Electronic and Cyber Defense
Vol. 14, No.2, 2026 (Serial No. 50) .pp.11-24
ISSN: 2322-4347 , E-ISSN: 2980-8979

A Persian Lipreading Model Based on the Combination of CNN and Transformer Networks

Zhina Shahidi Zandi [1], Vali Derhami [2*], AliMohammad Latif [3], MohammadTaghi Sadeghi [4]

[1] PhD Student, Yazd University, Yazd, Iran.
[2] Professor. Yazd University, Yazd, Iran ([*]Correspondence: vderhami@yazd.ac.ir)
[3] Professor. Yazd University, Yazd, Iran.
[4] Associate Professor, Yazd University, Yazd, Iran.

ARTICLE INFO

Article history:

Article Type: Research paper

Received: 20 February 2026
Revised: 12 May 2026
Accepted: 06 June 2026
Available online: 24 June 2026

Keywords:
Lip reading
Visual speech recognition
Convolutional neural networks (CNNs)
Feature extraction
Transformer

ABSTRACT

Speech is the primary medium of human communication and involves the interpretation of both auditory and visual information. Lipreading, or visual speech recognition, aims to infer spoken content by analyzing lip movements. With the growing importance of speech confidentiality in security-sensitive environments and the increasing vulnerability to acoustic eavesdropping, lipreading has attracted significant research attention in recent years. This paper presents a novel visual speech recognition framework that integrates convolutional neural networks (CNNs) with Transformer architectures to recognize speech from lip movement images of multiple speakers. In the proposed framework, CNNs extract discriminative spatial features, which are subsequently processed by a Transformer-based model to generate the corresponding textual output. The proposed method is evaluated on the Persian PAVID dataset. A key contribution of this work is viseme-level lip movement recognition, which improves generalization capability and enables continuous speech reconstruction. Experimental results demonstrate that the proposed approach reduces training time while enhancing overall recognition performance, achieving an accuracy of 89% in phoneme recognition on the test set and outperforming existing state-of-the-art methods.

Cite this article: Shahidi Zandi, Zh., Derhami, V., Latif, A., Sadeghi, M. A Persian Lipreading Model Based on the Combination of CNN and Transformer Networks. *Electronic and Cyber Defense*. 2026. 14(2):11-24. **DOI**: https://dor.isc.ac/dor/20.1001.1.23224347.1405.14.2.1.5

© Author(s) retain the copyright and full publishing rights

Publisher: Imam Hossein University.

«پدافند الکترونیک و سایبری»
سال چهاردهم، شماره ۲، تابستان ۱۴۰۵؛ ص ۲۴-۱۱
شاپا الکترونیکی: ۸۹۷۹-۲۹۸۰ شاپا چاپی: ۴۳۴۷-۲۳۲۲

ارائه‌ی یک مدل لب‌خوانی فارسی مبتنی بر ترکیب شبکه‌های پیچشی و ترنسفورمر

ژینا شهیدی زندی[۱]، ولی درهمی[۲*]، علی‌محمد لطیف[۳]، محمدتقی صادقی[۴]

[۱] دانشجوی دکتری، دانشگاه یزد، یزد، ایران.
[۲] استاد، دانشگاه یزد، یزد، ایران (نویسنده مسئول: ahmadifar@guilan.ac.ir)
[۳] استاد، دانشگاه یزد، یزد، ایران.
[۴] دانشیار، دانشگاه یزد، یزد، ایران.

چکیده

گفتار رایج‌ترین شکل ارتباط بین انسان‌ها است و شامل درک اطلاعات صوتی و تصویری می‌شود. لب‌خوانی فرایندی است که برای تشخیص گفتار از روی حرکات لب گوینده استفاده می‌شود. ازآنجاکه محرمانگی گفتار در محیط‌های امنیتی به دلایل مسائل شنود بسیار اهمیت دارد؛ بنابراین لب‌خوانی به یکی از مهم‌ترین مسائل امروزه تبدیل شده است. در این مقاله روشی نوین مبتنی بر ترکیب شبکه‌های پیچشی و شبکه‌ی ترنسفورمر در تشخیص گفتار از تصویر حرکات لب افراد مختلف ارائه می‌شود. در این ساختار ویژگی‌ها با استفاده از شبکه‌های پیچشی استخراج می‌گردند. سپس این ویژگی‌ها به‌عنوان ورودی به ساختار ترنسفورمر پیشنهادی برای تولید متن عبارت گفتاری در خروجی منتقل می‌شوند. الگوریتم پیشنهادی بر روی مجموعه‌دادگان پاوید در زبان فارسی ارزیابی گردیده است. نوآوری روش پیشنهادی بازشناسی حرکات لب در سطح واحدهای واج یا ویزم است. این رویکرد امکان تعمیم‌پذیری بهتر و بازسازی گفتار پیوسته را فراهم می‌سازد. علاوه بر این آزمایش‌ها نشان می‌دهد رویکرد پیشنهادی زمان آموزش را کاهش می‌دهد و عملکرد سامانه را ارتقا می‌بخشد. دقت روش پیشنهادی در تشخیص واج‌های داده‌های آزمون، ۸۹ درصد است که نسبت به مدل‌های دیگر دارای دقت بالاتری است.

مشخصات مقاله

تاریخچه مقاله:
نوع مقاله: علمی پژوهشی

دریافت: ۱۴۰۴/۱۲/۰۱
بازنگری: ۱۴۰۵/۰۲/۲۲
پذیرش: ۱۴۰۵/۰۳/۱۶
ارائه آنلاین: ۱۴۰۵/۰۴/۰۳

کلید واژه ها:
لب‌خوانی
بازشناسی دیداری گفتار
شبکه‌های عصبی پیچشی
استخراج ویژگی
ترنسفورمر

استناد: شهیدی زندی، ژینا، درهمی، ولی، لطیف، علی‌محمد، صادقی، محمدتقی. ارائه‌ی یک مدل لب‌خوانی فارسی مبتنی بر ترکیب شبکه‌های پیچشی و ترنسفورمر. (۱۴۰۵). *پدافند الکترونیکی و سایبری*، ۱۴(۲)، ۲۴-۱۱.

https://dor.isc.ac/dor/20.1001.1.23224347.1405.14.2.1.5

© نویسندگان حق نشر و حقوق کامل انتشار را برای خود محفوظ می‌دارند.

ناشر: دانشگاه جامع امام حسین(ع)

۱. مقدمه

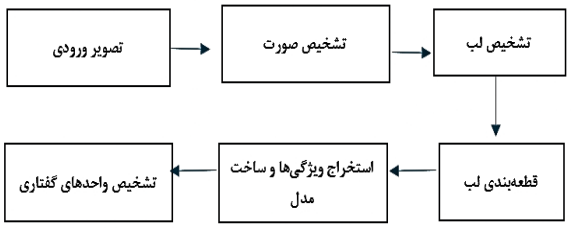
گفتار رایج‌ترین شکل ارتباط بین انسان‌ها است و شامل درک اطلاعات صوتی و تصویری می‌شود [۱]. کوچک‌ترین واحد دیداری گفتار در تصویر ویزم[1] است. ویزم معادل هر واج در فضای تصویر است. در واقع به شکل ظاهری لب در هنگام بیان یک واج، ویزم گفته می‌شود [۲]. واج کوچک‌ترین واحد آوایی زبان است. واج برای انتقال معنی در یک‌زبان استفاده می‌شود. واج‌ها با یکدیگر ترکیب می‌شوند و هجاها را می‌سازند. هجاها واحد سازنده کلمات هستند. متناظر هجاها در فضای تصویر ویسیلاب[2] نامیده می‌شود [۳].

لب‌خوانی روشی برای تشخیص گفتار از روی حرکات لب است [۴]. لب‌خوانی در محیط‌های امنیتی و محرمانه، مانند محیط‌های نظامی و دولتی، به دلایل مسائل شنود اهمیت دارد. در این فرایند، پردازش اطلاعات تصویری ناحیه لب برای شناسایی گفتار انجام می‌شود و هدف اصلی پژوهش، تشخیص گفتار در چنین محیط‌هایی از طریق لب‌خوانی است.

معروف‌ترین آزمایشی که در این حوزه انجام شده است همان آزمایش مک گارد و مک‌دونالد است. در این آزمایش نشان‌داده‌شده است که حتی افراد با شنوایی طبیعی به لب‌خوانی نیاز دارند و عدم توجه به آن افراد را در تشخیص گفتار می‌تواند دچار خطا کند [۵].

لب‌خوانی در حوزه سلامت برای کمک به افراد مبتلا به سرطان حنجره و معلولین گفتاری-حرکتی کاربرد دارد و به آن‌ها امکان ارتباط با اطرافیان را می‌دهد .علاوه بر این می‌توان از لب‌خوانی به‌جای استفاده از صفحه‌کلید بـرای ورود اطلاعات به رایانه یا برنامه‌های کاربردی گوشی همراه نیز استفاده شود [۶].

باتوجه‌به آنچه تاکنون بیان شد و اهمیت مسئله‌ی لب‌خوانی، مسئله‌ی موردنظر در این مقاله تشخیص گفتار و عبارات گوینده از روی تصاویر ویدئویی لب او است. ورودی ویدئویی است که از دوربین مقابل شخص گوینده در حین صحبت بر اساس قواعد تعیین شده و فهرست کلمات مشخص به‌دست‌آمده و خروجی عبارت گفتاری است. پژوهش‌های انجام شده در زمینه لب‌خوانی مبتنی بر استخراج ویژگی از ناحیه لب هستند. این ناحیه همواره شامل اطلاعات مفید گفتاری است و از طرفی دارای بردار ویژگی با بعد کم‌تری است. فرایند لب‌خوانی به‌طورکلی شامل چند مرحله است. شکل (۱) چارچوب کلی مراحل لب‌خوانی در پژوهش‌های انجام شده را نمایش می‌دهد. ابتدا قاب‌های[3] چهره‌ی بیانگر گفتار از ویدئوهای افراد گوینده نمونه‌برداری می‌شوند. سپس یک پیش‌پردازش انجام می‌شود تا بتوان ناحیه لب را از سایر اجزای چهره در قاب‌های تصویری از ویدئو تشخیص داد. این مرحله شامل تشخیص چهره، تعیین محل لب‌ها و استخراج ناحیه لب است. در ادامه ویژگی‌های ناحیه لب توسط روش‌های سنتی یا شبکه‌های عصبی عمیق استخراج می‌شوند. در این مرحله ویژگی‌های مؤثر و مرتبط یافت می‌شوند و کاهش بعد انجام می‌شود. در نهایت یک طبقه‌بند مناسب مبتنی بر ویژگی‌های استخراج شده آموزش داده می‌شود و تشخیص عبارت گفتاری از تصویر انجام می‌شود [۶, ۷].



