بهبود کارایی یک سیستم تشخیص احساس از گفتار به کمک شبکه مولد متخاصمی جهت کاربرد در روانشناسی بالینی

آرش شیلاندری[۱]، حسین مروی[۲]، حسین خسروی[۳]

تاریخ دریافت: ۱۴۰۰/۰۹/۲۲ تاریخ پذیرش: ۱۴۰۰/۱۰/۱۲

چکیده

مقدمه: فناوری تشخیص احساس از گفتار، می‌تواند به محققان کمک کند تا دریابند چه عواملی باعث می‌شود برخی از روان‌درمانگران درمان مؤثرتری نسبت به دیگران ارائه دهند، اطلاعاتی که می‌تواند برای بهبود تشخیص روش درمان استفاده شود. اگر بدانیم چه کسی می‌خواهد اقدام به خودکشی کند یا حداقل ریسک بالایی برای این کار دارد می‌توانیم پیشگیری کنیم و این دقیقاً همان چیزی است که علم روانشناسی سال‌هاست به آن نیاز دارد تا هزینه‌های درمان را کاهش دهد. ازاین‌رو، نیاز به تشخیص احساس از گفتار و پایگاه‌داده احساسی به‌شدت احساس می‌شود؛ ولی جمع‌آوری پایگاه‌داده با نمونه‌های زیاد نیازمند صرف چندین دهه است. افزایش داده و انتخاب ویژگی، از مفاهیم کلیدی در یادگیری ماشین هستند.

روش: هنگامی که داده‌های آموزشی در پایگاه‌داده متنوع نیستند و تعداد و تنوع آن‌ها در هر کلاس آموزشی محدود است، آموزش یک شبکه عصبی عمیق بدون آنکه پدیده اور فیتینگ اتفاق بیفتد، بیش از حد چالش‌برانگیز است. برای غلبه بر این چالش، این مقاله یک شبکه افزایش داده جدید، یعنی شبکه افزایش داده متخاصمانه و مبتنی بر شبکه‌های مولد متخاصمی را پیشنهاد می‌کند. این شبکه افزایش داده پیشنهادی، از یک شبکه مولد متخاصمی، یک رمزگذار خودکار و یک طبقه‌بند تشکیل شده است. این شبکه‌ها به طور خصمانه آموزش داده می‌شوند تا بردارهای ویژگی وابسته به هر کلاس را در فضای ویژگی‌ها ترکیب کنند، و سپس آن‌ها را به داده‌های موجود در پایگاه‌داده بیفزایند. برای داده‌های هر کلاس به‌صورت جداگانه یک شبکه مولد متخاصمی پیشنهاد شده است که از یک سو شباهت بین نمونه‌های واقعی و تولید شده را تضمین کند و از طرف دیگر باعث ایجاد تمایز عاطفی در بین نمونه‌های تولید شده بین کلاس‌های مختلف شود. برای رفع مشکل کوچک‌شدن بیش از حد گرادیان در فرآیند آموزش شبکه افزایش داده متخاصمی که منجر به عدم آموزش کافی شبکه‌های مولد و تشخیص‌دهنده و متوقف‌شدن فرایند آموزش پیش از یادگیری کامل توزیع داده‌ها در فضای ویژگی‌ها می‌شود، به‌جای استفاده از خطای متداول آنتروپی متقابل برای آموزش شبکه مولد متخاصمی، واگرایی و اسراستین برای تولید نمونه‌های مصنوعی باکیفیت بالا استفاده شده است.

یافته‌ها: عملکرد این مدل با استفاده از پایگاه‌داده احساسی برلین به‌عنوان مجموعه‌داده‌های آموزش، تست و ارزیابی شبکه مورد آزمایش قرار گرفته و مشخص شد که با ترکیب نمودن بردارهای ویژگی مصنوعی و بردارهای ویژگی واقعی، می‌توان مشکل کوچک‌شدن بیش از حد گرادیان و متعاقباً متوقف‌شدن ادامه روند آموزش شبکه را تا حد زیادی کاهش داد. نتایج به‌دست‌آمده نشان می‌دهد که داده‌های تولید شده توسط شبکه پیشنهادی می‌توانند در یک شبکه تشخیص احساس از سیگنال گفتار استفاده شوند و به این شبکه کمک کنند تا کلاسه‌بندی احساسی بهتری را انجام دهد.

واژه‌های کلیدی: پردازش گفتار، انتخاب ویژگی، افزایش داده، تشخیص احساس از گفتار، شبکه‌های مولد متخاصمی.

Improving the Efficiency of Speech Emotion Recognition System by Generative Adversarial Network in Clinical Psychology

Shilandari, *A., Marvi, H., Khosravi, H.

Abstract

Introduction: In the realm of psychotherapy, the utilization of speech emotion recognition technology holds promise in unraveling the factors that contribute to the varying effectiveness of psychotherapists. This valuable insight can significantly enhance the diagnosis and treatment methods employed. By identifying individuals at a heightened risk of suicide or displaying suicidal tendencies, we can take preventive measures, addressing a long-standing need within the field of psychology and ultimately reducing treatment expenses. Consequently, there is a pressing demand for the recognition of emotions through speech and the development of an extensive emotional database. However, amassing a substantial database with an ample number of samples would traditionally require several decades. To address this challenge, machine learning techniques such as data augmentation and feature selection play a pivotal role.

Methods: This paper introduces an innovative solution to address the challenge of training deep neural networks when the training data lacks diversity and is limited in each class. The proposed approach is an adversarial data augmentation network based on adversarial generative networks. This network consists of an adversarial generator network, an autoencoder, and a classifier. Through adversarial training, these networks combine feature vectors from each class in the feature space and integrate them into the database. Additionally, separate adversarial generative networks are proposed for each class, ensuring similarity between real and generated samples while creating emotional differentiation among different classes. To overcome the problem of excessive gradient reduction, which hinders proper training and halts the learning process before fully understanding the data distribution, the paper suggests using divergence and capture instead of mutual entropy error to generate high-quality synthetic samples.

Results: The model's performance was evaluated on the Berlin Emotional Database, serving as training, testing, and evaluation datasets. Combining artificial and real feature vectors effectively addressed the issue of excessive gradient shrinkage, resulting in a significant reduction in the network training process. The results demonstrated that the generated data from the proposed network can enhance speech signal emotion recognition, leading to improved emotional classification capabilities.

Keywords: Speech Emotion Recognition, Speech Feature Selection, Data Augmentation, Speech Emotion Recognition, Generative Adversarial Networks.

۱. نویسنده مسئول: دانشجوی دوره دکتری دانشکده مهندسی برق، دانشگاه صنعتی شاهرود، شاهرود، ایران **Shilandari@shahroodut.ac.ir**

۲. دانشیار دانشکده مهندسی برق، دانشگاه صنعتی شاهرود، شاهرود، ایران

۳. دانشیار دانشکده مهندسی برق، دانشگاه صنعتی شاهرود، شاهرود، ایران

مقدمه

معرفی: هوش مصنوعی علمی نوپا بوده که در عرصه‌های مختلف علوم راه خود را پیدا کرده است و علم روانشناسی نیز از توسعه بی‌تأثیر نبوده و در فرایند تحولی پیشرفته تلفیق هوش مصنوعی و علم روانشناسی یکی از لازمه‌های تکنولوژی در علوم شناختی است. تحقیقات هوش مصنوعی در روانشناسی مدل سازی و تولید هوشی مانند هوش انسان است. در این پژوهش به تشخیص احساس از سیگنال گفتار پرداخته شده و هدف اصلی این پژوهش ارائه روشی نوین جهت تشخیص احساسات در مدل‌سازی فرآیندهای داده کاوی و فرآیندهای ذهنی است. ادغام علوم شناختی و هوش مصنوعی یک شناخت عمیق از روش‌های شناختی و ارتباطی انسان ارائه می‌دهد. همچنین، در روش‌های خلاقانه و مهارت‌های فنی از راه‌حل‌ها و اپلیکیشن‌های مبتنی بر هوش مصنوعی در روانشناسی مهندسی استفاده می‌کنند. این موضوع به ویژه به ایجاد ارتباط بین علوم شناختی و هوش مصنوعی می‌پردازد که جهت بررسی عملکرد انسان و طراحی روانشناسی مهندسی استفاده می‌شود. ورود هوش مصنوعی به جنبه‌های مختلف زندگی، سبب بروز تحولات زیادی شده است که به صورت روزافزون در حال گسترش آن می‌باشیم. در این خصوص، علوم‌شناختی دانش لازم برای پیشرفت و توسعه فناوری‌های هوش مصنوعی را در اختیار فعالان این حوزه قرار می‌دهد. مطالعه مغز و روش تفکر او در حوزه‌های مختلف هوش مصنوعی مانند آموزش، پزشکی و غیره اهمیت دارد. دستگاه‌ها و ابزارهای پزشکی و مهندسی باید به گونه‌ای طراحی و تولید شوند تا بیشترین سازگاری را با مغز انسان داشته باشند. به عنوان مثال، اتومبیل‌های خودران باید به شکلی ساخته شوند که بتوانند به خوبی با عوامل انسانی تعامل داشته باشند و احساسات او را تشخیص دهند. شرکت‌هایی که در زمینه تولید نرم‌افزارها و تجهیزات خودکار بانکی فعالیت می‌کنند، نیاز به درک بالایی از تشخیص هویت انسان با استفاده از صدای او را دارند تا استفاده از تجهیزات مجهز به هوش مصنوعی که توسط آن‌ها تولید شده است، برای افراد مختلف آسان باشد. در خصوص شبکه‌های اجتماعی نیز باید به تمایلات کاربران و چگونگی احساسات انسانی توجه شود. بنابراین، به طور کلی در هر حوزه‌ای که نیاز به تعامل با انسان وجود داشته باشد، باید درک درستی از حالات روحی و عاطفی او داشته باشید. با توجه به اینکه هوش مصنوعی در رشته‌های مختلف و جنبه‌های گوناگون زندگی ما وارد شده است، استفاده از علوم شناختی بسیار مهم است. برای موفقیت‌آمیز بودن کاربرد هوش مصنوعی در هر صنعت، باید فهمید که چه چیزی برای مغز انسان نتیجه می‌دهد و می‌بایست بر اساس چه استانداردهایی عمل کرد. تشخیص احساسات در این میان از اهمیت بسیار بالایی جهت درک مفاهیم و حالات انسان برخوردار است. تحقیقات هوش مصنوعی شامل بخش‌های مختلفی است که بر مسائل خاص یا رویکردهای خاص تمرکز دارند که برای هر یک از آن‌ها از ابزار ویژه‌ای استفاده می‌شود. برنامه‌ریزی، یادگیری، استدلال، پردازش احساسات و توانایی حرکت و دستکاری اشیا از مسائل اصلی هوش مصنوعی است. رویکردها شامل روش‌های آماری، هوش محاسباتی و یادگیری ماشین است. الگوریتم‌های جستجو، روش‌های بهینه‌سازی و روش های مبتنی بر احتمال برخی از ابزارهای هوش مصنوعی هستند. دانشمندان علوم شناختی کوشش می‌کنند تا راهی برای درک ذهن و چگونگی تعاملات آن با جهان پیرامون پیدا کنند. برای این منظور از روش‌های علمی، شبیه‌سازی و مدل‌سازی استفاده می‌کنند و معمولا نتایج خروجی مدل‌ها را با جنبه‌های شناختی انسان مقایسه می‌کنند. بسیاری از دانشمندان علوم شناختی، دیدگاهی کارکرد گرایانه نسبت به ذهن دارند. به این معنی که حالات و فرآیندهای ذهنی باید توسط درک احساسات در آن‌ها توضیح داده شود. با توجه به این دیدگاه، حتی سیستم‌های غیر انسانی مانند ربات‌ها و رایانه‌ها را نیز می‌توان دارای شناخت دانست در صورتی‌که بتوانند احساسات را تشخیص داده و درک صحیحی از آن‌ها داشته باشند. استفاده از هوش مصنوعی در علوم‌شناختی باعث می‌شود تا درک عمیقی از شناخت انسان و به ویژه احساسات انسانی و ارتباطات ایجاد شود. این رشته بر این پایه ایجاد شده است که هوش انسانی را می‌توان به قدری دقیق توصیف کرد که بتوان ماشینی برای شبیه‌سازی آن تولید کرد. با پیشرفت‌هایی که در زمینه هوش مصنوعی و تولید سامانه‌های هوشمند حاصل شده، می‌توان بین شناخت طبیعی و شناخت مصنوعی پیوند برقرار کرد. به عنوان مثال، رایانه‌هایی تولید شده است که افراد معلول و کم‌توان

