استخراج ویژگی‌های مقاوم گفتاری زیر باندی با استفاده از شبکه‌های درهم‌پیچش چند دقتی

نوید نادری[۱]، کارشناس ارشد؛ بابک ناصرشریف[۲]، استادیار

۱- دانشکده مهندسی کامپیوتر - دانشگاه صنعتی خواجه‌نصیرالدین طوسی - تهران - ایران- navid.naderi@email.kntu.ac.ir
۲- دانشکده مهندسی کامپیوتر - دانشگاه صنعتی خواجه‌نصیرالدین طوسی - تهران- ایران- bnasersharif@kntu.ac.ir

چکیده: شبکه‌های عصبی درهم‌پیچش (CNN) به‌عنوان گروهی از شبکه‌های عصبی عمیق، در سال‌های اخیر کاربرد فراوانی در مدل‌سازی آکوستیک و همچنین استخراج ویژگی و مدل‌سازی توأم در بازشناسی گفتار یافته‌اند. در مقاله حاضر، پیشنهاد می‌شود تا از CNN برای استخراج ویژگی مقاوم به نویز استفاده شود، درحالی‌که ورودی CNN طیف سیگنال گفتار نویزی و خروجی هدف آن خروجی‌های متناظر تمیز از بانک فیلتر مل است. به‌این‌ترتیب CNN ویژگی‌های مقاوم به نویز را از طیف سیگنال گفتار استخراج می‌نماید. نقطه‌ضعف CNN در این روش آن است که تنها یک وضوح فرکانسی ثابت را به کار می‌گیرد. ازاین‌جهت، در این مقاله استفاده از چند شبکه عصبی درهم‌پیچش با اندازه‌های فیلتر درهم‌پیچش متفاوت، جهت مدل‌سازی تفاوت وضوح فرکانسی برای استخراج ویژگی از طیف سیگنال گفتار پیشنهاد می‌شود. روش پیشنهادی را شبکه عصبی درهم‌پیچش چند دقتی (MRCNN) نام‌گذاری کرده‌ایم. آزمایش‌ها روی دادگان Aurora2 نشان می‌دهند که CNN نسبت به شبکه باور عمیق در استخراج ویژگی مقاوم به نویز میانگین دقت بازشناسی را ۲۰ درصد بهبود می‌دهد. همچنین نتایج نشان می‌دهند که MRCNN میانگین دقت بازشناسی را نسبت به شبکه عصبی درهم‌پیچش استاندارد (تک دقتی) ۱ درصد بهبود می‌دهد.

واژه‌های کلیدی: شبکه عصبی درهم پیچش، بازشناسی مقاوم گفتار، تک دقتی، چند دقتی، بانک فیلتر مل.

Robust sub-band speech feature extraction using multiresolution convolutional neural networks

Navid Naderi[1], MSc; Babak Nasersharif[2], Assistant Professor

1-Computer Engineering Department, K.N.Toosi University of Technology, Tehran, Iran, Email: navid.naderi@email.kntu.ac.ir
2- Computer Engineering Department, K.N.Toosi University of Technology, Tehran, Iran, Email: bnasersharif@kntu.ac.ir

Abstract: Convolutional neural networks (CNNs), as a kind of deep neural networks, have been recently used for acoustic modeling and feature extraction along with acoustic modeling in speech recognition systems. In this paper, we propose to use CNN for robust feature extraction from the noisy speech spectrum. In the proposed manner, CNN inputs are noisy speech spectrum and its targets are denoised logarithm of Mel filter bank energies (LMFBs). Consequently, CNN extracts robust features from speech spectrum. The drawback of CNN in the proposed method is its fixed frequency resolution. Thus, we propose to use multiple CNNs with different convolution filter sizes to provide different frequency resolutions for feature extraction from the speech spectrum. We named this method as Multiresolution CNN (MRCNN). Recognition accuracy on Aurora 2 database, shows that CNNs outperform deep belief networks such that, CNN recognition accuracy has 20% relative improvement on average over DBN. However, results show that MRCNN recognition accuracy has 1% relative improvement on average over CNN.

Keywords: Convolutional neural network, Robust speech recognition, Single resolution, Multi-resolution, Mel filter bank.

تاریخ ارسال مقاله: ۱۳۹۶/۰۳/۱۰
تاریخ اصلاح مقاله: ۱۳۹۶/۰۵/۲۳، ۱۳۹۶/۰۸/۰۲ و ۱۳۹۶/۱۰/۱۷
تاریخ پذیرش مقاله: ۱۳۹۶/۱۰/۲۱
نام نویسنده مسئول: بابک ناصرشریف
نشانی نویسنده مسئول: ایران – تهران- سیدخندان-دانشگاه صنعتی خواجه‌نصیرالدین طوسی - دانشکده مهندسی کامپیوتر.

۱ -مقدمه

سیستم‌های خودکار بازشناسی گفتار، زمانی که در شرایط واقعی مورداستفاده قرار می‌گیرند با کاهش دقت مواجه می‌شوند. از مهم‌ترین عوامل کاهش دقت بازشناسی، می‌توان به وجود نویز محیطی، تفاوت در لحن گفتار، اثر اعوجاج کانال و عدم انطباق شرایط آموزش و آزمایش اشاره کرد. مقاوم‌سازی گفتار به معنای افزایش دقت شناسایی در هنگامی است که سیگنال گفتار به هر دلیلی تخریب شده است [۲،۱]. روش‌های مقاوم‌سازی گفتار در برابر نویز را به‌طورکلی می‌توان به سه دسته تقسیم کرد: روش‌های حذف نویز از سیگنال گفتار [۴،۳]، روش‌های حذف نویز از ویژگی‌های گفتار و روش‌های تطبیق مدل که دو روش اول جزو روش‌های پیش‌پردازش و روش آخر از روش‌های پس‌پردازش می‌باشد [۵،۲،۱].

روش‌های حذف نویز از ویژگی‌ها به دو دسته تقسیم می‌شوند. دسته اول روش‌هایی که به دنبال استخراج ویژگی‌های گفتاری جدید هستند که درروند استخراج ویژگی تغییراتی داده می‌شود، مانند ویژگی‌های PAC[۱] [۶]. روش‌های دسته دوم به دنبال بهبود و جبران اثر نویز در ویژگی‌های گفتاری متداول هستند؛ که در آن تبدیلی خطی یا غیرخطی بر روی بردار ویژگی اعمال می‌شود تا به‌این‌ترتیب ویژگی‌های نویزی به ویژگی‌های تمیز نگاشت شوند مانند هنجارسازی میانگین و واریانس ضرایب کپسترال و برابرسازی هیستوگرام[۲] و نیز نگاشت به کمک انواع شبکه‌های عصبی عمیق (DNNs[۳]) [۶،۴،۳].

در سال‌های اخیر، شبکه‌های عصبی عمیق کاربرد فراوانی در مدل‌سازی آکوستیک و نیز استخراج ویژگی از سیگنال گفتار یافته‌اند. از انواع شبکه‌های عصبی عمیق که در حوزهٔ پردازش گفتار مورداستفاده قرارگرفته‌اند می‌توان به شبکه‌های باور عمیق (DBNs[۴])، خود رمزگذار عمیق (DAE[۵]) و شبکه‌های عصبی درهم‌پیچش (CNNs[۶]) اشاره کرد [۸،۴،۳]. شبکه‌های باور عمیق، هم برای استخراج ویژگی مقاوم و هم برای حذف نویز از سیگنال گفتار مورداستفاده قرارگرفته‌اند [۹،۷،۴]. شبکه‌های خود رمزگذار نیز برای فشرده‌سازی و استخراج ویژگی و همچنین برای بهسازی سیگنال گفتار بکار رفته‌اند [۱۰،۳]. علاوه بر این شبکه باور عمیق نیز در مدل‌سازی آکوستیک به همراه مدل مخفی مارکف به طرز وسیعی بکار گرفته شده است [۱۱].

شبکه‌های عصبی درهم‌پیچش نوع به خصوصی از شبکه‌های عصبی است که از لایه‌های پیاپی درهم‌پیچش و ادغام[۷] به همراه حداقل یک لایه کاملاً متصل[۸] در بالای آن تشکیل شده است [۱۳،۱۲]. این نوع شبکه‌ها به دلیل ساختار درهم‌پیچش و ادغام در مقابل اعوجاج و جابه‌جایی‌های کوچک در طول زمان یا فرکانس مقاوم هستند و همچنین با اعمال درهم‌پیچش‌های محلی وابستگی‌ها را در همسایگی‌های کوچک‌تر در نظر می‌گیرند [۱۳،۱۲،۵].

شبکه‌های عصبی درهم‌پیچش به همراه مدل مخفی مارکف به‌منظور مدل‌سازی به کار گرفته‌شده‌اند. در این سیستم‌ها به‌منظور تخمین احتمالات مشاهده حالات مدل مخفی مارکف، از داده‌های حوزه طیف نگار[۹] [۱۴،۸،۵] و یا از داده‌های خام‌گفتاری [۱۵-۱۷] به‌عنوان ورودی شبکه استفاده شده است. هنگامی‌که از سیگنال‌های خام استفاده شده، خود شبکه نقش استخراج‌گر ویژگی را دارا می‌باشد.

این در حالی است که در برخی دیگر مقالات از این نوع شبکه‌ها به‌منظور استخراج ویژگی به‌تنهایی استفاده شده است. به‌عنوان نمونه در [۱۸]، CNN در کنار DBN به‌عنوان استخراج‌گر ویژگی برای یک سیستم بازشناسی گفتار پیوسته با واژگان بزرگ مورداستفاده قرارگرفته است. همچنین در [۱۹]، برای بازشناسی گفتار کسانی که مشکلات گفتاری دارند، از شبکه‌های گلوگاه درهم‌پیچش[۱۰] برای استخراج ویژگی بهره گرفته شده است.

