

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

زهره انصاری و سیدعلی سیدصالحی

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

چکیده

در این مقاله به معرفی شبکه‌های عصبی پیمانه‌ای عمیق و قابل رشد به منظور بهبود بازشناسی گفتار پیوسته پرداخته می‌شود. ساختار این شبکه‌ها و روش‌های پیش‌تعلیم معرفی شده برای آنها به‌گونه‌ای است که در عین هماهنگی با ساختار گفتار، در حافظه و محاسبات لازم صرفه‌جویی می‌شود. بدلیل قابلیت رشد این ساختارها، در تعلیم آنها اطلاعات فضایی-زمانی بردارهای بازنمایی در ورودی و اطلاعات فضایی-زمانی برجسب آوای آنها را در خروجی شبکه عصبی می‌توان انجمان کرد. شبکه تعلیم‌یافته با این ساختار انجمان‌گر فضایی-زمانی دوگانه، می‌تواند زیرفضای زنجیره‌های معتبر آوایی دادگان را 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 فارس‌داد و فارس‌داد بزرگ استفاده شد. نتایج آزمایش‌ها نشان می‌دهند که می‌توان دقت بازشناسی آوا را ببروی دادگان فارس‌داد تا ۷/۲٪ با استفاده از شبکه‌های عصبی پیمانه‌ای عمیق نسبت به مدل‌های مخفی مارکوف بالا برد؛ که با توجه آنها به ساختار فضایی-زمانی دوگانه این نتیجه تا ۵/۱٪ بهبود می‌یابد. بدلیل عدم وجود برجسب‌های آوایی برای دادگان بزرگ، یک روش تعلیم نیمه‌سرپرستی شده برای تعلیم شبکه‌های عصبی بررسی می‌یابد. این دادگان پیشنهاد شده است که می‌تواند به درصد بازشناسی قابل مقایسه‌ای با مدل‌های مخفی مارکوف دست یابد.

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

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، به صورت گسترشده از ساختارهای شبکه عصبی با یک یا دو لایه پنهان و همچنین، ترکیب شبکه‌های عصبی با HMM (HMM-HMM) در بازشناسی گفتار پرداخته می‌شد (وابیل و همکاران، ۱۹۸۹؛ مورگان و بورلارد، ۱۹۹۰؛ بنجیو، ۱۹۹۱؛ بورلارد و مورگان، ۱۹۹۴-۱۹۹۲؛ ترننتین 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.

۱- مقدمه

امروزه سامانه‌های بازشناسی گفتار خودکار^۱ (ASR) به صورت گسترشده در تعامل انسان با ماشین به کار می‌روند. در این سامانه‌ها، مدل‌های مخفی مارکوف^۲ (HMM) یا شبکه‌های عصبی^۳ (ANN) می‌توانند به عنوان مدل‌های صوتی برای مدل‌کردن بردارهای بازنمایی گفتاری به صورت واحدهای گفتاری (مثل آوا یا سه‌آوا^۴) استفاده شوند. در سامانه‌های ASR با مدل صوتی HMM^۵ در هر حالت^۶، مدل‌های GMM^۷ (برای مدل‌کردن بردارهای بازنمایی گفتاری مخلوط گوسی^۸) مطالعات زیادی نیز

^۱Automatic Speech Recognition Systems (ASR)

^۲Hidden Markov Models (HMM)

^۳Artificial Neural Networks (ANN)

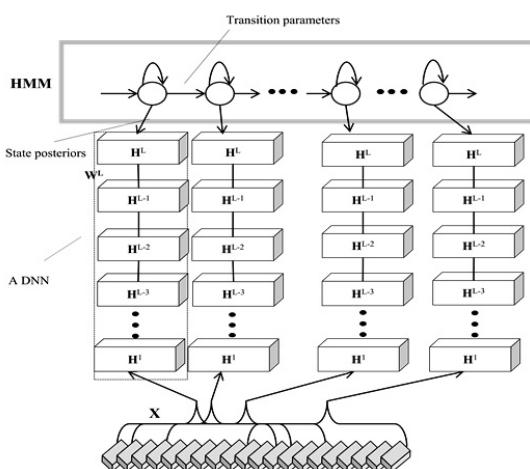
^۴Triphone

^۵State

^۶Gaussian Mixture Models (GMM)

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 بالا بردن کارآیی آنها در بازشناسی گفتار، پیشنهاد شده است (بریدل و داد، ۱۹۹۱؛ کونیگ، ۱۹۹۶؛ هه و همکاران، ۲۰۰۸؛ کینگسبری، ۲۰۰۹؛ محمد و همکاران، ۲۰۱۰؛ پرباوالکار و فوسلر-لوسیر، ۲۰۱۰؛ لواندووسکی و همکاران، ۲۰۱۴). اما در تمامی روش‌های معرفی شده 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 شبکه عصبی که بتواند اطلاعات توالی آوایی را به همراه اطلاعات صوتی، بهترین از روی دادگان استخراج کند، معرفی نشده است. در صورت معرفی چنین ساختاری، وظيفة 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

امروزه با پیشرفت‌هایی که در زمینه معرفی روش‌های یادگیری عمیق و ارائه ساخت‌افزارهای با سرعت پردازش بالا به دست آمده، امکان استفاده از شبکه‌های عصبی عمیق^۱ (DNN) در بازشناسی گفتار به وجود 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 بیلسس، ۲۰۱۲؛ عبدالحمید 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 بازیابی یک پنجره متحرک از توالی قاب‌های ورودی به دست می‌آورند؛ سپس، مدل‌های گرافیکی مثل HMM یا میدان تصادفی شرطی^۲ (CRF) برای مدل‌کردن وابستگی‌های زنجیره‌ای خطی توالی خروجی به دست آمده ترکیب می‌شوند. در واقع، در این ساختارها شبکه عصبی عمیق به عنوان مدل صوتی و مدل‌های گرافیکی اشاره شده به عنوان مدل‌های آوایی مورد استفاده قرار می‌گیرند و هر یک به صورت جداگانه تعلیم می‌بینند.

شبکه‌های عصبی عمیق به کارگرفته شده در مدل‌سازی صوتی به طور معمول دارای تعداد زیادی لایه‌های پنهان غیرخطی با تعداد نورون‌های یکسان در هر لایه هستند. لایه خروجی این شبکه‌ها نیز دارای تعداد زیادی نورون برابر با تعداد حالت‌های مدل‌های آوایی است. بنابراین، تعداد پارامترهای این شبکه‌ها بسیار زیاد است. با وجود معرفی روش‌های پیش تعلیم 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 معرفی شده، استفاده از شبکه‌های کانولوشنی است (لیکان و بنجیو، ۱۹۹۵)؛ اما هنوز موارد

^۱ Deep Neural Networks (DNN)

^۲ Conditional Random Field (CRF)

^۳ state



شبکه‌های کانولوشنی، روند پیش‌تعلیم معرفی شده می‌تواند در پیش‌تعلیم شبکه‌های کانولوشنی نیز مورد استفاده قرار بگیرد. با رشد دادن شبکه بازشناس آواز پیمانه‌ای پیش‌نهاش شده با قابلیت انجمن‌سازی فضایی-زمانی دوگانه به یک شبکه عصبی عمیق بازشناس لغت یک‌پارچه‌ای که می‌تواند در بازشناسی گفتار پیوسته با تعداد لغات زیاد^۲ مورد استفاده قرار بگیرد می‌توان دست یافته؛ 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 به طور معمول به کمک مدل‌های GMM-HMM که از قبل برروی این دادگان تعلیم یافته‌اند با مرزبندی تحمیلی گفتار^۳ به دست 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 به صورت جداگانه بر روی دادگان بزرگ را دارد. در همین اواخر پژوهش‌های زیادی در زمینه تعلیم NDNها 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 نیمه‌سرپرستی شده برای تنظیم شبکه عصبی عمیق با ساختار انجمن‌گر فضایی-زمانی دوگانه بر روی دادگان بزرگ معرفی می‌شود که به مدل‌های احتمالاتی از پیش‌تعلیم یافته نیازی ندارد و فقط از اطلاعات موجود بهره می‌گیرد.

در ادامه مقاله، در بخش دوم ساختار پیش‌نهاشی شبکه‌های عصبی پیمانه‌ای عمیق به منظور بازشناسی آواز معرفی می‌شود. در بخش سوم به توضیح مراحل رشد آنها به منظور 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 نیمه‌سرپرستی شده برای تنظیم این ساختار از شبکه‌های عصبی بر روی دادگان بزرگ که برچسب‌های آوایی برای آنها وجود ندارد، معرفی می‌شود. در بخش پنجم نتایج پیاده‌سازی‌های انجام شده بر روی دادگان گفتاری فارس‌داد و فارس‌داد بزرگ ارائه می‌شود. در نهایت، به بحث و نتیجه‌گیری نهایی پرداخته می‌شود.

