مقاله مروری

مجله ماشین‌بینایی و پردازش تصویر

دوره ۱۲، شماره ۴، بهار ۱۴۰۵

مروری بر مجموعه داده‌های صوتی و تصویری در بازشناسی تصویری گفتار پیوسته

مهشید السادات احصائی[۱]، اعظم باستان‌فرد[۲*]

چکیده

بررسی مجموعه داده‌های لب‌خوانی اولین چالش مهم در زمینه بازشناسی تصویری گفتار پیوسته است. گروهی از محققان برای بازشناسی گفتار و لب‌خوانی به جمع‌آوری مجموعه داده‌هایی جهت ارزیابی الگوریتم‌های پیشنهادی خود پرداخته‌اند. به گونه‌ای که به تناسب نیاز خود تنها برخی از ویژگی‌های داده‌ها را در نظر گرفته‌اند. چرا که داده‌های صوتی تصویری برای بازشناسی گفتار و لب‌خوانی دارای ویژگی‌های متفاوتی چون گفتار پیوسته و یا گفتار گسسته، زبان پایگاه داده‌ها، زاویه تصویربرداری از گویندگان است. محققان جهت پردازش گفتار و آغاز به کار نیاز به آمارهایی در رابطه با پایگاه داده‌های موجود دارند. چالش نداشتن آمار در زمینه داده‌های گفتار پیوسته انگیزه‌ای شد تا در این مقاله مجموعه داده‌های صوتی تصویری گفتار پیوسته معرفی شوند و نحوه جمع‌آوری آنها، تنظیمات ضبط، محیط ضبط و ویژگی‌های اصلی آنها مانند تعداد گویندگان، تعداد تکرار گفتارها، رزولوشن تصویر بررسی شوند. داده‌های مربوط به پایگاه داده‌های موجود بر حسب ویژگی‌های کمی و کیفی آنها دسته‌بندی و متناسب با این ویژگی‌ها آمارهای متفاوتی همچون درصد بومی بودن گویندگان، نسبت جنسیت گویندگان، میانگین سنی گویندگان، تعداد زوایای تصویربرداری از گویندگان و مدت زمان ضبط داده‌ها ارائه شده است. در پایان مزایای پایگاه داده‌های مورد مطالعه به همراه آدرس دسترسی به آنها لیست شده است.

کلید واژه ها

مجموعه داده صوتی و تصویری، گفتار پیوسته، لب‌خوانی خودکار، بازشناسی تصویری گفتار

۱- مقدمه

در دسترس بودن پایگاه داده‌هایی با کیفیت بالا برای مطالعه تشخیص گفتار و پژوهش‌های صوتی و تصویری به عامل اصلی در توسعه این پژوهش‌ها تبدیل شده است[۱۷۰]. در دهه نود پایگاه داده‌های اولیه برای سامانه‌های لب‌خوانی خودکار با کاربردهای متفاوتی شامل گفتار درمانی [۱۱۶] ایجاد شدند. این پایگاه داده‌ها شامل یکسری واژگان محدود با هدف شناسایی حروف الفبا مانند AVLetters[۴۱] و AVLetters2 [۴۲] و یا ارقام مانند CUAVE[۴]، BANCA[۵] بودند که محبوب‌ترین پایگاه داده رقمی CUAVE است. ولی با توسعه این سامانه‌ها و هدف شناسایی کلمات و جملات نیاز به توسعه پایگاه داده‌ها نیز احساس شد. به گونه‌ای که پایگاه داده‌های SAMPA [۴۰]، [۴۳] و [۴۴]، MODALITY [۴۵]، LRW[۴۶] و ISO-211 [۴۷] با هدف شناسایی کلمات جمع‌آوری شدند. گروه دیگری از پایگاه داده‌ها با هدف شناسایی اصطلاحات جمع‌آوری شدند که

این مقاله در ۱۲ مهر ۱۴۰۴ دریافت ودر ۸ بهمن ۱۴۰۴ پذیرفته شد

[۱] مهشید السادات احصائی، دانشکده مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد کرج، کرج، ایران، mahshid.ehsaei@iau.ac.ir

[۲*] اعظم باستان‌فرد (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد کرج، کرج، ایران، bastanfard@iau.ac.ir

بسیار محدودتر از جملات هستند مانند Grid [۷] و OuluVS [۸]. مجموعه داده‌های زیادی با زبان‌های متنوعی نیز با هدف شناسایی جملات ایجاد شدند که به عنوان مثال پایگاه داده‌های IBMViaVoice [۹] و LILiR [۱۴] از این گروه هستند. پایگاه داده‌هایی نیز وجود دارند که طیف وسیع‌تری از اهداف را در بر می‌گیرند، مثلا مجموعه داده AVOZES [۲۳] و AVA|| [۲۶] که شامل ارقام و جملات هستند و یا پایگاه داده OuluVS2 [۳۰] که شامل ارقام، اصطلاحات و جملات است. بنابراین به مرور زمان پایگاه داده‌های بعدی موارد گسترده‌تری شامل موضوعات، تکرار، روشنایی، حرکت سر و صورت و واژگان متنوعی را نیز مد نظر قرار دادند تا به واقعیت نزدیک‌تر شوند. از آنجایی که ایجاد پایگاه داده‌ها با در نظر گرفتن کلیه ابعاد، مسئله بسیار پیچیده‌ای است بنابراین هر یک از پایگاه داده‌های موجود بر روی یکسری از این ابعاد و ویژگی‌ها متمرکز شده‌اند و مابقی ویژگی‌ها را در نظر نگرفته‌اند. به طور مثال گروهی از پایگاه داده‌ها در محیط آزمایشگاهی ایجاد شده‌اند و دارای شرایط ضبط یکسان از جمله نور، فاصله تا دوربین و رنگ پس‌زمینه هستند مانند Grid [۷] و یا گروهی کاملا بالعکس از محیط‌های طبیعی به دست آمده‌اند که بسیاری از شرایط در حال تغییر است مانند LRS [۳] که بزرگ‌ترین پایگاه داده گفتار پیوسته شامل صدها هزار سخن توسط بیش از هزاران گوینده است. به طوری که ممکن است هر گوینده با لهجه خاصی صحبت کند که موضوع پر اهمیتی در گفتار است [۱۱۷].

گفتار پیوسته جایی است که توالی‌های ورودی شامل هم‌تولیدی از کلمات همسایه در یک ثانیه است [۴۸] و گوینده به طور طبیعی صحبت می‌کند به طوری که کلمات با مکث از یکدیگر جدا نمی‌شوند. در مقابل گفتار گسسته جایی است که کلمات می‌توانند در کمتر از یک ثانیه به طول انجامند [۴۸] و در واقع روش غیر طبیعی خواندن کلمات است به گونه‌ای که مکث مختصر بین آنها وجود دارد و پیدا کردن نقطه ابتدایی و انتهایی کلمات را آسان‌تر می‌کند و تلفظ هر کلمه تأثیری بر بقیه کلمات ندارد و از جمله پایگاه داده‌های گفتار گسسته می‌توان CUAVE [۴] را نام برد. بنابراین پایگاه داده‌ها به سوی پایگاه‌های بزرگ با توانایی زیاد برای آموزش سامانه‌های قوی لب‌خوانی خودکار و بازشناسی گفتار خودکار در حال رشد هستند [۴۹]. از طرفی دیگر گروهی از پایگاه داده‌ها به جمع‌آوری مجموعه‌ای از داده‌های وابسته به یک گوینده پرداخته‌اند مانند پایگاه داده RM-3000 [۲] که شامل مجموعه‌ای بزرگ از واژگان مختلفی است که توسط تنها یک گوینده بیان شده است. سامانه‌های وابسته به گوینده معمولاً با هدف شناسایی هویت گفتار یا گوینده بر اساس ویژگی‌های وابسته به گوینده ایجاد می‌شوند و در مقابل سامانه‌های مستقل از گوینده باید بتوانند بخشی از گفتار را بدون توجه به اینکه چه کسی صحبت می‌کند، تشخیص دهند [۵۹] و [۱۷۲]. که بر اساس این تعریف گروهی دیگر از پایگاه داده‌ها به جمع‌آوری مجموعه‌ای از داده‌ها از تعدادی از گویندگان پرداخته‌اند که مستقل از گوینده هستند، مانند پایگاه داده IBMSR [۵۸]. اکثر پایگاه داده‌های موجود به زبان انگلیسی تولید و طراحی شده‌اند ولی چندین پایگاه داده نیز به زبان‌های فارسی: AVA [۱۰۳]، AVA|| [۲۶]، آلمانی: SAMPA [۴۰]، [۴۳] و [۴۴]، GLips [۱۴۷]، اسپانیایی: VLRF [۲۲]، چک: UWB-07-ICAV [۱۲]، ژاپنی: CENSREC-1-AV [۶۰]، روسی: HAVRUS [۲۰]، فرانسوی: IV2 [۱۳]، هلندی: NDUTAVSC [۲۷]، ترکیب انگلیسی، هندی و بنگالی: MAVS [۱۴۵] و اندونزیایی: LUMINA [۱۴۶] موجود هستند.

از آنجایی که داده موضوع بسیار ارزشمند در قرن کنونی است، تعیین ویژگی‌های مناسب داده و جمع‌آوری داده‌ها از چالش‌های موجود پیش روی محققین است. از چالش‌ها و محدودیت‌های معمول پایگاه داده‌های گفتار صوتی تصویری می‌توان در دسترس بودن آنها، نداشتن آمار در زمینه داده‌ها، پوشش تعداد کمی از واج‌ها و ویزم‌ها و پوشش کلمات جدا شده مانند رقم یا حروف الفبا به جای گفتار پیوسته را نام برد و در تولید پایگاه داده‌ها نیز مواردی چون پیدا کردن افرادی که حوصله ضبط‌های طولانی و مکرر را داشته باشند، زمان‌بر بودن فرآیند تولید پایگاه داده‌ها و تعداد کم گویندگان را نام برد که تولید پایگاه داده‌ها را با مشکل مواجه کرده است. فرناندز لوپز و سوکنو [۴۹] با هدف بازبینی پژوهش‌های انجام شده در زمینه لب‌خوانی خودکار لیستی از پایگاه‌های اطلاعاتی سمعی بصری از سال ۲۰۰۷ تا ۲۰۱۷ برای لب‌خوانی ارائه داده‌اند و برای هر پایگاه داده فقط ویژگی‌های محدودی را مورد بررسی قرار داده‌اند. هدایتی‌پور و همکاران [۱] نیز در راستای مروری بر روش‌های لب‌خوانی، فهرستی از پایگاه‌های داده مورد استفاده در این زمینه را به همراه یکسری از ویژگی‌های آنها معرفی کرده‌اند. از آنجایی که محققین برای انجام پردازش داده‌ها نیاز به بررسی داده‌ها دارند و پایگاه داده‌های موجود به دلیل ذات داده‌ها دارای ویژگی‌های متعددی هستند ما در این مقاله پایگاه داده‌های گفتار پیوسته که گفتار طبیعی است از سال ۱۹۹۹ تا کنون را بر اساس وظیفه‌شان تقسیم کرده و ویژگی‌های کمی و کیفی آنها را تفکیک شده به طور نسبتا کامل مورد بررسی قرار داده و آمارهایی در این زمینه ارائه داده‌ایم. نقاط قوت و مزایای پایگاه داده‌های موجود لیست و آدرس دسترسی به آنها نیز ارائه شده است.

در ادامه این مقاله در بخش ۲ روش کار توضیح داده شده است. در بخش ۳ مجموعه داده‌های موجود در زمینه گفتار پیوسته شامل اصطلاحات و جملات مورد بررسی قرار می‌گیرند. پژوهش مورد مطالعه در بخش ۴ مورد بحث قرار می‌گیرد. در پایان در بخش ۵ جمع‌بندی و کارهای آینده بیان می‌شوند.

۲- روش کار

در این مطالعه، یک مرور شبه سیستماتیک با هدف گردآوری و تحلیل جامع پژوهش‌های کلیدی مرتبط با پایگاه داده‌های صوتی و

تصویری در حوزه گفتار پیوسته، شامل مشخصات فنی، ابعاد، زبان‌ها و کاربردهای آن‌ها که در جداول ویژه ثبت شدند، انجام شد. برای جستجوی منابع، پایگاه‌های داده علمی معتبر در حوزه کامپیوتر و علوم گفتار مانند IEEE Xplore, ACM Digital Library، Scopus و Google Scholar انتخاب و جستجوی گسترده‌ای با کلیدواژه‌های مرتبط شامل continuous speech, databases، speech corpora، recognition datasets انجام پذیرفت. این مرور بدون محدودیت زمانی صورت گرفت و کلیه مقالات مرتبط بدون توجه به سال انتشار، مورد بررسی قرار گرفتند. محدودیت زبان به مقالات منتشر شده به زبان انگلیسی و فارسی اعمال شد. از دیگر معیارهای ورود ارتباط مستقیم با موضوع بود که مقالات نامرتبط، تکراری یا کم‌کیفیت حذف شدند تا تحلیل تنها بر منابع معتبر و مرتبط متمرکز باشد. فرایند غربالگری شامل بررسی اولیه عناوین و چکیده‌ها و سپس مطالعه متن کامل مقالات مرتبط و قابل دسترس بود. ارزیابی کیفیت پایگاه‌ها بر اساس اعتبار منابع، گستردگی داده‌ها و قابلیت کاربرد آن‌ها در پژوهش‌های لب‌خوانی انجام گرفت. این فرآیند تحلیل، امکان درک عمیق‌تر و نظام‌مند از داده‌ها را فراهم می‌آورد و شکاف‌ها و روندهای پژوهشی به دقت شناسایی گردید و مبنای محکمی برای تحقیقات پیشرفته در حوزه بازشناسی گفتار و لب‌خوانی می‌باشد [۱۵۴].

در این مرور سیستماتیک، به منظور تحلیل جامع پایگاه داده‌های گفتار پیوسته، پرسش‌های پژوهشی زیر تعیین شدند که به عنوان چهارچوب اصلی تحقیق عمل می‌کنند:

- چه پایگاه داده‌های گفتار پیوسته‌ای تاکنون توسعه یافته‌اند و ویژگی‌های کلیدی آن‌ها چیست؟
- چه زبان‌ها و گویش‌هایی در این پایگاه‌های داده پوشش داده شده‌اند؟
- کاربردهای اصلی این پایگاه داده‌ها در حوزه‌های مختلف بازشناسی گفتار و پردازش زبان طبیعی چیست؟
- روند توسعه و به‌روزرسانی پایگاه‌های داده گفتار پیوسته چگونه بوده است؟
- چالش‌ها و محدودیت‌های موجود در جمع‌آوری و استفاده از داده‌های گفتار پیوسته کدام‌اند؟

۳- معرفی پایگاه داده‌های موجود در گفتار پیوسته

درک چگونگی پردازش و تفسیر گفتار انسان در شرایط نامطلوب یک چالش مهم علمی است. برای آزمایش و مقایسه نتایج منتشر شده توسط گروه‌های مختلف پژوهشی در زمینه بازشناسی گفتار و لب‌خوانی، مبنای مشترکی در قالب مجموعه داده‌های جامع گفتار صوتی تصویری طراحی شده است که از اهمیت بسیاری برخوردار هستند. بسیاری از این مجموعه داده‌های گفتار صوتی تصویری در دسترس عموم قرار ندارند. یکی از چالش‌های مهم پژوهشگران داشتن معیاری جهت شناسایی ویژگی‌های استاندارد و مدون و ارزیابی داده‌ها است. از جمله ویژگی‌هایی که یک پایگاه داده می‌بایست داشته باشد عبارتند از [۲۳]: تعداد زیادی از گویندگان به دلیل اهمیت آماری، پوشش گسترده‌ای از واج‌ها و ویزم‌ها، سطوح مختلف سر و صدای صوتی با نویز، تصاویر کامل چهره با رنگ، کلمات کوتاه و گفتار پیوسته با رونویسی، و قابلیت گسترش داده.

با مرور زمان، ایجاد مجموعه داده‌هایی که به واقعیت نزدیک باشد و از طریق آن بتوان مکالمات روزمره را لب‌خوانی کرد نیاز شد. لازم به ذکر است که می‌توانیم مجموعه داده‌های لب‌خوانی را بر اساس اجزایی که بر آنها تمرکز دارند، یعنی کاراکترها، ارقام، کلمات، عبارات و جملات دسته‌بندی کنیم [۱۶۹]. در ادامه پایگاه داده‌هایی که شامل عبارات مختصر، اصطلاحات و جملات هستند مورد بررسی قرار گرفته‌اند.

۳-۱- مجموعه داده‌های شامل اصطلاحات

طبق مطالعات انجام شده در این مقاله دو پایگاه داده وجود دارند که فقط شامل اصطلاحات هستند، پایگاه داده Grid[۷] و OuluVS[۸] که در ادامه تشریح می‌شوند.

۳-۱-۱- مجموعه داده Grid

این مجموعه داده در سال ۲۰۰۶ توسط انجمن آکوستیک آمریکا جمع‌آوری شده است.

جنبه‌های کمی: Grid[۷] با ۳۴ گوینده شامل ۱۶ زن و ۱۸ مرد جمع‌آوری شده است که شامل ۱۰۰۰ جمله ساده است. گویندگان کارمندان و دانشجویان گروه علوم کامپیوتر و گروه علوم ارتباطات انسانی دانشگاه شفیلد هستند که انگلیسی زبان اول آنها است. ۳ گوینده بیشتر زندگی خود را در انگلستان گذرانده‌اند، ۲ گوینده در اسکاتلند و یک نفر در جامائیکا متولد شدند. سن گویندگان بین ۱۸ تا ۴۹ سال است که میانگین سنی آنها ۲۷/۴ سال است.

جنبه‌های کیفی: سیگنال صوتی توسط میکروفن B&K مدل Nexus 2690 در شرایط آکوستیکی با فرمت TDT ضبط شده است. گویندگان برای بیان هر جمله ۳ ثانیه فرصت داشتند و می‌بایست با سبک طبیعی صحبت کنند. برای ضبط تصویر از دوربین فیلم‌برداری ویدئویی کَنون XM2 با سرعت ۲۵ فریم در ثانیه استفاده شده است. فیلم‌ها با پس‌زمینه آبی ساده با فرمت DV با استفاده از FFMPEG.2 به فرمت MPEG-1 تبدیل شدند. برای تولید روشنایی یکنواخت منابع نور در سراسر صورت قرار گرفته است. به طور متوسط، از هر ۱۰۰۰ گفتار ۵۷ سخن با خطا مواجه بودند که توسط گویندگان مجدد ضبط شدند که در مجموع، ۶۴۰ جمله و یا ۱/۹٪ از مجموعه داده دوباره ثبت شدند.

ژانگ و همکاران(روش LSTM و DCT) [۶۳]، آسل و همکاران(روش D-CNN۳ و Bi-GRU) [۶۴]، وند و همکاران (روش Feed-forward و LSTM) [۶۵](Feed-forward و

LSTM)[۶۶]، چیونگ و همکاران(CNN و + LSTM attention) [۳] و [۲۱]، سو و همکاران (۳-D CNN+highway و attention + Bi-GRU) [۶۷] نتایج لب‌خوانی خود را روی پایگاه داده Grid نشان داده‌اند که به ترتیب به نرخ بازشناسی ۸۲/۰۰٪، ۹۵/۲۰٪، ۷۹/۵۰٪، ۸۴/۷۰٪، ۹۷/۰۰٪ و ۹۷/۱۰٪ رسیدند. شاو و برکر [۶۸] و لن و همکاران [۶۹] نیز از این پایگاه داده برای استخراج ویژگی و طبقه بندی استفاده کردند که به ترتیب به کارایی ۵۸/۴۰٪ و ۶۵٪ رسیدند. گبی و همکاران [۸۰]، مورون و همکاران [۸۱]، افرت و همکاران [۸۲]، ایوانکو و همکاران [۸۶]، سارهان و همکاران [۸۷]، پارخ و همکاران [۱۰۵]، مرگام و همکاران [۱۱۹] و کیم و همکاران [۱۲۰]، چن و همکاران[۱۴۸]، زو و همکاران [۱۴۹]، شوئه و همکاران [۱۵۰] و کیو و همکاران [۱۵۱] نیز از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند. Grid از پایگاه داده‌هایی است که به طور وسیعی در تحقیقات مورد استفاده قرار گرفته است.

۲-۱-۳- پایگاه داده OuluVS

این پایگاه داده در سال ۲۰۰۹ جمع‌آوری شده است[۸].

جنبه‌های کمی: شامل ۸۱۷ دنباله از ده عبارت با تکرار یک تا پنج بار و ۲۰ گوینده متشکل از ۳ زن و ۱۷ مرد است. ۹ نفر از آنها عینک دارند و از چهار کشور مختلف هستند، بنابراین عادات تلفظ متفاوتی از سرعت صحبت کردن را دارند.

جنبه‌های کیفی: دوربین سونیDSR-200AP 3CCD با سرعت ۲۵ فریم در ثانیه با وضوح تصویر ۷۲۰ در ۵۷۶ پیکسل استفاده شده است. افراد بر روی یک صندلی در فاصله ۱۶۰ سانتی‌متری از دوربین نشستند. میانگین سایز تصویر دهان ۱۲۰ در ۷۰ است.

چیونگ و زیسرمن (CNN)[۴۶]،رکیک و همکاران [۷۰] و [۲۸]، پینگ‌پینگ و همکاران(روش SDF+STLF) [۷۱]، پتریدیس و پانتیک [۱۲۱]، مصباح و همکاران [۱۲۲]، جنگ و همکاران[۱۲۳]، شیراکاتا و سایتو[۱۲۴] و ژو و همکاران [۱۷۲] نتایج لب‌خوانی خود را روی مجموعه داده Grid نشان داده‌اند. ژائو و همکاران [۸] از این پایگاه داده که خود جمع‌آوری کرده‌اند برای استخراج ویژگی و طبقه‌بندی استفاده کردند که به کارایی ۶۲/۴٪ رسیدند.

۲-۳- مجموعه داده‌های شامل جملات

گویندگان در گروهی از پایگاه داده‌ها به منظور نزدیک شدن به واقعیت مجموعه‌ای از جملات را بیان کرده‌اند. که در ادامه به معرفی این پایگاه داده‌ها پرداخته خواهد شد.

۲-۱-۳- پایگاه داده IBMViaVoice

پایگاه داده IBMViaVoice[۹] در سال ۲۰۰۰ به زبان انگلیسی جمع‌آوری شده است.

جنبه‌های کمی: مجموعه داده شامل واژگان پیوسته گسترده‌ای تقریبا ۱۰۵۰۰ کلمه است و ۲۹۰ نفر این واژگان را بیان کردند و تا آن روز بزرگ‌ترین پایگاه داده صوتی، تصویری مستقل از گوینده بوده است که نمای از جلوی گویندگان را دارد. مدت کل پایگاه داده تقریباً ۵۰ ساعت است. در این پایگاه داده ۲۴۳۲۵ سخن همانند فرهنگ لغت تلفظ وجود دارد.

جنبه‌های کیفی: ویدئوها دارای رزولوشن ۷۰۴ در ۴۸۰ پیکسل با سرعت ۳۰ هرتز ضبط شده است. فایل صوتی با سرعت ۱۶ کیلوهرتز و بدون نویز با فرمت MPEG2 ضبط شده است. مجموعه آموزش شامل ۳۵ ساعت داده از ۲۳۹ موضوع است و تست یک پایگاه داده ۵ ساعتی از داده‌ها از ۲۵ نفر و یک مجموعه آزمون ۲/۵ ساعته از ۲۶ نفر است.

متیوس و همکاران [۷۲] از این مجموعه داده برای نشان دادن نتایج مدل خود استفاده کرده‌اند.

۲-۲-۳- مجموعه داده VIDTIMIT

مجموعه داده VIDTIMIT [۱۰] در سال ۲۰۰۲ به زبان انگلیسی با حجم ۳/۵ گیگابایت جمع‌آوری شده است.

جنبه‌های کمی: این مجموعه داده شامل ۴۳ گوینده که ۱۹ زن و ۲۴ مرد جملاتی کوتاه را بیان می‌کنند. در ۳ جلسه و با تاخیر ۷ روز بین جلسه ۱ و ۲، و ۶ روز تاخیر بین جلسه ۲ و ۳ در سال ۲۰۰۲ ضبط شده است. تأخیر بین این جلسات امکان تغییر در صدا، سبک مو، آرایش، لباس و خلق و خوی که در تلفظ تأثیر می‌گذارد را فراهم می‌کند. ضریب زوم دوربین پس از هر ضبط به طور تصادفی مختل می‌شود. برای هر نفر ۱۰ جمله وجود دارد که شش جمله اول به جلسه ۱ اختصاص داده می‌شود. دو جمله بعدی به جلسه ۲ و دو جمله باقی‌مانده به جلسه ۳ اختصاص داده می‌شود. دو جمله اول برای همه افراد یکسان است و هشت جمله بعدی برای هر فرد متفاوت است. میانگین مدت زمان هر جمله ۴/۲۵ ثانیه یا تقریباً ۱۰۶ فریم ویدیویی است. حرکت سر برای گویندگان مجاز است.

جنبه‌های کیفی: از دوربین فیلم‌برداری دیجیتال پال در یک محیط دفتر پر سر و صدا استفاده شد. که فیلم‌ها دنباله‌ای از تصاویر JPEG با وضوح ۳۸۴ در ۵۱۲ پیکسل ذخیره می‌شود. فایل صوتی نیز به صورت WAV مونو، ۱۶ بیتی، ۳۲ کیلوهرتز ذخیره می‌شود. اگر جلسه ۱ به عنوان بخش آموزش و جلسات ۲ و ۳ به عنوان بخش آزمون باشند جمله تکراری در این ۲ بخش وجود ندارد و بنابراین برای سامانه‌های مستقل از متن مناسب هستند.

کاپلتا و هارت [۷۳] از این پایگاه داده برای استخراج ویژگی و طبقه‌بندی استفاده کردند که به کارایی ۶۰/۱۰٪ رسیدند.

۳-۲-۳- مجموعه داده AV-TIMIT

مجموعه داده AV-TIMIT[۱۱] در سال ۲۰۰۴ به زبان انگلیسی شامل ۴ ساعت گفتار پیوسته با آوایی متعادل جمع‌آوری شده است.

جنبه‌های کمی: از ۲۲۳ گوینده مختلف شامل ۱۱۷ مرد و ۱۰۶ زن که همگی به جز ۱۲ نفر انگلیسی زبان بودند و فیلم‌هایی با وضوح بالا در دفتری اداری نسبتاً ساکت و دارای روشنایی کنترل شده در طول یک هفته ضبط شده است. افراد با سنین و قومیت‌های مختلف همین‌طور با و یا بدون ریش، عینک و کلاه هستند و حرکت سر برای گویندگان مجاز است. این پایگاه داده از نظر گفتار آوایی، غنی است که از ۴۵۰ جمله TIMIT-SX استفاده شده است تا پوشش مناسب آوایی زبان انگلیسی را در کم‌ترین کلمه ممکن ارائه دهد[۳۳]. هر گوینده ۲۰ جمله را می‌خواند که جمله اول برای همه یکسان است و بقیه جملات برای هر نوبت متفاوت است. ۲۳ نوبت از گفتار مختلف ایجاد شد که در هر یک از ۲۳ نوبت سخنرانی، حداقل ۹ گوینده مختلف صحبت کردند.

جنبه‌های کیفی: از میکروفن آرایه صوتی GN Netcom واقع در پشت صفحه کلید و دوربین فیلم‌برداری سونی DCR-VX2000 با کیفیت بالا که نمای روبه‌روی هر فرد را ضبط می‌کند استفاده شده است و پس‌زمینه افراد یک پرده آبی است. فیلم‌ها به صورت دیجیتال AVI با وضوح ۷۲۰ در ۴۸۰ و ۳۰ فریم در ثانیه ذخیره شد. فایل صوتی نیز در پرونده‌های WAV جداگانه با نمونه‌برداری ۱۶ کیلوهرتز ذخیره شده است. میانگین نسبت سیگنال به نویز در حین صحبت‌های فردی تقریباً ۲۵ دسی‌بل با انحراف استاندارد ۴/۵ دسی‌بل بود. رونوشت صوتی و تصویری نیز موجود است.

