مدلسازی وابسته به متن در بازشناسی گفتار پیوسته بر اساس درخت تصمیم‌گیری آوایی فارسی

سید حسین شمس[1]، سید محمد احدی[2]*

۱- فارغ التحصیل کارشناسی ارشد، دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر

۲- استادیار دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر

*** تهران، خیابان حافظ، کد پستی ۱۵۹۱٤**

sma@aut.ac.ir

(دریافت مقاله: اردیبهشت ۱۳۸۱، پذیرش مقاله: دی ۱۳۸۱)

چکیده – مدلسازی وابسته به متن[1] به عنوان شیوه‌ای مفید برای افزایش دقت مدلسازی در بازشناسی گفتار پیوسته مورد توجه است. معمول ترین شکل پیاده سازی این شیوه، استفاده از مدلهای سه آوایی است. با این همه، تعداد زیاد این مدلها موجب می‌شود که در عمل، آموزش سیستم با مشکلات زیادی همراه باشد و دستیابی به آموزش مقاوم[2] به سختی میسر گشته یا اصولاً مقدور نشود. یکی از شیوه‌های حل این مشکل، استفاده از روش گره زدن[3] پارامترها است. در این مقاله خوشه‌بندی[4] برای پارامترهای حالت مدل HMM صورت گرفته وحالتهای قرار گرفته در خوشه با یکدیگر گره زده شده‌اند تا در کل سیستم، تعداد پارامترها کاهش یافته و آموزش مقاوم حاصل شود. دو نوع دسته‌بندی یکی بر اساس پارامترهای مدلهای نهایی آموزش دیده و فاصله بین آنها و دیگری به کمک داده‌های آموزشی و بر مبنای درخت تصمیم گیری انجام شده است. در پیاده سازی روش اخیر به طراحی درخت تصمیم گیری بر اساس ویژگیهای آکوستیکی آواها در زبان فارسی و شباهتها و تفاوتهای آنها اقدام شده است. نتایج بدست آمده، موفقیت آمیز بودن هر دو روش را تأیید کرده است. با این حال، مزیت روش دوم در امکانپذیر ساختن تخمین پارامترهای مدلهای دیده نشده است.

کلید واژگان: مدلسازی وابسته به متن، بازشناسی گفتار پیوستهٔ فارسی، مدلهای مارکوف پنهان با چگالی پیوسته، گره زدن حالتها، درخت تصمیم گیری.

۱- مقدمه

استفاده از مدلهای مارکوف پنهان (HMM) در بازشناسی گفتار پیوسته با دایره بزرگ لغات، در سالهای اخیر به نتایج قابل قبولی منجر شده است. استفاده از مدلسازی بر اساس واحدهای واجی پایه، علی‌رغم دستیابی به نتایج قابل طرح در بازشناسی، در رسیدن به دقتهای بازشناسی بالا ـ حتی با

1. Context Dependent Modeling
2. Robust training
3. Tying
4. Clustering

استفاده از تعداد زیاد عناصر مخلوط در تابع چگالی احتمال خروجی ـ ناتوانی نشان داده است. یکی از شیوه‌های به کار رفته برای بهبود مدلسازی، به کار گرفتن مدلهای وابسته به متن بوده است [۱و۲].

در مدلسازی وابسته به متن، به‌جای استفاده از یک مدل به ازای هر واحد واجی پایه، با توجه به شرایط متنی که واحد پایه در آن واقع می‌شود، مدلهای مختلفی در نظر گرفته می‌شود. به این ترتیب، این امکان فراهم می‌شود که واجگونه‌های[1] هر واحد پایه بتوانند جداگانه مدل شده و در مجموع مدلسازی بسیار دقیق تری صورت گیرد. با این همه، این شیوهٔ مدلسازی موجب به‌وجود آمدن مشکلاتی نیز می‌گردد. مهمترین مشکل به‌وجود آمده، نبود داده آموزشی کافی برای آموزش مناسب تمامی واحدها، با توجه به تعداد بسیار زیاد آنها است. این مشکل، حتی برای دادگانهای گفتاری بسیار بزرگ نیز، با توجه به پراکندگی و احتمال پایین مشاهده بسیاری از شرایط متنی در داده آموزشی، ملاحظه می‌شود. بدیهی است هرچه شرایط متنی در نظر گرفته شده برای مدلسازی پیچیده تر باشد، شدت این مشکل نیز بیشتر می‌شود.

یکی از شیوه‌های پر استفاده در حل این مشکل گره زدن پارامترهای مدلها است [۳و٤]. گره زدن پارامترها می‌تواند موجب کاهش قابل توجه در پارامترهای سیستم و بنابراین بهبود آموزش شود. با این همه انتخاب مناسب پارامترها برای گره زدن و به‌کارگیری آلگوریتم گره زدن مناسب اهمیت زیادی در توفیق این شیوه دارد. پارامترهای مختلفی از سیستم در آلگوریتم گره زدن می‌توانند مورد استفاده قرار گیرد. در این میان گره زدن حالتها و عناصر مخلوط تابع چگالی احتمال مفیدتر واقع شده است.

در این تحقیق، برای نخستین بار از درخت تصمیم گیری که بر اساس قواعد آواشناسی زبان فارسی طراحی شده برای خوشه بندی پارامترها استفاده شده است. پارامترهای مورد استفاده در گره زدن، حالتهای HMM بوده اند. علاوه بر این، از شیوه خوشه بندی بر اساس داده[2] نیز برای خوشه بندی و گره زدن حالت‌ها در HMM استفاده شده و نتایج به‌دست آمده از دو روش در این مقاله مقایسه شده است.

