Journal of Modeling in Engineering 21 (2023) 163-172

Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

Research Article

Spoken Persian digits recognition using deep learning

Sahar Zarbafi[1], Kourosh Kiani[1,*], Razieh Rastgoo[1]

1. Faculty of Electrical and Computer Engineering, Semnan University, Semnan, Iran

*Corresponding Author: Kourosh.kiani@semnan.ac.ir

PAPER INFO

Paper history:
Received: 17 June 2023
Revised: 23 July 2023
Accepted: 21 August 2023

Keywords:
Spoken digits,
Classification,
Persian digits,
Mel spectrogram,
Dataset,
Transformer.

ABSTRACT

Classification of isolated digits is a fundamental challenge for many speech classification systems. While a lot of work has been done on spoken languages, limited research on spoken Persian digit data has been reported in the literature, and all researches have been done on the numbers in range 0 to 9. To this end, a huge dataset including a wider range of numbers has been collected with the participation of 145 people, including 75 men and 75 women. The collected dataset covers the numerical range from 0 to 599. After data preprocessing, the audio data is converted into a Mel spectrogram. To benefit from the recent advances in Deep Learning models, a Convolutional Neural Network (CNN) and a hybrid model including a transformer model along with a Long Short-Term Memory (LSTM) are proposed to extract the features and classify the data. The experimental results obtained from different analysis of the proposed models on the collected dataset indicate the validation accuracy of 98.03% for spoken Persian digits recognition.

DOI: https://doi.org/10.22075/jme.2023.30973.2472

How to cite this article:
Zarbafi, S., kiani, K., & Rastgoo, R. (2023). Spoken Persian digits recognition using deep learning. Journal of Modeling in Engineering, 21(74), 163-172. doi: 10.22075/jme.2023.30973.2472

تشخیص ارقام گفتاری فارسی با استفاده از شبکه‌های یادگیری عمیق

سحر زربافی[۱]، کوروش کیانی[۲،*]، راضیه راستگو[۳]

اطلاعات مقاله

نوع مقاله: پژوهشی
دریافت مقاله: ۱۴۰۲/۰۳/۲۷
بازنگری مقاله: ۱۴۰۲/۰۵/۰۱
پذیرش مقاله: ۱۴۰۲/۰۵/۳۰

واژگان کلیدی:
ارقام گفتاری،
طبقه‌بندی،
ارقام گفتاری فارسی،
طیف‌نگار مل،
پایگاه داده،
ترنسفورمر.

چکیده

طبقه‌بندی ارقام جدا شده چالش اساسی برای بسیاری از سیستم‌های طبقه‌بندی گفتار است. درحالی‌که مطالعات بسیاری بر روی زبان‌های گفتاری انجام شده است، تحقیقات محدودی در مورد داده‌های رقمی گفتاری فارسی گزارش شده است و تمامی تحقیقات مربوط به اعداد صفر تا ۹ بوده است. برای این منظور، پایگاه داده‌ی جامعی شامل بازه‌ی وسیعتری از اعداد با مشارکت ۱۴۵ نفر که شامل هفتاد نفر مرد و ۷۵ نفر زن هستند، جمع‌آوری گردیده است. پایگاه داده مذکور، بازه عددی صفر تا ۵۹۹ را پوشش می‌دهد. پس از پیش‌پردازش داده‌ها، داده‌های صوتی تبدیل به طیف‌نگار مل می‌گردند. در راستای بهره‌مندی از پیشرفت‌های اخیر در حوزه یادگیری عمیق، شبکه عصبی کانولوشنی و نیز یک مدل ترکیبی شامل مدل ترنسفورمر و حافظه کوتاه و بلند مدت جهت استخراج ویژگی و طبقه‌بندی داده‌ها مورد استفاده قرار می‌گیرد. نتایج تجربی حاصل از بررسی‌های مختلف مدل‌های پیشنهادی بر روی پایگاه داده جمع‌آوری شده حاکی از دقت اعتبارسنجی ۹۸/۰۳ درصد می‌باشد.

۱-مقدمه

توانایی برقراری ارتباط یکی از اساسی‌ترین جنبه‌های رفتار انسان است. انسان‌ها از طریق زبان‌های طبیعی به صورت کلامی و نوشتاری با یکدیگر ارتباط برقرار می‌کنند. قالب نوشتاری ارتباطات انسانی، توسط آوای انسان یعنی گفتار نمایش داده می‌شود. در این راستا، با پیشرفت در فن‌آوری-های زبان و گفتار، سیستم‌های تعاملی رایانه‌ای با کیفیت بالا ایجاد شده است [۱]. این سیستم‌ها، کاربردهای گسترده‌ای در آموزش، سرگرمی و تجارت دارند که منجر به افزایش تعامل میان انسان و ماشین با استفاده از زبان‌های طبیعی می‌گردند [۲]. همانند ارتباط انسان با انسان، حلقه تعامل با جریان اطلاعات بین رایانه و انسان تعریف می شود. شکل صوتی گفتار یا متنی زبان طبیعی، امکان برقراری ارتباط را فراهم می کند. در این میان، شکل صوتی گفتار یا شکل صوتی ارتباطات انسانی، راه مناسب‌تری برای برقراری ارتباطات انسانی است. این امر منجر به توسعه سیستم تشخیص گفتار گردیده است که به معنای فهم معنای گفتار انسان توسط رایانه می‌باشد. در این راستا، پیشرفت‌های اخیر در حوزه هوش مصنوعی نیز تأثیرگذار بوده است. الگوریتم های مختلف هوش مصنوعی و به خصوص یادگیری عمیق، کاربردهای مختلفی در حوزه‌های گوناگون داشته‌اند [۸-۲]. در این مقاله، الگوریتم‌های یادگیری عمیق جهت بهبود نتایج به کار گرفته شده‌اند. حوزه‌ی تحقیقاتی تشخیص گفتار حدود ۶۰ سال قدمت دارد. از زمان اختراع