می‌توانند از طریق برقراری ارتباط گفتاری آن‌ها را کنترل کنند و افکار موجود در ذهنشان را بدون استفاده از دست و با بیان در صفحه نمایش آن نشان دهند. در این‌گونه کاربردها، ماشین باید بتواند درک صحیحی از احساسات در گفتار انسان داشته باشد. در این زمینه، دستیارهای صوتی مانند الکسا و سیری کمک زیادی می‌کنند ولی هنوز نمی توانند احساسات شما را در دستورات صوتی که به آن‌ها داده می‌شود تشخیص دهند. این مقاله، روشی جهت مدل‌سازی تشخیص احساسات از سیگنال گفتار ارائه می‌دهد که می تواند احساسات را از سیگنال گفتار با دقت بسیار بالایی تشخیص دهد.

کاربردها

یکی از زیرمجموعه‌های هوش مصنوعی که یادگیری ماشین نام دارد و این مقاله به آن پرداخته است، مجموعه الگوریتم‌هایی هستند که می‌توانند از داده‌ها یاد بگیرند و ریسک خطر را مثلاً در حوزه خودکشی تخمین بزنند. فرض کنید بیمارستانی دو سال است که در حال داده جمع‌کردن از مراجعانی است که افکار خودکشی دارند و یا در احساسات آن‌ها احساس غم و اندوه به‌وفور تشخیص داده شده است. اگر شما اطلاعات دیگری مثلاً در مورد سن آن‌ها، میزان مصرف داروهای آن‌ها، طبقه اجتماعی آن و... داشته باشید این امکان وجود دارد که الگوی پنهانی باتوجه‌به متغیرهای جمع‌آوری شده شما وجود داشته باشد که بتواند متغیر وابسته شما یعنی اقدام به خودکشی را پیش‌بینی کند. این تقریباً یکی جدیدترین و نوآورانه‌ترین کارهایی است که در سال‌های اخیر انجام شده است. تعدادی از تحقیقات تخمین‌های خوبی از ریسک خودکشی باتوجه به تشخیص احساسات در گفتار افراد یا مکالمات آن‌ها با دوستان و نزدیکانشان به دست آورده و ارائه داده‌اند(۱).

به دلیل شیوع بالای افسردگی در دنیا و در ایران تحقیقات در این حوزه معمولاً جزو اولویت‌های پژوهشی سازمان‌ها و مراکز تحقیقاتی به‌حساب می‌آیند و به همین خاطر تعداد زیادی تحقیق در این حوزه موجود است که به دنبال یافتن الگوریتم‌ها و ویژگی‌هایی هستند که بهترین تخمین را از ابتلای افراد به افسردگی بدهند. برای بیماری‌های دوقطبی، اختلالات سلوک و ضداجتماعی و اختلالات اضطرابی در این زمینه کارهایی انجام شده است که با پیشرفت روزافزون این حوزه به یکی از کمک‌کننده‌ترین حوزه‌ها تبدیل خواهد شد. همچنین پردازش احساسات در گفتار می‌تواند کمک شایانی به پیشرفت تحقیقات در این زمینه نماید.

در بعضی موارد، افرادی که دچار ناراحتی‌های روانی هستند به روانشناسان مراجعه نمی‌کنند. این می‌تواند به دلیل وجود انگ در این باره باشد و یا می‌تواند به دلیل هزینه‌ها و... باشد که افراد برای گرفتن خدمت سلامت روان مراجعه نمی‌کنند. اما درعین‌حال این افراد امکان دارد در مثلاً شبکه‌های اجتماعی خود در مورد مشکل خود صحبت کنند. این برای روانشناسان خیلی مفید خواهد بود که به بخش بزرگی از دیتاها که جای دیگری امکان دارد باشند توجه کنند و از احساسات فرد در آن‌ها اطلاعات خوبی جهت شیوه مناسب درمان به دست آورند. در این زمینه دسته‌بندی‌هایی که حوزه یادگیری ماشین می‌تواند برای روانشناسان انجام دهد به‌شدت مفید خواهد بود. این که مثلاً در حوزه افسردگی دیتاها دسته‌بندی شوند تا افرادی که در دسته نیاز به کمک فوری قرار می‌گیرند سریعاً شناسایی و غربال شوند می‌تواند کمک بزرگی در این زمینه باشد (۲).

توجه به دیتاست‌ها و جمع‌آوری‌های هدف‌مند در این زمینه یکی کمبودهایی است که در حوزه روانشناسی در ایران داریم. این شاید به این دلیل است که وجود داده‌ها و اهمیت جمع‌آوری داده‌ها هنوز برای روانشناسان در ایران چیزی مهم به‌حساب نمی‌آید و هنوز اهمیت آن درک نشده است. بدون وجود یک سری داده عملاً نمی‌توان در حوزه‌های گوناگون از کاربردهای هوش مصنوعی کمک گرفت. این مقاله به مشکل کمبود تعداد نمونه‌ها در دیتاست احساسات گفتاری پرداخته و پس از معرفی جدیدترین روش افزایش داده‌ها که در مقاله دیگری آن را معرفی نمودیم، به رفع مشکل انتخاب ویژگی‌های کارآمد جهت افزایش داده در دیتاست به‌منظور کاهش محاسبات و استفاده در کاربردهای آنلاین هنگام مراجعه مریض در کلینیک مشاوره پرداخته شده و آن را برطرف نموده است.

روش‌های افزایش داده

کمبود داده[1] به‌عنوان یکی از مهم‌ترین چالش‌ها در سیستم‌های تشخیص احساس از گفتار شناخته می‌شود که عمدتاً در سه جنبه قابل‌بحث و بررسی است (۴۴): ۱. مشکل

