# آموزش وزنی مدل های وابسته به متن در بازشناسی گفتار پیوسته

سید حسین شمس
کارشناسی ارشد

سید محمد احدی
استادیار

دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر

## چکیده

در بازشناسی گفتار پیوسته مدل سازی وابسته به متن می تواند موجب افزایش دقت بازشناسی و در عین حال افزایش تعداد مدلها شود. این افزایش در تعداد مدلها می‌تواند مواجهه با کمبود داده های آموزشی را موجب گردد. روشهای مختلفی همانند گره زدن پارامترهای مدلها یا استفاده از تخمین MAP برای غلبه بر این مشکل به کار می‌روند. در این مقاله روشی جدید برای آموزش ارائه شده است که محاسن هر دو روش MAP و گره زدن را تواماً داراست و به نتایج بهتری نیز منجر می شود. مدلهای مورد استفاده از نوع سه آوائی می باشند.

## کلمات کلیدی

مدلهای مارکوف پنهان با چگالی پیوسته، بازشناسی گفتار پیوسته فارسی، مدلسازی وابسته به متن، آموزش بام ـ ولش، گره زدن حالتها، آموزش MAP .

# Weighted Training of Context Dependent Models for Continuous Speech Recognition

S.H. Shams
M.Sc. candidate

S.M. Ahadi
Assistant Professor

Electrical Engineering Department,
Amirkabir University of Technology

## Abstract

*Context dependent modelling is a well-known technique used to improve recognition rate in continuous speech recognition. However, it can also dramatically increase the overall system parameter count which usually leads to a grave problem known as sparse training data condition. Different techniques such as model parameter tying and MAP estimation are used to overcome this problem. In this paper, a new approach, namely weighted training, has been introduced which benefits from both MAP estimation and parameter tying strengths and also outperforms the widely used parameter tying approach. The triphone models are used in this implementation.*

## Keywords

*Continuous density hidden Markov models, Continuous Persian speech recognition, Context dependent modelling, Baum-Welch training, State tying, MAP estimation.*

## مقدمه

HMM ها معمولاً برای مدل کردن سیستمهای غیر خطی متغیر با زمان مورد استفاده قرار می‌گیرند. استفاده از مدلهای مارکوف پنهان در بازشناسی گفتار گسسته و پیوسته با دایره لغات بزرگ به نتایج قابل قبولی منجر شده است. انتظار می‌رود با استفاده از مدل‌سازی وابسته به متن [1]، به شرط آموزش مناسب، بتوان به نتایج بهتری دست پیدا نمود. مدل‌سازی وابسته به متن از آن جهت دارای اهمیت است که شرایط مختلف متنی را در مدل‌سازی آوای مورد نظر دخالت می‌دهد. به بیان دیگر، در این شیوه مدل‌سازی، اثر متن[2] نیز در مدل دخالت می‌کند و بنابراین در شرایط متنی مختلف، مدل‌های مختلفی مورد استفاده قرار می‌گیرند. در عمل نشان داده شده است که مشخصات آکوستیک هر واج به شدت تابع واج‌های پیش و پس از آن می‌باشد. این اثر حتی می‌تواند تا چند واج قبل و بعداز آن واج را هم پوشش دهد [1]. به همین دلیل برای غلبه بر این مشکل، بهترین راه حل مدل‌سازی آواها با توجه به شرایط متنی تشخیص داده شده است.

مدل‌سازی وابسته به متن، علیرغم مزیت بارزی که نسبت به مدل‌سازی بر اساس واج (ناوابسته به متن) دارد، از یک مشکل مهم رنج می‌برد و آن تعداد فوق العاده زیاد شرایط متنی است که هر واج می‌تواند در آن قرار گیرد. این امر منجر به افزایش فراوان تعداد مدل‌ها در مقایسه با شرایط عادی می‌گردد. این مسئله، علاوه بر افزایش حجم حافظه و پردازش در سیستم، باعث می‌شود که دستیابی به شرایط آموزش مقاوم[3] بسیار مشکل گردد. به طور معمول، در اغلب سیستم های بازشناسی گفتار، حتی در مواردی که دادگان‌های بسیار بزرگی برای آموزش سیستم در اختیار می‌باشد، بسیاری از مدل‌های وابسته به متن از مسئله کمبود داده آموزشی رنج می‌برند. برای غلبه بر این مشکل چند روش می‌تواند مورد استفاده قرار گیرد:

**الف-** برای جلوگیری از آموزش نامعتبر، تنها مدلهایی که میزان مناسبی داده آموزشی داشته باشند آموزش داده شوند. عیب این روش در این است که از کلیه اطلاعات استفاده نشده است و روش بهینه‌ای نیست.

**ب-** پس از بدست آوردن مدلهای اولیه، برای کم کردن پارامترهای سیستم از گره‌زدن استفاده شود. این کار می‌تواند موجب دستیابی به آموزش مقاوم شود. این شیوه و شیوه‌های مشابه در سیستم های بازشناسی زیادی پیاده سازی گردیده و به نتایج خوبی نیز منجر شده‌اند [2][3]. باید توجه داشت که هنگامی که خوشه‌بندی در این روش بر اساس داده صورت می‌پذیرد، چنانچه در ابتدا مدلها بعلت کمی داده آموزشی بخوبی آموزش داده نشده باشند، الگوریتم خوشه‌بندی لزوماً باعث گره خوردن مدلهایی که در واقعیت باید به هم شباهت داشته باشند نمی‌شود.

**ج-** آموزش MAP[4]: این شیوه آموزش که بر اساس تئوری بِیز[5] بنا گردیده، از یک مجموعه پارامتر پیشینه[6] بهره می‌برد. این روش آموزش به گونه‌ای عمل می‌کند که در صورت وجود داده آموزشی کافی مدل به سمت مدل بدست آمده از آموزش درستنمائی بیشینه (ML[7]) میل می کند و در صورت نبود دادهٔ آموزشی کافی به سمت مدل اولیه (پیشینه) تمایل دارد. استفاده از پارامترهای پیشینه موجب مقاوم شدن آموزش بویژه در شرایط داده آموزشی ناکافی می‌شود و از این رو MAP به عنوان شیوه‌ای موفق در آموزش در شرایط فوق محسوب شده و در روش‌های تطبیقی در بازشناسی گفتار نیز کاربرد فراوانی دارد [4][5]. مهم‌ترین مشکل در این روش بدست آوردن تخمین مناسب برای پارامترهای پیشینه است (رجوع به بخش ۵).

