علوم و مهندسي كامپيوتر
نشريه علمي پژوهشي انجمن كامپيوتر ايران
مجلد ۱، شماره ۳ (الف)، پائيز۱۳۸۲
صفحات۲۳-۳۲
مقاله عادی

مدل‌سازي آكوستيكي واحدهاي گفتاري با استفاده از روش بازنمايي توزيع GVQ

فربد رزازي* ابوالقاسم صيادیان**

*گروه مهندسی برق, دانشکده فنی , دانشگاه آزاد اسلامی، واحد علوم و تحقیقات, تهران, ایران
** دانشكده مهندسي برق, دانشگاه صنعتي اميركبير, تهران, ايران

چكيده

در اين مقاله، روشی برای توصيف توزيع آماري هر قطعه در مدل‌سازي آكوستيكي گفتار پيشنهاد شده است. اين بازنمايي كه در مدل‌هاي مختلف آكوستيكي قابل اعمال است، براي محاسبه احتمال يک بردار ويژگي، از يک توزيع گاوسي استفاده مي‌كند كه با شرايط محيطي واحد گفتاري مورد بازشناسي تطابق داشته باشد. اين توزيع گاوسي از ميان خوشه‌هاي گاوسي كه از پيش تخمين زده شده‌اند، انتخاب مي‌گردد. در اين زمينه، در مقاله حاضر الگوريتم‌هاي آموزش و بازشناسي مورد نياز تدوين، تحليل و آزموده شده‌اند و نشان داده شده است كه استفاده از روش آموزش GVQ روند آموزش را به نسبت قابل ملاحظه‌اي تسريع مي‌كند بدون آن‌كه نرخ بازشناسي نهايي را تغيير دهد. همچنين استفاده از روش GVQ زمينه را براي استفاده از روش‌هاي انتخاب تابع گاوسي در جهت تسريع بخش بازشناسي مهيا مي‌كند.

كلمات كليدي: بازشناسي گفتار، مدل‌سازي آكوستيكي گفتار، بازنمايي پارامتريک آماري، مدل‌سازي با مخلوط‌هاي گاوسي ، كوانتيزاسيون برداري، مدل كوانتيزاسيون برداري گاوسی

۱- مقدمه

مدل سازی آكوستيكی گفتار، هسته اصلي سيستم‌هاي بازشناسي گفتار به شمار مي‌رود. بازشناسي واحدهاي گفتار با استنتاج از ويژگي‌هاي طيفي استخراج شده از هر واحد گفتاري صورت مي‌پذيرد و لذا ارائه توصيفي قابل اعتماد از ويژگي‌هاي هر واحد گفتاري، اجتناب‌ناپذير خواهد بود.

پيچيدگي اين مدل‌سازي به توانايي بخش‌هاي ديگر سيستم بازشناسي گفتار وابسته است. به عنوان مثال، در صورتي كه ويژگي‌هاي استخراج شده از گفتار آنقدر تفكيک‌كننده باشد كه هر دنباله بردارهاي ويژگي مستقيماً به يک واحد زباني اشاره كند، نيازي به اين مدل‌سازي نيست. همچنين در صورتي كه مجموعه هدف خروجي آنقدر كوچک باشد كه اشتباه در بازشناسي اكوستيكي با مدل زباني جبران شود، مدل‌سازي اكوستيكي نقش حساس خود را از دست مي‌دهد. اما در عمل، نه ويژگي‌هاي موجود تفكيک‌پذيري كافي را ارائه مي‌دهند و نه كاربردهايي با مجموعه‌اي چنين محدود مورد علاقه هستند. لذا پيچيدگي زيادي به بخش مدل‌سازي آكوستيكي واحدهاي زباني تحميل مي‌گردد. در واقع، مدل‌سازي آكوستيكي، آنچه را كه توسط بخش‌هاي ديگر بازنمايي نشده است، بايد به شيوه‌اي آماري مدل كند. اين مسايل شامل تنوع در گوينده، محيط گفتار، وابستگي به متن گفتار، تلفظ، لهجه و احساسات بياني گوينده و حتي تنوع حاصل از دو بيان يک واحد گفتاري توسط يک گوينده مي‌شود.واضح است كه بيان چنين حجم پيچيدگي، تنها به صورتي آماري قابل مدل‌سازي است. به عبارت ديگر، در صورتي كه اطلاعاتي معين و غيرآماري در اين مسايل وجود داشته باشد، انتظار مي‌رود كه در بخش‌هاي قبلي سيستم بازشناسي گفتار نظير استخراج ويژگي‌هاي مناسب ، به صورتی معين و غير آماري بازنمايي شده باشد.از لحاظ آماري، با داشتن دنباله‌اي از بردارهاي مشاهده، بخش مدل‌سازي بايد بهترين دنباله واحدهاي زباني را بازشناسي كند. در صورتي كه بحث را به بازشناسي يک واحد زباني محدود كنيم، مدل آماري بايد واحدي زباني را انتخاب كند كه احتمال P(W|O) را حداكثر كند[۱].

$$W^* = \underset{W}{ArgMax}\, P(W \mid O) = \underset{W}{ArgMax}\, P(O \mid W).P(W) \qquad (۱)$$

در رابطه بالا، O نماينده بردارهاي ويژگي ورودي و W واحد زباني مورد بازشناسي است.

با توجه به این‌که در مقایسه بالا، احتمال مطلق بـــردارهاي ورودي نقشي در جستجوي بهترین واحد زباني ندارد، این احتمال از مخرج عبارت آخر حذف شده است. جمله P(O|W) بیانگر مشخصات آکوستیکي بردارهاي ویژگي یك واحد زباني است. جمله P(W) بیانگر احتمال حضور یك واحد زباني است و معمولاً به واحدهاي زباني مابعد و ماقبل بستگي دارد. نحوه محاسبه جمله اول به مدل‌سازي آکوستیکي گفتار و استخراج روشي براي محاسبه جمله دوم، به مدل‌سازي زباني مجموعه هدف مي‌انجامد.

مدل‌سازي آکوستیکي گفتار باید بتواند جنبه‌هاي مختلف پیچیدگي گفتار را توصیف کند. یکي از دو شاخه اصلي این پیچیدگي، تنوع بردارهاي ویژگي هر قطعه آکوستیکي است که توسط یک توزیع آماري توصیف مي‌شود. اگرچه معمولاً بردارهاي ویژگي متوالي در هر قطعه آکوستیکي، نه مستقل هستند و نه توزیع آنها در طول قطعه ثابت مي‌ماند، اما در مدل‌هاي کلاسیک این مدل‌سازي با فرض بردارهاي با توزیع یکسان و مستقل انجام مي‌شود. مدل‌هایي نیز براي تصحیح هر دو فرض پیشنهاد شده است [۲, ۳]

شاخه دوم، مدل‌سازي اثر کشش زماني در دنباله بردارهاي ویژگي هر قطعه آکوستیکي و شیوه قطعه‌بندي یک واحد زباني به قطعات متوالي است. این جنبه از مدل‌سازي گفتار با استفاده از مدل‌سازي مارکوف [۱] یا قطعه‌بندي نرم[۴, ۵, ۶] پي‌گیري مي‌شود و موضوع بررسي این مقاله نیست. جنبه‌هاي جانبي بسیاري نیز وجود دارد که مختص گفتار نیست و به شیوه مدل‌سازي آماري باز مي‌گردد. از جمله این جنبه‌ها، مي‌توان به تخمین دقیق پارامترها، توانایي تعمیم‌پذیري مدل، تعداد پارامترهاي مدل و حجم محاسبات و زمان بازشناسي اشاره کرد.

این مقاله به شاخه‌اي از بعد اول مدل‌سازي آکوستیکي گفتار مي‌پردازد. معمولاً توزیع بردارهاي ویژگي، توزیع پیچیده‌اي است که به صورت تحلیلي و با پارامترهاي محدود قابل بازنمایي نیست. اما ناگزیر با تقریبي به یکي از شیوه‌هاي توصیف پارامتریک برازش مي‌شود تا امکان محاسبه سریع توزیع در زمان بازشناسي فراهم گردد. همچنین توزیع حاصل که با نمونه‌هاي محدودي تخمین زده شده است، قابل تعمیم به بردارهایي باشد که در زمان آموزش حضور نداشته‌اند. پیچیدگي توزیع واقعي بردارهاي هر قطعه آکوستیکي از سویي و ملاحظات کاهش حجم محاسبات و تعداد پارامترها از سوي دیگر، مدلي را موفق مي‌دارد که آنچنان به طبیعت گفتار نزدیک باشد که با حداقل پارامتر و با حجم محاسبات کمتر بتواند توزیع آماري بردارها را بازنمایي کند.

واضح است که ارزیابي این شیوه بازنمایي توزیع به مدلي براي توصیف اثر کشش زماني و دیگر پدیده‌هاي آکوستیکي نیز نیازمند است. این بستر آزمون، در صورتي که از آن به صورت مقایسه‌اي و با استفاده از شیوه‌هاي بازنمایي مختلف پیاده‌سازي شود، مي‌توانـــد هر یک از شیـــوه‌هاي مـــدل‌سازي آکوستیکي موجـود (HMM[۱]، CMT- SSM[۲]، مدل قطعه‌بندي نرم یا هر مدل معمول دیگر) باشد.

