Journal of New Approach to Children's Education (JNACE) **Original Article**

Winter-2024, Volume 5, No 4 **DOI: 10.22034/NAES.2024.451265.1386**

ISSN: 2716-9766

A Study on Evaluating the Effect of Voice Activity Detection (VAD) Approach on Speech Emotion Recognition of Autistic Children

Seyed Mehdi Hoseini[1*]

1 Master of Computer Science, Scientific Computing, Department of Computer Science, Mazandaran University, Babolsar, Mazandaran, Iran

*** Corresponding author:** mehdihoseini.cs@gmail.com

Received: 2024-04-06 **Accepted:** 2024-05-17

Abstract

Background and Aim: Autism spectrum is a neurological disorder that manifests itself in the early years of a child's development. People with autism face challenges in regulating emotions and express their emotional states in different ways. The current research presents a vocal activity detection (VAD) system adapted to the voices of autistic children.

Methods: The proposed VAD system is a Recurrent Neural Network (RNN) with short-term memory (LSTM) cells. The data includes 25 English-speaking autistic children performing a structured learning activity and was collected as part of the DE-ENIGMA project.

Results: Our experiments show that the pediatric VAD system performs less well than our generic VAD system trained under the same conditions, as we obtain system performance characteristic curve under the curve (ROC-AUC) criteria of 0.662 and 0.850, respectively. The SER results show different performances between capacity and excitation, depending on the VAD system used, with a maximum match correlation coefficient (CCC) of 0.263 and a minimum root mean square error (RMSE) of 0.107.

Conclusion: Although the performance of SER models is generally low, the pediatric VAD system can lead to slightly improved results compared to other VAD systems and especially the VAD-less baseline, which supports the hypothesized importance of pediatric VAD systems in the context under discussion.

Keywords: Voice Activity Detection, Speech Emotion Detection, Recurrent Neural Network, Short-Term Memory Cells, Autism

How to Cite This Article: Hoseini, SM. (2024). A Study on Evaluating the Effect of Voice Activity Detection (VAD) Approach on Speech Emotion Recognition of Autistic Children. *JNACE*, 5(4): 194-206.

فصلنامه رویکردی نو بر آموزش کودکان | مقاله پژوهشی

زمستان ۱۴۰۲، دوره ۵، شماره ۴

DOI: 10.22034/NAES.2024.451265.1386

ISSN: 2716-9766

مطالعه ای بر ارزیابی تأثیر رویکرد تشخیص فعالیت صدا (VAD) بر تشخیص عواطف گفتاری کودکان اوتیستیک

سید مهدی حسینی [۱*]

[۱] کارشناس ارشد علوم کامپیوتر، گرایش محاسبات علمی، گروه علوم کامپیوتر، دانشگاه مازندران، بابلسر، مازندران، ایران

* نویسنده مسئول: mehdihoseini.cs@gmail.com

تاریخ دریافت مقاله: ۱۴۰۳/۰۱/۱۸ **تاریخ پذیرش مقاله:** ۱۴۰۳/۰۲/۲۸

چکیده

زمینه و هدف: طیف اوتیسم، اختلالی نورولوژیک است که خودش را در سال‌های اولیه رشد کودک نمایان می‌کند. افراد مبتلا به اوتیسم با چالش هایی در زمینه تنظیم احساسات مواجه هستند و حالات عاطفی خود را به روش های مختلف بیان می کنند. پژوهش فعلی یک سیستم تشخیص فعالیت صوتی (VAD) را ارائه می‌کند که با صداهای کودکان اوتیستیک سازگار شده است.

روش پژوهش: سیستم VAD ارائه شده یک شبکه عصبی بازگشتی (RNN) با سلول های حافظه کوتاه مدت (LSTM) است. داده ها شامل ۲۵ کودک اوتیستیک انگلیسی زبان است که یک فعالیت آموزشی ساختار یافته را انجام می دهند.

یافته ها: آزمایش‌های ما نشان می‌دهد که سیستم VAD کودک عملکرد کمتری نسبت به سیستم VAD عمومی ما دارد که تحت شرایط یکسان آموزش داده شده است، زیرا منحنی مشخصه عملکرد سیستم (ROC) را تحت معیارهای منحنی (ROC-AUC) به ترتیب ۰.۶۶۲ و ۰.۸۵۰ به دست می‌آوریم. نتایج SER عملکردهای متفاوتی را در بین ظرفیت و برانگیختگی، بسته به سیستم VAD مورد استفاده با حداکثر ضریب همبستگی تطابق (CCC)0.263 و حداقل ریشه میانگین مربعات خطا ۰.۱۰۷ نشان می‌دهد.

نتیجه‌گیری: اگرچه عملکرد مدل‌های SER به طور کلی پایین است، سیستم VAD کودک می‌تواند به نتایج کمی بهبود یافته در مقایسه با سایر سیستم‌های VAD و به ویژه تراز مبنا بدون VAD(VAD-less baseline) منجر شود، که از اهمیت فرضی سیستم‌های VAD کودک در زمینه مورد بحث حمایت می‌کند.

واژگان کلیدی: تشخیص فعالیت صدا، تشخیص عواطف گفتاری، شبکه عصبی بازگشتی، سلول‌های حافظه کوتاه مدت، اوتیسم

شیوه استناد به این مقاله: حسینی ، س م (۱۴۰۲) مطالعه ای بر ارزیابی تأثیر رویکرد تشخیص فعالیت صدا (VAD) بر تشخیص عواطف گفتاری کودکان اوتیستیک.*فصلنامه رویکردی نو بر آموزش کودکان* ، ۵(۴): ۲۰۶-۱۹۴.

مقدمه

شناخت عواطف فرایند شناسایی عواطف انسانی است. دقت افراد در تشخیص احساسات دیگران، معمولاً بسیار متفاوت است. استفاده از فناوری برای کمک به افراد در حوزه تشخیص احساسات، یک حوزه تحقیقاتی نسبتاً نوپا است. به‌طور کلی، این فناوری در صورتی که از روش‌های چندوجهی در پردازش داده‌ها استفاده کند، بسیار کارآمد خواهد بود.

تشخیص عواطف گفتار[۱] (SER) یکی از زیرشاخه‌های برجسته

محاسبات عاطفی است. وظایف SER مستمر[۲]، به ویژه در سناریوهای تعاملی، مانند تعاملات کودک و ربات به کمک ربات، می‌تواند مستعد ایجاد مصنوعات شنیداری و نمونه‌های محدود گفتار باشد و نیاز به تمایز بین نویز پس‌زمینه و نمونه‌های غنی از اطلاعات را ایجاد کند. بنابراین، سیستم‌های تشخیص فعالیت صوتی (VAD)[۳] معمولاً در وظایف SER برای حذف بخش‌های بدون صدا از سیگنال صوتی استفاده می‌شوند، به عنوان مثال در هارار[۴] و همکاران[۱] در سال ۲۰۱۷ ، القیفاری[۵] و همکاران[۲] در سال ۲۰۱۹ و آکچای و اوگوز[۶] [۳] در سال ۲۰۲۰ نمایش داده شده‌اند. با این حال، در سناریویی با بیش از یک گوینده، VAD به تنهایی ممکن است برای فیلتر کردن تمام اطلاعات غیر مرتبط در مورد وضعیت عاطفی یک سخنران خاص کافی نباشد.

اوتیسم یک وضعیت عصبی رشدی است که با مشکلات در ارتباطات اجتماعی و الگوهای محدود و تکراری رفتار، علایق یا فعالیت ها همراه است[۴]. تصویر بالینی اوتیسم ناهمگون (ناجور) است، از جمله تنوع در ویژگی‌های اوتیسم و مهارت‌های زبان گفتاری، و بیماری‌های همراهی که اغلب رخ می‌دهند[۵]، مانند اختلال اضطراب، اختلال نقص توجه–بیش فعالی، اختلال هماهنگی رشد یا اختلالات افسردگی[۹–۶].