در این تحقیق آموزش بگونه ای طرح ریزی شده که برای آموزش هر مدل از مدلهای دیگر شبیه به آن مدل هم استفاده گردد (از این جهت شباهت به گره زدن دارد) ولی بگونه‌ای که اطلاعات مربوط به هر مدل در آموزش خود آن مدل مؤثرتر واقع می‌شود. در اینجا نیز در صورت وجود داده آموزشی کافی مدل به سمت مدل آموزش یافته به روش معمول (ML) میل می‌کند و در صورت نبود دادهٔ آموزشی کافی، شبیه به مدل ناوابسته به متن میشود (از این نظر شبیه بهMAP است).

محاسن این روش بشرح زیر است:

الف _ کلیه مدلها آموزش می‌بینند.

ب _ در صورت وجود دادهٔ آموزشی کافی اثر مدلهای دیگر ناچیز می‌گردد.

ج _ تفاوت بین مدلها هر چند ناچیز در مدل نهایی ظاهر می‌شود.

د ـ آموزش مقاوم و دقیق خواهد بود.

می‌توان گفت که در این روش از محاسن روشهای MAP و گره زدن پارامترها تواماً استفاده می‌شود. پایگاه داده مورد استفاده در این تحقیق *فارس دات* (FARSDAT ) میباشد که یک دادگان با دایره کلمات متوسط است.

مقاله حاضر شامل نه بخش می‌باشد که بخش اول مقدمه و بخش دوم به مدلهای زیر لغوی می‌پردازند. بخش سوم به نحوه آموزش، بخش چهارم به نحوه گره زدن، بخش پنجم به روش MAP، بخش ششم به روش پیشنهادی و بخش هفتم به پیاده‌سازی اختصاص یافته‌اند. در بخش هشتم به نتایج پرداخته شده است. بخش نهم نیز نتیجه‌گیری و مسیر تحقیقات آینده را نشان می دهد.

## ۲ـ مدلهای زیر لغوی

انتخاب واحدهای پایه و مدل کردن آنها اولین قدم در طراحی سیستم بازشناسی می‌باشد . اگر عمل آموزش بطور کامل انجام گرفته باشد هر چه تنوع واحدها بیشتر شود نتیجه بازشناسی مطلوبتر خواهد بود. در عین حال افزایش تعداد مدل‌ها ممکن است به عدم آموزش مناسب منجر شود.

در این تحقیق در مرحله اول مدلها بر اساس واجهای زبان فارسی انتخاب شدند و سپس به مدل وابسته به متن سه آوائی[۸] تعمیم یافتند.

### ۲ـ۱ـ مدلسازی وابسته به متن [۶]

منظور از مدلسازی وابسته به متن این است که برای هر واحد زیر لغوی با توجه به واحدهای زیر لغوی قبل یا بعد (شرایط متنی) مدل جداگانه‌ای در نظر گرفته شود.

در زیر، گونه‌هایی از واحدهای پایه استفاده شده اعم از نابسته یا وابسته به متن معرفی میشوند.

**الف ـ مدلهای تک آوائی**[۹]: تنها یک مدل هر واج را در تمام محلهای وقوع مدل می‌کند. جدول شماره ۱ واجهای اصلی زبان فارسی و نمایش سمبلیک آنها را در این کاربرد نشان میدهد. مثال زیر به نحوه نمایش یک جمله بر این اساس می‌پردازد.

d@r baz ‘@st

_ d @ r + b a z + ‘ @ s t _

**ب ـ مدلهای دو آوائی**[۱۰]: هر واج با توجه به واج سمت راست یا سمت چپ بطور جداگانه مدل می‌شود. برای سکوت و فاصله تنها یک مدل (بدون توجه به متن) در نظر می‌گیرند و در صورتی که واج کناری سکوت و یا فاصله باشد بصورت تک آوائی مدل می‌شود.

به این ترتیب جمله بالا بصورت زیر نمایش داده می‌شود.

مدل‌سازی دو آوائی چپ:

_ d d<@ @<r + b b<a a<z + ‘ ‘<@ @<s s<t _

مدل سازی دو آوائی راست:

_ d>@ @>r r + b>a a>z z + ‘>@ @>s s>t t _

جدول (۱) نمایش سمبلیک واحدهای استفاده شده براساس واجهای زبان فارسی با استفاده از کاراکترهای ASCII.

| Phonetic Group | Phoneme | Example |
|---|---|---|
| Vowels | a | ketab |
| | @ | s@br |
| | e | yek |
| | u | xub |
| | i | $ir |
| | o | sorx |
| | w | lwh |
| Liquids | r | rah |
| | l | lazem |
| Glides | y | yek |
| Nasals | m | morq |
| | n | niru |
| Plosives | b | b@d |
| | p | por |
| | t | tond |
| | d | dust |
| | q | morq |
| | k | kuh |
| | g | goft |
| | ' | e'lam |
| Fricatives | f | fut |
| | s | sabun |
| | v | vorud |
| | x | xord |
| | z | ziba |
| | # | #ale |
| | $ | $ire |
| | h | hale |
| Affricates | j | javid |
| | c | c@medan |
| | + | فاصله بین کلمات |
| | _ | سکوت |

لازم به اشاره است که در نمایش قراردادی فوق، سمبل قرار گرفته در سمت راست علامت > و یا سمت چپ علامت < نشانگر واحد اصلی و سمبل دیگر نشانگر واحد متنی است.

