

تطبیق گوینده در بازشناسی گفتار پیوسته براساس تخمین MAP مبتنی بر تبدیل MLLR

سعید شریفیان* و سید محمد احدی**

دانشکده مهندسی برق، دانشگاه صنعتی امیر کبیر

(دریافت مقاله: ۸۲/۴/۴ - دریافت نسخه نهایی: ۸۲/۱۱/۴)

چکیده - روشهای مختلفی برای تطبیق گوینده در سیستمهای بازشناسی گفتار معرفی گردیده‌اند. در برخی روشها نظیر تخمین MAP تنها مدلهایی که داده آموزشی متناظرشان موجود باشد تازه سازی می‌شوند و برای بهبود قابل توجه دقت بازشناسی، داده آموزشی نسبتاً زیادی مورد نیاز است. در برخی دیگر نظیر MLLR که تعدادی تبدیلات عمومی بر روی خوشه‌های مدلهای اعمال می‌شود، برای دادگان کم آموزشی نتایج مطلوبی حاصل می‌شود، اما با افزایش دادگان، کارایی به حد اشباع می‌رسد. در این مقاله روش جدیدی مطرح می‌شود که از مزایای هر دو روش فوق برای دسترسی به کیفیت بالاتر بهره می‌برد. در این روش مدلهایی که داده آموزشی آنها موجود است به کمک تخمین MAP آموزش می‌بینند و برای مدلهایی که داده آموزشی (کافی) ندارند، با استفاده از روش MLLR مقادیر پیشینه مناسب برای تخمین MAP تأمین می‌شود. این روش، در عمل، بر روی یک سیستم آموزش دیده براساس دادگان فارس دات به نتایج بهتری نسبت به هر یک از دو روش MAP و MLLR دست یافته است.

واژگان کلیدی: مدلهای مارکوف پنهان (HMM)، بازشناسی گفتار پیوسته فارسی، تطبیق گوینده، تخمین MAP، تبدیل MLLR

Speaker Adaptation in Continuous Speech Recognition Using MLLR-Based MAP Estimation

S. Sharifian and S. M. Ahadi

Electrical Engineering Department, Amirkabir University, Hafez Avenue, Tehran 15914, Iran

Abstract: A variety of methods are used for speaker adaptation in speech recognition. In some techniques, such as MAP estimation, only the models with available training data are updated. Hence, large amounts of training data are required in order to have significant recognition improvements. In some others, such as MLLR, where several general transformations are applied to model clusters, the results are desirable for small training data, but with increasing training data, the performance improvement reaches the saturation level. In this paper, a new approach is introduced that makes use of the advantages of both mentioned techniques to improve the recognition rate. Here, the models with available training data are trained using MAP while

** - استادیار

* - کارشناس ارشد

for those with insufficient training data, appropriate prior parameters for MAP estimation are found using MLLR. This technique has yielded better performance in comparison to either MAP or MLLR, in a system based on FARSDAT speech corpus.

Keywords: Hidden Markov models (HMM), continuous Persian (Farsi) speech recognition, Speaker adaptation, MAP estimation, MLLR transformation.

فهرست علائم

$C(r)$	ماتریس قطری کوواریانس مقیاس شده گوسی λ در MLLR	φ	پارامتر رشد در MAP وزنی
$c_{ii}^{(r)}$	عنصر قطری λ از ماتریس قطری کوواریانس مقیاس شده گوسی λ در MLLR	$\gamma(t)$	درستنمایی بودن در یک حالت خاص از مدل در لحظه t
$d(.,.)$	فاصله دو حالت از مدل	λ	مجموعه پارامترهای مدل
O_t	داده‌های تطبیق در لحظه t	μ	بردار میانگین توزیع احتمال چند متغیره
$P(.)$	چگالی احتمالی	$\mu_{..}$	عضو بردار میانگین توزیع احتمال چند متغیره
$P(. .)$	چگالی احتمالی شرطی	$\hat{\mu}_j^{(q)}$	میانگین مدل اولیه قبل از آموزش، در MAP وزنی
P_r	درستنمایی لگاریتمی میانگین برای مشاهده λ	μ'	میانگین مدل آموزش دیده‌ای در آن دسته که داده آموزشی متناظر آن موجود است (در MAP وزنی)
w_j	ماتریس تبدیل $n \times (n+1)$ در MLLR	v_j	بردار میانگین گسترش یافته در MLLR
Z, G_i	پارامترهای کمکی تبدیل MLLR	Σ	ماتریس کوواریانس توزیع احتمال چند متغیره
$\alpha^{(q)r}, \beta^{(q)r}$	پارامترهای الگوریتم جلورونده-عقب رونده برای مشاهده λ از مدل q	σ	عضو ماتریس کوواریانس توزیع احتمال چند متغیره
	پارامترهای توزیع پیشینه $\mu_k, \tau_k, \alpha_k, \beta_k$		

۱- مقدمه

این گونه سیستمها، که با استفاده از گفتار یک گوینده خاص آموزش می‌بینند، از توانایی بالایی در بازشناخت گفتار گوینده مزبور برخوردارند. با این همه، چنانچه از چنین سیستمی برای بازشناسی گفتار گوینده دیگری استفاده شود توانایی سیستم به شدت افت می‌کند. در مقابل سیستمهای ناوابسته به گوینده^۱ (SI)، به خاطر آموزش دیدن با گفتار گویندگان مختلف^۲، قادر به بازشناخت گفتار طیف وسیعی از گویندگان هستند. به همین دلیل این گونه سیستمها مورد توجه بیشتر محققان، در مقایسه با سیستمهای وابسته به گوینده، قرار گرفته‌اند. با این همه، توانایی اینگونه سیستمها در بازشناسی گفتار یک گوینده عموماً کمتر از توانایی یک سیستم وابسته به گوینده آموزش دیده با گفتار همان گوینده است. لازم به اشاره است که این مطلب، نه تنها می‌تواند ناشی از عدم تطابق میان مشخصات گویندگان در دادگان آموزشی و مشخصات گوینده جدید باشد، بلکه تفاوت در

برقراری ارتباط طبیعی بین انسان و ماشین از جمله مواردی است که در سالهای اخیر مورد توجه بسیاری از محققان قرار گرفته است. با توجه به اینکه گفتار یکی از مهمترین روشهای برقراری ارتباط میان انسانهاست، تلاش بسیاری برای برقراری ارتباط گفتاری میان انسان و ماشین صورت گرفته است. بازشناسی گفتار، یکی از دو محور اصلی در این ارتباط است. مدل‌های مارکوف پنهان^۱ (HMMs) امروزه به عنوان پرکاربردترین روش برای برآوردن این منظور مورد استفاده قرار می‌گیرند. یکی از مسائلی که همواره فراروی این گونه سیستمها (و به طور کلی هر سیستم بازشناسی گفتاری) بوده، مسئله تنوع گویندگان و ویژگیهای گفتاری آنها بوده است. این مطلب باعث شده که در بسیاری از موارد، محققان برای پرهیز از این مشکل، به ایجاد سیستمهای وابسته به گوینده^۲ (SD) روی آورند.

