ارائهٔ درون‌یابی KNNGI[1] و مقایسهٔ آن با درون‌یابی FI[2] در بازشناسی گفتار

ابوالقاسم صیادیان*، کامبیز بدیع*، محمد شهرام معین** و نصرالله مقدم **
دانشکدهٔ برق دانشگاه صنعتی امیرکبیر
مرکز تحقیقات مخابرات ایران
دانشکدهٔ فنی و مهندسی، دانشگاه تربیت مدرس

(دریافت مقاله: ۸۱/۹/۱۷ – دریافت نسخه نهایی: ۸۳/۳/۲۵)

چکیده – مدل‌سازی آماری HMM رویکردی پرکاربرد در سیستمهای بازشناسی گفتار پیوسته و گسسته است. توزیع احتمال بردارهای مشاهدات هر حالت پنهان مدل، به دو روش پیوسته[3] یا گسسته[4] تخمین زده می‌شوند. عملکرد توزیع احتمال پیوسته (با مدل‌سازی GMM[5]) بالاتر از عملکرد توزیع احتمال گسسته (با مدل‌سازی VQ[6]) است. ولی چنانچه بخواهیم از رویکرد HMM برای بازشناسی گفتار گسسته با دایره لغات وسیع استفاده کنیم، هزینهٔ محاسباتی مرحلهٔ بازشناسی با افزایش تعداد لغات، به نحو چشمگیری افزایش می‌یابد. بدین لحاظ در بازشناسی گفتار گسسته با دایره لغات وسیع، از توزیع احتمال گسسته به منظور کاهش هزینهٔ محاسباتی و امکان پیاده‌سازی بی درنگ[7] استفاده می‌شود. برای جبران کاهش دقت و عملکرد مدل‌سازی DD-HMM، استفاده از درون‌یابی فازی FI مرسوم است. در این تحقیق روش درون‌یابی گوسی که دارای پشتوانه نظری قوی‌تر نسبت به FI است ارائه کرده‌ایم. کارایی دو روش درون‌یابی KNNGI و FI در بازشناسی ۱۵۰۰ کلمه فارسی مورد تحقیق و بررسی قرار دادیم. نتایج این تحقیق نشان می‌دهد که دقت و انعطاف‌پذیری درون‌یابی KNNGI بیشتر از روش FI است.

واژگان کلیدی : درون‌یابی گوسی، درون‌یابی فازی، مدل مارکف مخفی چگالی گسسته، بازشناسی تلفظ گسسته

Presentation of K Nearest Neighbor Gaussian Interpolation and Comparing it with Fuzzy Interpolation in Speech Recognition

A. Sayadiyan, K. Badi, M. Moin and N. Moghadam
Department of Electrical Engineering, AmirKabir University of Technology

Abstract: ***Hidden Markov Model is a popular statisical method that is used in continious and discrete speech recognition. The probability density function of observation vectors in each state is estimated with discrete density or continious density modeling. The performance (in correct word recognition rate) of continious density is higher than discrete density HMM, but its computation complexity is very high, especially in very large discrete utterance recognition problems. For real time implementation of very large discrete utterance recognition, we must use discrete density HMM (DDHMM). To increase the performance of DDHMM, one usual solution is fuzzy interpolation. In this study, we present a new method named Gaussian interpolation. We implemented and compared the performance of two types of interpolation methods for 1500 Persian speech command words. Results show that precision and flexibility of Gaussian interpolation is better thanthose of the fuzzy interpolation.***

Keywords: ***Gaussian Interpolation, Fuzzy Interpolation , Discrete Density HMM, Discrete utterance recogntion***

---

* – دانشیار ** – استادیار

فهرست علائم

a متغیر نماینده احتمال وقوع حالات شرطی	q متغیر نماینده حالت
b متغیر نماینده احتمال وقوع هر بردار مرجع در هر حالت	t متغیر نماینده توالی زمانی
c نام یک کلاس اکوسیتکی	V متغیر نماینده نام بردار کتاب کُد
d نام تابع فاصله اقلیدس	W متغیر نماینده وزن دادن
m متغیر نماینده مقدار فازی بودن	π متغییرنماینده احتمال وقوع حالت در شروع
N متغیر نماینده توزیع نرمال	α متغیر نماینده توابع عضویت نمایش فازی
o بردار ویژگی مشاهده شده	σ متغیر نماینده واریانس توزیع نرمال
p متغیر نماینده مقدار احتمال گوسی	