**ج ـ مدل های سه آوائی** : هر واج با توجه به واج سمت راست و سمت چپ بطور جداگانه مدل می‌شود. همانند قبل برای سکوت و فاصله بدون توجه به متن یک مدل در نظر می‌گیریم و در صورتی که واج کناری سکوت و یا فاصله باشد آنرا بصورت دو آوائی مدل می‌کنیم.

به این ترتیب جمله بالا بصورت زیر نمایش داده می‌شود.

_ d>@ d<@>r @<r + b>a b<a>z a<z + '>@ '<@>s @<s>t s<t _

توجه به این نکته لازم است که این گونه مدل‌سازی، سه آوائی داخل کلمه‌ای[11] نامیده می‌شود. شیوهٔ پیچیده‌تری به نام مدل‌سازی سه آوائی بین کلمه‌ای[12] نیز وجود دارد که در مرزهای کلمات از مدل دو آوائی استفاده نمی‌کند بلکه بهمراه فاصله، سکوت، آوای ابتدائی کلمهٔ بعدی و یا آوای انتهائی کلمهٔ قبلی یک مدل سه آوائی را بکار می‌گیرد. این شیوه موجب پیچیدگی فراوان در مدل‌سازی و بویژه در الگوریتم بازشناسی می‌گردد.

**د ـ مدلهای تعمیم یافته:** اگر چه ممکن است به نظر برسد که هیچ محدودیتی بر اینکه این فرآیند (افزایش تعداد واجهای کناری که در مدلسازی در نظر گرفته میشوند) تا کجا ادامه یابد وجود ندارد، در عمل عموماً مدلسازی به بیش از شرایط پنج آوائی[۱۳] (تا دو آوا از هر طرف) گسترش پیدا نمی‌کند.

## ۳ ـ آموزش

در مرحله آموزش ابتدا یک مدل اولیه مناسب برای واجها بدست می‌آوریم ( مدل تک آوائی ایجاد شده در [۷] ). سپس مدل تک آوائی را به عنوان مدل اولیه برای سه آوائی‌هایی که هسته آنها همان آوا است در نظر می‌گیریم . آنگاه مدلهای سه آوائی را آموزش می‌دهیم. بدیهی است که تعداد مدلهای بدست آمده از این مرحله به مراتب بیشتر از تعداد مدلهای اولیه (تک آوائی) می‌باشند.

### ۳ـ۱ ـ الگوریتم Baum-Welch و تخمین پارامترها

اگر $\alpha(t,i)$ احتمال مشاهده $t$ فریم اول باشد بطوری که در لحظه $t$ در حالت $I$ باشیم و $\beta(t,i)$ احتمال مشاهده $T=t-1$ فریم آخر باشد بطوری که در لحظه $t$ در حالت $i$ باشیم:

$$P(O(t)|\lambda)=\alpha(t,i) * \beta(t,i) \qquad (۱)$$

$$a_{ij}^{(q)} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha^{(q)r}(t,i) * a_{ij}^{(q)} * b_j^{(q)}(O_{t+1}^r) * \beta^{(q)r}(t+1,j)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha^{(q)r}(t,i) * \beta^{(q)r}(t,j)} \qquad (۲)$$

$$\mu_j^{(q)} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha^{(q)r}(t,j) * \beta^{(q)r}(t,j) * O_j^r}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha^{(q)r}(t,j) * \beta^{(q)r}(t,j)} \qquad (۳)$$

$$\Sigma_j^{(q)} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha^{(q)r}(t,j) * \beta^{(q)r}(t,j) * (O_j^r - \mu_j^{(q)})^2}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha^{(q)r}(t,j) * \beta^{(q)r}(t,j)} \qquad (۴)$$

که $R$ تعداد مشاهدات هر مدل و $T_r$ طول مشاهدات و $q$ شماره مدل است.

برای استفاده از این روابط در جملات بدون برچسب زمانی ابتدا مدل جمله از کنار هم قرار دادن تک تک مدلهای زیر لغوی بدست می‌آید. به این مدل FSN[۱۴] می‌گویند. سپس مقادیر جدید پارامترها با استفاده از روابط فوق تخمین زده می‌شود [۷][۸]. این الگوریتم باید بصورت مکرر روی داده‌های آموزشی اجرا گردد تا مقادیر مناسبی برای پارامترها بدست آید.

برای ساده نویسی از جایگزین‌های زیر در معادلات فوق استفاده خواهیم کرد.

$$an_{ij}^{(q)} = \sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r-1} \alpha^{(q)r}(t,i) * a_{ij}^{(q)} * b_j^{(q)}(O_{t+1}^r) * \beta^{(q)r}(t+1,j) \quad (۵)$$

$$ad_{ij}^{(q)} = \sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha^{(q)r}(t,i) * \beta^{(q)r}(t,j) \quad (۶)$$

$$\mu n_j^{(q)} = \sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha^{(q)r}(t,j) * \beta^{(q)r}(t,j) * O_j^r \quad (۷)$$

$$\mu d_j^{(q)} = \sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha^{(q)r}(t,j) * \beta^{(q)r}(t,j) \quad (۸)$$

$$\Sigma n_j^{(q)} = \sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha^{(q)r}(t,j) * \beta^{(q)r}(t,j) * (O_j^r - \mu_j^{(q)})^2 \quad (۹)$$

$$\Sigma d_j^{(q)} = \sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha^{(q)r}(t,j) * \beta^{(q)r}(t,j) \quad (۱۰)$$

## ٤ ـ گره زدن[۱۵]

همانگونه که گفتیم هرچه تنوع واحدها و بنابراین تعداد مدلها افزایش یابد، با شرط آموزش مناسب، نتیجه بازشناسی بهتر خواهد بود. اما با افزایش مدلها و با داده های آموزشی محدود امکان آموزش نامناسب پیش می‌آید. بنابراین باید با تمهیداتی ضمن کاهش تعداد پارامترها، امکان آموزش صحیح را برای پارامترها فراهم کنیم.

