



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

© 2019 Journal of New Approach to Children's Education (JNACE)



This work is published under CC BY-NC 4.0 license.

© 2022 The Authors.

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.





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 پذیرش مقاله: ۱۴۰۳/۰۲/۲۸

تاریخ دریافت مقاله: ۱۴۰۳/۰۱/۱۸

چکیده

زمینه و هدف: طیف اوتیسم، اختلالی نورولوژیک است که خودش را در سال‌های اولیه رشد کودک نمایان می‌کند. افراد مبتلا به اوتیسم با چالش‌هایی در زمینه تنظیم احساسات مواجه هستند و حالات عاطفی خود را به روش‌های مختلف بیان می‌کنند. 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 سلول‌های حافظه کوتاه مدت (LSTM) است. داده‌ها شامل ۲۵ کودک اوتیستیک انگلیسی زبان است که یک فعالیت آموزشی ساختار یافته را انجام می‌دهند. یافته‌ها: 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 مینا 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

تمامی حقوق نشر برای فصلنامه رویکردی نو بر آموزش کودکان محفوظ است.

شبهه 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

احساسات، یک حوزه تحقیقاتی نسبتاً نوپا است. به طور کلی، این فناوری در صورتی که از روش‌های چندوجهی در پردازش داده‌ها استفاده کند، بسیار کارآمد خواهد بود. تشخیص عواطف گفتاری^۱ (SER) یکی از زیرشاخه‌های برجسته

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

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 می‌دانیم، تاکنون هیچ تحقیقی به بررسی این موضوع پرداخته است که چگونه VAD می‌تواند چنین مدل‌سازی را بهبود بخشد.

در این مقاله، ما زیرمجموعه‌ای از داده‌های جمع‌آوری شده در پروژه DE-ENIGMA را بررسی می‌کنیم [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 میلمان) 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 پیاده‌سازی می‌کنیم که به‌طور خاص برای صداگذاری کودکان اوتیستیک در مجموعه داده‌ها 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 دست‌آزمایی است. WebRTC VAD بر اساس مدل‌های مخلوط گاوسی^{۱۳} (GMMs) و انرژی‌های 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 می‌دهیم. 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 محدودیت‌های رویکردهای خود را در بخش ۵ مورد بحث قرار می‌دهیم.

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، القیاری^۵ و همکاران [۲] در سال ۲۰۱۹ و آکچای و اوگوز^۳ [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 فعالیت‌ها 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-بیش‌فعالی، اختلال هماهنگی رشد یا اختلالات افسردگی [۶-۹].

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 کرده‌اند که چگونه فناوری اخیر از جمله هوش مصنوعی می‌تواند 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 بینش‌هایی در مورد قابلیت پیش‌بینی ربات در سناریوهای RCI با کودکان مبتلا به اوتیسم و همچنین پیش‌بینی شدت صفات مربوط به اوتیسم [۴] و تشخیص صداهای اکولالیک^۶ [۱۲] (به عنوان مثال، تکرار کلمات کودکان اوتیستیک بر اساس گفته‌های گفتاری طرفین مکالمه آنها) کمک کرد [۱۳].

شولر و همکاران وظیفه‌ای 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

۲. مجموعه داده

آزمایش‌های این دست‌نوشته بر اساس زیرمجموعه‌ای از داده‌های جمع‌آوری شده 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 می‌شد، تقسیم شدند. Zeno قادر به انجام حالات مختلف چهره مرتبط با احساسات است که توسط محقق از طریق یک رابط خارجی کنترل می‌شود. جلسات با چندین دوربین و میکروفون که زوایای مختلف اتاق را پوشش می‌دادند ضبط می‌شد.

هر کودک بین یک تا پنج جلسه روزانه (به طور متوسط ۳.۴)

جدول ۱: مروری بر سه پارتنیشن 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

تصمیمات خود را بر اساس ترکیبی از جریان‌های ویدئویی موجود به همراه یکی از ضبط‌های صوتی بومی دوربین‌های ویدئویی و همچنین به تصویر کشیدن شکل موج صوتی خام، مبتنی کنند. ابزار حاشیه‌نویسی به حاشیه‌نویسان اجازه می‌داد به نقاط دلخواه ضبط بپردازند. به طور کلی، هر جلسه توسط یک حاشیه‌نویس ارزیابی شد.