1. Data Shortage

اول فقدان بانک‌های اطلاعاتی گفتار عاطفی طبیعت‌گرایانه است. تعداد معدودی از بانک‌های اطلاعاتی با داده‌های گفتاری احساسی طبیعی که از موقعیت‌های واقعی در زندگی جمع‌آوری شده‌اند به دلیل برخی از مسائل قانونی و اخلاقی برای استفاده عمومی وجود دارند و یا حتی می‌توان گفت وجود ندارند (۳). علاوه بر این، گفتارهای احساسی در اکثر پایگاه‌های داده عمومی توسط بازیگران تولید و ضبط می‌شوند و سعی می‌شود در وضعیت احساسی موردنیاز باشند؛ ولی در نهایت بیان عاطفی آن‌ها ممکن است در مقایسه با موقعیت‌های دنیای واقعی متفاوت باشد یا در آن‌ها اغراق وجود داشته باشد. ۲. موضوع مهم بعدی حاشیه‌نویسی است. ازآنجاکه احساسات ابراز شده و احساسات مختلف، متفاوت هستند، حاشیه‌نویسی خارجی همیشه لازم است. منظور از حاشیه‌نویسی ابزاری کمکی است که به‌وسیله آن بتوان احساس گوینده را از گفتار او حدس زد و یا متوجه شد. برای نمایش احساس‌های مختلف معمولاً از مدل گسسته یا پیوستهٔ احساس استفاده می‌شود. در مدل گسسته احساس، از چند برچسب محدود به‌منظور شاخص‌گذاری احساس‌های مختلف استفاده می‌شود. به طور مثال در پایگاه‌دادهٔ احساسی برلین که منطبق بر مدل گسسته احساس است، به هر جمله یک برچسب متناظر با یکی از هفت احساس خشم، خستگی، انزجار، ترس، خوشحالی، عادی و ناراحتی اختصاص‌داده‌شده است. تعداد محدود برچسب‌ها در مدل گسسته باعث بروز مشکلاتی در بیان احساس‌های مختلف می‌شود. به طور مثال وقتی یک جمله با برچسب خشم یا ترس مشخص می‌شود، شدت این احساس‌ها در این برچسب مشخص نیست. علاوه‌برآن چگونگی گذار از این احساس به احساس دیگر در مدل گسسته مشخص نمی‌شود. در این راستا بسیاری از تحقیقات این حوزه بر روی موضوع تشخیص احساس پیوسته متمرکز شده‌اند (۴). ۳. گفتارهای موجود در اکثر پایگاه‌های داده به‌صورت نامتعادل در احساسات مختلف توزیع می‌شوند. به‌طورکلی، تعداد سخنان با احساسات خنثی بیشترین تعداد را در جملات یک گفتار دارد (۳). بااین‌حال، برای ارزیابی و آموزش بهتر یک طبقه‌بند، یک بانک اطلاعاتی متعادل موردنیاز است. علاوه بر این، اگر یک جمله دارای احساسات مختلف باشد، قضاوت انسان در مورد احساسات درک شده می‌تواند صرفاً بر اساس محتوای عاطفی جمله بدون تأثیر محتوای لغوی آن باشد (۳).

برای حل مشکل کمبود داده‌ها، مجموعه آموزشی را می‌توان با تکنیک‌های افزایش داده، افزایش داد. روش‌های سنتی افزایش داده به طور معمول داده‌های اصلی را تغییر داده و دچار مشکلاتی همچون اضافه‌کردن نویز و طنین‌اندازی به سیگنال‌های گفتار و برش، چرخش و غیره در تصاویر می‌نمایند و سپس داده‌های تبدیل شده را به داده‌های اصلی اضافه می‌کنند (۵). روش‌های پیشرفته‌تر افزایش داده مبتنی بر شبکه‌های مولد متخاصمی و انواع آن، شبکه‌های مولد متخاصمی مشروط[1] و یا رمزگذارهای خودکار متخاصمی[2] هستند.

یکی از روش‌های مؤثر برای تقویت و افزایش داده‌ها، استفاده از شبکه‌های مولد متخاصمی است که توسط گودفلو و همکاران در سال ۲۰۱۴ معرفی شده است (۶). در سال‌های اخیر، شبکه‌های مولد متخاصمی به‌عنوان یکی از موفق‌ترین رویکردها برای تولید نمونه شناخته شده‌اند. با استفاده از یک بازی مخالف بین یک شبکه تشخیص‌دهنده و یک شبکه مولد، شبکه‌های مولد متخاصمی آموزش می‌بینند تا نمونه‌هایی تولید کنند که از داده‌های واقعی قابل‌تشخیص نیستند. علاوه بر این، آنها دارای سه خصوصیت عمده هستند (۶): ۱. شبکه‌های مولد متخاصمی می‌توانند توزیع احتمال در مشکلات پیچیده دنیای واقعی را یاد بگیرند. ۲. شبکه‌های مولد متخاصمی می‌توانند با داده‌های ازدست‌رفته (داده‌های نویزی) آموزش داده شوند، جایی که برچسب‌های بسیاری از نمونه‌ها وجود ندارد و ۳. شبکه‌های مولد متخاصمی دارای خروجی‌های چند مدول هستند به این معنی که آن‌ها می‌توانند چندین جواب صحیح متفاوت تولید کنند و تنوع نمونه‌های تولید شده را افزایش دهند.

تاکنون روش‌های مؤثر و متعددی ارائه شده است که با معرفی ویژگی‌های جدید و کارا دقت سیستم‌های بازشناسی احساس را افزایش می‌دهند. در این مقاله، نرم‌افزار openSMILE(۷) جهت استخراج ویژگی‌ها استفاده گردیده و با روش‌های افزایش داده متخاصمی سعی در تولید

1. Conditional Generative Adversarial Networks
2. Adversarial Auto Encoders

بردارهای ویژگی جدید جهت افزایش نمونه‌های آموزش و متعادل نمودن پایگاه‌های داده شده است. همچنین، با استفاده از روش‌های انتخاب ویژگی، ویژگی‌های کمتر اثرگذار بر روی کارایی سیستم تشخیص احساس از گفتار حذف و ویژگی‌های مؤثر معرفی گردیدند. از چهار پایگاه‌داده متداول EMO-DB، eNTERFACE05، SAVEE و EMOVO جهت انجام آزمایش‌ها استفاده شده و علاوه بر این، تجزیه‌وتحلیل داده‌ها بر روی هر چهار پایگاه‌داده برای چهار احساس غمگین، عصبانی، خوشحال و خنثی ارائه گردیده است.

بخش ۳ مروری بر راه‌حل‌های رایج برای مشکل کمبود داده‌ها را ارائه می‌دهد. بخش ۴ طراحی شبکه پیشنهادی را توصیف می‌کند و تجزیه‌وتحلیل نظری را ارائه می‌دهد. بخش ۵ جزئیات آزمایش‌ها، از جمله توصیف داده‌ها، ویژگی‌ها، تنظیمات آزمایشی و پروتکل‌های ارزیابی را معرفی می‌کند. بخش ۶ نتایج تجربی را ارائه و تحلیل می‌کند. سرانجام، بخش ۷ نتیجه‌گیری و کارهای آینده را ارائه می‌دهد.

مطالعات مرتبط

کمبود داده‌ها می‌تواند باعث شود که یک مدل یادگیری ماشین قادر به یادگیری توزیع واقعی داده‌ها نباشد و در نهایت منجر به مشکل اور فیتینگ می‌شود. به‌عنوان‌مثال، در هنگام آموزش یک شبکه عصبی عمیق با نمونه‌های آموزشی کم در هر کلاس، اور فیتینگ رخ می‌دهد، درصورتی‌که هر نمونه هزاران ویژگی دارد. برای حل این مشکل، می‌توان از قاعده‌مندسازی برای ایجاد محدودیت در مدل استفاده کرد (۸). راه‌حل کلی دیگر، کاهش ابعاد با محدودیت پراکندگی است (۹). این رویکرد زمانی مؤثر خواهد بود که ویژگی‌های اضافی وجود داشته باشد. در غیر این صورت، اطلاعات مفید را از بین می‌برد و منجر به تخریب عملکرد می‌شود. همچنین، وجود بردارهای ویژگی کم‌اهمیت در آموزش شبکه عصبی، محاسبات را در این شبکه پیچیده و تأثیرگذاری نمونه‌ها را در آموزش بی‌اثر خواهد نمود؛ لذا انتخاب و حذف بردارهای ویژگی کم‌اهمیت موضوعی است که کمتر به آن در روش‌های افزایش داده به طور هم‌زمان پرداخته شده است.

برخی مقالات از روش‌های GMM و HMM استفاده می‌کنند تا توزیع ویژگی‌های آکوستیکی را یاد بگیرند و سپس از طبقه‌بندی بیزین یا اصل ماکزیمم بخت برای تشخیص احساس استفاده می‌کنند. برخی از مدل‌های پس‌زمینه جامع[1] استفاده می‌کنند تا از ویژگی‌های آکوستیکی، بردار ویژگی برای آموزش شبکه ماشین بردار پشتیبان بسازند که این روش بیشتر برای تشخیص گوینده استفاده می‌شود. برخی، روش‌های آماری را روی ویژگی‌های سطح پایین استفاده می‌کنند تا ویژگی‌های آماری سطح بالا را به دست آورند و سپس از ماشین بردار پشتیبان برای طبقه‌بندی ویژگی‌های کلی استفاده می‌شود. برخی از روش‌های KNN و یا درخت تصمیم‌گیری[2] استفاده می‌کنند که نیاز به تعداد ویژگی‌های زیاد و ویژگی‌های ساخته شده به‌صورت دستی دارد. بررسی چند نمونه از تحقیقات مهم این شاخه می‌تواند به نمایش بهتر مسئله و روشن‌شدن سمت‌وسوی پژوهش‌های مرتبط کمک کند.

در سال ۲۰۰۹، Rong و همکارانش از ویژگی‌های مبتنی بر فرکانس گام، انرژی، طول گفتار، ZCR و MFCC به‌منظور طبقه‌بندی احساس خشم، خوشحالی، ناراحتی و عادی استفاده نمودند. آن‌ها در این تحقیق با ابداع الگوریتم انتخاب ویژگی Ensemble random forest to trees و استفاده از درخت تصمیم‌گیری، بهترین نرخ تشخیص ۷۴/۵۸٪ را برای پایگاه داده‌ای که توسط خودشان طراحی شده بود به دست آوردند (۱۰). در همین سال Chandaka به‌منظور جداسازی احساس‌های خشم، خوشحالی، ناراحتی و عادی ویژگی‌های جدید مبتنی بر همبستگی متقابل[3] را از جملات پایگاه‌داده DES استخراج نموده و آن‌ها را به یک طبقه‌بند با ساختار درختی مبتنی بر ماشین‌های بردار پشتیبان اعمال نمود. وی بیشترین نرخ تشخیص ۸۴/۵۵٪ را به دست آورد (۱۱). همچنین در تحقیق دیگری، Altun با استخراج ویژگی‌های مبتنی بر فرکانس گام، نرخ صامت به مصوت، MFCC، LPC و چند ویژگی جدید با استفاده از زیر باندها[4]، به طبقه‌بندی جملات با احساس‌های خشم، خوشحالی، ناراحتی و عادی در پایگاه‌داده احساسی برلین پرداخت. بدین منظور

1. Universal Background Models (UBMs)
2. Decision Tree
3. Cross-Correlation
4. Sub-bands

با آزمایش چند نوع الگوریتم انتخاب ویژگی و دو نوع طبقه‌بند مبتنی بر ماشین‌های بردار پشتیبان، بهترین نرخ تشخیص ۷۳/۴٪ به دست آمد (۱۲).