۱- مقدمه

رویکرد آماری HMM در کاربردهای عملی بازشناسی گفتار، رویکردی غالب نسبت به سایر روشها از جمله ANN[۸] و DTW[۹] است. از ویژگیهای بارز رویکرد HMM استفاده بهینه از اطلاعات ویژگیهای طیفی و زبانی، در مدلسازی مؤثر و تشخیص صحیح وقایع اکوستیکی متفاوت و متمایز است. HMMها هم برای گفتار پیوسته[۱۰] و هم برای گفتار گسسته[۱۱] مورد استفاده قرار می‌گیرند. در کاربردهای بازشناسی گفتار پیوسته، علاوه بر مدل‌سازی اکوستیکی[۱۲] به مدل‌سازی زبانی[۱۳] نیز نیازمندیم. با توجه به اینکه در زبان فارسی پایگاه داده گفتاری جامع و پوشا برای مدل‌سازی ATP[۱۴] و همچنین پایگاه متنی جامع و پوشا برای مدل‌سازی SLM[۱۵] موجود نیست، توجه اصلی این تحقیق به کاربرد HMMها در بازشناسی گفتار گسسته معطوف شده است. هدف سیستم بازشناسی گفتار گسسته عبارت است از، تعیین کلاس یک سیگنال اکوستیکی دریافتی (با فرض اینکه محدوده تقریبی ابتدا و انتهای آن با دقت مناسبی مشخص است): همچنین پارامترهای مدل آماری M کلاس اکوستیکی متمایز، از قبل (در فاز آموزش) مشخص است. هر کلاس اکوستیکی می‌تواند یک فرمان یا درخواست صوتی یک تا حداکثر پنج کلمه‌ای باشد (در این تحقیق). در مدل‌سازی HMM، به منظور استفاده بهینه از اطلاعات زمانی به منظور متمایزسازی کلاسها، هر کلاس اکوستیکی توسط N حالت (سه حالت برای تشخیص واجها و $N \geq 5$ حالت برای تشخیص کلمات و عبارات) مدل‌سازی می‌شوند. اطلاعات آماری مربوط به قواعد توالی زمانی تولید حالتها، توسط ماتریس A (با بعد N×N) و بردار حالت اولیه π (با بعد N×۱) بازنمایی می‌شوند. تغییرات آماری ویژگیهای طیفی در داخل هر حالت هر کلاس توسط یک pdf[۱۶] پیوسته یا گسسته مدل‌سازی می‌شوند. روشهای تخمین (در فاز آموزش) و روشهای استفاده (در فاز بازشناسی) از ماتریس A و بردار π در هر دو روش CD-HMM و DD-HMM تقریباً یکسان است. وجه تمایز اصلی دو روش در نحوه مدل‌سازی pdf بردارهای طیفی در داخل حالات است. مدل‌سازی پیوسته pdf در سیستمهای بازشناسی پیوسته CSR[۱۷] که تعداد کلاسهای اکوستیکی کم و محدود است. (متناظر با تعداد واجهای هر زبان) (حدود ۳۰ واج برای فارسی و حدود ٤٤ واج برای زبان انگلیسی)، بسیار مورد توجه است. ولی در سیستمهای بازشناسی گفتار گسسته با تعداد کلاسهای اکوستیکی بالا (حدوداً بالای ۱۰۰۰ لغت)، استفاده از pdf پیوسته به ویژه در فاز بازشناسی از نظر پیاده‌سازی بی درنگ مقرون به صرفه و عملی نیست. در این تحقیق تعداد کلاسهای اکوستیکی بالاتر از ۱۰۰۰ لغت و برابر ۱۵۰۰کلمه است. در چنین مواردی از بازنمایی گسسته pdf (با استفاده از تکنیک VQ) یا از بازنمایی نیمه پیوسته[۱۸] استفاده می‌کند. چون بازنمایی گسسته دارای بالاترین سرعت در هنگام

پیاده‌سازی است (به علت امکان استفاده از ساختارهای مناسب برای جستجوی سریع در مرحله کدینگ VQ)، مورد توجه و عنایت این تحقیق قرار گرفته است. بنابراین روش پایه در این تحقیق DDHMM است. عملکرد مدل‌سازی DDHMM از نظر دقت پایینتر از مدل‌سازی DDHMM است. روش کلاسیک برای افزایش نسبی دقت روش بازشناسی DDHMM، استفاده از درون یابی فازی(FI) است. این روش‌ها در بخش(٤) مراجع[٨ و ٩] مورد بررسی قرار گرفته‌اند. در این تحقیق دو نوع درون یابی جدید به نامهای KNNGI (درون یابی گوسی K نزدیکترین همسایه) و KNNFI (درون یابی فازی K نزدیکترین همسایه) در بخش(٥) ارائه کرده و عملکرد آنها را در بخش(٥-٣) و بخش(٦) و (٧) مورد اریابی قرار دادیم. نتایج ارزیابی نشان می‌دهند که روش جدید درون یابی KNNGI دارای بهترین عملکرد در میان روشها بازشناسی DDHMM است.

٢- بررسی مدل‌سازی DD-HMM در بازشناسی گفتار گسسته

فرض می‌کنیم تعداد L کلاس اکوستیکی (کلمه یا عبارت) داریم. هر کلاس اکوستیکی توسط N حالت مدل‌سازی زمانی می‌شود. اطلاعات زمانی ماندن یا گذر از حالتی به حالتی دیگر برای هر کلاس به صورت زیر توصیف می‌شود:

$$A_l = \{a_{ij}(l)\} \quad \text{that} \quad a_{ij}(l) = P\{q_{t+1}(l) = j / q_t(l) = i\} \qquad (١)$$
$$l = ١ \text{ تا } L \qquad 1 \le i, j \le N$$

$$\pi_l = \{\pi_i(l)\} \quad \text{that} \quad \pi_i(l) = P\{q_{t=1}(l) = i\} \qquad (٢)$$
$$l = ١ \text{ تا } L \qquad 1 \le i \le N$$

فرض می‌کنیم بردارهای مشاهدات هر حالت توسط M بردار مرجع کد می‌شوند، در این صورت توزیع احتمال گسسته حالات به صورت زیر توصیف می‌شود:

$$B_l = \{b_i(l,k)\} \quad \text{that} \quad b_i(l,k) = P[o_t = v_k(l) / q_t(l) = i] \qquad (٣)$$
$$l = ١ \text{ تا } L \qquad 1 \le i \le N \qquad 1 \le k \le M$$

روشهای تخمین پارامترهای فوق‌الذکر از روی نمونه‌های آموزشی (در فاز آموزش) در مراجع متعدد از جمله [١] و [٥] قابل مطالعه است. در فاز بازشناسی با استفاده از الگوریتم برگشتی FP[19]، نمره (یا احتمال) هر کلاس با دریافت هر دنباله از بردار مشاهدات اکوستیکی $O_1O_2\ldots O_T$ به شرح زیر محاسبه می‌شود:

قدم اول: محاسبه در لحظه t=١

$$a_1(l,i) = \pi_i\, b_i(l,o_1) \qquad 1 \le i \le N \qquad (٤)$$

قدم دوم: مرحله تکراری

$$a_{t+1}(l,j) = \left\{\sum_{i=1}^{N} a_t(l,i).a_{ij}(l)\right\} b_j(l,O_{t+1})$$
$$1 \le t \le T-1 \qquad \text{و} \qquad 1 \le j \le N \qquad (٥)$$

قدم سوم: محاسبه نمره نهایی هر کلاس

$$P(l) = P(O_l,\ldots O_T/C_l) = \sum_{i=1}^{N} a_T(l,i) \qquad (٦)$$
$$1 \le t \le T-1 \qquad \text{و} \qquad ١ \le j \le N$$

مراحل فوق برای L تا ١ =l تکرار می‌شود. کلاسی که بیشترین P(l) را تولید کند به عنوان کلاس برنده اعلام می‌شود. البته از الگوریتم ویتربی [٥] نیز برای محاسبه نمره نهایی هر کلاس می‌توان استفاده کرد.

