درس:

پردازش سیگنال‌های دیجیتال

موضوع p24:

Introducing Translatotron: An End-to-End Speech-to-Speech Translation Model

استاد :

جناب آقای دکتر مهدی اسلامی

دانشجو:

حمیدرضا پورمحمّد

شماره دانشجویی :

۴۰۰۱۴۱۴۰۱۱۱۰۳۳

پاییز ۱۴۰۱

- **ما یک مشکل کم کارایی سالمندان در تشخیص خودکار گفتار (ASR) را از طریق تطبیق ویژگی‌های آگنوستیک با ASR بررسی می‌کنیم. بیشتر مجموعه داده‌های مدل‌های تشخیص گفتار از مجموعه داده‌های جمع‌آوری‌شده از سخنرانان بزرگسال تشکیل شده‌اند. در نتیجه، اکثر سیستم‌های تشخیص گفتار تجاری معمولاً روی سخنرانان بزرگسال عملکرد خوبی دارند. به عبارت دیگر، تنوع محدود سخنرانان در مجموعه داده‌های آموزشی، عملکرد غیرقابل اعتمادی را برای سخنرانان اقلیت (به عنوان مثال، افراد مسن) به دلیل دستیابی غیرممکن از داده‌های آموزشی ایجاد می‌کند. در پاسخ، این مقاله یک چارچوب تبدیل صدا مبتنی بر شبکه عصبی را برای تقویت تشخیص گفتار اقلیت پیشنهاد می‌کند. برای این منظور، ما یک مدل ترجمه صوتی شامل یک خوشه‌بندی واج‌شناسی بدون نظارت برای استخراج اطلاعات زبانی برای گفتار اقلیت در چارچوب مدل آکوستیک فعلی پیشنهاد می‌کنیم. پیشنهاد ما یک روش انطباق ویژگی طیفی است که می تواند در مقابل هر سیستم ASR تجاری یا باز قرار گیرد و از تغییر مستقیم تشخیص دهنده گفتار اجتناب شود. نتایج تجربی و تجزیه و تحلیل اثربخشی روش پیشنهادی ما را از طریق بهبود دقت تشخیص گفتار سالمندان نشان می‌دهد.**
- **شکل جدیدی از دستگاه ارتباطی تقویتی و جایگزین (AAC) برای افراد مبتلا به اختلال گفتاری شدید – کمک ارتباطی صدا خروجی صدای ورودی (VIVOCA) – شرح داده شده است. VIVOCA گفتار بی نظم کاربر را تشخیص می دهد و پیام هایی را ایجاد می کند که به گفتار مصنوعی تبدیل می شوند. توسعه سیستم با استفاده از روش‌های طراحی و توسعه کاربر محور انجام شد که الزامات کلیدی دستگاه را شناسایی و اصلاح کرد. یک روش جدید برای ساخت واژگان کوچک، تشخیص‌دهنده گفتار خودکار وابسته به سخنران با مقادیر کمتر داده‌های آموزشی، استفاده شد. آزمایش‌ها نشان داد که این روش در ایجاد عملکرد تشخیص خوب (متوسط دقت ۹۶ درصد) در گفتار بسیار بی‌نظم، حتی زمانی که گیجی در تشخیص افزایش می‌یابد، موفق است. تکنیک پیام‌سازی انتخابی عوامل مختلفی از جمله سرعت ساخت پیام و محدوده خروجی‌های پیام موجود را کاهش داد. VIVOCA در یک کارآزمایی میدانی توسط افراد مبتلا به دیزآرتری متوسط تا شدید مورد ارزیابی قرار گرفت و تأیید شد که می‌توانند از این دستگاه برای تولید خروجی گفتار قابل فهم از ورودی گفتار اختلال استفاده کنند. این کارآزمایی برخی مسائل را برجسته کرد که عملکرد و قابلیت استفاده دستگاه را در شرایط استفاده واقعی محدود می‌کند، با میانگین دقت تشخیص ۶۷ درصد در این شرایط. این محدودیت ها در کارهای آینده بررسی خواهند شد.**
- **در این مقاله، ما یک سیستم تشخیص گفتار سرتاسر برای افراد ژاپنی مبتلا به اختلالات بیانی ناشی از فلج مغزی آتتوید ارائه می‌کنیم. از آنجایی که بیان آنها اغلب ناپایدار یا نامشخص است، سیستم‌های تشخیص گفتار برای تشخیص گفتارشان تلاش می‌کنند. رویکردهای مبتنی بر یادگیری عمیق اخیر عملکرد امیدوارکننده ای را نشان داده اند. با این حال، این رویکردها به حجم زیادی از داده‌های آموزشی نیاز دارند و جمع‌آوری داده‌های کافی از چنین افراد دیزترمی دشوار است. این مقاله یک روش یادگیری انتقالی را پیشنهاد می‌کند که دو نوع دانش متناظر با مجموعه داده‌های مختلف را منتقل می‌کند: ویژگی وابسته به زبان (آوایی و زبانی) گفتار بدون اختلال و ویژگی مستقل از زبان گفتار دیزآرتریک. اولی از داده‌های گفتار غیر دیزارتریک ژاپنی و دومی از داده‌های گفتار دیزآرتریک غیر ژاپنی به‌دست می‌آید. در روش پیشنهادی، مدلی را با استفاده از گفتار دیزآرتریک ژاپنی و گفتار دیزآرتریک غیرژاپنی از قبل آموزش**

می‌دهیم و پس از آن، مدل را با استفاده از گفتار دیزآرتریک هدف ژاپنی تنظیم می‌کنیم. برای رسیدگی به داده‌های گفتاری دو زبان مختلف در یک مدل، از ماژول‌های رمزگشای خاص زبان استفاده می‌کنیم. نتایج تجربی نشان می‌دهد که رویکرد پیشنهادی ما می‌تواند به طور قابل‌توجهی عملکرد تشخیص گفتار را در مقایسه با سایر رویکردهایی که از داده‌های گفتاری اضافی استفاده نمی‌کنند، بهبود بخشد.