Bitouk و همکارانش در سال ۲۰۱۰ با استفاده از ویژگی‌های متداول عروضی و طیفی به طبقه‌بندی جملات با احساس‌های خشم، ترس، انزجار، خوشحالی، ناراحتی و عادی بر روی پایگاه‌های داده برلین و LDC پرداختند. آن‌ها با استفاده از یک روش انتخاب ویژگی بسته‌بند و ماشین بردار پشتیبان به بهترین نرخ تشخیص ۷۲/۵٪ دست پیدا کردند (۱۳). در این سال Yong و همکارش علاوه بر ویژگی‌های متداول عروضی، طیفی و کیفیتی گفتار، ویژگی‌های جدیدی مبتنی بر تئوری موسیقی ارائه نمودند. آن‌ها به‌منظور ارزیابی روش پیشنهادی خود از یک الگوریتم انتخاب ویژگی مبتنی بر SFFS[1] و طبقه‌بندهای مبتنی بر شبکه‌های عصبی، مدل مخفی مارکوف و ماشین‌های بردار پشتیبان به‌منظور جداسازی شش احساس خشم، خوشحالی، خستگی، ناراحتی، نگرانی و عادی بر روی پایگاه‌داده برلین استفاده نمودند و به بهترین نرخ تشخیص ۷۵٪ دست پیدا کردند (۱۴).

در سال ۲۰۱۱، Polzehl و همکارانش به‌منظور جداسازی احساس خشم نسبت به سایر احساس‌ها در پایگاه‌داده Germam Woz، English IVR، Germam IVR، ویژگی‌های مبتنی بر فرکانس گام، انرژی، MFCC و فرمنت‌ها را از داده‌ها استخراج نموده و با استفاده از الگوریتم انتخاب ویژگی مبتنی بر فیلتر نرخ بهره اطلاعاتی[2] و طبقه‌بندی مبتنی بر ماشین بردار پشتیبان به بهترین نرخ تشخیص ۷۹٪ دست پیدا کردند (۱۵). در همین سال Lee و همکارانش با استفاده از ویژگی‌های متداول عروضی و طیفی در جداسازی احساس‌های خشم، تأکید[3]، عادی مثبت[4] و آرامش[5] بر روی پایگاه‌داده AIBO و USC و IEMOCAP بهترین نرخ تشخیص ۴۱/۵۷٪ را به دست آوردند (۱۶). درحالی‌که Kockmann و همکارانش با معرفی یک گروه ویژگی طیفی جدید نرخ تشخیص ۷۰٪ را برای جداسازی همین احساس‌ها بر روی پایگاه‌داده AIBO دست آوردند(۱۷). Albornoz و همکارانش نیز با معرفی یک گروه ویژگی طیفی جدید به نرخ تشخیص ۸۸/۸۹٪ در جداسازی ۷ احساس در پایگاه‌داده برلین دست یافتند (۱۸). در همین سال، داود قراویان و همکارانش تشخیص احساس از گفتار را با استفاده از روش انتخاب ویژگی FCBF و شبکه عصبی فازی ARTMAP ارائه دادند. آن‌ها با استفاده از شبکه عصبی فازی ARTMAP، یک سیستم تشخیص احساس پیاده‌سازی کردند و برای رسیدن به این هدف، ویژگی‌های پایه مانند فرکانس پیچ، فرکانس فرمنت، انرژی، MFCCs، و ویژگی‌های مربوط به آن‌ها را استفاده نمودند و تأثیر استفاده از این ویژگی‌ها در بهبود عملکرد سیستم تشخیص احساس گفتار را بررسی نمودند و نشان دادند که با استفاده از ۲۵ ویژگی انتخاب شده با استفاده از روش FCBF، نرخ ۸۴/۹۷٪ برای تشخیص احساسات دست‌یافتنی است و با استفاده از FAMNN بهینه‌سازی‌شده (ساختار شبکه عصبی فازی ARTMAP)، دقت تشخیص احساسات افزایش می‌یابد و به بیش از ۸۷/۵۲٪ می‌رسد.

در سال ۲۰۱۲، منصور شیخان و همکارانش از شبکه ماشین بردار پشتیبان برای تشخیص احساسات گفتار با استفاده از روش انتخاب ویژگی ANOVA بهره بردند. آن‌ها نشان دادند که حتی با حذف ۲۲٪ از ویژگی‌ها، به طور متوسط دقت تشخیص احساسات را می‌توان تا ۲/۲٪ بهبود داد. همچنین، شبکه ماشین بردار پشتیبان پیشنهادی دقت تشخیص را حداقل به میزان ۸٪ در مقایسه با طبقه‌بندی‌کننده یکپارچه شبیه‌سازی‌شده بهبود می‌دهد. در همین سال، Bozkurt و همکارانش با وزن‌دهی به ضرایب مل کپسترول باتوجه‌به موقعیت فرمنت‌ها، ویژگی‌های جدیدی به نام WMFCC[6] را ابداع نموده و با استفاده از این ویژگی‌ها و یک طبقه‌بند مبتنی بر مدل مخفی مارکوف به طبقه‌بندی چهار احساس خشم، آرامش، تأکید و عادی بر روی پایگاه‌داده AIBO پرداختند. آن‌ها به‌منظور ارزیابی روش خود از نرخ تشخیص بدون وزن استفاده نموده و به عدد ۷۰٪ دست یافتند (۱۹). در یکی از تحقیقات شاخص صورت‌گرفته در این سال، Wu و همکارانش ویژگی‌های طیفی جدیدی را معرفی نموده و از این ویژگی‌ها به‌منظور

1. Sequential Floating Forward Selection (SFFS)
2. Information Gain Ratio
3. Emphatic
4. Neutral Positive
5. Rest
6. Weighted MFCC

طبقه‌بندی هفت احساس موجود در پایگاه‌داده برلین و تخمین احساس در فضای پیوسته بر روی پایگاه‌داده VAM استفاده نمودند. آن‌ها با استفاده از الگوریتم انتخاب ویژگی دومرحله‌ای فیلتر – بسته‌بند مبتنی بر معیار فیشر، الگوریتم جداساز خطی و جستجوی روبه‌جلو و طبقه‌بند مبتنی بر ماشین بردار پشتیبان به بهترین نرخ تشخیص ۸۵/۶٪ برای پایگاه‌داده برلین دست پیدا کردند. همچنین با استفاده از رگرسیون مبتنی بر ماشین‌های بردار پشتیان متوسط ضریب همبستگی ۷۳٪ را برای پارامترهای تخمین زده شده در فضای پیوسته احساس به دست آوردند (۲۰). در این سال Laukka و همکارانش سیستم خود را که به‌منظور بازشناسی احساس بر روی یک پایگاه‌داده طبیعی که با استفاده از مکالمات تلفنی ضبط شده طراحی نموده بودند آزمایش کرده و نتایج را با استفاده از ماتریس‌های تداخل گزارش نمودند (۲۱). پس از آن، طیف وسیعی از ویژگی‌های عروضی و طیفی به‌منظور جداسازی چهار احساس خشم، خوشحالی، ناراحتی و عادی بر روی پایگاه‌داده IEMOCAP مورد آزمایش قرار گرفتند. این تحقیقات با استفاده از الگوریتم‌های انتخاب ویژگی جستجوی روبه‌جلو و طبقه‌بندی مبتنی بر ماشین بردار پشتیبان، کارایی ویژگی‌های مختلف را در طبقه‌بندی هر احساس به‌خوبی با یکدیگر مقایسه نموده‌اند (۲۲).

در سال ۲۰۱۳، مروی و همکارش از ترکیب ویژگی‌های طیفی (ویژگی‌هایی که از طیف سیگنال دست می‌آیند مثل فرمنت‌ها، MFCC و PLP) و ویژگی‌های عروضی (ویژگی‌هایی که از آنالیز سیگنال در حوزه زمان دست می‌آیند و اغلب از منحنی فرکانس گام و انرژی سیگنال استخراج می‌شوند) استفاده کردند. آن‌ها از الگوریتم دومرحله‌ای شامل معیار فیشر و الگوریتم جداساز خطی به‌منظور کاهش ویژگی‌ها و همچنین به‌منظور تشخیص احساس از گفتار استفاده نمودند و نشان دادند که ترکیب ویژگی‌های عروضی و طیفی باعث افزایش متوسط نرخ تشخیص می‌شود و نرخ تشخیص در پایگاه‌داده درام را برای گویندگان مرد به ۴۷/۲۸٪ و نرخ تشخیص در گویندگان زن را به ۵۵/۷۴٪ افزایش داده است.

در مرجع (۲۳) ابتدا برای هر سگمنت یعنی مجموعه‌ای از فریم‌های پشت‌سرهم، با استفاده از شبکه عصبی عمیق یک توزیع احتمالی حالت احساسی مبتنی بر سگمنت تولید شده و سپس ازروی این توزیع‌های احتمالی، ویژگی‌های در سطح کلی به‌دست‌آمده است. این ویژگی‌های کلی به یک ELM داده شده (یک شبکه عصبی خاص تک‌لایه مخفی ساده و مؤثر) تا احساسات کلی به دست آید. علت استفاده از ELM در گام آخر این است که ELM ساده‌تر از DNN است و داده‌های کمتری برای آموزش نیاز دارد و در اینجا عملکرد خیلی بهتری از ماشین بردار پشتیبان دارد. نتایج عملی نشان می‌دهند که رویکرد پیشنهاد شده اطلاعات احساسی را به طرز مؤثری از ویژگی‌های سطح پایین به دست می‌آورد و منجر به افزایش ۲۰ درصدی دقت در مقایسه با رویکردهای جدید شده است.