شکل (۱). مراحل لب‌خوانی

باتوجه‌به مراحل نشان‌داده‌شده در شکل (۱)، روش‌های موجود در فرایند لب‌خوانی از ویژگی‌های مناسب برای تشخیص دیداری عبارت گفتاری استفاده می‌کنند. استخراج ویژگی نقش اساسی در فرایند لب‌خوانی دارد. روش‌های استخراج ویژگی به دودسته کلی روش‌های سنتی و روش‌های مبتنی بر شبکه‌های عصبی عمیق تقسیم می‌شوند. روش‌های سنتی شامل روش‌های مبتنی بر پیکسل [۸]، مبتنی بر شکل [۹] و یا ترکیبی از این روش‌ها است.

روش‌های سنتی استخراج ویژگی به دلیل نیاز به دانش خبرگی، زمان‌بر بودن و هزینه بالا مناسب نیستند. به همین دلیل شبکه‌های عصبی عمیق [۱۰-۱۲] جایگزین آن‌ها شده‌اند؛ زیرا با پردازش مستقیم ویدئوها، ویژگی‌های سطح بالا را به‌صورت خودکار استخراج کرده و دقت و انعطاف‌پذیری لب‌خوانی را افزایش می‌دهند. در سال‌های اخیر، شبکه‌های عصبی عمیق با انجام هم‌زمان استخراج ویژگی و طبقه‌بندی، تحول بزرگی در الگوریتم‌ها و کاربردهای لب‌خوانی ایجاد کرده‌اند.

اسائل و همکارانش اولین مدل لب‌خوانی در سطح جمله را بـه نـام LIPNET ارائه نمودند. ابتدا دنباله‌ای از قاب‌های تصاویر ویدئو به‌عنوان ورودی به شبکه داده می‌شوند. سپس از CNN[4] سه‌بعدی برای استخراج ویژگی بصری پویا و اطلاعات کوتاه‌مدت موجود در بعد زمان قاب‌های تصاویر ویدئویی استفاده می‌شود. در ادامه از واحد بازگشتی LSTM[5] برای ایجاد خروجی و نگاشت بین حرکت‌های لب و نمایش متن استفاده می‌شود. در پایان از تابع هزینه CTC[6] برای به‌دست‌آوردن اختلاف بین خروجی پیش‌بینی‌شده و خروجی واقعی استفاده می‌شود [۱۳].

1. Viseme.
2. Visyllable.
3. Frame.
4. Convolutional Neural Network.
5. Long Short Term Memory.
6. Connectionist Temporal Classification.

زو و همکارانش شبکه‌ای به نام LCANet بر پایه معماری CNN LSTM ارائه نمودند. این روش مبتنی بر سه‌گام اصلی است. در گام اول ناحیه دهان از قاب‌های تصاویر ویدئوی ورودی تشخیص داده می‌شود. در گام دوم ویژگی‌های مکانی و زمانی قاب‌های تصاویر ناحیه دهان با استفاده از رمزگذار CNN LSTM استخراج می‌شوند. در گام سوم رمزگشای Bi GRU از یک‌لایه بیشینه نرم برای محاسبه توزیع احتمال خروجی واحدهای توجه استفاده می‌کند. هر واحد توجه یک کاراکتر را در خروجی نمایش می‌دهد [۱۴].

ابرار و همکارانش از شبکه VGG از پیش آموزش داده شده، برای لب‌خوانی استفاده کردند. در این روش از تنظیم مجدد شبکه‌های از پیش آموزش‌یافته CNN به‌جای آموزش از ابتدا شبکه‌ها استفاده شده است. براین‌اساس از شبکه VGG از پیش آموزش داده شده روی مجموعه‌داده Imagenet استفاده شد. برای آموزش شبکه از یادگیری انتقالی و باز تنظیم لایه‌های انتهایی روی مجموعه‌دادگان MIRACL-VC1 استفاده گردید [۱۵].

پارخ و همکارانش یک مدل جدید برای استخراج ویژگی به نام CAE[1] یا شبکه‌های خود رمزگذار پیچشی معرفی نمودند. آموزش این شبکه در دو مرحله انجام می‌شود. ابتدا شبکه CAE برای باز تولید تصاویر ورودی با استفاده از معیار میانگین مربع خطا آموزش داده می‌شود. سپس بخش رمزگشای کانولوشن آن حذف می‌شود. ویژگی‌هایی که توسط بخش رمزگذار از تصاویر استخراج شده‌اند به شبکه LSTM برای طبقه‌بندی تصاویر داده می‌شوند [۱۶].

ریوا و همکارانش مدلی را برای لب‌خوانی ارائه نمودند. آن‌ها یک معماری یادگیری عمیق مبتنی بر ترکیب دو شبکه مجزا را معرفی نمودند. هریک از این دو شبکه شامل لایه‌های پیشرو و LSTM هستند. ابتدا ناحیه لب به‌اندازه ۸۰×۴۰ پیکسل از قاب‌های تصویر برش داده می‌شود. سپس به یکی از شبکه‌ها قاب‌های تصویر و به شبکه دیگر اختلاف پیکسل‌های تصویر در هر قاب با قاب قبلی به‌عنوان ورودی به شبکه داده می‌شود. در ادامه خروجی این دو شبکه به LSTM داده می‌شود. در پایان خروجی نهایی به یک‌لایه بیشینه نرم داده می‌شود تا با یکدیگر ادغام شوند [۱۷].

بیش‌تر پژوهش‌های لب‌خوانی از شبکه‌های بازگشتی[2] استفاده کرده‌اند، اما این شبکه‌ها در یادگیری وابستگی‌های بلندمدت با مشکل محوشدگی[3] و انفجار گرادیان[4] روبه‌رو هستند و به مقداردهی اولیه وزن‌ها حساس‌اند .در شبکه‌های بازگشتی به دلیل وجود تعداد زیاد لایه درصورتی‌که مقدار اولیه وزن‌ها کوچک باشد؛ مقدار گرادیان وزن‌ها در فرایند پس انتشار خطا[5] صفر می‌شود و درصورتی‌که مقدار اولیه وزن‌ها بزرگ باشد؛ مقادیر گرادیان وزن‌ها به‌صورت نمایی رشد می‌کند و منجر به انفجار گرادیان می‌گردد؛ بنابراین مقداردهی اولیه وزن‌ها در هنگام یادگیری توالی‌ها تأثیر بسیار زیادی در سرعت آموزش و همگرایی شبکه‌های بازگشتی دارد. این محدودیت‌ها سبب جایگزینی شبکه‌های بازگشتی با معماری ترنسفورمر گردید. ترنسفورمرها با استفاده از مکانیزم توجه چند سر، قادر به مدل‌سازی مستقیم وابستگی‌های بلندمدت بدون نیاز به پردازش ترتیبی هستند. این ویژگی موجب افزایش سرعت آموزش و قابلیت موازی‌سازی می‌شود و مشکل ناپدیدشدن گرادیان را برطرف می‌سازد.

هوانگ و همکارانش یک مدل لب‌خوانی را با استفاده از شبکه ترنسفورمر معرفی کردند. این مدل مبتنی بر سه‌گام اصلی است. در گام اول مجموعه‌دادگان پردازش می‌شود. در این مرحله هر ویدئو به قاب‌های تصاویر سازنده‌اش تبدیل می‌شود. سپس تشخیص چهره برای هر قاب از تصویر انجام می‌شود و ناحیه لب در هر یک از قاب‌های تصویر تعیین می‌گردد. در گام دوم استخراج ویژگی انجام می‌شود و از طریق یادگیری انتقالی پارامترهای شبکه VGG از پیش آموزش داده شده به اشتراک گذاشته می‌شوند. بدین ترتیب یادگیری مدل تسریع و بهینه می‌گردد. در پایان ویژگی‌های استخراج شده به شبکه ترنسفورمر برای طبقه‌بندی منتقل می‌شوند. دقت این مدل بر روی مجموعه‌داده GRID ۴۵/۸۱ درصد گزارش شده است که در مقایسه با سایر روش‌های اجراشده بر روی این مجموعه‌داده عملکرد قابل‌قبولی ارائه نمی‌دهد [۱۸]. مجموعه‌داده GRID شامل جملاتی است که از تعدادی کلمه محدود با ترتیب ثابت تشکیل شده‌اند. در این مجموعه فایل alignment زمان شروع و پایان هر کلمه را مشخص می‌کند. اکثر پژوهش‌های پیشین در حوزه‌ی زبان انگلیسی بر روی این مجموعه داده‌ی محدود و ساختارمند انجام شده‌اند.

ونگ و همکارانش روشی مبتنی بر ترکیب ترنسفورمر بینایی و CNN سه‌بعدی[6] را برای استخراج ویژگی‌های مکانی و زمانی قاب‌های تصاویر متوالی ویدئو در لب‌خوانی پیشنهاد کردند. CNN می‌تواند فقط یک تصویر را پردازش کند؛ ولی اطلاعات بین قاب‌های تصاویر متوالی از ویدئو در بعد زمان را در فرایند لب‌خوانی نمی‌تواند استخراج کند. درحالی‌که CNN سه‌بعدی اطلاعات موجود در بعد زمان را استخراج می‌کند. در این روش از CNN سه‌بعدی قبل از ترنسفورمر استفاده می‌شود. بدین ترتیب اطلاعات کوتاه‌مدت بین قاب‌های تصاویر متوالی ویدئویی توسط CNN سه‌بعدی استخراج می‌شوند. سپس برای استخراج ویژگی‌های مؤثر محلی و سراسری بین قاب‌های تصاویر متوالی ویدئویی از ترنسفورمرها استفاده می‌شود. ترنسفورمرها توسط مکانیزم توجه ویژگی‌های سراسری را به

1. Convolutional Auto Encoder.
2. Recurrent Neural Network.
3. Vanishing Gradient.
4. Exploding Gradient.
5. Back Propagation.
6. 3D Convolutional Vision Transformer (3DCVT).

دست می‌آورند. به‌این‌ترتیب ترکیبی از ترنسفورمر و CNN سه‌بعدی برای استخراج ویژگی‌های مکانی و زمانی مؤثر از قاب‌های تصاویر متوالی به کار می‌رود [۱۹].