مشکلات در مهارت‌های ارتباطی اجتماعی و شناخت و بیان احساسات در کودکان اوتیستیک می‌تواند تعامل با خانواده، همسالان و متخصصان را چالش‌برانگیز کند. با این حال، تنها چند پروژه تحقیقاتی بررسی کرده اند که چگونه فناوری اخیر از جمله هوش مصنوعی می تواند به درک بهتر نیازها و بهبود شرایط کودکان مبتلا به اوتیسم کمک کند: پروژه ASC-inclusion پلتفرمی را با هدف حمایت از کودکان در درک و بیان احساسات از طریق یک دنیای مجازی جامع[۱۰]، به عنوان مثال از طریق بازی ایجاد کرد[۱۱].

پروژه DE-ENIGMA1 بر درک بهتر رفتار و نیازهای کودکان اوتیسم در سناریوی تعامل ربات–انسان (RCI) به رهبری محقق تمرکز داشت و به بینش هایی در مورد قابلیت پیش بینی ربات در سناریوهای RCI با کودکان مبتلا به اوتیسم و همچنین پیش‌بینی شدت صفات مربوط به اوتیسم[۴] و تشخیص صداهای اکولالیک[۷] [۱۲] (به عنوان مثال، تکرار کلمات کودکان اوتیستیک بر اساس گفته های گفتاری طرفین مکالمه آنها) کمک کرد[۱۳].

شولر و همکاران وظیفه ای را برای تشخیص مبتنی بر گفتار کودکان مبتلا به اوتیسم و سایر اختلالات رشدی فراگیر معرفی کردند. به خصوص در زمینه SER برای افراد مبتلا به اوتیسم، داده‌ها بسیار کم به نظر می‌رسند که احتمالاً تا حدودی به دلیل هزینه‌های زمانی قابل توجه مورد نیاز برای جمع‌آوری چنین داده‌هایی از کودکان اوتیستیک است[۱۴].

از این مرور آثار مرتبط، آثار محدودی وجود دارد که با استراتژی‌های برچسب‌گذاری مداوم[۸]، احساسات کودکان اوتیستیک را مدل‌سازی می‌کند. تا آنجا که ما می دانیم، تاکنون هیچ تحقیقی به بررسی این موضوع پرداخته است که چگونه VAD می تواند چنین مدل سازی را بهبود بخشد.

در این مقاله، ما زیرمجموعه ای از داده های جمع آوری شده در پروژه DE-ENIGMA را بررسی می کنیم[۱۵]. داده‌های ارائه‌شده شامل حدود ۱۷ ساعت ضبط‌های صوتی و حاشیه‌نویسی‌های غنی[۹] از جمله حالت عاطفی درک شده به‌طور مداوم و دیاریشن سخن‌گو[۱۰] به صورت دستی است. داده‌ها چالش‌های متعددی را ایجاد می‌کنند که معمولاً با داده‌های مهارناشدنی مرتبط هستند، از جمله نویز (به عنوان مثال از حرکت روبات یا مبلمان) یا فاصله‌های مختلف تا میکروفون. علاوه بر این، یک چالش خاص در مجموعه داده فعلی ناشی از پراکندگی صداهای کودک در تعامل بین کودک، ربات و محقق است، زیرا چندین کودکی که در این مطالعه شرکت کردند، ارتباط گفتاری محدود و بدون محدودیت داشتند.

برخلاف کارهای رایج تشخیص احساسات مستمر، ما فرض کردیم که مدلی که به تنهایی بر صداهای کودک تمرکز می‌کند می‌تواند از مدل‌های دیگر بهتر عمل کند، زیرا انتظار داریم که صداهای کودک حاوی بیشترین اطلاعات در مورد حالات عاطفی کودکان باشد. به همین دلیل، در کار فعلی، ما یک سیستم VAD را پیاده سازی می کنیم که به طور خاص برای صداگذاری کودکان اوتیستیک در مجموعه داده ها آموزش داده شده است و عملکرد آن را در برابر یک سیستم VAD عمومی آموزش دیده (آموزش دیده بر روی تمام صداهای مجموعه داده ما) و همچنین اجرای VAD ارتباطات بلادرنگ وب[۱۱] (WebRTC) و حاشیه‌نویسی دستی دیاریشن سخنگو[۱۲]، برای کار SER در دست ارزیابی است. WebRTC VAD بر اساس مدل های مخلوط گاوسی[۱۳] (GMMs) و انرژی های ورود به سیستم شش باند فرکانسی[۱۴] است.

باقیمانده این مطالعه به شرح زیر تنظیم شده است. در بخش ۲، ما یک نمای کلی از مجموعه داده بررسی شده ارائه می دهیم. علاوه بر این، ما روش مبتنی بر یادگیری عمیق را برای هر دو وظیفه VAD و SER در بخش ۳ معرفی می‌کنیم. متعاقباً، نتایج تجربی را برای آزمایش‌های VAD جدا شده[۱۵] و همچنین وظیفه SER با یک سیستم ترکیبی VAD-SER در بخش ۴ ارائه می‌کنیم. در نهایت، قبل از پایان کار خود در بخش ۶، نتایج و محدودیت های رویکردهای خود را در بخش ۵ مورد بحث قرار می دهیم.

۲. مجموعه داده

آزمایش‌های این دست‌نوشته بر اساس زیرمجموعه‌ای از داده‌های جمع‌آوری‌شده در پروژه DE-ENIGMA Horizon 2020 است که در یک محیط مدرسه‌ای در بریتانیا و صربستان جمع‌آوری شده‌اند. در این کار، ما صرفاً بر روی داده‌های صوتی از مطالعه بریتانیایی پروژه تمرکز می‌کنیم، که تمام داده مرتبط برای آن موجود است. در اینجا، کودکان اوتیستیک فعالیت‌های آموزشی تشخیص هیجان را بر اساس برنامه آموزش ذهن‌خوانی به کودکان مبتلا به اوتیسم[۱۶]، تحت راهنمایی یک محقق انجام دادند. تأییدیه اخلاقی برای این مطالعه توسط کمیته اخلاق تحقیق در مؤسسه آموزش UCL و دانشگاه کالج لندن (REC 796) اعطا شد. کودکان به طور تصادفی به جلسات پژوهشگر یا جلساتی که توسط ربات انسان نما Zeno-R2 پشتیبانی می شد، تقسیم شدند. Zeno قادر به انجام حالات مختلف چهره مرتبط با احساسات است که توسط محقق از طریق یک رابط خارجی کنترل می شود. جلسات با چندین دوربین و میکروفون که زوایای مختلف اتاق را پوشش می دادند ضبط می شد.

هر کودک بین یک تا پنج جلسه روزانه (به طور متوسط ۳.۴) شرکت کرد که در مجموع ۸۴ جلسه با میانگین طول ۱۲.۴ دقیقه از ۲۵ کودک (۱۹ مرد، ۶ زن)، ۱۳ شرکت کننده در جلسات فقط محقق و ۱۲ شرکت کننده در جلسات با کمک ربات، با میانگین سنی ۸.۲ سال (انحراف استاندارد: ۲.۵ سال)، که توسط سه محقق مختلف (فقط یک محقق برای هر کودک) رهبری می‌شوند، انجام شد.