در این مقاله شیوه‌اي تحت عنوان کوانتیزاسیون برداري گاوسي (GVQ)[۳] پیشنهاد و ارزیابي شده است که در آن احتمال هر بردار تنها در یک تابع گاوسي سنجیده مي‌شود و این توزیع گاوسي از میان توابع گاوسي آموزش داده شده انتخاب مي‌گردد. در بخش دوم این مقاله ، به طور خلاصه شاخه‌هاي اصلي مطرح در بازنمایي پارامتریک توزیع هر قطعه آکوستیکي در پژوهش‌هاي پیشین بررسي و مقایسه مي‌شود. بخش سوم به ارائه مدل GVQ اختصاص دارد. ابتدا ایده اصلي این توصیف ارائه مي‌شود و سپس شیوه آموزش و استفاده از آن به صورت تحلیلي تبیین مي‌گردد. در بخش چهارم به ارزیابي تجـربي مـــدل در طبقه بنـــدي واج‌ها پرداخته مي‌شود و توانایي این بازنمایي عملاً نشان داده مي‌شود و در بخش پنجم مقاله جمع‌بندي مي‌گردد.

۲- شیوه‌هاي کلاسیک بازنمایي توزیع هر قطعه در مدل‌هاي آکوستیکي گفتار

مبناي توصیف‌هاي معمول براي توزیع خروجي هر قطعه آکوستیکي در شیوه‌هاي مختلف، یکسان است. همه این شیوه‌ها از این ایده استفاده مي‌کنند که فضاي بردارهاي ویژگي به ازاي هر قطعه ایستان، قابل خوشه‌بندي است و بیان مشخصات تحلیلي هر خوشه براي توصیف این فضا کافي خواهد بود. اگرچه شواهدي وجود دارد که به ازاي ویژگي‌هاي معمول و پر استفاده نظیر MFCC[۴] این فرض را مخدوش مي‌کند [۷], اما این فرض، دست‌کم روشي را براي تقسیم فضا به زیرفضاهایي تحلیلي ارائه مي‌دهد.

اگرچه اکثر روش‌ها براي مدل HMM پیشنهاد و ارزیابي شده‌اند، اما واضح است که ارائه نحوه بیان براي توزیع هر قطعه ، منحصر به HMM نیست، ولي اولاً ارائه و مبناي تحلیلي این روش‌ها از HMM آغاز گشته است و بیشتر مقالات از HMM به عنوان مبناي پیاده‌سازي استفاده کرده‌اند و ثانیاً اکثر روش‌هاي دیگر، تنها از شیوه بیان ترکیب مخلوط‌هاي گاوسي (GMM)[۵] که در CDHMM[۶] به کار برده مي‌شود، استفاده کرده‌اند و کمتر شیوه‌هاي دیگر توصیف توزیع از HMM پا فرا گذاشته‌اند. در ادامه به بیان نحوه توصیف توزیع در هر یک از شیوه‌هاي معمول پارامتري کردن آن پرداخته مي‌شود و روابط بازتخمین آن‌ها ارائه مي‌گردد. این روش‌ها با اندک تغییري به مدل‌هاي آماري دیگر نیز قابل تطبیق هستند.

الف- HMM با توزیع گسسته DDHMM[۷]

در مدل DDHMM، از مراکز خوشه‌ها به عنوان مشخصه هر خوشه استفاده مي‌شود. بنابراین توزیع خروجي هر حالت، توزیعي گسسته است که احتمال هر یک از مراکز خوشه‌ها را ارائه مي‌کند و توسط برداري به طول تعداد خوشه‌ها قابل توصیف است.

براي استفاده از مدل DDHMM، قبل از ورود دنباله بردارهاي مشاهدات به HMM ـ چه در مرحله آموزش و چه در مرحله بازشناسي ـ بردارهاي ویژگي باید به فضاي گسسته منتقل شوند. به همین سبب، پیش از آموزش HMM، با استفاده از روش LBG[۸], یک مدل VQ[۹] براي گسسته کردن فضاي بردارهاي ویژگي آموزش داده مي‌شود.

استفاده از فضاي گسسته، دقت بیان بردارهاي ویژگي و توانایي تعمیم مدل را کاهش مي‌دهد و بنابراین موجب افزایش خطاي بازشناسي مي‌شود. اما سادگي مدل باعث سرعت بیشتر در آموزش و بازشناسي DDHMM مي‌گردد.

پارامترهاي مدل DDHMM، ماتریس احتمال گذر حالات مدل مارکوف (ماتریس A) و ماتریس احتمال هر خوشه در هر حالت (ماتریس B) خواهد بود. پارامتر احتمال اولیه هر حالت معمولاً به صورت معین در نظر گرفته مي‌شود و فرض مي شود که گفتار در اولین قطعه گفتاري آغاز مي شود. البته احتمال حضور اولیه در هر یک از حالات قابل تصور مي‌باشد و پارامترهاي آن نیز قابل بازتخمین است. با توجه به فرض چپ به راست براي ترتیب قطعات گفتاري، ماتریس A به صورت بالا مثلثي خواهد بود.

$$
\begin{aligned}
a_{ij} &= P(s_t = j \mid s_{t-1} = i) \\
b_{ij} &= P(O_t = C_j \mid s_t = i) \\
P_i &= P(s_1 = i)
\end{aligned}
\qquad (۲)
$$

روابط بازتخمین در مدل DDHMM باید به تخمین درایه‌هاي ماتریس A و B بپردازد که خلاصه این روابط در زیر آورده شده است:

$$a_{ij} = \frac{P(s_{t-1}=i, s_t=j)}{P(s_{t-1}=i)} = \frac{\sum_{n=1}^{N_Q}\sum_{t=1}^{T}\alpha_{t-1}(i).a_{ij}.b_j(o_t).\beta_t(j)}{\sum_{n=1}^{N_Q}\sum_{t=1}^{T}\alpha_{t-1}(i).\beta_{t-1}(j)} \quad (۳)$$

$$b_j(k) = \frac{P(o_t=\nu_k, s_t=j)}{P(s_t=j)} = \frac{\sum_{n=1}^{N_Q}\sum_{t=1}^{T_n}\alpha_t(i).\beta_t(i).\delta(o_t,\nu_k)}{\sum_{n=1}^{N_Q}\sum_{t=1}^{T_n}\alpha_t(i).\beta_t(i)} \quad (۴)$$

$$\pi_i = P(s_1=i) = \frac{\sum_{n=1}^{N_Q}\alpha_0(i).\beta_0(i)}{\sum_{n=1}^{N_Q}\alpha_{T_n}(i)} \quad (۵)$$

در رابطه فوق، $\alpha_t(i)$ احتمال مستقیم تا لحظه t ام و با فرض حضور در حالت iام در این لحظه تعریف می‌شود. همچنین $\beta_t(i)$ احتمال معکوس بعد از لحظه t با فرض آغاز از لحظه i ام است. همچنین v_k نشان‌دهنده سمبل k

$$\alpha_t(i) = P(o_1 o_2 ... o_t, s_t = i)$$
$$\beta_t(i) = P(o_{t+1} o_{t+2} ... o_T \mid s_t = i) \quad (۶)$$

ب- HMM با توزیع پیوسته (CDHMM)

در مدل CDHMM ، توزیع چگالي احتمال خروجي هر حالت به صورت تابعي پیوسته بازنمایي مي‌گردد که موجب مي‌شود هر بردار مشاهده‌اي مجاز محسوب شود. به این ترتیب، خطاي حاصل شده از گسسته کردن فضا از میان مي‌رود. یک راه بازنمایي توزیع چگالي احتمال پیوسته با پارامترهاي کم، بازنمایي آن به صورت ترکیب خطي چند تابع پایه است. با توجه به این که توابع گاوسي، پایه‌اي براي فضاي توابع محسوب مي‌شود، مي‌توان از چند تابع گاوسي و ضرایب ترکیب آنها براي بازنمایي تابع پیوسته توزیع چگالي احتمال استفاده نمود.

$$P(O_t) = \sum_{m=1}^{M}\alpha_m N(O_t; \mu_m, \Sigma_m) \quad (۷)$$

این نحوه بازنمایي، GMM نامیده مي‌شود و علاوه بر HMM، در اکثر روش‌هاي آماري دیگر نیز مورد استفاده قرار مي‌گیرد.

انتخاب توابع گاوسی به عنوان پایه، از این انتظار ناشي شده است که با فرض استقلال بردارها، قضیه حد مرکزي، توزیع بردارها را به سمت توابع گاوسي سوق مي‌دهد و این تابع مي‌تواند در بازنمایي توزیع با حداقل تعداد توابع پایه و در نتیجه حداقل تعداد پارامتر کمک کند.

