مقاله پژوهشی

بهبود کیفیت سیگنال گفتار گسترش‌یافته با افزایش مقاوم‌سازی شبکه در مقابل نویز با استفاده از شبکه عصبی پیچشی

معصومه شفیعیان *
استادیار دانشکده فنی و مهندسی رسانه
دانشگاه صداوسیما
shafieian@iribu.ac.ir

محمدحسین محبی
کارشناسی ارشد مهندسی صدا
دانشگاه صداوسیما
mhchr777@gmail.com

تاریخ دریافت: ۱۴۰۳/۰۵/۳۱ تاریخ پذیرش: ۱۴۰۴/۰۳/۰۱

چکیده

هدف از سیستم افزایش پهنای‌باند صدا، افزایش محدوده فرکانسی سیگنال صدای ورودی سیستم است. چون با افزایش پهنای‌باند صدا در واقع کیفیت و وضوح صدا بیشتر خواهد شد، به این دلیل به این کار افزایش وضوح گفتار هم گفته می‌شود. یکی از کاربرد های بسیار مرسوم گسترش پهنای باند سیگنال صدا در سیستم تلفن است، که در آن پهنای باند سیگنال صدا زیر ۴کیلوهرتز است که سبب کاهش کیفیت صدای دریافت شده در گیرنده می‌شود. بنابراین هدف از گسترش پهنای‌باند صدا این است که صداهایی که گنگ و نامفهوم هستند، قابلیت درک بیشتری نسبت به قبل داشته باشند.

در این پژوهش، ابتدا به عنوان مرحله پیش‌پردازش، نویز جمع‌شونده، به عنوان عامل مخرب سیگنال، به مجموعه دادگان DAPS[۱] داده می‌شود و سپس با استفاده از روش نویزکاهی تبدیل‌موجک، سیگنال صدا نویزکاهی[۲] می‌شود و پس از آن سیگنال به شبکه عصبی پیچشی[۳] با تابع زیان هوبر[۴] داده می‌شود تا سیگنال باندپهن گسترش‌یافته تولید شود. در نهایت مدل پیشنهادی از لحاظ معیارهای SNR[۵] ، LSD[۶] ، PESQ[۷] و STOI[۸]، مورد ارزیابی قرار گرفت و با چندین روش مشابه در زمینه گسترش پهنای‌باند صدا، مقایسه شد. شبکه پیشنهادی توانست نسبت به سایر روش‌ها نظیر FFTNet[۹] مقدار ۲/۶ واحد برای LSD و مقدار ۴/۲ دسی‌بل برای مقدار SNR، همچنین نسبت به شبکه KUL[۱۰]، مقدار ۹ واحد برای LSD و مقدار ۳/۵ دسی‌بل برای SNR، در شرایط آزمایش با دادگان یکسان و میزان گسترش برابر، نتایج بهتری داشته باشد.

واژگان کلیدی: گسترش پهنای باند ، افزایش کیفیت گفتار، شبکه عصبی پیچشی ، تابع زیان هوبر، روش کاهش نویز تبدیل موجک

۱. مقدمه

تا چند سال پیش، کیفیت ارتباط از راه دور صوتی بوسیله انتخاب المان‌های طراحی در ۱۰۰ سال پیش محدود شده بود، که منجر به استفاده از نرخ نمونه‌برداری ۸ کیلوهرتز و در محدوده فرکانس عملی ۳۰۰-۴۰۰۰ هرتز شد. این فرکانس به اصطلاح باند باریک[۱۱] (NB)، کیفیت گفتار را به شدت محدود می‌کند. امروزه، مردم جهان به انتقال و گفتگو با «صدای HD[۱۲]» و «صدای UHD[۱۳]» علاقمند هستند. این هدف، مستلزم استفاده از کدکننده‌های باند پهن[۱۴] (WB) یا فوق باند پهن[۱۵] (SWB) است که به ترتیب از نرخ‌های نمونه برداری ۱۶ کیلوهرتز یا ۳۲ کیلوهرتز استفاده می‌کنند و مربوط به محدوده فرکانس ۵۰-۸۰۰۰ هرتز یا ۵۰-۱۴۰۰۰ هرتز یا حتی ۴۴۱۰۰ هرتز، بسته به نوع کاربرد، هستند.

با این حال، توسعه‌های WB و SWB همه جا موجود نیستند، زیرا نیاز به هزینه‌های قابل‌توجهی برای توسعه، آزمایش و استقرار زیرساخت های خدماتی دارند. علاوه بر این، تماس‌های انتها به انتها[۱۶] WB / SWB نیازمند دستگاه‌های ارتقا یافته در هر دو انتها هستند. فراهم کردن چنین زیر ساخت‌هایی سال‌های زیادی طول خواهد کشید تا پوشش کامل حاصل شود و ارتقا شبکه‌های خطوط تلفن ثابت به WB / SWB حتی بیشتر طول خواهد کشید؛ تا آن زمان تعداد قابل توجهی از تماس‌ها هنوز هم از باند باریک استفاده خواهند کرد.

هدف فن‌آوری گسترش / افزایش پهنای باند صدا [۱۷]، حل این مشکل از طریق تبدیل صداهای باند باریک به باند پهن و افزایش وضوح صداست. از این رو به روش‌های گسترش پهنای باند گفتار، روش‌های افزایش وضوح گفتار[۱۸] هم گفته می‌شود.

گسترش پهنای باند را می‌توان در حوزه فرکانس یا در حوزه زمان پیاده‌سازی کرد. پردازش حوزه فرکانس معمولاً طیف فریم ورودی را با استفاده از تبدیل سریع فوریه[۱۹] (FFT) محاسبه می‌کند. پردازش حوزه‌زمانی معمولاً شامل فیلترهای تطبیقی یا یک بانک فیلتر برای شکل دادن به طیف سیگنال گسترش‌یافته است. پردازش حوزه‌فرکانس از مزیت دسترسی مستقیم به نمایش طیفی برخوردار است، در حالی که پردازش در حوزه‌زمان این مزیت را دارد که تاخیر کلی کم‌تری ایجاد می‌کند و همچنین به دلیل عدم نیاز به محاسبات FFT، خطای احتمالی را کمینه می‌کند [۱].

گسترش پهنای باند که در پژوهش‌ها به عنوان وضوح بالای صوتی نیز از آن یاد می‌شود، به دلیل اهمیت آن در بسیاری از سیستم‌ها برای دهه‌ها موضوع مطالعه بوده است [۲-۱۰]. مطالعات اولیه بر تخمین پوش طیفی باند فرکانس بالا و استفاده از تحریک تولیدشده از باند فرکانس پایین برای بازیابی طیف فرکانس بالا متمرکز بود [۱۱]. تکنیک‌های سنتی مانند مدل‌های مخلوط گاوسی[۲۰] ، LPC[۲۱]، و HMM[۲۲] نیز مورد استفاده قرار گرفته‌اند [۱۲-۱۴]. با این حال، این روش‌ها به طور کلی در مقایسه با شبکه‌های عصبی عملکرد بدتری دارند [۱۵].

پیشرفت‌های اخیر در شبکه‌های عصبی عمیق، سیگنال‌های باند پهن را مستقیماً از سیگنال‌های باند باریک بدون نیاز به استخراج ویژگی تولید می‌کنند. به عنوان مثال، [۱۶] مدلی پیشنهاد کرد که درآن یک شبکه عصبی عمیق، با استفاده از لگاریتم طیفی به عنوان ویژگی‌های ورودی و خروجی برای انجام تبدیل

غیرخطی مورد نیاز، به عنوان یک تابع نگاشت آموزش داده شد. این شبکه عصبی متراکم با سه لایه پنهان به اندازه ۲۰۴۸ و تابع فعال‌سازی ReLU، نشان داد که در ۸۴ درصد موارد در مطالعه انجام‌شده، نسبت به مدل‌های مخلوط گاوسی برتر است.

در روش گسترش پهنای‌باند به روش FFTNet، سیگنال در حوزه فرکانس پردازش می‌شود و نمونه nام از روی n-1 نمونه قبلی تخمین زده می‌شود و به این صورت عمل می‌کند که ابتدا تعداد نمونه‌ها را به ۲ قسمت تقسیم می‌کند و سپس نمونه‌ها را نظیر به نظیر، با هم جمع می‌کند تا داده جدید تولید شود و مجدداً نمونه‌های جدید را به دو قسمت تقسیم می‌کند و نظیر به نظیر با هم جمع می‌کند تا زمانی که به یک داده واحد برسد. از این رو به این روش جمع تقسیم بر۲ (2SS[۲۳]) گفته می‌شود. در پژوهش [۱۷]، بجای 2SS از 3SS[۲۴] استفاده شده است، تا سرعت همگرایی افزایش یابد. این شبکه، یک شبکه پیچشی است، که پهنای‌باند سیگنال را از ۸۰۰۰ هرتزی، به ۴۴۱۰۰هرتزی، گسترش می‌دهد. برای آموزش شبکه از مجموعه داده DAPS استفاده شده است و برای قسمت ارزیابی ذهنی، از تارک مکانیکی آمازون[۲۵] استفاده شده است. در ارزیابی‌های عینی، روش FFTNet توانست عملکرد خوبی از لحاظ LSD و SNR داشته باشد ولی با این حال نسبت به روش معرفی شده در [۱۸] عملکرد ضعیف‌تری داشته است.