میکروفون، کانال انتقال و نویز محیط نیز می‌تواند مؤثر باشد.

برای جبران این عدم تطابقها و دستیابی به دقت بالاتر در بازشناسی گفتار یک گوینده جدید، روشهایی مطرح گردیده‌اند که می‌توانند در دو گروه کلی دسته بندی شوند:

۱- جبران سازی ویژگیهای استخراج شده از گفتار [۱]

۲- تطبیق مدل گوینده [۲ و ۳] که در آن پارامترهای مدل HMM تغییر می‌کنند.

هر چند که ترکیب این دو روش [۴] نیز می‌تواند به نتایج خوبی منتهی شود، در این مقاله ما تنها به بررسی عملکرد تطبیق مدل می‌پردازیم.

یک روش تطبیق خوب روشی است که بتواند با در اختیار داشتن داده آموزشی کمی از یک گوینده جدید دقت بازشناسی را (تا حد امکان) افزایش دهد. از آن مهمتر اینکه با موجود بودن داده آموزشی کافی (برای آموزش عادی سیستم با گفتار گوینده جدید)، دقتی بهتر یا معادل روشهای معمول آموزش، نظیر درست‌نمایی پیشینه^۵ (ML)، ارائه دهد. روشهای کمی وجود دارند که هر دو مورد فوق را در برگیرند. عمومترین این روشها، روش تطبیقی است که براساس تئوری بیز^۶ بنا شده است و تخمین MAP نامیده می‌شود [۲ و ۵]. در این روش پارامترهای مدل به عنوان متغیرهای تصادفی که تابع چگالی احتمال آنها موجود است در نظر گرفته می‌شوند. در این حالت تخمین MAP را می‌توان به دست آوردن مد تابع چگالی احتمال پیشینه^۷ با وجود داده آموزشی توصیف کرد. هنگامی که داده کمی از گوینده جدید برای آموزش وجود دارد نتایج به دست آمده از تخمین MAP می‌تواند بسیار بهتر از نتایج به دست آمده از تخمین ML باشد. با افزایش داده‌های آموزشی، کارایی تخمین MAP به سمت کارایی ML میل می‌کند و توانایی بازشناسی همانند یک سیستم SD را ارائه می‌کند. ضعف روش MAP در این است که تنها مدل‌هایی که داده آموزشی متناظر آنها وجود دارد تطبیق می‌شوند و این امر هنگامی که داده آموزشی کمی (مثلاً چند جمله، در یک سیستم بازشناسی گفتار پیوسته با دایره لغات متوسط) در اختیار است منجر به نتایج خوبی نمی‌شود.

روشهای دیگری نیز وجود دارند که اساساً با تخمین MAP تفاوت دارند. مهمترین این روشها، روشهای مبتنی بر تبدیل‌اند. از جمله این روشها می‌توان به ^۸ MLLR [۳]، ^۹ VFS [۶] و ^{۱۰} SM [۴] اشاره کرد. همچنان‌که از نام آنها برمی‌آید، در این روشها، برای دستیابی به تطبیق پارامترها، اقدام به یافتن یک تابع نگاشت یا تبدیل کلی یا جزئی بر اساس داده‌های تطبیق موجود می‌شود. در این خانواده از تکنیکها، برای رفع مشکل تعداد زیاد پارامترهای آزاد در مدل، این پارامترها به کمک گره زدن یا با اعمال محدودیتهایی بر روی آنها کاهش می‌یابند. این امر باعث می‌شود که با افزایش دادگان آموزشی پارامترها به سمت مقادیر مورد انتظار از روش درست‌نمایی پیشینه (ML) میل نکنند.

هیچ یک از روشهای مبتنی بر تبدیل و روش تخمین MAP به طور همزمان دو شرط بهبود بازشناسی با دادگان آموزشی کم و نزدیک شدن به تقریب ML در صورت آموزش با داده کافی را در بر نمی‌گیرند. به همین دلیل روشهای دیگری سعی کرده‌اند همزمان به این دو ویژگی دست یابند. از جمله این روشها، روشهای ^{۱۱} EMAP [۷]، ^{۱۲} SMAP [۹]، ^{۱۳} RMP [۱۰] و روش Quasi-Bayes با همبستگی میان بردارهای میانگین [۱۱] هستند که از تخمین MAP مشتق شده‌اند. در این روشها بر اساس همبستگی میان مدلها، همه مدلها تازه سازی می‌شوند به این ترتیب که اگر داده آموزشی متناظر با یک مدل موجود باشد تمام مدل‌هایی که با مدل فوق همبستگی دارند نیز با این داده آموزشی، آموزش می‌بینند. هر چند که این روشها در تئوری کاملاً عمومی‌اند اما در عمل احتیاج به تقریب زیادی دارند زیرا یافتن و استفاده از همبستگی بین مدلها، به ویژه در زمانی که داده آموزشی بسیار کم است، کار مشکلی است. به عنوان مثال در مراجع [۷ و ۸] مدلها به چندین زیر مدل تقسیم شده‌اند که تعداد این زیر مدلها به میزان داده آموزشی بستگی دارد.

از سوی دیگر توسعه برخی روشهای مبتنی بر تبدیل، مانند MLLR، به نحوی که تخمین MAP را در برگیرند، نظیر روش ^{۱۴} MAPLR [۱۲ و ۱۳]، نیز به نتایج خوبی منتهی شده است. در این روش ماتریس تبدیل پیشینه از مدل قبلی برای آموزش به همراه

ماتریس تبدیل به دست آمده از روش MLLR به مدل اعمال می‌شود. در مقاله حاضر روش جدیدی ارائه شده است که اساس آن نیز بر روشهای MLLR و MAP است و مزایای هر دو روش را در بر می‌گیرد. در این روش، دو شیوه MLLR و MAP به گونه دیگری توأمأ مورد استفاده قرار می‌گیرند به این نحو که ابتدا به کمک خوشه بندی مدل‌های HMM که به صورت سه آوایی^{۱۵} آموزش دیده‌اند مدل‌ها دسته بندی می‌شوند. سپس به کمک داده‌های آموزشی و روش MLLR برای هر خوشه یک ماتریس تبدیل به دست می‌آید. یک تبدیل عمومی هم برای خوشه‌هایی که هیچ‌گونه بردار آموزشی ندارند به دست می‌آوریم. حال هر مدلی که داده آموزشی آن موجود باشد به کمک تخمین MAP تطبیق می‌یابد و پارامترهای مدل‌هایی که داده آموزشی ندارند، با استفاده از مدل تبدیل یافته توسط ماتریس تبدیل MLLR به عنوان مقادیر کمکی و به کمک تخمینی مشابه MAP، تطبیق می‌شود. به این ترتیب، علاوه بر استفاده از مزایای هر دو روش، معایبی که هر یک از آنها به تنهایی دارند نیز پوشانده می‌شود. پایگاه داده مورد استفاده در این مقاله *فارس دات* است که یک دادگان با دایره کلمات متوسط است.

