مقاله مروری

مجله ماشین‌بینایی و پردازش‌تصویر
سال نهم، شماره چهارم، زمستان ۱۴۰۱

مروری بر پژوهش‌های لب‌خوانی خودکار:دادگان و روش‌ها

مهسا هدایتی‌پور[۱]، یاسر شکفته[۲]، محسن ابراهیمی مقدم[۲]

چکیده

لب‌خوانی یا فرآیند بازشناسی دیداری گفتار روش تشخیص گفتار از روی حرکات لب گوینده است. در سال‌های اخیر علاقه به لب‌خوانی خودکار رو به افزایش بوده و تحقیقات بسیاری در این زمینه انجام شده است و همچنان به عنوان یک موضوع تحقیقی پیشرو مطرح است. در این مقاله مروری بر تحقیقات انجام شده در زمینه لب‌خوانی خودکار ارائه شده است. فهرستی از پایگاه داده‌های مورد استفاده با در نظر گرفتن خصوصیاتی نظیر تعداد گوینده، محتوای گفتار و کیفیت تصاویر ارائه شده است. در این مقاله، تحقیقات متنوع لب‌خوانی در زمینه تشخیص اعداد، حروف، کلمات و جملات و روش‌های سنتی استخراج ویژگی و دسته‌بندی تا روش‌های جدید مبتنی بر یادگیری عمیق مرور شده‌اند. بعلاوه فعالیت‌های لب‌خوانی زبان فارسی شامل پایگاه داده‌های فراهم شده و تحقیقات مرتبط با جامعیت بیشتر معرفی شده است.

کلیدواژه‌ها

لب‌خوانی، بازشناسی دیداری گفتار،استخراج ویژگی، یادگیری ماشینی، یادگیری عمیق، شناسایی الگو

۱ مقدمه

گفتار عبارت است از رشته‌های آوایی که برطبق الگوهای خاص، سازمان‌یافته و برای ایجاد ارتباط به کار می‌روند. استفاده از گفتار یکی از مؤثرترین روش‌های ارتباطی بین انسان‌ها است. اختلال در گفتار می‌تواند شخص را در ارتباطات خود دچار مشکل سازد و از جمله مواردی که می‌تواند در گفتار اختلال ایجاد کند ناشنوایی یا کم‌شنوایی افراد، عدم توانایی صحبت‌کردن به‌دلیل از دست‌دادن حنجره، قرارگرفتن شخص در محیط‌های پرسروصدا و مواردی از این قبیل است. این موارد، افراد را در ارتباطات روزمره خود در درک کلمات و جملاتی که مخاطب بیان می‌کند، دچار مشکل می‌سازد.برخی از اندام‌های بدن به طور خاص در تولید آواهای زبان به کارگرفته می‌شوند. از جمله این اندام‌ها می‌توان به شش‌ها، نای، حنجره، گلو، حلق، حفره بینی، کام، زبان و دهان اشاره کرد که حالت تغییرپذیری شکل و حجم دهان عامل تعیین‌کننده بسیاری از مشخصه‌های آوایی صداهای گفتار می‌باشد. این تغییرات در شکل لب‌ها نیز نمودار می‌شود. واحدهای اساسی انتزاعی که به لحاظ نظری برای انتقال معانی در یک زبان نیاز است، را واج[۱] گویند. واج‌ها کوچکترین واحد صوتی (آوایی) زبان هستند. از ترکیب واج‌ها هجاها ساخته می‌شوند که واحدهای سازنده کلمات هستند. متناظر واج‌ها در فضای تصویری، ویزم‌ها[۲] هستند.ویزم‌ها، شکل ظاهری لب هنگام بیان واج و کوچک‌ترین واحد دیداری گفتار هستند[۱]. همچنین ویسیلاب[۳] متناظر تصویری هجاها در نظر گرفته می‌شود[۲]. از آن‌جا که شکل لب برای برخی از واج‌ها

این مقاله در مهر ماه ۱۴۰۰ دریافت، در بهمن ماه بازنگری و در فروردین ماه ۱۴۰۱ پذیرفته شد.

[۱] دانش‌آموخته کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز، دانشگاه شهید بهشتی، تهران، ایران
رایانامه:m.hedayatipour@mail.sbu.ac.ir
[۲ و ۳] گروه هوش مصنوعی رباتیک و رایانش شناختی ، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران
رایانامه: {y_shekofteh , m_moghadam}@sbu.ac.ir

نویسنده مسئول: مهسا هدایتی‌پور

DOR: dorl.net/dor/20.1001.1.23831197.1401.9.4.5.2

[1] Phoneme
[2]Viseme
[3]Visyllable

شبیه به هم به‌نظر می‌رسند؛ لذا ویزم‌ها را می‌توان بر اساس میزان شباهت تصویری دسته‌بندی کرد. این دسته‌بندی‌ها در جداولی موسوم به جدول نگاشت واج به ویزم مشخص می‌شود.

لب‌خوانی[۱] یا فرآیند بازشناسی دیداری گفتار[۲]، روش فهمیدن گفتار بوسیله تفسیر حرکات لب گوینده است. در این فرآیند با استخراج ویژگی‌های تصاویر لب و استفاده از آن‌هادر روش‌های شناسایی الگو[۳] و یادگیری ماشین[۴]، حرکات لب به صورت گفتار تفسیر می‌شود. استخراج این ویژگی‌ها با استفاده از روش‌های پردازش تصویر[۵] و یا توسط روش‌های یادگیری ماشین انجام می‌شود.

اولین سیستم لب‌خوانی خودکار در سال ۱۹۸۴ توسط Petajan در دانشگاه Illinois به وجود آمد. در این سیستم از حرکات لب در کنار صوت برای تشخیص بهتر گفتار در محیط‌های نویزی و حالتی که چند گوینده همزمان حرف می‌زدند، استفاده شد[۳]. در سال ۱۹۸۹ استفاده از شبکه‌های عصبی مصنوعی[۶] (ANN) برای استخراج ویژگی[۷] در لب‌خوانی لب‌خوانی توسط Yuhas استفاده شد[۴]. در سال ۱۹۹۳، Goldschen برای اولین بار از مدل‌های مخفی مارکوف[۸] (HMM) در حوزه لب‌خوانی استفاده کرد. در همان سال Silsbee از دانشگاه ایالتی تگزاس روشی برای تشخیص کلمات بر مبنای مشخصه کمی کردن بردار[۹] و مدل مخفی مارکوف پیاده سازی کرد[۵].

تلاش‌ها در جهت لب‌خوانی در بخش‌های مختلف پیگیری شده است: تشخیص حروف الفبا، تشخیص اعداد و کلمه و تشخیص جمله و گفتار پیوسته؛ که پایگاه داده‌های[۱۰] مورد نیاز برای هر یک فراهم شده است. این امر به‌دلیل تفاوت‌هایی بوده است که این حوزه‌ها با یکدیگر داشته‌اند[۶]. با توسعه تحقیقات در حوزه لب‌خوانی، نیاز به مجموعه دادگان استاندارد جهت کمک به فرآیند پژوهش و امکان مقایسه نتایج پژوهش‌های مختلف احساس شد. در سال ۱۹۹۸ مجموعه دادگان AVletters[۷] با حروف الفبای انگلیسی و سال ۱۹۹۹ مجموعه دادگان اعداد انگلیسی XM2VTS[۸] و در سال ۲۰۰۰ مجموعه دادگان جملات انگلیسی IBMViaVoice[۹] ساخته شد. از آن تاریخ به بعد مجموعه دادگان متعددی ساخته شده است که حروف، اعداد، جملات و عبارات یا ترکیبی از آن‌ها را پوشش می‌داد. این مجموعه دادگان شامل تعداد جملات متفاوت، تعداد گویندگان متفاوت و به زبان‌های مختلف بودند.

از سال ۲۰۱۵ شبکه‌های یادگیری عمیق[۱۱] ابتدا برای استخراج ویژگی[۱۰] و از سال ۲۰۱۶ [۶] به‌عنوان دسته‌بند[۱۲] نیز مورد استفاده قرار گرفتند.در برخی کاربردهای بازشناسی دیداری کلمات و جملات، با ترکیب اطلاعات مدل زبانی ساخته شده برمبنای اجزای گفتار و مدل‌های تصویری ساخته شده توسط دسته‌بندها، می‌توان تشخیص دقیق‌تری از محتوای دیداری گفتار داشت. به‌عنوان نمونه در مرجع[۱۱] از مدل زبانی خارجی مبتنی بر کلمات، در کنار دسته‌بند HMM استفاده شده است. گرچه در برخی ساختارها، مدل زبانی می‌تواند به صورت داخلی توسط بخشی از دسته‌بند آموخته شود ولی درمرجع[۱۲] نشان داده شده است که استفاده از اطلاعات مدل زبانی خارجی نسبت به مدل زبانی داخلی، نتایج بهتری داشته است.

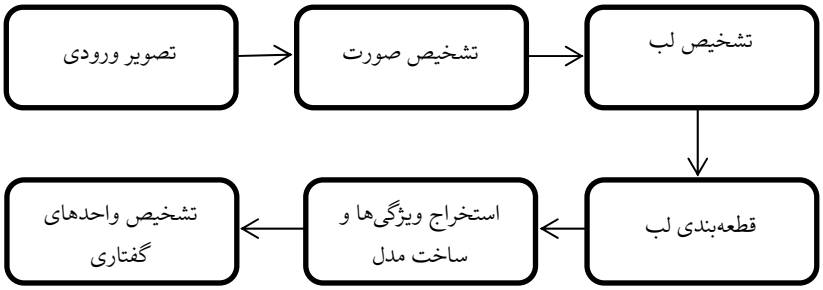
با نگاهی به تاریخچه توسعه فناوری لب‌خوانی از دهه هشتاد قرن بیست تاکنون مشخص می‌شود که لب‌خوانی، همچنان به‌عنوان یک حوزه تحقیقاتی برجسته در زمینه استخراج اطلاعات تصویری مورد توجه است. این توجه نه تنها منجر به افزایش حجم تحقیقات نظری شده است، بلکه کاربردهای عملی لب‌خوانی به صورت نرم‌افزار یا سخت‌افزار[۱۳] را در حال گسترش نشان می‌دهد. از جمله کاربردهای سیستم لب‌خوانی آن است که می‌تواند برای آموزش افرادی که دارای نقص شنوایی هستند و همچنین افرادی که با آن‌ها در ارتباط هستند، مفید باشد. همچنین یک سیستم لب‌خوانی می‌تواند به آژانس‌های اطلاعاتی کمک کند تا با استفاده از یک دوربین از محتویات یک مکالمه از فاصله دور اطلاع حاصل کنند، بدون آنکه اطلاعات صوتی در اختیار داشته باشند. یکی دیگر از کاربردها، استفاده از لب‌خوانی به جای صفحه کلید برای ورود اطلاعات به کامپیوتر است[۱۴].

در این مقاله ابتدا دانش موجود در زمینه لب‌خوانی و روش‌های مطرح و پرکاربرد در این زمینه، همراه با چالش‌های موجود در فرآیند بازشناسی دیداری گفتار بررسی می‌شود. سپس پایگاه داده‌های فراهم شده با قابلیت استفاده در لب‌خوانی که در پژوهش‌ها مورد استفاده بوده‌اند، معرفی شده و پس از آن مروری بر مقالات و پژوهش‌های مطرح در زمینه لب‌خوانی انجام شده است. در این مقاله، به منظور بررسی جامع‌تر لب‌خوانی زبان فارسی، پژوهش‌های زبان فارسی در بخشی جداگانه بررسی و معرفی شده‌اند.

۲ مراحل لب‌خوانی

مراحل لب‌خوانی را می‌توان به چند گام تقسیم کرد[۱۴]. این گام‌ها به صورت مرحله به مرحله در شکل ۱ نشان داده شده است. گام اول، اکتساب تصویر از ورودی است. در مرحله بعد صورت و لب با الگوریتم‌های تشخیص اجزا صورت

[1] Lip-Reading
[2] Visual Speech Recognition
[3] Pattern Recognition
[4] Machine Learning
[5] Image Proccesing
[6] Artifitial Neural Network
[7] Feature Extraction
[8] Hidden Markov Model
[9] Vector Quantization
[10] Database
[11] Deep Learning
[12] Classifier



شکل۱: مراحل لب‌خوانی

تشخیص‌داده شده و سپس با استفاده از روش‌های بخش‌بندی[1]، بخش‌بندی[1]، ناحیه لب از ناحیه غیر لب تفکیک می‌شود. پس از از جداسازی ناحیه لب، با استفاده از الگوریتم‌هایی نظیر تحلیل مؤلفه‌های اصلی[2] (PCA)، تحلیل تفکیک‌کننده خطی[3] (LDA)، تبدیل کسینوسی گسسته[4] (DCT)، الگوی باینری محلی[5](LBP)،محاسبه مشخصه‌های هندسی لب و شبکه‌های عصبی عمیق[6](DNN) بردار ویژگی مناسب استخراج می‌شود. مرحله بعدی در لب‌خوانی، آموزش دسته‌بند مناسب با کمک ویژگی‌های استخراج شده است. در این مرحله، بردار ویژگی‌هایی که بیانگر هر تصویر هستند با برچسب[7] مناسب آن داده به‌دسته‌بند آموزش داده می‌شود تا بتوان از آن در تشخیص تصویری گفتار استفاده کرد. از انواع دسته‌بندهای مورد استفاده می‌توان مدل مخفی مارکوف[8](HMM)، ماشین بردار پشتیبان[9] پشتیبان[9] (SVM) و دسته‌بندهای مبتنی بر شبکه‌های عصبی عمیق را نام برد.

۲-۱ روش‌های تشخیص ناحیه لب

در لب‌خوانی، پس از اکتساب تصویر ورودی، لازم است ناحیه لب در تصویر خام مشخص شود. کیفیت استخراج ناحیه لب در عملکرد سیستم لب‌خوانی تأثیر دارد. روش‌های تشخیص ناحیه لب شامل روش‌های مبتنی بر اطلاعات رنگ، روش‌های مبتنی بر ساختار صورت و روش‌های مبتنی بر مدل است[۱۵].

در روش‌های مبتنی بر اطلاعات رنگ از اختلاف رنگ ناحیه لب و پوست اطراف آن استفاده می‌شود. در این روش‌ها از مؤلفه‌های فضاهای رنگی مختلف استفاده می‌شود[۱۵]. در روش مبتنی بر ساختار صورت از مشخصات توزیع هر یک از اجزای صورت در ساختار صورت استفاده می‌شود. بدین صورت که موقعیت چشم، بینی و دهان نسبت به هم در افراد مختلف ثابت است و ناحیه لب می‌تواند در قیاس با صورت تشخیص داده شود. در روشی مانند ویولا-جونز [۱۶] ابتدا صورت با ویژگی‌های هار و دسته‌بند آدابوست تشخیص داده می‌شود و سپس ناحیه لب از آن استخراج می‌شود.

در روش‌های مبتنی بر مدل، یافتن ناحیه لب منطبق با شکل و ظاهر لب‌هااست. در الگوریتم snake یا ACM[10] [۱۷] ابتدا چند نقطه لب از طریق شرایط محدود شده قطعی به‌دست می‌آید. سپس با تعریف منحنی قابل تغییر شکل، شکل لب مشخص می‌شود. این منحنی با استفاده از ضرایب داخلی محدودیت انرژی و ضرایب خارجی محدودیت انرژی مشخص می‌شود.