مقاله حاضر شامل شش بخش می باشد که بخش اول مقدمه و بخش دوم به مسأله مدلهای زیر لغوی می‌پردازد. در بخش سوم نحوه آموزش و آموزش پذیری مطرح شده است. بخش چهارم به بررسی نحوه گره زدن پارامترها و بخش پنجم به پیاده سازی و بررسی نتایج بازشناسی اختصاص یافته و بخش ششم نیز نتیجه گیری و مسیر تحقیقات آینده را نشان می‌دهد.

۲- مدلسازی وابسته به متن؛ رویکردی در مدلسازی زیرلغوی

انتخاب واحدهای پایه و مدلسازی آنها نخستین گام در طراحی سیستم بازشناسی پیوسته است. اگر آموزش به‌طور کامل انجام شده 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 شد و سپس به مدل وابسته به متن سه آوایی[3] تعمیم یافتند.

منظور از مدلسازی وابسته به متن [۵] این است که برای هر واحد زیر لغوی با توجه به واحدهای زیر لغوی قبل یا بعد(شرایط متنی) مدل جداگانه ای در نظر گرفته شود. در زیر، گونه‌هایی از واحدهای پایه استفاده شده اعم بر نابسته یا وابسته به متن معرفی می‌شود.

1. Allophones

2. Data driven clustering

3. Triphones

الف ـ مدلهای تک آوایی[1]

فقط یک مدل از هر واج را در تمام محلهای وقوع مدلسازی می‌کند. جدول ۱ واجهای اصلی زبان فارسی و نمادهای سمبلیک آنها را که در این کاربرد به‌کار گرفته شده‌اند نشان می‌دهد.

جدول ۱ نمایش سمبلیک واحدهای استفاده شده بر اساس واجهای زبان فارسی با استفاده از کاراکترهایASCII .

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	s@lam
	v	v@hid
	x	xord
	z	ziba
	#	#ale
	$	$ire
	h	$@hid
Affricates	j	Javid
	c	c@medan
	+	فاصله بین کلمات
	_	سکوت

مثال زیر به نحوه نمایش یک جمله (در باز است) بر اساس نمادهای سمبلیک اشاره شده در بالا می‌پردازد.

d@r baz ‘@st

_ d @ r + b a z + ‘ @ s t _

ب ـ مدلهای دو آوایی[2]

هر واج با توجه به واج سمت راست یا سمت چپ به‌طور جداگانه مدل می‌شود. برای سکوت و فاصلهٔ بین کلمات، هریک فقط یک مدل را (بدون توجه به متن) در نظر می‌گیرند. برای نمایش ترکیبهای متنی راست و چپ، در این تحقیق، به‌ترتیب از علائم ">" و "<" استفاده می‌کنیم. به این ترتیب جمله بالا به‌صورت زیر نمایش داده می‌شود .

مدل سازی دو آوایی چپ[3]:

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

مدل سازی دو آوایی راست[4]:

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

در مدلسازی دوآوایی چپ، اولین واج (سمت چپ) و در مدلسازی دوآوایی راست، آخرین واج (سمت راست) به‌صورت تک آوایی مدل شده اند.

ج ـ مدل‌های سه آوایی

هر واج با توجه به واج سمت راست و سمت چپ به‌طور جداگانه مدل می‌شود . مانند قبل برای سکوت و فاصله، بدون توجه به متن، هریک فقط یک مدل در نظر می‌گیریم و در کنارهای کلمه از مدلسازی دو آوایی مناسب استفاده می‌کنیم.

به این ترتیب جمله بالا به‌صورت 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

1. Monophones

2. Biphones

3. Left biphones

4. Right biphones

آوایی داخل کلمه ای[1] نامیده می‌شود. شیوهٔ پیچیده تری به نام مدلسازی سه آوایی بین کلمه‌ای[2] نیز وجود دارد که در مرزهای کلمات از مدل دوآوایی استفاده نمی‌کند بلکه همراه با فاصله، سکوت یا آوای ابتدایی کلمهٔ، یک مدل سه آوایی را به‌کار می‌برد. این شیوه موجب پیچیدگی فراوان مدلسازی و بویژه در پیاده سازی الگوریتم بازشناسی مبتنی بر کلمه می‌شود.

د ـ مدلهای تعمیم یافته

این فرایند (افزایش تعداد واجهای کناری که در مدلسازی در نظر گرفته می‌شوند) می‌تواند به تعداد آواهای بیشتری از هر طرف نیز تعمیم داده شود. با این همه در عمل، با توجه به افزایش بیش از حد تعداد واحدها که مشکلات زیادی می‌تواند ایجاد کند، مدلسازی به بیش از شرایط پنج آوایی[3] (تا دو آوا از هر طرف) گسترش پیدا نمی‌کند.

۳- آموزش و آموزش پذیری

اولین گام در ایجاد سیستم بازشناسی گفتار، آموزش است. در این رویکرد، در مرحله آموزش ابتدا یک مجموعه مدل اولیه مناسب براساس واجها ایجاد کردیم(مدل تک آوایی ایجاد شده در [٦]). سپس مدل هر واج را به عنوان مدل اولیه‌ای برای سه آوایی‌هایی که هسته آنها همان واج است در نظر گرفتیم. این کار را ارث بری می‌نامیم . سپس مدلهای سه آوایی را آموزش دادیم. بدیهی است که تعداد مدلهای به‌دست آمده از این مرحله به مراتب بیشتر از تعداد مدلهای اولیه است.