در مرجع (۲۴) از روش اصلاح شده VQ برای کاهش ویژگی‌های آکوستیکی استفاده شده است. به این صورت که روش VQ و همچنین روش VQ تفاضلی روی ویژگی‌های آکوستیکی در سطح فریم استفاده شده است. استفاده از VQ تفاضلی پایداری ویژگی‌ها را با اضافه‌کردن دینامیک زمانی که حین آنالیز آماری از بین می‌رود تقویت می‌کند. مکانیزم استفاده شده در اظهارات مختلف ارزیابی شده است به‌منظور رسیدن به این هدف یک پایگاه‌داده محلی هم‌زمان با پایگاه‌داده EMO-DB با الگوریتم پیشنهاد شده استفاده شده است. همچنین پارامترهای طراحی طبقه‌بندی‌کننده و ویژگی‌های کاهش داده شده برای عملکرد بهینه فراهم شده است و در نهایت ویژگی‌های دیگری در سطح فریم و در سطح کل سخن اضافه شده تا بسته پیشنهادی تکمیل گردد. استفاده از روش‌های انتخاب ویژگی همیشه موفقیت طبقه‌بندی را افزایش نمی‌دهد. در یک مطالعه که روش‌های الگوریتم جداساز خطی و آنالیز مؤلفه‌های اصلی را مقایسه می‌کرد، مشاهده شد که استفاده از هر دو روش به طور هم‌زمان، منجر به نتیجه بهتری نسبت به هرکدام از آن‌ها می‌شود. زیرا آنالیز مؤلفه‌های اصلی روی داده‌های ناهمبسته مؤثرتر عمل می‌کند و الگوریتم جداساز خطی روی داده‌های با بعد کم بهتر عمل می‌کند. روش فیشر در کاهش سایز بهتر از آنالیز مؤلفه‌های اصلی عمل می‌کند. در مرجع (۲۵) ابعاد ویژگی را از ۲۷۶ به ۷۵ توسط روش SFFS کاهش دادند و تشخیص احساس به میزان ۲/۷٪ بهتر شد. در یک مطالعه از روش‌های مختلفی برای کاهش ویژگی از ۵۸ به

۱۸، ۲۸، ۳۱ و ۳۳ استفاده شد و دقت تشخیص احساس به ترتیب به میزان ۴/۵ و ۳ و ۲/۲ و ۰/۷ درصد افزایش پیدا کرد.

در مرجع (۲۶) عملکرد تشخیص هنگامی که ابعاد ویژگی توسط روش mRMR از ۳۸۰ به ۱۲۱ کاهش پیدا کرد، به میزان ۱/۵ درصد کم شد. در مرجع (۲۷) هنگامی که ابعاد ویژگی از ۵۵ به ۴۹ و ۴۵ و ۲۴ و ۸ توسط روش FCBF کاهش پیدا کرد تغییر در میزان تشخیص به ترتیب به میزان ۰/۹ و ۱/۱- و ۲/۳- و ۳/۴- درصد به دست آمد. در مرجع (۲۸) نرخ تشخیص با کاهش ویژگی از ۲۰۴ به ۸۷ توسط روش FCBF و طبقه‌بند ماشین بردار پشتیبان به میزان ٪۱/۵ افزایش پیدا کرد.

هو و همکاران (۲۹) از شبکه عصبی کانالوشن بسیار عمیق برای تولید ویژگی‌های اضافی برای آموزش مدل‌های صوتی استفاده کردند و دریافتند که افزایش داده به ساخت سیستم‌های تشخیص گفتار کمک بسیاری می‌کند.

در سال‌های اخیر، از شبکه‌های مولد متخاصمی برای تشخیص احساس از گفتار استفاده شده است. به‌عنوان‌مثال، چانگ و شفر از یک شبکه مولد متخاصمی عمیق[1] برای یادگیری نمایش متمایز از گفتار احساسی به روشی نیمه نظارت استفاده کردند (۳۰). این مقاله اما به کاربرد شبکه‌های مولد متخاصمی برای تولید داده‌های مصنوعی و انتخاب بهترین آن‌ها متمرکز شده است که هدف از آن تولید داده‌هایی است که توزیع داده‌های واقعی را گسترش می‌دهند تا عملکرد سیستم تشخیص احساس از گفتار را بهبود بخشند. به بیان دیگر، فرایند انتخاب ویژگی در داده‌های تولید شده سبب می‌شود که آن‌ها هدف‌مند به داده‌های اصلی اضافه گردند و سبب آموزش هر چه‌بهتر شبکه کلاسه بند احساس شوند و حجم محاسبات را کاهش و کارایی این شبکه را بیفزایند.

یک مشکل اساسی در آموزش شبکه‌های مولد متخاصمی، اطمینان از تعادل بین قابلیت شبکه مولد و شبکه تشخیص‌دهنده است. برای غلبه بر این مشکل، می‌توان از تکنیک آموزش تناوب دینامیکی[2] (۳۱) استفاده کرد. این استراتژی آموزشی، تعداد دوره‌های آموزش شبکه مولد و شبکه تشخیص‌دهنده را تغییر می‌دهد. در واقع، به‌جای بهینه‌سازی تعداد دوره‌های آموزشی، با تسهیل شبکه مولد جهت یادگیری توزیع داده‌های هدف، ثبات یادگیری را بهبود می‌بخشد. مشکل اساسی بعدی، انتخاب ویژگی‌های مؤثر در سیستم تشخیص احساس از گفتار است. اگر ویژگی‌ها به‌صورت هوشمندانه افزوده نشوند، حجم محاسبات بالا رفته و سیستم تشخیص احساس کند خواهد شد. همچنین ممکن است بردارهای ویژگی‌های افزوده شده، مرز بین دو کلاس را از بین برده و یا داده‌های مرزی تولید شوند که کار کلاسه‌بندی را مشکل و یا راندمان کلاسه‌بندی را پایین آورند. ازاین‌رو توجه به انتخاب ویژگی همیشه لازم خواهد بود.

ساختار کلی پیشنهاد شده

شکل ۱، ساختار کلی شبکه طبقه‌بندی‌کننده احساس پیشنهاد شده را برای طبقه‌بندی چهار احساس نمایش می‌دهد. از روش ترکیبی، جهت انتخاب ویژگی و از روش افزایش داده متخاصمی جهت افزایش داده‌ها و از ماشین بردار پشتیبان جهت طبقه‌بندی استفاده گردیده است.

پیش‌پردازش و استخراج ویژگی

تشخیص احساس از گفتار تاکنون به دلیل نبود پایگاه‌داده با داده‌های فراوان و در دسترس با مشکلات فراوانی روبه‌رو بوده است. عدم توانایی در انتخاب ویژگی‌های مهم و تأثیرگذار در شناسایی احساس، امکان استفاده از این سیستم‌ها را در کاربردهای آنلاین و هم‌زمان محدود می‌کند. همچنین وابسته بودن این سیستم به زبان، لهجه، سن، جنسیت، و نوع حالت گوینده از عمده مشکلات این روش بوده است.

1. DCGAN
2. Dynamic Alternation Training

شکل ۱) ساختار کلی شبکه تشخیص احساس پیشنهادی

جدول ۱) تعداد و نوع ویژگی‌های استخراج شده

نوع ویژگی	مشخصات ویژگی
الگوهای طیفی (۲۰۴ ویژگی):	میانگین تعداد الگوها در هر قاب برای هر باند فرکانسی تعداد نسبی هر الگو در باندهای فرکانسی و تعداد نسبی هر الگو در باندهای فرکانسی
انرژی هارمونیک‌ها (۷۸۰ ویژگی):	اعمال ۲۰ تابع آماری به E1 و E2 و مشتقات اول و دوم آن‌ها
ویژگی‌های عروضی (۲۴۱ ویژگی):	اعمال ۲۰ تابع آماری به منحنی فرکانس گام، منحنی انرژی، منحنی نرخ عبور از صفر، منحنی اپراتور انرژی تیگر و همچنین نسبت طول زمانی مصوت‌ها به صامت‌ها و کلیه مشتقات اول و دوم آن‌ها
ویژگی‌های طیفی (۹۶۰ ویژگی):	اعمال ۲۰ تابع آماری به ۱۲ ضریب اول MFCC، ۴ فرمنت اول و مشتقات اول و دوم آن‌ها
۲۰ تابع آماری شامل: مقدار کمینه، بیشینه، برد، میانگین، میانه، ۱۰٪ و ۲۵٪ صدک‌های اول، پنجم، دهم، بیست و پنجم، هفتاد و پنجم، نودم، نود و پنجم و نود و نهم، برد میان چارکی، واریانس، انحراف معیار، چولگی و کشیدگی.	

در این مقاله، پس از انجام پیش‌پردازش و حذف نویز از سیگنال گفتار، بردارهای ویژگی توسط نرم‌افزار openSMILE استخراج گردید. نرم‌افزار openSMILE(۷) برای استخراج ویژگی‌های احساسی در چالش Interspeech 2011 Speaker State Challenge(۳۲) مورداستفاده قرار گرفت و یک بردار ویژگی ۴۳۶۸ بعدی را برای هر گفتار ارائه نمود. این مجموعه ویژگی نسخه توسعه‌یافته‌ای از چالش Interspeech 2009 Emotion Challenge(۳۳) و چالش Paralinguistic Interspeech 2010(۳۴) بود. اولی عمدتاً در پرداختن به حالات عاطفی کوتاه‌مدت متمرکز است و دومی بویژگی‌های گوینده مانند سن و جنسیت می‌پردازد. چالش ۲۰۱۱ حالات احساسی کوتاه‌مدت و ویژگی‌های احساسی طولانی‌مدت را در نظر می‌گیرد. در نتیجه، این ویژگی‌ها می‌توانند حالات احساسی را به‌خوبی نشان دهند و تعداد زیاد ویژگی‌ها برای مطالعه مشکل کمبود داده‌ها در پایگاه‌داده، مناسب است.