۴-۲-۳- مجموعه داده UWB-07-ICAV

مجموعه داده UWB-07-ICAV [۱۲] مخفف شده دانشگاه وست بوهمیا برای UWB است و ۰۷ به دلیل اینکه ۷ سال ضبط آن طول کشید و ICAV نیز مخفف شده شرایط دارای اختلال در بازشناسی گفتار شنیداری صوتی است که در سال ۲۰۰۸ جمع‌آوری شده است و پایگاه داده قبلی یعنی UWB-05-HSCAVC [۵۰] را بزرگ می‌کند.

جنبه‌های کمی: این پایگاه داده شامل ۵۰ گوینده که نیمی از آنها مرد و نیمی دیگر از آنها زن هستند به زبان چک جمع‌آوری شده است و متوسط سن آنها ۲۲ سال است. پایگاه داده ۱۰۰۰۰ سخن را که به صورت پیوسته بیان می‌شود در قالب ۱۰۰۰۰ جمله مجزا ذخیره کرده است. هر گوینده ۲۰۰ جمله را تلفظ می‌کند که ۵۰ جمله برای همه یکسان است و بعد از بیان هر ۸ جمله نورپردازی تغییر می‌کند و بقیه جمله‌ها متفاوت هستند. همین‌طور بعد از بیان هر ۲۵ جمله نورپردازی تغییر می‌کند. هر گوینده پس از بیان یک جمله ۲ ثانیه مکث طولانی می‌کند. طول ضبط در هر جلسه به طور متوسط نیم ساعت است. همه فایل‌ها با برچسب‌های تصویری همراه هستند که منطقه دهان و ناحیه اطراف آن را مشخص می‌کند و تلفظ واقعی هر جمله نیز در فایل متنی رونویسی شده است. این مجموعه داده به مجموعه توسعه و ارزیابی تقسیم می‌شود که داده‌های توسعه شامل ۵۲ نفر است و داده‌های ارزیابی حدود ۳۰۰ نفر است که در بین این افراد ۷۷ نفر برای جلسه دوم و ۱۹ نفر برای جلسه سوم آمده‌اند.

جنبه‌های کیفی: برای ضبط ویدئوها از دو دوربین و دو میکروفن استفاده شده است. دوربین اول یک دوربین کَنون VCR MVX3i و داده‌ها بر روی نوار به مدت ۶۰ دقیقه در MPEG-4 رمزگذاری و در قالب AVI ذخیره شدند. برای گرفتن وضوح بهتر چهره، دوربین ۹۰ درجه چرخانده شده است. از نرم افزار AviSynth برای کم کردن فاصله فیلم استفاده کردند که از رزولوشن اصلی ۷۲۰ در ۵۷۶ پیکسل با ۲۵ فریم در ثانیه به همان وضوح تصویر با ۵۰ فریم در ثانیه رسید. در هنگام ضبط نور را تغییر دادند که نورپردازی متغیر باعث ایجاد سایه‌هایی بر روی صورت می‌شود. تصویربرداری نیز از نمای روبه‌رویی و بدون حرکت سر انجام می‌شود. داده‌های بصری از هر گوینده برای هر دوربین فیلم‌برداری ۵ گیگابایت فضا را اشغال می‌کند. دوربین دوم یک دوربین وب‌کم فیلیپس SPC 900NC با وضوح ۶۴۰ در ۴۸۰ با ۳۰ فریم در ثانیه بود که برای هر گوینده از وب‌کم حدود ۱۰۰ مگابایت فضا اشغال می‌شود. برای نورپردازی از دو چراغ هالوژن و دو چراغ دیجیتال استفاده شد. پس‌زمینه صحنه یکنواخت و سیاه است. هر جمله حاشیه نویسی شده است که به صورت تلفظ واقعی با نشانه آوایی است.

قسمت صوتی توسط دو میکروفن ضبط شده است که میکروفن اول میکروفن جهت دار AKG CK55L و دومی میکروفن پایه‌ای Sennheiser K6 که سمت چپ گوینده قرار داشت. اکثر داده‌ها با فرمت WAV با فرکانس نمونه‌برداری ۴۴ کیلوهرتز و وضوح ۱۶ بیت هستند. داده‌های صوتی برای یک گوینده تقریباً ۲۰۰ مگابایت فضای دیسک را اشغال می‌کند. در حین ضبط صدا از هدفون‌هایی که صدا ایجاد می‌کردند، برای تأثیرگذاری بر کیفیت یک سخنرانی استفاده شد.

۵-۲-۳- پایگاه داده IV2

سپس در سال ۲۰۰۸ پایگاه داده IV2[۱۳] به زبان فرانسوی جمع‌آوری شد.

جنبه‌های کمی: از ۳۰۰ گوینده ۷۷ نفر آن در دو جلسه که حدود یک ماه طول کشید و بقیه گویندگان در طول یک جلسه جملات را بیان کردند.

جنبه‌های کیفی: برای تصویربرداری از دو دوربین با کیفیت بالا و پایین استفاده می‌شود. دوربین دیجیتال با کیفیت بالا با عنوان DVCAM دارای رزولوشن ۷۸۰ در ۵۷۶ پیکسل با میانگین فاصله بین مرکز چشم ۹۰ پیکسل و دوربین وب‌کم با کیفیت پایین

با وضوح ۴۸۰ در ۴۴۰ پیکسل و میانگین فاصله بین مرکز چشم آن ۶۵ پیکسل تصویر و صوت را همزمان ثبت می‌کند. هر گوینده ۱۵ جمله را با مدت زمان حدود ۱۰۰ ثانیه در حالات خندان، انزجار و ترس بیان می‌کند. نسخه دوم این بانک اطلاعاتی شامل ۴۰۰۷ جلسه از ۴۶۶ سوژه است که در شرایط روشنایی کنترل شده به دست آمده است. این پایگاه داده از پنج حالت چهره خنثی، چشمان بسته، انزجار، خوشحالی و تعجب و دو حالت مختلف نورپردازی و سه حالت نما شامل جلو، نمای چپ و راست تشکیل شده است. در حین فیلم‌برداری نورپردازی از چپ به راست تغییر می‌کند. در این پایگاه داده مجموعه‌ای از ابرداده‌ها نیز موجود هستند شامل کلاس سنی، رنگ پوست، رنگ چشم، رنگ مو و وجود ریش، سبیل و عینک و اطلاعات کالیبراسیون سرهای استریوسکوپی است. به طور کلی داده‌های ضبط شده برای یک جلسه کامل حدود ۲/۲ گیگا بایت برای هر موضوع است.

۶-۲-۳- پایگاه داده LILiR

پایگاه داده LILiR [۱۴] در سال ۲۰۱۰ با همکاری آزمایشگاه پردازش گفتار و زبان در دانشگاه آنگلیای شرقی با آزمایشگاه پردازش سیگنال و گفتار دانشگاه سوریه جمع‌آوری شد که هدف ساخت یک مجموعه داده بزرگ انگلیسی چند گوینده صوتی و تصویری بود.

جنبه‌های کمی: این مجموعه داده شامل ۱۲ گوینده که ۷ نفر از آنها مرد و ۵ نفر زن هستند است و هر یک از آنها ۲۰۰ جمله از جملات مدیریت منابع[۳۴] را بیان می‌کنند. این بانک اطلاعاتی دارای واژگان تقریبی ۱۰۰۰ کلمه است و هر گوینده جملات خود را در یک جلسه ضبط می‌کند تا از اختلافات ظریف در روشنایی جلوگیری کند. جمع‌آوری داده‌ها در محیطی با شرایط کنترل روشنایی انجام شد.

جنبه‌های کیفی: این ویدئوها با استفاده از دوربین سه بعدی با کیفیت بالا با فرکانس نمونه‌برداری از ۲۵ فریم در ثانیه ضبط و تصاویر از روبه‌روی افراد ضبط شده است. برای فراهم کردن یک انتقال فریم نرم‌تر و نرخ فریم قابل مقایسه با صدا، فیلم اصلی از ۱۰۰ فریم در ثانیه نمونه‌برداری شده است.

شرمن چیس و همکاران [۷۴]، بودن [۷۵]، آکاکین و همکاران [۷۶] و اونگ و بودن [۷۷] از این پایگاه داده برای ارزیابی مدل‌هایشان استفاده کرده‌اند.

۷-۲-۳- پایگاه داده WAPUSK20

یکی دیگر از پایگاه داده‌های شامل جمله پایگاه داده WAPUSK20[۱۵] است که در سال ۲۰۱۰ از گروه GRID[۷] اتخاذ شده است و بیان جملات در شرایط واقع‌گرایانه در اتاق اداری معمولی اتفاق می‌افتد.

جنبه‌های کمی: از ۲۰ گوینده شامل ۹ زن و ۱۱ مرد برای بیان ۱۰۰ جمله منحصر به فرد استفاده شده است. این جملات شامل شش کلمه انگلیسی است که با استفاده از گزینه‌های دستور، رنگ، حرف اضافه، حرف، رقم و قید ترکیب می‌شوند. هر جمله در ۳ ثانیه بیان شده است و همه گویندگان به جز ۲ نفر جملات را در یک جلسه یک ساعتی بیان کردند. تنها ۲ نفر انگلیسی بومی هستند، یک نفر در اسپانیا، یک نفر در یونان، یک نفر در قزاقستان بزرگ شده‌اند و بقیه افراد آلمانی بومی هستند. میانگین سنی گویندگان ۲۹ سال است.

جنبه‌های کیفی: داده‌های صوتی با استفاده از یک جفت میکروفن OKTAVA MK 012 و یک جفت میکروفن Behringer ECM8000 ذخیره شده است. داده‌های تصویری از یک دوربین دید رنگی استریو، Bumblebee2 نوع BB2-03S2 با وضوح ۶۴۰ در ۴۸۰ پیکسل استفاده شده است که به صورت بی‌سیم به رایانه شخصی متصل می‌شود. یک تصویر نیز با دوربین وب لاجیتک‌کوئیک‌کم‌پرو ۹۰۰۰ که با USB به رایانه شخصی متصل است وضعیت سر گوینده را مشخص می‌کند. نرخ فریم داده‌های ویدئویی ۳۲ فریم در ثانیه است که حداکثر سرعت در بالاترین وضوح تصویر رنگ قابل دستیابی است. چهار کانال صوتی به صورت سیگنال ۳۲ بیتی با نرخ نمونه‌برداری ۱۶ kSamples / s ذخیره می‌شوند.

۸-۲-۳- پایگاه داده BL

پایگاه داده BL [۱۶] در سال ۲۰۱۱ به زبان فرانسوی جمع‌آوری شده است.

جنبه‌های کمی: شامل ۱۷ گوینده بومی بین ۲۳ تا ۴۸ سال که ۲۳۸ جمله از کلمات گفتاری پیوسته به زبان فرانسوی بیان می‌کنند. گویندگان دارای رژ لب آبی رنگ هستند. طول گفتار واقعی توسط گویندگان ۲۰ دقیقه است. فراوانی وقوع واج‌ها در پایگاه داده BL با فراوانی وقوع ۳۳ واج در زبان فرانسه مرتبط است. سیگنال صوتی به صورت دستی به جملات تقسیم می‌شود و سپس با یک فرایند حاشیه‌نویسی نیمه خودکار به واج‌ها تقسیم می‌شود. با استفاده از HTK[۵۲] از تراز اجباری استفاده و سپس مرزهای واج نیز به صورت دستی تنظیم شده است. در جلسه اول ۴ زن و ۴ مرد ۲۳۸ جمله را حدود ۳۴۰ دقیقه بیان می‌کنند. در جلسه دوم ۴ زن و ۵ مرد ۲۳۸ جمله را حدود ۱۸۰ دقیقه بیان می‌کنند.

جنبه‌های کیفی: دوربین‌ها و میکروفن‌ها بر روی میز نزدیک گوینده قرار دارند تا از نمای نزدیک با وضوح تصویر بالا و نسبت سیگنال به نویز مناسب سیگنال‌های صوتی استفاده کنند. یک نمایشگر استاندارد رایانه در جلوی گوینده قرار گرفته است که جملات را نشان می‌دهد. برای هر جلسه، کانال‌های صوتی با ۴۴/۱ کیلوهرتز با ۱۶ بیت نمونه‌برداری می‌شوند. از یک میکروفن AKG و یک میکروفن همه کاره Labtec استفاده شده است. فیلم‌ها در اتاق ضبط خاصی که عایق صدا با بازتاب کم است ضبط شده است. در جلسه اول داده‌های ویدئویی با دوربین رنگی جلویی کَنون MVX3i با ۲۵ فریم در ثانیه و وضوح تصویر ۵۷۶ در ۷۲۰

و دو میکروفن ضبط شده‌اند که می‌توانند برای تجزیه و تحلیل حرکات دو بعدی از دهان استفاده شود. داده‌های جمع‌آوری شده توسط جلسه دوم از دو دوربین کالیبره شده، یک دوربین عمق و دو میکروفن استفاده کرده‌اند و برای آنالیز سه بعدی استفاده می‌شود که دوربین رنگی نمای جانبی با سرعت ۳۰ فریم در ثانیه با وضوح تصویر ۶۴۰ در ۴۸۰ ضبط می‌شود. دوربین عمق، یک دوربین میکروسافت Kinect، ۳۰ تصویر در عمق در ثانیه را با وضوح ۶۴۰ در ۴۸۰ ضبط می‌کند. اتاق ضبط فقط با یک نور مصنوعی روشن می‌شود. هر فیلم با کدک MPEG4 فشرده شده است.

۹-۲-۳- پایگاه داده UNMC-VIER

پایگاه داده UNMC-VIER[۱۷] مجموعه گسترده‌ای از افراد با لهجه، رنگ پوست و موهای صورت را در خود جای داده و در ۶ روز در سال ۲۰۱۱ به زبان انگلیسی جمع‌آوری شده است.

جنبه‌های کمی: این پایگاه داده مشتمل بر ۱۲۳ نفر ۷۴ مرد و ۴۹ زن که دارای ترکیب قومی ۱۱۶ آسیایی، ۴ آفریقایی و ۳ اروپایی از دانشجویان کارشناسی، کارشناسی ارشد و کارمندان دانشگاه است. تغییر در میزان صدا و سرعت خواندن گفتار در حین ضبط وجود دارد. برای ضبط ویدیو نیز تغییرات بصری از قبیل حرکت سر، نور، بیان صورت، وضوح تصویر و پیش زمینه‌های پیچیده گنجانده شده است. فرایند جمع‌آوری داده‌ها در محیط کنترل شده و در محیط بدون کنترل تکرار می‌شود. محیط کنترل شده با پس‌زمینه آبی و روشنایی فلورسنت ثابت شده است. محیط کنترل نشده از سه مکان مختلف در یک اتاق برای داشتن شرایط مختلف نورپردازی و از نورپردازی طبیعی برای ایجاد تأثیر تغییر نور در چهره سوژه استفاده شده است زیرا نور به شدت تحت تأثیر آب و هوا (موقعیت خورشید، ابری یا بارانی) قرار دارد.

جنبه‌های کیفی: از دوربین‌های فیلم‌برداری با کیفیت بالا و دوربین‌های وب با کیفیت پایین در جمع‌آوری داده‌ها استفاده می‌شوند. از افراد خواسته می‌شود در حالی که سر خود را به سمت دوربین با زوایای ۰، ۳۰، ۶۰، ۹۰، ۱۲۰، ۱۵۰، ۱۸۰ درجه، بالا، پایین و جلوی دوربین فیلم‌برداری می‌چرخانند، ۰ تا ۹ را در هر دو محیط بخوانند. هر گوینده با دو صورت از قبیل چهره‌های شاد، اخمی، غمگین، غافلگیر، عصبانی، خواب آلود و خنثی که راحت است جملات را می‌خواند. برخی از افراد مایل به بیان بیش از دو صورت هستند. هر جمله توسط هر فرد ۵ تا ۱۴ بار با سرعت‌های مختلف تکرار می‌شود. ۱۴ فیلم در محیط کنترل شده و ۱۴ فیلم در محیط کنترل نشده از نمای روبه‌روی هر فرد یعنی ۲۸ فیلم وجود دارد که باید دو برابر شود زیرا وب کم در هر دو محیط به طور همزمان با دوربین فیلم‌برداری ضبط می‌شود. ۱۴ فیلم دیگر نیز برای نمای سمت چپ هر فرد در محیط کنترل شده وجود دارد. بنابراین مجموعا ۷۰ فیلم برای هر فرد در این پایگاه داده وجود دارد.

در محیط کنترل شده، از دو دوربین فیلم‌برداری پاناسونیک SDR-SW20 PAL استفاده شده است. یکی از دوربین‌های فیلم‌برداری از نمای روبه‌روی فرد و دیگری از نمای سمت چپ فرد تصویر می‌گیرد. وضوح تصویر دوربین‌های فیلم‌برداری ۷۰۸ در ۶۴۰ پیکسل است. فیلم با استفاده از وضوح نمونه‌برداری رنگی ۴:۲:۰ و ۲۵ فریم در ثانیه ضبط می‌شود. صدا با استفاده از نمونه‌های استریو ۱۶ بیتی با فرکانس ۴۸ کیلوهرتز با رمزگذاری mp2 ضبط می‌شود. وب‌کم مورد استفاده در محیط کنترل شده لاجیتک کوئیک‌کم پرو ۴۰۰۰ با وضوح تصویر ۳۲۰ در ۲۴۰ پیکسل است. این ویدئو با فرمت ۲۹ فریم در ثانیه با فرمت AVI خام گرفته شده است. صوت با استفاده از نمونه‌های مونو ۱۶ بیتی با فرکانس ۲۲ کیلوهرتز با رمزگذاری PCM ضبط می‌شود. در همین محیط از میکروفن داخلی وب کم با ۱۶ بیت در هر نمونه، نرخ نمونه‌برداری ۴۴ کیلو هرتز و حالت استریو برای ضبط گفتار به صورت جداگانه استفاده می‌شود.

از طرف دیگر، در محیط کنترل نشده، از یک دوربین فیلم‌برداری پاناسونیک پال SDR-S7 و یک سری لاجیتک‌کوئیک‌کم E3500 برای گرفتن نمای روبه‌روی فرد با وضوح تصویر ۳۲۰ در ۲۴۰ پیکسل با ۱۵ فریم در ثانیه با فرمت WWV خام استفاده می‌شود. جزئیات دوربین فیلم‌برداری مشابه مواردی است که در محیط کنترل شده استفاده می‌شود. صوت با استفاده از نمونه‌های مونو ۱۶ بیتی با فرکانس ۳۲ کیلوهرتز با رمزگذاری WMV2 ضبط می‌شود. فایل‌های صوتی گفتار به طور جداگانه توسط یک میکروفن کلیپ با تنظیمات یکسان با میکروفن مورد استفاده در محیط کنترل شده ضبط می‌شوند.

۱۰-۲-۳- پایگاه داده MOBIO

در سال ۲۰۱۲ پایگاه داده MOBIO [۱۸] به زبان انگلیسی به صورت منحصر به فرد جمع‌آوری شد چراکه جملات بیان شده توسط موبایلی که در دست گوینده قرار می‌گیرد ضبط می‌شود و در نتیجه محیطی کنترل نشده از شرایط نمایش و نور، روشنایی و پیش زمینه، تنوع بالا در کیفیت گفتار، و تنوع در محیط کسب از نظر آکوستیک دارد. این داده‌ها در ۶ سایت مختلف طی یک سال و نیم با افرادی که انگلیسی صحبت می‌کنند، ضبط شد.

جنبه‌های کمی: فیلم‌های این پایگاه داده شامل بیش از ۶۱ ساعت از داده‌های صوتی تصویری با ۱۲ جلسه مجزا است که معمولاً طی چند هفته از هم جدا می‌شوند. در مجموع، برای هر ۱۵۰ شرکت‌کننده ۱۹۲ نمونه منحصر به فرد صوتی و تصویری وجود دارد، این تقریباً دو برابر اندازه فاز اول بانک اطلاعاتی MOBIO[۱۸] است.

جنبه‌های کیفی: برای ضبط تنها اولین جلسه از لپ‌تاپ استاندارد ۲۰۰۸ مک‌بوک و بقیه داده‌ها از تلفن همراه نوکیا N93i استفاده شد. بر روی موبایل یک مدیر گفتگو نصب است که یکسری سوالات را از شرکت‌کنندگان می‌پرسد و پاسخ‌هایی را به

آنها نمایش می‌دهد تا پاسخ گویند. هر جلسه از مرحله اول شامل ۵ سؤال پاسخ کوتاه، ۵ سؤال کوتاه گفتار آزاد، ۱ متن از پیش تعریف شده و ۱۰ سؤال گفتار آزاد بود. فاز دوم کوتاه‌تر شد و شامل ۵ سؤال پاسخ کوتاه، ۱ متن از پیش تعریف شده و ۵ سؤال گفتار آزاد بود. در همه موارد از کاربران خواسته شد تا متنی از پیش تعریف شده یکسانی را به روشی طبیعی بخوانند که برای بیان بیش از ۱۰ ثانیه طراحی شده است و شرکت‌کنندگان در هنگام خواندن این جملات می‌توانند اصلاح خود را انجام دهند و یا کلمات دیگری را اضافه می‌کنند، همان‌طور که برای گفتار واقعی وجود دارد. ۵ سوال کوتاه عبارتند از:"نام شما چیست؟"، "آدرس شما چیست؟"، "تاریخ تولد شما چیست؟"، " شماره مجوز شما چیست؟"، و " شماره کارت اعتباری شما چیست؟" و سؤالات گفتار آزاد عبارتند از: سؤالات به طور تصادفی از لیستی از تقریباً ۴۰ سؤال انتخاب شده و پاسخ لازم نیست به سؤال مربوط باشد و فقط لازم بود که آنها تقریباً ۵ ثانیه صحبت کنند.

ترسادرن و همکاران [۷۸] از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۱۱-۲-۳- پایگاه داده RM-3000

پایگاه داده انگلیسی وابسته به گوینده با نام RM-3000[۲] در سال ۲۰۱۵ شامل ۳۰۰۰ جمله جمع‌آوری شد که توسط یک مرد بومی انگلیسی بیان شده است.

جنبه‌های کمی: این پایگاه داده ۳۰۰۰ جمله را به صورت تصادفی از ۸۰۰۰ جمله مجموعه مدیریت منابع [۳۵] گرفته است. مجموعه داده RM-3000 شامل ۴۵ واج منحصر به فرد، ۱۰۵۵۶ نشانه‌های واجی، ۹۷۹ کلمه منحصر به فرد، ۲۶۱۱۴ نشانه‌های کلمه، میانگین تعداد واج‌ها در هر جمله ۳۵/۱۹، میانگین تعداد کلمه‌ها در هر جمله ۸/۷ و میانگین تعداد واج‌ها در هر کلمه ۴/۰۴ است.

جنبه‌های کیفی: جملات در ۱۹ جلسه به مدت سه روز ثبت شد. فیلم‌ها با استفاده از یک دوربین Sanyo Xacti با کیفیت بالا با استفاده از اسکن پیش‌رونده با سرعت ۵۹/۹۴ فریم در ثانیه ضبط شدند. صدا با کیفیت بالا به طور همزمان با استفاده از میکروفن متصل به یقه پیراهن فرد در عمق ۱۶ بیتی و سرعت نمونه‌برداری از ۴۸ کیلوهرتز در کانال تک ضبط شد. فیلم‌ها با دوربین در حالت پرتره از روبه‌روی سوژه ضبط شده است و از ۱۹۲۰ پیکسل در ۱۰۸۰ پیکسل به ۳۶۰ در ۶۴۰ پیکسل یعنی یک سوم اندازه اصلی تقلیل می‌یابد. پس از ضبط فیلم‌ها برای استخراج فریم‌های تصویری پس از پردازش قرار گرفتند و یک مجموعه کوچک بین ۲۰ تا ۳۰ تصویر از هر جلسه با برچسب دستی برای تعریف علائم مشخص شد. برای بهبود دقت در روند ردیابی، مانند عکس‌برداری با داده‌های قبلی، علائم اضافی از جمله چشم، ابرو، بینی و خط فک تا پیشانی برچسب‌گذاری شده است. با استفاده از تصاویر دارای برچسب دستی و AAM، نقاط برجسته با استفاده از الگوریتم ترکیبی معکوس ردیابی می‌شوند.

هاول و همکاران [۴۷]، سانگسای و همکاران[۱۲۵] از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۱۲-۲-۳- پایگاه داده TCD-TIMIT

پایگاه داده بعدی TCD-TIMIT[۱۹] است که این مجموعه داده انگلیسی نیز در سال ۲۰۱۵ از جملات پایگاه داده TIMIT[۵۱] تولید شد که پوشش مناسبی از ویزم‌ها دارد و اکثر واج‌های موجود را پوشش می‌دهد. جملات TIMIT شامل دو جمله SA است تا لهجه گویندگان را برجسته کند؛ ۴۵۰ جمله SX شامل جفت واج‌های مختلف ممکن که به طور دستی طراحی شده است و ۱۸۹۰ جمله SI که از کتاب‌های نمایش‌نامه‌نویسی انتخاب شده است تا شامل جفت واج‌های غیر معمول متن‌ها باشد. هر گوینده در TIMIT[۵۱] ۱۰ جمله را بیان می‌کند دو جمله SA، پنج جمله SX و سه جمله SI. با آزمایش‌های مختلف نشان داده شد که ۹۸ جمله بیشترین تعداد جملاتی است که داوطلبان بدون دلسرد کردن بیان می‌کنند. برای ساختن متن‌های ۹۶ جمله‌ای متن‌های گوینده TIMIT به گروه‌های ۱۲ تایی تقسیم می‌شوند و حال هر گوینده ۸ جمله از هر ۱۲ گوینده TIMIT (۸×۱۲=۹۶) را بیان می‌کند.

جنبه‌های کمی: گویندگان ۶۲ نفر هستند که ۳ نفر زن متخصص لب‌خوانی با میانگین سنی ۶۰ سال هستند و ۵۹ نفر مابقی داوطلبان معمولی از دانشگاه هستند که ۳۲ نفر مرد و ۲۷ نفر زن با میانگین سنی ۲۴ سال با حداقل ۱۶ سال و حداکثر ۵۷ سال سن هستند.

جنبه‌های کیفی: از دو دوربین سونی PMW-EX3 استفاده شده است. یک دوربین تصویر از نمای روبه‌رو و دوربین دوم تصویر با زاویه ۳۰ درجه از سمت راست گوینده را می‌گیرد. همین‌طور میکروفن‌های خارجی موجود در آنها هم‌ترازی صوت و تصویر را نیز فراهم می‌کنند. رزولوشن این فیلم‌ها ۱۹۲۰ در ۱۰۸۰ پیکسل و سرعت ضبط ۳۰ فریم بر ثانیه است. از میکروفن بی‌سیم Shure PG185 که دارای فرستنده PG1 و گیرنده PG4 بود استفاده شد. اتاق ضبط عایق صدا نبود و سایر میکروفن‌ها نویزها را می‌گرفتند و در نهایت فیلم‌ها کم‌ترین سر و صدای خارجی را به خود گرفتند. میکروفن نزدیک به دهان و زاویه‌دار به سمت دهان هر گوینده قرار گرفته است. گویندگان هر جمله را با مکث ۲ ثانیه‌ای از جمله بعدی بیان کرده‌اند و سعی شده است که هر جمله با دهان بسته شروع و به پایان برسد. یک دقیقه فیلم از یکی از دوربین‌ها تقریباً ۲۶۰ مگابایت است. میانگین طول فیلم تقریبا ۱۵ دقیقه بود. پس از ضبط جملات ۵۹ گوینده، ۴۵۰ گیگابایت فیلم خام ضبط شده است.