۳-۱- آموزش براساس الگوریتم Baum-Welch و تخمین پارامترها

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

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

در این صورت روابط بازتخمین برای پارامترهای HMM و بر اساس الگوریتم Baum-Welch به این صورت بیان می‌شوند:

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

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

$$\Sigma_j^{(q)} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha^{(q)r}(t,j)*\beta^{(q)r}(t,j)*(o_j^r-\mu_j^{(q)})^2}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha^{(q)r}(t,j)*\beta^{(q)r}(t,j)} \qquad (٤)$$

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

برای استفاده از این فرمولها در جملات بدون برچسب زمانی ابتدا مدل جمله را از کنار هم قرار دادن تک تک مدلهای زیر لغوی به‌دست می‌آوریم. به این مدل FSN[4] می‌گویند. سپس مقادیر جدید پارامترها را تخمین می زنیم [۱]. این الگوریتم باید به‌صورت مکرر بر روی داده‌های آموزشی اجرا شود تا همگرایی نسبی ایجاد شده و مقادیر مناسبی برای پارامترها به‌دست آید. با توجه به تعداد بالای پارامترهای مدل در این حالت، در بسیاری موارد، تعداد

1. Word-internal triphone
2. Cross-word triphone
3. Quinphone
4. Finite State Network

دفعات دیدن مدل سه آوایی در متن آموزشی، صفر یا بسیار اندک است. لذا برای جلوگیری از آموزش نامناسب، ترتیبی داده شد که پارامترهای مدلهایی که کمتر از سه بار در داده آموزشی دیده شده‌اند، تغییر نیابند.

۳-۲- روشهای بهبود آموزش پذیری

برای افزایش دقت سیستم بازشناسی لازم است افزایش پارامترها به‌گونه‌ای باشد که امکان آموزش مناسب فراهم باشد. قابل قبول نخواهد بود که پارامتری فقط با معدودی نمونه آموزش داده شود. این مسأله در آموزش واریانس، خود را بهتر نشان می‌دهد. در عمل بین ده تا چند صد نمونه برای آموزش مقاوم میانگین و واریانس لازم است.

چندین روش برای غلبه بر کمبود داده آموزشی معرفی شده است که در اینجا به چند مورد از آنها اشاره می‌شود.

۳-۲-۱- برگشت به عقب[1]

هنگامی که اطلاعات کافی برای آموزش مدلی وجود ندارد، می‌توان با برگشت به عقب و استفاده از مدلی با وابستگی کمتر به متن ـ که دادهٔ کافی برای آن وجود دارد ـ آموزش را انجام داد. به‌عنوان مثال می‌توان از دوآوایی برای آموزش سه آوایی واز تک آوایی برای آموزش دو آوایی استفاده کرد. این کار تضمین می‌کند که برای آموزش هر مدل، دادهٔ کافی وجود خواهد داشت. در عین حال، وابستگی به متن مدلها کاهش می‌یابد [۵].

۳-۲-۲- هموارسازی[2]

در این روش پارامتر مدلهای با پیچیدگی بیشتر با توجه به مدلهای با پیچیدگی کمتر هموار می‌شود. یک راه برای انجام این کار، درونیابی بین مدل با وابستگی بیشتر و مدلهای با وابستگی کمتر است. این کار موجب می‌شود که وابستگی به متن در مدل رعایت شود در عین حال که سیستم مقاوم می‌شود.

۳-۲-۳- اشتراک گذاری[3]

روش دیگری برای افزایش مقاومت سیستم به کمبود داده آموزشی، به اشتراک گذاشتن مدل یا قسمتی از مدل در بین چند مدل مختلف است. این کار معقولی است زیرا ویژگیهای یک واج که در دو متن مختلف قرار دارد به همدیگر نزدیک است. این روش همچنین تضمین می‌کند که پارامترهای سیستم بخوبی آموزش می‌بینند، در عین حال که وابستگی مدلها به متن رعایت می‌شود.

در کلیه روشهای بالا لازم است تصمیم گرفته شود که کدام مدل و به چه مدلی برگشت کند یا با توجه به چه مدلهایی هموار شود یا کدام پارامترها با هم گره زده شوند.

در روش برگشت به عقب یک قاعده ساده وجود دارد و آن اینکه وابستگی مدلهای تک آوایی به متن نسبت به دو آوایی‌ها کمتراست. در عین حال این مدلها قابلیت آموزش بهتری نیز دارند. همچنین مدلهای دو آوایی وابستگی به متن کمتری نسبت به سه آوایی‌ها داشته و آموزش پذیرترند. اما ضعف این روش در این است که وابستگی به متن سیستم بسرعت کاهش می‌یابد که منجر به دقت کمتر می‌شود.

به‌طور مشابه قواعدی برای چگونگی هموار کردن پارامترها وجود دارد. به‌عنوان مثال می‌توان حالت اول سه آوایی‌ها را با استفاده از حالتهای اول دو آوایی‌ها هموار کرد و حالت آخر آنها را نیز با کمک حالتهای آخر دو آوایی‌ها هموار کرد.

اشتراک گذاری، آزادی عمل بیشتری نسبت به دو روش قبل دارد ودر عین آموزش مناسب، وابستگی به متن را بهتر رعایت می‌کند. به همین دلیل یکی از روشهای پراستفاده در سیستمهای بازشناسی امروزی، روش اشتراک گذاری یا به بیان دیگر گره زدن[4] است.

