Semnan University

Journal of Modeling in Engineering

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

ISSN: 2783-2538



Research Article

A Hybrid Approach for Speech Emotion Recognition: Data Augmentation and BiLSTM–Random Forest Integration

Ali Beiranvand [a], **Salman Karimi** [b,*]

[a] *Department of Electronics, Faculty of Engineering, Lorestan University, Khorramabad, Iran*
[b] *Department of Electronics, Faculty of Engineering, Lorestan University, Khorramabad, Iran*

UNCORRECTED PROOF

PAPER INFO

Paper history:

Received: 2022-03-13
Revised: 2022-05-12
Accepted: 2022-10-19

Keywords:

Data Augmentation;
Random Forest algorithm;
Speech Emotion Recognition;
BiLSTM neural network.

ABSTRACT

Speech Emotion Recognition (SER) is a significant field in speech signal processing and artificial intelligence, with broad applications in human-computer interaction, intelligent customer services, and emotional state detection. However, challenges such as the scarcity of diverse training data and the complexities of extracting effective features, limit the performance of SER systems. This paper presents a hybrid method based on Data Augmentation, a Bidirectional Long Short-Term Memory (BiLSTM) neural network, and the Random Forest algorithm to enhance the accuracy and reliability of the system. Initially, data augmentation techniques such as speed variation, noise addition, and pitch shifting are employed to generate synthetic samples. Subsequently, time-frequency features are extracted by the BiLSTM and passed to the Random Forest algorithm for final classification. This paper demonstrates that combining Data Augmentation with deep and traditional models can serve as a powerful approach to improving the accuracy and efficiency of SER systems. Evaluations of the proposed method on the expanded well-established EMODB database achieve an accuracy of 85.11%.

DOI: https://doi.org/10.22075/jme.2024.31282.2495

© 2024 Published by Semnan University Press.
This is an open access article under the CC-BY 4.0 license.(https://creativecommons.org/licenses/by/4.0/)

*** Corresponding author.**
E-mail address: Karimi.salman@lu.ac.ir

How to cite this article:

مقاله پژوهشی

رویکردی ترکیبی برای بازشناسی احساسات گفتار؛ داده افزایی و تلفیق مدل‌های BiLSTM–جنگل تصادفی

علی بیرانوند[۱]، سلمان کریمی[۲،*]

اطلاعات مقاله

دریافت مقاله: ۱۳۹۵/۰۳/۲۰
بازنگری مقاله: ۱۳۹۵/۰۶/۲۰
پذیرش مقاله: ۱۳۹۵/۰۷/۱۵

واژگان کلیدی:
افزایش داده،
الگوریتم جنگل تصادفی،
بازشناسی احساس گفتار،
شبکه BiLSTM .

چکیده

بازشناسی احساسات گفتار[۱] یکی از مهم‌ترین زمینه‌های پردازش سیگنال سیگنال و هوش مصنوعی است که کاربردهای گسترده‌ای در تعامل انسان-رایانه، خدمات رسانی هوشمند به مشتریان و تشخیص حالات عاطفی دارد. با این حال، چالش‌هایی مانند کمبود داده‌های آموزشی متنوع و پیچیدگی‌های استخراج ویژگی‌های مؤثر، عملکرد سیستم‌های شناسایی احساسات را محدود کرده است. در این مقاله، یک روش ترکیبی مبتنی بر داده افزایی[۲]، شبکه BiLSTM و الگوریتم جنگل تصادفی[۳] ارائه می‌شود تا دقت و قابلیت اطمینان سیستم شناسایی ارتقا یابد. در این راستا برای افزایش تعداد داده‌ها، ابتدا با بهره گیری از تکنیک‌هایی مانند تغییر سرعت، افزودن نویز و تغییر گام، تعداد نمونه‌های مورد استفاده را افزایش می‌دهیم. سپس، ویژگی‌های زمان-فرکانسی گفتار را توسط BiLSTM استخراج نموده و برای طبقه‌بندی نهایی به الگوریتم جنگل تصادفی منتقل می‌نماییم. این مقاله نشان می‌دهد که ترکیب داده افزایی با مدل‌های عمیق و سنتی می‌تواند به عنوان یک رویکرد قدرتمند در بهبود دقت و کارایی سیستم‌های بازشناسی احساسات گفتاری مورد استفاده قرار گیرد. ارزیابی روش پیشنهادی بر روی مجموعه داده‌ی توسعه یافته‌ی EMODB ، دقتی برابر ۸۵.۱۱٪ را ارائه می‌دهد.

DOI: https://doi.org/10.22075/jme.2024.31282.2495

© 2024 Published by Semnan University Press.
This is an open access article under the CC-BY 4.0 license.(https://creativecommons.org/licenses/by/4.0/)

۱-مقدمه[۴]

شناخت احساسات از طریق گفتار از یک حوزه تخصصی به بخشی مهم در تعامل انسان و رایانه تبدیل شده است. این سیستم‌ها با هدف تسهیل تعامل طبیعی با ماشین‌ها از طریق تعامل مستقیم صوتی به جای استفاده از دستگاه‌های سنتی به عنوان ورودی، طراحی شده‌اند تا محتوای کلامی‌را درک کرده و ارائه‌ی واکنش مناسب به شنوندگان انسانی را آسان‌تر کنند (Khalil 2019). به عنوان مثال، با استفاده از یک سیستم تشخیص احساسات گفتاری، می‌توان احساسات یک کودک را شناسایی کرد و متناسب با آن، بازی‌های مناسبی را برای او طراحی نمود. در سیستم‌های درمانی، یک پزشک به راحتی می‌تواند با استفاده از یک ساختار بازشناسی احساسات به عنوان ابزاری برای پیش‌بینی بیماری در فردی که از نظر فیزیولوژیکی افسرده است یا یک

[1] Speech Emotion Recognition (SER)
[2] Data Augmentation
[3] Random Forest (RF)

* پست الکترونیک نویسنده مسئول: Karimi.salman@lu.ac.ir
۱. دانشکده فنی و مهندسی، دانشگاه لرستان، خرم آباد، ایران
۲. دانشکده فنی و مهندسی، دانشگاه لرستان، خرم آباد، ایران

استناد به این مقاله: نحوه استناد فارسی در اینجا درج گردد.

UNCORRECTED PROOF

کودک مبتلا به اوتیسم استفاده کند. تشخیص احساسات انسانی از سیگنال‌های صوتی یکی از موثرترین و در عین حال چالش‌برانگیزترین تکنیک‌های مورد استفاده در حوزه پردازش گفتار است که ضمن پیچیدگی‌های فراوان، به دلیل دسترسی مستقیم به اصلی‌ترین ابزار انتقال پیام در انسان‌ها، یعنی گفتار، کاربردهای بسیار چشمگیری را ارائه می‌دهد (Dey, 2020).

با این حال، بیشتر پژوهش‌های موجود صرفاً بر افزایش دقت تشخیص احساس گفتار متمرکز شده‌اند و کاربردپذیری الگوریتم‌ها در پلتفرم‌های با منابع محدود را مورد توجه قرار نداده‌اند. در چنین محیط‌هایی، مدل‌هایی که دارای پارامترهای زیادی هستند و برای پیاده سازی نیازمند تعداد داده‌های زیاد و قدرت محاسباتی بالایی هستند، کارایی چندانی ندارند و نمیتوانند به صورت بهینه فعالیت کنند (Li 2024). تشخیص احساسات گفتاری یک پژوهش پیچیده همراه با چالش‌های متعدد است که به دلیل کاربردهای فراوان، محققان گوناگونی را به خود جذب کرده است. از جمله این زمینه‌ها می‌توان به بسترهایی نظیر نظارت صوتی، آموزش الکترونیکی، مطالعات بالینی، تشخیص دروغ، سرگرمی، بازی‌های رایانه‌ای و مراکز تماس اشاره نمود. در این راستا، علیرغم وجود حجم گسترده‌ای از تحقیقات و پیشرفت‌های چشمگیر سال‌های اخیر در حوزه بازشناسی احساسات گفتاری، هنوز به طور قطعی مشخص نیست که کدام روش مناسب‌ترین رویکرد در این حوزه را ارائه می‌کند (Er 2020). با این حال، توسعه و ارائه راه‌حل‌های نوآورانه برای کاربردهای گوناگون شناسایی احساسات همچنان با چالش‌هایی همراه است؛ زیرا تفاوت‌های موجود در زبان‌ها، فرهنگ‌ها، محیط‌ها و کیفیت گفتار بر عملکرد این سامانه‌ها تأثیر می‌گذارند. مطالعات پیشین در حوزه شناسایی احساسات عمدتاً بر داده‌های گفتاری در زبان‌هایی مانند انگلیسی، آلمانی و چینی تمرکز داشته‌اند و از مجموعه داده‌هایی نظیر IEMOCAP، RAVDESS، TESS و Emo-DB استفاده کرده‌اند (Jo 2025). با این حال، چالش‌های متعددی در این حوزه وجود دارد که از مهم‌ترین آن‌ها می‌توان به کمبود داده‌های آموزشی با برچسب دقیق، تنوع محدود داده‌ها و وجود نویز در سیگنال‌های گفتاری اشاره کرد. این چالش‌ها می‌توانند منجر به کاهش دقت و کارایی مدل‌های بازشناسی احساسات شوند.