* پست الکترونیک نویسنده مسئول: Kourosh.kiani@semnan.ac.ir
۱. دانشجوی کارشناسی ارشد، دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان
۲. دانشیار، دانشکده مهندسی برق و کامپیوتر ، دانشگاه سمنان
۳. استادیار، دانشکده مهندسی برق و کامپیوتر ، دانشگاه سمنان

اولین شناسایی گفتار در آزمایشگاه‌های بل در اوایل دهه ۱۹۵۰، پیشرفت‌ها و توسعه‌های جالب توجهی وجود داشته است [۹]. فورسبرگ شناسایی گفتار یا اصطلاحاً تشخیص خودکار گفتار را به‌عنوان فرآیند تفسیر گفتار انسان در رایانه تعریف کرده است. با این‌حال، ژورافسکی و مارتین [۱۰] شناسایی گفتار/تشخیص خودکار گفتار را از لحاظ فنی‌تر به‌عنوان ساختن سیستمی برای نگاشت سیگنال‌های صوتی به رشته‌ای از کلمات تعریف کرده‌اند. در سال‌های اخیر، تحقیقات گسترده‌ای در زمینه‌ی سیستم تشخیص خودکار گفتار صورت گرفته است [۱۱]. با این حال، محدودیت‌هایی همچون محیط‌های بدون سر و صدا، واژگان و زبان، پایین بودن میزان صحبت کردن و نیز وابستگی به بلندگو همچنان چالش برانگیز می‌باشند.

تشخیص گفتار به دلیل پیشرفت تکنولوژی روز به روز برای انسان مفیدتر می شود. تشخیص گفتار در زبان فارسی نیز در بسیاری از بخش ها استفاده می شود. تشخیص گفتار بنگالی، به‌ویژه تشخیص ارقام می‌تواند به فرمان‌های مبتنی بر گفتار برای دستگاه‌های اینترنت اشیا کمک کند. به عنوان مثال، یک سیستم کنترل ترافیک هوشمند می تواند از این امر بهره‌مند شود. طبقه‌بندی اعداد گفتاری فارسی نیز می‌تواند به تعامل انسان و رایانه کمک کند. در این راستا، دستیارهای صوتی به زبان فارسی و مبتنی بر هوش مصنوعی در دستگاه‌های تلفن همراه و نیز دیگر وسایل دیجیتال را می‌توان ایجاد کرد [۱۲]. بدین منظور، در این مقاله، ضمن جمع‌آوری پایگاه داده، مدلی هوشمند و مبتنی بر یادگیری عمیق ارائه گردیده است. نوآوری کار به شرح زیر می‌باشد:

- **پایگاه داده اعداد صوتی فارسی**: در این مقاله، برای اولین بار، یک پایگاه داده صوتی شامل بازه‌ی وسیعتری از اعداد با مشارکت ۱۴۵ نفر که شامل هفتاد نفر مرد و ۷۵ نفر زن هستند، جمع‌آوری گردیده است. پایگاه داده مذکور، بازه عددی صفر تا ۵۹۹ را پوشش می‌دهد.
- **مدل**: پس از پیش‌پردازش داده‌ها، داده‌های صوتی تبدیل به طیف‌نگار مل شده و برای یافتن ویژگی و طبقه‌بندی داده‌ها از شبکه عصبی کانولوشنی و نیز یک مدل ترکیبی شامل مدل ترنسفورمر و حافظه کوتاه و بلند مدت استفاده گردیده است.

ادامه‌ی این مقاله به صورت زیر می‌باشد:
در بخش ۲، مروری خلاصه بر کارهای گذشته صورت می گیرد. جزئیات پایگاه داده و نیز مدل پیشنهادی در بخش‌های ۳ و ۴ ارائه می‌گیرد. نتایج به دست آمده و نیز نتیجه‌گیری در بخش‌های ۵ و ۶ مورد بحث قرار می‌گیرد.

۲-مرور کارهای گذشته

تشخیص گفتار روشی است که سیگنال‌های گفتار انسان را به متن یا کلمات یا به هر شکلی که به راحتی توسط رایانه یا ماشین‌های دیگر قابل درک باشد، تبدیل می‌کند. مطالعات کمی در مورد سیستم‌های تشخیص رقم فارسی انجام شده است که اکثر آنها از مجموعه داده‌های کوچک استفاده می‌کنند[۱۳]. در این راستا، زبان‌های مختلفی مورد استفاده قرار گرفته‌اند. از جمله این زبان‌ها می‌توان به زبان انگلیسی، هندی، مَلیالَم و بوندلی اشاره نمود. اگرچه، تمام کارهای انجام شده در تشخیص اعداد در این زبان‌ها برای اعداد در بازه‌ی ۱ تا ۹ انجام شده است. به عنوان نمونه، شیونگ و همکارن [۱۴]، تشخیص گفتار مکالمه‌های مایکروسافت را برای سال ۲۰۱۶ گزارش کردند. آنها از شبکه‌های عصبی کانولوشنال و شبکه عصبی بازگشتی استفاده کردند. نرخ خطا برای یک سیستم منفرد متشکل از یکی از شبکه‌های مذکور، ٪۶/۹ و برای سیستم ترکیبی ٪۲/۶ می‌باشد. در مطالعه‌ای دیگر، گریوز و همکاران [۱۵]، دریافتند که شبکه‌های عصبی اسپایکی [1]در شبکه عصبی بازگشتی با حافظه‌ی طولانی کوتاه مدت[2] به طور قابل توجهی موثر عمل می‌کند. آنها در تشخیص هر رقمی با احتمال ٪۸۸ موفق بودند و زمانی که پیش پردازش‌های موثرتری انجام دادند، دقت تشخیص به ٪ ۹۹/۴ هم ارتقا یافت. گریوز و همکاران [۱۶]، شبکه عصبی بازگشتی با حافظه‌ی طولانی کوتاه مدت دوطرفه را با معماری‌های مختلف شبکه عصبی مقایسه کردند. آنها دریافتند که در مقایسه با شبکه عصبی بازگشتی و پرسپترون چند لایه[3] استاندارد، شبکه عصبی بازگشتی با حافظه‌ی طولانی کوتاه مدت بسیار سریعتر است. همچنین شبکه های دو جهته عملکرد بهتری نسبت به شبکه‌های یک طرفه دارند. ساکسنا و همکاران [۱۷]، روی ارقام هندی کار کردند و یک مدل مبتنی بر مدل پنهان مارکوف [4]ساختند.
نتایج تجربی حاکی از دقت ٪۹۴/۰۹ و ٪۸۵ بر روی

