خوشه‌بندی ویسیلاب‌های دو آوایی زبان فارسی در کاربرد لب‌خوانی

مهسا هدایتی‌پور * [۱]، یاسر شکفته[۲]، محسن ابراهیمی مقدم [۳]
دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی
1.m.hedayatipour@mail.sbu.ac.ir, 2.y_shekofteh@sbu.ac.ir, 3.m_moghadam@sbu.ac.ir

چکیده - لب‌خوانی، فرآیند بازشناسی دیداری گفتار است. در این فرآیند تشخیص واحدهای کوچک‌تر گفتاری می‌تواند مبنای تشخیص واحدهای بزرگ‌تر گفتاری باشد. یکی از چالش‌های این فرآیند، مشابه بودن تصاویر برخی از واحدهای گفتاری به علت جایگاه تولید یکسان در اندام‌های گفتاری است. بدین منظور در فرآیند لب‌خوانی، خوشه‌بندی تصاویر اجزای گفتار و ساختن جداول نگاشت از شکل شنیداری اجزای گفتار به شکل دیداری اجزای گفتار از اهمیت اساسی برخوردار است. از آنجا که بازشناسی دیداری هجاهای دو آوایی گام نوینی در لب‌خوانی زبان فارسی است، در این پژوهش سعی شده است برای بازشناسی دیداری بهینه هجاهای دو آوایی زبان فارسی جداول نگاشت مناسبی بر مبنای روش‌های یادگیری ماشین و یا دانش افراد خبره فراهم گردد. بیشینه دقت شناسایی برای نگاشت ۹ گانه از هجاهای دو آوایی به ویسیلاب به مقدار ۶۱/۸۱ درصد حاصل شده است. با مدل‌سازی این ویسیلاب‌ها توسط مدل مخفی مارکوف و تزریق مناسب اطلاعات مدل زبانی هجاها، دقت ۴۱/۱۸ درصد در شناسایی ۳۰ کلمه فارسی حاصل شده است.

کلید واژه- بازشناسی دیداری گفتار، لب‌خوانی، نگاشت هجا به ویسیلاب، هجاهای دو آوایی

۱- مقدمه

گفتار عبارت است از رشته‌های آوایی که برطبق الگوهای خاص سازمان یافته است. استفاده از گفتار یکی از مؤثرترین روش‌های ارتباطی بین انسان‌ها است. بنابراین اختلال در گفتار می‌تواند شخص را در ارتباطات خود دچار مشکل سازد و منجر به بروز مشکلات متعددی گردد. این اختلالات می‌تواند ناشی از ناشنوایی یا کم‌شنوایی افراد، عدم توانایی صحبت‌کردن به‌دلیل از دست‌دادن حنجره و یا حتی قرارگرفتن شخص در محیط‌های پر سروصدا و مواردی از این قبیل باشد. این موارد، افراد را در ارتباطات روزمره خود در درک کلمات و جملاتی که مخاطب بیان می‌کند، دچار مشکل می‌سازد.

برخی از اندام‌های بدن در تولید آواهای زبان به کارگرفته می‌شوند. از جمله این اندام‌ها می‌توان به شش‌ها، نای، حنجره، گلو، حلق، حفره بینی، کام، زبان و لب اشاره‌کرد که حالت تغییرپذیری شکل و حجم دهان عامل تعیین‌کننده بسیاری از مشخصه‌های آوایی صداهای گفتار می‌باشد. این تغییرات در شکل لب‌ها نیز نمودار می‌شود.

لب‌خوانی(Lip-Reading) یا فرآیند بازشناسی تصویری گفتار، روش فهمیدن گفتار بوسیله تفسیر حرکات لب گوینده است. در این فرآیند با استخراج ویژگی‌های تصاویر لب و استفاده از آن‌ها در روش‌های شناسایی الگو و یادگیری ماشین، حرکات لب به صورت گفتار تفسیر می‌شود. استخراج این ویژگی‌ها با استفاده از روش‌های پردازش تصویر و یا به صورت خودکار با روش‌های یادگیری ماشین انجام می‌شود.

سیستم لب‌خوانی می‌تواند برای آموزش افرادی که دارای نقص شنوایی هستند و همچنین افرادی که با آن‌ها در ارتباط هستند، مفید باشد. یک سیستم لب‌خوانی می‌تواند به آژانس‌های اطلاعاتی کمک کند تا با استفاده از دوربین از محتویات 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[۱].