ما داده ها را به شیوه ای مستقل از سخنران با توجه به کودکان تقسیم کردیم. از آنجایی که در مجموع سه محقق در مجموعه داده‌ها وجود داشت، هر کودک فقط با یک محقق تعامل داشت، تقسیم‌بندی داده‌های خود را بر اساس محققان گروه‌بندی می‌کنیم. ما این کار را انجام می‌دهیم تا از تطبیق بیش از حد مدل‌های یادگیری ماشینی خود بر روی ویژگی‌های گفتاری خاص افراد محققین که تا حد زیادی در صداسازی نقش دارند، جلوگیری کنیم. یک نمای کلی از پارتیشن ها در جدول ۱ آورده شده است. پارتیشن بندی برای هر دو نوع آزمایش استفاده می شود.

جدول ۱ : مروری بر سه پارتیشن مجموعه داده: آموزش، توسعه (dev.) و تست.

Partition	# Children	#Sessions	#researchers	Child vocalisations	Total vocalisations	Total duration
Train	12	41	1	1:26:39	6:42:15	9:43:34
Dev.	4	15	1	0:18:27	1:24:37	3:14:35
Test	9	28	1	0:32:42	2:35:21	4:22:03
Overall	25	84	3	2:17:49	10:42:14	17:20:13

جلسات از نظر داده های صوتی و تصویری، به دنبال یک پروتکل حاشیه نویسی (حاشیه‌نویسی) از پیش تعریف شده، بسیار حاشیه نویسی شدند. از جمله دستورالعمل‌هایی برای حواشی گوینده، نوع آواسازی، وقوع اکولالیا، نوع آوازهای غیرکلامی، و همچنین احساسات از نظر ظرفیت و برانگیختگی. برای مطالعه خود، از حاشیه‌نویسی‌های حواشی سخنران[۱۶]، مبدا برچسب‌ها[۱۷] برای تشخیص فعالیت صوتی، و همچنین حاشیه‌نویسی ظرفیت[۱۸] و برانگیختگی[۱۹] به عنوان برچسب‌هایی برای سیستم SER استفاده می‌کنیم.

۲.۱. حاشیه نویسی دیاریسیشن سخنران [۲۰]

دیاریسیشن سخنران یا Speaker Diarisation توسط انگلیسی زبانان مسلط با استفاده از ابزار حاشیه نویسی ELAN2 انجام شد. وظیفه برجسته کردن هر گونه صدای هر سخنران حاضر در جلسه بود، به عنوان مثال، کودک، محقق، هر فرد دیگر (به طور کلی یک معلم)، یا ربات Zeno. حاشیه نویسان[۲۱] می توانستند تصمیمات خود را بر اساس ترکیبی از جریان های ویدئویی موجود به همراه یکی از ضبط های صوتی بومی دوربین های ویدئویی و همچنین به تصویر کشیدن شکل موج صوتی خام، مبتنی کنند. ابزار حاشیه‌نویسی به حاشیه نویسان اجازه می داد به نقاط دلخواه ضبط بپرند. به طور کلی، هر جلسه توسط یک حاشیه نویس ارزیابی شد.

۲.۲. حاشیه نویسی احساسات

حاشیه نویسی احساسات در پایگاه داده با هدف ثبت ابعاد عاطفی ظرفیت و برانگیختگی[۲۲]، یعنی نمایش مداوم از میزان مثبت یا منفی (ظرفیت) و میزان خواب آلود یا برانگیختگی (برانگیختگی) یک حالت احساسی به نظر می رسد. ابعاد عاطفی معمولاً جایگزینی برای احساسات طبقه بندی شده مانند شادی، عصبانیت و غیره در هنگام ارزیابی حالات عاطفی افراد است. پنج ارزیاب خبره، همگی زبان مادری یا نزدیک به زبان انگلیسی، درک خود را از ارزش‌های ظرفیت و برانگیختگی بیان شده توسط کودکان در هر جلسه با در نظر گرفتن همان

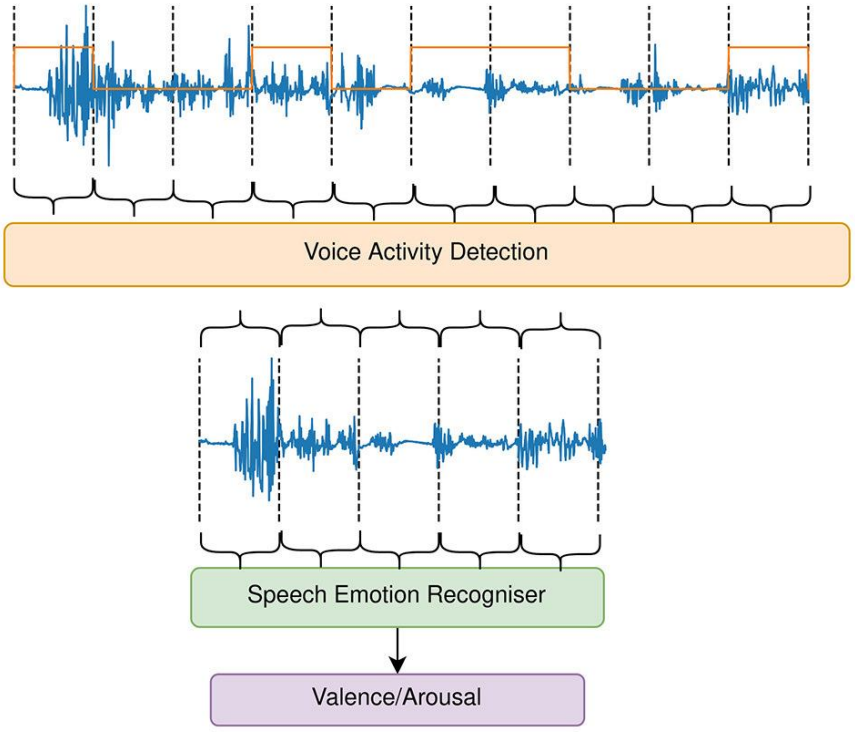
داده‌های ویدیویی و صوتی که در تکلیف حواشی گوینده بیان می‌کنند، شرح دادند.

برای فرآیند حاشیه نویسی، به ارزیاب ها جوی‌استیک[۲۳] (مدل Logitech Extreme 3D Pro) داده شد تا ظرفیت و برانگیختگی را به طور جداگانه حاشیه نویسی کنند. در حالی که حاشیه نویسان در حال تماشای ضبط جلسات بودند، موقعیت جوی‌استیک را تغییر دادند که به طور مداوم با نرخ نمونه برداری ۵۰ هرتز نمونه برداری می شد و درجه و علامت ظرفیت تخمینی یا مقادیر برانگیختگی را نشان می داد (مثبت در موقعیت بالا، منفی در یک موقعیت پایین).

حاشیه نویسی های حاشیه نویس های مختلف برای هر جلسه در یک دنباله استاندارد طلایی[۲۴] با استفاده از استاندارد طلایی تخمین‌گر وزنی ارزیاب[۲۵] (EWE) [۱۴] خلاصه می شود. استاندارد طلایی EWE معمولاً در وظایف تشخیص احساسات استفاده می‌شود[۱۸، ۱۷] و وزن‌های ویژه حاشیه‌نویس[۲۶] را بسته به همبستگی زوجی حاشیه‌نویسی‌ها[۲۷] در نظر می‌گیرد. برای آزمایش‌های خود، با محاسبه میانگین دوم[۲۸] بر روی حاشیه‌نویسی‌های استاندارد طلایی[۲۹]، تنها از یک برچسب احساس[۳۰] در هر ثانیه استفاده می‌کنیم.