۲.۲. حاشیه‌نویسی احساسات

حاشیه‌نویسی 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 می‌رسد. 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

جلسات از نظر داده‌های صوتی و تصویری، به دنبال یک پروتکل حاشیه‌نویسی (حاشیه‌نویسی) از پیش تعریف شده، بسیار حاشیه‌نویسی شدند. از جمله دستورالعمل‌هایی برای حواشی گوینده، نوع آوازسازی، 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. حاشیه‌نویسی دیاریسیشن سخنران^{۲۰}

دیاریسیشن سخنران یا Speaker Diarisation توسط انگلیسی‌زبانان مسلط با استفاده از ابزار حاشیه‌نویسی 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. حاشیه‌نویسان^{۲۱} می‌توانستند

آزمایش‌های خود، با محاسبه میانگین دوم^{۲۸} بر روی حاشیه‌نویسی‌های استاندارد طلایی^{۲۹}، تنها از یک برچسب احساس^{۳۰} در هر ثانیه استفاده می‌کنیم.

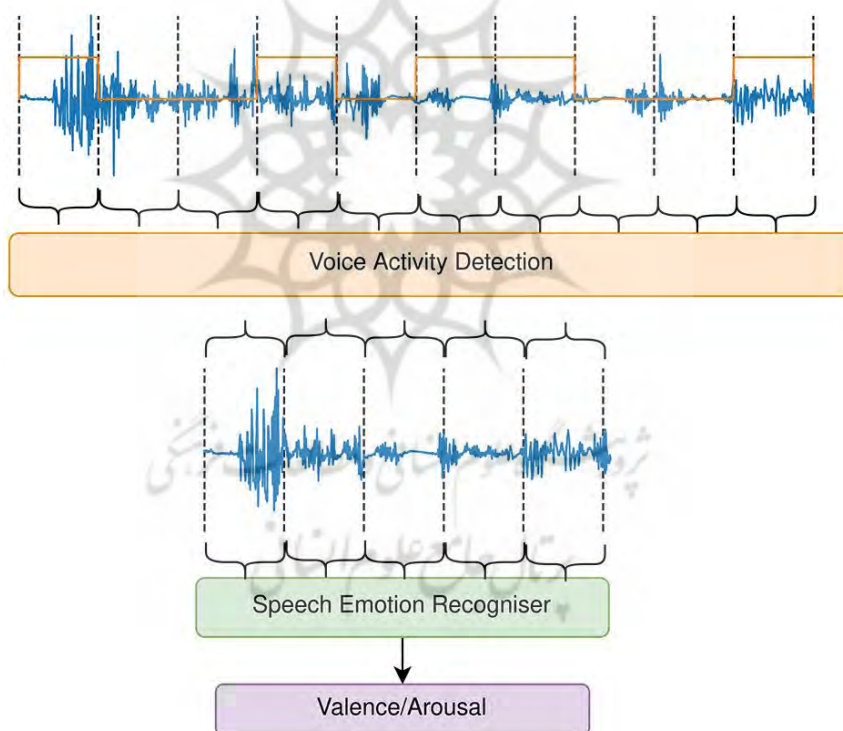
۳. روش پژوهش

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 تکه‌های صوتی طولانی ۱ ثانیه ارائه می‌شود و هدف آن برچسب زدن بخش‌هایی 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.

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

برای فرآیند حاشیه‌نویسی، به ارزیاب‌ها جوی‌استیک^{۳۳} (مدل Logitech Extreme 3D Pro) داده شد تا ظرفیت و برانگیختگی را به طور جداگانه حاشیه‌نویسی کنند. در حالی که حاشیه‌نویسان در حال تماشای ضبط جلسات بودند، موقعیت جوی‌استیک را تغییر دادند که به طور مداوم با نرخ نمونه برداری ۵۰ هرتز نمونه برداری می‌شد و درجه و علامت ظرفیت تخمینی یا مقادیر برانگیختگی را نشان می‌داد (مثبت در موقعیت بالا، منفی در یک موقعیت پایین).