انتخاب توابع گاوسي مناسب، مساله اصلي در تخمین پارامترهاي GMM است. پارامترهاي مدل GMM، پارامترهاي M مخلوط گاوسي (μ_i, Σ_i) و ضرایب اسکالـــر آنها (c_i) است که باید مورد تخمین قرار گیرد. در صـــورت داشتن مجموعه‌اي مشخص از بردارهاي ویژگي، تخمین این پارامترها با استفاده از الگوریتم EM [۱۰] انجام مي‌شود. بنابراین در صورت استفاده از آموزش وبتربي، با مشخص کردن ناحیه هر قطعه در دنباله مشاهـــدات مختلف، مجموعـــه بردارهاي هر قطعه معین مي‌شود و مي‌توان از روش EM براي بازتخمین پارامترهاي GMM استفاده کرد[۸]. واضح است که تخمین پارامترهاي a_{ij} و π_i در دو حالت پیوسته و گسسته یکسان خواهد بود.

$$P_t(j) = P(C_j \mid o_t) = \frac{c_{ij}.N(o_t; \mu_{ij}, \Sigma_{ij})}{\sum_{k=1}^{T} c_{ik}.N(o_t; \mu_{ik}, \Sigma_{ik})} \quad (۸)$$

$$c_{ij} = \frac{\sum_{t=1}^{T} P_t(j)}{\sum_{j=1}^{M}\sum_{t=1}^{T} P_t(j)} \quad (۹)$$

$$\mu_{ij} = \frac{\sum_{t=1}^{T} P_t(j).o_t}{\sum_{t=1}^{T} P_t(j)} \quad (۱۰)$$

$$\Sigma_{ij} = \frac{\sum_{t=1}^{T} P_t(j).(o_t-\mu_{ij}).(o_t-\mu_{ij})'}{\sum_{t=1}^{T} P_t(j)} \quad (۱۱)$$

در رابطه فوق، C_j نشان‌دهنده خوشه گاوسي زام مي باشد. در صورتي که از روش مستقیم براي محاسبه احتمال هر مشاهده استفاده شود، بردارهاي هر قطعه به صورت قطعي مشخص نیست و هر قطعه‌بندي با احتمال منتسب به آن، قطعه‌بندي مجاز شمرده مي‌شود و روابط بازتخمین پارامترهاي GMM توسط الگوریتم بام ولش انجام مي‌گردد. تنها تفاوت این دو روش در نحوه محاسبه احتمال هر خوشه با دانستن بردار ویژگي ورودي است که در روش دوم، میزان نقش هر قطعه‌بندي نیز در مجموع قطعه‌بندي‌هاي ممکن در نظر گرفته شده و مابقي روابط مانند الگوریتم EM است[۱].

$$P_t^{\,i}(j) = \frac{\alpha_t(i).\beta_t(i)}{\sum_{j=1}^{N_s}\alpha_t(i).\beta_t(i)} \cdot \frac{c_{ij}.N(o_t; \mu_{ij}, \Sigma_{ij})}{\sum_{k=1}^{T} c_{ik}.N(o_t; \mu_{ik}, \Sigma_{ik})} \quad (۱۲)$$

روش GMM از نگاه خوشه‌بندي، ترکیبي خطي از خوشه‌هایي است که هر یک توسط مرکز و واریانس خوشه‌ها بازنمایي شده‌اند و با توجه به طبیعت فضاي خوشه‌بندي ویژگي‌ها، انتظار مي‌رود که مکان مرکز خوشه‌ها و واریانس آنها نسبت به روش گسسته تغییر زیادي نداشته باشد و مي‌توان از مقدار به دست آمده در روش گسسته به عنوان مقادیر اولیه روابط بازتخمین استفاده نمود (روش k-means). استفاده از مدل پیوسته، خطاي مدل گسسته را کاهش مي‌دهد. اما بیان مدل پیوسته به صورت GMM با تعداد مخلوط‌هاي محدود خود به خطایي مي‌انجامد که اگرچه نسبت به روش گسسته کمتر است، اما انگیزه‌اي براي ارائه روش‌هاي دیگر بازنمایي توزیع پیوسته با پارامترهاي محدود ایجاد مي‌کند.

ج- HMM با توزیع نیمه‌پیوسته [۱۱] (SCDHMM)

همانطور که قبلاً ذکر شد، عیب اصلي مدل HMM با توزیع گسسته، استفاده از مراکز خوشه‌ها و حذف جزئیات آماري توزیع هر خوشه است. یک دیدگاه دیگر براي تصحیح این مدل، استفاده از خوشه‌هایي است که توسط میانگین و واریانس هر خوشه بازنمایي شده‌اند، به عبارت دیگر، فرض مي‌شود که هر خوشه توسط یک مدل گاوسي قابل توصیف است. در صورتي که کل فضاي ویژگي‌ها به این صورت خوشه‌بندي شده باشد، هر بردار به میزان احتمال حضور در هر خوشه، عنصر آن خوشه است. به این ترتیب، با دانستن احتمال حضور هر خوشه در خروجي هر حالت، احتمال هر بردار در هر حالت به صورت زیر قابل بیان است:

$$P(o_t \mid s_t) = \sum_{j=1}^{M} P(o_t \mid \nu_j).P(\nu_j \mid s_t)$$
$$= \sum_{j=1}^{M} b_j(s_t).N(o_t; \mu_j, \Sigma_j) \quad (۱۳)$$

در رابطه فوق، v_k خوشه k ام را نشان مي‌دهد. به عبارت دیگر، در این مدل استفاده از بازنمایي GMM با پارامترهایي صورت مي‌گیرد که حاصل از خوشه‌بندي کل فضاي ویژگي‌ها به تعدادي خوشه است و توابع گاوسین به ازاي هر حالت، جداگانه انتخاب نمي‌شوند. این ایده، همانند گسسته ساختن فضا پیش از ورود به مدل HMM در DDHMM مي‌باشد و از سوي دیگر، استفاده از توزیع پیوسته ترکیب توابع گاوسي، مدل را به مدل CDHMM شبیه مي‌کند.

در این روش، پارامترهاي خوشه‌بندي کل فضاي ویژگي‌ها به خوشه‌هاي گاوسي $\{N(o;\mu_j,\Sigma_j)\}$ به ازاي کل مدل‌ها به دست مي‌آید و به ازاي هر واحد زباني، احتمال حضور هر خوشه در هر قطعه $\{b_j(s_t)\}$ و احتمال گذر از هر حالت به حالات بعدي $\{a_{ij}\}$ مشخصه هر مدل به‌شمار مي‌رود. در[۹] روابط بازتخمین این پارامترها ارائه شده است. همچنین روش‌هایي براي تخمین همزمان مشخصات خوشه‌ها و پارامترهاي هر مدل پیشنهاد شده است [۹].

استفاده از مدل SCDHMM، مزیت سرعت مدل‌سازي DDHMM را با دقت CDHMM ترکیب مي‌کند. در[۱۰،۹] نشان داده شده است که دقت SCDHMM در بازشناسي واج، حتي از CDHMM مي‌تواند بیشتر است که این واقعیت با کاهش تعداد پارامترها نسبت به CDHMM قابل توجیه است.

در مرحله بازشناسي، براي کاهش حجم محاسبات، امکان استفاده از M مخلوط با بیشترین احتمال به ازاي هر قطعه وجود دارد که M مي‌تواند همان تعداد مخلوط‌هاي CDHMM باشد. به این ترتیب، تعداد محاسبه احتمال در توزیع گاوسي که بیشترین حجم محاسبات را به خود اختصاص مي‌دهد، کاهش مي‌یابد.

د- استفاده از VQ با خوشه‌هاي گاوسي براي DDHMM

یک تصحیح دیگر مدل DDHMM براي کاهش خطا، گسسته کردن فضا با فرض گاوسي بودن توزیع هر خوشه است[۱۱, ۱۲, ۱۳]. در این رویکرد، فرض مي‌شود که هر بردار ویژگي تنها متعلق به یک توزیع گاوسي است و با یافتن این توزیع مي‌توان اندیس آن‌را به عنوان مشخصه توزیع در مدل HMM گسسته به‌کار برد.

براي یافتن خوشه‌هاي با توزیع‌هاي گاوسي، روش آموزش VQ باید تصحیح گردد. معیار فاصله براي تشخیص خوشه مناسب براي هر بردار ویژگي باید احتمال این خوشه به شرط دانستن بردار ویژگي باشد.

$$C^* = \underset{j}{ArgMax}\, P(C_j \mid O_t) = \underset{j}{ArgMax}\, P(C_j).N(O_t;\mu_j,\Sigma_j) \qquad (۱۴)$$

در رابطه بالا، $P(C_j)$ به صورت نسبت تعداد عناصر خوشه j ام بر تعداد کل بردارها تخمین زده مي‌شود.

استفاده از این معیار براي خوشه‌بندي فضا، آموزش VQ را از LBG به روشي همانند EM مبدل مي‌کند. خروجي این سیستم خوشه‌بندي پس از آموزش، یک دنباله اندیس است که براي آموزش و بازشناسي با مدل DDHMM مناسب خواهد بود. اگرچه به نظر مي‌رسد که این شیوه راهي براي نجات سیستم DDHMM بوده است، اما ایده استفاده از یک تابع گاوسي به ازاي هر بردار ویژگي در مرحله آموزش، فرضي متناسب با طبیعت گفتار مي‌باشد که نشان از عدم ارگادیک بودن پدیده گفتار دارد.