۳. روش پژوهش

برای بررسی وظیفه SER مبتنی بر VAD، ما از دو مدل مجزا بر اساس استخراج ویژگی و شبکه‌های عصبی مکرر[۳۱] (RNN) با سلول‌های حافظه کوتاه‌مدت[۳۲] (LSTM) استفاده می‌کنیم. جزء اول، یک جزء VAD و دومی، یک جزء SER است. مدل VAD با تکه های صوتی طولانی ۱ ثانیه ارائه می شود و هدف آن برچسب زدن بخش هایی از سیگنال صوتی با یک صداسازی حاضر[۳۳] است. سپس مدل SER بر روی بخش‌های صوتی احتمالاً حاوی گفتار با هدف پیش‌بینی ظرفیت و برانگیختگی ابعاد عاطفی[۳۴] به صورت پیوسته آموزش داده می‌شود. تصویری از سیستم ترکیبی در شکل ۱ نشان داده شده است.



شکل ۱. استفاده متوالی از تشخیص فعالیت صوتی (VAD) و سیستم تشخیص احساسات گفتار (SER).

شکل ۱: استفاده متوالی از تشخیص فعالیت صوتی (VAD) و سیستم تشخیص احساسات گفتار (SER). سیستم VAD تکه‌های یک ثانیه‌ای سیگنال صوتی را حذف می‌کند، جایی که هیچ صدایی شناسایی نمی‌شود. هر تکه یک ثانیه ای باقیمانده به سیستم SER داده می شود تا مقادیر پیوسته برای ظرفیت و برانگیختگی را پیش بینی کند.

۳.۱. تشخیص فعالیت صوتی

از آنجایی که هدف سیستم VAD حذف هر چه بیشتر داده های کم عمق[۳۵] از داده های صوتی است، ما چندین روش را در اینجا با هم مقایسه می کنیم:

در سطح اول، ما سعی می کنیم تمام صداها را با سیستم های

VAD عمومی فیلتر کنیم، سیستمی که به طور خاص در مجموعه داده ما آموزش دیده است. دیگری اجرای سیستم WebRTC VAD، است که معمولاً به عنوان مقایسه برای سایر سیستم های VAD استفاده می شود، به عنوان مثال[۱۱، ۱۹].

امتیاز پرخاشگری(تجاوزکاری)[۳۶] WebRTC VAD برابر با یک تعیین شده است. علاوه بر این، ما از حاشیه نویسی های حقیقت‌پایه(عینی)[۳۷] برای همه صداها به عنوان یک استاندارد طلایی برای یک سیستم کلی VAD استفاده می کنیم.

در سطح دوم، ما سعی می‌کنیم فقط آوازهای کودکان را فیلتر کنیم، که احتمالاً حاوی بیشترین اطلاعات در مورد حالت عاطفی کودکان است. برای این کار، ما یک سیستم VAD کودک را بر روی مجموعه داده‌های ذکر شده در بالا آموزش می‌دهیم و از حاشیه‌نویسی‌های حقیقت پایه (عینی) برای صداهای کودک برای مقایسه بیشتر استفاده می‌کنیم. ارزیابی تأثیرات مختلف VADهای عمومی و VAD کودک مورد توجه بیشتر است، زیرا برخی اطلاعات در مورد وضعیت عاطفی کودکان را می توان از تعامل بین کودک و محقق بازیابی کرد. علاوه بر این، عملکرد بدتر سیستم VAD کودک در مقایسه با سیستم‌های عمومی VAD قوی‌تر می‌تواند منجر به تشخیص نویز محیط شود و بنابراین به طور بالقوه تأثیر منفی بر وظیفه SER دارد.

با توجه به مدت زمان بالقوه کوتاه[۳۸] صداسازی[۳۹]، ما ۱۳۰ عدد ComparE2016 LLD با اندازه فریم ۱۰ میلی ثانیه و اندازه پرش[۴۰] ۱۰ میلی ثانیه را از سیگنال صوتی خام با استفاده از جعبه ابزار openSMILE استخراج کردیم[۲۰]. سپس ویژگی‌های صوتی به یک RNN دو لایه دو جهته[۴۱] با سلول‌های LSTM و اندازه لایه مخفی[۴۲] ۱۲۸ واحد وارد می‌شوند و به دنبال آن یک لایه متراکم[۴۳] با یک نورون خروجی واحد[۴۴] که نشان‌دهنده اطمینان در تشخیص صدا است. معماری شبکه عصبی مشابه هاگرر[۴۵] [۲۱] است، اما بر اساس آزمایش های اولیه تنظیم شده است. ما از طول توالی (دنباله) ثابت ۱۰۰ نمونه در طول زمان آموزش استفاده می کنیم، به عنوان مثال، جریان صوتی به نمونه هایی با طول ۱ ثانیه بریده می شود. در حین آموزش این مساله رگرسیون، به هر فریم در صورت وجود گفتار برچسب ۱ و اگر وجود ندارد برچسب ۰ اختصاص داده می شود.

مدل های VAD برای ۸ دوره[۴۶] با اندازه دسته‌ای[۴۷] ۲۵۶ با استفاده از بهینه ساز Adam با نرخ یادگیری ۰.۰۱ و تابع ضرر[۴۸] میانگین مربعات خطا[۴۹] (MSE) داده شده است. ما تعداد نسبتاً کمی از دوره ها را بر اساس تعداد زیادی نمونه انتخاب می کنیم. با توجه به اینکه هر ثانیه ۱۰۰ عنصر دنباله‌ای[۵۰] را برای LSTM فراهم می کند، آموزش شامل حدود ۲۰۰۰ مرحله بهینه سازی است. برای ارزیابی سیستم VAD، منحنی مشخصه عملکرد گیرنده[۵۱] (ROC) را محاسبه می کنیم.

به عنوان مثال، ما آستانه اطمینان[۵۲] سیستم را تغییر می دهیم، که برای آن یک فریم به عنوان یک تشخیص، شناسایی می شود تا رابطه بین نرخ مثبت واقعی[۵۳] (TPR) و نرخ مثبت کاذب[۵۴] (FPR) را به تصویر بکشد. برای استنتاج، یک آستانه اطمینان را انتخاب می کنیم که با نرخ خطای برابر[۵۵] (EER) مطابقت دارد. به عنوان مثال، مقادیر مساوی FPR و TPR -1، که توسط تقاطع منحنی ROC و خط دوبخشی[۵۶] TPR + FPR = 1 مشاهده می شود. سپس سیستم VAD به عنوان یک مرحله پیش‌پردازش برای وظیفه SER استفاده می شود، به طوری که اگر حداقل ۲۵٪ از فریم های موجود در ۱ ثانیه بالاتر از آستانه اطمینان EER باشد، هر ثانیه از صدا به عنوان حاوی فعالیت صوتی طبقه بندی می شود.

۳.۲. تشخیص احساسات گفتار

برای کار SER ما از تکه‌های ۱ ثانیه‌ای صوتی استخراج‌شده با سیستم VAD استفاده می‌کنیم تا یک مقدار ظرفیت (و به ترتیب برانگیختگی) با ارزش پیوسته را در هر قطعه صوتی پیش‌بینی کنیم.

بنابراین سیستم VAD اعمال شده بر وظیفه SER، با انتخاب تکه های صوتی که با این فرضیه هدایت می شود که، صوتی با صداهای کودک حاوی بیشترین اطلاعات در مورد حالات عاطفی درک شده کودکان است و بنابراین منجر به عملکرد بالاتر در کار SER می شود، تأثیر می گذارد.