حاشیه‌نویسی‌های حاشیه‌نویس‌های مختلف برای هر جلسه در یک دنباله استاندارد طلایی^{۳۴} با استفاده از استاندارد طلایی تخمین‌گر وزنی ارزیاب^{۳۵} (EWE) [۱۴] خلاصه می‌شود. استاندارد طلایی EWE معمولاً در وظایف تشخیص احساسات استفاده می‌شود [۱۷، ۱۸] و وزن‌های ویژه حاشیه‌نویس^{۳۶} را بسته به همبستگی زوجی حاشیه‌نویسی‌ها^{۳۷} در نظر می‌گیرد. برای



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

۳.۱. تشخیص فعالیت صوتی

از آنجایی که هدف سیستم VAD حذف هر چه بیشتر داده‌های کم عمق^{۳۵} از داده‌های صوتی است، ما چندین روش را در اینجا با هم مقایسه می‌کنیم:

در سطح اول، ما سعی می‌کنیم تمام صداها را با سیستم‌های

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 ثانیه‌ای باقیمانده به سیستم SER داده می‌شود تا مقادیر پیوسته برای ظرفیت و برانگیختگی را پیش‌بینی کند.

کنیم. با توجه به اینکه هر ثانیه ۱۰۰ عنصر دنباله‌ای^{۵۰} را برای LSTM فراهم می‌کند، آموزش شامل حدود ۲۰۰۰ مرحله بهینه سازی است. برای ارزیابی سیستم VAD، منحنی مشخصه عملکرد گیرنده^{۵۱} (ROC) را محاسبه می‌کنیم. به عنوان مثال، ما آستانه اطمینان^{۵۲} سیستم را تغییر می‌دهیم، که برای آن یک فریم به عنوان یک تشخیص، شناسایی می‌شود 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 می‌کنیم 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 می‌شود. سپس سیستم VAD به عنوان یک مرحله پیش‌پردازش برای وظیفه SER استفاده می‌شود، به طوری که اگر حداقل ۲۵٪ از فریم‌های موجود در ۱ ثانیه بالاتر از آستانه اطمینان EER باشد، هر ثانیه از صدا به عنوان حاوی فعالیت صوتی طبقه بندی می‌شود.

۳.۲. تشخیص احساسات گفتار

برای کار SER ما از تکه‌های ۱ ثانیه‌ای صوتی استخراج شده 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 تکه‌های صوتی که با این فرضیه هدایت می‌شود 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 می‌شود، تأثیر می‌گذارد.

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 ویژگی‌های صوتی مبتنی بر متخصص استخراج می‌کنیم [۲۲]. دنباله ای از ویژگی‌های حاصل از یک جلسه، سپس به عنوان ورودی به مدل یادگیری عمیق ما شامل دو لایه RNN با سلول‌های LSTM و اندازه لایه پنهان ۱۲۸ واحد استفاده می‌شود. 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 آزمایش‌های اولیه دنبال می‌کنیم.

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 سیستم‌های VAD استفاده می‌شود، به عنوان مثال [۱۱].

امتیاز پرخاشگری (تجاوز کاری)^{۳۶} WebRTC VAD برابر با یک تعیین شده است. علاوه بر این، ما از حاشیه نویسی‌های حقیقت‌پایه (عینی)^{۳۷} برای همه صداها به عنوان یک استاندارد طلایی برای یک سیستم کلی VAD استفاده می‌کنیم.

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 می‌توان 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 نوزد 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 هاگر^{۴۵} [۲۱] است، اما بر اساس آزمایش‌های اولیه تنظیم شده است. ما از طول توالی (دنباله) ثابت ۱۰۰ نمونه در طول زمان آموزش استفاده می‌کنیم، به عنوان مثال، جریان صوتی به نمونه‌هایی با طول ۱ ثانیه بریده می‌شود. در حین آموزش این مساله رگرسیون، به هر فریم در صورت وجود گفتار برچسب ۱ و اگر وجود ندارد برچسب ۰ اختصاص داده می‌شود.

مدل‌های VAD برای ۸ دوره^{۴۶} با اندازه دسته‌ای^{۴۷} ۲۵۶ با استفاده از بهینه‌ساز Adam با نرخ یادگیری ۰.۰۱ و تابع ضرر^{۴۸} میانگین مربعات خطا^{۴۹} (MSE) داده شده است. ما تعداد نسبتاً کمی از دوره‌ها را بر اساس تعداد زیادی نمونه انتخاب می‌

(i) تشخیص فقط صداهای کودک، از جمله همپوشانی با صداهای دیگر و (ii) تشخیص هر گونه آواز، از جمله آوازه‌های همپوشانی. این دو رویکرد بر روی وظایف مربوطه ارزیابی می‌شوند. 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.