- تحریک الکتریکی و صوتی ترکیبی (EAS) تشخیص گفتار بهتری را نسبت به کاشت حلزون معمولی (CI) نشان داده است و عملکرد رضایت بخشی را در شرایط ساکت به همراه دارد. با این حال، هنگامی که سیگنال‌های نویز درگیر می‌شوند، سیگنال الکتریکی و سیگنال صوتی ممکن است مخدوش شوند، در نتیجه عملکرد تشخیص ضعیفی را به همراه دارد. برای سرکوب اثرات نویز، تقویت گفتار (SE) یک واحد ضروری در دستگاه‌های EAS است. اخیراً، یک الگوریتم تقویت گفتار حوزه زمان مبتنی بر شبکه‌های عصبی کاملاً کانولوشن (FCN) با تابع هدف مبتنی بر درک هدف کوتاه‌مدت (STOI) (به‌طور خلاصه FCN(S) نامیده می‌شود) به دلیل ساده بودن توجه فزاینده‌ای را به خود جلب کرده است. ساختار و اثربخشی بازیابی سیگنال های گفتاری تمیز از همتایان پر سر و صدا. با شواهدی که مزایای FCN(S) را برای گفتار عادی نشان می دهد، این مطالعه به ارزیابی توانایی آن در بهبود درک گفتار شبیه سازی شده EAS می پردازد. ارزیابی‌های عینی و آزمون‌های شنیداری برای بررسی عملکرد FCN(S) در بهبود درک گفتار گفتار عادی و صدادار در محیط‌های پر سر و صدا انجام شد. نتایج تجربی نشان می‌دهد که در مقایسه با روش سنتی حداقل میانگین مربع خطای SE و روش رمزگذاری خودکار SE حذف نویز عمیق، FCN(S) می‌تواند در فهم گفتار برای گفتار عادی و همچنین صداگذاری شده به دست آورد. این مطالعه، که اولین موردی است که رویکردهای یادگیری عمیق SE را برای EAS ارزیابی می‌کند، تأیید می‌کند که FCN(S) یک رویکرد SE موثر است که ممکن است به طور بالقوه در یک پردازنده EAS ادغام شود تا برای کاربران در محیط‌های پر سر و صدا مفید باشد.
- این مقاله احساسات گفتار و تشخیص طبیعی بودن را با استفاده از مدل‌های یادگیری عمیق با رویکردهای یادگیری چند وظیفه‌ای و یادگیری تک وظیفه‌ای ارزیابی می‌کند. مدل هیجانی ویژگی‌های ظرفیت، برانگیختگی و تسلط را که به عنوان احساسات بعدی شناخته می‌شوند، در خود جای می‌دهد. درجه‌بندی‌های طبیعی بودن در مقیاس نقطه‌ای ve به عنوان احساسات بعدی برچسب‌گذاری می‌شوند. یادگیری چندوظیفه ای هر دو عاطفه بعدی (به عنوان وظیفه اصلی) و نمرات طبیعی بودن (به عنوان یک کار کمکی) را به طور همزمان پیش بینی می کند. یادگیری تک تکلیفی یا هیجان بعدی (ظرفیت، برانگیختگی و تسلط) یا امتیاز طبیعی بودن را به طور مستقل پیش بینی می کند. نتایج با یادگیری چند وظیفه ای، بهبودی را نسبت به مطالعات قبلی در مورد یادگیری تک تکلیف برای تشخیص عواطف بعدی و پیش بینی طبیعی بودن نشان می دهد. در این مطالعه، یادگیری تک وظیفه ای هنوز نسبت به یادگیری چند وظیفه ای برای تشخیص طبیعی بودن برتری نشان می دهد. نمودارهای پراکنده نمرات پیش‌بینی احساسات و طبیعی بودن در برابر برچسب‌های واقعی در یادگیری چندکاره، فقدان مدل را نشان می‌دهد. نمی تواند نمرات پایین و بسیار بالا را پیش بینی کند. امتیاز پایین پیش‌بینی طبیعی بودن در این مطالعه احتمالاً به دلیل تعداد کم نمونه‌های نمونه گفتار غیرطبیعی است زیرا مجموعه داده MSP-IMPROV طبیعی بودن گفتار را ارتقا می‌دهد. این نتیجه که پیش‌بینی مشترک طبیعی بودن با احساسات به بهبود

عملکرد تشخیص احساسات کمک می‌کند، ممکن است در مدل تشخیص هیجان در کارهای آینده تجسم یابد.