مقاله حاضر مشتمل بر شش بخش است که پس از مقدمه، بخش دوم به بررسی اجمالی روشهای مبتنی بر تخمین MAP اختصاص یافته است. در بخش سوم به روش تطبیق MLLR و مسائل مربوط به آن پرداخته شده و بخش چهارم به طرح روش پیشنهادی این مقاله می‌پردازد. در بخش پنجم به مسائل مربوط به پیاده سازی و بررسی نتایج می‌پردازیم و در بخش ششم نیز نتیجه گیری و پیشنهادات ارائه شده است.

۲- تطبیق بیزین

۲-۱- تخمین MAP

تخمین MAP یا بیزین^{۱۶} از جمله روشهایی است که برای تطبیق پارامترهای HMM به کار می‌رود. تفاوت این روش با روش تخمین درستنمایی بیشینه در این است که در روش ML فرض بر این است که پارامترهای مدل ثابت ولی ناشناخته‌اند

حال آنکه در تخمین MAP این پارامترها متغیرهای تصادفی فرض می‌شوند که توزیع پیشینه^{۱۷} آنها مشخص است. این قابلیت روش MAP را در شرایط داده‌های آموزشی پراکنده^{۱۸} کارتر می‌کند چرا که اطلاعات پیشینه را نیز در تطبیق دخالت می‌دهد.

برای تطبیق گوینده در روش MAP از دو منبع داده استفاده می‌شود؛ یکی چگالی توزیع پیشینه پارامترها $P(\lambda)$ و دیگری داده‌های تطبیقی O_t . روش MAP تلاش در به دست آوردن مدلی دارد که برای آن

$$\lambda_{MAP} \propto \arg \max_{\lambda} P(O_t | \lambda) \cdot P(\lambda) \quad (1)$$

معادلات مربوط به تخمین بردارهای میانگین و واریانس توسط روش MAP برای یک HMM دارای چگالی مشاهدات تک گوسین چند متغیره به صورت زیر است [۲].

$$\mu'_k = \frac{\tau_k \mu_k + \sum_{t=1}^n \gamma_k(t) \cdot o_t}{\tau_k + \sum_{t=1}^n \gamma_k(t)} \quad (2)$$

$$\Sigma'_k = \frac{2\beta_k + \sum_{t=1}^n \gamma_k(t)(o_t - \mu'_k)^2 + \tau_k(\mu_k - \mu'_k)^2}{2\alpha_k - 1 + \sum_{t=1}^n \gamma_k(t)} \quad (3)$$

که در این معادلات $\mu_k, \tau_k, \alpha_k, \beta_k$ پارامترهای توزیع پیشینه‌اند [۲].

علی‌رغم اینکه مزیت عمده روش MAP بر ML، استفاده از پارامترهای پیشینه در تخمین است، آنچه در تخمین به شیوه MAP عمدتاً مشکل ساز است، تخمین مناسب پارامترهای پیشینه است، از آنجا که لزوم داشتن پارامترهای پیشینه مناسب انکار ناپذیر است، تلاش فراوانی برای یافتن پارامترهای پیشینه مناسب صورت گرفته که عموماً به راه‌های غیر تحلیلی منجر شده و هنوز یک راه تحلیلی مناسب برای حل این مسئله یافت نشده است [۱۴].

۲-۲- تخمین MAP وزنی

روش MAP وزنی روشی است که عمدتاً برای استفاده در کاربردهایی که دارای پارامتر آزاد زیاد بوده و از این رو تخمین

پارامترها در آنها دارای مشکل است طراحی شده است [۱۵]. مثالی از کاربرد این روش، تخمین پارامترها در یک سیستم وابسته به متن^{۱۹}، نظیر یک سیستم سه آوایی است. در این روش هم مشابه روش گره زدن، که در آموزش این گونه مدلها کاربرد فراوانی دارد، حالت‌های مشابه خوشه بندی می‌شوند. با مشاهده داده آموزشی متناظر با هر یک از مدل‌های موجود در یک دسته، مدل فوق با داده آموزشی جدید به کمک تطبیق MAP معمولی آموزش می‌بیند، اما مدل‌های دیگر هم دسته مدل فوق در صورت نداشتن داده آموزشی به کمک معادله زیر باز تخمین می‌شوند [۱۵]:

$$\mu_j^{(q)} = \frac{\tau_j((1-\varphi)\hat{\mu}_j^{(q)} + (\varphi \times \mu')) + \sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha^{(q)r}(t, j) \beta^{(q)r}(t, j) O_t^r}{\tau_j + \sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha^{(q)r}(t, j) \beta^{(q)r}(t, j)} \quad (4)$$

در معادله فوق $\hat{\mu}_j^{(q)}$ میانگین چگالی احتمال حالت مدل اولیه قبل از آموزش و μ' میانگین چگالی احتمال حالت مدل آموزش دیده‌ای در آن دسته است که داده آموزشی متناظر آن موجود است. φ نیز پارامتر رشد است که معمولاً مقداری بین صفر و کمتر از یک دارد و از معادله زیر قابل محاسبه است:

$$\varphi = \left| \frac{\hat{\mu}_j^{(q)} - \mu'}{\hat{\mu}_j^{(q)}} \right| \quad (5)$$

مقدار پارامتر φ در حقیقت نسبت تغییرات میانگین آموزش دیده و میانگین مدل اولیه است و به نوعی می‌توان آن را یک پارامتر برای وزن دهی براساس میزان تغییرات مدل بیان کرد. در صورت صفر شدن پارامتر φ تخمین MAP وزنی به سمت تخمین MAP معمولی میل خواهد کرد.

