بازشناسی مقاوم به نویز و تنوعات گفتار از طریق به اشتراک گذاشتن مؤلفه‌های مشترک

پروین زارعی اسکی کند و سیدعلی سیدصالحی

چکیده: یکی از روش‌های بهبود عملکرد سامانه‌های بازشناسی در برابر نویز و یا تنوعات ناخواسته، استخراج اطلاعات مشترک بین داده‌های مختلف ورودی می‌باشد. در مورد شبکه‌هایی که ظرفیت بسیار پایینی دارند امکان ذخیره‌سازی الگوها به‌صورت مفاهیم جداگانه وجود ندارد، لذا کیفیت بازشناسی شدیداً افت پیدا می‌کند. در این مقاله ساختاری ارائه شده است که بتواند زیرفضای مشترک بین داده‌های ورودی را استخراج کرده و آن را در میان گویندگان مختلف به اشتراک بگذارد. ساختار چندتکلیفی شبکه این امکان را فراهم می‌کند که این زیرفضا به‌صورت یک جاذب پیوسته واحد شکل بگیرد که این جاذب نسبت به تنوعاتی مانند تغییرات گوینده در فضای ورودی پویا می‌باشد. لذا داده‌های ورودی آغشته به نویز توسط یک نگاشت غیر خطی به یک مانیفولد در ابعاد پایین فیلتر می‌شوند که پویایی این مانیفولد مقاوم‌بودن آن را نسبت به تنوعاتی مثل تغییر گوینده تأمین می‌کند. اتصالات بازگشتی در طی روند تعلیم، یک جاذب پیوسته را در فضای ورودی شکل می‌دهند که کدهای گوینده به اطلاعات لازم جهت پویاسازی این جاذب تبدیل می‌شوند. پس از فرایند جذب‌شدن داده آغشته به نویز، عمل بازشناسی بر روی داده تمیز حاصله اعمال می‌شود. استخراج و به اشتراک گذاشتن مؤلفه‌های مشترک در این ساختار توانسته است کارایی جاذب‌ها را در بازشناسی مقاوم آوا تا حدود ۵٪ نسبت به مدل مشابه، بدون پویایی جاذب‌ها، در نسبت سیگنال به نویز dB ۰ بهبود بخشد.

کلید واژه: **استخراج مؤلفه‌های اساسی، بازشناسی گفتار مقاوم به نویز، به اشتراک گذاشتن مؤلفه‌های مشترک، جاذب پیوسته پویا، کاهش بعد غیر خطی.**

۱- مقدمه

بازشناسی گفتار در انسان نه تنها در برابر نویز مقاوم می‌باشد بلکه ایجاد تغییراتی در سیگنال گفتار مانند تغییر گوینده، انعکاس صدا، تغییر جهت و فاصله گوینده و غیره در عملکرد آن خللی ایجاد نمی‌کند. به‌عبارتی دیگر مغز انسان در بازشناسی به‌صورت پویا عمل کرده و خود را با هر گونه تغییر تطبیق می‌دهد. لذا در شرایط مختلف می‌تواند بازشناسی را با میزان صحت بالایی انجام دهد.

در مورد بازشناسی هوشمند گفتار، تاکنون تحقیقات وسیعی انجام شده است. سیستم‌های بازشناسی ارائه‌شده عملکرد بسیار خوبی در محیط‌های آزمایشگاهی و عاری از نویز دارند، ولی در مواجه با نویزهای غیر ایستان و یا سایر تنوعات ناخواسته قدرت بازشناسی لازم را ندارند. در پژوهش‌های اخیر پیشرفت‌های قابل توجهی در زمینه بازشناسی مقاوم الگو حاصل شده است ولی در قیاس با عملکرد مغز انسان هنوز ضعف‌های زیادی در ساختارهای ارائه‌شده وجود دارد [۱] و [۲].

برای دستیابی به سامانه‌های بازشناسی مقاوم، یکی از رویکردها استخراج مؤلفه‌های اساسی از الگو است که در این روش‌ها با استخراج مؤلفه‌های اساسی مقاوم، مانیفولد داده در ابعاد بالا به بعدهای ذاتی داده تصویر می‌شود. این نگاشت غیر خطی تا حد ممکن هرگونه نویز و تنوعات ناخواسته را از الگوی ورودی فیلتر می‌کند [۱] تا [۶]. یکی از روش‌های مؤثر در این زمینه استفاده از شبکه‌های کدکننده پالایش‌گر[1] است.

در این روش شبکه به‌گونه‌ای تعلیم می‌بیند که الگوی تمیز را از داده‌ای که به‌صورت تصادفی اعوجاج یافته است، بازسازی کند. در واقع شبکه تلاش می‌کند تا مقادیری از داده را که تخریب شده‌اند[2] از مقادیر موجود پیش‌بینی کند. بنابراین مؤلفه‌هایی که در لایه پنهان شکل می‌گیرند نسبت به نویز مقاوم می‌گردند [۷] و [۸]. اگرچه این روش تا حدی امکان بازشناسی مقاوم را فراهم می‌کند، ولی تنها در مواردی کاربرد دارد که نویز بخشی از سیگنال را تحت تأثیر قرار داده باشد. در حالی که در بیشتر موارد نویز و تنوعات ناخواسته بر روی کل سیگنال تأثیر می‌گذارند، در این حالت امکان بازیابی مقادیر تخریب‌شده سیگنال وجود ندارد [۷].

روش دیگر بر مبنای استخراج بازنمایی‌های تنک[3] از سیگنال است. از آنجایی که استخراج بازنمایی‌های فشرده‌تر از سیگنال قدرت تعمیم را افزایش می‌دهند، لذا بازنمایی‌های تنک در بازشناسی مقاوم الگو مؤثرتر عمل می‌کنند. به این ترتیب از طریق روش‌های آماری و فناوری اطلاعات تلاش می‌شود تا بازنمایی‌های تنک در لایه‌های پنهان شکل بگیرد. از آنجایی که این بازنمایی‌ها فشرده هستند، مؤلفه‌های مربوط به تنوعات ناخواسته و نویز از سیگنال تا حدودی فیلتر می‌شوند [۴]، [۵]، [۷]، [۹] و [۱۰].

یک روش در این راستا استفاده از ساختار رمزگذاری تنک[4] است. در این روش ساختارهای مختلف به‌صورت بدون سرپرستی با یکدیگر رقابت می‌کنند تا در نهایت کدهایی شکل بگیرد که بهتر بتواند ورودی را بیان کند. اشکال این روش ناپایداری کدهای حاصل است. به بیانی دیگر اندکی انحراف در الگوی ورودی تأثیر زیادی در کدهای بهینه می‌گذارد. این ناپایداری در لایه‌های بالاتر که این کدها به‌عنوان ورودی به آنها داده می‌شود، نتایج بدتری به همراه دارد. قدرت تعمیم در این روش که داده‌های مشابه با تفاوت بسیاری کد می‌شوند، کاهش می‌یابد [۵]، [۷] و [۱۰].

برای حل مشکل ناپایداری، روش‌هایی بر مبنای ترکیب دو روش کدکننده معمولی[5] و رمزگذاری تنک پیشنهاد شده است. در روش حاصل

این مقاله در تاریخ ۲۱ دی ماه ۱۳۸۹ دریافت و در تاریخ ۲۷ مهر ماه ۱۳۹۰ بازنگری شد.

پروین زارعی اسکی کند، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران، (email: parvin.zarei@aut.ac.ir).

سیدعلی سیدصالحی، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران، (email: ssalehi@aut.ac.ir).

1. Denoising Autoencoder
2. Missing Values
3. Sparse
4. Sparse Coding
5. Auto Encoder

با نام کدکننده تنک[1]، کدهای بهینه همانند روش کدگذاری تنک آزاد هستند. ولی یک اینکدر پارامتری نیز مانند روش‌های کدکننده معمولی تعریف می‌شود. الگوریتم تلاش می‌کند تا مانع فاصله‌گرفتن کدهای غیر پارامتری آزاد از خروجی اینکدر گردد. به این ترتیب کدها می‌توانند مزیت هر دو روش را داشته باشند. یعنی هم بتوانند به‌خاطر آزادبودن کدها ورودی را بهتر بازسازی کنند و همچنین از خروجی اینکدر فاصله نگیرند تا ناپایدار شوند [۷] و [۱۱].

مشکلی که در سامانه‌های استخراج ویژگی وجود دارد، در بازسازی ورودی از مؤلفه‌های تغییرناپذیر است. به بیانی دیگر تنها استفاده از ورودی‌های تغییرناپذیر برای بازسازی سیگنال کافی نیست. در یکی از روش‌ها بخش دیگری از سامانه بازشناسی به استخراج ویژگی‌های تغییرپذیر تعلق دارد. به‌عبارتی ویژگی‌های تغییرپذیر در بین ورودی‌ها به‌عنوان اطلاعات زاید حذف نمی‌شوند، بلکه از این ویژگی‌ها برای بازسازی کامل سیگنال و در نهایت استخراج بهتر ویژگی‌های تغییرناپذیر استفاده می‌شود [۴]. اشکال این روش از این جهت قابل بررسی است که از اطلاعات تفاوت بین ورودی‌ها برای بازشناسی استفاده نمی‌شود و پس از استخراج ویژگی‌های تغییرناپذیر، این اطلاعات بلااستفاده باقی می‌مانند. همچنین روش ارائه‌شده تنها برای تنوعاتی مانند شیفت، اشاره‌شده در مقاله، کاربرد دارد و قابل تعمیم به دیگر داده‌ها و یا تنوعات دیگر نیست.