فایل برچسب‌های سطح واج برای گفتار صوتی تولید شده است که از طریق نگاشت واج به ویزم این برچسب‌ها، فایلی برای گفتار بصری نیز ایجاد شده است. پایگاه داده اصلی TIMIT[۵۱]

شامل فایل برچسب‌های سطح کلمه‌ای و واجی برای همه جملات است. بیشتر کارهای منتشر شده روی TIMIT از مجموعه‌ای از واج‌های کاهش یافته استفاده می‌کنند و این مجموعه کاهش یافته در TCD-TIMIT استفاده می‌شود. فایل‌های برچسب واجی TIMIT توسط دانشمندان آوایی ایجاد شده است. هم‌ترازی اجباری برای داده‌های TCD-TIMIT با استفاده از ابزار P2FA که شامل مجموعه‌ای از مدل‌های سازگار با HTK[۵۲] و فایل‌های دیگر است و یک اسکریپت پایتون است که HTK را در حالت تراز اجباری تنظیم و فراخوانی می‌کند استفاده شد. اسکریپت پایتون یک فایل سطح واج تزار اجباری را برای یک کلیپ گفتار مشخص شده و فایل برچسب سطح کلمه خود تولید می‌کند. این کار با استفاده از فرهنگ لغت تلفظ انجام می‌دهد که شامل بیش از ۱۳۴۰۰۰ کلمه است. فرهنگ لغت تلفظ برای به دست آوردن رونوشت کلمات موجود در فایل سطح کلمه استفاده می‌شود. HTK[۵۲] همه رونوشت‌های متعدد یک کلمه را امتحان و دقیق‌ترین آنها را انتخاب می‌کند.

استرپو و همکاران [۷۹] و [۸۳]، گبی و همکاران [۸۰]، مورون و همکاران [۸۱]، افرت و همکاران [۸۲]، استرپو و همکاران[۸۳]، استرپو و هارته[۸۴] و سانگسای و همکاران [۸۵]، کیو و همکاران [۱۵۱]، از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

عبدالحمید و همکاران[۱۲۶]، سیلور و پتیل[۱۲۷]، ژانگ و همکاران[۱۲۸]، بومیک و مندل[۱۲۹]، هد و همکاران[۱۳۰]، فنگور و همکاران[۱۳۱]، و نعیم و بگ[۱۳۲] از پایگاه داده TIMIT استفاده کرده‌اند.

۱۳-۲-۳- پایگاه داده LRW

پایگاه داده LRW [۴۶] در سال ۲۰۱۶ با هدف شناسایی کلمات در سطح تک واژه‌ای به زبان انگلیسی معرفی شده است. از آرشیوهای ویدئویی برنامه‌های خبری و گفت‌وگوهای تلویزیونی بی‌بی‌سی انجام شده است.

جنبه‌های کمی: این پایگاه داده شامل ۵۰۰ کلمه متفاوت است که هر کلمه توسط صدها گوینده در ویدئوهای کوتاه تقریباً ۱/۱۶ ثانیه‌ای (۲۹ فریم با نرخ ۲۵ فریم بر ثانیه) بیان شده است. برای هر کلمه تا ۱۰۰۰ نمونه ادا شده توسط افراد مختلف وجود دارد که در مجموع حدود ۵۰۰,۰۰۰ نمونه ویدئو است. هدف اصلی پایگاه داده LRW شناسایی و تشخیص دقیق تک‌واژه‌ها از لب‌خوانی در محیط‌های طبیعی و بدون کنترل است، نه جملات طولانی یا ساختاری پیچیده.

جنبه‌های کیفی: فیلم‌ها با سرعت ۲۵ فریم بر ثانیه ضبط شده‌اند. طول ویدئوها ثابت و برابر ۲۹ فریم (حدود ۱/۱۶ ثانیه) می‌باشد و فرمت ویدئوها MP4 است. رزولوشن ویدئوها ثابت است ولی مقدار دقیق رزولوشن اعلام نشده است.

مصباح و همکاران [۱۲۲]، چیونگ و همکاران [۳]، افوراس و همکاران [۳۲]، زو و همکاران [۶۷]، مصباح و همکاران [۱۲۲]، پارخ و همکاران [۱۰۵]، وانگ و همکاران [۱۵۵]، مارتینز و همکاران[۱۵۶]، ما و همکاران [۱۵۷]، وکیم و همکاران [۱۱۸]، [۱۲۰]، [۱۳۸]، [۱۳۹] و چنگ و همکاران [۱۳۳] از این پایگاه داده استفاده کرده‌اند.

۱۴-۲-۳- مجموعه داده HAVRUS

مجموعه داده HAVRUS[۲۰] به زبان روسی در سال ۲۰۱۶ جمع‌آوری شده است.

جنبه‌های کمی: ۲۰ گوینده روسی تک زبانه بومی شامل ۱۰ مرد و ۱۰ زن بدون مشکل زبان یا شنوایی در ضبط‌ها شرکت کردند. هر یک از آنها ۲۰۰ عبارت روسی را تلفظ می‌کنند که ۱۳۰ عبارت برای آموزش ۲ متن غنی از نظر آوایی و ۷۰ عبارت برای تست برای هر گوینده. ۲۰ عبارت برای دستورالعمل‌های اطلاعات MIDAS در SPIIRAS[۵۳] و ۵۰ عبارت شماره تلفن هستند.

جنبه‌های کیفی: این مجموعه داده با فایل‌های ویدئویی بدون فشرده‌سازی با وضوح ۶۴۰ ×۶۴۰ پیکسل با ۲۰۰ فریم برثانیه ضبط شده است و فایل‌های صوتی نیز بدون فشرده‌سازی با PCM WAV تک کاناله با سرعت نمونه‌برداری ۱۶ کیلوهرتز است. همین‌طور دارای یک فایل متنی از زیرنویس‌های عبارات، کلمات، فونم‌ها و ویزم‌ها برای بخش آموزش است. برای داده‌های صوتی ضریب فرکانس مل از ۲۶ آنالیز بانکی فیلتر کانال از فریم‌های ۲۰ میلی ثانیه با گام ۵ میلی ثانیه محاسبه می‌شود که به عنوان پارامترهای آکوستیک ذخیره می‌شوند. ماژول پردازش سیگنال‌های ویدئویی بردارهای ویژگی تحریک کننده ۱۰ بعدی با فرکانس ۲۰۰ هرتز را به عنوان نتیجه تشخیص چهره و دهان چند مقیاس در فریم‌های ویدئویی با استفاده از طبقه‌بندی کننده‌های آبشار با AdaBoost و سپس استفاده از آنالیز مؤلفه‌های اصلی و آنالیز تفکیک خطی به ناحیه نرمال گرافیکی دهان محاسبه می‌کند. از دوربین سرعت بالای JAI Pulnix RMC-6740 با رزولوشن ۶۴۰ × ۴۸۰ و ۲۰۰ فریم بر ثانیه و یک میکروفن پویای Oktava MK-012 جهت ضبط اطلاعات استفاده شده است. مدت زمان هر فریم ویدیویی با سرعت ۲۵ فریم در ثانیه ۴۰ مگابایت است. ایوانکو و همکاران [۸۶] از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۱۵-۲-۳- مجموعه داده LRS

در سال ۲۰۱۷ مجموعه داده LRS[۳] منتشر شد که این مجموعه داده از برنامه‌های مختلف تلویزیونی بی‌بی‌سی به زبان انگلیسی بین سال‌های ۲۰۱۰ تا ۲۰۱۶ جمع‌آوری شده است.

جنبه‌های کمی: LRS با بیش از هزار گوینده شامل ۱۵۸۴ ساعت اخبار متشکل از ۵۰۴۹۳ جمله، ۱۹۹۷ ساعت برنامه با عنوان صبحانه متشکل از ۲۹۸۶۲ جمله، ۵۹۰ ساعت اخبار شبانه

متشکل از ۱۷۰۰۴ جمله همگی از شبکه یک بی‌بی‌سی و همین‌طور ۱۹۴ ساعت اخبار جهان متشکل از ۳۵۰۴ جمله و ۳۲۳ ساعت برنامه زمان پرسش متشکل از ۱۱۶۹۵ جمله از شبکه دو بی‌بی‌سی و ۲۷۲ ساعت برنامه جهان امروز شامل ۵۵۵۸ جمله از شبکه چهار بی‌بی‌سی که روی هم ۴۹۶۰ ساعت فیلم با ۱۱۸۱۱۶ جمله است. انتخاب این برنامه‌ها به دو دلیل است: (۱) طیف گسترده‌ای از گویندگان در اخبار و برنامه‌های بحث و گفتگو ظاهر می‌شوند، بر خلاف برنامه‌های درام با بازیگران ثابت؛ (۲) تغییرات شات کمتر رایج است، بنابراین جملات کامل با چهره‌های پیوسته وجود دارد.

جنبه‌های کیفی: ابتدا مرزهای شات تشخیص داده می‌شوند و تشخیص چهره مبتنی بر روش هیستوگرام شیب‌های جهت‌دار [۳۷] بر روی هر فریم ویدیو انجام می‌شود. تشخیص چهره از هر فرد نیز در سراسر فریم‌ها با استفاده از ردیاب KLT گروه‌بندی می‌شوند [۳۸].

برچسب زدن داده: زیرنویس در فیلم‌های بی‌بی‌سی همگام با صدا پخش نمی‌شوند. از آزمایشگاه Penn Phonetics Lab Forced Aligner[۶۱] و [۶۲] برای هم‌تراز کردن زیرنویس با سیگنال صوتی استفاده شد. خطاها در هم‌ترازی وجود دارد همان طور که زیرنویس کلمه به کلمه نیست، بنابراین برچسب‌های تراز شده با چک کردن در برابر سرویس تجاری گفتار به متن شرکت آی‌بی‌ام واتسون، فیلتر می‌شوند. فیلم‌ها به جملات یا عبارات با استفاده از نشانه‌گذاری در متن تقسیم می‌شوند. جملات با توقف کامل، کاما و علامت سوال از هم جدا می‌شوند؛ و به دلیل محدودیت‌های حافظه واحد پردازشگر گرافیکی به ۱۰۰ کاراکتر یا ۱۰ ثانیه تقسیم می‌شوند. یک مجموعه داده آموزش صوتی نیز با استفاده از جملات در برنامه‌های بی‌بی‌سی که توالی تصویر آنها در دسترس نیست نیز ایجاد کردند.

کورت‌نی و همکاران [۹۳]، چیونگ و همکاران[۲۱]، و چنگ و همکاران[۱۳۳] از این پایگاه داده استفاده کرده‌اند.

۱۶-۲-۳- مجموعه داده MV-LRS

مجموعه داده MV-LRS[۲۱] بر مبنای مجموعه داده LRS [۳] نیز در سال ۲۰۱۷ تولید شده است با این تفاوت که دارای ویدئوهایی است که نماهای مختلفی از گفتگو بین صفر تا ۹۰ درجه را شامل می‌شود یعنی از نمای روبه‌رو تا نیم‌رخ. از طرفی پایگاه داده LRS [۳] شامل اخبار است ولی پایگاه داده MV-LRS شامل برنامه‌های متنوعی از درام و برنامه‌هایی که مردم به طور واقعی با هم در حال گفتگو هستند. جملات این پایگاه داده شامل ۱۴۹۶۰ کلمه و ۷۴۵۶۴ سخن است.

چیونگ و زیسرمن [۲۱]، افوراس و همکاران [۳۲] و [۱۵۳]، ژائو و همکاران [۹۲] از این پایگاه داده برای آموزش مدل خود استفاده کرده‌اند.

۱۷-۲-۳- مجموعه داده VLRF

مجموعه داده VLRF [۲۲] نیز در سال ۲۰۱۷ به زبان اسپانیایی در یک محیط آزمایشگاهی ضبط شده است. در هر لحظه برای ضبط جملات دو نفر حضور دارند که یک نفر جملات را بیان می‌کند و دیگری در اتاقی مجزا با تماشای فیلم تنها تصویر نفر اول توسط یک صفحه تلویزیونی ۲۳ اینچی سعی در لب‌خوانی جملات گفته شده را دارد. که این کار امکان مقایسه لب‌خوانی توسط انسان و لب‌خوانی خودکار را فراهم می‌کند. هر جمله تا ۳ بار با گفتاری طبیعی و پیوسته تکرار می‌شود مگر آن که فرد لب‌خوان زودتر بتواند جمله را لب‌خوانی کند. از طرفی جملات بیان شده توسط گوینده و جملات حدس زده شده توسط لب‌خوان حاشیه‌نویسی می‌شوند. جلسات ضبط دارای ۴ سطح دشواری است به این ترتیب که ۳ سطح با ۶ جمله و یک سطح با ۷ جمله انجام می‌شود. در سطح اول جملات کوتاه با تعداد کلمات کم بین ۴ تا ۵ کلمه هستند که با افزایش سطح تعداد کلمات موجود در جمله تا حداکثر ۱۲ کلمه افزایش می‌یابد. جملات هر سطح نیز با هم بی ارتباط هستند. برای ایجاد انگیزه برای شرکت‌کنندگان و اطمینان از تمرکز آنها در پایان هر سطح نقش آنها با هم عوض می‌شود. در جمع‌آوری این مجموعه داده از افراد کم‌شنوا برای لب‌خوانی استفاده شده و نتایج به دست آمده برای آنها با افراد سالم مقایسه شده است.

جنبه‌های کمی: از ۲۴ داوطلب بزرگسال شامل ۳ مرد و ۲۱ زن استفاده کردند. سیزده نفر دانشجوی دانشگاه هستند، یکی معلم زبان امتحان UPF و ۱۰ شرکت‌کننده دیگر عضو فدراسیون انجمن والدین و ناشنوایان کاتالونیا (ACCAPS) هستند. داوطلبان به دو گروه افراد کم‌شنوا و افراد نرمال تقسیم شدند. ۱۵ شنونده با شنوایی عادی شامل ۱۴ زن و ۱ مرد که ۲ نفر از آنها بالای ۵۰ سال سن و سطح تحصیلات متفاوتی بودند و ۱۳ نفر مابقی در یک رده سنی و تحصیلات بودند. ۹ شرکت‌کننده کم‌شنوا هستند که شامل ۷ زن و ۲ مرد هستند و بالای ۳۰ سال سن دارند و از این ۹ نفر ۸ نفر در ابتدا دارای گفتار شنوایی بودند و سپس شنوایی خود را از دست می‌دهند ولی یک نفر از ابتدا ناشنوا بوده است. ۴ نفر از آنها نیز دارای سمعک و یا کاشت حلزون هستند. هر شرکت‌کننده ۲۵ جمله متفاوت از ۵۰۰ جمله را بیان می‌کند. در کل، ۱۰۲۰۰ کلمه وجود دارد (۱۳۷۴ یکتا). میانگین مدت زمان در هر جمله ۷ ثانیه و مدت کل پایگاه داده ۱۸۰ دقیقه و یا ۱۶۲۵۴۰ فریم است.

جنبه‌های کیفی: از یک دوربین پاناسونیک اچ‌پی‌ایکس۱۷۱ با یک سه پایه PRO6-HDV که در جلوی صندلی گوینده قرار گرفته است تا تصویر صورت تقریبا از روبه‌رو حاصل شود استفاده شد. میکروفن بر روی دوربین برای اطمینان از پوشش صدای گوینده نصب شده است. دوربین ضبط از نزدیک با سرعت ۵۰ فریم در ثانیه با وضوح ۷۲۰ × ۱۲۸۰ پیکسل و صدا در مونو ۴۸ کیلوهرتز با وضوح ۱۶ بیتی ثبت شده است. برای بدست آوردن نور یکنواخت و به حداقل رساندن سایه‌ها یا سایر مصنوعات

موجود در چهره گوینده، از دو Lumatek ultravight 1000W مدل ۵۳-۱۱ به همراه پانل‌های بازتابی استفاده شده است.

هر فریم که واج در آن لحظه تلفظ می‌شود برچسب زده می‌شود. از افزونه EasyAlign از Praat [۳۹] استفاده کردند که به شما امکان می‌دهد واج را در هر لحظه بر اساس جریان صوتی قرار دهید. به طور خاص، برنامه واج‌ها را به صورت نیمه خودکار قرار می‌دهد و معمولاً برای تطبیق مرزهای هر واج در موقعیت‌های دقیق‌تر، نیاز به مداخله دستی است. واج‌های بکار رفته بر اساس الفبای آوایی SAMPA [۴۰]، [۴۳] و [۴۴]، است که واژگان SAMPA برای زبان اسپانیایی از ۳۱ واج تشکیل شده است.

گومز و هینارجوس[۱۵۲] از این پایگاه داده استفاده کرده‌اند.

۱۸-۲-۳- مجموعه داده LRS3-TED

TED سازمانی غیرانتفاعی است که با شعار «گسترش ایده‌های ارزشمند» کنفرانس‌های سالیانه‌ای در موضوعات متنوع علمی، فناوری و فرهنگ برگزار می‌کند. سخنرانی‌ها کوتاه، تأثیرگذار و توسط افراد برجسته ارائه می‌شوند و ویدئوهای آن‌ها به‌صورت رایگان در دسترس عموم است. پایگاه داده LRS3-TED [۱۶۷] در سال ۲۰۱۸ از ویدئوهای این سخنرانی‌ها استخراج شده و به عنوان بزرگ‌ترین و غنی‌ترین پایگاه داده برای تشخیص گفتار بصری به زبان انگلیسی شناخته می‌شود و زیرمجموعه‌ای از LRS3 [۱۶۷] است .

جنبه‌های کمی: این مجموعه داده شامل بیش از ۴۰۰ ساعت ویدئو از سخنرانی‌های TED و TEDx می‌باشد. شامل حدود ۵۵۹۴ ویدئو از سخنرانی‌های متنوع با هزاران گوینده مختلف است.

جنبه‌های کیفی: ویدئوها با رزولوشن ۲۲۴×۲۲۴ پیکسل و نرخ ۲۵ فریم بر ثانیه آماده شده‌اند که صورت گویندگان به صورت کراپ شده استخراج شده است. صداهای همراه به صورت تک کاناله با نمونه‌برداری ۱۶ کیلوهرتز و عمق ۱۶ بیت ضبط شده‌اند.

لی و همکاران [۹۴]، ما و همکاران [۱۴۰]، [۱۶۱] و [۱۶۳]، پرجوال و همکاران[۱۴۲]، افوراس و همکاران [۱۵۳]، شی و همکاران [۱۶۲] از این پایگاه داده‌ها برای ارزیابی مدل خود استفاده کرده‌اند.

۱۹-۲-۳- مجموعه داده LRW-1000

مجموعه داده LRW-1000 [۱۶۶] در سال ۲۰۱۸ از برنامه‌های تلویزیونی خبری و گفتگوهای مرتبط با اخبار و وقایع جاری چین به زبان ماندارین جمع‌آوری شده است.

جنبه‌های کمی: این مجموعه داده شامل ۱۰۰۰ کلمه متفاوت است که در مجموع ۷۱۸,۰۱۸ نمونه دارد. داده‌ها از بیش از ۲۰۰۰ گوینده مختلف ضبط شده‌اند. این مجموعه دارای تنوع بسیار بالایی در مقیاس، وضوح تصویر، نویز پس‌زمینه و ویژگی‌های گوینده مانند زاویه دید، سن، جنسیت و آرایش است.

جنبه‌های کیفی: فیلم‌ها با سرعت ۲۵ فریم بر ثانیه ضبط شده‌اند. هر ویدیو از فریم‌های متعددی تشکیل شده که طول آن‌ها ثابت و برابر ۲۹ فریم، معادل ۱/۱۶ ثانیه است. فرمت ویدئوها MP4 است. رزولوشن ویدئوها ثابت است ولی مقدار دقیق رزولوشن اعلام نشده است.

وانگ و همکاران [۱۵۵]، مارتینز و همکاران[۱۵۶]، ما و همکاران [۱۵۷]، لن و همکاران[۱۵۸] و کیم و همکاران [۱۱۸] در تحقیقات خود از این پایگاه داده بهره برده‌اند.

۲۰-۲-۳- مجموعه داده CMLR

مجموعه داده CMLR [۱۵۹] که در سال ۲۰۲۰ معرفی شده، از برنامه‌های خبری ملی چین از سال ۲۰۰۹ تا ۲۰۱۸ به زبان چینی ماندارین جمع‌آوری شده است.

جنبه‌های کمی: داده‌های این مجموعه داده از ویدئوهای چهره‌های در حال صحبت همراه با زیرنویس استخراج شده‌اند، و روی جمله‌ها تمرکز دارد. جمله‌ها توسط ۱۱ گوینده شامل ۶ مرد و ۵ زن بیان شده است که حدود ۱۰۲,۰۷۲ جمله و حدود ۲۵,۶۳۳ عبارت که دربرگیرنده حدود ۳,۵۱۷ حرف چینی است را بیان می‌کنند.

جنبه‌های کیفی: برای برداشت ویدئوهای واضح‌تر به جای FFmpeg از OpenCV نسخه ۴ استفاده شده است. طبق استاندارد ضبط و پخش فیلم‌ها در چین سرعت ضبط ویدئوها ۲۵ یا ۳۰ فریم بر ثانیه (fps) است. رزولوشن ورودی تصاویر فریم به فریم ۱۲۸*۶۴ پیکسل است.

زو و همکاران [۱۴۹]، شوئه و همکاران [۱۵۰]، کیو و همکاران [۱۵۱]، لو و همکاران [۱۶۰]، و سان و همکاران [۱۶۴] نتایج تحقیقات خود را روی این مجموعه داده نشان داده‌اند.

۲۱-۲-۳- مجموعه داده NTSDB

مجموعه داده NTSDB [۱۶۸] در سال ۲۰۲۰ از برنامه‌های تلویزیونی چین مانند اخبار، سخنرانی و برنامه گفتگو از اینترنت به زبان ماندارین جمع‌آوری شده است.

جنبه‌های کمی: این پایگاه داده از برنامه‌های CCTV News و Logic Show انتخاب شده‌اند که در آنها گویندگان رو به دوربین فیلمبرداری هستند و تنها ویدیوهای تک نفره به صورت قطعه ۳ ثانیه‌ای برش خورده‌اند.

جنبه‌های کیفی: ویدئوها دارای نرخ فریم ویدئوها ۲۵ فریم در ثانیه هستند.

۲۲-۲-۳- مجموعه داده MAVS

مجموعه داده تلفن های هوشمند صوتی و تصویری چند زبانه MAVS [۱۴۵] در سال ۲۰۲۱ توسط ۱۰۳ گوینده بومی هندی به

۳ زبان زنده دنیا در ۳ جلسه مختلف برای هر گوینده که توسط ۵ گوشی تلفن همراه ضبط شده است، جمع‌آوری شده است.

جنبه‌های کمی: MAVS توسط ۷۰ مرد و ۳۳ زن هندی که میانگین سنی آنها ۲۷ سال است جمع‌آوری شده است. هر یک از گویندگان ۶ جمله را به ۳ زبان انگلیسی، هندی و بنگالی بیان می‌کنند که ۳ جمله برای همه افراد یکسان است و ۳ جمله دیگر برای هر گوینده منحصر به فرد است.

جنبه‌های کیفی: برای ضبط داده‌ها از پنج دستگاه تلفن هوشمند به نام‌های آیفون ۱۱، آیفون ۱۰، آیفون S6، سامسونگ S7 و سامسونگ S8 استفاده شده است. داده‌ها در ۳ جلسه مختلف ضبط شده‌اند. در جلسه اول هیچ نویزی وجود ندارد و دارای نور یکنواخت است. در جلسه دوم نویز کنترل شده با روشنایی یکنواخت اما متفاوت از جلسه اول است. و در جلسه سوم نویز کنترل نشده با پس‌زمینه طبیعی و نور غیریکنواخت است که بخش‌هایی از صورت گوینده تاریک است.

۲۳-۲-۳- مجموعه داده GLips

مجموعه داده GLips [۱۴۷] در سال ۲۰۲۲ به زبان آلمانی توسط شوبیرت و همکاران از بیش از ۱۰۰۰ ویدیو و زیرنویس از پارلمان هسیان آلمان، جمع‌آوری شده است. به طوری که زیرنویس‌ها به‌عنوان یک فایل متنی جداگانه در دسترس هستند. این مجموعه داده به گونه‌ای طراحی شده است که با پایگاه داده انگلیسی LRW[۴۶] سازگاری داشته باشد تا امکان استفاده در یادگیری انتقالی فراهم شود. هدف ساخت این پایگاه داده فراهم کردن داده‌های گسترده برای خواندن لب به زبان آلمانی و تحلیل‌های یادگیری عمیق است.

جنبه‌های کمی: این پایگاه داده شامل حدود ۲۵۰,۰۰۰ ویدئو عمومی از چهره گویندگان است. هر ویدئو برای خواندن لب در سطح کلمه به کار رفته و هر ویدئو یک کلمه را به مدت ۱,۱۶ ثانیه در بر دارد. پایگاه داده شامل ۵۰۰ کلمه متفاوت با طول بین ۴ تا ۱۸ کاراکتر است و هر کلمه دارای ۵۰۰ نمونه تصویری است.

جنبه‌های کیفی: فرمت ویدئوها H264 و به صورت MPEG-4 فشرده شده است. فیلم‌ها با رزولوشن ۲۲۴*۲۲۴ پیکسل و ۲۵ فریم بر ثانیه ضبط شد. دوربین‌ها از زوایای دید متنوع برای به دست آوردن نماهای چندگانه چهره استفاده کرده‌اند. داده‌ها شامل صدا و متادیتای متنی به صورت جداگانه بوده و همچنین فایل‌های کامل TextGrid که زمان‌بندی دقیق شروع و پایان هر کلمه یا بخش گفتار را مشخص می‌کنند.

۲۴-۲-۳- مجموعه داده CN-CVS

این مجموعه داده، بزرگ‌ترین مجموعه باز متنی زبان چینی برای پژوهش‌های گفتار تصویری محسوب می‌شود و در سال ۲۰۲۳ جمع‌آوری شده است [۱۷۱].

جنبه‌های کمی: بیش از ۲۵۵۰ نفر گوینده که شامل ترکیبی از زن و مرد با گستردگی سنی از نوجوان تا سالمند می‌شود. این تنوع باعث می‌شود گویندگان عمدتاً حرفه‌ای نباشند تا سبک‌های گفتاری متنوعی فراهم شود. مجموعاً بیش از ۲۰۰ هزار سخن با مدت زمان کل بیش از ۳۰۰ ساعت داده صوتی-تصویری است.

جنبه‌های کیفی: داده‌های این مجموعه شامل ۲ بخش مربوط به برنامه‌های خبری با دوربین‌های ثابت و زاویه مستقیم و بخشی دیگر مربوط به سخنرانی‌ها و برنامه‌های گفتاری اینترنتی با زوایا و شرایط ضبط متغیر است. صداها با کیفیت ۱۶ کیلوهرتز ضبط شده و تبدیل به اسپکتروگرام Mel شده‌اند. این مجموعه از نظر پوشش محتوای پیچیده، حجم، تنوع گویش، شرایط ضبط و تعداد گوینده گسترده است.

۲۵-۲-۳- مجموعه داده LUMINA

مجموعه داده LUMINA [۱۴۶] در سال ۲۰۲۴ به زبان اندونزیایی با ۱۴ گوینده بومی که هریک ۱۰۰۰ جمله را بیان می‌کنند، جمع‌آوری شده است.

جنبه‌های کمی: ۲۵۰۰ ترکیب از جملات که توسط گویندگان شامل ۹ مرد و ۵ زن که هر یک حداقل ۱۰۰۰ تا از آنها را بیان می‌کنند این مجموعه داده را تشکیل می‌دهد. جملات با توجه به پوشش هجاها در باهاسا اندونزی ایجاد شده‌اند و هر کلیپ ۳/۳ ثانیه با قرارگیری دقیق گوینده در جهت عمود بر دوربین ضبط شده است.

جنبه‌های کیفی: داده‌ها تحت شرایط خوب، در یک اتاق عایق صدای ۴×۴ متر با پس‌زمینه صفحه سبز ضبط شده است. دوربین Fujifilm XT-200 به صورت عمودی در مقابل گوینده در ۸۰ سانتی‌متر قرار گرفته است و ارتفاع دوربین با موقعیت دهان گوینده تنظیم می‌شود. برای نورپردازی Godox TL-4 و TaffStudi KS65 به صورت مورب در سمت راست و چپ اسپیکر قرار گرفته‌اند تا سایه‌های ناحیه صورت را از بین ببرند. پرومتر نیز جلوی گوینده در فاصله ۱۰۰ سانتی‌متری قرار می‌گیرد.