فعالیت‌هایی در این زمینه در زبان‌های مختلف صورت گرفته‌است. از جمله کارهایی که در زبان فارسی صورت گرفته می‌توان به تشخیص اعداد فارسی[۲]، تشخیص برخی از حروف الفبای فارسی[۳] و تشخیص تعداد محدودی از کلمات فارسی[۴] و [۵] اشاره‌کرد. در صورتیکه از مدل‌های مبتنی بر کلمات برای تشخیص کلمات استفاده شود، به ساخت مدل برای هر کلمه نیاز است که در این صورت برای پوشش تعداد زیادی از کلمات، تعداد مدل‌های مورد نیاز زیاد خواهد شد. اما در صورتیکه از واحدهای گفتاری سازنده کلمات استفاده شود، تنها به مدل‌سازی آن‌ها نیاز

خواهد بود که با توجه به اشتراک این واحدها در کلمات مختلف زبان، از تعداد مدل‌های مورد نیاز بسیار کاسته خواهد شد [۶].

در پژوهش‌هایی که تاکنون برای استفاده از واحدهای زبانی در لب‌خوانی زبان فارسی انجام شده[۷]، شناسایی ویزم‌ها(viseme) مورد مطالعه قرار گرفته است. از آنجا که براساس پژوهش‌های انجام شده در [۶]، تشخیص عبارات طولانی از عبارات کوتاه راحت‌تر و از صحت بیشتری برخوردار است، به همین دلیل تصمیم گرفته شد که تشخیص ویسیلاب‌ها(visyllable) به‌عنوان اجزای زبان جایگزین تشخیص ویزم‌ها شود. بدین منظور لازم است تا نگاشت جدیدی از هجا به ویسیلاب جایگزین نگاشت‌های مرسوم واج به ویزم شود. در بررسی‌های ما مشخص شد که ۲۹۴۶ کلمه از کلمات پیکره بی‌جن خان[۸] تنها از ترکیب هجاهای CV تشکیل شده‌اند. از طرفی دیگر، در بررسی هجاهای فارسی مشخص شده است که ۵۷/۷۵ درصد از کل هجاهای زبان فارسی ساخت هجایی CV دارند و بسامد هجاهای CVC و CVCC به ترتیب ۳۸/۰۹ درصد و ۴/۱۶ درصد است و در بسیاری موارد، هجاهای دیگر نیز هنگام گفتار به CV تبدیل می‌شوند. همچنین در بسیاری از مواقع هنگام تلفظ کلمات زبان فارسی، همخوان(consonant) پایانی از خوشه دو همخوانی حذف می‌شود و CVCC به CVC تبدیل می‌شود و با درج واکه(vowel)، هجاسازی مجدد صورت می‌گیرد و CVC نیز به CV تبدیل می‌شود، مثلا kargar به karegar تبدیل می‌شود. این شواهد همه حکایت از این موضوع دارد که هجای CV هجای مطلوب زبان فارسی است و در فرآیندهای واجی، هجاهای دیگر به آن تبدیل یا نزدیک می‌شود[۹]. با توجه به واقعیت‌های آماری مذکور و اهمیت زیاد هجاهای CV در زبان فارسی، تصمیم گرفته شد که خوشه‌بندی ویسیلاب‌های دو آوایی و تعیین جداول نگاشت هجاهای CV (تعداد آنها ۱۳۸ مورد در زبان فارسی است) به ویسیلاب به‌عنوان موضوع این پژوهش در نظر گرفته شود تا در کاربردهایی مانند بازشناسی دیداری گفتار مورد استفاده قرار گیرد.

ساختار این مقاله به این شرح است: در بخش دوم مقاله معرفی اجمالی بر روی واحدهای گفتاری و گفتار تصویری خواهیم داشت. در بخش سوم برخی ازکارهای مرتبط با جزئیات بیشتر آورده شده است. در بخش چهارم روش‌های پیشنهادی برای خوشه‌بندی ویسیلاب‌های دو آوایی و جداول نگاشت متناظر با هر روش آورده شده است و در بخش آخر جمع‌بندی و نتیجه‌گیری مقاله ارائه شده است.

۲- معرفی واحدهای گفتاری و گفتار تصویری

به صداهای متمایزی که توسط اندام‌های گفتاری انسان تولید می‌شوند، آوا می‌گویند. واج‌ها، آواهایی هستند که به‌عنوان واحدهای انتزاعی زبان، به خودی خود معنا ندارند ولی با تغییر در یک واژه منجر به تغییر معنا می‌شوند.

واج‌ها بطور کلی دو دسته‌اند: صامت (همخوان) و مصوت (واکه). همخوان‌ها که با علامت C نشان داده می‌شوند؛ صداهایی هستند که موقع بیان آن‌ها وضعیت دهان، زبان و لب‌ها لزوماً یکنواخت نیست و معمولاً مانع و یا انسدادی بر سر راه عبور هوای خروجی ایجاد می‌شود. واکه‌ها که با علامت V نشان داده می‌شوند؛ صداهایی هستند که موقع بیان آن‌ها وضعیت دهان، زبان و لب‌ها کاملاً یکنواخت است و هیچ مانعی بر سر راه هوای خروجی وجود ندارد[۱۰]. در زبان فارسی فقط همخوان‌ها می‌توانند در آغاز واژه قرار گیرند و این نکته عامل مهمی برای شناخت هجاهای فارسی از توالی واج‌ها با برچسب واکه و همخوان است. همچنین واکه‌ها به‌عنوان هسته هجاها شناخته می‌شوند و هیچگاه در ابتدای کلمه واقع نمی‌شوند[۱۱]. بطور کلی ۲۹ واج در زبان فارسی وجود دارند که شامل ۶ واکه و ۲۳ همخوان می‌باشند [۱۰].