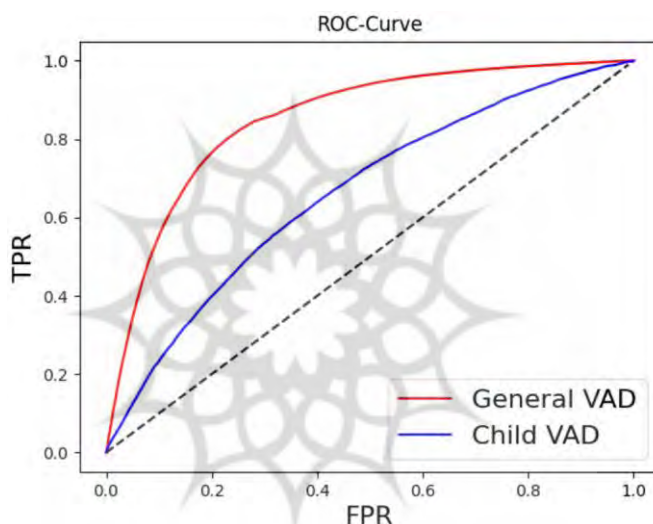
مدل‌های SER برای ۱۸۰ دوره با بهینه‌سازی کامل دسته‌ای^{۶۰} آموزش داده می‌شوند - هر جلسه یک دنباله تولید می‌کند - با استفاده از بهینه‌ساز Adam با نرخ یادگیری ۰.۰۰۰۱ و تابع ضرر MSE. تعداد بسیار بیشتری از دوره‌ها در مقایسه با آزمایش‌های VAD بر اساس بهینه‌سازی دسته‌ای کامل انتخاب می‌شوند، یعنی در هر دوره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

