شبکهٔ عصبی پیچشی با پنجره‌های قابل تطبیق برای بازشناسی گفتار

تکتم ذوقی و محمد مهدی همایون‌پور*

آزمایشگاه پردازش هوشمند داده‌های چندرسانه‌ای، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

در حالی‌که سامانه‌های بازشناسی گفتار به‌طور پیوسته در حال ارتقا هستند و شاهد استفادهٔ گسترده از آن‌ها هستیم، اما دقت این سامانه‌ها فاصلهٔ زیادی نسبت به توان بازشناسی انسان دارد و در شرایط ناسازگار این فاصله افزایش می‌یابد. یکی از علل اصلی این مسئله، تغییرات زیاد سیگنال گفتار است. در سال‌های اخیر، استفاده از شبکه‌های عصبی عمیق در ترکیب با مدل مخفی مارکف، موفقیت‌های قابل توجهی در حوزهٔ پردازش گفتار داشته است. این مقاله به‌دنبال مدل‌کردن بهتر گفتار با استفاده از تغییر ساختار در شبکهٔ عصبی پیچشی عمیق است؛ به‌نحوی که با تنوّعاتِ بیان گویندگان در سیگنال گفتار منطبق‌تر شود. در این راه، مدل‌های موجود و انجام استنتاج بر روی آن‌ها را بهبود و گسترش خواهیم داد. در این مقاله با ارائهٔ شبکهٔ پیچشی عمیق با پنجره‌های قابل تطبیق، سامانهٔ بازشناسی گفتار را نسبت به تفاوت بیان در بین گویندگان و تفاوت در بیان‌های یک گوینده مقاوم خواهیم کرد. تحلیل‌ها و نتایج آزمایش‌های صورت‌گرفته بر روی دادگان گفتار فارس‌دات و TIMIT نشان داد که روش پیشنهادی خطای مطلق بازشناسی واج را نسبت به شبکهٔ پیچشی عمیق به‌ترتیب به میزان ۱/۲ و ۱/۱ درصد کاهش می‌دهد که این مقدار در مسئلهٔ بازشناسی گفتار مقدار قابل توجهی است.

واژگان کلیدی: بازشناسی گفتار، شبکهٔ عصبی عمیق، شبکهٔ عصبی پیچشی، پنجره‌های قابل تطبیق.

Adaptive Windows Convolutional Neural Network for Speech Recognition

Toktam Zoughi & Mohammad Mehdi Homayounpour*

Laboratory for Intelligent Multimedia Processing, Department of Computer Engineering & Information Technology, Amirkabir University of Technology, Tehran, Iran

Abstract

Although, speech recognition systems are widely used and their accuracies are continuously increased, there is a considerable performance gap between their accuracies and human recognition ability. This is partially due to high speaker variations in speech signal. Deep neural networks are among the best tools for acoustic modeling. Recently, using hybrid deep neural network and hidden Markov model (HMM) leads to considerable performance achievement in speech recognition problem because deep networks model complex correlations between features. The main aim of this paper is to achieve a better acoustic modeling by changing the structure of deep Convolutional Neural Network (CNN) in order to adapt speaking variations. In this way, existing models and corresponding inference task have been improved and extended.

Here, we propose adaptive windows convolutional neural network (AWCNN) to analyze joint temporal-spectral features variation. AWCNN changes the structure of CNN and estimates the probabilities of HMM states. We propose adaptive windows convolutional neural network in order to make the model more robust against the speech signal variations for both single speaker and among various speakers. This model can better

* Corresponding author

* نویسندهٔ عهده‌دار مکاتبات

model speech signals. The AWCNN method applies to the speech spectrogram and models time-frequency varieties.
This network handles speaker feature variations, speech signal varieties, and variations in phone duration. The obtained results and analysis on FARSDAT and TIMIT datasets show that, for phone recognition task, the proposed structure achieves 1.2%, 1.1% absolute error reduction with respect to CNN models respectively, which is a considerable improvement in this problem. Based on the results obtained by the conducted experiments, we conclude that the use of speaker information is very beneficial for recognition accuracy.

Keywords: Speech recognition, deep neural network, Convolutional neural network, Adaptive windows convolutional neural network.

۱- مقدمه

هدف بازشناسی خودکار گفتار، دریافت گفتار ضبط‌شده به‌عنوان ورودی و تولید متن کلمات بیان‌شده به‌عنوان خروجی است. کاربردهای فراوان و روبه‌رشدی در زمینهٔ بازشناسی گفتار مانند نیاز به واسط گفتار در تلفن همراه و برچسب‌گذاری دادگان گفتاری و ویدئویی برای جستجوی آن‌ها وجود دارد. برای تمام این کاربردها نیازمند سامانه‌های بازشناسی گفتار خودکارِ قابل اعتماد هستیم. نتایج ارزشمندی در زمینه آموزش تمایزی، از قبیل روش‌های تخمین بیشینه اطلاعات متقابل[1] [48]، آموزش دسته‌بندی با خطای طبقه‌بندی کمینه[2] [51-50] و همچنین روش‌هایی بر مبنای مدل مخفی مارکف[3] [52] به‌دست آمده است. علاوه بر روش های ذکرشده، روش‌های جدیدی برای مدل‌کردن سامانهٔ صوتی مانند میدان‌های تصادفی شرطی (CRF[4]) و روش‌های CRF قطعه‌ای [6,53] ارائه شده‌اند. با وجود این‌که مسألهٔ بازشناسی گفتار بیش از پنجاه سال است با وسعت زیاد مطالعه می‌شود [8-7]، هنوز راه به‌نسبه طولانی برای حل کامل آن در پیش است. هم‌اکنون فناوری بازشناسی خودکار گفتار ، به‌طور گسترده مورد استفاده قرار می‌گیرد. با این وجود فاصله قابل تأملی به‌خصوص در شرایط ناساگار بین دقت انسان و ماشین وجود دارد [10-9].

۱-۱- نیاز برای ارائهٔ مدل بهتر

بر خلاف پیشرفت‌های قابل توجه‌ای که در دهه‌های اخیر در بازشناسی خودکار گفتار انجام شده، تغییرات چندانی در ساختار مدل گفتار انجام نشده است. پژوهش‌گران به‌تدریج به کاستی‌های موجود در مدل‌های گفتاری اشاره داشته [11] و نواقص روش‌های استاندارد مدل‌کردن گفتار را مورد نقد قرار داده‌اند [13-12]. بیشتر روش‌های عملی در بازشناسی گفتار ساختارهای موجود را پذیرفته و بدون تغییر ساختار از طریق سایر روش‌ها کاهش ملایمی در نرخ خطا داشته‌اند. این روش‌ها شامل روش‌های تطبیق‌پذیر (مانند بیشینهٔ درست‌نمایی رگرسیون خطی[5] [14] و هنجارسازی طول مجرای گفتار[6] [16-15])، آموزش تمایزی [51 , 17] و بهبود مدل زبانی [18] هستند. در دسترس‌بودن دادگان خیلی بزرگ و توانایی استفاده از کل این دادگان بسیار سودمند است. بیشتر سامانه‌های بازشناسی گفتار نیاز به سامانه‌های ارتقایافتهٔ رایانه‌ای دارند. با ارتقای قدرت محاسباتی، روش‌های جستجوی قوی‌تری برای بازشناسی گفتار قابل استفاده هستند.

سؤال اساسی در مسائل بازشناسی گفتار چگونگی مدل‌کردن صوت برای استخراج ویژگی‌های صوتی و درنهایت برچسب‌زنی مناسب فریم‌های گفتاری است. در سال‌های اخیر، شبکه‌های عصبی عمیق در ترکیب با مدل مخفی مارکف، مدلی موفق برای حل مسائل در حوزه پردازش گفتار به ارمغان آورده است. هدفی که در این مقاله دنبال می‌شود در جهت بر طرف‌کردن یکی از چالش‌های اساسی در مسائل بازشناسی گفتار و ارتقای سامانهٔ بازشناسی خودکار گفتار است. ویژگی‌های متفاوت مجاری گفتاری گویندگان، سبک‌های بیان گوناگون گویندگان، سرعت بیان، نوفه‌های محیطی و غیره باعث تغییرات زیادی در سیگنال گفتاری انسان می‌شود. درنظرنگرفتن این تغییرات باعث کاهش نرخ بازشناسی شده و همچنین باعث می‌شود، مدل ما به این تغییرات مقاوم نباشد. این مقاله با استفاده از شبکهٔ عصبی پیچشی قصد دستیابی به مدلی مقاوم به این نوع تنوعات را دارد.

در بخش بعد، روش‌های پیشین مرتبط با مقاله را در زمینهٔ بازشناسی گفتار مرور خواهیم کرد. در بخش ۳ روش پیشنهادی ما به نام AWCNN[7] معرفی و توصیف و در بخش

[1] Maximum Mutual Information (MMI)
[2] Minimum Classification Error (MCE)
[3] Hidden Markov Model (HMM)
[4] Conditional Random Fields (CRF)

[5] Maximum Likelihood Linear Regression (MLLR)
[6] Vocal Tract Length Normalization (VTLN)
[7] Adaptive windows convolutional neural network (AWCNN)

۴ نتایج ارزیابی و مقایسهٔ روش پیشنهادی با سایر روش‌های مطرح در این زمینه و هم‌چنین در بخش ۵ جمع‌بندی و نتیجه‌گیری ارائه می‌شود.

۲- مرور: روش‌های مدل‌کردن صوت در بازشناسی گفتار

در دو دههٔ اخیر مدل‌های مخفی مارکف از روش‌های غالب برای کارهای تشخیص گفتار با دادگان بزرگ به‌شمار می‌آیند [19]. در کاربردهای گفتار به‌طورمعمول از مدل مخفی مارکف برای مدل‌کردن خاصیت زمانی گفتار و از مدل مخلوط گاوسی برای مدل‌کردن ویژگی‌های فریم مورد نظر استفاده می‌شود. راه دیگر برای مدل‌کردن ویژگی‌های یک فریم در هر حالتِ مدل مخفی مارکف، استفاده از شبکه عصبی عمیق است. این شبکه‌ها تعدادی از فریم‌ها را به‌عنوان ورودی می‌گیرند و احتمال پسین بر روی حالات مدل مخفی مارکف را به‌عنوان خروجی شبکه عصبی به‌دست می‌دهند. پارامترهایی که در مدل مخفی مارکف نیاز به تخمین دارد، عبارتند از احتمالات اولیه حضور در حالات، احتمالات انتقال از یک حالت به حالت دیگر و نیز احتمالات پیشین یک مشاهده به شرط حضور در یک حالت خاص است. در روش‌های گذشته مدل مخفی مارکف، احتمال مشاهدات توسط مدل مخلوط گاوسی تخمین زده می‌شود [19-21]. در ادامه، این مدل‌ها را با سرنام GMM-HMM[1] نشان می‌دهیم. توانایی این مدل‌ها به‌خاطر محدودیت‌های توزیع مدل مخلوط گاوسی محدود می‌شود. برای رفع این مشکلات، مدل‌هایی از قبیل میدان تصادفی شرطی[2] و HCRF[3] [6,34,53] ارائه شده‌اند که از مدل‌های لگاریتم خطی[4] برای جایگزینی GMM-HMM استفاده می‌کند. نتایج به‌دست‌آمده نشان داده است، کارایی این روش ها، بهتر از روش GMM-HMM نیست [40]. در روش‌های جدیدتر مدل مخلوط گوسی با شبکه عمیق جایگزین می‌شود که نتایج مطلوبی داشته است [19-21, 35-36].