با توجه به محدود بودن رنج دینامیکی پارامترهای π, A در مدل‌سازی آماری هر کلاس، بسیاری از محققان اعتقاد دارند که نقش اصلی در بالا بردن عملکرد سیستم بازشناسی به بهبود مدل‌سازی و تخمین دقیق پارامترهای ماتریس B نهفته است. مدل‌سازی پیوسته عناصر ماتریس B توسط مدل GMM به عنوان بهترین راه حل شناخته شده است. اما به دلایلی که در این تحقیق بیان شد، از بازنمایی گسسته می‌خواهیم استفاده کنیم.

٣- روشهای طراحی کتاب کد برای مدل‌سازی گسسته pdf بردارهای طیفی

در بخش قبل بیان کردیم که برای محاسبه عناصر ماتریس B به تعدادی الگوی مرجع به نام کتاب کُد برای هر کلاس اکوستیکی نیازمندیم. فرض می‌کنیم $TV_1,\ldots TV_S$ مجموعه بردارهای آموزشی متعلق به یک حالت از یک کلاس اکوستیکی باشد. با استفاده از مجموعه بردارهای آموزشی و دو الگوریتم کلاسیک LBG[20] یا EM[21] می‌توان کتاب کد طراحی کرد [٧ و

۱۰ – ۱۳]. فـرض می‌کنیم تعداد کلمات کد هر کتاب (یا تعداد بـردارهای مـرجع) برابـر (M<<S)M باشـد. همچنـین فـرض مـی‌کنیم ویژگیهای طیفی ناهمبسته باشند (در اغلب کاربردهای عملی از ویژگی MFCC و مشتقات مرتبه اول و دوم آن استفاده مـی‌شـود کـه تقـریباً شرایط مطرح شده را دارا هستند). در این صورت خروجی هر دو الگوریتم LBG و EM شامل M بردارد میانگین Vi (M تا ۱ =i) M, (i= ۱ تا M) بردار واریانس VAR_i (i= ۱ تا M) (اگـر ویـژگیها ناهمبسته نباشند به جای بردار واریانس ماتریس کوواریانس تولید خواهد شد)، و تعداد M اسکالر وزنی W_i (M تـا ۱ =i) (بـا محـدودیت $\sum_{i=1}^{M} w_i = 1$) خواهـند بود. w_iها در الگوریتم EM به عنوان ضرایب مخلوط و در الگوریتم LBG به عنوان فرکانس نسبی برنده شدن هر کد ذکر می شوند. در مدل‌سازی کلاسیک DD-HMM صرفاً از بردارهای متوسط Vi (M تا i=1) در مـرحله کدیـنگ و بازشناسـی استفاده می شود. یعنی با دریافت هر بردار O_t آن را به یکی از M بردار V_i طبق منطق زیر نسبت می‌دهند:

$$\text{Code}(O_t) = V_k \quad \text{if} \quad k = \text{argmin}\,[d(V_1,O_t), d(V_2,O_t)\ldots d(V_M,O_t)] \qquad (۷)$$

d(X,Y) تابع فاصله اقلیدسی بین دو بردار Y,X است. بار اصلی هـزینه محاسـباتی مرحله کدینگ تعیین M تابع فاصله اقلیدسی d (O_t,V_i) (i= ۱ تـا M) اسـت. نکـته برجسـته در استفاده از روش بازشناسـی بـا چگالی گسسته آن است که، با استفاده از ساختارهای جسـتجوی سریع می‌توان فرایند تعیین کد برنده MDWC[۲۲] را به جای M بـار محاسـبه به $k.\log_2 M$ بار محاسبه تابع فاصله کاهش داد (k نوعاً بـین ۲ تا ٤ قرار دارد). البته کارایی ساختارهای جستجوی سریع موقعی قابـل ملاحظـه و قابل استفاده است که M باندازه کافی بزرگ باشد. این مزایا وقتی قابل توجه است که بخواهیم سیستم بازشناسی با دایره لغات وسیع و به صورت بی درنگ طراحی کنیم.

٤- اسـتفاده از بازنمایـی SCD[۲۳] و FVQ[۲٤] بـرای جبران کاهش دقت مدل‌سازی DD-HMM

اسـتفاده از چگالـی گسسـته اگرچه دارای مزایای سادگی و سـرعت پـیاده‌سازی به ویژه در فاز بازشناسی است، اما به دلیل چـندی کـردن گسسـته ویژگیهای طیفی به تعداد معینی سطوح، نویـز چـندی شـده کارایی سیستم را از نظر دقت بازشناسی به ویـژه بـرای تلفظها و گویندگانی که در مرحله آموزش شرکت نداشته اند، پایین می‌آورد. کاهش دقت سیستم بازشناسی چگالی گسسـته نسبت به چگالی پیوسته امری مسلم و ثابت شده برای محققـان اسـت. تـلاش دانشـمندان بـرای افـزایش کارایی مدل DD-HMM منـجر به ابـداع روشـهای SCD, FVQ شـد. ایـن روشـها در حـد قابـل قبولی نقطه ضعف DD-HMM را از نظر دقـت بازشناسـی برطـرف کـرده‌اند. در روش SCD، ابتدا یک کـتاب کد همانند حالت چگالی گسسته طراحی می‌شود. سپس بـا استفاده از مجموعه بردارهای آموزشی نسبت داده شده به هر کلمـه کـد، یـک چگالـی پیوسته با تعداد مخلوطهای محدودی (نـوعاً یـک تا شش) توسط مدل‌سازی GMM طراحی می‌شود. استفاده از مدل‌سازی SCD موجب افزایش دقت سیستم در حد مـدل‌سـازی پیوسته CD شده است. [۲ و ٤]. اگر چه استفاده از روش SCD بسـیار مفـید اسـت، ولـی چـون در عمل مجبوریم تعـداد کلمـات کـد کتاب کد اصلی (DD) را کم انتخاب کنیم، دیگـر قـادر نیسـتیم از کارایی مدل DD به صورت کامل (برای ایجـاد سـاختارهای جستجوی سریع) استفاده کنیم. در نتیجه از هـدف اصـلی خـود در به کارگیری مدل‌سازی گسسته pdf دور خواهیم شد. در روش FVQ علاوه بر استفاده از کلمه کد برنده (در مرحله بازشناسی) از تعداد I کد با کمترین فاصله نسبت به بردار مشاهده O_t نیز برای بالا بردن دقت سیستم به صورت زیر اسـتفاده مـی‌کنـند [۸ و ۹]. فرض می‌کنیم. $Vi_1,Vi_2,\ldots Vi_I$ کدهای برنده با کمترین فاصله نسبت به بردار مشاهده O_t باشند یعنی داریم:

$$d(v_{i1},o_t) \le d(v_{i2},o_t) \le \ldots d(v_{iI},o_t) \le d(v_j,o_t) \qquad (۸)$$
$$۱ \le j \le M$$
$$j \ne i_1,\ldots i_I$$

در روش کلاسـیک DD، مسـتقل از فواصـل O_t نسبت به سایر کـدها، نمـره نهایـی سیسـتم محاسـبه می‌شود. در روش FVQ، مقدار $b_j(O_t)$ به صورت زیر درون‌یابی می‌شود:

در مدل DD کلاسیک: $b_j(o_t) = b_j(V_{il})$

در مدل FVQ:

$$b_j(o_t) = \sum_{i=1}^{I} \alpha_i . b_j(o_t = V_i) \qquad (۹)$$

که در آن v_iها، کدهای برنده با قواعد ذکر شده‌اند.

ضرایب وزنی درون‌یابی α_iها(I تا ۱ =i) به صورت فازی به شرح زیر محاسبه می‌شوند:

$$\alpha_i = \frac{(d_i)^{\frac{1}{m}}}{\sum_{k=1}^{I} (d_k)^{\frac{1}{m}}} \qquad \text{that} \qquad \sum_{k=1}^{I} \alpha_i = 1 \qquad (۱۰)$$

که m مقدار فازی بودن[۲۵] سیستم (عددی بین ۱ تا ۲) و α_iها توابع عضویت در بازسازی $b_j(o_t)$ هستند [۹]. کارایی FVQ به مراتب بیشتر ازVQ (روش DD کلاسیک) است. همچنان که ملاحظه می‌شود، در این روش نیز مجبوریم کلیه توابع فاصله بین بردار مشاهده O_t و کلیه اعضای کتاب کد را به دست آورده و سپس آنها را مرتب کرده و کلمه کد برنده با تابع فاصله‌های متناظر را برای محاسبه α_i مورد استفاده دهیم. بنابراین نمی‌توانیم از ساختارهای جستجوی سریع برای اهداف این تحقیق استفاده کنیم.

۵- ارائه درون‌یابی KNNGI,[۲۶] KNNFI,[۲۷]

در روش KNNFI و KNNGI ساختارهایی به شرح زیر بر روی اعضای کتاب کد (در مرحله آموزش) ایجاد می‌کنیم. فرض می‌کنیم $(V_1,\ldots V_M)$ اعضای کتاب کد یک کلاس اکوستیکی باشند. برای هر بردار مرجع، تعداد k نزدیکترین همسایه بر مبنای کمترین فاصله به شرح زیر به دست می‌آوریم. فرض می‌کنیم $i_1,i_2,\ldots i_k$ تعداد k نزدکترین همسایه کلمه کد v_i باشند یعنی داریم:

$$d(v_i, v_{i_1}) \le d(v_i, v_{i_2}) \le \ldots d(v_i, v_{ik}) \qquad (۱۱)$$

فرایند تعیین k نزدیکترین همسایه را برای کلیه کلمات کد انجام می‌دهیم. ایندکس k نزدیکترین همسایه هر کلمه کد به همراه سایر اطلاعات کتاب کد را در فاز آموزش به دست آورده و ذخیره می‌کنیم. توجه می‌شود که فرایند تعیین ایندکس k نزدیکترین همسایه هر کلمه کد، صرفاً یکبار در مرحله آموزش محاسبه و ذخیره می‌شود.

۵-۱- ارائهٔ درون‌یابی KNNFI

فرض می کنیم o_t یک بردار مشاهده در فاز بازشناسی و v_i نزدیکترین بردار مرجع به آن باشد (که با ایجاد ساختارهای مناسب در فاز آموزش به طور سریع تعیین می‌شود.) فاصله بردار مشاهده o_t را با $v_i, v_{i1}, \ldots v_{ik}$ به شرح زیر تعیین می‌کنیم.

$d(o_t,v_i),\ d(o_t,v_{i1}),\ldots d(o_t,v_{ik})$

با استفاده از فواصل فوق توابع عضویت را به شرح زیر به دست می‌آوریم:

$$\alpha_j(o_t, v_i) = \frac{[d(o_t, v_{il})]^{\frac{1}{m}}}{\sum_{k=o}^{K+1} [d(o_t, v_{ik})]^{\frac{1}{m}}} \qquad (۱۲)$$

$$j=0,1,\ldots k \qquad , v_{io} \triangleq v_i$$

تعداد توابع عضویت برابر k+1 است. با استفاده از مقادیر k+1 تابع عضویت $\alpha_1,\ldots\alpha_k$ مقدار تابع احتمال $b_j(o_t)$ به شرح زیر درون‌یابی می‌شود:

$$b_j(o_t) = \sum_{l=0}^{k} \alpha_l \quad .b_j(v_{il}) \qquad (۱۳)$$

$$v_{io} \triangleq v_i$$

ملاحظه می‌شود که درون‌یابی صرفاً با استفاده از فاصله بردار مشاهده O_t از کد برنده و k نزدیکترین همسایه آن انجام می‌پذیرد. یعنی تعیین $b_j(o_t)$ به فواصل O_t از سایر کلمات کد بستگی ندارد. بنابراین در این روش نیز می‌توان از ساختارهای جستجوی سریع استفاده کرد.

