استخراج و مدل‌سازی واحدهای آوایی وابسته به بافت برای بهبود دقت بازشناسی گفتار پیوسته با روش دسته‌بندی واج‌ها

محمد بحرانی و حسین ثامتی

چکیده: در این مقاله برای بهبود دقت یک سیستم بازشناسی گفتار پیوسته فارسی، روش وابسته به بافت مناسبی پیشنهاد شده است. به دلیل بعضی محدودیت‌های موجود در سیستم بازشناسی، از ایدهٔ واحدهای آوایی چندگانه برای استخراج واحدهای آوایی وابسته به بافت استفاده گردیده است. بر اساس این ایده هر واج به چند نوع گوناگون دسته‌بندی می‌شود و هر دسته جداگانه مدل‌سازی می‌گردد. دسته‌بندی واج‌ها به صورت بی‌نظارت و با استفاده از الگوریتم k-means انجام شده است و برای محاسبه مرکز دسته‌ها روش کارایی پیشنهاد شده است. تعداد دسته مناسب برای هر واج با توجه به حجم داده‌های آموزشی آن واج و دقت بازشناسی واج در هنگام به‌کارگیری مدل‌های مستقل از بافت، حدس زده شده و سپس با روش‌های مبتنی بر سعی و خطا، تعداد دسته بهینه برای هر واج تعیین شده است. سپس هر دسته به عنوان یک واحد آوایی وابسته به بافت مدل‌سازی گردیده است. با استفاده از این مدل‌ها حدود ۲۲ درصد کاهش در نرخ خطای کلمات حاصل شده است.

کلید واژه: **بازشناسی گفتار پیوسته، دسته‌بندی، مدل مخفی مارکوف، مدل‌های وابسته به بافت.**

۱- مقدمه

یکی از مسائل مطرح در بازشناسی گفتار پیوسته، مسالهٔ مدل‌سازی اثر بافت[1] در گفتار می‌باشد. در ساختارهای مستقل از بافت، هر واحد آوایی (مثلا هر واج) به طور مستقل مدل می‌شود و واحدهای آوایی اطراف آن مورد توجه قرار نمی‌گیرند. در حالی که واحدهای آوایی در هنگام مجاورت با هم، اثر متقابل بر یکدیگر گذاشته و تغییراتی در طرز ادای آنها ایجاد می‌شود که این تغییرات بسته به نوع واج‌های مجاور، متفاوت است. به عبارت دیگر خصوصیات طیفی یک واج در گفتار کاملاً به بافت اطراف آن وابسته است؛ بنابراین برای مدل‌سازی دقیق و مناسب گفتار، به منظور بازشناسی، باید دگرگونی‌های حاصل از بافت را نیز در نظر گرفت و در واقع بافت صحبت را به نحوی در مدل‌سازی دخالت داد. بخشی از مسالهٔ مدل‌سازی اثر بافت، با انتخاب داده‌های آموزشی مناسب برای مدل تصادفی مورد استفاده حل می‌شود ولی لازم است روش‌های خاص برخورد با این مساله نیز به کارگیری شود تا دقت بازشناسی تا حد ممکن بالا رود. این روش‌ها استفاده از ساختارهای وابسته به بافت[2] در بازشناسی می‌باشند. تجربه نشان داده است که اطلاعات وابسته به بافت نقش مهمی در بازشناسی ایفا می‌کنند و با به‌کارگیری آنها میزان خطای بازشناسی به مقدار قابل توجهی کاهش می‌یابد.

روش‌های وابسته به بافت متنوعی برای مدل‌سازی اثرات بافت صحبت به کار رفته‌اند. در بیشتر این روش‌ها مدل‌سازی گفتار بر اساس واحدهای آوایی وابسته به بافت (مانند دو واجی، سه واجی، هجا، نیم‌هجا، واحدهای آوایی چندگانه[3] و ...) انجام گرفته و اثرات قابل توجهی بر افزایش دقت بازشناسی داشته است [۱] تا [۵]. هدف از مقالهٔ حاضر نیز بهبود دقت یک سیستم بازشناسی گفتار پیوسته فارسی با به کارگیری یک ساختار وابسته به بافت مناسب می‌باشد. در این مقاله، از ایدهٔ واحدهای آوایی چندگانه (مدل‌سازی واج‌گونه‌ها[4]) برای مدل‌سازی وابسته به بافت استفاده کرده‌ایم. مزیت استفاده از این روش اینست که می‌توان با توجه به حجم داده آموزشی در دسترس، تعداد واحدهای آوایی وابسته به بافت را انتخاب کرد. در ادامه در بخش ۲ به ایدهٔ واحدهای آوایی چندگانه و کارهای قبلی در این زمینه می‌پردازیم. در بخش ۳ جزئیات روش پیشنهادی را بررسی می‌کنیم. نتایج به کارگیری این روش در سیستم بازشناسی گفتار پیوسته در بخش ۴ آمده است. بخش ۵ نیز به جمع‌بندی و نتیجه‌گیری اختصاص دارد.

۲- واحدهای آوایی چندگانه

ایدهٔ کلی واحدهای آوایی چندگانه اینست که یک واج در کلمات و متن‌های مختلف، به حالت‌های گوناگونی تلفظ می‌شود. بنابراین می‌توان گفت که خصوصیات طیفی یک واج (واحد آوایی مستقل از بافت) در بافت‌های مختلف دچار تغییرات و دگرگونی‌هایی می‌شود که این تغییرات را می‌توان در چند دسته کلی جای داد. به عبارت دیگر هر واحد آوایی مانند p بسته به این که در چه متنی قرار دارد دارای چند حالت گوناگون خواهد بود (مانند p_1 ، p_2 ، p_3)؛ پس می‌توان به جای یک واحد آوایی مستقل از بافت، انواع گوناگون وابستگی به بافت آن را به عنوان واحدهای بازشناسی به کار برد [۱].

دلیل استفاده از واحدهای آوایی چندگانه در این مقاله (به جای روش‌های رایج‌تر وابسته به بافت مانند دوواجی و سه‌واجی) محدودیت‌های موجود در سیستم بازشناسی گفتار فارسی مورد استفاده [۶] می‌باشد. دو محدودیت عمده که با آن مواجه هستیم عبارتند از:

۱- دادگان گفتاری مورد استفاده (دادگان فارس‌دات) نسبتاً کوچک می‌باشد بنابراین برای به‌کارگیری دوواجی یا سه‌واجی به عنوان واحد بازشناسی با مشکل کمبود دادهٔ آموزشی مواجه هستیم [۷].

۲- سرعت بازشناسی سیستم با افزایش تعداد واحدهای آوایی کاهش می‌یابد بنابراین استفاده از واحدهای آوایی دوواجی و سه‌واجی، به دلیل

مقاله در تاریخ ۲۸ مرداد ماه ۱۳۸۳ دریافت و در تاریخ ۲۵ تیر ماه ۱۳۸۴ بازنگری شد. این تحقیق با حمایت مرکز صنایع نوین، وابسته به وزارت صنایع و معادن انجام گرفته است.