یکی از این روشها گره زدن پارامترها است. در بدست آوردن مدلهای سه آوائی جدید حالتها برای رخدادهای شبیه ولی جدا از هم هر آوا با هم گره زده می‌شوند. شکل ۱ پارامترهای مختلفی، که در یک سیستم مبتنی بر HMM میتوانند با هم گره زده شوند، را نشان می‌دهد. دایره‌ها پارامترهایی را نشان میدهند که میتوانند گره زده شوند [۲].

تقریباً کلیه پارامترهای HMM قابل گره زدن هستند. در فرآیند گره زدن ، پارامترهای گره زده شده در قسمتهای مختلف یکسان در نظر گرفته می‌شوند. مزیت گره زدن پارامترها در HMM این است که فرمولها ی اولیه بدون تغییر عمده قابل استفاده مجدد هستند و گره زدن همگرایی برنامه آموزش را نقض نمی‌کند. گره زدن باعث می شود پارامترهای مختلف بین مدلهای مختلف به اشتراک گذاشته شوند که در نتیجه با داده های آموزشی کم تخمین مقاومی از پارامترها بدست می‌آید [۹].

### ٤ـ۱ـ گره زدن حالت‌ها

حالتها گزینه خوبی برای گره زدن هستند، زیرا در عین حال که جزئی از هر مدل هستند می‌توانند مشخص‌کننده جزء زمانی هر آوا هم باشند [۳].

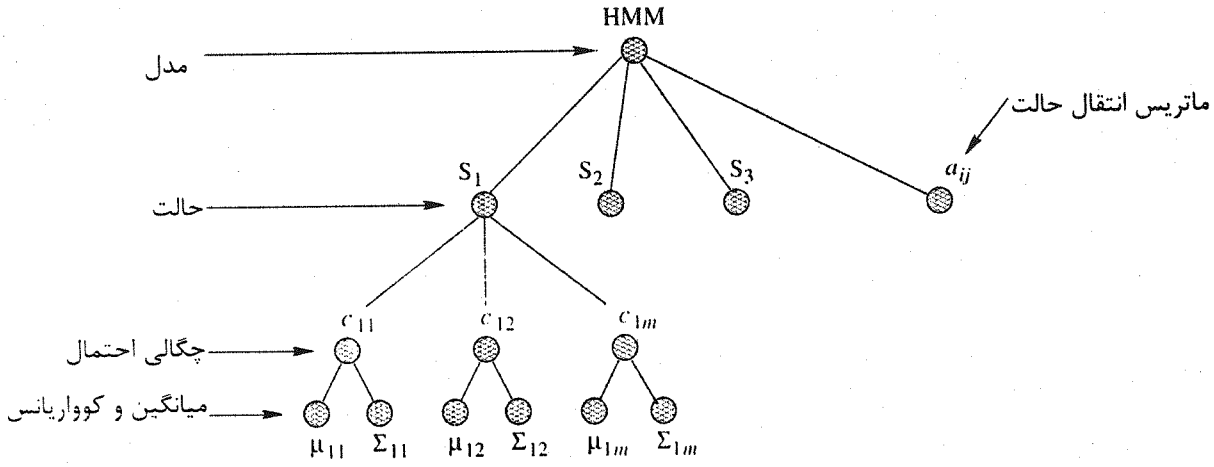
معادلات زیر چگونگی آموزش را در این حالت نشان می‌دهند.

$$a_{ij}^{(q)} = \frac{\sum_{n=1}^{N(q)} an_{ij}^{(M(q,n))} + an_{ij}^{(q)}}{\sum_{n=1}^{N(q)} ad_{ij}^{(M(q,n))} + ad_{ij}^{(q)}} \quad (۱۱)$$

$$\mu_j^{(q)} = \frac{\sum_{n=1}^{N(q)} \mu n_j^{(M(q,n))} + \mu n_j^{(q)}}{\sum_{n=1}^{N(q)} \mu d_j^{(M(q,n))} + \mu d_j^{(q)}} \qquad (۱۲)$$

$$\Sigma n_j^{(q)} = \frac{\sum_{n=1}^{N(q)} \Sigma n_j^{(M(q,n))} + \Sigma n_j^{(q)}}{\sum_{n=1}^{N(q)} \Sigma d_j^{(M(q,n))} + \Sigma d_j^{(q)}} \qquad (۱۳)$$

که $N(q)$ معرف تعداد حالاتی است که با حالت $q$ ام گره خورده اند و $M(q,n)$ ، $n$ امین حالتی است که با حالت $q$ گره خورده است.




**شکل (۱) ساختار سلسله مراتبی برای گره زدن پارامترها در HMM.**

### ٤ـ۱ـ۱ـ الگوریتم خوشه بندی[۱۶]

پیش از آنکه امکان گره زدن حالتها (یا بطور کلی پارامترها) در یک سیستم فراهم شود، باید مشخص نمود که کدام حالتها باید با یکدیگر گره بخورند. یک راه ممکن برای نیل به این هدف اجرای یک نوع خوشه‌بندی بر روی تمامی حالتهای سیستم است. در این شرایط معمولاً پارامترهای یک سیستم آموزش داده شده از نوع وابسته به متن مورد استفاده قرار می‌گیرند و خوشه‌بندی بر روی این پارامترها صورت می‌گیرد.

برای پیاده‌سازی عمل خوشه‌بندی می توان از چنین الگوریتمی استفاده نمود [۱۰].

۱ـ برای هریک از حالت‌های سیستم یک دسته حدا در نظر می‌گیریم.

۲ـ دو دسته‌ای که کمترین فاصله را دارند پیدا می‌کنیم.

۳ـ اگر این فاصله از مقدار از قبل تعیین شده‌ای کمتر بود این دو دسته را ادغام می‌کنیم و به قسمت ۲ برمی‌گردیم.

۴ـ پایان.

حالتهایی را که در پایان مراحل فوق در یک دسته قرار می‌گیرند با هم گره می‌زنیم.