۲۶-۲-۳- پایگاه داده Arman-AV

پایگاه داده Arman-AV [۱۷۴]، در سال ۲۰۲۳ به زبان فارسی به صورت جامع و چندمنظوره‌ای جمع‌آوری شده است.

جنبه‌های کمی: این پایگاه داده شامل حدود ۲۲۰ ساعت ویدئو از گفتار ۱۷۶۰ گوینده غیرحرفه‌ای است. مجموعه حاضر دربرگیرنده بیش از ۸۹ هزار جمله و تقریباً ۲/۵ میلیون واژه از گفتار گویندگان مختلف است. داده‌ها از سه منبع اصلی جمع‌آوری شدند: مصاحبه‌های بدون مجری، سریال‌ها و فیلم‌ها، و ویدئوهایUGC از طریق جستجوی کلمات کلیدی منتخب. در مرحله پردازش، با بهره‌گیری از روش‌های پیش‌پردازش نظیر تشخیص صحنه، شناسایی و ردیابی چهره، ویدئوها به بخش‌های

قابل استفاده برای سامانه‌های لب‌خوانی و تشخیص گفتار تفکیک و آماده‌سازی شده‌اند.

جنبه‌های کیفی: داده‌های این پایگاه شامل ویدئوهایی با فرمت mp4 با رزولوشن ۲۲۴×۲۲۴ پیکسل و نرخ ۲۵ فریم بر ثانیه ارائه می‌شود.

۳-۲-۲۷- مجموعه داده DVS-Lip

مجموعه داده DVS-Lip [۱۶۵]، در سال ۲۰۲۵ به زبان انگلیسی جمع‌آوری شده است.

جنبه‌های کمی: ۴۰ داوطلب، ۲۰ نفر از هر جنسیت، پنج توالی از کلمات که هر کدام یک ترتیب تصادفی از تمام کلمات موجود در واژگان است، را بیان کردند. در مجموع ۱۹۸۷۱ نمونه پس از حذف دستی فایل‌های رویداد آسیب‌دیده جمع‌آوری شد. از این تعداد، ۱۴۸۹۶ نمونه از ۳۰ داوطلب برای آموزش استفاده شد، در حالی که ۴۹۷۵ نمونه باقی‌مانده از ده داوطلب دیگر برای ارزیابی استفاده شد. جنسیت داوطلبان در هر دو مجموعه آموزشی و آزمایشی متعادل بود.

جنبه‌های کیفی: دوربین رویداد مورد استفاده برای ضبط، DAVIS346 (۳۴۶ × ۲۶۰) بود که قادر به خروجی تصاویر شدت و جریان‌های رویداد به طور همزمان و بدون هیچ تفاوتی در دیدگاه بود. از Montreal Forced Aligner1 برای تعیین مرزهای زمانی کلمات جداگانه استفاده شد. تصاویر با شدت (۲۵ فریم در ثانیه) ضبط شدند. برش مکانی روی داده‌های اصلی انجام شده است تا یک برش ۱۲۸ × ۱۲۸ با محوریت دهان حفظ شود.

تان و همکاران [۱۶۵] روش پیشنهادی خود را روی این مجموعه داده آزمایش کردند و به نتایج خوبی رسیدند.

۳-۲-۲۸- مجموعه داده DVS-LRW100

این مجموعه داده با استفاده از v2e [۱۷۵]، یک شبیه‌ساز دوربین رویداد جدید، برای تولید جریان‌های رویداد از بزرگترین پایگاه داده لب‌خوانی مبتنی بر ویدئو در سطح کلمه، LRW [۴۶]، در سال ۲۰۲۵ جمع‌آوری شده است.

جنبه‌های کمی: این پایگاه داده شامل ۱۰۷۶۶۴ سخن است.

جنبه‌های کیفی: از آنجایی که ویدیوهای موجود در پایگاه داده LRW نرخ فریم ۲۵ فریم در ثانیه دارند، به نرخ فریم بالاتری ارتقا داده شدند تا وضوح زمانی بالایی از رویدادهای شبیه‌سازی شده داشته باشیم.

تان و همکاران [۱۶۵] با جمع‌آوری این پایگاه داده روش پیشنهادی خود را نیز روی آن آزمایش کردند و به نتایج خوبی رسیدند.

۳-۳- مجموعه داده‌های ترکیبی

آخرین گروه از پایگاه داده‌های گفتار پیوسته، پایگاه داده‌های ترکیبی است که به منظور چندین هدف طراحی شده‌اند. به عنوان مثال پایگاه داده AV@CAR [۲۴] که به منظور شناسایی حروف الفبا، ارقام و جملات جمع‌آوری شده است و یا پایگاه داده MIRACL-VC [۲۸] که ترکیبی از کلمات و عبارات است. در ادامه مجموعه داده‌های ترکیبی به تفکیک شرح داده می‌شوند.

۳-۳-۱- مجموعه داده XM2VTSDB

مجموعه داده XM2VTSDB [۵۴] توسعه‌یافته M2VTS [۵۵] است که در سال ۱۹۹۹ به زبان فرانسوی جمع‌آوری شد و تفاوت اصلی آنها در اندازه بانک اطلاعاتی و تعداد ضبط‌های گرفته شده برای هر گوینده در طول هر جلسه است.

جنبه‌های کمی: این پایگاه داده توسط ۲۹۵ گوینده شامل ۸۸۵ سخن که با ۱۴۶۶ بار تکرار با هدف شناسایی ارقام ایجاد شد. هر گوینده ۳ جمله شامل دو دنباله عددی تصادفی و یک جمله که از نظر آوایی متعادل است را ۲ بار بیان می‌کنند.

جنبه‌های کیفی: با دوربین دیجیتال سونی VX1000E و دوربین VCR دیجیتال DHR1000UX تصویربرداری شده است. ویدئوی ۵۹ دقیقه‌ای با رزولوشن ۵۷۶ ×۷۲۰ و ۲۵ فریم در هر ثانیه است. هر فرد روبه‌روی دوربین روی صندلی نشسته و یک میکروفن به پیراهن او برای ضبط صدا وصل است که با وضوح نمونه‌برداری ۱۶ بیتی با فرکانس ۳۲ کیلوهرتز ضبط می‌شود. فیلم‌ها در فواصل یک ماه به مدت ۵ ماه گرفته شده است. ضبط در ۴ جلسه با تنظیمات یکسان صورت پذیرفت، و نور از دو طرف چپ و راست تابیده می‌شود و از یک پرده آبی به عنوان پس‌زمینه استفاده شد. بانک اطلاعاتی خام حاوی تقریبا ۳۰ ساعت ضبط فیلم دیجیتال است که به صورت دستی حاشیه نویسی شده است.

۳-۳-۲- مجموعه داده‌های AVOZES

مجموعه داده AVOZES [۲۳] در اوت ۲۰۰۰ به مدت یک هفته و آگوست ۲۰۰۱ به مدت ۲ روز در آزمایشگاه علوم کامپیوتر که عایق صوتی است در دانشگاه ملی استرالیا شامل ۲۰۰ سخن از ارقام و ۶۰ سخن از جملات با ۵۵ بار تکرار جمع‌آوری شده است.

جنبه‌های کمی: ۲۰ گوینده شامل ۱۰ زن و ۱۰ مرد ۲ ساعت ویدئو ایجاد کردند. علاوه بر ۲۰ گوینده ۴ نفر که پیشینه زبان دیگری هم داشتند استفاده شدند. به گونه‌ای که دو نفر از آنها زبان انگلیسی دارند (انگلستان، نیوزیلند) و سومین گوینده که آلمانی را به عنوان اولین زبان خود صحبت می‌کند، ولی یک سال را در انگلستان و دو سال در زمان ضبط را در استرالیا گذرانده بود در جلسه اول ضبط استفاده شدند و در جلسه دوم ضبط، نفر چهارم که انگلیسی غیر بومی با لهجه چینی که در زمان ضبط ۶ سال در استرالیا زندگی کرده بود استفاده شد. البته داده‌های افراد غیر بومی منتشر نشده است. شش گوینده عینک، سه نفر رژ لب، دو نفر هم ریش دارند. طراحی بدنه داده‌های AVOZES [۲۳] کلیه ویژگی‌ها و تقریبا همه واج‌های استرالیایی انگلیسی را پوشش می‌دهد که

جامع‌ترین مجموعه داده گفتاری صوتی تصویری در AuE تا به آن روز است.

جنبه‌های کیفی: میکروفن Sennheiser MKE 10-3 که به لباس گوینده با فاصله حدود ۲۰ سانتی متری زیر دهان وصل شده است با فرکانس۵۰Hz-20kHz صدای مونو را در نوار DV با ۴۸ کیلوهرتز ضبط می‌کند. فرمت داده‌های صوتی WAV هستند. دو دوربین فیلم‌برداری NTSC آنالوگ استاندارد در کنار هم روی یک دکل نصب شده‌اند و با فرکانس ۲۹/۹۷ هرتز تصاویر از روبه‌رو با سرعت ۳۰ فریم در هر ثانیه را با رزولوشن ۴۸۰ ×۷۲۰ با فرمت AVI ضبط می‌کنند. فاصله دوربین‌ها تا چهره در حدود ۶۰۰ ± ۵۰ میلی‌متر است که مربوط به مسافت یا عمق دوربینی است که دوربین‌ها برای آن کالیبره شدند. نورپردازی هم ثابت است و رنگ پس‌زمینه صفحه نمایش بین سبز تیره و آبی تیره در هر زمان نمایش سریع جابجا می‌شود. حرکت سر برای گویندگان مجاز است. این مجموعه داده در [۹۵] با جزئیات کامل شرح داده شده است.

سارگی و جوئك [۹۶]، چتی و واگنر [۹۷] و [۹۸] از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۳-۳-۳- مجموعه داده BANCA

مجموعه داده BANCA[۵] در سال ۲۰۰۳ به ۴ زبان انگلیسی، فرانسوی، ایتالیایی و اسپانیایی شامل عدد ۱۲ رقمی، نام، آدرس و تاریخ تولد جمع‌آوری شد.

جنبه‌های کمی: در مجموع ۲۰۸ نفر که برای هر زبان ۵۲ نفر شامل ۲۶ زن و ۲۶ مرد داده‌ها را بیان کردند. ویدئوها در محیط‌های کنترل شده، تخریب شده و نامطلوب در مدت ۳ ماه ضبط شده‌اند.

جنبه‌های کیفی: از دو دوربین یکی وب‌کم آنالوگ با کیفیت پایین و دیگری دوربین دیجیتال با کیفیت بالا با فرمت MPEG-7 ضبط استفاده شد و به همین تناسب از دو میکروفن ۱۲ بیتی و ۱۶ بیتی با فرکانس ۳۲ کیلوهرتز برای ضبط استفاده شده است.

مسر و همکاران [۹۹] از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۴-۳-۳- مجموعه داده AV@CAR

مجموعه داده AV@CAR[۲۴] در سال ۲۰۰۴ شامل ۸۰۰ سخن از حروف الفبا با ویدئوی ۱ ساعتی و ۶۰۰ سخن از ارقام با ویدئوی ۵۰ دقیقه‌ای و ۶۰۰۰ سخن از جملات با ویدئوی ۸ ساعتی با ۲۶ بار تکرار به زبان انگلیسی جمع‌آوری شده است.

جنبه‌های کمی: این مجموعه داده با ۲۰ گوینده بین ۲۵ تا ۵۰ سال متشکل از ۱۰ زن و۱۰ مرد است. کالبد صوتی تصویری این پایگاه داده را می‌توان به دو بخش اصلی تقسیم کرد. اولی در هنگام رانندگی داخل خودرو ضبط می‌شود و دومی در محیط بدون کنترل صدا جمع‌آوری می‌شود. بخش اتومبیل پایگاه داده از هفت کانال صوتی، یک کانال ویدیویی و اطلاعاتی در مورد سرعت خودرو، شرایط جاده، هوا، ترافیک و همچنین اطلاعات مربوط به گوینده و شرایط روشنایی تشکیل شده است. از طرف دیگر، ضبط‌های آزمایشگاهی با استفاده از پنج کانال صوتی و یک کانال ویدیویی، به علاوه یک جلسه تصاویر سه بعدی برای هر راننده نیز انجام شده است. هر گوینده در حین ضبط دارای حالت‌های مختلفی شامل خوشحالی، تعجب، خمیازه، خشم، انزجار، ترس و غم و اندوه است.

جنبه‌های کیفی: از میکروفن Q501T که برای محیط خودرو مناسب است برای ضبط صدا استفاده شده است. شش میکروفن در قسمت بالای اتومبیل قرار داده شده است که سه تای آنها بالای صندلی‌های جلو و مابقی در صندلی‌های عقب قرار دارند. برای جلسات آزمایشگاهی از میکروفن‌های برقی ارزان قیمت به همراه میکروفن آرایه صوتی VA 2000، میکروفن C477WR و یک Q501T استفاده شد. به منظور ضبط صدای دور، دو میکروفن برقی در گوشه‌های فوقانی یک کابین ۱/۸۶m x2.86x2.11 قرار دارند. آرایه NetCom و میکروفن Q501T در جلوی گوینده ۱ متر دور از او قرار گرفته‌اند. بخش صوتی بانک اطلاعاتی در ۱۶ کیلوهرتز و ۱۶ بیت برای هر کانال نمونه‌برداری شده است. برای پایگاه داده ویدئویی، یک دوربین کوچک Marshall ElectronicsUSA V-1204A که صفحه دوربین آن شامل شش LED مادون قرمز است تا نور کافی را تضمین کند در کنار آینه عقب قرار داده شده است تا چهره راننده را ضبط کند. این تصاویر به صورت سیاه و سفید، دیجیتالی شده با وضوح مکانی ۷۶۸ در۵۷۶ پیکسل، عمق پیکسل ۸ بیتی، سرعت ۲۵ فریم در ثانیه، با استفاده از DT3120 است. ضبط‌های آزمایشگاهی به دو بخش تقسیم می‌شوند. در حالت اول، گوینده هنگام تکرار کارهایی که در داخل ماشین انجام می‌شود، با همان دوربین مدل V-1204A و با شرایط روشنایی کنترل شده و نمای روبه‌روی صورت ضبط می‌شود. در بخش دوم، چندین تصویر با استفاده از سامانه ضبط سه بعدی از Vision RT Ltd به گوینده گرفته شده است. این سامانه ۶ تصویر همزمان از فرد می‌گیرد که ۴ تصویر برای بازسازی هندسه سه بعدی و ۲ تصویر باقی‌مانده برای گرفتن اطلاعات بافت سیاه و سفید و ایجاد یک توری سه بعدی از صورت است. تصاویر از نماهای روبه‌رو، سمت چپ، فوقانی، با عینک شفاف و با عینک آفتابی گرفته شده است.

۵-۳-۳- پایگاه داده AVICAR

همچنین در سال ۲۰۰۴ پایگاه داده AVICAR[۲۵] به زبان انگلیسی شامل ۲۶ سخن از حروف الفبا و ۱۳ سخن از ارقام و ۱۳۱۷ سخن از جملات با ۲۶ بار تکرار درون یک خودرو جمع‌آوری شد.

جنبه‌های کمی: این پایگاه داده از ۱۰۰ نفر متشکل از ۵۰ زن و ۵۰ مرد تشکیل شده است. حدود ۶۰٪ از گویندگان زبان مادری

انگلیسی آمریکایی دارند، در حالی که بقیه با آمریکای لاتین، اروپا، آسیای شرقی و آسیای جنوبی صحبت می‌کنند. گویندگان مجموعه‌ای از ۲۶ حرف الفبا، ۱۳ رقم گسسته، نام‌های گسسته و ۲۰۰ جمله شامل ۱۳۱۷ کلمه را بیان می‌کنند. به طور کلی ۵۹۰۰۰ گفتار ضبط شده است.

جنبه‌های کیفی: از ۸ میکروفن و ۴ دوربین استفاده شد. میکروفنهای تلفن همراه صوتی LM386 ارزان به قطر ۶ میلی متر در فاصله ۱/۵ اینچی قرار دارند. هفت کانال از هشت کانال صوتی از طریق کابل‌های محافظ بهADAT ارسال می‌شوند که هشت کانال صوتی را با وضوح ۱۶ بیت با سرعت نمونه‌برداری از ۴۸ کیلوهرتز به عنوان فایلهای WAV ذخیره ضبط می‌کند. هر دوربین از موقعیت‌های مختلف روی داشبورد هدف قرار گرفته شده است تا ناحیه صورت فرد را ضبط کند. چهار جریان ویدیویی توسط یک مالتی‌پلکسر ویدیویی که به یک دوربین فیلم‌برداری MiniDV ارسال می‌شود، ترکیب شده‌اند. یکی از دو کانال صوتی دوربین فیلم‌برداری برای ورودی میکروفن هشتم از آرایه میکروفن استفاده می‌شود. این پایگاه داده ویدئوی حدود ۳۳ ساعت دارای رزولوشن ۴۸۰ × ۷۲۰ و سرعت ۲۵ فریم در هر ثانیه دارد.

ناواراتنا و همکاران [۱۰۰] و [۱۰۱]، کلینشمیت و همکاران[۱۰۲]، و یون و همکاران [۱۷۳] از این پایگاه داده استفاده کرده‌اند.

۶-۳-۳- مجموعه داده IBMIH

این مجموعه داده IBMIH[۶] در سال ۲۰۰۴ شامل ارقام متصل و گفتار پیوسته واژگان بزرگ انگلیسی بر اساس دو حالت استودیو و هدست که نیاز به ردیابی چهره را برطرف می‌کند و تحت تاثیر تغییرات نور و سر نیست، با نویز و بدون نویز جمع‌آوری شد.

جنبه‌های کمی: در مجموع ۱۹۲ گوینده که ۷۹ نفر ارقام و ۱۱۳ نفر واژگان بزرگ را بیان کردند در این پایگاه داده همکاری داشته‌اند.

جنبه‌های کیفی: از دو دوربین یکی دوربین روی هدست و دیگری از روبه‌رو تصویربرداری شده است. داده‌ها با فرمت MPEG2 و با سرعت ۳۰ هرتز با رزولوشن ۷۲۰ در ۴۸۰ پیکسل ضبط شده‌اند. از یک میکروفن رومیزی با آکوستیک ۲۰ دسی‌بل با سرعت ۲۲ کیلوهرتز صدابرداری شده است.

۷-۳-۳- مجموعه داده VALID

مجموعه داده VALID[۵۷] برای تکمیل داده‌های XM2VTS[۵۴] در سال ۲۰۰۵ در یک در یک اتاق اداری واقع‌گرایانه، پر سر و صدا و دارای نور و صدای کنترل نشده به زبان انگلیسی جمع‌آوری شده است.

جنبه‌های کمی: ۱۰۶ نفر شامل ۲۹ زن و ۷۷ مرد دارای ترکیب قومی ۹۷ اروپایی و ۹ آسیایی است که متشکل از دانشجویان کارشناسی، کارشناسی ارشد و کارمندان کالج دانشگاه دوبلین هستند نام خود، یک دنباله عددی و یک جمله را بیان می‌کنند. ضبط در ۵ جلسه انجام شده است به طوری که جلسه اول، محیطی با نور و صدای کنترل شده و پس‌زمینه آبی رنگ است و ۴ جلسه دیگر با وجود سر و صدا و نور طبیعی در محیط‌های واقعی مانند خانه و دفاتر ضبط شده است.

جنبه‌های کیفی: کل پایگاه داده با دوربین دیجیتال کنون CCD XM13 پال با ۲۵ فریم بر ثانیه و با وضوح ۵۷۶ در ۷۲۰ پیکسل ضبط شده است. صدابرداری نیز با نمونه ۱۶ بیتی و فرکانس ۳۲ کیلوهرتز انجام شده است.

۸-۳-۳- مجموعه داده AVA

مجموعه داده AVA [۱۰۳] در سال ۲۰۰۹ برای اولین‌بار به زبان فارسی جمع‌آوری شده است. این پایگاه داده همه واج‌های فارسی را در هجاهای موجود و احتمالی در گفتار گسسته، جمله‌های رایج زبان فارسی و ارقام سریال را پوشش می‌دهد.

جنبه‌های کمی: از ۲ گوینده زن که دارای لهجه تهرانی هستند و هیچ یک از گویندگان دارای اختلال در بیان و اختلال صدا، آرایش و جراحی زیبایی نیستند، استفاده شده است. هر گوینده ۵۵۰۰ عبارت را بیان کرد که برای بیان این ۱۱۰۰۰ عبارت ۱۵ ساعت برای ۲ گوینده طول کشید.

جنبه‌های کیفی: از استودیو تلویزیونی IRIBU، دانشگاه پخش ایران، برای فیلم‌برداری داده‌ها با یک محیط حرفه‌ای مناسب انتخاب شد. سه دوربین دیجیتال یکی پاناسونیک N9000 برای ارجاع سریع استفاده می‌شود، و دو دوربین کَنون XL2. برای حذف سایه‌ها از چهره گوینده، چندین پروژکتور نورپردازی غیر از ۴ پروژکتور نور قابل حمل استفاده شده است. پس‌زمینه دوربین یک پرده آبی است. علاوه بر میکروفن های دوربین ها از یک میکروفن یقه AKG که زیر روسری گوینده نصب شده است هم استفاده می‌شود. در این پایگاه داده، ویدئو در قالب AVI، ۲۵ فریم در ثانیه، ۷۲۰ در ۵۷۶ پیکسل است؛ و صدا در قالب WAV، ۱۶ بیتی و ۴۸ کیلوهرتز است.

آقاالحمدی و همکاران [۱۱۳]، باستان‌فرد و همکاران [۱۱۴] از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۹-۳-۳- مجموعه داده | | AVA

مجموعه داده | | AVA [۲۶] در سال ۲۰۱۰ با ۱۴ گوینده به زبان فارسی جمع‌آوری شده است که نه تنها بیشتر ترکیبات آوایی مؤثر در زبان فارسی را در بر می‌گیرد، بلکه تأثیر هم‌تولیدی در اطلاعات بصری را نیز در نظر می‌گیرد.

جنبه‌های کمی: در این پروژه ۷ زن و ۷ مرد برای ضبط انتخاب شده‌اند که بین ۱۸ تا ۳۰ ساله و دارای لهجه تهرانی هستند. برای داشتن گفتار پیوسته نیز، از افراد خواسته شده است ۲۰ جمله رایج در فارسی معاصر را بگویند. همین‌طور اعداد به صورت جداگانه یا پیوسته تلفظ می‌شوند. برای رقم‌های پیوسته،

ابتدا گویندگان از ۰ تا ۲۱، ۳۰، ۴۰، ۱۰۰، ۲۰۰، ۱۰۰۰، میلیون و میلیارد و رقم‌های ۳۶۷، ۵۴۹ و ۸۲۱ را بشمارند. هر گوینده ۹۰ دقیقه فیلم‌برداری شده است که کل مرحله ضبط در ۵ جلسه، طی دو ماه انجام شده است.

جنبه‌های کیفی: از استودیو تلویزیونی IRIBU، دانشگاه پخش ایران، برای فیلم‌برداری داده‌ها با یک محیط حرفه‌ای مناسب انتخاب شد. سه دوربین دیجیتال یکی پاناسونیک N9000 برای ارجاع سریع استفاده می‌شود، و دو دوربین کَنون XL2. برای حذف سایه‌ها از چهره گوینده، چندین پروژکتور نورپردازی غیر از ۴ پروژکتور نور قابل حمل استفاده شده است. پس‌زمینه دوربین یک پرده آبی است. دو میکروفن استفاده می‌شود، یک میکروفن یقه AKG، در بالای سر بلندگو و میکروفن دیگر Road NT 1000 است. علاوه بر این، تمام دوربین‌ها به طور خودکار صدای خود را بر روی میکروفن خود ضبط می‌کنند. در این پایگاه داده، ویدئو در قالب AVI، ۲۵ فریم در ثانیه، ۷۲۰ در ۵۷۶ پیکسل است؛ و صدا در WAV، ۱۶ بیتی و ۴۸ کیلوهرتز است.

باستان‌فرد و همکاران[۱۱۳]، مقدم و همکاران[۱۱۵] از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۳-۳-۱۰- پایگاه داده NDUTAVSC

پایگاه داده NDUTAVSC[۲۷] در سال ۲۰۱۰ به زبان هلندی بزرگ‌ترین پایگاه داده صوتی تصویری تا آن روز است.

جنبه‌های کمی: این پایگاه داده با ۶۶ گوینده که ۲۰ زن و ۴۶ مرد هستند شامل رشته‌های ارقام با طول ثابت، رشته‌های حروف با طول تصادفی، توالی‌های کلمه تصادفی، گفتار پیوسته به مدت ۱۰ ساعت و ۳۸ دقیقه جمع‌آوری شده است.

جنبه‌های کیفی: از دو میکروفن برای ضبط صدا استفاده شده است که میکروفن اصلی در کنار گوینده در ارتفاع ۸۰ سانتی متری و دیگری در پهلوی گوینده به عنوان میکروفن محیط در ارتفاع ۱/۵ متری می باشد. از دو دوربین برای ضبط تصاویر در ارتفاع ۱/۲ متری استفاده شده است، یکی روبه‌رو و دیگری از پهلوی چپ گوینده تصویربرداری می‌کند. داده‌ها با سرعت ۱۰۰ هرتز ذخیره می‌شوند.

۳-۳-۱۱- مجموعه داده AGH AV

مجموعه داده AGH AV[۵۶] شامل سه نوع جمله مختلف از اعداد و دستورات که ۱۶۰ عبارت کوتاه از لیستی از محبوب‌ترین سوالات پرسیده شده برای دستیار مجازی، ۷ جمله مختلف از زبان گفتاری و بخش هایی از متون مانند مقالات، تعاریف و بخشی از داستان‌ها است و در سال ۲۰۱۲ به زبان لهستانی جمع‌آوری شده است.

جنبه‌های کمی: ۲۴ گوینده شامل ۱۱ زن و۱۳ مرد است. هر گوینده حدود ۱۰ دقیقه ضبط که در کل حدود ۴ ساعت است.

جنبه‌های کیفی: از یک دوربین HD با کیفیت فوق‌العاده با ۲۵ فریم در ثانیه استفاده شده است. در ضبط فیلم‌ها، تصویر فقط دارای چهره (نمای جلویی) در پس‌زمینه روشن با شرایط روشنایی نسبتا متغیر است.

۳-۳-۱۲- مجموعه داده MIRACL-VC

جنبه‌های کمی: مجموعه داده MIRACL-VC[۲۸] در سال ۲۰۱۴ توسط ۱۵ نفر شامل ۱۵۰۰ سخن از کلمات (۱۵ نفر × ۱۰ کلمه × ۱۰ بار) و ۱۵۰۰ سخن از عبارات (۱۵ نفر × ۱۰ عبارات × ۱۰ بار) جمع‌آوری شد. کلمات و عبارات ۱۰ بار تکرار شده‌اند.

جنبه‌های کیفی: فاصله بین گوینده و دوربین Kinect حدود ۱ متر است که تصاویری با سرعت ۱۵ فریم در ثانیه و رزولوشن ۶۴۰ در ۴۸۰ ضبط می‌کند.

دسای و همکاران [۱۰۴]، پارخ همکاران [۱۰۵]، گرگ و همکاران[۱۳۴]، هاشمی و همکاران[۱۳۵]، و آریپین و همکاران[۱۳۶] از این مجموعه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۳-۳-۱۳- پایگاه داده AusTalk

جنبه‌های کمی: همین‌طور در سال ۲۰۱۴ پایگاه داده دیگری با عنوان AusTalk[۲۹] و [۱۰۶] به زبان انگلیسی استرالیایی با ۱۰۰۰ گوینده از مکان‌های مختلف استرالیا حدود ۳۰۰۰ ساعت و در مجموع ۲۲ ترابایت داده جمع‌آوری شد. که شامل ۲۴۰۰۰ سخن از ارقام و ۹۶۶۰۰۰ سخن از کلمات و ۵۹۰۰۰ سخن از جملات با ۶ بار تکرار ایجاد شد.