²Large Vocabulary Continuous Speech Recognition (LVCSR)

³Force alignment

لایه‌های بعدی تنها از محدوده کوچکی از ورودی خود اطلاعات دریافت می‌کند. 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، به دلیل آنکه تمامی نورون‌های آن لایه به جستجو برای یافتن یک ویژگی یکسان در بخش‌های مختلف ورودی می‌پردازند، به مقاوم‌سازی شبکه عصبی نسبت به برچسب‌های غیر دقیق مکان آن ویژگی در قاب‌های گفتاری کمک می‌کند.

این ساختار از شبکه‌ها می‌تواند برای تعلیم اطلاعات توالی رشد داده شوند. در مواقعی که لازم است شبکه 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 دخالت دادن اطلاعات توالی آوایی قاب‌های گفتاری در تعلیم، ساختار این شبکه‌ها به گونه‌ای رشد داده می‌شود تا بتواند هم‌زمان با اطلاعات فضایی-زمانی قاب‌های گفتاری در ورودی شبکه عصبی، اطلاعات فضایی-زمانی برچسب‌های آنها را در خروجی انجمن کند. ساختار شبکه عصبی طراحی شده در این راستا که آن را ساختار انجمن‌گر فضایی-زمانی دوگانه^۱ (DST) می‌خوانیم، می‌تواند زیرفضای زنجیره‌های معتبر آوایی قاب‌ها در داده تعلیمی را یاد بگیرد. بنابراین، این شبکه می‌تواند به صورت خودسازمانده، توالی‌های خروجی نامعتبر را پالایش کند.

مسئله تعلیم شبکه‌های عصبی عمیق پیمانه‌ای معرفی شده یک مسئله غیر محدب 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، شبکه عصبی به صورت لایه‌به‌لایه تعلیم داده می‌شود. با توجه به ساختار پیمانه‌ای و وزن‌های رونوشت شده در زمان این شبکه، در هر لایه فقط وزن‌های مربوط به یک پیمانه تعلیم داده شده و وزن محاسبه شده برای بقیه پیمانه‌ها رونوشت می‌شود. این کار به میزان زیادی از حافظه مورد نیاز و حجم محاسبات لازم می‌کاهد. به دلیل شباهت این ساختار از شبکه‌های عصبی به

¹Double Spatio-Temporal Association (DST)

۲- معرفی شبکه‌های عصبی پیمانه‌ای عمیق بازشناس آوا

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 به شمار 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. به خصوص، در مواقعی که ساختارهای محلی مشابهی در نقاط مختلف ورودی وجود داشته باشد.

همان‌طور که در شکل (۲) نشان داده شده است، در شبکه پیمانه‌ای معرفی شده در این مقاله هر یک از آشکارسازهای لایه 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 فضای دیگر که به طور معمول دارای بعد بالاتری است، باعث می‌شود هر بردار بازنمایی به صورت جداگانه در مرحله نخست پردازش شود. این نگاشت در پژوهش‌های مختلفی در ساختار شبکه‌های عصبی بازشناس گفتار به کار رفته است (نژادقلی و سیدصالحی، ۱۳۷۹؛ یزدانی، ۱۳۸۰؛ کرمی، ۱۳۷۹) و منجر به کاهش زمان یادگیری نسبت به شبکه‌های تمام متصل در لایه نخست می‌شود. در صورتی که هر یک از نورون‌ها در دسته نورون‌های لایه پنهان نخست محدود به درنظر گرفتن اطلاعات مربوط به یک باند فرکانسی شوند، قدرت یادگیری شبکه عصبی بالاتر می‌رود. در این صورت، این نگاشت، با خاصیت تونوتوبیک دستگاه شنوایی انسان که به معنای حساس‌بودن مکان‌های مختلف در مغز به فرکانس‌های متفاوت سیگنال گفتار ورودی است، همانگ خواهد بود. (چن و همکاران، ۲۰۰۵؛ شکفته و الماس‌گنج، ۱۳۸۶).

بعد از نگاشت اطلاعات ورودی به لایه نخست، در لایه‌های بالاتر شبکه 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 برآسas دینامیک‌های کوچک‌تر آشکار کنند. درنهایت، تمامی اطلاعات استخراج شده در لایه‌های پایینی با یکدیگر ترکیب شده و بعد از عبور از یک لایه غیرخطی به خروجی می‌رسند.

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^۱ خوانده می‌شود بر روی قسمت‌های مجازی از ورودی عمل می‌کند تا زیر تکلیفی را از تکلیفی که بر عهده شبکه اصلی است، انجام دهد.

شبکه عصبی پیمانه‌ای عمیق پایه طراحی شده 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 می‌باشد. این شبکه با لغزیدن بر روی زنجیره بردارهای بازنمایی یا همان زنجیره قاب‌های ورودی، آواز شناسایی شده برای بردار بازنمایی مرکزی زنجیره ورودی در هر لحظه را در خروجی می‌دهد. ساختار این شبکه عصبی برآسas شبکه‌های عصبی تأخیر زمانی^۲ (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 مکعب مستطیل نشان داده شده و شامل تعدادی نورون می‌باشد، یک میدان دریافت^۳ محدود دارد. 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 دارد.

طراحی ساختار این شبکه عصبی به این صورت، علاوه بر قابلیت رشدیافتمن آن، به دلایل زیر انجام شده است: برآسas 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 لایه پایین‌تر برای پردازش اطلاعات آن بخش برقرار می‌شوند. دوم این که یکسان‌بودن وزن‌های یک لایه به آشکارسازهای آن لایه کمک می‌کند تا به جستجوی یک ویژگی مشترک در

¹ Module

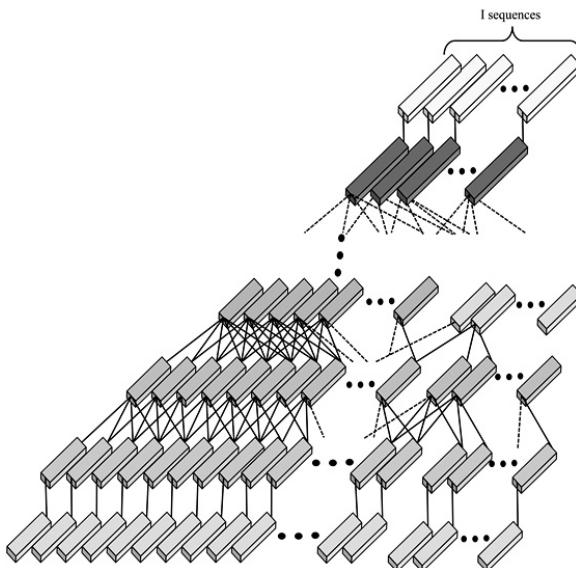
² Time Delay Neural Network (TDNN)

³ Receptive field



این شبکه بخواهد در هر لحظه یک توالی از قاب‌ها را بازشناسی کند، لازم است توسعه داده شود. با این کار، شبکه توسعه یافته، 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 مقایسه ساختار این شبکه با ساختار نشان داده شده 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): شبکه عصبی پیمانه‌ای با ساختار انجمن‌گر فضایی-زمانی دوگانه که از توسعه شبکه شکل (۲) بدست آمده است.

۴- تعلیم شبکه‌های عصبی پیمانه‌ای عمیق

در ادامه نحوه تعلیم شبکه‌های پیمانه‌ای عمیق معرفی شده که شامل شبکه‌های پیمانه‌ای پایه بازشناس آوا و ساختار توسعه یافته آنها به صورت ساختار انجمن‌گر فضایی-زمانی دوگانه می‌باشد، شرح داده می‌شود.