محمدبحرانی گروه هوش مصنوعی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، تهران، (email: bahrani@ce.sharif.edu).

حسین ثامتی، گروه هوش مصنوعی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، تهران، (email: sameti@sharif.edu).

1. Context
2. Context Dependent
3. Multiple Phone Units
4. Allophones

زیاد بودن تعداد آن‌ها، موجب کاهش شدید سرعت بازشناسی می‌شود.

روش‌های مختلفی برای پیدا کردن خودکار حالات گوناگون یک واج وجود دارد. یک روش مناسب برای دسته‌بندی[1] حالات گوناگون یک واج استفاده از یک سری قواعد آوایی استخراج شده از دانش آواشناسی می‌باشد. با استفاده از این قواعد می‌توان نوع تأثیرپذیری یک واج از واج‌های اطرافش را تعیین کرده و بر اساس آن واج را در یک دسته خاص جای داد [۸] و [۹]. برای به کارگیری این روش نیازمند دسترسی به قواعد آواشناسی هستیم. روش دیگری در [۱] پیشنهاد شده که به قواعد آواشناسی نیاز ندارد. در این روش ابتدا کل نمونه‌های یک واج در دادگان آموزشی جمع‌آوری شده و با استفاده آن‌ها یک مدل برای آن واج آموزش داده می‌شود سپس امتیاز شباهت هر کدام از نمونه‌های آموزشی با مدل به‌دست آمده سنجیده می‌شود؛ حال نمونه‌هایی که امتیاز شباهت[2] آن‌ها با مدل از حد خاصی کمتر است جدا شده و در یک دستهٔ جدید قرار می‌گیرند و برای آن‌ها مدل جداگانه‌ای آموزش داده می‌شود. این روند آنقدر ادامه می‌یابد تا زمانی که تعداد دسته‌های مختلف برای واج به حد مورد نظر برسد. در این حالت هر دسته بیانگر یک نوع مختلف از واج مورد نظر می‌باشد. عیب این روش اینست که واج‌هایی که امتیاز شباهت آن‌ها با مدل از حد خاصی کمتر است را در یک دسته قرار می‌دهد در حالی که این واج‌ها لزوماً واج‌های مشابهی نیستند و ممکن است پراکندگی زیادی نسبت به هم داشته باشند.

در این مقاله به دلیل عدم دسترسی به قواعد آوایی زبان فارسی، دسته‌بندی حالات گوناگون واج‌ها را به صورت بی‌نظارت و با استفاده از الگوریتم دسته‌بندی *k-means* [۱] انجام داده‌ایم. دلیل انتخاب این الگوریتم سادگی و سهولت پیاده‌سازی آن می‌باشد. در ادامه نحوهٔ به‌کارگیری این الگوریتم برای دسته‌بندی واج‌ها را شرح می‌دهیم.

۳- دسته‌بندی واج‌ها با استفاده از الگوریتم K-MEANS و مدل‌سازی آن‌ها

الگوریتم *k-means*، L بردار x_i $(i = ۱,...,L)$ را به K دسته نسبت می‌دهد طوری که پراکندگی در هر دسته مینیمم باشد. هر دسته با یک بردار معرف m_k $(k = ۱,\cdots,K)$ مشخص می‌شود که معمولاً مرکز[3] بردارهای منسوب به دسته k اُم می‌باشد.

۳-۱ استخراج دنبالهٔ ویژگی واج‌ها

برای دسته‌بندی واج‌ها باید هر نمونه واج از گفتار را با یک دنبالهٔ ویژگی[4] نمایش دهیم. برای این منظور لازم است دادگان گفتاری تقطیع واجی شده باشد، بدین معنی که واج‌های ادا شده در گفتار و مرز آنها در سیگنال گفتار دقیقاً مشخص باشد.

در ابتدا سیگنال گفتار مورد پیش‌پردازش قرار می‌گیرد بدین صورت که سیگنال گفتار به تعدادی فریم با طول مساوی (و با مقداری همپوشانی بین فریم‌ها) تقسیم شده و از هر فریم، پس از طی مراحلی، تعدادی ضریب بازنمایی (که در این سیستم ضرایب مل_ کپستروم می‌باشد) به عنوان بردار ویژگی استخراج می‌گردد. چون دادگان گفتاری تقطیع شده می‌باشد، فریم‌های ابتدایی و انتهایی هر واج را در سیگنال گفتار می‌توان مشخص کرد؛ بنابراین هر واج از تعدادی فریم تشکیل شده که هر فریم نیز با یک بردار ویژگی n بُعدی مشخص می‌گردد. پس می‌توان برای هر واج یک دنبالهٔ ویژگی به صورت دنباله‌ای از l بردار ویژگی فریم تعریف کرد که l مشخص کننده تعداد فریم‌های تشکیل دهنده واج می‌باشد. در این مقاله تعداد فریم تشکیل دهنده دنبالهٔ ویژگی یک واج را به عنوان «طول دنبالهٔ ویژگی» می‌شناسیم. تعداد فریم تشکیل دهنده هر واج بسته به میزان کشیدگی زمانی آن واج در گفتار متفاوت می‌باشد بنابراین دنبالهٔ ویژگی واج‌ها دارای طول‌های متفاوتی خواهند بود.

پس از پیش‌پردازش می‌توان به ازای هر یک از واج‌های زبان، دنبالهٔ ویژگی تمام نمونه‌های آن را در دادگان گفتاری جمع‌آوری کرد. اگر زبان دارای N واج باشد، N مجموعه دنبالهٔ ویژگی داریم که هر مجموعه مربوط به یکی از واج‌های زبان می‌باشد. حال می‌توان هر کدام از این مجموعه‌ها را به چند دسته تقسیم‌بندی کرد طوری که هر دسته بیانگر یکی از گوناگونی‌های آن واج باشد.

به دلیل غیر هم‌طول بودن دنباله‌های ویژگی واج‌ها، برای دسته‌بندی آن‌ها با استفاده از الگوریتم *k-means* با دو مشکل مواجه هستیم. مشکل اول این که نمی‌توان برای محاسبهٔ مرکز دسته‌ها از میانگین‌گیری معمولی دنباله‌ها استفاده کرد و مشکل دوم این که برای محاسبهٔ فاصله بین دنباله‌ها، توابع فاصله معمولی (مانند فاصله اقلیدسی) را نمی‌توان به کار برد. برای حل این مشکلات از روش انطباق زمانی پویا[5] (DTW) [۱] استفاده کرده‌ایم.