نوآوری اصلی این پژوهش در ارائه‌ی یک چارچوب ترکیبی و مرحله‌ای برای بازشناسی احساسات بر پایه‌ی تحلیل گفتار است که به‌طور هم‌زمان سه چالش اساسی این حوزه یعنی کمبود داده، استخراج ویژگی‌های زمانی مؤثر و تصمیم‌گیری پایدار در طبقه‌بندی احساسات را هدف قرار می‌دهد. در این چارچوب و در ساختار پیشنهادی مقاله، یکپارچه‌سازی هوشمند سه مؤلفه‌ی افزایش داده، یادگیری عمیق و یادگیری ماشین کلاسیک به‌صورت یک مسیر یکپارچه و منسجم انجام شده است. در ابتدا، به‌کارگیری هدفمند تکنیک‌های افزایش داده برای غنی‌سازی داده‌های گفتاری مورد استفاده قرار گرفت که با اعمال تغییراتی نظیر تغییر سرعت گفتار، تغییر زیروبم صدا و تغییر فرکانس پایه، مجموعه داده‌ی مورد استفاده از نظر تنوع صوتی گسترش یافت. این رویکرد نه‌تنها مشکل محدود بودن داده‌های احساسی را کاهش می‌دهد، بلکه باعث افزایش توان تعمیم مدل در شرایط واقعی و در مواجهه با گویندگان و الگوهای گفتاری متفاوت می‌شود.

در گام دوم، استفاده از شبکه‌ی دوطرفه‌ی مولفه‌ی حافظه بلند-کوتاه مدت (BiLSTM[5]) به‌عنوان هسته‌ی مرکزی استخراج ویژگی‌های زمانی-زمینه‌ای ارائه گردید. برخلاف بسیاری از روش‌ها که صرفاً از ویژگی‌های ایستا استفاده می‌کنند، BiLSTM قادر است وابستگی‌های زمانی بلندمدت موجود در سیگنال گفتار را در هر دو جهت زمانی مدل‌سازی کند. این قابلیت سبب می‌شود الگوهای پیچیده‌ی احساسی که در طول زمان شکل می‌گیرند، با دقت بیشتری استخراج گردند و نمایش نهفته‌ی غنی‌تری از سیگنال صوتی را تولید نمایند.

در سومین مرحله، به جداسازی فرآیند استخراج ویژگی از مرحله‌ی تصمیم‌گیری و ترکیب یادگیری عمیق با یک طبقه‌بند یادگیری ماشین کلاسیک می‌پردازیم. در این ساختار، خروجی‌های نهفته‌ی شبکه BiLSTM به عنوان بردارهای ویژگی سطح بالا به الگوریتم جنگل تصادفی داده می‌شوند تا فرآیند طبقه‌بندی نهایی را انجام دهند.

این طراحی باعث می‌شود مدل پیشنهادی به طور همزمان

[5] Bi-directional Long Short Term Memory

UNCORRECTED PROOF

از قدرت یادگیری شبکه‌های عمیق و در عین حال از پایداری و مقاومت الگوریتم جنگل تصادفی در برابر بیش‌برازش بهره‌مند شود. چنین ترکیبی موجب افزایش دقت و پایداری سیستم نسبت به استفاده‌ی مستقل از هر یک از این روش‌ها می‌شود. در نهایت، کارایی چارچوب پیشنهادی از طریق ارزیابی دادگان استاندارد EMO-DB مورد بررسی قرار خواهد گرفت. نتایج به دست آمده نشان می‌دهد که مدل پیشنهادی با دستیابی به دقت ۸۵.۱۱ درصد روی این داده عملکرد بهتری نسبت به بسیاری از روش‌های متداول مبتنی بر LSTM ، جنگل تصادفی و مدل‌های ترکیبی دیگر ارائه می‌دهد. دقت بالای نتایج حاصل نشان‌دهنده‌ی آن است که یکپارچه‌سازی فرآیندهای افزایش داده، استخراج ویژگی عمیق مبتنی بر BiLSTM و طبقه‌بندی مبتنی بر الگوریتم جنگل تصادفی، می‌تواند به‌عنوان یک راهکار مؤثر و مقاوم برای سیستم‌های بازشناسی احساسات گفتار در کاربردهای عملی مورد استفاده قرار گیرد.

در ادامه ساختار مقاله توضیح داده می‌شود. در بخشهای اول و دوم به ترتیب مقدمه‌ای درباره اهمیت موضوع و مروری بر پژوهش‌های انجام‌شده در این حوزه ارائه می‌شود. در بخش سوم، روش پیشنهادی پژوهش حاضر که منجر به بهبود دقت سیستم بازشناسی گفتار شده است، شرح داده می‌شود. در بخش چهارم، نحوه پیاده‌سازی و اجرای روش پیشنهادی توضیح داده می‌شود و در بخش پنجم، نتایج آزمایش‌های صورت گرفته مورد بررسی و تحلیل قرار می‌گیرد. در نهایت، در بخش ششم، جمع‌بندی و نتیجه‌گیری کلی مقاله ارائه خواهد شد.

۲- مروری بر پژوهش‌های پیشین

اگر ویژگی‌های اضافی از طریق تجمیع داده‌ها با استفاده از روش‌های مختلف، مانند اطلاعات زبانی و بصری، استخراج شوند می‌توانند به تقویت عملکرد طبقه‌بندها کمک کنند. با این حال، بهره‌گیری از این ویژگی‌ها به میزان اهمیت و دسترس‌پذیری آن‌ها وابسته است. پس از فرآیند استخراج، این ویژگی‌ها می‌توانند در اختیار طیف گسترده‌ای از الگوریتم‌های طبقه‌بندی قرار گیرند تا در ساختار تشخیص احساسات مورد استفاده قرار گیرند .در پژوهش‌های مختلف، الگوریتم‌های متعددی از حوزه یادگیری ماشین برای تحلیل و طبقه‌بندی احساسات بر اساس همبستگی‌های صوتی موجود در گفتار مورد بررسی قرار گرفته‌اند. از جمله مدل‌های متداول مورد استفاده در این زمینه می‌توان به طبقه‌بندهای متمایزکننده خطی، مدل‌های مخلوط گوسی (GMM[6])، مدل‌های پنهان مارکوف (HMM[7])، الگوریتم k-نزدیک‌ترین همسایه (kNN[8])، ماشین‌های بردار پشتیبان (SVM[9])، درخت تصمیم و شبکه‌های عصبی مصنوعی (ANN[10]) اشاره کرد. این مدل‌ها معمولاً برای طبقه‌بندی احساسات بر پایه ویژگی‌های صوتی استخراج‌شده از سیگنال‌های گفتاری به کار گرفته می‌شوند (Wani 2021). علاوه بر موارد فوق، شبکه‌های عصبی عمیق به دلیل توانایی بالا در فرآیند یادگیری به عنوان مدل‌های صوتی و کارآیی مناسب در نمایش سطح بالای ویژگی‌های خام و طبقه بندی داده‌ها، از جمله زمینه‌های مناسب برای بازشناسی احساسات گفتار بوده و به طور موثر مورد استفاده قرار می‌گیرند (Masters 2018; Hinton 2006).

در این مقاله، یک سیستم تشخیص خودکار احساسات از روی گفتار ارائه شده است که دو الگوریتم یادگیری ماشین یعنی درخت تصمیم و جنگل تصادفی را برای طبقه‌بندی احساسات استفاده می‌نماید. برای آموزش مدل‌ها، ویژگی‌هایی از جمله شدت صدا، فرکانس پایه و ضرایب کپسترال فرکانس مل (MFCC[11]) از سیگنال‌های صوتی استخراج شده‌اند. سیستم ما قادر است احساساتی مانند شادی، غم، عصبانیت و حالت خنثی را شناسایی کند. نتایج نشان می‌دهد که الگوریتم جنگل تصادفی در مقایسه با درخت تصمیم عملکرد بهتری داشته و می‌تواند به طور مؤثر برای تشخیص احساسات استفاده شود (Noroozi 2017). هدف اصلی این مقاله، بهبود دقت تشخیص هیجانات گفتار با استفاده از روش‌های یادگیری عمیق، به ویژه شبکه‌های LSTM و مکانیسم توجه[12] است(Xie 2019). در این مقاله یک ساختار نوین برای تشخیص احساسات گفتاری با نام حافظه کوتاه-بلند مدتِ پیچشی ارائه شده است که از بلوک‌های LSTM در ترکیب با شبکه‌های عصبی پیچشی

[6] Gaussian Mixture Model
[7] Hidden Markov Model
[8] K-Nearest Neighbor
[9] Suppoert Vector Machine
[10] Artificial Neural Network
[11] Mel-Frequency Cepstral Coefficient
[12] Attention Mechanism

استناد به این مقاله: نحوه استناد فارسی در اینجا درج گردد.