جنبه‌های کیفی: کلیه داده‌ها توسط پنج میکروفن و دو دوربین استریو ضبط شده است. این پایگاه داده شامل ارقام، کلمات و جملات است. آمار دقیق دموگرافیک مربوط به جنسیت، سن و سطح تحصیلات گویندگان در سرور داده‌ها در دسترس است. رزولوشن تصاویر به دست آمده ۶۴۰ در ۴۸۰ است.

۳-۳-۱۴- مجموعه داده OuluVS2

جنبه‌های کمی: مجموعه داده OuluVS2[۳۰] در سال ۲۰۱۵ با ۵۰ گوینده که در سه مرحله سه نوع جمله را بیان می‌کنند جمع‌آوری شد. در مرحله اول یک گوینده ده دنباله رقمی را فقط یک بار به طور پیوسته بیان می‌کند. در مرحله دوم، ده عبارت کوتاه روزانه انگلیسی مانند "سلام" و "از دیدار شما خوشحال شدم" سه بار بیان می‌کند. در مرحله سوم ده جمله فقط یک بار خوانده می‌شد. بیشتر گویندگان دانشجویان و کارکنان دانشگاه بودندکه هیچ یک انگلیسی زبان بومی نبودند. بلکه اروپایی، چینی، هندی/پاکستانی، عربی و آفریقایی بودند.

جنبه‌های کیفی: هر فرد روی صندلی در مقابل شش دوربین می‌نشیند. یک دوربین HD و یک دوربین پر سرعت HS از رو به

رو تصویربرداری میکنند. چهار دوربین HD دیگر در زوایای ۳۰ درجه، ۴۵ درجه، ۶۰ درجه و ۹۰ درجه در سمت راست فرد قرار دارند. از افراد خواسته شد حرکت سر را حفظ کرده و حالت صورت را خنثی کنند ولی با این حال حرکات طبیعی کنترل نشده سر در فیلم‌های ضبط شده مشاهده می‌شود.

ضبط در یک دفتر معمولی با نورپردازی مختلط و صداهای پس‌زمینه احتمالی انجام شده است. برای ضبط صدا و فیلم‌برداری از پنج دوربین GoPro Hero3 Black Edition استفاده شده است با وضوح تصویری ۱۹۲۰ در ۱۰۸۰ و سرعت ضبط ۳۰ فریم در ثانیه، سرعت بیت صوتی ۱۲۸ کیلوبیت بر ثانیه. یک دوربین PuxeLink PL-B774U نیز با وضوح ۴۸۰ در۶۴۰ و سرعت ضبط ۱۰۰ فریم در ثانیه از روبه‌روی افراد تصویربرداری می‌کند. فرمت ضبط داده‌ها MP4 است که حدود ۳۰ گیگابایت فضا اشغال می‌کند.

چیونگ و زیسرمن [۲۱]، [۴۶] و [۱۰۷]، پتریدیس و همکاران [۸۸]، [۱۰۸] و [۱۰۹]، فانگ و مك [۱۱۰]، لی و همکاران [۱۱۱]، سایتو و همکاران [۱۱۲] از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۱۵-۳-۳- پایگاه داده AV Digits

جنبه‌های کمی: پایگاه داده جدیدی که در سال ۲۰۱۸ به زبان انگلیسی جمع‌آوری شد AV Digits[۳۱] نام دارد که گویندگان آن از ۱۶ ملیت مختلف هستند. ۵۳ گوینده متشکل از ۱۲ زن و ۴۱ مرد با میانگین سنی ۲۶/۷ سال ارقام ۰ تا ۹ را ۵ بار با ۳ حالت معمولی، زمزمه و بدون صدا تکرار می‌کنند. ۳۹ گوینده شامل ۷ زن و ۳۲ مرد با میانگین سنی ۲۶/۳ سال نیز یکسری از اصطلاحات و کلمات را ۵ بار با ۳ حالت معمولی، زمزمه و بدون صدا تکرار می‌کنند.

جنبه‌های کیفی: این پایگاه داده از ۳ دوربین با وضوح ۱۲۸۰ در ۷۸۰ با ۳۰ فریم در ثانیه در یک محیط آزمایشگاهی ضبط شد که سه نمای روبه‌رو، ۴۵ درجه را ضبط می‌کنند. همچنین صوتی توسط ۳ دوربین با استفاده از میکروفن‌های داخلی ۴۴/۱ کیلوهرتز ضبط شده است.

پتریدیس و همکاران[۱۳۷] از این پایگاه داده برای ارزیابی مدل خود استفاده کرده‌اند.

۱۶-۳-۳- مجموعه داده LRS2-BBC

مجموعه داده LRS2-BBC [۳۲]، [۸۹] و [۱۵۳] که نسخه محدودتر و پیش انتخاب شده از LRS2 [۳۲] است، در سال ۲۰۱۸ به زبان انگلیسی توسط افوراس و همکاران از برنامه‌های متنوع تلویزیونی شبکه بی‌بی‌سی جمع‌آوری شده است. این مجموعه داده دارای ۱۰۰۰ ساعت متن گفتاری شامل واژگان گسترده‌ای از ۱۰۰۰ کلمه مختلف با بیش از یک مگابایت نمونه کلمه و بیش از ۱۰۰۰ گوینده مختلف که تنوع قابل توجهی از فرمت در سراسر برنامه‌ها از اخبار معمولی جایی که در آن یک گوینده تک به طور مستقیم در دوربین صحبت می‌کند وجود دارد تا بحث و گفتگویی که گویندگان هر کدام به یکدیگر نگاه می‌کنند و اغلب توجه خود را تغییر می‌دهند. این مجموعه داده شامل جملات و عباراتی است که گفتار آنها پیوسته است و هر کلمه ۲۹ بار تکرار می‌شود. مجموعه داده LRS2-BBC [۳۲]، [۱۵۳] دارای دو بخش مجزای آموزش و تست است.

افوراس و همکاران[۹۰] و [۹۱]، ژائو و همکاران [۹۲]، کورت‌نی و همکاران [۹۳]، لی و همکاران [۹۴]، کیم و همکاران[۱۳۸] و [۱۳۹]، ما و همکاران[۱۴۰]، [۱۶۱] و [۱۶۳]، چوی و همکاران[۱۴۱]، پرجوال و همکاران[۱۴۲]، هونگ و همکاران[۱۴۳]، و ژو و همکاران[۱۴۴] از این مجموعه داده‌ها برای ارزیابی مدل خود استفاده کرده‌اند.

در این بخش به بررسی و تشریح ویژگی‌های کمی و کیفی مجموعه داده‌ها به تفکیک محتوای آنها که شامل اصطلاحات، جملات و یا ترکیبی از موارد مختلف است پرداخته شد. در بخش آتی این ویژگی‌ها مقایسه خواهند شد و مورد بحث و نقد قرار می‌گیرند.

۴- مقایسه پایگاه داده‌های گفتار پیوسته

مجموعه داده‌هایی که در این مقاله مورد مطالعه قرار گرفتند دارای یکسری ویژگی‌های کمی و کیفی هستند که در قالب دو جدول به طور جداگانه مقایسه شده‌اند. از جمله ویژگی‌های کمی یک مجموعه داده می‌توان به تعداد کل گویندگان، تعداد گویندگان زن، تعداد گویندگان مرد، تعداد گویندگان بومی، میانگین سنی آنها، حجم داده‌ها و موارد این چنینی اشاره کرد. مشخصات ویژگی‌های کمی پایگاه داده‌های مورد بررسی در جدول (۱) نشان داده شده‌اند.

از طرفی دیگر، این مجموعه داده‌ها دارای یکسری ویژگی‌های کیفی نیز هستند که شامل شرایط محیط تصویربرداری از نظر کنترل شده و یا کنترل نشده، پیوسته و یا گسسته بودن گفتار بیان شده از سوی گویندگان، مشخصات فنی دوربین‌های تصویربرداری، مشخصات فنی میکروفن‌های صدابرداری، رزولوشن تصاویر ضبط شده و ویژگی‌هایی از این قبیل می باشد. این مجموعه ها از نظر ویژگی های مختلف در جدول (۲) مورد مقایسه قرار گرفته‌اند.

جدول (۱): مشخصات کمی مجموعه داده‌های بازشناسی گفتار پیوسته(اصطلاحات، جملات و ترکیبی)

ردیف	نام پایگاه داده	سال	حجم/ مدت کل پایگاه داده	طول زمان ضبط	تعداد کل گویندگان	تعداد گوینده زن	تعداد گوینده مرد	میانگین سنی گویندگان	زبان	بومی بودن	حرکت سر	تغییرات در ظاهر (ریش، عینک)	نما
۱	[۵۴]XM2VTS	۱۹۹۹	۳۰ ساعت	۵ ماه	۲۹۵	–	–	–	فرانسوی	–	مجاز	مجاز	روبه‌رو، چپ، راست، بالا و پایین
۲	[۹]IBMViaVoice	۲۰۰۰	–	–	۲۹۰	–	–	–	انگلیسی	–	–	–	روبه‌رو
۳	[۱۰]VIDTIMIT	۲۰۰۲	۳/۵ گیگابایت	–	۴۳	۱۹	۲۴	–	انگلیسی	–	مجاز	مجاز	–
۴	[۵]BANCA	۲۰۰۳	۱۴ ساعت	۳ ماه	۲۰۸	۱۰۴	۱۰۴	–	انگلیسی فرانسوی ایتالیایی اسپانیایی				
۵	[۲۳]AVOZES	۲۰۰۴	۲ ساعت	۹ روز	۲۰	۱۰	۱۰	۳۹/۵	استرالیایی انگلیسی	۲۰	مجاز	مجاز	روبه‌رو
۶	[۱۱]AV-TIMIT	۲۰۰۴	۴ ساعت	۱ هفته	۲۲۳	۱۰۶	۱۱۷	سن مختلف	انگلیسی	۲۱۱ نفر	مجاز	مجاز	روبه‌رو
۷	[۲۴]AV@CAR	۲۰۰۴	حدود ۹ ساعت و ۵۰ دقیقه	–	۲۰	۱۰	۱۰	۳۷/۵	اسپانیایی	–	–	–	روبه‌رو
۸	[۲۵]AVICAR	۲۰۰۴	حدود ۳۳ ساعت	–	۱۰۰	۵۰	۵۰	–	انگلیسی	۶۰	مجاز	مجاز	متغییر(۴ نما)
۹	[۶]IBMIH	۲۰۰۴	–	–	۱۹۲	–	–	–	انگلیسی	–	مجاز	مجاز	روبه‌رو
۱۰	[۵۷]VALID	۲۰۰۵	–	۱ ماه	۱۰۶	۲۹	۷۷	–	انگلیسی	۹۷ اروپایی و ۹ آسیایی	مجاز	مجاز	روبه‌رو و ۱۵ درجه به چپ، راست، بالا و پایین
۱۱	[۷]Grid	۲۰۰۶	حدود ۲۸ ساعت	–	۳۴	۱۶	۱۸	۲۷/۴	انگلیسی	۳۴	–	–	–
۱۲	UWB-07-[۱۲]ICAV	۲۰۰۸	حدود ۲۶۵ گیگابایت	۷ سال	۵۰	۲۵	۲۵	۲۲	چک	–	غیر مجاز	–	روبه‌رو
۱۳	[۱۳]IV2	۲۰۰۸	حدود ۶/۶ گیگابایت حدود ۵۰۰ ساعت	۱ ماه	۳۰۰	–	–	–	فرانسوی	–	–	مجاز	روبه‌رو چپ راست
۱۴	[۸]OuluVS	۲۰۰۹	حدود ۱۶ دقیقه	–	۲۰	۳	۱۷	–	انگلیسی	۰	–	مجاز	–
۱۵	[۱۰۳]AVA	۲۰۰۹	۱۶۰ گیگابایت/ ۱۵ ساعت	–	۲	۲	۰	–	فارسی	۲	غیر مجاز	غیرمجاز	نیم‌رخ و روبه‌رو
۱۶	[۲۶]AVA||	۲۰۱۰	۲۴ ساعت	۲ ماه	۱۴	۷	۷	۲۴	فارسی	۱۴	–	مجاز	روبه‌رو
۱۷	[۱۴]LILiR	۲۰۱۰	–	–	۱۲	۵	۷	–	انگلیسی	–	–	–	روبه‌رو
۱۸	[۱۵]WAPUSK20	۲۰۱۰	۲۰ ساعت	–	۲۰	۹	۱۱	۲۹	انگلیسی	۲ نفر	–	–	–
۱۹	[۲۷]NDUTAVSC	۲۰۱۰	۱۰ ساعت و ۳۸ دقیقه	–	۶۶	۲۰	۴۶	–	هلندی	–	–	–	روبه‌رو پهلو(۹۰)
۲۰	[۱۶]BL	۲۰۱۱	۳۴۰ دقیقه	–	۱۷	جلسه اول: ۴ / جلسه دوم: ۴	جلسه اول: ۴ / جلسه دوم: ۵	۳۵/۵	فرانسوی	۱۷	–	–	–

ردیف	نام پایگاه داده	سال	حجم/ مدت کل پایگاه داده	طول زمان ضبط	تعداد کل گویندگان	تعداد گوینده زن	تعداد گوینده مرد	میانگین سنی گویندگان	زبان	بومی بودن	حرکت سر	تغییرات در ظاهر (ریش، عینک)	نما
۲۱	UNMC-VIER[۱۷]	۲۰۱۱	–	۶ روز	۱۲۳	۴۹	۷۴	–	انگلیسی	۱۱۶ آسیایی، ۴ آفریقایی و ۳ اروپایی	مجاز	مجاز	زوایای ۳۰، ۶۰، ۹۰، ۱۲۰، ۱۵۰، ۱۸۰، بالا، پایین و روبه‌رو
۲۲	MOBIO [۱۸]	۲۰۱۲	بیش از ۶۱ ساعت	۱/۵ سال	۱۵۰	۵۱	۹۹	–	انگلیسی	–	–	–	–
۲۳	AGH AV[۵۶]	۲۰۱۲	۴ ساعت	–	۲۴	۱۱	۱۳	–	لهستانی	–		–	–
۲۴	MIRACL-VC [۲۸]	۲۰۱۴	–	–	۱۵	۱۰	۵	–	انگلیسی	–	–	–	–
۲۵	AusTalk[۲۹]	۲۰۱۴	حدود ۳۰۰۰ ساعت ۲۲ ترابایت	۳ سال	۱۰۰۰	–	–	–	استرالیایی انگلیسی	–	–	–	–
۲۶	RM-3000[۲]	۲۰۱۵	حدود ۴ ساعت	۳ روز	۱	۰	۱	–	انگلیسی	۱	غیر مجاز	غیر مجاز	رو به رو
۲۷	TCD-TIMIT[۱۹]	۲۰۱۵	۴۵۰ گیگابایت	–	۶۲	۳۰	۳۲	۴۲	انگلیسی	–	–	–	رو به رو ۳۰ درجه از راست
۲۸	OuluVS2[۳۰]	۲۰۱۵	حدود ۳۰ گیگابایت	–	۵۳	۱۳	۴۰	–	انگلیسی	۰	غیر مجاز	مجاز	۰، ۳۰، ۴۵، ۶۰ و ۹۰ درجه
۲۹	LRW[۴۶]	۲۰۱۶	جدود ۵۰۰ گیگابایت	–	۱۰۰+	–	–	–	انگلیسی	–	مجاز	مجاز	–
۳۰	HAVRUS[۲۰]	۲۰۱۶	–	–	۲۰	۱۰	۱۰	–	روسی	۲۰	–	–	–
۳۱	LRS[۳]	۲۰۱۷	۴۹۶۰ ساعت	۶ سال	۱۰۰۰+	–	–	–	انگلیسی	–	–	–	–
۳۲	MV-LRS [۲۱]	۲۰۱۷	حدود ۲۰ ساعت	–	۱۰۰۰+	–	–	–	انگلیسی	–	–	–	روبه‌رو تا نیم رخ
۳۳	VLRF [۲۲]	۲۰۱۷	۱۸۰ دقیقه ۱۶۲۵۴۰ فریم	–	۲۴	۲۱	۳	–	اسپانیایی	–	–	–	روبه‌رو
۳۴	AV Digits[۳۱]	۲۰۱۸	–	–	۵۳	۴۱	۱۲	۲۶/۷	انگلیسی	–	–	مجاز	روبه‌رو ۴۵ درجه نیم‌رخ
					۳۹	۷	۳۲	۲۶/۳					
۳۵	LRS2-BBC[۳۲]	۲۰۱۸	۱۰۰۰ ساعت	–	۱۰۰۰+	–	–	–	انگلیسی	–	مجاز	مجاز	متغییر
۳۶	LRS3-TED[۱۶۷]	۲۰۱۸	۴۰۰ ساعت	–	۱۰۰۰+	–	–	–	انگلیسی	–	–	–	–
۳۷	LRW-1000[۱۶۶]	۲۰۱۸	–	–	۲۰۰۰+	–	–	–	ماندارین	–	مجاز	مجاز	متعدد
۳۸	CMLR[۱۵۹]	۲۰۲۰	–	–	۱۱	۵	۶	–	ماندارین	–		مجاز	متعدد
۳۹	NTSDB[۱۶۸]	۲۰۲۰	–	–	–	–	–	–	ماندارین	–	–	مجاز	روبه‌رو
۴۰	MAVS[۱۴۵]	۲۰۲۱	–	–	۱۰۳	۳۳	۷۰	۲۷	انگلیسی، هندی و بنگالی	۱۰۳ بومی هندی	–	–	–

ردیف	نام پایگاه داده	سال	حجم/ مدت کل پایگاه داده	طول زمان ضبط	تعداد کل گویندگان	تعداد گوینده زن	تعداد گوینده مرد	میانگین سنی گویندگان	زبان	بومی بودن	حرکت سر	تغییرات در ظاهر (ریش، عینک)	نما
۴۱	GLips [۱۴۷]	۲۰۲۲	–	–	–	–	–	–	آلمانی	–	مجاز	مجاز	متعدد
۴۲	CN-CVS [۱۷۱]	۲۰۲۳	۳۰۰+	–	۲۵۵۰	–	–	–	ماندارین	–	مجاز	مجاز	متعدد
۴۳	Arman-AV[۱۷۴]	۲۰۲۳	۲۲۰ ساعت	–	۱۷۶۰	–	–	–	فارسی	–	مجاز	مجاز	متعدد
۴۴	LUMINA [۱۴۶]	۲۰۲۴	هر کلیپ ۳/۳ ثانیه	–	۱۴	۵	۹	–	اندونزیایی	۱۴	غیر مجاز	–	روبه‌رو
۴۵	DVS-Lip[۱۶۵]	۲۰۲۵	–	–	۴۰	۲۰	۲۰	–	انگلیسی	–	–	–	روبه‌رو

جدول (۲): مشخصات کیفی مجموعه داده‌های بازشناسی گفتار پیوسته (اصطلاحات، جملات و ترکیبی)

ردیف	نام پایگاه داده	پیوسته/ گسسته	شرایط محیطی	تجهیزات صوتی					تجهیزات تصویری				
				تعداد میکروفن	مدل میکروفن	سرعت ضبط	فرمت ضبط	نویز	تعداد دوربین	مدل دوربین	رزولوشن	سرعت ضبط	فرمت ضبط
۱	XM2VTS[۵۴]	پیوسته	آزمایشگاه	۱	–	۳۲ کیلو هرتز	–	–	۲	سونی VX1000E VCR	۵۷۶ در ۷۲۰	۲۵ فریم در ثانیه	–
۲	IBMViaVoice[۹]	پیوسته	مرکز تحقیقات IBM Thomas J. Watson	۱	–	۱۶ کیلو هرتز	MPEG2	ندارد	۱	–	۷۰۴ در ۴۸۰	۳۰ هرتز	–
۳	VIDTIMIT[۱۰]	پیوسته	دفتر کار	۱	–	۳۲ کیلو هرتز	WAV مونو	دارد	۱	PAL	۳۸۴ در ۵۱۲	–	JPEG
۴	BANCA[۵]	پیوسته	محیط‌های کنترل شده، تخریب شده و نامطلوب	۲	–	۳۲ کیلو هرتز	–	–	۲	وب‌کم آنالوگ / دیجیتال	۵۷۶ در ۷۲۰	۲۵ فریم در ثانیه	MPEG-7
۵	AVOZES[۲۳]	–	آزمایشگاه	۱	Sennheiser MKE 10-3	۴۸ کیلو هرتز	WAV	ندارد	۲	NTSC آنالوگ استاندارد	۴۸۰ در ۷۲۰	۳۰ فریم در ثانیه- ۲۹/۹۷ هرتز	AVI
۶	AV-TIMIT[۱۱]	–	دفتر اداری	۱	GN Netcom	۱۶ کیلو هرتز	WAV	ندارد	۱	SONY DCR-VX2000	۷۲۰ در ۴۸۰	۳۰ فریم در ثانیه	AVI
۷	AV@CAR[۲۴]	–	محیط خودرو	۶	Q501T (AKG, Austria)	۱۶ کیلو هرتز	–	–	۱	V-1204A (Marshall Electronics USA)	۷۶۸ در۵۷۶	۲۵ فریم در ثانیه	–
			آزمایشگاه	۳	VA 2000 (GN NetCom, Denmark) / C 477 W R (AKG, Austria) / Q501T (AKG, Austria)	۱۶ کیلو هرتز	–	–	۱	V-1204A (Marshall ElectronicsUSA)	۷۶۸ در۵۷۶	۲۵ فریم در ثانیه	–
۸	AVICAR[۲۵]	پیوسته و گسسته	درون خودرو	۸	LM386	۴۸ کیلو هرتز	WAV	دارد	۴	MiniDV	۷۲۰ در ۴۸۰	۲۵ فریم در ثانیه	–

ردیف	نام پایگاه داده	پیوسته/ گسسته	شرایط محیطی	تجهیزات صوتی					تجهیزات تصویری				
				تعداد میکروفن	مدل میکروفن	سرعت ضبط	فرمت ضبط	نویز	تعداد دوربین	مدل دوربین	رزولوشن	سرعت ضبط	فرمت ضبط
۹	IBMIH[۶]	پیوسته	محیط ایده‌آل شبیه استودیو و تحت نورپردازی یکنواخت	۱	دسک‌تاپ	۲۲ کیلو هرتز	–	دارد-ندارد	۲	هدست استودیو	۷۲۰ در ۴۸۰	۳۰ هرتز	MPEG2
۱۰	VALID[۵۷]	پیوسته	دفتر اداری	۱	–	۳۲ کیلو هرتز	PCM	دارد	۱	دیجیتال کنون 3CCD پال XM1	۵۷۶ در ۷۲۰	۲۵ فریم در ثانیه	PAL DV
۱۱	Grid[۷]	پیوسته	آزمایشگاه	۱	مدل B&K Nexus 2690	۵۰ کیلو هرتز	TDT	–	۱	Canon XM2		۲۵ فریم در ثانیه	MPEG-1
۱۲	UWB-07-ICAV[۱۲]	پیوسته	آزمایشگاه	۲	جهت دار AKG CK55L رومیزی Sennheiser K6	۴۴ کیلو هرتز	WAV	دارد	۲	VCR Canon MVX3i وب Philips SPC 900NC	۷۲۰ در ۵۷۶ ۶۴۰ در ۴۸۰	۵۰ فریم در ثانیه	AVI
۱۳	IV2[۱۳]	–	کابینی شبیه دستگاه فوتوماتون	–	–	–	–	–	۲	DVCAM WEBCAM	۷۸۰ در ۵۷۶ ۴۸۰ در ۴۴۰	–	–
۱۴	OuluVS[۸]	–	–	–	–	–	–	–	۱	SONY DSR-200AP 3CCD	۷۴۰ در ۵۷۶	۲۵ فریم در ثانیه	–
۱۵	AVA[۱۰۳]	پیوسته و گسسته	استودیوی ضبط	۱	AKG	۴۸ کیلو هرتز	WAVE	حداقل	۳	دو عدد Canon XL2 یک عدد PD150 SONY	۷۲۰ در ۵۷۶	۲۵ فریم در ثانیه	DV
۱۶	AVA||[۲۶]	پیوسته	استودیوی ضبط	۲	AKG Road NT 1000	۴۸ کیلو هرتز	WAV	–	۳	Panasonic N9000	۷۲۰ در ۵۷۶	۲۵ فریم در ثانیه	AVI
۱۷	LILiR[۱۴]	پیوسته	آزمایشگاه	۱	–	–	–	–	۱	Thomson Viper FilmStream	–	۲۵ فریم در ثانیه	–
۱۸	WAPUSK20[۱۵]	–	اتاق اداری معمولی	۴	۱ جفت OKTAVA MK 012 ۱ جفت Behringer ECM8000	۱۶ کیلو هرتز	–	دارد	۲	Bumblebee2 نوع BB2-03S2 Logitech (QuickCam Pro 9000)	۶۴۰ در ۴۸۰	۳۲ فریم در ثانیه	–
۱۹	NDUTAVSC[۲۷]	پیوسته	محیط کنترل شده با سطح نویز معقول و روشنایی مناسب	۲	–	–	–	–	۲	–	۶۴۰ در ۴۸۰	۱۰۰ فریم در ثانیه	–

ردیف	نام پایگاه داده	پیوسته/ گسسته	شرایط محیطی	تجهیزات صوتی					تجهیزات تصویری				
				تعداد میکروفن	مدل میکروفن	سرعت ضبط	فرمت ضبط	نویز	تعداد دوربین	مدل دوربین	رزولوشن	سرعت ضبط	فرمت ضبط
۲۰	BL[۱۶]	پیوسته	–	۲	AKG Labtec	۴۴/۱ کیلو هرتز	–	ندارد	–	Canon MVX3i	۵۷۶ در ۷۲۰	۲۵ فریم در ثانیه	–
		پیوسته	–	۲	AKG	۴۴/۱ کیلو هرتز	–	ندارد	–	رنگی جلویی	۶۴۰ در ۴۸۰	۳۰ فریم در ثانیه	–
		پیوسته	–	۲	Labtec	۴۴/۱ کیلو هرتز	–	ندارد	–	microsoft Kinect	۶۴۰ در ۴۸۰	۳۰ تصویر در عمق در ثانیه	
۲۱	UNMC-VIER[۱۷]	پیوسته و گسسته	محیط کنترل شده	۲	–	۴۸ کیلو هرتز	mp2	–	۳	۲ عدد Panasonic SDR-SW20 PAL	۷۰۸ در ۶۴۰ (نمای جلو و چپ)	۲۵ فریم در ثانیه	AVI
				۲	میکروفن داخلی وب‌کم	۲۲ کیلو هرتز	PCM			Logitech Quick Cam Pro 4000	۳۲۰ در ۲۴۰	۲۹ فریم در ثانیه	
			کنترل نشده	۲	–	۳۲ کیلو هرتز	WMV2	–	۲	Panasonic SDR-S7 PAL	۷۰۸ در ۶۴۰	۲۵ فریم در ثانیه	AVI
										Quick Cam E3500 Logitech	۳۲۰ در ۲۴۰	۱۵ فریم در ثانیه	WWV
۲۲	MOBIO [۱۸]	–	موبایل	–	–	–	–	دارد	۲	تلفن همراه نوکیا N93i	–	–	–
										لپ تاپ ۲۰۰۸ MacBook			
۲۳	AGH AV[۵۶]	پیوسته	کنترل نشده	–	–	–	–	–	–	دیجیتالHD	–	۲۵ فریم در ثانیه	–
۲۴	MIRACL-VC[۲۸]	–	–	–	–	–	–	–	۱	Kinect	۶۴۰ در ۴۸۰	۱۵ فریم در ثانیه	–
۲۵	AusTalk[۲۹]	–	–	۵	–	–	–	–	۲	–	–	–	–
۲۶	RM-3000[۲]	پیوسته	آزمایشگاه	۱	نصب روی یقه پیراهن	۴۸ کیلو هرتز	–	–	۱	Sanyo Xacti	۳۶۰ در ۶۴۰	۵۹/۹۴ فریم در ثانیه	–
۲۷	TCD-TIMIT[۱۹]	پیوسته	آزمایشگاه	۳	Shure PG185 و ۲ میکروفن دوربین	–	–	ندارد	۲	Sony PMW-EX3	۱۹۲۰ در ۱۰۸۰	۳۰ فریم در ثانیه	–
۲۸	OuluVS2[۳۰]	–	–	۶	میکروفن داخلی	۱۲۸ کیلوبیت بر ثانیه	–	دارد	۶	۵عدد GoPro Hero3 Black Edition	۱۹۲۰ در ۱۰۸۰	۳۰ فریم در ثانیه	MP4
										۱ عدد PuxeLink PL-B774U	۴۸۰ در۶۴۰	۱۰۰ فریم در ثانیه	
۲۹	LRW[۴۶]	پیوسته	بی‌بی‌سی	–	–	–	–	–	–	–	–	۲۵ فریم در ثانیه	۳۰ فریم در ثانیه
۳۰	HAVRUS[۲۰]	–	–	۱	Oktava MK-012	۱۶ کیلو هرتز	PCM WAV		۱	JAI Pulnix RMC-6740	۶۴۰ در ۴۸۰	۲۰۰ فریم در ثانیه	–
۳۱	LRS[۳]	پیوسته	بی‌بی‌سی	زیاد	–	–	–	–	–	–	–	–	–