هجا در زبان فارسی عبارت از یک رشته آوایی پیوسته است که از یک واکه و یک تا سه همخوان تشکیل می‌یابد. منظور از رشته آوایی پیوسته آن است که اجزای سازنده هجا طی یک فرآیند تولیدی بدون مکث تولید می‌گردند. این هجاها از نظر ساختمان دارای سه ترکیب زنجیری CV، CVC، CVCC هستند [۹]. هجاها نیز مانند واج‌ها، بر روی حرکات و شکل ظاهری لب اثر می‌گذارند. با ترکیب ۲۳ همخوان و ۶ واکه موجود در زبان فارسی، ۱۳۸ هجای دو آوایی موجود در زبان فارسی ساخته می‌شود.

متناظر با واج‌ها در گفتار، ویزم‌ها به‌عنوان واحدهای کوچک در حوزه دیداری گفتار محسوب می‌شوند. ویزم‌ها شکل ظاهری لب هنگام بیان واج و کوچک‌ترین واحد دیداری گفتار هستند[۱۲]. لب‌ها، دندان‌ها و زبان به‌عنوان نشانه‌های اولیه و حالت فک، چانه و بینی نیز به‌عنوان نشانه‌های بعدی در تشخیص دیداری گفتار محسوب می‌شوند. باید به این نکته اشاره‌کرد که در بازشناسی دیداری گفتار معمولاً فقط از اطلاعات حرکتی لب‌ها استفاده می‌شود. شکل ظاهری لب هنگام بیان هجاها، ویسیلاب نامیده می‌شود و به‌عنوان معادل دیداری هجا در نظر گرفته می‌شود [۱۳]. در بازشناسی دیداری گفتار می‌توان از تشخیص این ویسیلاب‌ها در ساختار کلمه استفاده کرد.

از آن‌جا که شکل ظاهری موقعیت لب‌ها برای برخی از واج‌ها شبیه به هم به‌نظر می‌رسند، لذا ویزم‌های زبان‌های مختلف مانند انگلیسی و یا فارسی را می‌توان بر اساس میزان شباهت تصویری دسته‌بندی‌کرد. این دسته‌بندی‌ها را در جداولی موسوم به جدول

نگاشت واج به ویزم مشخص می‌کنند [۷].

در واقع دسته‌بندی‌های مختلفی می‌توان برای واج‌های متناظر با یک ویزم یافت و از آنجا که تاکنون همه مدل‌های زبانی به صورت واج‌ها تعریف شده‌اند، چالش مهم انتخاب نگاشت مناسب از واج‌ها به ویزم‌ها است. در بررسی‌های [۱۴]، نشان داده شده است که برای هر گوینده ممکن است نگاشت متفاوتی وجود داشته باشد. در مقایسه جدول‌های نگاشت مختلف، دلایل زیر برای اختلاف بین نگاشت‌های واج به ویزم ذکر شده است: ۱) تفاوت شخصیت و هویت گویندگان که در طرز بیان گفتارشان ظاهر می‌شود. ۲) توانایی لب‌خوانی بین افراد مختلف متفاوت است و شخصی که بیشتر تمرین کرده است ویزم‌ها را بهتر تشخیص می‌دهد، به همین دلیل ممکن است جداول نگاشت متفاوتی به‌دست آید. ۳) نوع محتوا در اینکه همخوان‌ها چگونه روی لب ظاهر شوند تأثیر دارد. به این معنی که انتخاب مناسب محتوا، تمایز بین واج‌های غیرقابل تشخیص را آسان‌تر می‌کند. ۴) معیارهای خوشه‌بندی که برای گروه‌بندی واج‌ها در نظر گرفته شده است در محتوای خوشه‌ها تأثیر دارد. مثلا اگر آستانه شباهت برای قرار گرفتن در یک خوشه کم یا زیاد شود تعداد واج‌های همانند در یک خوشه زیاد و کم می‌شود. مجموعه این دلایل نشان می‌دهد که نگاشت‌های واج به ویزم وابستگی زیادی به گوینده دارند و انتخاب نگاشت مناسب یک چالش مهم در لب‌خوانی است. بدین صورت که نگاشتی که برای یک نوع از گفتار مناسب است، برای نوع دیگری مناسب نیست.

