معرفی شبکه‌های عصبی پیمانه‌ای عمیق با ساختار فضایی-زمانی دوگانه جهت بهبود بازشناسی گفتار پیوستهٔ فارسی

زهره انصاری و سیّدعلی سیّدصالحی

آزمایشگاه پردازش گفتار، دانشکدهٔ مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

در این مقاله به معرفی شبکه‌های عصبی پیمانه‌ای عمیق و قابل رشد به‌منظور بهبود بازشناسی گفتار پیوسته پرداخته می‌شود. ساختار این شبکه‌ها و روش‌های پیش‌تعلیم معرفی‌شده برای آنها به‌گونه‌ای است که درعین هماهنگی با ساختار گفتار، در حافظه و محاسبات لازم صرفه‌جویی می‌شود. به‌دلیل قابلیت رشد این ساختارها، در تعلیم آنها اطلاعات فضایی-زمانی بردارهای بازنمایی در ورودی و اطلاعات فضایی-زمانی برچسب آوایی آنها را در خروجی شبکه عصبی می‌توان انجمن کرد. شبکه تعلیم‌یافته با این ساختار انجمن‌گر فضایی-زمانی دوگانه، می‌تواند زیرفضای زنجیره‌های معتبر آوایی دادگان را یاد بگیرد. بنابراین، در ساختار خود زنجیره‌های خروجی نامعتبر را پالایش کرده و زنجیره‌های درست را می‌دهد. جهت بررسی عملکرد این ساختارها، از دو دسته دادگان گفتاری فارس‌دات و فارس‌دات بزرگ استفاده شد. نتایج آزمایش‌ها نشان می‌دهند که می‌توان دقت بازشناسی آوا را برروی دادگان فارس‌دات تا ۲/۷٪ با استفاده از شبکه‌های عصبی پیمانه‌ای عمیق نسبت به مدل‌های مخفی مارکوف بالا برد؛ که با توسعه آنها به ساختار فضایی-زمانی دوگانه این نتیجه تا ۵/۱٪ بهبود می‌یابد. به‌دلیل عدم وجود برچسب‌های آوایی برای دادگان بزرگ، یک روش تعلیم نیمه‌سرپرستی‌شده برای تعلیم شبکه‌های عصبی برروی این دادگان پیشنهاد شده است که می‌تواند به درصد بازشناسی قابل مقایسه‌ای با مدل‌های مخفی مارکوف دست یابد.

واژگان کلیدی: شبکه‌های عصبی عمیق، شبکه‌های عصبی پیمانه‌ای، پیش‌تعلیم، تعلیم نیمه‌سرپرستی شده، بازشناسی گفتار پیوسته.

۱- مقدّمه

امروزه سامانه‌های بازشناسی گفتار خودکار[1] (ASR) به‌صورت گسترده در تعامل انسان با ماشین به‌کار می‌روند. در این سامانه‌ها، مدل‌های مخفی مارکوف[2] (HMM) یا شبکه‌های عصبی[3] (ANN) می‌توانند به‌عنوان مدل‌های صوتی برای مدل‌کردن بردارهای بازنمایی گفتاری به‌صورت واحدهای گفتاری (مثل آوا یا سه‌آوا[4]) استفاده شوند. در سامانه‌های ASR با مدل صوتی HMM، در هر حالت[5] HMM، مدل‌های مخلوط گوسی[6] (GMM) برای مدل‌کردن بردارهای بازنمایی گفتاری مورد استفاده قرار می‌گیرند. مطالعات زیادی نیز برای بهبود این مدل‌ها (GMM-HMM) با هدف بالابردن درصد بازشناسی گفتار انجام شده است (کاپادیا و همکاران، ۱۹۹۳؛ مکدرموت و همکاران، ۲۰۰۷؛ جوانگ و همکاران، ۱۹۹۷).

در دو دهه گذشته، به‌صورت گسترده به استفاده از ساختارهای شبکه عصبی با یک یا دو لایهٔ پنهان و هم‌چنین، ترکیب شبکه‌های عصبی با HMM (ANN-HMM) در بازشناسی گفتار پرداخته می‌شد (وایبل و همکاران، ۱۹۸۹؛ مورگان و بورلارد، ۱۹۹۰؛ بنجیو، ۱۹۹۱؛ بورلارد و مورگان، ۱۹۹۳-۱۹۹۴؛ ترنتین و گوری، ۲۰۰۱؛ و سیدصالحی، ۱۳۷۴). اما عدم وجود سخت‌افزارهای با سرعت بالا و روش‌های یادگیری شبکه‌های عصبی با ساختار عمیق عملکرد ساختارهای شبکه عصبی را محدود می‌ساخت. به همین دلیل، مدل‌های GMM-HMM بیش از شبکه‌های عصبی مورد استفاده قرار گرفتند.

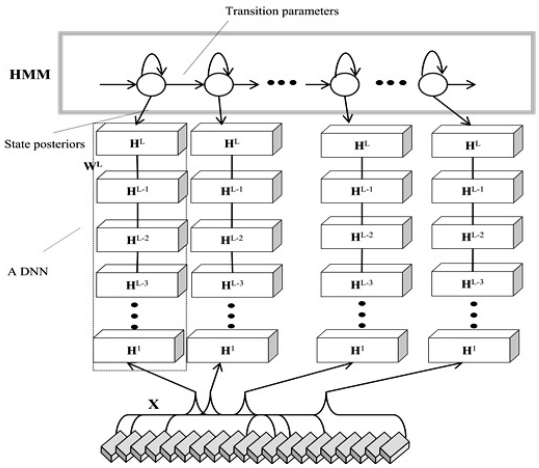
[1] Automatic Speech Recognition Systems (ASR)
[2] Hidden Markov Models (HMM)
[3] Artificial Neural Networks (ANN)
[4] Triphone
[5] State
[6] Gaussian Mixture Models (GMM)

امروزه با پیشرفت‌هایی که در زمینهٔ معرفی روش‌های یادگیری عمیق و ارائه سخت‌افزارهای با سرعت پردازش بالا به‌دست آمده، امکان استفاده از شبکه‌های عصبی عمیق[1] (DNN) در بازشناسی گفتار به‌وجود آمده است. استفاده از شبکه‌های عصبی عمیق در سامانه‌های گفتاری با دادگان‌های گفتاری متفاوت، نشان می‌دهد که این شبکه‌ها بهتر از GMMها در مدل‌سازی صوتی می‌توانند عمل کنند (سیده و همکاران، ۲۰۱۱؛ محمد و همکاران، ۲۰۱۲؛ دهل و همکاران، ۲۰۱۲؛ جیتلی و همکاران، ۲۰۱۲؛ یو و همکاران، ۲۰۱۳).

از جمله کاربردهای شبکه‌های عصبی عمیق در بازشناسی گفتار می‌توان به استفاده از ترکیب DNN-HMM در بازشناسی آوا اشاره کرد (دنگ و همکاران، ۲۰۱۲؛ محمد و همکاران، ۲۰۱۲؛ اندرو و بیلمس، ۲۰۱۲؛ عبدلحمید و همکاران، ۲۰۱۲). در این منابع و در بیشتر پژوهش‌های انجام شده در این مورد، همان‌طور که در شکل (۱) نشان داده شده است، ابتدا خروجی شبکه ایستای تعلیم یافته را به‌ازای یک پنجره متحرک از توالی قاب‌های ورودی به‌دست می‌آورند؛ سپس، مدل‌های گرافیکی مثل HMM یا میدان تصادفی شرطی[2] (CRF) برای مدل‌کردن وابستگی‌های زنجیره‌ای خطی توالی خروجی به‌دست آمده ترکیب می‌شوند. درواقع، در این ساختارها شبکهٔ عصبی عمیق به‌عنوان مدل صوتی و مدل‌های گرافیکی اشاره شده به‌عنوان مدل‌های آوایی مورد استفاده قرار می‌گیرند و هر یک به‌صورت جداگانه تعلیم می‌بینند.

شبکه‌های عصبی عمیق به‌کارگرفته‌شده در مدل‌سازی صوتی به‌طورمعمول دارای تعداد زیادی لایه‌های پنهان غیرخطی با تعداد نورون‌های یکسان در هر لایه هستند. لایهٔ خروجی این شبکه‌ها نیز دارای تعداد زیادی نورون برابر با تعداد حالت[3]های مدل‌های آوایی است. بنابراین، تعداد پارامترهای این شبکه‌ها بسیار زیاد است. با وجود معرفی روش‌های پیش‌تعلیم لایه‌به‌لایه (هینتون و همکاران، ۲۰۰۶؛ بنجیو و همکاران، ۲۰۰۷) برای هم‌گرایی این شبکه‌ها، زمان و حافظه زیادی به‌منظور تعلیم هر لایه مورد نیاز است. برای بهینه‌ساختن ساختارهای شبکه‌های عصبی عمیق با هدف هماهنگی بیشتر با ساختار گفتار و کاهش تعداد پارامترهای تعلیمی و زمان انجام محاسبات، یکی از راه‌کارهای معرفی‌شده، استفاده از شبکه‌های کانولوشنی است(لیکان و بنجیو، ۱۹۹۵)؛ اما هنوز موارد مهمی در مورد طراحی ساختارهای شبکه‌های عصبی عمیق در نظر گرفته نشده است.

از سوی دیگر، روش‌های مختلفی برای تعلیم اطلاعات توالی به شبکه‌های عصبی عمیق با هدف بالابردن کارآیی آنها در بازشناسی گفتار، پیشنهاد شده است (بریدل و داد، ۱۹۹۱؛ کونیگ، ۱۹۹۶؛ هه و همکاران، ۲۰۰۸؛ کینگسبری، ۲۰۰۹؛ محمد و همکاران،۲۰۱۰؛ پراباوالکار و فوسلر-لوسیر، ۲۰۱۰؛ لواندوسکی و همکاران، ۲۰۱۴). اما در تمامی روش‌های معرفی‌شده فرض بر این است که این شبکه‌ها در ترکیب با مدل‌های دیگری که نقش مدل آوایی را دارند به بازشناسی گفتار می‌پردازند. در هیچ یک از این روش‌ها یک ساختار یک‌پارچه شبکهٔ عصبی که بتواند اطلاعات توالی آوایی را به همراه اطلاعات صوتی، به‌تنهایی از روی دادگان استخراج کند، معرفی نشده است. در صورت معرفی چنین ساختاری، وظیفهٔ مدل‌سازی صوتی و هم‌چنین مدل‌سازی آوایی می‌تواند در یک ساختار یک‌پارچه شبکه عصبی انجام شود. بنابراین، این اطلاعات می‌توانند برای رسیدن به بازشناسی بهینه با یکدیگر تعامل داشته باشند.



(شکل- ۱): ساختار ترکیبی DNN-HMM استفاده شده در بسیاری از سامانه‌های بازشناسی گفتار. X توالی بردارهای ویژگی گفتاری است. مدل HMM متوالی بودن سیگنال گفتار را مدل می‌کند و مدل DNN یک مدل واحد است که در زمان‌های مختلف کپی شده است.

برای دست‌یافتن به دو هدف طراحی ساختار مناسب شبکه‌های عصبی به‌منظور بازشناسی گفتار و امکان تعلیم اطلاعات توالی به همراه اطلاعات صوتی به آنها، در این مقاله، شبکه‌های عصبی پیمانه‌ای عمیق قابل رشد معرفی می‌شود. در ساختار این شبکه‌ها، بعد از نگاشت بردارهای بازنمایی ورودی به یک لایه با بعد بالاتر، هر آشکارساز ویژگی در

[1] Deep Neural Networks (DNN)
[2] Conditional Random Field (CRF)
[3] state

لایه‌های بعدی تنها از محدودهٔ کوچکی از ورودی خود اطلاعات دریافت می‌کنند. هم‌چنین، ورودی‌های آشکارسازهای مختلف با یکدیگر هم‌پوشانی دارند. و از ایده به اشتراک گذاشتن وزن‌های آشکارسازهای مختلف در یک لایه مشابه شبکه‌های کانولوشنی استفاده شده است. به اشتراک گذاشتن وزن‌های یک لایه علاوه بر کاهش تعداد پارامترهای تعلیمی، به‌دلیل آنکه تمامی نورون‌های آن لایه به جست‌وجو برای یافتن یک ویژگی یکسان در بخش‌های مختلف ورودی می‌پردازند، به مقاوم‌سازی شبکهٔ عصبی نسبت به برچسب‌های غیردقیق مکان آن ویژگی در قاب‌های گفتاری کمک می‌کند.

این ساختار از شبکه‌ها می‌توانند برای تعلیم اطلاعات توالی رشد داده شوند. در مواقعی که لازم است شبکهٔ عصبی اطلاعات بیش‌تری را غیر از اطلاعات آوایی قاب مرکزی از ورودی استخراج کند (مانند استخراج اطلاعات مربوط به توالی آوایی قاب‌ها یا در سطح گسترده‌تر، استخراج اطلاعات مربوط به لغت)، می‌توان ساختار آنها را در تعداد لایه‌های پنهان و تعداد نورون‌های آنها با توجه به دانش موجود از تکلیف مورد نظر رشد داد. برای دخالت‌دادن اطلاعات توالی آوایی قاب‌های گفتاری در تعلیم، ساختار این شبکه‌ها به‌گونه‌ای رشد داده می‌شود تا بتواند هم‌زمان با اطلاعات فضایی-زمانی قاب‌های گفتاری در ورودی شبکهٔ عصبی، اطلاعات فضایی-زمانی برچسب‌های آنها را در خروجی انجمن کند. ساختار شبکهٔ عصبی طراحی‌شده در این راستا که آن را ساختار انجمن‌گر فضایی-زمانی دوگانه[1] (DST) می‌خوانیم، می‌تواند زیرفضای زنجیره‌های معتبر آوایی قاب‌ها در داده تعلیمی را یاد بگیرد. بنابراین، این شبکه می‌تواند به‌صورت خودسازمانده، توالی‌های خروجی نامعتبر را پالایش کند.