۳-۲ محاسبه فاصله بین دنباله ویژگی‌های واج‌ها با استفاده از DTW

الگوریتم DTW، دو دنباله X و Y که لزوماً از لحاظ کشش زمانی، هم طول نمی‌باشند را با هم منطبق[6] می‌کند طوری که فاصلهٔ دو دنباله منطبق شده کمینه باشد. بنابراین با به کارگیری DTW به عنوان تابع فاصله [۱۰] تا [۱۳] می‌توان فاصله دنباله‌های ویژگی دو واج را محاسبه نمود.

البته در کار ما هر یک از مؤلفه‌های دنباله‌ها خود یک بردار ویژگی n بعدی از ضرایب بازنمایی هستند که برای محاسبهٔ فاصلهٔ بین آن‌ها از تابع فاصلهٔ اقلیدسی استفاده می‌شود.

معیار DTW به عنوان تابع فاصله، متقارن نیست. برای متقارن کردن این معیار با استفاده از رابطهٔ زیر فاصله دو دنبالهٔ ویژگی واج را محاسبه می‌کنیم [۱] و [۱۰].

$$\delta_{DTW}(X,Y) = \frac{d_{DTW}(X,Y) + d_{DTW}(Y,X)}{۲} \qquad (۱)$$

۳-۳ محاسبهٔ مرکز دسته‌ها

بردار مرکز دسته‌ای از L بردار ویژگی مانند $\{x_i\}_{i=۱}^{L}$، برداری است که میانگین فواصل آن نسبت به بردارهای دسته، مینیمم ممکن باشد. به عبارت دیگر بردار مرکز $\bar{\mathbf{y}}$ با رابطه زیر تعریف می‌شود:

$$\bar{y} \triangleq \arg\min_{y} \frac{۱}{L}\sum_{i=۱}^{L} d(x_i, y) \qquad (۲)$$

لازم به ذکر است که بردار مرکز $\bar{\mathbf{y}}$ لزوما عضو مجموعه بردارهای دسته نمی‌باشد.

در هنگامی که با الگوهای گفتاری (دنباله‌های زمانی از بردارها) سروکار داریم، به دلیل تفاوت در طول زمانی دنباله‌ها، یافتن الگویی که

1. Clustering
2. Likelihood Score
3. Centroid
4. Feature Sequence
5. Dynamic Time Warping
6. Warp

میانگین فواصل را در دسته مینیمم کند، کار ساده و سرراستی نیست [۱]. بنابراین معمولاً سعی می‌شود تخمین مناسبی از این الگو به عنوان مرکز به کار رود. در این مقاله برای تخمین مرکز دنباله‌ها در هر دسته، از روش «انطباق و میانگین‌گیری» [۱] و [۱۰] و [۱۲] استفاده کرده‌ایم. در این روش به شیوه زیر عمل می‌کنیم:

ابتدا یکی از دنباله‌های دسته به عنوان دنباله محور[1] انتخاب می‌شود. معمولاً دنبالهٔ محور به این صورت انتخاب می‌شود که ابتدا یک ماتریس $L\times L$ از فواصل همهٔ جفت بردارهای موجود در دسته تشکیل می‌شود سپس برداری که میانگین فواصل آن با سایر بردارها از بقیه کمتر است به عنوان بردار محور انتخاب می‌گردد [۱] و [۱۰] و [۱۱] ولی در این مقاله دنبالهٔ محور را در هر دسته طوری انتخاب کرده‌ایم که طول آن میانگین (یا نزدیک به میانگین) طول دنباله‌های موجود در آن دسته باشد. دنباله‌ای که طول آن میانهٔ طول دنباله‌های دسته باشد نیز می‌تواند انتخاب مناسبی برای دنباله محور باشد [۱۲]. طول دنبالهٔ محور مشخص کننده طول دنبالهٔ مرکز دسته می‌باشد. چون دنبالهٔ مرکز در واقع به نوعی مشخص کننده میانگین دنباله‌های دسته است بنابراین طول آن را هم برابر میانگین طول دنباله‌های دسته تعیین کرده‌ایم. پس از مشخص کردن دنباله محور، هریک از دنباله‌های دسته را با دنباله محور با استفاده از روش DTW منطبق می‌کنیم. سپس به ازای هر مؤلفهٔ $\mathbf{p}_i$ از دنبالهٔ محور، تمام مؤلفه‌های دنباله‌های دیگر را که با $\mathbf{p}_i$ انطباق یافته‌اند در نظر گرفته، از آن‌ها میانگین‌گیری می‌کنیم و حاصل را به عنوان مولفهٔ $\mathbf{c}_i$ از دنباله مرکز قرار می‌دهیم. در این مقاله چون فاصله بین مؤلفه‌ها (که خود یک بردار n بعدی از ضرایب مل-کپستروم می‌باشند)، با تابع فاصلهٔ اقلیدسی محاسبه شده است، میانگین‌گیری بین آن‌ها به روش معمولی انجام می‌گیرد. می‌توان نشان داد [۷] که تخمین مرکز دسته‌ها به این روش نسبت به روش‌های معمول، که مبتنی بر ماتریس فاصله[2] می‌باشند، از نظر هزینهٔ محاسباتی مقرون به صرفه می‌باشد.

۳-۴ تعیین تعداد دسته‌ها برای هر واج

یکی از مشکلات به کارگیری روش دسته‌بندی بی‌نظارت، تعیین تعداد دسته‌ها برای هر واج می‌باشد. در غیاب اطلاعات آواشناسی هیچ روشی برای تعیین دسته مناسب برای هر واج وجود ندارد. به بیان دیگر نمی‌توان تعداد گوناگونی‌های هر واج را به طور قطعی تعیین کرد؛ ولی دو فاکتور را برای تعیین تعداد دسته‌های هر واج می‌توان در نظر گرفت. فاکتور اول فرکانس وقوع آن واج در دادگان آموزشی است. هر چه تعداد نمونه‌های یک واج خاص در دادگان بیشتر باشد می‌توان آن واج را به تعداد دستهٔ بیشتری تقسیم بندی کرد بدون این که نگران کمبود دادهٔ آموزشی برای آموزش مدل‌های مربوط به انواع گوناگون آن واج باشیم. برعکس در نظر گرفتن تعداد دستهٔ زیاد برای واج‌های با فرکانس وقوع کم، باعث می‌شود که تعداد نمونه‌ها در هر دسته اندک بوده و برای آموزش مدل ناکافی باشد.

فاکتور دیگر برای تعیین تعداد دسته مناسب برای هر واج، میزان خطای بازشناسی آن واج، در هنگامی که مدل سازی بر اساس تک‌واج‌ها انجام گرفته است، می‌باشد. هر چه یک واج با دقت کمتری بازشناسی شود می‌توان احتمال داد که پراکندگی نمونه‌های مختلف آن واج زیاد بوده است که باعث شده مدل‌سازی واج به خوبی انجام نگیرد. پس می‌توان هنگام دسته‌بندی، تعداد دستهٔ بیشتری برای آن واج در نظر گرفت. در عوض برای واجی که در حالت عادی با دقت بالایی شناسایی می‌گردد می‌توان از دسته‌بندی صرفنظر کرد یا تعداد دسته کمتری (نسبت به واج‌های دیگر) برای آن در نظر گرفت.