۳- مروری بر کارهای مرتبط

در زبان انگلیسی تحقیقات زیادی برای نگاشت واج به ویزم انجام شده و جداول نگاشت مختلفی به‌دست آمده است که در [۱۴] بررسی شده‌اند. در زبان فارسی تاکنون یک نگاشت واج به ویزم انجام شده است [۷]. بدین منظور ابتدا تصاویری که نماینده همخوان‌ها در هجاهای زبان فارسی بوده را از پایگاه داده AVA [۱۵] انتخاب کرده‌اند. سپس از روش تحلیل ویژه جهت استخراج ویژگی و خوشه‌بندی واج‌ها استفاده شده است. بدین صورت 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 است و به‌عنوان بردارهای ویژگی واج‌ها استفاده شده است. برای خوشه‌بندی، بردارهایی که فواصل اقلیدسی آن‌ها با یکدیگر از آستانه مشخصی کمتر بوده است در یک خوشه قرار گرفته‌اند. با اعمال این روش، ویزم‌های مرتبط با همخوان‌های زبان فارسی به ۷ خوشه بخش‌بندی شده‌اند. نتایج این خوشه‌بندی با استفاده از الفبای آوانگاری بین‌المللی(IPA) در جدول ۱ مشاهده می‌شود:

جدول ۱: خوشه‌بندی ویزم‌های فارسی [۷]

b, p, m	۱	۵	G, n, h, x, ʔ
f, v	۲	۶	r, l
d, t	۳	۷	tʃ, dʒ, s, z, ʃ, ʒ
k, g, j	۴		

۴- روش پیشنهادی

تاکنون برای بازشناسی دیداری هجاهای دو آوایی زبان فارسی، پژوهشی صورت نگرفته است. بازشناسی دیداری هجاها نیازمند ساخت جداول نگاشت هجا به ویسیلاب است. مشابه با نگاشت‌های واج به ویزم، به دلیل هم‌آوا بودن برخی از هجاها، باید هجاهای هم‌آوا را در یک خوشه قرار داد. برای این منظور ۴ روش در این بخش پیشنهاد و بررسی شده‌اند.

به منظور بررسی روش‌های پیشنهادی، ابتدا دادگان مناسبی فراهم شد. از تعداد ۴۰ نفر زن و مرد خواسته شد 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 شد. هر ویدیو با هجاهای مربوطه برچسب‌گذاری شد تا در بررسی روش‌های پیشنهادی استفاده شود.

در ابتدا ویدیوهای مربوط به هجاها را به دنباله‌ای از قاب‌ها تبدیل و سپس با استفاده از الگوریتم ویولا-جونز[۱۶]، ناحیه لب در این قاب‌ها تعیین شده است. نمونه‌هایی از تصاویر ناحیه لب در شکل ۱ آمده است:

شکل ۱: تصاویر ناحیه لب

گام بعدی؛ مشخص‌کردن و جداسازی قسمت لب در تصویر است. برای نیل به این مقصود از ترکیب مؤلفه‌های a و b از فضای رنگی Lab، مؤلفه H از فضای رنگی Psudo-Hue و مؤلفه U از فضای رنگی LUX به‌عنوان ورودی الگوریتم SFCM[۱۷] استفاده شده است تا بتوان پیکسل‌های تصاویر را به درستی به دو دسته لب

و غیرلب خوشه‌بندی کرد. خروجی این الگوریتم، تصویری سیاه و سفید از ناحیه غیرلب و لب خواهد بود.

این تصاویر شامل نواحی سفید به هم پیوسته هستند که لکه نامیده می‌شوند. نمونه‌ای از این تصاویر در شکل ۲ مشاهده می‌شود: سپس بزرگترین لکه از میان چندین لکه‌ی موجود در تصویر به‌عنوان بخش دهان برگزیده می‌شود. نمونه‌هایی از تصاویر قطعه‌بندی شده خروجی در شکل ۳ آمده است:

شکل ۲: تصاویر قطعه‌بندی شده ناحیه لب

شکل ۳: تصاویر بزرگ‌ترین لکه

پس از جداسازی ناحیه لب، نوبت به استخراج ویژگی‌های موردنظر می‌رسد. سه ویژگی ارتفاع، عرض و تعداد پیکسل‌های لب از ناحیه لب استخراج شده است.

جهت استخراج ویژگی‌های دیگر، الگوریتم LBP[۱۸] را روی دنباله قاب‌های موردنظر اعمال کرده و به این صورت برای هر قاب یک بردار ویژگی ۵۹×۱ ایجاد می‌شود. از ترکیب این بردار ویژگی با ویژگی‌های سه‌گانه فوق، برای هر قاب بردار ویژگی ۶۲ بعدی به‌دست می‌آید و برای هر دنباله قاب ویدیو، ماتریسی به ابعاد m×۶۲ که m تعداد قاب‌های آن ویدیو است، ایجاد می‌شود. سپس، دنباله ویژگی‌های به‌دست آمده برای هر دنباله قاب با روش zscore نرمال می‌شوند و تفاضل بین ویژگی ۶۲ بعدی هرقاب با قاب قبلی را محاسبه کرده و به‌عنوان یک ویژگی ۶۲ بعدی به بردار ویژگی‌های آن قاب اضافه می‌شود. نتیجه حاصل یک ماتریس ویژگی m×۱۲۴ برای هر دنباله قاب است که به‌عنوان ویژگی‌های پایه در تحلیل‌های این پژوهش استفاده شده است.