مسأله تعلیم شبکه‌های عصبی عمیق پیمانه‌ای معرفی شده یک مسأله غیرمحدب و پیچیده است. به این دلیل، در این مقاله روش پیش‌تعلیم متناسب با ساختار خاص این شبکه‌ها برای قراردادن آنها در یک نقطه شروع مناسب ارائه شده است. در این روش پیش‌تعلیم، شبکهٔ عصبی به‌صورت لایه‌به‌لایه تعلیم داده می‌شود. با توجه به ساختار پیمانه‌ای و وزن‌های رونوشت‌شده در زمان این شبکه، در هر لایه فقط وزن‌های مربوط به یک پیمانه تعلیم داده شده و وزن محاسبه‌شده برای بقیه پیمانه‌ها رونوشت می‌شود. این کار به میزان زیادی از حافظه مورد نیاز و حجم محاسبات لازم می‌کاهد. به‌دلیل شباهت این ساختار از شبکه‌های عصبی به شبکه‌های کانولوشنی، روند پیش‌تعلیم معرفی‌شده می‌تواند در پیش‌تعلیم شبکه‌های کانولوشنی نیز مورد استفاده قرار بگیرد. با رشد دادن شبکه بازشناس آوای پیمانه‌ای پیشنهاد شده با قابلیت انجمن‌سازی فضایی-زمانی دوگانه به یک شبکهٔ عصبی عمیق بازشناس لغت یک‌پارچه‌ای که می‌تواند در بازشناسی گفتار پیوسته با تعداد لغات زیاد[2] مورد استفاده قرار بگیرد می‌توان دست یافت؛ اما قبل از آن، لازم است این شبکه بر روی دادگان بزرگ تنظیم شود. در تمامی کاربردهای شبکه‌های عصبی عمیق در ترکیب با مدل‌های HMM یا CRF برای بازشناسی لغت نیز لازم است شبکه ابتدا بر روی دادگان مورد استفاده تنظیم شود. از آنجا که دادگان بزرگ فاقد برچسب‌های آوایی هستند، برچسب‌های تخمینی به‌طورمعمول به کمک مدل‌های GMM-HMM که از قبل برروی این دادگان تعلیم یافته‌اند با مرزبندی تحمیلی گفتار[3] به‌دست می‌آیند (دهل و همکاران، ۲۰۱۲؛ بورلارد و مورگان، ۱۹۹۳؛ هینتون و همکاران، ۲۰۱۲؛ سیناث و همکاران، ۲۰۱۱). که این خود مشکلات تعلیم یک مدل GMM-HMM به‌صورت جداگانه بر روی دادگان بزرگ را دارد. درهمین‌اواخر پژوهش‌های زیادی در زمینه تعلیم DNNها برروی دادگان بدون برچسب قابی و عدم استفاده از مدل‌های GMM به منظور مرزبندی تحمیلی انجام شده است (سنیور و همکاران، ۲۰۱۴؛ بکیانی و همکاران، ۲۰۱۴). در این مقاله نیز یک روش تعلیم نیمه‌سرپرستی‌شده برای تنظیم شبکه عصبی عمیق با ساختار انجمن‌گر فضایی-زمانی دوگانه بر روی دادگان بزرگ معرفی می‌شود که به مدل‌های احتمالاتی ازپیش‌تعلیم‌یافته نیازی ندارد و فقط از اطلاعات موجود بهره می‌گیرد.

در ادامهٔ مقاله، در بخش دوم ساختار پیشنهادی شبکه‌های عصبی پیمانه‌ای عمیق به‌منظور بازشناسی آوا معرفی می‌شود. در بخش سوم به توضیح مراحل رشد آنها به‌منظور دست یافتن به ساختار فضایی-زمانی دوگانه پرداخته می‌شود. در بخش چهارم، روند پیش‌تعلیم و تعلیم سرتاسری مورد استفاده برای این ساختارها توضیح داده می‌شود. هم‌چنین، روش تعلیم نیمه‌سرپرستی‌شده برای تنظیم این ساختار از شبکه‌های عصبی برروی دادگان بزرگ که برچسب‌های آوایی برای آنها وجود ندارد، معرفی می‌شود. در بخش پنجم نتایج پیاده‌سازی‌های انجام شده بر روی دادگان گفتاری فارس‌دات و فارس‌دات بزرگ ارائه می‌شود. درنهایت، به بحث و نتیجه‌گیری نهایی پرداخته می‌شود.

[1] Double Spatio-Temporal Association (DST)

[2] Large Vocabulary Continuous Speech Recognition (LVCSR)

[3] Force alignment

۲- معرفی شبکه‌های عصبی پیمانه‌ای عمیق بازشناس آوا

یک DNN، یک شبکه عصبی جلوسو با بیش از دو لایه پنهان غیرخطی بین ورودی و خروجی خود است. و یک شبکه پیمانه‌ای، یک شبکه متشکل از تعدادی شبکه‌های مستقل است که توسط یک‌سری اتصالات به یکدیگر مربوط می‌شوند. هر شبکه عصبی مستقل که یک پیمانه[1] خوانده می‌شود بر روی قسمت‌های مجزایی از ورودی عمل می‌کند تا زیر تکالیفی را از تکلیفی که بر عهده شبکهٔ اصلی است، انجام دهد.

شبکهٔ عصبی پیمانه‌ای عمیق پایه طراحی‌شده در این مقاله با هدف بازشناسی آوا، در شکل (۲) نشان داده شده است. این شبکه که در ادامه پژوهش‌های انجام‌شده توسط سیدصالحی در (سیدصالحی، ۱۳۸۴) پیشنهاد شده است، همانند دیگر شبکه‌های DNN شامل L لایه غیرخطی می‌باشند. این شبکه با لغزیدن بر روی زنجیره بردارهای بازنمایی یا همان زنجیره قاب‌های ورودی، آوای شناسایی شده برای بردار بازنمایی مرکزی زنجیره ورودی در هر لحظه را در خروجی می‌دهد. ساختار این شبکهٔ عصبی براساس شبکه‌های عصبی تأخیر زمانی[2] (TDNN)(وایبل و همکاران، ۱۹۸۹) طراحی شده است. همان‌طور که در شکل (۲) مشاهده می‌شود، هر یک از آشکارسازهای ویژگی در هر لایه که توسط یک مکعب‌مستطیل نشان داده شده و شامل تعدادی نورون می‌باشند، یک میدان دریافت[3] محدود دارند. هم‌چنین، وزن‌های اتصالات مربوط به تمامی آشکارسازها در یک لایه، یکسان در نظر گرفته شده است. مجموع هر آشکارساز با ورودی‌های مربوط به آن در اینجا، نقش یک پیمانه را دارند.

طراحی ساختار این شبکهٔ عصبی به این صورت، علاوه‌بر قابلیت رشدیافتن آن، به دلایل زیر انجام شده است: براساس آنچه در (وایبل و همکاران، ۱۹۸۹؛ لنگ و همکاران، ۱۹۹۰) آمده است، سیگنال‌های تمایزی در بازشناسی آوا طول کوتاهی دارند. بنابراین، اتصالات آشکارسازهای هر لایه به بخش کوچکی از توالی نورون‌های لایهٔ پایین‌تر برای پردازش اطلاعات آن بخش برقرار می‌شوند. دوم این‌که یکسان‌بودن وزن‌های یک لایه به آشکارسازهای آن لایه کمک می‌کند تا به جست‌وجوی یک ویژگی مشترک در بخش‌های مختلف ورودی بپردازند. این ویژگی می‌تواند در هر یک از بخش‌های ورودی وجود داشته باشد و نیازی به تعریف دقیق مکان رویداد آن ویژگی وجود ندارد. هم‌چنین، رونوشت کردن وزن‌ها یکی از راه‌های کاهش دادن پارامترهای شبکه‌های DNN به‌شمار می‌رود. بر این اساس، امکان استفاده از ساختارهای بزرگ‌تر و مفیدتر بدون اینکه به زمان و محاسبات زیادی برای تعلیم آنها نیاز باشد فراهم می‌شود. به‌خصوص، در مواقعی که ساختارهای محلی مشابهی در نقاط مختلف ورودی وجود داشته باشد.

همان‌طور که در شکل (۲) نشان داده شده است، در شبکهٔ پیمانه‌ای معرفی‌شده در این مقاله هر یک از آشکارسازهای لایهٔ پنهان نخست در هر زمان فقط از بردار ویژگی گفتاری مربوط به همان زمان ورودی می‌گیرند. بنابراین میدان دریافت آنها یک نمونه در زمان است. این نگاشت از فضای ورودی به فضایی دیگر که به‌طورمعمول دارای بعد بالاتری است، باعث می‌شود هر بردار بازنمایی به‌صورت جداگانه در مرحله نخست پردازش شود. این نگاشت در پژوهش‌های مختلفی در ساختار شبکه‌های عصبی بازشناس گفتار به‌کار رفته است (نژادقلی و سیدصالحی، ۲۰۰۹؛ یزدیان، ۱۳۸۰؛ کرمی، ۱۳۷۹) و منجر به کاهش زمان یادگیری نسبت به شبکه‌های تمام متصل در لایهٔ نخست می‌شود. درصورتی‌که هر یک از نورون‌ها در دسته نورون‌های لایهٔ پنهان نخست محدود به درنظرگرفتن اطلاعات مربوط به یک باند فرکانسی شوند، قدرت یادگیری شبکهٔ عصبی بالاتر می‌رود. در این صورت، این نگاشت، با خاصیت تونوتوپیک دستگاه شنوایی انسان که به معنای حساس‌بودن مکان‌های مختلف در مغز به فرکانس‌های متفاوت سیگنال گفتار ورودی است، هماهنگ خواهد بود (چن و همکاران، ۲۰۰۵؛ شکفته و الماس‌گنج، ۱۳۸۶).

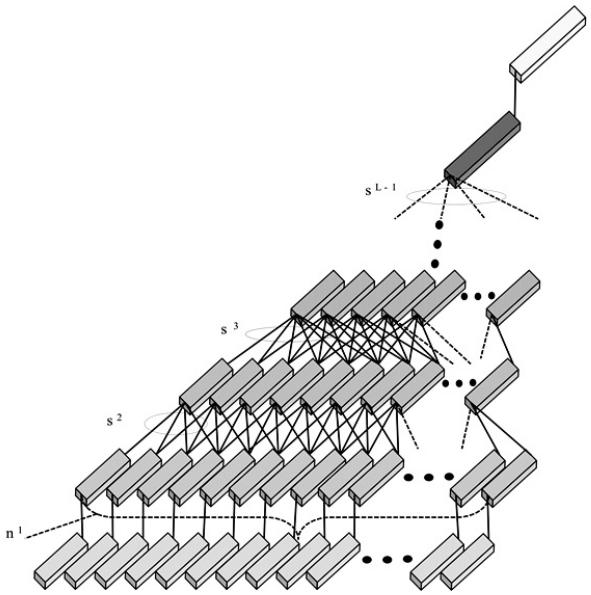
بعد از نگاشت اطلاعات ورودی به لایهٔ نخست، در لایه‌های بالاتر شبکهٔ عصبی، دینامیک‌های ورودی مورد تجزیه و تحلیل قرار می‌گیرند. هم‌چنین، در این شبکه میدان دریافت آشکارسازهای لایه‌های پایین‌تر، کوچک‌تر از میدان دریافت آشکارسازهای لایه‌های بالاتر است. چون هر چه آشکارسازها در لایه‌های بالاتری باشند، دینامیک‌های بزرگ‌تری از ورودی را می‌توانند براساس دینامیک‌های کوچکتر آشکار کنند. درنهایت، تمامی اطلاعات استخراج‌شده در لایه‌های پایینی با یکدیگر ترکیب شده و بعد از عبور از یک لایهٔ غیرخطی به خروجی می‌رسند.

[1] Module
[2] Time Delay Neural Network (TDNN)
[3] Receptive field

در این ساختار، برای به‌وجودآوردن امکان آشکارسازی ویژگی‌های مختلف از هر ورودی، هر آشکارساز دارای تعدادی نورون است.