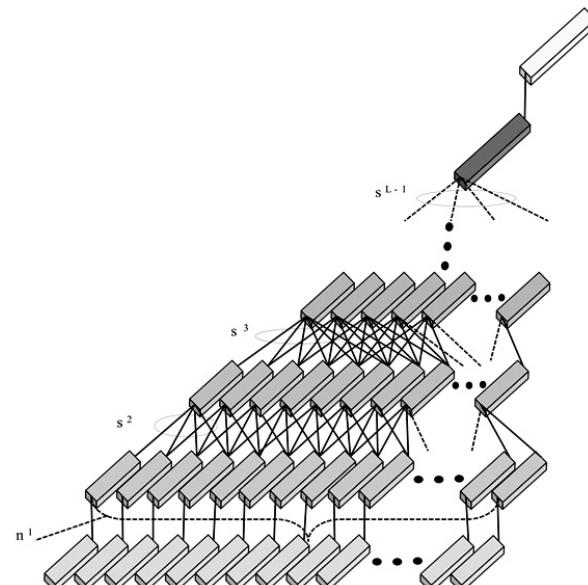
۴-۱- پیش تعلیم

روند پیش‌تعلیم معرفی شده همان‌طور که در شکل (۴) نشان داده شده است، به صورت لایه‌به‌لایه انجام می‌شود. هدف از به کار گرفتن این روند پیش‌تعلیم در این ساختار از شبکه‌های

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

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

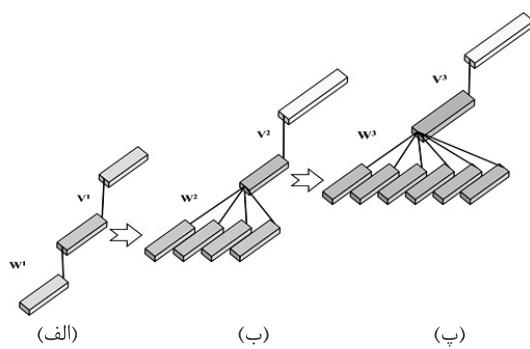
الگوریتم‌های تعلیم شبکه‌های عصبی که براساس اطلاعات تمایزی قاب‌ها طراحی شده‌اند، فقط از اطلاعات یک قاب و یا یک دسته از قاب‌ها برای بازشناسی آواز مربوط به قاب مرکزی ورودی استفاده می‌کنند. در این مقاله با توسعه ساختار شبکه عصبی پیمانه‌ای پایه معرفی شده در بخش قبل و با تعلیم مناسب آن، سعی می‌شود اطلاعات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



(شکل - ۲): ساختار پیمانه‌ای عمیق شبکه بازشناس آواز 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 وزن دار توالی‌های نورونی در لایه قبل به توالی‌های نورونی در لایه بعد است.

نحوه رشد این ساختار به این صورت است: شبکه بازشناس آواز 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

کنند. بنابراین، از یک ساختار شبکه عصبی خودانجمنی به صورت نشان داده شده در شکل ۴-۴(الف) استفاده می‌شود. در این شبکه، نگاشت ورودی به لایه پنهان توسط ماتریس وزن‌های W^1 و نگاشت لایه پنهان به خروجی توسط وزن‌های کمکی V^1 انجام می‌پذیرد. بعد از تعلیم این شبکه، لایه پنهان آن می‌تواند مؤلفه‌های فضای بازنمایی‌های ورودی را یاد بگیرد.



شکل-۴: روند پیش تعلیم شبکه عصبی با ساختار انجمان‌گر فضایی-زمانی دوگانه

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

$$E(W^l) = \frac{1}{2} \sum_{j=1}^J (d_j^l - z_j^l)^2 \quad (1)$$

در رابطه بالا / شماره لایه‌ای است که وزن مربوط به آن در هر مرحله بهینه می‌شود. برای محاسبه W^1 در رابطه بالا تعداد نورون‌های لایه ورودی و لایه خروجی شبکه خودانجمنی تک‌لایه‌ها پنهان است که برابر با تعداد پارامترهای هر بردار بازنمایی یا هر قاب است. بردار d^1 یا z^1 بردار بازسازی‌شده‌ای است که شبکه برای این بردار بازنمایی ساخته است.

مقدار W^1 محاسبه شده در این حالت، برای تمامی اتصالات مشابه در این لایه که همان نگاشت هر یک از بردارهای بازنمایی ورودی به فضای لایه پنهان نخست در ساختار شکل (۲) است، رونوشت می‌شود.

۴-۱-۳- پیش تعلیم وزن‌های مربوط به سایر لایه‌ها
برای پیش تعلیم وزن‌های لایه‌های پنهان بالاتر، همان‌طور که اشاره شد در هر مرحله شبکه‌های یک‌لایه پنهانی ساخته

عصبی این است که بتوان به صورت بهینه از اطلاعات دادگان تعلیمی در هر مرحله به منظور حفظ حداقل اطلاعات لازم برای بازشناسی استفاده کرد؛ لذا، برای تعلیم هر لایه، وزن‌های آن لایه به گونه‌ای تنظیم می‌شوند که گویی بازشناسی نهایی از روی خروجی آن به دست می‌آید. بنابراین، به صورت مرحله به مرحله وزن‌های مربوط به هر لایه با توجه به هدف بازشناسی مورد نظر به صورت مناسبی تعلیم می‌یابند. از وزن‌های به دست آمده در هر مرحله برای محاسبه ورودی لایه بعدی استفاده می‌شود. وزن‌ها باید به گونه‌ای محاسبه شوند تا حداقل اطلاعات تمایزی ورودی هر لایه را در جهت دست‌یافتن به خروجی مورد نظر استخراج کنند؛ اما در مورد پیش تعلیم وزن‌های لایه نخست شبکه‌های عصبی معرفی شده 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 به دست 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 به گونه‌ای تعلیم داده شوند که در تجزیه و تحلیل اطلاعات، اتلاف نداشته باشند. برای رسیدن به این هدف، لازم است ورودی‌های متمایز به صورت متمایز بیان شوند.

برای استفاده از روش پیش تعلیم لایه‌به‌لایه با توجه به ساختار شبکه عصبی بازشناس آواز قابل رشد معرفی شده با وزن‌های رونوشت شده در زمان و میدان دریافت محدود برای نورون‌های هر لایه، لازم است روند ویژه‌ای به کار گرفته شود. نحوه به کار گیری روش پیش تعلیم مورد نظر برای این ساختار از شبکه می‌تواند ایده‌ای برای چگونگی به کار گیری روش‌های پیش تعلیم مختلف برای شبکه‌های کانولوشنی باشد. در ادامه، نحوه پیاده‌سازی روش پیش تعلیم مورد استفاده آورده شده است.

۴-۱-۱- پیش تعلیم وزن‌های مربوط به لایه نخست
براساس آنچه در بخش پیش ارائه شد، وزن‌های لایه نخست (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 به صورت 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

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

۲-۴- تعلیم یک پارچه

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

$$E(\mathbf{W}) = \frac{1}{2} \sum_{i=1}^I \sum_{j=1}^J (d_{ij} - h_{ij}^l)^2 \quad (4)$$

در این رابطه \mathbf{W} تمامی وزن‌های مربوط به تمامی لایه‌ها در شبکه 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 انجمان گرفضایی-زمانی دوگانه برابر با تعداد قاب‌هایی است که به منظور تعلیم اطلاعات توالی آوازی آن قاب‌ها به شبکه در نظر گرفته شده است. h_{ij}^l خروجی نورون j از i نامین 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 خطأ طبق رابطه زیر و به صورت تکراری انجام بگیرد.

$$\Delta \mathbf{W}^l = -\gamma \frac{\partial E}{\partial \mathbf{W}^l} \quad (5)$$

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

می‌شود. ورودی این شبکه‌ها به کمک وزن‌های محاسبه شده در مراحل قبلی به دست می‌آید و خروجی آنها لایه بازشناسی برای بازشناسی بردار بازنمایی مرکزی ورودی آن شبکه در s هر زمان است. در این ساختار از شبکه‌ها، خروجی هر s آشکارساز در لایه پنهان $l-1$ (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 وزن‌های \mathbf{W}^l و \mathbf{V}^l و دوتابع غیرخطی در لایه پنهان و لایه خروجی، خروجی نهایی آنها محاسبه می‌شود. به عنوان 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 محاسبه وزن‌های مربوط به خروجی s^l توالی آشکارساز در لایه پنهان $l-1$ ($s^l > 1$) کافی است این وزن برای اتصالات مشابه رونوشت شود. در این صورت، خروجی مطلوب این شبکه، کد آوای مربوط به آشکارساز مرکزی دسته s^l تایی ورودی است که از روی برجسب آوایی دادگان به دست می‌آید. ماتریس وزن‌های \mathbf{W}^l نیز براساس تابع هزینه حداقل مربعات خطأ در رابطه (۱) و با استفاده از الگوریتم دلتا تعلیم داده می‌شوند. در این رابطه، I تعداد نورون‌های لایه بازشناسی خروجی شبکه‌های تک‌لایه پنهان است که بردار با تعداد آواها است. و بردار \mathbf{d}^l برجسب آوایی مربوط به بردار بازنمایی مرکزی ورودی شبکه است که به ازای آوای مربوطه مقدار یک و به ازای بقیه آواها مقدار صفر دارد. و \mathbf{z}^l خروجی‌ای است که شبکه برای این بردار بازنمایی مرکزی شناسایی کرده است که به صورت زیر محاسبه می‌شود.