۴-۱- ساخت جدول نگاشت

برای ساخت جداول نگاشت هجاهای دو آوایی به ویسیلاب دو آوایی، از بخشی از دادگان فراهم شده، به صورتی که در هر روش بیان گردیده، استفاده شده است. برای بررسی دقت جداول ساخته شده در همه روش‌ها، دادگان و مدل یکسان مطابق با آنچه در روش پایه ذکر شده است به کار برده شده است.

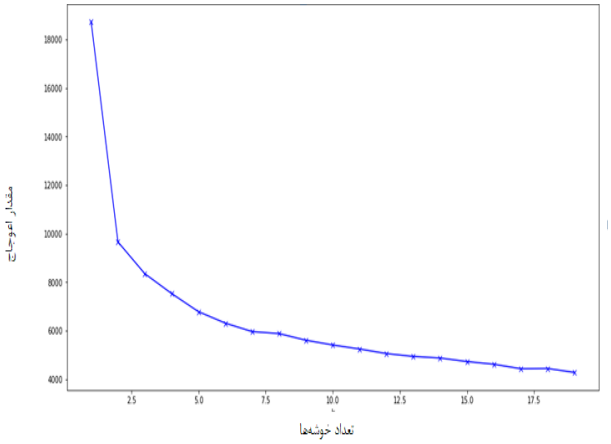
روش پایه- روش مبتنی بر نگاشت واج به ویزم: با استفاده از نتایج خوشه‌بندی ویزم‌های فارسی در [۷] و با ترکیب خوشه‌های ۷ گانه همخوان‌ها با ۶ واکه، جدول نگاشت هجا به ویسیلاب با ۴۲ خوشه ایجاد شد. سپس بردار ویژگی‌های نمونه‌ها با برچسب‌زنی مجدد ۴۲ گانه با مدل مخفی مارکوف با ۹ حالت و دو مدل مخلوط گاوسی مورد آموزش و آزمون قرار گرفته است. تعداد ۷۱۲۸ نمونه از ۲۵ نفر برای آموزش و ۱۸۰۷ نمونه از ۶ نفر برای آزمون استفاده شده است. در این روش دقت ۱۶/۰۳ درصد در تشخیص ویسیلاب‌ها حاصل شده است. نتیجه به‌دست آمده در این روش به‌عنوان مبنا جهت مقایسه با سایر جداول نگاشت هجا به ویسیلاب که در این پژوهش حاصل شده، استفاده شده است.

روش اول- روش مبتنی بر ماتریس سردرگمی غیر وابسته به گوینده: در این روش بر مبنای شباهت مدل‌های آموزش یافته متناظر با هجاها اقدام می‌شود. به این‌صورت که فرض می‌شود مدل‌های هجایی مشابه باید در یک خوشه قرار گیرند. در ابتدا هریک از ۱۳۸ هجا، معادل یک خوشه در نظر گرفته شده و مدل‌های متناظر با آن‌ها مورد آموزش و آزمون قرار گرفته‌اند. در تمام مراحل از مدل و داده‌های ذکر شده در روش پایه استفاده شده است. سپس از نتایج این آزمون، ماتریس سردرگمی(confusion matrix) ساخته شده و مقادیر هریک از سلول‌های ماتریس بر تعداد نمونه‌های موجود از کلاس هدف تقسیم شده است. سطر و ستون سلولی که بیشترین مقدار حاصل از این نرمال‌سازی در آن قرار داشته باشد بیانگر دو خوشه‌ای است که باید در هم ادغام شوند. مجدداً این مراحل با برچسب زنی مجدد برای نمونه‌ها، براساس خوشه‌های جدید به‌دست آمده، ادامه می‌یابد تا زمانی که هیچ خوشه تک هجایی باقی نماند. این امر در تعداد ۹ خوشه حاصل شده است. نگاشت حاصل در جدول ۲ مشاهده می‌شود. در مرحله نهایی، مقدار دقت شناسایی ویسیلاب‌ها ۴۰/۱۲ درصد بوده است.

روش دوم- روش مبتنی بر ماتریس سردرگمی وابسته به ۳ نفر: به دلیل آنکه وابستگی به گوینده یکی از چالش‌های جداول نگاشت می‌باشد، در آزمایشی دیگر از نمونه‌های ۳ نفر استفاده شد و همه مراحل مطابق آنچه در روش اول ذکر شد، انجام شد. تعداد خوشه‌های حاصل ۴ شد. نگاشت حاصل در جدول ۲ مشاهده می‌شود. در مرحله نهایی، مقدار دقت شناسایی ویسیلاب‌ها ۵۶/۸۸ درصد شد.