مدل‌های عمیق، کارایی خود را برای کاربردهای مختلف به اثبات رسانده‌اند. این مدل‌ها برای کدکردن و دسته‌بندی داده‌های گفتار، متن و تصویر کارائی خوبی نشان داده‌اند [19-23]. این ویژگی‌ها باعث شده‌است، بیشتر روش‌های خودکار بازشناسی گفتار به سمت روش‌های یادگیری عمیق سوق پیدا کنند. در روش‌های کنونی به جای استفاده از روش GMM-HMM، از روشی ترکیبی بر مبنای شبکهٔ عصبی عمیق و مدل مخفی مارکف برای بازشناسی آوا استفاده شده است [19-21].

در این روش به جای مدل مخلوط گاوسی از شبکهٔ عصبی عمیق که به‌صورت تمایزی آموزش دیده است، استفاده می‌شود. ترکیب شبکهٔ عصبی و مدل مخفی مارکف برای کاربرد بازشناسی خودکار گفتار از اواخر دهه ۱۹۸۰ شروع شد [33]. ساختارهای مختلف و الگوریتم‌های آموزشی متفاوتی در این راستا ارائه شده است. از شبکهٔ عصبی برای تخمین احتمالات پسینِ حالات مدل مخفی مارکف استفاده شده است، که با سرنام DNN-HMM[5] خوانده می‌شود [32-33]. در این مدل، خروجی شبکه عصبی برای تخمین احتمال پسین حالت‌های مدل مخفی مارکف پیوسته به شرط داشتن مشاهدات آکوستیکی، آموزش می‌بینند.

مدل‌های شبکهٔ عصبی عمیق باقی‌مانده[6] نیز از مدل‌های بسیار جدید در کار بازشناسی گفتار هستند که دقت بسیار خوبی نیز تا کنون ارائه داده‌اند [3,42-44]. در این مدل‌ها، خروجی شبکهٔ عصبی باقی‌مانده برای تخمین احتمال پسین حالت‌های مدل مخفی مارکف پیوسته به شرط داشتن مشاهدات آکوستیکی، آموزش می‌بینند. در برخی از مدل‌های ارائه‌شدهٔ نوین دیگر از مدل HMM استفاده نمی‌شود. این مدل‌ها به‌صورت یک‌پارچه آموزش می‌بینند[7]. مدل‌های یک‌پارچه همانند LSTM[8] به‌گونه‌ای طراحی شده‌اند که بُعد زمان را در ادامهٔ مدل توسط رویکردهایی همچون، CTC مدل خواهند کرد و درنتیجه چون خطای کل سامانه یک‌پارچه محاسبه می‌شود، دقتی بالاتر خواهند داشت [45-47]. در مدل‌هایی که از HMM استفاده می‌شود، یک‌بار شبکهٔ عصبی عمیق آموزش داده می‌شود و بار دیگر مدل HMM آموزش می‌بین؛ چون HMM زمان را در گفتار مدل خواهد کرد. درنتیجه چون خطای دو مدل (HMM و شبکه عصبی) متفاوت محاسبه می‌شود و بنابراین اگر مدل نخست خوب آموزش نبیند مدل دوم نیز چون بر اساس مدل نخست آموزش می‌بیند، خطای آن وارد مدل HMM خواهد شد. بنابراین به‌طورمعمول مدل‌هایی که به‌صورت یک‌پارچه آموزش می‌بینند، بهتر از مدل‌های بر مبنای HMM هستند [4,46-47].

[1] Gaussian Mixture Model Hidden Markov Model (GMM-HMM)
[2] Conditional Random Fields (CRF)
[3] Hidden Conditional Random Fields (HCRF)
[4] Log-linear models

[5] Deep Neural Network Hidden Markov Network (DNN-HMM)
[6] Deep Residual Learning
[7] End to End
[8] Long Short-Term Memory (LSTM)

۲-۱- شبکهٔ عصبی عمیق

شبکهٔ عصبی عمیق، یک شبکهٔ عصبی مصنوعی و دارای بیش از یک لایهٔ مخفی بین ورودی و خروجی است [19]. هر واحد مخفی، j، از تابع لاجستیک[1] یا تابعی غیرخطی برای نگاشت تمام ورودی‌هایش از لایه پایین، V_i، و فرستادن آن به خروجی که یک مقدار عددی است، استفاده می‌کند:

$$h_j = \sigma(\sum_i v_i W_{i,j} + c_j) \qquad (۱)$$

در شبکه‌های عصبی عمیق احتمال بیش‌برازش[2] وجود دارد که برای جلوگیری از بیش‌برازش در شبکهٔ عصبی راه حل‌هایی متنوعی وجود دارد [37]. شبکهٔ عصبی با تعداد زیادی لایه دارای پارامترهای زیادی است، که تشکیل یک مدل انعطاف‌پذیر را می‌دهد. این خاصیت شبکهٔ عصبی عمیق را قادر می‌سازد، روابط پیچیده و غیرخطی بین ورودی و خروجی را مدل کند [21]. یادگیری بهینه شبکه عصبی با تعداد زیادی لایه، بسیار مشکل است [25]. کارایی شبکه‌های عصبی جلوسوی چندلایه در استخراج مؤلفه‌های اساسی غیرخطی، حذف نوفه و بهسازی سیگنال گفتار نشان داده شده است [29-31].

نتایج به‌دست‌آمده از بازشناسی گفتار نشان می‌دهد که شبکهٔ عصبی عمیق با تعداد زیادی لایهٔ پنهان، بهبود بیشتری نسبت به مدل مخلوط گاوسی بر روی دادگان مرجع داشته است [19-21]. در مراجع [19-21,32] شبکهٔ عصبی عمیق در ترکیب با مدل مخفی مارکف، بر روی دادگان بزرگ بازشناسی گفتار[3] استفاده شده است. نتایج این پژوهش‌ها افزایش چشم‌گیر دقت بازشناسی، نسبت به روش آموزشی-تمایزی GMM-HMM بر روی این دادگان بوده است [19-23,35-36].

روش دیگری که در پژوهش‌های کنونی رقیب اصلی روش DNN-HMMs شد، CNN-HMM[4] است. این روش از ترکیب قدرت شبکهٔ عصبی عمیق پیچشی با مدل مخفی مارکف استفاده می‌کند. در مراجع [35-36,38-39] نشان داده شده است که در دادگان چالش‌برانگیز تجاری که برای کاربردهای واقعی طراحی شده‌اند، CNN-HMM به‌طور برجسته بر GMM-HMM و DNN-HMM غلبه می‌کند که در ادامه به معرفی آن پرداخته می‌شود.

[1] Logistic Function
[2] Overfitting
[3] Large Vocabulary Speech Recognition

۲-۲- شبکهٔ عصبی پیچشی

شبکهٔ عصبی پیچشی همانند شبکهٔ عصبی، یک لایهٔ ورودی و تعداد دلخواهی لایه مخفی دارد [35]. همان‌طورکه در شکل (۱) مشاهده می‌شود، تفاوتی که شبکهٔ عصبی پیچشی با شبکهٔ عصبی دارد، در لایهٔ مخفی آن است. شکل (۲) یک لایهٔ مخفی از شبکهٔ عصبی پیچشی را نشان می‌دهد. همان‌طور که مشاهده می‌شود، هر لایه از شبکهٔ عصبی پیچشی شامل یک لایهٔ پیچشی و یک لایهٔ ادغام است.

همان‌طور که در شکل (۱) نشان داده شده است، هر کدام از ویژگی‌های ورودی $O_i(i = 1, \dots, I)$ بر اساس ماتریس وزن‌ها $w_{i,j}$ $(i = 1, \dots, I\,; j = 1, \dots, J)$ به چند واحد در لایهٔ پیچشی $Q_j(j = 1, \dots, J)$ متصل می‌شوند. این نگاشت توسط اپراتور پیچشی در مباحث پردازش سیگنال نشان داده می‌شود [36]. نگاشت به لایهٔ پیچشی را به‌صورت زیر می‌توان نشان داد:

$$q_{j,m} = \sigma\left(\sum_{i=1}^{I}\sum_{n=1}^{F} o_{i,n+m-1} w_{i,j,n} + w_{0,j}\right), \quad (j = 1, \dots, J) \qquad (۲)$$

در این رابطه $o_{i,m}$، m اُمین واحد از i اُمین ویژگی ورودی است. F اندازه فیلتر است، که تعیین‌کنندهٔ تعداد باندهای فرکانسی در نگاشت ویژگی‌های ورودی به لایهٔ پیچشی است. رابطه (۲) را به‌کمک اپراتور پیچشی به‌صورت رابطه (۳) می‌توان نوشت:

$$Q_j = \sigma\left(\sum_{i=1}^{I} O_i * w_{i,j}\right) \quad (j = 1, \dots, J) \qquad (۳)$$

تعداد دسته نگاشت‌های ویژگی در لایهٔ پیچشی به‌طورمستقیم، تعداد ماتریس‌های وزن محلی مورد استفاده در نگاشت پیچشی را تعیین می‌کنند. در حالت پایه، شبکهٔ عصبی پیچشی پنجره‌ای مستطیل‌شکل، به‌طول تعداد فریم‌های ورودی و عرض F را مورد بررسی قرار می‌دهد. عمل‌گر پیچشی به‌نسبت F، نگاشتی با ابعاد پایین‌تر تولید می‌کند.

شبکهٔ عصبی پیچشی از دو جهت با شبکهٔ عصبی عمیقِ استاندارد که ارتباطات کامل بین لایه‌های مخفی دارد، متفاوت است. نخست این‌که لایهٔ پیچشی، ورودی خود را تنها از یک فضای محلی دریافت می‌کند. بدین معنی که هر واحد،

[4] Convolutional Deep Neural Network Hidden Markov Network (CNN-HMM)

ویژگی‌های محلی ورودی را دریافت می‌کند. دوم این‌که، واحدهای لایهٔ مخفی به چند دسته تقسیم می‌شوند و واحدهایی که در یک دسته قرار دارند، دارای وزن‌های مشترکی هستند که ورودی خود را از مکان‌های متفاوتی از ورودی در لایهٔ پایینی دریافت می‌کنند.