متعاقباً، ما ۸۸ ویژگی کاربردی را برای هر قطعه صوتی ۱ ثانیه ای بر اساس مجموعه پارامترهای صوتی حداقلی ژنو[۵۷] (eGeMAPS)، یک انتخاب جامع ویژگی های صوتی مبتنی بر متخصص استخراج می کنیم[۲۲]. دنباله ای از ویژگی های حاصل از یک جلسه، سپس به عنوان ورودی به مدل یادگیری عمیق ما شامل دو لایه RNN با سلول های LSTM و اندازه لایه پنهان ۱۲۸ واحد استفاده می شود. به دنبال آن یک لایه متراکم با ۱۲۸ نورون، یک واحد خطی اصلاح شده[۵۸] (ReLU) فعال سازی و نرخ انصراف[۵۹] ۰.۳.

یک لایه متراکم نهایی با یک نورون واحد، پیش‌بینی ظرفیت یا برانگیختگی کار ما را ارائه می‌کند. معماری شبکه یکسان به ترتیب برای ظرفیت و برانگیختگی به طور مستقل آموزش داده شده است. با روش شناسی خود، ما[۲۳] را با یک معماری مدل تنظیم شده بر اساس آزمایش های اولیه دنبال می کنیم.

مدل‌های SER برای ۱۸۰ دوره با بهینه‌سازی کامل دسته‌ای[۶۰] آموزش داده می‌شوند – هر جلسه یک دنباله تولید می‌کند – با استفاده از بهینه‌ساز Adam با نرخ یادگیری ۰.۰۰۰۱ و تابع ضرر MSE. تعداد بسیار بیشتری از دوره‌ها در مقایسه با آزمایش‌های VAD بر اساس بهینه‌سازی دسته‌ای کامل انتخاب می‌شوند، یعنی در هر دوره فقط یک مرحله بهینه‌سازی انجام می‌شود.

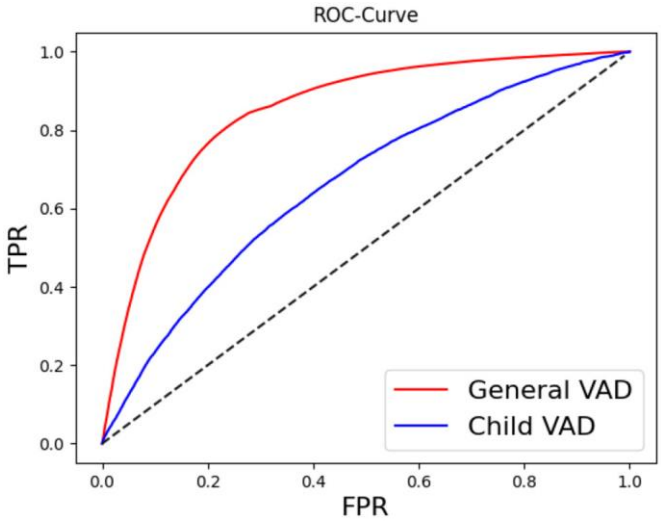
٤. آزمایشات

همه آزمایش‌ها در پایتون[۶۱] [۲۴]، و همچنین تنسورفلو[۶۲] [۲۵] برای مدل‌های یادگیری عمیق و آموزش پیاده‌سازی شده‌اند.

٤.١. تشخیص فعالیت صوتی

برای آزمایش‌های VAD خود، معماری را همانطور که در بخش ۳.۱ توضیح داده شد با دو هدف مختلف آموزش می‌دهیم:

(i) تشخیص فقط صداهای کودک، از جمله همپوشانی با صداهای دیگر و (ii) تشخیص هر گونه آواز، از جمله آوازهای همپوشانی. این دو رویکرد بر روی وظایف مربوطه ارزیابی می شوند. از این رو، هدف ما ارزیابی امکان‌سنجی آموزش یک سیستم VAD عمومی برای ویژگی‌ها و محدودیت‌های مجموعه داده‌های ما و بررسی بیشتر کار احتمالاً چالش‌برانگیزتر آموزش یک سیستم VAD تخصصی برای کودکان مبتلا به اوتیسم است. علاوه بر ارزیابی سیستم‌های VAD بر اساس عملکرد خام[۶۳] آن‌ها، ما تأثیر آن‌ها را بر وظیفه SER در بخش زیر ارزیابی می‌کنیم. ما منحنی‌های ROC را برای هر دو سیستم VAD کودک و سیستم VAD کلی در مورد وظایف مربوطه در شکل ۲ و همچنین EER و ناحیه زیر منحنی (AUC) در جدول ۲ گزارش می‌کنیم.

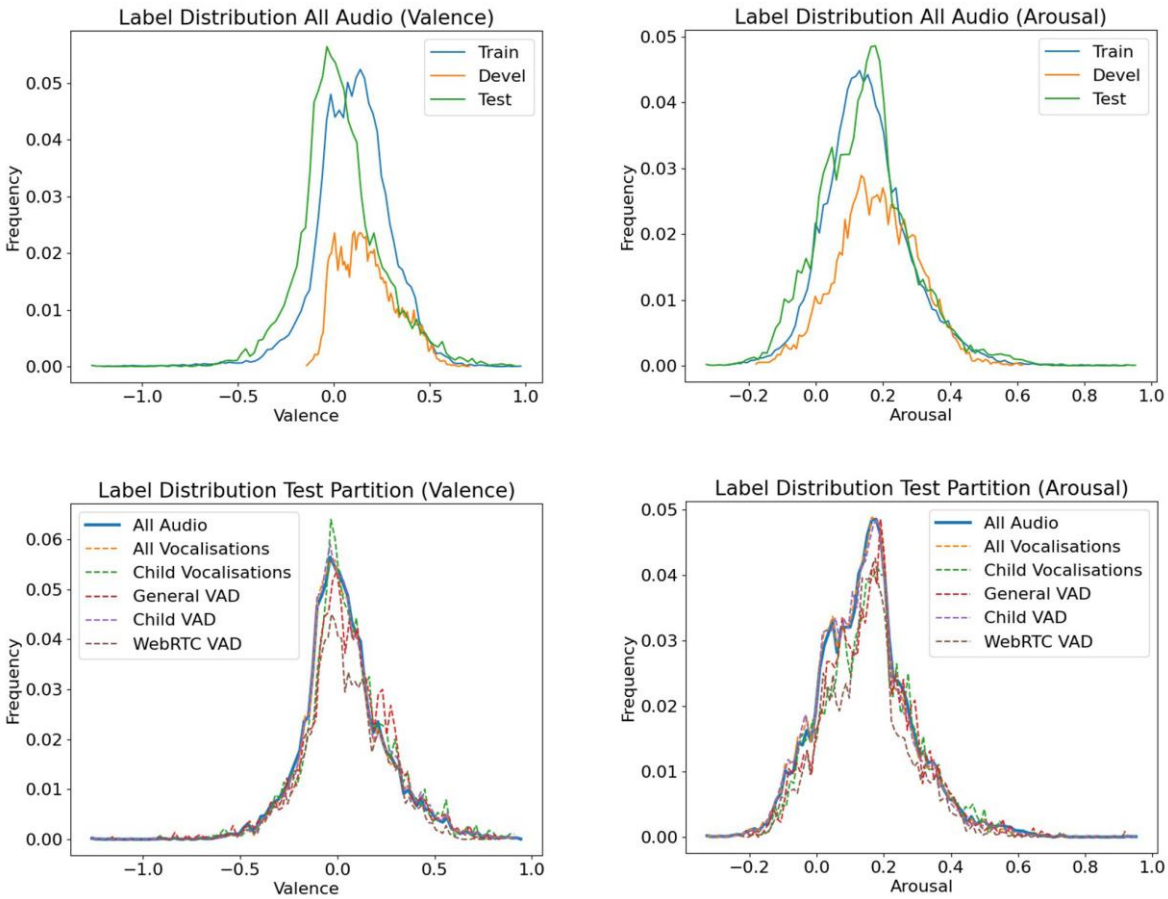


شکل ۲. مشخصه عملکرد گیرنده (ROC)

منحنی تشخیص فعالیت صوتی که برای صداهای کودک به طور خاص و برای همه صداها آموزش داده شده است.