ضعف مشترک همه این روش‌ها ایجادشدن کدهای متفاوت برای ورودی‌های مشابه است. به بیانی دیگر اشتراکات بین ورودی‌ها در بازشناسی در نظر گرفته نشده است. اگر در طراحی مدل‌های بازشناسی مؤلفه‌های مشترک بین داده‌ها لحاظ شوند، قدرت تعمیم سامانه بازشناسی افزایش می‌یابد. در نتیجه می‌توان با به اشتراک گذاشتن این ویژگی‌ها در مراحل بازنمایی مختلف به ساختارهایی دست پیدا کرد که حتی در الگوهای تعلیم داده شده در شبکه وجود ندارد [۷].

به این ترتیب سامانه توانایی خلق اطلاعات جدید با ترکیب اطلاعات تعلیم داده شده را پیدا می‌کند و در نتیجه قدرت تعمیم بیشتری خواهد داشت. از طرفی دیگر با به اشتراک گذاشتن اطلاعات مشترک می‌توان به حداکثر ظرفیت ذخیره‌سازی و سرعت پردازش دست یافت. در تحقیقات اخیر روش‌هایی بر مبنای ساختارهای سلسله مراتبی در راستای به اشتراک گذاشتن ویژگی‌ها پیشنهاد شده است.

یکی از روش‌ها بر مبنای شبکه SOM سلسله مراتبی است. این روش به‌گونه‌ای طراحی شده است که یک نگاشت ویژگی به اشتراک گذاشته شده[2]، جایگزین مجموعه نگاشت‌های مجاور می‌شود. در نتیجه اگر الگویی دچار اعوجاج و یا تغییراتی همانند شیفت شود، تنها یک نرون فعال می‌شود. بنابراین بازنمایی‌ها در برابر تنوعاتی مانند شیفت، مقیاس و اعوجاج مقاوم می‌شوند [۱۲].

در این مقاله سعی بر این است تا ساختاری از شبکه‌های عصبی بازگشتی ارائه شود تا مجموعه سیگنال‌های ناشی از گویندگان مختلف به‌صورت مفاهیم جداگانه به شبکه تعلیم داده نشوند، بلکه این داده‌ها دارای ویژگی‌هایی هستند که بین همه آنها مشترک می‌باشد. در نتیجه شبکه تنها یک مفهوم واحد از سیگنال گفتار یعنی پیام آن را یاد می‌گیرد، که این مفاهیم مشترک تحت تنوعاتی مانند تغییر گوینده می‌تواند تمام فضای ورودی را با پویایی خود پوشش دهد.

به بیانی دیگر در این ساختار، اطلاعات مربوط به پیام گفتار به‌عنوان مؤلفه‌های مشترک (غیر خطی) بین داده‌ها استخراج شده و مؤلفه‌های حاصل نسبت به تغییرات گوینده پویا می‌شوند. در نتیجه اطلاعات مشترک بین داده‌ها تحت یک زیرفضای واحد شکل می‌گیرد که این زیرفضا به‌صورت یک جاذب پیوسته[3] می‌باشد. به این ترتیب داده‌های آغشته به نویز تحت یک فرایند بازگشتی (دورزدن شبکه) به سمت جاذب پیوسته نگاشت می‌شوند. از آنجایی که این جاذب زیرفضای داده‌های تمیز می‌باشد، نویز به‌صورت غیر خطی از سیگنال فیلتر می‌شود. جاذب پیوسته حاصل نسبت به تغییرات گوینده پویا می‌باشد، در نتیجه اطلاعات مشترک بین گویندگان مختلف به اشتراک گذاشته می‌شود.

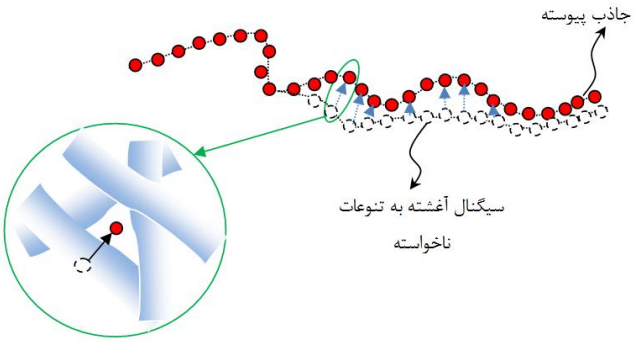
به این ترتیب شبکه نیاز به ایجاد جاذب‌های پیوسته جداگانه به‌ازای داده‌های مختلف را ندارد، لذا در مواردی که شبکه ابعاد کمتری نسبت به حجم دادگان دارد، ظرفیت ذخیره‌سازی شبکه افزایش پیدا می‌کند. با ترکیب مؤلفه‌های مشترک و مؤلفه‌های مربوط به تغییرات، می‌توان به داده‌هایی دست یافت که در مجموعه دادگان وجود ندارند و شبکه توانایی بازشناسی این داده ها را دارد. به این ترتیب قدرت تعمیم در شبکه افزایش پیدا می‌کند. نتایج نشان می‌دهد با این که ابعاد شبکه بسیار کوچک هستند ولی شبکه توانسته است تا حد ممکن با به اشتراک گذاشتن اطلاعات، مانیفولدهای پیچیده سیگنال گفتار را یاد بگیرد. به نظر می‌رسد علت قدرت تعمیم وسیع سامانه بازشناسی انسان در استخراج مفاهیم مشترک و به اشتراک گذاشتن آنها نهفته است. در شبکه طبقه‌بندی کننده، سیگنال‌های تمیز به‌عنوان داده‌های تعلیم و سیگنال‌های آغشته به نویز ایستان و غیر ایستان به‌عنوان داده تست به کار رفته‌اند.

در بخش دوم جنبه‌های نظری فیلترسازی غیر خطی سیگنال توسط دینامیک‌های جاذب مورد بررسی قرار می‌گیرد. بخش سوم نیز به معرفی دادگان مورد استفاده در شبکه پرداخته است. در بخش چهارم ساختار شبکه‌های عصبی پیشنهادی بررسی شده و به دنبال آن در بخش پنجم نتایج پیاده‌سازی الگوریتم بر روی دادگان واقعی آورده شده است. در نهایت بخش ششم شامل جمع‌بندی و نتیجه‌گیری می‌باشد.

۲- استفاده از جاذب‌های پیوسته برای فیلترسازی غیر خطی سیگنال

در این بخش نحوه عملکرد جاذب‌های پیوسته در پالایش سیگنال مورد بررسی قرار می‌گیرد. در ساده‌ترین بیان، جاذب برای یک سیستم دینامیک زیرمجموعه بسته A از فضای فاز است که با هر انتخابی برای نقطه شروع سیستم به A تکامل خواهد یافت. شبکه‌هایی بازگشتی که با قانون هبین[4] تعلیم می‌بینند، یک ویژگی جالب دارند که بعد از تعلیم وارد فاز بازیابی[5] می‌شوند. در چنین فازی اگر نسخه فازی الگوهای تعلیم داده شده به شبکه در ورودی قرار گیرند، حالت شبکه در هر گام تغییر می‌کند تا شبکه به نقطه‌ای برسد که فعالیت گره‌ها دیگر تغییر نکند. این حالت یک جاذب نقطه‌ای در سیستم دینامیکی این شبکه بازگشتی است [۱۳]. به بیانی دیگر هر کدام از نرون‌های لایه پنهان در فضای ورودی ابرصفحاتی ایجاد می‌کنند که توسط آنها فضای m بعدی ورودی چندی‌سازی می‌شود. از آنجایی که هر کدام از نواحی در خروجی شبکه یک مقدار واحد دارند، توسط این ابرصفحات فضای ورودی خوشه‌بندی شده و در خروجی شبکه تشخیص داده می‌شود که نمونه ورودی متعلق به

1. Sparse Autoencoder
2. Shared Feature Map
3. Continuous Attractor
4. Hebbian
5. Retrieval



شکل ۱: جذب‌شدن سیگنال آغشته به نویز و تنوعات ناخواسته توسط جاذب پیوسته. نمونه‌های سیگنال نویزی در هر گام به نمونه‌های تمیز که به‌صورت جاذب پیوسته شکل گرفته‌اند، نگاشته می‌شوند. این شکل یک فرم نمادین از مانیفولد یک‌بعدی سیگنال است، در حالت عمومی این مانیفولد چندبعدی می‌باشد.

کدام خوشه می‌باشد. در هنگام استفاده از تابع غیر خطی پله‌ای، هر کدام از نمونه‌های ورودی تنها متعلق به یک خوشه هستند. در صورتی که با استفاده از تابع غیر خطی نرم این نمونه می‌تواند به تمامی خوشه‌ها با درجه تعلق متفاوت نسبت داده شود که درجه تعلق نمونه ورودی به هر کدام از خوشه‌ها معمولاً بین صفر تا یک تغییر می‌کند [۱] و [۱۴]. وقتی نمونه ورودی آغشته به نویز می‌گردد، درجه تعلق آن به خوشه اصلی کمتر از نمونه تمیز می‌گردد. هر چقدر این اختلاف کمتر باشد بیانگر شباهت آن به نمونه اصلی است [۱].

