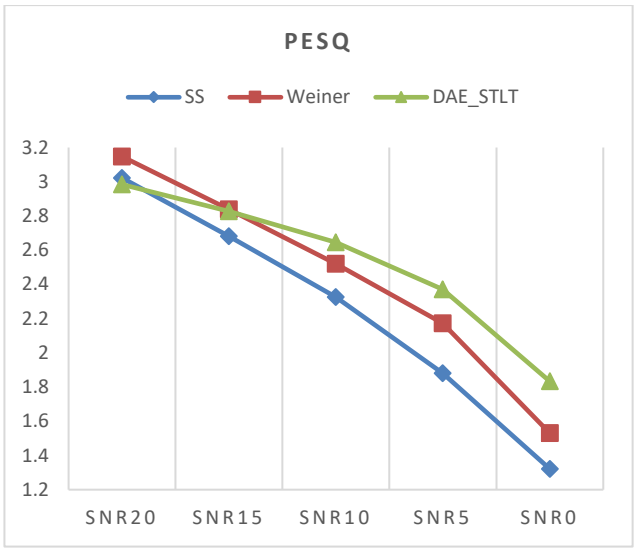
شکل ۱۰: مقایسه میانگین دقت بازشناسی گفتار بدون بهسازی (MFCC)، مدل زمان‌کوتاه(ST) با مدل پیشنهادی دو مرحله‌ای(STLT) در انواع نویز سه مجموعه تست A، B و C از Aurora-2 و در میزان اغشتگی به نویز ۰ تا ۲۰ دسیبل.

با توجه به نتایج بدست‌آمده در شکل ۱۰ مشخص است که روش پیشنهادی دو مرحله‌ای بهبود بیشتری را جهت بازشناسی گفتار نسبت به مدل زمان‌کوتاه ایجاد نموده است. همچنین مشخص است که دقت بازشناسی گفتار بدست‌آمده در مجموعه داده‌های تست B و C تفاوت زیادی با دقت بدست‌آمده در مجموعه داده تست A دارند که دلیل آن را می‌توان در تفاوت بین نوع نویزهای آموزش داده شده و استفاده شده در مرحله بهسازی دانست. چرا که نوع نویزهای بکاررفته در مجموعه داده تست A همان نوع نویزهای بکاررفته در آموزش است ولی نوع نویزهای بکاررفته در مجوعه تست B و C متفاوت است.

در نهایت بهسازی انجام شده با روش پیشنهادی دو مرحله‌ای با دو تا از روش‌های رایج بهسازی گفتار بنام روش تفاضل طیفی و روش فیلتر وینر مقایسه شده است که نتایج آن در شکل ۱۱ نشان داده شده است.

روش فیلتر وینر استفاده شده روش پیشنهاد شده توسط آقای S. Viahri و همکاران است که در سال ۲۰۱۶ پیشنهاد شده است و شامل پس‌پردازش بهسازی هارمونیک نیز می‌باشد [۲۶].

با توجه به نتایج بدست‌آمده در شکل ۱۱ مشخص است که دقت روش پیشنهادی (DDA_STLT) در نسبت‌های اغشتگی به نویز بالا کاملا از روش‌های رایج بهسازی گفتار مانند تفاضل طیفی (SS) و فیلتر وینر (Weiner) بهتر بوده است. در نسبت‌های اغشتگی به نویز پایین نیز اختلاف این روش‌ها با هم اندک است.



شکل ۱۱: مقایسه بهبود ایجادشده توسط روش بهسازی دو مرحله ای پیشنهادی مبتنی بر DDA با روش‌های بهسازی تفاضل طیف و فیلتر وینر.

۶- نتیجه‌گیری و بحث

فراگیری نگاشت بین داده نویزی شده و داده تمیز در بهسازی دادگان گفتاری و مقاوم‌سازی روش‌های بازشناسی گفتار از اهمیت بسیار بالایی برخوردار است که امروزه شبکه عصبی عمیق توانمندی خود را در ان زمینه به اثبات رسانده است.

در این مقاله روشی دو مرحله‌ای برای بهسازی گفتار با استفاده از شبکه خودرمزگذار عمیق کاهنده نویز ارائه شده است. در این روش ابتدا ویژگی‌های طیفی سیگنال گفتار استخراج و توسط شبکه خودرمزگذار عمیق کاهنده نویز تاثیرات نویز از آن حذف خواهد شد و سپس سیگنال گفتار اصلی با کمک طیف بهسازی شده ساخته خواهد شد.

در روش پیشنهادی در مرحله نخست بهسازی زمان‌کوتاه را انجام می‌دهد و سپس در مرحله دوم بهسازی زمان‌بلند. بهسازی زمان‌کوتاه برای حذف تخریب ایستان نویز و بهسازی زمان بلند برای حذف تخریب‌های غیرایستان می‌تواند موثر باشد.

در این مقاله سعی شده است تا با استفاده از کاهش تعداد زیرباندهای فرکانسی در فواصل زمانی بلند، امکان اعمال ورودی‌های زمان بلند برای شبکه عصبی خودرمزگذار عمیق کاهنده نویز فراهم گردد. در آزمایش‌های انجام‌شده مشخص گردید که استفاده از اطلاعات زمان‌بلند گفتاری اثری مثبت بر بهسازی گفتار دارد.