۳- رشد ساختار شبکهٔ عصبی پیمانه‌ای عمیق به ساختار انجمن‌گر فضایی-زمانی دوگانه

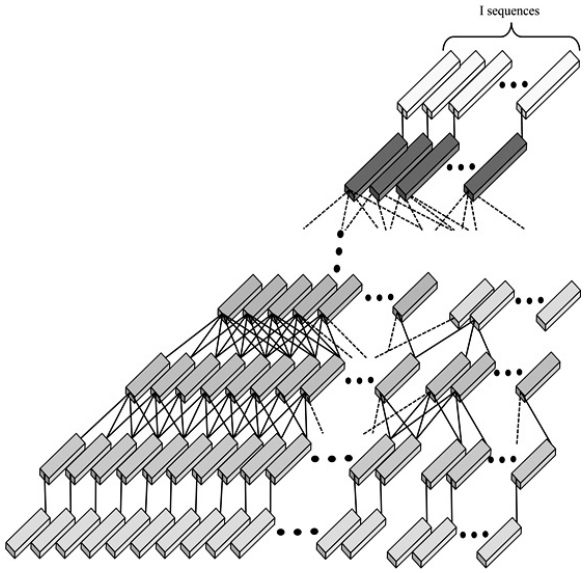
الگوریتم‌های تعلیم شبکه‌های عصبی که براساس اطلاعات تمایزی قاب‌ها طراحی شده‌اند، فقط از اطلاعات یک قاب و یا یک دسته از قاب‌ها برای بازشناسی آوای مربوط به قاب مرکزی ورودی استفاده می‌کنند. در این مقاله با توسعه ساختار شبکه عصبی پیمانه‌ای پایهٔ معرفی‌شده در بخش قبل و با تعلیم مناسب آن، سعی می‌شود اطلاعات توالی آوایی قاب‌ها به همراه اطلاعات تمایزی هر قاب در ساختار یک‌پارچه شبکه قرار داده شود. به این منظور، ساختار آن به‌صورت نشان‌داده‌شده در شکل (۳) توسعه داده می‌شود.



(شکل- ۲): ساختارپیمانه‌ای عمیق شبکه بازشناس آوای پایه. در این شکل هر آشکارساز که شامل تعدادی نورون است با یک مکعب مستطیل نشان داده شده است. هر خط در این شکل نشان دهنده اتصالات وزن‌دار توالی‌های نورونی در لایهٔ قبل به توالی‌های نورونی در لایهٔ بعد است.

نحوهٔ رشد این ساختار به این صورت است: شبکه بازشناس آوای اولیه با لغزیدن بر روی سیگنال گفتار با استفاده از اطلاعات زمینه قاب‌ها به بازشناسی قابی که در هر لحظه در مرکز ورودی شبکه قرار می‌گیرد، می‌پردازد. اگر این شبکه بخواهد در هر لحظه یک توالی از قاب‌ها را بازشناسی کند، لازم است توسعه داده شود. با این کار، شبکهٔ توسعه‌یافته، در ورودی توالی قاب‌هایی را که می‌خواهد بازشناسی کند به همراه محتوای راست و چپ آنها می‌بیند و در خروجی نیز توالی آواهای مربوط به این قاب‌ها را می‌دهد. با مقایسهٔ ساختار این شبکه با ساختار نشان‌داده‌شده در شکل (۱) مشاهده می‌شود که در این ساختار تا حدودی وظایف هر دو مدل DNN و HMM برآورده شده است.

می‌توان تعداد لایه‌های این شبکه را با توجه به سطح پیچیدگی اطلاعاتی که در نظر است استخراج شود افزایش داد.



(شکل- ۳): شبکهٔ عصبی پیمانه‌ای با ساختار انجمن‌گر فضایی-زمانی دوگانه که از توسعه شبکه شکل (۲) به‌دست آمده است.

۴- تعلیم شبکه‌های عصبی پیمانه‌ای عمیق

در ادامه نحوه تعلیم شبکه‌های پیمانه‌ای عمیق معرفی‌شده که شامل شبکه‌های پیمانه‌ای پایه بازشناس آوا و ساختار توسعه‌یافتهٔ آنها به‌صورت ساختار انجمن‌گر فضایی-زمانی دوگانه می‌باشد، شرح داده می‌شود.

۴-۱- پیش‌تعلیم

روند پیش‌تعلیم معرفی‌شده همان‌طور که در شکل (۴) نشان داده شده است، به‌صورت لایه‌به‌لایه انجام می‌شود. هدف از به‌کارگرفتن این روند پیش‌تعلیم در این ساختار از شبکه‌های

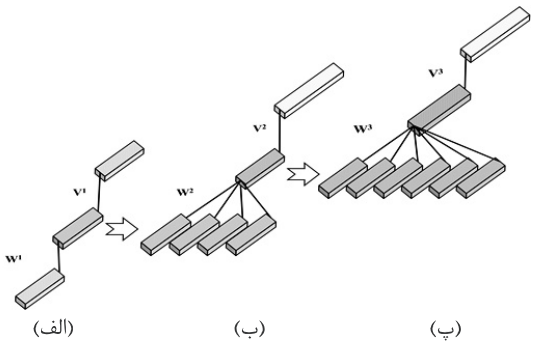
عصبی این است که بتوان به‌صورت بهینه از اطلاعات دادگان تعلیمی در هر مرحله به‌منظور حفظ حداکثر اطلاعات لازم برای بازشناسی استفاده کرد؛ لذا، برای تعلیم هر لایه، وزن‌های آن لایه به‌گونه‌ای تنظیم می‌شوند که گویی بازشناسی نهایی از روی خروجی آن به‌دست می‌آید. بنابراین، به‌صورت مرحله به مرحله وزن‌های مربوط به هر لایه با توجه به هدف بازشناسی مورد نظر به‌صورت مناسبی تعلیم می‌یابند. از وزن‌های به‌دست آمده در هر مرحله برای محاسبه ورودی لایهٔ بعدی استفاده می‌شود. وزن‌ها باید به‌گونه‌ای محاسبه شوند تا حداکثر اطلاعات تمایزی ورودی هر لایه را در جهت دست‌یافتن به خروجی مورد نظر استخراج کنند؛ اما در مورد پیش‌تعلیم وزن‌های لایهٔ نخست شبکه‌های عصبی معرفی‌شده باید گفت چون اطلاعات ورودی، بردارهای بازنمایی‌ای است که ما استخراج کرده‌ایم، ابتدا لازم است این اطلاعات توسط شبکه تجزیه و تحلیل شود؛ سپس، از اطلاعات به‌دست آمده توسط شبکه در این مرحله در جهت استخراج دانش لازم برای بازشناسی در مراحل بعد استفاده شود. بنابراین، وزن‌های مربوط به این لایه باید به‌گونه‌ای تعلیم داده شوند که در تجزیه و تحلیل اطلاعات، اتلاف نداشته باشند. برای رسیدن به این هدف، لازم است ورودی‌های متمایز به‌صورت متمایز بیان شوند.

برای استفاده از روش پیش‌تعلیم لایه‌به‌لایه با توجه به ساختار شبکهٔ عصبی بازشناس آوای قابل رشد معرفی‌شده با وزن‌های رونوشت شده در زمان و میدان دریافت محدود برای نورون‌های هر لایه، لازم است روند ویژه‌ای به‌کار گرفته شود. نحوهٔ به‌کارگیری روش پیش‌تعلیم مورد نظر برای این ساختار از شبکه می‌تواند ایده‌ای برای چگونگی به‌کارگیری روش‌های پیش‌تعلیم مختلف برای شبکه‌های کانولوشنی باشد. در ادامه، نحوه پیاده‌سازی روش پیش‌تعلیم مورد استفاده آورده شده است.

۴-۱-۱- پیش‌تعلیم وزن‌های مربوط به لایهٔ نخست

براساس آنچه در بخش پیش ارائه شد، وزن‌های لایهٔ نخست ($\mathbf{W}^1$) نقش تحلیل اطلاعات ورودی را بر عهده داشته و نقشی در استخراج دینامیک‌های ورودی ندارند.

بنابراین، لازم است وزن‌ها طوری تعلیم داده شوند تا بتوانند تمامی ورودی‌های متمایز را به‌صورت متمایز بیان کنند. به بیان دیگر، براساس آنچه که در (سیدصالحی و سیدصالحی، ۱۳۹۲) ارائه شده است، باید وزن‌ها طوری تعلیم داده شوند که بتوانند ورودی را بدون اتلاف بازسازی کنند. بنابراین، از یک ساختار شبکهٔ عصبی خودانجمنی به‌صورت نشان داده شده در شکل (۴-الف) استفاده می‌شود. در این شبکه، نگاشت ورودی به لایه پنهان توسط ماتریس وزن‌های $\mathbf{W}^1$ و نگاشت لایهٔ پنهان به خروجی توسط وزن‌های کمکی $\mathbf{V}^1$ انجام می‌پذیرد. بعد از تعلیم این شبکه، لایهٔ پنهان آن می‌تواند مؤلفه‌های فضای بازنمایی‌های ورودی را یاد بگیرد.



(شکل- ۴): روند پیش‌تعلیم شبکه عصبی با ساختار انجمن‌گر فضایی-زمانی دوگانه

در راستای محاسبه ماتریس وزن مربوط به هر یک از شبکه‌های یک‌لایه پنهان لازم است یک تابع هزینه تعریف شود. در اینجا تابع حداقل مربعات خطا که به‌صورت رابطه (۱) تعریف می‌شود، به‌عنوان تابع هزینه به‌کار می‌رود. و با استفاده از الگوریتم دلتا مقدار این تابع بهینه می‌شود.

$$E(\mathbf{W}^l) = \frac{1}{2}\sum_{j=1}^{J^l}\left(d_j^l - z_j^l\right)^2 \qquad (۱)$$

در رابطهٔ بالا l شمارهٔ لایه‌ای است که وزن مربوط به آن در هر مرحله بهینه می‌شود. برای محاسبه $\mathbf{W}^1$، J^1 در رابطهٔ بالا تعداد نورون‌های لایهٔ ورودی و لایهٔ خروجی شبکه خودانجمنی تک‌لایه‌ها پنهان است که برابر با تعداد پارامترهای هر بردار بازنمایی یا هر قاب است. بردار $\mathbf{d}^1$ یا خروجی هدف در اینجا بردار بازنمایی ورودی شبکه است و $\mathbf{z}^1$ بردار بازسازی‌شده‌ای است که شبکه برای این بردار بازنمایی ساخته است.

مقدار $\mathbf{W}^1$ محاسبه‌شده در این حالت، برای تمامی اتصالات مشابه در این لایه که همان نگاشت هر یک از بردارهای بازنمایی ورودی به فضای لایهٔ پنهان نخست در ساختار شکل (۲) است، رونوشت می‌شود.

۴-۱-۲- پیش‌تعلیم وزن‌های مربوط به سایر لایه‌ها

برای پیش‌تعلیم وزن‌های لایه‌های پنهان بالاتر، همان‌طور که اشاره شد در هر مرحله شبکه‌های یک‌لایهٔ پنهانی ساخته

می‌شود. ورودی این شبکه‌ها به کمک وزن‌های محاسبه‌شده در مراحل قبلی به‌دست می‌آید و خروجی آنها لایهٔ بازشناسی برای بازشناسی بردار بازنمایی مرکزی ورودی آن شبکه در هر زمان است. در این ساختار از شبکه‌ها، خروجی هر $\boldsymbol{s^l}$ آشکارساز در لایهٔ پنهان l-1 ($\boldsymbol{s^l}$) اندازه میدان دریافت آشکارسازهای لایه l یا همان تعداد تأخیرهای زمانی در هر پیمانه است) با یکدیگر ترکیب می‌شوند و بعد از اعمال ماتریس وزن های $\mathbf{W}^l$و $\mathbf{V}^l$ و دو تابع غیرخطی در لایهٔ پنهان و لایهٔ خروجی، خروجی نهایی آنها محاسبه می‌شود. به‌عنوان مثال، در شکل‌های (۴-ب) و (۴-پ) ساختار شبکه‌های یک لایه پنهان مربوط به تعلیم وزن‌های لایه‌های پنهان دوم و سوم شبکه پیمانه‌ای پیشنهادی نشان داده شده است. بنابراین، با محاسبهٔ وزن‌های مربوط به خروجی $\boldsymbol{s^l}$ توالی آشکارساز در لایهٔ پنهانl-1ام ($\mathbf{W}^l(l>1)$) کافی است این وزن برای اتصالات مشابه رونوشت شود. در این‌صورت، خروجی مطلوب این شبکه، کد آوای مربوط به آشکارساز مرکزی دسته $\boldsymbol{s^l}$ تایی ورودی است که از روی برچسب آوایی دادگان به‌دست می‌آید. ماتریس وزن‌های $\mathbf{W}^l$ نیز براساس تابع هزینه حداقل مربعات خطا در رابطه (۱) و با استفاده از الگوریتم دلتا تعلیم داده می‌شوند. در این رابطه، J^l تعداد نورون‌های لایهٔ بازشناسی خروجی شبکه‌های تک‌لایهٔ پنهان است که برابر با تعداد آواها است. و بردار $\mathbf{d^l}$ برچسب آوایی مربوط به بردار بازنمایی مرکزی ورودی شبکه است که به‌ازای آوای مربوطه مقدار یک و به‌ازای بقیه آواها مقدار صفر دارد. و $\mathbf{z^l}$ خروجی‌ای است که شبکه برای این بردار بازنمایی مرکزی شناسایی کرده است که به‌صورت زیر محاسبه می‌شود.

$$\mathbf{z^l} = f\left(f\left(\left(\sum_{m=1}^{s^l}\mathbf{H}_m^{l-1}\mathbf{W}_m^l\right)-\mathbf{b^l}\right)\mathbf{V}^l-\mathbf{c}^l\right) \quad (۲)$$

در رابطهٔ بالا $\mathbf{b^l}$و $\mathbf{c}^l$ به‌ترتیب، بردار بایاس مربوط به هر آشکارساز در لایهٔ l و بردار بایاس برای لایهٔ خروجی هستند. برای سادگی نمایش محاسبات، در روابط بعدی، بردارهای بایاس به سطر آخر ماتریس وزن‌های مربوطه اضافه شده‌اند. در این مقاله، توابع غیرخطی اعمال‌شده به تمامی لایه‌های پنهان و لایهٔ خروجی ($f(.)$) تابع سیگموید در نظر گرفته شده است که به‌صورت زیر تعریف می‌شود. این تابع مشهورترین تابع فعالیت است که توسط متخصصان حوزه گفتار مورد استفاده قرار می‌گیرد (یو و دنگ، ۲۰۱۴).

$$f(\theta)=\frac{1}{1+e^{-\theta}} \quad (۳)$$

و میدان دریافت مربوط به تمامی لایه‌ها (s^l) عدد فرد انتخاب شده است. با توجه به اینکه n^{l-1} تعداد آشکارسازها در لایهٔ l-1 است، $\mathbf{H}^{l-1}$ یک ماتریس n^{l-1} سطری از مقادیر خروجی آشکارسازها در لایه l-1 است. و $\mathbf{H}_m^{l-1}$ که m امین سطر از $\mathbf{H}^{l-1}$ را نشان می‌دهد، خروجی نورون‌های mامین آشکارساز در لایهٔl-1 است. بعد از محاسبه $\mathbf{W}^l$ که وزن ترکیب s^l آشکارساز در لایهٔ قبلی برای تولید خروجی آشکارسازها در لایهٔ l است، مقدار این وزن برای تمامی اتصالات مشابه رونوشت می‌شود.