قلی‌پور و همکارانش به بررسی و مقایسه دو معماری شبکه عصبی پیچشی سه‌بعدی و شبکه ترکیبی پیچشی دوبعدی با لایه حافظه کوتاه‌مدت پرداختند. آن‌ها برای مقایسه و طبقه‌بندی کلمات از یک مجموعه‌دادگان محدود ۲۵ کلمه‌ای در زبان فارسی و یک مجموعه‌دادگان انگلیسی به نام ouluvs2 استفاده نمودند. نتایج آزمایش‌ها برتری معماری دوم و قدرت بیشتر شبکه‌های حافظه کوتاه‌مدت را در مدل‌کردن پویایی زمانی نشان داد. در معماری دوم برای استخراج ویژگی از تصاویر از شبکه‌های عصبی پیچشی دوبعدی و برای مدل‌سازی پویایی زمانی از لایه‌های حافظه کوتاه‌مدت استفاده شد. ویژگی‌های استخراج شده به خروجی با ابعادی متناسب با تعداد واژگان نگاشت شد. خروجی یک‌لایه‌ی خطی با تابع بیشینه نرم است که برای محاسبه توزیع احتمال خروجی هر یک از گره‌ها استفاده گردید. هر گره یک کلمه را در خروجی نمایش می‌دهد [۲۰].

پژوهش‌های اندکی در حوزه‌ی لب‌خوانی در زبان فارسی انجام شده است. این پژوهش‌ها بر روی مجموعه‌دادگان اختصاصی با تعداد محدودی کلمه انجام شده‌اند که در دسترس عموم قرار ندارند. پژوهش حاضر با بهره‌گیری از مجموعه‌دادگان معتبر PAVID-CVS [۲۱] در زبان فارسی به بررسی و تحلیل دقیق روش پیشنهادی برای تشخیص کلمات گفتاری در این زبان می‌پردازد. مجموعه‌دادگان پاوید شامل ۱۳۸ هجا و ۵۵۲ کلمه است که توسط ۴۰ گوینده بیان می‌شوند. کلمات در این مجموعه‌داده به دودسته تقسیم می‌شود. دسته اول شامل ۵۲ کلمه است که هریک از گویندگان ۴ بار این کلمات را تکرار می‌کنند. دسته دوم شامل ۵۰۰ کلمه است. کلمه‌های دسته دوم به ۱۰ گروه ۵۰ کلمه‌ای تقسیم می‌شوند. هر گروه از این کلمه‌ها توسط ۴ گوینده مختلف بیان می‌شوند. این مجموعه‌دادگان امکان توسعه و ارزیابی روش پیشنهادی را با دقتی بالا فراهم می‌سازد. شکل (۲) نمایی از یک فیلم واقعی را در مجموعه‌دادگان پاوید نمایش می‌دهد.



شکل (۲). نمایی از یکی از فیلم‌های مجموعه پاوید

برخلاف بسیاری از پژوهش‌های پیشین که عمدتاً روی زبان‌های انگلیسی تمرکز داشته‌اند، این تحقیق با بهره‌گیری از داده‌های واقعی فارسی و تحلیل دقیق حرکات لب، الگوریتمی مناسب و بهینه برای تشخیص لب‌خوانی در زبان فارسی ارائه می‌دهد.

پژوهش‌های محدود پیشین در حوزه‌ی لب‌خوانی در زبان فارسی بر تشخیص کامل کلمات از روی تحلیل حرکات لب متمرکز بوده‌اند و از طبقه‌بندی کلمات در خروجی بهره می‌بردند. در پژوهش حاضر از رویکردی متفاوت بهره گرفته شده است. نوآوری اصلی این پژوهش بازشناسی حرکات لب در سطح واحدهای واج یا ویزم است. این رویکرد امکان تعمیم‌پذیری بهتر و بازسازی گفتار پیوسته را فراهم می‌سازد. در این پژوهش به دلیل اهمیت تشخیص واج در سامانه‌های لب‌خوانی و لزوم بررسی ارتباط قاب‌های ویدئویی دهان برای تشخیص دنباله‌ای از واج‌ها و تولید کلمات از ترنسفورمر استفاده می‌گردد.

در این مقاله ساختاری نوین مبتنی بر ترکیب شبکه‌های خود رمزگذار پیچشی و ترنسفورمر برای تشخیص گفتار از ویدئو در زبان فارسی ارائه می‌شود. در روش پیشنهادی ویژگی‌ها با استفاده از شبکه‌های خود رمزگذار پیچشی تولید می‌شوند. سپس ویژگی‌های استخراج شده به‌عنوان ورودی به ترنسفورمر برای طبقه‌بندی و تولید متن عبارت گفتاری در خروجی داده می‌شوند.

تمرکز اصلی این پژوهش بر ترکیب شبکه‌های خود رمزگذار پیچشی و ترنسفورمر برای بازشناسی دنباله‌ی واج‌ها است. دلیل این تمرکز، پتانسیل بالای شبکه‌های خود رمزگذار پیچشی در استخراج الگوهای پیچیده از ویدئو گوینده و ارائه‌ی یک روش مؤثر در فرایند لب‌خوانی بر مبنای ویدئوی گوینده است. این استخراج الگو موجب تسریع فرایند آموزش سامانه لب‌خوانی می‌گردد.

باتوجه‌به محدودیت‌ها و کمبود پژوهش‌ها در زمینه‌ی لب‌خوانی در زبان فارسی، این مقاله نخستین گام در توسعه سامانه‌های لب‌خوانی اختصاصی برای زبان فارسی محسوب می‌شود و می‌تواند مبنای مطالعات و کاربردهای آینده در این حوزه باشد.

از محاسن روش پیشنهادی کاهش بهینه‌ی زمان آموزش و ارتقای کیفی عملکرد سامانه است. رویکرد پیشنهادی گامی بلند در مسیر تکامل فناوری‌های لب‌خوانی خودکار است. این دستاورد، تحولی اساسی در توسعه‌ی سامانه‌های هوشمند تشخیص دیداری گفتار محسوب می‌شود. به طور خلاصه سهم علمی این مقاله به شرح زیر است: الف ارائه‌ی ساختار مبتنی بر ترنسفورمر به‌منظور تشخیص دنباله‌ای از واج‌ها و ترکیب آن‌ها جهت تولید کلمات ب) نخستین پیاده‌سازی سامانه لب‌خوانی مبتنی بر یادگیری عمیق در زبان فارسی ج) ارائه‌ی روشی برای تشخیص گفتار پیوسته.

ساختار مقاله به‌صورت زیر است. در بخش دوم ترنسفورمرها توضیح داده می‌شود. روش پیشنهادی تشخیص دیداری گفتار در بخش سوم ارائه می‌گردد. بخش چهارم نتایج شبیه‌سازی روش پیشنهادی بر روی ویدئوهای نمونه نمایش داده می‌شود. نتیجه‌گیری مقاله در بخش پنجم مطرح خواهد شد.

۲. ترنسفورمر

ترنسفورمرها، تحولی در حوزه‌ی یادگیری عمیق به وجود آوردند. ترنسفورمرها ابتدا در حوزه‌ی پردازش زبان طبیعی به کار می‌رفتند. اما به‌تدریج دامنه‌ی کاربرد آنها گسترش یافت و در سایر کاربردها ازجمله پردازش تصاویر ویدئویی مورداستفاده قرار گرفتند. معماری ترنسفورمر بر مبنای مکانیزم توجه است و از تکرار شبکه‌های پیچشی جلوگیری می‌کند. آزمایش‌ها در ترجمه ماشینی نشان می‌دهد که این مدل‌ها از نظر کیفیت برتر هستند، درحالی‌که موازی پذیری بیشتری دارند و به زمان قابل‌توجه کمتری برای آموزش نیاز دارند.

مدل‌های ترنسفورمر از دو بخش اصلی به نام‌های رمزگذار [۱] و رمزگشا[۲] تشکیل شده‌اند. بخش رمزگذار وظیفه‌ی پردازش ورودی‌ها و تولید بازنمایی‌های داخلی را برعهده دارد. این بخش از لایه‌های توجه چند سر[۳] و شبکه‌های پیش‌خور موضعی تشکیل شده است که به ترتیب پردازش می‌شوند. این ترکیب موجب استخراج ویژگی‌های پیچیده‌ای از ورودی‌ها می‌شود. بخش رمزگشا مسئول تولید خروجی‌ها بر اساس بازنمایی‌های داخلی تولید شده توسط رمزگذار است. این بخش نیز از لایه‌های توجه چند سر و شبکه‌های پیش‌خور تشکیل شده است. علاوه بر این، بخش رمزگشا با درنظرگرفتن بازنمایی‌های تولید شده در بخش رمزگذار، داده‌های ورودی را به خروجی‌های متناظر تبدیل می‌نماید [۲۲].

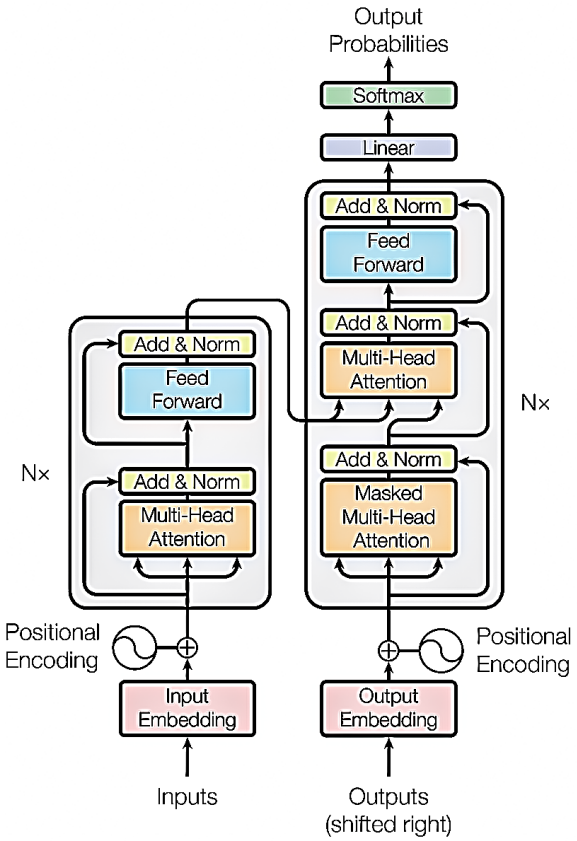
سازوکار توجه چند سر از چندین لایه توجه به خود[۴] تشکیل شده است. هر یک از این لایه‌ها با بهره‌گیری از سه ماتریس مجزای پرسش[۵]، کلید[۶] و ارزش[۷] اهمیت نسبی بخش‌های مختلف ورودی را محاسبه و بازنمایی‌های منحصربه‌فردی از داده‌های ورودی تولید می‌کنند.