- اثر لومبارد یکی از شناخته شده ترین اثرات نویز بر تولید گفتار است. گفتار با جلوه لومبارد در محیط های پر سر و صدا راحت تر از گفتار طبیعی معمولی قابل تشخیص است. تحقیقات قبلی ما نشان داد که مدل‌های سنتز گفتار ممکن است ویژگی‌های اثر لومبارد را حفظ کنند. در این مطالعه، ما چندین مدل گفتار، مانند هارمونیک، منبع و سینوسی را که برای گفتار لومبارد در زمینه تقویت گفتار اعمال می‌شوند، بررسی می‌کنیم. برای این منظور از ۱۰۰ گفتار طبیعی و ۱۰۰ با القای اثر لومبارد استفاده می شود. هدف این مطالعه بررسی این است که تا چه حد گفته‌های گفتاری بر اساس این مدل‌ها قابل تشخیص هستند و در چه آستانه سطح SNR (نسبت سیگنال به نویز) یک مدل خاص کار نمی‌کند. برای این منظور، مدل‌های سنتز شده و گفتار لومبارد با صداهای حماسی و ضبط صدای خیابان با SNRهای مختلف مخلوط می‌شوند. کیفیت این مدل ها با استفاده از شاخص های عینی و همچنین آزمون های ذهنی اندازه گیری می شود. از آنجایی که هیچ معیار استانداردی برای اعمال در گفتار تقویت‌شده وجود ندارد، یک معیار عینی برای ارزیابی کیفیت گفتار مدلی که ویژگی‌های گفتار لومبارد را ترکیب می‌کند، بر اساس یک بردار ویژگی، پیشنهاد شده‌است. سپس رویکرد ما با معیار استاندارد مورد استفاده در مخابرات و همچنین با نتایج آزمون ذهنی مقایسه می‌شود. بررسی‌های تجربی برتری مدل‌های منبع را نشان می‌دهد که برای سنتز گفتار لومبارد نسبت به سایر مدل‌های مورد استفاده استفاده می‌شود. همچنین، معیار پیشنهادی با نتایج ارزیابی ذهنی بیشتر از نتایج توصیه ITU-T P.563 مرتبط است. این با تجزیه و تحلیل آماری ANOVA بررسی شد.
- سیستم‌های اخیر ترکیب متن به گفتار (TTS) با موفقیت گفتار با کیفیت بالا را سنتز کرده‌اند. با این حال، درک گفتار TTS در محیط‌های پر سر و صدا کاهش می‌یابد زیرا اکثر این سیستم‌ها برای مدیریت محیط‌های پر سر و صدا طراحی نشده‌اند. چندین کار سعی کردند این مشکل را با استفاده از تنظیم دقیق آفلاین برای تطبیق TTS خود با شرایط پر سر و صدا برطرف کنند. برخلاف ماشین‌ها، انسان‌ها هرگز تنظیم دقیق آفلاین را انجام نمی‌دهند. در عوض، آن‌ها در مکان‌های پر سر و صدا با اثر لومبارد صحبت می‌کنند، جایی که به‌طور پویا تلاش‌های صوتی خود را برای بهبود شنیداری گفتارشان تنظیم می‌کنند. این توانایی توسط مکانیسم زنجیره گفتار پشتیبانی می شود که شامل بازخورد شنیداری از درک گفتار به تولید گفتار می شود. این مقاله یک رویکرد جایگزین برای TTS در محیط‌های پر سر و صدا پیشنهاد می‌کند که به اثر لومبارد انسانی نزدیک‌تر است. به طور خاص، ما Lombard TTS را در چارچوب زنجیره گفتار ماشینی پیاده سازی می کنیم تا گفتار را با سازگاری پویا ترکیب کنیم. TTS ما سازگاری را با تولید گفتار بر اساس بازخورد شنیداری انجام می‌دهد که شامل از دست دادن تشخیص خودکار گفتار (ASR) به عنوان معیار درک گفتار و پیش‌بینی نسبت گفتار به نویز (SNR) به عنوان اندازه‌گیری توان است. دو نسخه TTS مورد بررسی قرار می‌گیرند: TTS غیر افزایشی با بازخورد در سطح بیان و TTS افزایشی (ITTS) با بازخورد کوتاه‌مدت برای کاهش تاخیر بدون از دست دادن عملکرد قابل‌توجه. علاوه بر این، ما سیستم های TTS را در هر دو شرایط نویز استاتیک و پویا ارزیابی می کنیم. نتایج تجربی ما نشان می‌دهد که بازخورد شنیداری درک گفتار TTS را در نویز افزایش می‌دهد.