علاوه بر بازشناسی گفتار، از CNN در شاخه‌های دیگری نیز استفاده شده است. در [۲۰]، ویژگی‌های استخراج‌شده توسط CNN برای بازشناسی زبان، بکار رفته‌اند. از سوی دیگر، برای جداسازی سکوت از گفتار نیز، از ویژگی‌های استخراج‌شده به‌وسیله CNN کمک گرفته شده است [۲۱]. علاوه بر این، در حوزه پردازش تصویر از CNN به‌منظور مقابله با نویز جمع‌پذیر استفاده شده و کارایی آن در مقابل نویز جمع‌پذیر نشان داده شده است [۲۲].

در برخی کارها از ترکیب سایر شبکه‌های عصبی نظیر LSTM[۱۱] با CNN جهت استخراج ویژگی و بازشناسی گفتار استفاده شده است [۲۳،۲۴]. در [۲۵] یک لایه مخفی کاملاً متصل در ابتدای شبکه نقش بانک فیلتر مل را ایفا می‌کند که ورودی آن سیگنال خام است. خروجی این لایه به یک شبکه ماشین بولتزمن محدود شدهٔ درهم‌پیچش داده می‌شود.

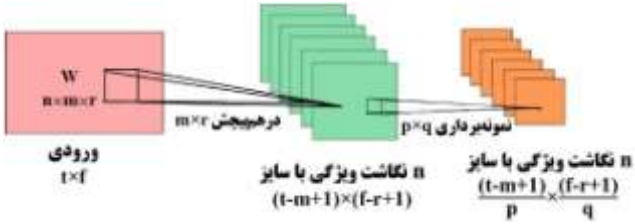
در این مقاله، از CNN برای استخراج ویژگی‌های مقاوم به نویز استفاده می‌شود. به‌این‌ترتیب که طیف سیگنال گفتار نویزی به‌عنوان ورودی به CNN داده می‌شود و در خروجی شبکه، مقادیر لگاریتم انرژی زیرباندهای مل (LMFB[۱۲]) تحویل گرفته می‌شود. به‌این‌ترتیب، CNN به‌عنوان یک بانک فیلتر، برای استخراج ویژگی مقاوم به نویز از طیف سیگنال گفتار بکار می‌رود. برای عملکرد بهتر CNN در نقش پیشنهادی، اندازه فیلتر درهم‌پیچش و اندازه ادغام و تعداد نورون‌ها در لایه‌ها نقش مهمی دارند که در کار حاضر به‌تفصیل موردبررسی قرارگرفته‌اند. از طرفی، اندازهٔ فیلتر درهم‌پیچش در CNN متداول در تمام زیر باندها یکسان است. درنتیجه یک CNN استاندارد نمی‌تواند کاهش لگاریتمی وضوح فیلتر مل را مدل کند. درنتیجه ساختار جدیدی به نام شبکه عصبی درهم‌پیچش چند دقتی (MRCNN) پیشنهاد می‌شود که در آن با استفاده از چند CNN با اندازه‌های فیلتر درهم‌پیچش متفاوت، چند وضوح فرکانسی متفاوت برای استخراج ویژگی‌ها استفاده می‌شود.

در ادامه مقاله، در بخش دوم شبکه عصبی درهم‌پیچش شرح داده خواهد شد. در بخش سوم روش پیشنهادی و ساختار شبکه معرفی خواهد شد. در بخش چهارم آزمایش‌ها و نتایج مورد بررسی قرار خواهند گرفت. درنهایت، بخش پنجم به جمع‌بندی اختصاص دارد.

۲ -شبکه عصبی درهم‌پیچش

یک شبکه عصبی درهم‌پیچش از سه بخش اصلی تشکیل می‌شود: لایه‌های درهم‌پیچش، لایه‌های ادغام و تعدادی لایه مخفی متصل کامل در بالای لایه‌های درهم‌پیچش و ادغام.

در شکل ۱ لایه درهم‌پیچش و ادغام با جزئیات بیش‌تر نشان داده شده‌اند. لایه درهم‌پیچش، ماتریس ورودی با ابعاد t×f را (که t نشانگر بعد زمان و f نشانگر بعد فرکانس می‌باشد) به‌عنوان ورودی دریافت می‌کند [۱۸]. لایه درهم‌پیچش متشکل از n نورون است؛ که هر نورون یک فیلتر است و ابعاد هر فیلتر m×r می‌باشد که با کل ورودی درهم‌پیچش می‌شود [۱۸،۱۴،۸]. هر فیلتر در بعد زمان - فرکانس بر روی ورودی (با فرض m<=t, r<=f) حرکت می‌کند و در خروجی هر نورون، نگاشت‌های ویژگی با ابعاد (t-m+1)×(f-r+1) مشاهده خواهد شد [۱۸]. این اشتراک‌گذاری وزن‌ها از طرفی باعث کاهش تعداد وزن‌های یادگیری نسبت به حالت کاملاً متصل می‌شود [۵] و از طرف دیگر توانایی مدل‌سازی وابستگی‌ها و ارتباطات محلی در سیگنال ورودی را فراهم می‌آورد [۱۸،۵].



شکل ۱: نحوه عملکرد لایه درهم‌پیچش و ادغام [۱۸]

پس از لایه درهم‌پیچش نوبت به لایه ادغام می‌رسد. در این لایه نمونه‌برداری‌های محلی از خروجی‌های لایه درهم‌پیچش انجام می‌شود. به‌صورت کلی روش‌های ادغام بسیاری تاکنون معرفی شده است و پرکاربردترین‌های آن‌ها عبارت‌اند از: ادغام میانگین[۱۳]، ادغام بیشینه[۱۴]، ادغام تصادفی[۱۵] [۲۶،۵].

کارایی ادغام بیشینه نسبت به ادغام میانگین در کاربردهای بازشناسی گفتار نشان داده شده [۵] و ادغام تصادفی بهبود چشم‌گیری را در پی نداشته است [۲۶]. ادغام بیشینه تغییرات محدود در حوزه زمان - فرکانس که در اثر حالات مختلف صحبت، تفاوت در گوینده و ... ایجاد می‌شود را از میان می‌برد [۱۸] و نسخه‌ای با وضوح[۱۶] پایین‌تر از خروجی‌های لایه درهم‌پیچش را ارائه می‌دهد [۸].

گاهی پس از لایه درهم‌پیچش از تابع فعالیت استفاده می‌شود. در برخی کارها از تابع فعالیت سیگموید [۱۹،۱۸،۱۴،۵]، در برخی موارد از تابع فعالیت تانژانت هیپربولیک [۲۰] و در برخی دیگر از تابع فعالیت ReLU[۱۷] [۲۲،۱۸] استفاده شده است. این در حالی است که در برخی مقالات تابع فعالیت پس از لایه ادغام اعمال شده است. به‌عنوان نمونه در [۲۶،۱۶] از تابع فعالیت تانژانت هیپربولیک بعد از لایه ادغام استفاده شده است.

پس از لایه‌های درهم‌پیچش و ادغام از یک یا چند لایه مخفی موسوم به لایه‌های کاملاً متصل استفاده می‌شود [۵].

در برخی کارها شبکه عصبی درهم‌پیچش درون ساختارهای دیگر تعبیه شده است. به‌عنوان‌مثال در [۲۷] یک شبکه عصبی درهم‌پیچش درون یک ساختار بازگشتی قرار داده شد، به این طریق یک شبکه عصبی درهم پیچش بازگشتی برای کارهای تشخیص واج و تشخیص احساسات مورداستفاده قرارگرفته است. در [۲۸] لایه شبکه عصبی درهم پیچش درون یک LSTM تعبیه شد. این لایه‌ها به همراه لایه‌های شبکه‌های عصبی درهم پیچش معمول به دنبال هم قرار داده شدند و شبکه بسیار عمیق تشکیل داده شد و ازآن‌جهت بازشناسی گفتار استفاده شد.

برخی کارها از روش ساده‌تری جهت بهره‌برداری از مزیت‌های شبکه‌های عصبی بازگشتی و شبکه‌های عصبی درهم‌پیچش استفاده کرده‌اند. به‌عنوان‌مثال در [۲۹] یک شبکه عصبی بازگشتی به دنبال یک شبکه عصبی درهم‌پیچش قرار داده شد.

در بیشتر کارها حداکثر از سه لایه درهم پیچش استفاده شده است. برخی کارها با کوچک در نظر گرفتن فیلترهای درهم‌پیچش و اندازه‌های ادغام از تعداد بیش‌تری لایه درهم‌پیچش و ادغام استفاده کرده‌اند که روش‌های پیشنهادی را شبکه‌های عصبی درهم‌پیچش خیلی عمیق نام‌گذاری کرده‌اند [۳۱،۳۰]. در [۳۰] اندازهٔ فیلتر درهم پیچش در لایه‌های اولیه کوچک در نظر گرفته شد و همین‌طور در لایه‌های اولیه از ادغام استفاده نشد. هدف از این تکنیک حفظ اطلاعات مفید برای لایه‌های بالاتر عنوان شده است. ورودی‌ها در این روش ضرایب LMFB در ۱۱ تا ۱۷ قاب گفتار بوده است. درحالی‌که در [۳۱] اندازهٔ فیلتر درهم‌پیچش و اندازهٔ ادغام در لایه اول بزرگ در نظر گرفته شد و دلیل آن کاهش بار محاسباتی در لایه‌های بالاتر درهم‌پیچش عنوان شده است. دلیل این تفاوت می‌تواند ناشی از تفاوت در ورودی این دو روش باشد زیرا که در [۳۱] ورودی شبکه داده‌های خام گفتار بوده است.