بازنمایی‌های گوناگون حاصل از لایه‌های توجه به خود با هم ترکیب می‌شوند تا یک بازنمایی جامع‌تر و غنی‌تر از ویژگی‌های ورودی ایجاد گردد. سپس با اعمال لایه‌های پیش‌خور، این نمایش‌های غنی‌شده به مقادیر خروجی مطلوب متناظر نگاشت می‌شوند. رابطه‌ی (۱) نحوه عملکرد توجه چند سر در معماری ترنسفورمر را نمایش می‌دهد [۲۲].

$$head_i = Attention(QW^Q, KW^K, VW^V)$$
$$= Softmax\left[\frac{QW_i^Q(KW_i^K)^T}{\sqrt{(d_k)}}\right]VW_i^V \quad (۱)$$
$$MultiHead(Q,K,V) =$$
$$concat(head_1, head_2, \dots, head_h)W_o$$

در رابطه (۱) W ماتریس وزن، K ماتریس کلید، Q ماتریس پرس‌وجو، V ماتریس ارزش و d_k ابعاد بردار کلید را نشان می‌دهد.

این مدل‌ها با استفاده از سازوکار توجه[۸]، توانایی مدل‌سازی روابط پیچیده و وابستگی‌های طولانی‌مدت بین عناصر داده را دارند. برخلاف شبکه‌های عصبی بازگشتی که به‌صورت متوالی داده‌ها را پردازش می‌کنند، ترنسفورمرها می‌توانند به‌صورت موازی عمل کنند. این امر موجب کاهش زمان آموزش و افزایش بازدهی در پردازش داده‌ها می‌شود. قابلیت ترنسفورمرها در درک روابط معنایی و ساختاری دادگان آن‌ها را به ابزاری قدرتمند تبدیل کرده است [۲۲]. شکل (۳) معماری ترنسفورمر اولیه که توسط واسوانی و همکارانش ارائه شد را نمایش می‌دهد.



شکل (۳). معماری ترنسفورمر [۲۲]

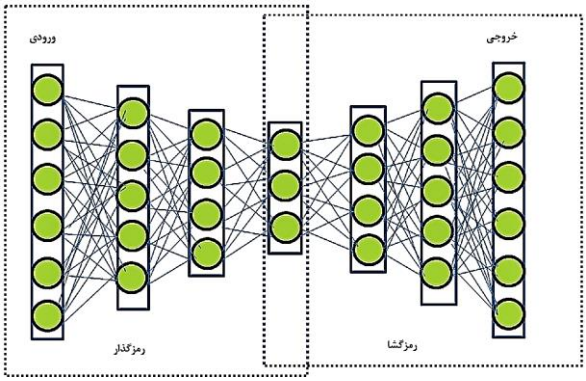
[1]. Encoder.
[2]. Decoder.
[3]. Multi-Head Attention.
[4]. Self-Attention.
[5]. Query.
[6]. Key.
[7]. Value.
[8]. Attention.

۳. روش پیشنهادی

اکثر پژوهش‌های اخیر در زمینه‌ی لب‌خوانی بر روی زبان انگلیسی متمرکز شده‌اند. ویژگی‌های منحصربه‌فرد زبان فارسی، شامل واج‌های خاص و حرکات دهانی متفاوت نیاز به توسعه‌ی روش‌های لب‌خوانی را در زبان فارسی ایجاد کرده است. پژوهش حاضر نخستین گام در توسعه سامانه‌های لب‌خوانی اختصاصی برای زبان فارسی به شمار می‌رود. تشخیص دقیق واج‌ها در سامانه‌های لب‌خوانی اهمیت فراوانی دارد، زیرا واج‌ها واحد بنیادین زبان را تشکیل می‌دهند و حرکت‌های متمایز لب در تصویر را ایجاد می‌کنند. ازآنجاکه لب‌خوانی مبتنی بر تحلیل حرکت‌های بصری صورت و لب‌هاست، تشخیص صحیح واج‌ها منجر به بهبود دقت در سامانه‌های لب‌خوانی می‌گردد. به‌عبارت‌دیگر، تشخیص دقیق واج‌ها باعث کاهش ابهام بین حرکت‌های مشابه لب و افزایش قابلیت تفکیک کلمات می‌گردد. تشخیص دقیق واج برای سامانه‌های لب‌خوانی به‌ویژه در زبان فارسی، حیاتی است. اکثر پژوهش‌های انجام شده در زمینه‌ی لب‌خوانی از شبکه‌های بازگشتی[1] استفاده کرده‌اند. همان‌طور که اشاره شد شبکه‌های بازگشتی با مشکل محوشدگی گرادیان[2] و انفجار گرادیان در هنگام یادگیری توالی‌های بلندمدت مواجه هستند که توانایی یادگیری توالی‌های طولانی را کاهش می‌دهد. این محدودیت‌ها موجب شد تا معماری ترنسفورمر، با بهره‌گیری از مکانیزم توجه و قابلیت پردازش موازی، به‌عنوان جایگزینی مؤثر برای شبکه‌های بازگشتی مطرح گردد.

در این پژوهش با بهره‌گیری از ترکیب شبکه خود رمزگذار پیچشی و معماری ترنسفورمر روشی نوین برای لب‌خوانی در زبان فارسی ارائه می‌شود. برای تسریع و عملکرد بهتر ابتدا ویژگی‌های هریک از قاب‌های ویدئویی توسط شبکه‌ی خود رمزگذار پیچشی استخراج می‌شود. سپس ویژگی‌های استخراج شده به‌عنوان ورودی به شبکه‌ی ترنسفورمر برای طبقه‌بندی داده می‌شوند.

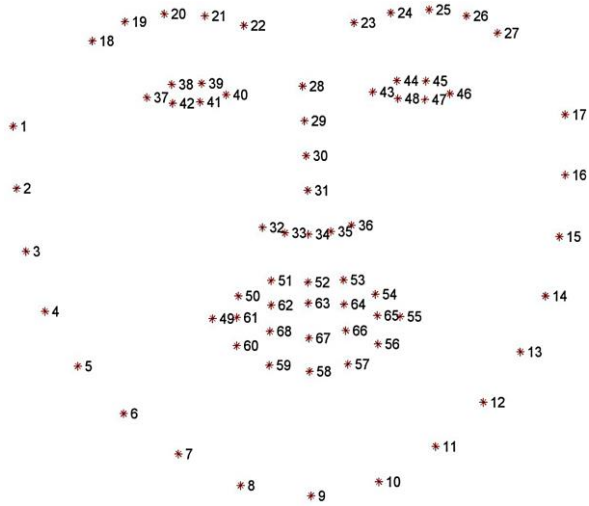
شبکه‌های خود رمزگذار از دو بخش رمزگذار و رمزگشا تشکیل شده‌اند. بخش رمزگذار تصویر ورودی را برای استخراج ویژگی از چندلایه شبکه عصبی عبور می‌دهد و بردار ویژگی را استخراج می‌کند. بخش رمزگشا برای بازسازی تصویر اصلی در خروجی از روی بردار ویژگی استفاده می‌شود. تابع هزینه به‌صورت اختلاف تصویر ورودی و خروجی تعریف می‌شود. در طی فرایند آموزش تابع هزینه بایستی کاهش یابد. لایه وسط گلوگاه نامیده می‌شود که بردار ویژگی استخراج شده را نمایش می‌دهد و از آن برای دسته‌بندی استفاده می‌شود [۲۳]. شبکه‌های خود رمزگذار یک روش یادگیری هوشمند بدون ناظر است. این شبکه یک الگوریتم یادگیری عمیق است که از دو شبکه‌ی متقارن با چندلایه کم‌عمق ساخته شده است. شکل (۴) ساختار خود رمزگذار پیچشی را نمایش می‌دهد.



شکل (۴). نمایی از ساختار خود رمزگذار

۳-۱. پیش‌پردازش داده

برای آموزش شبکه‌ی خود رمزگذار پیچشی، ابتدا ویدئو به قاب‌های سازنده‌اش تبدیل می‌شود. تشخیص چهره برای هر قاب از ویدئو انجام می‌شود و نمونه قاب‌های ویدئویی بدون چهره حذف می‌شوند. سپس از هر قاب ناحیه اطراف لب‌ها برش داده می‌شود. برای شناسایی چهره و جداسازی ناحیه لب از ردیاب نقطه‌ی صورت Dlib [۲۴] استفاده می‌شود. این ردیاب ۶۸ نقطه روی صورت را تشخیص می‌دهد. شکل (۵) نمایی از این ۶۸ نقطه را نشان می‌دهد.

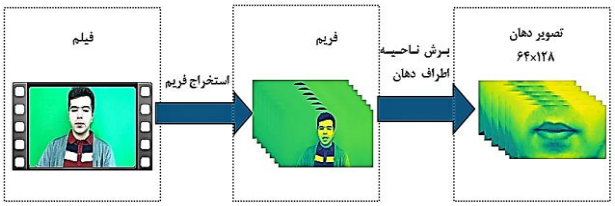


شکل (۵). نمایی از ۶۸ نقطه ردیابی شده در چهره [۲۵]

الگوریتم Dlib مبتنی بر روش هیستوگرام گرادیان‌های جهت‌دار[3] و ماشین بردار پشتیبان[4] است که با استخراج ویژگی‌های لبه و ساختارهای محلی تصویر، توانایی تشخیص دقیق و سریع چهره را در شرایط گوناگون روشنایی دارد. الگوریتم Dlib به دلیل دقت بالا و کارایی مطلوب، به طور گسترده‌ای در پردازش تصویر و بینایی ماشین مورداستفاده قرار می‌گیرد. این الگوریتم قادر به شناسایی چهره در تصاویر با وضوح مختلف است.

بعد از شناسایی لب‌ها توسط ردیاب Dlib ناحیه اطراف لب با ابعاد (۱۲۸×۶۴) برش داده می‌شود. شکل (۶) استخراج ناحیه اطراف لب در قاب‌های ویدئویی را نشان می‌دهد.

[1]. Recurrent Neural Network.
[2]. Vanishing Gradient.
[3]. Histogram Of Oriented Gradients.
[4]. Support Vector Machine.