WaveNet یک شبکه عصبی پیچشی است که برای تولید صدا و شکل موج خام سیگنال استفاده می‌شود. تفاوت WaveNet و FFTNet این است که، WaveNet بطور کلی بر روی سیگنال حوزه زمان پردازش می‌کند، ولی FFTNet سیگنال را با تبدیل فوریه به حوزه فرکانس می‌برد و پردازش‌ها در حوزه فرکانس روی سیگنال انجام می‌شود. در [۱۹] با استفاده از WaveNet سیگنال‌های با فرکانس ۸ کیلوهرتز به ۲۴ کیلوهرتز گسترش یافت و از مجموعه‌داده LibriTTS[26] جهت آموزش شبکه استفاده شده است. نتایج و سیگنال‌های تولیدی حاصل از WaveNet، نسبت به AMR-WB[۲۷]، وضوح و کیفیت بالاتری داشت که با توجه به استانداردهای GSM[28]، برای سیستم‌های موبایلی بسیار مناسب است و قابلیت پیاده‌سازی بر روی پردازنده‌های موبایلی را دارد.

با الهام از الگوریتم‌های وضوح تصویر، که از تکنیک‌های یادگیری ماشین برای درون‌یابی تصویری با وضوح پایین به تصویری با وضوح بالاتر استفاده می‌کنند، در [۲۰] یک U-Net پیچشی شامل بلوک‌های نمونه‌کاهی و نمونه‌افزایی متوالی با اتصالات پرش پیشنهاد شده است. بر اساس آن، یک جزء شبکه عصبی به نام مدولاسیون خطی ویژگی‌های زمانی (TFiLM[۲۹]) در [۲۱] معرفی شد. TFiLM با ترکیب عناصر رویکردهای پیچشی و بازگشتی[۳۰] در یک ساختار U-Net مانند، وابستگی‌های ورودی بلندمدت[۳۱] را در ورودی‌های متوالی ثبت می‌کند. علاوه بر این، فعال‌سازهای یک مدل پیچشی را با استفاده از اطلاعات بلندمدت ثبت‌شده توسط یک شبکه عصبی بازگشتی مدوله می‌کند. یک نوع بلوک آنلاین از مدل مدولاسیون خطی ویژگی زمانی (TFiLM) برای دست‌یابی به گسترش پهنای‌باند توسط [۲۲] پیشنهاد شد. این ساختار، رکن اصلی U-Net مدل TFiLM را برای کاهش زمان استنتاج ساده کرده و برای کاهش افت عملکرد از یک مبدل[۳۲] کارآمد در گلوگاه استفاده می‌کند. همچنین از پیش‌آموزش خود نظارت شده و افزایش دادگان برای

افزایش کیفیت سیگنال‌های با پهنای‌باند گسترش‌یافته و کاهش حساسیت نسبت به روش‌های نمونه‌کاهی استفاده می‌کند.

مدولاسیون خطی ویژگی مبتنی بر توجه (AFiLM[۳۳]) [۲۳] یک شبکه با معماری U-Net مانند برای وضوح بالای صوتی پیشنهاد کرد که پیچش و توجه به خود را ترکیب می‌کند. AFiLM به جای شبکه‌های عصبی بازگشتی از مکانیزم توجه‌به‌خود برای مدوله‌کردن فعال‌سازهای مدل پیچشی استفاده می‌کند.

در روش گسترش پهنای‌باند با استفاده از شبکه GAN[۳۴]، [۲۴] با استفاده از آموزش شبکه مولد تخاصمی با تابع زیان تخاصمی[۳۵]، برای نخستین بار، پهنای باند سیگنال گفتار باندباریک، به سیگنال باند-پهن، گسترش یافت. برای آموزش شبکه از مجموعه داده VCTK[۳۶] استفاده شده است. این شبکه، پهنای-باند سیگنال را از ۸۰۰۰ هرتزی، به ۱۶۰۰۰هرتزی، تحت عنوان ۲برابر گسترش‌باند و از ۴۰۰۰هرتزی به ۱۶۰۰۰هرتزی تحت عنوان ۴برابر، گسترش می‌دهد. این شبکه در ارزیابی‌های عینی و ذهنی، از لحاظ معیار ارزیابی ادراکی کیفیت صدا (PESQ[۳۷])، عملکرد خوبی داشت و توانست به عدد قابل قبول ۴/۳۲ برسد. اما این روش در ارزیابی LSD و SNR، نتوانست به خوبی معیار PESQ برسد.

شبکه‌های مولد تخاصمی (GANs) در [۲۵] به کار گرفته شدند، که در آن جزء سوم اضافی در خط TTS[۳۸] برای نمونه‌افزایی عصبی پیشنهاد شده است. این روش صدا با وضوح پایین‌تر (۱۶-۲۴ کیلوهرتز) را به صدای با وضوح کامل (۴۴/۱ کیلوهرتز) تبدیل می‌کند. علاوه بر این، محققان استفاده از یک مدل احتمالی انتشار[۳۹] برای وضوح بالای صوتی را مورد بررسی قرار دادند که بر اساس رمزگذارهای صوتی عصبی طراحی شده است [۲۶].

در روش گسترش پهنای‌باند به روش تحلیل دو مرحله‌ای [۲۷] محققان با استفاده از شبکه TCN[۴۰] و CRN[۴۱] در بخش اول جهت بازسازی دامنه، سپس استفاده از یک شبکه Wave – U – Net و استفاده از تبدیل‌فوریه چند دقتی[۴۲] به منظور بهره‌مندی از مزایای هم زمان پردازش حوزه زمان – فرکانسی، سعی در بهبود کیفیت سیگنال گسترش‌یافته داشتند. پردازش حوزه فرکانس جهت بازسازی دامنه سیگنال و پردازش حوزه فرکانس جهت بازسازی فاز سیگنال بکار رفت. این شبکه، پهنای باند سیگنال را از ۸۰۰۰ هرتزی، به ۱۶۰۰۰هرتزی، گسترش می‌دهد. شبکه با تابع زیان MSE[۴۳] و مجموعه داده VBC[۴۴] آموزش دید. این روش توانست سیگنال باند باریک را بهتر نسبت به روش‌های DNN[۴۵] و شبکه TFiLM، بازسازی کند و در معیارهای ارزیابی PESQ و LSD عملکرد بهتری داشته باشد. نتایج نشان داد، استفاده از پردازش همزمان زمان-فرکانسی، روی LSD نتیجه مطلوبی در پی داشته است و در ضمن میزان SNR به صورت نسبی حفظ شد.

در روش گسترش پهنای باند بلادرنگ[۴۶] [۲۸] از یک شبکه سبک جهت پردازش حوزه فرکانس، به نام SEANet[۴۷] که یک شبکه مولد تخاصمی است، به منظورگسترش پهنای باند استفاده شده است. از مزایای این شبکه می‌توان به پردازش سریع و استفاده از پردازنده‌های معمول، حتی در حد پردازنده‌های موبایل جهت پردازش اشاره کرد. دیگر مزایای مهم این شبکه تاخیر بسیار کم است، که این ویژگی به بلادرنگ نمودن پردازش کمک می‌کند. علت این امر، استفاده از تابع

فعال ساز ELU و همچنین استفاده از تابع زیان تخاصمی بصورت همزمان در شبکه است. این شبکه، پهنای‌باند سیگنال را از ۸۰۰۰ هرتزی، به ۱۶۰۰۰هرتزی، گسترش می‌دهد. جهت پردازش، از مجموعه داده VCTK استفاده شده است. نتایج حاصله بیانگر این بود که، شبکه SEANet بصورت معمولی، عملکرد بهتری نسبت به همین شبکه در حالت بلادرنگ دارد.

در روش گسترش پهنای‌باند با استفاده از تابع زیان زمان-فرکانسی با استفاده از یک شبکه پیچشی (پیچشی) خودکدگذار[۴۸] [۲۹]، از یک تابع زیان زمان-فرکانسی نیز استفاده شده و پهنای‌باند سیگنال باند-باریک گسترش‌یافته است. این شبکه، پهنای باند سیگنال را از ۸۰۰۰ هرتزی، به ۱۶۰۰۰هرتزی، گسترش می‌دهد. این شبکه با استفاده از مجموعه داده TIMIT[49] آموزش دید و تست شد، که توانست از لحاظ معیارهای LSD و SNR نسبت به روش‌های دیگر برتری داشته باشد.

یک مدل مولد شکل‌موج مبتنی‌بر Glow برای ایجاد وضوح[۵۰] بالای صوتی در [۳۰] پیشنهاد شد. به طور خاص، ادغام WaveNet و Glow به طور مستقیم احتمال دقیق صدای هدف با وضوح بالا (HR[۵۱]) را مشروط به اطلاعات با وضوح پایین (LR[۵۲]) به حداکثر می‌رساند. برای استخراج اطلاعات صوتی از صدای با وضوح پایین، یک رمزگذار صوتی LR و یک رمزگذار STFT[53] پیشنهاد شدند که اطلاعات LR را به ترتیب از حوزه زمان و حوزه فرکانس رمزگذاری می‌کنند.