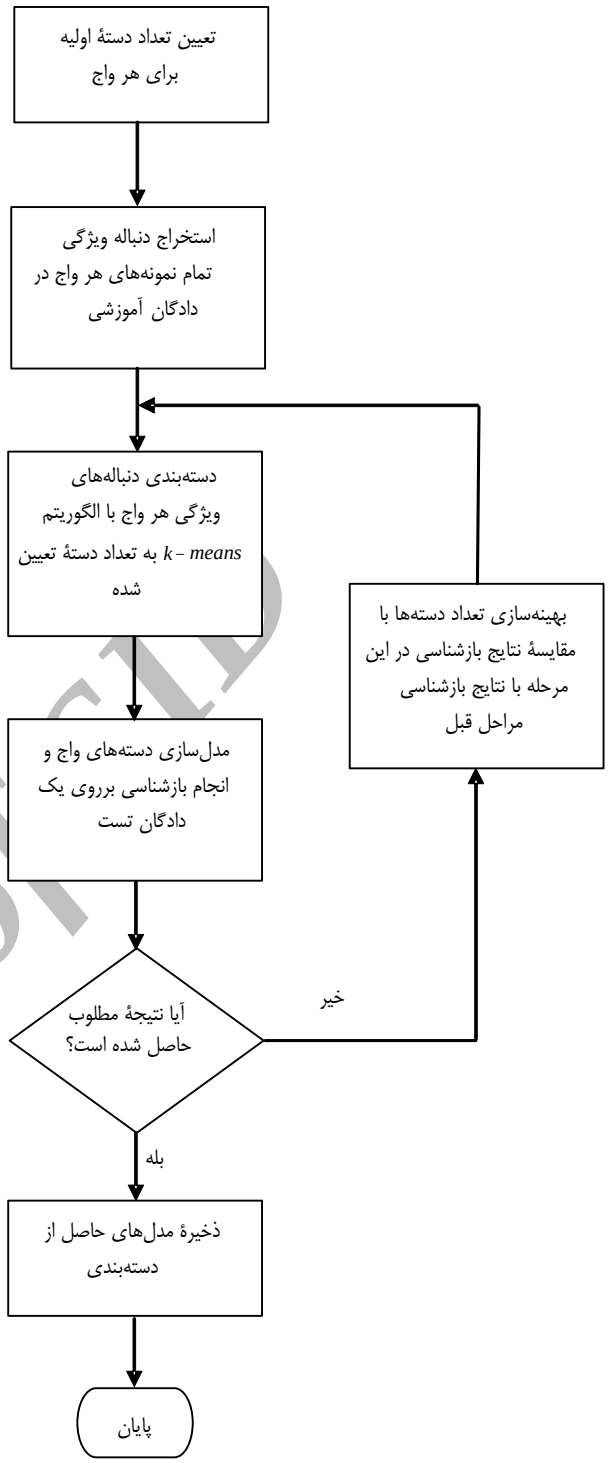
در این مقاله تعداد دستهٔ مناسب برای هر واج با توجه به دو فاکتور مذکور تعیین شده است (تعداد دسته‌ها اغلب بین ۱ تا ۵ دسته می‌باشد)؛ سپس تعداد دسته‌ها بهینه‌سازی گردیده است. روشی که برای بهینه سازی تعداد دسته‌ها به کار برده‌ایم مبتنی بر سعی و خطا می‌باشد. در این روش ابتدا برای هر واج یک تعداد دسته اولیه در نظر گرفته و واج‌ها را بر اساس آن دسته‌بندی می‌کنیم؛ سپس مدل‌های مربوط به هر دسته را آموزش داده و با استفاده از این مدل‌ها بازشناسی را بر روی یک دادگان تست انجام می‌دهیم. حال میزان خطای بازشناسی هر واج را با میزان خطای بازشناسی همان واج در حالتی که مدل‌سازی بر اساس واج‌های دسته‌بندی نشده انجام گرفته است مقایسه می‌کنیم. بر اساس این که به کارگیری مدل‌های حاصل از دسته‌بندی چقدر در میزان خطای بازشناسی یک واج تأثیر داشته است، تصمیم‌گیری در مورد بهتر کردن تعداد دسته‌های آن واج انجام می‌گیرد؛ بدین صورت که برای واج‌هایی که میزان خطای بازشناسی آن‌ها پس از دسته‌بندی کاهش یافته است تعداد دسته‌ها را افزایش می‌دهیم (البته تا حدی که در هر دسته داده آموزشی کافی موجود باشد) و برای واج‌هایی که میزان خطای بازشناسی آن‌ها پس از دسته بندی افزایش یافته و یا بدون تغییر مانده است تعداد دسته‌ها را کاهش می‌دهیم و یا از دسته بندی آن واج صرفنظر می‌کنیم. این روند می‌تواند تا رسیدن به یک دسته‌بندی بهینه چند مرحله تکرار شود و در هر مرحله، نتایج بازشناسی حاصل از به کارگیری دسته‌بندی جدید با نتایج مراحل قبل مقایسه گردند. لازم به ذکر است که حداقل بودن تعداد کل دسته‌ها (تعداد کل واحدهای آوایی) حائز اهمیت است زیرا کمتر بودن تعداد واحدهای آوایی در سرعت فرآیند بازشناسی تأثیر مثبت خواهد داشت.

۳-۵ مدل‌سازی وابسته به بافت با استفاده از واج‌های دسته‌بندی شده

پس از تعیین تعداد دسته مناسب برای هر واج، الگوریتم *k-means* را برای دسته‌بندی دنباله‌های ویژگی مربوط به نمونه‌های مختلف آن واج به کار می‌بریم. به جای تابع فاصله از DTW استفاده می‌کنیم و برای محاسبهٔ مرکز دسته‌ها روش «انطباق و میانگین‌گیری» را به کار می‌بریم. مقداردهی اولیه دنباله‌های مرکز می‌تواند بر اساس دسته‌بندی تصادفی دنباله‌های ویژگی صورت گیرد؛ بدین صورت که ابتدا دنباله‌های ویژگی به K دسته به طور تصادفی تقسیم بندی می‌گردند، سپس مرکز این دسته‌ها به عنوان مقدار اولیه دنباله‌های مرکز در نظر گرفته می‌شود. تصادفی بودن دسته‌بندی اولیه ممکن است در اجراهای مختلف الگوریتم *k-means* دسته‌بندی‌های نهایی متفاوتی تولید کند. برای اجتناب از جواب‌های متفاوت، باید الگوریتم همیشه از یک دسته‌بندی اولیهٔ ثابت آغاز گردد. برای این منظور می‌توان در ابتدا دنباله‌های ویژگی واج‌ها را به یک ترتیب خاص مرتب کرده (مثلاً به ترتیب رخ دادن آن‌ها در دادگان گفتاری) سپس آن‌ها را به K دسته مساوی تقسیم نمود.