شکل (۶). استخراج قاب‌های لب از ویدئو

۲-۳. استخراج ویژگی لب: قاب‌های لب در ویدئو که در شکل (۶) نمایش داده شده‌اند به‌عنوان داده‌های ورودی برای آموزش شبکه‌ی خود رمزگذار پیچشی مورد استفاده قرار می‌گیرند. در این پژوهش از شبکه خود رمزگذار پیچشی برای استخراج ویژگی هریک از قاب‌های لب در ویدئو استفاده می‌شود. ساختار رمزگذار و رمزگشای شبکه‌ی خود رمزگذار پیچشی به‌کاررفته در این مقاله در شکل‌های (۷) و (۸) نمایش‌داده‌شده است.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 64, 128, 8)	80
max_pooling2d (MaxPooling2D)	(None, 32, 64, 8)	0
conv2d_1 (Conv2D)	(None, 32, 64, 8)	584
max_pooling2d_1 (MaxPooling2D)	(None, 16, 32, 8)	0
conv2d_2 (Conv2D)	(None, 16, 32, 8)	584
max_pooling2d_2 (MaxPooling2D)	(None, 8, 16, 8)	0
conv2d_3 (Conv2D)	(None, 8, 16, 8)	584
flatten (Flatten)	(None, 1024)	0

Total params: 1,832
Trainable params: 1,832
Non-trainable params: 0

شکل (۷). نمایی از رمزگذار پیچشی

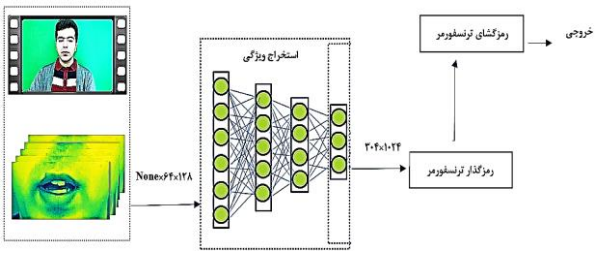
Model: "sequential_1"

Layer (type)	Output Shape	Param #
reshape (Reshape)	(None, 8, 16, 8)	0
conv2d_transpose (Conv2DTranspose)	(None, 8, 16, 8)	584
up_sampling2d (UpSampling2D)	(None, 16, 32, 8)	0
conv2d_transpose_1 (Conv2DTranspose)	(None, 16, 32, 8)	584
up_sampling2d_1 (UpSampling2D)	(None, 32, 64, 8)	0
conv2d_transpose_2 (Conv2DTranspose)	(None, 32, 64, 3)	219
up_sampling2d_2 (UpSampling2D)	(None, 64, 128, 3)	0
conv2d_transpose_3 (Conv2DTranspose)	(None, 64, 128, 1)	28

Total params: 1,415
Trainable params: 1,415
Non-trainable params: 0

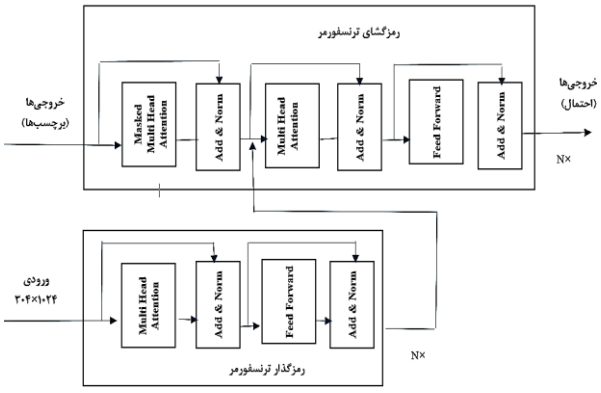
شکل (۸). نمایی از رمزگشای پیچشی

پس از آموزش شبکه‌ی خود رمزگذار پیچشی بخش رمزگذار آن می‌تواند ویژگی‌های مکانی قاب‌های تصویر را استخراج کند. ویژگی‌های استخراج شده به‌عنوان ورودی به ترنسفورمر برای طبقه‌بندی و تولید متن عبارت گفتاری در خروجی منتقل می‌شوند. معماری ترنسفورمر با شناسایی وابستگی‌های زمانی میان ویژگی‌های استخراج شده از قاب‌های تصویر قادر است ارتباطات طولانی، پیچیده و غیرخطی بین آن‌ها را تشخیص دهد و متن عبارت گفتاری را در خروجی تولید نماید. شکل (۹) معماری روش پیشنهادی را نمایش می‌دهد.



شکل (۹). معماری کلی روش پیشنهادی

معماری ترنسفورمر در این مقاله از ساختار مبتنی بر رمزگذار و رمزگشا تشکیل شده است. روش پیشنهادی با تطبیق معماری ترنسفورمر، مدلی کارآمد برای فرایند لب‌خوانی توسعه داد. شکل (۱۰) معماری ترنسفورمر به‌کاررفته در این روش را نمایش می‌دهد.



شکل (۱۰). معماری ترنسفورمر

همان‌طور که در شکل (۱۰) ملاحظه می‌شود ورودی مدل ترنسفورمر یک ماتریس ویژگی با ابعاد ۳۰۴×۱۰۲۴ است. هر بردار ۱۰۲۴ بعدی نمایانگر ویژگی‌های استخراج‌شده از یک قاب تصویر است. باتوجه‌به آزمایش‌های انجام شده حداکثر طول توالی ویدئو ورودی ۳۰۴ قاب در نظر گرفته شده است. معماری ترنسفورمر با بهره‌گیری از قابلیت‌های خودتوجهی وابستگی‌های بلندمدت در دنباله‌های ویدئویی را به‌خوبی مدلسازی می‌کند.

۴. نتایج شبیه‌سازی

در این تحقیق، برای پیاده‌سازی روش پیشنهادی از زبان برنامه‌نویسی پایتون استفاده شده است. کتابخانه Pandas برای پردازش پایگاه‌داده و کتابخانه TensorFlow برای طراحی

معماری‌های یادگیری عمیق پیشنهادی مورداستفاده قرار گرفت. اجرای تمام کدها با استفاده از سرور مجازی GPU 3090 شرکت هوش مصنوعی اهورا انجام شده است.

توسعه‌ی مدل پیشنهادی ترکیب شبکه‌ی خود رمزگذار پیچشی و ترنسفورمر به سه‌فاز تقسیم می‌شود. در گام اول، ابتدا هریک از شبکه‌ها به‌صورت مجزا بر روی یک مجموعه‌داده‌ی متناسب آموزش داده می‌شود. در گام دوم، هریک از شبکه‌های مجزا به‌صورت پیوسته بر روی مجموعه‌داده‌ی اعتبارسنجی ارزیابی می‌شود تا کیفیت تعمیم هریک از شبکه‌ها بر روی داده‌های دیده نشده بررسی شود. در گام سوم، زمانی که فرایند آموزش خاتمه یافت، مدل بر روی یک مجموعه‌داده‌ی آزمون ارزیابی می‌شود تا معیارهای ارزیابی نهایی مدل محاسبه شود.

یکی از چالش‌ها در حوزه‌ی لب‌خوانی وابستگی به گوینده است. زیرا افراد مختلف دارای لهجه‌های خاص هستند و حرکات لب متفاوتی دارند و تنوع رفتار بر نوع صحبت کردن افراد تأثیرگذار است. به‌عنوان مثال حالات خلقی متفاوت مانند عصبانیت بر طرز بیان افراد تاثیر گذار خواهد بود. چالش دیگر این پژوهش زیاد بودن توالی دادگان ورودی و ابعاد بالای قاب‌های تصاویر ویدئویی در حوزه‌ی لب‌خوانی است؛ بنابراین آموزش سامانه‌های لب‌خوانی با افزایش قابل‌توجه پیچیدگی و زمان آموزش مواجه است. در این پژوهش ابتدا برای رفع چالش‌های موجود و تسریع فرایند آموزش از شبکه‌های خود رمزگذار پیچشی برای استخراج ویژگی هریک از قاب‌های ویدئو در ورودی استفاده می‌شود. سپس ویژگی‌های استخراج شده به‌عنوان ورودی به ترنسفورمر برای طبقه‌بندی ویزم‌ها و تولید متن عبارت گفتاری در خروجی داده می‌شوند.

قاب‌های ویدئو در این مجموعه به‌صورت تصاویر رنگی هستند. این قاب‌ها با استفاده از تابع cvtColor در کتابخانه OpenCV به تصاویر خاکستری تبدیل می‌شوند. این تبدیل موجب کاهش ابعاد داده، کاهش پیچیدگی محاسباتی و حذف اطلاعات رنگی غیرضروری برای تشخیص چهره می‌شود؛ بنابراین تمرکز بر ویژگی‌های ساختاری تصویر افزایش می‌یابد. به‌منظور شناسایی نواحی چهره در تصاویر خاکستری، از ردیاب Dlib استفاده شد. مقادیر پیکسل‌های تصاویر ورودی بر عدد ۲۵۵ تقسیم می‌شوند تا دامنه‌ی مقادیر در بازه‌ی [۰,۱] قرار گیرد. این عملیات نرمال‌سازی منجر به کاهش واریانس داده‌های ورودی و تسهیل در فرایند بهینه‌سازی با استفاده از گرادیان نزولی می‌شود. نرمال‌سازی منجر به پایداری و یادگیری بهتر مدل‌ها در فرایند آموزش می‌گردد.

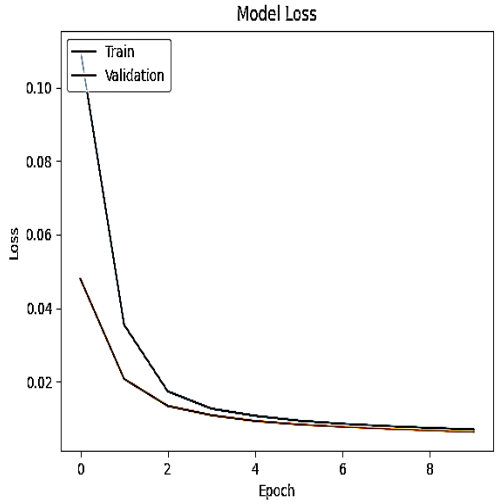
برای آموزش شبکه‌ی خود رمزگذار ابتدا مجموعه‌ای شامل ۱۰۴۰ فیلم از دادگان پاوید انتخاب شد. ۸۰ درصد از مجموعه‌ی انتخابی برای آموزش شبکه خود رمزگذار پیچشی و ۲۰ درصد برای اعتبارسنجی اختصاص یافت. این رویکرد باهدف نظارت بر روند آموزش و جلوگیری از پدیده بیش برازش در طول فرایند یادگیری استفاده شد. معیار ارزیابی در آموزش شبکه خود رمزگذار پیچشی تابع هزینه برای بازسازی قاب‌های ویدئویی است.

فهرست پارامترهای شبکه‌ی خود رمزگذار پیچشی در جدول (۱) نمایش‌داده‌شده است.

جدول (۱). پارامترهای شبکه خود رمزگذار پیچشی

Optimizer	adam
Loss	mse
Number of epochs	۱۰
learning rate	۰/۰۰۱

شکل (۱۱) نمودار هزینه شبکه‌ی خود رمزگذار پیچشی را در طول فرایند آموزش و ارزیابی نمایش می‌دهد.