۴. آزمایشات

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

۱.۴. تشخیص فعالیت صوتی

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

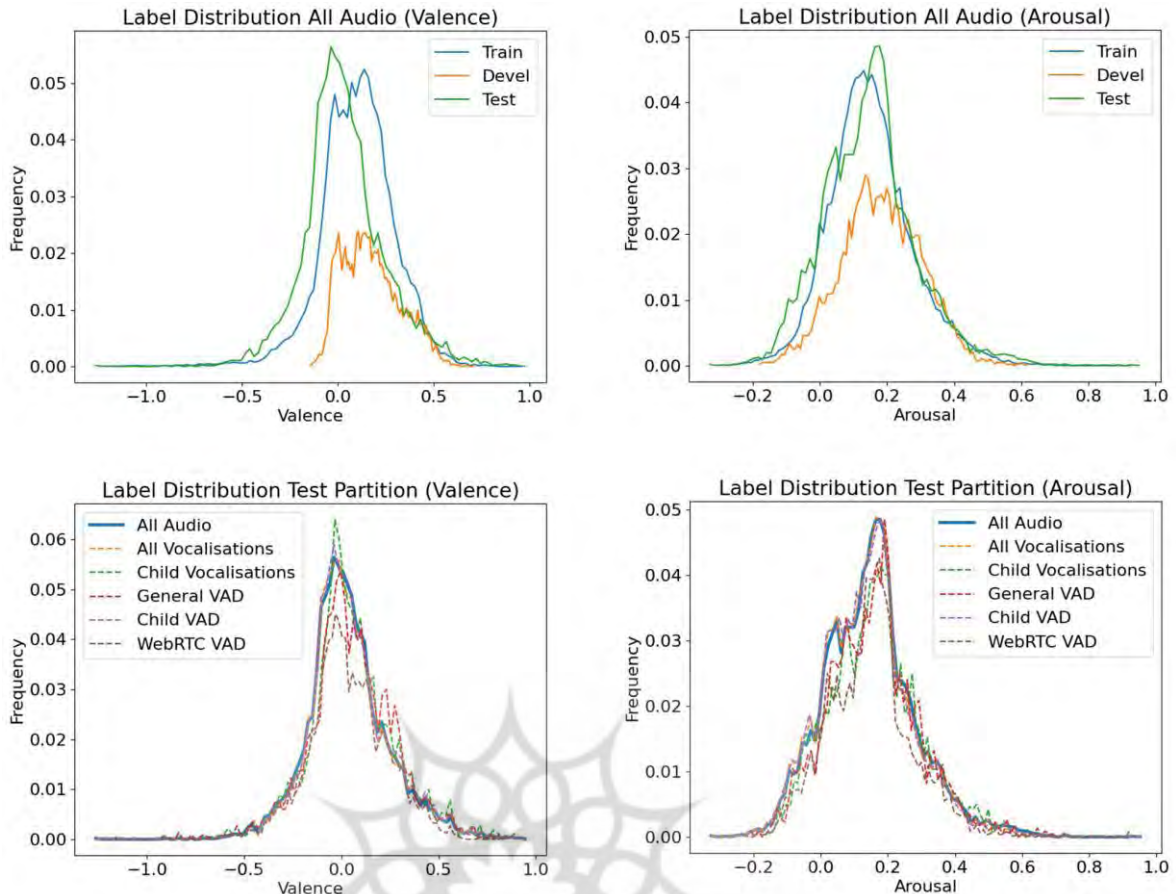


شکل ۲. مشخصه عملکرد گیرنده (ROC)

منحنی تشخیص فعالیت صوتی که برای صداهای کودک به طور خاص و برای همه صداهای 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 برچسب‌های ظرفیت (سمت چپ) و برچسب‌های برانگیختگی (در سراسر) با در نظر گرفتن تمام داده‌های صوتی بدون پیش‌پردازش VAD در پارتیشن‌های مختلف (بالا)، و همچنین توزیع‌های تنظیم شده پارتیشن آزمایشی پس از پیش‌پردازش از طریق سیستم‌های مختلف VAD و حاشیه نویسی صدا (پایین).

۲.۴. 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 می‌کنیم تا اگر ۲۵ فریم از ۱۰۰ فریم در یک ثانیه دارای اطمینان پیش‌بینی^{۶۴}، بالاتر از آستانه EER باشند، تکه‌های ۱ ثانیه را از ضبط‌های جلسه استخراج کنیم. اگر WebRTC VAD یک فعالیت صوتی را برای حداقل ۰.۲۵ ثانیه از صدا پیش‌بینی کند، به روشی مشابه، تکه‌های ۱ ثانیه استخراج می‌شوند. به همین ترتیب، ما از حاشیه نویسی‌های حقیقت پایه صداهای کودک و همچنین حاشیه نویسی‌های حقیقت پایه همه گویندگان برای تقلید از یک VAD کودک کامل و یک سیستم VAD کلی کامل استفاده می‌کنیم. به عنوان پایه، ما از صدا بدون هیچ گونه پیش‌پردازش مبتنی بر VAD (همه صدا) استفاده می‌کنیم. شکل ۳ توزیع^{۶۵} مقادیر ظرفیت و برانگیختگی را در بین پارتیشن‌ها و همچنین توزیع تنظیم شده پارتیشن آزمایشی^{۶۶} پس از فیلتر کردن، از طریق

سیستم‌های VAD و حاشیه نویسی صدا را نشان می‌دهد. برای ارزیابی، ما از ریشه میانگین مربعات خطا^{۶۷} (RMSE) و همچنین ضریب همبستگی تطابق^{۶۹} (CCC) بر اساس لین^{۶۸} (۱۹۸۹) [۲۶] استفاده می‌کنیم که بین دو توزیع x و y به صورت تعریف شده است.

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

با ضریب همبستگی^{۷۰} ρ ، و همچنین میانگین μ و انحراف استاندارد^{۷۱} σ توزیع مربوطه. از آنجایی که CCC به عنوان معیاری برای دنباله‌ها طراحی شده است و دارای ضعف ذاتی برای دنباله‌ها و دنباله‌های کوتاه با تغییرات اندک است، ما هنگام محاسبه CCC همه پیش‌بینی‌ها و برچسب‌ها را از یک پارتیشن داده به دو دنباله مربوطه ترکیب می‌کنیم. 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

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 قابل توجهی بیشتر است.