۴-۲- تعلیم یک‌پارچه

بعد از پیش‌تعلیم شبکهٔ عصبی عمیق پیمانه‌ای که برای هر دو حالت پایه و توسعه‌یافته با ساختار انجمن‌گر فضایی-زمانی دوگانه به‌صورت اشاره شده در زیر بخش (۴-۱) قابل اجراست، وزن‌های محاسبه‌شده در ساختار اصلی شبکه قرار داده می‌شود؛ سپس، این شبکه به‌صورت یک‌پارچه و سرتاسری برای تنظیم دقیق وزن‌ها تعلیم داده می‌شود. به این منظور، از الگوریتم کاهش گرادیان برای بهینه‌سازی تابع هزینهٔ حداقل مربعات خطای محاسبه‌شده به‌صورت زیر استفاده می‌شود.

$$E(\mathcal{W})=\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(d_{ij}-h_{ij}^L\right)^2 \quad (۴)$$

در این رابطه $\mathcal{W}$ تمامی وزن‌های مربوط به تمامی لایه‌ها در شبکهٔ پیمانه‌ای عمیق را نشان می‌دهد. I تعداد دسته نورون‌های خروجی است. هر دسته نورون، به تعداد آواها نورون دارد و به بازشناسی قاب مرکزی در ورودی خود می‌پردازد. این تعداد در شبکه پیمانه‌ای پایه برابر یک و در شبکه با ساختار انجمن‌گر فضایی-زمانی دوگانه برابر با تعداد قاب‌هایی است که به‌منظور تعلیم اطلاعات توالی آوایی آن قاب‌ها به شبکه در نظر گرفته شده است. h_{ij}^L خروجی نورون jام از iامین دسته از نورون‌های لایه آخر است و d_{ij} مقدار هدف این نورون را نشان می‌دهد.

جهت تعلیم سرتاسری وزن‌ها با استفاده از الگوریتم کاهش گرادیان، تنظیم وزن‌ها باید در جهت عکس گرادیان خطا طبق رابطه زیر و به‌صورت تکراری انجام بگیرد.

$$\Delta\mathbf{W}^l=-\gamma\frac{\partial E}{\partial\mathbf{W}^l} \quad (۵)$$

در این رابطه γ ضریب یادگیری است. در الگوریتم پس‌انتشار خطای استاندارد، گرادیان خطا نسبت به وزن هر لایه طبق رابطهٔ (۶) محاسبه می‌شود.

$$\frac{\partial E}{\partial \mathbf{W}^l} = \frac{\sum_{m=1}^{n^l}\left[\mathbf{H}_m^{l-1}, \mathbf{H}_{m+1}^{l-1}, \dots, \mathbf{H}_{m+s^l}^{l-1}, 1\right]' \mathbf{Er}_m^l}{n^l} \quad (۶)$$

در مورد لایهٔ نخست، $\mathbf{H}^0$ همان توالی بردارهای بازنمایی ورودی ($\mathbf{X}$) و در مورد لایه آخر، $\mathbf{H}^L$ برابر خروجی I دسته از نورون‌های خروجی است.

در این رابطه، $\left[\mathbf{H}_m^{l-1}, \mathbf{H}_{m+1}^{l-1}, \dots, \mathbf{H}_{m+s^l}^{l-1}, 1\right]$ بردارهای تشکیل شده از قراردادن خروجی‌های مربوط به s^l آشکارساز در لایهٔ *l-1* و نورون بایاس است که از طریق $\mathbf{W}^l$ به هر یک از آشکارسازهای نورونی در لایهٔ l متصل هستند. چون وزن $\mathbf{W}^l$ برای تمامی اتصالات مشابه یکسان است، از گرادیان خطا نسبت به آن میانگین‌گیری می‌شود. با توجه به اینکه تابع فعالیت مورد استفاده برای تمامی لایه‌ها تابع سیگموید است، $\mathbf{Er}_m^L$ یا خطای هر یک از نورون‌های مربوط به دسته mام در لایه خروجی L به‌صورت نشان داده شده در رابطهٔ (۷) محاسبه می‌شود.

$$\mathbf{Er}_m^L = \mathbf{H}_m^L.(\mathbf{1} - \mathbf{H}_m^L).(\mathbf{D}_m - \mathbf{H}_m^L) \quad (۷)$$

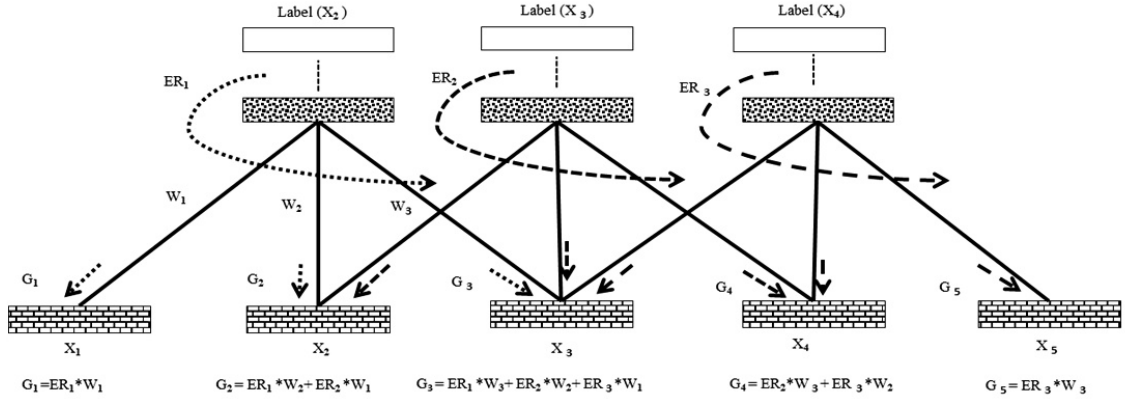
در این رابطه، (.) ضرب هر درایه از دو بردار با طول برابر را نشان می‌دهد. و $\mathbf{1}$ بیان‌کنندهٔ بردار با درایه‌های یک است. بر این اساس، خطای پس انتشار شده به لایه‌های پایین‌تر به‌صورت زیر به‌دست می‌آید.

$$\mathbf{Er}_m^l = \mathbf{H}_m^l.(\mathbf{1} - \mathbf{H}_m^l).\mathbf{G}_m^l \quad l < L \quad (۸)$$

چون در این ساختار از شبکهٔ عصبی، ورودی‌های آشکارسازهای لایهٔ پنهان بالاتر با یکدیگر هم‌پوشانی دارند، بایستی خطاهای پس‌انتشارشده به هر آشکارساز در لایهٔ پنهان پایین‌تر با یکدیگر جمع شوند. در این حالت، $\mathbf{G}_m^l$ طبق رابطهٔ زیر محاسبه می‌شود:

$$\mathbf{G}_m^l = \sum_{p=1}^{s^{l+1}} \sum_{q=1}^{n^{l+1}} \left(\mathbf{Er}_q^{l+1}(\mathbf{W}_p^{l+1})'\right) \times \delta((m+1)-(p+q)) \quad (۹)$$

خطای هر آشکارساز q در لایه بالاتر ($l+1$)، به تعداد مشخصی واحد آشکارساز در لایهٔ پایین‌تر (l) پس‌انتشار می‌شود. که این تعداد همان میدان دریافت آشکارسازهای لایه $l+1$ (s^{l+1}) است. $\mathbf{W}_p^{l+1}$ در این رابطه، وزن اتصالات مربوط به pامین آشکارساز در لایه l به هر آشکارساز در لایهٔ $l+1$ است. چون وزن‌ها برای تمامی آشکارسازهای لایه $l+1$ یکسان است، $\mathbf{W}_p^{l+1}$ به q وابسته نیست. $\delta(.)$ در رابطهٔ بالا، تابع دیریکله است که وقتی مقدار $p+q-1$ برابر با m می‌شود یک می‌دهد و در غیر این‌صورت خروجی آن صفر است. در شکل (۵) به‌صورت نمایشی طریقه پس‌انتشار خطا از لایهٔ خروجی به لایهٔ قبل از آن رسم شده است. در این شکل میدان دریافت هر آشکارساز در لایهٔ خروجی برابر ۳ است. همان‌طور که در این شکل مشاهده می‌شود، هم‌پوشانی ورودی‌های آشکارسازهای لایهٔ بالاتر باعث می‌شود خطای مربوط به خروجی چند آشکارساز به ورودی هر آشکارساز منعکس شود. این حالت، در یادگیری اطلاعات توالی خروجی توسط هریک از وزن‌ها مؤثر است.



(شکل- ۵): طریقه پس انتشار خطا در شبکه پیمانه‌ای با ساختار کپی شده در زمان و میدان دریافت محدود برای آشکارسازهای هر لایه

در حالت کلی، روند تعلیم شبکه با ساختار فضایی-زمانی دوگانه را می‌توان به‌صورت پس انتشار خطای رابطه (۱۰) در نظر گرفت. در حالی که برای تعلیم در سطح قاب، خطای پس‌انتشار شده به‌صورت $(\{\mathbf{d}-\mathbf{h}^L\}|\mathbf{X}(t))$ می‌باشد.

$$\left(\{\mathbf{D}_1 - \mathbf{H}_1^{\mathrm{L}}, \dots, \mathbf{D}_{\mathrm{I}} - \mathbf{H}_{\mathrm{I}}^{\mathrm{L}}\} | \mathbf{X}(\mathrm{t}: \mathrm{t}+\mathrm{I})\right) \quad (۱۰)$$

۴-۳- تعلیم نیمه‌سرپرستی‌شده شبکه با ساختار انجمن‌گر فضایی-زمانی دوگانه

در بخش مقدمه اشاره شد که برای استفاده از شبکهٔ عصبی تعلیم‌یافته با ساختار انجمن‌گر فضایی-زمانی دوگانه (DST) در سامانهٔ بازشناسی گفتار با تعداد لغات زیاد، لازم است این شبکه بر روی دادگان بزرگ تنظیم شود. بیشتر دادگان بزرگ توسط مجموعه لغاتی که شامل حاشیه‌نویسی هر لغت به‌صورت واجی[1] هستند، ارائه می‌شوند و فاقد برچسب‌های آوایی قاب‌ها هستند (ون بائل و همکاران، ۲۰۰۷)؛ لذا، در اینجا یک الگوریتم برای تعلیم نیمه‌سرپرستی‌شده این شبکه بر روی دادگان بزرگ معرفی می‌شود. این روش به مرزبندی تحمیلی گفتار توسط مدل‌های GMM-HMM از پیش تعلیم یافته بر روی این دادگان نیاز ندارد.