UNCORRECTED PROOF

(CNN[۱۳]) تشکیل شده است (Kwon 2020). علاوه بر این، سعی می‌شود با استفاده از روش‌های یادگیری عمیق، ویژگی‌های صوتی مصنوعی را برای طبقه‌بندی احساسات در یادگیری پیشرفته استفاده نمود. این عملکرد، تشخیص قوی‌تری نسبت به طبقه بندی کم عمق متشکل از یادگیری ماشین سنتی ارائه می‌نماید (Yang 2021). این مقاله بررسی کارایی ترکیب شبکه‌های عصبی پیچشی و ترنسفورمر[۱۴] در بازشناسی احساسات از روی گفتار است. ترکیب فوق به منظور کاهش زمان آموزش، افزایش دقت شناسایی و کاهش نیاز به داده‌های برچسب‌دار طراحی شده است. ترکیب CNN و ترنسفورمر به دلیل توانایی در استخراج ویژگی‌های مکانی-زمانی، عملکرد بهتری در بازشناسی احساسات نشان داد. این روش می‌تواند برای کاربردهای واقعی با داده‌های محدود و نویزی مفید باشد (Pourebrahim 2023). در این مقاله، مجموعه داده‌ی EMO-DB برای طبقه‌بندی صداهای گفتاری مورد استفاده قرار گرفته است. برای طبقه‌بندی هفت احساس مختلف موجود در این مجموعه داده، از الگوریتم ماشین بردار پشتیبان استفاده شده است. در این فرآیند، ویژگی‌های MFCC از تمامی۵۳۵ نمونه گفتاری موجود در مجموعه‌داده استخراج شده و ارزیابی آماری جداگانه‌ای بر روی این ویژگی‌ها انجام شده است. در مراحل آموزش و آزمایش مدل، از روش اعتبارسنجی متقاطع ده‌بخشی[۱۵] استفاده شده است. همچنین، تحلیل عملکرد سیستم با بهره‌گیری از ماتریس درهم‌ریختگی و معیار دقت انجام شده که در نهایت دقتی معادل ۶۸٪ به دست آمده است (Kandali 2008; Milton 2013). ساختار مدل‌های یادگیری عمیق، مانند شبکه‌های عصبی پیچشی و حافظه کوتاه-بلند مدت، به دلیل قابلیت‌های بالا، به طور گسترده در تشخیص احساس گفتار استفاده می‌شود (de Pinto et al. 2020; Anvarjon 2020). در این مقاله یک ساختار شبکه عصبی برای گرفتن گفتار با طول متغیر برای بازشناسی احساسات ارائه شده است. در روش مورد استفاده، CNN برای نشان دادن ویژگی‌های طیف نگاری گفتار استفاده شده و شبکه عصبی مکرر (RNN[۱۶])، بخش‌های گفتار با طول متغیر را مدیریت می‌کنند (Ma et al. 2018). پس از ایشان، ژانگ و همکاران تکنیکی بر پایه‌ی مدل Alex-Net برای بازشناسی احساسات ارائه کردند که از پیش آموزش برای نمایش ویژگی‌ها و از ماشین بردار پشتیبان برای طبقه بندی احساسات گفتار استفاده نمودند (Zhang et al. 2018). به طور مشابه، لیو و همکاران از ترکیب دو ساختار CNN و LSTM برای شناسایی احساسات گفتار بهره بردند (Liu et al 2018). در پژوهشی دیگر، چن و ژائو یک مدل چند وجهی به نام STSER را پیشنهاد کردند که به صورت همزمان متن و صوت را می‌گیرد. در این ساختار، ویژگی‌ها به عنوان ورودی CNN مورد استفاده قرار گرفته و پس از اعمال تغییرات مورد نظر، برای یادگیری ویژگی‌های محلی استفاده می‌شوند. پس از آن، BiLSTM برای تشخیص وابستگی‌های طولانی مدت مرتبط با احساسات گفتار استفاده می‌شود (Chen and Zhao 2020).

۳-روش پیشنهادی

در این پژوهش، روش نوین برای بازشناسی احساسات گفتار با استفاده از ترکیب تشبکه عصبی حافظه بلند-کوتاه مدت دوطرفه و الگوریتم جنگل تصادفی ارائه شده است. یکی از چالش‌های اصلی در بازشناسی احساسات از طریق سیگنال‌های گفتاری، کمبود داده‌های آموزشی با برچسب دقیق و تنوع محدود در داده‌ها است. برای غلبه بر این مشکل، از تکنیک داده افزایی استفاده شده است که شامل ایجاد نمونه‌های جدید از داده‌های موجود با اعمال تغییراتی مانند تغییر سرعت، گام و تغییر فرکانس پایه سیگنال گفتاری می‌باشد (Ardan Dal Ri, Cifariello Ciardi, and Conci 2023). این تکنیک به افزایش تنوع داده‌ها و کاهش مشکل بیش‌برازش در مدل‌های یادگیری عمیق کمک می‌کند.

در این روش، ابتدا ویژگی‌های صوتی شامل MFCC، انرژی و فرکانس پایه از سیگنال‌های گفتاری استخراج شده، سپس داده‌های افزایش‌یافته به شبکه BiLSTM داده می‌شوند. شبکه BiLSTM به دلیل قابلیت مدل‌سازی وابستگی‌های زمانی دوطرفه، قادر است الگوهای پیچیده زمانی موجود در سیگنال‌های گفتاری را به‌طور مؤثر تشخیص دهد. پس از

[13] Convolutional Neural Network

[14] Transformer
[15] 10-fold cross validation

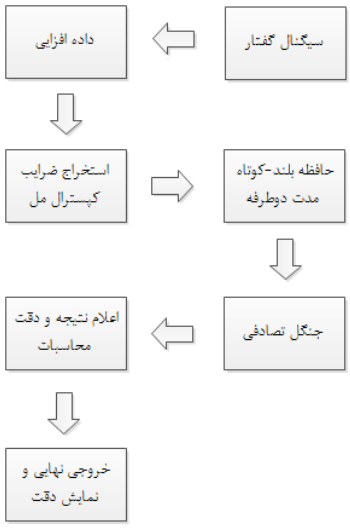
[16] Recurrent Neural Network

UNCORRECTED PROOF

آن، خروجی شبکه BiLSTM به عنوان ورودی برای الگوریتم جنگل تصادفی استفاده می‌شود که از یک سو به دلیل قابلیت بالای آن در دسته‌بندی و از سوی دیگر مقاومت مناسب آن در برابر عوامل مخرب مانند نویز، به عنوان لایه تصمیم‌گیری نهایی عمل می‌کند.

نتایج آزمایش‌ها روی مجموعه داده‌های معیار نشان می‌دهد که استفاده از تکنیک افزایش داده در ترکیب با BiLSTM و الگوریتم جنگل تصادفی منجر به بهبود قابل توجهی در دقت و کارایی سیستم می‌شود. دقت کلی سیستم پیشنهادی در تشخیص احساسات (شامل خشم، شادی، غم، تعجب و حالت خنثی) به ۸۵.۱۱٪ برای دادگان EMO-DB می‌رسد که نسبت به روش‌های سنتی و حتی استفاده تنها از BiLSTM یا الگوریتم جنگل تصادفی، عملکرد به مراتب بهتری را نشان می‌دهد.

نتایج این تحقیق نشان می‌دهد که ترکیب تکنیک افزایش داده با مدل‌های عمیق و سنتی می‌تواند به عنوان یک رویکرد قدرتمند در بهبود دقت و کارایی سیستم‌های بازشناسی احساسات گفتاری مورد استفاده قرار گیرد. این روش می‌تواند در کاربردهای مختلفی مانند ربات‌های اجتماعی، سیستم‌های پشتیبانی مشتری و واسطهای تعامل انسان-ماشین مفید باشد. بلوک دیاگرام روش پیشنهادی در شکل (۱) نشان داده شده است.



شکل ۱- بلوک دیاگرام طرح پیشنهادی

۳-۱ داده افزایی و استخراج ویژگی‌ها

در این مقاله برای ارزیابی روش‌های بازشناسی احساسات گفتاری، دو پایگاه داده استاندارد نقش محوری ایفا می‌کنند EMO-DBکه شامل ۵۳۵ گفتار آلمانی‌زبان از ۱۰ گوینده حرفه‌ای و هفت احساس کلاسیک است و RAVDESS، که ۱۴۴۰ نمونه انگلیسی‌زبان با محتوای لغوی ثابت و هشت دسته احساسی را فراهم می‌آورد (Zhao, Mao, and Zhang 2023).

ابتدا داده‌های گفتاری از مجموعه داده‌های معیار EMODB به طور جداگانه فراخوانی می‌شوند. این داده‌ها شامل ضبط‌های صوتی افراد هستند که حاوی احساسات مختلفی هستند. دادگان EMODB یکی از معروف‌ترین مجموعه‌های صوتی-احساسی است که در حوزه‌های پردازش زبان طبیعی، شناسایی احساسات از روی صدا و رباتیک اجتماعی استفاده می‌شود. این مجموعه توسط دانشگاه صنعتی برلین ایجاد شده و شامل ضبط‌های صوتی از بازیگران آلمانی‌زبان است که حالات احساسی مختلف را شبیه‌سازی کرده‌اند. احساسات پوشش داده شده در این دادگان عبارتند از خشم، بیزاری ،ترس، شادی، غم، تعجب و حالت خنثی هستند (Burkhardt et al. 2005).

سیگنال‌های گفتاری خام قبل از استخراج ویژگی‌ها نیاز به پیش‌پردازش دارند. پیش پردازش سیگنال گفتاری به فریم‌های کوچک تقسیم می‌شود. سپس برای غلبه بر مشکل کمبود داده‌های آموزشی و افزایش تنوع داده‌ها، از تکنیک افزایش داده برای تقویت داده ها استفاده می‌شود. در حوزه افزایش داده در پردازش سیگنال‌های صوتی، تغییر گام یکی از تکنیک‌های رایج برای افزایش تنوع داده‌ها و بهبود عملکرد مدل‌های یادگیری ماشین است. این کار با تغییر خصوصیات فونتیک و فیزیولوژیکی صدا (مانند تفاوت‌های بین گویندگان مختلف) به مدل‌ها کمک می‌کند تا بهتر به تغییرات طبیعی در داده‌های واقعی مقاومت نشان دهند (Park 2019). برای این مطالعه، ما از این مجموعه دادگان EMODB شناخته شده استفاده کرده و نقش تثبیت شده، این دادگان را به بستری ایده آل برای ارزیابی پایداری سیستم پیشنهادی ما تبدیل کرده است. سیگنال‌های گفتاری خام قبل از استخراج ویژگی‌ها نیاز به پیش‌پردازش دارند. پیش پردازش سیگنال گفتاری به فریم‌های کوچک تقسیم می‌شود. برای اطمینان از اینکه ویژگی هایی که استخراج می‌کنیم تمیز و قابل اعتماد هستند، ابتدا همه ضبط ها را پیش پردازش می‌کنیم. این شامل حذف نویز پس زمینه، نرمال سازی سطح دامنه برای یکنواختی در کلیپ ها و تقسیم جملات طولانی تر به واحدهای قابل مدیریت است. این گام حیاتی، اطلاعات صوتی نامربوط را حذف می‌کند و پایه ای پایدار برای محاسبه دقیق ویژگی

UNCORRECTED PROOF

ها فراهم می‌آورد. یکی از موانع اصلی در بازشناسی احساسات، کمبود داده های متنوع و برچسب خورده گفتار احساسی است. برای غلبه بر این مشکل و جلوگیری از برازش بیش از حد مدل به نمونه های محدود، استراتژی ترکیبی افزایش داده ها را پیاده سازی می‌کنیم. با گسترش مصنوعی اما واقع بینانه مجموعه داده‌های مورد استفاده، مدل را آموزش می‌دهیم که بر محتوای احساسی گفتار تمرکز کند، نه ویژگی‌های جانبی یک ضبط یا گوینده خاص. سیستم پیشنهادی به گونه‌ای طراحی شده است که تغییرپذیری طبیعی در محیط‌های واقعی را شبیه سازی کند. بر این اساس، درخواست مورد نظر عبارتند از:

- تغییر مقیاس زمانی: تغییر سرعت گفتار بدون تغییر زیر و بمی و تقلید از تفاوت در سرعت گفتار.
- تغییر زیر و بم: تغییر فرکانس پایه[17] برای تقریب تفاوت‌های صوتی در سن، جنسیت و فیزیولوژی.
- تزریق نویز کنترل شده: افزودن نویز سفید با نسبتهای مختلف سیگنال به نویز برای بررسی مقاومت در برابر شرایط صوتی ناقص.