همان‌طور که در شکل (۱) مشاهده می‌شود، عمل‌گر ادغام[1] به لایهٔ پیچشی اعمال می‌شود. لایهٔ ادغام[2] نیز دارای چند دسته ویژگی است. اندازهٔ دسته‌های ویژگی در این لایه از اندازهٔ دسته‌ها در لایهٔ پیچشی کوچکتر، ولی تعداد دسته‌ها در هر دو لایه یکسان است. هدف لایهٔ ادغام کاهش ابعاد هر دسته نگاشت ویژگی‌ها است. این بدین معناست که واحدهای این لایه، ویژگی‌های عمومی‌تری از لایه پایین را شامل می‌شوند و چون این عمومیت روی فرکانس‌های متفاوت قرار دارد، نسبت به تغییرات کوچک در باندهای فرکانسی مقاوم خواهند بود [36]. این کاهش با اعمال تابع ادغام به تعداد زیادی نواحی محلی در لایهٔ پیچشی با اندازه‌ای برابر اندازهٔ ادغام[3] به‌دست خواهد آمد. برای تابع ادغام از عمل‌گرهای بیشینه یا میانگین استفاده می‌شود. تابع ادغام به هر کدام از دستهٔ ویژگی‌ها به‌صورت جداگانه اعمال خواهد شد. زمانی‌که تابع ادغام[4] بیشینه مورد استفاده قرار می‌گیرد، لایهٔ ادغام به‌صورت رابطه (۴) تعریف می‌شود:

$$p_{i,m} = \max_{n=1}^{G} q_i(m-1)\times s + n \qquad (۴)$$

در این رابطه G اندازهٔ ادغام و s اندازهٔ شیفت و تعداد هم‌پوشانی دو پنجره در حین ادغام را بیان می‌کند [35].

۱-۲-۲- روال آموزش شبکهٔ عصبی پیچشی

آموزش وزن‌ها در لایهٔ پیچشی با اعمال تغییراتی در الگوریتم پس‌انتشار خطا امکان‌پذیر است. برای این‌که الگوریتم آموزش را در شبکهٔ پیچشی ارائه کنیم، در ابتدا عمل‌گر پیچشی را در رابطه (۳) به شکلی مشابه شبکهٔ عصبی عمیق نشان می‌دهیم، بنابراین الگوریتم آموزش، مشابه شبکهٔ عصبی عمیق استاندارد قابل اجرا است.

شکل (۳-الف) ماتریس وزن $\widehat{W}$ را نشان می‌دهد که این ماتریس تُنُک است. ماتریس $\widehat{W}$ از تکرار ماتریس پایه W در رابطه (۵) تشکیل شده است. ماتریس پایه W از ترکیب تمامی ماتریس‌های وزن محلی تشکیل شده است. این ماتریس W دارای I.F سطر است که در آن F نشان‌دهندهٔ اندازه فیلتر و هر باند شامل I سطر برای I دسته ویژگی ورودی و هم‌چنین W دارای J ستون است که نشان‌دهندهٔ وزن J دسته نگاشت ویژگی در لایه پیچشی است. یک بردار تک‌ردیفه $\hat{o}$ از تمامی ویژگی‌های ورودی به شکل رابطهٔ (۶) تشکیل شده است.

$$W = \begin{bmatrix} w_{1,1,1} & w_{1,2,1} & \cdots & w_{1,J,1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1,1} & w_{I,2,1} & \cdots & w_{I,J,1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1,2} & w_{I,2,2} & \cdots & w_{I,J,2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1,F} & w_{I,2,F} & \cdots & w_{I,J,F} \end{bmatrix} \qquad (۵)$$

$$\hat{o} = [v_1|v_2|\ldots|v_M] \qquad (۶)$$

در این رابطه v_m یک بردار سطری شامل مقادیر mاُمین باند فرکانسی برای تمام I ویژگی و M تعداد باندهای فرکانسی در لایهٔ ورودی است. بنابراین خروجی لایهٔ پیچشی که در رابطهٔ (۳) محاسبه شد، می‌تواند به شکل رابطه (۷) محاسبه شود:

$$\hat{q} = \sigma(\hat{o}\widehat{W}) \qquad (۷)$$

رابطه (۷) دارای شکل ریاضی مشابه شبکهٔ عصبی عمیق است. بنابراین وزن‌های لایهٔ پیچشی با استفاده از الگوریتم پس‌انتشار خطا به‌روزرسانی و برای به‌روزرسانی $\widehat{W}$ از رابطهٔ (۸) استفاده می‌شود:

$$\Delta\widehat{W} = \varepsilon.\hat{o}'e \qquad (۸)$$

نحوهٔ رفتار با وزن‌های مشترک در لایهٔ پیچشی با شبکهٔ عصبی عمیق که هیچ اشتراک‌گذاری در وزن‌های آن وجود ندارد، اندکی متفاوت است. در شبکهٔ پیچشی وزن‌های مشترک در حین به‌روزرسانی، مطابق رابطه (۹) جمع خواهند شد:

$$\Delta w_{i,j,n} = \sum_{m} \Delta\widehat{W}_{i+(m+n-2)\times I,j+(m-1)\times J} \qquad (۹)$$

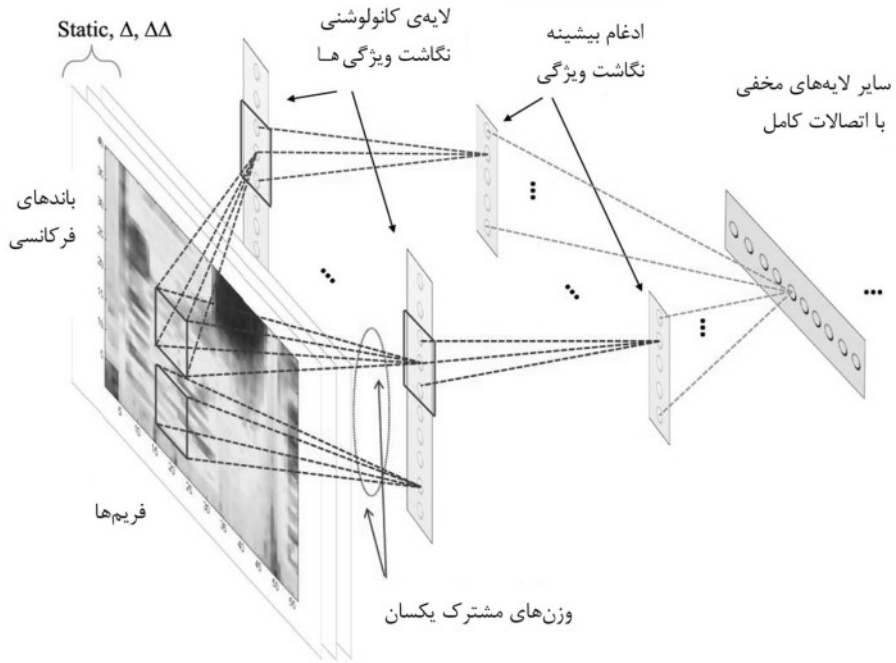
در این رابطه I و J به‌ترتیب نشان‌دهندهٔ تعداد نگاشت ویژگی‌ها در لایهٔ ورودی و پیچشی هستند. به‌علت آن‌که در لایهٔ ادغام هیچ وزنی نداریم، بنابراین نیازی به آموزش این لایه نیست؛ اما سیگنال خطا می‌بایست از طریق تابع ادغام به لایه‌های پایین انتقال یابد. در حالت ادغام بیشینه‌ها، سیگنال خطا از میان واحدهای ادغام، فقط به واحدی با بیش‌ترین میزان فعالیت انتقال می‌یابد. میزان خطای انتقال‌یافته به لایهٔ پیچشی مطابق رابطه (۱۰) قابل محاسبه است.

[1] Pooling operation
[2] Pooling ply
[3] Pooling size
[4] Max-pooling

$$e_{i,n}^{low} = \sum_{m} e_{i,m}.\delta(u_{i,m} + (m-1) \times s - n) \quad (۱۰)$$

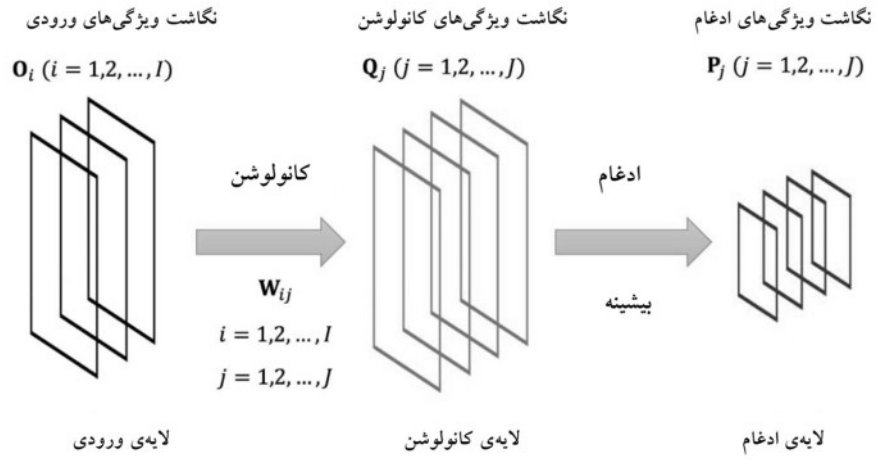
در این رابطه $\delta(x)$ تابع دلتا نامیده می‌شود و چنانچه x=1 باشد، دارای مقدار یک و در غیر این‌صورت مقدار آن صفر است. $u_{i,m}$ اندیس واحد با بیشترین مقدار از بین واحدهای ادغام است که به‌صورت رابطهٔ (۱۱) تعریف می‌شود:

$$u_{i,m} = argmax_{n=1}^{G}\ q_i(m-1) \times s + n \quad (۱۱)$$



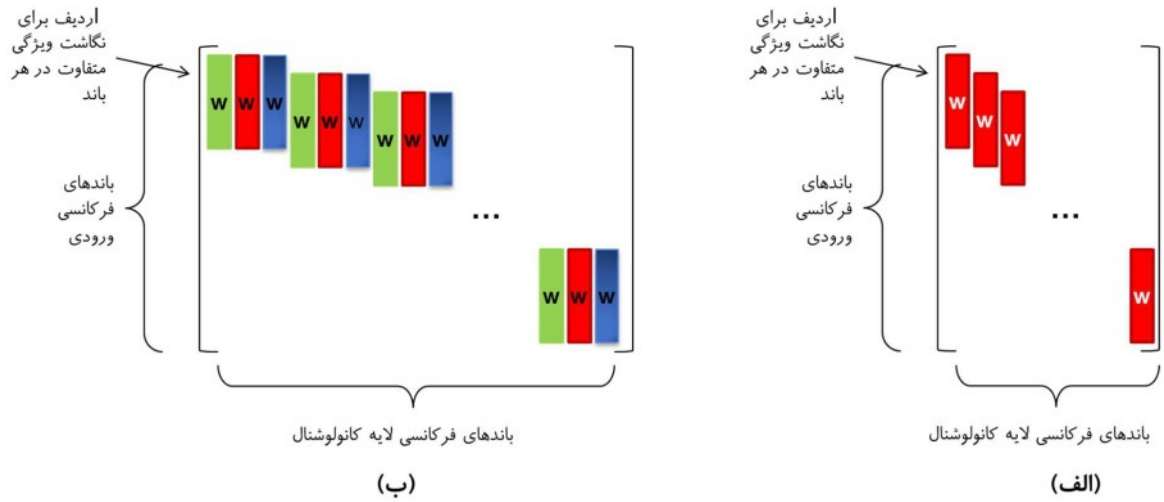
(شکل-۱): نمایی از شبکهٔ عصبی پیچشی، اپراتور پیچشی در طول باندهای مختلف فرکانسی اعمال شده است [39]

(Figure-1): A representation of Convolutional deep Neural Network (CNN), the convolution operator is applied along different frequency bands [39]



(شکل-۲): یک لایه از شبکهٔ عصبی پیچشی شامل یک لایهٔ پیچشی و یک لایه ادغام است [39]

(Figure-2): A convolutional neural network layer contains a convolution layer and one pooling layer [39]



(شکل-۳): نمایش اشتراک محدودشدهٔ وزن‌ها و اتصالات محلی در ماتریس وزن‌های لایهٔ پیچشی در قالب یک ماتریس تنک. عناصری که نشان داده نشده‌اند، صفر هستند.