در شکل (۶) الگوریتم تعلیم نیمه‌سرپرستی‌شده پیشنهادی آورده شده است. در پیاده‌سازی این الگوریتم فرض می‌شود دادگان کوچک‌تری که دارای مشخصات آوایی مشابهی با دادگان بزرگ هستند و شامل برچسب‌های آوایی قاب‌های گفتاری نیز می‌باشند، وجود دارد؛ لذا ابتدا شبکهٔ عصبی با ساختار DST بر روی دادگان کوچک‌تر به‌خوبی تعلیم داده می‌شود؛ سپس، از شبکهٔ تعلیم‌یافته برروی دادگان کوچک‌تر برای بازشناسی بردارهای بازنمایی بدون برچسب دادگان بزرگ و استخراج اطلاعات آوایی برای هر قاب استفاده می‌شود؛ لذا شبکهٔ عصبی با ساختار DST تعلیم‌یافته بر روی دادگان کوچک در طول هر لغت از دادگان بزرگ که زمان آغاز و پایان آن مشخص است، لغزانده می‌شود. آواهای بازشناسی‌شده توسط این شبکه برای هر قاب از iامین لغت ورودی در آرایه $\boldsymbol{\beta}_{u_i}$ قرار داده می‌شود؛ اما این شبکه اشتباهاتی نیز در بازشناسی دارد. بنابراین، استفاده از دانش شبکهٔ عصبی تعلیم‌یافته بر روی دادگان کوچک‌تر به‌تنهایی کافی نیست و لازم است از اطلاعات دیگری در کنار این اطلاعات استفاده شود. از سوی دیگر، می‌توان از حاشیه‌نویسی واجی هر لغت برای استخراج برچسب آوایی هر قاب استفاده کرد. در این راستا، می‌توان از روی برچسب‌های آوایی دادگان کوچک‌تر، طول متوسط هر یک ازآواها ($|ph_m|$) را از تقسیم تعداد قاب‌های اختصاص‌داده‌شده به آن آوا به تعداد کل قاب‌های این دادگان به‌دست آورد؛ سپس، براساس زمان شروع و پایان هر لغت در دادگان بزرگ و هم‌چنین براساس حاشیه‌نویسی واجی این لغت، تعداد قاب‌هایی که به هر آوا می‌توان اختصاص داد، محاسبه می‌شود. در این حالت، فرض شده است که آواهای بیان‌شده همان واج‌های نوشته‌شده هستند. به این ترتیب، یک برچسب‌دهی آوایی دیگر به‌ازای هر قاب گفتاری از این دادگان به‌دست می‌آید. در این روش برچسب‌دهی، اگر لغت iام به‌صورت دنبالهٔ واجی $\{ph_1, ph_2, \dots, ph_{\bar{l}_{ui}}\}$ که $\bar{l}_{u_i}$ تعداد واج‌های مربوط به این لغت است، بیان شده باشد و طول لغت $|u_i|$ باشد، دنبالهٔ برچسب آوایی آن به‌صورت رابطه (۱۱) به‌دست می‌آید. در این دنباله هر آوا به اندازه $\bar{n}_{ph_m}$ قاب تکرار می‌شود.

$$\boldsymbol{\alpha}_{u_i} = \{ph_1, \dots, ph_1, ph_2, \dots, ph_2, \dots, ph_{\bar{l}_{ui}}, \dots, ph_{\bar{l}_{ui}}\}$$
$$\bar{n}_{ph_m} = \frac{|ph_m|}{\sum_{j=1}^{\bar{l}_{ui}} |ph_j|} \times |u_i| \quad (۱۱)$$

همان‌طور که می‌دانیم، محل گذر از یک آوا به آوای دیگر در توالی برچسب‌های آوایی تخمین‌زده‌شده برای هر قاب براساس طول متوسط آواها نمی‌تواند خیلی دقیق باشد. بنابراین، لازم است به‌صورت بهینه از اطلاعات توالی برچسب‌های به‌دست آمده از هر دو روش نتیجهٔ بازشناسی شبکهٔ عصبی و طول متوسط آوا استفاده کرد.

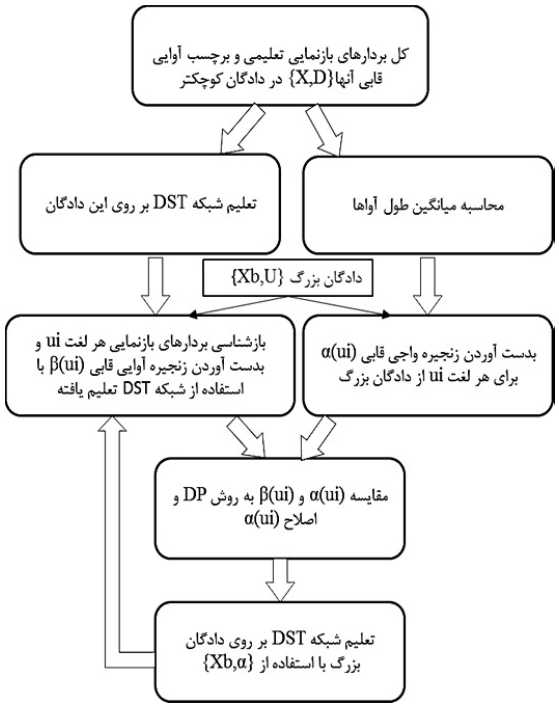
برچسب‌دهی براساس طول متوسط آواها (آرایه $\boldsymbol{\alpha}$)، تا حدودی آواهای درست یک لغت را در بر دارد و برچسب‌های به‌دست‌آمده از نتایج بازشناسی شبکه DST (آرایه $\boldsymbol{\beta}$)، اطلاعات مناسب‌تری را در مورد نقطه گذر بین آواها به‌دست می‌دهد. بنابراین، برای اشتراک اطلاعات موجود و به‌دست‌آوردن برچسب‌های آوایی بهینه برای هر لغت، برچسب‌های به‌دست‌آمده در آرایهٔ $\boldsymbol{\alpha}$ با استفاده از برنامه‌ریزی پویا[2] با برچسب‌های به‌دست‌آمده در آرایهٔ $\boldsymbol{\beta}$ مقایسه می‌شوند و بهترین انطباق این دو آرایه به‌دست می‌آید.

نتیجهٔ پیاده‌سازی برنامه‌ریزی پویا نشان می‌دهد که برچسب‌های موجود در آرایهٔ $\boldsymbol{\alpha}$ در چه زمان طول تخمین‌زده‌شده برای هر آوا را کمتر یا بیشتر از طول آوای اداشده پیش‌بینی کرده‌اند. بر این اساس، برچسب‌گذاری اولیه تصحیح شده و به‌عنوان برچسب‌های جدید برای لغات ذخیره می‌شوند. از برچسب‌های به‌دست آمده در این مرحله جهت تعلیم شبکهٔ عصبی با ساختار DST برروی دادگان بزرگ استفاده می‌شود. بعد از چند تکرار از تعلیم شبکه براساس برچسب‌های به‌دست آمده، دوباره از شبکهٔ

[1] Phonemic transcription

[2] Dynamic Programming (DP)

تعلیم‌یافته تا این مرحله برای بازشناسی بردارهای بازنمایی دادگان بزرگ و به‌دست‌آوردن آرایه برچسب آوایی جدید استفاده می‌شود. این روند تا هم‌گرایی تعلیم شبکه ادامه پیدا می‌کند.



(شکل- ۶): الگوریتم تعلیم نیمه‌سرپرستی‌شده برای شبکهٔ عصبی با ساختار DST بر روی دادگان بزرگ. Xb در این شکل مجموعه بردارهای بازنمایی تعلیمی از دادگان بزرگ و U مجموعه لغات تعلیمی است.

۵- پیاده‌سازی و نتایج

جهت ارزیابی عملکرد شبکه‌های عصبی پیمانه‌ای عمیق قابل رشد، از دادگان گفتاری فارسی فارس‌دات و فارس‌دات بزرگ (بیجن خان و همکاران، ۱۹۹۴؛ شیخ‌زادگان و بیجن خان، ۱۳۸۵) استفاده شد.

در این مقاله از بردارهای بازنمایی LHCB[1] که لگاریتم انرژی بانک فیلترهای هنینگ با باند بحرانی قرار داده شده بر روی تبدیل فوریه سیگنال در مقیاس بارک هستند؛ استفاده می‌شود. این بازنمایی‌ها برای تعلیم به شبکه‌های TDNN در بسیاری از منابع مناسب تشخیص داده شده‌اند (سیدصالحی، ۱۳۷۴؛ رحیمی نژاد، ۱۳۸۲؛ انصاری، ۱۳۸۲؛ نژادقلی و سیدصالحی، ۲۰۰۹). این بازنمایی‌ها برای شبکه‌های عصبی به‌کار رفته در بازشناسی گفتار در مقایسه با بردارهای بازنمایی MFCC که به‌طورمعمول برای بازشناسی گفتار استفاده می‌شوند، مناسب‌تر هستند.

طبقه‌بندی‌کننده‌های آماری مثل HMMها به‌منظور آنکه بتوانند بردارهای ویژگی گفتاری که المان‌های آنها به‌واسطه خصوصیات طیفی گفتار و هم‌چنین همپوشانی فیلتربانک‌ها با یکدیگر همپوشانی دارند را به دقت مدل کنند، لازم است از ماتریس‌های کوواریانس کامل که هم از لحاظ محاسباتی حجیم هستند و هم به میزان داده تعلیم زیادی نیاز دارند استفاده کنند. به این دلیل، با اعمال تبدیل کسینوسی گسسته[2] (DCT) در مرحله آخر از استخراج بازنمایی‌های MFCC المان‌های این بردارهای ویژگی غیرمرتبط شده و برای طبقه‌بندی کننده‌های آماری با ماتریس کوواریانس قطری مناسب می‌شود (تن و لیندبرگ، ۲۰۰۸). ولی برای شبکه‌های عصبی این تبدیل قسمتی از اطلاعاتی که مربوط به توالی بردارهای بازنمایی است را حذف می‌کند.

هر قاب از بردارهای بازنمایی LHCB استفاده شده در این مقاله شامل ۱۸ پارامتر می‌باشد که تقریباً از هر ۲۳ میلی‌ثانیه از سیگنال گفتار با گام پیش‌روی نصف طول این پنجره استخراج شده است. این بازنمایی‌ها به کمک هنجارسازی نرم ۱ طوری هنجار شده‌اند که میانگین هر مؤلفه از آنها روی کل دادگان تعلیمی صفر و نرم ۱ آنها که میانگین قدر مطلق بردارهای بازنمایی متمرکز شده است برابر یک شود.

لازم به ذکر است در تمامی نتایج بازشناسی آوای گزارش شده در این مقاله، از الگوریتم ویتربی برای رمزگشایی و محاسبه دقت بازشناسی آوا مشابه روند اشاره شده در (پینتو و همکاران، ۲۰۱۱) استفاده شده است. در این روش، فرض می‌شود که هر آوا توسط یک HMM چند حالته مدل شده است. احتمال پسین حالت‌های هر مدل آوایی با یکدیگر برابر هستند. این احتمالات در هر زمان، از خروجی شبکه عصبی بازشناس آوا به ازای بردار بازنمایی مرکزی ورودی در آن زمان محاسبه می‌شود. با مقیاس کردن این احتمالات، مقدار درست‌نمایی انتشاری[3] هر حالت به منظور استفاده در الگوریتم ویتربی به‌دست می‌آید. پارامترهای انتقال بین این آواها نیز از مدل بایگرم آوایی[4] که در مورد دادگان فارس‌دات و فارس‌دات بزرگ به ترتیب از روی برچسب‌های آوایی کل دادگان فارس‌دات و برچسب‌های واجی کل دادگان فارس‌دات بزرگ و با استفاده از جعبه ابزار SRILM محاسبه می‌شود، به‌دست می‌آید. با توجه به اینکه در تعلیم شبکه عصبی با ساختار انجمن‌گر فضایی-زمانی

[1] Logarithm of Hanning Critical Band filter banks (LHCB)

[2] Discrete Cosine Transform (DCT)

[3] Emission likelihood

[4] Phonetic bigram

دوگانه اطلاعات توالی خروجی به‌کار گرفته شد، انتظار می‌رود وزن‌های شبکه در حین تعلیم اطلاعاتی در مورد توالی آواها به‌دست آورده باشند. بنابراین، برای بررسی اینکه به چه میزان شبکه توانسته است اطلاعات توالی آوایی را استخراج کند، ابتدا در الگوریتم ویتربی مورد استفاده برای گذر بین آواهای مختلف مقدار احتمال صفر در نظر گرفته شد. که معادل با عدم استفاده از احتمالات مدل زبانی بایگرم آوایی در سامانه بازشناسی گفتار است. سپس، جهت مقایسه بهترین نتایج بازشناسی این مدل در مقایسه با مدل شبکه عصبی پیمانه‌ای پایه و با مدل HMM، مدل زبانی بایگرم آوایی از روی دادگان تخمین زده و در محاسبات ویتربی دخالت داده شد. نتیجه کدگشایی توسط الگوریتم ویتربی با اعمال پارامتر جریمه درج بهبود می‌یابد. این پارامتر می‌تواند از تغییر سریع خروجی الگوریتم بین آواهای مختلف جلوگیری کند. این پارامتر به‌صورت زیر با اصلاح پارامترهای گذر بین آواها در الگوریتم ویتربی مورد استفاده قرار می‌گیرد.

$$\tilde{a}_{rt} = GSF * a_{rt} + IP \qquad (۱۲)$$

در این رابطه، a_{rt} پارامتر گذر بین آوای r به آوای t است که با استفاده از مدل بایگرم آوایی جایگذاری می‌شود. GSF ضریب تاثیر مدل بایگرم آوایی در الگوریتم ویتربی است که برای عدم استفاده از این مدل، مقدار این پارامتر صفر می‌شود. و IP پارامتر جریمه درج است.

در ادامه به توضیح هر یک از دادگان و ارائه نتایج آزمایش‌های انجام شده با استفاده از هر یک از آنها پرداخته می‌شود. درصد دقت بازشناسی آوای نهایی طبق رابطه زیر و با استفاده از معیار NIST محاسبه می‌شود:

$$Phone\ Recognition\ Rate = \frac{K - Del - Ins - Sub}{K} \qquad (۱۳)$$

در این رابطه، K تعداد کل آواهای بردار مرجع، Del تعداد آواهای حذف‌شده، Ins تعداد آواهای درج‌شده و Sub تعداد آواهای جانشین شده است.