تئوری استفاده از چنین روشی برای تطبیق بر این اساس است که در تخمین MAP صورت رابطه را می‌توان به دو قسمت تجزیه کرد. یکی از این دو قسمت میزان درست‌نمایی داده آموزشی در مدل فعلی است و دیگری $\tau_j \hat{\mu}_j^{(q)}$ است که بیانگر میزان مشارکت میانگین توزیع پیشینه یعنی $\hat{\mu}_j^{(q)}$ در تخمین MAP است. این بخش در حقیقت میزان استفاده از

اطلاعات پیشینه را برای به دست آوردن میانگین جدید مشخص می‌کند. پارامتر $\hat{\mu}_j^{(q)}$ که در اینجا بیانگر اطلاعات پیشینه است، در MAP معمولی از میانگین چگالی احتمال حالت مدل SI قابل محاسبه است. بدیهی است که با نزدیکتر کردن این پارامتر به مدل متناظر SD می‌توان تخمین بهتری را به دست آورد. از آنجا که مدل‌های دسته بندی شده در یک گروه دارای مقادیر نزدیک میانگین‌اند، می‌توان میزان تغییرات یک مدل را که برای آن داده آموزشی وجود دارد به سایر مدل‌های هم گروه آن تعمیم داد و فرض کرد که تمام میانگین‌های موجود در آن مدل به یک نسبت تغییر می‌کنند. با این فرض $\hat{\mu}_j^{(q)}$ موجود در تخمین MAP معمولی را به دو بخش میانگین قبلی مدل SI و میانگین جدید مدل آموزش دیده در گروه، که با نسبت وزنی خاصی که از میزان تغییرات مدل SI و SD ناشی شده است ادغام می‌شوند، تبدیل می‌کنیم.

۳- روش MLLR

۳-۱- الگوریتم خوشه بندی

پیش از آنکه به تبدیل MLLR بپردازیم باید به نحوی دسته‌های تبدیل را تعیین کرد. این عمل به این خاطر انجام می‌شود که پارامترهایی که شبیه به هم هستند با هم در یک گروه قرار گیرند و یک تبدیل یکتا برای همه آنها استفاده شود. روشهای زیادی برای خوشه بندی وجود دارند که می‌توانند بر اساس یک درخت تصمیم‌گیری و یا دوری و نزدیکی بردارهای مدل از یکدیگر بنا شوند. از آنجا که مدل مورد استفاده در این کاربرد سه آوایی است، برای کاهش پارامترها از گره زدن حالتها استفاده شده است. از آنجا که مدل هر آوا از سه حالت و هر حالت از یک تک گوسین تشکیل شده است خوشه بندی عملاً به خوشه کردن گوسینها محدود شده است. برای خوشه بندی می‌توان از چنین الگوریتمی استفاده کرد [۱۶].

ابتدا برای هر یک از حالتها یک دسته جدا در نظر می‌گیریم. سپس دو دسته‌ای که کمترین فاصله را دارند می‌یابیم و اگر این فاصله از مقدار از قبل تعیین شده‌ای کمتر بود آن دو دسته را

است. پس از به دست آمدن میانگینهای جدید ($\hat{\mu}_j$) عملیات بازشناسی توسط این میانگینها انجام می‌گیرد. تبدیلهای به نحوی محاسبه می‌شوند که درستنمایی داده‌های آموزشی را بیشینه کنند. اگر T فریم از داده آموزشی $O = o_1, o_2, \dots, o_T$ موجود باشد احتمال حضور در حالت j ام در زمان t در حالی که داده O تولید شده است را با $\gamma_j(t)$ نمایش می‌دهیم.

$$\gamma_j(t) = \frac{f(O, \theta_t = j | \lambda)}{f(O | \lambda)} \quad (8)$$

از $\gamma_j(t)$ از الگوریتم Forward-Backward به سادگی برای داده‌های آموزشی محاسبه می‌شود. با فرض اینکه ماتریس کوواریانس قطری باشد و ماتریس تبدیل w_j بین R گوسی، $J_R \cdot J_0$ ، مشترک باشد، ماتریس تبدیل w_j ستون به ستون از رابطه $w_i = G_i^{-1} z_i$ به دست می‌آید. در این رابطه z_i ستون i ام از ماتریس Z است که به صورت

$$Z = \sum_{t=1}^T \sum_{j=1}^R \gamma_{jr}(t) \sum_{j_r=1}^{-1} o_t v'_{j_r} \quad (9)$$

است و G_i نیز به صورت

$$G_i = \sum_{r=1}^R c_{ii}^{(r)} v_{jr} v'_{jr} \quad (10)$$

نوشته می‌شود که $c_{ii}^{(r)}$ عنصر قطری i ام از ماتریس قطری کوواریانس مقیاس شده گوسی r ام در کلاس گوسیهای R است که از رابطه

$$C^{(r)} = \sum_{t=1}^T \gamma_{jr}(t) \sum_{j_r=1}^{-1} \quad (11)$$

به دست می‌آید. تازه سازی پارامترها به کمک روابط فوق یک مرحله از تطبیق MLLR است. اگر تغییرات در پارامترها با یک مرحله تطبیق به نحوی باشد که احتمال پیشینه اشغال آن حالت افزایش یابد، آن گاه با تکرار روش MLLR میزان درستنمایی مدل افزایش می‌یابد.

ادغام می‌کنیم و مجدداً دو دسته دیگر انتخاب می‌کنیم. این کار را آن قدر تکرار می‌کنیم تا الگوریتم همگرا شود.

برای یافتن فاصله دسته‌هایی که دارای چندین عضو هستند فاصله تمام اعضای هر دسته را با تمام عضوهای دسته دیگر محاسبه می‌کنیم و سپس از آنها میانگین می‌گیریم. در این الگوریتم از رابطه دیورژانس، به شرح زیر برای به دست آوردن فاصله بین دسته‌ها استفاده کرده‌ایم [۱۷].

$$d(S_1, S_2) = \left(\frac{1}{N} \sum_{i=1}^N \frac{(\mu_{1i} - \mu_{2i})^2}{\sigma_{1i} \sigma_{2i}} \right)^{\frac{1}{2}} \quad (6)$$

۳-۲- تطبیق MLLR

در این بخش به بررسی روش MLLR و روابط مورد استفاده در آن می‌پردازیم. جزئیات بیشتر در این خصوص در مرجع [۳] یافت می‌شود.