۲-۲ روش‌های استخراج ویژگی

پس از مشخص شدن ناحیه لب، قدم بعدی استخراج ویژگی‌های مناسب جهت استفاده در فرآیند دسته‌بندی است.

در مرحله استخراج ویژگی، می‌توان از روش‌های سنتی یا روش‌های نوین مبتنی بر یادگیری عمیق استفاده کرد. روش‌های سنتی استخراج ویژگی شامل روش‌های مبتنی بر پیکسل، روش‌های مبتنی بر شکل و یا ترکیبی از این دو است.

۲-۲-۱روش‌های استخراج ویژگی مبتنی بر پیکسل

در روش‌های مبتنی بر پیکسل، پیکسل‌های تصویر ناحیه لب به‌عنوان ویژگی در نظر گرفته می‌شود و از آنجا که دامنه ویژگی‌ها بزرگ است از روش‌های کاهش بعد نظیر PCA، LDA، DCT و غیره برای کاهش بعد استفاده می‌شود. همچنین در روش‌های جریان نوری[11] تغییرات پیکسل‌ها در هر قاب تصویر نسبت به قاب قبلی، اطلاعات حرکتی لب‌ها را مشخص می‌کند که به عنوان ویژگی در نظر گرفته می‌شود. در برخی دیگر از روش‌های مبتنی بر پیکسل از ویژگی‌های محلی پیکسل‌ها استفاده می‌شود مانند LBP برای تصویر تک قاب دو بعدی و LBPTOP برای چند قاب تصویر از ویدیو که عامل زمان را هم در نظر می‌گیرد.

۲-۲-۲روش‌های استخراج ویژگی مبتنی بر شکل

در روش‌های مبتنی بر شکل از ویژگی‌های هندسی نظیر ارتفاع، عرض و مساحت لب و یا شکل خطوط خارجی لب

[1]Segmentation
[2]Principal Component Analysis
[3]Linear Discriminant Analysis
[4]Discrete Cosine Transform
[5]Local Binary Pattern
[6]Deep Neural Network
[7]Label
[8]Hidden Markov Model
[9]Support Vector Machine
[10]Active Contour Model
[11]Optical Flow

استفاده می‌شود. خطوط خارجی لب با الگوریتمی مانند ACMو با یافتن نقاط کلیدی لب و قرار دادن آن‌ها در یک بردار ویژگی انجام می‌شود[۱۵].

۲-۲-۳ روش‌های استخراج ویژگی ترکیبی

در بسیاری از تحقیقات لب‌خوانی از ترکیبی از ویژگی‌های استخراجی مبتنی بر پیکسل و شکل در قالب یک بردار ویژگی استفاده می‌شود.

۲-۲-۴ روش‌های استخراج ویژگی مبتنی بر یادگیری عمیق

در روش‌های مبتنی بر یادگیری عمیق، مرحله استخراج ویژگی به عنوان جزیی از فرآیند یادگیری و دسته‌بندی درون معماری‌های شبکه عصبی مورد استفاده مانند CNN، در نظر گرفته می‌شود. هر چند در برخی پژوهش‌ها[۱۸] از شبکه‌های خودرمزگذار[1] به عنوان استخراج‌گر ویژگی برای دسته‌بندهای سنتی استفاده شده است. در این شبکه‌ها ابتدا تصویر ورودی در عبور از چند لایه از شبکه عصبی به بردار ویژگی تبدیل می‌شود. این بخش از شبکه، رمزگذار[2] نامیده می‌شود. در بخش دیگر که رمزگشا[3] نامیده می‌شود از این بردار ویژگی برای بازسازی تصویر تصویر خروجی استفاده می‌شود. تفاوت تصویر خروجی و ورودی به‌عنوان تابع خطای شبکه در نظر گرفته می‌شود که طی فرآیند آموزش باید کاهش یابد. از بردار ویژگی به‌دست آمده در بخش میانی شبکه که گلوگاه[4] نامیده می‌شود به‌عنوان بردار ویژگی ویژگی در دسته‌بندها استفاده می‌شود.

۲-۳ دسته‌بندی

مرحله بعدی در لب‌خوانی، آموزش مدل‌های مناسب با کمک ویژگی‌های استخراج شده است. در این مرحله، بردار ویژگی‌هایی که بیانگر هر تصویر هستند با برچسب مناسب آن داده به‌دسته‌بند آموزش داده می‌شود تا بتوان ازمدل‌های آموزش یافته، در تشخیص تصویری گفتار استفاده کرد. از انواع دسته‌بندهای مورد استفاده می‌توان دسته‌بند KNN[5]، HMM، SVM و دسته‌بندهای مبتنی بر شبکه‌های عصبی عمیق را نام برد.

دسته‌بند KNN، نمونه داده‌ها را براساس کمترین فاصله(اقلیدسی) از نمونه‌های برچسب خورده، دسته‌بندی می‌کند.دسته‌بند SVM با نگاشت داده‌ها به فضای مناسب، فاصله بین دسته‌ها را بیش‌تر و آن‌ها را متمایزتر می‌کند. دسته‌بند HMM به صورت دنباله‌ای زنجیره‌ای از حالت‌ها براساس مشاهدات و احتمال وقوع آن‌ها در نظر گرفته می‌شود و به همین دلیل برای کاربرد لب‌خوانی که تشخیص دنباله‌ای از تصاویر است، مناسب است. مدل‌های مخلوط گوسی[6](GMM) را می‌توان به‌عنوان تخمین تابع توزیع احتمال مشاهدات برای هر حالت HMM مورد استفاده قرار داد.

شبکه‌های عمیق در ابتدا به عنوان دسته‌بند برای بردارهای ویژگی استخراج شده به روش‌های سنتی به کار رفته‌اند. در مراحل بعدی شبکه‌های عمیق برای استخراج ویژگی و دسته‌بندی در مدل‌های انتها به انتها[7] به کار رفته‌اند. شبکه CNN به‌عنوان مدل قدرتمند استخراج‌گر ویژگی‌های تصویری و همچنین برای وظایف دسته‌بندی به کار رفته است. این شبکه شامل لایه‌های پیچشی[8] و پولینگ است. لایه‌های پیچشی حاصل‌ضرب داخلی بین یک فیلتر خطی و داده‌های تصویر را محاسبه می‌کند و سپس با اعمال یک تابع فعال‌سازی غیرخطی[9] نظیر سیگموئید[10]، خروجی به‌دست می‌دهد. لایه‌های پولینگ برای کاهش پارامترهای مدل و اندازه تصویر استفاده می‌شود. شبکه‌های RNN[11] و LSTM[12] به‌عنوان شبکه‌هایی که برای مدل کردن دنباله‌ها مناسب هستند در نظر گرفته می‌شوند. این شبکه‌ها شامل سلول‌هایی هستند که در هر سلول ارتباطات داخلی به نام دروازه وجود دارد که آن را به مانند یک واحد حافظه موقت نمایان می‌کند. شبکه‌های LSTM جایگزین شبکه‌های RNN که مشکل محو شدگی گرادیان داشت، شده‌اند. در شبکه‌های LSTM هر سلول شامل سه دروازه[13] ورودی، خروجی و فراموشی است که ارتباطات جمعی و ضربی بین آن‌ها برقرار است تا جریان دائمی خطا را برقرار کند و از مشکل محو شدگی گرادیان جلوگیری کند. در پاره‌ای از معماری‌های شبکه‌های عمیق جهت لب‌خوانی از شبکه‌های پیش آموزش یافته مانند VGGNet[۱۹] و ResNet[۲۰] استفاده شده است.

در کاربردهای لب‌خوانی ممکن است شبکه‌های عمیق مانند CNN و LSTM به صورت ترکیبی به کار روند و شبکه‌های CNN-LSTM را به وجود آورد. در این شبکه‌ها، خروجی شبکه CNN به عنوان ویژگی‌های استخراج شده از تصاویر به LSTM فرستاده می‌شود تا دسته‌بندی انجام شود. نوع دیگر شبکه‌های عمیق Bi-LSTM است که همان شبکه‌های LSTM دوسویه هستند. این شبکه‌ها قادرند برخلاف شبکه‌های LSTM و HMM که خروجی آن‌ها فقط برمبنای اطلاعات گذشته است،

[1] Auto Encoder
[2] Encoder
[3] Decoder
[4] Bottleneck
[5] K-nearest Neighbors

[6] Gaussian Mixture Model
[7] End-to-End
[8] Convolutional
[9] Non-Linear Activation Function
[10] Sigmoid
[11] Recurrent Neural Networks
[12] Long Short Term Memory
[13] Gate

خروجی‌هایی برمبنای محتوای گذشته و آینده یک دنباله تولید کنند.

۳ چالش‌های لب‌خوانی

چالش‌های لب‌خوانی را از دو منظر چالش‌های کیفیت تصاویر و چالش‌های ساخت مدل می‌توان بررسی کرد.

۳-۱ چالش‌های کیفیت تصاویر

از منظر کیفیت استخراج ویژگی از تصویر، نوردهی نامناسب در ضبط تصاویر، نویز تصاویر و موقعیت نامناسب صورت، چالش به حساب می‌آیند. در برخی موارد هم شکل ظاهری ریش، سبیل یا آرایش‌های مرسوم می‌تواند در استخراج ناحیه لب و استخراج ویژگی‌ها مشکلاتی را ایجاد کند[۱۴]. جهت اجتناب از این چالش دادگان تصاویر پیش از ورود به سیستم لب‌خوانی مورد پیش پردازش قرار می‌گیرند تا در صورت نامناسب بودن جهت استخراج ناحیه لب از آنها صرف‌نظر شود و از ورود دادگان نویز به سیستم جلوگیری شود[۲۱].

۳-۲ چالش‌های ساخت مدل

از منظر ساخت مدل برای تشخیص گفتار از حرکات لب، موارد زیر به‌عنوان چالش محسوب می‌شوند:

۱) در نگاشت واج به ویزم که در سیستم‌های لب‌خوانی به کار می‌رود، ابهامات فراوانی وجود دارد چرا که صداهای مختلف در هنگام ادا شدن، شکل لب‌های یکسانی را ایجاد می‌کنند که قابل تمایز نیست[۲۲].

۲) برخی آواها[1] که در فضای میانی دهان ایجاد می‌شوند و برخی دیگر از آواها که در عمق دهان و یا حلق ایجاد می‌شوند، اثرات ناچیزی در نمای خارجی تصاویر لب دارند[۲۳].

۳) برخی واج‌ها بسته به محل قرارگرفتن در کلمه اثرات تصویری متفاوتی نمایش می‌دهند، مثلاً تصویر لب هنگام ادای 'n' در 'banana' با 'nothing' و همچنین "م" در "شما" با "شمع" متفاوت است[۲۳].

۴) در هر زبان کلمات زیادی وجود دارند که هنگام بیان، اثرات همانندی در شکل لب ایجاد می‌کنند مانند 'white' و 'queen' در انگلیسی و "خواست" و "داشت" در فارسی. این امر دقت لب‌خوانی را بسیار کاهش می‌دهد.

۵) سرعت‌های مختلف در گفتار افراد مختلف، تشخیص عبارات را دچار مشکل می‌کند.

۶) ممکن است کلمات توسط گوینده اشتباه تلفظ شوند[۱۴].

۷) مطالعاتی که روی عملکرد انسانی لب‌خوانی انجام شده است، نشان می‌دهد حتی در مورد افراد ناشنوایی که به طور ویژه در لب‌خوانی توانمند شده‌اند در مواجهه با افرادی که پیش‌تر آن‌ها را ندیده‌اند، نرخ صحت عملکردشان کاهش می‌یابد[۱۴]. این امر، نشان دهنده وابستگی به گوینده، در سیستم‌های لب‌خوانی است.

۴ پایگاه داده‌های لب‌خوانی

پایگاه داده‌های صوتی و تصویری بسیاری برای کاربردهای مختلف در بسیاری از زبان‌ها ساخته شده‌اند. تفاوت ساختار این دادگان در کیفیت محتوا، تعداد گویه‌ها[2]، تعداد گویندگان، جنسیت گویندگان، بازه سنی گویندگان، شرایط محیط ضبط و کیفیت ضبط مجموعه دادگان است. نمونه‌هایی از مجموعه داده‌های موجود در زبان‌های مختلف غیر فارسی در جدول۱ شرح داده شده‌اند.

تاکنون برای گفتار زبان فارسی مجموعه دادگان صوتی و تصویری به شرح زیر معرفی شده‌اند:

پایگاه داده صوتی و تصویری AVA[۲۴] برای گفتار درمانی و یادگیری زبان فارسی و طبقه‌بندی ویزم‌های فارسی معرفی شده است. این پایگاه داده شامل کلیه عبارات به شکل CV، CVC، VC، VCV، CVCC است. همچنین ۲۰ جمله از آزمون لب‌خوانی سارا [۲۵] که در آن‌ها تعداد واج‌ها متوازن شده‌اند در این پایگاه داده موجود است. همه ویدیوها با نرخ ۲۵ قاب در ثانیه و با تفکیک‌پذیری ۵۷۶×۷۲۰ پیکسل ضبط شده‌اند. این دادگانتوسط ۲ گوینده زن بیان شده‌است.

پایگاه داده صوتی- تصویری AVA II[۲۶] برای کاربردهای لب‌خوانی ساخته شده است. دادگان موجود در آن، همان دادگان موجود در پایگاه داده AVA است. تعداد گویندگان به ۷ زن و ۷ مرد افزایش یافته است. تصاویر این دادگان در یک استودیوی حرفه‌ای و با ۳ دوربین و نرخ ۲۵ قاب در ثانیه و تفکیک‌پذیری ۵۷۶×۷۲۰ پیکسل ضبط شده‌اند.

پایگاه داده صوتی- تصویری SFAVD[۲۷] با هدف کمک به ساخت سر سخنگو در انیمیشن‌ها فراهم شده است. این پایگاه داده شامل ۶۰۰ جمله فارسی است که هریک شامل ۵ تا ۲۰ کلمه است. در انتخاب کلمات و جملات، قواعد زبان‌شناسی و چگونگی تأثیر آن‌ها در شکل لب در نظر گرفته شده است. این کلمات در چهار مرحله از پایگاه داده‌های PEYKARE[۲۸]، FARSDAT[۲۹]انتخاب شده‌اند. جملات انتخابی همه کلمات پرکاربرد، آواها، diaphoneها و هجاهای فارسی را شامل می‌شوند. همه ویدیوها با نرخ ۳۰ قاب در ثانیه ضبط شده‌اند. این پایگاه داده توسط یک گوینده مرد بیان شده است.