(Figure-3): Limited weight sharing and local connectivity in the convolutional layers shown as sparse matrix. This representation does not show zero elements.

۳- روش پیشنهادی

در دههٔ کنونی شبکه‌های عمیق از جدیدترین و قوی‌ترین ابزارهای پردازش گفتار در مقایسه با سایر روش‌ها محسوب می‌شوند [25-19]. هدف اصلی این مقاله بهبود ساختار و برطرف‌کردن مشکلات آموزشی شبکهٔ پیچشی عمیق به‌منظور پردازش گفتار است.

شبکهٔ عصبی پیچشی با در نظرگرفتن ویژگی‌های محلی، به اشتراک‌گذاری وزن‌ها و ادغام بیشینه‌ها، باعث کاهش تعداد پارامترها برای آموزش شبکه می‌شود. درنهایت باعث مقاوم‌شدن مدل و کاهش بیش‌برازش خواهد شد. هر کدام از این ویژگی‌ها منجر به بهبود کارایی بازشناسی گفتار خواهد شد. شبکهٔ عصبی پیچشی با بررسی سیگنال در یک پنجرهٔ زمان-فرکانس بر روی طیف‌نگار[1] سیگنال گفتار، امکان درنظرگرفتن ویژگی‌های محلی را به‌صورت توأمان فراهم می‌کند. این مسأله باعث می‌شود که درصورت اطلاع از باندهای فرکانسی که تحت تأثیر نوفه قرار گرفته‌اند و یا لحظات زمانی نوفه‌ای‌بودن سیگنال، امکان بررسی متفاوت باندهای فرکانسی و لحظات زمانی مختلف فراهم شود.

ورودی شبکهٔ عصبی پیچشی می‌بایست دوبعدی در نظر گرفته شود، چون پنجره‌گذاری و سایر عملیات در این شبکه بر اساس دو بعد است. ورودی در نظر گرفته شده در مقالات [35-36,39] نمودار طیف‌نگار است. به جای نمودار طیف‌نگار از ورودی‌های دو بعدی دیگری نیز می‌توان استفاده کرد. در این مقاله نیز از طیف‌نگار سیگنال گفتار به عنوان ورودی استفاده می‌شود؛ زیرا طیف‌نگار سیگنال گفتار اطلاعات بُعد زمان و فرکانس را به‌صورت هم‌زمان می‌تواند نمایش دهد.

۳-۱- درنظرگرفتن تفاوت‌ها در بیان یک واج

یکی از مشکلات اساسی بازشناسی گفتار، تفاوت در بیان واج‌های یکسان توسط افراد مختلف و حتی بیان متفاوت یک واج یکسان به‌عنوان مثال /a/ توسط همان شخص است. بنابراین نمی‌توان به‌سادگی واج‌های یکسان را در دسته‌های یکسان قرار داد. روش پایه برای کار بازشناسی گفتار مدل GMM-HMM است [33]. مشکل اساسی این روش در نظر گرفتن احتمالات انتقال از یک حالت به حالت دیگر و نیز احتمالات پسین یک مشاهده به‌صورت فرمولی ثابت است. این امر باعث می‌شود در مقابل طول بیان متفاوت یک واج به‌خوبی عمل نکند [26,41]؛ اما روش‌هایی مبتنی بر مدل مخفی مارکوف مانند HSMM و ESHMM ارائه شده‌اند که حالات مدل HMM را گسترش می‌دهند [26,41]. HSMMها با استفاده از تابع چگالی احتمالِ طول یک حالت، می‌توانند طول گویش را مدل کنند [28]. در ESHMMها هر حالت با HMM دیگری جایگزین می‌شود؛ بنابراین طول تابع چگالی احتمال هر حالت برابر است با طول تمام توابع چگالی احتمالی که به آن حالت منتسب شده است [41]. برای درنظرگرفتن طول بیان‌های متفاوت در بازشناسی گفتار روش‌هایی بر پایه SCRF[2] مطرح هستند [27]. این روش‌ها در تلاش‌اند با

[1] Spectrogram

[2] Segmental Conditional Random Fields (SCRF)

استخراج ویژگی‌هایی که بیان‌گر طول بیان متفاوت هستند و افزودن آن به مدل SCRF، کارایی سامانه‌های بازشناسی را ارتقا بخشند.

روشی که در این مقاله بیان می‌شود، بر پایهٔ شبکهٔ عصبی پیچشی است. شبکهٔ عصبی پیچشی و به اشتراک گذاشتن وزن‌ها این امکان را به شبکه می‌دهد که از پنجره‌هایی با مقیاس‌های متفاوت استفاده شود. برای این منظور مطابق شکل (۴) پنجره‌هایی با طول متفاوت می‌توان در نظر گرفت؛ هر پنجره می‌تواند مدل‌کننده بیانی متفاوت از یک واج یکسان در حوزه زمان-فرکانس باشد. برای مثال پنجره با طول کم نشان‌دهندهٔ بیان سریع یک واج است. همچنین، پنجره با طول زیاد نشان‌دهندهٔ بیان آرام واج است. بنابراین گستردگی طول بیان‌ها با درنظرگرفتن طول پنجره متفاوت مدل می‌شوند. در ادامهٔ این مقاله شبکهٔ عصبی پیچشی با پنجره‌های قابل تطبیق با سرنام AWCNN[1] نامیده می‌شود؛ در صورتی‌که یک واج در بازهٔ زمانی طولانی‌تری بیان شود، پنجرهٔ بلندتری نسبت به بیان همان واج در زمان کوتاه‌تر نیاز دارد. در ادامه توضیح داده خواهد شد چگونه با این شیوه، تنوعات بیان‌های گوناگون در بیان یک واج مدل می‌شود.

AWCNN-HMM -۳-۲

شبکهٔ عصبی پیچشی مبنا (CNN)، شامل یک لایهٔ پیچشی و یک لایهٔ ادغام است. در CNN بعد از لایهٔ ورودی تعدادی لایهٔ پیچشی تعریف و در ادامهٔ آن تعدادی لایه با اتصالات کامل تعریف می‌شود. شکل (۴) ساختار کلی مدل پیشنهادی را نشان می‌دهد. این مدل مانند سایر شبکه‌های عمیق شامل لایهٔ ورودی، تعدادی لایهٔ پیچشی، لایهٔ ادغام و تعدادی لایه با اتصالات کامل است. ساختار شبکهٔ عصبی پیچشی با پنجره‌های قابل تطبیق (AWCNN) یک لایهٔ پیچشی و دو لایهٔ ادغام است. همچنین لایهٔ پیچشی AWCNN با شبکهٔ عصبی پیچشی معمول (CNN[2]) متفاوت است. در CNN، هر دسته در لایهٔ پیچشی دارای یک ستون است؛ اما همان‌طورکه در شکل (۴) دیده می‌شود، هر سطر در لایهٔ پیچشی AWCNN دارای چندین واحد[3] یا ستون است. هر واحد در این لایه، به پنجره‌ای با طول متفاوت تعلق دارد. هر پنجرهٔ طیف‌نگار برای طول بیان متفاوتی از آوای یک گوینده در نظر گرفته شده است. بنابراین به تعداد پنجره‌ها با طول متفاوت در لایهٔ پیچشی ستون خواهیم داشت. به‌عنوان مثال درصورتی‌که سه طول مختلف برای پنجره‌ها در نظر گرفته شود، به تبع آن سه ستون در لایهٔ پیچشی خواهیم داشت. nW اندیس پنجره‌ها است. ستون nWام ورودی خود را از پنجرهٔ nWام دریافت می‌کند. همان‌طورکه در شکل (۴) نشان داده شده است، ویژگی‌های ورودی $O_i(i = 1, \dots, I(nW))$، ابعاد متفاوت دارند که $I(nW)$ به معنای طول پنجره‌ای است که برای پنجرهٔ شماره nWام در نظر گرفته شده است. هر کدام از ویژگی‌هایِ ورودی O_i بر اساس ماتریس وزن‌ها $w_{i,j}$ $(i = 1, \dots, I\ ; j = 1, \dots, J)$ به چند واحد در لایهٔ پیچشی $Q_j(j = 1, \dots, J)$ متصل می‌شوند. نگاشت به لایهٔ پیچشی در مدل پیشنهادی را به‌صورت رابطهٔ (۱۲) می‌توان نشان داد:

$$q_{j,m,nW} = \sigma\left(\sum_{i=1}^{I(nW)}\sum_{n=1}^{F} o_{i,n+m-1}w_{i,j,n} + w_{0,j}\right), \quad (j = 1, \dots, J; nW = 1, \dots, NW) \qquad (۱۲)$$

در این رابطه $o_{i,m}$، mُمین واحد از iُمین ویژگی ورودی است. F اندازه فیلتر است، که تعیین‌کنندهٔ تعداد باندهای فرکانسی در نگاشت ویژگی‌های ورودی به لایهٔ پیچشی است. رابطه (۱۲) را به‌کمک اپراتور پیچشی به‌صورت رابطه (۱۳) می‌توان نوشت:

$$Q_{j,nW} = \sigma\left(\sum_{i=1}^{I(nW)} O_i * w_{i,j}\right) \quad (j = 1, \dots, J; nW = 1, \dots, NW) \qquad (۱۳)$$

تعداد دسته نگاشت‌های ویژگی در لایهٔ پیچشی به‌طور مستقیم تعداد ماتریس‌های وزن محلی مورد استفاده در نگاشت پیچشی را تعیین می‌کنند. در حالت پایه، شبکهٔ عصبی پیچشی، پنجره‌ای مستطیلی‌شکل و به‌طول I و عرض F را مورد بررسی قرار می‌دهد.