شکل (۱۱). نمودار هزینه شبکه خود رمزگذار پیچشی

شکل (۱۱)، منحنی خطای آموزش مدل خود رمزگذار پیچشی پیشنهادی را بر اساس معیار میانگین مربع خطا نشان می‌دهد. ابتدا خطا حدود ۰/۱۲ است که بیانگر خطای بازسازی نسبتاً بالا ناشی از مقداردهی اولیه‌ی تصادفی وزن‌های شبکه است. در تکرارهای بعدی، مقدار خطا به‌سرعت و به‌صورت یکنواخت کاهش می‌یابد و پس از ۴ بار تکرار آموزش به مقدار ۰/۰۱ می‌رسد. رابطه‌ی (۲) نحوه‌ی محاسبه‌ی معیار میانگین مربع خطا را نمایش می‌دهد.

$$MSE = \frac{\sum(y_i - \hat{y_i})}{n} \qquad (۲)$$

پارامتر y_i میزان واقعی و $\hat{y_i}$ میزان تخمینی متناظر با هر پیکسل از تصویر و n تعداد کل پیکسل‌های قاب است. کاهش سریع میزان خطای آغازین در شکل (۱۱) نشان می‌دهد که شبکه‌ی پیشنهادی به طور مؤثری ویژگی‌های مکانی ناحیه‌ی لب را آموخته و ویژگی‌های سطح پایین معنادار این ناحیه را در مدت‌زمان کوتاهی استخراج می‌کند.

همگرایی نرم و بدون نوسان، پایداری عددی و پیکربندی بهینه‌ی انتخاب‌شده (بهینه‌ساز adam با نرخ یادگیری ۰/۰۰۱) را تأیید می‌کند.

مقدار کمینه‌ی نهایی mse نشان‌دهنده‌ی موفقیت شبکه‌ی خود رمزگذار پیچشی در بازسازی قاب‌های تصویر خروجی و استخراج بردارهای ویژگی مؤثر است که برای مراحل شناسایی و طبقه‌بندی قابل استفاده‌اند. همچنین، همگرایی سریع در چهار تکرار بیانگر کارایی بالای آموزش مدل پیشنهادی است.

جدول (۲) و (۳) به ترتیب مشخصات لایه‌های رمزگذار و رمزگشای شبکه خود رمزگذار پیچشی را نمایش می‌دهد.

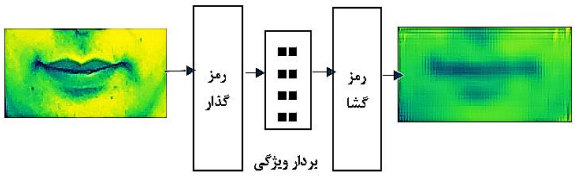
جدول (۲). مشخصات لایه‌های رمزگذار

لایه	ابعاد لایه	نوع فعال‌ساز	تعداد فیلتر	اندازه کرنل
Input	۶۴×۱۲۸×۱	-	-	-
Conv 2D	۶۴×۱۲۸×۸	relu	۸	۳×۳
Maxpooling	۳۲×۶۴×۸	-	-	۲×۲
Conv2D	۳۲×۶۴×۸	relu	۸	۳×۳
Maxpooling	۱۶×۳۲×۸	-	-	۲×۲
Conv2D	۱۶×۳۲×۸	relu	۸	۳×۳
Maxpooling	۸×۱۶×۸	-	-	۲×۲
Conv2D	۸×۱۶×۸	relu	۸	۳×۳

جدول (۳). مشخصات لایه‌های رمزگشا

لایه	ابعاد لایه	نوع فعال‌ساز	تعداد فیلتر	اندازه کرنل
Conv2DTranspose	۸×۱۶×۸	relu	۸	۳×۳
UpSampling2D	۱۶×۳۲×۸	-	-	۲×۲
Conv2DTranspose	۱۶×۳۲×۸	relu	۸	۳×۳
UpSampling2D	۳۲×۶۴×۸	-	-	۲×۲
Conv2DTranspose	۳۲×۶۴×۳	relu	۳	۳×۳
UpSampling2D	۶۴×۱۲۸×۳	-	-	۲×۲
Conv2DTranspose	۶۴×۱۲۸×۱	relu	۱	۳×۳

شکل (۱۲) نمایی از شبکه خود رمزگذار پیچشی را نمایش می‌دهد که ویژگی‌های مرتبط با ناحیه لب را استخراج نموده و قاب‌های ویدئویی این ناحیه را با دقت بالا بازسازی می‌کند.



شکل (۱۲). شبکه خود رمزگذار پیچشی

شکل (۱۲) نشان می‌دهد که معماری اتوانکودر از عملکرد مطلوبی در بازسازی تصاویر برخوردار است و بردار ویژگی ایجاد شده، کارایی لازم برای حفظ اطلاعات کلیدی را دارد.

پس از آموزش شبکه خود رمزگذار پیچشی، بخش رمزگذار آن قادر به استخراج ویژگی‌های مکانی از قاب‌های تصویر خواهد بود. تعداد قاب‌های ویدئویی در مجموعه‌دادگان پاوید یکسان نیست. حداکثر تعداد قاب‌های ورودی از مجموعه‌دادگان پاوید بر اساس آزمایش‌های انجام شده، برابر با ۳۰۴ قاب تعیین شده است. باتوجه‌به اینکه مدل ترنسفورمر در پیاده‌سازی خود انتظار دارد همه داده‌ها دارای ابعاد یکسان باشند، لذا برای هر ویدئو یک ماتریس با ابعاد ۳۰۴×۱۰۲۴ تعریف می‌شود که بعد ویژگی‌های استخراج شده برای هر قاب برابر با ۱۰۲۴ است. ویدیوهایی که زمان کمتری دارند برای یکسان سازی ابعاد آنها از پرکردن با صفر[1] استفاده می‌شود.

برای آموزش ترنسفورمر مجموعه‌داده متوازنی از کلمات انتخاب می‌شود. ویژگی هریک از قاب‌های ویدئویی با استفاده از بخش رمزگذار شبکه خود رمزگذار پیچشی آموزش داده شده، استخراج می‌شود؛ بنابراین برای هر قاب از ویدئو یک بردار ویژگی با ابعاد ۱۰۲۴ به دست می‌آید. بردارهای ویژگی استخراج شده به ترتیب در سطرهای ماتریس مذکور قرار می‌گیرند و به‌عنوان ورودی به مدل ترنسفورمر منتقل می‌شود. سپس متن گفتاری متناظر در خروجی تولید می‌شود.

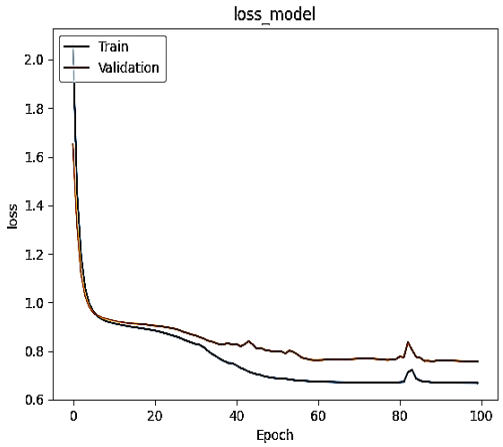
جدول (۴) فهرست پارامترهای معماری ترنسفورمر را نمایش می‌دهد. انتخاب بسیاری از فراپارامترها در جدول (۴) براساس بازخوردهای به‌دست‌آمده از عملکرد مدل پیشنهادی در چند اجرای اولیه انجام شده است. در این میان، تعداد hidden units برابر با ۱۰۲۴ در نظر گرفته شده است، زیرا بُعد بردارهای ورودی ترنسفورمر (بردار ویژگی‌های استخراج شده توسط بخش رمزگذار شبکه خود رمزگذار) نیز ۱۰۲۴ است. این انتخاب به منظور حفظ سازگاری ابعاد در معماری مدل پیشنهادی صورت گرفته است.

[1]. Zero Padding.

جدول (۴). پارامترهای معماری ترنسفورمر در روش پیشنهادی

Learning rate	۰/۰۰۰۰۱
Number of encoders	۴
Number of decoders	۴
Number of heads	۴
Number of epochs	۱۰۰
Number of feed forwards	۵۱۲
Number of hidden units	۱۰۲۴
Batch_size	۸
Loss	CategoricalCrossentropy
label_smoothing	۰/۱

شکل (۱۳) نمودار هزینه معماری ترنسفورمر را در روش پیشنهادی طی فرآیند آموزش نشان می‌دهد.



شکل (۱۳). نمودار هزینه ترنسفورمر در طول فرایند آموزش

نمودار نزولی تابع هزینه مدل ترنسفورمر در شکل (۱۳) روند همگرایی مطلوب و پایدار را در فرآیند آموزش نشان می‌دهد. ابتدا تابع هزینه با شیب تندی کاهش یافته که نشان‌دهنده یادگیری سریع الگوهای پایه توسط مدل ترنسفورمر است. این امر حاکی از کارایی بالای معماری ترنسفورمر در استخراج ویژگی‌های زمانی مؤثر داده‌ها است. در مرحله میانی (تکرار ۲۰ تا ۴۰)، کاهش هزینه با شیب ملایم‌تری ادامه یافته که بیانگر یادگیری تدریجی الگوهای پیچیده‌تر و روابط غیرخطی در داده‌ها است. در نهایت، در مرحله همگرایی (تکرار ۵۰ تا ۶۰)، مدل به نقطه بهینه خود نزدیک شده است.

پایداری فرایند آموزش و عدم بروز بیش برازش از طریق اختلاف کم بین مقادیر تابع هزینه آموزش و اعتبارسنجی مشهود است. این پایداری را می‌توان به عوامل متعددی از جمله مکانیزم توجه چند سر، اتصالات باقیمانده، و تنظیم مناسب فراپارامترها نسبت داد.

در این مطالعه، به دلیل محدودیت سخت‌افزاری و حافظه آزمایش‌ها بر روی ۵۰ کلمه از مجموعه‌دادگان پاوید انجام شد. مرجع [13] برای مقایسه کارایی روش پیشنهادی مورد بررسی قرار گرفت. جدول (5) فهرست پارامترهای روش مرجع [13] را نمایش می‌دهد.