- فناوری گفتار کمکی به دلیل ماهیت آسیب‌دیده گفتار دیزآرتریک، مانند صدای نفس‌گیر، گفتار تیره، مصوت‌های مخدوش و صامت‌ها، یک کار چالش برانگیز است. یادگیری تعبیه‌های فشرده و متمایز برای گفتارهای دیزآرتریک برای اختلال در تشخیص گفتار ضروری است. ما یک رویکرد مبتنی بر هیستوگرام حالت‌ها (HoS) را پیشنهاد می‌کنیم که از مدل مارکوف پنهان-شبکه عصبی عمیق (DNN-HMM) برای یادگیری جاسازی‌های فشرده و متمایز مبتنی بر شبکه استفاده می‌کند. بهترین توالی حالت انتخاب شده از شبکه کلمه برای نشان دادن گفتار دیزارتریک استفاده می‌شود. . سپس یک طبقه‌بندی‌کننده مبتنی بر مدل متمایز برای تشخیص این تعبیه‌ها استفاده می‌شود. عملکرد رویکرد پیشنهادی با استفاده از سه مجموعه داده، یعنی ۱۵ کلمه مشابه صوتی، مجموعه داده های ۱۰۰ کلمه ای رایج پایگاه داده UA-SPEECH، و مجموعه داده ۵۰ کلمه ای پایگاه داده TORGO ارزیابی می شود. رویکرد مبتنی بر HoS پیشنهادی به طور قابل توجهی بهتر از مدل سنتی مارکوف پنهان و DNNHMM- عمل می کند. رویکردهای مبتنی بر هر سه مجموعه داده توانایی تشخیص و فشرده بودن تعبیه‌های مبتنی بر HoS پیشنهادی منجر به بهترین دقت تشخیص گفتار مختل می‌شود.
- این بررسی وضعیت تحقیق رابط گفتار خاموش (SSI) را خلاصه می کند. SSI ها بر سیگنال های زیستی غیر صوتی تولید شده توسط بدن انسان در طول تولید گفتار تکیه می کنند تا هر زمان که ارتباط کلامی معمولی امکان پذیر نباشد یا مطلوب نباشد، امکان برقراری ارتباط را فراهم می کند. در این بررسی، ما بر اولین مورد تمرکز می‌کنیم و آخرین تحقیقات SSI را با هدف ارائه روش‌های ارتباطی جایگزین و تقویتی جدید برای افراد مبتلا به اختلالات گفتاری شدید ارائه می‌کنیم. SSI ها می توانند سیگنال های زیستی مختلفی را برای برقراری ارتباط بی صدا به کار گیرند، مانند ضبط الکتروفیزیولوژیک فعالیت عصبی، ضبط الکترومیوگرافی (EMG) حرکات دستگاه صوتی یا ردیابی مستقیم حرکات مفصل با استفاده از تکنیک های تصویربرداری. بسته به اختلال، برخی از تکنیک‌های حسی ممکن است برای گرفتن اطلاعات مربوط به گفتار مناسب‌تر از دیگران باشند. به عنوان مثال، روش‌های EMG و تصویربرداری برای بیماران حنجره‌برداری شده، که مجرای صوتی آن‌ها تقریباً دست نخورده باقی می‌ماند، اما پس از برداشتن تارهای صوتی قادر به صحبت کردن نیستند، اما برای افراد فلج شدید ناموفق هستند، مناسب هستند. از سیگنال‌های زیستی، SSIها پیام مورد نظر را با استفاده از الگوریتم‌های تشخیص خودکار گفتار یا سنتز گفتار رمزگشایی می‌کنند. علیرغم پیشرفت های قابل توجه در سال های اخیر، اکثر SSI های امروزی فقط در تنظیمات آزمایشگاهی برای کاربران سالم تایید شده اند. بنابراین، همانطور که در این مقاله مورد بحث قرار گرفت، تعدادی از چالش‌ها باقی مانده است که در تحقیقات آینده قبل از ارتقای SSIs به برنامه‌های کاربردی در دنیای واقعی، باید مورد توجه قرار گیرند. اگر بتوان با موفقیت به این مسائل رسیدگی کرد، SSI های آینده با بازگرداندن قابلیت های ارتباطی، زندگی افراد مبتلا به اختلالات گفتاری شدید را بهبود خواهند بخشید.
- نویزهای محیطی می تواند تهدیدی برای عملکرد پایدار سیستم های تشخیص گفتار فعلی باشد. بنابراین، ایجاد مجموعه ای از ویژگی های جلویی که قادر به شناسایی گفتار در نسبت سیگنال به نویز کم باشد، ضروری است. در این مقاله، یک ویژگی همجوشی قوی پیشنهاد شده است که می تواند اطلاعات گفتار را به طور کامل مشخص کند. برای به دست آوردن ضرایب مغزی گوش حلزونی (CFCC)، یک ویژگی جدید

ابتدا توسط تابع غیرخطی قانون قدرت استخراج می‌شود که می‌تواند ویژگی‌های شنوایی گوش انسان را شبیه‌سازی کند. سپس فناوری بهبود گفتار به قسمت جلویی استخراج ویژگی معرفی می‌شود و ویژگی استخراج‌شده و تفاوت مرتبه اول آن‌ها در ویژگی‌های ترکیبی جدید ترکیب می‌شوند. یک ویژگی انرژی ضریب مغزی اپراتور انرژی (Teager (TEOCC نیز استخراج می‌شود و با ویژگی‌های ترکیبی فوق الذکر ترکیب می‌شود تا مجموعه ویژگی‌های همجوشی را تشکیل دهد. سپس تجزیه و تحلیل مؤلفه اصلی (PCA) برای انتخاب ویژگی و بهینه‌سازی مجموعه ویژگی اعمال می‌شود، و مجموعه ویژگی‌های ملی در افراد غیر خاص، کلمات جدا شده و سیستم تشخیص گفتار کوچک واژگانی استفاده می‌شود. در نهایت، یک آزمایش مقایسه ای از تشخیص گفتار برای تأیید مزایای مجموعه ویژگی پیشنهادی با استفاده از یک ماشین بردار پشتیبانی (SVM) طراحی شده است. نتایج تجربی نشان می‌دهد که مجموعه ویژگی پیشنهادی نه تنها نرخ تشخیص بالا و عملکرد ضد نویز عالی را در تشخیص گفتار نشان می‌دهد، بلکه می‌تواند اطلاعات شنیداری و انرژی را در سیگنال‌های گفتار به طور کامل مشخص کند.