[1] Utterance

پایگاه داده صوتی- تصویری PAVID-CVs[۳۰] شامل همه هجاهای CV زبان فارسی و ۵۵۲ کلمه پرتکرار زبان فارسی که این هجاها را شامل می‌شوند، است. این کلمات برمبنای پرتکرار بودن در پیکره بی‌جن خان[۲۸] و به کمک پیکره آوایی FARANET انتخاب شده‌اند. قواعد زبان‌شناسی و توازن هجاها در کلمات انتخابی رعایت شده است. این دادگان توسط ۲۰ گوینده زن و ۲۰ گوینده مرد در شرایط محیطی شبیه به زندگی روزمره بیان شده است. همه ویدیوها با یک دوربین و با نرخ ۵۰ قاب در ثانیه و تفکیک‌پذیری ۱۰۸۰*۱۹۲۰ پیکسل ضبط شده‌اند. در جدول ۲ مجموعه داده‌های موجود در زبان فارسی مشخص شده است. همچنین در شکل‌های ۲ و ۳ نمونه هایی از تصاویرپایگاه داده PAVID-CVs و GRIDنشان داده شده است([۳۰]، [۳۱]).

۵ مروری بر پژوهش‌های لب‌خوانی زبان‌های غیر فارسی

در این بخش سعی شده است تا کارهای متنوع و شاخص در زمینه لب‌خوانی بررسی شود. تنوع کارهای انجام شده از نظر حیطه کار اعم از لب‌خوانی اعداد و حروف تا کلمات و جملات و همچنین کارهای متآخر در زمینه لب‌خوانی گفتار پیوسته مدنظر قرار گرفته است. همچنین تنوع ویژگی‌های استخراج شده و روش‌های دسته‌بندی در کارهای بررسی شده مورد توجه بوده است تا مرور نسبتاً جامعی از کارهای انجام شده و روش‌های مورد استفاده در لب‌خوانی، فراهم شود.

جدول ۱ : مشخصات برخی از پایگاه داده های موجود در زبان‌های مختلف

پایگاه داده	سال	ارجاع	زبان	محتوا	تعداد کلاس	گویه‌ها	تعداد گوینده(نفر)	تفکیک‌پذیری	قاب در ثانیه
M2VTS [۳۲]	۱۹۹۷	۲۵۱	فرانسوی	عدد	۱۰	۸۸۵	۳۷	۷۲۰×۵۷۶	۲۵
AVLetters [۷]	۱۹۹۸	۶۲۴	انگلیسی	الفبا	۲۶	۷۸۰	۱۰(۵ زن و ۵ مرد)	۳۷۶×۲۸۸	۲۵
AV-CMU [۳۳]	۱۹۹۸	۶۵	انگلیسی	کلمه	۷۸	۷۸۰	۱۰(۳ زن و ۷ مرد)	۷۲۰×۴۸۰	۳۰
XM2VTS [۸]	۱۹۹۹	۱۸۴۹	انگلیسی	عدد	۱۰	۸۸۵	۲۹۵	۷۲۰×۵۷۶	۲۵
Dutch [۳۴]	۲۰۰۲	۲۹	هلندی	کلمه جمله عدد عبارت	۱۴ ۱۰ ۳ ۵	-	۸(۱ زن و ۷ مرد)	۳۸۴×۲۸۸	۲۵
VIDTIMIT [۳۵]	۲۰۰۲	۸۰	انگلیسی	جمله	۳۴۶	۴۳۰	۴۳(۱۹ زن و ۲۴ مرد)	۵۱۲×۳۸۴	۲۵
CUAVE [۳۶]	۲۰۰۴	۳۵۱	انگلیسی	عدد	۱۰	۷۰۰۰	۳۶(۱۷ زن و ۱۹ مرد)	۷۲۰×۴۸۰	۳۰
AVICAR [۳۷]	۲۰۰۴	۲۱۶	انگلیسی	حرف عدد جمله	۳۶ ۱۳ ۱۳۱۷+	۵۹۰۰۰	۱۰۰(۵۰ زن و ۵۰ مرد)	۷۲۰×۴۸۰	۳۰
AV@CAR [۳۸]	۲۰۰۴	۵۴	اسپانیایی	الفبا عدد جمله	۲۶ ۱۰ ۲۵۰	۸۰۰ ۶۰۰ ۶۰۰۰	۲۰(۱۰ زن و ۱۰ مرد)	۷۶۸×۵۷۶	۲۵
HITBi-CAV [۳۹]	۲۰۰۵	۷	چینی	جمله	۲۰۰	۶۰۰۰	۱۰	۲۵۶×۲۵۶	۲۵
UWB-05-HSAVC [۴۰]	۲۰۰۵	۲۲	چک	جمله	۲۰۰	۲۰۰۰۰	۱۰۰	۷۲۰×۵۷۶	۲۵
GRID [۳۱]	۲۰۰۶	۱۰۰۰	انگلیسی	عبارت	۵۱	۳۴۰۰۰	۳۴(۱۶ زن و ۱۸ مرد)	۷۲۰×۵۷۶	۲۵
AVLetters2 [۴۱]	۲۰۰۸	۸۸	انگلیسی	الفبا	۲۶	۹۱۰	۵	۱۹۲۰×۱۰۸۰	۵۰
IV2 [۴۲]	۲۰۰۸	۲۶	فرانسوی	جمله	۱۵	۴۵۰۰	۳۰۰	۷۸۰×۵۷۶	۲۵
UWB-07-ICAV	۲۰۰۸	۲۰	چک	جمله	۷۵۵۰	۱۰۰۰۰	۵۰	۷۲۰×۵۷۶	۵۰

پایگاه داده	سال	ارجاع	زبان	محتوا	تعداد کلاس	گویه‌ها	تعداد گوینده(نفر)	تفکیک‌پذیری	قاب در ثانیه
[۴۳]									
Ouluvs [۴۴]	۲۰۰۹	۲۹۸	انگلیسی	عبارت	۱۰	۱۰۰۰	۲۰	۷۲۰×۵۷۶	۲۵
CENSREC-1-AV [۴۵]	۲۰۱۰	۴۲	ژاپنی	عدد	۱۰	۳۲۳۴	۴۲ (۲۰ زن و ۲۲ مرد)	۷۲۰×۴۸۰	۳۰
NDUTAVSC [۴۶]	۲۰۱۰	۲۲	آلمانی	عدد	۶۹۰۷	۶۹۰۷	۶۶ (۲۰ زن و ۴۶ مرد)	۶۴۰×۴۸۰	۱۰۰
				کلمه					
				جمله					
LTS5 [۴۷]	۲۰۱۱	۴۳	فرانسوی	عدد	۱۰	۱۸۰	۲۰	۱۹۲۰×۱۰۸۰	۲۵
BL [۴۸]	۲۰۱۱	۱۴	فرانسوی	جمله	۲۳۸	۴۰۴۶	۱۷ (۸ زن و ۹ مرد)	۶۴۰×۴۸۰	۳۰
MIRACL-VCI [۴۹]	۲۰۱۴	۴۸	انگلیسی	کلمه	۱۰	۱۵۰۰	۱۵	۶۴۰×۴۸۰	۱۵
				عبارت					
Aus Talk [۵۰]	۲۰۱۴	۲۹	انگلیسی	عدد	۱۰	۲۴۰۰۰	۱۰۰۰	۶۴۰×۴۸۰	۴۸
				کلمه	۹۶۶	۹۶۶۰۰۰			
				جمله	۵۹	۵۹۰۰۰			
Ouluvs2 [۵۱]	۲۰۱۵	۱۱۳	انگلیسی	عدد	۱۰	۱۵۹۰	۵۳ (۱۳ زن و ۴۰ مرد)	۱۹۲۰×۱۰۸۰	۳۰
				عبارت					
				جمله	۵۳۰	۵۳۰			
TCD-TIMIT [۵۲]	۲۰۱۵	۱۵۶	انگلیسی	جمله	۵۹۵۴	۶۹۱۳	۶۲	۱۹۲۰×۱۰۸۰	۳۰
BBC LRW [۵۳]	۲۰۱۶	۴۰۳	انگلیسی	کلمه	۵۰۰	۴۰۰۰۰۰	۱۰۰۰	۲۵۶×۲۵۶	۲۵
HAVRUS [۵۴]	۲۰۱۶	۲۹	روسی	جمله	۱۵۳۰	۴۰۰۰	۲۰ (۱۰ زن و ۱۰ مرد)	۶۴۰×۴۶۰	۲۰۰
LRS2 [۱۲]	۲۰۱۷	۵۱۳	انگلیسی	جمله	۱۴۹۶۰	۷۴۵۶۴	۱۰۰۰	۱۶۰×۱۶۰	۲۵
VLRF [۵۵]	۲۰۱۷	۲۴	اسپانیایی	جمله	۱۳۷۴+	۱۰۲۰۰+	۲۴ (۲۱ زن و ۳ مرد)	۱۲۸۰×۷۲۰	۵۰
AV-Digits [۵۶]	۲۰۱۸	۳۸	انگلیسی	عدد	۱۰	۷۹۵	۵۳	۱۲۸۰×۷۸۰	۳۰
				عبارت		۵۸۵۰	۹		
AVSD [۵۷]	۲۰۱۹	۳	عربی	کلمه	۱۰	۱۱۰۰	۲۲ (۱۴ زن و ۸ مرد)	۱۹۲۰×۱۰۸۰	۳۰

جدول ۲: مشخصات پایگاه داده‌های موجود در زبان فارسی

پایگاه داده	سال	ارجاع	محتوا	تعداد کلاس	گویه‌ها	تعداد گوینده(نفر)	تفکیک‌پذیری	قاب در ثانیه
AVA [۲۴]	۲۰۰۹	۱۵	هجا، عدد، عبارت و جمله	-	-	۲ زن	۷۲۰×۵۷۶	۲۵
AVA II [۲۶]	۲۰۱۰	۱۲	هجا، عدد، عبارت و جمله	-	-	۱۴ (۷ زن و ۷ مرد)	۷۲۰×۵۷۶	۲۵
SFAVD [۲۷]	۲۰۱۳	۴	جمله	۶۰۰	۶۰۰	۱ مرد	-	۳۰
PAVID-CVs [۳۰]	۲۰۲۱	۰	هجاهای CV	۱۳۸	۲۲۰۸۰	۴۰ (۲۰ زن و ۲۰ مرد)	۱۹۲۰*۱۰۸۰	۵۰
			کلمه	۵۵۲	۱۶۳۲۰			

وابستگی یاعدم وابستگی به گوینده در نتایج پژوهش‌ها نیز دراین بررسی مورد توجه قرار گرفته است. عدم وابستگی به گوینده یا مستقل از گوینده حالتی است که دادگان مورد استفاده در مرحله آزمون مدل‌ها مربوط به فرد یا افرادی باشد که دادگان آن‌ها در مرحله ساخت مدل‌ها استفاده نشده باشد. در صورتی که دادگان آموزش و آزمون مربوط به فرد یا افراد یکسان باشد، وابستگی به گوینده محسوب می‌شود. بررسی‌ها در دو بخش به شرح زیر انجام شده است:

۵-۱ پژوهش‌های لب‌خوانی حروف و اعداد

در مرجع[۵۸] با استفاده از دسته‌بند HMM بر روی دادگان AVLetters2 و استفاده از ویژگی‌های ترکیبی فیلترهایSeive و PCA به دقت ۸۳/۰۰% و با استفاده از ویژگی به‌دست آمده از مدل AAM به دقت ۸۵/۰۰% رسیده‌اند.

در مرجع[۵۹] با استفاده از یک سیستم انتها به انتها مبتنی بر هم ترازی جنگل تصادفی خمینه[1] (RFMA) برای تشخیص حروف در مجموعه دادگان AVLetters و AVLetters2 به ترتیب به دقت‌های ۶۹/۶۰% و ۹۱/۸۰% رسیده‌اند.

در مرجع[۶۰] با استفاده از روش‌های DCT-PCA و DWT-PCA به ترتیب ۳۵ و ۴۲ ویژگی را از یک پایگاه‌داده چینی استخراج کرده‌اند. این پایگاه داده صوتی تصویری شامل ۳۷ کاراکتر پرکاربرد چینی بوده که توسط ۴ گوینده (۲ مرد و ۲ زن) بیان شده است و هر یک از کاراکترها را ۱۰ مرتبه تکرار کرده‌اند. همه ویدیوها با نرخ ۲۵ قاب در ثانیه و با تفکیک‌پذیری۵۷۶×۷۲۰ پیکسل ضبط شده‌اند. تصاویر از جلوی صورت و قسمت بالای بدن با پس زمینه ساده ضبط شده‌اند. با اعمال دسته‌بند HMM بر روی هر یک از ویژگی‌های به‌دست آمده در حالت وابسته به گوینده به ترتیب به دقت ۷۲/۸۰% و ۷۷/۴۰% رسیده‌اند.

در مرجع[۶۱] با استفاده از مدل سهمی خطوط خارجی لب، ۵ ویژگی از تصاویر پایگاه‌داده AV-CMU و ۱۷۷۰ ویژگی با روش STLBP[2] از پایگاه‌داده AVLetters استخراج شده است. همچنین ۳ ویژگی هندسی لب از پایگاه داده AV-UNR که توسط نویسندگان مقاله ساخته شده، استخراج شده است. این پایگاه داده صوتی تصویری شامل ۱۰ کلمه است که توسط ۱۶ گوینده بیان شده و هرکلمه را ۲۰ مرتبه به صورت تصادفی تکرار کرده‌اند. این کلمات متناظر اسپانیایی مفاهیمی مانند بالا، پایین، چپ، راست و ... است. ویدیوها با نرخ ۶۰ قاب در ثانیه و با تفکیک‌پذیری ۴۸۰×۶۴۰ پیکسل ضبط شده‌اند. با استفاده از دسته‌بند جنگل تصادفی[3] به ترتیب در هر پایگاه‌داده به دقت ۶۵/۶۸%، ۶۴/۸۹% و ۷۲/۲۸% دست یافته‌اند. در شبکه‌ای به نام LIPNET[۶۲] که با استفاده از ۳ لایه STCNN، ۳ لایه پولینگ، ۲ لایه Bi-GRU، یک شبکه عصبی چندلایه و CTC loss به صورت انتها به انتها[4] پیاده سازی شده، بر روی پایگاه داده GRID، دو ارزیابی صورت گرفته است. در ارزیابی اول ۲ مرد و ۲ زن برای آزمون و بقیه افراد برای آموزش و در ارزیابی دوم از هر گوینده ۲۵۵ جمله برای آزمون و بقیه جملات برای آموزش استفاده شده است. به ترتیب در ارزیابی اول نرخ خطای کلمه[5] به ۱۱/۴۰% و نرخ خطای کاراکتر[6] به ۶/۴۰% و در ارزیابی دوم نرخ خطای کلمه به ۴/۸۰% و نرخ خطای کاراکتر به ۱/۹۰% رسیده است.

در مرجع[۶۳] با معرفی سیستم RTMRBM[7] که مبتنی بر ماشین‌های بولتزمن محدود شده است و توانایی استخراج اطلاعات معنایی از داده‌ها را دارد، به همراه SVM برای تشخیص حروف انگلیسی روی مجموعه دادگان AVLetters و AVLetters2 به دقت ۶۴/۶۳% و ۳۱/۲۱% رسیده‌اند.