در شبکه خودانجمنی دارای اتصال بازگشتی از خروجی به ورودی، نمونه‌های تعلیم به‌عنوان نقاط تعادل پایدار در شبکه شکل می‌گیرند. در هنگام تست شبکه وقتی نمونه نویزی به‌عنوان ورودی شبکه قرار می‌گیرد، میزان نویز یا اعوجاج خروجی شبکه کمتر از نویز ورودی خواهد بود. حال اگر خروجی حاصل دوباره در ورودی شبکه قرار بگیرد، بار دیگر میزان نویز و اعوجاج آن کاهش می‌یابد. با تکرار این عمل بازگشتی، نمونه‌ای که از وضعیت اصلی خود در فضای ورودی به علت نویز و یا تنوعات خارج شده است، در هر گام به نمونه اصلی نزدیک‌تر می‌شود. این روند تا جایی ادامه پیدا می‌کند که فاصله اقلیدسی خروجی‌های متوالی شبکه کمتر از یک مقدار تعریف شده باشد [۱] و [۱۵]. به بیانی دیگر داده‌های تعلیم به‌عنوان جاذب در فضای ورودی شکل می‌گیرند و با قراردادن داده‌های آغشته به نویز و یا تنوعات ناخواسته به‌عنوان ورودی، به زیرفضای داده‌های تمیز جذب می‌شوند.

تعداد حالت‌های جاذب در شبکه متناسب با تعداد گره‌ها می‌باشد. لذا این امکان وجود دارد که تعداد جاذب‌ها را با افزایش تعدا گره‌ها افزایش دهیم. این گره‌های پیوسته می‌توانند یک مانیفولد پیوسته از جاذب‌های نقطه‌ای را شکل دهند. این مدل‌ها شبکه‌های عصبی با جاذب‌های پیوسته (CANNs)[1] نامیده می‌شوند. این شبکه‌ها همان شبکه‌هایی با جاذب‌های نقطه‌ای هستند که جاذب‌های آن ساختار ویژه‌ای دارند. وقتی سیگنال گفتار که از مجموعه‌ای از نقاط تشکیل شده است به‌عنوان ورودی به شبکه تعلیم داده می‌شود، با قرارگرفتن جاذب‌های نقطه‌ای در کنار یکدیگر یک جاذب پیوسته شکل می‌گیرد که در بر گیرنده مسیر سیگنال گفتار در فضای ورودی می‌باشد [۱] (شکل ۱).

جاذب‌های پیوسته توانایی یادگیری مانیفولدهای غیر خطی و پیچیده را دارند و از این طریق امکان پالایش غیر خطی داده‌های ورودی را فراهم می‌کنند [۱] و [۱۶].

1. Continuous Attractor Neural Networks

۳- مجموعه دادگان

دادگان استفاده‌شده در این تحقیق مربوط به ۸۰۰ جمله از سیگنال گفتار ۴۰ فرد مختلف از مجموعه دادگان فارس دات می‌باشد. با توجه به این که سامانه‌های بازشناسی استفاده‌شده در این تحقیق سامانه‌های بازشناسی گفتار پیوسته می‌باشد، دادگان به‌کار رفته شده نیز سیگنال پیوسته هستند. برای تعلیم شبکه از ۴۰۰ جمله از سیگنال گفتار بیان‌شده توسط ۴۰ گوینده مختلف در طی دو جلسه استفاده شده است. هر سیگنال گفتار ۱۰ جمله مختلف می‌باشد که این جملات در گویندگان مختلف نیز تفاوت می‌کند. این داده‌ها نه تنها در جنسیت متفاوت هستند بلکه بیشتر تنوعات از جمله تغییر در لهجه، بلندی صدا، سرعت بیان و غیره را در بر می‌گیرند که به‌صورت تصادفی از مجموعه دادگان انتخاب شده‌اند.

از آنجایی که جاذب پیوسته زیرفضای داده‌های بدون نویز می‌باشد، از سیگنال‌های تمیز برای تعلیم شبکه استفاده شده است؛ که این سیگنال‌ها ۴۰۰ جمله جلسه دوم دادگان فارس دات می‌باشد. این دادگان شامل تمامی آواهای فارسی می‌باشد در نتیجه می‌توان مدل را با دقت بالایی ارزیابی کرد.

برای تست ساختار ارائه‌شده از ۴۰۰ جمله بیان‌شده توسط ۴۰ گوینده که آغشته به نویز جمعی در نسبت سیگنال به نویزهای متفاوت هستند، استفاده شده است. این بررسی در دو نوع نویز ایستان و غیر ایستان انجام داده شده است که نویز ایستان از نوع نویز سفید و نویز غیر ایستان مربوط به نویز ثبت‌شده در خیابان می‌باشد.

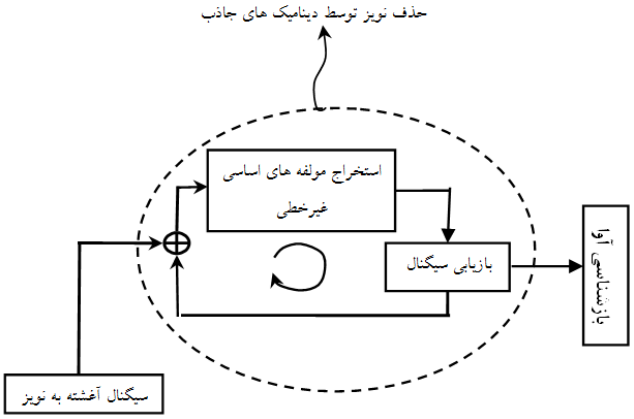
داده‌ها برای قرارگرفتن در ورودی شبکه در چندین مرحله آماده‌سازی می‌شوند. اولین مرحله استخراج ویژگی سطح پایین از سیگنال‌هاست. منظور از سطح پایین، استخراج ویژگی‌ها به‌صورت مستقیم از ورودی است که حاوی اطلاعات خام از داده‌ها است. از بین روش‌های مختلف استخراج ویژگی بهترین نتایج مربوط به بازنمایی‌های طیفی می‌باشد. در این تحقیق از روش لگاریتم انرژی بانک فیلترهای مجذور هنینگ استفاده شده است. این روش در تحقیقات قبلی نتایج مطلوبی در بازشناسی آوا به همراه داشته است [۱]، [۱۷] و [۱۸].

در این روش بردار ویژگی هر فریم به کمک بانک فیلترها به‌دست می‌آید. به این ترتیب که در هر بازه زمانی مشخص، فیلتر اعمال شده و انرژی خروجی هر فیلتر به‌عنوان پارامتر خروجی در نظر گرفته می‌شود. در ابتدا سیگنال گفتار با استفاده از پنجره هنینگ به بازه‌های زمانی کوتاه‌مدت تقسیم می‌شود. با استفاده از تبدیل فوریه زمان کوتاه STFT طیف فرکانسی سیگنال به‌دست می‌آید.

برای به‌دست آوردن اطلاعات مفیدتر، بانک فیلترهای متشکل از ۱۸ فیلتر میان‌گذر در مقیاس بارک که با یکدیگر هم‌پوشانی دارند، بر روی این طیف اعمال می‌شود. برای تبدیل فرکانس از مقیاس هرتز به مقیاس بارک از رابطه زیر استفاده شده است

$$f_{bark} = 6 \ln\left[\frac{f_{HZ}}{600} + \sqrt{\left(\frac{f_H}{600}\right)^2 + 1}\right] \qquad (1)$$

که f_{bark} فرکانس در مقیاس بارک و f_H فرکانس در مقیاس هرتز است. از آنجایی که از ۱۸ فیلتر استفاده شده است، برای هر فریم ۱۸ پارامتر به‌دست می‌آید. در آخرین مرحله بر روی خروجی‌های فیلتر، لگاریتم اعمال می‌شود. لگاریتم‌گیری باعث می‌شود که ویژگی‌های حاصل کمتر به تغییرات دینامیکی حساسیت نشان دهند. از آنجایی که مقدار انرژی بسیار کوچک بوده و حتی در موارد سکوت برابر با صفر نیز می‌شود، انرژی



شکل ۲: بلوک دیاگرام عملکرد شبکه با اتصالات بازگشتی در حذف نویز توسط دینامیک‌های جاذب در آن. در این مدل شبکه به‌گونه‌ای تعلیم می‌بینند که عمل بازشناسی را پس از فیلترسازی نویز انجام دهد. بخش خودانجمنی مدل وظیفه استخراج مؤلفه‌های اساسی غیر خطی از داده را به عهده دارد. این مؤلفه‌ها با نگاشت غیر خطی سیگنال ورودی به زیرفضایی با ابعاد پایین و بازسازی دوباره ورودی از این مؤلفه‌ها حاصل می‌شود. در نتیجه این فرایند تکراری سیگنال نویزی به حالت‌های پایدار که به شکل جاذب پیوسته شکل گرفته‌اند، نگاشته شده و نویز از سیگنال ورودی به تدریج پاکسازی می‌شود.

را با یک مقدار مثبت جمع کرده و سپس لگاریتم را اعمال می‌کنیم. پارامترهای بازنمایی به‌صورت زیر به‌دست خواهند آمد

$$C_j = \log(a + E_j) \qquad (۲)$$

مراحل استخراج پارامترهای بازنمایی را می‌توان به‌صورت زیر خلاصه کرد:

۱) انتخاب یک فریم از سیگنال گفتار به طول ۱۰۲۴ نمونه.
۲) حذف مقدار DC فریم.
۳) اعمال پنجره هنینگ بر روی فریم.
۴) محاسبه تبدیل فوریه ۱۰۲۴نقطه‌ای از فریم پنجره‌گذاری شده.
۵) محاسبه طیف توان.
۶) اعمال فیلترهای مجذور هنینگ بر روی طیف توان.
۷) محاسبه لگاریتم خروجی هر فیلتر به‌عنوان یک پارامتر بازنمایی.

پس از استخراج ویژگی بر روی پارامترهای حاصل عمل هنجارسازی انجام می‌شود.