نتایج آزمایش‌های انجام شده در این مقاله بر روی مجموعه دادگان Aurora-2 نشان داده است که روش پیشنهادی می‌تواند از نظر بهسازی گفتار و معیار PESQ نسبت به فیلتر وینر در شرایط اغشتگی به نویز بالا به میزان ۰/۳ بهبود ایجاد کند و هم از نظر دقت بازشناسی خودکار گفتار نسبت به شبکه خودرمزگذار عمیق کاهنده نویز با ورودی زمان‌کوتاه حدود ۲٪ بهبود ایجاد کند.

هرچند این روش به‌عنوان یک روش بهســـازی تک کانال معرفی و ارزیابی شـــده اســـت اما قابلیت تعمیم به بهســـازی چندکاناله را نیز دارا می‌باشد که می‌تواند به‌عنوان یک کار آتی در نظر گرفته شود.

مراجع

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process., vol. 27, no. 2, pp. 113–120, 1979.

[2] K. K. Ravi and P. V. Subbaiah, “A survey on speech enhancement methodologies,” Int. J. Intell. Syst. Appl., vol. 8, no. 12, p. 37, 2016.

[3] V. Sunnydayal, N. Sivaprasad and T. K. Kumar, “A survey on statistical based single channel speech enhancement techniques,” Int. J. Intell. Syst. Appl., vol. 6, no. 12, p. 69, 2014.

[4] I. Cohen and B. Berdugo, “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” IEEE Signal Process. Lett., vol. 9, no. 1, pp. 12–15, 2002.

[5] I. Cohen, “Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator,” IEEE Signal Process. Lett., vol. 9, no. 4, pp. 113–116, 2002.

[6] Y. Ephraim and I. Cohen, “Recent advancements in speech enhancement,” Circuits Signals Speech Image Process., 2006.

[7] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Acoust. Speech Signal Process., vol. 33, no. 2, pp. 443–445, 1985.

[8] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Trans. Acoust. Speech Signal Process., vol. 32, no. 6, pp. 1109–1121, 1984.

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

[10] D. Wang, “Time-Frequency masking for speech separation and Its potential for hearing aid design,” Trends Amplif., vol. 12, no. 4, pp. 332–353, 2008.

[۱۱] حسین شایقی، علی قاسمی، «پیش‌بینی قیمت روزانه برق با شبکه عصبی بهبودیافته مبتنی بر تبدیل موجک و روش آشوبناک جستجوی گرانشی»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۵، شماره ۴، شماره صفحه ۱۰۳-۱۱۳، زمستان ۱۳۹۴.

[۱۲] فرید کربلایی، حمیدرضا شعبانی، رضا ابراهیم‌پور، «ارزیابی برون‌خط پایداری گذرا به وسیله تعیین دقیق CCT با استفاده از شبکه عصبی با ورودی‌های مبتنی بر توابع انرژی»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۶، شماره ۱، شماره صفحه ۲۷۷-۲۸۵، زمستان ۱۳۹۵.

[13] Y. Xu, J. Du, L. R. Dai and C. H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 23, no. 1, pp. 7–19, Jan. 2015.

[14] F. Weninger et al., “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in International Conference on Latent Variable Analysis and Signal Separation, 2015, pp. 91–99.

[15] B. Li, Y. Tsao and K. C. Sim, “An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition.,” in Proceedings of Interspeech 2013, pp. 3002–3006, 2013.

[16] Z. Chen, S. Watanabe, H. Erdogan and J. R. Hershey, “Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks,” Unkn. J., vol. 2015–January, pp. 3274–3278, 2015.

[17] L. Deng and D. Yu, " Deep learning: methods and applications", Foundations and Trends® in Signal Processing: Vol. 7: No. 3–4, pp 197-387, 2014.

[18] L. Dehyadegary, S. Ali Seyyedsalehi and I. Nejadgholi, “Nonlinear enhancement of noisy speech, using continuous attractor dynamics formed in recurrent neural networks,” Neurocomputing, vol. 74, no. 17, pp. 2716–2724, Oct. 2011.

[19] S. Tan and K. C. Sim, “Learning utterance-level normalisation using Variational Autoencoders for robust automatic speech recognition,” in 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 43–49, 2016.

[20] X. Lu, Y. Tsao, S. Matsuda and C. Hori, “Speech enhancement based on deep denoising autoencoder.,” in Interspeech, pp. 436–440, 2013.

[21] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Process. Lett., vol. 21, no. 1, pp. 65–68, 2014.

[22] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” science, vol. 313, no. 5786, pp. 504–507, 2006.

[23] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P.-A. Manzagol, “Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion,” J Mach Learn Res, vol. 11, pp. 3371–3408, Dec. 2010.

[24] T. Gao, J. Du, Y. Xu, C. Liu, L.-R. Dai and C.-H. Lee, “Improving deep neural network based speech enhancement in low SNR environments,” in International Conference on Latent Variable Analysis and Signal Separation, pp. 75–82, 2015.

[25] D. Pearce, H. Hirsch and E. E. D. Gmbh, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in in ISCA ITRW ASR2000, pp. 29–32, 2000.

[26] S. Vihari, A. S. Murthy, P. Soni and D. C. Naik, “Comparison of speech enhancement algorithms,” Procedia Comput. Sci., vol. 89, no. Supplement C, pp. 666–676, Jan. 2016.

زیرنویس‌ها

[1] Spectral Subtraction

[2] Wiener Filter

[3] Minimum Mean Square Error

[4] Variational Autoencoder

[5] Auto-Encoder

[6] Denoising Auto-Encoder

[7] Encoder

[8] Decoder

[9] Manifold

[10] Feature Learning

[11] Deep Denoising Auto-Encoder

[12] Stereo Data

[13] Mini-batch size