1. Backing off
2. Smoothing
3. Sharing
4. Tying

٤- روشهای موجود درگره زدن پارامترها و روش پیشنهادی

برای تصمیم گیری در مورد چگونگی گره زدن از دو روش به‌طور عمده استفاده می‌شود [٥]:

٤-١- پایین به بالا[1]

در این روش در ابتدا برای هر پارامتر ـ یا دسته ای از آنها که امکان گره خوردن با یکدیگر دارند ـ یک دسته جداگانه در نظر می‌گیریم. سپس فرایند ادغام دسته‌ها برای رسیدن به دسته‌های کمتر اما آموزش پذیرتر پیاده می‌شود.

٤-٢- بالا به پایین[2]

در این شیوه، ابتدا فرض می‌کنیم تمام پارامترهای مورد نظر در یک گروه قرار دارند. سپس مراحل جداسازی برای رسیدن به دسته‌های نهایی انجام می‌شود.

هر یک از این روشها می‌توانند بر اساس داده یا بر اساس ویژگیهای آکوستیکی عمل کنند. با این حال، معمولاً از روش پایین به بالا در گره زدن بر اساس داده استفاده می‌شود، در حالی که روش بالا به پایین در دسته بندی بر اساس ویژگیهای آکوستیکی مورد استفاده قرار می‌گیرد.

شکل ١ پارامترهای مختلفی را که در سیستم مبتنی بر HMMهای با چگالی مشاهدات پیوسته قابلیت گره زده شدن دارند نشان می‌دهد [٧]. بالاترین سطح برای گره زدن، سطح مدل است. مدلهای سه آوایی ایجاد شده بر اساس این روش در اصطلاح، سه آوایی‌های عمومیت یافته[3] نامیده می‌شوند. این روش به نتایج بهتری در مقایسه با هموار کردن می‌رسد.

گره زدن در سطوح پایین تر، از جمله حالتها، چگالی‌های احتمال و پارامترهای چگالی‌های احتمال نیز ممکن است. در این میان، حالتها گزینه خوبی برای گره زدن هستند زیرا در عین حال که جزئی از هر مدل هستند، می‌توانند مشخص کننده جزء زمانی هر آوا نیز باشند [٨]. در عمل، گره زدن حالتها نسبت به گره زدن مدلها به نتایج بهتری منجر شده است [٥].

شیوه‌های به‌کار گرفته شده در این تحقیق نیز مبتنی بر گره زدن حالتها است.

٤-٣- گره زدن بر اساس روش پایین به بالا

همانگونه که اشاره شد، در این روش ابتدا فرض می‌شود مدلها متفاوت هستند. سپس برای رسیدن به آموزش مقاوم، از گره زدن استفاده می‌شود. این کار با بررسی مدلهای اولیه و دسته بندی مدلها (یا اجزای آنها) انجام می‌شود. باید ترتیبی اتخاذ شود تا اعضای هر دسته علاوه برآموزش پذیری، شباهت کافی داشته باشند تا به مدل دقیقی برسیم. در عمل دو شرط بررسی می‌شود؛ یکی اینکه اعضای هر دسته تاحد امکان به یکدیگر شبیه باشند و دیگر اینکه تک تک دسته‌ها به تعداد کافی نمونهٔ آموزشی داشته باشند.

دسته‌بندی در دو مرحله انجام می‌شود:

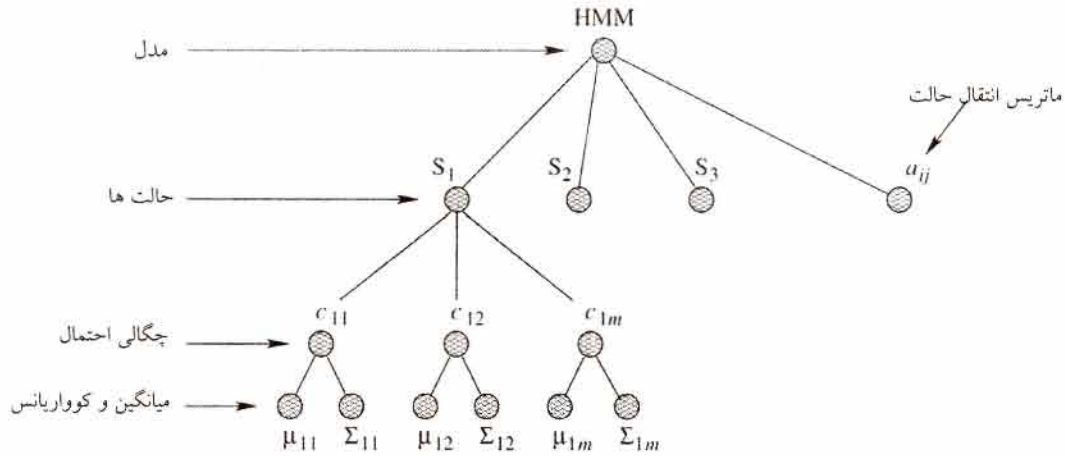
الف ـ مرحله اول شامل یک روند تکرار شونده است و در آن دسته‌های با شباهت زیاد یکی می‌شوند. برای این کار بر اساس معیاری فاصله بین دسته‌ها محاسبه می‌شود و در صورت کمتر بودن از سطح آستانه از پیش تعیین شده، دو دسته یکی می‌شوند. این مرحله هنگامی که کمترین فاصلهٔ بین دسته‌ها از مقدار از پیش تعیین شده بیشتر شود، پایان می‌یابد.

ب ـ دسته‌هایی که تعداد نمونهٔ آموزشی آنها از مقدار تعیین شده ای کمتر باشد با نزدیکترین دسته ادغام می‌شوند.