برای سیگنال‌های گفتار در مجموع دو نوع هنجارسازی طولی و عرضی بر روی سیگنال گفتار مفید می‌باشد. در روش هنجارسازی عرضی، میانگین پارامترهای بازنمایی هر فریم محاسبه شده و سپس این مقدار میانگین از هر یک از پارامترها کاسته می‌شود. در روش هنجارسازی طولی هر پارامتر بازنمایی نسبت به تغییراتی که در کل دادگان دارد، هنجارسازی می‌شود. در این روش ابتدا بردار میانگین پارامترها بر کل دادگان محاسبه می‌شود و در نهایت مقدار حاصل از تک تک پارامترها کاسته می‌شود. سپس میانگین مقادیر حاصل محاسبه شده و تمامی بردارها بر این مقدار حاصل تقسیم می‌شود. در این تحقیق از بازنمایی طولی استفاده شده است.

این نحوه هنجارسازی موجب می‌شود با هم دامنه شدن پارامترهای بازنمایی نسبت به هم، ارزش همه مؤلفه‌ها در برابر مدل بازشناخت تقریباً یکسان شود و مدل بهتر بتواند عمل طبقه‌بندی را روی پارامترهای هنجارسازی انجام دهد. لازم به ذکر است که هر کدام از فریم‌ها در مجموعه دادگان فارس دات برچسب‌دهی شده‌اند.

به هنگام تست شبکه نیز در هر بار ۱۴ فریم از سیگنال آغشته به نویز به‌عنوان ورودی به شبکه داده می‌شود و خروجی آن پس از چندین بار دورزدن پردازش در شبکه با برچسب‌ها مقایسه می‌شود. خروجی‌های مطلوب به‌صورت بردارهای ۳۵بیتی تعریف شده‌اند که بیت متعلق به یک آوای خاص یک و مقادیر بقیه بیت‌ها برابر صفر در نظر گرفته شده‌اند.

۴- روش‌ها

همان‌طور که اشاره شد، ورودی ساختارهای پیشنهادی ۱۴ فریم متوالی از بردار بازنمایی سیگنال گفتار است. در خروجی تشخیص داده می‌شود که این فریم‌ها مربوط به کدام یک از ۳۵ آوای فارسی هستند. خروجی‌های مطلوب نیز به‌صورت بردارهای ۳۵بیتی تعریف شده‌اند که بیت متعلق به یک آوای خاص یک و مقادیر بقیه بیت‌ها برابر صفر در نظر گرفته شده‌اند. تابع فعالیت همه نرون‌ها تابع دوقطبی تانژانت هیپربولیک انتخاب شده است. همچنین تعلیم شبکه در تمامی روش‌ها با الگوریتم پس‌انتشار خطا انجام می‌شود.

۴-۱ ساختار شبکه عصبی بازگشتی جاذب

ساختار پیشنهادی در تحقیق قبلی [۱] برای بازشناسی مقاوم به نویز سیگنال گفتار بر مبنای یک ساختار چندتکلیفی[1] می‌باشد که در این ساختار ابتدا سیگنال ورودی توسط اتصال بازگشتی موجود در لایه پنهان پالایش شده و سپس عمل بازشناسی انجام می‌شود (شکل ۲). در این مدل شبکه تلاش می‌کند تا سیگنال نویزی را به زیرفضای داده‌های تمیز به‌صورت غیر خطی کاهش بعد دهد [۱]. بخش خودانجمنی مدل نقش تصویر سیگنال ورودی به مؤلفه‌های اساسی یا به بیان دیگر ابعاد اصلی گفتار تمیز را به عهده دارد. در نتیجه مؤلفه‌های شکل‌گرفته در لایه پنهان زیرفضای کاهش بعد یافته سیگنال ورودی است [۱۹] و [۲۰]. تلاش برای نگاشت غیر خطی ورودی به ابعاد اساسی تا حد زیادی سیگنال را از نویز و تنوعات ناخواسته پاکسازی می‌نماید [۱] و [۱۴].

به این ترتیب در هنگام تعلیم، در نتیجه اتصال بازگشتی، زیرفضای داده‌های تمیز به‌عنوان جاذب پیوسته شکل می‌گیرد. بستر جذب این جاذب پیوسته کل فضای ورودی خواهد بود که تمام سیگنال‌های نویزی و تغییریافته توسط تنوعات را در بر می‌گیرد [۱]. بلوک دیاگرام عملکرد شبکه در شکل ۲ نشان داده شده است.

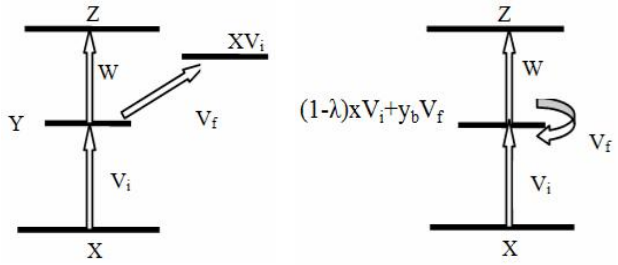
در ساختار ارائه‌شده ماتریس وزن لایه ورودی سعی در ایفای دو نقش را دارد. از یک طرف به‌گونه‌ای تنظیم می‌شود تا ورودی مناسب برای شبکه خودانجمنی را فراهم کند، از طرف دیگر عملکرد بازشناسی آوا را در خروجی شبکه بهبود بخشد. به همین دلیل در اصلاح این وزن دو خطا نقش دارد که یکی از آنها اختلاف خروجی اتصال بازگشتی با حاصل‌ضرب سیگنال ورودی در ماتریس وزن لایه اول و دیگری مربوط به اختلاف خروجی شبکه با مقادیر مطلوب در نظر گرفته شده است. این شبکه دارای ۶۴ نرون در لایه پنهان می‌باشد که ساختار آن در شکل ۳ نمایش داده شده است [۱].

در ادامه نحوه تعلیم شبکه مورد بررسی قرار گرفته است. خروجی لایه پنهان y از مجموع دو مقدار حاصل می‌شود که ضریب دخالت هر کدام از این مقادیر را با λ مشخص می‌کنیم

$$y = f\left((1-\lambda)xV_i + \lambda y_b V_f\right) \qquad (۳)$$

در این رابطه y_b مقدار خروجی لایه پنهان در تکرار قبل دورزدن در حلقه بازگشتی است و x سیگنال ورودی می‌باشد. مقدار λ برابر با ۰/۷ در

1. Multi-Task Learning



(ب) (الف)

شکل ۳: (الف) ساختار شبکه عصبی بازگشتی و (ب) ساختار شبکه با اتصال بازگشتی بازشده بیانگر نحوه تعلیم آن. وزن‌های اتصال بازگشتی V_f به‌گونه‌ای تنظیم می‌شوند که ورودی لایه پنهان در گام قبلی xV_i را در خروجی بخش خودانجمنی بسازند. به این ترتیب جاذب پیوسته در فضای ورودی شکل می‌گیرد. x بیانگر ورودی شبکه و V_i، W، V_f وزن‌های اتصالات هستند و z نشان‌دهنده خروجی نهایی مدل می‌باشد که یک کد ۳۵بیتی است. در نهایت y_b خروجی لایه پنهان در گام قبلی است.

نظر گرفته شده است. به این ترتیب سهم بیشتری از ورودی از اتصال بازگشتی ناشی می‌شود. این کار شبکه را نسبت به نویز مقاوم‌تر می‌کند؛ به دلیل این که شبکه به‌گونه‌ای تعلیم می‌بیند تا از داده اعوجاج‌یافته ورودی تمیز را استخراج کند. به این ترتیب مسیر جذب سیگنال نیز به شبکه تعلیم داده می‌شود.

همان‌طور که اشاره شد، برای اصلاح وزن‌ها دو خطا در شبکه پس‌انتشار می‌شود

$$E_{1n} = \sum_{i=1}^{l} (d(n,i) - z(n,i))^2 \qquad (۴)$$

$$E_{2n} = \sum_{i=1}^{m} [\sum_{j=1}^{m} x(n,j) V_{i_{ji}} - \sum_{j=1}^{m} y_b(n,j) V_{f_{ji}}]^2 \qquad (۵)$$

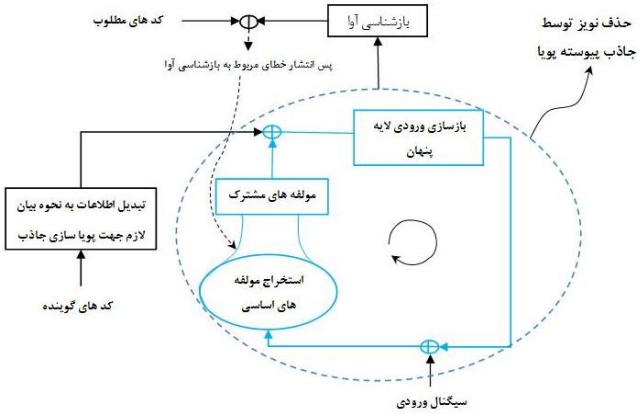
در این رابطه $d(n,i)$ بیانگر مقادیر مطلوب تعریف در زمان n و m تعداد نرون‌ها در لایه پنهان است.

این شبکه با ۱۰ جمله از سیگنال گفتار یک فرد تعلیم می‌بیند و سپس با ۱۰ جمله دیگر سیگنال گفتار همین گوینده که آغشته به نویز ایستان و غیر ایستان هستند، تست می‌شود. نتایج گزارش‌شده نشان می‌دهد که درصد صحت بازشناسی با این روش نسبت به شبکه بدون اتصال بازگشتی به خصوص در نسبت سیگنال به نویز پایین بهبود قابل توجه‌ای دارد [۱].