۳ -روش پیشنهادی

چنانچه در مقدمه اشاره شد، CNN قبلاً برای مدل‌سازی آکوستیک در ترکیب با مدل مخفی مارکف و نیز استخراج ویژگی از سیگنال گفتار خام مورداستفاده قرارگرفته است. در این مقاله پیشنهاد می‌شود که از CNN برای استخراج ویژگی‌های مقاوم در برابر نویز محیط، استفاده شود. به‌این‌ترتیب که لگاریتم طیف توان یک قاب از سیگنال گفتار به‌عنوان ورودی به CNN داده می‌شود و در خروجی لگاریتم انرژی فیلتر بانک مل از شبکه تحویل گرفته می‌شود. در تحقیقات قبلی DBN نیز به همین منظور مورداستفاده قرارگرفته است [۷]. در CNN استاندارد، اندازهٔ فیلتر درهم پیچش و ادغام ثابت می‌باشد درنتیجه روش شبکه عصبی درهم‌پیچش چند دقتی برای یادگیری بانک فیلتر مل به‌عنوان روش پیشنهادی دیگری معرفی می‌شود.

۳ ـ۱ ـروش پیشنهادی: شبکه عصبی درهم‌پیچش تک دقتی

برای یادگیری بانک فیلتر، اندازهٔ فیلتر درهم‌پیچش بر اساس پهنای باند بانک فیلتر مل تنظیم شده است؛ اما با توجه به این‌که ما از حالت FWS استفاده می‌کنیم، اندازهٔ فیلتر درهم‌پیچش در همه زیر باندها یکسان است و رفتار لگاریتمی پهنای باند بانک فیلتر مل را دارا نیست.

در جدول ۱، محدوده فرکانسی فیلترهای مل در نرخ نمونه‌برداری ۸ کیلوهرتز نشان داده شده است که مشخصاً در فرکانس‌های پایین‌تر وضوح بیش‌تری نسبت به فرکانس‌های بالاتر وجود دارد.

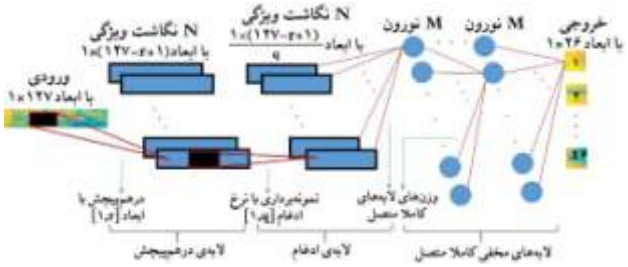
با توجه به محدوده فرکانسی فیلترهای مل و توزیع آن در بازه‌های فرکانسی مختلف، در کار حاضر اندازه فیلترهای درهم‌پیچش مطابق با جدول ۲ تعریف و در نظر گرفته شده است. اندازه فیلترهای درهم‌پیچش طوری تعریف‌شده‌اند که بر اساس پهنای باند فیلترهای مل، بیشترین وضوح و وضوح متوسط فرکانسی در آن‌ها لحاظ شود. با توجه به این‌که اندازه فیلتر درهم‌پیچش برای همه فیلترها ثابت و مساوی است، وضوح پایین فرکانسی در اندازه فیلترها در نظر گرفته نشده است.

جدول ۱: محدوده فرکانسی فیلترهای مل برای نرخ نمونه‌برداری ۸ کیلوهرتز

شماره فیلتر	فرکانس شروع (Hz)	فرکانس پایان (Hz)	شماره فیلتر	فرکانس شروع (Hz)	فرکانس پایان (Hz)
۱	۰	۱۰۶	۱۴	۱۰۵۱	۱۳۱۶
۲	۵۱	۱۶۵	۱۵	۱۱۷۹	۱۴۶۳
۳	۱۰۶	۲۲۸	۱۶	۱۳۱۶	۱۶۲۲
۴	۱۶۵	۲۹۶	۱۷	۱۴۶۳	۱۷۹۱
۵	۲۲۸	۳۶۹	۱۸	۱۶۲۲	۱۹۷۳
۶	۲۹۶	۴۴۷	۱۹	۱۷۹۱	۲۱۶۹
۷	۳۶۹	۵۳۱	۲۰	۱۹۷۳	۲۳۷۸
۸	۴۴۷	۶۲۱	۲۱	۲۱۶۹	۲۶۰۳
۹	۵۳۱	۷۱۷	۲۲	۲۳۷۸	۲۸۴۴
۱۰	۶۲۱	۸۲۱	۲۳	۲۶۰۳	۳۱۰۳
۱۱	۷۱۷	۹۳۲	۲۴	۲۸۴۴	۳۳۸۱
۱۲	۸۲۱	۱۰۵۱	۲۵	۳۱۰۳	۳۶۸۰
۱۳	۹۳۲	۱۱۷۹	۲۶	۳۳۸۱	۴۰۰۰

جدول ۲: اندازهٔ فیلترهای درهم‌پیچش پیشنهادی و پهنای باند متناظر آن‌ها در نرخ نمونه‌برداری ۸ کیلوهرتز

اندازهٔ فیلتر درهم‌پیچش	پهنای باند (Hz)	اندازهٔ فیلتر درهم‌پیچش	پهنای باند (Hz)
۱×۳	۹۳	۱×۸	۲۵۰
۱×۴	۱۲۵	۱×۱۵	۴۶۸
۱×۵	۱۵۶	۱×۱۸	۵۶۲
۱×۶	۱۸۷	۱×۲۰	۶۲۵
۱×۷	۲۱۸	۱×۲۲	۶۸۷



شکل ۲: ساختار CNN پیشنهادی
(در بهترین ساختار: N=240، r=6، q=3، دولایه مخفی و M=500)

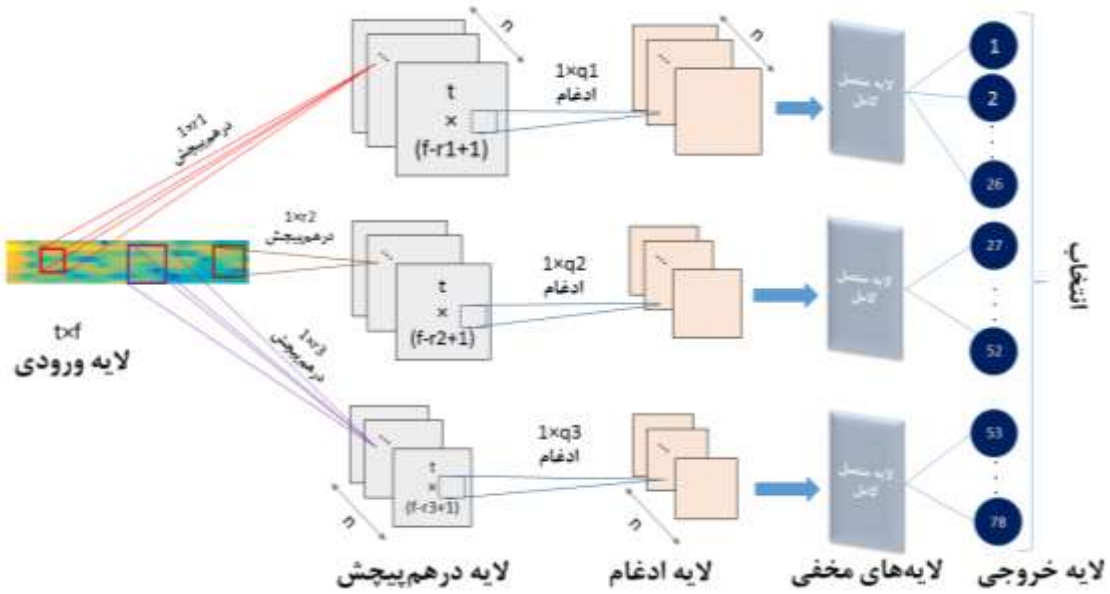
در شکل ۲ ساختار کامل CNN پیشنهادی به همراه ورودی و خروجی استفاده شده در این مقاله نشان داده شده است. در این شکل، N تعداد نورون‌های لایه درهم‌پیچش، r اندازهٔ فیلتر درهم‌پیچش، q اندازهٔ ادغام و M تعداد نورون در هر لایه مخفی می‌باشد. با توجه به آزمایش‌های انجام‌شده، اندازهٔ بهینهٔ فیلتر درهم‌پیچش ۶×۱ و اندازهٔ بهینهٔ ادغام ۳×۱ به‌دست آمده است. همچنین تعداد نورون‌های بهینه در لایه درهم‌پیچش ۲۴۰ و در لایه کاملاً متصل تک لایه و دولایه ۵۰۰ در نظر گرفته شده است. به‌کارگیری بیش از دولایه مخفی موجب افت نتایج شده است.

انتخاب ساختار فوق بر اساس تحقیقات و مقالات قبلی پردازش گفتار صورت گرفته است که در آن‌ها تعداد نورون‌های لایه درهم‌پیچش را از ۴۰ تا ۵۱۲ و تعداد نورون‌های لایه‌های کاملاً متصل را در حالت تک لایه، ۵۰۰ و در حالت دو یا سه لایه ۱۰۰۰ در نظر گرفته‌اند [۵-۲۵].

۳ ـ۲ ـروش پیشنهادی: شبکه عصبی درهم‌پیچش چند دقتی

همان‌طور که گفته شد، اندازهٔ فیلتر درهم‌پیچش در شبکه عصبی درهم‌پیچش متداول در تمام زیر باندها یکسان است. درنتیجه یک CNN استاندارد نمی‌تواند ساختار لگاریتمی وضوح فیلتر مل را مدل کند. در این بخش ساختار جدیدی به نام شبکه عصبی درهم‌پیچش چند دقتی (MRCNN) پیشنهاد می‌شود که در آن با استفاده از چند CNN با اندازه‌های فیلتر درهم‌پیچش متفاوت، بر مشکل وضوح فرکانسی متفاوت غلبه می‌کنیم. به‌این‌ترتیب که برای هر وضوح فرکانسی متوسط، کم و زیاد در بانک فیلتر مل یک یا چند شبکه عصبی درهم‌پیچش در نظر گرفته شود. ساختار MRCNN در شکل ۳ نشان داده شده است.