در مرجع[۶۴] با استفاده از تعداد پیکسل‌های تشخیص داده شده از خطوط خارجی[8] لب؛ ویژگی‌هایی مانند ارتفاع، عرض، محیط و نسبت کادر محدوده لب را از ویدیوهای مربوط بهواکه‌های انگلیسی استخراج کرده‌اند. هر کدام از این واکه‌ها توسط یک گوینده، حداقل ده مرتبه تکرار شده‌اند. این ویدیوها با نرخ ۳۰ قاب در ثانیه ضبط شده‌اند. بااعمال ضریب همبستگی پیرسون روی ویژگی‌های به‌دست‌آمده از این تصاویر به دقت ۸۰/۰۰% رسیده‌اند.

شبکه‌ای به نام LCANet بر پایه معماری CNN-LSTM[9] LSTM[9] معرفی شده[۶۵] که از سه بخش تشکیل شده است. ابتدا، سه تصویر پیاپی به یک شبکه سه بعدی CNN وارد می‌شود تا اطلاعات تصویری و همچنین اطلاعات زمانی کوتاه مدت، استخراج شود. سپس دو لایه شبکه بزرگراه[10] برای انتقال مستقیم برخی اطلاعات ورودی به خروجی استفاده شده است. در نهایت، ویژگی‌های استخراج شده به شبکه‌ای ترکیبی از Bi-GRU همراه با مکانیزم توجه[11] داده می‌شود تا اطلاعات زمانی طولانی‌مدت نیز به‌دست آید. با اعمال این شبکه روی کاراکترها و کلمات از پایگاه‌داده انگلیسی GRID به ترتیب به دقت‌های ۹۸/۷۰% و۹۷/۱۰% دست یافته‌اند.

در مرجع[۶۶] از لب‌خوانی به عنوان روشی برای ورود کلمه عبور در سامانه احراز هویت در یک سیستم تلفن همراه استفاده شده است. با استفاده از مختصات نقاط بالا، پایین وگوشه‌های چپ و راست لب به عنوان ویژگی برای چهار حرف انگلیسی از مجموعه دادگان AVLetters با دسته‌بند SVM به دقت ۶۸/۷۵% رسیده‌اند.

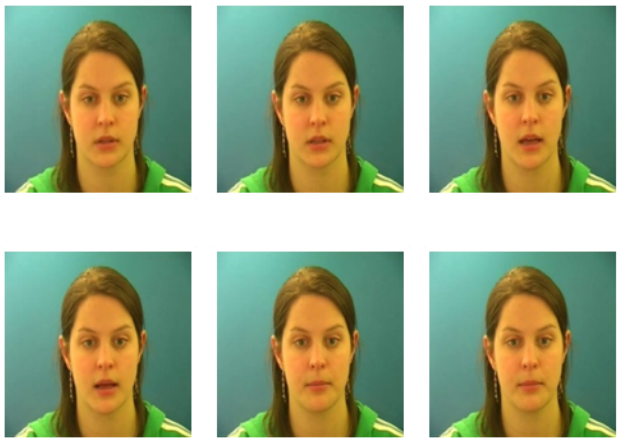
در مرجع[۶۷] با استفاده از ویژگی‌های استخراجی از مدل AAM و دسته‌بند HMM روی دادگان CUAVE به دقت ۸۳/۰۰% رسیده‌اند.

در مرجع[۶۸] از ویژگی‌های DCT و DWT جهت استخراج ویژگی از پایگاه‌داده صوتی تصویری متشکل از ۳۳ عدد انگلیسی استفاده شده است. این اعداد توسط افراد ناشنوای مرد و زن که در محدوده سنی ۱۱ تا ۱۸ سال هستند، بیان شده

[1] Random Forest Manifold Alignment
[2] Spatial atemporal Local Binary Pattern
[3] Random Forest
[4] End-to-End
[5] Word Error Rate
[6] Character Error Rate
[7] Recurrent Temporal Multimodal Restriced Boltzman Machines
[8] Contour
[9] Convolutional Neural Network-Long Short Term Memory
[10] Highway
[11] Attention

است. این کلمات معمولاً برای دریافت پول از مشتری در پیشخوان مغازه‌ها استفاده می‌شوند. ویدیوها با نرخ ۲۵ قاب در ثانیه و با تفکیک‌پذیری ۶۰۰×۹۶۰ پیکسل ضبط شده است. این کلمات شامل اعداد ۱ تا ۱۹ (۱۹ کلمه)، ۲۰، ۳۰ تا ۱۰۰ (۹ کلمه)، thousand، lakh، rupee، rupees، paise (۵ کلمه) هستند. با استفاده از دسته‌بند HMM بر روی ویژگی‌های به‌دست آمده از روش‌های مذکور به ترتیب به دقت‌های ۹۱/۰۰% و ۹۷/۰۰% دست یافته‌اند.

شکل ۲: نمونه‌هایی از تصاویر پایگاه داده PAVID-CVs[۳۰]



شکل ۳: نمونه‌هایی از تصاویر پایگاه داده GRID[۳۱]

در مرجع[۶۹] از ویژگی‌های هندسی مبتنی بر خطوط لب و ویژگی‌های مبتنی بر ناحیه مانند DCT، DWT و PCA جهت استخراج ویژگی از پایگاه‌داده‌ای که شامل ۱۰ عدد انگلیسی (صفر تا ۹) است، استفاده شده است. در این پایگاه‌داده ۱۱ نفر (۶ مرد و ۵ زن) هر عدد را ۲۰ مرتبه تحت شرایط نوری یکنواخت تکرار می‌کنند. ویدیوها با نرخ بیان تقریباً مشابه ضبط شده‌اند. با استفاده از دسته‌بند HMM به دقت میانگین ۹۲/۱۲% برای این اعداد دست یافته‌اند.

در مرجع[۷۰] از ۳ ویژگی هندسی لب مانند ارتفاع، عرض و محیط جهت استخراج ویژگی از پایگاه داده In-house db استفاده شده است. این پایگاه‌داده از ۱۶ نفر (۸ مرد و ۸ زن) که در محدوده سنی ۲۲ تا ۲۵ سال هستند، تشکیل شده است. از همه گویندگان خواسته‌شده که اعداد صفر تا ۹ را بطور متوالی و تصادفی ۶ مرتبه تکرار کنند. ویدیوها با نرخ ۲۵ قاب در ثانیه و تفکیک‌پذیری ۴۸۰×۷۲۰ پیکسل و در شرایط نور طبیعی ضبط شده‌اند. در این پایگاه‌داده ویدیوها با لهجه انگلیسی هندی در پس زمینه آبی ضبط شده و هیچ حرکت سری وجود ندارد. با استفاده از دسته‌بند HMM از نوع ارگودیک، در حالت وابسته به گوینده میانگین دقت زن‌ها و مردها به ترتیب ۷۸/۳۳% و ۷۵/۲۵% و دقت کل ۷۶/۶۰% است. از طرفی دیگر آزمایش‌هایی بر روی پایگاه‌داده CUAVE انجام شده که در حالت وابسته به گوینده میانگین دقت زن‌ها و مردها به ترتیب ۷۹/۶۰% و ۷۷/۸۰% و دقت کل ۷۸/۳۳% شده است.

در مرجع[۷۱] از شبکه‌های عصبی عمیق بولتزمن[1]، DCT و LDA جهت استخراج ویژگی از ویدیوهای مربوط به دنباله اعداد انگلیسی موجود درپایگاه داده AusTalk استفاده شده است. با اعمال دسته‌بند HMM بر روی ترکیبی از این ویژگی‌ها به دقت ۶۹/۱۰% رسیده‌اند.

در مرجع[۷۲] از ویژگی‌های هندسی لب و ویژگی‌های استخراج شده از شبکه‌های عصبی عمیق گلوگاه بر روی پایگاه‌داده CUAVE استفاده شده است. با استفاده از این ویژگی‌ها در دسته‌بندهای GMM-HMM و DNN-HMM به ترتیب به دقت‌های ۶۳/۴۰% و ۶۴/۹۰% دست یافته‌اند.

در مرجع[۷۳] دو دنباله ویژگی محاسبه شده است. در دنباله اول، ۵۰ ویژگی از تصاویر خام دهان و در دنباله دوم، ۵۰ ویژگی از تصاویر تفریق شده دهان استخراج می‌شود. سپس مشتق اول و دوم هر یک از این دنباله ویژگی‌ها محاسبه می‌شود. این ویژگی‌ها از ویدیوهای مربوط به اعداد صفر تا ۹ انگلیسی از پایگاه‌داده CUAVE به‌دست آمده‌اند. در نهایت، این دو دنباله از ویژگی‌ها به دو شبکه LSTM داده شده و سپس با ترکیب اطلاعات این دو دنباله به‌عنوان ورودی به یک شبکه Bi-LSTM، به دقت۶۰/۷۸% دست یافته‌اند.

در مرجع[۷۴] از مختصات نقاط لب به‌دست‌آمده از روش ASM، ضرایب DCT و ویژگی‌های هندسی جهت استخراج ویژگی از ویدیوهای مربوط به اعداد صفر تا ۹ انگلیسی از پایگاه‌داده CUAVE استفاده شده است. با اعمال دسته‌بند HMM به دقت ۶۱/۵۰% و با دسته‌بند شبکه باور عمیق[2] در حالت وابسته به گوینده و مستقل از گوینده در تشخیص واج‌ها به ترتیب به دقت‌های ۷۷/۶۵% و۷۳/۴۰% دست یافته‌اند. همچنین نرخ تشخیص کلمه در حالت وابسته به گوینده و مستقل از گوینده به‌ترتیب ۸۰/۲۵% و ۷۶/۹۱% است.

در مرجع[۵۶] از شبکه خودرمزگذار جهت استخراج ویژگی از تصاویر پایگاه‌داده AV Digits استفاده شده است. این پایگاه‌داده شامل اعداد صفر تا ۹ و عبارات کوتاه انگلیسی

[1] Boltzman Deep Neural Network
[2] Deep Belief Network

می‌باشد که توسط ۵۳ نفر (۴۱ مرد و ۱۲ زن) و در سه حالت عادی، زمزمه و بی‌صدا بیان می‌شوند. گویندگان هر یک از اعداد و عبارات را پنج مرتبه بطور تصادفی بیان می‌کنند. ویدیوها با نرخ ۳۰ قاب در ثانیه و تفکیک‌پذیری ۷۸۰×۱۲۸۰ پیکسل ضبط شده‌اند. با اعمال دسته‌بند Bi-LSTM بر روی اعداد و عبارات در حالت مستقل از گوینده به ترتیب به دقت‌های ۶۹/۷۰% و ۶۸/۰۰% دست یافته‌اند.

در جداول ۳و ۴خلاصه بررسی‌های انجام شده در تشخیص حروف و اعداد جهت مقایسه ارائه شده است.

بررسی‌های انجام شده نشان می‌دهد که بیشترین ویژگی‌های مورد استفاده در روش‌های سنتی تشخیص حروف و اعداد در لب‌خوانی، DCT و یا ترکیب DCTبا تبدیل‌های نظیر LDAیا PCAو AAMبوده است. از دسته بند HMMدر بیش‌تر پژوهش‌ها استفاده شده است. همچنین در برخی پژوهش‌ها از دسته‌بندهایی نظیر جنگل تصادفی و یا SVMنیز استفاده شده است. به منظور مقایسه روش‌ها لازم است عملکرد آن‌ها را روی مجموعه دادگان یکسان بررسی کرد. مجموعه دادگان CUAVEبرای اعداد و AVLetters2برای حروف در روش‌های سنتی پرکاربرد بوده‌اند. در مجموعه دادگان CUAVE که مخصوص شناسایی اعداد است بیش‌ترین دقت در مرجع[۶۷] با استفاده از ویژگی AAMو دسته‌بند HMMبه میزان ۸۳/۰۰% به‌دست آمده است. در مجموعه دادگان AVLetters2در مرجع[۵۹] با استفاده از ویژگی AAMو دسته‌بند RFMAبه دقت ۹۱/۸۰% دست یافته‌اند. مقایسه نتایج جداول نشان می‌دهد که در روش‌های سنتی تشخیص اعداد و حروف با وجود استفاده بیش‌تر از DCTو ترکیبات آن، بهترین دقت‌ها با ویژگی AAMدر دسته‌بند HMMبه‌دست آمده است.

در روش‌های یادگیری عمیق بررسی روی مجموعه دادگان AVLettersنشان می‌دهد که بیش‌ترین دقت به میزان ۶۴/۶۳% در مرجع[۶۳] با استفاده از سیستمی مبتنی بر ماشین‌های بولتزمن با قابلیت استخراج اطلاعات معنایی از داده‌ها به‌دست آمده است؛ که البته این نتیجه در مقایسه با نتیجه حاصل از سیستم مبتنی بر RFMAدر مرجع[۵۹] که ۶۹/۹۰% بوده است، کمتر است. بنابراین برای تشخیص حروف، سیستم‌های سنتی بهتر از روش‌های یادگیری عمیق عمل کرده‌اند که البته ممکن است به دلیل کم بودن حجم دادگان برای روش‌های یادگیری عمیق باشد.

۵-۲ پژوهش‌های لب‌خوانی کلمه و جمله

در مرجع[۷۵] از ضرایب دو بعدی تبدیل فوریه به‌عنوان ویژگی استخراج شده از پایگاه‌داده استفاده شده است. این پایگاه‌داده شامل ۴۴ کلمه و عدد است که توسط یک گوینده بیان شده‌اند. ویدیوهای این پایگاه‌داده با نرخ ۲۵ قاب در ثانیه و تفکیک‌پذیری ۲۴۰×۳۲۰ پیکسل ضبط شده‌اند. با استفاده از دسته‌بند HMM، به دقت ۶۰/۰۰% رسیده‌اند.

در مرجع [۷۶] از روش DCT-PCA سراسری[1] و DCT-PCA بلوکی[2] جهت استخراج ویژگی از پایگاه‌داده صوتی تصویری HIT Bi-CAVDB که به زبان چینی است، استفاده شده است. این پایگاه‌داده توسط مؤسسه چینی فناوری دوحالته[3] هاربین ساخته شده و شامل تصاویر صورت در روشنایی طبیعی است. در این پایگاه داده ۲۰۰ کلمه توسط ۱۰ گوینده در سه نوبت تصویربرداری بیان شده است، این کلمات ۹۶ هجای چینی را پوشش می‌دهند. همه ویدیوها با نرخ ۲۵ قاب در ثانیه و تفکیک‌پذیری ۲۵۶×۲۵۶ پیکسل ذخیره شده‌اند. با اعمال دسته‌بند شبه‌پیوسته HMM[4] بر روی هر یک از ویژگی‌ها به ترتیب به بیشینه دقت ۷۷/۱۰% و ۷۶/۹۰%در حالت وابسته به گوینده دست یافته‌اند.

در مرجع[۷۷] از بردارهای ویژگی موجود در پایگاه داده ویژگی‌های PLF-08، استفاده شده است. این پایگاه داده شامل ۳۵۷۳۰ نمونه است که ۱۴ نفر ۳۰ کلمه را ۸۵ مرتبه بیانکرده‌اند و برای هر قاب تصویر ۱۲ ویژگی استخراج و در پایگاه داده،ذخیره شده است. در حالت وابسته به گوینده با استفاده از دسته‌بندهای HMM، KNN وANN به ترتیب دقت‌های ۹۱/۴۱%، ۹۷/۰۴% و ۹۷/۴۵% گزارش شده است. در حالتمستقل از گوینده نیز با استفاده از دسته‌بندهای HMM، KNNو ANN به ترتیب دقت‌های ۶۹/۴۵%، ۷۸/۹۰% و ۸۲/۵۸% گزارش شده است.