روش سوم- روش مبتنی بر الگوریتم خوشه‌بندی: در این روش از الگوریتم خوشه‌بندی kmeans استفاده شده است. به این ترتیب که برای نمونه داده‌های یک نفر به ازای تعداد ۱ تا ۲۰ خوشه متفاوت، فرآیند خوشه‌بندی هجاها انجام شده است. سپس میانگین

کمترین فاصله داده‌ها از مرکز خوشه‌ها به‌عنوان مقدار اعوجاج محاسبه می‌شود. نمودار اعوجاج خوشه‌ها در شکل ۴ مشاهده می‌شود. با استفاده از روش elbow، تعداد خوشه بهینه، ۹ به‌دست آمده است. نگاشت حاصل در جدول ۲ مشاهده می‌شود. سپس با استفاده از مدل و نمونه‌های ذکر شده در روش پایه و برچسب‌زنی نمونه‌ها طبق جدول نگاشت به‌دست آمده، دقت شناسایی ویسیلاب‌ها در این روش ۱۸/۱۵ درصد شد.



شکل ۴: نمودار اعوجاج خوشه‌ها در روش سوم

روش چهارم- روش مبتنی بر دانش انسانی: در این روش از دانش افراد خبره برای خوشه‌بندی ویسیلاب‌های دو آوایی استفاده شده است. برای این منظور در آزمایشی از فردی کاملاً ناشنوا برای تشخیص هجاها استفاده شد. بدین منظور نمونه‌های دو تایی از ۱۳۸ هجا که توسط یک نفر بیان شده بود به تعداد ۲۷۶ نمونه بود به صورت تصادفی و نامرتب به فرد مذکور نشان داده شد و تشخیص او ثبت شد. سپس عنوان واقعی هر هجا و عنوانی که توسط فرد ناشنوا تشخیص داده شده بود در یک خوشه قرار گرفتند. از این روش، تعداد خوشه‌های جدول نگاشت به ۹ رسید. نگاشت حاصل در جدول ۲ مشاهده می‌شود. سپس با استفاده از مدل و نمونه‌های ذکر شده در روش پایه و برچسب زنی نمونه‌ها طبق جدول نگاشت به‌دست آمده، دقت شناسایی ویسیلاب‌ها در این روش ۶۱/۸۱ درصد شد.

۵- استفاده از نگاشت هجا به ویسیلاب برای تشخیص کلمه

همانطور که ذکر شد هدف از ایجاد جداول نگاشت هجا به ویسیلاب، استفاده از آن برای ساخت مدل‌های دقیق‌تر تشخیص هجا و نهایتاً استفاده از این مدل‌ها به منظور تشخیص کلمه است. بدین منظور نمونه های ۳۰ کلمه از دادگان ۳ نفر انتخاب شده است. این کلمات با استفاده از جدول نگاشت روش چهارم برچسب زده شده‌اند. از چهار نمونه هر نفر، یک نمونه برای آزمون و سه نمونه برای آموزش انتخاب شده است. در فرآیند آموزش و آزمون از ابزار HTK و توابع آن استفاده شده است. آموزش با مدل مخفی مارکوف ۷ حالته با دو مدل مخلوط گاوسی انجام شده است. در مرحله آزمون از مدل زبانی ۲-گرم مبتنی بر ۳۰ کلمه در رمزگشای(decoder) HVite استفاده شده است و شبکه‌بندهای مربوطه نیز ایجاد شده‌اند. دقت این ۳۰ کلمه براساس ۹ کلاس (بر اساس خوشه‌بندی روش چهارم) ۲۷/۰۶ درصد حاصل شده است. سپس با استفاده از تابع HLRescore و محاسبه مجدد شبکه‌بندها و با تزریق اطلاعات مدل زبانی ۳-گرم حاوی اطلاعات توالی هجاها، به دقت ۴۱/۱۸ درصد براساس ۱۳۸ کلاس و در حالت وابسته به گوینده دست یافته‌ایم.