یک روش وضوح بالای گفتار مبتنی بر رمزگذارصوتی عصبی (NVSR[۵۴]) توسط [۳۱] پیشنهاد شد، که قادر به مدیریت ورودی‌های با وضوح مختلف و نسبت‌های گوناگون نمونه‌افزایی است. NVSR از یک ساختار سیستم پشت‌سرهم پیروی می‌کند که شامل یک ماژول گسترش پهنای‌باند mel، یک ماژول رمزگذار صوتی عصبی و یک ماژول پس‌پردازش است.

یک مدل مولد مبتنی بر انتشار برای ایجاد وضوح بالای صوتی مقاوم در طیف گسترده‌ای از انواع نمونه‌های صوتی به نام AudioSR طراحی شده که در [۳۲] پیشنهاد شده است. AudioSR به طور خاص برای افزایش کیفیت جلوه‌های صوتی، موسیقی و گفتار طراحی شده است.

هدف اصلی گسترش پهنای‌باند سیگنال صدا، ساخت قسمت بالای فرکانسی سیگنال باندباریک، در محدوده ۴ تا ۱۶کیلوهرتز یا حتی فرکانس‌های بالاتر تا ۲۴ یا ۴۴/۱ کیلوهرتز است که این به معنی بهبود کیفیت سیگنال است. در این مقاله، هدف اینست که با شبکه عصبی پیچشی و همچنین استفاده از روش‌های کاهش نویز، با پیگیری اهداف:

- استفاده از آموزش چندمرحله‌ای شبکه عصبی پیچشی در گسترش پهنای‌باند از فرکانس ۸۰۰۰ به ۴۴۱۰۰ هرتز
- کاهش نویز مجموعه داده با استفاده از روش تبدیل موجک و روش آستانه‌گذاری
- افزایش مقاومت شبکه در برابر نویز با استفاده از تابع زیان هوبر در شبکه عصبی پیچشی

بهبودی در نتایج پژوهش‌های پیشین، حاصل شود.

۲. روش پیشنهادی

در این پژوهش، مشابه پژوهش‌های سال‌های اخیر، با فرض مناسب بودن و تعدد نسبتاً زیاد روش‌های گسترش پهنای‌باند صدا از ۴۰۰۰ به ۸۰۰۰ هرتز، از

این قسمت صرف‌نظر شده و محدوده جدیدتر گسترش یعنی، از فرکانس ۸۰۰۰ به ۴۴۱۰۰ هرتز مورد مطالعه قرار می‌گیرد و سعی خواهد شد تا در کیفیت سیگنال‌های صدا در محدوده ۸۰۰۰ هرتزی که کیفیت چندان مناسبی ندارند، با گسترش به محدوده فرکانس بالاتر (۴۴۱۰۰ هرتز)، بهبود حاصل شود.

یکی از چالش‌های مهم در زمینه پردازش سیگنال و شبکه‌های عصبی و یادگیری ماشین، افزایش مقاومت در برابر نویز و یا کاهش نویز موجود در سیگنال است. وجود نویز، همواره به عنوان عامل مخرب سیگنال و شبکه، اجتناب ناپذیر است و به نوعی باید شبکه در برابر نویز مقاوم شود تا از اثر آن برروی شبکه کاسته شود. از سوی دیگر، ممکن است بسیاری از سیگنال‌هایی که برای گسترش پهنای‌باند مورد استفاده قرار می‌گیرند، در شرایط نویزی ضبط شده باشند یا اینکه سیگنال‌های ضبط شده توسط دستگاه‌های قدیمی باشند که به مرور زمان دچار آسیب شده و حالت نویزی‌گونه به خود گرفته باشند. پس راهکارهای مقابله با نویز در این پژوهش به دو دسته تقسیم می‌شوند:

۱. از بین بردن نویز موجود در مجموعه داده
۲. افزایش مقاومت شبکه در برابر نویز

نویز موجود در دادگان با استفاده از روش‌های نویزکاهی، کاهش داده می‌شود. روش مورد استفاده این پژوهش در این زمینه، استفاده از روش تبدیل موجک و روش آستانه‌گذاری است. جهت افزایش مقاومت شبکه در برابر نویز، از تابع زیان هوبر استفاده می‌شود که در واقع ترکیبی از توابع MSE و MAE[۵۵] است و خاصیت مقاومت در برابر نویز دارد.

ساختار کلی روش پیشنهادی جهت گسترش پهنای‌باند سیگنال صدا، از دو بخش اصلی تشکیل شده است:

۱. به عنوان مرحله پیش‌پردازش: کاهش نویز موجود در مجموعه داده با استفاده از روش آستانه کاهش نویز موجک[۵۶].
۲. استفاده از شبکه عصبی پیچشی جهت گسترش پهنای‌باند سیگنال صدا.

۳. آماده سازی دادگان

در این پژوهش، از دادگان DAPS استفاده شده است. این مجموعه داده، دارای ۱۵ نسخه شامل نسخه‌های حرفه‌ای و نسخه‌های در محیط واقعی است. هر نسخه شامل حدود ۴/۵ ساعت گفتار است که توسط ۲۰ گوینده مختلف گفته‌شده است. گویندگان با نام‌های F1، F2، F3 و ... نشان دهنده گویندگان زن و M1، M2، M3 و ... نشان دهنده گویندگان مرد هستند.

برای استفاده از این مجموعه داده، ابتدا داده مجموعه F1 (به عنوان مثال) به بخش‌های کوچک‌تر ۲ثانیه‌ای تقسیم می‌شوند، که حدود ۸۰۰۰ داده ۲ ثانیه‌ای تولید می‌شود. با این کار، هم در بخش نویزکاهی، به دلیل کاهش طول سیگنال، و هم در بخش آموزش و تست شبکه، می‌توان سرعت پردازش شبکه را در دوره[۵۷]ها افزایش داد. دلیل انتخاب F1، نزدیک کردن شرایط آموزش و تست، به شرایط موجود در پژوهش [۱۷] ، به جهت ارزیابی و مقایسه بهتر است.

در مرحله بعدی، از آنجایی که در علم مخابرات عموماً نویزی‌کردن سیگنال با نرخ SNR، ۱۵ الی ۲۰ دسی‌بل انجام می‌شود [۳۳]، در این پژوهش نیز، دادگان با نسبت سیگنال به نویز ۱۵dB، نویزی شده و سپس

دادگان آماده ورود به الگوریتم نویزکاهی و سپس آموزش توسط شبکه عصبی است.

۴. استفاده از روش نویزکاهی به عنوان مرحله پیش‌پردازش

استفاده از داده‌های تمیز در شبکه، جهت گسترش پهنای باند سیگنال، کمی ایده‌آل‌گرایانه به نظر می‌رسد. زیرا در واقعیت بسیاری از داده‌هایی که هدف گسترش پهنای باند آنهاست، ممکن است داده‌هایی باشند که پیشتر در محیطی نویزی (هر چند با نسبت سیگنال بر نویز زیاد) ضبط شده‌اند و بر اساس شرایط امکان ضبط آن تنها یک‌بار فراهم بوده است؛ یا با توجه به ضعف تکنولوژی در زمینه ضبط صدا در گذشته، صداها، نویزی ضبط شده‌اند. پس روش نویزکاهی، به عنوان پیش‌پردازش در زمینه گسترش پهنای‌باند، می‌تواند برای گسترش پهنای‌باند داده‌های واقعی که حامل نویزی هر چند کم هستند، مفید واقع شود.

روش‌های بسیاری در سال‌های اخیر برای کاهش نویز مورد استفاده قرار گرفته‌اند، از جمله: استفاده از شبکه عصبی، روش‌های شکل‌دهی پرتو[۵۸]، روش تفریق طیفی[۵۹] و غیره. یکی دیگر از این روش‌ها، استفاده از روش کاهش نویز به روش تبدیل موجک است. چون در این پژوهش از یک شبکه در جهت گسترش پهنای‌باند (که در بخش بعدی توضیح داده می‌شود) استفاده شده‌است، از شبکه عصبی، جهت نویزکاهی استفاده نشده است.

۴-۱. نویزی کردن سیگنال

هدف از نویزی کردن سیگنال، این است که سیگنال، مشابه سیگنال‌هایی شود که بعنوان مثال، در محیط‌های پر سروصدا ضبط شده‌اند، یا صداهای ذخیره‌شده در فایل‌های آرشیویی که روی سخت‌افزارهای آنالوگ ذخیره‌شده‌اند و به مرور زمان و آسیب‌های خارجی، حالت نویزگونه به خود گرفته‌اند. بنابراین، چون دسترسی به طیف وسیعی از این مجموعه داده‌ها برای تست در شبکه وجود نداشت، سعی شد شبیه‌سازی حالت نویزی‌گونه برروی مجموعه داده‌های تمیز انجام شود.

لذا با اضافه کردن نویز سفید جمع‌شونده گاوسی[۶۰] به داده‌های تمیز، با نسبت سیگنال به نویز مشخص، تا حدودی می‌توان به این هدف رسید و خروجی شبکه را به واقعیت نزدیک نمود. بنابراین، سیگنال تمیز با نسبت سیگنال بر نویز ۱۵dB، نویزی شد. اگرچه نویز انواع گوناگونی دارد، مانند: نویز خیابانی، نویز دفترکار، نویز بوق[۶۱]و غیره، ولی در این پژوهش به عنوان گام نخست در نویزی‌کردن سیگنال در پژوهش‌های گسترش پهنای‌باند، صرفاً از نویز سفید جمع شونده گاوسی استفاده شد.