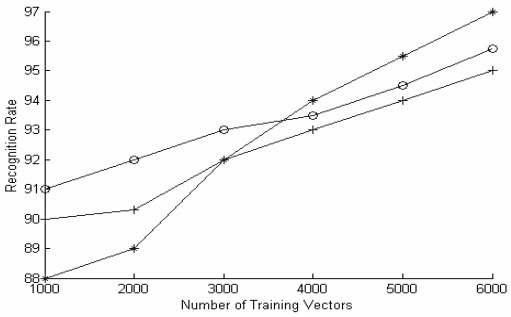
هـ - مقایسه روش‌هاي مختلف توصیف توزیع خروجي هر قطعه

براي مقایسه شیوه‌هاي مختلف بازنمایي توزیع قطعات، باید از واحدهاي زباني یکسان و بانک‌هاي اطلاعاتي وسیع، یکسان و استاندارد استفاده نمود. متاسفانه چنین مقایسه‌اي در منابع به‌خصوص در مورد طبقه‌بندي واج‌ها وجود ندارد و در هر یک از این منابع، تنها به بررسي انواع مورد علاقه با ابعادي محدود پرداخته شده است. یک مقایسه مناسب و قابل اعتماد در زمینه بازشناسي واج‌هاي زبان انگلیسي به صورت مستقل از متن در [۹] آمده است. اما متاسفانه نتایج به دست آمده بر روي بانک اطلاعاتي غیر استاندارد انجام شده است و حتي از مجموعه واج‌هاي استاندارد TIMIT نیز استفاده نشده است. در این آزمون، مدل HMM به صورت پیوسته، گسسته و نیمه‌پیوسته با هم مقایسه شده و براي پیاده‌سازي مدل CDHMM از ۵ مخلوط به ازاي هر حالت و براي مدل‌هاي DDHMM و SCDHMM از ۱۳۲ خوشه براي کل واج‌ها استفاده گردیده است. خلاصه این نتایج در جدول ۱ گردآوري شده است.

جدول ۱- مقایسه نرخ بازشناسي در توصیف‌هاي مختلف توزیع هر حالت مدل HMM

مدل	DDHMM	CDHMM	SCDHMM
نرخ بازشناسي	٪۵۳/۲	٪۵۸/۵	٪۶۰/۰

از مقایسه نتایج ارائه شده میان سه روش معمول مدل‌سازي توزیع در HMM به نظر مي‌آید که روش SCDHMM، روش دقیق‌تري در بازشناسي واج باشد. بهبود نتایج نسبت به DDHMM، نتیجه بازنمایي دقیق‌تر خوشه‌ها و بهبود نتایج نسبت به CDHMM، احتمالاً نتیجه دقت بهتر در تخمین پارامترها به علت کاهش آنها است. مقایسه دیگر انجام شده بین سه روش گسسته، پیوسته و نیمه‌پیوسته براي بازشناسي کلمات در [۱۰] آمده است.



شکل ۱- تغییرات نرخ بازشناسي در روش‌هاي مختلف توصیف توزیع هر قطعه در بازشناسي کلمات نسبت به تغییرات تعداد بردارهاي آموزشي (o: مدل نیمه‌پیوسته، +: مدل گسسته، *: مدل پیوسته)

این مقایسه در شکل ۱ به صورت منحني تغییرات نرخ خطا نسبت به تعداد نمونه‌هاي آموزشي رسم شده است. چنانچه مشخص است، در تعداد نمونه‌هاي کم، توصیف نیمه‌پیوسته بهتر از دو روش دیگر عمل مي‌کند، در حالي‌که با افزایش نمونه‌هاي آموزشي و کاهش نیاز به استفاده از پارامترهاي مشترک بین قطعات اکوستیکي، دقت سیستم پیوسته بیشتر مي‌شود. این منحني مي‌تواند معیاري براي انتخاب روش توصیف توزیع با توجه به تعداد نمونه‌هاي آموزشي ارائه دهد.

همچنین روش کتاب کد گاوسي براي DDHMM در مقاله‌هاي [۱۲،۱۱] مورد آزمون و مقایسه قرار گرفته است. این آزمون‌ها بر روي دو مجموعه از کلمات با لغت‌نامه وسیع انجام شده است که خلاصه این نتایج در جدول ۲ آمده است. مقایسه نتایج به دست آمده نشان مي‌دهد که در این مورد، سیستم GVQ-DDHMM نتیجه دقیق‌تري نسبت به روش‌هاي دیگر مدل‌سازي ارائه کرده است.

نکته قابل توجه این‌که در نتایج این آزمون، مدل نیمه‌پیوسته، مزیت خود را نسبت به CDHMM از دست داده است که احتمالاً ناشي از اختلاف در شرایط دو آزمون بوده است. این روش تاکنون در مورد طبقه‌بندي و بازشناسي واج‌ها آزموده نشده است.

جدول ۲- مقایسه نرخ بازشناسي در توصيف‌هاي مختلف توزيع هر حالت مدل HMM

مدل	GVQ-HMM	CDHMM	SCDHMM
نرخ بازشناسي(بانک اطلاعاتي اول)	٪۹۸/۳	٪۹۷/۹	٪۹۷/۶
نرخ بازشناسي(بانک اطلاعاتي دوم)	٪۹۵/۲	٪۹۴/۶	٪۹۳/۸

در هر دو آزمون اخير نشان داده شده است که در صورتي که طراحي کتاب کد توابع گاوسي با سرپرستي باشد، نتايج بهتري به دست مي آيد. به عبارت ديگر، در صورتي که هر واج به صورت جداگانه با يک يا چند تابع گاوسي مدل شود و کليه توابع به دست آمده از اين روش، مجموعه کتاب کد را تشکيل دهند، نرخ بازشناسي افزايش مي‌يابد. اين رويکرد به صورت ضمني به نزديک شدن نحوه مدل‌سازي‌هاي مذکور به CDHMM مي‌انجامد، زيرا مشخصه مدلCDHMM اين است که به ازاي هر حالت، توابع گاوسي جداگانه‌اي طراحي شود. به نظر مي‌آيد اين بهبود در نرخ بازشناسي، نشانگر توانايي مدل CDHMM در مقابل روش‌هاي ديگر است و عدم کارايي اين مدل بيشتر از عدم دقت در تخمين پارامترها ناشي مي‌شود تا نحوه گسسته کردن فضا. در اکثر منابع متاخرتر نيز بيشتر توجه به مدل CDHMM متمرکز شده است[۱۰].

۳- مدل GVQ

۳-۱ ايده اصلي

اکثر مدل‌هاي آماري از تعداد محدودي تابع گاوسي به عنوان پايه فضاي توابع براي تقريب توزيع چگالي احتمال هر درايه بردار ويژگي استفاده مي‌کنند. اين رويکرد به عنوان رويکردي عمومي هر شکل تابعي را مي‌تواند به تعدادي تابع گاوسي پايه تجزيه کند که اين توابع و ضرايب آنها با معيار افزايش احتمال توليد داده‌هاي آموزشي با الگوريتم EM بهينه مي‌شوند.

در رويکرد VQ، فضا به تعدادي خوشه گسسته مي‌شود و بردار ورودي تنها به يکي از خوشه‌ها وابسته است. در اين رويکرد، پس از انتساب هر بردار ويژگي به يک خوشه، شماره آن خوشه به عنوان مشخصه بردار ويژگي جانشين بردار ويژگي مي‌شود و مشخصات خوشه به عنوان مشخصات بردار ويژگي در مدل‌سازي به کار مي‌رود. اين مشخصات معمولاً به انديس خوشه منحصر است و مدل‌سازي‌هاي آماري ديگر بر مبناي دنباله انديس‌هاي خوشه‌ها انجام مي‌پذيرد. معيارهايي که براي انتساب يک بردار به يک خوشه آزموده شده است، فاصله اقليدسي و فاصله اقليدسي نرماليزه شده نسبت به واريانس خوشه است [۱۴].

عيب اصلي استفاده از VQ به اين شيوه، خطاي حاصل از گسسته‌سازي فضا و از دست دادن اطلاعات آماري مکان دقيق بردار ويژگي ورودي در فضاي ويژگي‌هاست. اين خطا، نرخ بازشناسي مدل‌هاي مبتني بر VQ را محدود مي‌کند.

در رويکرد GVQ، هر بردار ويژگي تنها به يک خوشه آکوستيکي تعلق دارد.

در مرحله بازشناسي، پس از انتخاب نزديک‌ترين خوشه به بردار ويژگي، به جاي جانشين ساختن انديس خوشه، از احتمال حضور بردار مذکور در خوشه منتخب به عنوان احتمال آن بردار در قطعه مورد بررسي λ_k استفاده مي‌شود.

$$P(O_i \mid \lambda_k) = \underset{j}{ArgMax}\, b_j(k).N(O_i;\mu_{jk},\Sigma_{jk}) \qquad (۱۵)$$

انتخاب يک خوشه از ميان خوشه‌هاي مختلف به رويکرد VQ شباهت دارد. از سوي ديگر، استفاده از احتمال حضور بردار ويژگي در يک قطعه از مدل GMM الهام گرفته شده است.