در مجموعه ویژگی‌های چالش Interspeech 2011 از ویژگی‌های سطح پایین مثل انرژی فریم میانگین مربع[۱]، ضرایب کپسترال مل فرکانسی[۲]، نرخ عبور از صفر[۳]، احتمال صدا[۴] و فرکانس‌های اساسی[۵] استفاده شد(۳۵). برای استخراج این ویژگی‌ها، فریم‌های ۲۵ میلی‌ثانیه با هم‌پوشانی ۱۰ میلی‌ثانیه از سیگنال گفتار در نظر گرفته شدند. سپس ویژگی‌های سطح پایین توسط عملکردهای آماری مانند حداکثر، حداقل، دامنه، انحراف معیار و... برای استخراج ویژگی‌های سطح بالا پردازش شدند. با استفاده از این عملکردهای آماری به ویژگی‌های سطح پایین، می‌توان ویژگی‌های فریم، فریم شده را به ویژگی‌های بیانی[۶] تعمیم داد. جزئیات بیشتر را می‌توان در وب‌سایت openSMILE[۷] یافت. برای هر چهار مجموعه پایگاه‌داده، ویژگی‌های با واریانس صفر حذف شدند و بقیه ویژگی‌ها به طور مستقل توسط z-norm نرمال گردید. تعداد و نوع این ویژگی‌ها برای هر نمونه، ۲۱۸۵ ویژگی مطابق جدول ۱ است.

شکل ۲، ساختار شبکه افزایش داده متخاصمی پیشنهاد شده را نشان می‌دهد که شامل یک شبکه رمزگذار خودکار E(x)، یک شبکه رمزگشا R(E(x)) ، یک شبکه مولد G(z,y) و یک شبکه تشخیص‌دهنده D(h) است. شبکه افزایش داده

1. Root-Mean-Square (RMS) Frame Energies
2. Mel-Frequency Cepstrum Coefficients (MFCCs)
3. Zero-Crossing Rates
4. Voice Probabilities
5. Fundamental Frequencies
6. Utterance-Level Features
7. https://www.audeering.com/openSMILE /

متخاصمی برای دستیابی به سه هدف عمده طراحی شده است. در هدف اول، توزیع داده‌هایی را می‌آموزد که اطلاعات احساسی در آن‌ها پنهان هستند. در هدف دوم، سعی در تطبیق توزیع پسین $p(\hat{h}|z,y)$ با توزیع پسین $p(h|x)$ می‌نماید و در هدف سوم، سعی می‌نماید خطاهای بازسازی بین x و $\hat{x}$ را به حداقل ممکن برساند. هر چهار شبکه در شکل ۱ برای دستیابی به این اهداف به طور متخاصمانه آموزش داده شده‌اند. رمزگذار E برای یادگیری توزیع m بعدی داده‌های h که حاوی احساسات هستند، آموزش‌دیده است. به طور هم‌زمان، آشکارساز می‌آموزد که بردارهای احساس را از فضای ویژگی‌ها در فضای اصلی بازسازی کند. شبکه مولد، نمونه‌های گرفته شده از توزیع گاوسی M و برچسب‌های احساسی رمزگذاری شده را به‌عنوان ورودی گرفته و در فضای ویژگی، نمونه‌های تقلبی تولید می‌کند و هدف آن تولید نمونه‌هایی است که از نمونه‌های واقعی در فضای ویژگی قابل‌تشخیص نیستند، یعنی $p(h|x) \approx p(\hat{h}|z,y)$. شبکه تشخیص‌دهنده برای تشخیص اینکه یک بردار ویژگی از داده‌های واقعی ناشی می‌شود یا در شبکه مولد تولید شده است بهینه می‌شود. مزیت تولید نمونه در فضای ویژگی بجای تولید نمونه در فضای اصلی این است که می‌توان از تولید بردارهای با ابعاد بالا جلوگیری کرد.

برای آموزش شبکه پیشنهادی، تلفات تعریف شده به شرح زیر به حداقل ممکن می‌رسد:

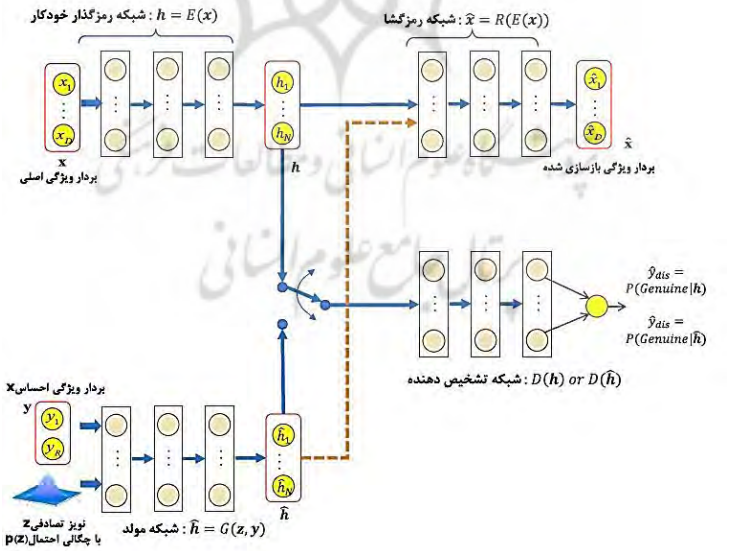
$$\mathcal{L}_D^{(ADAN)} = -\mathbb{E}_{x\sim P_{data}(x)}\{\log D(\mathrm{E}(\mathbf{x}))\} - \mathbb{E}_{z\sim P_z(z)}\{\log(1 - D(\mathrm{G}(\mathbf{z},\mathbf{y})))\} \quad (۱)$$

$$\mathcal{L}_R^{(ADAN)} = \mathbb{E}_{x\sim P_{data}(x)}\left\{\|\mathbf{x} - R(\mathrm{E}(\mathbf{x}))\|^2\right\} \quad (۲)$$

$$\mathcal{L}_E^{(ADAN)} = \mathbb{E}_{x\sim P_{data}(x)}\left\{\|\mathbf{x} - R(\mathrm{E}(\mathbf{x}))\|^2 - \sum_{k=1}^{K} y_{emo}^{(k)} \log C(\mathrm{E}(\mathbf{x}))_k\right\} \quad (۳)$$

$$\mathcal{L}_G^{(ADAN)} = \mathbb{E}_{z\sim P_z(z)}\left\{\log(1 - D(\mathrm{G}(\mathbf{z},\mathbf{y}))) - \alpha \sum_{k=1}^{K} y_{emo}^{(k)} \log C(\mathrm{G}(\mathbf{z},\mathbf{y}))_k\right\} \quad (۴)$$

که k نشان‌دهنده عنصر k اُم یک بردار است و G مخفف شبکه مولد، R برای شبکه آشکارساز، E برای شبکه رمزگذار، D برای شبکه تشخیص‌دهنده است. αسهم خطای طبقه‌بندی را در تلفات شبکه مولد را تعیین می‌کند.



شکل۲) ساختار شبکه افزایش داده متخاصمی پیشنهادی. این شبکه شامل یک شبکه رمزگذار خودکار در بالا سمت چپ، یک شبکه رمزگشا در بالا سمت راست، یک شبکه مولد در پایین سمت چپ و یک شبکه تشخیص‌دهنده در پایین سمت راست شکل است. تمامی شبکه‌ها از نوع شبکه عصبی عمیق هستند و خط نقطه‌چین فقط برای افزایش داده‌ها و پس از آموزش شبکه استفاده می‌شود

شبکه افزایش داده متخاصمی واسراستین(۳۶) برای غلبه بر مشکل ازبین‌رفتن و کوچک‌شدن گرادیان پیشنهاد شده است. با کوچک‌شدن بیش از حد گرادیان، فرایند اصلاح وزن و آموزش شبکه عملاً متوقف می‌شود. باتوجه به دو توزیع احتمال Pr و Pg، فاصله واسراستین به این صورت تعریف می‌شود:

$$W_1(\mathbb{P}_r, \mathbb{P}_g) = \sup_{\|f\|_L \le 1} \mathop{\mathbb{E}}_{x\sim \mathbb{P}_r}\{f(x)\} - \mathop{\mathbb{E}}_{\tilde{x}\sim \mathbb{P}_g}\{f(\tilde{x})\} \quad (۵)$$

جایی که fL≤ ۱ نشان می‌دهد که f محدودیت ۱-Lipschitz را برآورده می‌کند. برش وزن و مجازات گرادیان دو روش معمول برای اعمال محدودیت ۱-Lipschitz است. در روش برش وزن، اگر وزن‌ها از محدوده مورد انتظار بیشتر و یا کمتر شود، آن‌ها را به حداقل یا حداکثر مقداری خاص تبدیل خواهیم نمود و در روش مجازات گرادیان، جریمه گرادیان بر اساس محدودیت لیپشیتس است که از این واقعیت ناشی می‌شود که اگر شیب‌ها در همه‌جا حداکثر یک باشند، توابع یک لیپشیتس هستند. اختلاف مربع آن‌ها از عدد یک به‌عنوان جریمه گرادیان استفاده می‌شود. بااین‌حال، طبق(۳۷)، برش وزن می‌تواند فضای جستجوگر تابع f را باریک کند و منجر به یک راه‌حل غیربهینه شود. برای غلبه بر محدودیت‌های برش وزن، مجازات گرادیان اعمال شد(۳۸). بااین‌حال، در شرایط محدود بودن تعداد داده‌ها، برآورده کردن محدودیت k-Lipschitz برای کل حوزه داده دشوار است. با این ملاحظات، وو و همکاران(۳۷) واگرایی جدیدی را برای دیورژانس و اسراستین پیشنهاد کردند که می‌تواند بدون اعمال محدودیت لیپشیتس، فاصله تقریبی واسراستین را محاسبه نماید که به شرح زیر است:

$$L_{DIV} = \mathop{\mathbb{E}}_{x\sim \mathbb{P}_r}\{f(x)\} - \mathop{\mathbb{E}}_{\tilde{x}\sim \mathbb{P}_g}\{f(\tilde{x})\} + \lambda \mathop{\mathbb{E}}_{\tilde{x}\sim \mathbb{P}_u}[\|\nabla f(\tilde{x})\|^p] \quad (۶)$$