برای ارتباط بین لایهٔ پیچشی و لایهٔ ادغام میانگین، تابع ادغام به سطر نخست در دستهٔ یک از لایهٔ پیچشی اعمال می‌شود و میانگین این سطر به لایهٔ بعد انتقال پیدا خواهد کرد؛ سپس مطابق شکل (۴) این عمل‌گر به سطر دو از همین دسته اعمال می‌شود و این عمل‌گر به تمامی سطرها و سپس تمامی دسته‌ها به همین شیوه اعمال خواهد شد. هر عنصر در لایهٔ ادغام میانگین مدل پیشنهادی به‌صورت رابطه (۱۴) قابل محاسبه است:

$$t_{j,m} = r\sum_{nW=1}^{NW} q_{j,m,nW} \qquad (۱۴)$$

[1] Adaptive Windows Convolutional Neural Network (AWCNN)

[2] Convolutional neural network

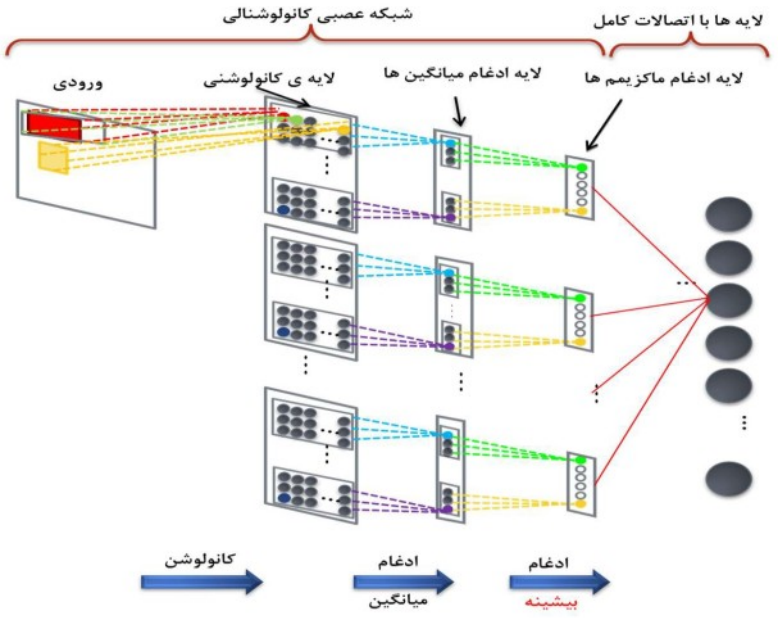
[3] Unit

در این رابطه r ضریب نرمال‌سازی است. برای عمل‌گر ادغام از عمل‌گرهای بیشینه یا میانگین استفاده می‌شود. تابع ادغام به هر کدام از دستهٔ ویژگی‌ها به‌صورت جداگانه اعمال خواهد شد.

در مرحلهٔ بعد تابع ادغام بیشینه به لایهٔ ادغام میانگن اعمال می‌شود. تابع ادغام بیشینه نیز به‌صورت رابطه (۱۵) تعریف می‌شود:

$$p_{j,m} = max_{n=1}^{G}\ t_j(m-1) \times s + n \qquad (۱۵)$$

این فرمول درواقع نحوهٔ محاسبه هر عنصر در لایهٔ ادغام بیشینه را نشان می‌دهد. در این رابطه G اندازهٔ پنجرهٔ ادغام و s اندازهٔ شیفت و تعداد هم‌پوشانی دو پنجره را در حین ادغام بیان می‌کند.

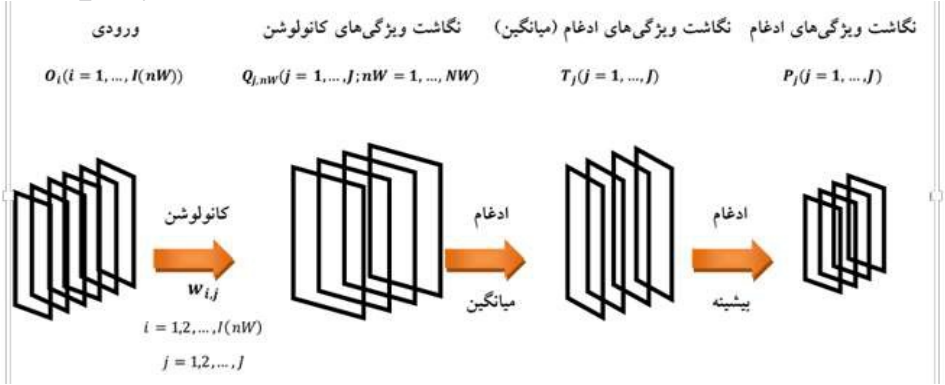


(شکل-۴): نمایش شبکه عصبی عمیق پیشنهادی AWCNN

(Figure-4): A representation of the proposed method (AWCNN)

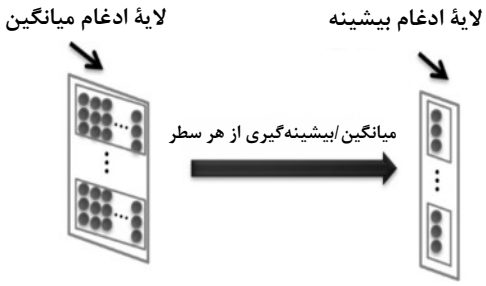
همان‌طور که در شکل (۵) نشان داده شده، اپراتور پیچشی به پنجره‌های در نظر گرفته‌شده در لایهٔ ورودی اعمال می‌شود و نتیجهٔ آن به لایه بعد منتقل می‌شود. این نگاشت توسط اپراتور پیچشی در مباحث پردازش سیگنال مطابق رابطه (۱۳) انجام می‌شود. در لایهٔ پیچشی با اعمال عملگر ادغام میانگین مطابق رابطه (۱۴) نتیجه به لایهٔ بعد انتقال پیدا می‌کند. در لایهٔ ادغام بیشینه نیز اپراتور ادغام بیشینه مطابق رابطه (۱۵) اعمال خواهد شد.

ارتباط بین لایهٔ ادغام میانگین و ادغام بیشینه نیز در شکل (۶) نشان داده شده است. بعد از لایهٔ ادغام بیشینه لایه‌های پیچشی دیگری را می‌توان تعریف کرد که خروجی لایهٔ پیچشی نخست به‌عنوان ورودی آن در نظر گرفته می‌شود و نحوه آموزش آن نیز به همین صورت است.



(شکل-۵): نمایی از یک لایه از شبکه پیشنهادی AWCNN که شامل یک لایهٔ پیچشی و دو لایهٔ ادغام است

(Figure-5): An illustration of one layer of the proposed method containing one convolution layer and two pooling layers



(شکل-۶): نحوه تغییر لایه ادغام میانگین به لایه ادغام بیشینه

(Figure-6): Mean-pooling layer conversion to max-pooling layer

همچنین همان‌طور که در شکل (۴) نشان دادیم، نتیجه این عملیات می‌تواند به لایه‌ای با اتصالات کامل یا دوباره به لایه پیچشی دیگری انتقال پیدا کند. نتیجه حاصل از لایه ادغام بیشینه می‌تواند به لایه پیچشی دیگری یا لایه با اتصالات کامل انتقال پیدا کند. درنهایت بعد از تعریف لایه‌ها در شبکه پیشنهادی AWCNN، کل شبکه به‌طور یک‌پارچه آموزش می‌بینند که نحوهٔ آموزش آن در بخش ۲-۲-۱ شرح داده شد. با توجه به تغییراتی که در مدل پیشنهادی نسبت به ساختار اولیه شبکه‌های عصبی پیچشی ایجاد شده است، این مدل اطلاعات مربوط به تفاوت در بیان در بین گویندگان و تفاوت در بیان‌های یک گوینده را می‌تواند مدل کند.

۴- ارزیابی و مقایسه

به‌منظور ارزیابی کارایی روش پیشنهادی، آن را با زبان پایتون و به‌کمک ابزار Kaldi و کتابخانه NumPy، Theano پیاده‌سازی و با الگوریتم‌هایی چون CNN-HMM [12] و DNN-HMM [13] مقایسه خواهیم کرد. Theano یک کتابخانه برای تعریف، بهینه‌سازی و اجرای عبارات ریاضی شامل آرایه‌های چندبعدی مانند ماتریس‌ها به‌صورت بهینه است. یکی از ویژگی‌های این کتابخانه تعامل با کتابخانهٔ NumPy و همچنین استفاده از GPUبرای تسریع عملیات مشتق‌گیری عبارات ریاضی است. این کتابخانه اغلب شامل جدیدترین الگوریتم‌ها و معماری شبکه‌های عمیق است. NumPy یکی از بسته‌های محاسبات علمی در پایتون است. این کتابخانهٔ پایتون امکان ایجاد آرایه‌های چندبعدی، مانند ماتریس‌ها را در اختیار ما قرار می‌دهد. یک آرایهٔ NumPy ویژگی‌هایی مانند، محاسبات ریاضی، تغییر اندازهٔ آرایه‌ها، مرتب‌سازی آرایه‌ها، انتخاب عناصر در آرایه‌ها و محاسبات آماری را در اختیار برنامه‌نویس می‌گذارد. علت مقایسهٔ این الگوریتم‌ها این است که هر دو در ترکیب با HMM استفاده شده‌اند و شبکه‌های مورد استفاده در هر دو الگوریتم از نوع عمیق هستند. برای ارزیابی این روش‌ها از دو مجموعه داده TIMIT و فارس‌دات استفاده کرده‌ایم. به‌علت حجم مناسب دادگان TIMIT اغلب روش‌های پیشنهادی جهانی بر روی آن ارزیابی شده‌اند. بنابراین امکان مقایسه روش ما با طیف وسیعی از دیگر روش‌ها فراهم می‌شود. ما الگوریتم‌ها را از دید خطای بازشناسی که عامل مهم در کار بازشناسی است، مقایسه کرده‌ایم. همه آزمایش‌ها بر روی رایانه‌ای دارای پردازنده 2.88GHz هشت هسته‌ای و حافظه RAM به ظرفیت ۳۲ گیگابایت انجام شده است. GPU استفاده شده در این رایانه GeForce GTX 780 است که حافظه این‌GPU، ۶۱۴۴ مگا بایت و دارای ۲۳۰۴ هسته پردازشی کودا[1] است.

۴-۱- ارزیابی روش پیشنهادی بر روی دادگان TIMIT

تمام آزمایش‌های این قسمت بر روی مجموعه هسته اصلی آزمون در دادگان TIMIT انجام شده است. ویژگی‌های این مجموعه‌داده‌ها در جدول (۱) نشان داده شده‌اند. همان‌گونه که این جدول نشان می‌دهد، این مجموعه‌داده‌ها ویژگی‌های متنوعی دارند. ستون آخر این جدول تعداد کل ساعات داده‌های آموزش و آزمون را نشان می‌دهد.