شرط توقف الگوریتم، ثابت ماندن مرکز دسته‌ها در دو تکرار متوالی در نظر گرفته شده است. برای این منظور در هر تکرار الگوریتم *k-means*، فواصل دنباله مرکز دسته‌ها نسبت به دنباله‌های مرکز متناظرشان در تکرار قبلی محاسبه می‌شود و اگر همهٔ فاصله‌ها از یک حد آستانهٔ کوچک مانند ε کمتر بودند به معنی همگرا شدن روند دسته‌بندی و پایان الگوریتم می‌باشد. به عبارت دیگر، اگر $C_k^{(i)}$ مرکز دسته kاُم در تکرار i از الگوریتم

1. Pivot
2. Distance Matrix



شکل ۱: مراحل کلی استخراج و مدل‌سازی واحدهای آوایی وابسته به بافت به روش دسته‌بندی واج‌ها.

باشد، باید شرط زیر برای تمام دسته‌ها برقرار شود تا الگوریتم متوقف گردد

$$\delta_{DTW}\left(C_k^{(i-1)}, C_k^{(i)}\right) < \varepsilon \quad , k = 1,2,\ldots,k \qquad (3)$$

در غیر این صورت تکرار الگوریتم ادامه می‌یابد تا زمانی که شرط توقف برقرار شود.

لازم به ذکر است که محاسبه مرکز دسته‌ها به روش ذکر شده، میانگین فواصل را در هر دسته مینیمم نمی‌کند [۱]؛ زیرا به دلیل متفاوت بودن کشش زمانی دنباله‌ها یافتن دنباله‌ای که میانگین فواصل دنباله‌های دسته نسبت به آن مینیمم باشد کار سرراستی نیست؛ بنابراین همگرایی الگوریتم تضمین شده نمی‌باشد. به همین دلیل یک ماکزیمم تعداد تکرار نیز در نظر گرفته شده تا در صورت همگرا نشدن، الگوریتم پس از رسیدن به این ماکزیمم تکرار متوقف گردد.

پس از اعمال الگوریتم دسته‌بندی، واج مورد نظر به K دسته تقسیم‌بندی می‌شود که هر دسته را می‌توان به عنوان یکی از حالات گوناگون آن واج در نظر گرفت. در واقع هر دسته بیانگر یک واحد آوایی وابسته به بافت می‌باشد و دنباله‌های ویژگی که در یک دسته قرار دارند داده‌های آموزشی را برای مدل‌سازی آن واحد آوایی فراهم می‌سازند. به عنوان مثال اگر واج /a/ به سه دسته تقسیم شود سه واحد آوایی /a1/، /a2/ و /a3/ تولید خواهد شد که هر یک بیانگر یکی از گوناگونی‌های واج /a/ در متن‌های مختلف می‌باشند. حال به جای آن که یک مدل برای واج /a/ تشکیل گردد، برای هر کدام از سه نوع گوناگون واج /a/ مدل جداگانه‌ای تشکیل می‌شود. مدل‌سازی می‌تواند به طور مستقیم با استفاده از دنباله‌های ویژگی موجود در هر دسته صورت گیرد. تعداد کل مدل‌هایی که باید آموزش داده شوند برابر با تعداد کل واحدهای آوایی وابسته به بافت و به عبارت دیگر برابر با تعداد کل دسته‌های مربوط به همهٔ واج‌ها می‌باشد.

برای بازشناسی واج‌ها در مرحله جستجو از این مدل‌ها استفاده می‌شود؛ بنابراین خروجی بخش بازشناسی دنباله‌ای از واحدهای آوایی وابسته به بافت خواهد بود. با توجه به این که واحد آوایی بازشناسی شده از انواع گوناگون کدام واج باشد، واج اصلی خروجی مشخص می‌گردد. به عنوان مثال اگر دنباله واحدهای آوایی بازشناسی شده <a3, b1, c2> باشد، دنباله واجی خروجی <a, b, c> خواهد بود.

لازم به ذکر است که چون در تعیین تعداد دسته‌ها برای یک واج به تعداد وقوع آن واج در دادگان توجه داریم مشکل کمبود داده آموزشی پیش نمی‌آید. علاوه بر این چون تعداد دسته در نظر گرفته شده برای هر واج معمولاً کمتر از ۵ دسته می‌باشد، تعداد کل واحدهای آوایی خیلی زیاد نمی‌شود؛ بنابراین بازشناسی می‌تواند با سرعت نسبتاً قابل قبولی انجام گردد.

روندنمای شکل ۱ مراحل کلی استخراج و مدل‌سازی واحدهای آوایی وابسته به بافت به روش دسته‌بندی واج‌ها را نشان می‌دهد.

٤- نتایج به دست آمده

٤-١ نتایج به دست آمده از روند دسته‌بندی واج‌ها

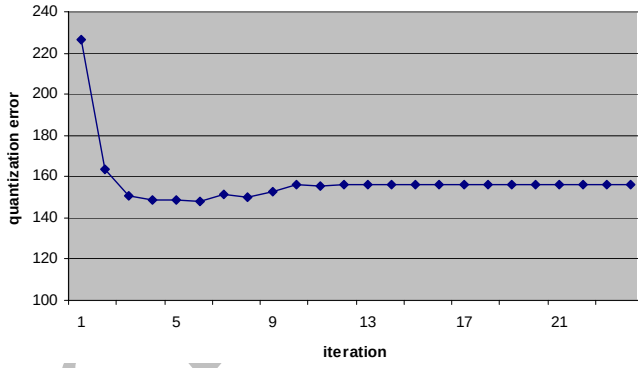
برای تست الگوریتم دسته‌بندی واج‌ها، از دادگان فارس‌دات استفاده شده است. در این دادگان سیگنال‌های گفتاری ضبط شده به طور دستی در سطح واج تقطیع و برچسب دهی شده‌اند، بنابراین دادگان مذکور برای استخراج دنباله ویژگی واج‌ها و دسته‌بندی آن‌ها مناسب می‌باشد.

در مرحله پیش‌پردازش، از هر فریم گفتار بردار ویژگی استخراج می‌گردد. در آزمایشهای ما، طول فریم‌ها ۲۰ میلی‌ثانیه و همپوشانی بین فریم‌های مجاور، ۱۲ میلی‌ثانیه در نظرگرفته شده است. بردار ویژگی استخراج شده نیز شامل ۱۲ ضریب مل-کپستروم $(C_0 - C_{11})$ همراه با مشتق‌های زمانی اول و دوم آن‌ها می‌باشد. به این ترتیب دنباله ویژگی یک واج با طول l فریم، دنباله‌ای از l بردار ویژگی ۳۶ بُعدی از ضرایب مل-کپستروم و مشتق‌های آن‌ها خواهد بود.