$$\mathbf{z}^l = f \left(f \left(\left(\sum_{m=1}^{s^l} \mathbf{H}_m^{l-1} \mathbf{W}_m^l \right) - \mathbf{b}^l \right) \mathbf{V}^l - \mathbf{c}^l \right) \quad (2)$$

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

$$f(\theta) = \frac{1}{1 + e^{-\theta}} \quad (3)$$

پنهان پایین‌تر با یکدیگر جمع شوند. در این حالت، G_m^l طبق رابطه زیر محاسبه می‌شود:

$$G_m^l = \sum_{p=1}^{s^{l+1}} \sum_{q=1}^{n^{l+1}} (\text{Er}_q^{l+1} (W_p^{l+1})') \times \delta((m+1) - (p+q)) \quad (9)$$

خطای هر آشکارساز q در لایه بالاتر ($l+1$), به تعداد مشخصی واحد آشکارساز در لایه پایین‌تر (l) پس‌انتشار می‌شود. که این تعداد همان میدان دریافت آشکارسازهای می‌شود. W_p^{l+1} در این رابطه، وزن اتصالات لایه $l+1$ (s^{l+1}) است. W_p^{l+1} در این رابطه، وزن اتصالات مربوط به p امین آشکارساز در لایه l به هر آشکارساز در لایه $l+1$ است. چون وزن‌ها برای تمامی آشکارسازهای لایه $l+1$ می‌شوند. W_p^{l+1} به q وابسته نیست. (\cdot) در رابطه بالا، تابع دیریکله است که وقتی مقدار $p+q-1$ برابر با m می‌شود یک می‌دهد و در غیر این صورت خروجی آن صفر است. در شکل (۵) به صورت نمایشی طریقه پس‌انتشار خطای لایه خروجی به لایه قبل از آن رسم شده است. در این شکل میدان دریافت هر آشکارساز در لایه خروجی برابر ۳ است. همان‌طور که در این شکل مشاهده می‌شود، هم‌پوشانی ورودی‌های آشکارسازهای لایه 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.

$$\frac{\partial E}{\partial \mathbf{W}^l} = \frac{\sum_{m=1}^{n^l} [\mathbf{H}_m^{l-1}, \mathbf{H}_{m+1}^{l-1}, \dots, \mathbf{H}_{m+s^l}^{l-1}, 1]' \mathbf{E} \mathbf{r}_m^l}{n^l} \quad (6)$$

در مورد لایه 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.

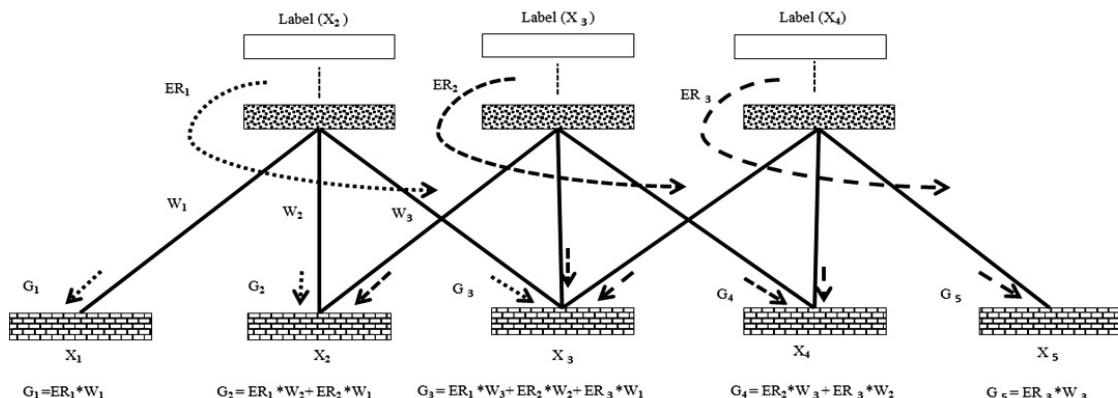
در این رابطه، $[\mathbf{H}_m^{l-1}, \mathbf{H}_{m+1}^{l-1}, \dots, \mathbf{H}_{m+s^l}^{l-1}, 1]$ بردارهای تشکیل شده از قراردادن خروجی‌های مربوط به s^l آشکارساز در لایه $l-1$ و نورون بایاس است که از طریق \mathbf{W}^l به هر یک از آشکارسازهای نورونی در لایه l متصل هستند. چون وزن \mathbf{W}^l برای تمامی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

$$\text{Er}_m^l = \mathbf{H}_m^l \cdot (\mathbf{1} - \mathbf{H}_m^l) \cdot (\mathbf{D}_m - \mathbf{H}_m^l) \quad (7)$$

در این رابطه، (\cdot) ضرب هر درایه از دو بردار با طول برابر را نشان می‌دهد. و $\mathbf{1}$ بیان‌کننده بردار با درایه‌های یک است. بر این اساس، خطای پس‌انتشار شده به لایه‌های پایین‌تر به صورت زیر به دست می‌آید.

$$\text{Er}_m^l = \mathbf{H}_m^l \cdot (\mathbf{1} - \mathbf{H}_m^l) \cdot \mathbf{G}_m^l \quad l < L \quad (8)$$

چون در این ساختار از شبکه عصبی، ورودی‌های آشکارسازهای لایه پنهان بالاتر با یکدیگر هم‌پوشانی دارند، بایستی خطاهای پس‌انتشار شده به هر آشکارساز در لایه



(شکل - ۵): طریقه پس‌انتشار خطای در شبکه پیمانه‌ای با ساختار کمی شده در زمان و میدان دریافت محدود برای آشکارسازهای هر لایه

(۱۰) در نظر گرفت. در حالی که برای تعلیم در سطح قاب، خطای پس‌انتشار شده به صورت $(\mathbf{X}(t) - \mathbf{h}^L) \{ \mathbf{X}(t) - \mathbf{h}^L \}$ می‌باشد.

در حالت کلی، روند تعلیم شبکه با ساختار فضایی-زمانی دوگانه را می‌توان به صورت پس‌انتشار خطای رابطه



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 بیان شده همان واج‌های نوشته شده هستند. به این ترتیب، یک برچسب‌دهی آوایی دیگر به‌ازای هر قاب گفتاری از این دادگان به‌دست می‌آید. در این روش برچسب‌دهی، اگر لغت \bar{u}_i به صورت دنباله واجی $\{ph_1, ph_2, \dots, ph_{\bar{l}_{ui}}\}$ که \bar{l}_{ui} تعداد واج‌های مربوط به این لغت است، بیان شده باشد و طول لغت $|u_i|$ باشد، دنباله برچسب آوایی آن به صورت رابطه (۱۱) به‌دست می‌آید. در این دنباله هر آوا به اندازه \bar{n}_{ph_m} قاب تکرار می‌شود.

$$\alpha_{u_i} = \{ph_1, \dots, ph_1, ph_2, \dots, ph_2, \dots, ph_{\bar{l}_{ui}}, \dots, ph_{\bar{l}_{ui}}\}$$

$$\bar{n}_{ph_m} = \frac{|ph_m|}{\sum_{j=1}^{\bar{l}_{ui}} |ph_j|} \times |u_i| \quad (11)$$

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

برچسب‌دهی براساس طول متوسط آواها (آرایه α)، تا حدودی آواهای درست یک لغت را در بر دارد و برچسب‌های به‌دست آمده از نتایج بازشناسی شبکه DST (آرایه β)، اطلاعات مناسب‌تری را در مورد نقطه گذر بین آواها به‌دست می‌دهد. بنابراین، برای اشتراک اطلاعات موجود و به‌دست آوردن برچسب‌های آوایی بهینه برای هر لغت، برچسب‌های به‌دست آمده در آرایه α با استفاده از برنامه‌ریزی پویا^۲ با برچسب‌های به‌دست آمده در آرایه β مقایسه می‌شوند و بهترین انطباق این دو آرایه به‌دست می‌آید.