با این که این ساختار قابلیت ویژه‌ای در حذف نویز دارد ولی در هنگام تعلیم آن با تعداد زیادی از گویندگان؛ به‌علت محدودیت تعداد ابرصفحات و در نتیجه کمبود ظرفیت ذخیره‌سازی؛ عملکرد آن به شدت افت پیدا می‌کند. به‌علت این که شبکه ظرفیت یادگیری سیگنال‌های ناشی از گویندگان مختلف را به‌صورت مفاهیم جداگانه ندارد. به‌عبارتی شبکه قادر نیست جاذب‌های متعددی در فضای ورودی به‌ازای گویندگان مختلف تشکیل دهد. لذا این راه حل به ذهن می‌رسد که به‌جای تعلیم جاذب‌های متعدد به شبکه، تنها یک جاذب پیوسته در فضای ورودی شکل بگیرد که این جاذب با تنوعات سیگنال پویا می‌گردد. یعنی شبکه به‌گونه‌ای تعلیم ببیند که یک فیلتر غیر خطی تطبیقی داشته باشیم.

۲-۴ ساختار پیشنهادی برای استخراج و به اشتراک گذاشتن مؤلفه‌های مشترک

اگر اشتراکات بین الگوهای ورودی در نظر گرفته نشوند، شبکه مجبور می‌شود تا برای بازشناسی آنها را به‌صورت کاملاً مجزا فرض کرده و برای تفکیک آنها از ابرصفحات جداگانه‌ای استفاده کند. این عمل باعث می‌شود تا شبکه نیاز به ظرفیت ذخیره‌سازی بالایی داشته باشد. از آنجایی که



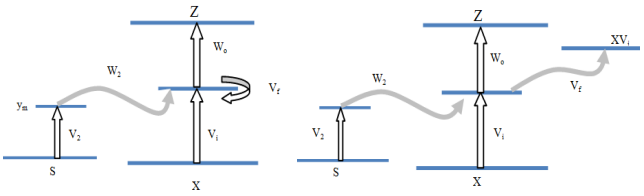
شکل ۴: بلوک دیاگرام عملکرد مدل پیشنهادی جهت استخراج و به اشتراک گذاشتن مؤلفه‌های مشترک. همانند شبکه مرجع بخش خودانجمنی نقش استخراج مؤلفه‌های غیر خطی و فرم‌دادن جاذب پیوسته در فضای ورودی را به عهده دارد. برای تنظیم وزن‌ها، دو خطا در شبکه پس‌انتشار می‌شود. اولین خطا مربوط به خطای بازشناسی آوا در خروجی شبکه است که با پس‌انتشار شدن این خطا آواهای یکسان بیان شده توسط گویندگان مختلف به صورت باسرپرستی خوشه بندی می‌شوند. بخش پویاکننده جاذب نیز با دریافت کدهای گوینده بهترین مقادیر را برای تطبیق مدل با تغییرات گوینده تولید می‌کند.

شبکه توانایی پوشش تمام تنوعات ممکن را ندارد، در نهایت در بازشناسی دچار مشکل می‌گردد. لذا در این تحقیق روشی ارائه شده است تا بتوان مؤلفه‌های مشترک بین داده‌ها را استخراج کرده و با به اشتراک گذاشتن این مؤلفه‌ها طیف وسیعی از تنوعات را پوشش دهیم.

در این ساختار بخش خودانجمنی شبکه در شکل‌گیری جاذب پیوسته نقش دارد. کدهای گوینده نیز از طریق یک شبکه جداگانه به اطلاعات لازم جهت پویاسازی این جاذب پیوسته تبدیل می‌شوند. بلوک دیاگرام عملکرد شبکه در شکل ۴ آورده شده است. خروجی‌های مطلوب شبکه به‌صورت کدهای ۳۵بیتی تعریف شده‌اند. خطای بازشناسی آوا که اختلاف خروجی شبکه با این مقادیر مطلوب است، جهت اصلاح وزن‌ها در شبکه پس انتشار می‌شود. این عمل باعث می‌شود تا آواهای یکسان از گویندگان مختلف در یک خوشه قرار بگیرند. از آنجایی که توسط بخش خودانجمنی مؤلفه‌های اساسی سیگنال استخراج شده است، توسط خطای پس‌انتشار شده این مقادیر به سمت مؤلفه‌های مشترک اساسی هدایت می‌شوند. در نهایت تلاش در این است که تنها یک زیرفضای مشترک به‌ازای گویندگان مختلف شکل بگیرد.

به این ترتیب ابتدا سیگنال ورودی به زیرفضای مؤلفه‌های اساسی نگاشت می‌شود و همچنین به‌طور هم‌زمان اطلاعات لازم جهت پویاسازی جاذب پیوسته از طریق یک شبکه جداگانه از کدهای گوینده استخراج می‌گردد. برای پویاکردن جاذب این اطلاعات به مؤلفه‌های مشترک استخراج‌شده اضافه می‌شود. بخش خودانجمنی مدل تلاش می‌کند تا زیرفضای مربوط به هر گوینده را بازسازی کند که این کار با قراردادن مقدار xV_i به‌عنوان خروجی مطلوب انجام می‌شود. در واقع تنها یک زیرفضای واحد در فضای ورودی شکل می‌گیرد که تنها حاوی اطلاعات پیام سیگنال گفتار است که این زیرفضا توسط اطلاعات گوینده پویا می‌شود. از آنجایی که تمامی قسمت‌های مدل به‌طور هم‌زمان تعلیم می‌بینند، شبکه تلاش می‌کند تا در هر گام با تنظیم بهتر پارامترهای شبکه پویاکننده جاذب، امکان به اشتراک گذاشتن این زیرفضای مشترک را فراهم کند.

اتصال بازگشتی علاوه بر این که در استخراج مؤلفه‌های اساسی سیگنال نقش دارد؛ باعث می‌شود که زیرفضای حاصل به‌صورت یک



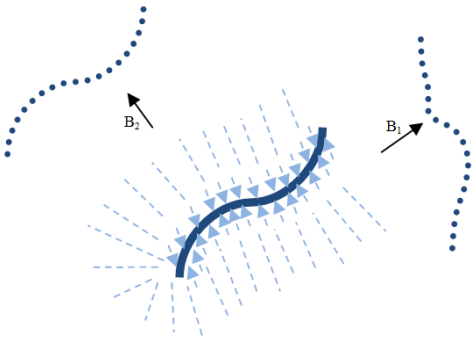
شکل ۵: ساختار شبکه پیشنهادی جهت استخراج و به اشتراک گذاشتن مؤلفه‌های مشترک. S نشان‌دهنده کدهای گوینده است که مدل به‌گونه‌ای تعلیم می‌بیند که این کدها را توسط ماتریس‌های V_2 و W_2 به اطلاعات لازم جهت پویاسازی این مؤلفه‌ها تبدیل نماید.

جاذب پیوسته عمل نماید. در نتیجه سیگنال آغشته به نویز به‌صورت غیر خطی به این زیرفضا نگاشت می‌شود. در نهایت پس از حذف نویز عمل بازشناسی انجام می‌شود.

از آنجایی که بخش زیادی از ورودی لایه پنهان از مسیر اتصالات بازگشتی ناشی می‌شود، شبکه یاد می‌گیرد تا در هر بار دور زدن، سیگنال تمیز را از سیگنال اعوجاج‌یافته استخراج کند. به این ترتیب مؤلفه‌های شکل‌گرفته در لایه پنهان نسبت به اعوجاج مقاوم‌تر خواهند بود. به‌طور خلاصه مدل از سه شبکه با نقش‌های متفاوت که از طریق اتصالات وزن‌دار به یکدیگر مرتبط شده‌اند، ساخته شده است. شبکه اصلی که آواشناس نام‌گذاری شده، نقش بازشناسی آوا از مؤلفه‌های استخراج‌شده در لایه پنهان را به عهده دارد. این شبکه تلاش می‌کند که با خوشه‌بندی آواهای مشترک از گوینده‌های مختلف، اطلاعات مشترک را استخراج کند. شبکه دوم که در وزن لایه ورودی با شبکه آواشناس مشترک است، سیگنال ورودی را به زیرفضای مؤلفه‌های اصلی به‌صورت غیر خطی کاهش بعد می‌دهد، اتصال بازگشتی این شبکه در پدیدآمدن دینامیک‌های جاذب در فضای ورودی نقش دارد. در نهایت شبکه سوم که به نام شبکه پویاکننده جاذب نام‌گذاری شده، از طریق اتصالات وزن‌دار W_2 به لایه پنهان شبکه آواشناس متصل می‌شود. این شبکه تلاش می‌کند تا کدهای گوینده را به اطلاعات لازم جهت پویاسازی جاذب واحد شکل‌گرفته در لایه پنهان تبدیل نماید. در نتیجه مؤلفه‌های مشترکی که توسط شبکه آواشناس استخراج شده‌اند، با کمک اطلاعاتی از گوینده که شبکه پویاکننده جاذب فراهم می‌کند، توسط بخش خودانجمنی در بین داده‌های مختلف به اشتراک گذاشته می‌شوند.

شبکه دارای ۳۵ نرون در لایه خروجی است که این تعداد بر اساس تعداد آواهای فارسی تعیین شده است. لذا خروجی نهایی شبکه به‌صورت یک بردار ۳۵ بیتی خواهد بود که بیتی که دارای ماکزیمم مقدار است بیانگر یک آوای مشخص خواهد بود. خروجی‌های مطلوب نیز به‌صورت بردارهای ۳۵ بیتی تعریف شده‌اند که بیت متعلق به یک آوای خاص یک و مقادیر بقیه بیت‌ها برابر صفر در نظر گرفته شده‌اند.