[1] Spiking Neural Networks (SNN)
[2] Long-Short Term Memory (LSTM)
[3] Multilayer perceptron (MLP)
[4] Hidden Markov Model (HMM)

داده های آموزشی و آزمایشی می‌باشد. دیکسیت و همکاران [۱۸]، بر روی ارقام بوندلی کار کردند. در این کار، از الگوریتم ضرایب پیشگویی خطی[1] و اصلاح شده ضرایب کپسترال مل[2] استفاده شده است. علی و همکاران [۱۹]، بر روی دسته بندی k نزدیک‌ترین همسایه [3] زبان پشتو کار کردند. آنها از ضرایب کپسترال مل برای استخراج ویژگی و از k نزدیک ترین همسایه نیز برای طبقه‌بندی استفاده گردیده است. نتایج تجربی، دقت طبقه‌بندی ۷۶/۸ درصد را نشان می‌دهند. محمد و همکاران [۲۰]، از مدل پنهان مارکوف [4]و ضرایب کپسترال مل استفاده کردند. نتایج تجربی مدل پیشنهادی نشان می‌دهد که ارقام صفر تا ۵، دقت بالاتر از ٪ ۹۵ و ارقام ۶ تا ۹ دقت تقریبی ٪ ۹۰ را نشان می‌دهند. علاوه بر این، آنها دریافتند که به دلیل لهجه‌های مختلف، مدل پیشنهادی در دو جفت رقم ۶ با ۹ و نیز ۷ با ۸ دچار اشتباه می‌گردد. سامون و همکاران [۲۱]، سه معماری شبکه‌های عصبی کانولوشنال برای تشخیص دستورالعمل‌های گفتار کوتاه بنگالی ارائه کرده‌اند. این سه مدل شامل مدل شبکه عصبی کانولوشنال مبتنی بر ضرایب کپسترال مل، مدل شبکه عصبی کانولوشنال خام و مدل شبکه عصبی کانولوشنال از پیش آموزش‌دیده با استفاده از یادگیری انتقالی می‌باشند که به ترتیب دقت‌های تشخیص ٪۷۴، ٪۷۱ و ٪۷۳ را به دست آورده‌اند.

غانتی و همکاران [۲۲]، از ضرایب کپسترال مل برای استخراج ویژگی و از کوانتیزه کردن برای کاهش ابعاد و ایجاد یک کد موثر استفاده کرده‌اند. در مدل پیشنهادی، پیچش زمانی پویا ۵ و نیز یک طبقه‌بندی کننده بر اساس حداقل فاصله استفاده شده است. آن ها دریافتند که رویکرد پردازش گفتار مبتنی بر ضرایب کپسترال مل محدودیتهایی را در حضور نویز برای سناریوهای تشخیص رقم گفتاری مستقل از سخنران نشان می‌دهد. گوپتا و همکاران [۲۳]، روی زبان بنگالی کارکردند. برای این منظور، آنها از ویژگی های ضرایب کپسترال مل و از تحلیل مولفه‌های اصلی۶ برای کاهش ابعاد استفاده کردند. ماشین بردار پشتیبان، جنگل تصادفی و پرسپترون چندلایه نیز مورد استفاده قرار گرفته است که بهترین دقت به دست آمده برابر با ۹۰ درصد بوده است. سانتانا و همکاران [۲۴]، یک سیستم تشخیص آسیب شناسی گفتار را پیشنهاد داده‌اند که به طور خودکار سیستم صوتی بیماران را تجزیه و تحلیل می‌کند. در این راستا، شبکه‌های یادگیری عمیق برای طبقه‌بندی سیگنال‌های گفتاری، برای تمایز بین سیگنال صوتی که طبیعی یا آسیب دیده است، استفاده شده است. الگوریتم لونبرگ مارکوارت برای طبقه‌بندی سیگنال‌های صوتی و نیز ماشین بولتزمن محدود شده برای پیاده‌سازی طبقه‌بندی یادگیری عمیق سیگنال‌های صوتی به کار گرفته شده است که منجر به دستیابی به دقت طبقه‌بندی ٪۹۸ و ٪۹۲ به ترتیب برای این الگوریتم‌ها گردیده است. گو و همکاران [۲۵]، یک چارچوب یادگیری عمیق چندوجهی جدید را معرفی کرده‌اند که توانایی استخراج ویژگی‌های سطح بالا از داده‌های متنی- صوتی را به منظور طبقه‌بندی گفتار دارا می‌باشد. سیستم پیشنهادی در یک محیط پزشکی واقعی آزمایش شده است تا به عنوان مرجعی برای تحقیقات آینده باشد. زمانی که ۶ هدف مختلف شناسایی شد، مدل به دقت متوسط ٪۸۳.۱۰ دست یافته است. مدل ارائه شده در اینجا بهتر از مدل‌های موجود که از ویژگی‌های متفاوتی استفاده می‌کنند، عمل کرده است.