هر عنصر مخلوط گوسین که نماینده یک حالت در مدل HMM تک گوسین است توسط بردار میانه μ_j و ماتریس کوواریانس Σ_j مشخص می‌شود. در روش MLLR میانگینهای مدل SI به میانگینهای نا معلوم اما قابل تخمین مدل تطبیق یافته، یعنی $\hat{\mu}_j$ ، توسط یک تبدیل رگرسیون خطی نگاشت می‌شوند. این نگاشت از داده‌های آموزشی به طریق زیر به دست می‌آید:

$$\hat{\mu}_j = w_j \cdot v_j \quad (7)$$

که در آن w_j یک ماتریس تبدیل $n \times (n+1)$ است که n بعد بردارهای میانگین بوده و v_j بردار میانگین گسترش یافته به صورت $v_j = [1, \mu_{j1}, \dots, \mu_{jn}]'$ است.

اگر برای هر مدل یک w به دست آید، آنگاه این روش تطبیق به همان روش استاندارد Baum-Welch تبدیل می‌شود. اما برای اینکه با استفاده از داده کم تمام مدلها تطبیق یابند، کلاسهای رگرسیونی تعریف می‌شوند که شامل پارامترهای مدل گره خورده‌اند. برای هر کلاس یک ماتریس تبدیل به دست می‌آید که از داده‌های آموزشی متعلق به آن کلاس مشتق شده

۴- بهبود مدل‌های پیشینه در تخمین MAP مبتنی بر تبدیل MLLR² (MPIM)

علی‌رغم استفاده از دو شیوه MAP و MLLR، روش پیشنهادی MPIM، از نظر نحوه استفاده از این دو شیوه، کاملاً با روش MAPLR [۱۸] متفاوت است. در روش MAPLR اطلاعات مربوط به توزیع احتمال پیشینه ماتریسهای تبدیل مورد استفاده قرار می‌گیرد، به این معنی که اقدام به تخمین ماتریسهای انتقال MLLR به همراه پارامترهای مدل HMM، در چارچوب تخمین MAP، می‌شود، حال آنکه در روش پیشنهادی MPIM از مدل‌های تطبیق یافته به کمک ماتریس تبدیل MLLR برای بهبود مدل‌های پیشینه در تخمین MAP استفاده می‌شود.

این روش تنها بر روی میانگینهای توزیعی گوسین در مدل‌ها اعمال می‌شود. شیوه کار به این صورت است که ابتدا مدل‌های اولیه بر اساس شباهت و نزدیکی گوسین‌ها، که با معیار دیورژانس سنجیده می‌شود، خوشه بندی می‌شوند. سپس برای هر خوشه به دست آمده به شرط وجود حداقل سه نمونه داده آموزشی در آن خوشه یک ماتریس تبدیل w_1 به کمک روش MLLR به دست می‌آید. یک ماتریس w_g (ماتریس تبدیل کلی) نیز با استفاده از تمام داده‌های آموزشی تشکیل می‌شود. حال با استفاده از ماتریسهای تبدیل w_1 به دست آمده بردارهای میانگین جدید را برای توزیعی همه مدل‌ها حساب می‌کنیم. برای به دست آوردن میانگینهای جدید توزیعیها برای مدل‌های خوشه‌هایی که داده آموزشی کافی ندارند، از w_g استفاده می‌کنیم. با این حال، میانگینهای به دست آمده توسط ماتریسهای تبدیل مستقیماً جایگزین میانگینهای قبلی نمی‌شوند. در اینجا به کمک تخمین MAP مدل‌هایی که داده آموزشی متناظر آنها وجود دارد توسط آن داده‌ها آموزش می‌بینند و مدل‌هایی که داده آموزشی متناظر ندارند از مدل‌های تبدیل یافته برای بهبود مدل‌های پیشینه در تخمین MAP استفاده می‌کنند.

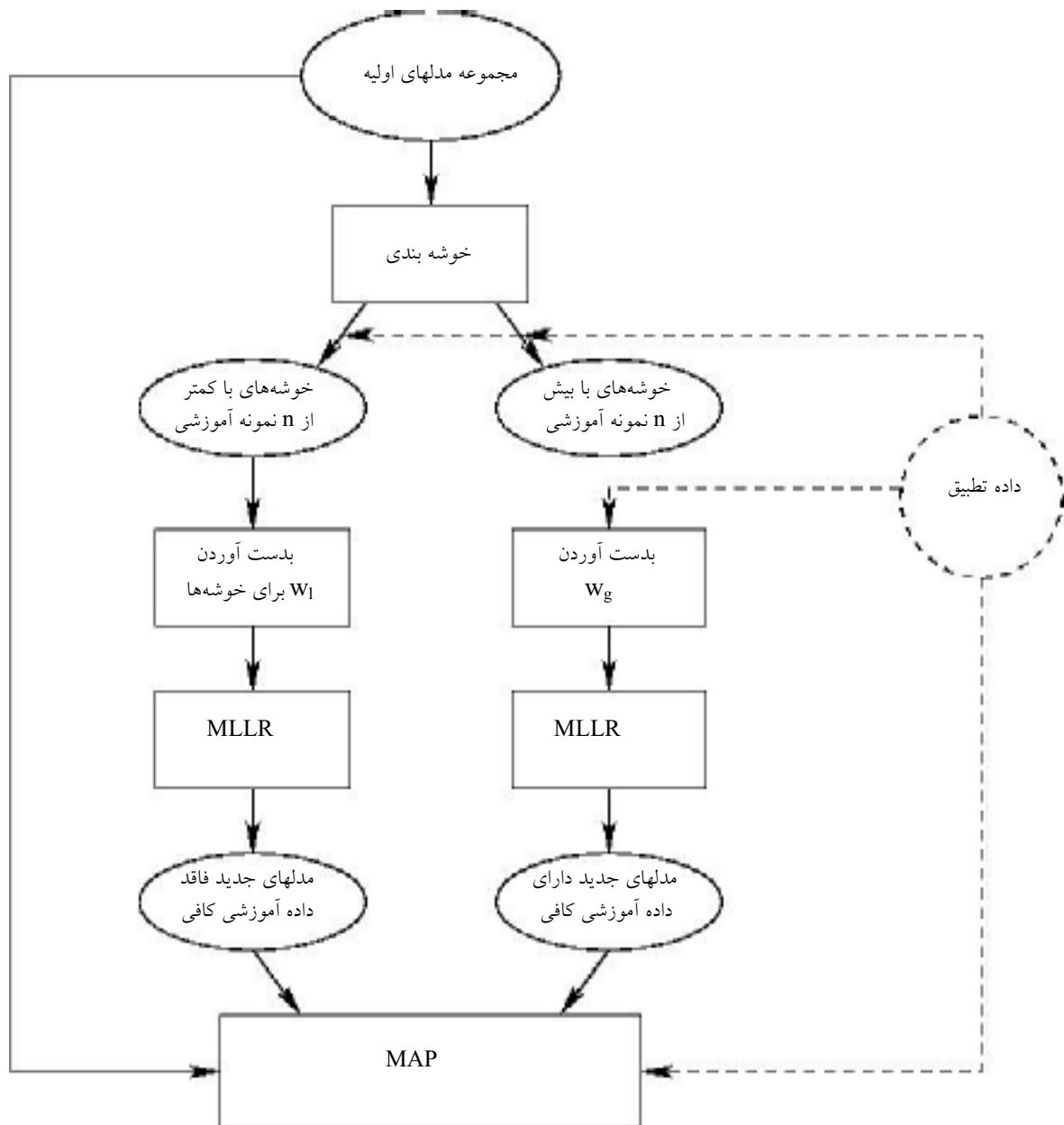
به کمک این روش با هر تعداد داده آموزشی ورودی، هر چند کم، تمام مدل‌های موجود به نوعی آموزش دیده و تغییر می‌یابند و چون از تخمین MAP برای تطبیق استفاده می‌کنند