نتیجه پیاده‌سازی برنامه‌ریزی پویا نشان می‌دهد که برچسب‌های موجود در آرایه α در چه زمان طول تخمین‌زده شده برای هر آوا را کمتر یا بیشتر از طول آوای اداشه پیش‌بینی کرده‌اند. بر این اساس، برچسب‌گذاری اولیه تصحیح شده و به عنوان برچسب‌های جدید برای لغات ذخیره می‌شوند. از برچسب‌های به‌دست آمده در این مرحله جهت تعلیم شبکه عصبی با ساختار DST برروی دادگان بزرگ استفاده می‌شود. بعد از چند تکرار از تعلیم شبکه براساس برچسب‌های به‌دست آمده، دوباره از شبکه

²Dynamic Programming (DP)

$$(\{D_1 - H_1^L, \dots, D_I - H_I^L\}|X(t: t+I)) \quad (10)$$

۳-۴- تعلیم نیمه‌سرپرستی شده شبکه با ساختار انجمان گرفضایی-زمانی دوگانه

در بخش مقدمه اشاره شد که برای استفاده از شبکه عصبی تعلیم‌یافته با ساختار انجمان گرفضایی-زمانی دوگانه (DST) در سامانه 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 به صورت واجی^۱ هستند، ارائه می‌شوند و فاقد برچسب‌های آوایی قاب‌ها هستند (ون باائل و همکاران، ۲۰۰۷)؛ لذا، در اینجا یک الگوریتم برای تعلیم نیمه‌سرپرستی شده این شبکه بر روی دادگان بزرگ معرفی می‌شود. این روش به مرزیندی تحمیلی گفتار توسط مدل‌های GMM-HMM از پیش تعلیم یافته بر روی این دادگان نیاز ندارد.

در شکل (۶) الگوریتم تعلیم نیمه‌سرپرستی شده 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 شبکه عصبی با ساختار DST بر روی دادگان کوچک‌تر به خوبی تعلیم داده می‌شود؛ سپس، از شبکه تعلیم‌یافته بر روی دادگان کوچک‌تر برای بازشناسی بردارهای بازنمایی بدون برچسب دادگان بزرگ و استخراج اطلاعات آوایی برای هر DST قاب استفاده می‌شود؛ لذا شبکه 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 بازشناسی شده توسط این شبکه برای هر قاب از نامین لغت ورودی در آرایه β_{u_i} قرار داده می‌شود؛ اما این شبکه اشتباهاتی نیز در بازشناسی دارد. بنابراین، استفاده از دانش شبکه عصبی تعلیم‌یافته بر روی دادگان کوچک‌تر به‌نهایی 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 ازاواها (ph_m) را از تقسیم تعداد قاب‌های اختصاص‌داده شده به آن آوا به تعداد کل قاب‌های این دادگان به‌دست آورد؛ سپس، براساس زمان شروع و پایان هر لغت در دادگان بزرگ و

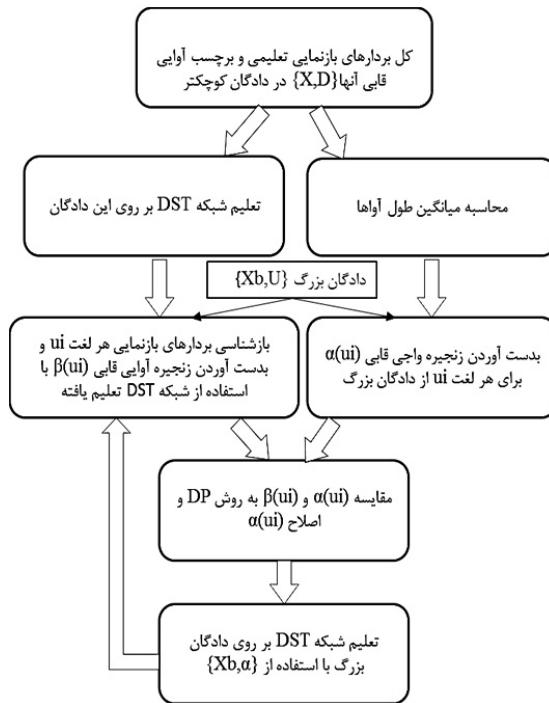
¹Phonemic transcription

طبقه‌بندی کننده‌های آماری مثل HMM‌ها به منظور 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 حجمی هستند و هم به میزان داده تعلیم زیادی نیاز دارند استفاده کنند. به این دلیل، با اعمال تبدیل کسینوسی گستته^۲ (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 LHCN 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 طریق 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 بازنمایی آوای گزارش شده در این مقاله، از الگوریتم ویتری براي رمزگشایی و محاسبه دقت بازنمایی 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 بازنمایی آوا به ازای بردار بازنمایی مرکزی وروی در آن زمان محاسبه می‌شود. با مقیاس کردن این احتمالات، مقدار درستنمایی انتشاری^۳ هر حالت به شبکه استفاده در الگوریتم ویتری به دست می‌آید. پارامترهای انتقال بین این آواها نیز از مدل بایکرم آوایی^۴ که در مورد دادگان فارس‌داد و فارس‌داد بزرگ به ترتیب از روی برچسب‌های آوایی کل دادگان فارس‌داد و برچسب‌های واجی کل دادگان فارس‌داد بزرگ و با استفاده از جعبه ابزار SRILM محاسبه می‌شود، به دست می‌آید. با توجه به اینکه در تعلیم شبکه عصبی با ساختار انجمان گر فضایی-زمانی

تعلیم یافته تا این مرحله برای بازنمایی بردارهای بازنمایی دادگان بزرگ و به دست آوردن آرایه برچسب آوایی جدید استفاده می‌شود. این روند تا هم‌گرایی تعلیم شبکه ادامه پیدا می‌کند.



(شکل - ۶): الگوریتم تعلیم نیمه‌سرپرستی شده برای شبکه 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 فارس‌داد و فارس‌داد بزرگ (بیجن خان و همکاران، ۱۹۹۴؛ شیخزادگان و بیجن خان، ۱۳۸۵) استفاده شد.

در این مقاله از بردارهای بازنمایی^۱ LHCN 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 به کار رفته در بازنمایی گفتار در مقایسه با بردارهای بازنمایی MFCC که به طور معمول برای بازنمایی گفتار استفاده می‌شوند، مناسب‌تر هستند.

^۱Discrete Cosine Transform (DCT)

^۲Emission likelihood

^۳Phonetic bigram

^۴Logarithm of Hanning Critical Band filter banks (LHCN)



انتخاب شد. انتخاب گویندگان مربوط به هر دو دسته به صورت تصادفی انجام شده است. هر کدام از ۳۰۴ گوینده تعداد ۲۰ جمله را در دو جلسه بیان کرده اند. بنابراین، در مجموع، ۵۹۴۰ جمله برای تعلیم شبکه و ۱۴۰ جمله برای آزمون شبکه مورد استفاده قرار گرفته شد. گفتارهای ضبط شده به صورت دستی برای هر لغت و آواز جداسازی و برچسب دهی شده اند.

۱-۱-۵- پیاده سازی ساختار شبکه عصبی پیمانه ای و رشد آن به صورت ساختار DST بر روی این دادگان

ابتدا ساختار شبکه عصبی پیمانه ای عمیق بازشناس آواز پایه نشان داده شده در شکل (۲) با خصوصیات آورده شده در جدول (۱) طراحی شد.

(جدول-۱): خصوصیات ساختاری شبکه عصبی پیمانه ای پایه

	تعداد آشکار سازها	تعداد هر آشکار ساز	تعداد گره ها در زمانی هر پیمانه	تعداد تأخیر	لایه
-	۱۸	۲۲			وروودی
۱	۶۴	۲۲			لایه نخست
۹	۵۱۲	۱۵			لایه دوم
۱۵	۶۲	۱			لایه سوم
۱	۳۶	۱			لایه خروجی

طراحی این ساختار با توجه به تجربیات به دست آمده از پژوهش های پیشین بر روی دادگان فارس دات انجام شده است. براساس این پژوهش ها، در (کرمی، ۱۳۷۹) نشان داده شده است که برای فراهم آوردن اطلاعاتی در مورد محتوای صوتی هر قاب برای یک شبکه بازشناس آواز، یک پنجره مستشكل از ۲۵-۷ قاب باید به عنوان ورودی آن در نظر گرفته شود. هم چنین، با مقایسه نتایج بازشناسی آواز به دست آمده در (انصاری، ۱۳۸۲) با نتایج (نژادقلی و سید صالحی، ۲۰۰۹) می توان به این نتیجه رسید که یک شبکه TDNN که لایه پنهان نخست آن به صورت نیمه متصل با لایه ورودی ارتباط دارد سریع تر از نمونه تمام متصل آن هم گرا می شود، بدون آنکه کارایی خود را از دست داده باشد. از آنجایی که بردارهای ویژگی استخراج شده از سیگنال گفتاری دارای طول ۱۸ برای هر قاب می باشند، در این مقاله نورون های بیشتری از بعد هر یک از بردارهای ویژگی در لایه پنهان نخست در نظر گرفته شده اند تا بتوانند ویژگی های مختلفی را از هر قاب گفتاری آشکار کنند. در (نژادقلی و سید صالحی، ۹۲۰۰؛ توحیدی پور و همکاران، ۲۰۱۲) نورون برای پردازش هر قاب در نظر گرفته شده اند؛ اما از آنجاکه دادگان