- **سنتز گفتار راه درازی را پیموده است، زیرا مدل‌های فعلی تبدیل متن به گفتار (TTS) اکنون می‌توانند گفتار طبیعی با صدای انسان را تولید کنند. با این حال، بیشتر تحقیقات TTS بر استفاده از داده‌های گفتار بزرگسالان متمرکز است و کار بسیار محدودی روی سنتز گفتار کودکان انجام شده است. این مطالعه یک خط لوله آموزشی برای مدل‌های پیشرفته TTS عصبی تنظیم شبکه (SOTA) با استفاده از مجموعه داده‌های گفتار کودک ایجاد و تأیید کرد. این رویکرد یک کار تنظیم مجدد TTS چند بلندگو را برای ارائه یک خط لوله یادگیری انتقالی اتخاذ می کند. یک مجموعه داده گفتار کودک در دسترس عموم پاک شد تا زیرمجموعه‌ای کوچک‌تر از تقریباً ۱۹ ساعت ارائه شود که اساس آزمایش‌های تنظیم شبکه ما را تشکیل داد. هر دو ارزیابی ذهنی و عینی با استفاده از یک MOSNet از پیش آموزش دیده برای ارزیابی عینی و یک چارچوب ذهنی جدید برای ارزیابی‌های میانگین امتیاز نظر (MOS) انجام شد. ارزیابی های ذهنی به MOS 3.95 برای درک گفتار، ۳.۸۹ برای طبیعی بودن صدا و ۳.۹۶ برای سازگاری صدا دست یافتند. ارزیابی عینی با استفاده از یک MOSNet از پیش آموزش دیده همبستگی قوی بین صدای واقعی و مصنوعی کودک را نشان داد. شباهت گوینده نیز با محاسبه شباهت کسینوس بین جاسازی‌های گفته‌ها تأیید شد. یک مدل تشخیص خودکار گفتار (ASR) نیز برای ارائه مقایسه نرخ خطای کلمه (WER) بین صدای واقعی و مصنوعی کودک استفاده می‌شود. مدل TTS آموزش‌دیده ملی قادر به سنتز گفتار کودکانه از نمونه‌های صوتی مرجع به‌مدت ۵ ثانیه بود.**
- **بسیاری از آثار بر روی الگوریتم های تشخیص احساسات گفتار متمرکز شده اند. با این حال، بیشتر بر انتخاب مناسب ویژگی‌های صوتی گفتار تکیه می‌کنند. در این مقاله، ما یک الگوریتم جدید تشخیص احساسات را پیشنهاد می‌کنیم که بر هیچ ویژگی صوتی گفتاری تکیه نمی‌کند و اطلاعات جنسیت گوینده را ترکیب می‌کند. هدف ما این است که از اطلاعات غنی از داده‌های خام گفتار، بدون هیچ مداخله‌ای مصنوعی بهره ببریم. به طور کلی، سیستم های تشخیص احساسات گفتار به انتخاب دستی ویژگی های آکوستیک سنتی مناسب به عنوان ورودی طبقه بندی شده برای تشخیص احساسات نیاز دارند. با استفاده از الگوریتم های یادگیری عمیق، شبکه به طور خودکار اطلاعات مهم را از سیگنال گفتار خام برای لایه طبقه بندی انتخاب می کند تا تشخیص احساسات را انجام دهد. می‌تواند از حذف اطلاعات احساسی که**

نمی‌توانند مستقیماً به‌عنوان مشخصه صوتی گفتار مدل‌سازی شوند، جلوگیری کند. ما همچنین اطلاعات جنسیت گوینده را به الگوریتم پیشنهادی اضافه می‌کنیم تا دقت تشخیص را بیشتر بهبود ببخشیم. الگوریتم پیشنهادی یک شبکه عصبی کانولوشنال باقیمانده (R-CNN) و یک بلوک اطلاعات جنسیتی را ترکیب می‌کند. داده های گفتاری خام به طور همزمان به این دو بلوک ارسال می شود. شبکه R-CNN اطلاعات احساسی لازم را از داده های گفتاری به دست می آورد و مقوله احساسی را طبقه بندی می کند. الگوریتم پیشنهادی بر روی سه پایگاه داده عمومی با سیستم‌های زبانی مختلف ارزیابی می‌شود. نتایج تجربی نشان می دهد که الگوریتم پیشنهادی به ترتیب ۵.۶٪، ۷.۳٪ و ۱.۵٪ بهبودهایی در دقت در زبان ماندارین، انگلیسی و آلمانی در مقایسه با الگوریتم های با بالاترین دقت موجود دارد. به منظور تأیید تعمیم الگوریتم پیشنهادی، از پایگاه‌های اطلاعاتی FAU و eNTERFACE استفاده می‌کنیم، در این دو پایگاه داده مستقل، الگوریتم پیشنهادی می‌تواند به ترتیب به دقت ۸۵.۸ و ۷۱.۱ درصد نیز دست یابد.