جدول (۵). فهرست پارامترها مرجع [۱۳]

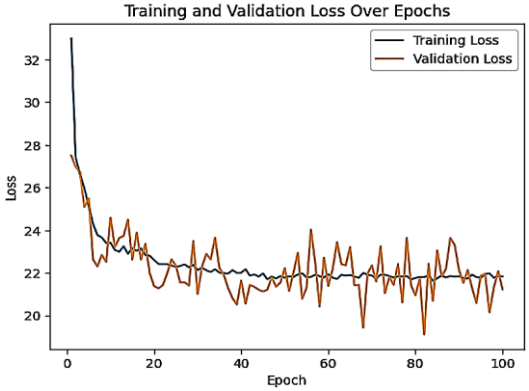
Batch _size	۲
dropout	۰/۱
Num_epochs	۱۰۰
Learning rate	۰/۰۰۱

در پیاده‌سازی مرجع [۱۳] به منظور کاهش حجم محاسبات و جلوگیری از اشغال بیش از حد حافظه، اندازه‌ی دسته برابر با ۲ تنظیم شد. جدول (۶) مشخصات لایه‌های روش لب‌خوانی مرجع [۱۳] را نمایش می‌دهد.

جدول (۶). مشخصات لایه‌های روش مرجع [۱۳]

لایه	ابعاد لایه	نوع فعال‌ساز	تعداد فیلتر	اندازه کرنل
Input	۳۰۴×۶۴×۱۲۸×۱	-	-	-
Conv 3D	۳۰۴×۶۴×۱۲۸×۱۲۸	relu	۱۲۸	۳×۳×۳
Maxpooling	۳۰۴×۳۲×۶۴×۱۲۸	-	_	۱×۲×۲
Conv3D	۳۰۴×۳۲×۶۴×۲۵۶	relu	۲۵۶	۳×۳×۳
Maxpooling	۳۰۴×۱۶×۳۲×۲۵۶	-	_	۱×۲×۲
Conv3D	۳۰۴×۱۶×۳۲×۷۵	relu	۷۵	۳×۳×۳
Maxpooling	۳۰۴×۸×۱۶×۷۵			
Bidirectional LSTM	۳۰۴×۵۱۲	tanh (پیش‌فرض)	_	_
Bidirectional LSTM	۳۰۴×۵۱۲	tanh (پیش‌فرض)	_	_
Dense	۳۰۴×۳۸	softmax	_	_

شکل (۱۴) نمودار هزینه مرجع [۱۳] را در طول فرایند آموزش روی مجموعه‌دادگان پاوید نمایش می‌دهد.



شکل (۱۴). نمودار هزینه مرجع [۱۳] در طول فرایند آموزش

باتوجه‌به نوسانات زیاد خطای اعتبارسنجی و همگرایی زودهنگام منحنی خطای آموزش در شکل (۱۴) ملاحظه می‌شود، مدل مرجع [۱۳] نتوانسته است ویژگی‌های دیداری مؤثری را از داده‌های ورودی استخراج کند. این مسئله موجب شده تا مدل در تعمیم‌پذیری به داده‌های جدید عملکرد ضعیفی داشته باشد. به‌عبارت دیگر، سیستم تنها تا حدی محدود الگوهای موجود در داده‌های آموزشی را یاد گرفته و در بازشناسی واج‌ها یا کلمات جدید دچار خطا شده است. نتایج حاصل از اجرای مدل لب‌خوانی توسط مرجع [۱۳] و روش پیشنهادی در تکرار ۱۰۰ آموزش روی مجموعه‌دادگان فارسی پاوید به ترتیب در شکل‌های (۱۵) و (۱۶) نمایش‌داده‌شده است.

```
Original: ZAle
Prediction: xorAni
~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Original: CAboki
Prediction: xorui
~~~~~~~~~~~~~~~~~~~~~~~~~~~~
wer_score:1.0
cer_score:1.0
```

شکل (۱۵). نتایج پیش‌بینی روش مرجع [۱۳]