بعد از چندین بار آزمایش، تعداد نرون‌های لایه‌های پنهان برابر با ۶۴ نرون در نظر گرفته شده است تا شبکه پایه حداقل تعداد ابرصفحات لازم برای جداسازی آواهای سیگنال گفتار تنها یک گوینده را داشته باشد. تابع فعالیت تمامی نرون‌ها نیز تابع غیر خطی سیگموئید دوقطبی است. مقادیر ورودی شبکه پویاکننده جاذب کدهایی است که به‌صورت باینری متناسب با هر گوینده تعریف شده‌اند. لذا تعداد نرون‌های لایه ورودی بخش پویاکننده جاذب برابر با ۶ و تعداد نرون‌های لایه ورودی شبکه اصلی برابر با ۲۵۲ نرون در نظر گرفته شده‌اند که هر ۱۴ فریم از سیگنال ورودی یک الگو ورودی را شکل می‌دهند. ساختار مدل در شکل ۵ ساده‌سازی شده است.



شکل ۶: نمادی از پویایی جاذب‌ها توسط مؤلفه‌های تفاوت.

در این ساختار شبکه به‌گونه‌ای تعلیم می‌بیند تا مؤلفه‌های مشترک شکل‌گرفته در لایه پنهان توسط اطلاعات گوینده، در گویندگان مختلف به اشتراک گذاشته شود. به این ترتیب تنها یک جاذب واحد توسط شبکه خودانجمنی برای تمامی گویندگان ایجاد می‌شود که این جاذب در فضای ورودی پویا می‌باشد. در نهایت مدلی حاصل می‌شود که می‌تواند اشتراکات مجموعه سیگنال‌های گفتار با تنوعات را استخراج کند و همچنین متناسب با این مؤلفه‌های تفاوت، پویایی لازم را برای جبران آنها داشته باشد. از این طریق بازشناسی نسبت به تغییرات ناخواسته مقاوم گردد.

به بیانی دیگر می‌توان گفت که حرکت داده‌ها در فضای ورودی به سمت قعر بستر جذب، تحت کنترل قرار می‌گیرد. یعنی این حرکت از پیش تعیین شده می‌باشد. این عمل باعث می‌شود داده‌هایی که از ناحیه خود خارج می‌شوند به سمت جاذب به درستی حرکت کنند و از هرگونه نویز و تنوعات ناخواسته کاملاً پالایش شوند که این کار با استخراج مؤلفه‌های تفاوت سیگنال‌ها و جبران آنها توسط جاذب پویا انجام می‌پذیرد.

در واقع در این ساختار، سیگنال گفتار به‌صورت مستقل از گوینده و عاری از هر نویز در یک جاذب پیوسته واحد شکل گرفته و با پویایی این جاذب واحد، مؤلفه‌های تفاوت ناشی از گویندگان مختلف جبران می‌شود. با این کار از تعدد جاذب‌های پیوسته جلوگیری می‌شود (شکل ۶).

برای اصلاح وزن‌ها دو خطا در شبکه پس‌انتشار می‌شود که اولین خطا مربوط به خطای بازشناسی آوا، یعنی مجموع مربع اختلاف خروجی نهایی مدل z با مقادیر مطلوب d می‌باشد

$$E_{1n} = \sum_{i=1}^{l} (d(n,i) - z(n,i))^2 \qquad (۶)$$

برای به‌دست آوردن مقادیر لایه پنهان، یک تابع غیر خطی بر روی مجموع سه مقدار حاصل‌ضرب بردار ورودی در وزن لایه ورودی xV_i، خروجی شبکه پویاکننده جاذب b_j و خروجی اتصال بازگشتی در گام قبلی اعمال می‌شود. طبق رابطه زیر

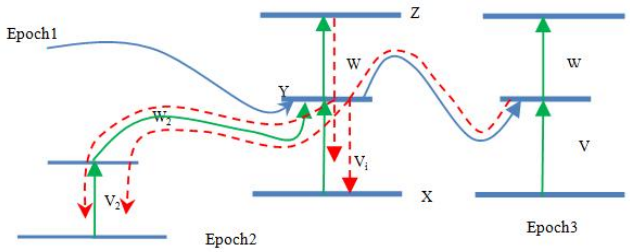
$$y = f\left(\sum_i \sum_j ((1-\lambda) x_i V_i + b_j + \lambda y_{b_i} V_f)\right) \qquad (۷)$$

که در این رابطه b_j خروجی شبکه پویاکننده جاذب است که از روابط زیر حاصل می‌شود

$$y_m = f(sV_2) \qquad (۸)$$

$$b_j = y_m W_2 \qquad (۹)$$

که s بیانگر کدهای باینری متناسب با هر گوینده است و همچنین V_2 و W_2 به‌ترتیب ماتریس وزن‌های لایه اول و دوم شبکه پویاکننده جاذب را نشان می‌دهند.



شکل ۷: نحوه پس‌انتشار خطا در شبکه پیشنهادی، خطوط نقطه‌چین نشان‌دهنده مسیر پس‌انتشار خطا و خطوط ممتد وزن‌های شبکه می‌باشند. برای تنظیم وزن‌ها دو خطا در شبکه پس‌انتشار می‌شود، خطای بازشناسی آوا و خطای بازسازی ورودی لایه پنهان. وزن‌های اتصال بازگشتی به‌گونه‌ای تعلیم می‌بینند که بتوانند ورودی را با استفاده از اطلاعات گفتار و اطلاعات گوینده فراهم‌شده بازسازی کنند.

خطای دیگری که در شبکه پس‌انتشار می‌شود مربوط به خطای بازسازی در بخش خودانجمنی است

$$E_{۲n} = \sum_{i=۱}^{m}[\sum_{j=۱}^{m} x(n,j)V_{i_{ji}} - \sum_{j=۱}^{m} y_b(n,j)V_{f_{ji}}]^۲ \qquad (۱۰)$$

که در این رابطه y_b مقدار خروجی لایه پنهان در تکرار قبل دورزدن در حلقه بازگشتی است. سیگنال خطایی که به لایه پنهان پس‌انتشار می‌شود تحت رابطه زیر محاسبه می‌شود

$$\delta_y = (۱-y^۲)(\delta_z W + ۰/۰۰۱(xV_i + b_j - \tilde{y})V_f) \qquad (۱۱)$$

δ_z بردار خطای مربوط به بازشناسی آوا، W ماتریس وزن لایه خروجی شبکه و $\tilde{y}$ ورودی لایه پنهان در گام جاری می‌باشد که از رابطه زیر حاصل می‌شود

$$\tilde{y} = (۱-\lambda)xV_i + \lambda\tilde{y}_b V_f + b_j \qquad (۱۲)$$

$(۱-y^۲)$ نیز مشتق تابع سیگموئید دوقطبی است. در نهایت $\tilde{y}_b$ خروجی لایه پنهان در گام قبلی و λ ضریب خطای بازسازی بوده که برابر با ۰/۷ در نظر گرفته شده است. مسیر پس‌انتشار شدن خطا در شکل ۷ نمایش داده شده است.

لازم به ذکر است که مقادیر ورودی شبکه پویاکننده جاذب کدهایی است که به‌صورت باینری متناسب با هر گوینده تعریف شده‌اند. در نهایت با استفاده از الگوریتم پس‌انتشار خطا وزن‌ها اصلاح می‌شود. تعلیم شبکه تا جایی ادامه پیدا می‌کند تا به حداقل خطای در نظر گرفته شده برای بازشناسی دست پیدا کنیم.

۵- ارزیابی شبکه‌ها

۵-۱ نتایج

برای ارزیابی این شبکه‌ها از ۴۰۰ جمله جلسه اول مجموعه دادگان فارس دات که آغشته به نویز ایستان و غیر ایستان جمعی می‌باشند، استفاده شده است. همان‌طور که قبلاً نیز اشاره شد، هر ۱۰ جمله از تعداد زیادی فریم تشکیل شده که در هر بار ۱۴ فریم به‌عنوان ورودی شبکه قرار می‌گیرند. به‌ازای این ۱۴ فریم یک کد مطلوب تعریف شده است و در نهایت تعداد دفعاتی که پاسخ شبکه به‌ازای تقریباً ۲۷۰۰۰ فریم صحیح بوده به‌صورت درصد صحت بیان می‌شود. اعداد بیان‌شده در جدول نتایج ۱ و ۲ میانگین درصد صحت‌های به‌دست آمده به‌ازای ۴۰ گوینده مختلف می‌باشد.

البته لازم به ذکر است که با توجه به این که هدف این مقاله تلاش

جدول ۱: میانگین نتایج شبکه‌ها به سیگنال آغشته به نویز ایستان ۴۰ گوینده.

SNR	میانگین نتایج شبکه بازگشتی جاذب	میانگین نتایج مدل پیشنهادی
۰ db	٪۳۵/۳۷	٪۴۰/۷۵
۵ db	٪۴۲/۸۲	٪۴۶/۸۴
۱۰ db	٪۴۸/۱۱	٪۵۱/۴۲
۱۵ db	٪۵۲/۰۹	٪۵۴/۰۵
۲۰ db	٪۵۴/۳۳	٪۵۵/۵۸
سیگنال تمیز	٪۵۵/۶۹	٪۵۶/۶۲

جدول ۲: میانگین نتایج شبکه‌ها به سیگنال آغشته به نویز غیر ایستان ۴۰ گوینده.