به‌طورکلی از دو یا سه CNN استفاده شده و از خروجی‌های CNN ها به دو صورت بهره گرفته شده است. در روش اول، تمام خروجی‌های CNN ها به دنبال هم قرار داده شدند و درنتیجه هنگامی‌که از دو CNN استفاده شده، بردار ویژگی به اندازهٔ ۵۲ و هنگامی‌که از سه CNN استفاده شده، بردار ویژگی به اندازهٔ ۷۸ تشکیل شده است که این بردارهای ویژگی برای آموزش سیستم GMM-HMM بکار رفته‌اند.



شکل ۳: بلوک دیاگرام شبکه عصبی درهم‌پیچش چند دقتی پیشنهادی

در روش دوم، از هر CNN تعدادی خروجی به نحوی انتخاب شده است که درنهایت بردار ویژگی به اندازهٔ ۲۶ (تعداد متناظر بانک‌های فیلتر مل) برای آموزش سیستم GMM-HMM بکار رود. این انتخاب با توجه به اندازهٔ فیلتر درهم‌پیچش و همچنین پهنای باند زیر باندهای فیلتر مل صورت گرفته است. به این صورت که از شبکه‌های با وضوح بالاتر (فیلتر درهم‌پیچش کوچک‌تر) خروجی‌های اول آن‌ها و از شبکه‌های با وضوح پایین‌تر خروجی‌های آخر آن‌ها انتخاب شده است.
هنگامی‌که از سه CNN در سیستم MRCNN استفاده شده (مانند شکل ۳)، MRCNN3 نام گرفته و بر اساس جدول ۲ اندازهٔ فیلتر درهم‌پیچش در فرکانس‌های پایین، متوسط و بالا متفاوت انتخاب شده است. به این صورت که در شکل ۳ خواهیم داشت: r1<r2<r3.

هنگامی‌که از دو CNN در سیستم MRCNN استفاده شده، MRCNN2 نام‌گرفته و تنها فرکانس‌های پایین و بالا در نظر گرفته شده است، درنتیجه CNN میانی در شکل ۳ با اندازهٔ فیلتر درهم‌پیچش r2 حذف خواهد شد. درنتیجه خواهیم داشت: r1<r3.

۴ -آزمایش‌ها و نتایج

در این پژوهش مجموعه داده مورد ارزیابی، مجموعه Aurora2 می‌باشد [۳۲]. این مجموعه داده حاوی گفتار متصل متشکل از ارقام انگلیسی تا ۷ رقم با نرخ نمونه‌برداری ۸ کیلوهرتز می‌باشد. در دادگان Aurora2 دو مجموعه مجزای آموزشی و آزمایشی مشخص شده است. داده‌های آموزش به دو دسته تقسیم می‌شوند: داده‌های تمیز و داده‌های نویزی. داده‌های آزمایش نیز به سه دسته A، B و C تقسیم می‌شوند که هر یک هم شامل گفتار تمیز و هم شامل گفتار نویزی با نویزهای مشخص شده برای هر مجموعه در نسبت‌های سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل می‌باشند.

مجموعه گفتار نویزی A شامل نویزهای جمع‌پذیر مترو، ماشین، همهمه و نمایشگاه به همراه فیلتر کانال G.712 می‌باشد. مجموعه گفتار نویزی B شامل نویزهای جمع‌پذیر رستوران، خیابان، فرودگاه و ایستگاه قطار به همراه فیلتر کانال G.712 می‌باشد. مجموعه گفتار نویزی C شامل نویزهای جمع‌پذیر مترو و خیابان است که هر یک با نویزهای مجموعه A و B مشترک هستند؛ اما در مجموعه C از فیلتر کانال MIRS استفاده شده که متفاوت از نویز کانال استفاده شده در مجموعه A و B است.

طول قاب و شیفت قاب به ترتیب ۲۵ و ۱۰ میلی‌ثانیه و تعداد فیلترهای مل در نظر گرفته‌شده ۲۶ عدد است. مدل آکوستیک مورداستفاده مدل مخفی مارکف با ۱۶ حالت است که برای هر حالت، مخلوط گوسی ۳ مؤلفه‌ای در نظر گرفته شده است و با مجموعه گفتار تمیز Aurora2 آموزش دیده است که شامل ۸۴۴۰ فایل گفتاری است. برای پیاده‌سازی CNN جعبه‌ابزار CNTK[۱۸] میکروسافت بکار رفته است [۳۳].

برای آموزش CNN کلمات مجزای نویزی موجود در مجموعه آموزش چندحالته استفاده شده است که فقط نویزهای مجموعه A، شامل نویزهای مترو، همهمه، ماشین و نمایشگاه را در برمی‌گیرد و داده‌های هدف، از گفتار تمیز متناظر تأمین‌شده‌اند. در این حالت مجموعه آموزش CNN شامل پنجاه‌هزار قاب گفتاری است. برای آزمایش CNN و سیستم بازشناسی گفتار، از نویزهای مجموعه B شامل نویزهای رستوران، خیابان، فرودگاه و ایستگاه قطار و نویز کانال متفاوت و نویزهای جمع‌پذیر در مجموعه C شامل مترو و خیابان استفاده شده است. به‌دلیل وجود مقادیر مثبت و منفی در ورودی و خروجی، هم در لایه درهم‌پیچش و هم در لایه مخفی، تابع فعالیت تانژانت هیپربولیک بکار گرفته شده است. تابع فعالیت خروجی خطی

می‌باشد و از گرادیان نزولی تصادفی (SGD[19]) و از اندازهٔ دسته‌های کوچک[20] بین ۱۰۰ تا ۵۰۰ برای آموزش شبکه استفاده شده است. همچنین از گام پنجم به بعد ضریب ممان[21] نیز بکار رفته است.

نرخ آموزش در ۳ گام[22] اول لایه درهم‌پیچش ۰٫۱، لایه مخفی ۰٫۰۱ و لایه خروجی ۰٫۰۰۰۱ انتخاب شده است که از گام چهارم تا گام دهم هر یک تا ۰٫۰۰۳ برابر نرخ آموزش اولیه کاهش می‌یابند. با توجه به آزمایش‌های انجام شده، بهترین تعداد گام‌های آموزش ۱۰ انتخاب شده است.

۴ ـ ۱ ـ نتایج شبکه عصبی درهم‌پیچش تک دقتی

۴ ـ ۱ ـ ۱ ـ تأثیر اندازهٔ فیلتر درهم‌پیچش

برای تعیین اندازهٔ بهینهٔ فیلترهای درهم‌پیچش، اندازه‌های تعریف شده در جدول ۲ موردبررسی قرارگرفته‌اند. اندازهٔ ادغام در تمام اندازه‌های فیلتر ۳×۱ بوده است و از ۲۴۰ نورون برای لایه درهم‌پیچش و یک لایه ۵۰۰ نورونی برای لایه مخفی استفاده شده است. همان‌طور که میانگین نتایج بازشناسی بر روی سه مجموعه آزمایشی A، B و C بر روی نسبت‌های سیگنال به نویز ۰ تا ۲۰ دسی‌بل در جدول ۳ نشان می‌دهند، اندازهٔ فیلتر درهم‌پیچش ۶×۱ معادل پهنای باند ۱۸۷ هرتز (معادل پهنای باند فیلتر شماره ۹ مل در جدول ۱) بهترین نتایج بازشناسی را به دست داده است.

۴ ـ ۱ ـ ۲ ـ تأثیر اندازهٔ ادغام

یکی از پارامترهای مهم در مقاوم‌سازی سیستم در مقابل جابه‌جایی‌های کوچک فرکانسی، ادغام می‌باشد. در جدول ۴ برای بهترین اندازهٔ فیلتر درهم‌پیچش (۶×۱) با ساختار مشابه بخش (۴-۱-۱-) تأثیر اندازه‌های ادغام مختلف نشان داده شده است. از جدول می‌توان مشاهده کرد که بهترین اندازهٔ ادغام ۳×۱ و ۴×۱ می‌باشد. در آزمایش‌های بعدی ما ۳×۱ بهتر از ۴×۱ عمل کرده بنابراین ما ۳×۱ را به‌کاربرده‌ایم.

۴ ـ ۱ ـ ۳ ـ تأثیر تعداد نورون‌های لایه درهم‌پیچش

با بهترین حالت به‌دست‌آمده از بخش‌های (۴-۱-۱ و ۴-۱-۲) تعداد نورون‌های لایه درهم‌پیچش مختلف مورد آزمایش قرارگرفته و در جدول ۵ نتایج ثبت شده است. از جدول ۵ می‌توان نتیجه گرفت که ۲۴۰ بهترین تعداد نورون برای لایه درهم‌پیچش می‌باشد.

۴ ـ ۱ ـ ۴ ـ تأثیر تعداد لایه‌های مخفی

به‌منظور بررسی تأثیر اندازه و تعداد لایه‌های مخفی، بهترین پارامترها و تنظیمات به‌دست آمده در بخش‌های (۴-۱-۱ تا ۴-۱-۳) با یک تا سه لایه مخفی با ۵۰۰ و ۱۰۰۰ نورون و همچنین دو لایه مخفی با ۲۰۴۸ نورون موردبررسی قرارگرفته و نتایج در جدول ۶ به ثبت رسیده است. بهترین نتیجه با دولایه مخفی که هر یک ۵۰۰ نورون دارند، به‌دست آمده است.