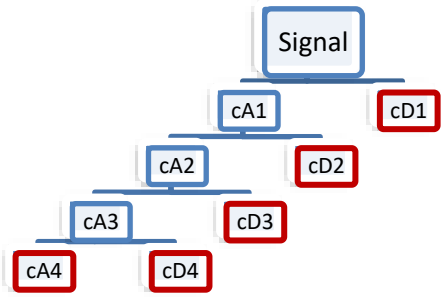
۴-۲. اعمال تبدیل موجک

پس از نویزی کردن سیگنال، تبدیل موجک از سیگنال گرفته می‌شود. موجک استفاده شده در تبدیل، از خانواده سیملت[۶۲] و نوع آن Sym4 می‌باشد. دلیل استفاده از این موجک، تقارن نسبی و همچنین قابلیت بازشدگی آن در زمان نسبت به بقیه موجک‌ها است؛ در نتیجه برای تحلیل فرکانس‌های بالا مناسب‌تر است. بطور کلی ۷ درجه سیملت تعریف شده است و در این پژوهش از Symlet4 استفاده شده است. بدیهی است هر چه از سیملت درجه بالاتری استفاده شود، دقت

تحلیل موجک، بیشتر می‌شود و در نتیجه، سرعت پردازش کمتر خواهد شد.

درجه یا سطح [۶۳] هر موجک نشان‌دهنده تعداد نمونه‌هایی است که موجک با آن تقریب زده می‌شود. هرچه درجه موجک بالاتر باشد، تقریب موجک بهتر و در نتیجه شکل موجک، صاف تر خواهد بود. در این مقاله از درجه ۴ تبدیل موجک استفاده شد، پس در ۴ سطح تبدیل موجک بر روی سیگنال اعمال می‌شود و ضرایب تبدیل موجک، شامل۴ سیگنال تحت عنوان ضرایب جزئی[۶۴] (cD) و یک سیگنال تحت عنوان ضریب تقریب[۶۵] (cA) تولید می‌شود. در هر مرحله، تبدیل موجک روی ضریب تقریب، اعمال می‌شود و مجدد ضریب جزئی و ضریب تقریب مرحه بعد تولید می‌شود. پس از بدست آوردن ضرایب تبدیل موجک و تجزیه شدن سیگنال به ضرایب cA و cD، این ضرایب، که به نوعی سیگنال‌های کوچک‌تر از سیگنال اصلی نویزی هستند، به مرحله آستانه‌گذاری می‌روند تا نویز آن‌ها حذف شوند. به دلیل این‌که ضرایب جزئیات، نشان‌دهنده قسمت فرکانس بالا در سیگنال هستند، در نتیجه می‌توان گفت که در اثر حذف و یا تخریب تعدادی از ضرایب جزئیات، بخشی از طیف فرکانسی فیلتر می‌شود. اگر در سیگنال، نویزهای فرکانس بالای زیادی وجود داشته باشند، می‌توان با آستانه‌گذاری، شدت این نویزها را کاهش داد و یا به صورت قابل-ملاحظه‌ای، حذف نمود. پس در واقع:

- ضرایب تقریب نشان‌دهنده خروجی فیلتر پایین‌گذر (فیلتر میانگین‌گیر) در تبدیل موجک گسسته هستند.
- ضرایب جزئی نشان‌دهنده خروجی فیلتر بالاگذر (فیلتر مشتق‌گیر) در تبدیل موجک گسسته هستند.



شکل ۱. مراحل تشکیل ضرایب تبدیل موجک

۴-۲-۱. آستانه‌گذاری

در این پژوهش، با استفاده از روش جذر لگاریتم[۶۶] سیگنال[۳۴]، مقادیر آستانه مطابق رابطه ۱ برای ضرایب موجک تعیین می‌شوند.

$$th_j = \sigma_j \sqrt{2\log\left(N_j\right)} \qquad (۱)$$

که در آن N_j، سیگنال نویزی است و σ_j مقدار انحراف میانگین [۶۷] است که از رابطه ۲، محاسبه می‌شود.

$$\sigma_j = \frac{\text{median}(|\omega_j|)}{0.6745} \qquad (۲)$$

که در آن ω مقدار ضرایب موجک در مقیاس jاست.

در نهایت مقادیر آستانه به ضرایب تبدیل موجک اعمال می‌شود و پس از آستانه‌گذاری روی ضرایب و اعمال تبدیل معکوس موجک، سیگنال نویزکاهی شده، استخراج می‌شود.

۵. طراحی شبکه جهت گسترش پهنای باند سیگنال صدا

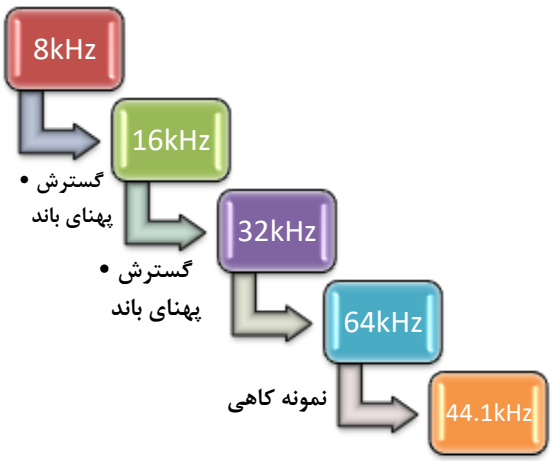
شبکه شامل دو بخش اصلی است که هر بخش بصورت جداگانه مورد بررسی قرار خواهد گرفت: ۱) فاز آموزش یا یادگیری و ۲) فاز آزمایش یا تست.

۵-۱. طراحی بخش آموزش شبکه

به منظور افزایش کارایی زمانی شبکه در این بخش از آموزش مرحله‌ای استفاده شده است. در آموزش مرحله‌ای، مرحله به مرحله خروجی را رصد نموده تا بتوان بهترین نتیجه را از شبکه گرفت. چنانچه خروجی هر مرحله مناسب نبود، نیاز به آموزش مجدد از اول نیست و تنها همان قسمت را تغییر داده و مجدد خروجی گرفته می‌شود و به نوعی وضعیت شبکه بصورت مداوم پایش می‌شود.

منظور از آموزش مرحله‌ای شبکه این است که، بجای اینکه پهنای باند ۸۰۰۰ هرتزی مستقیماً به ۴۴۱۰۰ هرتزی گسترش‌یابد، ابتدا سیگنال ۸۰۰۰ هرتزی با گسترش بوسیله شبکه، به ۱۶۰۰۰ هرتزی و سپس از ۱۶۰۰۰ به ۳۲۰۰۰ و در نهایت از ۳۲۰۰۰ به ۶۴۰۰۰ هرتزی گسترش داده می‌شود و در این بین وزن‌های هر شبکه جداگانه ذخیره شده که برای مرحله‌ی بعد گسترش، مجدد از همان وزن‌های مرحله قبل استفاده شود که هم میزان حافظه کوتاه‌مدت سیستم[۶۸] استفاده شده کاهش یابد، و هم سرعت آموزش افزایش پیدا کند. درنهایت با نمونه‌کاهی از ۶۴۰۰۰ به ۴۴۱۰۰ هرتز، هدف اصلی که تولید سیگنال گسترش‌یافته ۴۴۱۰۰ هرتزی است، حاصل می‌گردد. در شکل ۲، آموزش مرحله‌ای شبکه نشان‌داده‌شده است.

معماری شبکه جهت ۲ برابر گسترش پهنای باند سیگنال صدا یا گفتار در شکل ۳، نشان داده شده است.



شکل ۲. آموزش مرحله ای شبکه بر حسب پهنای باند

ابتدا مجموعه داده به سیگنال‌های کوچک با طول زمانی ۲ ثانیه (همراه با سکوت جهت محاسبه توان نویز به روش VAD[۶۹] برای ارزیابی) تقسیم می‌شود، حال این مجموعه داده کلی با نرخ اعتبارسنجی متقابل[۷۰] ٪۲۰ به شبکه برای آموزش داده می‌شود. شبکه با استفاده از الگوریتم بهینه‌ساز آدام[۷۱] بهینه می‌شود. به منظور یادگیری بهتر شبکه، از روش کاهش نمایی نرخ یادگیری[۷۲] با مقدار اولیه 10^{-2} استفاده شده است که این مقدار پله پله جهت بهینه‌سازی بهتر، کاهش می‌یابد. همچنین از تابع فعال‌ساز PReLU در هر لایه بعد از هر پیچش استفاده شده است.