قسمت بالای شکل ۳ نشان می‌دهد که تفاوت زیادی بین توزیع برجسب در آموزش و پارتیشن آزمایشی برای وظیفه SER وجود ندارد. با این حال پارتیشن توسعه^{۷۵} به طور قابل توجهی منحرف می‌شود. قسمت پایین شکل ۳ نشان دهنده تفاوت در توزیع برجسب 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 برجسب دارد و بنابراین نباید هیچ مزیت قابل توجهی مرتبط با برجسب 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 پیش‌پردازش VAD کودک به دست می‌آید، حتی بهتر از پیش‌پردازش بر اساس حاشیه‌نویسی‌های حقیقت پایه. به نظر می‌رسد این نتایج با این فرضیه مطابقت دارند که تنها در نظر گرفتن صداهای کودک می‌تواند عملکرد سیستم‌های SER را برای کودکان اوتیستیک بهبود بخشد و همچنین عملکرد سیستم معقول VAD کودک را پیشنهاد می‌کند. 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 دست است. محدودیت‌های بیان نتایج مورد بحث در اینجا

شکل ۲ و جدول ۲ نشان می‌دهد که هم یک سیستم فعالیت صوتی عمومی و هم یک سیستم فعالیت صوتی ویژه کودک با عملکرد بالاتر را می‌توان از داده‌های موجود آموزش داد. با این حال، سیستم VAD عمومی در مقایسه با سیستم مخصوص کودکان عملکرد آشکارا برتری را نشان می‌دهد. یکی از دلایل آشکار این امر از خود مجموعه داده ناشی می‌شود. جدول ۱ نشان می‌دهد که مجموعه داده بیش از چهار برابر بیشتر حاشیه نویسی را برای VAD عمومی در مقایسه با سیستم VAD کودک ارائه می‌دهد که منجر به یک وظیفه VAD کودک نامتعادل تر می‌شود. 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 می‌کند.

جدول ۳ بیشتر نشان می‌دهد که تمام سیستم‌های 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^{۷۶} 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^{۷۷} در نظر گرفته شود.

به نظر می‌رسد 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

استخراج نتایج همکاری نمودند، اعلام می‌دارند.

تعارض منافع

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

واژه نامه

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

باید در نظر گرفته شود، زیرا پیشرفت‌های کوچک همراه با عملکرد کلی پایین مدل‌های SER همیشه در معیارهای ارزیابی بررسی شده 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 روش‌ها همیشه در دسترس نباشند، به عنوان مثال، کودکان از فوکوس دوربین خارج می‌شوند یا برای مدت طولانی سکوت می‌کنند. 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.

۶. 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 فعالیت صوتی^{۸۰} می‌تواند 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 مدل‌های SER، به احتمال زیاد ناشی از چالش‌های کار در حال انجام باشد، بیان نتایج را ضعیف می‌کند. 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 شرکت‌کنندگان 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

واژه نامه		واژه نامه
74. out-of-the-box	۷۴. خارج از جعبه	۳۱. recurrent neural networks
75. development partition	۷۵. پارتیشن توسعه	۳۲. long short-term memory
76. autistic child vocalisations	۷۶. صداگذاری کودکان اوتیستیک	۳۳. vocalisation present
77. obot-assisted intervention	۷۷. مداخله با کمک ربات	۳۴. affective dimensions
78. improve programme success	۷۸. بهبود موفقیت برنامه	۳۵. information-shallow data
79. noise-heavy	۷۹. کیفیت پر نویز	۳۶. aggressiveness score
80. voice activity	۸۰. فعالیت صوتی	۳۷. ground truth
81. continuous SER	۸۱. SER مداوم	۳۸. potentially short duration
		۳۹. vocalisations
		۴۰. hop size
		۴۱. two-layer bi-directional
		۴۲. hidden layer
		۴۳. dense layer
		۴۴. single output neuron
		۴۵. Hagerer
		۴۶. epochs
		۴۷. batch size
		۴۸. loss
		۴۹. mean square error
		۵۰. sequence elements
		۵۱. receiver operating characteristic
		۵۲. confidence threshold
		۵۳. true positive rate
		۵۴. false positive rate
		۵۵. equal-error-rate
		۵۶. bisectional line
		۵۷. Minimalistic Acoustic Parameter Set
		۵۸. rectified linear unit
		۵۹. dropout rate
		۶۰. full batch optimisation
		۶۱. Python
		۶۲. Tensorflow
		۶۳. raw performance
		۶۴. prediction confidence
		۶۵. distribution
		۶۶. test partition's adjusted distribution
		۶۷. root mean squared error
		۶۸. Lin
		۶۹. concordance correlation coefficient
		۷۰. correlation coefficient
		۷۱. standard deviation
		۷۲. correctness of the detections
		۷۳. human target annotations

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:

- [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.
- A meta-analysis. *J. Abnormal Child Psychol.* 2019; 47: 165-175. doi: 10.1007/s10802-018-0402-1
- [9] Zaloski 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 multimodal 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.

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