SNR	میانگین نتایج شبکه بازگشتی جاذب	میانگین نتایج مدل پیشنهادی
۰ db	٪۳۴/۰۴	٪۳۷/۶۷
۵ db	٪۳۹/۶۵	٪۴۳/۵۷
۱۰ db	٪۴۴/۸۳	٪۴۸/۲۸
۱۵ db	٪۴۹/۰۷	٪۵۲/۵۱
۲۰ db	٪۵۲/۶۷	٪۵۴/۶۹
سیگنال تمیز	٪۵۵/۶۹	٪۵۶/۶۲

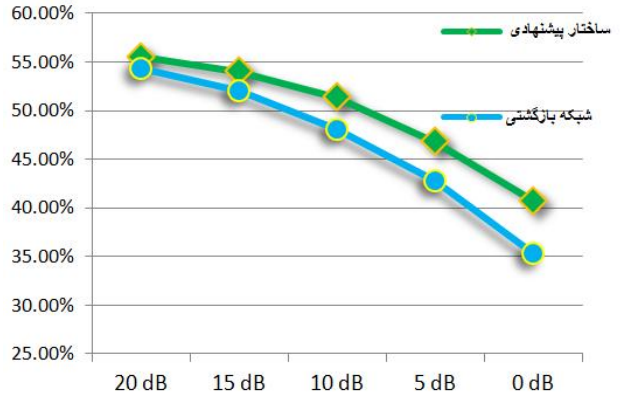
برای پویاسازی جاذب‌ها می‌باشد، لازم است ساختار پیشنهادی با شبکه‌ای با ساختار مشابه بدون اتصالات لازم برای پویاسازی جاذب‌ها مقایسه شود تا بتوان میزان موفق‌بودن روش را در به اشتراک گذاشتن مؤلفه‌های مشترک بررسی کرد.

نتایج گردآوری شده در جداول عملکرد قوی‌تر مدل پیشنهادی را نسبت شبکه بازگشتی، نشان می‌دهد. بدیهی است که با کاهش نسبت سیگنال به نویز احتمال خطا در بازشناسی افزایش می‌یابد و شبکه در بازشناسی دچار مشکل می‌شود. در این حالت، جاذب‌های پیوسته نویز و تنوعات ناخواسته را تا حدودی فیلتر نموده و درصد بازشناسی را افزایش می‌دهند. لذا تأثیر عملکرد جاذب‌های پیوسته در نسبت سیگنال به نویز پایین پررنگ‌تر می‌باشد.

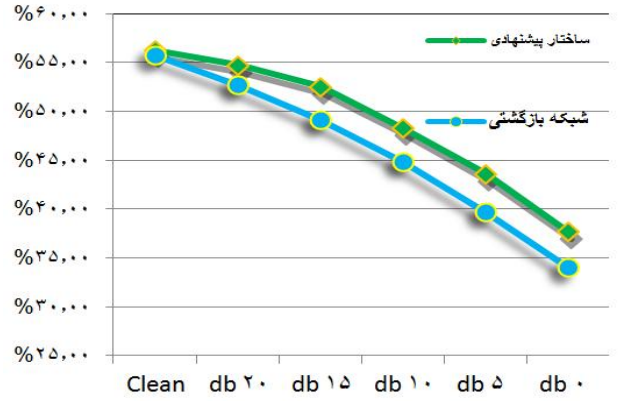
همان‌طور که می‌بینیم در مورد مدل ارائه‌شده، فاصله درصد صحت بازشناسی داده‌های با نویز بالا نسبت به درصد صحت بازشناسی داده‌های با نویز پایین بسیار کم است. به بیانی دیگر شبکه داده‌های با نسبت سیگنال به نویز ۰ db را با تفاوت تنها ۶ درصد نسبت به سیگنال تمیز بازشناسی کرده است. یعنی می‌توان نتیجه گرفت که جاذب‌ها در حذف نویز بسیار خوب عمل کرده‌اند و مشکل شبکه در بازشناسی، مربوط به جبران تغییر گوینده سیگنال ورودی است.

نکته قابل توجه دیگر این که طبق منحنی‌های رسم‌شده در شکل ۸ و ۹ با کاهش نسبت سیگنال به نویز، میزان بهبود صحت بازشناسی آوا در مدل پیشنهادی افزایش یافته است. به بیانی دیگر اختلاف بین دو مدلی که با یکدیگر مقایسه شده‌اند با کاهش نسبت سیگنال به نویز افزایش می‌یابد.

با افزایش درصد نویز سیگنال ورودی، وابستگی عملکرد مدل به جاذب‌ها بیشتر می‌شود. لذا از آنجایی که با به اشتراک‌گذاشتن مؤلفه‌های مشترک عملکرد جاذب‌ها بهبود یافته است، تفاوت عملکرد دو مدل در نسبت سیگنال به نویز پایین پررنگ‌تر می‌شود و این بیانگر عملکرد قوی‌تر جاذب‌ها با به اشتراک‌گذاشتن مؤلفه‌های مشترک است. بنابراین همان‌طور که انتظار می‌رفت مدل پیشنهادی نسبت به شبکه بازگشتی در نسبت سیگنال به نویز پایین بهتر عمل می‌نماید.



شکل ۸: نمودار ارزیابی میانگین نتایج شبکه‌ها به سیگنال آغشته به نویز ایستان ۴۰ گوینده با تغییر نسبت سیگنال به نویز.



شکل ۹: نمودار ارزیابی میانگین نتایج شبکه‌ها به سیگنال آغشته به نویز غیر ایستان ۴۰ گوینده با تغییر نسبت سیگنال به نویز.

۵-۲ بررسی اهمیت آماری نتایج

برای بررسی اهمیت آماری از روش توضیح داده شده در ذیل استفاده شده است.

برای مقایسه عملکرد دو الگوریتم A_1 و A_2 که سری آواهای $\{u_i\}=u_1,\ldots,u_n$ را بازشناسی می‌کنند، از این روش استفاده می‌شود. میزان خطای واقعی این دو الگوریتم که نامعلوم است با p_1 و p_2 نشان داده می‌شود. هدف بررسی این مطلب است که آیا شواهد کافی آماری برای اثبات رابطه $p_1>p_2$ ، $p_1<p_2$ و یا $p_1=p_2$ وجود دارد. متغیر X_i^j به‌صورت زیر تعریف می‌شود که اگر آوای u_i را الگوریتم j درست تشخیص داده مقدار X_i^j برابر با ۱ و در غیر این صورت برابر با صفر قرار داده می‌شود. S^j نیز برابر با تعداد دفعاتی است که الگوریتم j آوا را درست تشخیص داده است

$$S^j=\sum_{i=1}^{n}X_i^j \qquad (۱۳)$$

منطقی است اگر فرض شود S از یک توزیع دوجمله‌ای $B(n,p_j)$ پیروی می‌کند، به‌دلیل این که خطاها از یکدیگر مستقل هستند. تخمین ماکزیمم احتمال p_j از رابطه زیر به‌دست می‌آید

$$\hat{P}_j=\frac{S^j}{n} \qquad (۱۴)$$

واریانس p_j برابر با σ_j^2 می‌باشد که از رابطه زیر حاصل می‌شود

$$\sigma_j^2=\frac{P_j(1-P_j)}{n} \qquad (۱۵)$$

هدف ارزیابی فرضیه باطل H_0 می‌باشد

$$H_0: p_1=p_2=p \qquad (۱۶)$$

با توجه به این که $d=p_1-p_2=0$ است. طبق فرضیه H_0، تخمین ماکزیمم احتمال d برابر با $\hat{P}_1-\hat{P}_2$ با واریانس σ_d^2 است

$$\sigma_d^2=Var(P_1-P_2) \qquad (۱۷)$$

اگر $\hat{P}_1$ و $\hat{P}_2$ مستقل باشند این معادله می‌تواند به‌صورت زیر نوشته شود

$$\sigma_d^2=Var(P_1)-Var(P_2)=\sigma_1^2+\sigma_2^2 \qquad (۱۸)$$

اگر فرضیه H_0 صحیح باشد، σ_d^2 می‌تواند به‌صورت زیر تخمین زده شود

$$\sigma_d^2=\frac{2\hat{P}(1-\hat{P})}{n} \qquad (۱۹)$$

که در این رابطه تخمین ماکزیمم احتمال P برابر با $\hat{P}_1+\hat{P}_2/2$ می‌باشد. اگر n به اندازه کافی بزرگ باشد و فرضیه H_0 صحیح باشد، توزیع آماری w به سمت توزیع نرمال با میانگین صفر و واریانس یک گرایش دارد

$$W=\frac{(\hat{P}_1-\hat{P}_2)}{\sqrt{\frac{2\hat{P}(1-\hat{P})}{n}}} \qquad (۲۰)$$

برای تست فرضیه باطل مقدار $P=2\Pr(Z\geq|W|)$ محاسبه می‌شود که Z یک مقدار تصادفی با توزیع نرمال میانگین ۰ و واریانس ۱ است. اگر p محاسبه‌شده کمتر از مقدار خاص انتخاب‌شده α باشد، فرضیه H_0 رد می‌شود (مقادیر معمول α، ۰/۰۵، ۰/۰۱ و ۰/۰۰۱ می‌باشد). در صورتی که این فرض رد نشود می‌توان نتیجه گرفت که تفاوت بازدهی این دو الگوریتم تصادفی است و این اختلاف اهمیت آماری ندارد [۲۱].