۵-۱- دادگان گفتاری فارس‌دات

این مجموعه شامل سیگنال گفتاری پیوسته و تمیز از ۳۰۴ گویندهٔ زن و مرد است که در سن، لهجه و میزان تحصیلات با یکدیگر متفاوتند (بیجن خان و همکاران، ۱۹۹۴). این دادگان به‌طورتقریبی مشابه با دادگان TIMIT در زبان انگلیسی است. از این میزان داده، گفتار مربوط به ۲۹۷ نفر برای تعلیم شبکه و گفتار مربوط به ۷ نفر دیگر برای آزمون انتخاب شد. انتخاب گویندگان مربوط به هر دو دسته به‌صورت تصادفی انجام شده است. هر کدام از ۳۰۴ گوینده تعداد ۲۰ جمله را در دو جلسه بیان کرده‌اند. بنابراین، در مجموع، ۵۹۴۰ جمله برای تعلیم شبکه و ۱۴۰ جمله برای آزمون شبکه مورد استفاده قرار گرفته شد. گفتارهای ضبط‌شده به‌صورت دستی برای هر لغت و آوا جداسازی و برچسب‌دهی شده‌اند.

۵-۱-۱- پیاده‌سازی ساختار شبکهٔ عصبی پیمانه‌ای و رشد آن به‌صورت ساختار DST برروی این دادگان

ابتدا ساختار شبکهٔ عصبی پیمانه‌ای عمیق بازشناس آوای پایه نشان داده شده در شکل (۲) با خصوصیات آورده‌شده در جدول (۱) طراحی شد.

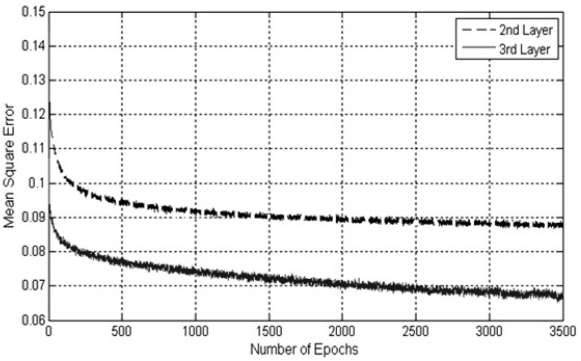
(جدول-۱):خصوصیات ساختاری شبکهٔ عصبی پیمانه‌ای پایه

لایه	تعداد آشکارسازها	تعداد گره‌ها در هر آشکارساز	تعداد تأخیر زمانی هر پیمانه
ورودی	۲۳	۱۸	-
لایهٔ نخست	۲۳	۶۴	۱
لایهٔ دوم	۱۵	۵۱۲	۹
لایهٔ سوم	۱	۶۲	۱۵
لایهٔ خروجی	۱	۳۶	۱

طراحی این ساختار با توجه به تجربیات به‌دست‌آمده از پژوهش‌های پیشین برروی دادگان فارس‌دات انجام شده است. براساس این پژوهش‌ها، در (کرمی، ۱۳۷۹) نشان داده شده است که برای فراهم‌آوردن اطلاعاتی در مورد محتوای صوتی هر قاب برای یک شبکهٔ بازشناس آوا، یک پنجرهٔ متشکل از ۷-۲۵ قاب باید به‌عنوان ورودی آن درنظر گرفته شود. هم‌چنین، با مقایسه نتایج بازشناسی آوای به‌دست‌آمده در (انصاری، ۱۳۸۲) با نتایج (نژادقلی و سیدصالحی، ۲۰۰۹) می‌توان به این نتیجه رسید که یک شبکهٔ TDNN که لایهٔ پنهان نخست آن به‌صورت نیمه‌متصل با لایهٔ ورودی ارتباط دارد سریع‌تر از نمونه تمام‌متصل آن هم‌گرا می‌شود، بدون آنکه کارایی خود را از دست داده باشد. از آنجایی‌که بردارهای ویژگی استخراج‌شده از سیگنال گفتاری دارای طول ۱۸ برای هر قاب می‌باشند، در این مقاله نورون‌های بیشتری از بعد هر یک از بردارهای ویژگی در لایهٔ پنهان نخست در نظر گرفته شده‌اند تا بتوانند ویژگی‌های مختلفی را از هر قاب گفتاری آشکار کنند. در (نژادقلی و سیدصالحی، ۲۰۰۹؛ توحیدی پور و همکاران، ۲۰۱۲) ۳۲ نورون برای پردازش هر قاب در نظر گرفته شده‌اند؛ اما از آنجاکه دادگان

تعلیمی به‌کارگرفته‌شده در این مقاله بیشتر است، تعداد نورون‌ها به ۶۴ نورون افزایش داده شد. علاوه‌بر آن، این شبکه به‌صورتی طراحی شده است که لایه‌های پایین‌تر ویژگی‌های پویایی کوتاه‌تر و لایه‌های بالاتر ویژگی‌های پویایی بلندتری را کشف کنند. بنابراین، اطلاعات استخراج‌شده از ۹ قاب متوالی در لایهٔ پنهان نخست توسط یک سری از نورون‌ها (۵۱۲ نورون) در لایهٔ پنهان دوم پردازش می‌شود. پردازش اطلاعات این پویایی‌های کوتاه به‌منظور آشکارکردن اطلاعات اصوات متمایزگز که در بازشناسی آوا مؤثر است به‌کار می‌رود. از این‌رو پانزده توالی از آشکارسازها در لایهٔ پنهان دوم در نظر گرفته شده‌اند تا اطلاعات مربوط به توالی‌های کوتاه از نورون‌های لایهٔ پنهان نخست را پردازش کنند. سپس، ویژگی‌های استخراج‌شده در این لایه توسط ۶۲ نورون در لایهٔ پنهان سوم پردازش می‌شوند. در این لایه پویایی‌های بلندتر با پهنای پانزده توالی از آشکارسازهای لایهٔ پنهان پایین‌تر پردازش می‌شوند. و درنهایت بازشناسی آوا در لایهٔ خروجی به‌دست می آید.

این شبکهٔ عصبی طبق روند اشاره‌شده در زیربخش (۴-۱) به‌صورت لایه‌به‌لایه پیش‌تعلیم داده شد؛ سپس، به اندازه هزار تکرار، با استفاده از روابط اشاره شده در زیربخش (۴-۲) و با در نظر گرفتن $I=1$ تعلیم سرتاسری یافت. در شکل (۷) نمودار کاهش خطای تعلیم شبکه‌های یک لایهٔ پنهان ساخته شده برای تعلیم وزن‌های لایه‌های پنهان دوم و سوم برای تعداد تکرار یکسان نشان داده شده است. همان‌طور که در این شکل مشاهده می‌شود، خطای حداقل مربعات خطا در تعلیم وزن‌های لایه سوم پایین‌تر از مقدار آن در تعلیم وزن‌های لایهٔ دوم است. دلیل آن را می‌توان در اطلاعات چکیده‌تر و مفیدتر به‌دست‌آمده در لایهٔ سوم به‌منظور بازشناسی نهایی دانست.



شکل (۷): نمودار کاهش خطای تعلیم برای تعلیم دو شبکه یک‌لایهٔ پنهان ساخته‌شده برای پیش‌تعلیم وزن‌های لایه‌های دوم و سوم از شبکهٔ عصبی پیمانه‌ای عمیق

در ادامه، ساختار شبکهٔ عصبی پیمانه‌ای پیشنهادشده در جدول (۱) تعداد پیمانه‌های خود در هر لایه را به‌صورتی رشد می‌دهد که خروجی آن هم‌زمان بتواند پانزده توالی قابی آوایی را ببیند. مشخصات ساختاری شبکه توسعه‌یافته به‌صورت ساختار انجمن‌گر فضایی-زمانی دوگانه در جدول زیر آمده است.

(جدول- ۲): خصوصیات ساختاری شبکه عصبی پیمانه‌ای با ساختار انجمن‌گر فضایی-زمانی دوگانه

لایه	تعداد آشکارسازها	تعداد گره‌ها در هر آشکارساز	تعداد تأخیر زمانی هر پیمانه
لایه ورودی	۳۷	۱۸	-
لایهٔ نخست	۳۷	۶۴	۱
لایهٔ دوم	۲۹	۵۱۲	۹
لایهٔ سوم	۱۵	۶۲	۱۵
لایه خروجی	۱۵	۳۶	۱

انتخاب پانزده توالی ($I=15$) به این دلیل انجام شده است که بدون نیاز به عمیق‌ترکردن شبکه بتوان به مقایسه عملکرد ساختار شبکه عصبی پیمانه‌ای پایه با ساختار فضایی-زمانی دوگانه پرداخت. با انتخاب توالی‌های بزرگتر چون اطلاعات سطح بالاتری نیاز است توسط شبکه یادگیری شود، لازم است شبکه عمیق‌تر شود.

به منظور رشددادن شبکهٔ پیمانه‌ای پایه تا درنظرگرفتن پانزده توالی در خروجی، زنجیره‌ای از ۳۷ قاب گفتاری به ورودی اعمال می‌شود. در این حالت تعداد آشکارسازها در لایهٔ پنهان نخست به ۳۷ آشکارساز افزایش می‌یابد؛ سپس، در لایهٔ پنهان دوم به‌منظور پردازش اطلاعات استخراج شده از هر توالی متشکل از ۹ آشکارساز در لایهٔ پایین‌تر، ۲۳ توالی نورونی در این لایه در نظر گرفته شده است. در لایهٔ بعدی، اطلاعات هر پانزده توالی از خروجی نورون‌ها در لایهٔ پنهان پایینی توسط پانزده توالی آشکارساز در این لایه پردازش می‌شود؛ سپس، اطلاعات هر یک از آشکارسازهای ۶۲ نورونی در این لایه برای دست‌یابی به خروجی آوایی مورد بررسی قرار می‌گیرد. بنابراین، در خروجی شبکه ۱۵ توالی از ۳۶ نورون در نظر گرفته شده است.

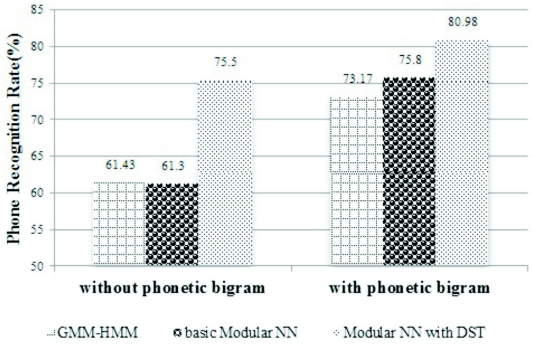
با قراردادن وزن‌های پیش‌تعلیم یافته به‌صورت لایه به لایه در شبکه، این ساختار به اندازه یکصد تکرار به‌صورت سرتاسری تعلیم داده شد.

چون توسعهٔ ساختار شبکهٔ بازشناس آوای پایه به ساختار DST با هدف قراردادن اطلاعات توالی آوایی قاب‌ها

در وزن‌های آن است، برای استفاده در آزمون دیگر نیازی به وجود چندین دسته نورونی در خروجی این شبکهٔ عصبی نیست. بنابراین، با برداشتن پیمانه‌های اضافه‌شده در تعلیم به‌منظور درنظر گرفتن توالی آوایی در خروجی، ساختار شبکه DST در آزمون به ساختار شبکهٔ عصبی پیمانه‌ای پایه کاهش یافت؛ اما با این تفاوت که وزن‌های این شبکه شامل اطلاعات توالی آوایی است.

جهت مقایسه عملکرد شبکه‌های عصبی بازشناس آوای پیمانه‌ای معرفی‌شده با مدل‌های احتمالاتی، یک مدل GMM-HMM که مجموعه‌ای از مدل‌های HMM چندحالته ساخته‌شده برای هر آوا است، با استفاده از معیار بیشینهٔ درست‌نمایی[1] تعلیم داده شد. در اینجا برای هر حالت از مدل‌های HMM از مدل‌های GMM با هشت گوسین برای دست‌یافتن به بهترین نتیجهٔ بازشناسی استفاده شد. به‌دلیل فرض استقلال شرطی در مدل‌های HMM، با هدف بالابردن کارآیی این مدل‌ها لازم است از بازنمایی‌های مستقل‌شده به‌عنوان ورودی مدل استفاده کرد. به‌علاوه، استفاده از مشتقات زمانی مرتبه نخست و دوم این بازنمایی‌ها برای اعمال اطلاعات دینامیک گفتار در تعلیم مدل HMM لازم است. بنابراین، برای تعلیم و آزمون این مدل از بردارهای ویژگی ایستا-پویای MFCC با بعد ۳۹ استفاده شد.

در نمودار شکل (۸) درصد دقت بازشناسی آوای به‌دست آمده از شبکه تعلیم یافته با استفاده از ساختار DST در مقابل شبکه عصبی عمیق پیمانه‌ای اولیه و مدل GMM-HMM نشان داده شده است.