مامیربایف و همکاران [۲۶] تجزیه و تحلیل مقایسه‌ای از پنج الگوریتم طبقه‌بندی گفتار ارائه داده‌اند. بر اساس نتایج بررسی‌ها، استفاده از یک شبکه عصبی پرسپترون چندلایه و نیز شبکه عصبی چند لایه عمیق منجر به دقت‌های ٪ ۹۳ و نیز ٪ ۹۹/۶۵ گردیده است. زاده و همکاران [۲۷]، مدلی مبتنی بر شبکه عصبی کانولوشنال عمیق برای استخراج ویژگی‌ها از سیگنال گفتار با استفاده از ضرایب کپسترال مل ارائه داده‌اند. نتایج تجربی بر روی پایگاه داده نامشخص حاکی از دقت ٪ ۸۴.۱۷ برای آزمایش مدل می باشد که معادل ٪ ۷.۳۲ بهبود در مقایسه با کارهای موجود است. رویکرد پیشنهادی در تشخیص رقم جدا شده پشتو می‌باشد. داوودی و همکاران [۲۸]، در یک مطالعه اخیر در مورد تشخیص گفتار یک کلمه‌ای دری، از شبکه‌های عصبی کانولوشنال برای تشخیص کلمات جدا شده در گفتار دری استفاده کردند. ضرایب کپسترال مل برای استخراج ویژگی در طول آموزش استفاده شده است که منجر به دقت ٪ ۸۸/۲ گردیده است که نشان می دهد

[1] Linear predictive coding (LPC)
[2] Frequency Cepstral Coefficients (MFCC)
[3] K-Nearest Neighbors (KNN)
[4] Hidden Markov Model (HMM)
[5] Dynamic Time Warping (DTW)
[6] Principal Component Analysis (PCA)

روش پیشنهادی کلمات تجسم شده را با دقت بالایی پیش بینی می‌کند. مارکولا و همکاران [۲۹]، با استفاده از روش تحلیل استرس صوتی، رویکرد جدیدی به نام «تشخیص دروغ» برای طبقه‌بندی گفتار پیشنهاد کرده‌اند. آنها از شبکه حافظه بلند مدت کوتاه برای تجزیه و تحلیل و طبقه‌بندی گفتار یک فرد به عنوان معتبر یا غیر معتبر استفاده کرده‌اند. بهترین مدل شبکه عصبی در روش پیشنهادی دقت ۷۲.۵ درصد را نشان داده است. نتایج تجربی از دیدگاه شناسایی الگوهایی در صدای افراد تحت استرس، حائز اهمیت می‌باشد.

در این مقاله، برای اولین بار، یک پایگاه داده صوتی شامل بازه‌ی وسیعتری از اعداد با مشارکت ۱۴۵ نفر که شامل هفتاد نفر مرد و ۷۵ نفر زن هستند، جمع‌آوری گردیده است. پایگاه داده مذکور، بازه عددی صفر تا ۵۹۹ را پوشش می‌دهد. پس از پیش‌پردازش داده‌ها، داده‌های صوتی تبدیل به طیف‌نگار مل شده و برای یافتن ویژگی و طبقه‌بندی داده‌ها از شبکه عصبی کانولوشنی و نیز یک مدل ترکیبی شامل مدل ترنسفورمر و حافظه کوتاه و بلند مدت استفاده گردیده است.

۳- پایگاه داده جمع‌آوری شده

پایگاه داده ارائه شده در این مقاله، با مشارکت ۱۴۵ نفر که شامل ۷۰ نفر مرد و ۷۵ نفر زن هستند، جمع‌آوری شده است. مشارکت‌کنندگان، عدد صفر تا ۵۹۹ را با گوشی تلفن همراه خود و یا از طریق رایانه شخصی ضبط کرده‌اند. هر فرد مشارکت‌کننده هر عدد را ۱۰ بار به‌صورت مجزا ضبط کرده است. در مجموع، شش هزار داده صوتی جمع‌آوری گردیده است. سن افراد بین ۲۰ تا ۳۰ سال بوده است و از شهرهای مختلف ایران و با لهجه‌های مختلف هستند. تلاش شده است تا حداقل نویز محیط در هنگام ضبط موجود باشد. داده‌های جمع‌آوری شده مورد بررسی دقیق قرار گرفته و نامگذاری شده‌اند.

۴- مدل پیشنهادی

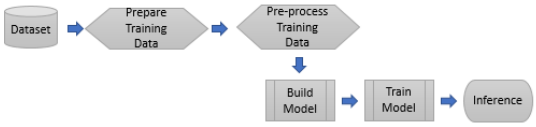
دیاگرام کلی روش پیشنهادی در شکل (۱) نشان داده شده است. این دیاگرام شامل مراحل زیر می‌باشد:

- **جمع‌آوری پایگاه داده**: در این مرحله، پایگاه داده معرفی شده در بخش ۳، جمع‌آوری می‌گردد.
- **پیش‌پردازش پایگاه داده**: به منظور استفاده از داده های جمع‌آوری شده، پیش‌پردازش‌هایی برای مرتب‌سازی و برچسب‌گذاری داده‌ها انجام می‌گردد. شکل (۲) مراحل مختلف پیش‌پردازش بر روی داده های ورودی را نشان می‌دهد. فایل صوتی را با فرمت ".wav" خوانده و بارگذاری می شود، داده ها را برچسب گزاری می شوند، باید همه صداها را استاندارد کرده و به یک نرخ نمونه‌برداری تبدیل شود تا همه آرایه‌ها دارای ابعاد یکسان باشند همهٔ صداها را با نرخ نمونه‌برداری ۲۲۰۵۰ ذخیره می‌کنیم. برخی از فایل‌های صوتی مونو هستند (یعنی ۱ کانال صوتی) درحالی‌که برخی از آنها استریو هستند (یعنی ۲ کانال صوتی). ازآنجایی‌که مدل انتظار دارد همه موارد دارای ابعاد یکسان باشند، تمام صداها مونو می شود، با اضافه‌کردن صفر برای اضافه‌کردن طول دیتاهای صوتی و یا با حذف‌کردن داده‌های صوتی طولانی داده‌های صوتی را یکسان می‌شود. صداها را به طیف‌نگار مل تبدیل می شود. آنها ویژگی‌های اساسی صدا را به تصویر می‌کشند و اغلب مناسب‌ترین راه برای واردکردن داده‌های صوتی در مدل‌های یادگیری عمیق هستند.
- **ساخت مدل**: در این مرحله، مدل شبکه عصبی کانولوشنی و نیز مدل مبتنی بر ترنسفورمر و شبکه حافظه کوتاه و طولانی مدت، ساخته می‌شود تا در مرحله بعد مورد آموزش قرار گیرد.
- **آموزش مدل**: در این مرحله، آموزش مدل با استفاده از داده‌های جمع‌آوری شده، انجام می‌گردد. بعد از انجام پیش‌پردازش نوبت به استخراج ویژگی‌ها می‌رسد. ویژگی طیف‌نگار مل از داده‌ها گرفته می‌شود. طیف‌نگار مل از اعمال تبدیل فوریه بر روی سیگنال‌های صوتی ایجاد می‌گردد. تبدیل فوریه سیگنال را به فرکانس‌های تشکیل‌دهنده آن تجزیه می‌کند و دامنه هر فرکانس موجود در سیگنال را نمایش می‌دهد. در طیف‌نگار مل، مدت زمان سیگنال صوتی به بخش‌های زمانی کوچک‌تر تقسیم می‌گردد و سپس تبدیل فوریه بر روی هر بخش اعمال می‌گردد تا فرکانس‌های موجود در آن بخش تعیین گردد. پس از آن، تبدیل فوریه برای همه بخش‌ها با هم ترکیب می‌گردند.

یکی از راه‌های محاسبه تبدیل فوریه استفاده از تکنیکی به نام تبدیل فوریه گسسته است. محاسبه تبدیل فوریه گسسته، هزینه‌بر می‌باشد. بنابراین، در

عمل از الگوریتم تبدیل فوریه سریع استفاده می‌شود که روشی کارآمد برای پیاده‌سازی تبدیل فوریه گسسته است. با این‌حال، تبدیل فوریه سریع، اجزای فرکانس کلی را برای کل سری زمانی سیگنال صوتی به طورکلی ارائه می‌دهد و نحوه تغییر اجزای فرکانس سیگنال صوتی را مشخص نمی‌کند. به عنوان مثال، امکان دیدن قسمت اول صدا که دارای فرکانس‌های بالا است و قسمت دوم که دارای فرکانس‌های پایین است داده نمی‌شود. برای رفع این چالش و نیز به دست آوردن نمای دانه‌ای بیشتر و دیدن تغییرات فرکانس در طول زمان، از الگوریتم تبدیل فوریه کوتاه‌مدت استفاده می‌شود. الگوریتم تبدیل فوریه کوتاه‌مدت نوع دیگری از تبدیل فوریه است که سیگنال صوتی را با استفاده از یک پنجره زمانی کشویی به بخش‌های کوچک‌تر تقسیم می‌کند، تبدیل فوریه سریع را در هر بخش می‌گیرد و سپس آنها را ترکیب می‌کند. بنابراین قادر است تغییرات فرکانس را به همراه زمان ثبت کند.

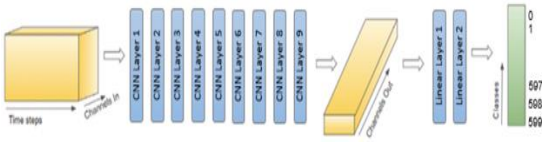
- **آزمایش مدل**: پس از آموزش مدل، آزمایش مدل بر روی داده‌ها صورت می‌گیرد.



شکل ۱: فلوچارت کلی روش پیشنهادی.

۴-۱- معماری مدل شبکه عصبی کانولوشنال

از آنجایی‌که داده‌های ما اکنون از تصاویر طیف‌نگاری تشکیل شده‌اند، یک معماری طبقه‌بندی با استفاده از شبکه عصبی کانولوشنی برای پردازش آنها ساخته می‌شود. این معماری دارای چهار بلوک کانولوشن است که نقشه‌های ویژگی را ایجاد می‌کنند. سپس این داده‌ها به فرمت مورد نیاز تغییر شکل می‌دهند تا بتوان آنها را در لایه طبقه‌بندی‌کننده خطی وارد کرد که در نهایت پیش‌بینی‌های ۶۰۰ کلاس را در خروجی ارائه می‌دهند.



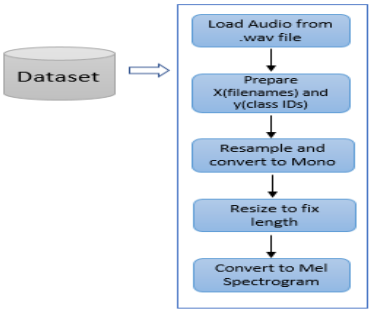
شکل ۲: مدل پیشنهادی اول.