مدل SCDHMM نيز به ترکيب مدل‌هاي VQ و GMM مي‌پردازد [۹].

در اين مدل، تلاش اصلي به کاهش تعداد پارامترهاي GMM معطوف شده و خوشه‌هاي مورد استفاده در GMM طراحي شده به ازاي هر قطعه، از مجموعه خوشه‌هايي انتخاب مي‌شود که براي بازنمايي کل فضاي بردارهاي ويژگي طراحي شده‌اند.

$$P(O_i \mid \lambda_k) = \sum_{j=1}^{M} b_j(k).N(O_i;\mu_{jk},\Sigma_{jk}) \qquad (۱۶)$$

در اين مدل، احتمال انتخاب خوشه j ام توسط قطعه k ام ($b_j(k)$) و پارامترهاي خوشه‌هاي فضاي ويژگي‌ها مشخصه مدل هر قطعه در نظر گرفته مي‌شود. بنابراين ضريب $b_j(k)$ به ازاي يک قطعه ثابت است. در مقابل در مدل GVQ اين ضريب توسط بردار ويژگي ورودي انتخاب مي‌شود.

توصيف GMM و VQ کلاسيک به صورت ضمني بردارهاي نمونه يک قطعه را ارگاديک فرض مي‌کند. در اين توصيف رفتار آماري کليه بردارهاي ويژگي با يک ترکيب خطي بيان مي‌شود. در حالي که در عمل اين‌گونه نيست و خوشه‌هاي آکوستيکي مختلفي به يک واج منتسب مي‌شوند و عملاً در يک گفتار خاص، تنها يکي از خوشه‌ها ظاهر شده است.

مدل GVQ فرض ارگاديک بودن دنباله بردارهاي ويژگي را مورد بازبيني قرار داده است. انتخاب نزديکترين خوشه (يا خوشه‌ها) توسط مدل GVQ به معناي عدم تاثير کل خوشه‌ها در بازنمايي يک قطعه از هر واحد زباني است و در يک بردار ورودي مشخص، تنها خوشه‌اي تاثير داده مي‌شود که بردار ويژگي متعلق به آن است.

پارامترهاي مختلفي موجب خدشه‌دار شدن فرض ارگاديک بودن در دنباله بردارهاي ويژگي يک واحد زباني مي‌گردد. مدل‌سازي همزمان لهجه‌هاي مختلف، افراد با سنين مختلف يا با جنسيت غيريکسان از دلايل اصلي ايجاد خوشه‌هاي مختلف است، در حالي که در يک قطعه از واحد زباني ورودي، تنها يکي از اين شرايط اتفاق افتاده است.

عدم ارگاديک بودن دنباله ورودی در HMM قطعه‌اي[۱۲](SHMM) نيز مورد توجه بوده است [۱۶،۱۵]. در اين مدل، پارامترهاي توزيع هر قطعه (معمولاً ميانگين آن) متناسب با بردارهاي ويژگي قطعه ورودي تخمين زده مي‌شود. در مقابل در GVQ انتخاب تابع گاوسي از ميان تعداد محدودي تابع گاوسي از پيش تعيين‌شده انجام مي‌گيرد و اين انتخاب به ازاي هر بردار ويژگي تکرار مي‌شود.

شيوه انتخاب گاوسين[۱۳](GS) رويکرد ديگري شبيه GVQ است که با خاستگاهي متفاوت مطرح شده است [۱۴, ۱۷, ۲۰]. اين ايده با تمرکز بر تسريع روش‌هاي بازشناسي کلمات با لغت‌نامه وسيع، زمان بازشناسي را با محاسبه ترکيب GMM در تعداد اندکي از توابع گاوسي انتخاب شده کاهش مي‌دهد. اين شيوه بر اين مبنا پايه‌گذاري شده است که در عمل تعداد کمي از خوشه‌ها در محاسبه تابع درست‌نمايي نقش غالب دارند و محاسبه به شيوه GS بدون کاهش قابل ملاحظه نرخ بازشناسي، زمان بازشناسي را به صورت جدي تحت تاثيـــر قـــرار مي‌دهد. اين شيوه، روند آموزش مدل را تغيير نمي‌دهد و تنها به عنـوان رونـــدي تکميلي براي تسريع مرحله بازشناسي مورد استفاده واقع مي‌شود. با توجه به خاستگاه اين روش، اين شيوه در بازشناسي واج آزموده نشده است و همچنين روند آموزش مدل تغيير نيافته است.

۳-۲ مدل تحليلي GVQ

استفاده از مدل GVQ نيازمند تعريف معيار فاصله مناسب براي انتخاب تابع گاوسي و شيوه‌اي براي آموزش پارامترهاي خوشه‌هاي گاوسي است. معيار مورد استفاده براي اين مدل احتمال هر خوشه به شرط حضور بردار ورودي بوده است.

$$P(C_i \mid X_j) = \frac{P(X_j \mid C_i).P(C_i)}{P(X_j)} \qquad (۱۷)$$

در رابطه فوق، C_i خوشه مورد نظر و X_j بردار ورودي مي‌باشد. انتخاب بهترین تابع گاوسي با جستجوي خوشه با بیشترین احتمال صورت مي پذیرد.

$$C_{j^*} = \underset{i}{ArgMax}(P(C_i \mid X_j) = \underset{i}{ArgMax}(P(X_j \mid C_i).P(C_i)) = \underset{i}{ArgMax}\, N(X_j;\mu_i,\Sigma_i).\alpha_i \qquad (۱۸)$$

که در رابطه فوق، α_i نشان‌دهنده احتمال یک خوشه در میان خوشه‌هاي دیگر و μ_i و Σ_i به ترتیب میانگین و واریانس خوشه i ام است.

آموزش GVQ بر مبناي جستجوي بهترین خوشه به ازاي هر بردار آموزشي و تصحیح پارامترهاي خوشه‌هاي گاوسي بر مبناي عناصر منتسب به آن‌ها صورت مي‌پذیرد. این فرایند آن‌قدر ادامه مي‌یابد که رشد تابع درست‌نمایي کل مجموعه بازشناسي از $\sqrt[3]{10}$ کمتر شود. این آستانه با توجه به این استنتاج شده است که به صورت تجربي کارایي آستانه ۱۰ به ازاي کل یک واج نشان داده شده است [۲۱] و در این پژوهش هر واج به سه قطعه هم‌طول تقسیم شده است. تابع درست‌نمایي کل مجموعه آموزشي به صورت زیر تعریف مي‌شود:

$$P = \prod_{j=1}^{N_o} N(X_j;\mu_{j^*},\Sigma_{j^*}).\alpha_{j^*} \qquad (۱۹)$$

N_o نماینده تعداد کل بردارهاي آموزشي هر قطعه است. بیان عبارت فوق به صورت لگاریتمي مي تواند نحوه خوشه‌بندي را بهتر بازنمایي کند:

$$LLF = \log(P) = \sum_{j=1}^{N_o} \log\left(\frac{\alpha_{j^*}}{\left|\Sigma_{j^*}\right|}\right) + \sum_{j=1}^{N_o}\sum_{k=1}^{p}\left(\frac{(O_{j^*}^k - \mu_{j^*}^k)}{\sigma_{j^*}^k}\right)^2 \qquad (۲۰)$$

در رابطه بالا از مقادیر ثابتي که در فرایند بهینه‌سازي تاثیر ندارند، صرف‌نظر شده است. p و O_j^k به ترتیب بعد و درایه k ام بردار ویژگي ورودي را نشان مي‌دهد. مقایسه این رابطه با VQ استاندارد نشان‌دهنده استفاده از معیار خوشه‌بندي فاصله اقلیدسي نرمالیزه‌شده نسبت به واریانس است که یک جمله به صورت $\sum_{j=1}^{N_o}\log\left(\frac{\alpha_{j^*}}{\left|\Sigma_{j^*}\right|}\right)$ به آن افزوده شده است. عبارت اخیـــر هنگامي حداکثـر مي‌شود که:

$$\alpha_1 = \alpha_2 = \ldots = \alpha_M \qquad (۲۱)$$

بنابراین جمله افزوده شده سعي دارد بـــردارهاي ویژگي را به صـــورتي عادلانه در خوشه‌ها پخش کند به صورتي که جمعیت بردارهاي هر خوشه یکسان باشد. این توصیف به صورت ضمني بیان مي‌دارد که حتي اگر فضاي بردارهاي ویژگي داراي خوشه‌هاي واضحي نباشد [۷], این شرط، خوشه‌بندي مناسبي را براي بازنمایي فضاي ویژگي‌ها پیشنهاد مي‌کند. در عمل نیز نشان داده شده است که استفاده از معیار فاصله فوق موجب خوشه‌بندي مناسب فضاي ویژگي خواهد شد.