در این تحقیق برای سنجش اهمیت آماری نتایج، تنها برای نسبت سیگنال به نویز db ۰ ایستان را بررسی می‌کنیم. برای این که فرض مستقل‌بودن $\hat{P}_1$ و $\hat{P}_2$ معتبر باشد، لازم است که دو الگوریتمی که با هم مقایسه می‌شوند بر روی داده‌های یکسان تست نشوند. لذا برای الگوریتم اول اهمیت آماری نتایج تست بر روی گوینده شماره ۱ تا گوینده شماره ۲۰ و برای الگوریتم دوم از گوینده شماره ۲۱ تا ۴۰ را بررسی می‌کنیم. الگوریتم A_1 و A_2 به‌ترتیب تعداد خطای ۲۷۶۲۵۵ و ۲۶۰۷۹۰ را ایجاد می‌کنند. از آنجایی که مقدار n برابر با ۴۴۸۹۲۰ می‌باشد، مقادیر $\hat{P}_1=0/615$ و $\hat{P}_2=0/581$ حاصل می‌شود. لذا مقدار $\hat{P}$ برابر با ۰/۵۹۸ خواهد بود و با قراردادن در (۲۳) مقدار W به‌دست خواهد آمد. از آنجایی که $z=|W|$ می‌باشد مقدار P به‌صورت زیر به‌دست خواهد آمد

$$P'=\int_{-\infty}^{z}\phi(t)\,dt\rightarrow P=2\int_{z}^{\infty}\phi(t)\,dt=2(1-p')=0/0006 \qquad (۲۱)$$

از آنجایی که مقدار p بسیار کمتر از α می‌باشد لذا فرض H_0 صحیح نبوده و بهبود در نتایج توسط الگوریتم پیشنهادی از اهمیت آماری بالایی برخوردار است.

۶- بحث و نتیجه‌گیری

همان‌طور که در بخش مقدمه نیز اشاره شد، وقتی سامانه دارای ابعاد و ظرفیت ذخیره‌سازی پایینی است شبکه با مشکل عدم توانایی در تفکیک‌پذیری مواجه می‌شود به‌دلیل این که توانایی ذخیره‌سازی داده‌هایی با ابعاد بالا به‌صورت مفاهیم جداگانه را ندارد. با استخراج ویژگی‌های مشترک، سامانه بازشناسی تنها زیرفضای واحد اشتراکات را یاد می‌گیرد

که این زیرفضا متناسب با تنوعات قابلیت به اشتراک گذاشته شدن بین داده‌های مختلف را دارد. در این مقاله تلاش کردیم تا با استفاده از یک ساختار چندتکلیفی یک جاذب پیوسته واحد که زیرفضای اشتراکات بین داده‌ها یعنی اطلاعات مربوط به پیام سیگنال گفتار است، در فضای ورودی شکل بگیرد که این زیرفضا قابلیت به اشتراک گذاشته شدن بین داده‌های مختلف را دارد. از آنجایی که این زیرفضای مشترک یک جاذب پیوسته واحد می‌باشد، با دورزدن شبکه سیگنال‌های نویزی به‌صورت غیر خطی به این زیرفضا نگاشت می‌شوند. به بیانی دیگر داده‌های تست به ابعاد اساسی خود کاهش بعد می‌یابند. لازم به ذکر است که در این ساختار بر خلاف بسیاری از سامانه‌های بازشناسی گفتار اطلاعات گوینده حذف نمی‌شود بلکه از این اطلاعات بهره گرفته تا بتوانیم بازشناسی دقیق‌تری داشته باشیم. به این صورت که اطلاعات گوینده تعیین می‌کنند که سیگنال آغشته به نویز به کدام زیرفضا جذب شود. در نتیجه خطاهایی در بازشناسی که به دلیل تنوعات ایجاد شده‌اند به حداقل می‌رسد. لذا با این که شبکه از نظر تعداد ابرصفحات و تعداد لایه‌ها با محدودیت مواجه است، می‌تواند سیگنال‌های گفتار ناشی از گویندگان مختلف را تا حد ممکن با کیفیت بالایی استخراج کند.

همان‌طور که در بخش نتایج نیز مشاهده کردیم این ساختار توانسته است درصد صحت بازشناسی را تا حدود ۵ درصد در نسبت سیگنال به نویز dB ۰ افزایش دهد. به نظر می‌رسد که برای بهبود بیشتر عملکرد شبکه، نیاز به داشتن شبکه آواشناس با کارایی بالاتری هستیم تا بتواند ویژگی‌های داده‌های ورودی را در مراحل مختلف استخراج و با ترکیب آنها در بازشناسی آوا موفق‌تر عمل کند. همان‌گونه که سامانه بازشناسی در انسان نیز برای تشخیص از ویژگی‌های در سطح بالاتر مانند مفاهیم و معانی بهره می‌گیرد که رسیدن به چنین اهدافی نیازمند طراحی ساختارهای سلسله مراتبی است.

مراجع

[1] L. Dehyadegary, S. A. Seyyedsalehi, and I. Nejadgholi, "Nonlinear enhancement of noisy speech, using continuous attractor dynamics formed in recurrent neural networks," *J. Neurocomputing in Press*, vol. 74, no. 17, pp. 2716-2724, Jun. 2011.

[2] M. P. Ghaemmaghami, F. Razzazi, H. Sameti, S. Dabbaghchian, and B. BabaAli, "Noise reduction algorithm for robust speech recognition using MLP neural network," in *2nd Asia-Pacific Conf. on Computational Intelligence and Industrial Applications, IEEE*, vol. 2, pp. 377-380, Nov. 2009.

[3] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. of IEEE Computer Vision and Pattern Recognition Conf.,* pp. 1735-1742, Oct. 2006.

[4] M. Ranzato and Y. LeCun, "A sparse and locally shift invariant feature extractor applied to document images," in *Proc. of IEEE Int. Conf. on Document Analysis and Recognition,* vol. 2, pp. 1213-1217, Sep. 2007.

[5] H. Wersing and E. Korner, "Learning optimized features for hierarchical models of invariant object recognition," *Neural Computation*, vol. 15, no. 7, pp. 1559-1588, Jul. 2003.

[6] Y. Wu, X. Liu, and W. Mio, "Learning representations for object classification using multi-stage optimal component analysis," *Neural Networks*, vol. 21, pp. 214-221, Dec. 2008.

[7] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1-127, 2009.

[8] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. of the 25th Int. Conf. on Machine Learning*, vol. 307, pp. 1096-1103, 2008.

[9] M. Ranzato, F. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. of IEEE Conf. one Computer Vision and Pattern Recognition,* 8 pp., Jun. 2007.

[10] M. Ranzato, Y. L. Boureau, and Y. LeCun, *Sparse Feature Learning for Deep Belief Networks*, MIT Press, 2008.

[11] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311-3325, Dec. 1997.

[12] S. Aly, N. Tsuruta, and R. Taniguchi, "Feature map sharing hypercolumn model for shift invariant face recognition," *Artificial Life and Robotics*, vol. 14, no. 2, pp. 271-274, May 2009.

[13] T. P. Trappenberg, "Continuous attractor neural networks," in *Recent Developments in Biologically Inspired Computing*, L. N. de Castro and F. J. Von Zuben, eds., IDEA Group Publishing, 2003.

[14] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645- 678, May 2005.

[15] I. B. Ciocoiu, "Invariant pattern recognition using analog recurrent associative memories," *Neurocomputing*, vol. 73, no. 1-3, pp. 119-126, Dec. 2009.

[16] Z. Hu, X. Fan, Y. Song, and D. Liang, "Joint trajectory tracking and recognition based on bi - directional nonlinear learning," *Image and Vision Computing*, vol. 27, no. 9, pp. 1302-1312, Aug. 2009.

[۱۷] ک. کریمی، *به‌کارگیری مشخصات گوینده در جهت بهبود کیفیت مدل‌های بازشناخت گفتار*، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی امیرکبیر، دانشکده مهندسی پزشکی، ۱۳۸۱.

[۱۸] م. ولی و س. ع. سیدصالحی، "ارزیابی کارایی دو بازنمایی MFCC و LHCB در بازشناسی مقاوم به تنوعات گفتار مستقیم و تلفنی،" *دهمین کنفرانس سالانه انجمن کامپیوتر ایران*، جلد ۱، صص. ۳۱۲–۳۰۵، آذر ۱۳۸۳.

[19] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans., Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[20] J. Zhang and S. Z. Li, "Adaptive nonlinear auto - associative modeling through manifold learning," *Lecture Notes in Computer Science*, vol. 3518, pp. 599-604, May 2005.

[21] L. Gillick and S. Cox, "Some statistical issues in the comparision of speech recognition algorithms," in *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing,* vol. 1, pp. 532-535, *Glasgow*, UK, May 1989.

پروین زارعی اسکی کند در سال ۱۳۸۶ مدرک کارشناسی مهندسی پزشکی خود را از دانشگاه آزاد اسلامی واحد علوم و تحقیقات و در سال ۱۳۸۹ مدرک کارشناسی ارشد مهندسی پزشکی خود را از دانشگاه صنعتی امیرکبیر دریافت نمود. زمینه‌های پژوهشی مورد علاقه ایشان پردازش سیگنال، مدل سازی عملکرد مغز و هوش مصنوعی است.

سید علی سیدصالحی مدرک کارشناسی خود را در مهندسی برق از دانشگاه صنعتی شریف در سال ۱۳۶۱، کارشناسی ارشد را در مهندسی برق از دانشگاه صنعتی امیرکبیر در سال ۱۳۶۷ و دکتری خود را در مهندسی برق- بیوالکتریک از دانشگاه تربیت مدرس در سال ۱۳۷۴ دریافت نموده است. وی در حال حاضر دانشیار دانشکده مهندسی پزشکی دانشگاه صنعتی امیرکبیر می‌باشد. زمینه‌های پژوهشی مورد علاقه ایشان پردازش و بازشناسی گفتار، شبکه‌های عصبی مصنوعی و زیستی، مدل‌سازی عملکرد مغز و پردازش خطی و غیرخطی سیگنال است.