- در فرآیند زمانی بیان احساسات، برخی فواصل، اطلاعات هیجانی برجسته تری را نسبت به سایرین در خود جای می دهند. در این مقاله، با معرفی یک ماژول ادغام زمانی دقیق در معماری شبکه عصبی عمیق (DNN)، یک چارچوب ساده اما موثر تشخیص احساسات گفتار (SER) ارائه می‌کنیم که قادر است به طور خودکار بخش‌های برجسته احساسی را در حالی که سرکوب می‌کند، برجسته کند. وجود موارد کمتر مرتبط برای یک گفتار ورودی، توالی ویژگی استخراج شده توصیفگرهای سطح پایین (LLD) به طور مساوی به چندین بخش زمانی با هم همپوشانی تقسیم می‌شوند و ویژگی‌های سطح قطعه با انجام عملکردها بر روی LLDهای هر بخش محاسبه می‌شوند. سپس این ویژگی‌های سطح بخش وارد یک مدل DNN می‌شوند که احتمالات احساسات و همچنین نمایش فشرده‌تر هر بخش را خروجی می‌دهد. یک ماژول ادغام زمانی دقیق، متشکل از یک DNN کمکی و یک مدل مخلوط گاوسی (GMM)، برای یادگیری وزن‌های برجستگی احساسی بخش‌های زمانی مختلف از نمایش‌های متراکم پیشنهاد شده است، که سپس به احتمالات احساسی در سطح بخش برای پیش بینی سطح گفتار ملی قابل‌توجه، ماژول ادغام زمانی دقیق و معماری DNN برای انتزاع ویژگی‌ها را می‌توان به طور مشترک تنها با استفاده از برچسب‌های سطح بیان آموزش داد، در حالی که بدون هیچ گونه اطلاعات نظارتی در سطح قاب یا سطح بخش. نتایج تجربی روی سه مجموعه داده احساسات منتشر شده عمومی RML، EMO-DB، و IEMOCAP نشان می‌دهد که چارچوب پیشنهادی عملکرد پیشرفته‌ای را در SER به دست می‌آورد.
- روش‌های فیزیولوژیکی و روان‌فیزیکی امکان بررسی گسترده مسیرهای عصبی صعودی (آوران) از گوش به مغز در پستانداران و نقش آن‌ها در افزایش سیگنال‌ها در نویز را فراهم می‌کنند. با این حال، علاقه بیشتری به نزولی (وابران) برهای عصبی در مسیر شنوایی پستانداران وجود دارد. این مسیر وابران از طریق سیستم olivocochlear عمل می کند، پردازش شنوایی را با عصب دهی حلزون اصلاح می کند و توانایی انسان را برای تشخیص صداها در پس زمینه های پر سر و صدا افزایش می دهد. درک موثر گفتار ممکن است به یک تعامل پیچیده بین ثابت‌های زمانی وابران و انواع نویز پس‌زمینه بستگی داشته باشد. در این مطالعه، یک مدل شنوایی با پردازش الهام‌گرفته از وابران، قسمت جلویی یک سیستم تشخیص خودکار گفتار (ASR) را فراهم کرد، که به عنوان ابزاری برای ارزیابی تشخیص گفتار با تغییرات ثابت‌های زمانی (۵۰ تا ۲۰۰۰ میلی‌ثانیه) و استفاده می‌شود. نوع نویز پس زمینه (نویز مدوله نشده و مدوله شده). با

فعال‌سازی وابران، حداکثر بهبود تشخیص گفتار (برای هر دو نوع نویز) برای نسبت سیگنال به نویز در حدود ۱۰ دسی‌بل، مشخصه موقعیت‌های گوش دادن به گفتار در دنیای واقعی، رخ داد. بهبود گفتار خالص به دلیل فعال سازی وابران (NSIEA) در نویز مدوله شده کمتر از نویز تعدیل نشده بود. برای نویز بدون تعدیل، NSIEA با افزایش زمان ثابت افزایش یافت. برای نویز مدوله‌شده، NSIEA برای ثابت‌های زمانی تا ۲۰۰ میلی‌ثانیه افزایش یافت، اما برای ثابت‌های زمانی طولانی‌تر، مطابق با زمان‌های مدولاسیون پاکت گفتار که برای تشخیص گفتار در نویز مدوله‌شده مهم است، مشابه باقی ماند. این مدل درک ما را از فعل و انفعالات پیچیده مربوط به تشخیص گفتار در نویز بهبود می بخشد و می تواند برای شبیه سازی مشکلات درک گفتار در نویز به عنوان یک نتیجه از انواع مختلف کاهش شنوایی استفاده شود.