ردیف	نام پایگاه داده	پیوسته/ گسسته	شرایط محیطی	تجهیزات صوتی					تجهیزات تصویری				
				تعداد میکروفن	مدل میکروفن	سرعت ضبط	فرمت ضبط	نویز	تعداد دوربین	مدل دوربین	رزولوشن	سرعت ضبط	فرمت ضبط
۳۲	MV-LRS [۲۱]	پیوسته	بی‌بی‌سی	–	–	–	–	–	–	–	۱۶۰در ۱۶۰	۲۵ فریم در ثانیه	–
۳۳	VLRF [۲۲]	پیوسته	آزمایشگاه	–	–	۴۸ کیلو هرتز	–	–	–	HP Panasonic HPX 171	۷۲۰ در ۱۲۸۰	۵۰ فریم در ثانیه	–
۳۴	AV Digits[۳۱]	–	–	۳	میکروفن داخلی	۴۴/۱ کیلو هرتز	–	–	۳	–	۱۲۸۰ در ۷۸۰	۳۰ فریم در ثانیه	–
۳۵	LRS2-BBC[۳۲]	پیوسته	بی‌بی‌سی	–	–	–	–	–	–	–	–	–	–
۳۶	LRS3-TED[۱۶۷]	پیوسته	TED,TEDx	–	–	۱۶ کیلو هرتز	–	–	–	–	×۲۲۴ ۲۲۴	۲۵ فریم در ثانیه	–
۳۷	LRW-1000[۱۶۶]	پیوسته	اخبار چین	–	–	–	–	دارد	–	–	–	۲۵ فریم در ثانیه	Mp4
۳۸	CMLR[۱۵۹]	پیوسته	اخبار چین	–	–	–	–	–	–	–	۶۴ در ۱۲۸	۲۵ فریم در ثانیه	–
۳۹	NTSDB[۱۶۸]	پیوسته	برنامه‌های تلویزیونی	–	–	–	–	–	–	–	–	۲۵ فریم در ثانیه	–
۴۰	MAVS[۱۴۵]		محیط‌های بدون نویز، نویز کنترل شده و نویز کنرل نشده	–	–	–	–	دارد–ندارد	۵	آیفون ۱۱، آیفون ۱۰، آیفون S6، سامسونگ S7 و سامسونگ S8	–	–	–
۴۱	GLips [۱۴۷]	پیوسته	پارلمان هسیان	–	–	–	–	–	–	–	۲۲۴در ۲۲۴	۲۵ فریم در ثانیه	H264
۴۲	CN-CVS [۱۷۱]	پیوسته	وب	–	–	۱۶ کیلو هرتز	–	ندارد	–	–	–	–	–
۴۳	Arman-AV[۱۷۴]	پیوسته	مصاحبه‌ها، سریال‌ها و فیلم‌ها، و ویدئوهای UGC	–	–	–	–	–	–	–	۲۲۴در ۲۲۴	۲۵ فریم در ثانیه	mp4
۴۴	LUMINA [۱۴۶]	–	اتاق عایق صدا	۱	Saramonic Blink500	۱۶ کیلو هرتز	wav	دارد	۱	Fujifilm XT-200	۱۰۸۰در ۱۹۲۰	۵۰ فریم در ثانیه	mp4
۴۵	DVS-Lip[۱۶۵]	–	–	–	–	–	–	–	–	DAVIS346	۲۶۰در ۳۴۶	–	–

در دسترس بودن مجموعه داده‌ها و خلاصه ای از مزایای هر یک از آن‌ها می‌تواند کمک شایانی به محققان در انتخاب پایگاه داده مناسب برای پژوهش‌های خود ارائه دهد. در جدول (۳)، نام پایگاه داده‌ها، مزایای کلیدی هر یک، تناسب آن‌ها با روش‌های یادگیری عمیق بر اساس ارزیابی ادبیات موجود و مشخصات گزارش‌شده شامل حجم، تنوع، کیفیت برچسب‌گذاری و عملکرد مدل‌های یادگیری عمیق، و همچنین آدرس اینترنتی دسترسی به آن‌ها به تفکیک ارائه شده است. این ارزیابی بر پایه معیارهای استاندارد انجام شده و جهت‌گیری‌های آتی را برای بهبود تهیه مجموعه داده‌ها برجسته می‌سازد.

جدول (۳): ویژگی‌های کلیدی، مزایا و منابع دسترسی مجموعه‌داده‌ها

ردیف	نام پایگاه داده	مزایا	مناسب برای روش‌های یادگیری عمیق	آدرس دسترسی به پایگاه داده
۱	[۵۴]XM2VTS	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر	خیر	http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb
۲	[۹]IBMViaVoice	۱. دسترسی به پایگاه داده	خیر	https://download.cnet.com/IBM-ViaVoice/3000-2074_4-37608.html
۳	[۱۰]VIDTIMIT	۱. دسترسی به پایگاه داده	بله (محدود)	*http://www.idiap.ch/~sanders/vidtimit/* http://conradsanderson.id.au/vidtimit/#downloads
۴	[۵]BANCA	۱. دسترسی به پایگاه داده ۲. بالانس تعداد گویندگان زن و مرد ۳. متشکل از ۴ زبان زنده دنیا	خیر	http://banca.ee.surrey.ac.uk
۵	[۲۳]AVOZES	۱. دسترسی به پایگاه داده ۲. بالانس تعداد گویندگان زن و مرد ۳. دارای پوشش کامل واج‌ها و ویزم‌های زبان استرالیایی انگلیسی	بله (محدود)	http://users.cecs.anu.edu.au/~roland/avozes.html
۶	[۱۱]AV-TIMIT	۱. دارای پوشش آوایی غنی در کم‌ترین کلمه ممکن زبان انگلیسی ۲. مناسب روش‌های شبکه عصبی عمیق	بله	-
۷	[۲۴]AV@CAR	۱. بالانس تعداد گویندگان زن و مرد	بله (محدود)	-
۸	[۲۵]AVICAR	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر ۳. بالانس تعداد گویندگان زن و مرد ۴. مناسب روش‌های شبکه عصبی عمیق	بله	http://www.ifp.uiuc.edu/speech/AVICAR/
۹	[۶]IBMIH	۱. با وجود هدست نیاز به ردیابی چهره ندارد	خیر	-
۱۰	[۵۷]VALID	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر	بله (محدود)	http://ee.ucd.ie/validdb/
۱۱	[۷]Grid	۱. دسترسی به پایگاه داده ۲. مناسب روش‌های شبکه عصبی عمیق	بله	https://www.grid.ac/downloads
۱۲	UWB-07-[۱۲]ICAV	۱. دسترسی به پایگاه داده ۲. بالانس تعداد گویندگان زن و مرد	خیر	keyx9csi@nottingham.edu.my
۱۳	[۱۳]IV2	۱. زوایای متفاوت تصویر	خیر	-
۱۴	[۸]OuluVS	۱. دسترسی به پایگاه داده	بله (محدود)	http://www.cse.oulu.fi/CMV/Downloads
۱۵	[۱۰۳] AVA	۱. زوایای متفاوت تصویر ۲. دارای پوشش کامل واج‌ها و ویزم‌های زبان فارسی ۳. مناسب روش‌های شبکه عصبی عمیق	بله	-
۱۶	[۲۶]AVA||	۱. بالانس تعداد گویندگان زن و مرد ۲. در نظر گرفتن اثر هم‌تولیدی ۳. مناسب روش‌های شبکه عصبی عمیق	بله	-
۱۷	[۱۴]LILiR	۱. دسترسی به پایگاه داده	بله (محدود)	http://www.ee.surrey.ac.uk/Projects/LILiR/datasets.html
۱۸	[۱۵]WAPUSK20	۱. دسترسی به پایگاه داده	خیر	http://www.emsp.tu-berlin.de/forschung/AVSR.
۱۹	[۲۷]NDUTAVSC[	۱. زوایای متفاوت تصویر	بله (محدود)	-
۲۰	[۱۶]BL	۱. دسترسی به پایگاه داده ۲. دارای فراوانی وقوع واج‌های زبان فرانسوی	خیر	http://bl-database.inria.fr/
۲۱	UNMC-[۱۷]VIER	۱. زوایای متفاوت تصویر	بله (محدود)	-
۲۲	[۱۸] MOBIO	۱.دسترسی به پایگاه داده ۲. مناسب روش‌های شبکه عصبی عمیق	بله	http://www.idiap.ch/dataset/mobio
۲۳	[۵۶]AGH AV	۱. دسترسی به پایگاه داده ۲. مناسب روش‌های شبکه عصبی عمیق	خیر	https://www.agh.edu.pl/en/

ردیف	نام پایگاه داده	مزایا	مناسب برای روش‌های یادگیری عمیق	آدرس دسترسی به پایگاه داده
۲۴	MIRACL-VC [۲۸]	۱. دسترسی به پایگاه داده	بله (محدود)	https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1
۲۵	[۲۹]AusTalk	۱. دسترسی به پایگاه داده ۲. تنوع گویندگان ۳. مناسب روش‌های شبکه عصبی عمیق	بله	https://austalk.edu.au/
۲۶	[۲]RM-3000	۱. دارای پوشش گسترده‌ای از واج‌های زبان انگلیسی	خیر	-
۲۷	TCD-[۱۹]TIMIT	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر ۳. مناسب روش‌های شبکه عصبی عمیق ۴. دارای پوشش اکثر واج‌ها و ویزم‌های زبان انگلیسی	بله	www.mee.tcd.ie/~sigmedia/Resources
۲۸	[۳۰]OuluVS2	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر ۳. مناسب روش‌های شبکه عصبی عمیق	بله	http://www.ee.oulu.fi/research/imag/OuluVS2/
۲۹	[۴۶]LRW	۱. دسترسی به پایگاه داده ۲. تنوع گویندگان ۳. مناسب روش‌های شبکه عصبی عمیق	بله	https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html
۳۰	[۲۰]HAVRUS	۱. بالانس تعداد گویندگان زن و مرد	بله (محدود)	-
۳۱	[۳]LRS	۱. دسترسی به پایگاه داده ۲. تنوع گویندگان ۳. مناسب روش‌های شبکه عصبی عمیق	بله	http://www.robots.ox.ac.uk/~vgg/
۳۲	MV-LRS [۲۱]	۱. زوایای متفاوت تصویر ۲. تنوع گویندگان ۳. مناسب روش‌های شبکه عصبی عمیق	بله	-
۳۳	VLRF [۲۲]	۱. دسترسی به پایگاه داده ۲. پوشش ۳۱ واج زبان اسپانیایی ۳. مناسب روش‌های شبکه عصبی عمیق	بله	vlrf.database@upf.edu
۳۴	[۳۱]AV Digits	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر	بله (محدود)	https://ibug-avs.eu/
۳۵	[۳۲]LRS2-BBC	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر ۳. تنوع گویندگان ۴. مناسب روش‌های شبکه عصبی عمیق	بله	https://cloud.google.com/speech-to-text http://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html
۳۶	LRS3-[۱۶۷]TED	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر ۳. تنوع گویندگان ۴. مناسب روش‌های شبکه عصبی عمیق	بله	https://www.robots.ox.ac.uk/~vgg/data/lip_reading/
۳۷	LRW-[۱۶۶]1000	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر ۳. مناسب روش‌های شبکه عصبی عمیق	بله	https://vipl.ict.ac.cn/resources/databases/201810/t20181017_32714.html
۳۸	[۱۵۹]CMLR	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر ۳. تنوع گویندگان ۴. مناسب روش‌های شبکه عصبی عمیق	بله	https://www.vipazoo.cn/CMLR.html
۳۹	[۱۶۸]NTSDB	۱. مناسب روش‌های شبکه عصبی عمیق	بله (محدود)	-
۴۰	[۱۴۵]MAVS	۱. دسترسی به پایگاه داده ۲. متشکل از ۳ زبان زنده دنیا ۳. مناسب روش‌های شبکه عصبی عمیق	بله	https://docs.google.com/forms/d/e/1FAIpQLSfTMqnQj8KNoUi1Ms1tx8Ewgil2l4wAAJVaKUJs6VkWfjAo4w/viewform?usp=sf_link
۴۱	GLips [۱۴۷]	۱. زوایای متفاوت تصویر ۲. مناسب روش‌های شبکه عصبی عمیق	بله	-

ردیف	نام پایگاه داده	مزایا	مناسب برای روش‌های یادگیری عمیق	آدرس دسترسی به پایگاه داده
۴۲	CN-CVS [۱۷۱]	۱. دسترسی به پایگاه داده ۲. زوایای متفاوت تصویر ۳. مناسب روش‌های شبکه عصبی عمیق	بله	http://index.cslt.org/mediawiki/index.php/CN-CVS
۴۳	Arman-Av [۱۷۴]	۱. زوایای متفاوت تصویر ۲. تنوع لهجه‌ها ۳. تنوع گویندگان ۴. مناسب روش‌های شبکه عصبی عمیق	بله	-
۴۴	LUMINA [۱۴۶]	۱. دسترسی به پایگاه داده ۲. پوشش هجاها در زبان باهاسای اندونزی ۳. مناسب روش‌های شبکه عصبی عمیق	بله	https://data.mendeley.com/datasets/8fw93k4rny/4
۴۵	DVS-Lip[۱۶۵]	۱. دسترسی به پایگاه داده ۲. بالانس تعداد گویندگان زن و مرد ۳. مناسب روش‌های شبکه عصبی عمیق	بله (محدود)	https://sites.google.com/view/event-based-lipreading

محققان از این مجموعه داده‌ها در کاربردهای متنوعی از جمله بازشناسی گفتار، انتخاب ویژگی‌ها [۷۱]، ردیابی ویژگی [۷۷]، تأیید مشترک چهره و صدا [۷۸]، جداسازی و بهبود صدا [۸۰]، بهبود گفتار [۸۱]، شناسایی افراد [۹۷]، امنیت بیومتریک [۹۸]، سیستم‌های تایید چهره[۹۹] و تشخیص ناحیه لب [۱۰۱] بهره جسته‌اند. نتایج پژوهش‌های مبتنی بر این مجموعه داده‌ها، به ویژه مطالعات متمرکز بر بازشناسی گفتار و لب‌خوانی، بر اساس نرخ خطای کاراکتر، نرخ خطای کلمه و دقت تشخیص در جدول‌های (۴)، (۵) و (۶) بررسی شده است.

سختی مجموعه داده‌ها منحصراً بر اساس بازه دقت؛ بهترین دقت گزارش شده توسط مدل‌های مرجع استاندارد در مقالات تعیین گردید: مجموعه داده‌هایی با دقت کمتر از ۶۰٪ در دسته زیاد، دقت بین ۶۰-۸۰٪ در دسته متوسط و دقت بالای ۸۰٪ در دسته کم قرار گرفتند.

جدول (۴) نرخ خطای کاراکتر (%) پژوهش‌ها در مجموعه داده‌های بازشناسی تصویری گفتار پیوسته

ردیف	نام پایگاه داده	تعداد پژوهش‌های ارجاع داده شده	پژوهش‌ها	نرخ خطای کاراکتر	کمترین نتیجه	بیشترین نتیجه
۱	[۷]Grid	۱۴۵۶	سارهان و همکاران [۸۷]	۱/۴	۱/۱۶	۱/۴
			چن و همکاران[۱۴۸]	۱/۱۶		
۲	MV-LRS [۲۱]	۱۶۶	چیونگ و زیسرمن [۲۱]	۴/۵	۴/۵	۴/۵
۳	CMLR[۱۵۹]	۱۰۰	ژائو و همکاران [۹۲]	۳۱/۲۷	۸/۰	۳۱/۲۷
			شوئه و همکاران [۱۵۰]	۲۲/۵۴		
			لو و همکاران [۱۶۰]	۹/۹۱		
			ما و همکاران [۱۶۳]	۸/۰		

جدول (۵) نرخ خطای کلمه(%) پژوهش‌ها در مجموعه داده‌های بازشناسی تصویری گفتار پیوسته

ردیف	نام پایگاه داده	تعداد پژوهش‌های ارجاع داده شده	پژوهش‌ها	نرخ خطای کلمه		کمترین نتیجه	بیشترین نتیجه
۱	IBMViaVoice[۹]	۳۸۰	متیوس و همکاران [۱۴۹]	DCT	۵۸/۱	۵۸/۱	۶۴/۰
				AAM	۶۴/۰		
۲	[۷]Grid	۱۴۵۶	سارهان و همکاران [۸۷]	۳/۳		۰/۳۶۳	۹/۶۴
			مرگام و همکاران [۱۱۹]	SD	۱/۳		
				SI	۸/۶		
			کیم و همکاران [۱۲۰]	A + C	۴/۰		
				A + P + C	۳/۸		
			چن و همکاران[۱۴۸]	۲/۷۱			
			زو و همکاران [۱۴۹]	SD	۱/۴۵		
				SI	۹/۶۴		
			شوئه و همکاران [۱۵۰]	۰/۸۷			

ردیف	نام پایگاه داده	تعداد پژوهش‌های ارجاع داده شده	پژوهش‌ها	نرخ خطای کلمه		کمترین نتیجه	بیشترین نتیجه
			کیو و همکاران [۱۵۱]	۰/۵۹۳	SD		
				۰/۳۶۳	SI		
۳	LRW[۴۶]	۹۹۸	کیم و همکاران [۱۳۸]	۱۳/۸۶			
۴	MV-LRS [۲۱]	۱۶۶	چیونگ و زیسرمن [۲۱]	۵۶/۱		۵۶/۱	۵۶/۱
۵	LRS2-BBC[۳۲]	۱۱۶۰	افوراس و همکاران [۳۲]	۴۸/۳		۱/۵	۵۹/۷
			ژائو و همکاران [۹۲]	۶۵/۳			
			لی و همکاران [۹۴]	۳۳/۰	Visual		
				۵۹/۷	Audio		
				۲۱/۸	Audio-Visual		
			ما و همکاران [۱۴۰]	۱/۵	Visual		
				۱۴/۶	Audio		
				۱/۵	Audio-Visual		
			افوراس و همکاران [۱۵۳]	۸/۲			
			ما و همکاران [۱۶۱]	۳/۹	Visual		
				۳۷/۹	Audio		
				۳/۷	Audio-Visual		
			ما و همکاران [۱۶۳]	۲۵/۵			
			پرجوال و همکاران[۱۴۲]	۲۲/۶			
			ژو و همکاران[۱۴۴]	۲۴/۳	Visual		
				۲/۳	Audio-Visual		
۶	LRS3-TED[۱۶۷]	۶۴۷	افوراس و همکاران [۳۲]	۵۸/۹		۰/۹	۶۷/۳
			لی و همکاران [۹۴]	۳۳/۲	Visual		
				۶۷/۳	Audio		
				۲۲/۷	Audio-Visual		
			ما و همکاران [۱۴۰]	۱/۰	Visual		
				۱۹/۱	Audio		
				۰/۹	Audio-Visual		
			ژو و همکاران[۱۴۴]	۲۶/۲	Visual		
				۱/۲	Audio-Visual		
			ما و همکاران [۱۶۱]	۱/۳	Visual		
				۳۰/۴	Audio		
				۱/۲	Audio-Visual		
			ما و همکاران [۱۶۳]	۳۱/۵			
			پرجوال و همکاران[۱۴۲]	۳۰/۷			
			افوراس و همکاران [۱۵۳]	۷/۲			
			شی و همکاران [۱۶۲]	۲۶/۹			
۷	LRW-1000[۱۶۶]	۲۴۸	وانگ و همکاران [۱۵۵]	۵۷/۱		۵۷/۱	۵۷/۱
۸	CMLR[۱۵۹]	۱۰۰	زو و همکاران [۱۴۹]	۲۷/۷۹	SD	۰/۲۸	۴۳/۱۸
				۴۳/۱۸	SI		
			کیو و همکاران [۱۵۱]	۰/۳۶	SD		
				۰/۲۸	SI		
			سان و همکاران [۱۶۴]	۴/۰			

جدول (۶) ارزیابی سختی مجموعه داده‌های بازشناسی تصویری گفتار پیوسته بر اساس دقت تشخیص(%)

ردیف	نام پایگاه داده	تعداد پژوهش‌های ارجاع داده شده	پژوهش‌ها	دقت		کمترین دقت	بیشترین دقت	سختی
۱	VIDTIMIT[۱۰]	۱۳۵	کاپلتا و هارت [۷۳]	۶۰/۱۰		۶۰/۱۰	۶۰/۱۰	متوسط
۲	AVICAR[۲۵]	۲۵۸	ناواراتنا و همکاران[۱۰۰]	Visual	۳۱/۸۹	۳۱/۸۹	۷۲/۵۸	متوسط
				Audio	۶۴/۶۹			
				Audio-Visual	۶۴/۶۹			
			کلینشمیت و همکاران[۱۰۲]	۷۲/۵۸				
			یون و همکاران [۱۷۳]	۳۷/۸۷				
۳	Grid[۷]	۱۴۵۶	ژانگ و همکاران [۶۳]	۸۲/۰۰		۵۵/۱۲	۹۷/۱۰	کم
			آسل و همکاران [۶۴]	۹۵/۲				
			وند و همکاران [۶۵]	Eigenlips + SVM	۷۰/۶			
				HOG + SVM	۷۱/۳			
				LSTM	۷۹/۶			
			وند و همکاران [۶۶]	۸۴/۷				
			چیونگ و همکاران [۲۱]	۹۷/۰۰				
			سو و همکاران [۶۷]	۹۷/۱۰				
			ایوانکو و همکاران [۸۶]	GMMCHMM	۵۵/۱۲			
				DNN-HMM	۷۱/۳۴			
				End-to-end	۸۴/۳			
			سارهان و همکاران [۸۷]	-				
			پارخ و همکاران [۱۰۵]	۸۴/۸۰				
۴	OuluVS[۸]	۳۹۹	چیونگ و زیسرمن [۴۶]	۹۱/۴		۵۲/۷	۹۳/۷۲	کم
			رکیک و همکاران [۷۰]	۹۳/۲				
			رکیک و همکاران [۲۸]	SD	۹۳/۲			
				SI	۶۸/۳			
			پینگ‌پینگ و همکاران[۷۱]	۸۷/۵۵				
			پتریدیس و پانتیک [۱۲۱]	۸۱/۸				
			مصباح و همکاران [۱۲۲]	۹۳/۷۲				
			شیراکاتا و سایتو[۱۲۴]	۵۲/۷				
			ژو و همکاران [۱۷۲]	۸۱/۳				
۵	LILiR[۱۴]	۱۰۸	آکاکین و همکاران [۷۶]	۹۸/۲		۹۸/۲	۹۸/۲	کم
۶	MIRACL-VC[۲۸]	۹۴	رکیک و همکاران [۷۰]	۹۶/۰		۵۲/۹	۹۸/۰	کم
			رکیک و همکاران[۲۷]	اصطلاح	کلمه			
				۹۶/۰	۹۵.۴	SD		
				۷۹/۲	۶۳.۱	ID		
			پارخ همکاران [۱۰۵]	SD	۹۸/۰			
				ID	۶۳/۲۲			
			گرگ و همکاران[۱۳۴]	۴۴/۵				
			هاشمی و همکاران[۱۳۵]	۵۲/۹				
			آریپین و همکاران[۱۳۶]	۹۰/۶۷				
۷	RM-3000[۲]	۱۵	هاول و همکاران [۴۷]	۶۶/۳		۶۶/۳	۸۴/۶۷	کم
			سانگسای و همکاران[۱۲۵]	۸۴/۶۷				
۸	TCD-TIMIT[۱۹]	۳۱۹	استرپو و همکاران [۸۳]	۶۶/۲۷		۰/۳۷۲	۶۶/۲۷	متوسط

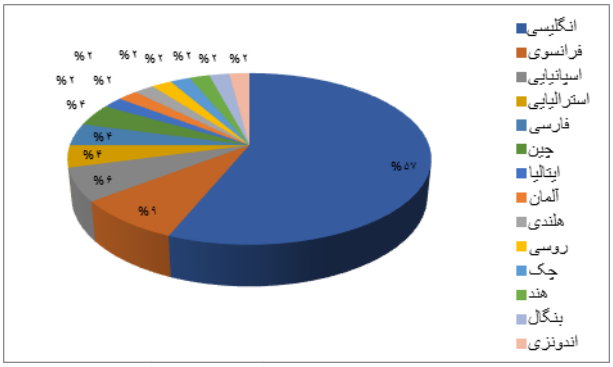
ردیف	نام پایگاه داده	تعداد پژوهش‌های ارجاع داده شده	پژوهش‌ها	دقت		کمترین دقت	بیشترین دقت	سختی
			استرپو و هارته[۸۴]	۴۶/۸۰				
			سانگسای و همکاران [۸۵]	۴۸/۷۴	SD			
				۴۲/۹۷	ID			
			کیو و همکاران [۱۵۱]	۰/۳۷۲	SD			
				۰/۳	ID			
۹	[۳۰]OuluVS2	۲۰۶	چیونگ و زیسرمن [۲۱]	۹۷/۰		۵۵/۸۶	۹۷/۱	کم
			چیونگ و زیسرمن [۴۶]	۹۷/۱				
			پتریدیس و همکاران [۸۸]	۸۳/۲				
			چیونگ و زیسرمن [۱۰۷]	۵۵/۸۶				
			پتریدیس و همکاران [۱۰۸]	۸۸/۳				
			پتریدیس و همکاران [۱۰۹]	۸۵/۳				
			فانگ و مك [۱۱۰]	۸۸/۵				
			لی و همکاران [۱۱۱]	۸۵/۴۱				
			سایتو و همکاران [۱۱۲]	۸۵/۶۱				
			جنگ و همکاران[۱۲۳]	۹۰/۹				
۱۰	[۴۶]LRW	۹۹۸	چیونگ و همکاران [۳]	۹۷		۵۵/۸۶	۹۷/۱	کم
			زو و همکاران [۶۷]	۹۷/۱				
			مصباح و همکاران [۱۲۲]	۵۵/۸۶				
			چنگ و همکاران[۱۳۳]	۸۳/۲				
			وانگ و همکاران [۱۵۵]	۸۸/۳				
			مارتینز و همکاران[۱۵۶]	۸۵/۳				
			ما و همکاران [۱۵۷]	۸۸/۵				
			کیم و همکاران [۱۱۸]	۸۵/۴۱				
			پارخ و همکاران [۱۰۵]	۸۵/۶۱				
			کیم و همکاران [۱۲۰]	۸۹/۳۳	C			
				۸۹/۹۹	C +P			
				۸۹/۷۵	A + P + C			
			کیم و همکاران [۱۳۸]					
			کیم و همکاران [۱۳۹]	۸۸/۵				
۱۱	[۲۰]HAVRUS	۴۷	ایوانکو و همکاران [۸۶]	۴۵/۱۸	GMMCHMM	۱/۱۳	۴۵/۱۸	زیاد
				۲۵/۵۷	DNN-HMM			
				۱/۱۳	End-to-end			
۱۲	[۲۲] VLRF	۶۵	گومز و هینارجوس[۱۵۲]	۵۹/۷		۵۹/۷	۵۹/۷	زیاد
۱۳	[۳۱]AV Digits	۸۵	پتریدیس و همکاران[۱۳۷]	۷۰/۸		۷۰/۸	۷۰/۸	متوسط
۱۴	[۳۲]LRS2-BBC	۱۱۶۰	کورت‌نی و همکاران [۹۳]	۸۵/۲		۵۹/۶	۸۵/۲	کم
			چنگ و همکاران[۱۳۳]	۵۹/۶				
۱۵	[۱۶۷]LRS3-TED	۶۴۷	کورت‌نی و همکاران [۹۳]	۸۷/۱		۸۷/۱	۸۷/۱	کم
۱۶	[۱۶۶]LRW-1000	۲۴۸	مارتینز و همکاران[۱۵۶]	۴۱/۴		۴۱/۴	۹۰/۷۱	کم
			ما و همکاران [۱۵۷]	۴۶/۶				
			لن و همکاران[۱۵۸]	۹۰/۷۱				
			کیم و همکاران [۱۱۸]	۵۰/۸۲				

مجموعه‌داده‌ها در زبان‌های متنوعی جمع‌آوری شده‌اند که درصد فراوانی آن‌ها در نمودار (۱) نمایش داده شده است. بر اساس بررسی‌های انجام‌شده، در این مقاله فقدان مجموعه داده‌های مناسب در برخی از زبان‌ها از پرداختن به الگوریتم‌های مختلف در آن زبان‌ها جلوگیری می‌کند. بررسی نمودار (۱) نشان می‌دهد که زبان انگلیسی متنوع‌ترین مجموعه داده‌ها را در بر می‌گیرد و بالتبع تحقیقات گسترده‌تری در زمینه پردازش گفتار و لب‌خوانی صورت پذیرفته است.