۵-۲- ارائهٔ درون یابی KNNGI

بردارهای مرجع $V_1,\ldots V_M$ (یا کلمات کد) نوعاً توسط الگوریتم LBG یا EM طراحی می‌شوند. در پایان اجرای الگوریتم LBG (یا EM)، علاوه بر بردارهای مرجع، بردارهای واریانس متناظر نیز محاسبه می‌شوند (با فرض قطری بودن ماتریس کوواریانس بردارها). در روش درون یابی KNNFI از اطلاعات واریانس هر کلمه کد هیچ گونه استفاده‌ای نمی‌شود. اما می‌دانیم که واریانس هر بردار مرجع، نماینده چگونگی پراکندگی بردارهای آموزشی است که آن بردار مرجع نماینده آنهاست. سؤال اساسی برای محققان این نوشتار این بود که چگونه از این اطلاعات اضافی برای بهبود عملکرد فرایند درون‌یابی استفاده کنیم. با توجه به کارایی مدلهای GMM، ایده

زیر را مورد توجه قرار دادیم:

فرض می‌کنیم v_i کد برنده با مشاهده بردار O_t باشد، همچنین $v_{i1},...v_{ik}$ تعداد k نزدیکترین همسایه بردار مرجع v_i باشد. ضرایب وزنی درون‌یابی را به شرح زیر پیشنهاد و مورد استفاده قرار دادیم:

$$p_j \triangleq N(o_t, v_{ij}, \sigma_{ij}) \qquad (۱۴)$$

N(-) نماینده تابع گوسی با بردار میانگین v_{ij} و بردار واریانس σ_{ij} است.

$$\alpha_j \triangleq \frac{p_j}{\sum_{l=o}^{K} p_l} \qquad \text{that} \qquad j=0 \quad \text{تا} \quad k \qquad (۱۵)$$

ملاحظه می‌شود که در درون‌یابی KNNGI، نیز صرفاً از فاصله گوسی بردار مشاهده o_t و کد برنده o_t و k نزدیکترین همسایه آن استفاده شده است. بنابراین استفاده از درون‌یابی KNNGI به ساختار جستجوی سریع خدشه‌ای وارد نمی‌کند.

۵-۳ بررسی کارایی محاسباتی KNNFI و KNNGI؛

با توجه به اینکه روشهای ارائه شده در این تحقیق برای سیستمهای بازشناسی با دایره لغات وسیع طراحی شده‌اند، در نتیجه پیاده سازی هر دو مدل بازشناسی (CDHMM و DDHMM) به صورت گره زدن مخلوطها (TM)[۲۸] انجام پذیرفته است. در مدلCDHMM از۵۱۲ مخلوط گوسی و برای مدل DDHMM از ۱۰۲٤ بردار VQ برای بازنمایی pdf حالات تمامی کلمات مربوط به تمامی گویندگان استفاده کردیم. برای محاسبه مقدار درست نمائی هر فریم دریافتی در مدل DDHMM ، به حداقل ۵۱۲×۳p عمل (ضرب و جمع و تقسیم) و ۵۱۲ محاسبه تابع اکسپانسیل و ۵۱۲ عمل ضرب نیازمندیم. (p بعد بردار ویژگی است که در این تحقیق برابر ۲۵ است). فرض می‌کنیم هزینه محاسباتی هر عمل ضرب دو برابر عمل جمع و هر عمل تقسیم چهار برابر عمل جمع و محاسبه تابع اکسپانسیل ده برابر عمل جمع باشد. بنابراین مجموع تعداد عملیات محاسباتی برای تعیین نمره یک فریم دریافتی متناسب با ۱۰۲٤× ۹۳ عمل جمع می‌شود. برای حالت DDHMM، مجموع هزینه محاسباتی متناسب با ۱۰۲٤ × ۵۰ عمل جمع خواهد شد. البته در تخمین هزینه محاسباتی فوق الذکر، از وارد کردن حجم محاسبات مربوط به تحلیل هر فریم و مقایسه پویا در محاسبه نمرات کلاسهای اکوستیکی، تشخیص ابتدا و انتهای کلمه و ...، به علت مشترک بودن فرایندهای فوق الذکر در هر دو مدل CDHMM و DDHMM صرف‌نظر شده است. در محاسبات تحلیلی فوق، مشاهده می‌شود که هزینه محاسباتی DDHMM حدود نصف CDHMM است. این نسبت کاهش محاسباتی، ارزش توجه و به کارگیری مدل DDHMMدر مقابل CDHMM را ندارد (با توجه به مقدار افت دقت بازشناسی مدل DDHMM در مقابل مدل CDHMM). هزینه محاسباتی تخمین زده شده برای مدل DDHMM با فرض استفاده از ساختار جستجوی کامل است. در سیستمهای کاربردی از ساختارهای جستجوی سریع، مانند ساختار باینری و یا ساختار تلفیقی باینری و جستجوی تصادفی k نزدیکترین همسایه استفاده می‌کنند. درحالت دوم، تعداد عملیات برای محاسبه نمره یک فریم متناسب با ۲p. (log۲۱۰۲٤)k. می‌شود. (k بین ۲ تا ٤ بوده و در این تحقیق k=۳ حاصل شده است). بنابراین در مدل‌سازی DDHMM با ساختار تلفیقی جستجوی سریع، هزینه محاسباتی برای تعیین نمره یک فریم متناسب با ۱۰۲٤ × ۱/۵ عمل جمع می‌شود. مقایسه هزینه محاسباتی مدل DDHMM با ساختار تلفیقی جستجوی سریع و مدل CDHMM، علت توجه محققان به مدل‌سازی DDHMM را به خوبی آشکار می‌کند (کاهش هزینه محاسباتی به نسبت 1/62).

دو روش درون یابی KNNFI و KNNGI که برای جبران کاهش دقت بازشناسی مدل DDHMM در طی این تحقیق ارائه و پیاده سازی شده‌اند، از ساختار تلفیقی جستجوی سریع مدل DDHMM (برای کاهش بار محاسباتی) استفاده می‌کنند. بنابراین هزینه محاسباتی آنها تقریباً معادل روش DDHMM با ساختار تلفیقی جستجوی سریع است. در روشهای ارائه شده (درفاز بازشناسی)، ابتدا بردار مرجع برنده (از میان ۱۰۲٤ بردار مرجع) را به روش جستجوی سریع تعیین می‌کنیم. آن گاه

فاصله فازی یا فاصله گوسی بردار فریم دریافتی و k بردار مرجع نزدیکترین همسایه بردار مرجع برنده را محاسبه می‌کنیم.