۴-۲- مرحله آموزش مدل شبکه عصبی کانولوشنال

در طی آموزش شبکه، نرخ یادگیری به‌صورت پویا با پیشرفت آموزش تغییر می‌کند که معمولاً به آموزش اجازه می‌دهد در دوره‌های کمتری همگرا شود. مدل برای ۸۰ تکرار آموزش داده می‌شود که در هر تکرار، دسته‌ای از داده‌ها مورد پردازش قرار می‌گیرند. معیار ارزیابی مدل، دقت طبقه‌بندی می‌باشد.

۴-۳- معماری مدل ترکیبی ترنسفورمر و حافظه کوتاه و طولانی مدت

مدل ترنسفورمر از یک رمزگذار و یک رمزگشا تشکیل شده است. رمزگذار ترنسفورمر طیف‌نگار مل را به عنوان ورودی می‌گیرد تا نمایشی غنی‌تری از ویژگی‌ها ایجاد می‌کند. مدل پیشنهادی، ترکیبی از مدل ترنسفورمر و شبکه حافظهٔ کوتاه و طولانی مدت می‌باشد که نقاط قوت این دو مدل را برای طبقه‌بندی اعداد رقمی صوتی با استفاده از ویژگی‌های طیف مل ترکیب می‌کند. معماری پیشنهادی شامل یک لایه رمزگذار از مدل ترنسفورمر و به دنبال آن یک لایه شبکه حافظهٔ کوتاه و طولانی مدت و نیز یک لایه خروجی متراکم برای طبقه‌بندی اعداد صوتی رقمی فارسی است که در شکل (۳) نمایش داده شده است.

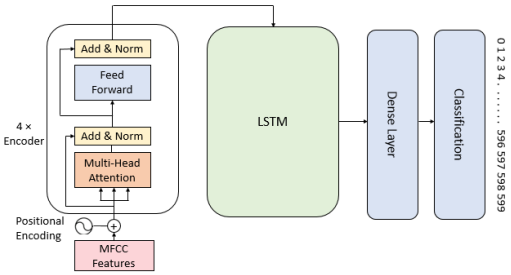


شکل ۳: پیش پردازش های صورت گرفته بر روی داده ها.

۴-۴- مرحله آموزش مدل ترکیبی ترنسفورمر و شبکه حافظهٔ کوتاه و طولانی مدت

با تعریف توابع بهینه‌ساز و ضرر مناسب برای مدل، آموزش به‌صورت پویا نرخ یادگیری را با پیشرفت آموزش تغییر می دهد که معمولاً منجر به همگرا شدن آموزش در دوره‌های کمتری می‌شود. مدل برای ۷۰ دوره به صورت پویا آموزش داده می‌شود که در هر تکرار، دسته‌ای از داده‌ها مورد پردازش قرار می‌گیرد. مشابه با مدل شبکه عصبی کانولوشنی، معیار ارزیابی مدل، دقت طبقه‌بندی می‌باشد.

شکل (۴) معماری پیشنهادی مدل ترکیبی ترنسفورمر و شبکه حافظهٔ کوتاه و طولانی مدت را نشان می‌دهد



شکل ۴: مدل پیشنهادی دوم.

۵- نتایج تجربی

در این بخش، جزئیات مربوط به پیاده‌سازی مدل و نیز نتایج به دست آمده از مدل بر روی پایگاه داده جمع‌آوری شده، ارائه داده می‌شود.

پیاده‌سازی مدل پیشنهادی بر روی یک پردازنده Intel(R) Xeon(R) E5-2699 (۲ پردازنده) با ۵۰ گیگابایت رم با سیستم عامل Microsoft Windows 10 و نرم‌افزار Python و نیز محیط اجرایی Jupyter Notebook با پردازنده گرافیکی NVIDIA Tesla K80 انجام گردیده است. کتابخانه‌های TensorFlow و Keras در پیاده‌سازی مدل استفاده شده‌اند. بهینه‌ساز Adam با سایز مینی‌بچ ۵۰، نرخ یادگیری تطبیقی و همچنین در مجموع ۲۰۰ دوره با توقف اولیه در مدل استفاده شده است.

برخی از فایل‌های صوتی مونو هستند (یعنی ۱ کانال صوتی) درحالی‌که برخی از آنها استریو هستند (یعنی ۲ کانال صوتی). از آنجایی‌که مدل انتظار دارد همه داده‌ها دارای ابعاد یکسان باشند، تمام داده‌های صوتی تبدیل به حالت مونو می‌گردند. برخی از فایل‌های صوتی با نرخ نمونه ۴۸۰۰۰ هرتز نمونه‌برداری می‌شوند، درحالی‌که اکثریت آنها با نرخ ۴۴۱۰۰ هرتز نمونه‌برداری می‌شوند. این بدان معناست که ۱ ثانیه صدا دارای اندازه آرایه ۴۸۰۰۰ برای برخی از فایل‌های صوتی است، درحالی‌که این اندازه آرایه برای سایر فایل‌ها، کوچک‌تر و برابر با ۴۴۱۰۰ می‌باشد. در این راستا، نیازمند یکسان‌سازی همه داده‌ها با نرخ نمونه‌برداری مشابه می‌باشیم تا همه آرایه‌ها دارای ابعاد یکسان باشند. به همین دلیل، همهٔ داده‌های صوتی با نرخ نمونه‌برداری ۲۲۰۵۰ ذخیره می‌شوند. در نتیجه، اندازه تمام نمونه‌های صوتی تغییر اندازه داده می‌شوند تا با افزایش مدت‌زمان آنها، با بی‌صدا کردن یا با کوتاه‌کردن آنها، طول یکسانی داشته باشند. با اضافه‌کردن صفر برای افزایش طول داده‌های صوتی یا با حذف‌کردن بخشی از داده‌های صوتی طولانی، فرآیند یکسان‌سازی طول داده‌های صوتی انجام می‌گردد.