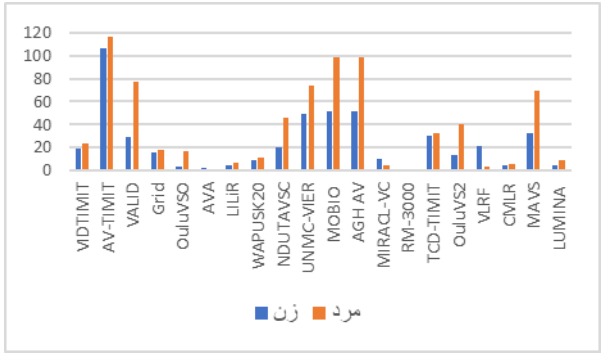
همین‌طور در نمودار (۲) درصد فراوانی تعداد گویندگان زن و مرد از مجموعه داده‌های گفتار پیوسته‌ای که در مقاله مربوطه به تعداد گویندگان زن و مرد در آن مجموعه داده اشاره کرده است، نشان داده شده است.

همان‌طور که در نمودار (۳) مشاهده می‌شود مجموعه داده‌هایی که توازن تعداد گویندگان زن و مرد را رعایت کرده‌اند، به همراه تعداد گویندگان نشان داده شده است.

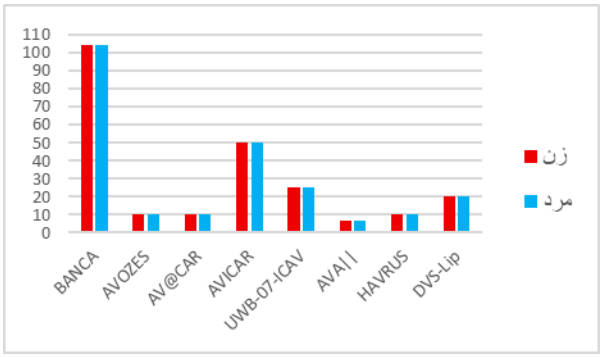
حجم مجموعه داده‌های گفتار پیوسته مورد مطالعه در این مقاله که اطلاعات آن‌ها بر حسب گیگابایت داده شده است و بیشتر از ۳ گیگابایت هستند در نمودار (۴) و آن دسته از مجموعه داده‌هایی که اطلاعات آن‌ها بر حسب مدت زمان هستند، برای آن‌هایی که کمتر از ۱۰۰ ساعت هستند در نمودار (۵) و بیشتر از ۱۰۰ ساعت در نمودار (۶) نمایش داده شده است.



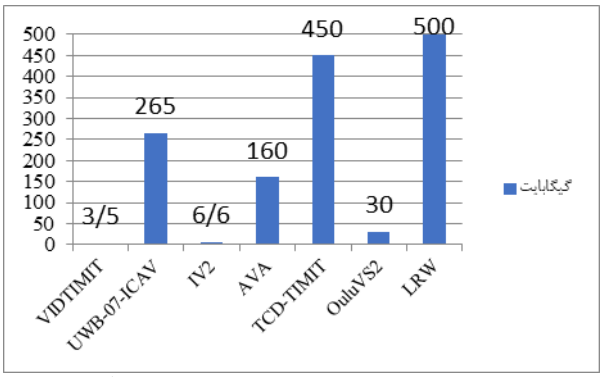
نمودار (۱): درصد فراوانی زبان پایگاه داده‌های گفتار پیوسته



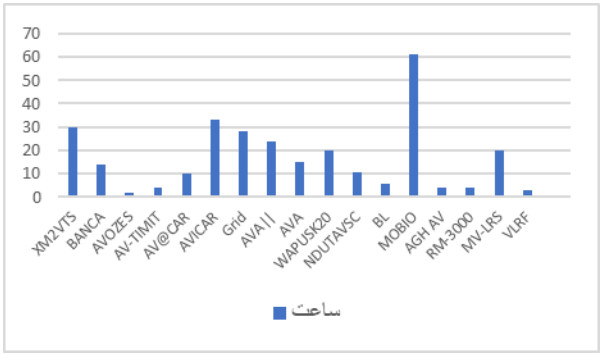
نمودار (۲): درصد فراوانی تعداد گویندگان زن و مرد در مجموعه داده‌ها



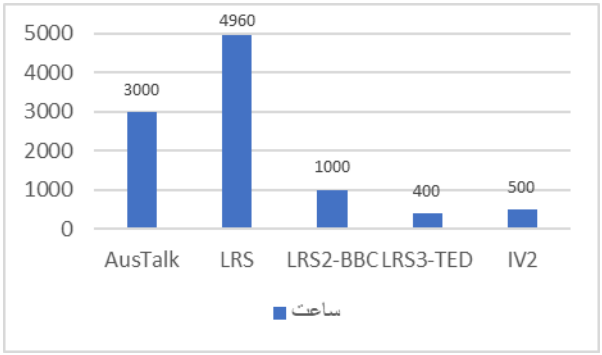
نمودار (۳): مجموعه داده‌هایی با توازن جنسیتی



نمودار (۴): مقایسه حجم تعدادی از مجموعه داده‌های گفتار پیوسته



نمودار (۵): مدت زمان (ساعت) تعدادی از مجموعه داده‌های گفتار پیوسته



نمودار (۶): مدت زمان (ساعت) تعدادی از مجموعه داده‌های بزرگ گفتار پیوسته

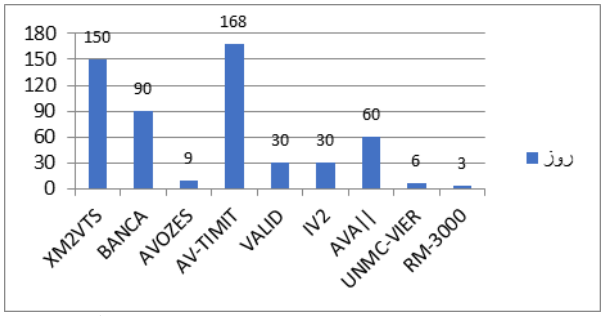
مطالعه حجم و مدت زمان این مجموعه داده‌ها به انتخاب روش مناسب برای حل مسائل بر روی این مجموعه داده‌ها به محققان کمک می‌کند. چرا که روش‌های سنتی نیاز به حجم زیادی از داده‌ها ندارند و بالعکس الگوریتم‌های مبتنی بر شبکه عصبی عمیق به حجم گسترده‌ای از داده‌ها نیازمند هستند و بدین ترتیب مجموعه داده TCD-TIMIT[۱۹] با ۴۵۰ گیگابایت داده و یا LRS[۳] با ۴۹۶۰ ساعت داده می‌توانند گزینه مناسبی برای روش‌های عمیق باشند.

از طرفی دیگر مدت زمانی که طول کشیده است تا داده‌ها در هر یک از مجموعه داده‌های مورد مطالعه ضبط و جمع‌آوری شوند متفاوت هستند که بر اساس اطلاعاتی که در رابطه با این مجموعه داده‌ها وجود دارد، جهت بررسی آنها می‌توان نمودار (۷) را مشاهده کرد.

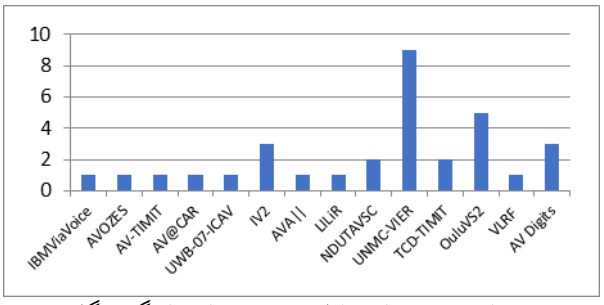
در هر یک از مجموعه داده‌ها از تعداد دوربین‌های متفاوتی جهت تصویربرداری استفاده شده است. به طوری که گروهی از مجموعه داده‌ها تنها با یک دوربین نمای روبه‌رو را ضبط کرده‌اند، از جمله آنها می‌توان VLRF[۲۲] و LILiR[۱۴] را نام برد. گروهی از مجموعه داده‌ها از ۲ دوربین یا بیشتر برای تصویربرداری استفاده کرده‌اند تا کاربردهای وسیع‌تری را پشتیبانی کنند. از جمله مجموعه داده‌هایی که از ۲ دوربین استفاده کرده‌اند می‌توان TCD-TIMIT[۱۹] را نام برد که تصویر از روبه‌رو و تصویر با زاویه ۳۰ درجه از سمت راست گوینده را دارند. برای بررسی بیشتر مجموعه داده‌هایی که تعداد دوربین‌های ضبط ویدئو در مقاله مربوطه تعیین شده است، می‌توان نمودار (۸) را مشاهده کرد.

جهت بررسی میانگین سنی گویندگانی که در جمع‌آوری این مجموعه داده‌ها همراهی داشته‌اند می‌توان نمودار (۹) را مشاهده کرد.

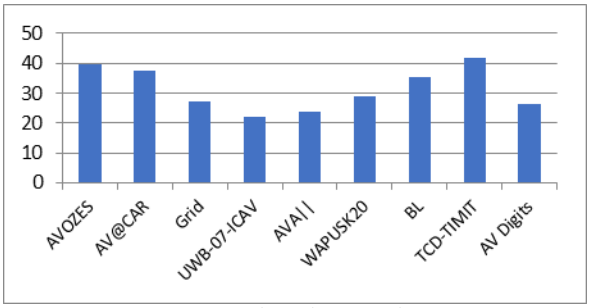
از آن جایی که از افراد مختلفی برای جمع‌آوری داده‌ها استفاده شده است گروهی از شرکت کنندگان، افراد بومیِ زبان مورد نظر برای ضبط هستند و گروهی دیگر افراد غیر بومی هستند تا تلفظ‌های مختلفی از جملات به دست آیند. برای مثال در مجموعه داده AV-TIMIT[۱۱] از ۲۲۳ نفر گوینده، ۲۱۱ نفر بومی هستند. درصد بومی بودن گویندگان در مجموعه داده‌هایی که اطلاعات آن در مقالات مربوطه ذکر شده است، در نمودار (۱۰) نشان داده شده است.



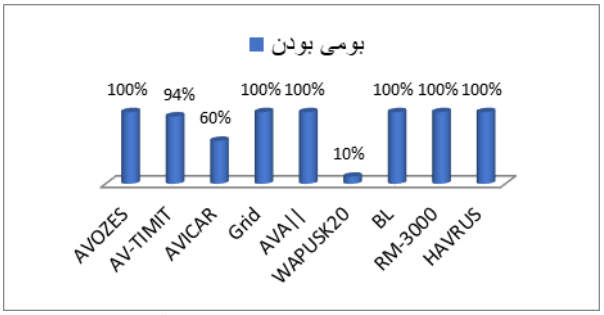
نمودار (۷): مدت زمان ضبط داده‌ها در مجموعه داده‌های گفتار پیوسته



نمودار (۸): تعداد زوایای تصویربرداری از گویندگان



نمودار (۹): میانگین سنی گویندگان در مجموعه داده‌ها



نمودار (۱۰): درصد افراد بومی در پایگاه داده

۵- بحث

در این مطالعه، بررسی ۴۵ مجموعه داده صوتی و تصویری گفتار پیوسته به صورت جامع انجام شد تا تنوع، کیفیت و نقاط قوت و ضعف هر پایگاه داده مشخص گردد. نتایج نشان داد که هر پایگاه داده ویژگی‌های منحصر به فردی دارد که می‌تواند بر انتخاب و کاربرد آن در پژوهش‌های مختلف تأثیرگذار باشد. وجود تفاوت‌های فراوان در ساختار، و حجم، اهمیت تطابق پایگاه داده با هدف پژوهش و مدل‌های مورد استفاده را برجسته می‌کند.

مقایسه نتایج این بررسی با مطالعات پیشین نشان می‌دهد که بسیاری از تحقیقات پیشین به بررسی محدودتری از پایگاه داده‌ها پرداخته‌اند و گستردگی بررسی فعلی موجب شده است که دیدی جامع‌تر نسبت به وضعیت کنونی داده‌ها فراهم گردد. این امر می‌تواند به پژوهشگران کمک کند تا با آگاهی بیشتر نسبت به انتخاب مجموعه داده مناسب تصمیم‌گیری کنند و محدودیت‌های هر کدام را مد نظر قرار دهند.

عوامل متعددی مانند کیفیت داده‌ها، تنوع گونه‌های زبانی، و شرایط ضبط بر کارایی مدل‌های یادگیری ماشینی تأثیرگذار است که در این بررسی مورد توجه قرار گرفت. با این حال، محدودیت‌هایی نیز وجود دارد؛ از جمله اینکه استاندارد یکنواختی برای ارزیابی کامل داده‌ها وجود ندارد و برخی پایگاه داده‌ها به دلیل عدم دسترسی یا مستندسازی ناکافی به‌طور کامل تحت بررسی قرار نگرفته‌اند.

همچنین، پژوهش‌های ترکیبی که از چندین پایگاه داده استفاده می‌کنند می‌تواند به بهبود عملکرد مدل‌ها و افزایش قابلیت تعمیم نتایج کمک کند.

در نهایت، این بررسی جامع بر اهمیت انتخاب درست پایگاه داده در تحقیقات مرتبط با حوزه لب‌خوانی و پردازش گفتار تأکید دارد و نقشی کلیدی در پیشرفت دانش و توسعه روش‌های نوین ایفا می‌کند. این مطالعه می‌تواند به عنوان مرجعی معتبر برای پژوهشگران و توسعه‌دهندگان مدل‌ها در این زمینه مورد استفاده قرار گیرد.

۶- نتیجه‌گیری

در این مقاله، پیشرفت پایگاه داده‌های صوتی تصویری از سال ۱۹۹۹ تاکنون مورد بررسی قرار گرفت که با تغییر فناوری از پایگاه داده‌های کوچک و محدود به سمت پایگاه داده‌های بزرگ و واقع‌گرایانه سوق پیدا کرد. در واقع، مجموعه داده‌های صوتی تصویری که شامل گفتار پیوسته بوده و صرفاً مشتمل بر اصطلاحات یا جملات بودند، تا مجموعه داده‌هایی که ترکیبی از اصطلاحات، جملات و اعداد را شامل می‌شدند، به‌طور جامع مورد بحث و بررسی قرار گرفتند. از آنجایی که ایجاد مجموعه داده‌ها با در نظر گرفتن کلیه ابعاد، مسئله بسیار پیچیده‌ای است بنابراین هر یک از پایگاه داده‌های موجود بر روی یکسری از این ابعاد و ویژگی‌ها متمرکز شده‌اند و مابقی ویژگی‌ها را در نظر نگرفته‌اند. همین طور در این تحقیق با تمرکز بر روی مجموعه داده‌های صوتی تصویری گفتار پیوسته، یک مرور شبه سیستماتیک با هدف گردآوری و تحلیل جامع پژوهش‌های کلیدی مرتبط، انجام شده است. این مقاله به محققانی که در جستجوی مجموعه داده‌های صوتی-تصویری مناسب برای پیشبرد اهداف پژوهشی و ارزیابی نتایج تحقیقات خود در حوزه موضوعات مرتبط با گفتار پیوسته، از جمله لب‌خوانی، پردازش گفتار و مدل‌های زبانی هستند، کمک می‌کند. همچنین، محققان می‌توانند با بررسی دقیق ویژگی‌های هر مجموعه داده و شناخت جامع مزایا و چالش‌های مربوطه، مجموعه داده‌ای را که بیشترین تطابق را با اهداف پژوهشی آنان دارد، انتخاب نموده و بدین ترتیب مسیر مطالعات خود را بهینه‌سازی کنند. چرا که در این مقاله ویژگی‌های کمی، کیفی و نقاط قوت مجموعه داده‌های گفتار پیوسته موجود به تفکیک در قالب جداول و نمودارها، مورد بررسی قرار گرفته و به تصویر کشیده شده است که متناسب با نیاز پژوهشگران، قابل مقایسه و تصمیم‌گیری جهت استفاده در پژوهش‌های پردازش گفتار و لب‌خوانی است. برای تحقیقات آتی، پیشنهاد می‌شود بر توسعه معیارهای استاندارد ارزیابی مجموعه داده‌ها مانند تناسب با معماری‌های یادگیری عمیق نوین، تنوع داده‌ها و کیفیت برچسب‌گذاری تمرکز شود. همچنین، تدوین مجموعه داده‌های واقعی از پلتفرم‌های اجتماعی، با هدف افزایش تناسب با مدل‌های چندوجهی حال حاضر، مورد تأکید قرار گیرد.

مراجع

[۱] مهسا هدایتی‌پور، یاسر شکفته، محسن ابراهیمی‌مقدم، مروری بر پژوهش‌های لب‌خوانی خودکار: دادگان و روش‌ها، مجله ماشین بینایی و پردازش تصویر، سال نهم، شماره چهارم، زمستان ۱۴۰۱.

[2] D. L. Howell, Confusion Modelling for Lip-Reading, Ph.D. thesis, University of East Anglia, 2015.

[3] J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, “Lip reading sentences in the wild”, in: Proc. Conference on Computer Vision and Pattern Recognition, 2017, pp. 3444–3453. Son Chung, J.; Senior, A.; Vinyals, O.; Zisserman, A. Lip Reading Sentences in the Wild. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

[4] E. K. Patterson, S. Gurbuz, Z. Tufekci and J. N. Gowdy, “CUAVE: A new audio-visual database for multimodal human-computer interface research”, in: Proc. International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 2002, pp. 2017–2020.

[5] E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree and et al., “The BANCA database and evaluation protocol”, in: Proc. International Conference on Audio- and Video-Based Biometric Person Authentication, (AVBPA) 2003. Lecture Notes in Computer Science, vol 2688. Springer, pp. 625–638.

[6] J. Huang, G. Potamianos, J. Connell and C. Neti, “Audio-visual speech recognition using an infrared headset”, Speech Communication, Vol. 44, No.1-4,2004, pp. 83–96.

[7] M. Cooke, J. Barker, S. Cunningham and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition”, Journal of the Acoustical Society of America, Vol. 120, No. 5, 2006, pp. 2421–2424.

[8] G. Zhao, M. Barnard and M. Pietikainen, “Lipreading with local spatiotemporal descriptors”, IEEE Transactions on Multimedia, Vol.11, No.7, 2009, pp. 1254–1265.

[9] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison and A. Mashari, “Audio visual speech recognition”, Tech. rep., IDIAP, 2000.

[10] C. Sanderson, “The VidTIMIT database”, Tech. rep., IDIAP, 2002.

[11] T. J. Hazen, K. Saenko, C.-H. La and J. R. Glass, “A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments”, in: Proc. International Conference on Multimodal Interfaces, 2004, pp. 235–242.

[12] J. Trojanova, M. Hruz, P. Campr and M. Zelezn y, “Design and recording of Czech audio-visual database with impaired conditions for continuous speech recognition”, in: Proc. International Conference on Language Resources and Evaluation, 2008.

[13] D. Petrovska-Delacretaz, S. Lelandais, J. Colineau, L. Chen, B. Dorizzi, M. Ardabilian, E. Krichen, M.-A. Mellakh, A. Chaari, S. Guerfi and et al., “The IV 2 multimodal biometric database (including iris, 2D, 3D, stereoscopic, and talking face data), and the IV 2-2007 evaluation campaign”, in: Proc. International Conference on Biometrics: Theory,

International Conference on Language Resources and Evaluation, 2014
[30] I. Anina, Z. Zhou, G. Zhao and M. Pietikainen, “OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis”, in: Proc. International Conference on Automatic Face and Gesture Recognition, IEEE, Vol. 1, 2015, pp. 1–5.
[31] S. Petridis, J. Shen, D. Cetin and M. Pantic, “Visual-only recognition of normal, whispered and silent speech”, in: Proc. International Conference on Acoustics, Speech and Signal Processing (in press), 2018.
[32] T. Afouras, J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, “Deep Audio-Visual Speech Recognition”, IEEE Transactions, 22 December 2018. arXiv:1809.02108
[33] V. Zue, S. Seneff, and J. Glass, “Speech database development: TIMIT and beyond. Speech Communication”, vol. 9, no. 4, 1990, pp. 351-356.
[34] W. Fisher, G. Doddington, and K. Goudie-Marshall, “The DARPA speech recognition research database: specifications and status”, In Proceedings of the DARPA Speech Recognition Workshop, 1986.
[35] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, “The darpa 1000-word resource management database for continuous speech recognition”, International Conference on IEEE In Acoustics, Speech, and Signal Processing, ICASSP-88, 1988, pp. 651-654.
[36] A. A. Karpov and A.L. Ronzhin, “Information enquiry kiosk with multimodal user interface”, Pattern Recogn. Image Analy. Vol. 19, No. 3, 2009, pp. 546–558.
[37] D. E. King, “Dlib-ml: A machine learning toolkit”. The Journal of Machine Learning Research, Vol. 10, 2009, pp. 1755–1758.
[38] C. Tomasi and T. Kanade, “Selecting and tracking features for
image sequence analysis”, Robotics and Automation, 1992.
[39] P. Boersma and et al., “Praat, a system for doing phonetics by computer”, Glot international, vol. 5, 2002, pp. 341–345.
[40] J. C. Wells and et al., “Sampa computer readable phonetic alphabet”, Handbook of standards and resources for spoken language systems, vol. 4, 1997.
[41] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox and R. Harvey, “Extraction of visual features for lipreading”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, 2002, pp. 198–213.
[42] S. J. Cox, R. Harvey, Y. Lan, J. L. Newman and B. J. Theobald, “The challenge of multispeaker lip-reading”, in: Proc. International Conference on Auditory-Visual Speech Processing, 2008, pp. 179–184.
[43] J. Wells, “Sampa computer readable phonetic alphabet”, 2003.
[44] I. Shdaifat,R. Grigat and D. Langmann, “A System for Automatic Lip Reading”, in International Conference on Audio-Visual Speech Processing St. Jorios (ISCA) AVSP, France, 4-7 September 2003.
[45] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus and M. Szykulski, “An audio-visual corpus for multimodal automatic speech recognition”, Journal of Intelligent Information Systems, 2017, pp. 1–26.
[46] J. S. Chung and A. Zisserman, “Lip reading in the wild”, in: Proc. Asian Conference on Computer Vision, 2016, pp. 87–103.
[47] Dominic Howell, Stephen Cox and Barry Theobald, “Visual Units and Confusion Modelling for Automatic Lip-reading”, Image and Vision Computing Journal Elsevier, vol. 51, no. C, pp. 1-12, July 2016.
[48] Joon Son Chung and Andrew Zisserman, “Learning to lip read words by watching videos”, Computer Vision and Image Understanding Journal Elsevier, Vol. 173, August 2018, pp. 76-85

Applications and Systems, 2008, pp. 1–7. 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems, 2008, pp. 1-7, doi: 10.1109/BTAS.2008.4699323.
[14] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong and R. Bowden, “Improving visual features for lip-reading”, in: Proc. International Conference on Auditory-Visual Speech Processing, 2010
[15] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler and R. Orglmeister, “WAPUSK20 - A database for robust audiovisual speech recognition”, in: Proc. International Conference on Language Resources and Evaluation, 2010.
[16] Y. Benezeth, G. Bachman, G. Le-Jan, N. Souviraa-Labastie and F. Bimbot, BL-Database: A French Audiovisual Database for Speech Driven Lip Animation Systems”, Ph.D. thesis, INRIA, 2011.
[17] Y. W. Wong, S. I. Chng, K. P. Seng, L.-M. Ang, S. W. Chin, W. J. Chew and K. H. Lim, “A new multi-purpose audiovisual UNMC-VIER database with multiple variabilities”, Pattern Recognition Letters, Vol. 32, No. 13, 2011, pp.1503–1510.
[18] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, ¨ J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy and et al., “Bi-modal person recognition on a mobile phone: using mobile phone data”, in: Proc. International Workshop on Multimedia and Expo, 2012, pp. 635–640.
[19] N. Harte and E. Gillen, “TCD-TIMIT: An audio-visual corpus of continuous speech”, IEEE Transactions on Multimedia, Vol. 17, No. 5, 2015, pp. 603–615.
[20] V. Verkhodanova, A. Ronzhin, I. Kipyatkova, D. Ivanko, A. Karpov and M. Zelezny, “HAVRUS corpus: high-speed recordings of audio-visual Russian speech”, in: Proc. International Conference on Speech and Computer, 2016, pp. 338–345.
[21] J. S. Chung and A. Zisserman, “Lip reading in profile”, in: Proc. British Machine Vision Conference, 2017.
[22] A. Fernandez-Lopez, O. Martinez and F. M. Sukno, “Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database”, in: Proc. International Conference on Automatic Face and Gesture Recognition, 2017, pp. 208–215.
[23] R. Goecke and J. B. Millar, “The audio-video australian english speech data corpus AVOZES”, in: Proc. International Conference on Spoken Language Processing, 2004, pp. 2525–2528.
[24] A. Ortega, F. Sukno, E. Lleida, A. F. Frangi, A. Miguel, L. Buera and E. Zacur, “AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition”, in: Proc. International Conference on Language Resources and Evaluation, 2004, pp. 763–767.
[25] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu and T. S. Huang, “AVICAR: audio-visual speech corpus in a car environment”, in: Proceedings of Interspeech, 2004.
[26] A.Bastanfard, M.Fazel, A.A. Kelishami and M.Aghaahmadi, “The Persian Linguistic Based Audio-Visual Data Corpus: AVA II, Considering Coarticulation”, S. Boll et al. (Eds.): MMM 2010, LNCS 5916, 2010, pp. 284–294.
[27] A. G. Chitu, K. Driel and L. J. Rothkrantz, “Automatic lip reading in the Dutch language using active appearance models on high speed recordings”, in: Proc. International Conference on Text, Speech and Dialogue, 2010, pp. 259–266.
[28] A. Rekik, A. Ben-Hamadou and W. Mahdi, “A new visual speech recognition approach for RGB-D cameras”, in: Proc. International Conference on Image Analysis and Recognition, 2014, pp. 21–28.
[29] D. Estival, S. Cassidy, F. Cox and D. Burnham, “AusTalk: an audiovisual corpus of Australian English”, in: Proc.

of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 548–555.

[68] Shao, X.; Barker, J. Stream weight estimation for multistream audio–visual speech recognition in a multispeaker environment. Speech Commun. 2008, 50, 337–353. [CrossRef]

[69] Lan, Y.; Harvey, R.; Theobald, B.; Ong, E.J.; Bowden, R. Comparing visual features for lipreading. In Proceedings of the International Conference on Auditory-Visual Speech Processing 2009, Norwich, UK, 10–13 September 2009; pp. 102–106.