ایندکس k نزدیکترین همسایه هر بردار مرجع در فاز آموزش تعیین می‌شود. بنابراین در فاز بازشناسی به بار محاسباتی اضافی نیازی نداریم. مقدار بهینه k (تعداد همسایه‌ها) بین ۵ تا ٦ گزارش می‌شود. در این تحقیق پس از آزمایشهای تجربی فراوان k را برابر ٦ در نظر گرفتیم.

٦- نتایج پیاده‌سازی

مدلهای درون‌یابی ارائه شده در بازشناسی ۱۵۰۰ کلمه فارسی ناوابسته به گوینده[۲۹] مورد استفاده و ارزیابی قرار گرفتند. پایگاه داده گفتاری شامل ۳۰ مرد و ۲۰ زن بوده است. از هر گوینده حدود ٦۰ تلفظ برای هر کلمه ضبط کردیم. (ده تلفظ طی ٦ روز متفاوت). فرکانس نمونه‌برداری ۸ kHZ و نمونه‌ها به صورت ۱٦ بیتی ضبط شده‌اند. داده‌های گفتاری توسط میکروفون برای کیفیت خوب در محیط عادی ضبط شده‌اند. تعداد حالات هر کلمه در مدل‌سازی HMM را برابر پنج (N=۵) قرار دادیم. ویژگیهای مورد استفاده ۱۲ ضریب MFCC[۳۰] و مشتق مرتبه اول آن و مشتق لگاریتم گین یعنی مجموعاً ۲۵ ویژگی بوده‌اند. روش تحلیل برای استخراج ویژگیهای طیفی روش PLP[۳۱] [۵] بوده است. ویژگیهای فوق به ازای هر ٦/۲۵ میلی ثانیه یکبار استخراج شده‌اند. برای هموار کردن کانتور ویژگیها، از یک فیلتر غیرخطی MM(۷و۳)[۳۲] استفاده کردیم. برای تخمین پارامترهای آماری مدل HMM، از الگوریتم بام ولش [۵ و ٦] استفاده کردیم. تعداد اعضاء کتاب کد برای DDHMM برابر ۱۰۲٤ بردار و برای CDHMM برابر ۵۱۲ مخلوط در نظر گرفته شدند. چندین آزمون به شرح زیر انجام پذیرفته است:

۱- پیاده‌سازی مدل CD-HMM
۲- پیاده‌سازی مدل DD-HMM بدون درون‌یابی
۳- پیاده‌سازی مدل DD-HMM با درون‌یابی KNNFI
٤- پیاده‌سازی مدل DD-HMM با درون یابی KNNGI

جداول (۱) و (۲) نتایج بازشناسی برای چهار نوع مدل‌سازی فوق‌الذکر را نشان می‌دهد. در آزمون مربوط به نمونه‌های آموزشی، $\frac{2}{3}$ گویندگان مرد و زن برای آموزش و $\frac{1}{3}$ گویندگان باقیمانده برای آزمایش مورد استفاده قرار گرفته‌اند. فرایند فوق پنج مرتبه با تغییر گویندگان آموزشی (به صورت تصادفی) انجام پذیرفته است. نتیجه اعلام شده متوسط و انحراف معیار این آزمون است. با مروری به نتایج چهار آزمون انجام پذیرفته به جمع‌بندی زیر می‌رسیم:

۱- دقت بازشناسی بر روی نمونه‌های آزمایشی همیشه کمتر از نمونه‌های آموزشی است. اما قدرت سیستم طراحی شده می‌بایستی به گونه‌ای باشد که افت دقت در حد قابل قبول باشد (حدود یک تا حداکثر دو درصد) افت دقت بر روی نمونه‌های آزمایشی در مدل CD-HMM قابل قبول بوده ولی برای مدل DD-HMM (بدون درون‌یابی) غیرقابل قبول است.

۲- قضاوت نهایی برای بررسی عملکرد یک سیستم بازشناسی، بررسی نتایج دقت بازشناسی بر روی نمونه‌های آزمایشی است. بدین لحاظ جدول (۲) نتایج بازشناسی دو روش درون یابی KNNFI و KNNGI را بر روی داده‌های آزمایشی نشان می‌دهد. ملاحظه می‌شود، که دقت بازشناسی دو روش درون‌یابی ارائه شده در مقایسه با مدل DD-HMM به شدت بهبود یافته و به دقت سیستم CD-HMM نزدیک شده است.

۳- دقت بازشناسی درون یابی KNNGI بر روی نمونه‌های آزمایشی به طور متوسط حدود ۰/۷۹ درصد بهتر از دقت بازشناسی درون‌یابی KNNFI است.

٤- دقت بازشناسی DD-HMM با درون‌یابی KNNGI در حدود دقت بازشناسی CD-HMM (با حدود ۰/۵ درصد افت دقت) است بنابراین با اطمینان کافی می‌توان از آن برای سیستمهای کاربردی استفاده کرد.