جایی که Pu اندازه‌گیری احتمال رادون است، λ تأثیر اصطلاح گرادیان را روی تابع هدف کنترل می‌کند و p مربوط به فضای Lp برای تابع f است. علاوه بر این، λ و p باید به‌صورت λ>۰ و p>۱ باشند تا اطمینان حاصل شود، همان‌طور که در(۳۷) ثابت شده است، LDIV در(۸) یک واگرایی متقارن است. با درنظرگرفتن(۶) برای شبکه‌های افزایش داده متخاصمی، توابع خطا در شبکه مولد و شبکه تشخیص‌دهنده عبارت‌اند از:

$$\mathcal{L}_D^{(WADAN)} = \mathbb{E}_{p(x,z,\tilde{x},y)}\{D(E(x) - D(G(z,y)) + \lambda[\|\nabla_{\tilde{x}}\|^p]\} \quad (۷)$$

$$\mathcal{L}_G^{(WADAN)} = \mathbb{E}_{p(x,y,z)}\left\{D(G(z,y)) - a\sum_{k=1}^{K} y_{emo}^{(k)} \log C(G(z,y))_k\right\} \quad (۸)$$

توابع خطای دیگر همانند(۱)-(۳) خواهند بود. ساختار شبکه افزایش داده متخاصمی واسراستین نیز همانند شکل ۱ است. تفاوت دیگر این دو شبکه در این است که لایه نهایی شبکه تشخیص‌دهنده در حالت اول از تابع فعال‌ساز سیگموئید[1] استفاده می‌کند، درحالی‌که در شبکه‌های افزایش داده متخاصمی واسراستین، در لایه نهایی، شبکه تشخیص‌دهنده از تابع فعال‌ساز خطی[2] بهره می‌گیرد.

پس از آموزش شبکه افزایش داده متخاصمی و یا شبکه افزایش داده متخاصمی واسراستین، شبکه مولد به شبکه رمزگشای R (که با نقطه‌چین در شکل ۱ مشخص شده است) برای افزایش داده متصل شده است. با واردکردن برچسب‌های احساسات و بردارهای تصادفی گاوسی z به شبکه مولد، می‌توان نمونه‌های مصنوعی را از خروجی شبکه رمزگشا به دست آورد.

آزمایش‌ها

دقت[3]، صحت[4] و امتیاز اف یک[5] سه سنجه متداولی هستند که علاوه بر صحت دسته‌بندی برای مسائل دسته‌بندی یک مدل شبکه عصبی مورداستفاده قرار می‌گیرند. سه سنجه اضافه دیگری که کمتر متداول اما محبوب هستند عبارت‌اند از: ضریب کاپای کوهن[6]، MCC[7] و ماتریس درهم ریختگی[8]. از معیار صحت جهت معیار ارزیابی آزمایش‌ها استفاده گردیده است. شبیه‌سازی‌ها با استفاده از Keras و

1. Sigmoid Activation Function
2. Linear Activation

3. Precision
4. Recall
5. F1-Score
6. Cohen's Kappa
7. Matthews Correlation Coefficient
8. Confusion Matrix

Tensorflow2.2 در نرم‌افزار پایتون Python 3.8 (64-bit) انجام گردیدند.

آزمایش‌ها این مقاله بر روی پایگاه‌داده گفتار احساسی برلین[1](۳۹) انجام شده است. پایگاه‌داده گفتار احساسی برلین، یک مجموعه‌داده کوچک است که شامل ۸۰۰ جمله است که به هفت کلاس احساس تقسیم شده‌اند. تمام گفته‌ها توسط ده بازیگر حرفه‌ای ضبط شده است. بُعد بردارهای ویژگی نمونه‌برداری شده روی ۱۰۰ قرار داده شد، یعنی $M = dim(h) = dim(\hat{h}) = 100$. پارامتر α در(۴) برابر یک قرار گرفت. برای مقابله با مشکل عدم توازن تعداد داده‌ها در هر کلاس، وزن‌های مختلف در هر کلاس برای خطای شبکه اختصاص‌یافته است. اگر بزرگ‌ترین کلاس دارای *nm* نمونه باشد و کلاس k دارای *nk* نمونه باشد، وزن اختصاص داده شده به این کلاس *(nm/nk)* خواهد بود. وقتی از کل مجموعه پایگاه‌داده هیجان‌انگیز احساسی در اعتبارسنجی استفاده می‌شود، پارامتر α در(۴) و(۸) روی ۰/۱ تنظیم گردید. بر اساس تجربه، مقادیر ضرایب λ و p به ترتیب روی ۱۰ و ۵ تنظیم شدند. هنگامی‌که شبکه همگرا شد، شبکه مولد برای تولید نمونه‌های مصنوعی به شبکه آشکارساز متصل می‌شود. برچسب‌های احساسی و بردارهای تصادفی گاوسی z به ورودی شبکه مولد داده می‌شود. بردارهای نهان مصنوعی خروجی از مولد سپس به شبکه رمزگشا (نقطه‌چین در شکل ۱) منتقل می‌شوند تا داده‌های افزوده شده را در فضای اصلی تولید کنند. در این مقاله، ده مجموعه افزوده ایجاد شد که هر یک از آن‌ها دارای اندازه و توزیع برچسب مشابه مجموعه اصلی هستند. سپس داده‌های مصنوعی به مجموعه آموزش اولیه برای انتقال احساسات افزوده شد. اجزای موجود در شبکه افزایش داده متخاصمی، شبکه‌های عصبی عمیق کاملاً متصل[2] با دولایه پنهان هستند. تعداد نورون‌های مخفی برای شبکه رمزگذار و شبکه رمزگشا ۸۰۰ است، درحالی‌که برای قسمت‌های باقیمانده ۱۰۰ است. تعداد لایه‌های مخفی باتوجه‌به تعداد نورون‌های ورودی و تعداد نورون‌های خروجی انتخاب می‌شود به‌طوری‌که پدیده اورفیت[3] در شبکه رخ نداده است. ReLU به جز آخرین لایه DNN ها و شبکه افزایش داده متخاصمی، به‌تمامی لایه‌ها اعمال شد. همچنین، از الگوریتم خاویر[4](۴۰) برای مقداردهی اولیه وزن شبکه DNN و از بهینه‌ساز آدام[5](۴۱) با نرخ یادگیری ۰/۰۰۰۱ برای آموزش آن‌ها استفاده شد. DNN ها با استفاده از Tensorflow 2.1 در نرم‌افزار پایتون پیاده‌سازی گردیدند.

یافته‌ها

جدول ۲ بالاترین میزان درصد WA و UAR را با روش‌های مختلف نشان می‌دهد. این نتایج با استفاده از مقادیر مختلف داده‌های افزوده شده، به‌دست آمده‌اند. این جدول نشان می‌دهد که پس از افزایش داده‌ها مبتنی بر روش افزایش داده متخاصمی، می‌توانیم شبکه DNN بهتری را برای تشخیص احساسات از گفتار آموزش دهیم. این عملکرد را می‌توان با استفاده از آموزش طبقه‌بند با نمونه‌های تقویت شده تولید شده از شبکه مولد متخاصمی واسراستین که بسیار بهتر از نتیجه به‌دست آمده از(۴۲) به وسیله کلاسه بند SVM با استفاده از ویژگی‌های دست‌ساز است، بهبود بخشید. UAR در این روش حتی از چن و همکاران(۴۳) که از CRNN های سه‌بعدی برای تولید ویژگی‌ها استفاده نمودند بالاتر است.

1. Berlin Database of Emotional Speech (EmoDB)
2. Fully Connected (FC) Neural Networks
3. Overfitting
4. Xavier Algorithm
5. Adam Optimizer

جدول۲) مقایسه نتایج تکنیک‌های مختلف افزایش داده و تشخیص احساس از گفتار

روش	کلاسه بند	WA%	UAR%
افزودن نویز	DNN	۸۲/۰۶	۸۰/۷۵
افزودن نویز	SVM	۸۱/۱۲	۸۰/۲۵
SMOTE	DNN	۸۲/۴۳	۸۱/۵۱
SMOTE	SVM	۸۰/۸۳	۷۹/۵۱
شبکه مولد متخاصمی سازگار با چرخه	DNN	۸۳/۵۵	۸۲/۲۰
شبکه مولد متخاصمی سازگار با چرخه	SVM	۸۱/۵۰	۸۰/۳۰
شبکه مولد متخاصمی سازگار با چرخه + فاصله واسراستین	DNN	۸۴/۴۹	۸۳/۳۳
شبکه مولد متخاصمی سازگار با چرخه + فاصله واسراستین	SVM	۸۱/۰۷	۸۰/۰۸
(2D- ACRNN (43	DNN	–	۷۹/۳۸
(3D- ACRNN (43	DNN	–	۸۲/۸۲