جدول ۲: نگاشت‌های هجا به ویسیلاب براساس نمادهای IPA

روش	خوشه‌های ویسیلاب و هجاهای درون هر خوشه
روش اول	{po ،tɑ ،ho ،bu ،ʒu ،du ،ʔo ،pu ،tʃu ،xo ،vɑ ،ʒɑ، zu ،nu ،ʔu ،qo ،gu ،ʃe ،lo ،ʒo ،ko ،ʧo ،to ،se ،ʧi، bo ،ru ،dʒu ،lu ،so ،su ،ju ،dʒo ،hu ،tu ،xu ،ku، {vu ،fu ،ʧɑ ،mu ،zo ،no ،go ،ro ،jo ،do، hi ،qe ،xe ،je ،ni ،ʔe ،te ،la ،nɑ ،ʔɑ ،hɑ ،ne ،vi}، {kɑ ،ri ،gɑ ،qɑ ،le ،he ،li ،ge ،qɑ ،hɑ ،ʔi ،de، dʒɑ ،ʃo ،ʃɑ ،ki ،qu ،qi ،ti ،di ،ji ،gi ،si ،zi ،ʃu ،fo}، {dʒe ،rɑ، {mɑ ،sɑ ،xɑ ، mo}، pi ،ba ،pe ،va ،fe ،fɑ ،fa ،dɑ ،pɑ ،bɑ ،ra ،zɑ ، da}، tʃe ،na ،zɑ ،re ،pa ،ve ،ma ،vo ،bi ،me ،be ،fi ،mi، {sɑ ،tɑ ،kɑ ،ke ،ze، {gɑ ،jɑ ،jɑ ،lɑ}، {ʒe ،dʒi ،ʒi ،ʃi}، {ʃɑ ،dʒɑ ،ʒɑ ،tʃɑ}، {xɑ ،xi ،ʔɑ}
روش دوم	so ،dʒo ،no ،lo ،ho ،bo ،ʔo ،mu ،do ،po ،hɑ ،fɑ}، nu ،fu ،to ،ʒo ،tu ،ʃo ،ko ،jo ،go ،xo ،ro ،zo ،qo، ʔu ،ru ،ku ،lu ،pu ،tʃo ،ʒu ،tʃu ،ʃu ،su ،dʒu ،bu، lɑ ،bɑ ،zɑ ،sɑ ،nɑ ،xɑ ،qɑ ،vo ،zu ،du ،qu ،hu، vu ،gu ،gɑ ،kɑ ،xu ،ʔɑ ،jɑ ،dʒɑ ،ʃɑ ،ʒɑ ،rɑ ،dɑ، {ju ،mɑ ،vɑ ،pɑ ،fo، dʒɑ ،ka ،na ،ba ،sa ،ʔa ،la ،qa ،ha ،ʃi ،tʃi ،dʒi}، ni ،ra ،ʃɑ ،le ،pa ،ʒe ،ʃe ،tʃe ،tʃɑ ،ja ،ʒɑ ،tʃɑ ،dʒe، li ،za ،ʒi ،gɑ ،ti ،tɑ ،ke ،xɑ ،qe ،ne ،tɑ ،he ،xe ،pe، {ʔe، ri ،ji ،ʔi ،pi ،bi ،ge ،qi ،di ،ki ،re ،xi ،hi ،mo ،zi}، {mi ،ze ،je ،se ،da ،gi ،de ،si ،te، {ma ،be ،me ،ve ،vi ،fi ،va ،fe ،fa}
روش سوم	{ʃi ،ji ،za ،ze ،ja ،fi ،la ،di ،tʃu ،tʃi ،mi ،ma ،li ،ti}، {dʒe ،dʒɑ ،sɑ ،ʃu، mo ،ʔɑ ،bi ،be ،pi ،pe ،pa ،ka ،mɑ ،te ،pu ،tɑ}، vu ،me ،gi ،ge ،gɑ ،fo ،ki ،dɑ ،de ،fɑ ،da ،bu ،ʔɑ، {to ،vi ،tu ،vo ،xi ،vɑ ،ve، {ho ،dʒɑ ،ʃɑ ،ru ،no ،tɑ ،xo}،

	{fu ،qu ،su ،so ،je ،jɑ ،ju ،jo ،fe ،tʃo ،mu ،ʔu}، {dʒo}، {he ،hi ،qi ،qa ،ʔe ،ʔi ،nu ،ni}، {gɑ ،ne ،ko ،du ،do ،ʒo ،zu ،ʒu ،lɑ ،lu ،lo ،le ،ʒi}، {ro ،rɑ ،ra ،qe ،qo ،ʃe ،go ،si ،ʃa ،zɑ ،zi ،xu ،kɑ}، {xa ،xe ،xɑ ،zo ،dʒi ،hu ،re}، {ri ،gu ،se ،ʒa ،ʒe ،tʃe ،ke}، {ha ،ʃo ،ʒɑ ،fa ،dʒu ،tʃɑ ،tʃa ،ba ،bɑ ،nɑ ،na ،ku}، {va}، {qɑ ،bo ،ʔo ،pɑ ،po ،hɑ ،sɑ}
روش چهارم	{pi ،bi ،ba ،mi ،be ،me ،pa ،pe ،ma ،ha ،ʔa}، {rɑ ،nɑ ،gɑ ،zɑ ،ja ،tɑ ،qɑ ،dɑ ،xɑ ،kɑ ،hɑ ،ʔɑ}، {ga ،je ،sa ،ra ،ti ،la ،ka ،za ،qa ،na ،da ،lɑ ،ta ،sɑ}، {dʒɑ ،ʃi ،dʒa ،ji ،re ،ʔi ،he ،di ،hi ،ni ،de ،si ،se}، {dʒi ،ʃa ،tʃa ،ri ،xa ،qe ،ʔe ،ge ،xi ،qi ،li ،te ،le ،ʒa}، {dʒe ،tʃe ،xe ،ne ،ki ،zi ،ze ،ʃɑ ،jɑ ،tʃɑ ،ʒi ،ʃe ،ke}، {gi ،ʒe ،ʒɑ ،tʃi}، {pɑ ،mɑ ،bɑ}، {ʔu ،nu ،ʒu ،dʒu ،pu ،mu ،su ،ru ،po ،bu ،ju ،bo}، {mo ،ʃu ،ku ،lu ،gu ،du ،tu ،to ،ro ،so ،no ،zo ،xu}، {go ،tʃo ،tʃu ،ʃo ،dʒo ،xo ،qo ،lo ،jo ،ko ،ʔo ،ho}، {do ،hu ،ʒo ،qu}، {va ،fa}، {fɑ ،vɑ}، {vi ،fi ،ve ،fe}، {vu ،fu ،vo ،fo}، {zu}