در آزمون دیگری، برنامه‌ای به زبان C (در محیط VC++6) برای بازشناسی گفتار ۱۵۰۰ کلمه فارسی به دو روش CD-HMM و DD-HMM (با دورن‌یابی KNNGI) نوشته شده است. سیستم بازشناسی به روش DD-HMM (با درون‌یابی KNNGI) به طور متوسط حدود ۰/۷۲ ثانیه پس از دریافت

جدول۱- نتایج دقت بازشناسی نا وابسته به گوینده مدل CDHMM و DDHMM

نوع مدل	نمونه‌های آموزشی	نمونه‌های آزمایشی
CDHMM	۹۸/۲۱ (۱/۰۵)	۹۷/٦٥ (۱/۸٥)
DDHMM	۹۹/۱۸ (۰/۵۱)	۸۹/۲۷ (۲/۸۲)

جدول۲- نتایج دقت بازشناسی ناوابسته به گوینده مدل DD-HMM با درون‌یابی KNNGI و KNNFI

نوع درون‌یابی	نمونه‌های آموزشی	نمونه‌های آزمایشی
KNNGI	۹۸/۳۱ (۰/۹۷)	۹۷/۱٦ (۱/۹۲)
KNNFI	۹۸/٤۸ (۰/۹۱)	۹٦/۳۷ (۲/۱۲)

کلمه مورد نظر نتایج بازشناسی را اعلام می‌کند سیستم بازشناسی به روش CD-HMM به طور متوسط حدود دو دقیقه پس از دریافت کلمه مورد نظر نتایج بازشناسی را اعلام می‌دارد. برنامه‌ها بر روی رایانه پنتیوم ٤ (۲/٤ GHZ) با حافظه ۵۱۲ mbyte اجرا شدند. بنابراین از نظر اجرای بی درنگ، به نظر می‌رسد که روش CD-HMM قابلیت کاربرد چندانی برای دایره لغات وسیع ندارد. ولی روش DD-HMM (با درون‌یابی KNNGI) دارای قابلیت کاربرد در سیستمهای کاربردی با اجرای بی درنگ است.

۷- بررسی کارایی درون یابی KNNGI و KNNFI

مدل CDHMM، توزیع آماری بردارهای ویژگی داخل هر حالت هر کلاس اکوستیکی را توسط یکpdf پیوسته و پارامتریک، به نام GMM بازنمایی می‌کند. تعداد مخلوطهای توزیع پارامتریک GMM هر حالت، بستگی به تنوع گفتار و تعداد گوینده‌ها، متغیر است. توزیع پیوسته و پارامتریک GMM، توازن مناسبی بین قدرت متمایزسازی و قدرت تعمیم پذیری کلاسهای اکوستیکی ایجاد می‌کند. با افزایش تعداد بردارهای کتاب کد، قادریم قدرت متمایزسازی مدل DDHMM را بیشتراز مدلCDHMM کنیم. این توانمندی مدل DDHMM در اغلب تحقیقات محققان دیگر نیز گزارش شده است. متأسفانه در اغلب کاربردهای عملی، به قدرت متمایزسازی و قدرت تعمیم پذیری به طور همزمان نیازمندیم. علت اصلی پایین بودن قدرت تعمیم پذیری مدل DDHMM، بازنمایی گسسته فضای pdf توسط تعداد محدودی بردار (اعضای کتاب کد) است. در روش CDHMM به علت پیوسته بودن بازنمایی فضای pdf، بردارهایی که در فاز آموزش دیده نشده‌اند نیز دارای مقادیر احتمال مناسبی برای همان کلاس‌اند. این توانمندی به علت گسسته شدن فضای pdf در مدل DDHMM در حد مناسبی برقرار نیست. البته مدل DDHMM نسبت به مدل CDHMM دارای دو برتری اساسی است؛ ۱- قدرت متمایز سازی بیشتر بر روی نمونه‌های آموزشی ۲- سرعت اجرای بسیار سریعتر. تنها مشکل آن پایین بودن قدرت تعمیم پذیری است. (یعنی پایین بودن دقت بازشناسی بر روی نمونه‌های آزمایشی که قبلاً در آموزش شرکت نداشته‌اند). روش درون یابی فازی FI که قبلاً توسط محققان دیگر [۸ و ۹] ارائه شده است، تا حدودی نقطه ضعف مدل DDHMM را از نظر قدرت تعمیم پذیری برطرف کرده است. روش درون یابی فازی کلاسیک، از ساختارهای جستجوی سریع استفاده نمی‌کند. زیرا به ازای هر بردار تست ورودی، فاصله فازی آن با تمام بردارهای مرجع محاسبه شده و سپس تعداد ۵ تا ٦ بردار مرجع که فاصله فازی آن از بقیه اعضای کتاب کد کمتر است، برای درون یابی نهایی انتخاب می‌شوند. بنابراین درون یابی فازی کلاسیک، قابلیت پیاده‌سازی سریع مدل DDHMM را کم رنگ می‌کند. در روش KNNFI که در این تحقیق ارائه کردیم، ضمن استفاده از خواص درون یابی فازی برای افزایش قابلیت تعمیم پذیری

مـدل DDHMM، بـه سـرعت پیاده سازی مدل DDHMM نیز خدشـه‌ای وارد نمی‌کند. زیرا در روش KNNFI، از ساختارهای جسـتجوی سـریع مـی‌تـوان بـــراحتـی اسـتـفاده کرد. بنابراین درون‌یابـی KNNFI در مقابـل روش FI کلاسیک، دارای قدرت پـیاده‌سـازی سریعتر است(که یکی از اهداف اصلی این تحقیق بوده است).

درون یابی KNNGI که نیز در این تحقیق ارائه و پیاده‌سازی شده است، ضمن حفظ قابلیت پیاده‌سازی سریعتر(همانند درون یابـی KNNFI) مقـدار مناسـبی نیـز بـه دقت بازشناسی سیستم افـزوده اسـت. ایـن حقـیقت در نتایج مندرج در جدول (۲) به خوبـی آشـکار اسـت. همـان طـوری که از جدول (۲) مشاهده می‌شود، دقت درون یابی KNNGI حدود ۰/۷۹ درصد بیشتر از درون یابی KNNFI شده است. این افزایش دقت معادل کاهش ۲۲ درصـدی خطـای سیسـتم است. دلیل تحلیلی کاهش خطا، اسـتفاده از پـــارامـتـر نرمالیـزه کنـنده واریـانس و تابع تحلیلی گوسـی می‌تواند باشد. همان طور که می‌دانیم، واریانس نماینده پراکندگـی نمـونه‌هـا حـول مقدار میانگین است. در درون یابی فـازی، مسـتقل از واریانس خوشه، فاصله بردار دریافتی با مرکز خوشـه محاسبه شده و از یک فاصله اقلیدسی نیز برای محاسبه فاصـله فـازی اسـتفاده مـی‌شود. با مقایسه روابط (۱۲) و (۱٤) می‌توان گفت که علت دوم برتری درون یابی KNNGI، استفاده از تابع فاصله گوسی به جای تابع فاصله اقلیدسی است. بنابراین در مجمـوع مـی‌تـوان گفـت کـه مـدل DDHMM بـه همـراه درون‌یابـی KNNGI جایگـزین مناسـبی برای دسترسی به دقت بـالا و پـیاده‌سـازی سـریع سیستمهای بازشناسی با دایره لغات وسیع است.