بحث

در این مقاله روشی جهت تشخیص احساس از گفتار ارائه گردید که می‌تواند در قالب یک اپلیکیشن تلفن همراه تغییرات احساسی روزانه در گفتار را تشخیص دهد و بر اساس آن‌ها، میزان سلامت روحی فرد را مشخص کند. برای مثال، اگر جملاتی که فرد به کار می‌برد، از یک الگوی منطقی پیروی نکنند، می‌توانند نشانه‌ای مهم از وجود اسکیزوفرنی باشند. کمبود داده می‌تواند از دستیابی به نتیجه مطلوب در آموزش شبکه عصبی عمیق جلوگیری کند که این مسئله یک مشکل جدی در تشخیص احساس از گفتار با استفاده از شبکه‌های عصبی عمیق است؛ بنابراین، یک شبکه جدید برای افزایش داده جهت تولید نمونه‌های مصنوعی ارائه شد که نمونه‌های تولید شده را در فضای نمونه‌های اصلی جای می‌دهد. به‌جای بردارهای حاوی ویژگی‌های احساس در فضای با ابعاد بالا، روش پیشنهادی می‌تواند فضایی از احساسات ایجاد کند که نمونه‌های ساخته شده را در فضای اصلی بازسازی کند. نتایج نشان می‌دهد که روش پیشنهادی می‌تواند بر مشکل کوچک‌شدن گرادیان در شبکه‌های مولد متخاصمی معمولی غلبه کند و نمونه‌های احساسی جدیدی تولید کند که با ترکیب آن‌ها با نمونه‌های اصلی برای تبدیل احساس از سیگنال گفتار مفید واقع شوند. تأثیر سایر تکنیک‌های افزایش داده، مانند تکثیر مشاهدات[1]، تبدیل داده‌ها و SMOTE، بر روی مدل‌های غیرخطی متفاوت است. SMOTE می‌تواند نمونه‌های مصنوعی قابل توجهی از داده‌های با کلاس‌های کم تولید کند. افزودن نویز به نمونه‌های اصلی هنگامی‌که نمونه‌های اصلی قابل‌تشخیص هستند می‌تواند کمک کند؛ اما هنگامی‌که با یکدیگر ادغام می‌شوند به عملکرد آسیب می‌رساند. شبکه افزایش داده متخاصمی پیشنهادی می‌توانند نمونه‌های ارزشمند و جدیدی ایجاد کنند که به بهبود تشخیص احساسات از گفتار با شبکه‌های عصبی عمیق کمک نمایند. در مقایسه با سایر تکنیک‌های افزایش داده، روش پیشنهاد شده می‌تواند عملکرد بهتری داشته باشد. مدل پیشنهادی می‌تواند بر مشکلات آموزش شبکه‌های مولد متخاصمی استاندارد غلبه کند. وجود دو ورودی مجزا در شبکه تشخیص‌دهنده به جلوگیری از کوچک‌شدن تدریجی گرادیان و عدم تعادل در داده‌های آموزش کمک می‌کند. استراتژی تولید نمونه‌های تقلبی احساس در فضای ویژگی نمونه‌های اصلی و به دنبال آن بازسازی نمونه‌ها در فضای اصلی، شبکه مولد را در فریب شبکه تشخیص‌دهنده کمک می‌کند. همچنین اطمینان حاصل می‌شود که نمونه‌های تولید شده می‌توانند توزیع واقعی داده‌ها را دنبال کنند.

تشکر و قدردانی

به‌عنوان نویسنده مقاله، تمایل دارم از تلاش‌های بی‌دریغ و راهنمایی‌های ارزشمند و حمایتی که از سوی آقای دکتر نادر سلیمانی، دانشیار گروه مدیریت آموزشی دانشگاه آزاد اسلامی واحد گرمسار، در هنگام تدوین مقاله‌ام دریافت کردم، صمیمانه تشکر و قدردانی کنم.

1. Duplicating the Observations

منابع

1. A. Belouali1, S. Gupta, V. Sourirajan, N. Allen, and A. Alaoui, "Acoustic and language analysis of speech for suicidal ideation among US veterans," in BioData Mining., 2021, pp. 1–17.
2. AM. Chekroud, RJ. Zotti, Z. Shehzad, R. Gueorguieva, and MK. Johnson, "Cross-trial prediction of treatment outcome in depression: a machine learning approach," in The Lancet Psychiatry., 2016, pp. 243–250.
3. L. Breiman, Classification and Regression Trees. CRC Press, 2017.
4. J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, 2009.
5. M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
6. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Mar. 2003.
7. T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," 2017, arXiv:1702.05538. [Online]. Available: https://arxiv.org/abs/1702.05538
8. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. "Generative adversarial nets". In: Advances in neural information processing systems. 2014.
9. F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in Proc. 21st ACM Int. Conf. Multimedia MM, 2013, pp. 835–838.
10. J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, 2009.
11. S. Chandaka, A. Chatterjee, and S. Munshi, "Support vector machines employing cross-correlatign for emotional speech recognition," Measurement, vol. 42, no. 4, pp. 611–618, 2009.
12. H. Altun and G. Polat, "Boosting selection of speech-related features to improve performance of multi-class SVMs in emotion detection," Expert Syst. Appl., vol. 36, no. 4, pp. 8197–8203, 2009.
13. D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," Speech Commun., vol. 52, no. 7, pp. 613–625, 2010.
14. B. Yang and M. Lugger, "Emotion recognition from speech signals using new. harmony features," Signal Processing, vol. 90, no. 5, pp. 1415–1423, 2010.
15. T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, "Anger recognition in speech using acoustic and linguistic cues," Speech Commun., vol. 53, no. 9, pp. 1198–1209, 2011.
16. M. Abdelwahab and C. Busso, "Study of Dense Network Approaches for Speech Emotion Recognition," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5084–5088.
17. M. Kockmann, L. Burget, and J. "Honza" Černocký, "Application of speaker- and language identification state-of-the-art techniques for emotion recognition," Speech Commun., vol. 53, no. 9, pp. 1172–1185, 2011.
18. E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," Computa Speech Lang., voly 25, no. 3, pp. 556–570, 2011.
19. E. Bozkurt, E. Erzin, Ç. E. Erdem, and A. T. Erdem, "Formant position based weighted spectral features for emotion recognition," Speech Commun., vol. 53, no. 9, pp. 1186–1197, 2011.
20. S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," Speech Commun., vol. 53, no. 5, pp. 768–785, 2011.
21. P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, and K. Elenius, "Expression of effect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation," Comput. Speech Lang., vol. 25, no. 1, pp. 84–104, 2011.
22. H. Pérez-Espinosa, C. A. Reyes-García, and L. Villaseñor-Pineda, "Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model," Biomed. Signal Process. Control, vol. 7, no. 1, pp. 79–87, 2012.
23. K. Han, D. Yu, and I. Tashev, Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. 2014.
24. H. Palo and M. Mohanty, "Modified-VQ Features for Speech Emotion Recognition," J. Appl. Sci., vol. 16, pp. 406–418, Sep. 2016.
25. B. Schuller, R. Müller, M. Lang, and G. Rigoll, Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. 2005.
26. I. Luengo, E. Navas, and I. Hernáez, "Feature Analysis and Evaluation for Automatic Emotion Identification in Speech," Multimedia, IEEE Trans., vol. 12, pp. 490–501, Nov. 2010.
27. D. Gharavian, M. Sheikhan, and F. Ashoftedel, "Emotion recognition improvement using normalized formant supplementary features by a hybrid of DTW-MLP-GMM model," Neural Comput. Appl., vol. 22, no. 6, pp. 1181–1191, 2013.
28. X. Zhao, S. Zhang, and B. Lei, "Robust emotion recognition in anoisy speech via sparse representation," Neural Comput. Appl., vol. 24, Jun. 2013.
29. H. Hu, T. Tan, and Y. Qian, "Generative adversarial network-based data augmentation for

noise-robust speech recognition," in Proc. IEEE Int. Conf. Acoustic., Speech Signal Process. (ICASSP), Apr. 2018, pp. 5044–5048.
30. J– Chang, S. Scherer. "Learning representations of emotional speech with deep convolutional generative adversarial networks". In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2017.
31. Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller, "Snore- GANs: Improving automatic snore sound classification with synthesized data," IEEE J. Biomed. Health Information., vol. 24, no. 1, pp. 300–310, Jan. 2020.
32. B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in Proc. Interspeech, Sep. 2011, pp. 3201–3204.
33. B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in Proc. Interspeech, Sep. 2009, pp. 312–315.
34. B. S.h uller et al., "The INTERSPEECH 2010 paralinguistic challenge," in Proct Interspeech, Sep. 2010, pp. 2794–2797.
35. J. R. Deller Jr, J. G. Proakis, and J. H. L. Hansen, Discrete-time Processing of Speech Signals. Basingstoke, U.K.: Macmillan Pub, 1993.
36. M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in Proc. Int. Conf. Mach. Learn., Aug. 2017, pp. 214–223.
37. J. Wu, Z. Huang, J. Thoma, D. Acharya, and L. Van Gool, "Wasserstein divergence for GANs," in Proc. Eur. Conf. Compute. Vis. (ECCV), Sep. 2018, pp. 653–668.
38. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in Proc. Adv. Neural Inf. Process. Syst., I. Guyon, U. V. Lux burg, S. Bagnio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. New York, NY, USA: Curran Associates, 2017, pp. 5767–5777.
39. F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in Proc. 9th Eur. Conf. Speech Commun. Technol., 2005, pp. 1–4.
40. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proc. 13th Int. Conf. Artif. Intel. Statist., 2010, pp. 249–256.
41. D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conftm Learn. Represent. (ICLR), 2015, pp.1–15.
42. I. Luengo, E. Navas, and I. Hernaez, "Feature analysis and evaluation for automatic emotion identification in speech," IEEE Trans. Multimedia, vol. 12, no. 6, pp. 490–501, Oct. 2010.
43. M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
44. A. Shilandari, H. Marvi, H. Khosravi and W. Wang "Speech emotion recognition using data augmentation method by cycle-generative adversarial networks," in Journal of Signal, Image, and Video Processing., DOI: https://doi.org/10.1007/s11760-022-02156-9, 2022.