[70] Rekik, Ahmed & Ben-Hamadou, Achraf & Mahdi, Walid. (2015). An adaptive approach for lip-reading using image and depth data. Multimedia Tools and Applications. 75. 10.1007/s11042-015-2774-3

[71] Pingping Wu, Hong Liu, Xiaofei Li, Ting Fan, Xuewu Zhang. A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion. IEEE Transactions on Multimedia, Institute of Electrical and Electronics Engineers, 2016, 18 (3), pp.326-338. ff10.1109/TMM.2016.2520091ff. ff

[72] Iain Matthews, Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, A COMPARISON OF MODEL AND TRANSFORM-BASED VISUAL FEATURES FOR AUDIO-VISUAL LVCSR, IEEE International Conference on Multimedia and Expo, ICME 2001,

[73] Cappelletta, L.; Harte, N. Viseme definitions comparison for visual-only speech recognition. In Proceedings of the 2011 19th European Signal Processing Conference, Barcelona, Spain, 29 August–2 September 2011; pp. 2109–2113.

[74] Tim Sheerman-Chase, Eng-Jon Ong and Richard Bowden, "Feature Selection of Facial Displays for Detection of Non Verbal Communication in Natural Conversation", in natural conversation, IEEE International Workshop on Human-Computer Interaction, Kyoto, 2009

[75] R. Bowden, LILir Twotalk Corpus, University of Surrey, 20108 http://www.ee. surrey.ac.uk/Projects/LILiR/twotalk_corpus.

[76] Hatice Çınar Akakın, Bülent Sankur, Robust classification of face and head gestures in video, Image and Vision Computing, Volume 29, Issue 7, 2011, Pages 470-483, ISSN 0262-8856,

[77] Eng-Jon Ong, Richard Bowden, Robust Facial Feature Tracking Using Shape-Constrained Multiresolution-Selected Linear Predictors, IEEE Transactions on Pattern Analysis and Machine Intelligence 33(9):1844 – 1859, October 2011

[78] Philip Tresadern, Chris McCool, Norman Poh, Pavel Matejka, Abdenour Hadid, Christophe Levy, Tim Cootes, and Sebastien Marcel. Mobile biometrics (mobio): Joint face and voice verification for a mobile platform. IEEE pervasive computing, 2012.

[79] George Sterpu, Christian Saam, and Naomi Harte. 2018. Attention-based Audio-Visual Fusion for Robust Automatic Speech Recognition. In Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18). Association for Computing Machinery, New York, NY, USA, 111–115.DOI:https://doi.org/10.1145/3242969.3243014

[80] A. Gabbay, A. Ephrat, T. Halperin and S. Peleg, "Seeing Through Noise: Visually Driven Speaker Separation And Enhancement," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 3051-3055, doi: 10.1109/ICASSP.2018.8462527.

[81] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff and L. Badino, "Face Landmark-based Speaker-independent Audio-visual Speech Enhancement in Multi-talker Environments," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal

[49] Fernandez-Lopez and Adriana and Federico Sukno. "Survey on automatic lip-reading in the era of deep learning", Image Vision Comput, Vol. 78, 2018, pp.53-72.

[50] P. Cısar, M. Zelezny, Z. Krnoul, J. Kanis, J. Zelinka, and L. Muller, "Design and recording of czech speech corpus for audio-visual continuous speech recognition", In Auditory Visual Speech Processing Workshop, 2005.

[51] Garofolo, John S. and et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus", LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[52] S. J. Young and et al., The HTK Book, Version 3.4. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[53] A.A. Karpov and A.L. Ronzhin, "Information enquiry kiosk with multimodal user interface", Pattern Recogn. Image Analy, Vol. 19, No. 3, 2009, pp. 546–558.

[54] K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre, "XM2VTSDB: The extended M2VTS database", in: Proc. International Conference on Audio and Video-based Biometric Person Authentication, Vol. 964, 1999, pp. 965–966.

[55] The M2VTS database; http://www.tele.ucl.ac.be/ M2VTS/m2fdb.html.

[56] M. Igras, B. Ziołko and T. Jadczyk, "Audiovisual database of Polish speech recordings", Studia Informatica, Vol. 33, No. 2B, 2012, pp. 163–172.

[57] N. A. Fox, B. A. OMullane and R. B. Reilly, "VALID: A new practical audio-visual database, and comparative results", in: Proc. International Conference on Audio-and Video-Based Biometric Person Authentication, 2005, pp. 777–786.

[58] P. J. Lucey, G. Potamianos and S. Sridharan, "Patch-based analysis of visual speech from multiple views, in: Proc. International Conference on Auditory-Visual Speech Processing, 2008.

[59] A. Torfi, S. M. Iranmanesh and N. NasrAbadi, "3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition", IEEE Computer Vision and Pattern Recognition Journal, Vol. 5, Issue: 99, 18 Jun 2017, pp. 22081 – 22091.

[60] S. Tamura, C. Miyajima, N. Kitaoka, T. Yamada, S. Tsuge, T. Takiguchi, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda and et al., "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition", in: Proc. International Conference on Auditory-Visual Speech Processing, 2010.

[61] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", the Journal of the Acoustical Society of America, Vol. 87, No. 4, 1990, pp. 1738–1752.

[62] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus", Journal of the Acoustical Society of America, Vol. 123, No. 5, 2008, pp. 38-78.

[63] Zhang, X.; Xu, Y.; Abel, A.K.; Smith, L.S.; Watt, R.; Hussain, A.; Gao, C. "Visual Speech Recognition with Lightweight Psychologically Motivated Gabor Features". Entropy 2020, 22, 1367. https://doi.org/10.3390/e22121367

[64] Assael, Y.M.; Shillingford, B.; Whiteson, S.; De Freitas, N. LipNet: End-to-End Sentence-level Lipreading. arXiv 2016, arXiv:1611.01599.

[65] Wand, M.; Koutník, J.; Schmidhuber, J. Lipreading with long short-term memory. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 6115–6119.

[66] Wand, M.; Schmidhuber, J.; Vu, N.T. Investigations on End-to-End Audiovisual Fusion. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 3041–3045.

[67] Xu, K.; Li, D.; Cassimatis, N.; Wang, X. LCANet: End-to-end lipreading with cascaded attention-CTC. In Proceedings

[96] Saragih, Jason & Goecke, Roland. (2006). Learning active appearance models from image sequences. Proc. VisHCI 2006, Volume 56 of CRPIT. 56.

[97] Chetty, G & Wagner, M 2007, Spatiotemporal person authentication based on multilevel fusion. in M Cree (ed.), Proceedings of the Image and Vision Computing New Zealand Conference. Proceedings of Image and Vision Computing New Zealand, University of Canterbury, New Zealand, pp. 248-253

[98] Chetty, Girija & Wagner, Michael. (2008). A robust spatio-temporal face modelling approach using 3D multimodal fusion for biometric security applications. Proc SPIE. 6944. 10.1117/12.778631.

[99] K. Messer et al., "Face authentication test on the BANCA database," Proceedings of the 17th International Conference on Pattern Recognition IEEE, 2004. ICPR 2004., 2004, pp. 523-532 Vol.4, doi: 10.1109/ICPR.2004.1333826.

[100] Navarathna, Rajitha & Dean, David & Lucey, Patrick & Sridharan, Sridha & Fookes, Clinton. (2010). Recognising Audio-Visual Speech in Vehicles using the AVICAR Database. Australasian International Conference on Speech Science and Technology (SST).

[101] R. Navarathna, P. Lucey, D. Dean, C. Fookes and S. Sridharan, "Lip detection for audio-visual speech recognition in-car environment," 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010), 2010, pp. 598-601, doi: 10.1109/ISSPA.2010.5605429.

[102] Kleinschmidt, Tristan and Dean, David and Sridharan, Sridha and Mason, Michael (2007) A Continuous Speech Recognition Evaluation Protocol for the AVICAR Database. In Proceedings International Conference On Signal Processing and Communication Systems, Gold Coast, Australia.

[103] Bastanfard, Azam & Fazel, Maryam & Abdi, Alireza & Aghaahmadi, Mohammad. (2009). A comprehensive audio-visual corpus for teaching sound Persian phoneme articulation. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics. 169 - 174. 10.1109/ICSMC.2009.5346591.

[104] Dhairya Desai, Priyesh Agrawal, Priyansh Parikh, Piyush Kumar Soni, 2020, Visual Speech Recognition, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020)

[105] Dharin Parekh, Ankitesh Gupta, Shharrnam Chhatpar, Anmol Yash Kumar, Manasi Kulkarni, "Lip Reading Using Convolutional Auto Encoders as Feature Extractor", Computer Vision and Pattern Recognition, 31 May 2018, arXiv:1805.12371v1

[106] Alghowinem, S., Wagner, M., & Göcke, R. (2013). AusTalk — The Australian speech database: Design framework, recording experience and localisation. *2013 8th International* Conference on Information Technology in Asia (CITA), 1-7.

[107] Chung, J.S.; Zisserman, A. Out of time: Automated lip sync in the wild. In Asian Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2016; pp. 251–263. 23.

[108] Petridis, S.; Wang, Y.; Li, Z.; Pantic, M. End-to-End Multi-View Lipreading. In Proceedings of the British Machine Vision Conference, London, UK, 4–7 September 2017.

[109] S. Petridis, Y. Wang, Z. Li, and M. Pantic, ''End-to-end audiovisual fusion with LSTMs,'' 2017, arXiv:1709.04343. [Online]. Available: http://arxiv.org/abs/1709.04343 [118] S.

[110] Fung, I.; Mak, B. End-to-end low-resource lip-reading with maxout CNN and LSTM. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2511–2515.

Processing (ICASSP), 2019, pp. 6900-6904, doi: 10.1109/ICASSP.2019.8682061.

[82] Ariel Ephrat, Tavi Halperin, Shmuel Peleg, "Improved Speech Reconstruction From Silent Video", Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 455-462

[83] G. Sterpu, C. Saam and N. Harte, "Can DNNs Learn to Lipread Full Sentences?", 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 16-20, doi: 10.1109/ICIP.2018.8451388.

[84] George Sterpu, Naomi Harte, "Towards Lipreading Sentences with Active Appearance Models", Presented at The 14th International Conference on Auditory-Visual Speech Processing (AVSP 2017), Image and Video Processing (eess.IV); Audio and Speech Processing (eess.AS), arXiv:1805.11688v1

[85] Kwanchiva Thangthai, Helen L Bear, Richard Harvey, "Comparing phonemes and visemes with DNN-based lipreading", BMVC Lipreading Workshop 2017, arXiv:1805.02924v1

[86] Ivanko, D. & Ryumin, D. (2021). A NOVEL TASK-ORIENTED APPROACH TOWARD AUTOMATED LIP-READING SYSTEM IMPLEMENTATION. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLIV-2/W1-2021. 85-89. 10.5194/isprs-archives-XLIV-2-W1-2021-85-2021.

[87] A. M. Sarhan, N. M. Elshennawy and D. M. Ibrahim, "Hlr-net: a hybrid lip-reading model based on deep convolutional neural networks," Computers, Materials & Continua, vol. 68, no.2, pp. 1531–1549, 2021.

[88] Stavros Petridis, Yujiang Wang, Pingchuan Ma, Zuwei Li, Maja Pantic, "End-to-end visual speech recognition for small-scale datasets", Pattern Recognition Letters,Volume 131, 2020, Pages 421-427, ISSN 0167-8655, ELSEVIER

[89] BBC and Oxford University. 2017. The BBC-Oxford Multi-View Lip Reading Sentences 2 (LRS2) Dataset. http://www.robots.ox.ac.uk/~vgg/data/lip_reading_ sentences/. (2017). Online, Accessed: 11 August 2018.

[90] Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman, "Deep Lip Reading: a comparison of models and an online application", Computer Vision and Pattern Recognition (cs.CV), 15 Jun 2018, arXiv:1806.06053v1

[91] Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman, "ASR is all you need: cross-modal distillation for lip reading", Computer Vision and Pattern Recognition (cs.CV); Sound (cs.SD); Audio and Speech Processing (eess.AS), 28 Nov 2019 arXiv:1911.12747v2 [cs.CV]

[92] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, Mingli Song1, "Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers", in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 04, 2020, pp. 6917–6924.

[93] Courtney L., Sreenivas R. (2020) Using Deep Convolutional LSTM Networks for Learning Spatiotemporal Features. In: Palaiahnakote S., Sanniti di Baja G., Wang L., Yan W. (eds) Pattern Recognition. ACPR 2019. Lecture Notes in Computer Science, vol 12047. Springer, Cham. https://doi.org/10.1007/978-3-030-41299-9_24

[94] Lee Y-H, Jang D-W, Kim J-B, Park R-H, Park H-M. Audio–Visual Speech Recognition Based on Dual Cross-Modality Attentions with the Transformer Model. *Applied Sciences*. 2020; 10(20):7263. https://doi.org/10.3390/app10207263

[95] Goecke, R., & Millar, J. (2004). A Detailed Description of the AVOZES data corpus. In S. Cassidy, F. Cox, R. Mannell, & S. Palethorpe (Eds.), Proceedings of the 10th Australian International Conference on Speech Science and Technology (pp. 486-491). ASSTA. https://assta.org/proceedings/sst/2004/proceedings/papers/sst2004-354.pdf

[126] Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACMTransAudioSpeechLangProcess22(10):1533–154

[127] Sailor HB, Patil HA (2016) Novel unsupervised auditory Filterbank learning using convolutional RBM for speech recognition. IEEE/ACMTransAudioSpeechLangProcess24(12):2341–2353

[128] Ravanelli M, Serdyuk D, Bengio Y (2018) Twin Regularization for online speech recognition. arXiv:180405374

[129] Bhowmik T, Mandal SKD (2016) Deep neural network based phonological feature extraction for Bengali continuous speech. In: Signal and information processing (IConSIP), pp 1–5

[130] Hegde RM, Murthy HA, Gadde VRR (2007) Significance of the modified group delay feature in speech recognition. IEEE Trans Audio Speech Lang Process 15(1):190–202

[131] Souheil Fenghour, Daqing chen and Perry Xiao, Decoder-Encoder LSTM for Lip Reading, (2019) ICSIE '19: Proceedings of the 2019 8th International Conference on Software and Information Engineering, April 2019, Pages 162–166

[132] Saad Naeem, Omer Beg, STRATA: Word Boundaries & Phoneme Recognition From Continuous Urdu Speech using Transfer Learning, Attention, & Data Augmentation, arXiv - EE - Audio and Speech Processing (IF), Pub Date : 2022-04-16, DOI: arxiv-2204.07848

[133] Shiyang Cheng, Pingchuan Ma, Georgios Tzimiropoulos, Stavros Petridis, Adrian Bulat, Jie Shen, Maja Pantic, Towards Pose-invariant Lip-Reading, Computer Vision and Pattern Recognition (cs.CV), 2019, arXiv:1911.06095

[134] A. Garg, J. Noyola, and S. Bagadia. (2016). Lip reading using CNN and LSTM. Technical report Stanford University - CS231n project report.

[135] S. Nadeem Hashmi, H. Gupta, D. Mittal, K. Kumar, A. Nanda and S. Gupta. (2018). A Lip Reading Model Using CNN with Batch Normalization. 11th International Conference on Contemporary Computing.

[136] Aripin, Aripin and Setiawan, Abas, Indonesian Lip-Reading Recognition Using Long-Term Recurrent Convolutional Network. Available at SSRN: https://ssrn.com/abstract=4444973 or http://dx.doi.org/10.2139/ssrn.4444973

[137] S. Petridis, J. Shen, D. Cetin and M. Pantic. (2018). Visual-only recognition of normal, whispered and silent speech. Proc. International Conference on Acoustics, Speech and Signal Processing

[138] Minsu Kim, Joanna Hong, and Yong Man Ro, "Lip-to-speech synthesis in the wild with multi-task learning", In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1-5. IEEE,2023.

[139] Minsu Kim, Jeong Hun Yeo, and Yong Man Ro, "Distinguishing homophenes using multi-head visual-audio memory for lip reading," In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 1174-1182, 2022.

[140] Pingchuan Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis and M. Pantic, "Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096889

[141] Jeongsoo Choi, Minsu Kim, and Yong Man Ro, "Intelligible lip-to-speech synthesis with speech units", arXiv preprint arXiv:2305.19603, 2023

[111] Lee, D.; Lee, J.; Kim, K.E. "Multi-view automatic lip-reading using neural network". In Asian Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2016; pp. 290–302.

[112] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikäinen, "Concatenated frame image based cnn for visual speech recognition", in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 277–289.

[113] Aghaahmadi, Mohammad & Dehshibi, Mohammad Mahdi & Bastanfard, Azam & Fazlali, Mahmood. (2013). Clustering Persian viseme using phoneme subspace for developing visual speech application. Multimedia Tools and Applications. 65. 521-541. 10.1007/s11042-012-1128-7.

[114] Bastanfard, Azam & Aghaahmadi, Mohammad & Abdi, Alireza & Fazel, Maryam & Moghadam, Maedeh. (2009). Persian Viseme Classification for Developing Visual Speech Training Application. 1080-1085. 10.1007/978-3-642-10467-1_104.

[115] Moghadam, Maedeh, Azam Bastanfard, and Mohammad Mahdi Dehshibi. (2011). "Toward Clustering Persian Vowel Viseme: A New Clustering Approach based on HMM.", International Conference on Signal Acquisition and Processing, Volume 2, IEEE

[116] Bastanfard A., Rezaei N.A., Mottaghizadeh M., Fazel M. (2010) A Novel Multimedia Educational Speech Therapy System for Hearing Impaired Children. In: Qiu G., Lam K.M., Kiya H., Xue XY., Kuo CC.J., Lew M.S. (eds) Advances in Multimedia Information Processing - PCM 2010. PCM 2010. Lecture Notes in Computer Science, vol 6298. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15696-0_65

[117] R. Mahdavi, A. Bastanfard and D. Amirkhani, "Persian Accents Identification Using Modeling of Speech Articulatory Features", 2020 25th International Computer Conference, Computer Society of Iran (CSICC), 2020, pp. 1-9, doi: 10.1109/CSICC49403.2020.9050139.

[118] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro, "Multi-modality associative bridging through memory: Speech sound recollected from face video", In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 296-306, 2021.

[119] D. Kumar Margam, R. Aralikatti, T. Sharma, A. Thanda, P. A K, S. Roy and S. M. Venkatesan. (2019). Lip Reading with 3D-2D-CNN BLSTMHMM and word-CTC models. arXiv:1906.12170

[120] Minsu Kim, Hyung-Il Kim, and Yong Man Ro, "Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition", arXiv preprint arXiv:2302.08102, 2023.

[121] S. Petridis, M. Pantic. (2016). Deep complementary bottleneck features for visual speech recognition, Proc. International Conference on Acoustics, Speech and Signal Processing.

[122] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa and M. Daoudi, (2019). Lip reading with hahn convolutional neural networks. Image and Vision Computing.

[123] D.-W. Jang, H.-I. Kim, C. Je, R.-H. Park, and H.-M. Park. (2019). Lip reading using committee networks with two different types of concatenated frame images. IEEE Access, vol. 7.

[124] T. Shirakata and T. Saitoh, "Lip Reading Experiments for Multiple Databases using Conventional Method," 2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Hiroshima, Japan, 2019, pp. 409-414, doi: 10.23919/SICE.2019.8859932.

[125] K. Thangthai, R. Harvey, S. Cox and B.-J. Theobald. (2015). Improving lip-reading performance for robust audiovisual speech recognition using DNNs. Proc. International Conference on Auditory-Visual Speech Processing

Distilled and Efficient Models", 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p.p 7608-7612, 10.1109/ICASSP39728.2021.9415063

[158] Tian Lan; Jun Song; Zhiwei Hou; Kang Chen; Shuping He; Hai Wang, "Event-Triggered Fixed-Time Sliding Mode Control for Lip-Reading-Driven UAV: Disturbance Rejection Using Wind Field Optimization", in IEEE Transactions on Automation Science and Engineering, vol. 22, pp. 9090 - 9103, 19 November 2024, 10.1109/TASE.2024.3496939

[159] Ya Zhao, Rui Xu, and Mingli Song. 2020. A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading. In Proceedings of the 1st ACM International Conference on Multimedia in Asia (MMAsia '19). Association for Computing Machinery, New York, NY, USA, Article 32, 1–6. https://doi.org/10.1145/3338533.3366579

[160] Dawei Luo, Dongliang Xie, Yuqing Zhang, Wanpeng Xie, Baosheng Sun, "Hard sample semantic reconstruction for mandarin visual speech recognition," Digital Signal Processing, vol. 160, p. 105066, 2025

[161] P. Ma, S. Petridis and M. Pantic, "End-To-End Audio-Visual Speech Recognition with Conformers," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 7613-7617, doi: 10.1109/ICASSP39728.2021.9414567.

[162] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, Abdelrahman Mohamed, "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction," 2022, arXiv:2201.02184

[163] Pingchuan Ma, Stavros Petridis, Maja Pantic, "Visual Speech Recognition for Multiple Languages in the Wild," 2022, arXiv:2202.13084

[164] Baosheng Sun, Dongliang Xie, Tiantian Duan, Variable Structure and Modeling Units for Chinese Lipreading, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 36, No. 15, pp. 2256021, ;kvg aniT jovdf

[165] G. Tan, Z. Wan, Y. Wang, Y. Cao and Z. -J. Zha, "Tackling Event-Based Lip-Reading by Exploring Multigrained Spatiotemporal Clues," in IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 5, pp. 82 79-8291, May 2025, doi: 10.1109/TNNLS.2024.3440495.

[166] Yang, S., Zhang, Y., Feng, D., Yang, M.,Wang, C., Xiao, J., Long, K., Shan, S., Chen, X.: LRW-1000: a naturally-distributed largescale benchmark for lip reading in the Wild. In: 2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)

[167] Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition", 2018, arXiv:1809.00496

[168] Xuejuan Chen, Jixiang Du1, Hongbo Zhang, "Lipreading with DenseNet and resBi-LSTM", 2020, Signal, Image and Video Processing, Springer, Vol. 14, No. 5, p.p. 981-989, 10.1007/s11760-019-01630-1

[169] Marzieh Oghbaie, Arian Sabaghi, Kooshan Hashemifard, Mohammad Akbari, "When deep learning deciphers silent video: a survey on automatic deep lip reading," Multimedia Tools and Applications, 22 March 2025

[170] Dmitry Ryumin and Alexandr Axyonov and Elena Ryumina and Denis Ivanko and Alexey Kashevnik and Alexey Karpov, "Audio–visual speech recognition based on regulated transformer and spatio–temporal fusion strategy for driver assistive systems", Expert Systems with Applications, Vol. 252, 2024, p.p. 124159

[171] C. Chen, D. Wang and T. F. Zheng, "CN-CVS: A Mandarin Audio-Visual Dataset for Large Vocabulary Continuous Visual to Speech Synthesis," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal

[142] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. “Sub-word level lip reading with visual attention”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5162‑5172, 2022.

[143] Joanna Hong, Minsu Kim, Daehun Yoo, and Yong Man Ro. “Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition”, arXiv preprint arXiv:2207.06020, 2022.

[144] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and FuruWei. "Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning". IEEE Transactions on Multimedia, 2023.

[145] Hareesh Mandalapu and Aravinda Reddy P N and Raghavendra Ramachandra and K Sreenivasa Rao and Pabitra Mitra and S R Mahadeva Prasanna and Christoph Busch, "Multilingual Audio-Visual Smartphone Dataset And Evaluation", mandalapu 2021, arXiv eprint arXiv:2109.04138, 2021

[146] Setyaningsih ER, Handayani AN, Irianto WSG, Kristian Y, Chen CTSL. LUMINA: Linguistic unified multimodal Indonesian natural audio-visual dataset. *Data Brief*. 2024; 54:110279. Published 2024 Mar 1. doi:10.1016/j.dib.2024.110279

[147] Gerald Schwiebert, Cornelius Weber, Leyuan Qu, Henrique Siqueira, Stefan Wermter, "A Multimodal German Dataset for Automatic Lip Reading Systems and Transfer Learning", arXiv:2202.13403v3 [cs.CV] 11 May 2022

[148] Weicong Chen, Xu Tan, Yingce Xia,Tao Qin, Yu Wang, Tie-Yan Liu, "DualLip: A System for Joint Lip Reading and Generation," in MM '20: Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, USA, 2020.

[149] Xue, Feng and Li, Yu and Liu, Deyin and Xie, Yincen and Wu, Lin and Hong, Richang, "LipFormer: Learning to Lipread Unseen Speakers Based on Visual-Landmark Transformers," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 9, pp. 4507-4517, 2023

[150] Feng Xue, Peng Li, Yu Li, Shujie Li,, "WPELip: enhance lip reading with word-prior information," Multimedia Systems, vol. 31, no. 2, 28 January 2025

[151] Leyuan Qu, Cornelius Weber, Stefan Wermter, "LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 2, pp. 2772-2782, 2024

[152] David Gimeno-Gómez and Carlos-D. Martínez-Hinarejos, "Continuous lipreading based on acoustic temporal alignments," EURASIP Journal on Audio, Speech, and Music Processing, vol. 25, 2024.

[153] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman, "Deep Audio-Visual Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 8717-8727, 2022.

[154] Braun, V. and Clarke, V. (2006) ‘Using thematic analysis in psychology’, Qualitative Research in Psychology, vol. 3, no.2 pp. 77–101. doi: 10.1191/1478088706qp063oa.

[155] Huijuan Wang, Boyan Cui, Quanbo Yuan, Gangqiang Pu, Xueli Liu, Jie Zhu, "Mini-3DCvT: a lightweight lip-reading method based on 3D convolution visual transformer," The Visual Computer, vol. 41, no. 3, pp. 1957 - 1969, 11 June 2024.

[156] Martinez, Brais and Ma, Pingchuan and Petridis, Stavros and Pantic, Maja, "Lipreading Using Temporal Convolutional Networks", 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p.p. 6319-6323, 10.1109/ICASSP40776.2020.9053841

[157] Ma, Pingchuan and Martinez, Brais and Petridis, Stavros and Pantic, Maja, " Towards Practical Lipreading with

Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095796

[172] Z. Zhou, G. Zhao and M. Pietikäinen, "Towards a practical lipreading system," *CVPR 2011*, Colorado Springs, CO, USA, 2011, pp. 137-144, doi: 10.1109/CVPR.2011.5995345.

[173] Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. Hunag, “Lipreading by locality discriminant graph,” in International Conference on Image Processing, pp. 325–328, San Antonio, TX, 2007.

[174] Javad Peymanfard and Samin Heydarian and Ali Lashini and Hossein Zeinali and Mohammad Reza Mohammadi and Nasser Mozayani, "A multi-purpose audio-visual corpus for multi-modal Persian speech recognition: The Arman-AV dataset", Expert Systems with Applications, vol. 238, pp. 121648, 2024.

[175] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic DVS events," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW),* p. 1312–1321, Jun. 2021.

مهشید السادات احصائی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه آزاد قزوین در سال ۱۳۸۵ دریافت کرد؛ سپس در سال ۱۳۸۸ موفق به اخذ مدرک کارشناسی ارشد مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه آزاد قزوین شد. وی در حال حاضر نیز، دانشجوی مقطع دکترا در دانشگاه آزاد کرج در رشته مهندسی کامپیوتر نرم‌افزار است. ایشان از سال ۱۳۸۸ تا کنون عضو هیئت علمی دانشگاه آزاد قزوین بوده و زمینه‌های پژوهشی مورد علاقه ایشان پردازش چندرسانه‌ای و یادگیری عمیق است.

اعظم باستان‌فرد پس از اخذ مدرک کارشناسی در رشته ریاضی کاربردی در کامپیوتر در سال ۱۳۷۷، کارشناسی ارشد و دکتری خود را در رشته مهندسی کامپیوتر در سال‌های ۱۳۸۰ و ۱۳۸۳ از دانشگاه صنعتی توکیو ژاپن دریافت کرد. در سال ۱۳۸۴ موفق به انجام دوره پسادکترا از دانشگاه ژنو در موضوع محاسبه تصاویر پوست انسان مجازی شد. از سال ۱۳۸۵ ایشان عضو هیئت علمی دانشگاه آزاد کرج بوده و زمینه پژوهشی ایشان پردازش چندرسانه‌ای است.