- مزایای تشخیص نقطه پایانی گفتار (EPD) از ویژگی‌های حالت رمزگشا (DSF) سیستم تشخیص خودکار گفتار آنلاین (ASR). با این حال، DSFها از طریق فرآیند رمزگشایی ASR به دست می‌آیند، که می‌تواند بسیار گران باشد، به‌ویژه در سناریوهای با منابع محدود مانند دستگاه‌های تعبیه‌شده. برای پرداختن به این مشکل، این مقاله یک پیش‌بینی‌کننده پایان بیان (EOU) مبتنی بر مدل زبان (LM) را پیشنهاد می‌کند که برای تعیین احتمالات چارچوبی نشانه EOU مشروط به تاریخچه کلمه قبلی به‌دست‌آمده از ۱-best آموزش داده شده است. فرضیه رمزگشایی سیستم ASR به روشی انتها به انتها بدون فرآیند رمزگشایی واقعی در مرحله آزمون. علاوه بر این، یک استراتژی جدید EPD سرتاسری برای ترکیب دانش مدل‌سازی صوتی مبتنی بر تعبیه آوایی (PE) و دانش مدل‌سازی زبان مبتنی بر پیش‌بینی‌کننده EOU در یک رویکرد EPD مبتنی بر تعبیه ویژگی‌های صوتی (AFE) ارائه شده است. چارچوب EPD مبتنی بر شبکه‌های عصبی مکرر (RNN) الگوریتم EPD پیشنهادی بر اساس RNN های گروهی ساخته شده است که به طور مستقل برای سه بخش آموزش داده شده اند که عبارتند از پیش بینی EOU مبتنی بر LM، EPD مبتنی بر AFE و مدل آکوستیک مبتنی بر PE (AM) مطابق با هر هدف. مجموعه RNN ها در سطح آخرین لایه های پنهان به هم متصل می شوند و سپس به طبقه بندی EPD مبتنی بر شبکه های عصبی عمیق کاملا متصل (DNN) متصل می شوند که مطابق با هدف نهایی EPD آموزش داده شده است. پس از آن، آنها به طور مشترک در مرحله دوم آموزش DNN دوباره آموزش داده می شوند تا خطای نقطه پایانی پایین تر را ایجاد کنند. چارچوب EPD پیشنهادی از نظر دقت نقطه پایانی و میزان خطای کلمه برای وظایف CHiME-3 و مقیاس بزرگ ASR ارزیابی شد. نتایج تجربی نشان می‌دهد که الگوریتم EPD پیشنهادی به طور موثری از رویکردهای EPD معمولی بهتر عمل می‌کند.
- فناوری تشخیص گفتار انتها به انتها این مشکل را حل می کند که هر جزء مستقل است و مدل ها نمی توانند به طور مشترک در مدل تشخیص گفتار سنتی بهینه شوند. این مولفه‌هایی مانند مدل صوتی، مدل زبان و واحد رمزگشایی مدل ترکیبی را در یک شبکه عصبی واحد ترکیب می‌کند، که می‌تواند از نقص‌های ذاتی چندین ماژول جلوگیری کند و پیچیدگی مدل تشخیص گفتار را تا حد زیادی کاهش می‌دهد. در این تحقیق، یک سیستم تشخیص گفتار Amdo-Tibetan بر اساس مدل Listen، Attend و Spell (LAS) توسط فناوری تشخیص گفتار end-to-end ساخته شده است. این می تواند تبدیل مستقیم از دنباله گفتار Amdo-Tibetan به دنباله کاراکتر مربوطه را درک کند و تا حد زیادی دشواری ساخت مدل تشخیص گفتار Amdo-Tibetan را کاهش می دهد. برای بهبود بیشتر عملکرد سیستم پیشنهادی،

بهبودهای زیر انجام شده است: در مرحله اول، مکانیسم توجه چند سر برای بهبود دقت تراز بین بردارهای حالت رمزگشا و رمزگذار معرفی شده است. ثانیا، تکنیک صاف کردن برچسب برای حل مشکل بیش از حد استفاده می شود. ثالثاً، یک مدل زبان **N-gram** با مدل **LAS** ترکیب می‌شود تا دقت تشخیص گفتار را افزایش دهد و معیار حداکثر اطلاعات متقابل (**MMI**) برای آموزش متمایز استفاده می‌شود. و در نهایت، یادگیری انتقالی برای غلبه بر مشکل داده های آموزشی ناکافی استفاده می شود. نتایج تجربی نشان می‌دهد که مدل پیشنهادی می‌تواند به طور قابل‌توجهی عملکرد تشخیص گفتار **Amdo-Tbetan** را افزایش دهد.