در مرجع[۷۹] از ویژگی‌های به‌دست‌آمده از افکنش[5] عمودی و افقی جهت استخراج ویژگی از پایگاه‌داده‌ای که شامل ۳ کلمه اندونزیایی است، استفاده شده است. این پایگاه‌داده شامل تصاویر۱۰ نفر (۵ مرد و ۵ زن) است که هر کلمه را دو مرتبه تکرار می‌کنند. این ویدیوها با نرخ ۲۵ قاب در ثانیه و تفکیک‌پذیری ۴۸۰×۶۴۰ پیکسل و از ناحیه دهان افراد ضبط شده‌اند. با استفاده از شبکه عصبی انتشار رو به عقب به دقت ۶۷/۰۰% دست یافته‌اند.

در مرجع[۸۰] از بردار ویژه تصاویر لب به‌دست‌آمده توسط PCA و هیستوگرام گرادیان[6] جهت استخراج ویژگی از داده‌های ۱۹ نفر از پایگاه‌داده GRID استفاده شده است. با اعمال دسته‌بند SVM در حالت وابسته به گوینده به دقت ۷۱/۳۰% دست‌یافته‌اند. از طرفی دیگر با استفاده از روش انتها به انتها که مبتنی برشبکه‌ای متشکل از یک لایه پیش‌خور و دو لایه LSTM است به دقت ۷۹/۶۰% دست یافته‌اند.

---

[1] Entire Manner
[2] Block Manner
[3] Bimodal
[4] Semi-Continues
[5] Projection
[6] Histogram Of Gradient

جدول ۳: خلاصه‌ای از پژوهش‌های انجام شده در لب‌خوانی حروف زبان‌های مختلف

سال	مرجع	مدل		پایگاه داده	شرایط	دقت(درصد)
		ویژگی	دسته‌بند			
۲۰۱۷	[۶۴]	ویژگی هندسی لب	ضریب همبستگی پیرسون	دادگان اختصاصی	وابسته به گوینده	۸۰/۰۰
۲۰۱۴	[۶۰]	DCT-PCA	HMM	چینی	وابسته به گوینده	۷۲/۸۰
		DWT-PCA				۷۷/۴۰
۲۰۱۸	[۶۵]	CNN-LSTM	Bi-GRUهمراه با مکانیزم توجه	GRID	مستقل از گوینده	۹۸/۷۰
۲۰۱۶	[۶۲]	شبکه انتها به انتها (STCNN، Bi-GRU، شبکه عصبی چندلایه و CTC loss)		GRID	ارزیابی اول	نرخ خطا ۶/۴۰
					ارزیابی دوم	نرخ خطا ۱/۹۰
۲۰۰۸	[۵۸]	ویژگی‌های ترکیبی فیلترهای Seiveو PCA	HMM	AVLetters2	مستقل از گوینده	۸۳/۰۰
		AAM				۸۵/۰۰
۲۰۱۳	[۵۹]	شبکه انتها به انتها (مبتنی بر RFMA)		AVLetters	مستقل از گوینده	۶۹/۶۰
				AVLetters2		۹۱/۸۰
۲۰۱۴	[۶۱]	STLBP	جنگل تصادفی	AVLetters	–	۶۴/۸۹
۲۰۱۶	[۶۳]	RTMRBM	SVM	AVLetters	مستقل از گوینده	۶۴/۶۳
				AVLetters2		۳۱/۲۱
۲۰۱۷	[۶۶]	مختصات نقاط بالا، پایین وگوشه‌های چپ و راست لب	SVM	AVLetters	مستقل از گوینده	۶۸/۷۵
۲۰۱۸	[۷۸]	انتها به انتها (CNN)		TULAVD (زبان چک)	وابسته به گوینده	۶۵/۸۰

جدول ٤: خلاصه‌ای از پژوهش‌های انجام شده در لب‌خوانی اعداد زبان‌های مختلف

سال	مرجع	مدل		پایگاه داده	شرایط	دقت(درصد)
		ویژگی	دسته‌بند			
۲۰۱۱	[۶۸]	DCT	HMM	دادگان اختصاصی	–	۹۱/۰۰
		DWT				۹۷/۰۰
۲۰۱٤	[۶۹]	ویژگی هندسی لب، DCT، DWTو PCA	HMM	دادگان اختصاصی	–	۹۲/۱۲
۲۰۱٤	[۷۰]	ویژگی هندسی لب	HMM	دادگان اختصاصی	وابسته به گوینده	دقت زن‌ها ۷۸/۳۳
						دقت مردها ۷۵/۲۵
						۷٦/٦۰
				CUAVE	وابسته به گوینده	دقت زن‌ها ۷۹/٦۰
						دقت مردها ۷۷/۸۰
						۷۸/۳۳
۲۰۰۹	[۶۷]	AAM	HMM	CUAVE	مستقل از گوینده	۸۳/۰۰
۲۰۱۷	[۷۲]	ویژگی هندسی لب و شبکه عصبی عمیق گلوگاه	GMM-HMM	CUAVE	مستقل از گوینده	٦۳/٤۰
			DNN-HMM			٦٤/۹۰
۲۰۱۷	[۷۳]	۵۰ ویژگی از تصاویر دهان و ۵۰ ویژگی از تصاویر تفریق شده و مشتق اول و دوم	Bi-LSTM	CUAVE	مستقل از گوینده	۶۰/۷۸
۲۰۱۸	[۷۴]	ASM، DCTو ویژگی هندسی لب	HMM	CUAVE	–	٦۱/۵۰
			شبکه باور عمیق		وابسته به گوینده	۸۰/۲۵
					مستقل از گوینده	۷٦/۹۱
۲۰۱۸	[۵۶]	شبکه خودرمزگذار	Bi-LSTM	AV Digits	مستقل از گوینده	٦۹/۷۰
۲۰۱۵	[۷۱]	شبکه عصبی عمیق بولتزمن، DCTو LDA	HMM	AusTalk	–	٦۹/۱۰

در مرجع[۸۱] از ویژگی‌های LBP-TOP جهت استخراج ویژگی از پایگاه‌داده هندی استفاده شده است. این پایگاه‌داده شامل ۱۰ کلمه هندی بوده که توسط ۲۰ نفر بیان شده است و هر کلمه را ۲۰ مرتبه تکرار کرده‌اند. ویدیوها با دوربین رایانه[1] و با نرخ ۱۵ قاب در ثانیه ضبط شده‌اند. در هنگام تصویربرداری، نور از روبه‌رو به صورت افراد تابیده شده است. با استفاده از شبکه عصبی چندلایه انتشار رو به عقب[2] به دقت ۹۷/۰۰% دست یافته‌اند.

در مرجع[۸۲] از ۴ بردار ویژگی شامل ویژگی‌های پویای استخراج شده از شبکه عصبی گلوگاه، ویژگی‌های ایستای به‌دست‌آمده از CNN، ویژگی‌های DCT و ترکیب این ویژگی‌های پویا و ایستا روی پایگاه‌داده منتخب، استفاده شده است. این پایگاه‌داده شامل ۲۸۳۶ کلمه ژاپنی انتخاب شده از پایگاه‌داده گفتار چینیATR[۸۳]است. ویدیوها با نرخ ۶۰ قاب در ثانیه از یک گوینده مرد فیلمبرداری شده‌اند. با استفاده از ۴ بردار ویژگی مذکور به صورت جداگانه، در دسته‌بند HMM، به ترتیب به دقت‌های ۷۱/۷۶% و ۵۸/۹۴% و ۵۱/۵۸% و ۵۶/۴۸% دست یافته‌اند.

در مرجع[۲۲] از مدل‌های AAM و ASM جهت استخراج ویژگی‌های شکل و ظاهر از پایگاه‌داده‌ای صوتی تصویری استفاده شده است. در این پایگاه‌داده از ۱۲ گوینده (۷ مرد و ۵ زن) استفاده شده است و هر گوینده ۲۰۰ جمله را از پیکره RM انتخاب و بیان کرده‌است. تعداد کل لغات حدود ۱۰۰۰ کلمه است. ویدیوها از ۵ زاویه مختلف و با تفکیک‌پذیری ۱۰۸۰×۱۹۲۰ پیکسل ضبط شده‌اند. ویژگی‌ها با روش نرمال‌سازی میانگین نرمال شده و با اعمال LDA و رگرسیون خطی احتمال حداکثر فضای ویژگی[3] و دسته‌بند HMM-DNN، در حالت وابسته به گوینده به بیشینه دقت ۶۴/۰۰% دست یافته‌اند.

در مرجع[۷۸] از ویژگی‌های DCT، LBPTOP، HOGTOP و ویژگی‌های زمانی مکانی به‌دست آمده از شبکه‌های عصبی عمیق برروی پایگاه‌داده صوتی تصویری TULAVDاستفاده شده است. این پایگاه داده شامل ۵۰ کلمه و ۱۰۰ جمله به زبان چک بوده و برطبق تعادل آوایی انتخاب شده‌اند. در این پایگاه‌داده، تصویربرداری از ۵۴ گوینده (۳۱ مرد و ۲۳ زن) در محدوده سنی بین ۲۰ تا ۷۰ سال انجام شده است. ویدیوها با نرخ ۳۰ قاب در ثانیه و تفکیک‌پذیری ۴۸۰×۶۴۰ پیکسل ضبط شده‌اند. نتایج موجود نشان می‌دهند که با استفاده از شبکه عصبی عمیق به بهبود ۲۲/۰۰% نسبت به حالتی که تنها از داده صوتی استفاده شده، دست یافته‌اند. با استفاده ازویژگی‌های زمانی-مکانی شبکه‌های عصبی عمیق همراه با استفاده از مدل‌های زبانی به دقت ۹۲/۳۰% برای کلمات در حالت وابسته به گوینده دست یافته‌اند. دقت تشخیص ویزم نیز ۶۵/۸۰% گزارش شده است.

در مرجع[۸۴] از شبکه خودرمزگذار پیچشی جهت استخراج ویژگی از ۳ پایگاه‌داده استفاده شده است. با اعمال دسته‌بند LSTM بر روی ۹ کلمه ساده و ۲۷ کلمه ابهام‌آمیز از پایگاه‌دادهBBC's LRW به ترتیب به دقت‌های ۸۵/۶۱% و ۷۷/۴۲%، بر روی ۱۰ کلمه از پایگاه‌داده MIRACLE_VCI در حالت وابسته به گوینده و مستقل از گوینده به ترتیب به دقت‌های ۹۸/۰۰% و ۶۳/۲۲% و بر روی ۵ نفر از پایگاه داده GRID به دقت ۸۴/۸۰% دست یافته‌اند.

در مرجع[۸۵] برای تشخیص دیداری گفتار پیوسته از مجموعه دادگان شامل ۳۸۸۰ ساعت از ویدیوهای یوتیوب استفاده شده است. در این روش از یک معماری یادگیری عمیق به نام ویژن به واج[4] (V2P) که تصاویر ویدیویی خام را به دنباله‌ای از کلمات تبدیل می‌کند، استفاده شده است. این اولین بار است که مدل‌های برمبنای فونم(واج) با تکنیک‌های رمزگشای فونم به کلمه ترکیب می‌شوند. معماری V2P شامل یک ماژول شبکه پیچیده سه بعدی برای استخراج ویژگی‌های فضایی-زمانی[5] از ویدیو و یک ماژول زمانی برای تقویت این ویژگی‌ها در طی زمان است. خروجی این شبکه توزیعی از فونم‌ها است. در این روش، نرخ خطای تشخیص کلمه به ۴۰/۹۰% رسیده است که بهترین دقت تاکنون است.

درمرجع[۸۶] از دو شبکه استفاده شده که هر یک از چند لایه پیش‌خور و لایه‌های LSTM تشکیل شده است. در ابتدا ناحیه لب به صورت تصاویری با ابعاد ۸۰×۴۰ پیکسل از قاب‌ها استخراج می‌شود. به یکی از شبکه‌ها این تصاویر و به شبکه دیگر تفاضل پیکسل‌های هر تصویر با تصویر قبلی به‌عنوان ویژگی‌های حرکتی[6]، به‌عنوان ورودی داده شده‌اند. سپس سپس خروجی این دو شبکه به شبکه LSTM داده شده است. در نهایت خروجی به یک لایه softmax داده می‌شود تا یک مدل انتها به انتها فراهم شود. نتایج این مدل بر روی دادگان GRID بررسی شده است. بهترین نتایج در حالتی که شبکه‌ها شامل دو لایه پیش‌خور با ۲۵۶ گره و یک لایه LSTM با ۲۵۶ سلول بوده‌اند، به‌دست آمده‌است. دقت در حالتی‌که مدل با یک گوینده آموزش یافته و با یک گوینده دیگر آزمون شده، دقت آزمون ۲۶/۳۰% بوده است. در حالتیکه مدل با ۸ گوینده آموزش یافته و با یک گوینده آزمون شده، دقت آزمون ۵۵/۳۰% گزارش شده است. این نتایج نشان می‌دهد که وابستگی به گوینده همچنان چالشی اساسی در لب‌خوانی است.

[1] Webcam
[2] Back Propagation
[3] Feature-Space Maximum Likelihood Linear Regression(FMLLR)
[4] Vision to Phoneme
[5] Spatial-Temporal
[6] Motion

در مرجع[۸۷] با بکارگیری شبکه‌های پیچشی زمانی[1] (TCN) به جای CRU دوسویه، در تشخیص کلمات مجزا روی مجموعه دادگان LRW و LRW1000 دقت‌های ۸۵/۳۰% و ۴۱/۴۰% حاصل شده است. معماری استفاده شده در این شبکه برپایه ResNet18 است که لایه اول آن با یک شبکه پیچشی سه بعدی جایگزین شده است.

در مرجع[۸۸] با افزودن معماری متراکم[2] به TCN و ساخت معماری DC-TCN دقت‌های ۸۸/۳۶% و ۴۵/۶۵% در LRW و LRW1000 برای تشخیص کلمات به دست آمده است. در معماری اتصال متراکم از خروجی لایه‌های کانولوشن برای تغذیه لایه‌های بعدی و انتقال ویژگی‌های استخراج شده به آن‌ها استفاده شده است.

در مرجع[۸۹] با استفاده از ویژگی‌های تبدیل DCT در دسته‌بند HMM روی دادگان GRID برای تشخیص عبارات انگلیسی به دقت ۵۷/۰۰% رسیده‌اند.

در مرجع[۹۰] با استفاده از ضرایب DCT سه بعدی به‌دست‌آمده از پایگاه داده ViDTIMIT و اعمال دسته‌بند سه بعدی HMM به دقت ۹۶/۰۰% دست یافته‌اند.

در مرجع[۹۱] با معرفی یک شبکه CNN استخراج ویژگی مبتنی بر VGG-M به نام SyncNet و دسته‌بند LSTM روی دادگان Ouluvs2 به دقت ۹۴/۱۰% رسیده‌اند.