برای یافتن فاصله دسته هایی که دارای چندین عضو هستند فاصله تمام اعضاء هر دسته را با تمام عضوهای دسته دیگر محاسبه می‌کنیم وسپس از آنها میانگین می‌گیریم. در این آلگوریتم از رابطه دیورژانس، به شرح زیر برای بدست آوردن فاصله بین دسته‌ها استفاده کرده‌ایم [۱۱].

$$d(S_1,S_2)=\left(\frac{1}{N}\sum_{i=1}^{N}\frac{(\mu_{1i}-\mu_{2i})^2}{\sigma_{1i}\sigma_{2i}}\right)^{\frac{1}{2}} \qquad (۱۴)$$

## ۵ ـ تخمین MAP

تخمین MAP که گاه تحت عنوان روش بیزین[۱۷] در آموزش HMMها نیز نامیده می‌شود، بر اساس تخمین بیشینهٔ پسین[۱۸] عمل می‌کند. در این روش، در تخمین پارامترهای HMMها علاوه بر مقادیر مشاهدات از مقادیر پیشینه نیز کمک گرفته می‌شود. به این ترتیب، بر خلاف تخمین درستنمائی بیشینه (ML) که در شرایط داده‌های آموزشی پراکنده[۱۹] به تخمینهای غیر قابل اعتماد منجر می‌شود، در این شرایط می‌توان تخمینهایی نه چندان دور از واقعیت بدست آورد. البته رسیدن به این هدف مشروط بر انتخاب پارامترهای پیشینه مناسب است.

فرمولهای آموزش MAP برای یک HMM دارای چگالی مشاهدات تک گوسین بصورت زیر است [۹].

$$a_{ij}^{(q)}=\frac{(\eta_{ij}-1)+an_{ij}^{(q)}}{\sum_{k=1}^{n}(\eta_{ik}-1)+ad_{ij}^{(q)}} \qquad (۱۵)$$

$$\mu_j^{(q)}=\frac{\tau_j\hat{\mu}_j^{(q)}+\mu n_j^{(q)}}{\tau_j+\mu d_j^{(q)}} \qquad (۱۶)$$

$$\Sigma_j^{(q)}=\frac{u_j^{(q)}+\tau_j(\hat{\mu}_j^{(q)}-\mu_j^{(q)})^2+\Sigma n_j^{(q)}}{(\gamma_j-p)+\Sigma d_j^{(q)}} \qquad (۱۷)$$

که در این روابط $p$ بعد بردار میانگین ( و ماتریس کواریانس) و مقادیر $\eta$، $\tau$، $\gamma$، $\hat{\mu}_j^{(q)}$ و $u_j^{(q)}$ پارامترهای توزیع پیشینه می‌باشند که در این میان $\tau$ و $\hat{\mu}_j$ پارامترهای توزیع گوسین چند متغیرهٔ پیشینه، $\gamma$ و $u_j$ پارامترهای توزیع و یشارت پیشینه[۲۰] و $\eta$ پارامتر توزیع دیریشله[۲۱] پیشینه می‌باشند [۹].

آنچه در تخمین به شیوهٔ MAP عمدتاً مشکل‌ساز است، تخمین مناسب پارامترهای پیشینه می‌باشد. از آنجا که مزیت اصلی تخمین MAP بر تخمین ML، استفاده از پارامترهای پیشینه است که در شرایط کمبود دادهٔ آموزشی از تخمین نامناسب پارامترهای مدل جلوگیری می‌نمایند، لزوم داشتن پارامترهای پیشینهٔ مناسب انکار ناپذیر است. لازم به تذکر است که علیرغم تلاشی که در این زمینه صورت گرفته، یک راه تحلیلی مناسب برای حل این مسأله یافت نشده است [۱۲].

## ۶ـ آموزش وزنی

تلاش در آموزش مدلهای وابسته به متن در اینجا بر این نکته متمرکز گشته که تا حد امکان از شباهت و نزدیکی مدلها در

حین آموزش و برای جبران مشکل کمبود داده‌ها استفاده نمود. بـه همیـن منظـور روابـط آمـوزش (بـاز تخمیـن) پارامترهـای CDHMM به شرح زیر تغییر داده شده‌اند.

$$a_{ij}^{(q)} = \frac{\varphi \sum_{n=1}^{N(q)} an_{ij}^{(M(q,n))} + an_{ij}^{(q)}}{\varphi \sum_{n=1}^{N(q)} ad_{ij}^{(M(q,n))} + ad_{ij}^{(q)}} \qquad (۱۸)$$

$$\mu_j^{(q)} = \frac{\varphi \sum_{n=1}^{N(q)} \mu n_j^{(M(q,n))} + \mu n_j^{(q)}}{\varphi \sum_{n=1}^{N(q)} \mu d_j^{(M(q,n))} + \mu d_j^{(q)}} \qquad (۱۹)$$

$$\Sigma n_j^{(q)} = \frac{\varphi \sum_{n=1}^{N(q)} \Sigma n_j^{(M(q,n))} + \Sigma n_j^{(q)}}{\varphi \sum_{n=1}^{N(q)} \Sigma d_j^{(M(q,n))} + \Sigma d_j^{(q)}} \qquad (۲۰)$$

که $N(q)$ معرف تعداد حالاتی است که با حالت $q$ ام شبیه درنظر گرفته شده‌اند و $M(q,n)$، $n$امین حالتی اسـت کـه شـبیه بـه حالت $q$ است. $\varphi$ پارامتری است بین ۰ و ۱ که میزان اثر مدلهای دیگر را در مدل $q$ تعیین می‌کند.

بطوری که مشاهده می‌شود، روابط فوق به هر دو روش گره زدن و MAP شباهتهایی دارند. در واقع اگر در این روابط $\varphi$ برابر یک قرار داده شود به فرمولهای گره زدن می‌رسیم، و هر چه این پارامتر کوچکتر شود نتیجه به سمت مدل سه آوائی گـره نخورده نزدیک می‌شود. این روش مشکل عمده MAP یعنی تعداد زیادی پارامتر آزاد که تاثیر عمده ای بر نتیجه آموزش داشته و باید به درستی تخمین زده شوند را ندارد و تنها یک پارامتر است که باید تعیین شود. در حـالی کـه محاسـن روش MAP از جهت کمک گرفتن از پارامترهای مدل های شبیه (که در آنجا مقادیر پیشینه نامیده می شوند) را داراسـت. در عیـن حـال، بـر خلاف شیوه گره زدن پارامترها، تفاوت بین مدلها را، حتی در پایین ترین سطوح، هم حفظ می‌کند.