پس از یکسان‌سازی طول داده‌های صوتی، این داده‌ها به طیف‌نگار مل تبدیل می‌شوند که یکی از موثرترین روش‌ها جهت واردکردن داده‌های صوتی در مدل‌های یادگیری عمیق هستند. در این راستا، ویژگی‌های اصلی صدا در نظر گرفته می‌شوند.

پس از فراخوانی داده صوتی، فرآیند نمونه‌برداری داده با فرکانس ۲۲/۰۵ کیلوهرتز انجام می‌شود که مدت‌زمان آن حدود ۳ ثانیه یا ۶۶۱۵۰ = ۲۲۰۵۰ * ۳ می‌باشد. اگر صدا دارای ۱ کانال باشد، شکل آرایه (۱, ۶۶۱۵۰) خواهد بود. وقتی صدا به یک طیف‌نگار مل تبدیل می‌شود، داده دارای ابعاد زیر خواهد بود:

(۱، ۶۴، ۱۳۰) = (تعداد_کانال‌ها، باندهای فرکانس مل، مراحل_زمان)

عدد ۱۳۰ از تقسیم ۶۶۱۵۰ بر اندازه پنجره حاوی تعدادی از نمونه‌ها به دست می‌آید.

تقسیم داده‌ها برای آموزش و آزمایش مدل به طور تصادفی انتخاب می‌شوند. در نهایت، ۸۰ درصد داده‌ها برای آموزش و ۲۰ درصد نیز برای آزمایش مدل انتخاب می‌شوند. آموزش شبکه عصبی کانولوشنی در ۱۰۰ دوره[1] انجام می‌شود. نرخ یادگیری[2] در هنگام شروع آموزش ۰.۱ است و در دوره‌های بالاتر نرخ یادگیری در ۰.۱ ضرب خواهد شد. دسته‌ای از تصاویر با ابعاد

(batch_sz ،num_channels ،Mel freq_bands ،time_steps)

به مدل وارد می‌شود (۶۴، ۱، ۱۲۸، ۱۳۰). خروجی هر لایه در مدل شبکه عصبی کانولوشنی پیشنهادی به طور خلاصه در جدول ۱ نشان داده شده است. مدل ترکیبی پیشنهادی شامل ترنسفورمر و حافظه کوتاه و طولانی مدت با استفاده از بهینه‌ساز Adam با نرخ یادگیری ۰.۰۰۱ برای ۸۰ دوره آموزش داده می‌شود. رفتار مدل ترکیبی پیشنهادی در طی آموزش و ارزیابی در شکل (۵) قابل مشاهده می‌باشد. دقت مدل ترکیبی ۹۸ /۰۳ و مدل مبتنی بر شبکه‌های عصبی

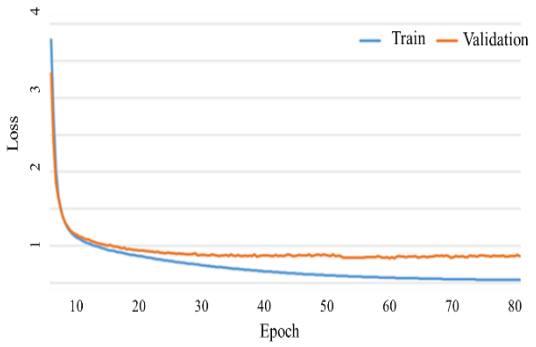
[1] Epoch

[2] Learning rate

کانولوشنی ۹۲/۰۱ بدست آمده است. همانگونه که در این شکل مشاهده می‌گردد، مدل ترکیبی پیشنهادی، پس از تعدادی تکرار، به رفتار پایداری می‌رسد.

جدول ۱: نمایش خروجی هر لایه

	Name	Type	Shape
0	conv2d	Conv2D	(None, 128, 126, 64)
1	batch_normalization	BatchNormalization	(None, 128, 126, 64)
2	max_pooling2d	MaxPooling2D	(None, 64, 63, 64)
3	conv2d_1	Conv2D	(None, 62, 61, 32)
4	batch_normalization_1	BatchNormalization	(None, 62, 61, 32)
5	max_pooling2d_1	MaxPooling2D	(None, 31, 31, 32)
6	conv2d_2	Conv2D	(None, 30, 30, 32)
7	batch_normalization_2	BatchNormalization	(None, 30, 30, 32)
8	max_pooling2d_2	MaxPooling2D	(None, 15, 15, 32)
9	flatten	Flatten	(None, 7200)
10	dense	Dense	(None, 64)
11	dense_1	Dense	(None, 600)



شکل ۵: رفتار مدل ترکیبی پیشنهادی در طی آموزش و ارزیابی

۶- نتیجه‌گیری

یکی از چالش‌های اساسی برای بسیاری از سیستم‌های طبقه‌بندی گفتار، طبقه‌بندی ارقام صوتی رقمی می‌باشد. علیرغم مطالعات بسیار بر روی زبان‌های گفتاری، تحقیقات محدودی در مورد داده‌های رقمی گفتاری فارسی در ادبیات گزارش شده است. تمامی تحقیقات مربوط به اعداد صفر تا ۹ بوده است. در این مقاله، جهت مقابله با این چالش، پایگاه داده‌ی جامعی شامل بازه‌ی وسیعتری از اعداد با مشارکت ۱۴۵ نفر که شامل هفتاد نفر مرد و ۷۵ نفر زن هستند، جمع‌آوری گردیده است. بازه‌ی اعداد صفر تا ۵۹۹ در این پایگاه داده پوشش داده شده است. پس از پیش‌پردازش داده‌ها و تبدیل آنها به طیف‌نگار مل، شبکه عصبی کانولوشنی و نیز یک مدل ترکیبی شامل مدل ترنسفورمر و حافظه کوتاه و بلند مدت استفاده گردیده است. دقت اعتبارسنجی ۹۸/۰۳ درصد بر روی پایگاه داده جمع‌آوری شده به دست آمده است. برای انجام کارهای بعدی، پیشنهاد می‌شود پایگاه داده با ترکیب بیشتری از افراد با لهجه و گویش‌های متفاوت و نیز رده‌های سنی بیشتر، غنی‌تر گردد. علاوه بر این، استفاده از سایر ویژگی‌های صوتی نیز می‌تواند مورد استفاده قرار گیرد.