در نهایت سیگنال گسترش یافته توسط تابع زیان هوبر با سیگنال اصلی مقایسه می‌شود و اختلاف دو سیگنال تحت عنوان مقدار زیان، بدست می‌آید. همانطور که پیشتر بیان شد، در این پژوهش از تابع زیان هوبر، که نخستین بار توسط [۳۵] معرفی و مورد استفاده قرار گرفت و ثابت شد که این تابع زیان نسبت به نویز مقاوم است، استفاده شده‌است. تابع زیان هوبر مطابق رابطه 3 تعریف می‌شود:

(۳)

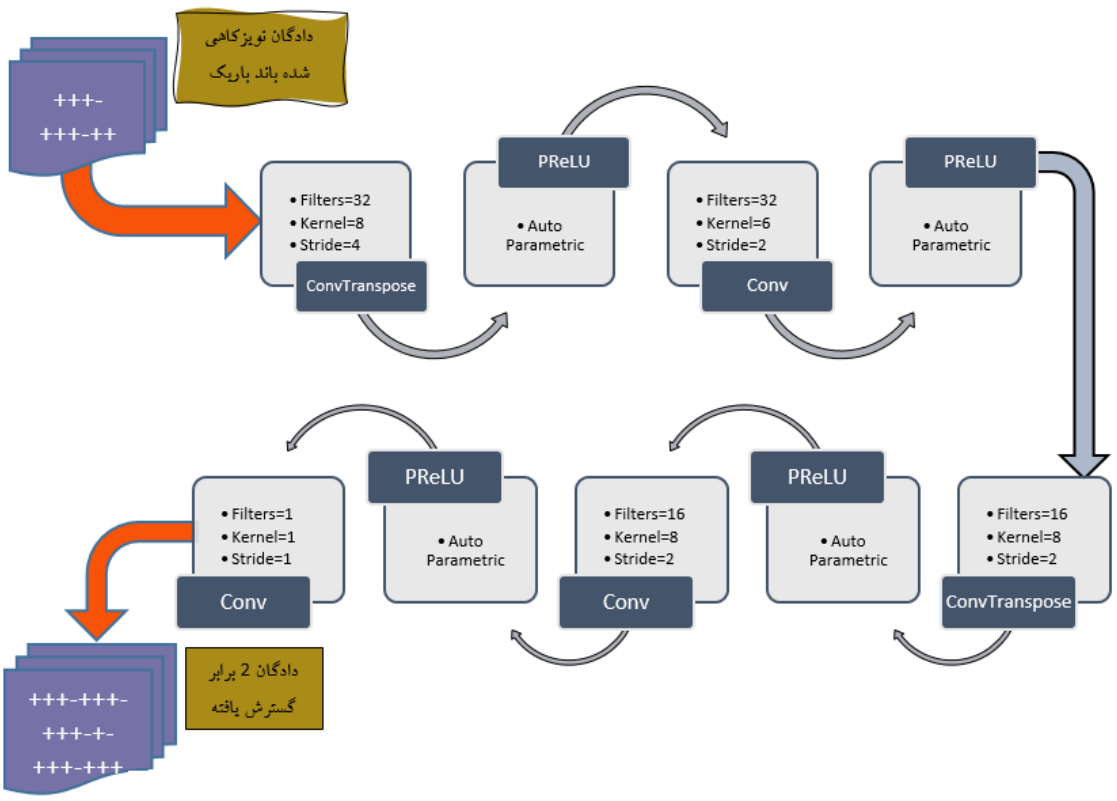
$$L_\delta(y-f(x)) = \begin{cases} \frac{1}{2}(y-f(x))^2, & |y-f(x)| \le \delta \\ \delta|y-f(x)| - \frac{1}{2}\delta^2, & |y-f(x)| > \delta \end{cases}$$

که در آن δ یک ابر متغیر[۷۳] است.

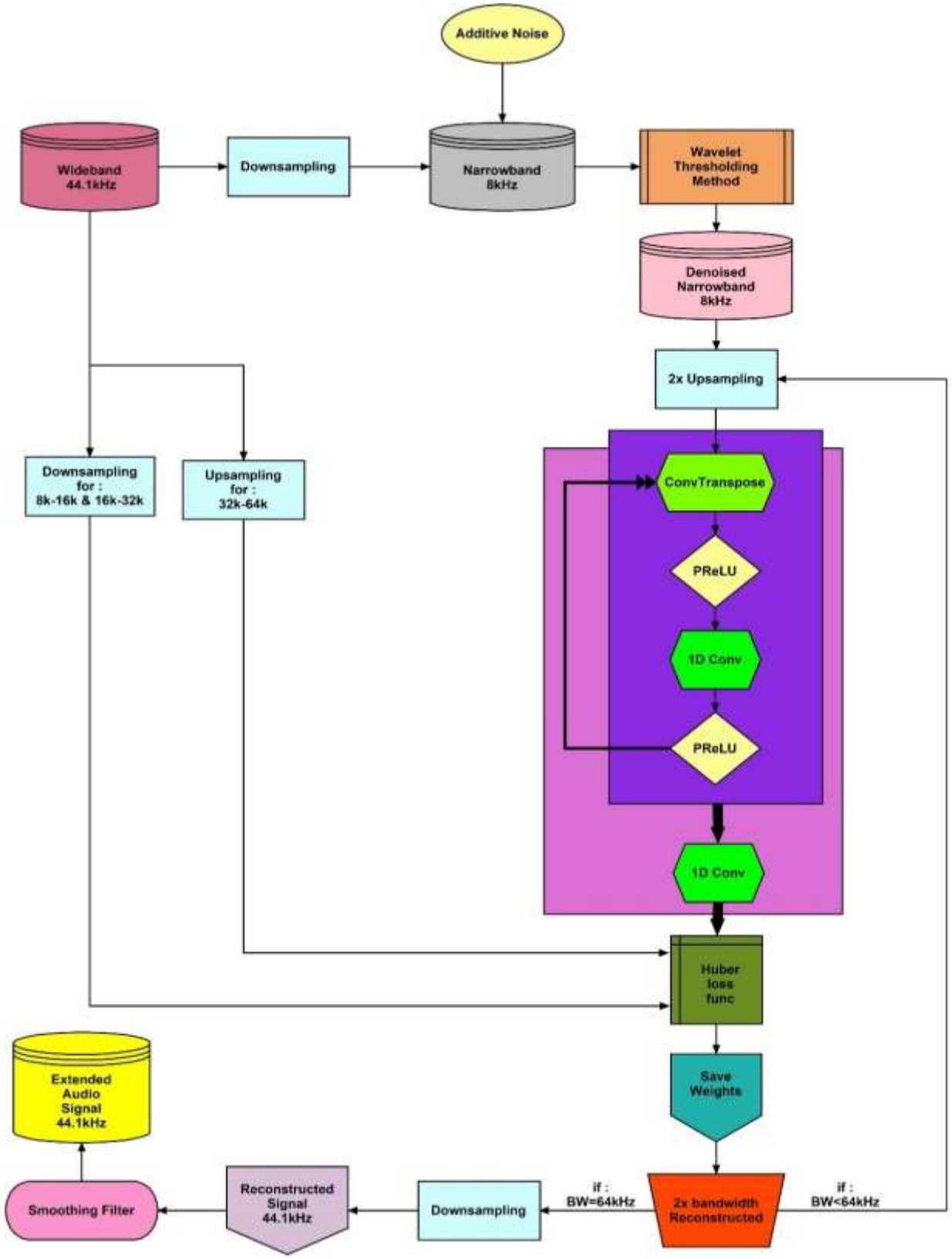
استفاده از تابع زیان هوبر موجب شد، میزان نویزگیری شبکه کاهش یابد و در ضمن همگرایی شبکه افزایش یابد. در جدول ۱، نتیجه استفاده از تابع زیان هوبر و روش نویزکاهی، نشان داده شده است. سیگنال نویزی وقتی داخل شبکه می‌شود، میزان نویز آن افزایش می‌یابد و نویز داخل شبکه، تقویت می‌شود. اما به هنگام استفاده از تابع زیان هوبر، بدلیل مقاومت این تابع مقابل نویز، میزان نویز پذیری شبکه کاهش چشمگیری پیدا کرد و همراه با نویزکاهی تبدیل موجک، این‌دو به نوعی مکمل یکدیگر عمل کردند و باعث بهبود کیفیت سیگنال تولیدی شدند.

جدول ۱. میزان تغییرات SNR، در تغییر تابع زیان [یافته‌های پژوهش]

	استفاده از روش کاهش نویز موجک	عدم استفاده از روش کاهش نویز موجک
استفاده از تابع زیان MAE	۱۶/۸	۱۳/۲
استفاده از تابع زیان هوبر	۱۸/۱	۱۵/۴



شکل ۳. معماری شبکه پیشنهادی به منظور گسترش پهنای باند سیگنال



شکل ۴. ساختار کلی شبکه پیشنهادی به منظور نویزکاهی و گسترش پهنای باند سیگنال

به همین ترتیب و با همین ساختار، شبکه ۱۶۰۰۰ هرتز به ۳۲۰۰۰ هرتز و همچنین ۳۲۰۰۰ به ۶۴۰۰۰ هرتزی تعریف می‌شوند و وزن‌های شبکه مرحله قبل به مرحله بعد منتقل می‌شود تا سرعت شبکه در مراحل آموزش افزایش یابد.