مزیت دیگر این روش نسبت به MAP این است که عباراتی که در صورت یا مخرج با هم جمع می‌شـوند همنـوع هسـتند و در ضمن بنوعی بیانگر میزان نمونه‌های آموزشی استفاده شده در تخمین هستند. برای هر مدل، هر چـه کـه تعـداد نمونـه‌های آموزشی بیشتر باشد، مقدار بدست آمده از روابط (۵) تا (۱۰) بیشتر خواهد بود. تفاوت دیگـر ایـن روش بـا روش گـره زدن بـر اساس داده در این است که مستقیماً از مدل تک آوائی به مدل سه آوائی می‌رسیم و نیازی به دو مرحله آموزش، یکـی پیـش از مرحله خوشه بندی و گره زدن و دیگری بعد از آن نمی‌باشد. به این ترتیب، روش آموزش وزنی ضمن ارائـه مزایـای روش گـره زدن، جایگزین مناسبی نیز برای روش MAP می‌باشد.

## ۷ـ پیاده سازی

سیستم بازشناسی موجود با استفاده از داده‌های *فارس دات* آموزش داده شـده و آزمـایش گردیده‌انـد . ایـن پایگـاه داده در شکل اصلی خود متشکل از ۶۰۰۰ ادای جمله است که از ۳۰۰ گوینده بدست آمده است . هر گوینده بطور تصـادفی ۲۰ جملـه از ۴۰۰ جمله موجود را خوانده است که ۲ جمله بین تمام گوینده‌ها مشترک است. با توجه به اینکه گوینده‌ها از نقـاط مختلـف کشور و با لهجه‌های مختلف انتخاب شده‌اند، جملات مربوط به گوینده‌های تـهرانی اسـتخراج گردیـده کـه پـس از بررسـی‌ها و

انتخاب نهایی حدود ۲۷۰۰ جمله بدست آمده است. این جملات به دو دسته ۱۸۰۰ و ۹۰۰ تایی تقسیم شده و از دسته اول برای آموزش و از دسته دوم برای تست استفاده گردید [۱۳].

پایگاه داده فارس دات حدود ۱۲۰۰ کلمه مختلف را شامل میشود و در آن حدود ۲۴۰۰ سه آوائی مجزا وجود دارد. با توجه به اینکه تعداد آواهای پایه در زبان فارسی، در این تحقیق، ۳۰ در نظر گرفته شده است، با یک محاسبه ساده، تعداد سه آوائی‌ها باید به ۲۷۰۰۰ بالغ گردد. در این شرایط، تنها با در نظر گرفتن گروه‌های همخوان و واکه و نیز ساختار هجائی زبان فارسی، از ۸ ترکیب ممکن سه گانه واکه و همخوان، تنها ۴ ترکیب CCV، CVC، VCC وVCV در زبان فارسی ممکن است. به این ترتیب، با توجه به تعداد ۷ واکه و ۲۳ همخوان موجود، تعداد سه آوائی‌های حاصل به ۱۲۲۳۶ کاهش خواهد یافت. البته در عمل، بسیاری از این ترکیب‌های سه آوائی ممکن نمی‌باشند [۱۴] و لذا تعداد واقعی سه آوائی‌ها در زبان فارسی به کسری از عدد فوق محدود می‌گردد. با این حال، و با توجه به اینکه مطالعه زیادی در زبان فارسی بر روی سه آوائی‌ها صورت نگرفته، آمار دقیقی در این خصوص در منابع مختلف یافت نشده و احتمال می‌رود که تعداد واقعی سه آوائی‌ها در زبان فارسی به بیش از دو برابر تعداد موجود در دادگان فارس دات هم برسد.

محدود بودن حجم دادگان فارس دات و پراکنده بودن سه آوائی‌ها در آن (از نظر تعداد دفعات وقوع) به خوبی نشان‌دهنده این واقعیت است که آموزش مستقیم آنها مواجه با مشکل داده پراکنده بوده و نتیجه مطلوبی نخواهد داشت. به عنوان مثال، از میان بیش از ۲۴۰۰ سه آوائی بدست آمده از فارس دات، ۲۳۶ سه آوائی کمتر از سه نمونه آموزشی و ۸۱۱ سه آوائی کمتر از پنج نمونه آموزشی داشته‌اند. آمار کاملتری از تعداد دفعات وقوع سه آوائی ها در متن آموزشی در [۱۵] ارائه گردیده است. همین امر به خوبی بستر لازم را برای پیاده سازی آلگوریتمی که در شرایط داده پراکنده بتواند با استفاده از اطلاعات سایر مدل‌ها اقدام به آموزش مقاوم مدل‌های با داده آموزشی کم نماید فراهم می کند. بنابراین، آلگوریتم آموزش وزنی، آلگوریتمی کاملاً مناسب برای آموزش سه آوائی ها در شرایط حاضر می‌باشد.

هدف از آموزش مدلها استفاده از آنها برای بازشناسی می باشد. لذا طراحی و پیاده سازی یک الگوریتم بازشناسی قوی و در عین حال سریع اهمیت فراوانی دارد. روشی که جهت بازشناسی بکار گرفته شده است موسوم به الگوریتم انتقال نشانه[۲۲] می‌باشد [۷]. مدلهای زبان مورد استفاده برای آزمایش سیستم طراحی شده، زوج کلمه[۲۳] و بدون گرامر[۲۴] می‌باشند.