دو گانه اطلاعات توالی خروجی به کار گرفته شد، انتظار می رود وزن های شبکه در حین تعلیم اطلاعاتی در مورد توالی آواها بدست آورده باشند. بنابراین، برای بررسی اینکه به چه میزان شبکه توانسته است اطلاعات توالی آوازی را استخراج کند، ابتدا در الگوریتم ویتری بورد استفاده برای گذر بین آواهای مختلف مقدار احتمال صفر در نظر گرفته شد. که معادل با عدم استفاده از احتمالات مدل زبانی با یک گرم آوازی در سامانه بازشناسی گفتار است. سپس، جهت مقایسه بهترین نتایج بازشناسی این مدل در مقایسه با مدل شبکه عصبی پیمانه ای پایه و با مدل HMM، مدل زبانی با یک گرم آوازی از روی دادگان تخمین زده و در محاسبات ویتری دخالت داده شد. نتیجه کد گشایی توسط الگوریتم ویتری با اعمال پارامتر جریمه درج بهبود می یابد. این پارامتر می تواند از تغییر سریع خروجی الگوریتم بین آواهای مختلف جلوگیری کند. این پارامتر به صورت زیر با اصلاح پارامترهای گذر بین آواها در الگوریتم ویتری مورد استفاده قرار می گیرد.

$$\tilde{a}_{rt} = GSF * a_{rt} + IP \quad (12)$$

در این رابطه، a_{rt} پارامتر گذر بین آواز r به آواز t است که با استفاده از مدل با یک گرم آوازی جایگذاری می شود. GSF ضریب تاثیر مدل با یک گرم آوازی در الگوریتم ویتری است که برای عدم استفاده از این مدل، مقدار این پارامتر صفر می شود. و IP پارامتر جریمه درج است. در ادامه به توضیح هر یک از دادگان و ارائه نتایج آزمایش های انجام شده با استفاده از هر یک از آنها پرداخته می شود. در صد دقت بازشناسی آواز نهایی طبق رابطه زیر و با استفاده از معیار NIST محاسبه می شود:

$$Phone Recognition Rate = \frac{K - Del - Ins - Sub}{K} \quad (13)$$

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

۱-۵- دادگان گفتاری فارس دات

این مجموعه شامل سیگنال گفتاری پیوسته و تمیز از ۳۰۴ گوینده زن و مرد است که در سن، لهجه و میزان تحصیلات با یکدیگر متفاوتند (بیجن خان و همکاران، ۱۹۹۴). این دادگان به طور تقریبی 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

در ادامه، ساختار شبکه عصبی پیمانه‌ای پیشنهاد شده در جدول (۱) تعداد پیمانه‌های خود در هر لایه را به صورتی رشد می‌دهد که خروجی آن هم‌زمان بتواند پانزده توالی قابی آوایی را بینند. مشخصات ساختاری شبکه توسعه یافته به صورت ساختار انجمن گرفضایی-زمانی دوگانه در جدول زیر آمده است.

(جدول-۲): خصوصیات ساختاری شبکه عصبی پیمانه‌ای با ساختار انجمن گرفضایی-زمانی دوگانه

تعداد تأخیر زمانی هر پیمانه	تعداد گره‌ها در هر آشکارساز	تعداد آشکارسازها	لایه
-	۱۸	۳۷	لایه ورودی
۱	۶۴	۳۷	لایه نخست
۹	۵۱۲	۲۹	لایه دوم
۱۵	۶۲	۱۵	لایه سوم
۱	۴۶	۱۵	لایه خروجی

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

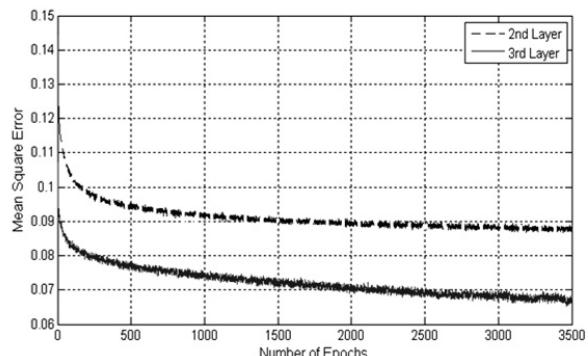
به منظور رسیدادن شبکه پیمانه‌ای پایه تا در نظر گرفتن پانزده توالی در خروجی، زنجیره‌ای از ۳۷ قاب گفتاری به ورودی اعمال می‌شود. در این حالت تعداد آشکارسازها در لایه پنهان نخست به ۳۷ آشکارساز افزایش می‌یابد؛ سپس، در لایه پنهان دوم به منظور پردازش اطلاعات استخراج شده از هر توالی متشکل از ۹ آشکارساز در لایه پایین‌تر، ۲۳ توالی نورونی در این لایه در نظر گرفته شده است. در لایه بعدی، اطلاعات هر پانزده توالی از خروجی نورون‌ها در لایه پنهان پایینی توسعه 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 ۴۶ نورون در نظر گرفته شده است.

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

چون توسعه ساختار شبکه بازشناس آوای پایه به ساختار DST با هدف قراردادن اطلاعات توالی آوایی قاب‌ها

تعلیمی به کار گرفته شده در این مقاله بیشتر است، تعداد نورون‌ها به ۶۴ نورون افزایش داده شد. علاوه‌بر آن، این شبکه به صورتی طراحی شده است که لایه‌های پایین‌تر ویژگی‌های پویایی کوتاه‌تر و لایه‌های بالاتر ویژگی‌های پویایی بلندتر را کشف کنند. بنابراین، اطلاعات استخراج شده از ۹ قاب متولی در لایه پنهان نخست توسط یک سری از نورون‌ها (۵۱۲ نورون) در لایه پنهان دوم پردازش می‌شود. پردازش اطلاعات این پویایی‌های کوتاه به منظور آشکار کردن اطلاعات اصوات متمایز‌گر که در بازشناسی آواز مؤثر است به کار می‌رود. از این‌رو پانزده توالی از آشکارسازها در لایه پنهان دوم در نظر گرفته شده‌اند تا اطلاعات مربوط به توالی‌های کوتاه از نورون‌های لایه پنهان نخست را پردازش کنند. سپس، ویژگی‌های استخراج شده در این لایه توسط ۶۲ نورون در لایه پنهان سوم پردازش می‌شوند. در این لایه پویایی‌های بلندتر با پنهانی پانزده توالی از آشکارسازهای لایه پنهان پایین‌تر پردازش می‌شوند. و در نهایت بازشناسی آواز در لایه خروجی به دست می‌آید.

این شبکه عصبی طبق روند اشاره شده در زیربخش (۱-۴) به صورت لایه‌به‌لایه پیش‌علیم داده شد؛ سپس، به اندازه هزار تکرار، با استفاده از روابط اشاره شده در زیربخش (۲-۴) و با در نظر گرفتن $I=1$ تعلیم سرتاسری یافت. در شکل (۷) نمودار کاهش خطای تعلیم شبکه‌های یک لایه 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 خطای تعلیم وزن‌های لایه سوم پایین‌تر از مقدار آن در تعلیم وزن‌های لایه دوم است. دلیل آن را می‌توان در اطلاعات چکیده‌تر و مفیدتر به دست آمده در لایه سوم به منظور بازشناسی نهایی دانست.



شکل (۷): نمودار کاهش خطای تعلیم برای تعلیم دو شبکه یک لایه پنهان ساخته شده برای پیش‌علیم وزن‌های لایه‌های دوم و سوم از شبکه عصبی پیمانه‌ای عمیق

همان طور که در این جدول مشاهده می شود، توسعه ساختار شبکه عصبی پیمانه ای به صورت ساختار انجمان گر فضایی- زمانی دوگانه توانسته است، درصد بازنگشتنی آوا را بهبود دهد. مدل های صوتی مثل HMM یا شبکه عصبی پیمانه ای عمیق پایه به گونه ای تعلیم یافته اند که هر بار یک زنگیره در فضای ورودی را به یک نقطه در فضای آواهای خروجی نگاشت دهنند. بنابراین، در زمان آزمون، هر زنگیره بردار ویژگی ورودی به نقطه ای در فضای خروجی نگاشته می شود که می تواند به زنگیره های مختلفی از آواها در خروجی تعلق داشته باشد؛ اما اگر در تعلیم مدل صوتی، زنگیره معتبر خروجی نیز به آن تعلیم داده شود، خاصیت انجمان بودن زنگیره الگوهای خروجی به میدان می آید. درنتیجه، همان طور که در شکل (۹-ب) نشان داده شده است، می توان به کمک اطلاعات آواهای اطراف به طبقه بندی صحیح تری از آوای وسط در خروجی دست یافت و خروجی شبکه به طبقه ای که دنباله خروجی به دست آمده در ازای آن معتبر نیست، برده نمی شود؛ که این حالت در تعلیم شبکه عصبی GMM-HMM اعمال شده است. هم چنان، ساختار شبکه عصبی انجمان گر فضایی-زمانی دوگانه توانسته است تا حدی نقش مدل سازی آوازی را نیز اجرا کند؛ زیرا بدون استفاده از اطلاعات مدل آوازی در رمزگشایی، درصد بازنگشتنی قابل قبولی به دست داده است. هم چنان، با توجه به این جدول مشاهده می شود که شبکه عصبی پیمانه ای پایه در مقایسه با مدل GMM-HMM توانسته است در صورت استفاده از مدل بایگرم آوازی، بهتر عمل کند؛ دلیل آن، طراحی دقیق تر ساختار این شبکه و توانایی استخراج اطلاعات مفید تر از دادگان ورودی از طریق تعلیم مناسب آن است.