پس از تولید سیگنال با پهنای‌باند ۶۴۰۰۰، چون هدف تولید سیگنال با پهنای‌باند ۴۴۱۰۰ هرتز بود و همچنین با دانستن این موضوع که سیگنال ۶۴۰۰۰ هرتزی، حاوی نمونه‌های سیگنال ۴۴۱۰۰ هرتزی نیز هست، از سیگنال نمونه‌کاهی می‌شود تا به ۴۴۱۰۰ هرتزی برسد که این سیگنال هدف نهایی این پژوهش است.

۵-۲. طراحی بخش آزمایش شبکه

در بخش آزمایش شبکه پیچشی، شبکه آموزش داده شده با وزن‌های آن در حافظه ذخیره می‌شود و سپس داده‌های باند باریک تست (٪۲۰ داده‌ها که جهت تست نگه داشته‌شد) به شبکه وارد می‌شوند و سیگنال گسترش یافته باند پهن، تولید می‌شود.

۶. ارزیابی

در این پژوهش از هردو ارزیابی عینی و ذهنی جهت بررسی میزان کارایی شبکه پیشنهادی استفاده شده است.

۶-۱. ارزیابی عینی

برای ارزیابی و مقایسه بهتر روش پیشنهادی با سایر روش‌ها، هر ۴ روش ارزیابی SNR ، STOI ، PESQ و LSD پیاده‌سازی شد. بدیهی است که مقایسه همزمان همه روش‌ها، در همه‌ی پژوهش‌ها بصورت همزمان، عملاً ممکن نیست و صرفاً روش‌هایی با روش پیشنهادی مقایسه می‌شوند که از مجموعه دادگان DAPS استفاده کرده‌اند.

درجدول ۲، مقایسه نتایج پژوهش با برخی روش‌ها که صرفاً برروی مجموعه دادگان DAPS انجام شده‌اند، آورده شده‌اند. برخی روش‌ها، گسترش پهنای‌باند را در محدوده ۸۰۰۰ به ۱۶۰۰۰ ، انجام داده‌اند ولی روش شبکه پیشنهادی، از ۸۰۰۰ به ۴۴۱۰۰ است. در هر حال چون گسترش پهنای‌باند در محدوده‌های مختلفی قابل انجام است (مانند ۴۰۰۰ به ۸۰۰۰، ۸۰۰۰ به ۱۶۰۰۰، ۴۰۰۰ به ۱۶۰۰۰، ۸۰۰۰ به ۲۴۰۰۰، ۸۰۰۰ به ۴۴۱۰۰) و هر شبکه، سبک طراحی مختص به خود را دارد، نمی‌توان شبکه طراحی‌شده را با همه روش‌های موجود بصورت همزمان، قیاس کرد.

- روش گسترش ۲ مرحله‌ای توانست، برای ارزیابی SNR ، مقدار ۲۳/۰۱ را برای شبکه تک‌مرحله‌ای CRN و مقدار ۲۲/۸۹ را برای شبکه تک مرحله‌ای TCN ثبت کند. سپس با ۲ مرحله‌ای کردن شبکه و پردازش حوزه‌های زمان-فرکانسی، این اعداد به ترتیب به مقادیر ۲۴/۰۲ و ۲۳/۷۱ بهبود پیدا کرد که پردازش زمان – فرکانسی سیگنال، بخاطر تحلیل در هر دو حوزه، از سرعت کم‌تری برخوردار است و شبکه عملاً سنگین است [۲۷].
- در روش استفاده از تابع زیان زمان-فرکانسی، ارزیابی شبکه با استفاده از دو مجموعه داده انجام شده است، ابتدا با استفاده از مجموعه داده DAPS مقدار SNR بدست آمده، ۲۱/۴ بدست آمد. سپس شبکه با استفاده از دادگان VCTK تست مجدد شد که مقدار SNR، ۱۹/۳ بدست آمد [۲۹].
- در روش استفاده از شبکه FFTNet، در حالت تک سخنگو از دادگان DASP استفاده شده است.

مقدار SNR در این شبکه۲۰/۴ بدست آمده است[۱۷].

- در شبکه شناخته شده Kuleshov [۱۸] که یک شبکه بازگشتی[۷۴] است و جزء بهترین روش‌ها در زمینه گسترش پهنای‌باند است، شبکه در حالات مختلف برروی دادگان DAPS ارزیابی شد، در بهترین حالت توانست میزان SNR را، ۲۱/۱ بدست آورد [۱۸].

جدول ۲. ارزیابی عینی برحسب SNR [یافته‌های پژوهش]

روی دادگان با SNR=15 dB	**روی دادگان بدون نویز**					
گسترش از ۸۰۰۰ به ۴۴۱۰۰	گسترش از ۸۰۰۰ به ۴۴۱۰۰			گسترش از ۸۰۰۰ به ۱۶۰۰۰		
شبکه پیشنهادی	**شبکه پیشنهادی**	Kul RNN [۱۸]	FFTNet CNN [۱۷]	**شبکه پیشنهادی**	T-F loss CNN [۲۹]
۱۸/۱	**۲۴/۶**	۲۱/۱	۲۰/۴	**۲۲/۷**	۲۱/۸

خلاصه ارزیابی PESQ ، صرفاً برروی مجموعه دادگان DAPS ، در جدول ۳ آورده شده است:

جدول ۳. ارزیابی عینی برحسب PESQ [یافته‌های پژوهش]

روی دادگان با SNR=15 dB	**روی دادگان بدون نویز**		
گسترش از ۸۰۰۰ به ۴۴۱۰۰	گسترش از ۸۰۰۰ به ۴۴۱۰۰		
شبکه پیشنهادی	**شبکه پیشنهادی**	Kul RNN [۱۸]	FFTNet CNN [۱۷]
۳/۶۱	**۳/۶۸**	۲/۸۹	۳/۳

- در روش گسترش دو مرحله‌ای، مقدار PESQ برای شبکه TCN، ۴/۰۱ و برای شبکه CRN، مقدار۴/۱ در شرایط **بدون حضور نویز** بدست آمد[۲۷].
- در روش استفاده از شبکه مولد تخاصمی، مقدار PESQ برای این شبکه و شرایط **بدون نویز**، مقدار ۳/۴ بدست آمده است[۲۴].
- در روش شبکه بازگشتی KUL، مقدار PESQ در شرایط **بدون حضور نویز**، مقدار ۲/۸۹ بدست آمده است[۱۸].

معیار ارزیابی شنوایی زمان کوتاه (STOI)، یک روش ارزیابی عینی است که بر اساس یک ضریب همبستگی بین دو سیگنال، میزان شنوایی سیگنال را از نویز، برحسب درصد بیان می‌کند. این ارزیابی در زمینه گسترش پهنای باند، آنچنان مورد استفاده قرار نمی‌گیرد، زیرا معمولاً روش‌های گسترش پهنای‌باند برروی دادگان غیر نویزی پیاده‌سازی می‌شدند.

نتایج ارزیابی STOI از روش پیشنهادی در این پژوهش بصورت زیر است:

- در حالت نویزی: ارزیابی انجام شده از مدل طراحی

شده، به روش STOI، میزان ۸۹/۱٪ بدست آمد [یافته‌های پژوهش].

- درحالت بدون نویز: ارزیابی انجام شده از مدل طراحی شده، به روش STOI، میزان ۹۹/۲٪ بدست آمد [یافته‌های پژوهش].

معیار فاصله لگاریتمی طیف‌ها (LSD) یک روش ارزیابی سیگنال است که در آن لگاریتم طیف دو سیگنال با یکدیگر مقایسه می‌شود. هر چه تفاضل آن کمتر باشد، بدین معنی است که سیگنال‌ها به یکدیگر نزدیک‌تر هستند. لگاریتم طیف‌ها طبق رابطه ۴ محاسبه می‌شود[۲۷]:

$$\mathrm{LSD} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left[\log\frac{S(\omega)^2}{\hat{S}(\omega)^2}\right]^2} \quad (۴)$$

ارزیابی انجام شده برروی شبکه طراحی شده بیانگر اینست که:

- میزان LSD برای روش پیشنهادی، در شرایط نویزی، مقدار ۱/۳۱ بدست آمده است [یافته‌های پژوهش].
- میزان LSD برای روش پیشنهادی، در شرایط بدون نویز، مقدار ۱/۲۳ بدست آمده است [یافته‌های پژوهش].

نتایج ارزیابی سایر روش‌ها در زمینه گسترش پهنای باند بر اساس معیار ارزیابی LSD عبارتند از:

- میزان LSD در روش شبکه دو مرحله‌ای برای شبکه TCN، مقدار ۱/۲۶ و برای CRN، مقدار ۱/۲۴ بدست آمد [۲۷].
- در روش استفاده از تحلیل زمان-فرکانسی، میزان LSD برابر ۱/۵۷ بدست آمد[۲۹].
- در استفاده از روش FFTNet، نتیجه ۳/۹۲ برای LSD بدست آمد که مقدار چندان مناسبی نیست[۱۷].
- برای شبکه مولد تخاصمی، میزان LSD تقریباً ۱۰/۲۴ بدست آمد که این عدد نیز مناسب نیست[۲۴].