## ۸ ـ نتایج

در پیاده‌سازی از یک توزیع گوسین برای هر حالت استفاده شده است. بدیهی است که استفاده از مخلوط گوسین موجب بهبود نتیجهٔ بازشناسی خواهد شد. علت عدم استفاده از آن در اینجا به اشباع رسیدن سیستم بازشناسی بویژه با مدل زبان زوج کلمه می‌باشد. برای آموزش وزنی، سه‌آوائی‌های با هسته مشترک را شبیه به هم گرفته‌ایم و ضریب $\varphi$ برابر ۰/۰۰۵ قرار داده شده است. همچنین در بدست آوردن مدل سه آوائی گره زده نشده، برای جلوگیری از آموزش نامناسب، مدلهائی که کمتر از سه نمونه آموزشی داشته‌اند دست نخورده باقی مانده‌اند. مدل های سه آوائی گره زده شده بر اساس آلگوریتم بحث شده در بخش ۴-۱ بدست آمده‌اند.

جدول ۲ نشان دهندهٔ نتایج بدست آمده از آزمایش‌های بازشناسی گفتار می باشد. نتایج اعلام شده برای آزمایش‌های بدون گرامر با استفاده از حدود ۱۲۰ جملهٔ آزمایشی بدست آمده‌اند و مبین دقت بازشناخت کلمه[۲۵] میباشند. این جملات از ابتدای مجموعهٔ آزمایشی و به ترتیب انتخاب و مورد استفاده قرار گرفته‌اند. علت استفاده از تنها ۱۲۰ جمله برای آزمایش در این مرحله، بکارگرفته نشدن مدل زبان می‌باشد که باعث طولانی شدن فرایند آزمایش (زمان بازشناسی) می‌گردد. بنابراین به کار گرفتن ۹۰۰ جمله در آزمایش‌ها مشکلات عملی به همراه دارد. بررسی‌های انجام شده نشان دهنده این واقعیت می‌باشند که تفاوت چندانی بین نتایج بدست آمده از ۱۲۰ و ۹۰۰ جمله در این حالت وجود ندارد و نتایج بدست آمده کاملاً قابل اعتماد می‌باشند.

**جدول (۲) نتایج بازشناسی**

| مدل | تعداد حالت‌ها | مدل زبان | دقت بازشناسی |
|---|---|---|---|
| تک آوائی | ۹۶ | بدون گرامر | ٪۴۶/۲ |
| سه آوائی بدون گره زدن | ۷۲۷۸ | | ٪۶۰/۱ |
| سه آوائی گره خورده | ۱۴۰۰ | | ٪۶۷/۸ |
| سه آوائی با آموزش وزنی | ۷۲۷۸ | | ٪۷۲/۹ |
| تک آوائی | ۹۶ | زوج کلمه | ٪۹۳/۱ |
| سه آوائی بدون گره زدن | ۷۲۷۸ | | ٪۷۲/۰ |
| سه آوائی گره خورده | ۱۴۰۰ | | ٪۹۴/۱ |
| سه آوائی با آموزش وزنی | ۷۲۷۸ | | ٪۹۶/۵ |

نتایج بخش بدون گرامر بخوبی حاکی از برتری شیوهٔ آموزش پیشنهادی بر روش معمول آموزش یعنی درستنمائی بیشینه است. در حالیکه مدل‌های سه‌آوائی، پیش از گره زدن، بوضوح نسبت به مدل‌های تک‌آوائی بهتر عمل نموده‌اند، گره‌زدن حالت در آنها موجب بهبودی بیش از پیش نتایج گردیده است. با این همه، شیوهٔ پیشنهادی آموزش توانسته در این شرایط حتی به نتایج برتری نیز دست یابد. اعمال مدل زبان زوج کلمه، طبیعتاً موجب بهبودی در نتایج بدست آمده از کلیهٔ آزمایش‌های فوق می‌گردد. با این‌همه، این بهبودی در سیستم مبتنی بر مدل‌های تک‌آوائی بیش از حد انتظار است. علت این امر پایین بودن ضریب "سرگشتگی"[۲۶] در دادگان تصور می‌شود. این مطلب می‌تواند دقت بازشناسی را بشدت افزایش دهد. با این‌همه، در استفاده از مدل‌های سه‌آوائی گره نخورده، با توجه به تعداد زیاد مدل‌ها و اینکه بسیاری از آنها بخوبی آموزش ندیده‌اند، احتمال خطا در بازشناخت بعضی کلمات زیادتر است و محدودیت انتخاب کلمات بعدی در بسیاری موارد از اصلاح نتیجه جلوگیری می‌نماید. با این حال، در این شرایط هم روش پیشنهادی آموزش به بهترین نتایج دست یافته است.

ذکر یک نکته نیز در اینجا لازم است و آن اینکه ستون تعداد حالت ها در جدول ۲ نشان‌دهنده کل تعداد حالت‌های موجود در سیستم است. به عنوان مثال، برای مدلهای تک آوائی، ۳۲ مدل سه حالته (۹۶ حالت) و برای سه آوائی‌های بدون گره زدن، ۲۴۲۶ مدل سه حالته (۷۲۷۸ حالت) وجود دارند. همین تعداد مدل و حالت در سه آوائی‌های با آموزش وزنی نیز دیده می شود. در مورد سه آوائی‌های گره خورده، با توجه به فرایند گره زدن حالت‌ها، ۷۲۷۸ حالت منطقی به ۱۴۰۰ حالت فیزیکی کاهش یافته است. لازم به اشاره است که تعداد حالت‌های فیزیکی بدست آمده کاملاً تابع سطح آستانه مورد استفاده در آلگوریتم خوشه‌بندی بخش ۴-۱-۱ است.

**جدول (۳) نتایج بازشناسی به ازاء ضرایب مختلف تضعیف در آموزش وزنی.**

| $\varphi$ | ۱ | ۰/۱ | ۰/۰۵ | ۰/۰۲ | ۰/۰۱ | ۰/۰۰۵ | ۰/۰۰۱ |
|---|---|---|---|---|---|---|---|
| NG | ٪۴۶/۲۴ | ٪۶۵/۲۶ | ٪۶۸/۱۹ | ٪۷۱/۱۱ | ٪۷۱/۶۶ | ٪۷۲/۹۴ | ٪۶۸/۳۷ |
| WP | ٪۹۴/۱۶ | ٪۹۶/۷۳ | ٪۹۶/۹۳ | ٪۹۷/۴۵ | ٪۹۷/۳۲ | ٪۹۶/۵۵ | ٪۹۱/۳۱ |

جدول (۳) نتایج بازشناسی را به ازاء ضرایب تضعیف مختلف نشان می‌دهد. همانگونه که دیده می‌شود با کوچک شدن ضریب تضعیف ابتدا دقت سیستم افزایش می‌یابد، اما بعد همانطور که انتظار داریم رو به کاهش می‌نهد. در هر دو حالت بدون گرامر و استفاده از گرامر زوج کلمه با این وضعیت مواجه هستیم ولی میزان پارامتر $\varphi$ برای دستیابی به نرخ بازشناسی بیشینه در این دو حالت متفاوت است.