الگوریتم خوشه بندی زیر این مراحل را بخوبی نشان می‌دهد [٩].

١- برای هر حالت یک دسته جدا را در نظر می‌گیریم.

1. Bottom up approach
2. Top down approach
3. Generalized triphones



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

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

۳- اگر این فاصله از مقدار از پیش تعیین شده ای (δ) کمتر بود این دو دسته را ادغام می‌کنیم و به قسمت ۲ برمی‌گردیم.

٤- دسته‌ای را که کمترین نمونهٔ آموزشی را دارد همراه با دسته‌ای که کمترین فاصله را با آن دارد مشخص می‌کنیم.

۵- اگر تعداد نمونه‌های آموزشی دستهٔ اول از مقدار مشخصی کمتر بود دو دسته را یکی می‌کنیم و به مرحلهٔ ٤ می‌رویم.

٦- پایان.

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

برای یافتن فاصله دسته‌هایی که دارای چندین عضو هستند، فاصله تمامی اعضای هر دسته را با تمامی عضوهای دسته دیگر محاسبه می‌کنیم وسپس از آنها میانگین می‌گیریم. در این آلگوریتم از رابطه دیورژانس به شرح زیر، برای به‌دست آوردن فاصله بین دسته‌ها استفاده کرده‌ایم [۱۰].

$$d(S_1,S_2)=\left(\frac{1}{N}\sum_{i=1}^{N}\frac{(\mu_{1i}-\mu_{2i})^2}{\sigma_{1i}\sigma_{2i}}\right)^{\frac{1}{2}} \qquad (۵)$$

که N بعد ضرایب مربوط به هر فریم است و μ_{ji} و σ_{ji} بترتیب عناصرام میانگین و واریانس توزیع‌های احتمال دو حالت مورد نظر است.

گره زدن حالتها موجب کاهش چشمگیر تعداد پارامترها می‌شود، در عین حال که تعداد مدلهای متفاوت (از نظر فیزیکی) کاهش چشمگیری نمی‌یابد، زیرا لزومی ندارد که تمامی حالتهای مدل به‌گونه‌ای یکسان گره بخورند.

٤-٤- گره زدن بر اساس روش بالا به پایین

گره زدن بر اساس روش پایین به بالا ـ که بر اساس پارامترهای مدلهای آموزش دیده موجود عمل می‌کند ـ دارای این محدودیت است که به وجود نمونه‌هایی از هر مدل در دادهٔ آموزشی احتیاج دارد. بنابراین قابلیت تخمین پارامترهای مدلهایی را که ممکن است در حین بازشناسی

رخ دهـند اما در دادهٔ آموزشی دیده نشده‌اند ندارد. همچنین بـرای مدلهایـی که به تعداد بسیار کم دیده شده اند، مکانیزم گـره زدن قـابل اعـتماد نخواهـد بـود. ایـن مشکل با داشتن پایگـاه داده ای که به تعداد لازم دادهٔ آموزشی برای هر مدل داشـته باشـد رفع می‌شود، اما این امر فقط برای دادگان‌های بـا دایـرهٔ لغـات نسـبتاً کوچـک مقـدور است. در ضمن در شـرایطی کـه بخواهـیم از مدلهای بین کلمه‌ای استفاده کنیم، این مشکل بسیار حاد می‌شود.

روش اشـاره شـده قبلـی، در ایـن کاربرد، اصطلاحاً *گره زدن بر اساس داده*[1] نیز نامیده می‌شود. این شیوه پیش از این نـیز در بازشناسی گفتار فارسی مورد استفاده قرار گرفته [٦] و در ایـنجا فقـط به عنوان مقایسه، نتایج بازشناسی به کمک آن ارائـه خواهـد شـد. ایـن شیوه، برای آنکه بتواند بدرستی پـیاده‌سـازی شـود، نیاز دارد که در ابتدا مدلها به شکل نسبتاً خوبـی آمـوزش دیده باشند تا خوشه بندی پارامترهای مدل بـه حقیقـت نزدیکـتر باشد. این امر با این واقعیت که اصولاً گـره زدن بـرای رفـع مشکل کمبود داده‌های آموزشی مدلها صورت می‌گیرد در تناقض است. به این ترتیب، روش فوق، بویـژه در خوشـه‌بـندی پارامترهایـی که دارای دادهٔ آموزشی محدودنـد، بـه‌صـورت مطلوبـی عمـل نمـی‌کـند. شـرایط هنگامـی‌بدتـر خواهـد بـود که در هنگام آزمایش با مدلهایی مواجـه شویم که در میان داده‌های آموزشی موجود نبوده‌اند. ایـن قبـیل مدلهـای سه آوایی، اصطلاحاً *سه آوایی‌های دیده نشده*[2] نیز نامیده می‌شوند.

اسـتفاده از روش بالا به پایین بر اساس درخت تصمیم گـیری[3] و بـر مبنای پرسشهای آواشناسی به ما اجازه می‌دهد کـه در عیـن دسـته‌بـندی مطلـوب، مدلهایـی را کـه در دادهٔ آموزشـی دیـده نشـده انـد نـیز در گروه مناسبی قرار دهیم. روش بـه‌کـار رفـته در ایـن مقالـه نیز بر اساس همین شیوه طراحی شده و پس از ایجاد یک درخت تصمیم‌گیری آوایی مبتنـی بـر ویژگیهای زبان فارسی، برای خوشه‌بندی آواها از آن استفاده می‌کند.