جدول ۲. نرخ خطای برابر (EERs) و ناحیه زیر منحنی (AUC) برای سیستم تشخیص فعالیت صدای کودک و سیستم تشخیص فعالیت صوتی عمومی که در کار مربوطه ارزیابی شده است.

VAD System	EER	AUC
Child VAD	0.381	0.662
General VAD	0.215	0.850



شکل ۳. توزیع برچسب های ظرفیت (سمت چپ) و برچسب های برانگیختگی (در سراسر) با در نظر گرفتن تمام داده های صوتی بدون پیش‌پردازش **VAD** در پارتیشن های مختلف (بالا)، و همچنین توزیع های تنظیم شده پارتیشن آزمایشی پس از پیش‌پردازش از طریق سیستم های مختلف **VAD** و حاشیه نویسی صدا (پایین).

۲.٤. تشخیص احساسات گفتار

همانطور که در بخش ۳ توضیح داده شد، ما از سیستم VAD کودک خود و سیستم کلی VAD آموزش داده شده در بخش قبل استفاده می کنیم تا اگر ۲۵ فریم از ۱۰۰ فریم در یک ثانیه دارای اطمینان پیش بینی[۶۴]، بالاتر از آستانه EER باشند، تکه های ۱ ثانیه را از ضبط های جلسه استخراج کنیم. اگر WebRTC VAD یک فعالیت صوتی را برای حداقل ۰.۲۵ ثانیه از صدا پیش بینی کند، به روشی مشابه، تکه های ۱ ثانیه استخراج می شوند. به همین ترتیب، ما از حاشیه نویسی های حقیقت پایه صداهای کودک و همچنین حاشیه نویسی های حقیقت پایه همه گویندگان برای تقلید از یک VAD کودک کامل و یک سیستم VAD کلی کامل استفاده می کنیم. به عنوان پایه، ما از صدا بدون هیچ گونه پیش‌پردازش مبتنی بر VAD (همه صدا) استفاده می کنیم. شکل ۳ توزیع[۶۵] مقادیر ظرفیت و برانگیختگی را در بین پارتیشن ها و همچنین توزیع تنظیم شده پارتیشن آزمایشی[۶۶] پس از فیلتر کردن، از طریق سیستم های VAD و حاشیه نویسی صدا را نشان می دهد.

برای ارزیابی، ما از ریشه میانگین مربعات خطا[۶۷] (RMSE) و همچنین ضریب همبستگی تطابق[۶۹] (CCC) بر اساس لین[۶۸] (۱۹۸۹) [۲۶] استفاده می کنیم که بین دو توزیع x و y به صورت تعریف شده است.

$$\mathrm{CCC}(x,y)=\frac{\rho(x,y)\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}, \qquad (1)$$

با ضریب همبستگی[۷۰] ρ، و همچنین میانگین μ و انحراف استاندارد[۷۱] σ توزیع مربوطه.

از آنجایی که CCC به عنوان معیاری برای دنباله ها طراحی شده است و دارای ضعف ذاتی برای دنباله ها و دنباله های کوتاه با تغییرات اندک است، ما هنگام محاسبه CCC همه پیش بینی ها و برچسب ها را از یک پارتیشن داده به دو دنباله مربوطه ترکیب می کنیم. نتایج برای ظرفیت و برانگیختگی در جدول ۳ خلاصه شده است.

جدول ۳. نتایج کار تشخیص احساسات گفتار (SER).

VAD System	# Samples detected	Valence (CCC/RMSE)		Arousal (CCC/RMSE)	
		Dev	Test	Dev	Test
	17,944	0.200/0.201	0.021/0.245	0.201/0.121	0.168/0.138
	40,013	0.012/0.160	0.117/0.260	0.100/0.120	0.154/0.142
	29,918	0.140/0.183	0.063/0.224	0.263/0.107	0.098/0.152
	10,961	0.153/0.169	0.085/0.277	0.182/0.115	0.145/0.143
	47,184	-0.032/0.160	0.120/0.231	0.166/0.114	0.105/0.156
	62,370	0.133/0.162	0.024/0.231	0.093/0.122	0.049/0.152

۵. بحث

شکل ۲ و جدول ۲ نشان می دهد که هم یک سیستم فعالیت صوتی عمومی و هم یک سیستم فعالیت صوتی ویژه کودک با عملکرد بالاتر را می توان از داده های موجود آموزش داد. با این حال، سیستم VAD عمومی در مقایسه با سیستم مخصوص کودکان عملکرد آشکارا برتری را نشان می دهد. یکی از دلایل آشکار این امر از خود مجموعه داده ناشی می شود. جدول ۱ نشان می دهد که مجموعه داده بیش از چهار برابر بیشتر حاشیه نویسی را برای VAD عمومی در مقایسه با سیستم VAD کودک ارائه می دهد که منجر به یک وظیفه VAD کودک نامتعادل تر می شود. علاوه بر این، به نظر می‌رسد که وظیفه آموزش یک سیستم VAD تخصصی و متمرکز بر کودکان اوتیستیک عموماً چالش‌برانگیزتر باشد، زیرا این مدل نه تنها نیاز به تشخیص ویژگی‌های معمول گفتار دارد، بلکه باید بین ویژگی‌های گفتاری گویندگان نیز تمایز قائل شود.

یعنی مدل باید الگوهای مشترکی را در صداسازی کودکان با سطوح زبانی مختلف پیدا کند و آن‌ها را از الگوهای موجود در صدای محققان متمایز کند. سطوح مختلف زبانی کودکان درگیر در مطالعه، و نیز شیوه‌های بیان منحصربه‌فرد آنها، به احتمال زیاد کشف ویژگی‌های مشترک را دشوار می‌کرد.

جدول ۳ بیشتر نشان می دهد که تمام سیستم های VAD در نظر گرفته شده دارای حساسیت تا حد زیادی متفاوت هستند. منظور ما از واژه حساسیت در این زمینه، تعداد کل رویدادهای تشخیص صدا مستقل از صحت تشخیص[۷۲] است. حساسیت سیستم کودک VAD، با هدف تشخیص صداهای واقعی کودک، می‌تواند بسیار بالا با تقریباً دو برابر تعداد رویدادهای شناسایی شده در مقایسه با تعداد حاشیه‌نویسی‌های هدف انسانی[۷۳] در نظر گرفته شود.

به نظر می رسد دو سیستم VAD باقیمانده به طور طبیعی بسیار حساس تر از VAD کودک هستند، زیرا هدف آنها فیلتر کردن صداهای کودک نیست. با این حال، هر دو WebRTC VAD خارج از جعبه[۷۴](مبتکرانه) ، و همچنین VAD عمومی آموزش‌دیده ما، هر دو به نظر می‌رسد حساسیت کمتری نسبت به حاشیه‌نویسی حقیقت پایه همه صداهای سخنران نشان می‌دهند و انحراف رویدادهای تشخیص WebRTC به طور قابل‌توجهی بیشتر است.

قسمت بالای شکل ۳ نشان می دهد که تفاوت زیادی بین توزیع برچسب در آموزش و پارتیشن آزمایشی برای وظیفه SER وجود ندارد. با این حال پارتیشن توسعه[۷۵] به طور قابل توجهی منحرف می شود. قسمت پایین شکل ۳ نشان دهنده تفاوت در توزیع برچسب احساسات در مجموعه آزمایشی است که توسط پیش‌پردازش از طریق رویکردهای مختلف VAD ایجاد شده است. اگرچه انتخاب سیستم VAD تنها تأثیر کمی بر توزیع برچسب دارد و بنابراین نباید هیچ مزیت قابل توجهی مرتبط با برچسب به هیچ یک از آزمایش‌های SER حاصل شود، اما همچنان بر قابلیت مقایسه نتایج تأثیر می‌گذارد زیرا داده‌های آزمایشی را تغییر می‌دهد. .