## ۹ ـ نتیجه گیری و پیشنهادات

بررسی‌های انجام شده نشان دهندهٔ این واقعیتند که استفاده از مدلهای وابسته به متن نتایج بهتری از مدلهای نابسته به متن دارند. با این همه در مدلسازی وابسته به متن تعداد پارامترها می‌تواند بشدت افزایش یابد. بنابراین، برای رسیدن به نتیجه بهتر باید تمهیداتی برای مقاوم کردن آموزش اندیشید.

روش آموزش وزنی پیشنهادی، که بر اساس استفاده از پارامترهای مدل‌های هم خانواده عمل می‌نماید، هم در حالت بدون گرامر و هم در شرایط اعمال مدل زبان (در اینجا مدل زوج کلمه) به نتایج بهتری نسبت به روش معمول آموزش دست یافته است. این نتایج در مقایسه با دقت حاصل از یک سیستم با حالت های گره زده شده (روش معمول مورد استفاده امروزه در سیستم های بازشناسی گفتار در سطح جهان) نیز حکایت از برتری نسبی روش آموزش وزنی دارند. پیشنهادات زیر برای ادامه کار مفید به نظر می‌رسند:

الف ـ استفاده از توزیع گوسین با تعداد عناصر مخلوط بزرگتر.

ب ـ استفاده از پایگاه داده کاملتر.

ج ـ استفاده از مدلهای زبان پیچیده‌تر (مشروط به رعایت بند ب).

د ـ استفاده از مدلهای زیر لغوی وابسته به متن بین کلمه‌ای.

ه ـ تحقیق بیشتر در مورد اینکه چه مدلهایی را شبیه به هم در نظر بگیریم. برای این کار می‌توان همانند گره زدن عمل کرد، ولی در پایان از روابط (۱۸) تا (۲۰) برای آموزش استفاده کرد.

و ـ گره زدن پارامترها پس از آموزش وزنی و استفاده از این روش برای آموزش مجدد.

ز ـ تحقیق در مورد بدست آوردن پارامتر $\varphi$ بصورت تحلیلی. احتمالاً این کار همانند روش MAP قابل انجام است.

ح ـ استفاده از آموزش وزنی در کاربردهائی نظیر تطبیق گوینده.

## زیرنویس‌ها

1- Context Dependent Modeling
2- Context
3- Robust
4- Maximum a Posteriori
5- Bayes
6- Prior
7- Maximum Likelihood
8- Triphones
9- Monophones
10- Biphones
11- Word-internal triphones
12- Cross-word triphone modeling
13- Quinphones
14- Maximum Likelihood
15- Finite State Network
16- Tying
17- Clu16- stering
18- Bayesian
19- Maximum a Posteriori
20- Sparse Training Data
21- Wishart distribution
22- Dirichlet distribution
23- Token passing
24- Word-pair
25- No-gram
26- Word recognition accuracy
27- Perplexity

# مراجع


[1] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev and S.J. Young, " The Development of the 1994 HTK Large Vocabulary Speech Recognition System", In Proc. ARPA Spoken Language Technology Workshop, Bartoncreek, 1995.

[2] S.J. Young, "The general use of Tying in Phoneme-Based HMM Speech Recognisers", Proc. ICASSP-93, Vol. 1, pp. 569-572, San Francisco.

[3] S.J. Young, J.J. Odell and P.C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modelling", In ARPA Human Language Technology Workshop, pp. 307-312, Plainsboro, March 1994.

[4] J-L. Gauvain, and C-H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models", Speech Communication, vol. 11, 1992, pp.205-213.

[5] J. Ming, P. O'Boyle, M. Owens and F.J. Smith, "A Bayesian Approach for Building Triphone Models for Continuous Speech Recognition", IEEE Trans. Speech and Audio Proc., vol. 7, No. 6, Nov. 1999, pp. 678-684.

[6] J.J. Odell, "The Use of Context in Large Vocabulary Speech Recognition", PhD Thesis, Cambridge University Engineering Department, 1995.

[۷] م. پرویزی ، س. م. احدی، "بازشناسی گفتار فارسی پیوسته با دایره کلمات متوسط"، مجموعه مقا لات کنفرانس مهندسی برق ایران، ۱۳۸۰.

[8] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, New Jersey, Prentice-Hall, 1993.

[9] J-L. Gauvain and C-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, April 1994.

[10] S.M. Ahadi, "Reduced Context Sensitivity in Persian Speech Recognition via Syllable Modeling", In Proc. SST-2000, pp. 492-497, Canberra.

[11] HTK: Hidden Markov Model Toolkit Ver. 1.4, Reference Manual, Cambridge University Engineering Department, p. 76, 1992.

[12] Q. Huo and C. Chan, "Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition", Technical Report, Department of Computer Science, University of Hong Kong, September 1992.

[13] S.M.Ahadi, "Recognition of Continuous Persian Speech Using a Medium-Sized Vocabulary Speech Corpus", in Proc. EUROSPEECH'99, Budapest.

[۱۴] ی. ثمره، آواشناسی زبان فارسی، ویرایش دوم، مرکز نشر دانشگاهی، تهران ۱۳۸۰.

[۱۵] س. ح. شمس، "مدل سازی وابسته به متن در بازشناسی گفتار پیوسته فارسی"، پایان نامه کارشناسی ارشد، دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر، ۱۳۸۰.