۸- جمع‌بندی و نتیجه‌گیری

هـدف ایـن تحقـیق بازشناسی گفتار گسسته (با دایره لغات وسـیع) با امکان پیاده‌سازی بی درنگ (با تأخیری در حدود نیم تـا حداکثـر یک ثانیه) بر روی رایانه‌های PC قابل دسترس بوده است. با توجه به این که مدل‌سازی HMM روش غالب و موفق در سیسـتمهای بازشناسـی گفتار پیوسته و گسسته است، از این روش بـرای کاربرد ذکر شده استفاده کردیم. دو رویکرد چگالی پیوسته CD و چگالی گسسته DD، برای بازنمایی pdf بردارهای طیفی مورد استفاده قرار می‌گیرند. رویکرد اول CD دارای دقت بازشناسی مناسب است. متأسفانه به علت حجم بالای محاسبات (بـرای دایـره لغـات وسـیع) امکـان استفاده از آن در بازشناسی بـی‌درنگ با رایانه‌های PC موجود و قابل دسترس وجود ندارد. رویکـرد دوم DD اگـرچه دارای دقت پایینتری (نسبت به CD) است ولی پیاده‌سازی بی درنگ آن بر روی pcهای قابل دسترس امکانپذیر است. روش کلاسیک برای افزایش دقت رویکرد دوم (DD) (بـا حفـظ سـاختارهای جسـتجوی سـریع) اسـتفاده از درون‌یابـی فـازی FI اسـت. در ایـن تحقـیق روش درون‌یابـی جدیـدی بـه نـام KNNGI ارائـه و مـورد اسـتفاده قـرار دادیم. درون‌یابـی گوسـی، ضـمن بـالا بـردن دقت بازشناسی (در حد رویکـرد اول CD) بـه راحتـی بـر روی رایانـه‌هـای موجود به صـورت بـی درنـگ قابل پیاده‌سازی است. پیاده‌سازی دو مدل درون‌یابـی بـرای بازشناسـی ناوابسـته بـه گویـنده ۱۵۰۰ کلمه مجـزای فارسی نشان می‌دهد که عملکرد دقت روش درون‌یابی گوسـی بیشـتر از درون‌یابـی فـازی و در حد رویکرد اول یعنی CDHMM است.

واژه نامه

1. K nearest neighbor Gaussian interpolation
2. fuzzy interpolation
3. continious density HMM
4. discrete density HMM
5. Gaussian mixture model
6. vector quantization
7. real time
8. artifitial neural net
9. dynamic time warpping
10. continious speech
11. discrete (utterance) speech
12. acoustic modeling
13. language modeling
14. acoustic to phonetic
15. statistical LM
16. probability density function
17. continious speech recognition
18. semi continious
19. forward procedure
20. Linde, Bazo, Gray
21. expectation maximization
22. minimum distance winer codeword
23. semi continious density
24. fuzzy vector quantizition

25. fuzziness
26. K nearest neighbor fuzzy interpolation
27. KNN Gaussian interpolation
28. TIED mixture
29. speaker dependent
30. mel frequncy cepstrum coeffitient
31. perceptual linear prediction
32. mean median filter

مراجع

1. Rabiner, L., and Juang, B.H., *Fundamentals of Speech Recognition*, Prentice Hall Inc, 1993.
2. Huang, X. D., "Phoneme Classification Using Semi Continuous Hidden Markov Models," *IEEE Transactions on signal processing*, Vol. 40, No. 5, pp.45 – 56, May 1992.
3. Zhang, Y., Togneri, R., and Alder, M., "Phoneme Based Vector Quantization in a Discrete HMM Speech Recognizer," *IEEE Transaction on Speech and Audio Processing*, Vol. 5, No.1, January 1997.
4. Furui, S., and Sondhi, M. M, *Advances in Speech Signal Proceesing*, Marcel Dekker, Inc., 1992.
5. Becchetti, C., Ricotti, L. P., *Speech Recognition, Theory and C Implementation*, John wiley & Sons, Ltd, 1999.
6. Jurafsky, D., Martin, J. H., *Speech and Language Processing*, Prentice Hall. Inc., 2000.
7. Linde, Y., Buzo, A,. Gray, R. M., "An Algorithm for Vector Quantizer design," *IEEE Transaction on Communication*, Vol. Com 28, pp. 84-95, 1980.
8. Tseng, H. P., Sabin, M., Lee, E., "Fuzzy Vector Quantization applied to Hidden Markov Modeling," *In proc, ICASSP*-87, 1987.
9. Schwartz, R., Kimball, O., kubala, F., Feng, M., Chow, Y., Barry, C., Makhoul, J., "Robust Smothing Methods for Discrete Hidden Markiov Models," *In proc, ICASSP*-89, 1989.
10. Dempster, A. P., Laird, N. M., Rubin, D. B., "Maximum Likelihood from Incomplete Data Via the LM Algorithm," *J. Roy. Stat. Soc. Ser.* B(methodological), Vol. 39, pp. 1-38, 1977.
11. Render, R. A., Walke, H. F. "Mixture Densities, Maximum Likelihood, and the EM Algorihm," *SIAM Rer.*, Vol. 26, pp. 195-239, 1984.
12. Bailey, T. L., Elkan, C., "Fitting a Mixture Model by Expectation Maximization to Discove Motifs in Biopolymers," *proc. of Biology*, pp. 28-30, AAAI PRESS, (1994).
13. Gauvain, J. L., Lee, C. H., "Maximum a Posteriori Estimation Gausian Mixture Observations of Mardov Chain," *IEEE Trans. On speech and Audio processing*, No. 2, pp. 291-298, Apr. 1994.