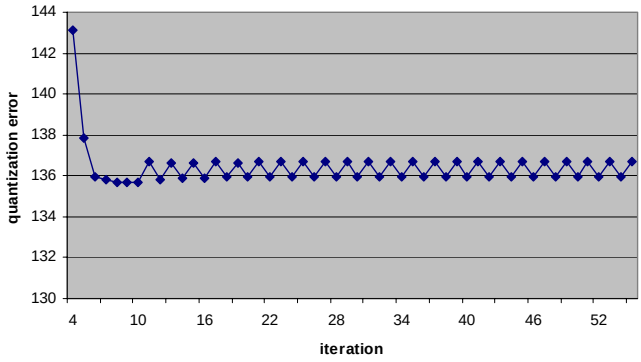
برای تست الگوریتم دسته‌بندی، به ازای هر واج زبان فارسی، الگوریتم دسته‌بندی را (با تعداد دسته‌های از پیش تعیین شده) بر روی مجموعهٔ دنباله ویژگی تمام نمونه‌های آن واج در دادگان فارس‌دات اعمال کرده‌ایم.

جدول ۱: تعداد تکرار الگوریتم دسته‌بندی برای تقسیم واج‌ها به ۳ دسته.

واج	p	b	t	d	c	k	;	g	q	]	'	,	f	v	s	z
تکرار	۲۲	۳۱	۱۲	۲۱	inf	inf	inf	۱۱	۱۱	۱۷	۱۳	۱۴	۱۵	۲۱	۲۰	۹
واج	.	[	x	h	l	r	m	n	y	i	e	a	u	o	/	
تکرار	۲۲	۲۸	۱۹	inf	۲۳	۲۰	۲۳	inf	۱۵	۲۴	۱۷	inf	inf	۳۶	۲۱	



شکل ۲: همگرایی الگوریتم دسته بندی برای تقسیم واج /i/ به ۳ دسته.



شکل ۳: عدم همگرایی الگوریتم دسته‌بندی برای تقسیم واج /i/ به ۴ دسته.

جدول ۲: درصد دقت و درصد صحت بازشناسی واج‌ها در حالت بدون دسته‌بندی.

صحت بازشناسی	دقت بازشناسی	خطای جایگزینی	خطای حذف	خطای درج	تعداد واحدهای آوایی
٪۸۱/۲۰	٪۷۶/۱۶	٪۹/۰۹	٪۹/۷۱	٪۵/۰۵	۴۴

جدول ۳: نتایج بازشناسی پس از دسته بندی واج‌ها به ۳ دسته (دسته بندی ۱).

صحت بازشناسی	دقت بازشناسی	خطای جایگزینی	خطای حذف	خطای درج	تعداد واحدهای آوایی
٪۸۴/۵۶	٪۷۸/۳۴	٪۷/۸۸	٪۷/۵۶	٪۶/۲۲	۱۲۰

جدول ۱ تعداد تکرار لازم برای همگرایی الگوریتم دسته‌بندی در تقسیم هر واج به ۳ دسته را نشان می‌دهد. همان طور که مشاهده می‌شود برای اکثر واج‌ها، روند دسته‌بندی پس از چند مرحله تکرار (معمولاً کمتر از۳۰ تکرار) همگرا شده است. اما به دلیل تضمین نبودن همگرایی الگوریتم، روند دسته‌بندی برای بعضی از واج‌ها حتی با تعداد تکرار زیاد نیز به همگرایی نرسیده است. در این موارد الگوریتم پس از رسیدن به یک ماکزیمم تعداد تکرار (در آزمایش‌های ما ۱۰۰ تکرار) متوقف شده است. در موارد عدم همگرایی، روند دسته‌بندی پس از مدتی به یک حالت نوسانی می‌رسد و در نتیجه شرط همگرایی هیچ وقت ارضا نمی‌شود. در این موارد باید تعداد دسته‌ها برای واج مورد نظر را تغییر داد.

شکل ۲ همگرایی الگوریتم دسته‌بندی را برای دسته‌بندی نمونه‌های واج /i/ در دادگان فارس‌دات به ۳ دسته نشان می‌دهد. همان طور که مشاهده می‌شود، الگوریتم پس از حدود ۲۴ مرحله تکرار، همگرا شده است. شکل ۳ نیز عدم همگرایی الگوریتم دسته‌بندی را برای تقسیم واج /i/ به ۴ دسته نشان می‌دهد. در این جا الگوریتم از مرحله یازدهم به بعد به حالت نوسانی رسیده است.

۴-۲ نتایج حاصل از به‌کارگیری واج‌های دسته‌بندی شده در بازشناسی گفتار پیوسته

برای مدل‌سازی واحدهای آوایی حاصل از دسته‌بندی، مدل‌های مخفی مارکوف با استفاده از الگوریتم *segmental k-means* [۱] آموزش داده شدند. هر مدل شامل ۶ حالت و هر حالت شامل تلفیقی از ۴۴ توزیع گوسی می‌باشد. مجموعهٔ آموزش شامل کل دادگان فارس‌دات می‌باشد که ۱۴۰ جمله از آن برای مجموعهٔ آزمون کنار گذاشته شده است. در مرحله بازشناسی نیز از الگوریتم جستجوی «ویتربی شعاعی» [۶]، برای یافتن بهترین دنباله واحدهای آوایی استفاده شده است.

درصد انواع خطاها (درج، حذف و جایگزینی) و همچنین میانگین درصد صحت و درصد دقت بازشناسی واج‌ها (در حالت مستقل از بافت)، در جدول ۲ آمده است.

در مرحله بعد، آموزش مدل‌ها برای واحدهای آوایی حاصل از دسته‌بندی واج‌ها (واحدهای آوایی وابسته به بافت) انجام گرفت. برای آزمودن میزان تأثیر به کارگیری این واحدهای آوایی در دقت بازشناسی و همچنین تعیین تعداد دستهٔ بهینه برای واج‌ها، دو سری آزمایش انجام گرفت. در آزمایشهای سری اول، در ابتدا تعداد دسته‌ها برای تمام واج‌ها (به جز بستارها) یکسان و برابر با ۳ انتخاب شد (در اکثر آزمایش‌ها، واج‌های بستاری مورد دسته‌بندی قرار نگرفته‌اند زیرا این واج‌ها را به عنوان گوناگونی‌هایی از یک واج شبه سکوت در نظر گرفته‌ایم). البته روند دسته بندی برای بعضی از واج‌ها به ازای ۳ دسته، به همگرایی کامل نرسید و در نتیجه تعداد دسته‌ها برای این واج‌ها، افزایش یا کاهش داده شد. جدول ۳ نتایج بازشناسی را در این حالت نشان می‌دهد (دسته‌بندی ۱). همان طور که مشاهده می‌شود، میانگین درصد صحت و درصد دقت واج‌ها به ترتیب حدود ۲,۲ و ۳,۴ درصد نسبت به حالتی که مدل‌سازی بر اساس واج صورت گرفته، افزایش یافته است. تعداد کل واحدهای آوایی وابسته به بافت (تعداد کل دسته‌ها) در این حالت ۱۲۰ واحد می‌باشد.