بـه ایـن مـنظور، بـرای هـر واج و بـرای هر حالت مدل واجگونـه مـربوط، یـک درخـت ایجاد می‌شود. این درخت تعییـن مـی‌کـند کـه پارامترهای هر حالت با کمک حالتهای کـدام مدلهـا سـاخته می‌شوند. برای نمایش بهتر، درخت را بـه‌صـورت وارونـه (ریشه درخت در بالا) نمایش می‌دهیم. شـکل ۲ درخـت تصـمیم گـیری نمونـه پیشنهادی را نشان می‌دهد.

بـرای ساخت هر مدل از ریشه شروع می‌کنیم و با توجه بـه هـر پرسش به گره مربوط می‌رویم تا به گرهی برسیم که دیگـر تقسـیم انجـام نشـود. ایـن گـره مشخص کننده نقطه اشتراک حالت مورد نظر از مدل حاضر، با حالتهای متناظر از سایر مدلها است.

ایـن روش چندیـن مزیـت نسبت به روش پایین به بالا دارد:

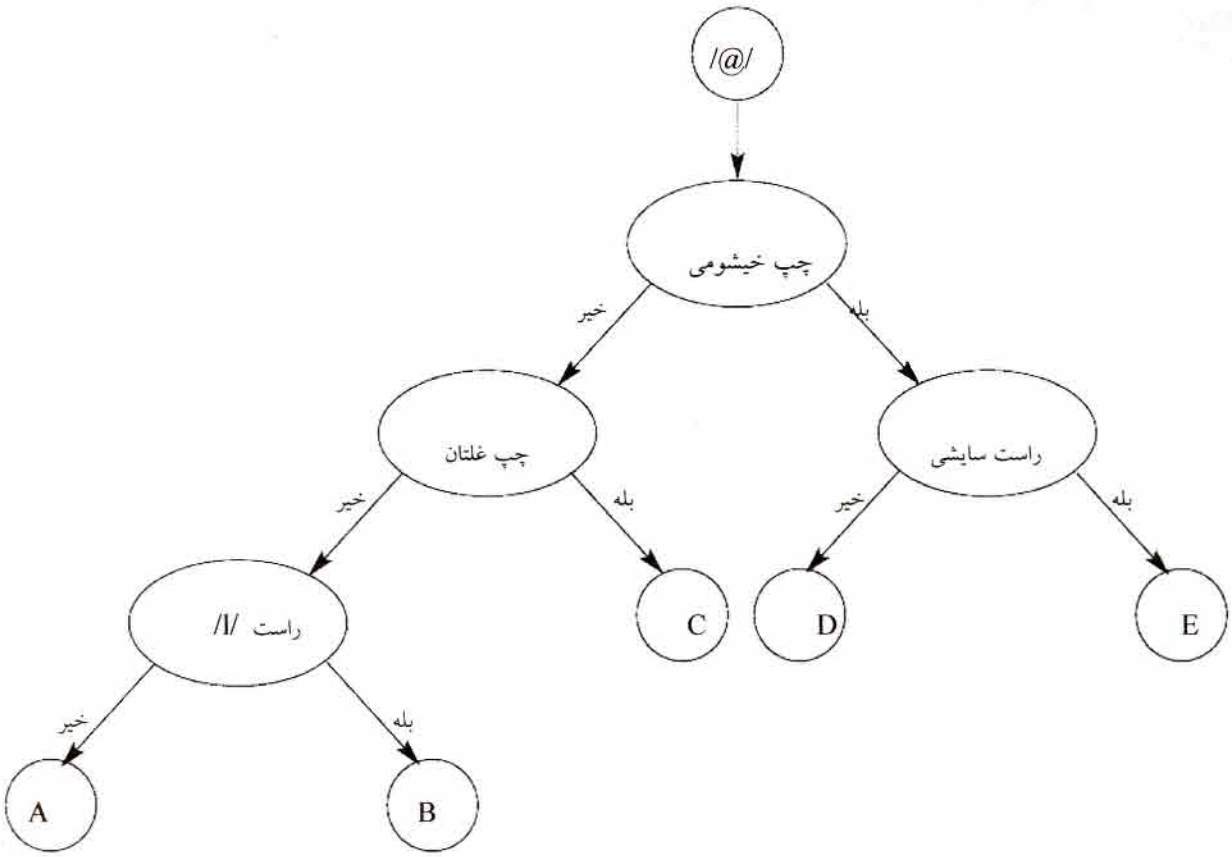
الـف ـ شـکل سلسـله مراتبی درخت تضمین می‌کند که بـرای هـر مدل امکان تخمین پارامترهای آن وجود دارد، در صورتی که ممکن است دادهٔ آموزشی هم نداشته باشد.

ب ـ اطلاعـات خارجـی آواشناسـی می‌تواند در ساخت درخت استفاده شود.

ج ـ الگوریتم تقسیم می‌تواند به‌گونه‌ای اجرا شود که در هـر گـره دادهٔ آموزشـی کافـی داشته باشیم. این کار ما را از اعمـال جداگانـه شروط اضافی که ممکن است منجر به گره خوردن نامناسب شود، بی نیاز می‌کند.

د ـ وابسـتگی‌هـای بیشـتر را مـی‌تـوان بـه‌راحتـی و بـا گسـترش پرسشـها در مدلسـازی اعمـال کـرد. به‌عنوان مثال

1. Data driven tying
2. Unseen triphones
3. Decision tree



شکل ۲ بخش نمونه ای از یک درخت تصمیم گیری.