این تغییرات هم به صورت فردی و هم به صورت ترکیبی اعمال می‌شوند و یک فضای آموزشی صوتی متنوع و بسیار بزرگ‌تر ایجاد می‌کنند که به مدل پیشنهادی کمک می‌نمایند در گستره‌ی وسیع‌تری تعمیم یابد.

۳-۱-۱ ثبت امضای صدا: استخراج ویژگی‌ها

این فرآیند شامل ایجاد نمونه‌های جدید از داده‌های موجود با اعمال تغییرات زیر است: تغییر سرعت سیگنال گفتاری، تغییر فرکانس پایه و افزودن نویز. هر یک از تغییرات فوق به دو صورت جداگانه و ترکیبی روی داده‌ها اعمال می‌شوند تا تنوع داده‌ها افزایش یابد. پس از تقویت داده‌ها، ویژگی‌های صوتی از سیگنال‌های گفتاری استخراج می‌شوند. یکی از مهم‌ترین ویژگی‌ها MFCC می‌باشد. این ویژگی به‌عنوان یکی از معیارهای اصلی برای تشخیص الگوهای فرکانسی در سیگنال‌های صوتی شناخته می‌شوند (Chaudhari and Jadhav 2019). از نمونه‌های گفتاری تمیز و تقویت شده، مجموعه‌ای از ویژگی‌های صوتی استخراج می‌گردد که به طور جمعی امضای احساسی گفتار را توصیف می‌کنند. سنگ بنای این فرآیند ضرایب MFCC هستند که به طور مؤثری طیف توان کوتاه مدت صدا را مدل سازی می‌کند و در ثبت کیفیت صدا بسیار عالی عمل می‌کنند. برای ارائه تصویری کامل‌تر، در این مقاله ضرایب کپسترال را با ویژگی‌های آوایی که حامل نشانه‌های احساسی حیاتی هستند، تلفیق می‌کنیم:

- فرکانس پایه (F0 / زیر و بمی): ردیابی خطوط صوتی و تغییرات مرتبط با احساسات مختلف (مثلا بالا و متغیر برای احساس شادی، پایین و تخت برای احساس غم).
- انرژی سیگنال (بلندی صدا): اندازه گیری دینامیک دامنه و شدت گفتار در طول زمان.

این مجموعه ویژگی چندوجهی که شامل ویژگی‌های طیفی، صوتی و انرژی است، تعداد ویژگی‌های مناسبی را تشکیل می‌دهد که مدل تلفیقی BiLSTM-جنگل تصادفی پیشنهادی این مقاله از آن بدست خواهد آمد.

۳-۲ مدل سازی دینامیک زمانی با استفاده از شبکه‌ی BiLSTM

شبکه‌های عصبی LSTM و BiLSTM از جمله معماری‌های پرکاربرد در حوزه پردازش سری‌های زمانی هستند. این شبکه‌ها به‌طور خاص برای مدل‌سازی وابستگی‌های بلندمدت در داده‌های ترتیبی طراحی شده‌اند. در کاربردهایی مانند تشخیص احساسات از سیگنال‌های گفتاری، که الگوهای زمانی نقش مهمی‌ایفا می‌کنند، استفاده از BiLSTM بهتر است (Zhang et al. 2021). ساختار احساس در گفتار ذاتاً متوالی است و در طول زمان شکل گرفته و تغییر می‌کند. برای مدل سازی این دینامیک‌های زمانی، از حافظه دوطرفه بلند-کوتاه مدت BiLSTM استفاده می‌کنیم. این الگوریتم به ما اجازه می‌دهد تا نحوه‌ی تکامل ویژگی‌ها را تفسیر کنیم و زمینه و جریان لازم برای شناسایی دقیق احساسات را ثبت نماییم.

۳-۲-۱ منطق دوطرفه بودن، معماری شبکه و فرمول بندی

یک بینش کلیدی در بازشناسی احساسات گفتاری این است که معانی احساسی اغلب وابسته به زمینه هستند. اهمیت یک نشانه صوتی مکث، اوج زیر و بمی‌یا تغییر شدت

[17] Pitch Shifting

UNCORRECTED PROOF

می‌تواند به آنچه قبل و بعد از آن می‌آید بستگی داشته باشد. یک LSTM استاندارد و یک طرفه اطلاعات را به طور دقیق از گذشته به آینده پردازش می‌کند و توانایی آن را برای استفاده از این زمینه «آینده» محدود می‌کند. معماری BiLSTM این محدودیت را با پردازش توالی در هر دو جهت به طور همزمان برطرف می‌کند. این دیدگاه دوجهته تضمین می‌کند که هر لحظه از گفتار با آگاهی کامل از کل زمینه اطراف تفسیر شود.

دنباله بردارهای ویژگی صوتی X=(x1,x2,…,xT) که در آن هر xt نمایانگر یکی از ویژگی‌های بدست آمده از ضرایب MFCC، F0 و انرژی است، به عنوان ورودی عمل می‌کند. لایه BiLSTM از دو LSTM مستقل تشکیل شده است: یک LSTM جلو که دنباله را به ترتیب زمانی پردازش می‌کند و یک LSTM عقب که آن را به صورت معکوس پردازش می‌کند. مکانیزم اصلی LSTM برای هر جهت شامل دروازه‌های زیر در گام زمانی t است:

- دروازه فراموشی: تصمیم می‌گیرد چه اطلاعاتی را از سلول حذف کند.

$$f_t = \sigma(W_f \cdot [h_{(t-1)}, x_t] + b_f) \qquad (۱)$$

در معادله (۱)، f_t دروازه فراموشی است که این دروازه تعیین می‌کند چه مقدار از حالت سلول قبلی باید حفظ یا حذف شود. مقدار f_t از طریق تابع سیگموید σ محاسبه می‌شود. در این رابطه، W_f ماتریس وزن دروازه فراموشی و b_f بردار بایاس آن است. همچنین $[h_{(t-1)}, x_t]$ نشان‌دهنده الحاق بردار حالت پنهان گام قبلی $h_{(t-1)}$ و بردار ورودی در زمان جاری x_t است.

- دروازه ورودی: تعیین می‌کند چه اطلاعات جدیدی ذخیره شود.

$$i_t = \sigma(W_i \cdot [h_{(t-1)}, x_t] + b_i) \qquad (۲)$$

در معادله (۲)، دروازه ورودی با نماد i_t معرفی شده است. این دروازه مشخص می‌کند چه مقدار از اطلاعات جدید باید در حالت سلول ذخیره شود. در این رابطه، W_i ماتریس وزنی دروازه ورودی و b_i بردار بایاس آن هستند.

$$C_t = tanh(W_C \cdot [h_{(t-1)}, x_t] + b_C) \qquad (۳)$$

در معادله (۳)، مقدار منتخب حالت سلول با نماد C_t محاسبه می‌شود. این مقدار از طریق تابع تانژانت هایپربولیک به دست می‌آید. W_c ماتریس وزن تولید محتوای جدید سلول و b_c بردار بایاس متناظر با آن است.

- به روزرسانی وضعیت سلول: اطلاعات گذشته و جدید را ترکیب می‌کند.

$$C_t = f_t * C_{(t-1)} + i_t * C_t \qquad (۴)$$

در معادله (۴)، به‌روزرسانی حالت سلول با نماد C_t در این مرحله، اطلاعات قبلی و اطلاعات جدید با یکدیگر ترکیب می‌شوند.

- دروازه خروجی: تصمیم می‌گیرد کدام بخش از وضعیت سلول را خروجی دهد.

$$o_t = \sigma(W_o \cdot [h_{(t-1)}, x_t] + b_o) \qquad (۵)$$

در معادله (۵)، دروازه خروجی با نماد o_t محاسبه می‌شود. این دروازه تعیین می‌کند چه بخشی از حالت سلول برای تولید خروجی شبکه مورد استفاده قرار گیرد.

$$h_t = o_t * tanh(C_t) \qquad (۶)$$

در معادله (۶)، حالت پنهان شبکه با نماد h_t محاسبه می‌شود. این مقدار خروجی واقعی واحد LSTM در زمان t محسوب می‌شود و از ترکیب دروازه خروجی و حالت سلول به دست می‌آید.

برای گذر رو به جلو، این محاسبات حالت پنهان را به دست می دهند $\overrightarrow{h_t}$. برای پاس به عقب، که به دنباله معکوس اعمال می‌شود، آن ها به دست $\overleftarrow{h_t}$ می‌آیند. نمایش نهایی و غنی از زمینه در هر گام زمانی، الحاق هر دو حالت جهتی است:

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \qquad (۷)$$

در رابطه (۷)، h_t که «بردار ادغام شده» نامیده می‌شود، الگوهای زمانی «منتهی به» و «از هر لحظه» را در بر می‌گیرد و به مدل امکان می‌دهد امضاهای پیچیده‌ای مانند واکنش تأخیری یا تغییر کیفیت صدا را تشخیص دهد.