در مرحله بعد بهینه‌سازی تعداد دسته‌ها با توجه به میزان تأثیر دسته‌بندی بر دقت بازشناسی هر واج و همچنین فرکانس وقوع هر واج در دادگان آموزشی صورت گرفت. در این مرحله دقت و صحت بازشناسی

جدول ۴: نتایج بازشناسی پس از دو مرحله بهینه‌سازی تعداد دسته‌ها (دسته بندی ۲).

صحت بازشناسی	دقت بازشناسی	خطای جایگزینی	خطای حذف	خطای درج	تعداد واحدهای آوایی
٪۸۵/۹۷	٪۷۹/۶۳	٪۷/۰۱	٪۷/۰۳	٪۶/۳۴	۱۱۵

جدول ۵: نتایج بازشناسی پس از دسته‌بندی بر اساس دو فاکتور مورد نظر (دسته بندی ۳).

صحت بازشناسی	دقت بازشناسی	خطای جایگزینی	خطای حذف	خطای درج	تعداد واحدهای آوایی
٪۸۵/۷۳	٪۷۸/۷۶	٪۷/۳۷	٪۶/۹۰	٪۶/۹۷	۱۰۴

جدول ۶: نتایج بازشناسی پس از یک مرحله بهینه‌سازی (دسته بندی ۴).

صحت بازشناسی	دقت بازشناسی	خطای جایگزینی	خطای حذف	خطای درج	تعداد واحدهای آوایی
٪۸۵/۹۹	٪۷۹/۵۷	٪۶/۹۰	٪۷/۱۱	٪۶/۴۲	۱۰۶

واج‌ها نسبت به «دسته‌بندی ۱» کمی کاهش یافت بنابراین یک مرحله بهینه‌سازی مجدد تعداد دسته‌ها صورت گرفت (دسته‌بندی ۲) که نتایج آن در جدول ۴ آمده است. در این حالت مشاهده می‌شود که میانگین دقت بازشناسی در حدود ۱/۳ درصد، نسبت به «دسته‌بندی ۱»، افزایش یافته ضمن آن که تعداد کل دسته‌ها نیز کمتر شده است.

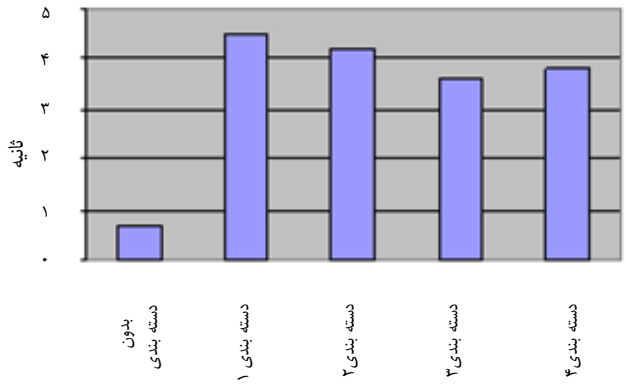
در سری دوم از آزمایش‌ها، در ابتدا تعداد دسته‌ها برای هر واج با توجه به میزان خطای بازشناسی واج در حالت مستقل از بافت (بدون دسته‌بندی) و همچنین تعداد وقوع آن واج در دادگان، تعیین شد. جدول ۵ نتایج بازشناسی را در این آزمایش نشان می دهد. (دسته بندی ۳). پس از یک مرحله بهینه‌سازی، به «دسته‌بندی ۴» رسیدیم که نتایج حاصل از آن تقریباً برابر با نتایج «دسته بندی ۳» می‌باشد ولی تعداد کل دسته‌ها در آن کمتر است (جدول ۶).

در این مقاله، «دسته‌بندی ۴» را با توجه به دقت بازشناسی حاصل و تعداد کل دسته‌ها در آن به عنوان دسته‌بندی بهینه انتخاب کرده‌ایم. هرچند نمی‌توان ادعا کرد که تعداد دسته‌ها در این حالت نیز کاملاً بهینه است ولی در آزمایش‌های مختلف دیگری که انجام شد، نتیجه بهتری حاصل نگردید.

یکی از نتایجی که از آزمایش‌های مختلف با تعداد دسته‌های گوناگون و بهینه‌سازی تعداد دسته‌ها می‌توان گرفت اینست که افزایش (و یا کاهش) دقت بازشناسی هر واج، پس از دسته‌بندی، علاوه بر این که به تعداد دسته‌های خود واج بستگی دارد، به تعداد دسته‌های واج‌های دیگر نیز وابسته است. به عبارت دیگر نمی‌توان تأثیر یک دسته‌بندی خاص برای یک واج را بر دقت بازشناسی آن واج به طور مطلق در نظر گرفت و باید به تعداد دسته‌های مربوط به واج‌های دیگر نیز توجه داشت.

٤-۳ نتایج بازشناسی در سطح کلمه

هدف نهایی بازشناسی گفتار، به دست آوردن دنبالهٔ کلمات می‌باشد. برای تبدیل دنبالهٔ واجی حاصل از بازشناسی به دنبالهٔ کلمات از روش جستجوی همزمان واج و کلمه [۱۴] استفاده کرده‌ایم. در این روش، در مرحله جستجو، همزمان با شکل گرفتن دنبالهٔ واجی، بهترین دنبالهٔ کلمات متناظر با دنبالهٔ واجی نیز، با جستجو در یک درخت واژگان، شکل می‌گیرد. با استفاده از این روش، بازشناسی در سطح کلمه بر روی مجموعه آزمون قبلی انجام شده است. اندازهٔ مجموعهٔ واژگان مورد



شکل ۴: متوسط زمان بازشناسی جملات مجموعه آزمون، در سطح کلمه.

جدول ۷: نتایج بازشناسی در سطح کلمه در آزمایشهای مختلف.