جدول ۳: تأثیر اندازهٔ فیلتر درهم‌پیچش در میانگین دقت بازشناسی بر روی نسبت‌های سیگنال به نویز 0-20 dB

اندازهٔ فیلتر درهم‌پیچش	دقت بازشناسی			
	مجموعه A	مجموعه B	مجموعه C	میانگین
۱×۳	۶۱/۵۳	۵۷/۳۰	۵۶/۵۶	۵۸/۴۶
۱×۴	۶۰/۱۰	۵۹/۹۰	۵۵/۸۰	۵۸/۶۰
۱×۵	۶۰/۶۶	۵۷/۴۵	۵۶/۶۰	۵۸/۲۴
۱×۶	۶۱/۳۱	۵۸/۷۱	۵۶/۸۷	۵۸/۹۷
۱×۷	۶۰/۷۹	۵۸/۲۷	۵۵/۷۹	۵۸/۲۸
۱×۸	۶۰/۹۰	۵۷/۸۴	۵۵/۳۷	۵۸/۰۴

جدول ۴: تأثیر اندازهٔ ادغام برای اندازهٔ فیلتر درهم‌پیچش ۶×۱

اندازهٔ ادغام	دقت بازشناسی			
	مجموعه A	مجموعه B	مجموعه C	میانگین
۱×۲	۶۰/۸۳	۵۸/۷۷	۵۵/۲۹	۵۸/۳۰
۱×۳	۶۱/۳۱	۵۸/۷۱	۵۶/۸۸	۵۸/۹۷
۱×۴	۶۱/۴۷	۵۸/۹۹	۵۶/۴۳	۵۸/۹۶

جدول ۵: تأثیر تعداد نورون‌های لایه درهم‌پیچش در اندازه فیلتر ۶×۱ و اندازه ادغام ۳×۱

تعداد نورون‌های لایه درهم‌پیچش	دقت بازشناسی			
	مجموعه A	مجموعه B	مجموعه C	میانگین
۱۲۰	۶۰/۹۰	۵۸/۶۷	۵۵/۷۸	۵۸/۴۵
۲۰۰	۶۰/۶۸	۵۸/۵۷	۵۵/۷۵	۵۸/۳۳
۲۲۰	۶۰/۶۸	۵۸/۷۴	۵۵/۰۰	۵۸/۱۴
۲۴۰	۶۱/۳۱	۵۸/۷۱	۵۶/۸۷	۵۸/۹۷
۲۵۶	۶۰/۵۱	۵۸/۶۴	۵۵/۴۵	۵۸/۲۰
۳۰۰	۶۰/۵۷	۵۸/۳۷	۵۵/۱۱	۵۸/۰۲
۵۱۲	۶۰/۸۲	۵۸/۷۹	۵۴/۶۲	۵۸/۰۸

جدول ۶: تأثیر تعداد لایه‌های مخفی در CNN با تعداد نورون‌های لایه درهم‌پیچش ۲۴۰، اندازهٔ فیلتر ۶×۱، اندازهٔ ادغام ۳×۱

تعداد نورون‌های لایه مخفی	دقت بازشناسی			
	مجموعه A	مجموعه B	مجموعه C	میانگین
۱×۵۰۰	۶۱/۳۱	۵۸/۷۱	۵۶/۸۷	۵۸/۹۷
۲×۵۰۰	۶۳/۲۹	۵۹/۸۰	۵۸/۸۹	۶۰/۶۶
۳×۵۰۰	۶۲/۸۹	۵۹/۳۷	۵۸/۳۵	۶۰/۲۰
۱×۱۰۰۰	۶۰/۷۹	۵۸/۵۱	۵۵/۵۸	۵۸/۲۹
۲×۱۰۰۰	۶۰/۹۰	۵۸/۶۲	۵۵/۳۹	۵۸/۳۰
۳×۱۰۰۰	۶۰/۵۱	۵۷/۲۳	۵۶/۴۱	۵۸/۰۵
۲×۲۰۴۸	۵۹/۱۵	۵۳/۱۷	۵۱/۵۱	۵۴/۶۱

بر اساس جدول‌های (۳ تا ۶)، می‌توان گفت اندازهٔ فیلتر درهم‌پیچش ۶×۱ و اندازهٔ ادغام ۳×۱ در ساختار شبکه با ۲۴۰ نورون درهم‌پیچش و دولایه مخفی با ۵۰۰ نورون، بهترین عملکرد را در یادگیری یک بانک فیلتر مقاوم به نویز داشته است.

۴ ۲ -مقایسه عملکرد CNN و DBN

در [۷]، از DBN برای استخراج ویژگی‌های طیفی مقاوم از سیگنال‌های نویزی و اعوجاج یافته استفاده شده است. بدین منظور با استفاده از DBN سه نگاشت ویژگی انجام گرفته است: نگاشت طیف نگار نویزی به طیف نگار تمیز، نگاشت LMFB نویزی به LMFB تمیز و نگاشت طیف نگار نویزی به LMFB. از بین این سه مورد بهترین نتیجه از نگاشت طیف نگار به LMFB به‌دست آمده است. ازآنجاکه در کار حاضر نیز ما از CNN برای استخراج LMFB از طیف سیگنال نویزی استفاده می‌کنیم، بنابراین می‌توان کارآیی DBN و CNN را در این روش استخراج ویژگی با هم مقایسه نمود. به همین دلیل، در این بخش، بهترین نتیجه به‌دست آمده از بخش ۴-۱-۴ را با نتایج شبکه DBN که برای استخراج ویژگی از دادگان Aurora2 آموزش دیده است، مقایسه خواهیم کرد.

تلاش شده است که DBN ساختاری مشابه با CNN داشته باشد. برای آموزش DBN از گفتار نویزی مجموعه A استفاده شده است که داده‌های هدف (۲۶ فیلتر مل) از گفتار تمیز متناظر تأمین‌شده‌اند. همچنین از دو لایه مخفی با ۵۰۰ نورون و توابع فعال‌ساز تانژانت هیپربولیک استفاده شده است. سطر DBN (1 frame) در جدول ۷ بیانگر این ساختار است که در ورودی آن تنها از طیف یک قاب گفتار استفاده شده است. سطر DBN (3 frame) در این جدول، نشان‌دهنده نتایج شبکه باور عمیقی است که از طیف سه قاب متوالی گفتار به‌عنوان ورودی شبکه استفاده کرده و دادگان آموزشی آن، هم شامل دادگان تمیز و هم شامل دادگان نویزی است.

همان‌طور که در جدول ۷ مشاهده می‌شود، به‌طور میانگین CNN در مقابل DBN بسیار بهتر عمل می‌کند تا آنجا که حتی با استفاده از تک‌قاب ورودی و دادگان آموزشی نویزی عملکرد بهتر از DBN با سه قاب ورودی و مجموعه دادگان آموزشی نویزی و تمیز دارد. طبیعتاً این امر قابل‌انتظار است؛ زیرا DBN فقط یک نگاشت کور بدون توجه به خصوصیات فرکانسی طیف انجام می‌دهد، درحالی‌که CNN با در نظر گرفتن اندازه فیلتر درهم‌پیچش مناسب و عملکرد ادغام به خصوصیات طیف گفتار توجه نشان می‌دهد.

۴ ۳ -نتایج شبکه عصبی درهم‌پیچش چند دقتی

در جدول ۸ و جدول ۹ (به ترتیب) نام‌های اختصاری برای تمام روش‌های پیشنهادی برای سیستم‌های با دو سطح دقت (MRCNN 2) و سیستم‌های با سه سطح دقت (MRCNN 3) نشان داده شده است. عددی که در پرانتز بعد از MRCNN 2 و MRCNN 3 آمده، شماره روش استفاده شده را بیان می‌کند که شامل اطلاعاتی مثل اندازهٔ فیلتر و تعداد ویژگی‌هاست. عبارات HR، MIR و LR به ترتیب خلاصه‌سازی عبارات: CNN با وضوح بالا، CNN با وضوح میانی و CNN با وضوح پایین می‌باشند. در ستون‌های HR، MIR و LR اندازهٔ فیلتر و تعداد ویژگی انتخابی در پرانتز جلوی آن برای CNN موردبحث در آن ستون نوشته‌شده است.

جدول ۷: مقایسه عملکرد DBN و CNN در یادگیری بانک فیلتر

سیستم استخراج‌گر ویژگی	دقت بازشناسی			
	مجموعه A	مجموعه B	مجموعه C	میانگین
LMFB+CMVN	۳۲/۸۴	۳۶	۳۳/۱۳	۳۳/۹۹
DBN (1 frame)	۴۸/۶۳	۴۹/۳۴	۴۸/۰۸	۴۸/۶۸
DBN (3 frames)	۵۸/۲۰	۵۸/۰۴	۵۷/۱۲	۵۷/۷۹
CNN	۶۳/۲۹	۵۹/۸۰	۵۸/۸۹	۶۰/۶۶

جدول ۸: تنظیمات مختلف CNN با دو سطح دقت

نام مدل	HR	LR	بعد ویژگی
Conc.	۱×۲۰ (۲۶)	۱×۶ (۲۶)	MRCNN_2 (1)
Sel.	۱×۲۰ (۷)	۱×۶ (۱۹)	MRCNN_2 (2)
Sel.	۱×۱۸ (۷)	۱×۶ (۱۹)	MRCNN_2 (3)
Conc.	۱×۲۲ (۷)	۱×۶ (۱۹)	MRCNN_2 (4)
Sel.	۱×۲۰ (۶)	۱×۶ (۲۰)	MRCNN_2 (5)
Sel.	۱×۲۰ (۸)	۱×۶ (۱۸)	MRCNN_2 (6)

جدول ۹: تنظیمات مختلف CNN با سه سطح دقت

بعد ویژگی	روش انتخاب	HR	MIR	LR	نام مدل
۷۸	Conc.	۱×۶ (۲۶)	۱×۸ (۲۶)	۱×۲۰ (۲۶)	MRCNN_3 (1)
۷۸	Conc.	۱×۴ (۲۶)	۱×۶ (۲۶)	۱×۲۰ (۲۶)	MRCNN_3 (2)
۲۶	Sel.	۱×۴ (۱۲)	۱×۶ (۷)	۱×۲۰ (۷)	MRCNN_3 (3)

برای مثال، (۱۹) ۶×۱ به معنای آن است که اندازهٔ فیلتر درهم‌پیچش این CNN برابر ۶×۱ و تعداد ویژگی‌های انتخابی از این CNN برابر ۱۹ است. ستون بعد ویژگی، بعد ویژگی‌های به‌دست آمده از سیستم MRCNN را نشان می‌دهد. ستون روش انتخابی، روش انتخاب ویژگی از CNN های مختلف را مشخص می‌کند که دو حالت دارد: .Conc به معنای روش الحاق تمام خروجی‌های CNN با وضوح بالا،CNN با وضوح میانی و CNN با وضوح پایین و .Sel به معنای انتخاب خروجی‌های هر CNN و الحاق خروجی‌های انتخابی از هرکدام از CNN های یادشده است.