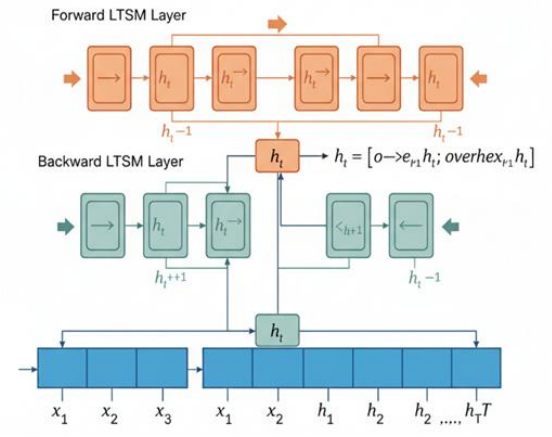
۳-۲-۲ ادغام درون چارچوب ترکیبی

در معماری ترکیبی پیشنهادی، BiLSTM به عنوان یک استخراج کننده پیشرفته ویژگی زمانی عمل می‌کند (Kim et al. 2020). این توالی، ویژگی‌های خام صوتی را به یک مجموعه‌ی سطح بالا و زمینه‌ای تبدیل می‌کند که معمولاً به عنوان خروجی نهایی BiLSTM، به صورت نمایشی تجمیعی در طول زمان در نظر گرفته می‌شود. این جاسازی برای طبقه بندی مستقیم استفاده نمی‌شود و به جای آن به عنوان ورودی پیشرفته به طبقه‌بند جنگل تصادفی منتقل می‌گردد. این امر، هم افزایی قدرتمند و

UNCORRECTED PROOF

مکملی در سیستم پیشنهادی ایجاد می‌کند. در این خصوص BiLSTM در مدل‌سازی و فشرده‌سازی وابستگی‌های بلندمدت از داده‌های صوتی افزوده شده مهارت دارد، در حالی که جنگل تصادفی از این جاسازی‌های غنی برای انجام طبقه بندی قوی و مبتنی بر مجموعه بهره می‌برد. مسیر افزونه گسترده در اینجا حیاتی است و توالی‌های آموزشی متنوعی را فراهم می‌کند که BiLSTM بتواند الگوهای زمانی قابل تعمیم را بدون بیش برازش بیاموزد . شکل (۳) مسیر پردازش BiLSTM را نشان می‌دهد.

معماری شبکه پیشنهادی BiLSTM که در جدول (۱) ارائه شده است، شامل چندین لایه متوالی برای پردازش و طبقه‌بندی داده‌های گفتاری است. در ابتدا، لایه ورودی دنباله‌ای[18] قرار دارد که وظیفه دریافت بردارهای ویژگی استخراج‌شده از سیگنال گفتار، مانند MFCC و سایر ویژگی‌های صوتی را بر عهده دارد و آن‌ها را به‌صورت یک توالی زمانی به شبکه وارد می‌کند. پس از آن، یک لایه برای کاهش بیش‌برازش و افزایش توان تعمیم‌پذیری مدل به کار گرفته شده است. در ادامه، لایه اصلی شبکه یعنی BiLSTMقرار دارد که با پردازش داده‌ها در دو جهت زمانی (رو به جلو و معکوس)، وابستگی‌های بلندمدت موجود در سیگنال گفتار را استخراج کرده و نمایش غنی‌تری از الگوهای زمانی تولید می‌کند. خروجی این لایه مجدداً از یک لایه دورانداز[19] عبور می‌کند تا از وابستگی بیش از حد مدل به برخی نرون‌ها جلوگیری شود. سپس، یک لایه‌ی تماماً متصل[20] قرار دارد که نمایش‌های استخراج‌شده توسط BiLSTMرا به فضای کلاس‌های احساسی نگاشت می‌کند. در نهایت، لایه بیشینه نَرم[21] برای محاسبه احتمال تعلق هر نمونه به کلاس‌های مختلف احساسی به کار می‌رود و لایه Classification Output بر اساس بیشترین احتمال، برچسب نهایی احساس را تعیین می‌کند. این ساختار لایه‌ای امکان استخراج مؤثر ویژگی‌های زمانی و انجام طبقه‌بندی دقیق احساسات گفتاری را فراهم می‌سازد.



شکل ۳-معماری شبکه LSTM دوطرفه (BiLSTM).

جدول ۱- لایه‌های شبکه LSTM دوطرفه (BiLSTM).

لایه‌های شبکه پیشنهادی	نوع
sequenceInputLayer	ورودی دنباله‌ای
dropoutLayer	دورانداز
bilstmLayer	حافظه کوتاه-بلند مدت دوطرفه
dropoutLayer	دورانداز
fullyConnectedLayer	تماماً متصل
softmaxLayer	بیشینه نَرم
classificationLayer	خروجی طبقه‌بند

۳-۳ جنگل تصادفی؛ لایه تصمیم گیری مقاوم

پس از استخراج جاسازی‌های زمانی سطح بالا توسط BiLSTM، به یک طبقه بند نهایی نیاز داریم که هم مقاوم و هم قابل تفسیر باشد. در این مقاله برای این اجرای این نقش از الگوریتم جنگل تصادفی استفاده می‌کنیم. به عنوان یک روش قدرتمند مجموعه‌ای، جنگل تصادفی به واسطه دقت بالا، مقاومت در برابر بیش برازش و مدیریت طبیعی فضاهای ویژگی پیچیده شناخته می‌شود و آن را به شریک ایده آل برای استخراج عمیق ویژگی‌های BiLSTM تبدیل می‌کند.

۳-۳-۱ یکپارچه سازی با موتور ویژگی BiLSTM

در معماری ترکیبی پیشنهادی، جریان داده به صورت زیر است: BiLSTM به عنوان موتور ویژگی زمانی و جنگل تصادفی به عنوان تصمیم گیرنده مبتنی بر مجموعه عمل می‌کنند. در تعبیه‌های دنباله‌ای زمینه‌ایِ تولید شده توسط BiLSTM، معمولاً یک بردار با طول ثابت که گفتار را

[18] Sequence Input
[19] Dropout
[20] Fully Connected
[21] Softmax

UNCORRECTED PROOF

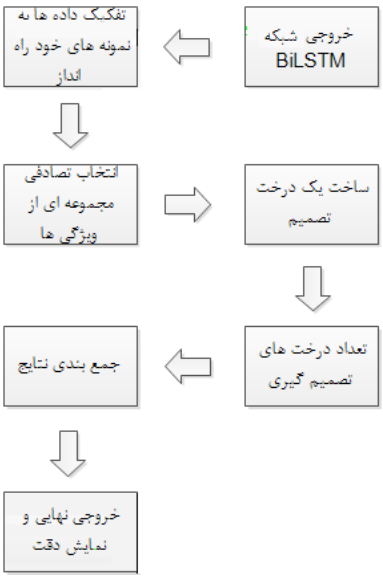
خلاصه می‌کند، به عنوان بردار ویژگی ورودی به جنگل تصادفی داده می‌شود.

این طراحی با توضیحات زیر از نقاط قوت هر دو جزء بهره می‌برد:

۱) BiLSTM در انجام وظیفه‌ی پیچیده‌ی مدلسازی و فشرده‌سازی وابستگی‌های زمان-بلند و متوالی سازیِ داده‌های صوتی خام و افزوده شده به یک مجموعه ویژگی متمایز مهارت دارد.

۲) در مرحله بعد، جنگل تصادفی بر روی فضای ویژگی بهینه شده در مرحله اول عمل می‌کند و با استفاده از مجموعه فرامین تصمیم‌گیری غیرخطی خود، طبقه بندی نهایی را از طریق مکانیزم رأی گیری اکثریت انجام می‌دهد.

در فرآیند پیشنهادی، هم افزایی دو زیرسیستم معرفی شده، به صورت ذیل است: مؤلفه یادگیری عمیق، الگوهای زمانی پیچیده را استخراج می‌کند، در حالی که مؤلفه‌ی کلاسیک یادگیری ماشین، یک طبقه‌بندی قوی و قابل تعمیم به همراه تفسیر الگوهای بدست آمده را ارائه می‌دهد. این تلفیق، مستقیماً به چالش اصلی ساخت یک سیستم بازشناسی احساسات که علاوه بر دقت، در برابر تغییرپذیری گفتار در دنیای واقعی مقاوم نیز می‌باشد، پاسخ می‌دهد. شکل (۴) نقش مؤلفه‌ی جنگل تصادفی را در معماری کلی مدل ترکیبی نشان می‌دهد.



شکل ۴- شماتیک الگوریتم جنگل تصادفی.

معادله ریاضی الگوریتم جنگل تصادفی در رابطه‌ی (۸) نشان داده شده است.

$$\hat{y} = \sum_{i=1}^{T} I\left(f_i(x) \right) \qquad (۸)$$

که در آن T تعداد درخت‌ها و $f_i(x)$ پیش بینی درخت i-ام و I تابع شاخص است.

۳-۴-معیار ارزیابی

برای ارزیابی دقیق عملکرد مدل ترکیبی پیشنهادی، از معیارهای ارزیابی استانداردی استفاده می‌کنیم که برای ارزیابی وظایف طبقه بندی، به ویژه در شناسایی احساسات گفتار چندکلاسی ضروری هستند. این شاخص‌ها معیارهایی کمی و قابل تفسیر از دقت، قابلیت اطمینان و پایداری مدل در کلاس‌های مختلف احساسی ارائه می‌دهند.