(شکل-۸): درصددقت بازشناسی آوای به‌دست آمده توسط شبکهٔ عصبی پیمانه‌ای با ساختار DST، شبکهٔ عصبی پیمانه‌ای اولیه و مدل GMM-HMM، محاسبه‌شده توسط الگوریتم ویتربی در دو حالت استفاده از مدل بایگرم آوایی و بدون آن برای دادگان فارس دات

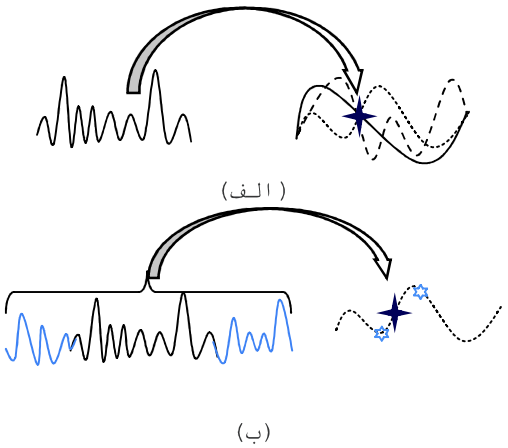
همان‌طور که در این جدول مشاهده می‌شود، توسعه ساختار شبکه عصبی پیمانه‌ای به‌صورت ساختار انجمن‌گر فضایی-زمانی دوگانه توانسته است، درصد بازشناسی آوا را بهبود دهد. مدل‌های صوتی مثل HMM یا شبکهٔ عصبی پیمانه‌ای عمیق پایه به‌گونه‌ای تعلیم یافته‌اند که هر بار یک زنجیره در فضای ورودی را به یک نقطه در فضای آواهای خروجی نگاشت دهند. بنابراین، در زمان آزمون، هر زنجیره بردار ویژگی ورودی به نقطه‌ای در فضای خروجی نگاشته می‌شود که می‌تواند به زنجیره‌های مختلفی از آواها در خروجی تعلق داشته باشد؛ اما اگر در تعلیم مدل صوتی، زنجیرهٔ معتبر خروجی نیز به آن تعلیم داده شود، خاصیت انجمن‌بودن زنجیرهٔ الگوهای خروجی به میدان می‌آید. درنتیجه، همان‌طور که در شکل (۹-ب) نشان داده شده است، می‌توان به کمک اطلاعات آواهای اطراف به طبقه‌بندی صحیح‌تری از آوای وسط در خروجی دست یافت و خروجی شبکه به طبقه‌ای که دنبالهٔ خروجی به‌دست آمده در ازای آن معتبر نیست، برده نمی‌شود؛ که این حالت در تعلیم شبکهٔ عصبی DST اعمال شده است. هم‌چنین، ساختار شبکه عصبی انجمن‌گر فضایی-زمانی دوگانه توانسته است تا حدی نقش مدل‌سازی آوایی را نیز اجرا کند؛ زیرا بدون استفاده از اطلاعات مدل آوایی در رمزگشایی، درصد بازشناسی قابل قبولی به‌دست داده است. هم‌چنین، با توجه به این جدول مشاهده می‌شود که شبکه عصبی پیمانه‌ای پایه در مقایسه با مدل GMM-HMM توانسته‌است در صورت استفاده از مدل بایگرم آوایی، بهتر عمل کند؛ دلیل آن، طراحی دقیق‌تر ساختار این شبکه و توانایی استخراج اطلاعات مفیدتر از دادگان ورودی از طریق تعلیم مناسب آن است.

۵-۲- دادگان گفتاری فارس‌دات بزرگ

این دادگان شامل سیگنال گفتار مربوط به یکصد گوینده با تفاوت‌های سن، لهجه و تحصیلات است. در این دادگان، هر گوینده ۲۵ صفحه از روزنامه با موضوعات مختلف را می‌خواند. در کل، هر فرد حدود چهارهزار لغت از متون مختلف روزنامه‌ای را بیان کرده است. این متون شامل موضوعات مختلف از جمله سیاسی، اقتصادی، فرهنگی، ورزشی و ادبی است (شیخ زادگان و بیجن خان، ۱۳۸۵). از این میان، گفتار مربوط به ۹۵ گوینده که دربردارنده حدود ۵۷ ساعت گفتار است به‌عنوان دادهٔ تعلیم و گفتار مربوط به پنج نفر دیگر که شامل حدود سه ساعت گفتار است برای

[1] Maximum Likelihood

آزمون مورد استفاده قرار گرفت. در ادامه به توصیف آزمایش‌های انجام‌شده با استفاده از این دادگان پرداخته می‌شود.



(شکل- ۹): الف) نگاشت سیگنال گفتار به یک نقطه در فضای خروجی که می‌تواند مربوط به زنجیره‌های آوایی متفاوت باشد. این نگاشت توسط بسیاری از مدل‌های صوتی مثل HMM انجام می‌شود. ب) نگاشت سیگنال گفتار به تنها یک زنجیره از آواهای خروجی که توسط شبکه با ساختار DST انجام می‌شود.

۵-۲-۱- تعلیم نیمه‌سرپرستی‌شده شبکهٔ عصبی با ساختار انجمن‌گر فضایی-زمانی دوگانه بر روی فارس‌دات بزرگ

با توجه به آنچه در زیربخش (۴-۳) اشاره شد، برای تنظیم شبکهٔ عصبی پیمانه‌ای عمیق با ساختار انجمن‌گر فضایی-زمانی دوگانه بر روی دادگان بزرگ از شبکهٔ عصبی تعلیم‌یافته برروی دادگان فارس‌دات استفاده می‌شود. با توجه به وجود برچسب‌های زمانی برای هر لغت در دادگان بزرگ، شبکهٔ عصبی DST در طول لغت حرکت می‌کند و دنباله آوایی $\boldsymbol{\beta}_{u_i}$ نشان داده شده در شکل (۶) را ایجاد می‌کند. هم‌چنین، دنباله آوایی $\boldsymbol{\alpha}_{u_i}$ براساس میانگین طول آواهای محاسبه‌شده از روی دادگان فارس‌دات و حاشیه‌نویسی‌های واجی هر لغت در فارس‌دات بزرگ به‌دست می‌آید؛ سپس، الگوریتم برنامه‌ریزی پویا برای مقایسه $\boldsymbol{\alpha}_{u_i}$ با $\boldsymbol{\beta}_{u_i}$ و اصلاح آن براساس نتایج به‌دست آمده به‌کار می‌رود.

نکتهٔ قابل توجه این است که برای تعلیم شبکهٔ عصبی پیمانه‌ای برروی دادگان فارس‌دات از برچسب‌های آوایی بهره گرفته شده است. بنابراین، نتایج بازشناسی شبکه به‌صورت آوایی است؛ اما برچسب‌دهی براساس میانگین طول آواها براساس حاشیه‌نوشته‌های واجی لغات انجام گرفته است. بنابراین، اختلافاتی بین آرایه‌های به‌دست‌آمده از دو روش وجود دارد که مجبور به پذیرش این خطای برچسب‌دهی هستیم؛ ولی خطاهای حذف و درج که بیشتر به‌دلیل عدم مرزبندی صحیح برچسب‌ها است تا حد خوبی اصلاح می‌شوند. این روند تعلیم شبکه براساس برچسب‌های به‌دست‌آمده و استفاده از آن برای تولید برچسب جدید دوبار تکرار می‌شود.

در جدول (۳) درصد بازشناسی واج به‌دست‌آمده توسط شبکه DST تعلیم‌یافته برروی دادگان فارس‌دات بزرگ به روش نیمه‌سرپرستی‌شده نشان داده شده است. جهت بررسی اثر روش برچسب‌دهی ارائه‌شده، در آزمایش دیگری، تعلیم شبکهٔ DST برروی دادگان بزرگ فقط با استفاده از برچسب‌های به‌دست آمده از میانگین طول آواها انجام گرفت. هم‌چنین نتایج به‌دست آمده با نتیجه بازشناسی واج مدل HMM تعلیم‌یافته برروی این دادگان در جدول زیر مقایسه شده است.

(جدول-۳): درصد دقت بازشناسی واج بر روی دادگان فارس‌دات بزرگ با استفاده از شبکهٔ عصبی با ساختار فضایی-زمانی تعلیم‌یافته به‌صورت نیمه‌سرپرستی‌شده در مقایسه با مدل HMM

مدل بازشناس آوا + روش تعلیم بر روی دادگان بزرگ	درصد بازشناسی واج
مدل DST تعلیم یافته برروی دادگان فارس دات	۵۳
مدل DST + تعلیم نیمه‌سرپرستی شده بر اساس میانگین طول آواها	۶۴
مدل DST + تعلیم نیمه‌سرپرستی شده بر اساس ترکیب اطلاعات میانگین طول آواها و نتایج بازشناسی شبکه DST	۶۹/۶
مدل HMM + بیشترین درست‌نمایی	۷۳

با مقایسه نتایج به‌دست‌آمده مشاهده می‌شود که با استفاده از روش نیمه‌سرپرستی‌شده اشاره شده، نسبت به وقتی که تنها از میانگین طول آواها استفاده می‌شود، اطلاعات تمایز آوایی بهتری به شبکه تعلیم داده شده است. هم‌چنین، درصد دقت بازشناسی واج به‌دست آمده با به‌کارگیری مدل HMM بالاتر از درصد بازشناسی به‌دست‌آمده توسط مدل شبکهٔ عصبی با ساختار DST است. دلیل این تفاوت را می‌توان به این خاصیت مدل‌های HMM نسبت داد که برخلاف شبکه‌های عصبی، قابلیت تعلیم بر روی دادگانی را که برچسب‌های زمانی ندارند، دارند؛ لذا با تعلیم اولیه مدل‌های HMM برای هر آوا برروی دادگان

فارس‌دات، ادامه تعلیم آنها برروی دادگان بزرگ با مشکلی مواجه نمی‌شود. به‌طوری که با دادن اطلاعات مربوط به دنبالهٔ لغات و فرهنگ واژگان تلفظ آنها بدون آنکه زمان‌بندی لغات مشخص باشد می‌توان پارامترهای مدل‌های HMM را بازتخمین کرد.

۶- بحث و نتیجه‌گیری

در این مقاله، یک ساختار شبکهٔ عصبی عمیق پیمانه‌ای قابل رشد به‌منظور بازشناسی آوا معرفی شد. این شبکه علاوه‌بر هماهنگی با ساختار پردازش سلسله‌مراتبی گفتار در مغز انسان، می‌تواند جهت پردازش اطلاعات پیچیده‌تر رشد یافته و عمیق‌تر شود؛ که این خاصیت نیز با ساختار لازم برای پردازش اطلاعات پیچیده در مغز هماهنگ است. در طراحی ساختار این شبکه برخلاف شبکه‌های پیمانه‌ای متداول که هر پیمانه به‌صورت مستقل عمل می‌کند، پیمانه‌ها در تعامل با یکدیگر می‌باشند. تعامل پیمانه‌ها، امکان پردازش اطلاعات دینامیک گفتار در طول سلسله‌مراتب را برای استخراج اطلاعات غنی‌تر به‌وجود می‌آورد. درواقع، این شبکهٔ عصبی به‌دلیل ساختار طراحی‌شدهٔ خود که مبتنی بر اطلاعات گفتاری است، با استخراج الگوهای زمانی بهتر از ورودی در لایهٔ نخست و سپس ترکیب این الگوها با هدف استخراج اطلاعات دینامیک به‌ترتیب از ساده‌ترین تا پیچیده‌ترین آنها توانسته است اطلاعات بیشتری در مورد قاب‌های ورودی به‌دست آورد.

شبکه عصبی پیمانه‌ای رشدیافته به‌صورت ساختار انجمن‌گر فضایی-زمانی دوگانه همان‌طور که در شکل (۸) نشان داده شده است می‌تواند به درصد بازشناسی واج بالاتری در مقایسه با مدل HMM برسد. بالابودن کارایی این شبکه به توانایی آن در پالایش‌کردن توالی‌های آوایی نامعتبر در لایهٔ خروجی مربوط می‌شود. پالایش زنجیره‌های خروجی غلط توسط این شبکه براساس این فرضیه که شبکه‌های عصبی با توجه به ظرفیت خود مؤلفه‌های غیرخطی فضای دادگان تعلیمی را یاد می‌گیرند و نمونه‌های دیگر را به کمک مؤلفه‌های یادگرفته‌شده به فضای دانش خود نگاشت می‌دهند استوار است. شبکهٔ عصبی انجمن‌گر فضایی-زمانی دوگانه به‌منظور دخالت‌دادن مؤلفه‌های مربوط به ارتباطات انجمنی الگوها در فضای ورودی با فضای خروجی در عملکرد شبکه طراحی شده است. در این حالت، شبکه براساس مؤلفه‌های مختلفی که در تعلیم یادگرفته‌است، هر توالی ورودی را مورد بررسی قرار داده و ورودی را به یکی از توالی‌های خروجی یاد گرفته شده می‌برد. در این حالت، در مقایسه با زمانی که شبکه فقط یک طبقهٔ مربوط به هر الگوی ورودی را یاد می‌گیرد، اطلاعات توالی کمک می‌کند نمونه‌هایی که به‌صورت موضعی نوفه‌ای شده‌اند به طبقه غلطی برده نشوند؛ بلکه اطلاعات مربوط به توالی خروجی‌ها که در ساختار شبکه شکل گرفته و یادگیری شده آنها را به طبقه‌ای که زنجیره آواهای مربوط به آن در زمان تعلیم به شبکه تعلیم داده شده است، می‌برد و درنتیجه به بازشناسی صحیح‌تر خروجی منجر می‌شود.

همان‌طور که در جدول (۳) نشان داده شده است، شبکهٔ عصبی با ساختار DST با تعلیم نیمه‌سرپرستی‌شده بر روی دادگان فارس‌دات بزرگ توانست به درصد بازشناسی واج قابل مقایسه‌ای با نتیجه بازشناسی مدل HMM دست یابد. با وجود آنکه مدل‌های HMM برای هر آوا قابلیت تعلیم بر روی دادگانی را که برچسب‌های زمانی ندارند دارند؛ اما در فرآیند تعلیم این مدل‌ها وجود یک متخصص چه در مرحله تولید واژه‌نامه‌های تلفظی لغات و چه در مرحله گره زدن مدل‌های وابسته به محتوا بسیار مؤثر است. درصورتی‌که با تعلیم نیمه‌سرپرستی‌شده شبکهٔ عصبی برروی دادگان بزرگ بدون استفاده از اطلاعات مدل HMM و در نتیجه به‌کار گرفتن دانش یک متخصص می‌توان به درصد بازشناسی قابل مقایسه‌ای با مدل HMM دست یافت.