۶- نتیجه‌گیری

در این پژوهش خوشه‌بندی تصاویر هجاهای دو آوایی زبان فارسی با ۴ روش مختلف انجام شد و جداول نگاشت هجا به ویسیلاب متناسب با هر روش ارایه شد. همچنین نتایج بهترین جدول نگاشت در بازشناسی ۳۰ کلمه فارسی مورد استفاده قرار گرفت و نتایج حاصله ارایه شد. بررسی‌های این پژوهش نشان می‌دهد که چالش وابستگی به گوینده همچنان چالش مهمی در لب‌خوانی و یافتن نگاشت هجا به ویسیلاب است و نگاشت حاصل از نمونه‌های کمتر در روش‌های دوم و سوم قابلیت تعمیم به نمونه‌های بیشتر را ندارد. نتیجه دیگر اینکه، در ۳ روش مختلف، تعداد خوشه بهینه ۹ حاصل شد. نتایج مشخص می‌کند که توزیع هجاها در خوشه‌های مختلف از الگوی یکسانی پیروی نمی‌کند. جداول نگاشت به‌دست آمده همراه با اطلاعات زبانی دیگر مانند مدل‌های زبانی، می‌تواند در بازشناسی دیداری گفتار مفید واقع شود. دقت بالای عامل انسانی در خوشه‌بندی ویسیلاب‌ها با سابقه طولانی شخص در لب‌خوانی و استفاده از تجربیاتش قابل توجیه است. به‌نظر می‌رسد نگاشت حاصل از عامل انسانی بتواند در پژوهش‌های بعدی در زمینه لب‌خوانی زبان فارسی و سایر پژوهش‌های مرتبط مانند آموزش ناشنوایان مورد استفاده قرار گیرد.

مراجع

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

[2] 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.

[۳] برخان, مسعود؛ فتاح علیزاده و وفا میهمی، ۱۳۹۶، طراحی و پیاده سازی یک سیستم جهت تشخیص خودکار حروف زبان فارسی از طریق لب خوانی با روش های پردازش تصویر نهمین کنفرانس فناوری اطلاعات و دانش(IKT 2017) تهران، دانشگاه صنعتی امیرکبیر

[4] K. M. Talarposhti and M. K. Jamei, “An Efficient Model for Lip-reading in Persian Language Based on Visual Word and Fast Furrier Transform Combined with Neural Network,” vol. 8, no. 2, p. 22, 2017.

[5] 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.

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

[7] 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.

[8] 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.

[۹] اسلامی, محرم, رحیمی, اسلامی و سودابه, “نظام آوایی زبان فارسی در آینهٔ آمار,” زبان و زبان‌شناسی, vol. 9, no. 18, pp. 65–90, 2013.

[۱۰] محمدمهدی همایون‌پور, پژوهشنامه تبدیل متن به گفتار. شورای عالی اطلاع رسانی, شماره ۲۳۷۸۰۱۱, ۱۳۹۰.

[۱۱] یدالله ثمره, آواشناسی زبان فارسی: آواها و ساخت آوایی هجا. مرکزنشردانشگاهی, ۱۳۹۱.

[12] 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.

[13] 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.

[14] H. L. Bear and R. Harvey, “Phoneme-to-viseme mappings: the good, the bad, and the ugly,” *Speech Communication*, vol. 95, pp. 40–67, 2017.

[15] 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.

[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] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen, “Fuzzy c-means clustering with spatial information for image segmentation,” *computerized medical imaging and graphics*, vol. 30, no. 1, pp. 9–15, 2006.

[18] 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.