۱) دقت کلی[۲۲]، معیاری بنیادی است که نسبت تمام جملات آزمون که به درستی طبقه بندی شده‌اند به تعداد کل نمونه‌ها را نشان می‌دهد. این معیار خلاصه‌ای سطح بالا از عملکرد مدل ارائه داده و به صورت زیر محاسبه می‌شود:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (۹)$$

که در آن:

- TP[۲۳] تعداد نمونه‌های کلاس مورد نظر که کلاس آن‌ها به درستی شناسایی شده است.
- TN[۲۴] تعداد نمونه‌های سایر کلاس‌ها که کلاس آن‌ها بدرستی کلاس هدف شناسایی نشده است.
- FP[۲۵] تعداد نمونه‌های سایر کلاس‌ها که کلاس آن‌ها به غلط کلاس هدف شناسایی شده است.
- FN[۲۶] تعداد نمونه‌های کلاس مورد نظر که کلاس آن‌ها به غلط شناسایی شده است.

۲) دقت کلاسی[۲۷] نشان می‌دهد که از بین تمام مواردی که مدل یک کلاس خاص (مثلاً شادی) را پیش‌بینی کرده، چند درصد آن‌ها واقعاً به همان کلاس تعلق داشته‌اند. به عبارت دیگر، «دقت کلاسی» اندازه‌گیری اطمینان از پیش‌بینی‌های مثبت است:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \qquad (۱۰)$$

[22] Accuracy
[23] True Positive
[24] True Negative
[25] False Positive
[26] False Negative
[27] Precision

UNCORRECTED PROOF

که در آن c نشان‌دهنده یک کلاس عاطفی خاص (مثل خشم یا شادی) است. TP_c تعداد پیش‌بینی‌های درست برای کلاس c و FP_c تعداد مواردی که سایر کلاس‌ها به اشتباه به‌عنوان کلاس c طبقه‌بندی کرده‌اند، می‌باشد.

۳) بازیابی کلاسی[۲۸] یا حساسیت[۲۹]، نشان می‌دهد که از بین تمام نمونه‌های واقعیِ یک کلاس خاص، چند درصد آن‌ها به‌درستی شناسایی شده‌اند.

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (۱۱)$$

که در آن FN_c تعداد حالاتی است که نمونه‌های واقعی کلاس c به اشتباه به عنوان سایر کلاس‌ها طبقه‌بندی شده‌اند.

۴-روش اجرا

محاسبات و شبیه سازی‌های مورد اشاره در این مقاله در بستر نرم افزار Matlab 2022 b پیاده‌سازی گردیده‌اند. در این بخش جزئیات سیستم، پیکربندی مدل، پردازش داده‌ها، روند بهبود نتایج و روش‌های آموزشی مورد استفاده برای ارزیابی چارچوب ترکیبی پیشنهادی، شرح داده می‌شود.

الگوریتم مورد نظر بر روی پایگاه داده استاندارد EMODB پیاده‌سازی شده است. در ابتدا، پس از استخراج توالی ویژگی‌های صوتی (MFCCs، F۰، انرژی)، مقادیر بدست آمده نرمال سازی شده و به پنجره‌های همپوشان تقسیم می‌شوند. برای این منظور، بردارهای ویژگی، از دنباله‌های صوتی ۲۰ فریمی که با یکدیگر ۱۰ فریم همپوشانی دارند و طول هر یک از این فریم‌ها ۳۰ میلی ثانیه می‌باشد، استخراج می‌گردند. این شیوه‌ی بخش‌بندی، ضمن حفظ کارایی محاسباتی، دینامیک زمانی معناداری را ایجاد می‌کند.

در این مقاله، برای رفع کمبود داده‌ها و افزایش پایداری مدل، یک خط لوله جامع پیاده سازی نمودیم. بر این اساس، از هر نمونه آموزشی دادگان EMODB، ۳۰۰ نمونه‌ی جدید، با حفظ احساس پایه موجود در داده‌ی اصلی، ایجاد گردید. برای تولید داده‌های جدید، فرآیندهای ذیل اجرا گردیدند:

- تغییر گام گفتار: با احتمال ۰.۵ اعمال گردید.
- کشش زمانی: با استفاده از بازه ضریب کشش [۰.۳،+۰.۳-] ثانیه روی همه نمونه‌ها اعمال شد.
- نویز افزایشی: با نسبت سیگنال به نویز کنترل شده، به صورت تصادفی در بازه [۲۰، ۴۰-] دسی بل بر روی همه نمونه‌ها اعمال گردید.

این تبدیلات در طول فرآیند آموزش اعمال شدند تا مدل در معرض مجموعه‌ای متنوع و چالش‌برانگیز از توالی‌های صوتی قرار گیرد و بدین ترتیب تنوع واقعی داده‌ها شبیه‌سازی شود. هسته اصلی استخراج ویژگی در این پژوهش یک شبکه BiLSTM است. به منظور جلوگیری از نشت داده و تضمین استقلال گوینده، نمونه‌های مربوط به هر گوینده به‌طور کامل تنها در یکی از مجموعه‌های آموزش یا آزمایش قرار داده شدند. همچنین از میان داده‌های آموزشی، ۱۵٪ به‌عنوان مجموعه اعتبارسنجی در نظر گرفته شد تا برای نظارت بر فرآیند آموزش شبکه BiLSTM و جلوگیری از بیش‌برازش مورد استفاده قرار گیرد. در نهایت، توزیع داده‌ها به این صورت تعیین شد: ۶۸٪ برای آموزش، ۱۲٪ برای اعتبارسنجی و ۲۰٪ برای آزمایش. علاوه بر این، لایه‌بندی داده‌ها به‌منظور حفظ نسبت مناسب هر کلاس احساسی در تمامی‌بخش‌ها اعمال شد. عملکرد مدل نیز با استفاده از یک طرح ارزیابی سخت‌گیرانه و مستقل از گوینده سنجیده شد؛ به‌طوری‌که مدل بر اساس گفته‌های گویندگانی که در مرحله آموزش حضور نداشتند مورد آزمایش قرار گرفت. معماری BiLSTMLayer، با ۳۰۰ واحد مخفی طراحی گردید و به منظور جلوگیری از بیش برازش، در قبل و بعد از آن از یک لایه‌ی dropoutLayer (با نرخ ۰.۵-) استفاده نمودیم. در این مقاله، پارامترهای مقدارپذیر لایه فوق به شرح ذیل تنظیم گردیدند: وزن‌های ورودی (۲۴۰۰×۴۰۰)، وزن‌های بازگشتی (۲۴۰۰×۳۰۰) و سوگیری (۲۴۰۰×۱).

در ادامه، مدل با استفاده از ابرپارامترهای زیر آموزش داده شد: اندازه مینی بچ ۵۱۲، نرخ یادگیری اولیه ۰.۰۰۵ با دوره افت ۲ دوره‌ای و حداکثر ۵ دوره آموزش. خروجی‌های لایه‌ی نهایی BiLSTM به عنوان بردار ویژگی ورودی در طبقه‌بند جنگل تصادفی، که آخرین لایه تصمیم گیری در الگوریتم پیشنهادی است، مورد استفاده قرار می‌گیرد.

در این پژوهش، جنگلی متشکل از ۳۰۰ درخت تصمیم گیری ساخته شد که هر کدام با استفاده از معیار ناخالصی

[28] Recall

[29] Sensitivity

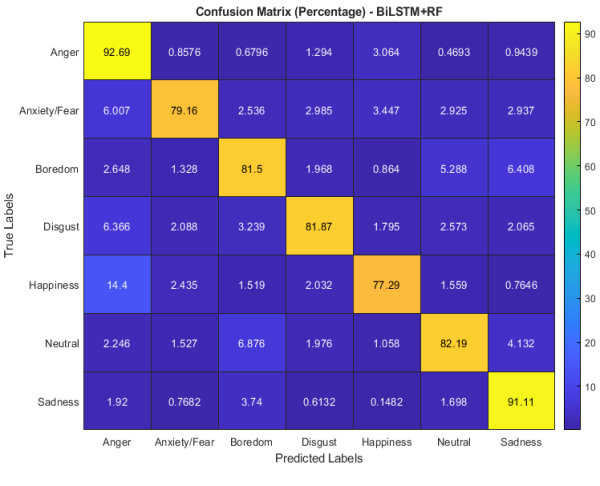
UNCORRECTED PROOF

جینی ایجاد گردیده‌اند. برای هر درخت، یک نمونه بوت استرپ از داده‌های آموزشی و زیرمجموعه‌ای تصادفی از ویژگی‌ها استفاده شد تا تنوع درون مجموعه تضمین شود. تصمیم گیری در خصوص احساس داده‌ی هدف با رأی گیری اکثریت در همه شاخه‌ها تعیین می‌شود. شکل (۵) آموزش شبکه عصبی BiLSTM را برای دادگان ارتقا یافته‌ی EMODB نشان می‌دهد که در آن دقت کلی، به عنوان معیار اصلی گزارش شده است.



شکل ۵- آموزش شبکه عصبی BiLSTM برای دادگان EMO-DB

عملکرد نهایی سیستم از طریق ماتریس درهم ریختگی شکل (۶) برای دادگان EMODB بیان گردیده است که تجزیه و تحلیل دقیق پیش‌بینی‌ها در مقابل برچسب‌های واقعی را ارائه می‌دهد و امکان تحلیل سردرگمی بین احساسات را فراهم می‌کند.



شکل ۶- ماتریس در هم ریختگی در دادگان EMO-DB به صورت درصد

جدول (۲)، مقایسه‌ی عملکرد مدل پیشنهادی و سایر مدل‌های مشابه را نشان می‌دهد. در این راستا، جدول (۳) دقت و بازیابی کلاسی مدل پیشنهادی بر روی دادگان توسعه یافته‌ی EMODB را مبتنی بر هر کلاس نشان می‌دهد.