(جدول-۱): ویژگی‌های مجموعه داده‌های TIMIT مورد استفاده در آزمایش‌ها

(Table-1): Characteristics of TIMIT dataset used in the experimentations

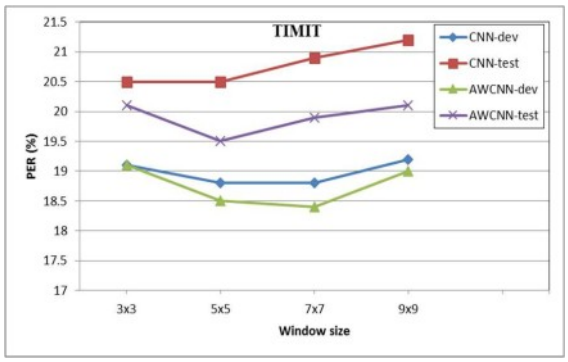
مجموعه داده	تعداد گویندگان	تعداد جملات	تعداد ساعات داده
آموزش	462	3696	3.14
هسته مرکزی آزمون	24	192	0.16
مجموعه کامل آزمون	168	1344	0.81

آزمایش‌های مختلفی برای ارزیابی روش پیشنهادی انجام شده است، که در ادامه به‌تفصیل بیان می‌شود. در ادامه به مقایسهٔ دو الگوریتم CNN-HMM و AWCNN-HMM از نظر خطای بازشناسی واج و توصیف نتایج آزمایش‌های انجام‌شده بر روی دادگان TIMIT می‌پردازیم.

در آزمایش نخست هر دو الگوریتم، CNN و AWCNN

[1] Cuda

مورد بررسی قرار گرفته‌اند. هر دو الگوریتم در ترکیب با مدل HMM استفاده شده است که در ادامه به‌ترتیب با سرنام CNN-HMM و AWCNN-HMM ذکر خواهند شد. در هر دو الگوریتم شبکهٔ عصبی با دو لایهٔ پیچشی و سه لایهٔ با اتصالات کامل آموزش می‌بینند. در الگوریتم CNN-HMM شبکهٔ عصبی عمیق پیچشی که در مرجع [35-36] اشاره شده، به‌عنوان روش پایه مورد استفاده قرار گرفته است. در این آزمایش ابتدا اندازهٔ پنجره در روش CNN-HMM، ۳×۳ است. همچنین اندازهٔ پنجره در روش AWCNN-HMM، سه اندازهٔ متفاوت ۳×۳ و ۵×۵ و ۷×۷ در نظر گرفته می‌شود؛ ولی برای سهولت در نمایش، در شکل‌ها اندازهٔ پنجرهٔ کوچک آن یعنی ۳×۳ نشان داده می‌شود. بنابراین در شکل (۷)، محور افقی با عدد، ۵×۵ برای روش AWCNN-HMM به معنی سه پنجره با اندازه‌های ۵×۵، ۷×۷ و ۹×۹ است. درحالی‌که عدد ۵×۵ در روی محور، برای روش CNN-HMM نمایان‌گر یک شبکهٔ پیچشی با اندازهٔ پنجرهٔ ۵×۵ است. در روش AWCNN-HMM سه پنجره با اندازه‌های متفاوت برای بررسی سه نوع بیان متفاوت از یک واج در نظر گرفته شده است. این سه پنجرهٔ متفاوت برای درنظرگرفتن سه طول متفاوت از یک بیان است. شکل (۷) به بررسی تأثیر افزایش اندازهٔ پنجره‌ها در خطای بازشناسی واج می‌پردازد. محور افقی در این شکل، اندازهٔ پنجره را نشان می‌دهد و محور عمودی خطای بازشناسی واج در دادگان TIMIT را نشان می‌دهند. هر چه مقدار خطای بازشناسی، کمتر شود، نتیجه مطلوب‌تر است؛ زیرا نشان می‌دهد الگوریتم به چه میزان برچسب صحیح به هر داده نسبت داده است. همان‌طورکه در شکل (۷) مشاهده می‌شود، درنظرگرفتن تفاوت بیان‌ها در الگوریتم پیشنهادی باعث افزایش کارایی آن شده است. همان‌طورکه در این شکل نشان داده شده است، الگوریتم ما به‌طورتقریبی در تمام موارد، بر روی مجموعه داده TIMIT نسبت به روش CNN-HMM نتایج بهتری کسب کرده است. علت به‌دست‌آمدن نتایج بهتر با توجه به ساختار مدل پیشنهادی درنظرگرفتن سه طول متفاوت از یک پنجره و بنابراین درنظرگرفتن یک واج با طول‌های متنوع است؛ بنابراین اگر این واج کمی کشیده‌تر یا اندکی سریع‌تر بیان شود در مدل پیشنهادی لحاظ خواهد شد. همان‌طورکه در شکل (۷) مشاهده می‌شود، در صورتی‌که اندازهٔ پنجره ۵×۵ باشد، نسبت به سایر اندازهٔ پنجره‌ها نتایج بهتری حاصل خواهد شد. چون با درنظرگرفتن اندازه‌های خیلی کوچک (به‌عنوان مثال ۳×۳) برای پنجره‌ها اطلاعات لازم برای یک واج در نظر گرفته نخواهد شد. از طرفی با درنظرگرفتن اندازهٔ خیلی بزرگ برای پنجره‌ها احتمال ورود اطلاعات غیرضروری به سامانه زیاد خواهد بود و باعث افزایش خطای بازشناسی واج خواهد شد.



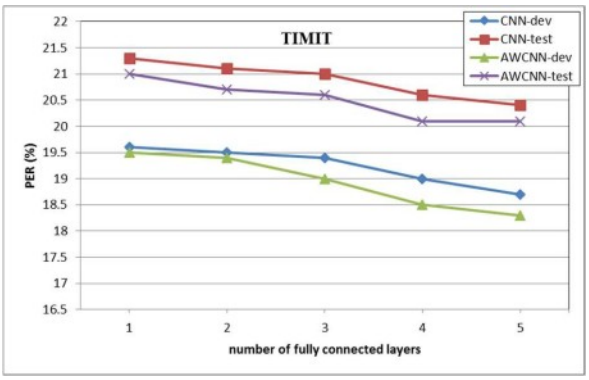
(شکل-۷): بررسی کارایی دو الگوریتم (CNN و AWCNN) و تاثیر افزایش اندازه پنجره‌ها در خطای بازشناسی واج

(Figure-7): The effect of the different window sizes on the error rates of CNN and AWCNN algorithms

در آزمایش دوم تأثیر افزایش تعداد لایه‌های با اتصالات کامل مورد بررسی قرار می‌گیرد. در این آزمایش نیز دو الگوریتم، CNN و AWCNN مورد بررسی قرار گرفته‌اند. هر دو الگوریتم شبکهٔ عصبی با دو لایهٔ پیچشی و تعداد متفاوتی لایه با اتصالات کامل آموزش می‌بینند. اندازهٔ پنجره روش AWCNN-HMM، سه اندازهٔ ۵×۵ و ۷×۷ و ۹×۹ است و اندازهٔ پنجره در روش CNN-HMM، ۵×۵ در نظر گرفته شده است. محور افقی نشان‌دهندهٔ تعداد لایه‌ها با اتصالات کامل و محور عمودی نشان‌دهندهٔ خطای بازشناسی واج است. برای مثال عدد دو در محور افقی شکل (۸) نشان می‌دهد دو لایه با اتصالات کامل در مدل پیشنهادی استفاده شده است. شکل (۸) به بررسی تأثیر افزایش تعداد لایه‌های با اتصالات کامل در خطای بازشناسی واج می‌پردازد. همان‌طورکه در شکل (۸) مشاهده می‌شود، با افزایش عمقِ شبکهٔ عصبی با اتصالات کامل نتایج بهتری حاصل خواهد و خطای بازشناسی کاهش می‌یابد. علت‌به‌دست‌آمدن نتایج بهتر با توجه به ساختار مدل پیشنهادی این است که با افزایش عمق، امکان بررسی اطلاعات و تجزیه وتحلیل داده‌ها به‌صورت سلسله‌مراتبی در لایه‌های بالاتر به‌وجود خواهد آمد. در لایه‌های پایین‌تر اطلاعات عمومی‌تر وجود دارد؛ اما لایه‌های بالاتر حاوی اطلاعات تمایزی هستند. همان‌طور که در شکل (۸) مشاهده می‌شود، افزایش تعداد لایه‌های مخفی سبب افزایش دقت بازشناسی خواهد شد.

در آزمایش بعد الگوریتم‌های متفاوت مانند: CNN-HMM، AWCNN-HMM و DNN-HMM را مقایسه

خواهیم کرد. این نتایج در جدول (۲) قابل مشاهده هستند. در جدول (۲)، LWS نشان‌دهندهٔ اشتراک‌گذاری محدودشدهٔ وزن‌ها است که توضیح مفصل آن در مراجع [35-36] آمده است. J: تعداد نگاشت ویژگی‌ها در لایه‌های ادغام، f: اندازهٔ پنجره پیچشی، p1: اندازهٔ پنجره ادغام نخست، p2: اندازهٔ پنجره ادغام دوم، s: میزان جابه‌جایی پنجره را نشان می‌دهند. پارامتر آخر نیز تعداد لایه‌ها با اتصالات کامل را نشان می‌دهد. به‌عنوان مثال ۳×۱۰۲۴ به معنای سه لایه با اتصالات کامل و دارای ۱۰۲۴ نرون در هر لایه است. تعداد لایه‌های شبکهٔ پیچشی به همراه لایه‌ها با اتصالات کامل در ستون سوم جدول نشان داده شده است. برای مثال 3 CON+ 4 FC به معنای وجود سه لایهٔ پیچشی و چهار لایهٔ با اتصالات کامل است.