خلاصه این ارزیابی‌ها، صرفاً برروی مجموعه دادگان DAPS در جدول۴، نشان داده شده است:

جدول۴. ارزیابی عینی برحسب LSD [یافته‌های پژوهش]

روی دادگان با SNR=15dB	روی دادگان بدون نویز					
گسترش از ۸۰۰۰ به ۴۴۱۰۰	گسترش از ۸۰۰۰ به ۴۴۱۰۰			گسترش از ۸۰۰۰ به ۱۶۰۰۰		
روش پیشنهادی	روش پیشنهادی	FFTNet [۱۷]	Kul RNN [۱۸]	روش پیشنهادی	T-F loss CNN [۲۹]	
۱/۳۱	**۱/۲۳**	۳/۹۲	۱۰/۳	**۱/۴۶**	۱/۵۷	

۶-۲. ارزیابی ذهنی

برای ارزیابی ذهنی از روش میانگین نمرات (MOS[75]) استفاده شد. بدین طریق که سیگنال تولید شده توسط شبکه همراه با سیگنال باند پهن یا تمیز موجود در دادگان و همچنین سیگنال باند پهن نویزی را به ۲۵ نفر که اصلاحاً Oracle نامیده می‌شوند، داده می‌شود و از هر کدام خواسته می‌شود تا به این سیگنال‌ها با توجه به وضوح و کیفیت سیگنال، نمره‌ای بین ۰-۵

تخصیص دهند. سپس از کل نمرات میانگین گرفته می‌شود و عدد بدست‌آمده تحت عنوان ارزیابی MOS مشخص می‌گردد. نتایج MOS شبکه طراحی‌شده و همچنین سیگنال باند پهن موجود در دادگان و سیگنال نویزی باند پهن یا تمیز در جدول ۵ آورده شده‌اند.

جدول۵. ارزیابی ذهنی به روش MOS به ازای گسترش از ۸۰۰۰ به ۴۴۱۰۰ هرتز [یافته‌های پژوهش]

	سیگنال نویزی باندباریک 8 KHz	سیگنال گسترش یافته 44/1 KHz	سیگنال باند پهن 44/1 KHz
روش پیشنهادی	2/2	4/1	4/6
FFTNet [۱۷]	-	3/39	4/54
KUL [۱۸]	-	3/54	4/5

۷. بحث و نتیجه‌گیری

در این پژوهش یک سیستم متشکل از یک شبکه عصبی پیچشی و یک الگوریتم حذف نویز به منظور افزایش پهنای باند سیگنال صدا، طراحی شد. پس از نویزی کردن دادگان، الگوریتم نویزکاهی روی آن‌ها اعمال و سپس شبکه با استفاده از دادگان نویزکاهی شده، آموزش و مورد آزمایش و تست قرار گرفت. نکته مهم در افزایش مقاومت شبکه در برابر نویز، استفاده از تابع زیان هوبر بود. برای ارزیابی روش پیشنهادی از ارزیابی عینی و ذهنی استفاده شد. ازآنجاکه ارزیابی‌های عینی برپایه محاسبات ریاضی هستند، دارای نتایج قابل استنادتری می‌باشند. عموماً در زمینه گسترش پهنای باند سیگنال صدا، از ارزیابی‌های LSD ، PESQ و SNR استفاده می‌شود. شایان ذکر است که روش‌های گسترش پهنای باند پیشین که با نتایج این پژوهش مورد مقایسه قرار گرفته‌اند، برروی سیگنال‌های بدون نویز انجام شده، درحالیکه در این پژوهش از نویز خارجی برروی دادگان استفاده شده است. از این‌رو، علاوه بر معیارهای ارزیابی دیگر، از معیار ارزیابی STOI که نوعی سنجش میزان اثرگذاری نویز بر سیگنال است و میزان درک و شنوایی سیگنال را بر اساس قابلیت‌های شنوایی انسان می‌سنجد نیز استفاده شد. سپس این معیارها با اعداد بدست آمده در پژوهش‌های مختلف مقایسه گردید.

بنابراین، در ادامه، نتایج پژوهش انجام‌شده را می‌توان در قالب اثر یکایک بلوک‌های پیشنهادی بر کیفیت سیگنال گسترش یافته طبقه‌بندی نمود.

۷-۱. استفاده از الگوریتم نویزکاهی

بطورکلی نویزی‌کردن سیگنال، هدفی بود در جهت واقع بینانه کردن ورودی و خروجی شبکه، زیرا همانگونه که پیشتر توضیح داده شد، استفاده از مجموعه دادگان بدون نویز، نوعی ایده‌آل‌گرایی است. چنانچه بنابر استفاده مفید از شبکه باشد، شبکه باید قادر باشد تا داده‌های نویزی که در اثر گذر زمان دچار تخریب شده‌اند و یا در محیطی نویزی ضبط شده‌اند را نیز گسترش دهد. لذا استفاده از الگوریتم نویزکاهی، برای دادگان واقعی، چه در مرحله آموزش و چه در مرحله تست و آزمایش، اهمیت بسزایی دارد. الگوریتم‌های نویزکاهی زیادی تعریف شده هستند، می‌توان به عنوان یک پیشنهاد بسیار موثر، از شبکه عصبی جهت نویزکاهی نیز استفاده نمود؛ در این پژوهش از روش موجک بهره گرفته‌شد. چون دادگان مرحله آموزش، با میزان سیگنال به نویز ۱۵ دسی‌بل، تخریب شدند، نمی‌توان انتظار داشت که الگوریتم‌های نویزکاهی، دادگان را همانند اول بسازد. ولی می‌توان با تغییر روش، نتایج مناسب‌تری کسب کرد. نتیجه استفاده از الگوریتم نویزکاهی و عدم استفاده از آن، به

عنوان مرحله پیش پردازش، در جدول ۶ و ۷ آمده است.

جدول ۶. مقادیر SNR به ازای استفاده از الگوریتم نویزکاهی [یافته‌های پژوهش]

	عدم استفاده از روش کاهش نویز موجک	استفاده از روش کاهش نویز موجک
SNR	۱۵/۴	**۱۸/۱**

جدول۷. مقادیر STOI به ازای استفاده از الگوریتم نویزکاهی [یافته‌های پژوهش]

	عدم استفاده از روش کاهش نویز موجک	استفاده از روش کاهش نویز موجک
STOI (%)	۶۸/۶	**۸۹/۱**

۷-۲. استفاده از تابع زیان هوبر

هدف استفاده از تابع زیان هوبر، افزایش مقاوم‌پذیری شبکه در مقابل نویز بود. تابع زیان هوبر، چون ترکیبی از MAE و MSE است، به ازای یک مقدار آستانه بین این دو تابع متغیر است. این خاصیت سبب می‌شود تا چنانچه داده‌های شبکه دارای نویز باشند و تابع زیان هم مقدار بالایی دارد، به سرعت به تابع زیان دیگری تغییر پیدا کند تا همگرایی شبکه دچار اشکال نشود. بنابراین استفاده از تابع زیان هوبر به همراه الگوریتم نویزکاهی تبدیل موجک، هر دو باهم توانستند مقدار سیگنال بر نویز را افزایش دهند، با وجود اینکه داده‌ها نویزی بودند. از لحاظ همگرایی، شبکه با هر ۳ نوع تابع MSE و MAE و هوبر، تقریباً سرعت همگرایی یکسانی داشت، ولی در هوبر این سرعت کمی بیشتر بود. در جدول ۸، نتایج استفاده از تابع زیان هوبر، نشان داده شده است.

جدول ۸. نتایج استفاده از تابع زیان هوبر در معیار SNR [یافته‌های پژوهش]

	عدم استفاده از دادگان نویزی	استفاده از دادگان نویزی و روش کاهش نویز موجک
استفاده از تابع زیان MAE	۲۱/۴	۱۶/۸
استفاده از تابع زیان هوبر	۲۴/۶	۱۸/۱

۷-۳. استفاده از روش آموزش چند مرحله‌ای

یکی از نوآوری‌های دیگر این پژوهش، استفاده از آموزش چند مرحله‌ای شبکه است. آموزش چند مرحله‌ای، به دلیل چند قسمتی کردن شبکه، سرعت همگرایی شبکه را تا۴۰٪ افزایش داد، به طوری که شبکه در ۱۰ دوره اول خود در هر مرحله تقریباً به ۹۰٪ همگرایی خود رسید. این روش کمک بسزایی در استفاده از GPUهای رایگان مهیا شده در بستر سرور گوگل کولب[۷۶] کرد و باعث ذخیره‌شدن زمان اجرا[۷۷] برای مراحل بعدی شد. بنابراین برتری دیگر شبکه طراحی‌شده نسبت به سایر روش‌ها، زمان آموزش کم و سرعت همگرایی بالای شبکه است. بطور میانگین برای آموزش هر مرحله از شبکه، ۲۰دوره برای کسب نتیجه مناسب کافی است.

منابع

[1] Prasad, N., and T. Kishore Kumar. "Bandwidth extension of speech signals: A comprehensive review." *International Journal of Intelligent Systems and Applications* 8, no. 2 (2016): 45-52.

[2] Epps, Julien, and W. Harvey Holmes. "A new technique for wideband enhancement of coded narrowband speech." In *1999 IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No. 99EX351)*, pp. 174-176. IEEE, 1999.