طبق جدول ۳، بهترین نتایج آزمایش برای برانگیختگی در آزمایش‌های SER ما، با پیش پردازش VAD کودک به دست می‌آید، حتی بهتر از پیش‌پردازش بر اساس حاشیه‌نویسی‌های حقیقت پایه. به نظر می رسد این نتایج با این فرضیه مطابقت دارند که تنها در نظر گرفتن صداهای کودک می تواند عملکرد سیستم های SER را برای کودکان اوتیستیک بهبود بخشد و همچنین عملکرد سیستم معقول VAD کودک را پیشنهاد می کند. با این حال، این تجزیه و تحلیل فقط مقدار مشخصی را برای مجموعه توسعه برانگیختگی و حتی کمتر برای آزمایش‌های ظرفیت دارد، که تمایل به دستیابی به عملکرد پایین‌تری در وظایف SER صوتی در مقایسه با آزمایش‌های برانگیختگی دارند. با این وجود، سیستم‌های مبتنی بر VAD در بیشتر آزمایش‌ها از سیستم بدون VAD بهتر عمل می‌کنند، که نشان‌دهنده مزیت آشکار سیستم‌های مبتنی بر VAD برای کار در دست است. محدودیت‌های بیان نتایج مورد بحث در اینجا

باید در نظر گرفته شود، زیرا پیشرفت‌های کوچک همراه با عملکرد کلی پایین مدل‌های SER همیشه در معیارهای ارزیابی بررسی‌شده سازگار نیستند.

کار آینده باید تأثیر یک سیستم VAD ویژه کودک را در رویکرد تشخیص احساسات چند وجهی بررسی کند. با توجه به سناریوهای پیچیده ناشی از جلسات با کودکان اوتیستیک، اجتناب ناپذیر است که همه روش ها همیشه در دسترس نباشند، به عنوان مثال، کودکان از فوکوس دوربین خارج می شوند یا برای مدت طولانی سکوت می کنند. تشخیص و در نظر گرفتن آن روش‌های گمشده، به عنوان مثال در قالب یک سیستم VAD که به ترکیب ویژگی‌های وزنی کمک می‌کند، ممکن است تأثیر قابل‌توجهی بر رفتار مدل داشته باشد و حتی به توضیح تصمیم‌های رویکردهای کاربردی کمک کند.

٦. نتیجه گیری

با این پژوهش، امکان‌سنجی و کاربرد یک سیستم VAD را که به طور خاص، در مورد صداگذاری کودکان اوتیستیک[۷۶] آموزش داده شده بود، برای وظایف SER در جلسات مداخله با کمک ربات[۷۷] برای کودکان اوتیسم، به منظور بهبود موفقیت برنامه[۷۸] برای کودکان مبتلا به اوتیسم مورد بحث قرار دادیم. با توجه به اندازه و همچنین کیفیت پر نویز[۷۹] مجموعه داده، ما نشان دادیم که مولفه فعالیت‌صوتی[۸۰] می تواند با عملکرد معقول آموزش داده شود، در حالی که از یک سیستم VAD عمومی آموزش داده شده یکسان پایین تر است. نتایج ما بیشتر نشان می‌دهد که استفاده از سیستم‌های VAD، و به‌ویژه سیستم‌های VAD کودک، می‌تواند منجر به بهبود جزئی SER مداوم[۸۱] برای کودکان اوتیسم شود. حتی اگر به طور کلی عملکرد پایین در مدل های SER، به احتمال زیاد ناشی از چالش های کار در حال انجام باشد، بیان نتایج را ضعیف می کند. تحقیقات بیشتر بر اساس این کار، استفاده از سیستم‌های VAD کودک را به عنوان مبنایی برای استراتژی‌های داده از دست رفته در وظایف SER چندوجهی بررسی خواهد کرد.

موازین اخلاقی

در این مطالعه اصول اخلاق در پژوهش شامل اخذ رضایت آگاهانه از شرکت کنندگان و حفظ اطلاعات محرمانه آنها رعایت گردیده است.

تشکر و قدردانی

پژوهشگران مراتب قدردانی و تشکر خود را از کلیه شرکت‌کنندگان این پژوهش که با استقبال و بردباری، در روند استخراج نتایج همکاری نمودند، اعلام می‌دارند.

تعارض منافع

نویسندگان این مطالعه هیچ گونه تعارض منافعی در انجام و نگارش آن ندارند.

واژه نامه

۱. تشخیص عواطف گفتار	1. Speech emotion recognition
۲.SER مستمر	2. Continuous SER
۳. سیستم‌های تشخیص فعالیت صوتی	3. Voice activity detection
۴. هارار	4. Harár
۵. القیفاری	5. Alghifari
۶. آکچای و اوگوز	6. Akçay and Oğuz
۷. صداهای اکولالیک	7. echolalic vocalisations
۸. استراتژی‌های برچسب‌گذاری مداوم	8. continuous labelling strategies
۹. حاشیه‌نویسی‌های غنی	9. rich annotations
۱۰. دیاریشن سخن‌گو	10. speaker diarisation
۱۱. ارتباطات بلادرنگ وب	11. Web Real-Time Communication
۱۲. حاشیه‌نویسی دستی دیاریشن سخنگو	12. manual speaker diarisation annotations
۱۳. مدل های مخلوط گاوسی	13. Gaussian mixture models
۱۴. سیستم شش باند فرکانسی	14. log energies of six frequency bands
۱۵.VAD جدا شده	15. isolated VAD
۱۶. حاشیه‌نویسی‌های حواشی سخنران	16. speaker diarisation annotations
۱۷. مبدا برچسب‌ها	17. origin of the labels
۱۸. حاشیه‌نویسی ظرفیت	18. arousal annotations
۱۹. برانگیختگی	19. valence
۲۰. حاشیه نویسی دیاریسیشن سخنران	20. Speaker Diarisation Annotation
۲۱. حاشیه نویسان	21. annotators
۲۲. ابعاد عاطفی ظرفیت و برانگیختگی	22. emotional dimensions valence and arousal
۲۳. جوی‌استیک	23. joystick
۲۴. استاندارد طلایی	24. gold standard sequence
۲۵. ارزیاب	25. evaluator weighted estimator
۲۶. وزن‌های ویژه حاشیه‌نویس	26. annotator-specific weights
۲۷. همبستگی زوجی حاشیه‌نویسی‌ها	27. pairwise correlation of the annotations
۲۸. میانگین دوم	28. second-wise average
۲۹. حاشیه‌نویسی‌های استاندارد طلایی	29. gold standard annotations
۳۰. برچسب احساس	30. emotion label