آزمایش	درصد دقت	درصد صحت
بدون دسته بندی	۶۷/۹	۷۸/۰
دسته بندی ۱	۷۲/۵	۸۲/۹
دسته بندی ۲	۷۴/۲	۸۲/۹
دسته بندی ۳	۷۱/۷	۸۱/۸
دسته بندی ۴	۴۷/۷	۸۳/۲

استفاده در حدود ۱۰۹۰ کلمه می‌باشد. جدول ۷ نتایج بازشناسی در سطح کلمه را که با استفاده از مدل‌های حاصل از دسته‌بندی‌های مختلف به دست آمده‌اند، نشان می‌دهد. همان طور که مشاهده می‌شود با استفاده از مدل حاصل از دسته‌بندی ۴ (دسته‌بندی بهینه) حدود ۷ درصد افزایش در دقت بازشناسی داشته‌ایم (از حدود ۶۸ به حدود ۷۵ درصد)، که معادل ۲۲ درصد کاهش در نرخ خطا می‌باشد.

در شکل ۴ متوسط زمان بازشناسی جملات مجموعهٔ آزمون، در سطح کلمه، در آزمایش‌های مختلف آمده است. لازم به ذکر است که کلیهٔ آزمایش‌ها بر روی یک کامپیوتر پنتیوم ۴ (۲/۶ GHz)، با ۵۱۲ مگابایت حافظه، انجام گرفته است.

۵- جمع‌بندی و نتیجه‌گیری

در این مقاله روشی برای استخراج واحدهای آوایی وابسته به بافت با استفاده از دسته‌بندی بدون نظارت واج‌ها ارائه شد. مزیت این روش تعداد نسبتاً کم واحدهای آوایی تولید شده (نسبت به روشهایی مانند به کارگیری دوواجی و سه‌واجی) می‌باشد. تاثیر به کارگیری این روش، کاهش ۲۲ درصدی در نرخ خطای بازشناسی کلمات بود. استفاده از چنین روش دسته‌بندی، در بازشناسی گفتار پیوسته مستقل از گوینده برای زبان فارسی سابقه نداشته است و نتایج به دست آمده در این مقاله بدان جهت اهمیت دارد که قبل از دسته‌بندی نیز نتایج گزارش شده بالاترین دقتی است که تاکنون در کارهای مشابه در زبان فارسی ارائه شده است و با وجود این، روش دسته‌بندی مورد نظر در این مقاله توانسته است خطای بازشناسی را باز هم کاهش دهد.

عیب عمده این روش دشوار بودن تعیین تعداد دسته بهینه برای هر واج می‌باشد. روش بهینه‌سازی به کار رفته در این مقاله مبتنی بر سعی و خطا بود که مستلزم صرف وقت و انجام آزمایش‌های بسیاری است. استفاده از الگوریتم‌هایی که بتوانند، با در نظر گرفتن معیارهای خاصی، به طور اتوماتیک یک دسته‌بندی بهینه برای مجموعه‌ای از الگوها پیدا کنند، در این مورد می‌تواند بسیار مؤثر باشد.

مراجع

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, 1993.

[2] L. Villarrubia, L. H. Gomez, J. M. Elvira, and J. C. Torrecilla, "Context-dependent units for vocabulary-independent spanish speech recognition," in *Proc. ICASSP 96*, vol. 1, pp. 451-454, Georgia, 1996.

[3] W. Reichl and W. Chou, "Robust decision tree state tying for continuous speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 5, pp. 555-566, Sep. 2000.

[4] J. Zhang, F. Zheng, J. Li, C. Luo, and G. Zhang, "Improved context-dependent acoustic modeling for continuous chinese speech recognition," in *Proc. EuroSpeech 2001*, vol. 3, pp. 1617-1620, Sep. 2001.

[5] A. Ganapathiraju, J. Hamaker, J. Picone, and M. Ordowski, "Syllable-based large vocabulary continuous speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 4, pp. 358-366, May 2001.

[۶] ب. باباعلی، بررسی روش‌های هرس‌کردن برای بهبود عملکرد یک سیستم بازشناسی گفتار پیوسته مبتنی بر مدل مخفی مارکوف، پایان‌نامهٔ کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۸۲.

[۷] م. بحرانی، *به کارگیری ساختارهای وابسته به بافت در بازشناسی گفتار پیوسته مبتنی بر مدل مخفی مارکوف*، پایان‌نامهٔ کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۸۲.

[8] J. Ferreiros and J. M. Pardo, "Improving continuous speech recognition in Spanish by phone-class semi continuous HMMs with pausing and multiple pronunciations," *Speech Communication*, vol. 29, no. 1, pp. 65-76, Sep. 1999.

[۹] ا. غلامپور، *بازشناسی مستقل از گوینده واج‌های فارسی در صحبت پیوسته*، پایان نامه دکترا، دانشکده مهندسی برق، دانشگاه صنعتی شریف، ۱۳۷۹.

[10] T. Oates, M. D. Schmill, and P. R. Cohen, "A method for clustering the experiences of a mobile robot that accords with human judgements," in *Proc. 17th National Conf. on Artificial Intelligence*, pp. 846-851, 2000.

[11] V. Vuori and J. Laaksonen, "A comparison of techniques for automatic clustering of handwritten characters," in *Proc 16th Int. Conf. on Pattern Recognition*, vol. 3, pp. 168-171, Quebec, 2002.

[12] D. Bakhash, Hierarchical Clustering and Sequence Averaging for Improved Efficiency and Accuracy of On-Line Chinese Character Recognition, On-line: http://www.mit.edu:8001/people/cadet/Clustering/node1.html, 2003.

[13] J. Picone, "Duration in context clustering for speech recognition," *Speech Communication*, vol. 9, no. 2, pp. 119-128, Apr. 1990.

[14] S. Ortmanns, A. Eiden, and H. Ney, "Improved lexical tree search for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 2, pp. 817-820, Seattle, May 1998.

محمد بحرانی ی در سال ۷۹ مدرک کارشناسی مهندسی کامپیوتر ـ سخت‌افزار خود را از دانشگاه شیراز و در سال ۸۲ مدرک کارشناسی ارشد مهندسی کامپیوتر ـ هوش مصنوعی خود را از دانشگاه صنعتی شریف دریافت نمود و هم‌اکنون دانشجوی دکترای هوش مصنوعی در دانشگاه صنعتی شریف می‌باشد. زمینه‌های علمی مورد علاقه ایشان عبارتند از: پردازش و بازشناسی گفتار، پردازش زبان طبیعی، پردازش تصویر و الگوریتم‌های تکاملی.

حسین ثامتی تحصیلات خود را در مقاطع کارشناسی مهندسی برق ـ الکترونیک و کارشناسی ارشد مهندسی برق ـ سخت افزار کامپیوتر به ترتیب در سالهای ۱۳۶۴ و ۱۳۶۸ از دانشگاه صنعتی شریف و در مقطع دکتری مهندسی برق در سال ۱۳۷۳ از دانشگاه واترلوی کانادا به پایان رسانده است و هم‌اکنون استادیار دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پردازش، بازشناسی، بهسازی و فشرده‌سازی گفتار، پردازش زبان طبیعی، پنهان‌نگاری در گفتار و پردازش علایم دیجیتال.