می‌توان در دسته بندی، پرسشهای مربوط به جنس گوینده، سن، لهجه و ... اعمال کرد. ساختار درخت تضمین می‌کند که در صورت وجود دادهٔ کافی و تأثیر قابل توجه این عوامل، جداسازی مناسب صورت گیرد.

درخت بر اساس یک روش بهینه محلی ایجاد می‌شود. ابتدا از ریشه شروع می‌کنیم و گره ریشه را می‌سازیم و کلیهٔ حالتهای اولیه را در آن قرار می‌دهیم. بعد از اینکه هر گره ایجاد شد، پرسش بهینه به‌صورتی انتخاب می‌شود که در دسته‌های ایجاد شده به حداکثر افزایش درستنمایی برسیم و در ضمن در گره‌های ایجاد شده دادهٔ کافی برای آموزش موجود باشد. هنگامی‌که هیچ گرهی امکان تقسیم شدن نداشته باشد کار ما به پایان می‌رسد.

با توجه به روند اشاره شده در بالا، باید معیاری را برای محاسبه شباهت بین حالتها در دست داشته باشیم[۱۱و۱۲]. محاسبهٔ کامل درستنمایی بیشینه (ML)[1] بسیار وقت گیر است و پیاده سازی آن ممکن نیست، اما داشتن تخمینی ازدرستنمایی لگاریتمی[2] برای داده‌های آموزشی با توجه به فرضهای زیر امکان پذیر است:

الف ـ تناظر بین مشاهدات و حالتها در طی مراحل دسته بندی تغییر نمی‌کند.

ب ـ درستنمایی را می‌توان با میانگین گیری ساده از درستنمایی لگاریتمی مشاهدات ـ که با احتمال نظیر بودن مشاهده با حالت مورد نظر وزن دهی شده ـ محاسبه کرد.

1. Maximum Likelihood
2. Log likelihood

بنابراین:

$$L = \sum_{f\in F}\sum_{s\in S} \ln(\Pr(o_f, \mu_s, \Sigma_s))\gamma_s(o_f) \approx \ln(\Pr(O,S)) \quad (٦)$$

که S مشخص کنندهٔ مجموعه حالتهایی است که در یک دسته قرار گرفته اند و F کل فریمهای دادهٔ آموزشی است. با ساده سازی عبارت فوق به عبارت زیر می‌رسیم [۵].

$$L = -\frac{1}{2}\left(\ln\left[(2\pi)^n|\Sigma(S)|\right]+n\right)\sum_{s\in S}\sum_{f\in F}\gamma_s(o_f) \quad (٧)$$

که $|\Sigma(S)|$ دترمینان ماتریس کواریانس است.

برای گره با حالتهای S که به دو گره با حالتهای $S_{yes}(q)$ و $S_{no}(q)$ با توجه به پرسش q تقسیم شده، افزایش درستنمایی را به‌صورت زیر به‌دست می‌آوریم:

$$\Delta l_q = L(S_{yes}(q)) + L(S_{no}(q)) - L(S) \quad (٨)$$

مقادیر $\sum_{f\in F}\gamma_s(o_f)$ را برای هر حالت یک بار محاسبه و ذخیره می‌کنیم. همچنین اطلاعات لازم برای محاسبهٔ $\Sigma(S)$ را با توجه به روابط Baum-Welch از روی داده‌های آموزشی استخراج و ذخیره می‌کنیم. به این ترتیب برنامه دسته‌بندی در مدت زمان کوتاهی قابل اجرا خواهد بود.

٤-١-٤- پرسشها

پرسشهایی که در ساخت درخت تصمیم گیری به‌کار رفته بر اساس مطالعات آواشناسی زبان فارسی [۱۳] انتخاب شده است. بر این اساس پرسشهای ساده دو جوابی (بله، خیر) استخراج شده است. هر پرسش در مورد واج سمت راست یا چپ سؤال می‌کند. به‌عنوان مثال در شکل ۲ پرسش " آیا واج سمت چپ خیشومی است ؟" به‌عنوان اولین پرسش مطرح شده است.

پرسشها به چهار دسته تقسیم می‌شود:

الف ـ پرسشهای عمومی

ب ـ پرسشهای واکه‌ها

ج ـ پرسشهای همخوانها

د ـ سایر پرسش‌ها (پرسشهای پیچیده تر).

هر پرسش شامل یک مجموعهٔ واجی است که به واج سمت چپ یا سمت راست اعمال می‌شود. در ضمن هر واج به‌تنهایی نیز به‌عنوان پرسش در نظر گرفته شده است.

بر این اساس برای ساخت درخت تصمیم گیری ۱۵۰ پرسش استخراج شده است. تمام پرسشها برای هر سه حالت هر واج و برای هر دو واج کناری به‌طور جداگانه اعمال می‌شوند. جدول ۲ نمونه‌هایی از این پرسشها را نشان می‌دهد.

جدول ۲ نمونه هائی از پرسشهای درخت تصمیم گیری.