ابتدا به‌صورت جداگانه دو روش پیشنهادی را ارزیابی می‌کنیم و درنهایت بهترین ساختارهای هر دو روش را با یکدیگر و با شبکه عصبی درهم‌پیچش تک دقتی موردقیاس قرار می‌دهیم.

۴ ۳ ۱ -نتایج MRCNN برای حالت الحاق تمامی خروجی‌ها

به‌منظور تمرکز بر روی اندازه‌های فیلتر متناسب با پهنای باند زیر باندهای مل، در این بخش بر روی زیر باندهای با وضوح بالا متمرکز می‌شویم. نتایج سیستم‌های CNN مختلف برای حالت الحاق تمام ویژگی‌ها در جدول ۱۰ نشان داده شده است. از فیلترهای درهم‌پیچش

با وضوح بالا برای فرکانس‌های ۰-۲ کیلوهرتز و از فیلترهای درهم‌پیچش با وضوح پایین برای ۲-۴ کیلوهرتز استفاده شده است. اندازهٔ فیلتر درهم‌پیچش برای وضوح پایین بین ۱۸×۱ تا ۲۲×۱ و برای وضوح بالا بین ۴×۱ تا ۸×۱ در نظر گرفته شده است.

بر اساس دو سطر اول جدول ۱۰، به نظر می‌رسد که اندازهٔ فیلتر درهم‌پیچش ۶×۱ انتخاب مناسبی برای وضوح بالا باشد. چنانچه در دو سطر آخر جدول ۱۰ نشان داده شده است، پیشنهاد می‌شود که از این اندازهٔ فیلتر برای CNN با وضوح بالا استفاده شود؛ بنابراین، برای بهترین مدل یافت شده، دو لایه مخفی استفاده شده و در سطر آخر جدول ۱۰ گزارش‌شده است.

همان‌طور که در جدول ۱۰ مشاهده می‌شود، CNN با دو سطح دقت کارایی بهتری در استخراج ویژگی مقاوم در مقایسه با سه سطح دقت دارد؛ که در شکل ۴ بهترین ساختار MRCNN برای حالت الحاق تمامی خروجی‌ها به‌صورت بلوک دیاگرام نشان داده شده است.

۴ ۳ ۲ -نتایج MRCNN برای حالت الحاق منتخب خروجی‌ها

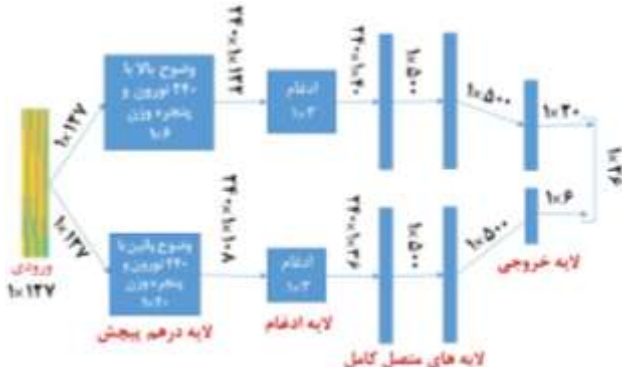
در این بخش ابتدا بر روی اندازهٔ فیلتر درهم‌پیچش در وضوح پایین متمرکز می‌شویم، درحالی‌که استراتژی انتخاب را تغییری نداده و یکسان در نظر می‌گیریم. پس‌ازآن، بر روی استراتژی انتخاب متمرکز خواهیم شد. باآنکه در بخش ۴-۳-۱ برتری CNN با دو سطح دقت نسبت به CNN با سه سطح دقت نشان داده شده است، در این بخش CNN با سه سطح دقت نیز مورد آزمایش قرار می‌گیرد.

جدول ۱۰: میانگین دقت بازشناسی برای CNN چند دقتی در حالت الحاق تمامی خروجی‌ها

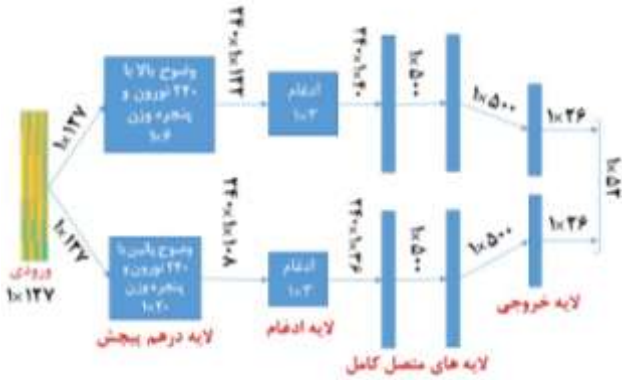
نام مدل	تعداد لایه‌های مخفی	دقت بازشناسی			
		مجموعه A	مجموعه B	مجموعه C	میانگین
MRCNN_3 (1)	۱	۶۱/۴۷	۵۸/۸۶	۵۷/۰۵	۵۹/۱۳
MRCNN_3 (2)	۱	۶۱/۴۸	۵۸/۷۱	۵۷/۳۶	۵۹/۱۸
MRCNN_2 (1)	۱	۶۱/۹۲	۵۹/۳۸	۵۸/۰۰	۵۹/۷۷
MRCNN_2 (1)	۲	۶۳/۵۷	۵۹/۸۰	۵۸/۸۸	۶۰/۷۵

جدول ۱۱: میانگین دقت بازشناسی برای CNN چند دقتی در حالت الحاق خروجی‌های منتخب

نام مدل	تعداد لایه‌های مخفی	دقت بازشناسی			
		مجموعه A	مجموعه B	مجموعه C	میانگین
MRCNN_2 (2)	۱	۶۱/۶۶	۵۸/۷۷	۵۸/۱۹	۵۹/۵۴
MRCNN_2 (3)	۱	۶۱/۶۲	۵۸/۷۸	۵۷/۶۳	۵۹/۴۳
MRCNN_2 (4)	۱	۶۱/۳۵	۵۸/۵۱	۵۷/۷۱	۵۹/۱۹
MRCNN_2 (5)	۱	۶۱/۶۰	۵۸/۶۸	۵۸/۰۵	۵۹/۴۴
MRCNN_2 (6)	۱	۶۰/۹۶	۵۸/۲۵	۵۶/۵۶	۵۸/۵۹
MRCNN_3 (3)	۱	۶۱/۳۶	۵۸/۳۶	۵۷/۶۳	۵۹/۱۲
MRCNN_2 (2)	۲	۶۳/۶۷	۶۰/۱۲	۵۹/۸۳	۶۱/۲۱
MRCNN_2 (5)	۲	۶۳/۷۲	۶۰/۱۳	۵۹/۸۶	۶۱/۲۴



شکل ۴: بلوک دیاگرام بهترین ساختار شبکه عصبی درهم‌پیچش چند دقتی در حالت الحاق تمامی خروجی‌ها



شکل ۵: بلوک دیاگرام بهترین ساختار شبکه عصبی درهم‌پیچش چند دقتی در حالت الحاق خروجی‌های منتخب

همان‌طور که در جدول ۱۱ مشاهده می‌شود، CNN با دو سطح دقت و استراتژی انتخاب ویژگی ۲۰ خروجی اول CNN با اندازهٔ فیلتر ۶×۱ و ۶ خروجی آخر CNN با اندازهٔ فیلتر ۲۰×۱، بهترین نتیجه را در میان تمام استراتژی‌های انتخاب دارد. پس از یافتن بهترین ساختارها در حالت تک لایه مخفی، آن‌ها را با دو لایه مخفی مورد آزمایش قراردادیم که در دو سطر آخر ذکرشده‌اند.

همان‌طور که در جدول ۱۰ و جدول ۱۱ مشاهده می‌شود بهترین نتایج را سیستم MRCNN_2 (5) ارائه می‌دهد که متشکل از دو CNN با اندازهٔ فیلتر درهم‌پیچش ۶×۱ و ۲۰×۱ و استراتژی انتخاب ۲۰ ویژگی از CNN اول و ۶ ویژگی از CNN دوم و دو لایه مخفی با ۵۰۰ نورون می‌باشد؛ که بهترین ساختار MRCNN برای حالت الحاق خروجی‌های منتخب به‌صورت بلوک دیاگرام در شکل ۵ نشان داده شده است.

۴ ۴ -مقایسه روش‌های پیشنهادی

به‌منظور بررسی عملکرد روش CNN تک دقتی و MRCNN در استخراج ویژگی‌های مقاوم به نویز، بهترین نتایج به‌دست آمده از CNN تک دقتی و MRCNN با روش‌های استاندارد دیگر از جمله LMFB+CMVN و DBN در سه قاب مجاور در جدول ۱۲ مورد مقایسه قرارگرفته‌اند. همان‌طور که در جدول ۱۲ مشاهده می‌شود

شبکه عصبی درهم‌پیچش چند دقتی با دو سطح دقت به میزان قابل‌توجهی میانگین دقت بازشناسی را افزایش می‌دهد.