مراجع

[1] P. Sanderson, "Cognitive work analysis and the analysis, design, and evaluation of human-computer interactive systems," in Proceedings 1998 Australasian Computer Human Interaction Conference. OzCHI'98 (Cat. No. 98EX234). 1998.

[2] A. Gunawan, "English digits speech recognition system based on hidden Markov models," in Proceedings of International Conference Computer. 2010.

[3] R. Rastgoo and V. Sattari Naeini, "A neurofuzzy QoS-aware routing protocol for smart grids," 22nd Iranian Conference on Electrical Engineering (ICEE), 2014, pp. 1080-1084.

[4] Rastgoo, R. and Sattari Naeini, V. Tuning parameters of the QoS-aware routing protocol for smart grids using genetic algorithm. Applied Artificial Intelligence, Vol. 30, No. 1, 2016, pp. 52-76.

[5] N. Majidi, K. Kiani, and R. Rastgoo, "A deep model for super-resolution enhancement from a single image," Journal of AI and Data Mining, Vol. 8, No. 4, 2020, pp. 451-460.

[6] K. Kiani, R. Hematpour, and R. Rastgoo, "Automatic grayscale image colorization using a deep hybrid model," Journal of AI and Data Mining, Vol. 9, No. 3, 2021, pp. 321-328.

[7] R. Rastgoo, and V. Sattari-Naeini, "Gsomcr: Multi-constraint genetic-optimized qos-aware routing protocol for smart grids. Iranian Journal of Science and Technology, "Transactions of Electrical Engineering," Vol. 42, 2018, pp. 185-194.

[8] R. Rastgoo, and K. Kiani, "Face recognition using fine-tuning of Deep Convolutional Neural Network and transfer learning," Journal of Modeling in Engineering, Vol. 17, No. 58, 2019, pp. 103-111.

[9] Y. Xu, "English speech recognition and evaluation of pronunciation quality using deep learning," Mobile Information Systems, Vol. 20, No. 2, 2022, pp. 1-12.

[10] M.K. Scheuerman, J.M. Paul, and J.R. Brubaker, "How computers see gender: An evaluation of gender classification in commercial facial analysis services," in Proceedings of the ACM on Human-Computer Interaction, 2019, pp. 1-33.

[11] Li, H., et al., "A convolutional neural network cascade for face detection," in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[12] R. Sharmin, S.K. Rahut, and M.R. Huq, "Bengali spoken digit classification: A deep learning approach using convolutional neural network," Procedia Computer Science, Vol. 171, 2020, pp. 1381-1388.

[13] O. Sen, and P. Roy, "A convolutional neural network based approach to recognize bangla spoken digits from speech signal," in 2021 International Conference on Electronics, Communications and Information Technology (ICECIT). 2021.

[14] W. Xiong, et al., "The Microsoft 2017 conversational speech recognition system," in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018.

[15] A. Graves, N. Beringer, and J. Schmidhuber, "A comparison between spiking and differentiable recurrent neural networks on spoken digit recognition," in The 23rd IASTED International Conference on modelling, identification, and control. 2004.

[16] A. Graves, and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," Neural networks, Vol. 18, 2005, pp. 602-610.

[17] D.P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[18] A. Dixit, A. Vidwans, and P. Sharma, "Improved MFCC and LPC algorithm for bundelkhandi isolated digit speech recognition," in 2016 international conference on electrical, electronics, and optimization techniques (ICEEOT), 2016.

[19] Z. Ali, et al., "Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN," International Journal of Speech Technology, Vol. 18, No. 2, 2015, pp. 271-275.

[20] G. Muhammad, Y.A. Alotaibi, and M.N. Huda, "Automatic speech recognition for Bangla digits," in 2009 12th International Conference on Computers and Information Technology. 2009.

[21] S.A. Sumon, et al., "Bangla short speech commands recognition using convolutional neural networks," in 2018 international conference on bangla speech and language processing (ICBSLP). 2018.

[22] S.K. Ghanty, S.H. Shaikh, and N. Chaki, "On recognition of spoken Bengali numerals," in 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM). 2010.

[23] A. Gupta, and K. Sarkar, "Recognition of spoken bengali numerals using MLP, SVM, RF based models with PCA based feature summarization," Int. Arab J. Inf. Technol., Vol. 15, No. 2, 2018, pp. 263-269.

[24] D.S.S. Megala, "Detection And Classification Of Speech Pathology Using Deep Learning," International journal of scientific & technology research, Vol. 8, No. 12, 2019.

[25] Y. Gu, et al., "Speech intention classification with multimodal deep learning," in Canadian conference on artificial intelligence. 2017.

[26] O. Mamyrbayev, et al., "Voice identification using classification algorithms," Intelligent System and Computing, 2019.

[27] B. Zada, and R. Ullah, "Pashto isolated digits recognition using deep convolutional neural network," Heliyon, Vol. 6, No. 2, 2020, pp. 3372.

[28] M. Dawodi, et al., "Dari speech classification using deep convolutional neural network," in 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2020.

[29] F.M. Marcolla, R. de Santiago, and R.L. Dazzi, "Novel Lie Speech Classification by using Voice Stress," in ICAART, Vol. 2, 2020.