در مرجع[۹۲] روی دادگان Ouluvs2 با استفاده از ترکیب ویژگی‌های DCT و PCA و دسته‌بند HMMبه دقت ۶۳/۰۰% و با استفاده از ترکیب DCT و LDA به دقت ۷۴/۰۰% رسیده‌اند. همچنین با استفاده از مقادیر خام پیکسل در دسته‌بندمدل‌های نهفته متغیر[3](LVM)دقت ۷۳/۰۰% به‌دست آمده است. با استفاده از ویژگی‌های استخراج شده CNN در دسته‌بند LSTM، دقت ۸۱/۱۰% گزارش شده است.

در مرجع[۹۳] با استفاده از ترکیب قاب‌های بهم چسبانیده و معماری مختلف شبکه‌های از پیش ساخته‌شدهNIN، AlexNet و GoogleNet به ترتیب به دقت‌های ۸۱/۱۰%، ۸۲/۸۰% و ۸۵/۶۰%روی دادگان Ouluvs2 به‌دست آمده است.

در مرجع[۹۴] با استفاده از ویژگی‌های استخراجی یک شبکه CNN و دسته‌بند LSTM همراه با مکانیزم توجه روی دادگان LRW برای تشخیص کلمات به دقت ۷۶/۲۰% و روی دادگان GRID به دقت ۹۷/۰۰% رسیده‌اند.

در مرجع[۹۵] با استفاده از یک معماری انتها به انتها شامل شبکه خودرمزگذار ۳ لایه کاملاً متصل و یک لایه خطی گلوگاه همرا با دسته‌بند Bi-LSTM، روی دادگان Ouluvs2 به دقت ۹۴/۷۰% رسیده‌اند.

در مرجع[۹۶] از شبکه maxout CNN جهت استخراج ویژگی از ویدیوهای ده عبارت کوتاه انگلیسی از پایگاه‌داده Ouluvs2 استفاده شده است. با استفاده از این ویژگی‌ها در شبکهmaxout CNN-maxout LSTM به دقت ۸۷/۶۰% دست یافته‌اند.

در مرجع[۹۷] از شبکه ResNet CNN جهت استخراج ۱۲۸ ویژگی از پایگاه‌دادهی TCD-TIMIT استفاده شدهاست. با اعمال شبکه دنباله به دنباله مبتنی بر توجه[4]، که شامل رمزگذار و رمزگشا مبتنی بر شبکه‌های بازگشتی است، دقت نسبت به حالتی که تنها از داده صوتی استفاده شده است، بهبود یافته است. در حالت بدون نویز، نرخ خطای کاراکتر از ۱۹/۱۶% به ۱۷/۷۰% کاهش یافته است و در حالتی که نسبت سیگنال به نویزdb5 بوده، نرخ خطای کاراکتر از ۴۶/۵۲% به ۳۲/۶۸% کاهش یافته است.

در مرجع[۹۸] از خودرمزگذار عمیق[5] جهت استخراج ویژگی‌های ایستا از پایگاه‌داده انگلیسی TCD-TIMIT استفاده شده است. با اعمال دسته‌بند DNN-HMM، در حالت وابسته به گوینده به دقت ۵۷/۳۶% و در حالت مستقل از گوینده به دقت ۵۳/۸۳% دست یافته‌اند.

در مرجع[۹۹] از دو نوع از تصاویر قاب‌های بهم چسبانیده[6] چسبانیده[6] حاصل از ویدیوهای ده عبارت از پایگاه‌داده Ouluvs2 استفاده شده است. نوع اول با کنارهم قراردادن دنباله تصاویر لب از یک ویدیو در یک تصویر و نوع دوم با کنار هم قراردادن تصاویر ۱۶ بخش[7] پراهمیت از تصاویر لب در تصویری دیگر ایجاد شده است. با واردکردن این تصاویر به دسته‌بندی که متشکل از ۸ شبکه از نوع quarter VGG-M است، در حالت وابسته به گوینده به دقت ۹۰/۹۰% دست‌یافته‌اند.

در جداول ۵ و ۶ خلاصه بررسی های انجام شده در تشخیص کلمات و جملات جهت مقایسه ارائه شده است.

در روش‌های سنتی لب‌خوانی کلمه و عبارات از ویژگی‌ها و دسته‌بندهای مختلف روی مجموعه دادگان مختلف استفاده شده است. برای مجموعه دادگان پرکاربرد GRIDبا استفاده از ویژگی‌های DCTو دسته‌بند HMMدر مرجع[۸۹] دقت ۵۷/۰۰% گزارش شده است. در مرجع[۶۵] با استفاده از شبکه ترکیبی CNN-LSTMبرای استخراج ویژگی و شبکه GRUدوسویه همراه با مکانیزم توجه به دقت ۹۸/۷۰% رسیده‌اند که پیشرفت بسیار قابل توجهی به نظر می‌رسد. برای دیگر مجموعه دادگان پرکاربرد مانند Ouluvs2در مرجع[۹۱] با ترکیب یک شبکه CNNو دسته‌بند LSTMبه دقت ۹۴/۱۰% دست یافته‌اند که نسبت به دقت گزارش شده در مرجع[۹۲] که

[1] Temporal Convolutional Networks
[2] Densely Connected
[3] Latent Variable Models
[4] Attention-Based seq2seq Network
[5] Deep Auto-Encoder
[6] Concatenated Frame Images
[7] Landmark

با استفاده از ترکیب ویژگی‌های DCT و LDA و دسته‌بند HMM به میزان ۷۴/۰۰% به‌دست آمده بوده، بهبودی در حدود ۲۰ درصد داشته است.

۶ مروری بر پژوهش‌های لب‌خوانی زبان فارسی

تاکنون فعالیت‌های اندکی در زمینه لب‌خوانی گفتار فارسی انجام شده است. این تحقیقات مبتنی بر دادگان خودساخته مرتبط با هر پژوهش بوده است و هنوز دادگان مرجعی برای لب‌خوانی فارسی مورد استفاده قرار نگرفته است. دامنه این پژوهش‌ها بیشتر شناسایی حروف و اعداد و معدودی از کلمات بوده است. مروری بر این پژوهش‌ها به شرح زیر است:

در پایان نامه ۱۳۸۴ در دانشگاه تربیت مدرس [۱۰۰] یک سیستم بازشناسی تصویری گفتار با شش کلمه بالا، پایین، چپ، راست، جلو و عقب ساخته شده است. در هر قاب تصویر، موقعیت نقطه وسط مرز بیرونی لب بالایی و نقطه وسط مرز درونی لب پایینی مشخص شده است. با در نظر گرفتن مرز بیرونی این دو نقطه از رشته تصاویر مربوط به ادای هر کلمه یک سیگنال زمانی ساخته شده و برای بازشناسی کلمات از روش DTW[1] استفاده می‌شود. ۴ گوینده هریک از کلمات را ۶ بار ادا کرده‌اند و مجموعه دادگانی شامل ۱۴۴ نمونه جمع‌آوری شد. دقت بازشناسی کلمات این مجموعه ۸۰/۳۰% گزارش شده است.

در مرجع[۱۰۱] اولین اقدام لب‌خوانی خودکار در زبان فارسی، جهت تشخیص شش واکه زبان فارسی انجام شده است. در این پژوهش از نمونه دادگانی شامل ۴۲ کلمه تک هجایی فارسی که توسط دو زن و شش مرد بیان شده، استفاده شده است. با استخراج ویژگی‌های عرض و ارتفاع قطعه لب مربوط به ده قاب میانی نمونه ویدیوها و ارسال آن‌ها به شبکه عصبی دولایه شامل ۱۰ ورودی و ۲۵ نورون مخفی و ۶ خروجی به عنوان دسته‌بند، دقت به میزان ۷۰/۵۰% به‌دست آمده است.

درمرجع[۱۰۲] یک مدل ۱۶ نقطه‌ای برای استخراج خطوط خارجی لب با استفاده از ASM[2] پیشنهاد شده است. با روش خوشه‌بندی فازی، یک نگاشت احتمالی از تصاویر رنگی به‌دست آمده و به کمک یک تابع هزینه، ناحیه لب با احتمال بالاتر از ناحیه غیرلب جدا می‌شود. سپس مدل لب با تغییر پارامترها بر روی این نگاشت احتمالی انطباق داده می‌شود. پارامترهای نهایی به‌عنوان بردار ویژگی هر قاب در نظر گرفته شده و چون طول دنباله قاب‌ها متفاوت است، برای یکسان سازی ابعاد بردار ویژگی، از روش درون‌یابی و برای کاهش بعد بردار ویژگی از روش‌های PCA و FLD استفاده می‌شود. برای هر نمونه، بردار ویژگی ۶×۲۰ به‌دست آمده است. مجموعه دادگان جمع‌آوری شده شامل تصاویر۲۰ گوینده (۱۶ مرد و ۴ زن) از گروه سنی ۲۰ تا ۵۰ سال که دارای شکل لب و رنگ پوست متفاوتی بوده‌اند. ۵ نفر از این گویندگان ریش و سبیل داشته‌اند. ضمناً گویندگان هیچ‌گونه آرایشی ندارند. هریک از گویندگان ۶ عدد فارسی (یک تا شش) را ۱۰ مرتبه ادا کرده‌اند. دادگان جمع‌آوری شده از لحاظ شرایط نوری حین تصویربرداری از تنوع مناسبی برخوردار می‌باشد. ویدیوها با تفکیک‌پذیری۱۲۰×۱۰۰ پیکسل ضبط شده‌اند و صورت و لب گویندگان مختلف در حین ادای کلمات چرخش و جابجایی داشته است. برای دسته‌بندی از ۶ مدل HMM متناظر با ۶ عدد فارسی استفاده شده است و به دقت ۸۹/۰۰% دست یافته‌اند.

در مرجع[۱۰۳] ابتدا تصاویری که نماینده همخوان‌ها در هجاهای زبان فارسی بوده را از پایگاه‌داده AVA[۲۴] به صورت دستی انتخاب کرده‌اند. سپس از روش تحلیل ویژه[3]جهت استخراج ویژگی و خوشه‌بندی واج‌ها استفاده شده است. بدین‌صورت که برای هر واج، یک بردار شاخص مقادیر ویژه محاسبه شده است. این بردار با محاسبات تحلیل مقادیر ویژه بر روی تصاویر نرمال شده و برگزیدن برداری که بیش‌ترین تمایز را با بردارهای شاخص واج‌های دیگر ایجاد کرده، انتخاب شده است. سپس، فاصله اقلیدسی بردارهای واج‌ها با این بردارهای شاخص محاسبه شده و به‌عنوان بردارهای ویژگی واج‌ها استفاده شده است. برای خوشه‌بندی در روش پیشنهادی، بردارهایی که فواصل اقلیدسی آن‌ها با یکدیگر از آستانه مشخصی کمتر بوده است در یک خوشه قرار گرفته‌اند. این مقدار آستانه برابر با دو سوم ماکزیمم فاصله بین بردارها در نظرگرفته شده است. با اعمال این روش، ویزم‌های زبان فارسی به هفت خوشه، بخش‌بندی شده‌اند. نتایج این خوشه‌بندی در جدول ۷ مشاهده می‌شود.

جدول ۵: خوشه‌بندی ویزم‌های فارسی[۱۰۳]

شماره خوشه	ویزم‌های خوشه	شماره خوشه	ویزم‌های خوشه
۱	b, p, m	۵	G, n, h, x, ʔ
۲	f, v	۶	r, l
۳	d, t	۷	tʃ, dʒ, s, z, ʃ, ʒ
٤	k, g, j		

[1]Dynamic Time Warping
[2]Active Shape Model
[3] Eigen Analysis

جدول ٦: خلاصه‌ای از پژوهش‌های انجام شده در لب‌خوانی کلمات زبان‌های مختلف

سال	مرجع	مدل		پایگاه داده	شرایط	دقت(درصد)
		ویژگی	دسته‌بند			
۲۰۰٦	[۷۶]	DCT-PCA سراسری	HMM	HIT Bi-CAVDB	وابسته به گوینده	۷۷/۱۰
		DCT-PCA بلوکی				۷٦/۹۰
۲۰۱۷	[۹۴]	CNN	LSTM	GRID	مستقل از گوینده	۹۷/۰۰
۲۰۱۸	[۶۵]	CNN-LSTM	Bi-GRUهمراه با مکانیزم توجه	GRID	مستقل از گوینده	۹۷/۱۰
۲۰۱٦	[۶۲]	شبکه انتها به انتها (STCNN، Bi-GRU، شبکه عصبی چندلایه و CTC loss)		GRID	ارزیابی اول	نرخ خطا۱۱/٤۰
					ارزیابی دوم	نرخ خطا٤/۸۰
۲۰۲۰	[۸۶]	شبکه انتها به انتها (پیش‌خور و LSTM)		GRID	مستقل از گوینده	۵۵/۳۰
۲۰۱۹	[۸۴]	شبکه خودرمزگذار پیچشی	LSTM	GRID	۵ نفر	۸٤/۸۰
۲۰۱٦	[۸۰]	PCA و هیستوگرام گرادیان	SVM	GRID	وابسته به گوینده	۷۱/۳۰
		انتها به انتها (پیش‌خور و LSTM)				۷۹/٦۰
۲۰۱٤	[۶۱]	ویژگی هندسی لب	جنگل تصادفی	AV-CMU	شبه وابسته به گوینده	۶۵/۶۸
				AV-UNR		۷۲/۲۸
۲۰۱۱	[۷۷]	PLF-08	HMM	PLF-08	وابسته به گوینده	۹۱/٤۱
			KNN			۹۷/۰٤
			ANN			۹۷/۴۵
			HMM		مستقل از گوینده	٦۹/٤۵
			KNN			۷۸/۹۰
			ANN			۸۲/۵۸
۲۰۱٦	[۸۱]	LBP-TOP	شبکه عصبی انتشار رو به عقب	پایگاه‌داده هندی	-	۹۷/۰۰
۲۰۱٦	[۸۲]	شبکه عصبی گلوگاه	HMM	ATR	-	۷۱/۷٦
		CNN				۵۸/۹٤
		DCT				۵۱/۵۸
		ویژگی‌های پویا و ایستا				۵٦/٤۸
۲۰۱۵	[۷۹]	افکنش عمودی و افقی	شبکه عصبی انتشار رو به عقب	پایگاه داده اندونزیایی	-	٦۷/۰۰
۲۰۰۱	[۷۵]	ضرایب دو بعدی تبدیل فوریه	HMM	دادگان اختصاصی	-	٦۰/۰۰
۲۰۱۸	[۷۸]	انتها به انتها (CNN)		TULAVD (زبان چک)	وابسته به گوینده	۹۲/۳۰
۲۰۱٦	[۲۲]	AAM و ASM	HMM-DNN	دادگان اختصاصی	وابسته به گوینده	٦٤/۰۰
۲۰۱۹	[۸۵]	انتها به انتها (شبکه یادگیری عمیق vision to phoneme)		دادگان اختصاصی	مستقل از گوینده	نرخ خطا ٤۰/۹۰
۲۰۱۷	[۹۴]	CNN	LSTM	LRW	مستقل از گوینده	۷۶/۲۰
۲۰۲۰	[۸۷]	شبکه پیچشی زمانی		LRW	مستقل از گوینده	۸۵/۳۰
				LRW1000		٤۱/٤۰
۲۰۲۱	[۸۸]	TCN و DC-TCN		LRW	مستقل از گوینده	۸۸//۳٦
				LRW1000		٤۵/٦۵
۲۰۱۹	[۸۴]	خودرمزگذار پیچشی	LSTM	LRW	۹ کلمه ساده	۸۵/٦۱
					۲۷ کلمه ابهام آمیز	۷۷/٤۲
				MIRACLE_VCI	وابسته به گوینده	۹۸/۰۰
					مستقل از گوینده	٦۳/۲۲