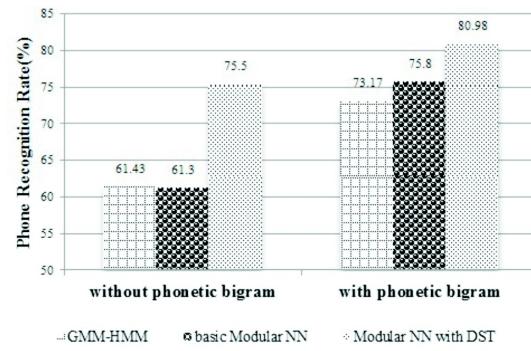
۲-۵- دادگان گفتاری فارس دات بزرگ

این دادگان شامل سیگنال گفتار مربوط به یکصد گوینده با تفاوت های سن، لهجه و تحصیلات است. در این دادگان، هر گوینده ۲۵ صفحه از روزنامه با موضوعات مختلف را می خواند. در کل، هر فرد حدود چهار هزار لغت از متون مختلف روزنامه ای 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 در بردارنده حدود ۵۷ ساعت گفتار است به عنوان داده تعلیم و گفتار مربوط به پنج نفر دیگر که شامل حدود سه ساعت گفتار است برای

در وزن های آن است، برای استفاده در آزمون دیگر نیازی به وجود چندین دسته نورونی در خروجی این شبکه عصبی نیست. بنابراین، با برداشت پیمانه های اضافه شده در تعلم به منظور در نظر گرفتن توالی آوابی در خروجی، ساختار شبکه DST در آزمون به ساختار شبکه عصبی پیمانه ای پایه کاوش یافت؛ اما با این تفاوت که وزن های این شبکه شامل اطلاعات توالی آوابی است.

جهت مقایسه عملکرد شبکه های عصبی بازنگشتنی آوازی پیمانه ای معرفی شده با مدل های احتمالاتی، یک مدل GMM-HMM که مجموعه ای از مدل های HMM چند حالت ساخته شده برای هر آوا است، با استفاده از معیار بیشینه درست نمایی^۱ تعلیم داده شد. در اینجا برای هر حالت از مدل های HMM از مدل های GMM با هشت گوسین برای دست یافتن به بهترین نتیجه بازنگشتنی استفاده شد. به دلیل فرض استقلال شرطی در مدل های HMM، با هدف بالا بردن کارآیی این مدل ها لازم است از بازنگشتنی های مستقل شده به عنوان ورودی مدل استفاده کرد. به علاوه، استفاده از مشتقات زمانی مرتبه نخست و دوم این بازنگشتنی ها 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 بازنگشتنی آوازی به دست آمده از شبکه تعلیم یافته با استفاده از ساختار DST در مقابل شبکه عصبی عمیق پیمانه ای اولیه و مدل GMM- HMM نشان داده شده است.



(شکل-۸): درصد دقت بازنگشتنی آوازی به دست آمده توسط شبکه عصبی پیمانه ای با ساختار DST. شبکه عصبی پیمانه ای اولیه و مدل GMM-HMM، محاسبه شده توسط الگوریتم ویترینی در دو حالت استفاده از مدل بایگرم آوازی و بدون آن برای دادگان فارس دات

^۱ Maximum Likelihood

وجود دارد که مجبور به پذیرش این خطای برچسبدهی هستیم؛ ولی خطاهای حذف و درج که بیشتر به دلیل عدم مرزبندی صحیح برچسبها است تا حد خوبی اصلاح می‌شوند. این روند تعلیم شبکه براساس برچسبهای به دست آمده و استفاده از آن برای تولید برچسب جدید دوبار تکرار می‌شود.

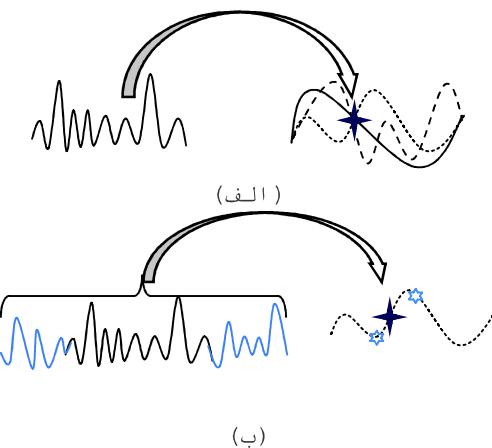
در جدول (۳) درصد بازنگاری واژه به دست آمده توسط شبکه DST تعلیم یافته بر روی دادگان فارس دات بزرگ به روش نیمه سرپرستی شده نشان داده شده است. جهت بررسی اثر روش برچسبدهی ارائه شده، در آزمایش دیگری، تعلیم شبکه DST بر روی دادگان بزرگ فقط با استفاده از برچسبهای به دست آمده از میانگین طول آواها انجام گرفت. همچنین نتایج به دست آمده با نتیجه بازنگاری واژه مدل HMM تعلیم یافته بر روی این دادگان در جدول زیر مقایسه شده است.

(جدول-۳): درصد دقت بازنگاری واژه بر روی دادگان فارس دات بزرگ با استفاده از شبکه عصبی با ساختار فضایی-زمانی

تعلیم یافته به صورت نیمه سرپرستی شده در مقایسه با مدل HMM	
مدل بازنگاری آوا + روش تعلیم بر روی دادگان بازنگاری واژه	دادگان بزرگ
۵۳	مدل DST تعلیم یافته بر روی دادگان فارس دات
۶۴	مدل DST + تعلیم نیمه سرپرستی شده بر اساس میانگین طول آواها
۶۹/۶	مدل DST + تعلیم نیمه سرپرستی شده بر اساس ترکیب اطلاعات میانگین طول آواها و نتایج بازنگاری شبکه
۷۳	مدل HMM + بیشترین درست نمایی

با مقایسه نتایج به دست آمده مشاهده می‌شود که با استفاده از روش نیمه سرپرستی شده 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 بازنگاری واژه به دست آمده با به کارگیری مدل HMM بالاتر از درصد بازنگاری به دست آمده توسط مدل شبکه عصبی با ساختار DST است. دلیل این تفاوت را می‌توان به این خاصیت مدل های HMM نسبت داد که برخلاف شبکه های عصبی، قابلیت تعلیم بر روی دادگانی را که برچسبهای زمانی ندارند، دارند؛ لذا با تعلیم اولیه مدل های HMM برای هر آوا بر روی دادگان

آزمون مورد استفاده قرار گرفت. در ادامه به توصیف آزمایش های انجام شده با استفاده از این دادگان پرداخته می شود.



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

۱-۲-۵- تعلیم نیمه سرپرستی شده شبکه عصبی با ساختار انجمان گر فضایی-زمانی دوگانه بر روی فارس دات بزرگ

با توجه به آنچه در زیربخش (۳-۴) اشاره شد، برای تنظیم شبکه عصبی پیمانه ای عمیق با ساختار انجمان گر فضایی-زمانی دوگانه بر روی دادگان بزرگ از شبکه عصبی تعلیم یافته بر روی دادگان فارس دات استفاده می شود. با توجه به وجود برچسبهای زمانی برای هر لغت در دادگان بزرگ، شبکه عصبی DST در طول لغت حرکت می کند و دنباله آوایی β_{u_i} نشان داده شده در شکل (۶) را ایجاد می کند. همچنین، دنباله آوایی α_{u_i} براساس میانگین طول آواهای محاسبه شده از روی دادگان فارس دات و حاشیه نویسی های واجی هر لغت در فارس دات بزرگ به دست می آید؛ سپس، الگوریتم برنامه ریزی پویا برای مقایسه α_{u_i} با β_{u_i} و اصلاح آن براساس نتایج به دست آمده به کار می رود.