مدل‌های به دست آمده هیچ گاه اشیاع نمی‌شوند و همچنین، آنچنان که در بخش بعد خواهیم دید با دادگان کم، دقت بازشناسی به نحو چشمگیری افزایش می‌یابد. نمودار جعبه‌ای (۱) مراحل اجرای این الگوریتم را نشان می‌دهد.

اساس کار به این صورت است که در رابطه MAP وزنی، به جای استفاده از مدل آموزش دیده توسط روش MAP از مدل تبدیل یافته توسط ماتریس تبدیل به دست آمده از روش MLLR استفاده می‌گردد. بر این اساس، فرض می‌شود که $\hat{\mu}_j = w_j \times v_j$ بردار میانگین حالت زام تبدیل یافته توسط روش MLLR باشد که در آن w_j تبدیل متناظر با آن حالت و v_j بردار میانگین گسترش یافته مدل SI است. همچنین اگر میانگین مدل SI متناظر با آن حالت با $\hat{\mu}_j^{(q)}$ نمایش داده شود، آن گاه می‌توان روش وزنی MAP و MLLR را به صورت رابطه (۴) برای تطبیق میانگینهای حالت زام نوشت که در آن $\mu_j^{(q)}$ میانگین تطبیق یافته با روش MAP و MLLR است. در این حالت می‌توان پارامتر ϕ را مطابق با شیوه به کار رفته در روش MAP وزنی به صورت رابطه (۵) نوشت. در این رابطه وزن ϕ نشان دهنده میزان مشارکت هر یک از مدل‌های SI و مدل تطبیق یافته توسط روش MLLR در تعیین اطلاعات پیشینه مدل است. در این رابطه هر چه مقدار ϕ به سمت یک نزدیکتر شود، مدل به سمت تطبیق MAP معمولی میل می‌کند و نقش مدل اولیه SI در تخمین MAP بارزتر می‌شود. اگر ϕ به سمت صفر میل کند، میانگین تبدیل یافته توسط روش MLLR به عنوان میانگین احتمال پیشینه مدل در تخمین MAP خودنمایی می‌کند. عملاً با زیاد شدن مقدار ϕ عملیات تطبیق کند می‌شود و نیاز به داده آموزشی بیشتری دارد ولی در عوض به اشیاع نمی‌رود و با افزایش دادگان تطبیقی قابلیت بهینه شدن مدل را فراهم می‌آورد.

این خصوصیت دقیقاً معادل خصوصیت روش MAP است.

در مقابل با کاهش میزان ϕ می‌توانیم نقش تبدیل MLLR را در تطبیق افزایش دهیم. یعنی با میزان داده کمتر آموزشی سریعاً مدل‌ها به سمت مدل SD میل می‌کنند و نرخ بازشناسی نسبت به تخمین MAP افزایش قابل ملاحظه‌ای می‌یابد اما سیستم به



شکل ۱- روش تطبیق پیشنهادی (MPIM).

دیده است [۱۵]. برای آموزش این سیستم از حدود ۱۸۰۰ جمله که از گویندگان دارای لهجه تهرانی در این دادگان به دست آمده استفاده شده است. هر گوینده در حدود ۲۰ جمله را ادا کرده که دو جمله بین تمام گوینده‌ها مشترک است. در این سیستم، ابتدا سیگنال گفتار از نرخ اصلی نمونه برداری ۴۴/۱ kHz به ۱۶ kHz کاهش نرخ داده می‌شود. پس از انجام عملیات پردازشی ۱۲ ضریب LP کپستروم، ۱۲ ضریب دلتا

اشباع می‌رود. در عمل شروع به اشباع رفتن سیستم از صد جمله آموزشی به بعد است که برای عملیات تطبیق سریع بسیار مناسب است.

۵- پیاده سازی و نتایج

سیستم SI مورد استفاده برای تطبیق گوینده سیستم HMM سه آوایی است که با استفاده از دادگان فارس دات^{۲۱} آموزش

کپستروم و یک ضریب لگاریتم انرژی از هر فریم ۲۵ میلی ثانیه‌ای استخراج می‌شود. برای هر آوا از سه حالت استفاده شده است که هر حالت از یک گوسین تشکیل یافته است. لازم به اشاره است که سیستم فوق به علت قابل دسترس بودن مورد استفاده واقع شده و بدیهی است که بخش‌های مختلف آن و بویژه بخش استخراج ویژگی‌ها در این سیستم بازنشاسی گفتار می‌توانند از شیوه‌های مناسبتری بهره ببرند.

برای آموزش این سیستم از گره زدن حالتها استفاده شده است. پس از اعمال الگوریتم خوشه بندی ۸۰ خوشه به دست آمد. برای آزمون الگوریتمها از سه گوینده مرد با لهجه معمولی استفاده شد. از هر گوینده ۳۰۰ جمله ضبط شد که در دو گروه آموزش و تطبیق (۲۰۰ جمله) و آزمایش (۱۰۰ جمله) طبقه‌بندی شد. سپس نتایج بازنشاسی برای هر گوینده توسط سیستم SI و تطبیق با روشهای MAP و MLLR و روش تلفیقی پیشنهادی MPIM به ازای ۵ و ۲۰ و ۵۰ و ۱۰۰ جمله آموزشی به دست آمد. لازم به اشاره است که استفاده از رابطه (۵) در تعیین پارامتر ϕ به جواب قابل قبولی منجر نگردید. علت این امر احتمالاً آن است که در روش MLLR مقادیر بدست آمده برای μ_j^i ممکن است با مقدار قبلی مدل SI یعنی $\hat{\mu}_j^{(q)}$ تفاوت زیادی داشته باشند. در نتیجه استفاده از وزنهای ثابت به جای وزنهای متغیر آزمایش گردید که به نتایج بسیار خوبی منتهی شد. مقدار بهینه‌ای که برای ϕ بدست آمده و کلیه آزمایشات براساس آن می‌باشد مقدار $\phi = 0.3$ است. در جدول (۱) متوسط نتایج بازنشاسی برای سه گوینده مشاهده می‌شود. این نتایج با استفاده از گرامر زوج کلمه به دست آمده‌اند. این در حالی است که مدل SD دقت بازنشاسی معادل ۹۶٪ را داراست که با ۲۰۰ جمله آموزشی حاصل شده است.