[3] Vaseghi, Saeed, Esfandiar Zavarehei, and Qin Yan. "Speech bandwidth extension: Extrapolations of spectral envelop and harmonicity quality of excitation." In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 3, pp. III-III. IEEE, 2006.

[4] Han, Jinyu, Gautham J. Mysore, and Bryan Pardo. "Language informed bandwidth expansion." In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1-6. IEEE, 2012.

[5] Seo, Hyunson, Hong-Goo Kang, and Frank Soong. "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise." In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6087-6091. IEEE, 2014.

[6] Per Ekstrand. "Bandwidth extension of audio signals by spectral band replication." In *Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA02)*. Citeseer, 2002.

[7] Larsen, Erik, and Ronald M. Aarts. *Audio bandwidth extension: application of psychoacoustics, signal processing and loudspeaker design*. John Wiley & Sons, 2005.

[8] Jax, Peter, and Peter Vary. "On artificial bandwidth extension of telephone speech." *Signal Processing* 83, no. 8 (2003): 1707-1719.

[9] Qian, Yasheng, and Peter Kabal. "Wideband speech recovery from narrowband speech using classified codebook mapping." In *Australian Int. Conf. Speech Science, Technology*, pp. 106-111. 2002.

[10] Nour-Eldin, Amr H., and Peter Kabal. "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech." In *Interspeech*, pp. 53-56. 2008.

[11] Iser, Bernd, and Gerhard Schmidt. "Bandwidth extension of telephony speech." In *Speech and Audio Processing in Adverse Environments*, pp. 135-184. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[12] Bachhav, Pramod B., Massimiliano Todisco, Moctar Mossi, Christophe Beaugeant, and Nicholas Evans. "Artificial bandwidth extension using the constant Q transform." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5550-5554. IEEE, 2017.

[13] Bradbury, Jeremy. "Linear predictive coding." *Mc G. Hill* (2000).

[14] Tokuda, Keiichi, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. "Speech synthesis based on hidden Markov models." *Proceedings of the IEEE* 101, no. 5 (2013): 1234-1252.

[15] Abel, Johannes, and Tim Fingscheidt. "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, no. 1 (2017): 71-83.

[16] Li, Kehuang, and Chin-Hui Lee. "A deep neural network approach to speech bandwidth expansion." In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4395-4399. IEEE, 2015.

[17] Feng, Berthy, Zeyu Jin, Jiaqi Su, and Adam Finkelstein. "Learning bandwidth expansion using perceptually-motivated loss." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 606-610. IEEE, 2019.

[18] Kuleshov, Volodymyr, S. Zayd Enam, and Stefano Ermon. "Audio super resolution using neural networks." *arXiv preprint arXiv:1708.00853* (2017).

[19] Gupta, Archit, Brendan Shillingford, Yannis Assael, and Thomas C. Walters. "Speech bandwidth extension with wavenet." In *2019 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, pp. 205-208. IEEE, 2019.

[20] Kuleshov, Volodymyr, S. Zayd Enam, and Stefano Ermon. "Audio super resolution using neural networks." *arXiv preprint arXiv:1708.00853* (2017).

[21] Birnbaum, Sawyer, Volodymyr Kuleshov, Zayd Enam, Pang Wei W. Koh, and Stefano Ermon. "Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations." *Advances in Neural Information Processing Systems* 32 (2019).

[22] Nguyen, Viet-Anh, Anh HT Nguyen, and Andy WH Khong. "Tunet: A block-online bandwidth extension model based on transformers and self-supervised pretraining." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161-165. IEEE, 2022.

[23] Rakotonirina, Nathanaël Carraz. "Self-attention for audio super-resolution." In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1-6. IEEE, 2021.

[24] Eskimez, Sefik Emre, and Kazuhito Koishida. "Speech super resolution generative adversarial network." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3717-3721. IEEE, 2019.

[25] Kumar, Rithesh, Kundan Kumar, Vicki Anand, Yoshua Bengio, and Aaron Courville. "NU-GAN: High resolution neural upsampling with GAN." *arXiv preprint arXiv:2010.11362* (2020).

[26] Lee, Junhyeok, and Seungu Han. "Nu-wave: A diffusion probabilistic model for neural audio upsampling." *arXiv preprint arXiv:2104.02321* (2021).

[27] Lin, Ju, Yun Wang, Kaustubh Kalgaonkar, Gil Keren, Didi Zhang, and Christian Fuegen. "A Two-Stage Approach to Speech Bandwidth Extension." In *Interspeech*, pp. 1689-1693. 2021.

[28] Li, Yunpeng, Marco Tagliasacchi, Oleg Rybakov, Victor Ungureanu, and Dominik Roblek. "Real-time speech frequency bandwidth extension." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 691-695. IEEE, 2021.

[29] Wang, Heming, and Deliang Wang. "Time-frequency loss for CNN based speech super-resolution." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 861-865. IEEE, 2020.

[30] Zhang, Kexun, Yi Ren, Changliang Xu, and Zhou Zhao. "WSRGlow: A glow-based waveform generative model for audio super-resolution." *arXiv preprint arXiv:2106.08507* (2021).

[31] Liu, Haohe, Woosung Choi, Xubo Liu, Qiuqiang Kong, Qiao Tian, and DeLiang Wang. "Neural vocoder is all you need for speech super-resolution." *arXiv preprint arXiv:2203.14941* (2022).

[32] Liu, Haohe, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D. Plumbley. "AudioSR: Versatile audio super-resolution at scale." In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1076-1080. IEEE, 2024.

[33] Hidayat, Risanuri, Agus Bejo, Sujoko Sumaryono, and Anggun Winursito. "Denoising speech for MFCC feature extraction using wavelet transformation in speech recognition system." In *2018 10th international conference on information technology and electrical engineering (ICITEE)*, pp. 280-284. IEEE, 2018.

[34] Ali, M. A., and P. M. Shemi. "An improved method of audio denoising based on wavelet transform." In *2015 international conference on Power, Instrumentation, Control and Computing (PICC)*, pp. 1-6. IEEE, 2015.

[35] Huber, Peter J. "Robust estimation of a location parameter." In *Breakthroughs in statistics: Methodology and distribution*, pp. 492-518. New York, NY: Springer New York, 1992.

پی‌نوشت

[1] Device and Produced Speech (DAPS)
[2] Denoising Methods
[3] Convolutional Neural Network
[4] Huber loss function
[5] Signal-to-Noise Ratio (SNR)
[6] Log-Spectral Distortion (LSD)
[7] Perceptual Evaluation of Speech Quality (PESQ)
[8] Short-Time Objective Intelligibility (STOI)
[9] Fast Fourier Transform network (FFTNet)
[10] Kuleshov (KUL)
[11] Narrow Band (NB)
[12] High Definition (HD)
[13] Ultra High Definition (UHD)
[14] Wide Band (WB)
[15] Super Wide Band (SWB)
[16] End to End
[17] Speech Bandwidth Extension / Expansion (BWE)
[18] Speech Super Resolution (SSR)
[19] Fast Fourier Transform (FFT)
[20] Gaussian Mixture Model (GMM)
[21] Linear Predictive Coding (LPC)
[22] Hidden Markov Model (HMM)
[23] Two Split Summation (2SS)
[24] Three Split Summation (3SS)
[25] Amazon Mechanical Turk (MTurk)
[26] Libri Text to Speech
[27] Adaptive Multi-Rate Wideband (AMR-WB)
[2828] Global System for Mobile Communications
[29] Temporal Feature-Wise Linear Modulation (TFiLM)
[30] Convolutional Recurrent
[31] Long range
[32] Transformer
[33] Attention-based Feature-Wise Linear Modulation
[34] Generative Adversarial Networks
[35] Adversarial Loss
[36] Voice Cloning Toolkit (VCTK)

[37] Perceptual Evaluation of Speech Quality (PESQ)
[38] Text to Speech
[39] Diffusion
[40] Temporal Convolutional Network (TCN)
[41] Convolutional Recurrent Network (CRN)
[42] Multi-Resolution Short-Time Fourier Transform (MSTFT)
[43] Mean Squared Error (MSE)
[44] Valentini-Botinhao Corpus (VBC)
[45] Deep neural networks (DNN)
[46] Real-Time
[47] Speech Enhancement Network (SEANet)
[48] Autoencoder
[49] Texas Instruments/Massachusetts Institute of Technology
[50] Resolution
[51] High Resolution
[52] Low Resolution
[53] Short Time Fourier Transform
[54] Neural Vocoder-based Speech Super-Resolution (NVSSR)
[55] Mean Absolute Error (MAE)
[56] Wavelet Threshold Denoising
[57] Epoch
[58] Beamforming
[59] Spectral Subtraction
[60] Additive White Gaussian Noise (AWGN)
[61] Horn Noise
[62] Symlet
[63] Level
[64] Detail Coefficients (cD)
[65] Approximation Coefficients (cA)
[66] Root Logarithm
[67] Mean Absolute Deviation (MAD)
[68] Random Access Memory (RAM)
[69] Voice Activity Detector (VAD)
[70] Cross-Validation Rate
[71] Adam Optimization Algorithm
[72] Exponential Decay Learning Rate
[73] Hyper-Parameter
[74] Recurrent Neural Network
[75] Mean Opinion Score
[76] Google Colaboratory
[77] Runtime