نکته قابل توجه این است که برای تعلیم شبکه عصبی پیمانه ای بر روی دادگان فارس دات از برچسبهای آوایی بهره گرفته شده است. بنابراین، نتایج بازنگاری شبکه عصبی به صورت آوایی است؛ اما برچسبدهی براساس میانگین طول آواها براساس حاشیه نویسی های واجی لغات انجام گرفته است. بنابراین، اختلافاتی بین آرایه های به دست آمده از دو روش





ورودی را مورد بررسی قرار داده و ورودی را به یکی از توالی های خروجی یاد گرفته شده می برد. در این حالت، در مقایسه با زمانی که شبکه فقط یک طبقه مربوط به هر الگوی ورودی را یاد می گیرد، اطلاعات توالی کمک می کند نمونه هایی که به صورت موضعی نویشهای شده اند به طبقه غلطی برده نشوند؛ بلکه اطلاعات مربوط به توالی خروجی ها که در ساختار شبکه شکل گرفته و یادگیری شده آنها را به طبقه ای که زنجیره آواهای مربوط به آن در زمان تعلیم به شبکه تعلیم داده شده است، می برد و درنتیجه به بازنگاری صحیح تر خروجی منجر می شود.

همان طور که در جدول (۳) نشان داده شده است، شبکه عصبی با ساختار DST با تعلیم نیمه سرپرستی شده بر روی دادگان فارس دات بزرگ توانست به درصد بازنگاری واج قابل مقایسه ای با نتیجه بازنگاری مدل HMM دست یابد. با وجود آنکه مدل های HMM برای هر آواز قابلیت تعلیم بر روی دادگانی را که برچسب های زمانی ندارند دارند؛ اما در فرآیند تعلیم این مدل ها وجود یک متخصص چه در مرحله تولید واژه نامه های تلفظی لغات و چه در مرحله گره زدن مدل های وابسته به محظوظ بسیار مؤثر است. درصورتی که با تعلیم نیمه سرپرستی شده شبکه عصبی بر روی دادگان بزرگ بدون استفاده از اطلاعات مدل HMM و در نتیجه به کار گرفتن دانش یک متخصص می توان به درصد بازنگاری قابل مقایسه ای با مدل HMM دست یافت.

تعلیم نیمه سرپرستی شده این شبکه تا این حد برای استفاده به عنوان شبکه پایه ای برای توسعه یافتن به منظور بازنگاری لغت می تواند کافی باشد؛ زیرا اطلاعات سطح بالاتر واژگانی در شبکه توسعه یافته می تواند برای تنظیم دقیق تر وزن های شبکه بازنگاری واج مورد استفاده قرار گیرد. این نتیجه با چگونگی بازنگاری واج در انسان که اطلاعات واژگانی در سیاری از موارد به کمک بازنگاری واج می آیند، هماهنگ است (ساموئل، ۲۰۱۱).

در ادامه کار، این ساختار از شبکه های عصبی طوری رشد داده می شوند که بتوانند به صورت یک پارچه و بدون نیاز به واژگان و مدل زبانی به بازنگاری لغت بپردازند. در این راستا لازم است شبکه عصبی طراحی شده بر روی کل دنباله واژی مربوط به هر گفته از گفتار هر فرد تعلیم داده شود تا بتواند اطلاعات صوتی و زبانی لغات را به صورت یکجا یاد بگیرد. برای رسیدن به این هدف علاوه بر عمیق تر کردن شبکه موجود لازم است اتصالات بازگشتی نیز در ساختار آن در نظر گرفته شود تا بتواند زنجیره های واژی طولانی مدت

فارس دات، ادامه تعلیم آنها بر روی دادگان بزرگ با مشکلی مواجه نمی شود. به طوری که با دادن اطلاعات مربوط به دنباله لغات و فرهنگ واژگان تلفظ آنها بدون آنکه زمان بندی لغات مشخص باشد می توان پارامترهای مدل های HMM را باز تخمین کرد.

۶- بحث و نتیجه گیری

در این مقاله، یک ساختار شبکه عصبی عمیق پیمانه ای قابل رشد به منظور بازنگاری آواز معرفی شد. این شبکه علاوه بر هماهنگی با ساختار پردازش سلسله مراتبی گفتار در مغز انسان، می تواند جهت پردازش اطلاعات پیچیده تر رشد یافته و عمیق تر شود؛ که این خاصیت نیز با ساختار لازم برای پردازش اطلاعات پیچیده در مغز هماهنگ است. در طراحی ساختار این شبکه برخلاف شبکه های پیمانه ای متداول که هر پیمانه به صورت مستقل عمل می کند، پیمانه ها در تعامل با یکدیگر می باشند. تعامل پیمانه ها، امکان پردازش اطلاعات دینامیک گفتار در طول سلسله مراتب را برای استخراج اطلاعات غنی تر به وجود می آورد. درواقع، این شبکه عصبی به دلیل ساختار طراحی شده خود که مبتنی بر اطلاعات گفتاری است، با استخراج الگوهای زمانی بهتر از ورودی در لایه نخست و سپس ترکیب این الگوهای با هدف استخراج اطلاعات دینامیک به ترتیب از ساده ترین تا پیچیده ترین آنها توانسته است اطلاعات بیشتری در مورد قابه های ورودی به دست آورد.

شبکه عصبی پیمانه ای رشد یافته به صورت ساختار انجمن گر فضایی - زمانی دوگانه همان طور که در شکل (۸) نشان داده شده است می تواند به درصد بازنگاری واج بالاتری در مقایسه با مدل HMM برسد. بالابودن کارایی این شبکه به توانایی آن در پالایش کردن توالی های آوایی نامعتبر در لایه خروجی مربوط می شود. پالایش زنجیره های خروجی غلط توسط این شبکه بر اساس این فرضیه که شبکه های عصبی با توجه به ظرفیت خود مؤلفه های غیرخطی فضای دادگان تعلیمی را یاد می گیرند و نمونه های دیگر را به کمک مؤلفه های یادگرفته شده به فضای دانش خود نگاشت می دهند استوار است. شبکه عصبی انجمن گر فضایی - زمانی دوگانه به منظور دخالت دادن مؤلفه های مربوط به ارتباطات انجمنی الگوها در فضای ورودی با فضای خروجی در عملکرد شبکه طراحی شده است. در این حالت، شبکه بر اساس مؤلفه های مختلفی که در تعلیم یادگرفته است، هر توالی

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

۱۳۷۹

یزدیان، م.، طراحی مدل بازناسن گفتار توسط شبکه‌های عصبی برپایه پردازش وقایع گستته 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., Sheikhdadegan, 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-dependent pre-trained deep neural networks for large-vocabulary speech recognition.Audio, Speech,

را مدل کند. علاوه بر آن لازم است توابع هدف مناسبی برای تعلیم سرتاسری شبکه‌های رشدیابته بهمنظور افزایش درصد بازناسنی لغت تولیدشده تعریف شود.

تشکر و قدردانی

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

۷- مراجع

انصاری، ل.. مدل‌سازی اثرات هم تولیدی آواها در یک مدل شبکه عصبی بازناخت 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 بازناسن گفتار، ۱۳۸۲، ش. ۵۵

سیدصالحی، س.ز.. سیدصالحی، س.ع.. روش پیش تعلیم سریع بر مبنای کمینه سازی خطاب 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.. بازناخت 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 بازناسن گفتار بر اساس شبکه عصبی به منظور بازناسنی دادگان گفتاری فارسی با تعداد لغات زیاد، گزارش پژوهشی، پژوهشکده پژوهش هوشمند علائم، ۱۳۸۴

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

شيخ زادگان، ج.. بیجن خان، م.. دادگان‌های گفتاری زبان فارسی، دومین کارگاه پژوهشی زبان فارسی و رایانه، ۱۳۸۵، ص. ۲۴۱-۲۴۷

فصل نهم





- 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
- Tohidpour, 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. In Proc. 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.



زهرا 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 دکترا در بیوالکتریک در دانشگاه صنعتی امیرکبیر است. زمینه‌های پژوهشی مورد علاقه ایشان پردازش و بازناسی گفتار، شناسایی الگو، شبکه‌های عصبی مصنوعی و یادگیری عمیق می‌باشد.

نشانی رایانامه ایشان عبارت است از:

z_ansari@aut.ac.ir



سید علی سید صالحی 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 دکترا 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 بازناسی گفتار، شبکه‌های عصبی مصنوعی و زیستی، مدل‌سازی عملکرد مغز و پردازش خطی و غیرخطی سیگنال است.

نشانی رایانامه ایشان عبارت است از:

ssalehi@aut.ac.ir

فصل نهم