نتایج ارائه شده در جدول (۱) حاوی نکات جالبی‌اند. تطبیق MLLR به خوبی با میزان کم داده تطبیق توانسته است بهبودی مناسبی کسب کند، اگرچه با افزایش تعداد جملات، این میزان بهبودی نیز افزایش یافته ولی این افزایش چندان زیاد نیست. برای MAP با تعداد بسیار کم جملات تطبیق، حتی کاهش دقت

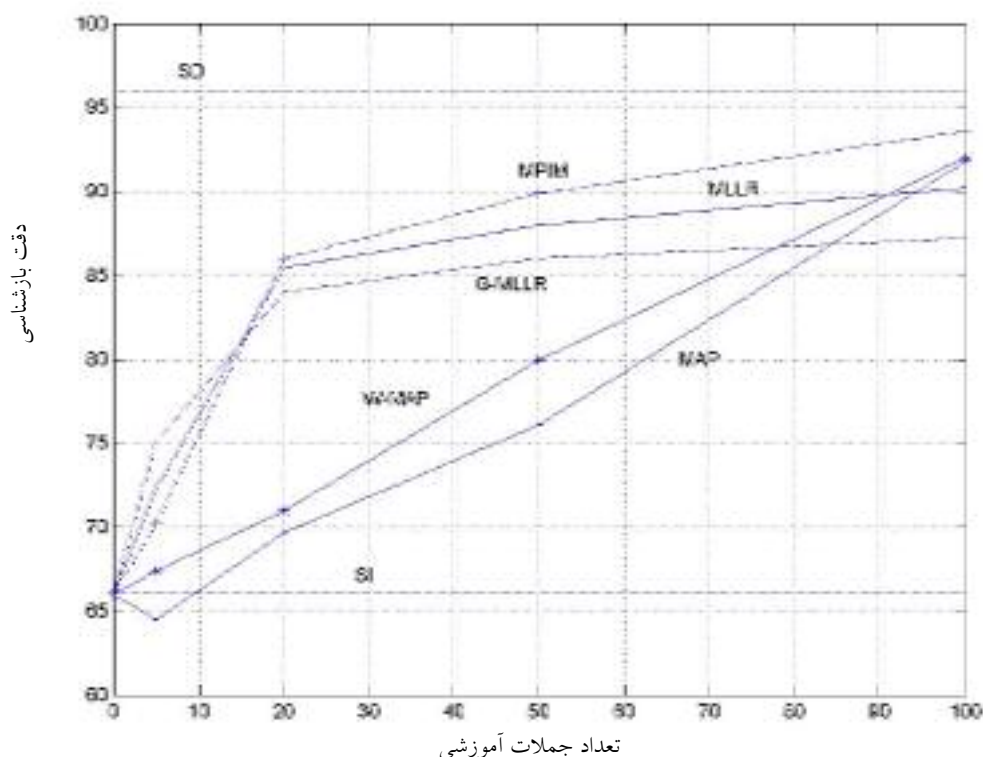
بازنشاسی نیز دیده می‌شود که ناشی از پراکندگی زیاد داده آموزشی و تخمین نادرست ناشی از آن برای برخی پارامترهاست. در مقابل با افزایش داده تطبیق، دقت بازنشاسی افزایش می‌یابد و با ۱۰۰ جمله تطبیق، دقت نسبت به MLLR نیز فزونی می‌گیرد که در واقع نشان دهنده خاصیت MAP در گرایش تدریجی تخمین به سمت تخمین ML با افزایش داده است. نکته قابل توجه در این جدول آن است که روش MPIM با ۵۰ جمله تطبیق حدود ۸۰٪ و با ۱۰۰ جمله تطبیق حدود ۹۲٪ فاصله بین دو سیستم SI و SD را پیموده، حال آنکه این مقادیر برای روش MAP به ترتیب حدود ۳۴٪ و ۸۶٪ و برای روش MLLR به ترتیب حدود ۷۳٪ و ۸۱٪ است.

شکل (۲) روند بهبود نرخ بازنشاسی را برای چند روش تطبیق گوینده با افزایش تعداد جملات تطبیق نشان می‌دهد. در اینجا نتایج برای روشهای MAP، MAP و زنی، MLLR، مLLR عمومیته یافته (G-MLLR) و MPIM نشان داده شده‌اند. همچنانکه ملاحظه می‌شود، روشهای مبتنی بر MLLR، در مقایسه با روشهای مبتنی بر MAP، با تعداد جملات کم، به بهبودی قابل ملاحظه‌ای در نرخ بازنشاسی دست می‌یابند، حال آنکه این روشها، با افزایش تعداد جملات تطبیق، کارایی خود را از دست داده و نمی‌توانند همانند MAP مرتباً با افزایش داده آموزشی به سمت توانایی یک سیستم وابسته به گوینده گرایش یابند. در مقابل، روش MPIM از هر دو مزیت برخوردار بوده و در حجم کم داده از قابلیت‌های MLLR و در حجم زیاد داده از تواناییهای MAP بهره می‌برد.

در مجموع می‌توان گفت که روش تلفیقی پیشنهادی با توجه به استفاده از مزیت‌های MLLR در اعمال تابع انتقال عمومی و MAP در آموزش مناسب، به دقتی بیش از هر دو روش فوق در جملات تطبیقی متوسط تا زیاد دست می‌یابد. لازم به اشاره است که کاهش دقت MAP در ۵ جمله تطبیقی، تاثیر منفی خود را بر روی روش تلفیقی پیشنهادی نیز گذاشته است. با اعمال شرط حداقل نمونه دیده شده برای تطبیق پارامترها در MAP می‌توان این مشکل را نیز تا حد زیادی مرتفع نمود.

جدول ۱- میانگین نتایج به دست آمده از اجرای الگوریتمها

روش	جمله ۵	جمله ۲۰	جمله ۵۰	جمله ۱۰۰
SI	۶۶/۰			
SD	۹۶/۰			
MLLR	۷۲/۳	۸۵/۵	۸۸/۰	۹۰/۲
MAP	۶۴/۵	۶۹/۷	۷۶/۱	۹۱/۸
MPIM	۷۰/۲	۸۶/۰	۸۹/۹	۹۳/۶



شکل ۲- مقایسه بهبودی در نرخ بازشناسی با افزایش داده تطبیق برای روش MPIM در مقایسه با چندین روش دیگر تطبیق گوینده.