جدول ۲- تحلیل عملکرد مدل‌ها

نوع مدل	دقت مدل(درصد)
الگوریتم جنگل تصادفی	۶۶.۶
LSTM	۷۳.۷
کانولوشن مبتنی بر انتخاب ویژگی	۷۷.۶
پیچشی+LSTM	۷۷.۱
ماشین بردار پشتیبان	۷۵.۷۹
پیچشی+ ترنسفورمر	۸۰.۹۴
مدل پیشنهادی	۸۵.۱۱

جدول۳- دقت کلاسی و بازیابی کلاسی مدل پیشنهادی بر روی دادگان EMODB در هر کلاس

احساس	دقت کلاسی(%)	دقت کلی(%)
خشم	۸۳.۶۰	**۹۲.۵۷**
ترس/ نگرانی	۸۶.۱۹	**۸۱.۰۰**
خستگی	۸۳.۴۶	**۷۹.۴۵**
انزجار	۸۵.۱۵	**۸۲.۲۰**
شادی	۸۵.۵۲	**۷۷.۸۹**
خنثی	۸۲.۷۷	**۸۰.۶۰**
غم	۸۷.۴۵	**۹۰.۸۰**

۵- تحلیل نتایج و بحث

تحلیل نتایج بر اساس جدول (۲) و مقایسه‌ی آن با عملکرد سایر مدل‌ها نشان می‌دهد که ساختار پیشنهادی عملکرد بهتری نسبت به سایر روش‌ها داشته است. نتایج این جداول نشان می‌دهد که سیستم ارائه شده دقتی برابر ٪۸۵.۱۱ بر روی دادگان EMODB ارائه نموده و پایداری قابل‌توجهی از خود نشان داده است که طبق معیارهای گفته شده عملکردی فوق العاده محسوب می‌گردد. این یافته‌ها به وضوح بیانگر اثربخشی یکپارچه‌سازی، افزایش هوشمند تعداد داده‌ها، یادگیری عمیق برای استخراج ویژگی و یادگیری ماشین کلاسیک برای طبقه‌بندی در توسعه سیستم‌های بازشناسی عملی و مقاوم در محیط‌های واقعی است. شکل (۶) ماتریس در هم ریختگی مدل پیشنهادی روی دادگان EMODB را به صورت درصد نشان می‌دهد. جدول (۳) عملکرد مدل پیشنهادی را در تشخیص هر یک از هفت کلاس احساسی پایگاه داده EMODB نشان

UNCORRECTED

می‌دهد. این جدول دو معیار اصلی را گزارش می‌کند؛ بازیابی: نسبت نمونه‌های واقعی هر کلاس که به درستی شناسایی شده‌اند و دقت: نسبت پیش‌بینی‌های مثبت هر کلاس که واقعاً درست بوده‌اند. تحلیل تفصیلی این پارامترها بر اساس کلاس‌های احساسی به صورت زیر است:

- خشم: بازیابی ۹۲/۵۷٪ (بالاترین مقدار در بین همه کلاس‌ها) و دقت ۸۳/۶۰٪. خشم، بالاترین بازیابی را دارد که نشان می‌دهد سیستم اکثر نمونه‌های واقعی خشم را شناسایی می‌کند. با این حال، دقت پایین‌تر (نسبت به بازیابی) نشان می‌دهد که برخی از نمونه‌های دیگر احساسات به اشتباه به عنوان خشم طبقه‌بندی شده‌اند.
- غم: بازیابی ۹۰/۸۰٪ (دومین مقدار بالا) و دقت ۸۷/۴۵٪ (بالاترین مقدار در بین همه کلاس‌ها)؛ غم هم بازیابی بالا و هم دقت بسیار بالایی دارد. این نشان می‌دهد سیستم نه تنها بیشتر نمونه‌های واقعی غم را شناسایی کرده است، بلکه پیش‌بینی‌های غم آن نیز بسیار قابل اعتماد است.
- اضطراب/ترس: بازیابی ۸۱/۰۰٪ و دقت ۸۶/۱۹٪. عملکردی متعادل با دقت نسبتاً بالا؛ این احساسات ممکن است دارای ویژگی‌های صوتی مشترک با سایر احساسات باشند.
- تنفر :بازیابی ۸۲/۲۰٪ و دقت ۸۵/۱۵٪. عملکرد خوب و متعادل، مشابه با اضطراب/ترس دارد.
- خستگی: بازیابی۷۹/۴۵٪ و دقت ۸۳/۴۶٪ پایین‌ترین بازیابی در بین همه کلاس‌ها. خستگی ممکن است دارای ویژگی‌های صوتی کم‌مشخصه‌تر یا مشابه با حالت خنثی باشد.
- خنثی: بازیابی ۸۰/۶۰٪ و دقت۸۲/۷۷٪. عملکرد متوسط؛ تشخیص حالت خنثی اغلب چالش‌برانگیز است زیرا فاقد نشانه‌های عاطفی بارز است.
- شادی: بازیابی ۷۷/۸۹٪ (دومین مقدار پایین) و دقت ۸۵/۵۲٪. جالب توجه است که شادی با وجود دقت بالا (۸۵/۵۲٪)، بازیابی نسبتاً پایینی دارد (۷۷/۸۹٪). این ممکن است نشان دهد سیستم در تشخیص شادی عملکردی محافظه‌کارانه دارد.

با توجه به توضیحات فوق و الگوهای شناسایی‌شده، در روش پیشنهادی احساسات با شدت بالا بهتر تشخیص داده می‌شوند. در این میان خشم و غم (احساسات با برانگیختگی بالا) بهترین عملکرد را دارند که این امر به دلیل ویژگی‌های صوتی متمایز این احساسات است (مانند زیروبمی‌بالا، انرژی زیاد). در بررسی‌های صورت گرفته به نظر می‌رسد احساسات مثبت چالش‌برانگیزتر هستند. شادی و خستگی (که می‌توانند به عنوان حالات کم‌انرژی در نظر گرفته شوند) پایین‌ترین بازیابی را دارند. این یافته با پژوهش‌های قبلی همسوست که نشان می‌دهند احساسات مثبت اغلب تنوع بیان بیشتری دارند.

۶-نتیجه گیری

این پژوهش یک چارچوب ترکیبیِ سه مرحله‌ای برای بازشناسی احساسات گفتاری ارائه و به‌صورت جامع اعتبارسنجی می‌نماید. برخلاف رویکردهای متداول که تنها بر یک مؤلفه نظیر پیش‌پردازش، یادگیری عمیق، یا طبقه‌بندی متمرکز می‌شوند، راهکار پیشنهادی این مقاله ترکیب هوشمندانه‌ای از افزایش داده هدفمند، شبکه BiLSTM و طبقه‌بند جنگل تصادفی را در یک خط مشی یکپارچه پیاده‌سازی می‌کند. این تلفیقِ هم‌افزا، نه‌تنها دقت را بهبود می‌بخشد، بلکه پایداری سیستم در شرایط صوتی چالش‌برانگیز را نیز به‌طور چشمگیری ارتقا می‌دهد. این یافته‌ها دو نکته کلیدی را برجسته می‌سازد: ۱) موثر بودن هم‌زمانی فرآیندهای یادگیری ماشین و داده افزایی در ساخت سیستم‌های عملی شناسایی احساسات؛ ۲) تعمیم‌پذیری زبانی و فرهنگی معماری پیشنهادی؛ چرا که بدون آموزش مجدد، عملکرد قابل‌مقایسه‌ای در دو فضای زبانی متمایز نشان می‌دهد.

به طور خلاصه در این مقاله:

۱) یک خط تولید جامع برای افزایش داده‌ها طراحی گردید که با استفاده از تغییر زیر و بمی، کشش زمانی و نویز سفید کنترل شده‌ی افزایشی، مجموعه آموزش EMODB را به طور مصنوعی گسترده و متنوع‌تر نمودیم. این موضوع در جلوگیری از برازش بیش از حد و آموزش مدل برای تمرکز بر ویژگی‌های مرتبط با احساسات در میان تغییرات صوتی حیاتی بود.

UNCORRECTED PROOF

۲) یک مدل هم‌افزا طراحی نمودیم که در آن حافظه‌ی بلند-کوتاه مدت دوطرفه، BiLSTM، به عنوان موتور ویژگی زمانی عمل نموده و جاسازی‌های زمینه‌ای سطح بالا را از دنباله‌های گفتار استخراج می‌کند. سپس این جاسازی‌ها، به یک طبقه‌بند جنگل تصادفی منتقل می‌شوند که تصمیم گیری مبتنی بر کلاس داده‌ها در آن، پایداری استثنایی و قدرتمنی را در ایجاد تمایز برای بازشناسی احساسات فراهم می‌کند و تصمیم‌گیری نهایی را پایدار می‌سازد.

۳) ارزیابی‌های جامع صورت گرفته بر روی پایگاه داده توسعه یافته‌ی EMODB، عملکرد قدرتمند چارچوب پیشنهادی را تأیید می‌کند. نتایج بدست آمده نشان می‌دهد که سیستم پیشنهادی دقتی برابر ۸۵.۱۱٪ ارائه می‌کند.

علیرغم موفقیت‌های چشم گیر روش پیشنهادی، این تکنیک نیز همانند سایر پژوهش‌ها با محدودیت‌هایی همراه است. از معایب روش پیشنهادی می‌توان به سرعت نسبتا پایین فرآیند یادگیری حافظه‌ی بلند-کوتاه مدت دوطرفه اشاره نمود که با بهره‌گیری از سیستم‌های پردازنده پرسرعت مانند GPU می‌توان نسبت به رفع آن اقدام نمود. در این خصوص لازم به ذکر است که این شبکه پس از یادگیری، از سرعت بالایی در فرآیند آزمایش برخوردار است و بدین واسطه بهره گیری از آن در فرآیند شناسایی، پس از آموزش، هیچ گونه خللی ایجاد ننموده و میتواند در کاربردهای زمان-واقعی نیز مورد استفاده قرار گیرد.