تعلیم نیمه‌سرپرستی‌شده این شبکه تا این حد برای استفاده به‌عنوان شبکه پایه‌ای برای توسعه‌یافتن به‌منظور بازشناسی لغت می‌تواند کافی باشد؛ زیرا اطلاعات سطح بالاتر واژگانی در شبکهٔ توسعه‌یافته می‌تواند برای تنظیم دقیق‌تر وزن‌های شبکهٔ بازشناس واج مورد استفاده قرار گیرد. این نتیجه با چگونگی بازشناسی واج در انسان که اطلاعات واژگانی در بسیاری از موارد به‌کمک بازشناسی واج می‌آیند، هماهنگ است (ساموئل، ۲۰۱۱).

در ادامهٔ کار، این ساختار از شبکه‌های عصبی طوری رشد داده می‌شوند که بتوانند به‌صورت یک‌پارچه و بدون نیاز به واژگان و مدل زبانی به بازشناسی لغت بپردازند. در این راستا لازم است شبکهٔ عصبی طراحی‌شده برروی کل دنباله واجی مربوط به هر گفته از گفتار هر فرد تعلیم داده شود تا بتواند اطلاعات صوتی و زبانی لغات را به‌صورت یکجا یاد بگیرد. برای رسیدن به این هدف علاوه‌بر عمیق‌ترکردن شبکه موجود لازم است اتصالات بازگشتی نیز در ساختار آن در نظر گرفته شود تا بتواند زنجیره‌های واجی طولانی مدت

را مدل کند. علاوه‌بر آن لازم است توابع هدف مناسبی برای تعلیم سرتاسری شبکه‌های رشدیافته به‌منظور افزایش درصد بازشناسی لغت تولیدشده تعریف شود.

تشکر و قدردانی

بدین‌وسیله از آقای دکتر شکفته و پژوهش‌گران گروه پردازش صوت و زبان طبیعی پژوهشکده پردازش دادهٔ پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی به‌خاطر همکاری در انجام آزمون‌های مربوط به مدل‌های HMM برروی دادگان فارس دات بزرگ قدردانی می‌شود.

۷- مراجع

انصاری، ل.، مدلسازی اثرات هم تولیدی آواها در یک مدل شبکه عصبی بازشناخت گفتار، پایان نامه کارشناسی ارشد، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۲

رحیمی نژاد، م.، سیدصالحی، س.ع.، مقایسه و ارزیابی کارایی انواع روش‌های استخراج پاراکترهای بازنمایی و هنجارسازی در بازشناسی مستقل از گوینده گفتار، ۱۳۸۲، ش. ۵۵

سیدصالحی، س.ز.، سیدصالحی، س.ع.، روش پیش تعلیم سریع برمبنای کمینه سازی خطا برای همگرایی یادگیری شبکه‌های عصبی با ساختار عمیق، دو فصل نامه پردازش علائم و داده‌ها، ۱۳۹۲، دوره ۱۹، ش.۱، ص. ۱۳-۲۶

سیدصالحی، س.ع.، بازشناخت گفتار پیوسته فارسی با استفاده از مدل عملکردی مغز انسان در درک گفتار، پایان نامه دکترا، دانشکده فنی و مهندسی، دانشگاه تربیت مدرس، ۱۳۷۴

سیدصالحی، س.ع.، طراحی سامانه بازشناسی گفتار بر اساس شبکه عصبی به منظور بازشناسی دادگان گفتاری فارسی با تعداد لغات زیاد، گزارش پژوهشی، پژوهشکده پرازش هوشمند علائم، ۱۳۸۴

شکفته، ی.، الماس گنج، ف.، بهبود بازشناسی گفتار با استفاده از شبکه‌های عصبی دربرگیرنده الگوهای زمانی، کنفرانس بین المللی فناوری اطلاعات (IKT 2007)، دانشگاه فردوسی مشهد، ۱۳۸۶

شیخ زادگان، ج.، بیجن خان، م.، دادگان‌های گفتاری زبان فارسی، دومین کارگاه پژوهشی زبان فارسی و رایانه، ۱۳۸۵، ص. ۲۴۷-۲۶۱

کرمی، ش.، به‌کارگیری اطلاعات موجود در نواحی گذرای مرز واج‌ها به منظور افزایش توان مدل‌های شبکه عصبی بازشناس گفتار مستقل از گوینده، پایان نامه کارشناسی ارشد، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۷۹

یزدیان، م.، طراحی مدل بازشناسی گفتار توسط شبکه‌های عصبی برپایه پردازش وقایع گسسته سیگنال گفتار، پایان نامه کارشناسی ارشد، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۰

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In Proc. ICASSP. 2012, March. pp. 4277-4280

Andrew, G., &Bilmes, J.. Sequential deep belief networks. In Proc. ICASSP. 2012, March. pp. 4265-4268

Bacchiani, M., Senior, A., & Heigold, G.. Asynchronous, Online, GMM-free Training of a Context Dependent Acoustic Model for Speech Recognition. In Proc. Fifteenth Annual Conference of the International Speech Communication Association. 2014

Bengio, Yoshua. Artificial neural networks and their application to sequence recognition. PhD Dissertation. McGill University.1991

Bengio, Y., Lamblin, P., Popovici, D. & Larochelle, H.. Greedy layer-wise training of deep networks. In Proc. NIPS. 2007. pp. 153-160

Bijankhan, M., Sheikhzadegan, J., Roohani, M. R., Samareh, Y., Lucas, C., &Tebyani, M..FARSDAT-The speech database of Farsi spoken language. In Proc. of the Australian Conference on Speech Science and Technology. 1994, December.Vol. 2. pp. 826-830

Bourlard, H., & Morgan, N..Continuous speech recognition by connectionist statistical methods.Neural Networks, IEEE Transactions on. 1993. Vol.4. No. 6. pp. 893-909

Bourlard, H. A., & Morgan, N. Connectionist speech recognition: a hybrid approach. Springer.1994. Vol. 247

Bridle, J. S., & Dodd, L..An Alphanet approach to optimising input transformations for continuous speech recognition. In Proc. ICASSP. 1991, April.pp. 277-280

Chen, B., Zhu, Q., Morgan, N.. Tonotopic multilayer perceptron a neural network for learning long term temporal features for speech recognition. In Proc. ICASSP. 2005. pp. 945-948

Dahl, G. E., Yu, D., Deng, L., &Acero, A..Context-depen-dent pre-trained deep neural networks for large-vocabulary speech recognition.Audio, Speech,

Mohamed, A. R., Dahl, G. E., & Hinton, G..Acoustic modeling using deep belief networks.Audio, Speech, and Language Processing, IEEE Transactions on. 2012. Vol.20. No. 1. pp. 14-22

Morgan, N., &Bourlard, H.. Continuous speech recognition using multilayer perceptrons with hidden Markov models. In Proc. ICASSP.1990. pp. 413-416

Nejadgholi, I., Seyyedsalehi, S.A..Nonlinear normalization of input patterns to speaker variability in speech recognition neural networks. Neural Computing and Applications.2009. Vol. 18.pp. 45-55

Pinto, J., Garimella, S., Hermansky, H., &Bourlard, H..Analysis of MLP-based hierarchical phoneme posterior probability estimator. .Audio, Speech, and Language Processing, IEEE Transactions on. 2011. Vol. 19. No. 2. pp. 225-241

Prabhavalkar, R., &Fosler-Lussier, E..Backpropagation training for multilayer conditional random field based phone recognition. In Proc. ICASSP. 2010, March. pp. 5534-5537

Sainath, T. N., Kingsbury, B., Ramabhadran, B., Fousek, P., Novak, P., & Mohamed, A. R.. Making deep belief networks effective for large vocabulary continuous speech recognition. In Proc. Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on. 2011, December. pp. 30-35

Samuel. A. G.. Speech perception. Annual review of psychology. 2011. Vol. 62. pp. 49-72

Seide, F., Li, G., & Yu, D..Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. In Proc.INTERSPEECH. 2011, August. pp. 437-440

Senior, A., Heigold, G., Bacchiani, M., & Liao, H.. GMM-free DNN training. In Proc. ICASSP. 2014. pp. 5639-5643

Tan, Z. H., & Lindberg, B. (Eds.). Automatic speech recognition on mobile devices and over communication networks. Springer Science & Business Media. 2008

Tohidypour, H. R., Seyyedsalehi, S. A., Behbood, H., & Roshandel, H.. A new representation for speech frame recognition based on redundant wavelet filter banks. Speech Communication. 2012. Vol. 54. No. 2. pp. 256-271

Trentin, E., &Gori, M.. A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing. 2001. Vol. 37. No. 1. pp. 91-126

Van Bael, Ch.,Boves, L., van den Heuvel, H. & Strik, H.. Automatic phonetic transcription of large speech corpora. Computer speech and Language. 2007. Vol.21. No. 4. pp. 652-668

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J..Phoneme recognition using time-delay neural networks.Acoustics, Speech and Signal Processing, IEEE Transactions on. 1989. Vol. 37. No. 3. pp. 328-339

and Language Processing, IEEE Transactions on. 2012. Vol. 20. No. 1. pp. 30-42

Deng, L., Yu, D., & Platt, J.. Scalable stacking and learning for building deep architectures. InProc. ICASSP. 2012, March.pp. 2133-2136

He, X., Deng, L., & Chou, W.. Discriminative learning in sequential pattern recognition. Signal Processing Magazine, IEEE. 2008. Vol. 25. No. 5. pp. 14-36

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B.. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE. 2012. Vol. 29. No. 6.
pp. 82-97

Hinton, G., Osindero, S. & Teh, Y.. A fast learning algori-thm for deep belief nets. Neural Comput., 2006. Vol. 18. pp. 1527-1554

Jaitly, N., Nguyen, P., Senior, A. W., &Vanhoucke, V. .Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition. In INTERSPEECH.2012, September

Juang, Biing-Hwang, Wu Hou, and Chin-Hui Lee. Minimum classification error rate methods for speech recognition.Speech and Audio Processing, IEEE Transactions on. 1997. Vol. 5. No. 3. pp. 257-265

Kapadia, S., V. Valtchev, and S. J. Young. MMI training for continuous phoneme recognition on the TIMIT database. In Proc. ICASSP.1993. Vol. 2

Kingsbury, B..Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In Proc. ICASSP. 2009, April. pp. 3761-3764

Konig, Y..Remap: recursive estimation and maximization of a posteriori probabilities in transition-based speech recognition. PhD Dissertation. University of California. Berkeley.1996

Lang, K. J., Waibel, A. H., & Hinton, G. E.. A time-delay neural network architecture for isolated word recognition. Neural networks,1990. Vol. 3. No. 1. pp. 23-43

LeCun, Y., &Bengio, Y..Convolutional networks for images, speech, and time series.The handbook of brain theory and neural networks.3361.1995

Lewandowski, N., Droppo, J., Seltzer, M., & Yu, D. Phone sequence modeling with recurrent neural networks. In Proc. ICASSP. 2014

McDermott, E., Hazen, T. J., Le Roux, J., Nakamura, A., &Katagiri, S..Discriminative training for large-vocabulary speech recognition using minimum classification error.Audio, Speech, and Language Processing, IEEE Transactions on. 2007. Vol. 15. No. 1.pp.203-223

Mohamed, A. R., Yu, D., & Deng, L.. Investigation of full-sequence training of deep belief networks for speech recognition. In INTERSPEECH. 2010, September. pp. 2846-2849

Yu, D., Deng, L., &Seide, F..The deep tensor neural network with applications to large vocabulary speech recognition.Audio, Speech, and Language Processing, IEEE Transactions on,2013. Vol.21. No. 2. pp. 388-396

Yu, D., Deng, L. Automatic Speech recognition: A deep learning approach. Springer. 2014.

زهره انصاری مدرک کارشناسی خود را در رشته مهندسی پزشکی- بیوالکتریک از دانشگاه اصفهان در سال ۱۳۸۶ و کارشناسی ارشد را در همان رشته از دانشگاه صنعتی امیرکبیر در سال ۱۳۸۹ دریافت کرده است. وی هم‌اکنون دانشجوی دکترای بیوالکتریک در دانشگاه صنعتی امیرکبیر است. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش و بازشناسی گفتار، شناسایی الگو، شبکه‌های عصبی مصنوعی و یادگیری عمیق می‌باشد.

نشانی رایانامهٔ ایشان عبارت است از:

z_ansari@aut.ac.ir

سیّد علی سیّدصالحی 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 کرده است. وی در حال حاضر دانشیار دانشکده مهندسی پزشکی دانشگاه صنعتی امیرکبیر است. زمینه‌های پژوهشی مورد علاقه ایشان پردازش و بازشناسی گفتار، شبکه‌های عصبی مصنوعی و زیستی، مدل‌سازی عملکرد مغز و پردازش خطی و غیرخطی سیگنال است.

نشانی رایانامهٔ ایشان عبارت است از:

ssalehi@aut.ac.ir