جدول ۷: خلاصه‌ای از پژوهش‌های انجام شده در لب‌خوانی جملات/عبارت‌های زبان‌های مختلف

سال	مرجع	مدل		پایگاه داده	شرایط	دقت(درصد)
		ویژگی	دسته‌بند			
۲۰۰۹	[۶۷]	DCT	HMM	GRID	مستقل از گوینده	۵۷/۰۰
۲۰۱۸	[۵۶]	شبکه خودرمزگذار	Bi-LSTM	AV Digits	مستقل از گوینده	۶۸/۰۰
۲۰۱۶	[۹۱]	SyncNet	LSTM	Ouluvs2	مستقل از گوینده	۹۴/۱۰
۲۰۱۶	[۹۲]	DCT+PCA	HMM	Ouluvs2	مستقل از گوینده	۶۳/۰۰
		DCT+LDA				۷۴/۰۰
		مقادیر خام پیکسل	LVM			۷۳/۰۰
		CNN	LSTM			۸۱/۱۰
۲۰۱۶	[۹۳]	ترکیب قاب‌های بهم چسبانیده	NIN	Ouluvs2	مستقل از گوینده	۸۱/۱۰
			AlexNet			۸۲/۸۰
			GoogleNet			۸۵/۶۰
۲۰۱۷	[۹۵]	شبکه انتها به انتها (شبکه خودرمزگذار و Bi-LSTM)		Ouluvs2	مستقل از گوینده	۹۴/۷۰
۲۰۱۸	[۹۶]	شبکه maxout CNN	شبکه maxout CNN-maxout LSTM	Ouluvs2	مستقل از گوینده	۸۷/۶۰
۲۰۱۹	[۹۹]	قاب‌های بهم چسبانیده	quarter VGG-M	Ouluvs2	مستقل از گوینده	۹۰/۹۰
۲۰۱۱	[۹۰]	DCT	HMM	ViDTIMIT	مستقل از گوینده	۹۶/۰۰
۲۰۱۸	[۹۷]	ResNet CNN	شبکه دنباله به دنباله مبتنی بر توجه	TCD-TIMIT	وابسته به گوینده	در حالت بدون نویز، نرخ خطای کاراکتر ۱۷/۷۰
۲۰۱۸	[۹۸]	شبکه خودرمزگذار عمیق	DNN-HMM	TCD-TIMIT	وابسته به گوینده	۵۷/۳۶
					مستقل از گوینده	۵۳/۸۳

در مرجع[۲۳] از اندازه‌های ناحیه لب، ناحیه دندان، ناحیه زبان و نسبت ارتفاع به عرض لب به‌عنوان ویژگی‌های استخراج شده از پایگاه‌داده استفاده شده است. این پایگاه داده شامل ۳۱ کلمه فارسی بوده و هر کلمه حداقل ده مرتبه در شرایط مختلف نور و فاصله توسط ۳ مرد و ۲ زن بیان شده‌اند. ویدیوها با نرخ۳۰ قاب در ثانیه و تفکیک‌پذیری ۱۰۸۰×۱۹۲۰ پیکسل ضبط شده‌اند. در ابتدا قاب‌هایی که مقادیر ویژگی‌های آن‌ها نسبت به قاب قبلی بدون تغییر بوده، حذف شده‌اند. سپس با کنار هم قرار دادن ویژگی‌های قاب‌های باقیمانده، بردار ویژگی مربوط به هر ویدیو تشکیل می‌شود. با اعمال تبدیل فوریه بر روی هر ویژگی، ضرایب ۹ گانه‌ای به‌دست آمده‌است. این ضرایب به‌دست آمده برای هر یک از ۴ ویژگی در کنار هم قرار داده شده و یک بردار ویژگی ۳۶ بعدی برای ویدیوی هر کلمه ساخته شده است.سپس این بردار ویژگی به یک شبکه عصبی چهار لایه داده شده است. متوسط صحت اعلام شده برای ۳۱ کلمه ۸۶/۸۰% بوده است.

در مرجع[۱۰۴] یک سیستم لب‌خوانی غیر برخط[1] بر روی تلفن همراه برای تشخیص ۶ کلمه فارسی پیاده‌سازی شده است. این کلمات توسط ۳ گوینده بیان شده و هر کلمه را ۳۵ مرتبه تکرار کرده‌اند. در این سیستم، ناحیه لب را با استفاده از روش ویولا جونز تشخیص داده و بعد ازجداسازی ناحیه موردنظر، مختصات چهار نقطه اصلی محدوده لب (بالا، پایین، چپ و راست) به‌عنوان ویژگی استخراج شده است. ویدیوها با دوربین ۲ مگاپیکسلی لپتاپ و با تفکیک‌پذیری ۴۸۰×۶۴۰ پیکسل در یک محیط با شرایط نوری ثابت و با قرار گرفتن دوربین روبه‌روی فرد ضبط شده است. با اعمال دسته‌بند SVM روی این ویژگی‌ها به دقت ۷۴/۰۰% دست یافته‌اند.

در مرجع[۱۰۵] از شبکه‌های عصبی مصنوعی با ۴ لایه برای تشخیص ۱۲ واج زبان فارسی که نمود واضحی در شکل لب‌ها داشته‌اند، استفاده شده است. با استفاده از مجموعه دادگان شامل ۲ زن و ۸ مرد به عنوان گوینده که هرکدام یک تکرار از این واج‌ها را بیان کرده‌اند، میانگین فاصله بین نقاط بالا و پایین لب را به عنوان مشخصه هر واج در نظر گرفته و با استفاده از شبکه عصبی انتشار رو به عقب به میانگین دقت ۹۴/۶۰% در حالت غیروابسته به گوینده دست یافته‌اند.

[1] Offline

جدول ۸: خلاصه‌ای از پژوهش‌های انجام شده در لب‌خوانی زبان فارسی

سال	مرجع	مدل		پایگاه داده	محتوای تشخیص داده شده	دقت(درصد)
		ویژگی	دسته‌بند			
۲۰۰۵	[۱۰۰]	موقعیت نقطه وسط مرز بیرونی لب بالایی و نقطه وسط مرز درونی لب پایینی	DTW	دادگان اختصاصی	کلمه	۸۰/۳۰
۲۰۰۶	[۱۰۱]	ویژگی‌های عرض و ارتفاع قطعه لب	شبکه عصبی دولایه	دادگان اختصاصی	واکه	۷۰/۵۰
۲۰۰۸	[۱۰۲]	ASM	HMM	دادگان اختصاصی	عدد	۸۹/۰۰
۲۰۱۳	[۱۰۳]	روش تحلیل ویژه	روش پیشنهادی	AVA	حروف	۷ خوشه ویزم‌های فارسی
۲۰۱۷	[۲۳]	اندازه‌های نواحی لب، دندان و زبان و نسبت ارتفاع به عرض لب	شبکه عصبی	دادگان اختصاصی	کلمه	۸۶/۸۰
۲۰۱۹	[۱۰۴]	مختصات چهار نقطه اصلی محدوده لب	SVM	دادگان اختصاصی	کلمه	۷۴/۰۰
۲۰۱۹	[۱۰۵]	میانگین فاصله بین نقاط بالا و پایین لب	شبکه عصبی	دادگان اختصاصی	واج	۹۴/۶۰
۲۰۲۰	[۱۰۶]	ارتفاع، عرض، تعداد پیکسل‌های لب و ویژگی ۵۹ بعدی LBP	HMM	دادگان اختصاصی	هجا	۵ روش خوشه‌بندی هجا

در مرجع[۱۰۶] با توجه به هم آوا بودن برخی از هجاهای CV، هجاهای هم آوا در یک خوشه قرار گرفته‌اند. این جداول نگاشت هجا به ویسیلاب شامل پنج روش خوشه‌بندی برای هجاهای CV است. به منظور بررسی روش‌های پیشنهادی، ابتدا دادگان مناسبی فراهم شده است. از تعداد ۴۰ نفر زن و مرد خواسته شده که هریک از ۱۳۸ هجای CV زبان فارسی را ۴ مرتبه بیان کنند و تصاویر آن‌ها با نرخ ۵۰ قاب در ثانیه و تفکیک پذیری ۱۰۸۰ ×۱۹۲۰ پیکسل ضبط شده است. پس از جداسازی ناحیه لب، سه ویژگی ارتفاع، عرض، تعداد پیکسل‌های لب و ویژگی ۵۹ بعدی LBP[۱۰۷] از ناحیه لب استخراج شده است. سپس، بردار ویژگی‌های ۶۲ بعدی به‌دست آمده برای هر دنباله قاب با روش zscore نرمال شده و تفاضل بین ویژگی ۶۲ بعدی هرقاب با قاب قبلی، محاسبه شده و به‌عنوان یک ویژگی ۶۲ بعدی به بردار ویژگی‌های آن قاب اضافه می‌شود. نتیجه حاصل یک ماتریس ویژگی ۱۲۴×m برای هر دنباله قاب است که به‌عنوان ویژگی‌های پایه در تحلیل‌های این پژوهش استفاده شده است. سپس بردار ویژگی‌های نمونه‌ها با برچسب‌زنی مجدد براساس خوشه‌ها، از HMM با ۹ حالت و دو مدل مخلوط گاوسی جهت آموزش و آزمون استفاده شده است.

خلاصه‌ای از پژوهش‌های لب‌خوانی زبان فارسی در جدول ۸ آمده است. پژوهش‌های انجام شده در زبان فارسی اندک هستند و به دلیل آنکه از مجموع دادگان یکسانی استفاده نکرده‌اند، قابل مقایسه نیستند.

۷ جمع‌بندی و نتیجه‌گیری

در این مقاله به بررسی روش‌های به‌کار رفته در بازشناسی دیداری گفتار و مجموعه دادگان مورد استفاده در مراجع مورد بررسی پرداخته شد. همچنین کارهای انجام شده در لب‌خوانی زبان فارسی به صورت ویژه مورد بررسی قرار گرفت. بررسی‌ها در مورد حروف، کلمات، اعداد و جملات انجام شد تا به درک تنوع کارهای انجام شده و مطالعه مفیدتر برای خوانندگان و پژوهشگران گرامی کمک کند. مرور پژوهش‌های انجام شده نشان داد که کارهای انجام شده در سیر زمانی از حروف و اعداد به شناسایی کلمات و جملات رسیده است. روش‌های مورد استفاده در استخراج ویژگی‌ها از ویژگی‌های هندسی لب، ویژگی‌های استخراج شده از پیکسل‌های تصویر با الگوریتم‌هایی نظیر LDA، PCA، DCTو DWTو ویژگی‌های مبتنی بر مدل‌های آماری نظیر AAMو ASMبه سمت استخراج ویژگی‌های مبتنی بر شبکه‌های یادگیری عمیق نظیر خودرمزگذار و CNNحرکت کرده‌اند. همچنین دسته‌بندها از دسته‌بندهای کلاسیک نظیر SVM، جنگل تصادفی و HMMبه دسته‌بندهای مبتنی بر RNNنظیر Bi-LSTM، LSTMو معماری‌های جدیدتر شبکه‌های CNNتغییر یافته‌اند. در روش‌های سنتی، دسته‌بند HMMنسبت به سایر دسته‌بندها نتایج بهتری داشته است که با توجه به ویژگی این دسته‌بند در مدل کردن توالی گام‌های زمانی قابل انتظار است. در شناسایی کلمات و عبارات، معماری مبتنی بر یادگیری عمیق، پیشرفت قابل توجهی نسبت به روش‌های سنتی نشان داده‌اند که به نظر می‌رسد آینده پژوهش‌های لب‌خوانی را تحت تأثیر خود قرار دهند. هر چند روش‌های یادگیری عمیق بهبود قابل ملاحظه‌ای در شناسایی کلمات و عبارات داشته‌اند، اما استفاده از HMMدر شناسایی حروف و اعداد، همچنان نتایج خوبی داشته است. معماری یادگیری عمیق مورد استفاده در پژوهش‌های اخیر به سمت استفاده بیش‌تر از شبکه‌های LSTMیک سویه و دو سویه همراه با مکانیزم توجه و در ترکیب با سایر معماری‌های شبکه‌های عمیق نظیر CNNسوق یافته است. با پیشرفت‌های حاصل از

یادگیری عمیق، در کارهای اخیر تمایل به بازشناسی دیداری گفتار پیوسته بیش‌تر شده است و این زمینه همچنان به عنوان یک مسئله باز، طرف توجه تحقیقات در آینده خواهد بود. در زمینه لب‌خوانی زبان فارسی، پژوهش‌های بسیار اندکی انجام شده است که پیشرفت در این زمینه توجه بیش‌تر پژوهشگران را طلب می‌کند.

مراجع

[1] A. Bastanfard, M. Aghaahmadi, M. Fazel, M. Moghadam, and others, "Persian viseme classification for developing visual speech training application," in *Pacific-Rim Conference on Multimedia*, 2009, pp. 1080–1085.

[2] Y. Pei and H. Zha, "Stylized synthesis of facial speech motions," *Computer Animation and Virtual Worlds*, vol. 18, no. 4–5, pp. 517–526, 2007.

[3] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition (Speech Reading)," PhD Thesis, University of Illinois at Urbana-Champaign, 1984.

[4] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71, 1989.

[5] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 337–351, 1996.

[6] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing*, vol. 78, pp. 53–72, 2018.

[7] 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, pp. 198–213, 2002.

[8] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Second international conference on audio and video-based biometric person authentication*, 1999, vol. 964, pp. 965–966.

[9] C. Neti *et al.*, "Audio visual speech recognition," IDIAP, 2000.

[10] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[11] H. L. Bear and R. Harvey, "Decoding visemes: Improving machine lip-reading," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2016, pp. 2009–2013.

[12] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6447–6456.

[13] M. Kawamura, N. Kakita, T. Osaki, K. Sugahara, and R. Konishi, "On the hardware realization of lip reading system," in *SICE 2003 Annual Conference* (*IEEE Cat. No. 03TH8734*), 2003, vol. 3, pp. 2452–2457.

[14] N. Akhter and A. Chakrabarty, "A Survey-based Study on Lip Segmentation Techniques for Lip Reading Applications," 2016.

[15] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, "A Survey of Research on Lipreading Technology," *IEEE Access*, 2020.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, 2001, vol. 1, p. I–I.

[17] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[18] K. Paleček, "Extraction of features for lip-reading using autoencoders," in *International Conference on Speech and Computer*, 2014, pp. 209–216.

[19] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] S. Pujari, S. Sneha, R. Vinusha, P. Bhuvaneshwari, and C. Yashaswini, "A Survey on Deep Learning based Lip-Reading Techniques," in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks* (*ICICV*), 2021, pp. 1286–1293.