این ملاحظات، مسیر روشنی را برای پژوهش‌های آینده ترسیم می‌کند:

۱) یکی از مهم‌ترین مسیرهای تحقیقاتی آینده، گسترش ارزیابی مدل در سناریوهای چندزبانه و چندپایگاهی است. در این پژوهش که معماری پیشنهادی بر روی پایگاه داده EMODB مورد ارزیابی قرار گرفت، نتایج نشان‌دهنده عملکرد مناسب آن بوده است. با این حال، بررسی عملکرد این معماری بر روی مجموعه داده‌های متنوع‌تر و زبان‌های مختلف می‌تواند دید جامع‌تری درباره میزان تعمیم‌پذیری مدل فراهم کند. انجام آزمایش‌های میان-دادگانی[30]، به‌گونه‌ای که مدل بر روی یک پایگاه داده آموزش داده شده و بر روی پایگاه داده‌ی دیگر مورد آزمون قرار گیرد، می‌تواند توانایی مدل در انتقال دانش و سازگاری با شرایط داده‌ای متفاوت را به‌طور دقیق‌تری مشخص کند. چنین مطالعاتی به توسعه سیستم‌های بازشناسی احساس مقاوم‌تر و قابل استفاده در محیط‌های واقعی کمک خواهد کرد.

۲) توسعه و بهبود روش‌های داده‌افزایی می‌تواند به‌عنوان یکی از مسیرهای مهم تحقیقاتی آینده در نظر گرفته شود. در این پژوهش از تکنیک‌هایی نظیر تغییر زیر و بمی‌صدا، کشش زمانی و افزودن نویز کنترل‌شده برای افزایش تنوع داده‌های آموزشی استفاده شده است. در مطالعات آینده می‌توان روش‌های پیشرفته‌تر داده‌افزایی را مورد بررسی قرار داد؛ به‌عنوان مثال شبیه‌سازی شرایط صوتی مختلف، استفاده از نویزهای محیطی واقعی، یا به‌کارگیری روش‌های یادگیری مولد برای تولید نمونه‌های صوتی جدید. این رویکردها می‌توانند به افزایش تنوع داده‌ها، بهبود تعمیم‌پذیری مدل و کاهش مشکل بیش‌برازش در سیستم‌های یادگیری عمیق کمک کنند.

۳) بررسی و توسعه معماری‌های پیشرفته‌تر برای استخراج ویژگی‌های زمانی می‌تواند به بهبود بیشتر عملکرد سیستم منجر شود. در این پژوهش، از شبکه BiLSTM برای مدل‌سازی وابستگی‌های زمانی سیگنال گفتار استفاده نمودیم. در مقالات آتی می‌توان عملکرد این معماری را با مدل‌هایی مانند شبکه‌های مبتنی بر مکانیسم توجه[31]، معماری‌های انتقال[32] و یا ترکیب شبکه‌های پیچشی با مدل‌های توالی مقایسه کرد. همچنین بررسی ساختارهای ترکیبی دیگر و تنظیم دقیق ابرپارامترهای شبکه می‌تواند به دستیابی به

[30] cross-corpus
[31] attention
[32] transformer

UNCORRECTED PROOF

نمایش‌های غنی‌تر ویژگی‌ها و در نتیجه افزایش دقت سیستم کمک کند.

۴) بررسی طبقه‌بندهای جایگزین و روش‌های ترکیبی در مرحله تصمیم‌گیری نیز می‌تواند یکی دیگر از مسیرهای تحقیقاتی آینده باشد. در این پژوهش از الگوریتم جنگل تصادفی به‌عنوان طبقه‌بند نهایی استفاده شده است که به دلیل ساختار مجموعه‌ای خود از پایداری مناسبی در مقابل عوامل ناخواسته برخوردار بود. در پژوهش‌های بعدی می‌توان عملکرد این طبقه‌بند را با روش‌های دیگری نظیر ماشین بردار پشتیبان، شبکه‌های عصبی چندلایه یا الگوریتم‌های تقویتی مقایسه کرد و یا از روش‌های ترکیبی چندطبقه‌بندی‌کننده برای افزایش دقت و پایداری سیستم استفاده نمود.

۵) در نهایت، پیاده‌سازی و ارزیابی سیستم در کاربردهای عملی و محیط‌های واقعی نیز از دیگر جهت‌گیری‌های مهم برای تحقیقات آینده محسوب می‌شود.

بسیاری از سیستم‌های بازشناسی احساسات در محیط‌های آزمایشگاهی و با داده‌های کنترل‌شده ارزیابی می‌شوند، در حالی که در کاربردهای واقعی شرایط صوتی، پیچیده‌تر و دارای تغییرات بیشتری هستند. بنابراین، بررسی عملکرد مدل پیشنهادی در سناریوهایی نظیر ربات‌های اجتماعی، سامانه‌های پشتیبانی مشتری، دستیارهای صوتی و رابط‌های تعامل انسان و ماشین می‌تواند نقش مهمی در انتقال این فناوری از محیط‌های پژوهشی به کاربردهای عملی ایفا کند. به طور کلی، توسعه آزمایش‌های چندپایگاهی، بهبود روش‌های داده‌افزایی، استفاده از معماری‌های عمیق پیشرفته‌تر و ارزیابی سیستم در شرایط واقعی می‌تواند مسیرهای مؤثری برای ادامه این پژوهش و ارتقای عملکرد سیستم‌های بازشناسی احساس از گفتار در آینده فراهم سازد.

مراجع

Khalil, R. A. 2019. “Speech Emotion Recognition Using Deep Learning Techniques: A Review.” IEEE Access 7: 117327–117345.

Dey, A. 2020. “A Hybrid Meta-Heuristic Feature Selection Method Using Golden Ratio and Equilibrium Optimization Algorithms for Speech Emotion Recognition.” IEEE Access 8: 200909–200922.

Li, H. 2024. “A Lightweight Multi-Scale Model for Speech Emotion Recognition.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 32: 3045–3058.

Er, M. B. 2020. “A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features.” IEEE Access 8: 221822–221831.

Jo, A.-H. 2025. “Classification of Speech Emotion State Based on Feature Map Fusion of TCN and Pretrained CNN Model from Korean Speech Emotion Data.” IEEE Access 13: 12345–12356.

Wani, T. M. 2021. “A Comprehensive Review of Speech Emotion Recognition Systems.” IEEE Access 9: 47798–47814.

Noroozi, F. 2017. “Vocal-Based Emotion Recognition Using Random Forests and Decision Tree.” International Journal of Speech Technology 20 (2): 239–247.

Xie, Y. 2019. “Speech Emotion Classification Using Attention-Based LSTM.” In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6725–6729.

Kwon, S. 2020. “CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network.” Mathematics 8 (12): Article 2133.

Yang, L. 2021. “Speech Emotion Analysis of Netizens Based on Bidirectional LSTM and PGCDBN.” IEEE Access 9: 65432–65441.

Pourebrahim, Y. 2023. “Speech Emotion Recognition Using a Combination of Transformer and Convolutional.” Journal of Intelligent Processing in Electrical Technology 13 (52): 45–58. (In Persian).

Kandali, A. B., A. Routray, and T. K. Basu. 2008. “Emotion Recognition from Assamese Speeches Using MFCC Features and GMM Classifier.” In Proceedings of the IEEE Region 10 Conference (TENCON), 1–5.

UNCORRECTED PROOF

Milton, A., S. S. Roy, and S. T. Selvi. 2013. “SVM Scheme for Speech Emotion Recognition Using MFCC Feature.” International Journal of Computer Applications 69 (9): 34–39.

de Pinto, M. G., M. Polignano, P. Lops, and G. Semeraro. 2020. “Emotions Understanding Model from Spoken Language Using Deep Neural Networks and Mel-Frequency Cepstral Coefficients.” In Proceedings of the IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), 1–5.

Anvarjon, T., Mustaqeem, and S. Kwon. 2020. “Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features.” Sensors 20 (18): Article 5212.

Ma, X., Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai. 2018. “Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms.” In Proceedings of Interspeech, 3683–3687.

Zhang, S., S. Zhang, T. Huang, and W. Gao. 2018. “Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching.” IEEE Transactions on Multimedia 20 (6): 1576–1590.

Liu, Z.-T., M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan. 2018. “Speech Emotion Recognition Based on Feature Selection and Extreme Learning Machine Decision Tree.” Neurocomputing 273: 271–280.

Chen, M., and X. Zhao. 2020. “A Multi-Scale Fusion Framework for Bimodal Speech Emotion Recognition.” In Proceedings of Interspeech, 374–378.

Masters, T. 2018. “Supervised Feedforward Networks.” In Deep Belief Nets in C++ and CUDA C, Vol. 1, 9–89. Berkeley, CA: Apress.

Hinton, G. E., S. Osindero, and Y.-W. Teh. 2006. “A Fast Learning Algorithm for Deep Belief Nets.” Neural Computation 18 (7): 1527–1554.

Ardan Dal Ri, F., F. Cifariello Ciardi, and N. Conci. 2023. “Speech Emotion Recognition and Deep Learning: An Extensive Validation Using Convolutional Neural Networks.” IEEE Access 11: 118785–118798. https://doi.org/10.1109/ACCESS.2023.3326671.

Burkhardt, F., et al. 2005. “A Database of German Emotional Speech.” In Proceedings of Interspeech, 1517–1520.

Zhao, Y., X. Mao, and Z. Zhang. 2023. “Noise-Robust Speech Emotion Recognition via Multi-Task Learning and Cross-Corpus Adaptation.” IEEE Transactions on Affective Computing 14 (2): 1324–1336. https://doi.org/10.1109/TAFFC.2022.314159.

Park, D. S. 2019. “A Simple Data Augmentation Method for Automatic Speech Recognition.” In Proceedings of Interspeech, 2613–2617.

Chaudhari, S. S., and A. V. Jadhav. 2019. “MFCC and Its Variants for Speech and Speaker Recognition: A Review.” Circuits, Systems, and Signal Processing 38: 523–548.

Zhang, S., et al. 2021. “Speech Emotion Recognition Using Deep Neural Networks.” IEEE Access 9: 15212–15224.

Kim, J., et al. 2020. “Robust Speech Emotion Recognition Using Bidirectional LSTM Networks.” IEEE Transactions on Affective Computing 13 (1): 200–210.

UNCORRECTED PROOF