واژه نامه	
۳۱. شبکه‌های عصبی مکرر	31. recurrent neural networks
۳۲. سلول‌های حافظه کوتاه‌مدت	32. long short-term memory
۳۳. صداسازی حاضر	33. vocalisation present
۳۴. برانگیختگی ابعاد عاطفی	34. affective dimensions
۳۵. داده های کم عمق	35. information-shallow data
۳۶. تجاوزکاری	36. aggressiveness score
۳۷. حقیقت‌پایه(عینی)	37. ground truth
۳۸. مدت زمان بالقوه کوتاه	38. potentially short duration
۳۹. صداسازی	39. vocalisations
۴۰. اندازه پرش	40. hop size
۴۱. دو لایه دو جهته	41. two-layer bi-directional
۴۲. لایه مخفی	42. hidden layer
۴۳ لایه متراکم	43. dense layer
۴۴. نورون خروجی واحد	44. single output neuron
۴۵. هاگرر	45. Hagerer
۴۶. دوره	46. epochs
۴۷. اندازه دسته‌ای	47. batch size
۴۸. تابع ضرر	48. loss
۴۹. میانگین مربعات خطا	49. mean square error
۵۰. عنصر دنباله‌ای	50. sequence elements
۵۱. مشخصه عملکرد گیرنده	51. receiver operating characteristic
۵۲. آستانه اطمینان	52. confidence threshold
۵۳. نرخ مثبت واقعی	53. true positive rate
۵۴. نرخ مثبت کاذب	54. false positive rate
۵۵. نرخ خطای برابر	55. equal-error-rate
۵۶. خط دوبخشی	56. bisectional line
۵۷. پارامترهای صوتی حداقلی ژنو	57. Minimalistic Acoustic Parameter Set
۵۸. یک واحد خطی اصلاح شده	58. rectified linear unit
۵۹. نرخ انصراف	59. dropout rate
۶۰. بهینه‌سازی کامل دسته‌ای	60. full batch optimisation
۶۱. پایتون	61. Python
۶۲. تنسورفلو	62. Tensorflow
۶۳. عملکرد خام	63. raw performance
۶۴. اطمینان پیش بینی	64. prediction confidence
۶۵. توزیع	65. distribution
۶۶. پارتیشن آزمایشی	66. test partition's adjusted distribution
۶۷. میانگین مربعات خطا	67. root mean squared error
۶۸. لین	68. Lin
۶۹. ضریب همبستگی تطابق	69. concordance correlation coefficient
۷۰. ضریب همبستگی	70. correlation coefficient
۷۱. انحراف استاندارد	71. standard deviation
۷۲. مستقل از صحت تشخیص	72. correctness of the detections
۷۳. حاشیه‌نویسی‌های هدف انسانی	73. human target annotations

واژه نامه	
۷۴. خارج از جعبه	74. out-of-the-box
۷۵. پارتیشن توسعه	75. development partition
۷۶. صداگذاری کودکان اوتیستیک	76. autistic child vocalisations
۷۷. مداخله با کمک ربات	77. obot-assisted intervention
۷۸. بهبود موفقیت برنامه	78. improve programme success
۷۹. کیفیت پر نویز	79. noise-heavy
۸۰. فعالیت‌صوتی	80. voice activity
۸۱. SER مداوم	81. continuous SER

فهرست منابع

[1] Harár P, Burget R, Dutta MK. "Speech emotion recognition with deep learning," in 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN) (Noida: IEEE), 2017; 137-140.

[2] Alghifari MF, Gunawan TS, Qadri SA, Kartiwi M, Janin Z. On the use of voice activity detection in speech emotion recognition. Bull. Elect. Eng. Inf. 2019;8: 1324-1332. doi: 10.11591/eei.v8i4.1646

[3] Akçay MB, Oğuz K. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. 2020; 116: 56-76. doi: 10.1016/j.specom.2019.12.001

[4] Baird A, Amiriparian S, Cummins N, Alcorn AM, Batliner A, Pugachevskiy S, et al. "Automatic Classification of autistic child vocalisations: a novel database and results," in Proceedings of the Interspeech 2017 (Stockholm), 2017; 849-853.

[5] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5), Arlington, VA: APA. 2013.

[6] Kopp S, Beckung E, Gillberg C. Developmental coordination disorder and other motor control problems in girls with autism spectrum disorder and/or attention-deficit/hyperactivity disorder. Res. Develop. Disabil. 2010; 31: 350-361. doi: 10.1016/j.ridd.2009.09.017

[7] Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J. Autism spectrum disorder. Lancet 2018; 392: 508-520. doi: 10.1016/S0140-6736(18)31129-2

[8] Hudson CC, Hall L, Harkness KL. Prevalence of depressive disorders in individuals with autism spectrum disorder:

A meta-analysis. J. Abnormal Child Psychol. 2019; 47: 165-175. doi: 10.1007/s10802-018-0402-1

[9] Zaboski BA, Storch EA. Comorbid autism spectrum disorder and anxiety disorders: a brief review. Future Neurol. 2018; 13: 31-37. doi: 10.2217/fnl-2017-0030

[10] Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, et al. "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in Proceedings Interspeech 2013, 14th Annual Conference of the International Speech Communication Association (Lyon). 2013.

[11] Nahar R, Kai A. "Effect of data augmentation on dnn-based vad for automatic speech recognition in noisy environment," in 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE) Kobe, 2020; 368-372.

[12] Amiriparian S, Baird A, Julka S, Alcorn A, Ottl S, Petrović S, et al. "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," in Proceedings of the Interspeech 2018 (Hyderabad), 2018; 2334-2338

[13] Schadenberg BR, Reidsma D, Evers V, Davison DP, Li JJ, Heylen DK, et al. Predictable robots for autistic children-variance in robot behaviour, idiosyncrasies in autistic children's characteristics, and child–robot engagement. ACM Trans. Comput. Human Interact. 2021; 28: 1-42. doi: 10.1145/3468849

[14] Schuller BW. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 61, 2018; 90-99. doi: 10.1145/3129340

[15] Shen J, Ainger E, Alcorn A, Dimitrijevic SB, Baird A, Chevalier P, et al. Autism data goes big: A publicly-accessible multi-modal database of child interactions for behavioural and machine learning research. In International Society for Autism Research Annual Meeting (Kansas City, MO). 2018.

[16] Howlin P, Baron-Cohen S, Hadwin J. Teaching Children With Autism to Mind-Read: A Practical Guide for Teachers and Parents. Chichester: J. Wiley & Sons Chichester. 1999.

[17] Ringeval F, Schuller B, Valstar M, Cummins N, Cowie R, Tavabi L, et al. "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop AVEC '19 (New York, NY: Association for Computing Machinery), 2019; 3-12.

[18] Ringeval F, Schuller B, Valstar M, Gratch J, Cowie R, Scherer S, et al. "Avec 2017: real-life depression, and affect recognition workshop and challenge," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17 (New York, NY: Association for Computing Machinery), 2017; 3-9.

[19] Salishev S, Barabanov A, Kocharov D, Skrelin P, Moiseev M. "Voice activity detector (vad) based on long-term mel frequency band features," in Text, Speech, and Dialogue, eds P. Sojka, A. Horák, I. Kopeček, and Pala, K. (Cham: Springer International Publishing), 2016; 352-358.

[20] Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans. Affect. Comput. 2015; 7: 190-202. doi: 10.1109/TAFFC.2015.2457417

[21] Hagerer G, Pandit V, Eyben F, Schuller B. "Enhancing lstm rnn-based speech overlap detection by artificially mixed data," in Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio, Erlangen. 2017.

[22] Eyben F, Wöllmer M, Schuller B. "Opensmile: the munich versatile and fast open-source audio feature extractor," in MM '10 (New York, NY: Association for Computing Machinery), 2010; 1459-1462.

[23] Stappen L, Baird A, Rizos G, Tzirakis P, Du X, Hafner F, et al. "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild," in Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop MuSe'20, 2020; 35-44.

[24] Van Rossum G, Drake FL. Python 3 Reference Manual. (Scotts Valley, CA: CreateSpace). 2009.

[25] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale Machine Learning on Heterogeneous Systems. Available online at: https://www.tensorflow.org/ (accessed December, 2015; 13: 2021).

[26] Lin L. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45: 255-268. doi: 10.2307/2532051