شکل ۶ میانگین دقت بازشناسی کلمه برای CNN، MRCNN،DBN در سه قاب مجاور و LMFB+CMVN را بر روی میانگین نسبت‌های سیگنال به نویز ۰ تا ۲۰ دسی‌بل برای هر نوع نویز

جدول ۱۲: مقایسه میانگین دقت بازشناسی CNN تک دقتی و MRCNN

نام مدل	دقت بازشناسی			
	مجموعه A	مجموعه B	مجموعه C	میانگین
CNN	۶۳/۲۹	۵۹/۸۰	۵۸/۹۰	۶۰/۶۶
MRCNN_2 (5)	۶۳/۷۲	۶۰/۱۳	۵۹/۸۶	۶۱/۲۴
LMFB+CMVN	۳۲/۸۴	۳۶	۳۳/۱۳	۳۳/۹۹
DBN (3 frames)	۵۸/۲۰	۵۸/۰۴	۵۷/۱۲	۵۷/۷۹

نشان می‌دهد. همان‌طور که در شکل ملاحظه می‌شود، روش MRCNN در تمام انواع نویز، حتی در مجموعه C که دارای نویز کانال است، بهترین کارایی را نسبت به روش‌های دیگر داشته است البته در برخی نویزها مانند نویز ماشین و فرودگاه روش‌های MRCNN و CNN عملکرد یکسانی داشته‌اند.

شکل ۷ میانگین دقت بازشناسی کلمه در نویزهای مختلف را برای هر سطح نویز برای CNN، MRCNN، DBN در سه قاب مجاور و LMFB+CMVN نشان می‌دهد. همان‌طور که در شکل ۷ مشاهده می‌شود، در نویز شدید دو روش CNN و MRCNN در یک سطح عمل کرده‌اند. این در حالی است که در نویزهای متوسط و ضعیف MRCNN نسبت به CNN تک دقتی عملکرد بهتری داشته است.

۵ ۴ -تحلیل نتایج

با توجه به نتایج به‌دست آمده از بخش ۴-۲- به نظر می‌رسد وضوح درهم‌پیچش ۶×۱ معادل با پهنای باند ۱۸۷ هرتز (که وضوحی متوسط به حساب می‌آید) مناسب‌ترین اندازهٔ درهم پیچش در شبکه عصبی درهم‌پیچش است. همچنین مشاهده شد که ۲۴۰ نورون درهم‌پیچش بهترین نتیجه را ارائه می‌دهد که دلیل آن را می‌توان به وجود ۲۴۰ نمای[۲۳] مختلف از وزن‌ها (فیلترها) با وضوح ۶×۱ ارجاع داد. این در حالی است که در شبکه‌های باور عمیق با یک نما ورودی بررسی می‌شود، درنتیجه همان‌طور که در نتایج مشهود است شبکه عصبی درهم‌پیچش نسبت به یک شبکه باور عمیق با ساختاری مشابه میانگین خطای بازشناسی کم‌تری دارد. به‌طورکلی شبکه عصبی درهم‌پیچش در تمام نویزها بهتر از شبکه باور عمیق عمل کرده، اما به‌طور مشخص در نویزهای مجموعه A بهتر از سایر مجموعه‌ها عمل کرده است که می‌توان دلیل آن را به وجود این نوع نویز در داده‌های آموزشی نسبت داد.

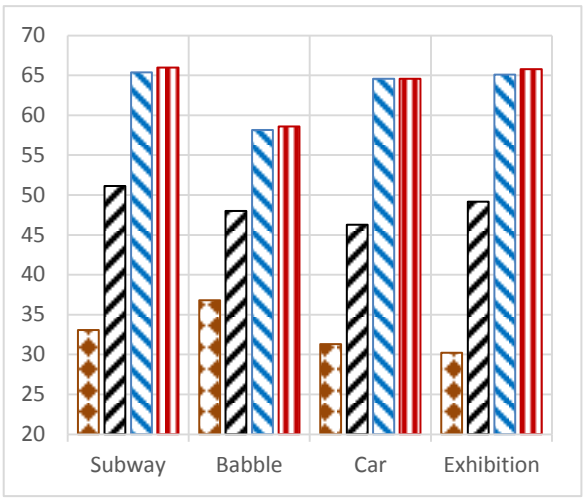
در بررسی روش‌های درهم‌پیچش چند دقتی مشاهده شد که دو سطح دقت عملکردی بهتر از سه سطح دقت دارد. به نظر می‌رسد که دو سطح نمای ۲۴۰ نورونی جهت مدل‌سازی اختلاف وضوح کفایت می‌کند و نیاز به سه سطح نما نیست. همچنین در مقایسه با شبکه عصبی درهم‌پیچش تک دقتی، روش چند دقتی عملکردی بهتر داشته است؛ بنابراین می‌توان نتیجه گرفت که ۴۸۰ نمای مختلف با وضوح متفاوت عملکرد بهتری نسبت به ۲۴۰ نمای مختلف با وضوح یکسان داشته است. البته باید در نظر داشت که در زمان استفاده از روش چند دقتی با دو سطح دقت، هر دو شبکه به‌صورت مجزا عمل می‌کنند و بنابراین زمان اجرای مراحل استخراج ویژگی نسبت به یک سطح دقت مقداری افزایش می‌یابد؛ اما این میزان افزایش در زمان آزمایش یک سیستم بازشناسی گفتار قابل‌توجه نیست.

همچنین در بخش ۴-۴- مشاهده شد که در نویز شدید MRCNN نسبت به CNN بهبودی نداشته است. به نظر می‌رسد نویز به میزانی شدید است که ویژگی‌های گفتاری اطلاعات مفید زیادی ندارند و درنتیجه استفاده از دو سطح دقت نیز نمی‌تواند اطلاعات مقاوم‌تری استخراج نماید. در نویزهای ماشین و فرودگاه هم MRCNN نسبت به CNN بهبودی نداشته است و دلیل آن را می‌توان به شکل و ساختار نویزها مربوط دانست که در فرکانس‌های بالا نویز کم‌تر و در فرکانس‌های پایین نویز شدیدتر است.

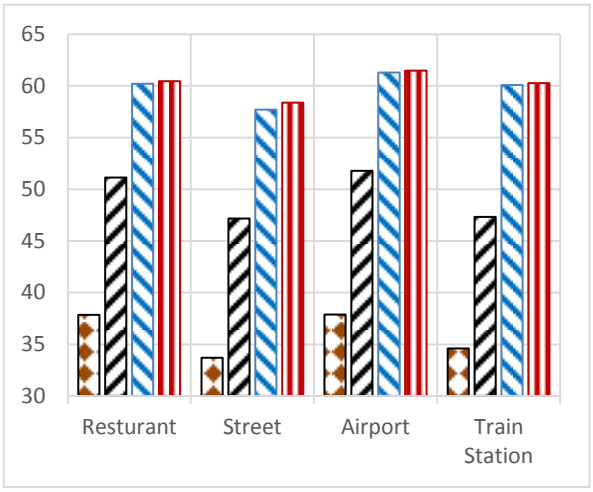
۵ -نتیجه‌گیری

در مقاله حاضر، پیشنهاد گردید که از شبکه عصبی درهم‌پیچش برای یادگیری یک بانک فیلتر مقاوم به نویز و درنتیجه استخراج ویژگی‌های گفتاری مقاوم به نویز استفاده شود. به دلیل ثابت بودن اندازهٔ فیلتر درهم‌پیچش در شبکه‌های عصبی درهم‌پیچش استاندارد امکان مدل کردن تفاوت وضوح بالا و پایین امکان‌پذیر نیست درنتیجه جهت مقابله با مشکل نام‌برده روش شبکه‌های عصبی درهم‌پیچش چند دقتی نیز پیشنهاد شد که در آن از دو یا سه سطح دقت در فیلترهای درهم‌پیچش استفاده شده است. فیلترهای درهم‌پیچش با وضوح پایین تا وضوح متوسط فرکانسی، اندازه‌های مختلف ادغام، ساختارهای متفاوت برای CNN، روش‌های مختلف ادغام خروجی‌های CNN در روش چند دقتی مورد آزمایش قرار گرفتند. بر اساس این آزمایش‌ها، ساختار شبکه با دو سطح دقت به‌صورت: انتخاب ۲۰ خروجی اول CNN با اندازه فیلتر درهم‌پیچش ۶×۱ با ۲۴۰ نورون، اندازه ادغام ۳×۱ و دو لایه مخفی با ۵۰۰ نورون و انتخاب ۶ خروجی آخر CNN با اندازهٔ فیلتر درهم‌پیچش ۲۰×۱ با ۲۴۰ نورون، اندازه ادغام ۳×۱ و دو لایه مخفی با ۵۰۰ نورون به‌دست آمده است.

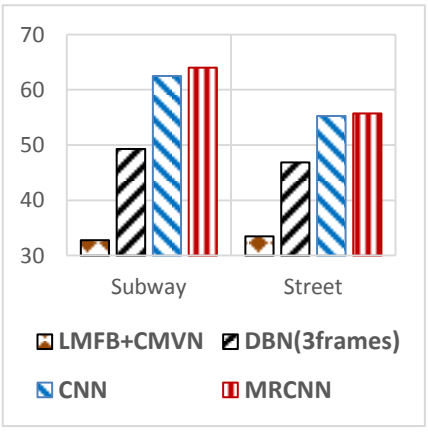
نتایج بیانگر آن‌اند که شبکه عصبی درهم‌پیچش در یادگیری این بانک فیلتر مقاوم به نویز بسیار بهتر از شبکه‌های باور عمیق عمل می‌کند. از طرف دیگر شبکه عصبی درهم‌پیچش چند دقتی نسبت به تک دقتی بهتر عمل می‌کند. این نتیجه بر اساس توجه مستقیم شبکه درهم‌پیچش به خصوصیات فرکانسی طیف گفتار قابل پیش‌بینی است؛ زیرا شبکه باور عمیق فقط یک نگاشت کور را یاد می‌گیرد و به‌طور مستقیم به ویژگی‌های فرکانسی گفتار توجهی ندارد. در کارهای آتی، قصد آن است که روشی با ساختار عمیق‌تر برای استخراج ویژگی مقاوم‌تر پیشنهاد شود.