(شکل-۸): مقایسهٔ دو الگوریتم (CNN و AWCNN) و بررسی تأثیر افزایش تعداد لایه‌های با اتصالات کامل

(Figure-8): The effect of different number of fully connected layers on CNN and AWCNN algorithms

(جدول-۲): خطای بازشناسی واج بر روی دادگان TIMIT

(Table-2): Phone Error Rate (PER%) on TIMIT dataset

ردیف	مدل	تعداد لایه‌ها	خطای داده آزمون (PER٪)	خطای داده توسعه (PER٪)
1	CNN-HMM {2×LWS(j:75 p1:3 s:1 f:5) + 3×1024}	2Conv + 3FC	20.5	18.8
2	AWCNN–HMM {2×LWS(j:75 p1:3 p2:4 s:1 f:5) + 3×1024}	2Conv + 3FC	19.5	18.5
3	DNN-HMM {pre-training + 4×1024}	4FC	22.9	21.0
4	CNN-HMM {2×LWS(j:147 p1:3 s:1 f:7)+10×1024}	2Conv + 10FC	20.8	19
5	AWCNN–HMM {2×LWS(j:147 p1:3 p2:4 s:1 f:7)+10×1024}	2Conv + 10FC	20.5	19
6	CNN-Res–HMM{6×LWS(j:75 p:1×1 s:1 f:3) + 3×1024}	6Cov + 3Fc	20.3	18.5
7	AWCNN-Res–HMM{6×LWS(j:75 p1:1×1 p2:1×3 s:1 f:5+ 3×1024}	6Cov + 3Fc	19.4	18.0

در این آزمایش برای روش پیشنهادی هر دو ادغام، بیشینه در نظر گرفته شده است. در جدول (۲) از مقایسه سطر یک و دو مشاهده می‌شود روش پیشنهادی AWCNN-HMM، خطای مطلق روش مطرح CNN-HMM را به میزان یک درصد کاهش داده است؛ قابل ذکر است، هر دو مدل دارای دو لایهٔ پیچشی و سه لایه با اتصالات کامل هستند. این میزان بهبود، قابل قبول برای کار بازشناسی گفتار به‌شمار می‌آید. همچنین با توجه به سطرهای دو و سه در جدول، روش پیشنهادی با دو لایهٔ پیچشی و سه لایه با اتصالات کامل، به میزان ۳/۴ درصد کاهش خطای مطلق نسبت به روش DNN-HMM با چهار لایه را داشته است. سطرهای چهار و پنج در این جدول نشان می‌دهد، درصورتی‌که تعداد لایه‌های با اتصالات کامل به ده برسد، همچنان روش AWCNN-HMM بر روش CNN-HMM برتری دارد؛ ولی افزایش زیاد تعداد لایه‌ها سبب می‌شود گردش اطلاعات به‌خصوص در لایه‌های بالایی به‌خوبی صورت نپذیرد. همچنین مقایسهٔ سطرهای دو و پنج نشان می‌دهد، الگوریتم با ده لایه اتصالات کامل بهتر از سه لایه نشده است؛ اما درصورتی‌که از روش‌های اصلاحی همچون پیش‌آموزش استفاده شود [5,19-20] خطای این روش‌ها کاهش خواهد یافت. در سطرهای شش و هفت جدول به بررسی تأثیر اعمال شبکهٔ عصبی عمیق باقیمانده[1] به مدل پیشنهادی و همچنین مدل CNN می‌پردازیم. شبکهٔ عصبی باقی‌مانده با سرنام Res به CNN و AWCNN اضافه شده است. با مقایسهٔ سطرهای شش و هفت مشاهده می‌شود با اعمال مدلِ AWCNN-Res-HMM با شش لایهٔ پیچشی و سه لایه با اتصالات کامل، خطای مطلق نسبت به CNN-Res-HMM با ساختار مشابه، ۰/۹ درصد کاهش می‌یابد. از این آزمایش و جدول (۲) می‌توان نتیجه گرفت روش پیشنهادی در ترکیب با مدل باقی‌مانده، AWCNN-Res-HMM، با شش لایهٔ پیچشی و سه لایه با اتصالات کامل، بهترین نتیجه نسبت به سایر ساختارها را به‌دست می‌دهد.

[1] Deep Residual Learning

۲-۴- ارزیابی روش پیشنهادی بر روی دادگان فارس‌دات

برای ارزیابی عملکرد روش پیشنهادی در زبان فارسی، تعدادی آزمایش بر روی دادگان فارس‌دات [1,49] انجام شده است. این دادگان مشابه دادگان TIMIT هستند. دادگان فارس‌دات شامل سیگنال گفتاری از ۳۰۴ گویندهٔ زن و مرد است [1]. مشخصات این دادگان در جدول(۳) آورده شده است. همان‌طور که در جدول (۳) مشاهده می‌شود، از این دادگان ۳۹۹۴ جمله برای آموزش، ۴۷۵ جمله به‌عنوان دادگان توسعه و ۲۸۷ جمله به‌عنوان دادگان آزمون در نظر گرفته شده است [2]. برای پیاده‌سازی روش‌های پایه بر روی فارس‌دات از جعبه ابزار کلدی استفاده شده است [2].

(جدول-۳): ویژگی‌های مجموعه داده‌های فارس‌دات مورد استفاده در آزمایش‌ها [2]

(Table-3): Characteristics of FarsDat dataset used in the experimentations [2]

مجموعه داده	تعداد گویندگان	تعداد جملات
آموزش	224	3994
توسعه	50	475
آزمون	30	287

در جدول (۴) به مقایسهٔ الگوریتم‌های متفاوت مانند: CNN-HMM، AWCNN-HMM و DNN-HMM بر روی دادگان فارس‌دات می‌پردازیم. شاخصه‌های موجود در جدول (۴) مشابه جدول (۲) می‌باشند. تعداد لایه‌های شبکهٔ پیچشی به‌همراه لایه‌ها با اتصالات کامل در ستون سوم جدول نشان داده شده است. در این آزمایش نیز برای روش پیشنهادی، هر دو ادغام، بیشینه در نظر گرفته شده است. در جدول (۴) از مقایسه سطرهای یک و دو مشاهده می‌شود روش پیشنهادی AWCNN-HMM، خطای مطلق روش مطرح CNN-HMM را به میزان ۰/۵ درصد کاهش داده است؛ قابل ذکر است هر دو مدل دارای هفت لایهٔ پیچشی و سه لایه با اتصالات کامل هستند. همچنین با توجه به سطرهای دو و سه در جدول، روش پیشنهادی با هفت لایهٔ پیچشی و ۳ لایه با اتصالات کامل، به میزان ۰/۷ درصد کاهش خطای مطلق نسبت به روش DNN-HMM با ده لایه را داشته است. در سطرهای چهار و پنج جدول به بررسی تأثیر اعمال شبکهٔ عصبی عمیق باقی‌مانده به مدل پیشنهادی و همچنین مدل CNN می‌پردازیم. شبکهٔ عصبی باقی‌مانده با سرنام Res به CNN و AWCNN اضافه شده است. با مقایسهٔ سطرهای چهار و پنج مشاهده می‌شود با اعمال مدلِ AWCNN-Res-HMM با هفت لایهٔ پیچشی و سه لایه با اتصالات کامل، خطای مطلق نسبت به CNN-Res-HMM با ساختار مشابه، ۰/۸ درصد کاهش می‌یابد.

در سطر شش جدول (۴)، مدلِ AWCNN-Res-HMM با ۱۹ لایه پیچشی، ۰/۲ درصد خطای مطلق کمتری نسبت به AWCNN-Res-HMM با ۷ لایه پیچشی دارد. که این مسأله نشان می‌دهد با توجه به کوچک‌بودن دادگان افزایش تعداد لایه‌های پیچشی باعث بهبود مختصری در آموزش شبکه و همچنین بهبود نتایج می‌شود. درصورتی‌که حجم دادگان بیشتر باشد، انتظار می‌رود افزایش لایه‌های شبکه باعث بهبود بیشتری در نتایج شود.

همان‌طور که در آزمایش‌ها دیده می شود، الگوریتم پیشنهادی AWCNN، در بیشتر موارد خطای بازشناسی واج کمتری نسبت به روش CNN دارد. این مسأله در همهٔ آزمایش‌ها صادق است. آزمایش‌ها نشان می‌دهند، الگوریتم پیشنهادی دقت بازشناسی بالاتری نسبت به روش‌های مطرح دارد، چون اندازه‌های متفاوت در پنجره‌ها امکان بررسی طول متفاوت بیان‌ها و تنوعات بیان آواها را فراهم خواهد کرد. با آمدن هر واج جدید، عملیات مدل‌کردن واج با پنجره‌هایی متفاوت به‌طور هم‌زمان صورت می‌پذیرد. علاوه‌براین در الگوریتم پیشنهادی، درصورتی‌که یک آوا کشیده‌تر یا سریع‌تر بیان شود با درنظرگرفتن پنجره‌هایی به‌ترتیب بلندتر و یا کوتاه‌تر طول آوا در مدل لحاظ خواهد شد.

ستون هفت از جدول (۴) زمان آموزش مدل‌های عمیق با ساختار متفاوت را مورد بررسی قرار می‌دهد. همان‌طورکه مشاهده می‌کنیم، زمان آموزش روش پیشنهادی نسبت به سایر روش‌ها بیشتر است؛ اما به‌طورمعمول برای مقایسه الگوریتم‌ها زمان **آزمون** را مد نظر قرار می‌دهند. چراکه در هنگام **آموزش** پارامترها و وزن‌ها آموزش داده می‌شود و بعد از مرحلهٔ آموزش پارامترهای مدل و وزن‌ها دیگر به‌روزرسانی نمی‌شوند. بنابراین بعد از آموزش مدل و پارامترها فقط از مدلِ آموزش داده شده، استفاده می‌کنیم. همان‌طور که در جدول (۴) ستون شش مشاهده می‌شود، زمان آزمون تمام روش‌ها به‌جز DNN بسیار نزدیک به هم است. به‌عنوان مثال با مقایسهٔ سطرهای چهار و پنج جدول مشاهده می‌شود، زمان آزمون برای مدل پیشنهادی سه دقیقه بیشتر از زمان آزمون CNN-Res-HMM با ساختار مشابه، است که این مقدار ناچیز است.

(جدول-۴): خطای بازشناسی واج بر روی دادگان فارس‌دات

(Table-4): Phone Error Rate (PER%) on FarsDat dataset

ردیف	مدل	تعداد لایه‌ها	خطای داده آزمون (٪PER)	خطای داده توسعه (٪PER)	زمان آزمون (min)	زمان آموزش (min)
1	CNN-HMM {7×LWS(j:75 p1:3 s:1 f:5) + 3×1024}	7Conv + 3FC	21.7	21.5	52	144.6
2	AWCNN–HMM {7×LWS(j:75 p1:3 p2:4 s:1 f:5) + 3×1024}	7Conv + 3FC	21.2	21.9	59	258
3	DNN-HMM {pre-training + 10×1024}	10FC	21.9	22.2	1	21
4	CNN-Res–HMM{7×LWS(j:75 p:1×1 s:1 f:3) + 3×1024}	7Cov + 3Fc	21.2	21.2	57	191
5	AWCNN-Res–HMM{7×LWS(j:75 p1:1×1 p2:1×3 s:1 f:5+ 3×1024}	7Cov + 3Fc	20.4	20.3	60	410
6	AWCNN-Res–HMM{19×LWS(j:75 p1:1×1 p2:1×3 s:1 f:5) + 3×1024}	19Cov + 3Fc	20.2	19.8	89	630

۵- نتیجه‌گیری

شبکه‌های عصبی عمیق، از بهترین ابزارهای مدل‌سازی صوت هستند. روش‌های عمیق در ترکیب با مدل مخفی مارکف تأثیر شگرفی در افزایش دقت بازشناسی گفتار دارند. در این مقاله روش جدیدی با تغییر ساختار شبکهٔ عصبی پیچشی ارائه شد تا احتمالات پسین و سایر احتمالات مدل مخفی مارکف توسط این روش جدید تخمین زده شود. بازشناسی خودکار گفتار به‌علت تغییرات ویژگی‌های گویندگان و تغییرات زیاد سیگنال گفتار، کار بسیار دشواری است. در این مقاله با ارائهٔ نسخه جدید از شبکهٔ عصبی پیچشی سامانهٔ بازشناسی گفتار نسبت به تغییرات ویژگی‌های گویندگان و تفاوت در نحوه بیان آن‌ها مقاوم می‌شود. همچنین در این مقاله، گفتار با استفاده از شبکهٔ عصبی پیچشی با پنجره‌های قابل تطبیق در ترکیب با مدل مخفی مارکف مدل می‌شود. در اینجا شبکه عصبی پیچشی با تغییرات پیشنهادشده بر روی طیف‌نگار سیگنال گفتار اعمال شد. تغییرات پیشنهادی منجر شد که مدل به تغییرات بیان‌های متفاوت یک واج توسط گویندگان مختلف و تغییرات زمانی و فرکانسی جزئی که در هنگام بیان واج یکسان به‌وجود می‌آید، تا حد زیادی مقاوم شود. آزمایش‌های صورت‌گرفته بر روی مجموعه داده‌ها نشان‌دهندهٔ کارایی روش پیشنهادی نسبت به روش‌های ارائه‌شدهٔ قبلی است. نتیجه‌ای که از این آزمایش‌ها و تحلیل‌ها حاصل می‌شود. منعطف‌بودن مدل پیشنهادی نسبت به تغییرات بیان‌های متفاوت و درنظرگرفتن آن‌هاست که منجر به ارائهٔ مدل بهتری نسبت به مدل‌های گذشته شد.