```
Word Error Rate: 0.2500
-----------------------------------------------
char Error Rate: 0.1571
Target    : <paZuhi>
Prediction: <topogerAfi>
-----------------------------------------------
Target    : <viZenAme>
Prediction: <viZenAme>
-----------------------------------------------
Target    : <jodAnaSodani>
Prediction: <jodAnaSodani>
-----------------------------------------------
Target    : <hojumi>
Prediction: <xoruji>
-----------------------------------------------
Target    : <CAboki>
Prediction: <CAboki>
-----------------------------------------------
Target    : <robubi>
Prediction: <robubi>
-----------------------------------------------
Target    : <saduqi>
Prediction: <saduqi>
-----------------------------------------------
Target    : <kupe>
Prediction: <kupe>
-----------------------------------------------
epoch 99 _on_epoch_end_displayoutputs
```

شکل (۱۶). نتایج پیش‌بینی مجموعه‌دادگان پاوید توسط روش پیشنهادی

همانطور که در شکل (۱۵) و (۱۶) ملاحظه می‌شود؛ روش پیشنهادی در تشخیص کلمات موفق بوده است در حالی‌که عملکرد روش مرجع [۱۳] در تشخیص کلمات مجموعه‌دادگان پاوید پس از ۱۰۰ دور آموزش کاملاً ناموفق بود. از طرف دیگر زمان آموزش روش مرجع [۱۳] ده شبانه روز به طول انجامید در حالی‌که آموزش روش پیشنهادی کمتر از یک روز به زمان نیاز داشت. این دستاورد مهم روش پیشنهادی حاصل از کاهش سایز ورودی ترنسفورمر و استخراج ویژگی‌های مناسب توسط شبکه‌ی خودرمز گذار پیچشی است.

یکی از معیارهای رایج برای ارزیابی عملکرد سامانه‌های تشخیص خودکار گفتار نرخ خطای حرف[1] است. این معیار مشابه نرخ خطای کلمه[2] عمل می‌کند، با این تفاوت که به‌جای کلمه، حرف را در نظر می‌گیرد. رابطه (۳) نحوه محاسبه‌ی نرخ خطای حرف را نمایش می‌دهد.

$$CER = (S + D + I) / N \times 100 \quad (۳)$$

پارامتر S تعداد حرف‌هایی است که در متن پیش‌بینی به‌اشتباه جایگزین حرف‌های متن مرجع شده‌اند. پارامتر D تعداد حرف‌هایی است که در متن مرجع وجود دارند، اما در متن پیش‌بینی حذف شده‌اند. پارامتر I تعداد حرف‌های اضافی است که در متن پیش‌بینی وجود دارند، اما در متن مرجع نیستند و پارامتر N تعداد کل حرف‌های متن مرجع است.

برای ارزیابی کارایی روش پیشنهادی و مرجع [۱۳] از دو معیار نرخ خطای حرف و نرخ خطای کلمه استفاده می‌شود. جدول (۷) نرخ خطای حرف و نرخ خطای کلمه روش پیشنهادی و مرجع [۱۳] را بر روی مجموعه‌دادگان پاوید نمایش می‌دهد.

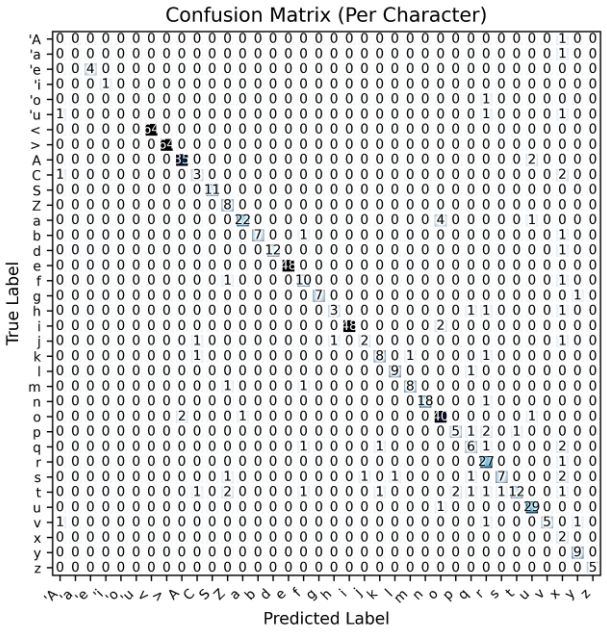
جدول (۷). نرخ خطای حرف و کلمه‌ی روش پیشنهادی و مرجع [۱۳]

	روش پیشنهادی	مرجع [۱۳]
نرخ خطای کلمه (WER)	۰/۲۵	۱
نرخ خطای حرف (CER)	۰/۱۵۷۱	۰/۹۸۴

نتایج ارزیابی‌ها نشان می‌دهد که ترکیب شبکه‌های خود رمزگذار پیچشی با ترنسفورمر در روش پیشنهادی به دلیل ادغام ویژگی‌های محلی و مکانی با ویژگی‌های زمانی می‌تواند عملکرد مطلوبی در تشخیص عبارت گفتاری در ویدئو داشته باشد.

ماتریس درهم‌ریختگی[3] روش پیشنهادی در مرحله‌ی آزمون بر روی دادگان پاوید در شکل (۱۷) نمایش‌داده‌شده است.

[1]. Character Error Rate.
[2]. Word Error Rate.
[3]. Confusion Matrix.



شکل (۱۷). ماتریس درهم ریختگی روش پیشنهادی

محور افقی و عمودی شکل (۱۷) به ترتیب برچسب‌های پیش بینی شده و واقعی را برای هریک از واج‌ها در زبان فارسی نمایش می‌دهد. ماتریس درهم ریختگی شکل (۱۷) رفتار معماری پیشنهادی را در مواجهه با شباهت‌های دیداری واج‌های فارسی نشان می‌دهد. همانطور که در شکل (۱۷) ملاحظه می‌شود، مدل پیشنهادی در شناسایی واج‌های مشابه دیداری موفق بوده است. رابطه‌ی (۴) نحوه‌ی محاسبه‌ی دقت روش پیشنهادی را نشان می‌دهد.

$$Accuracy = \sum_i C_{ii} \Big/ \sum_i \sum_j C_{ij} \quad (۴)$$

پارامتر C ماتریس درهم ریختگی، C_{ii} عناصر روی قطر اصلی (عناصری که به‌درستی طبقه‌بندی شده‌اند)، C_{ij}تمام عناصر ماتریس را نشان می‌دهند. دقت روش پیشنهادی مطابق رابطه‌ی (۴) و شکل (۱۷) بر روی دادگان پاوید برابر با ۸۹ درصد است که در مقایسه با روش مرجع [۱۳] بسیار قابل توجه است.

نتایج حاصل از روش پیشنهادی در مقایسه با روش مرجع [۱۳] تأییدکننده این ادعا است که روش پیشنهادی عملکرد مناسبی در زبان فارسی دارد.

پژوهش حاضر نخستین گام در توسعه سامانه‌های لب‌خوانی اختصاصی برای زبان فارسی به شمار می‌رود که با تشخیص ویزم‌ها و ترکیب آن‌ها لب‌خوانی را در سطح کلمه انجام می‌دهد. پژوهش‌های محدود دیگر در حوزه لب‌خوانی در زبان فارسی از طبقه‌بندی کلمات روی مجموعه‌دادگان اختصاصی استفاده کرده‌اند.

۵. نتیجه‌گیری

در این مقاله یک ساختار ترکیبی جدید مبتنی بر شبکه‌های خود رمزگذار پیچشی و ترنسفورمر برای تشخیص گفتار از تصویر طراحی شد. شبکه‌های خود رمزگذار پیچشی به‌عنوان یک روش یادگیری هوشمند بدون ناظر برای استخراج ویژگی به کار می‌روند. در روش پیشنهادی ابتدا از رمزگذار شبکه خود رمزگذار پیچشی برای استخراج ویژگی‌های مناسب استفاده می‌شود. سپس ویژگی‌های استخراج شده به‌عنوان ورودی به ترنسفورمر منتقل می‌شوند. بخش رمزگذار ترنسفورمر ویژگی‌های مرتبط با بعد زمان را به دست می‌آورد. بخش رمزگشای ترنسفورمر برای طبقه‌بندی و تولید متن عبارت گفتاری در خروجی به کار می‌رود. تولید متن به شناسایی و تفسیر ویژگی‌های بصری استخراج شده و تبدیل آن‌ها به متن متناظر کلمات یا عبارات می‌گویند؛ بنابراین ویژگی‌های استخراج شده به نمایش‌های متنی نگاشت می‌شوند. در این مرحله سامانه کلمات یا عبارات مرتبط را بر اساس ویژگی‌های آموخته شده طبقه‌بندی می‌کند. نوآوری اصلی این مقاله تشخیص حرکات لب در سطح واحدهای واج یا ویزم است. معماری پیشنهادی، امکان تشخیص گفتار پیوسته را فراهم کرده و از قابلیت تعمیم‌پذیری بالاتری نیز برخوردار است. نتایج نشان داد، ایده ارائه شده جهت استفاده از خود رمزگذار نه‌تنها توانست ویژگی‌های مناسب را استخراج کند؛ بلکه باعث کاهش سایز ورودی به ترنسفورمر شد. این کاهش سایز ورودی به مدل ترنسفورمر سبب دستاورد مهم روش پیشنهادی در کاهش زمان آموزش فرایند لب‌خوانی گردید. علاوه بر این کاهش سایز اثرات مطلوبی در کاهش حجم محاسبات و دقت نهایی داشت. به‌عنوان نتیجه دیگر، آزمایش‌ها نشان داد که ساختار شبکه ترنسفورمر پیشنهادی با تنظیمات انجام شده می‌تواند عملکرد مناسبی در زبان فارسی داشته باشد. مقایسه نتایج نرخ خطای حرف و کلمه حاصل از پیاده‌سازی مرجع [۱۳] و روش پیشنهادی بر روی مجموعه‌دادگان پاوید کارایی روش پیشنهادی را در مقایسه با روش دیگر نشان می‌دهد.

به‌عنوان کارهای آینده، قدم‌های زیر می‌تواند دنبال شود: الف) جمع‌آوری داده‌های بیشتر ب) تعیین ساختارهایی که از شبکه آموزش‌دیده این مقاله بتوانند برای تشخیص جمله با لب‌خوانی اقدام کنند. ج) بهینه‌سازی شبکه از لحاظ حجم محاسباتی برای اینکه بتواند روی یک گوشی تلفن همراه اجرا شود.

قدردانی:

اجرای کدها و آزمایش‌های این تحقیق با مساعدت شرکت هوش مصنوعی اهورا بر روی پلتفرم پردازش گرافیکی که آن شرکت در اختیار آزمایشگاه هوش محاسباتی و رباتیک دانشگاه یزد قرار داده است، انجام شده است. بدین‌وسیله از آن شرکت محترم تشکر و قدردانی می‌شود. همچنین از جناب آقای دکتر یاسر شکفته عضو محترم هیئت‌علمی دانشگاه شهید بهشتی تهران و سرکار خانم مهندس مهسا هدایتی پور برای در اختیار قراردادن مجموعه‌دادگان پاوید کمال تشکر به عمل می‌آید.

۵. مراجع

[1] S. Jeon, A. Elsharkawy, and M. S. Kim, “Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition,” Sensors, vol. 22, no. 1, Art. no. 1, Jan. 2022. https://doi.org/10.3390/s22010072.

[2] A. Bastanfard, M. Aghaahmadi, A. A. kelishami, M. Fazel, and M. Moghadam, “Persian Viseme Classification for Developing Visual Speech Training Application,” in Advances in Multimedia Information Processing - PCM 2009, P. Muneesawang, F. Wu, I. Kumazawa, A. Roeksabutr, M. Liao, and X. Tang, Eds., Berlin, Heidelberg: Springer, 2009, pp. 1080–1085. https://doi.org/10.1007/978-3-642-10467-1_104.

[3] Y. Pei and H. Zha, “Stylized synthesis of facial speech motions,” Computer Animation and Virtual Worlds, vol. 18, no. 4–5, pp. 517–526, Sep. 2007. https://doi.org/10.1002/cav.186.

[4] T. Zhang, L. He, X. Li, and G. Feng, “Efficient End-to-End Sentence-Level Lipreading with Temporal Convolutional Networks,” Applied Sciences, vol. 11, no. 15, Art. no. 15, Jan. 2021. https://doi.org/ 10.3390/app11156975.

[5] H. Mcgurk and J. Macdonald, “Hearing lips and seeing voices,” Nature, vol. 264, no. 5588, pp. 746–748, Dec. 1976. https://doi.org/10.1038/264746a0.

[6] N. Akhter and A. Chakrabarty, “A Survey-based Study on Lip Segmentation Techniques for Lip Reading Applications,” 2016. Accessed: Jun. 22, 2025. [Online]. Available: https://www.semanticscholar.org/paper/A-Survey-based-Study-on-Lip-Segmentation-Techniques-Akhter-Chakrabarty/457a8efb4d9b71770ccddf0c74b88179a097e500

[7] D. Yu, “The application of manifold based visual speech units for visual speech recognition,” PhD Thesis, Dublin City University, 2008. Accessed: Jun. 22, 2025. [Online]. Available: https://doras.dcu.ie/598/

[8] W. C. Yau, D. K. Kumar, and S. P. Arjunan, “Visual recognition of speech consonants using facial movement features,” Integr. Comput.-Aided Eng., vol. 14, no. 1, pp. 49–61, Jan. 2007.

[9] E. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, “An improved automatic lipreading system to enhance speech recognition,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, in CHI ’88. New York, NY, USA: Association for Computing Machinery, May 1988, pp. 19–25. https://doi.org/10.1145/57167.57170.

[10] A. Maroosi, E. Zabbah, and H. Ataei Khabbaz, “Network Intrusion Detection using a combination of artificial neural networks in a hierarchical manner,” Electronic and Cyber Defense, vol. 8, no. 1, pp. 89–99, 2020. (In Persian)

[11] A. Dolat Khah, M. Asadpour, R. Hashempour, and behnam dorostkar, “Violent behavior detection in surveillance cameras using convolutional and memory neural networks,” Electronic and Cyber Defense, vol. 12, no. 4, p., 2025. (In Persian)

[12] S. Hourali, F. Hourali, and A. Pakzad, “Cyber Threat Information Extraction using Deep Learning and Knowledge Representation,” Electronic and Cyber Defense, vol. 13, no. 2, p., 2025. (In Persian)

[13] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, “LipNet: End-to-End Sentence-level Lipreading,” Dec. 16, 2016, arXiv: arXiv:1611.01599. https://doi.org/ 10.48550/arXiv.1611.01599.

[14] K. Xu, D. Li, N. Cassimatis, and X. Wang, “LCANet: End-to-End Lipreading with Cascaded Attention-CTC,” Mar. 13, 2018, arXiv: arXiv:1803.04988. https://doi.org/ 10.48550/arXiv.1803.04988.

[15] M. A. Abrar, A. N. M. N. Islam, M. M. Hassan, M. T. Islam, C. Shahnaz, and S. A. Fattah, “Deep Lip Reading-A Deep Learning Based Lip-Reading Software for the Hearing Impaired,” in 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC) (47129), Nov. 2019, pp. 40–44. https://doi.org/10.1109/R10-HTC47129.2019.9042439.

[16] D. Parekh, A. Gupta, S. Chhatpar, A. Y. Kumar, and M. Kulkarni, “Lip Reading Using Convolutional Auto Encoders as Feature Extractor,” May 31, 2018, arXiv: arXiv:1805.12371. https://doi.org/10.48550/arXiv.1805.12371.

[17] 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), May 2020, pp. 4407–4411. https://doi.org/10.1109/ICASSP40776.2020.9053535.

[18] H. Huang, Ch. Song, J. Ting, T. Tian, Ch. Hong, Zh. Di, D. Gao, “A Novel Machine Lip Reading Model,” Procedia Computer Science, vol. 199, pp. 1432–1437, Jan. 2022. https://doi.org/10.1016/j.procs.2022.01.181.

[19] H. Wang, G. Pu, and T. Chen, “A Lip Reading Method Based on 3D Convolutional Vision Transformer,” IEEE Access, vol. 10, pp. 77205–77212, 2022. https://doi.org/10.1109/ACCESS.2022.3193231.

[20] A. Gholipour, H. Mohammadzade, A. Ghadami, and A. Taheri, “Automatic Lip Reading of Persian Words by a Robotic System Using Deep Learning Algorithms,” Iran J Sci Technol Trans Electr Eng, vol. 48, no. 4, pp. 1519–1538, Dec. 2024. https://doi.org/10.1007/s40998-024-00756-4.

[21] M. Hedayatipour, Y. Shekofteh, and M. E. Moghaddam, “PAVID-CV s: Persian Audio-Visual Database of CV syllables,” in 2021 29th Iranian Conference on Electrical Engineering (ICEE), May 2021, pp. 470–473. https://doi.org/10.1109/ICEE52715.2021.9544268.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, “Attention Is All You Need,” Aug. 02, 2023, arXiv: arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762.

[23] K. Paleček, “Extraction of Features for Lip-reading Using Autoencoders,” in Speech and Computer, A. Ronzhin, R. Potapova, and V. Delic, Eds., Cham: Springer International Publishing, 2014, pp. 209–216. https://doi.org/10.1007/978-3-319-11581-8_26.

[24] D. E. King, davisking/dlib-models. (Oct. 29, 2025). C++. Accessed: Oct. 31, 2025. [Online]. Available: https://github.com/davisking/dlib-models

[25] “Facial landmarks with dlib, OpenCV, and Python - PyImageSearch.” Accessed: Oct. 31, 2025. [Online]. Available: https://pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/