در مرحله بازشناسي نیز تابع درست‌نمایي به صورت ضرب احتمال بردارهاي ویژگي در خوشه‌هاي انتخاب شده محاسبه مي‌گردد:

$$LLF = \prod_{j=1}^{T} \underset{i}{ArgMax}(N(X_j;\mu_i,\Sigma_i)) \qquad (۲۲)$$

که در حالت GS با Q مخلوط گاوسي، از ترکیب خطي Q نزدیکترین خوشه نزدیک به بردار O_j استفاده مي‌شود:

$$LLF = \prod_{j=1}^{T} \frac{1}{\sum_{q=1}^{Q}\alpha_q} \sum_{q=1}^{Q} \alpha_q.N(X_j;\mu_q,\Sigma_q) \qquad (۲۳)$$

لازم به ذکر است که انتخاب مدل گاوسي با توجه به احتمال هر خوشه (α_i) انجام پذیرفته است، چنانچه در مرحله آموزش هم از این ضریب استفاده شده است، اما در مرحله بازشناسي با انتخاب خوشه، احتمال هر بردار تنها در خوشه منتخب محاسبه مي‌شود.

این مدل‌سازي جانشین مناسبي براي مدل عمومي GMM در دو مرحله آموزش و بازشناسي به‌شمار مي‌رود. در مرحله آموزش، آزمایش نشان‌دهنده همگرایي سریع‌تر تخمین پارامترهاي مدل است و طبیعت غیرخطي انتخاب تابع گاوسي در GVQ، شیوه‌اي مناسب را براي مدل‌سازي رفتار غیرارگادیک گفتار در مرحله بازشناسي فراهم مي‌آورد. خوشه‌هاي حاصل از GVQ می تواند مقدار اولیه مناسبی نیز برای آغاز الگوریتم باز تخمین EM در GMM به‌شمار رود. همچنین استفاده از GVQ، مي‌تواند زمینه را براي کاهش زمان بازشناسي با استفاده از GS ایجاد کند.

مشکل اصلي این روش، همگرایي تضمین نشده آموزش GVQ است به صورتي که نمي‌توان از افزایش همیشگي تابع درست‌نمایي در هر مرحله بازتخمین اطمینان داشت. در حالي که آموزش GMM از این مزیت برخوردار است.اما در عمل در تعداد اندکي از موارد، کاهش تابع درست‌نمایي در روند بازتخمین مشاهده شده که مقدار این کاهش نیز اندک بوده است.

۳-۳ آزمون و ارزیابي توصیف GVQ در مدل قطعه‌بندي نرم

مدل پیشنهادي بر روي بانک اطلاعاتي TIMIT مورد ارزیابي قرار گرفته است. بانک اطلاعاتي TIMIT که توسط شرکت TI[۱۴] و دانشگاه MIT[۱۵] تهیه شده است [۲۳،۲۲] و اداره استانـــدارد آمـــریکا (NIST)[۱۶] آنـــرا تایید کـــرده است [۲۴], یک بانک اطلاعـاتي از گفتـــار پیوسته انگلیسي مي‌باشد که از روي متن خوانـــده شده است. کلیه جملات، کلمات، و واج‌هاي این بانک اطلاعاتي داراي برچسب زماني مي‌باشد.این بانک اطلاعاتي شامل ۶۳۰۰ جمله از ۶۳۰ گوینده و با ۸ لهجه معمول آمریکاي شمالي است.

۷۰٪ گوینده‌ها مرد و ۳۰٪ آنها زن هستند. ۱۰ جمله توسط هر گوینده ادا مي‌شود که ۲ جمله آن میان کلیه گوینده‌ها مشترک است. در مجموع ۲۳۴۲ جمله متمایز در TIMIT وجود دارد که شامل ۲ جمله مشترک میان همه گوینده‌ها (جملات با پیشوند sa)، ۴۵۰ جمله مشترک میان دسته‌هاي ۷ نفري گوینده‌ها (جملات با پیشوند sx) و ۱۸۹۰ جمله تک گوینده (جملات با پیشوند si) است. بانک اطلاعاتي TIMIT به دو بخش آمـوزش و آزمـون تقسیم شده است.

این دو بخش چه از لحاظ گویندگان و چـه از لحـــاظ جملات گفته شده مجـــزا هستند بخش آمـوزش شامل ۴۶۲ گوینـــده و بخش آزمـــون شامل ۱۶۸ گوینـــده است. در کلیه آزمایش‌هاي انجام شده براي تخمین پارامترها از کل مجموعه آمـوزش و در مرحله بازشناسي از کل مجموعه آزمون استفاده شده است. مجموعاً در مجموعه آموزش ۱۴۲۹۱۰ آوا و در مجموعه آزمـــون ۵۱۹۶۷ آوا وجود دارد.

این بانک به اطلاعاتي با ۶۱ آوا، آوانویسي شده است. در جدول ۳ این آواها و تقسیم‌بندي آنها دسته‌هاي مختلف آوایي ذکر شده است.

جدول ۳- آواهای انگلیسی در TIMIT

دسته آوايي	نام در TIMIT	معادل فارسي	دسته آوايي	نام در TIMIT	معادل فارسي
vowels	iy	i	Nasals	em	
	ih			en	
	eh	e		eng	
	ey			nx	
	ae	@	Fricatives	s	s
	aa	a		sh	$
	aw			z	z
	ay			zh	#
	ah			f	f
	ao			th	
	oy	o		v	v
	ow	w		dh	
	uh		Affricatives	jh	j
	uw			ch	c
	ux	u	Stops	b	b
	er			d	d
	ax			g	g
	ix			p	p
	axr			t	t
	ax-h			k	k
Semivowels	l	l		dx	
	r	r		q	
	w		Pauses And Silences	pau	
	y	y		epi	
	hh	h		bcl	
	hv			dcl	
	el			gcl	
Nasals	m	m		pcl	
	n	n		Tcl	
	ng			kcl	
				h#	

همچنين آواهايي كه معـــادل آنها در زبـــان فارسي وجود دارد، با * علامت‌گذاري شده‌اند. واضح است كه واج‌هاي "خ" و "ق" معادلي در زبان انگليسي ندارند، ولي معادل مابقي واج‌هاي فارسي در TIMIT وجود دارد.در نگارش دنباله واج‌هاي زبان انگليسي از علائم كمتري نسبت به آواها استفاده مي‌شود. در واقع آواهـــا واحـــدهاي آكوستيكي زبان را نشان مي‌دهند، اما يک واج ممكن است به صورت چند آوا ادا شود. جدول ۴ تناظر تقسيم‌بندي آواهاي TIMIT را به ۳۹ كلاس واجي زبان انگليسي نشان مي‌دهد. اين ۳۹ دسته، كلاس‌هاي معمول مورد استفاده در ارزيابي سيستم‌هاي طبقه‌بندي و بازشناسي واج مي‌باشد.

در پياده‌سازي سيستم، در بخش پيش‌پردازش، سيگنال صحبت با نرخ ۱۶KHzنمونه‌برداري مي‌شود و از فيلتر پيش‌تاكيد عبور مي‌كند.سپس دنبالـــه

جدول ۴- تقسيم بندی آواهایTIMIT به ۳۹ کلاس واجی انگلیسی

شماره کلاس	کلاس	شماره کلاس	کلاس
۱	k	۲۱	aa ao
۲	l el	۲۲	ae
۳	m em	۲۳	ah ax ax-h
۴	n en nx	۲۴	aw
۵	ng eng	۲۵	ay
۶	ow	۲۶	b
۷	oy	۲۷	ch
۸	p	۲۸	d
۹	r	۲۹	dh
۱۰	s	۳۰	dx
۱۱	sh zh	۳۱	eh
۱۲	t	۳۲	er axr
۱۳	th	۳۳	ey
۱۴	uh	۳۴	f
۱۵	uw ux	۳۵	g
۱۶	v	۳۶	hh hv
۱۷	w	۳۷	ih ix
۱۸	y	۳۸	iy
۱۹	z	۳۹	jh
۲۰	bcl dcl gcl kcl pcl tcl epi q pau h#		

نمونـــه‌هاي زماني به قاب‌هاي به طول ۱۶ms (۲۵۶ نقطه) با شكل همينگ و با همپوشاني ۵۰٪ تقسيم مي‌شود. در صورتي كه تعداد نقاط، مضرب صحيحي از ۱۲۸ نباشد، به طول واحد زباني به ميزان كافي سكوت (نمونه با اندازه صفر) اضافه مي‌گردد.

در بخش استخراج ويژگي، از هر قاب ۱۲ ضريب ويژگي كپستروم درمقياس مل (MFCC) (C_1 تا C_{12}) و ويژگي‌هاي ديناميكي درجه يک متناظر با آن در طول واحد زباني (ΔC_1 تا ΔC_{12}) استخراج مي‌گردد.

ويژگي‌هاي محاسبه‌شده C_i^t و ΔC_i^t همراه با ويژگي ديناميكي تغييرات لگاريتم انرژي قاب كه به صورت رابطه فوق محاسبه مي‌شود، بردار ويژگي ۲۵ عضوي نهايي را مي‌سازند. هر يک از درايه‌هاي بردارهاي ويژگي روي كل مجموعه بانک اطلاعاتي آموزشي نسبت به ميانگين و واريانس همسان‌سازي[۱۷] شده‌اند.