۶- مراجع

6- References

[۱] ج .شیخ زادگان ،م. بی‌جن خان، "دادگان گفتاری زبان فارسی"، *دومین کارگاه پژوهشی زبان فارسی و رایانه، ۱۳۸۵، ص ۲۶۱-۲۴۷.*

[1] Bi Jan Khan, J. Sheykhzadegan, “Persian speech dataset”, in *Machine Translation in Persian*, 2006, pp. 261-247.

[۲] ب .باباعلی، "پایه گذاری بستری نو و کارآمد در حوزه بازشناسی گفتار فارسی"، *پردازش علائم و داده‌ها، دوره ۱۳، ش. ۳، ص. ۱-۱۳، ۱۳۹۴.*

[2] B. BabaAli, “A state-of-the-art and efficient framework for Persian speech recognition”, *Signal and Data Processing,* Vol. 13, No. 3, pp. 1-13, 2015.

[۳] س .ز .سیدصالحی و س.ع. سیدصالحی، "بهبود مدل تفکیک‌کننده منیفلدهای غیرخطی به‌منظور بازشناسی چهره با یک تصویر از هر فرد"، *پردازش علائم و داده‌ها، دوره ۱۲، ش. ۱، ص. ۳-۱۶، ۱۳۹۴.*

[3] S. Z. Seyyedsalehi, and A. Seyyedsalehi, "Improving the nonlinear manifold separator model to the face recognition by a single image of per person." *Signal and Data Processing*, Vol. 12, No.1, pp. 3-16, 2015.

[۴] ز .انصاری و ع .سید صالحی، "معرفی شبکه های عصبی پیمانه ای عمیق با ساختار فضایی-زمانی دوگانه جهت بهبود بازشناسی گفتار پیوسته فارسی"، *پردازش علائم و داده‌ها، دوره ۱۳، ش. ۱، ص. ۳۹-۵۶، ۱۳۹۵.*

[4] Z. Ansari, and A. Seyyedsalehi, "Deep Modular Neural Networks with Double Spatio-temporal Association Structure for Persian Continuous Speech Recognition." *Signal and Data Processing*, Vol. 13, No.1, pp. 39-56, 2016.

[۵] س .ز .سیدصالحی و س .ع .سیدصالحی، "روش پیش تعلیم سریع بر مبنای کمینه‌سازی خطا برای همگرائی یادگیری شبکه‌های عصبی با ساختار عمیق"، *پردازش علائم و داده‌ها*، دوره ۱۰، ش. ۱، ص. ۱۳-۲۶، ۱۳۹۲.

[5] S. Z. Seyyedsalehi, and A. Seyyedsalehi, "A new fast pre training method for training of deep neural network." *Signal and Data Processing*, Vol. 10, No.1, pp. 13-26, 2013.

[6] Y. Hifny and S. Renals, "Speech recognition using augmented conditional random fields," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 2, pp. 354–365, 2009.

[7] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic Recognition of Spoken Digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, p. 637–642, 1952.

[8] L. Rabiner and B. Juang, Fundamentals of Speech Recognition: Prentice Hall, vol. 22. 1993.

[9] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 7, pp. 1–15, 1997.

[10] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Communication*, vol. 49, no. 5, pp. 336–347, 2007.

[11] M. Ostendorf, "Moving Beyond the 'Beads-on-a-String' Model of Speech," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 1999, pp. 79–83.

[12] H. Bourlard, H. Hermansky, and N. Morgan, "Towards increasing speech recognition error rates," *Speech Communication*, vol. 18, no. 3, pp. 205–231, 1996.

[13] N. Morgan, Q. Zhu, and A. Stolcke, "Pushing the envelope-aside," *Signal Processing Magazine*, vol. 22, no. 5, pp. 81–88, 2005.

[14] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[15] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, vol. 1, pp. 356–1996.

[16] L. Welling, S. Kanthak and H. Ney, "Improved methods for vocal tract normalization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999, vol. 2, pp. 761–764.

[17] D. Povey. Discriminative Training for Large Vocabulary Speech Recognition. PhD thesis, Cambridge University, 2003.

[18] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 1996, pp. 310–318.

[19] G. E. Dahl, D. Yu, L. Deng and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[20] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[21] A. R. Mohamed, G. Hinton and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4273–4276.

[22] R. Salakhutdinov and G. Hinton, "An Efficient Learning Procedure for Deep Boltzmann Machines," *Neural Computation*, vol. 24, no. 8, pp. 1967–2006, 2012.

[23] R. Salakhutdinov, "Learning deep generative models," PHD thesis, Toronto, Ont., Canada, 2009.

[24] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[25] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[26] P. Ramesh and J. G. Wilpon, "Modeling state durations in hidden Markov models for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992, vol. 1, pp. 381–384.

[27] P. N. Justine T. Kao, Geoffrey Zweig, "Discriminative duration modeling for speech recognition with segmental conditional random fields," in *ICASSP*, 2011. PP. 4476-4479.

Transactions on Speech and Audio Processing, vol. 22, no. 10, pp. 1533–1545, 2014.

[40] G. Heigold, "A log-linear discriminative modeling framework for speech recognition," PhD dissertation, Aachen, Germany, 2010.

[41] M. Russell and A. Cook, "Experimental evaluation of duration modelling techniques for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1987, vol. 12, pp. 2376–2379.

[42] H. Lee and H. Kwon, "Going Deeper with Contextual CNN for Hyperspectral Image Classification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4843–4855, 2017.

[43] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[45] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2013, no. 3, pp. 6645–6649.

[46] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, 2016, pp. 167–174.

[47] W. Song and J. Cai, "End-to-End Deep Neural Network for Automatic Speech Recognition," *CS224d: Deep Learning for Natural Language Processing*, pp. 1–8, 2015.

[48] S. Kapadia, V. Valtchev and S. J. Young, "MMI training for continuous phoneme recognition on the TIMIT database," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993, vol. 2, pp. 491–494.

[49] M. Bijankhan, J. Sheikhzadegan, and M. R. Roohani, Y. Samareh, "FARSDAT- the speech database of farsi spoken language," in *proceedings Australian conference on speech science and technology*, 1994, vol. 2, pp. 826–830.

[50] B. H. Juang, W. Chou, and C. H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[51] E. McDermott, T. J. Hazen, J. Roux, A. Nakamura and S. Katagiri, "Discriminative

[28] S. Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2. pp. 215–243, 2010.

[29] S. J. Rennie, P. Fousek, and P. L. Dognin, "factorial hidden restricted boltzmann machines for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4297–4300.

[30] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7596–7599.

[31] A. Maas and Q. Le, "Recurrent Neural Networks for Noise Reduction in Robust ASR.," in *Interspeech*, 2012, pp. 3–6.

[32] H. Bourlard and N. Morgan, "Continuous speech recognition by connectionist statistical methods," *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 893–909, 1993.

[33] A. J. Robinson, G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, and D. A. G. Williams, "Connectionist speech recognition of Broadcast News," *Speech Communication*, vol. 37, no. 1–2, pp. 27–45, 2002.

[34] Y. H. Sung and D. Jurafsky, "Hidden conditional random fields for phone recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding,* 2009, pp. 107–112.

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

[36] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-R. Mohamed, G. Dahl, and B. Ramabhadran, "Deep Convolutional Neural Networks for Large-scale Speech Tasks.," *Neural networks*, vol. 64, pp. 39–48, Sep. 2014.

[37] T. N. Sainath, B. Kingsbury, H. Soltau and B. Ramabhadran, "Optimization techniques to improve training speed of deep neural networks for large speech tasks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 11, pp. 2267–2276, 2013.

[38] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representtations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, vol. 2008, pp. 1–8.

[39] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE*

Training for Large-Vocabulary Speech Recognition Using Minimum Classification Error," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 203–223, 2007.

[52] F. Sha and L. Saul, "Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, vol. 1, pp. 265–268.

[53] G. Zweig, P. Nguyen, D. Van Compernolle, K. Demuynck, L. Atlas, P. Clark, G. Sell, M. Wang, F. Sha, H. Hermansky, D. Karakos, A. Jansen, S. Thomas, S. Bowman and J. Kao, "Speech recognition with segmental conditional random fields," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5044–5047.

تکتم ذوقی تحصیلات خود را در مقطع کارشناسی ارشد رشته مهندسی کامپیوتر (هوش مصنوعی) در دانشگاه شیراز (۱۳۸۸) به پایان رساند. وی هم اکنون دانشجوی مقطع دکترای کامپیوتر (هوش مصنوعی) در دانشگاه صنعتی امیرکبیر است. موضوعات مورد علاقه ایشان بازشناسی گفتار، پردازش سیگنال، پردازش صوت و پردازش تصویر است.

نشانی رایانامهٔ ایشان عبارت است از:

tzoughi@aut.ac.ir

محمدمهدی همایون‌پور تحصیلات خود را در مقطع کارشناسی در رشته برق (الکترونیک) در دانشگاه صنعتی امیرکبیر (۱۳۶۶) وکارشناسی ارشد را در رشته برق (مخابرات) در دانشگاه خواجه نصیرالدین طوسی (۱۳۶۹) و دکترای خود را در رشته برق در دانشگاه پاریس ۱۱، فرانسه (۱۳۷۴) به پایان رساند. وی هم‌اکنون دانشیار دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه امیرکبیر است. موضوعات مورد علاقهٔ ایشان پردازش سیگنال، پردازش صوت و گفتار، پردازش زبان طبیعی، یادگیری ماشین و چندرسانه‌ای است.

نشانی رایانامهٔ ایشان عبارت است از:

homayoun@aut.ac.ir