الف) مجموعه A

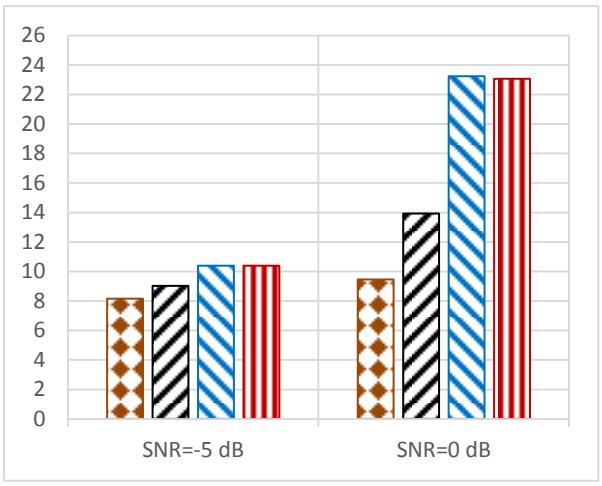


ب) مجموعه B

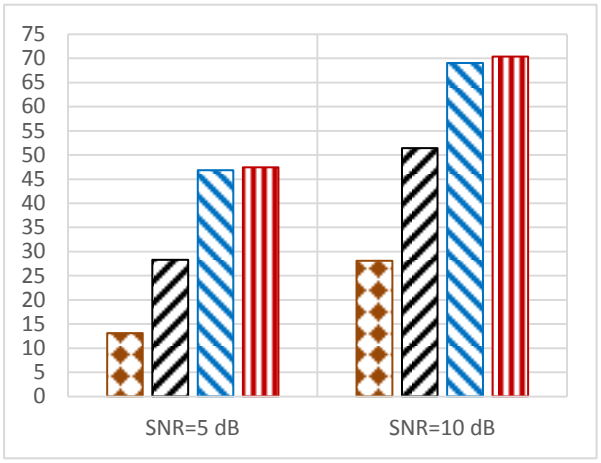


ج) مجموعه C

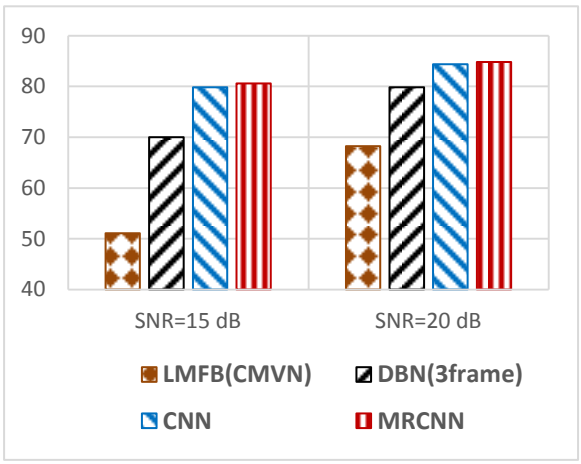
شکل ۶: میانگین دقت بازشناسی (0-20 DB) برحسب نوع نویز، برای CNN، MRCNN، LMFB+CMVN و DBN در حالت سه قاب



الف) نویز شدید



ب) نویز متوسط



ج) نویز ضعیف

شکل ۷: میانگین دقت بازشناسی برحسب نسبت سیگنال به نویز، برای CNN، MRCNN، LMFB+CMVN و DBN در حالت سه قاب

مراجع

[۱] فائزه بنی‌اردلان، احمد اکبری، بابک ناصرشریف، «حذف نویز و استخراج ویژگی‌های گلوگاه در سطح زیرباند توسط شبکه‌های خودرمزگذار عمیق برای بازشناسی گفتار»، *کنفرانس پردازش سیگنال و سیستم‌های هوشمند، دانشگاه صنعتی امیرکبیر*، دوره اول، ۱۳۹۴.

[۲] مجتبی غلامی‌پور، بابک ناصرشریف، «مقاوم‌سازی ویژگی‌های مل کپستروم نسبت به نویز با استفاده از شبکه باور عمیق»، *کنفرانس پردازش سیگنال و سیستم‌های هوشمند، دانشگاه صنعتی امیرکبیر*، دوره اول، ۱۳۹۴.

[۳] مجتبی حاجی آبادی، عباس ابراهیمی مقدم، حسین خوش بین، «حذف نویز صوتی مبتنی بر یک الگوریتم وفقی نوین»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۳، ص: ۱۳۹-۱۴۷، پائیز ۱۳۹۵.

[۴] مسعود گراوانچی‌زاده، ساناز قائمی سردرودی، «بهبود کیفیت گفتار مبتنی بر بهینه‌سازی ازدحام ذرات با استفاده از ویژگیهای ماسک‌گذاری سیستم شنوائی انسان»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۳، ص: ۲۸۷-۲۹۷، پاییز ۱۳۹۵.

[5] O. Abdel-Hamid, A. r. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1533-1545, 2014.

[6] S. Ikbal and H., Bourlard, "Phase autocorrelation derived robust speech features" *in Proc. ICASSP*, vol. 2, pp. 133-136, 2003.

[7] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," *in Proc. Interspeech*, pp. 2484-2488, 2015.

[8] O. Abdel-Hamid, A. r. Mohamed, H. Jiang and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in 2012 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4277-4280, 2012.

[9] J. Du, Q. Wang, T. Gao, Y. Xu, L. Dai and C.H. Lee, "Robust Speech Recognition with Speech Enhanced Deep Neural Networks", *Interspeech*, pp. 616-620, 2014.

[10] X. Feng, Y. Zhang and J. Glass. "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition" *In Proc. ICASSP*, pp. 1759-1763, 2014.

[11] A. Mohamed, G.E. Dahl and G. Hinton, "Acoustic Modeling Using Deep Belief Networks", Audio, Speech and Language Processing, *IEEE Transactions on*, Vol. 20, pp. 14-22, 2011.

[12] T. N. Sainath, A.-r. Mohamed, B. Kingsbury and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8614-8618, 2013.

[13] O. Abdel-Hamid, L. Deng and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *Interspeech*, pp. 3366-3370, 2013.

[14] J.-T. Huang, J. Li and Y. Gong, "An analysis of convolutional neural networks for speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4989-4993, 2015.

[15] D. Palaz,, R. Collobert and M. Magimai Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Interspeech*, pp. 1766-1770, 2013.

[16] D. Palaz, M. M. Doss and R. Collobert, "Convolutional Neural Networks-based continuous speech recognition using raw speech signal," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4295-4299, 2015.

[17] D. Palaz, and R. Collobert, "Analysis of cnn-based speech recognition system using raw speech as input," in *Proceedings of Interspeech*, 2015.

[18] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-r. Mohamed, G. Dahl, et al., "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39-48, 2015.

[19] Y. Takashima, T. Nakashika, T. Takiguchi and Y. Ariki, "Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pp. 1411-1415, 2015.

[20] A. Lozano-Diez, R. Zazo-Candil, J. Gonzalez-Dominguez, D. T. Toledano and J. n. Gonz?lez-Rodr?guez, "An end-to-end approach to language identification in short utterances using convolutional neural networks," *in INTERSPEEC*H, 2015.

[21] S. Thomas, S. Ganapathy, G. Saon and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2519-2523, 2014.

[22] R. Yeh, M. Hasegawa-Johnson and M. N. Do, "Stable and symmetric filter convolutional neural network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2652-2656, 2016.

[23] T. N. Sainath, O. Vinyals, A. Senior and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," *in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4580-4584, 2015.

[24] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson and O. Vinyals, "Learning the speech front-end with raw waveform cldnns*," in Proc. Interspeech*, 2015.

[25] T. N. Sainath, B. Kingsbury, A.-r. Mohamed and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *Automatic Speech Recognition and Understanding (ASRU)*, 2013 IEEE Workshop on, pp. 297-302, 2013.

[26] T. N. Sainath, B. Kingsbury, A.-r. Mohamed, G. E. Dahl, G. Saon, H. Soltau, et al., "Improvements to deep convolutional neural networks for LVCSR," *in Automatic Speech Recognition and Understanding (ASRU)*, 2013 IEEE Workshop on, pp. 315-320, 2013.

[27] Y. Zhao, X. Jin, X. Hu, "Recurrent convolutional neural network for speech processing.", *in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[28] Y. Zhang, W. Chan, N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition.", in *2017*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

[29] K. Choi, G. Fazekas, M. Sandler, K.Cho, "Convolutional recurrent neural networks for music classification", in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[30] Y. Qian, M. Bi, T. Tan and K. Yu, "Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263-2276, Dec. 2016.

[31] W. Dai, C. Dai, S. Qu, J. Li, S. Dos, " very deep convolutional neural networks for raw waveforms", in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[32] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *in ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[33] A. Agarwal, E. Akchurin, et al., "An Introduction to Computational Networks and the Computational Networks Toolkit", *microsoft technical reports*, 2016.

زیرنویس‌ها

[1] Phase autocorrelation feature
[2] Histogram equalization
[3] Deep Neural Networks
[4] Deep belief networks
[5] Deep auto-encoders
[6] Convolutional Neural Networks
[7] pooling
[8] Fully Connected layer
[9] spectrogram
[10] Convolutive Bottleneck Networks
[11] long short-term memory
[12] Log Mel Filter Bank
[13] Mean-pooling (average-pooling)
[14] Max-pooling
[15] stochastic pooling
[16] resolution
[17] rectified linear unit
[18] Microsoft Cognitive Toolkit
[19] Stochastic Gradient Descent
[20] mini batch
[21] momentum
[22] Epoch
[23] View