[22] I. Almajai, S. Cox, R. Harvey, and Y. Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2016, pp. 2722–2726.

[23] K. Mirzaei Talarposhti and M. Khaki Jamei, "An Efficient Model for Lip-reading in Persian Language Based on Visual Word and Fast Furrier Transform Combined with Neural Network," *Journal of Advances in Computer Research*, vol. 8, no. 2, pp. 103–124, 2017.

[24] A. Bastanfard, A. A. Kelishami, M. Fazel, and M. Aghaahmadi, "A comprehensive audio-visual corpus for teaching sound persian phoneme articulation," in *2009 IEEE International Conference on Systems, Man and Cybernetics*, 2009, pp. 169–174.

[25] G. and Movallali, "'Sara Lipreading Test' Development, Standardization and Evaluation in a Group with Acquired Hearing-Impairment," *Archives of Rehabilitation*, vol. 1, no. 3, 2001,

[26] A. Bastanfard, M. Fazel, A. A. Kelishami, and M. Aghaahmadi, "The Persian linguistic based audio-visual data corpus, AVA II, considering coarticulation," in *International Conference on Multimedia Modeling*, 2010, pp. 284–294.

[27] Z. Naraghi and M. Jamzad, "SFAVD: Sharif Farsi audio visual database," in *The 5th Conference on Information and Knowledge Technology*, 2013, pp. 417–421.

[28] M. Bijankhan, J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi, "Lessons from building a Persian written corpus: Peykare," *Language resources and evaluation*, vol. 45, no. 2, pp. 143–164, 2011.

[29] M. Bijankhan, J. Sheikhzadegan, and M. R. Roohani, "FARSDAT-The speech database of Farsi spoken language," 1994.

[30] M. Hedayatipour, Y. Shekofteh, and M. Ebrahimi Moghaddam, "PAVID-CVs: Persian Audio-Visual Database of CV syllables," *The 29th Iranian Conference on Electrical Engineering*, 2021.

[31] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[32] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database (release 1.00)," in *International Conference on Audio-and Video-Based Biometric Person Authentication*, 1997, pp. 403–409.

[33] F. J. Huang and T. Chen, "Real-time lip-synch face animation driven by human voice," in *1998 IEEE Second Workshop on Multimedia Signal Processing (Cat. No. 98EX175)*, 1998, pp. 352–357.

[34] J. C. Wojdel, P. Wiggers, and L. J. Rothkrantz, "An audio-visual corpus for multimodal speech recognition in dutch language," 2002.

[35] C. Sanderson, "The vidtimit database," IDIAP, 2002.

[36] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *2002 IEEE International conference on acoustics, speech, and signal processing*, 2002, vol. 2, p. II–2017.

[37] B. Lee *et al.*, "AVICAR: Audio-visual speech corpus in a car environment," 2004.

[38] A. Ortega *et al.*, "AV@ CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition," 2004.

[39] H. Minghui, "Bimodal database and its material segmentation for lip-reading recognition on sentence," *Computer Engineering and Applications*, vol. 3, 2005.

[40] P. Císař, M. Železnỳ, Z. Krňoul, J. Kanis, J. Zelinka, and L. Müller, "Design and recording of Czech speech corpus for audio-visual continuous speech recognition," 2005.

[41] S. J. Cox, R. W. Harvey, Y. Lan, J. L. Newman, and B.-J. Theobald, "The challenge of multispeaker lip-reading.," in *AVSP*, 2008, pp. 179–184.

[42] D. Petrovska-Delacrétaz *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 *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, 2008, pp. 1–7.

[43] J. Trojanová, M. Hrúz, P. Campr, and M. Železnỳ, "Design and recording of czech audio-visual database with impaired conditions for continuous speech recognition," 2008.

[44] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[45] S. Tamura *et al.*, "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," presented at the International Conference on Auditory-Visual Speech Processing, 2010.

[46] 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 *International Conference on Text, Speech and Dialogue*, 2010, pp. 259–266.

[47] V. Estellers and J.-P. Thiran, "Multipose audio-visual speech recognition," in *2011 19th European Signal Processing Conference*, 2011, pp. 1065–1069.

[48] Y. Benezeth, G. Bachman, G. Le-Jan, N. Souviraà-Labastie, and F. Bimbot, "BL-Database: A French audiovisual database for speech driven lip animation systems," PhD Thesis, INRIA, 2011.

[49] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "A new visual speech recognition approach for RGB-D cameras," in *International conference image analysis and recognition*, 2014, pp. 21–28.

[50] M. Wagner *et al.*, "The big australian speech corpus (the big asc)," in *SST 2010, Thirteenth Australasian International Conference on Speech Science and Technology*, 2011, pp. 166–170.

[51] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *2015 11th IEEE*

International Conference and Workshops on Automatic Face and Gesture Recognition (*FG*), 2015, vol. 1, pp. 1–5.

[52] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[53] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016, pp. 87–103.

[54] V. Verkhodanova, A. Ronzhin, I. Kipyatkova, D. Ivanko, A. Karpov, and M. Železnỳ, "HAVRUS corpus: high-speed recordings of audio-visual Russian speech," in *International Conference on Speech and Computer*, 2016, pp. 338–345.

[55] 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 *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG 2017*), 2017, pp. 208–215.

[56] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2018, pp. 6219–6223.

[57] L. A. Elrefaei, T. Q. Alhassan, and S. S. Omar, "An Arabic visual dataset for visual speech recognition," *Procedia Computer Science*, vol. 163, pp. 400–409, 2019.

[58] S. Cox, R. Harvey, Y. Lan, J. Newman, and B. Theobald, "The Challenge of Multispeaker Lip-Reading," in *International Conference on Auditory-Visual Speech Processing*, 2008, pp. 179–184.

[59] Y. Pei, T.-K. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 129–136.

[60] D. Wu and Q. Ruan, "Lip reading based on cascade feature extraction and HMM," in *2014 12th International Conference on Signal Processing* (*ICSP*), 2014, pp. 1306–1310.

[61] L. D. Terissi, M. Parodi, and J. C. Gómez, "Lip reading using wavelet-based features and random forests classification," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 791–796.

[62] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: Sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, vol. 2, no. 4, 2016.

[63] D. Hu, X. Li, and others, "Temporal multimodal learning in audiovisual speech recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3574–3582.

[64] B.-S. Lin, Y.-H. Yao, C.-F. Liu, C.-F. Lien, and B.-S. Lin, "Development of novel lip-reading recognition algorithm," *IEEE Access*, vol. 5, pp. 794–801, 2017.

[65] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-end lipreading with cascaded attention-CTC," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG 2018*), 2018, pp. 548–555.

[۶۶] لسانی, ف. سادات, ف. قزوینی, فرانک ,دیانت, "لب‌خوانی: روش جدید احراز هویت در برنامه‌های کاربردی گوشی‌های تلفن همراه اندروید," ۲۰۱۷.

[67] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, 2009.

[68] N. Puviarasan and S. Palanivel, "Lip reading of hearing impaired persons using HMM," *Expert Systems with Applications*, vol. 38, no. 4, pp. 4477–4481, 2011.

[69] X. Liu and Y.-M. Cheung, "An exemplar-based hidden Markov model with discriminative visual features for lipreading," in *2014 Tenth International Conference on Computational Intelligence and Security*, 2014, pp. 90–93.

[70] S. S. Morade and S. Patnaik, "A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition," *optik*, vol. 125, no. 18, pp. 5181–5186, 2014.

[71] C. Sui, M. Bennamoun, and R. Togneri, "Listening with your eyes: Towards a practical visual speech recognition system using deep boltzmann machines," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 154–162.

[72] M. H. Rahmani and F. Almasganj, "Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features," in *2017 3rd International Conference on Pattern Recognition and Image Analysis* (*IPRIA*), 2017, pp. 195–199.

[73] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMs," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2017, pp. 2592–2596.

[74] F. Vakhshiteh, F. Almasganj, and A. Nickabadi, "Lip-reading via deep neural networks using hybrid visual features," *Image Analysis & Stereology*, vol. 37, no. 2, pp. 159–171, 2018.

[75] K. Yu, X. Jiang, and H. Bunke, "Sentence lipreading using hidden Markov model with integrated grammar," *International journal of pattern recognition*

and artificial intelligence, vol. 15, no. 01, pp. 161–176, 2001.

[76] X. Hong, H. Yao, Y. Wan, and R. Chen, "A PCA based visual DCT feature extraction method for lip-reading," in *2006 International Conference on Intelligent Information Hiding and Multimedia*, 2006, pp. 321–326.

[77] J. Shin, J. Lee, and D. Kim, "Real-time lip reading system for isolated Korean word recognition," *Pattern Recognition*, vol. 44, no. 3, pp. 559–571, 2011.

[78] K. Paleček, "Experimenting with lipreading for large vocabulary continuous speech recognition," *Journal on Multimodal User Interfaces*, vol. 12, no. 4, pp. 309–318, 2018.

[79] F. Arifin, A. Nasuha, and H. D. Hermawan, "Lip reading based on background subtraction and image projection," in *2015 International Conference on Information Technology Systems and Innovation* (*ICITSI*), 2015, pp. 1–3.

[80] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2016, pp. 6115–6119.

[81] N. Rathee, "Investigating back propagation neural network for lip reading," in *2016 International Conference on Computing, Communication and Automation* (*ICCCA*), 2016, pp. 373–376.

[82] Y. Li, Y. Takashima, T. Takiguchi, and Y. Ariki, "Lip reading using a dynamic feature of lip images and convolutional neural networks," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science* (*ICIS*), 2016, pp. 1–6.

[83] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[84] D. Parekh, A. Gupta, S. Chhatpar, A. Yash, and M. Kulkarni, "Lip reading using convolutional auto encoders as feature extractor," in *2019 IEEE 5th International Conference for Convergence in Technology* (*I2CT*), 2019, pp. 1–6.

[85] B. Shillingford *et al.*, "Large-Scale Visual Speech Recognition" *Proc. Interspeech 2019*, pp. 4135–4139, 2019.

[86] M. Riva, M. Wand, and J. Schmidhuber, "Motion Dynamics Improve Speaker-Independent Lipreading," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 4407–4411.

[87] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 6319–6323.

[88] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2857–2866.

[89] D. Kolossa, S. Zeiler, A. Vorwerk, and R. Orglmeister, "Audiovisual speech recognition with missing or unreliable data.," in *AVSP*, 2009, pp. 117–122.

[90] K. Y. Min and L. H. Zuo, "A lip reading method based on 3-D DCT and 3-D HMM," in *Proceedings of 2011 International Conference on Electronics and Optoelectronics*, 2011, vol. 1, pp. V1-115.

[91] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*, 2016, pp. 251–263.

[92] D. Lee, J. Lee, and K.-E. Kim, "Multi-view automatic lip-reading using neural network," in *Asian conference on computer vision*, 2016, pp. 290–302.

[93] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikäinen, "Concatenated frame image based cnn for visual speech recognition," in *Asian Conference on Computer Vision*, 2016, pp. 277–289.

[94] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2017, pp. 3444–3453.

[95] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-End Multi-View Lipreading," 2017.

[96] I. Fung and B. Mak, "End-to-end low-resource lip-reading with maxout CNN and LSTM," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2018, pp. 2511–2515.

[97] G. Sterpu, C. Saam, and N. Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 111–115.

[98] K. Thangthai and R. W. Harvey, "Building Large-vocabulary Speaker-independent Lipreading Systems.," in *INTERSPEECH*, 2018, pp. 2648–2652.

[99] D.-W. Jang, H.-I. Kim, C. Je, R.-H. Park, and H.-M. Park, "Lip Reading Using Committee Networks With Two Different Types of Concatenated Frame Images," *IEEE Access*, vol. 7, pp. 90125–90131, 2019.

[۱۰۰] س. صمدیان،ع. دفتریان, "لب خوانی مجموعه محدودی از کلمات فارسی," کارشناسی ارشد, وزارت علوم، تحقیقات و فناوری - دانشگاه تربیت مدرس, ۱۳۸٤ .

[101] V. S. Sadeghi and K. Yaghmaie, "Vowel recognition using neural networks," *IJCSNS International Journal*

of Computer Science and Network Security, vol. 6, no. 12, pp. 154–158, 2006.

[102] R. Shalbaf, M. Vafadoost, A. Shalbaf, and R. Kahnamouei, "Recognition of Six Digits from Lip Movement Using Color Image," in *4th Kuala Lumpur International Conference on Biomedical Engineering 2008*, 2008, pp. 221–225.

[103] M. Aghaahmadi, M. M. Dehshibi, A. Bastanfard, and M. Fazlali, "Clustering Persian viseme using phoneme subspace for developing visual speech application," *Multimedia tools and applications*, vol. 65, no. 3, pp. 521–541, 2013.

[104] F. S. Lesani, F. Fotouhi Ghazvini, and R. Dianat, "Developing an Offline Persian Automatic Lip Reader as a New Human–Mobile Interaction Method in Android Smart Phones," *Journal of Circuits, Systems and Computers*, vol. 28, no. 08, p. 1950132, 2019.

[105] M. Barkhan, F. Alizadeh, and V. Maihami, "Designing and implementing a system for Automatic recognition of Persian letters by Lip-reading using image processing methods," *Journal of Advances in Computer Engineering and Technology*, vol. 5, no. 2, pp. 71–80, 2019.

[۱۰۶] م. هدایتی‌پور, ی. شکفته ,م. ابراهیمی‌مقدم, "خوشه‌بندی ویسیلاب‌های دوآوایی زبان فارسی در کاربرد لب‌خوانی ",یازدهمین کنفرانس فناوری اطلاعات و دانش,(IKT 2020)

[107] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

مهسا هدایتی پور دارای مدرک کارشناسی ارشد در رشته مهندسی کامپیوتر (گرایش هوش مصنوعی و رباتیکز) از دانشگاه شهید بهشتی تهران می‌باشد. زمینه‌های پژوهشی مورد علاقه ایشان پردازش تصویر، بینایی ماشین، شناسایی الگو و یادگیری ماشین است.

یاسر شکفته در سال‌های ۱۳۸۷ و ۱۳۹۲ مدارک کارشناسی ارشد و دکترای خود را در رشته مهندسی پزشکی(گرایش بیوالکتریک) از دانشگاه صنعتی امیرکبیر اخذ کرد. ایشان از سال ۱۳۹۵ تاکنون عضو هیئت علمی دانشکده مهندسی و علوم کامپیوتر دانشگاه شهید بهشتی است. زمینه‌های پژوهشی مورد علاقه ایشان پردازش گفتار و زبان طبیعی، شناسایی الگو و سامانه‌های پویا و آشوبی است.

محسن ابراهیمی مقدم مدارک کارشناسی ارشد و دکترای خود را در رشته مهندسی نرم‌افزار از دانشگاه صنعتی شریف تهران اخذ کرد. در حال حاضر، استاد دانشکده مهندسی و علوم کامپیوتر دانشگاه شهید بهشتی تهران است. زمینه‌های تحقیقاتی مورد علاقه ایشان پردازش تصویر، بینایی ماشین، ساختمان داده‌ها و طراحی الگوریتم است.