اين همسان‌سازي، توزيع هر درايه را در مجموعه واج‌ها به صورت توزيعي با ميانگين صفر و واريانس يک تبديل مي‌كند و ارزش نسبي درايه‌هاي مختلف را يكسان مي‌سازد. در مرحله انطباق با مدل ها, مدل پيشنهادی اين مقاله و همچنين مدل های استاندارد جايگزين گرديده و با هم مقايسه شده اند. در كليه مدل ها از توابع گاوسی با ماتريس كوواريانس قطری به عنوان پايه فضای توابع استفاده شده است. به عبارتی ديگر, فرض شده است كه درايه های مختلف بردارهای ويژگی از هم مستقل اند. استفاده از ماتريس كوواريانس قطری در توابع گاوسی محاسبه تابع درست‌نمايي و روند تخمين پارامترها را چه از لحاظ تحليلي و چه از لحاظ محاسباتي تسهيل كرده است. خروجي بلوک انطباق با مدل‌ها، واجي است كه مدل آن بيشترين انطباق را با واحـــدهاي زباني ورودي داشته است. در اين پژوهش، اين تصميم‌گيري بـــدون اعمـــال هيچگونه مدل‌زباني بـــراي توالي واج‌ها استخراج شده است تا نقش مدل آكوستيكي به‌تنهايي قابل بررسي باشد.

مدل GVQ پیشنهادي بر روي مدل آکوستیکي قطعه‌بندي نرم مورد ارزیابي قرار گرفته است. در اولین آزمون، روش آموزش و بازشناسي GVQ با روش استاندارد مقایسه شده است. جدول ۵، نتیجه این ارزیابي را به ازاي تغییر شیوه آموزش از EM به GVQ و همچنین تغییر شیوه بازشناسي از GMM به انتخاب گاوسي به ازاي هر بردار نشان مي‌دهد.

جدول ۵- مقایسه نرخ طبقه بندی واج در دو روش توصیف توزیع GVQ و GMM

روش آموزش	روش بازشناسي	مردان	زنان	متوسط
GVQ	1-Best Gaussian	۶۲/۵۴	۶۲/۱۴	۶۲/۴۲
GVQ	GMM	۶۲/۶۱	۶۲/۲۳	۶۲/۵۰
EM	1-Best Gaussian	۶۲/۷۲	۶۲/۵۶	۶۲/۶۷
EM	GMM	۶۲/۵۰	۶۳/۱۰	۶۲/۶۸

جدول ۶- مقایسه زمان آموزش (به ثانیه) در دو روش توصیف توزیع GVQ و EM (همپوشاني حداکثر)

روش آموزش	مردان	زنان
GVQ	۳۵۴۲۸	۱۱۴۶۳۷
EM	۱۱۰۹۸۰	۲۵۷۹۳۳

این آزمون با روش آموزش و بازشناسي قطعه‌بندي نرم و با حداکثر میزان همپوشاني قطعات و استفاده از ۸ تابع گاوسي به ازاي هر بردار ویژگي انجام پذیرفته است. ردیف آخر جدول نشان‌دهنده نرخ بازشناسي حاصل توسط شیوه‌هاي آموزش EM و بازشناسي GMM است. ردیف سوم، نرخ بازشناسي را به ازاي تغییر روش آموزش از EM به GVQ و ردیف دوم، نرخ بازشناسي را به ازاي تغییر روش بازشناسي از GMM به انتخاب بهترین تابع گاوسي نشان مي‌دهد. ردیف اول نیز نشانگر تاثیر همزمان هر دو تغییر است.

دیده مي‌شود که نرخ بازشناسي تغییر چشمگیري نداشته است. در مجموعه مردان بهترین نرخ بازشناسي با استفاده از آموزش EM و بازشناسي GVQ حاصل شده، در صورتي‌که فاصله این بهبود با بدترین شرایط تنها ۰/۲٪ است. در مجموعه زنان نیز بهترین نرخ بازشناسي توسط شیوه استاندارد حاصل شده است که نزدیک ۱٪ با بدترین شرایط متفاوت است.

در مجموع مي‌توان اینگونه جمع‌بندي کرد که نرخ بازشناسي در این شرایط تغییر جدي نکرده است، اما کارایي به دست آمده در جدول ۶ و در زمان آموزش قابل مشاهده است که همگرایي الگوریتم آموزش در GVQ نزدیک ۳ برابر سریعتر از روش EM حاصل شده است. همچنین استفاده از GVQ امکان استفاده از روش هاي انتخاب گاوسین را مهیا مي‌کند که به تسریع بازشناسي مي‌انجامد. کاربرد دیگر GVQ مي‌تواند تخمین مقادیر اولیه الگوریتم EM باشد. این کاربرد روي واج‌هاي TIMIT آزموده شده است و نتایج آن در جدول ۷ آورده شده است. استفاده از GVQ به عنوان مقادیر اولیه حدود ۰/۵٪ نتایج نهایي را بهبود بخشیده است. این میزان اگرچه چشمگیر نیست، اما نشان‌دهنده آن است که الگوریتم EM از مقادیر اولیه بهتري بهره گرفته است.

جدول ۸ همین آزمون را به ازاي میزان همپوشاني بهینه قطعات در قطعه‌بندي نرم نشان مي‌دهد. در این همپوشاني، (OLC=۰/۳) بهبود کمتري (حدود ۰/۱۵٪) در نرخ بازشناسي دیده مي‌شود که نشانگر این واقعیت است که نرخ بازشناسي واج تقریباً اشباع شده و تغییر آن به سختي امکان‌پذیر است.

جدول ۷- مقایسه نرخ طبقه بندی واج (بهمپوشانی حداکثر- بازشناسی GMM)

روش آموزش	مردان	زنان	متوسط
GVQ	۶۲/۶۱	۶۲/۲۳	۶۲/۵۰
EM	۶۲/۵۰	۶۳/۱۰	۶۲/۶۸
GVQ+ EM	۶۲/۸۹	۶۳/۷۹	۶۲/۱۶

جدول ۸- مقایسه نرخ طبقه بندی واج (بهمپوشانی حداکثر- بازشناسی GMM)

روش آموزش	مردان	زنان	متوسط
GVQ	۶۲/۰۲	۶۴/۷۵	۶۴/۹۴
EM	۶۴/۸۵	۶۴/۶۴	۶۴/۷۹
GVQ+ EM	۶۴/۹۱	۶۴/۶۸	۶۴/۸۴

جدول ۹- مقایسه نرخ بازشناسي در استفاده از تعداد توابع گاوسي متفاوت در بازشناسي (آموزش GVQ - همپوشاني بهینه)

روش بازشناسي	مردان	زنان	متوسط
1-best Gaussian	۶۴/۹۴	۶۴/۷۹	۶۴/۹۰
2-best Gaussians	۶۴/۹۷	۶۴/۸۲	۶۴/۹۳
4-best Gaussians	۶۵/۰۲	۶۴/۸۲	۶۴/۹۶
8-best Gaussians	۶۵/۰۲	۶۴/۷۵	۶۴/۹۴

جدول ۱۰- نرخ بازشناسي مدل GVQ تغییر یافته در قطعه‌بندي نرم

روش آموزش	مردان	زنان	متوسط
GVQ تغییر یافته	۱۵/۳۴	۱۱/۸۵	۱۴/۲۹

جدول ۹، توانایي روش انتخاب گاوسین را به ازاي تعداد توابع گاوسي مختلف بررسي مي‌کند. به‌نظر مي‌آید تغییرات نرخ بازشناسي ناچیز و قابل صرف‌نظر است. این نتیجه از این جهت حائز اهمیت است که نشان مي‌دهد در طبقه‌بندي واج‌ها، یک تابع گاوسي براي بازنمایي هر بردار ویژگي کافي است و افزایش تعداد توابع گاوسي، به نرخ بازشناسي نمي‌افزاید. این نتیجه زمینه را براي تسریع سرعت بازشناسي با استفاده از روش انتخاب گاوسین مهیا مي‌کند.

یک روش دیگر نیز در زمینه مدل‌سازي غیرارگادیک دنباله گفتار، مورد آزمون قرار گرفته است. در این روش که ایده اصلي آن از SHMM الهام گرفته شده است،یکي از هشت خوشه برای کل قطعه انتخاب می شود و تابع درست نمایی کل قطعه بر مبنای یک خوشه منتخب محاسبه می گردد. این روش که GVQ تغییر یافته نام‌گذاري شده است، بر روي مجموعه آزمون TIMIT آزموده شده و به نرخ‌هاي بازشناسي بسیار پایینی منتهي شده است (جدول ۱۰).

این نتایج نامقبول حاکي از عدم امکان مدل‌سازي مناسب یک قطعه آکوستیکي از گفتار با یک خوشه گاوسي است. به عبارت دیگر اگرچه یک تابع گاوسي براي توصیف هر بردار ویژگي در یک قطعه آکوستیکي کافي است، ولي خوشه غالب از ابتدا تا انتهاي یک قطعه ثابت نمي‌ماند و خوشه مناسب یک قطعه در طول زمان قطعه، از خوشه‌اي به خوشه دیگر مهاجرت مي‌کند. این نتیجه، فرض SHMM را در مورد مدل‌سازي دینامیک درون‌قطعه‌اي به صورت جدي مشکوک مي‌نمایاند و لزوم استفاده از ایده مهاجرت را تقویت می کند.