- ابزارهای کمکی برای تشخیص اختلال در گفتار به دلیل اختلالات عصبی در حال ظهور هستند و این یک کار نسبتاً پیچیده است. یک سیستم تشخیص گفتار با اختلالات هوشمند به افراد مبتلا به اختلال گفتار کمک می کند تا تعاملات خود را با دنیای خارج بهبود بخشند. سخنرانان ناتوان در تلفظ کلمات مشکل دارند که منجر به محتوای گفتاری جزئی یا ناقص می شود. سیستم های تشخیص خودکار گفتار موجود به دلیل تغییرات خاص گوینده که به شدت اختلالات عصبی بستگی دارد، برای تشخیص گفتار مختل موثر نیستند. در این کار، ما دو رویکرد مهم یعنی مدل مارکوف پنهان-شبکه عصبی عمیق و رویکرد اطلاعات متقابل حداکثر بدون شبکه را برای تشخیص موثر گفتار آسیب‌دیده در زبان تامیل بررسی کرده‌ایم. نمونه‌های آموزشی و آزمایشی از افراد مبتلا به اختلالات عصبی مختلف در سطوح مختلف درک مانند بالا، متوسط، پایین و بسیار پایین جمع‌آوری می‌شوند. دقت تشخیص با استفاده از دو مجموعه داده یعنی ۲۰ کلمه مشابه صوتی و ۵۰ کلمه بدنه گفتار مختل در تامیل ارزیابی و مقایسه می‌شود.
- تا به امروز، فناوری تشخیص گفتار برای اکثر زبان‌ها با موفقیت در دستگاه‌های ارتباطی بی‌سیم استفاده شده است. با این حال، تبتی به عنوان یک زبان اقلیت، منابع بسیار محدودی برای تشخیص خودکار گفتار مرسوم دارد. فاقد داده های کافی، واحدهای زیر واژه، واژگان، و فهرست واژه ها برای برخی از گویش ها است. در این مقاله، ما یک مدل چندکاره پایان به انتها برای انجام همزمان تشخیص محتوای گفتار تبتی، شناسایی گویش و تشخیص گوینده ارائه می‌کنیم. این مدل از پردازش فرهنگ لغت تلفظ و تقسیم بندی کلمات برای گویش های جدید اجتناب می کند و در عین حال امکان آموزش سه کار را در یک مدل واحد فراهم می کند. ما چارچوب تشخیص چند وظیفه ای را بر اساس **WaveNet-CTC** می سازیم. اطلاعات گویش و شناسه بلندگو در خروجی برای آموزش استفاده می شود. نتایج تجربی نشان می‌دهد که روش ما عملکرد بهتری در مقایسه با یک مدل ویژه کار دارد.
- در سیستم های تعامل انسان و ماشین، تشخیص احساسات گفتار نقش کلیدی ایفا می کند. تشخیص احساسات طبقه بندی شده در چند دهه اخیر پیشرفت زیادی کرده است، اما تشخیص احساسات از گفتار خود به خودی هنوز بسیار چالش برانگیز است. هدف این مقاله بررسی تشخیص هیجان از گفتار خود به خود در مدل سه بعدی است. هر بعد نشان دهنده یک ویژگی اولیه و عمومی از یک احساس است. سطوح میانی هر بعد در این مقاله معرفی شد. شبکه **LSTM** به دلیل اثربخشی آن در تشخیص عواطف گفتار برای تخمین ابعاد استفاده شد. در آزمایشات از پایگاه داده **IEMOCAP** استفاده می کنیم و دقت ۳۰-۳۵ درصد است. ماتریس‌های سردرگمی نشان می‌دهند که روش ما به مکان بعدی متمرکزتری منجر می‌شود. بعلاوه،

ابعاد در بازشناسی عواطف مقوله ای اعمال شد. این نشان می‌دهد که افزایش سطوح بعد می‌تواند امکان تخمین بعد را فراهم کند و نشان می‌دهد که می‌توان تشخیص عواطف گفتاری را با ابعاد ارتقا داد.

- سیستم‌های پرکاربرد تشخیص گفتار خودکار (ASR) به طور تجربی در مطالعات مختلف نشان داده شده‌اند که ناعادلانه هستند و نرخ خطای بالاتری برای برخی از گروه‌های کاربران نسبت به سایرین دارند. به عنوان مثال، تغییر جنسیت، سن، تحصیلات یا نژاد آنها) نباید توزیع احتمال را در رونویسی‌های گفتار به متن ممکن تغییر دهد. در پارادایم انصاف خلاف واقع، همه متغیرهای مستقل از وابستگی گروه (مثلاً متنی که توسط گوینده خوانده می شود) بدون تغییر باقی می مانند، در حالی که متغیرهای وابسته به وابستگی گروه (مثلاً صدای گوینده) به طور خلاف واقع اصلاح می شوند. از این رو، ما با آموزش ASR برای به حداقل رساندن تغییر در احتمالات نتیجه آن علیرغم تغییر خلاف واقع در ویژگی های جمعیت شناختی فرد، به عادلانه بودن ASR نزدیک می شویم. با شروع از معیار شانس برابر متضاد فردی، ما تسهیلاتی را برای آن فراهم می‌کنیم و عملکرد آنها را برای سیستم‌های ASR سرتاسر مبتنی بر طبقه‌بندی زمانی اتصالگرا (CTC) مقایسه می‌کنیم. ما آزمایش‌های خود را روی مجموعه زبان منطقه‌ای آفریقایی آمریکایی (CORAAL) و مجموعه داده‌های LibriSpeech انجام می‌دهیم تا تفاوت‌های ناشی از جنسیت، سن، تحصیلات و نژاد را در نظر بگیریم. ما نشان می‌دهیم که با آموزش خلاف واقع، می‌توانیم میانگین نرخ خطای شخصیت را کاهش دهیم و در عین حال شکاف عملکردی کمتری بین گروه‌های جمعیتی و انحراف معیار خطای کمتری در بین افراد به دست آوریم.

۵ پیشنهاد برای بهبود و ارتقای روش مورد استفاده:

۱- قرار دادن سوکتی در مغر برای فهیمدن صحبت‌های همدیگر بدون صحبت کردن!
۲- سوکتی که حتی کسایی که زبان آن‌ها متفاوت است فهمیده شود.
۳- سوکتی که حتی کسایی که لال هستن هم فهمیده شود.
۴- سوکتی که فرکانس آن با فرکانس شخصی که میخواهیم با آن صحبت کنیم یکی باشد تا صحبت‌هایمان را جز آن شخص، کسی متوجه نشود.
۵- سوکتی که نویز اطراف را نگیرد تا مکالمه از فاصله دور هم میسر باشد.