پراستفاده‌ترین زمینه برای اینگونه الگوریتم‌ها، تطبیق گوینده با نظارت یکباره است که عموماً به صورت off-line انجام می‌گیرد و این افزایش حجم محاسبات مشکل عمده‌ای برای آن تلقی نمی‌گردد.

۶- نتیجه‌گیری و پیشنهادات

در این مقاله ضمن طرح روشهای مختلف تطبیق گوینده نظیر MAP و MLLR و روش تلفیقی پیشنهادی MPIM، نتایج پیاده‌سازی و اجرا ارائه گردید. نتایج به دست آمده حکایت از

مسئله پیچیدگی الگوریتم تطبیق نیز در اینجا قابل اشاره است. بدست آوردن دو گونه ماتریس تبدیل و اعمال دو گانه MLLR، اگرچه حجم محاسبات را به دقتاً دو برابر تبدیل افزایش نمی‌دهد، ولی در مجموع باعث افزایش نسبی حجم محاسبات می‌گردد. افزوده شدن MAP به این مجموعه نیز به نوبه خود موجب افزایش میزان محاسبات می‌گردد. در مجموع می‌توان گفت که این الگوریتم، نسبت به هر دو روش MAP و MLLR از حجم محاسبات بیشتری برخوردار است. با این همه، این مطلب نمی‌تواند در عمل چندان مشکل آفرین باشد چرا که

گوسین به جای حالت‌های تک مخلوطی، پیاده‌سازی الگوریتم خوشه‌بندی دقیقتر و استفاده از ساختار درختی در خوشه‌بندی همانند روشی که در SMAP پیاده‌سازی شده است پیشنهاد می‌شود.

افزایش دقت بازشناسی با استفاده از این روش دارد به طوری که در تعداد جملات متوسط به بالا نسبت به هر دو روش MLLR و MAP بهبودی حاصل شده است. برای بهبود هر چه بیشتر نتایج و ادامه کار استفاده از حالت‌هایی با چند مخلوط

واژه نامه

- | | | |
|---|---|---|
| 1. hidden Markov models | 10. stochastic matching | 17. prior |
| 2. speaker dependent | 11. extended MAP | 18. sparse training data |
| 3. speaker independent | 12. structural MAP | 19. context-dependent |
| 4. speaker pooling | 13. regression-based model prediction | 20. MLLR-based prior improvement for MAP estimation |
| 5. maximum likelihood | 14. maximum <i>a posteriori</i> linear regression | 21. FarsDat |
| 6. Bayes | 15. triphone | 22. generalized MLLR |
| 7. posterior | 16. Bayesian | |
| 8. maximum likelihood linear regression | | |
| 9. vector field smoothing | | |

مراجع

- Lee, C-H., "On Stochastic Feature and Model Compensation Approaches to Robust Speech Recognition," *Speech Communication*, vol. 25, pp. 29-47, 1998.
- Gauvain, J-L. and Lee, C-H., "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291-298, 1994.
- Leggetter, C.J. and Woodland, P.C., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous-Density Hidden Markov Models," *Comput. Speech Lang.*, vol. 9, pp. 171-185, 1995.
- Sankar, A. and Lee, C-H., "A maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 190-202, May 1996.
- Lee, C-H., Lin, C-H. and Juang, B-H., "A Study on Speaker Adaptation of the Parameters of Continuous-Density Hidden Markov Models," *IEEE Trans. Acoustic, Speech, Signal Processing*, vol. ASSP-39, pp. 806-814, Apr. 1991.
- Ohkura, K., Sugiyama, M., and Sagayama, S., "Speaker Adaptation Based on Transfer-Vector-Field Smoothing with Continuous Mixture Density HMMs," in *Proc. Int. Conf. Spoken Language Processing '92*, pp. 369-372, 1992.
- Zavaliagos, G., and Schwartz, R., "Maximum *a Posteriori* Adaptation for Large-Scale HMM Recognizers," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '95*, pp. 725-728, Detroit, 1995.
- Stern, R.M. and Lasry, M.J., "Dynamic Speaker Adaptation for Feature-Based Isolated Word Recognition," *IEEE Trans. Acoustic, Speech, Signal Processing*, vol. ASSP-35, pp. 751-763, June 1997.
- Shinoda, K., and Lee, C-H., "A Structural Bayes Approach to Speaker Adaptation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, No. 3, March 2001.
- Ahadi, S.M., and Woodland, P.C., "Combined Bayesian and Predictive Techniques for Rapid Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 11, No. 3, July 1997.
- Lee, C-H., "On-Line Adaptive Learning of the Correlated Continuous-Density Hidden Markov Model for Speech Recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 386-397, July 1998.
- Siohan, O., Chesta, C., and Lee, C-H., "Hidden Markov Model Adaptation Using Maximum *a Posteriori* Linear Regression," In *Proc. Workshop Robust Methods Speech Recognition in Adverse Conditions*, Tampere, Finland, pp. 147-150, May 1999.
- Siohan, O., Chesta, C., and Lee, C-H., "Joint Maximum *A Posteriori* Adaptation of Transformation and HMM Parameters". *IEEE Transactions on Speech and Audio Processing*, vol. 9, No. 4, May 2001.
- Huo, Q., and Chan, C., "Bayesian Adaptive Learning of Parameters of Hidden Markov Model for Speech Recognition," Technical Report, Department of Computer Science, University of Hong Kong, 1992.
- شمس، س. ح. و احدی، س. م. "روش آموزش وزنی وابسته

به متن در گفتار پیوسته فارسی، نشریه علمی پژوهشی
/میرکبیر، شماره ۵۳، سال ۱۴، ص ۹۲-۸۰ زمستان ۱۳۸۱.

16. Ahadi, S.M., "Reduced Context Sensitivity in Persian Speech Recognition via Syllable Modeling," in *Proc. SST-2000*, pp. 492-497, Canberra.
17. HTK Hidden Markov Model Toolkit ver. 3.1, Reference Manual, Cambridge University Engineering Department, p. 76, 2001.

18. Siohan, O., Myrvoll, T-A. and Lee, C-H., "Structural Maximum *a Posteriori* Linear Regression for Fast HMM Adaptation" In *Workshop on Automatic Speech Recognition: Challenges for the new Millenium*, Paris, France, Sept.2000, ISCA ITRW ASR2000.