۴- جمع‌بندي و نتيجه‌گيري

در اين مقاله روش GVQ براي مدل‌سازي توزيع آماري هر قطعه آكوستيكي گفتار پيشنهاد و آزموده شده است. مبناي اين بازنمايي تاكيد بر اين واقعيت بوده است كه پيچيدگي توزيع آماري بردارهاي ويژگي هر قطعه آكوستيكي حاصل تنوع در گوينده، جنسيت، لهجه، و شرايط محيطي است و اين تنوع‌ها همگي در يک گفتار خاص به وقوع نمي‌پيوندد، بلكه هر بردار با معلوم بودن شرايط فوق از توزيعي ساده پيروي مي‌كند. بنابراين با محدود ساختن توزيع آماري، نه تنها از حجم محاسبات چه در زمان آموزش و چه در زمان بازشناسي كاسته شده است، بلكه انطباق بهتر با شرايط طبيعي گفتار موجب شده است كه اين ساده‌سازي باعث كاهش نرخ بازشناسي نشود.

در اين زمينه روش آموزش EM و معيار بهينه‌سازي آن تعميم داده شده و نشان داده شده است كه معيار حاصل فضا را به شيوه‌اي عادلانه‌تر از معيارهاي پيشين به خوشه‌هاي گاوسي مي‌شكند. نتايج تجربي آزمون‌هاي انجام شده نيز مؤيد موفقيت اين مدل‌سازي است.

در مقابل، ديده مي‌شود كه يک توزيع گاوسي ثابت، براي توصيف يک قطعه آكوستيكي كافي نيست. اين نتيجه كه به وضوح با فرض HMM قطعه‌اي در تضاد است. ايده مسير مهاجرت را از خوشه‌اي به خوشه ديگر در طول يک قطعه مطرح مي‌كند كه پژوهش در اين زمينه در حال انجام است. همچنين ايده GVQ با استفاده از روش‌هاي انتخاب تابع گاوسي مي‌تواند به صورت قابل ملاحظه‌اي سرعت بازشناسي را افزايش دهد. در اين زمينه نيز پژوهش ادامه دارد.

مراجع

[1] L. R. Rabiner, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMM To Segment Models : A Unified View of Stochastic Modeling of Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 360-378, 1996.

[3] M. Ostendorf, and S. Roucos, "A Stochastic Segment Model for Phoneme Based Continuous Speech Recognition," *IEEE Trans. on Audio Speech and Signal Processing*, vol. 37, no . 12, pp. 1857-1869,1989.

[4] F. Razzazi, and A. Sayadiyan, "Evaluation of Soft Segment Modeling on a Phoneme Recognition System," 10th *IEEE Intl Conf. on Electronics, Circuits and Systems*, ICECS 2003, 2003.

[۵] ف . رزازي، و الف . صيادیان، **ارائه يک مدل قطعه بندي نرم براي مدل‌سازي كشش زماني در سيستم هاي بازشناسي گفتار**, دهمين كنفرانس مهندسي برق ايران، شيراز، ارديبهشت ۱۳۸۲.

[6] F. Razzazi, and A. Sayadiyan, " Soft Segment Modeling, A Robust Duration Modeling for Speech Recognition", 9th *Intl Speech Science and Technology Conf., SST2003*, 2002.

[7] T. Kinnunen, I. Karkkainen, and P. Franti, "Is Speech Data Clustered? – Statistical Analysis of Cepstral Features", *Proc. of Eurospeech*, vol. 4, pp. 2627-2630, 2001.

[8] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. on Patt. Anal. and Machine Intell.* , vol. PAMI-5, pp. 179-190, 1983.

[9] X. D. Huang, "Phoneme Classification Using Semi continuous Hidden Markov Models," *IEEE Trans. on Signal Processing*, vol. 40, no. 5, pp. 1062-1067, 1992.

[10] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*: *A Guide to Theory, Algorithm, and System Development*, Prentice- Hall, 2001.

[11] Y. Zhang, M. Alder, and R. Togneri, "Using Gaussian Mixture Modeling in Speech Recognition," *Proc. of IEEE ICASSP* pp. I.613-I.616, 1994.

[12] Y. Zhang, R. Togneri, and M. Alder, "Phoneme Based Vector Quantization in a Discrete HMM Speech Recognizer, " *IEEE Trans. on Speech and Audio Processing*, vol. 5, no.1, pp. 26- 32, 1997.

[13] Y. Zhang, and M. Alder, " An Improved HMM/VQ Training Procedure for Speaker Independent Isolated Word Recognition," *Intl Symp. on Speech, Image Processing and Neural Network*, pp. 722-725, 1994.

[14] E. Bocchieri, "Vector Quantization for Efficient Computation of Continuous Density Likelihoods," *Proc. of IEEE ICASSP*, pp. 692-695, 1993.

[15] W. J. Holmes, and M. J. Russel, "Modeling Speech Variability with Segmental HMMs," *Proc. of IEEE ICASSP*, pp. 447-450, 1996.

[16] W. J. Holmes, and M. J. Russel, "Experimental Evaluation of Segmental HMMS," *Proc. of IEE ICASSP*, pp.536-539, 1995.

[17] K. M. Knill, M. J. F. Gales, and S. J. Young "Use of Gaussian Selection in Large Vocabulary Continuous Speech Recognition Using HMMs," *Proc. ICSLP*, pp. 470-473, 1996.

[18] A. Lee, T. Kawahara, and K. Shikano, "Gaussian Mixture Selection Using Context-Independent HMM," *Proc. of IEEE ICASSP*, pp. 69-72, 2001.

[19] A. Lee, T. Kawahara, and K. Shikano, "Gaussian Mixture Selection Using Context-Independent HMM," *Information Processing Society of Japan Journal*, 2002.

[20] D. Paul, "An Investigation of Gaussian Shortlists," *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 209-212, 1999.

[21] C. Becchotti, and L.P. Ricotti, *Speech Recognition, Theory and C++ Implementation*, Wiley, 1999.

[22] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA Speech Recognition database: Specification and Status", *Proc. of the Speech Recognition Workshop*, Palo Alto, pp. 93-96,1986.

[23] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallette, and N. Dehlgren, "DARPA TIMIT Acoustic-phonetic Continuous Speech Corpus CD-ROM," *Technical Report NISTIR 4930, National Institute of Standards and Technology*, 1993.

[24] L. Lamel, R. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. of the Speech Recognition Workshop*, Report no. SAIC-86/1546, pp. 100-109, 1986.

[1]Hidden Markov Model
[2] Constrained Mean Trajectory Stochastic Segment Modeling
[3] Gaussian Vector Quantization
[4] Mel Frequency Cepstrum Coefficient
[5] Gaussian Mixture Model
[6] Contirous Density HMM
[7] Discrete Density HMM
[8] Loyde-Buzzo-Gray
[9] Vector Quantization
[10] Expectation-Maximiziation
[11] Semi-Continuaus Density HMM
[12] Segmental HMM
[13] Gaassian Selection
[14] Texas Instruments
[15] Massachusett Institute of Tecnology
[16] National Institute of Standards and Technologies
[17] Normalization

فربد رزازي کارشناسي و کارشناسي ارشد خود را از دانشکده مهندسي برق دانشگاه صنعتي شريف، در گرايش مخابرات رشته مهندسي برق به ترتيب در سال‌هاي ۱۳۷۳ و ۱۳۷۶ به پايان رسانده است و دکتراي خود را از دانشکده مهندسي برق دانشگاه صنعتي اميرکبير، در گرايش مخابرات سيستم رشته مهندسي برق در سال ۱۳۸۲ اخذ کرده است. وي از سال۱۳۷۷ تاکنون عضو هيات علمي دانشگاه آزاد اسلامي– واحد علوم و تحقيقات تهران است. زمينه تحقيقاتي او پردازش سيگنال‌هاي ديجيتال، پردازش گفتار و بازشناسي آماري الگو مي‌باشد. آدرس پست الکترونيکی نامبرده عبارتست از :

farbod _ razzazi@ yahoo.com

ابوالقاسم صیادیان کارشناسی خود را در رشته مهندسی الکترونیک در سال ۱۳۵۹ از دانشگاه صنعتی امیرکبیر و کارشناسی ارشد خود را در رشته مهندسی برق – مخابرات (گرایش سیستم) در سال ۱۳۶۶ از دانشگاه صنعتی اصفهان و دکترای خود را در رشته مهندسی مخابرات (گرایش سیستم) در سال ۱۳۷۳ از دانشگاه تربیت مدرس اخذ نمود.

نامبرده از سال ۱۳۷۰ تاکنون عضو هیات علمی دانشکده مهندسی برق دانشگاه صنعتی امیرکبیر می باشد. زمینه تحقیقاتی نامبرده پردازش سیگنال (به ویژه صوت، گفتار، تصویر)، رمزنگاری و کدینگ می باشد. آدرس پست الکترونیکی نامبرده عبارتست از :

eeas35@cic.aut.ac.ir