پرسش	مجموعهٔ واجی
پرسشهای عمومی	
واکه	@ a e i o u w
همخوان انفجاری	‘ b d g k p q t
همخوان خیشومی	m n
همخوان سایشی	# $ f h s v x z
همخوان انفجاری – سایشی	c j
همخوان روان و لرزشی	l r y
پرسشهای واکه ها	
واکهٔ بسیط	@ a e l o u
واکهٔ پیشین	@ e i
واکهٔ پسین	a o u w
واکهٔ بسته	i u
واکهٔ متوسط	e o w
واکهٔ باز	@ a
پرسشهای همخوانها	
همخوان باز	# $ f h l m n s v x y z
همخوان بسته	‘ b c d g j k p q r t
همخوان روان	l y
همخوان سایشی واکدار	# v z
همخوان سایشی بیواک	$ f h s x

۵- پیاده‌سازی سیستم پیشنهادی و نتایج

۵-۱- پیاده‌سازی سیستم پیشنهادی

سیستم بازشناسی موجود با استفاده از داده‌های به‌دست آمده از دادگان فارس دات آموزش داده و آزمایش شده است [۱٤]. این پایگاه داده در شکل اصلی خود متشکل از ٦۰۰۰ ادای جمله است که از ۳۰۰ گوینده به‌دست آمده است. هر گوینده به‌طور تصادفی ۲۰ جمله از ٤۰۰ جمله موجود را خوانده است که ۲ جمله بین تمام گوینده‌ها مشترک است. با توجه به این که گوینده‌ها از نقاط مختلف کشور و با لهجه‌های مختلف انتخاب شده‌اند، جملاتی که مربوط به گوینده‌های تهرانی بوده استخراج شده که پس از انتخاب نهایی حدود ۲۷۰۰ جمله به‌دست آمده است. این جملات به دودسته ۱۸۰۰ و۹۰۰ تایی تقسیم شده اند و از دسته اول برای آموزش و از دسته دوم برای آزمایش استفاده شده است. بخش به‌دست آمده از پایگاه داده فارس دات حدود ۱۲۰۰ کلمه مختلف را شامل می‌شود ودر آن حدود ۲٤۰۰ سه آوایی مجزا وجود دارد.

هدف از آموزش مدلها استفاده از آنها برای بازشناسی است، لذا طراحی و پیاده سازی الگوریتم بازشناسی قوی و در عین حال سریع، اهمیت فراوانی دارد. روشی که برای بازشناسی به‌کار گرفته شده موسوم به الگوریتم انتقال نشانه[1] است [٦]. این الگوریتم که بر اساس الگوریتم ویتربی پیاده سازی شده از شیوه جستجوی شعاعی بهره می‌برد که امکان کاهش فضای عظیم جستجوی ایجاد شده در بازشناسی گفتار پیوسته را فراهم می‌کند. این کار موجب کاهش فراوان حجم حافظه مورد نیاز و نیز زمان بازشناسی می‌شود. علاوه بر این امکان بهره جستن از نوعی مدل زبان احتمالی نیز در فرایند بازشناسی وجود دارد که به نوبه خود می‌تواند سرعت و دقت بازشناسی را به‌مراتب بهتر کند.

در پیاده سازی سیستم از HMMهای سه حالته بدون انتقال پرشی به ازای هر سه آوایی استفاده شده است. تنها استثنای مدل، فاصله بین کلمات است که دارای انتقال پرشی نیز هست.

برای آموزش سه آواییها از مدلهای اولیه مبتنی بر تک آواییها استفاده کرده ایم. به این ترتیب، مدل ابتدایی هر سه آوایی، مدل آموزش دیده تک آوایی بوده است. سپس به آموزش مدلهای سه آوایی براساس داده‌های آموزشی موجود اقدام کردیم. برای جلوگیری از تخمینهای نامناسب، به مدلهای دارای نمونه‌های آموزشی کمتر از میزان از پیش تعیین شده، اجازه آموزش داده نشد. درعین حال برای پیشگیری از مقادیر بسیار کوچک واریانس، از مقدار *کف واریانس*[2] استفاده کردیم.

پس از این مرحله به گره زدن مدلهای سه آوایی اقدام شد. برای خوشه بندی و گره زدن ، از درخت تصمیم گیری اشاره شده استفاده شد. در این مرحله نیز شرط حداقل نمونه‌های آموزشی برای مقداردهی پارامترها اعمال شده است.

بررسی صورت گرفته روی جملات استفاده شده از دادگان فارس دات ، نشان دهنده این واقعیت است که از مجموع مدلهای دیده شده در داده‌های آموزشی و آزمایشی، ۲۰ سه آوایی اصولا نمونه‌ای در داده‌های آموزشی نداشته‌اند (سه آوایی دیده نشده). همچنین، حدود ٦۳٪ از مدلها (بیش از ۱۵۲۰ مدل از بین حدود ۲٤۳۰ مدل موجود)، کمتر از ۱۰ نمونه آموزشی داشته اند. از آنجا که حداقل نمونه‌های آموزشی برای انجام آموزش مدل ، برابر ۱۰ در نظر گرفته شده است. این تعداد زیاد مدلها، در مرحله آموزش دست نخورده باقی می‌مانند و بنابراین نیاز اساسی به استفاده از چنین روشهائی برای گره زدن پارامترها کاملا آشکار می‌شود. لازم است اشاره شود که این مسأله در برخورد با دادگانهای طبیعی تر و گسترده تر، به‌مراتب

1. Token passing

2. Variance floor

شدیدتر خواهد بود.

۵-۲- نتایج پیاده‌سازی سیستم بازشناسی

بررسی نتایج به‌دست آمده و ارائه شده در جدول ۳ بخوبی نشان دهندهٔ کارآیی روشهای مطرح شده است. همچنانکه ملاحظه می‌شود، استفاده از مدلهای سه آوایی ضمن آنکه به‌طور کلی باعث بهبودی بازشناسی نسبت به استفاده از مدلهای تک آوایی می‌شود، به علت افزایش فراوان در تعداد مدلها، ممکن است به شرایط آموزش ناکافی[1] منجر شده و در نتیجه به کاهش در دقت بازشناسی نیز منجر شود. این امر مستقیماً با میزان دادهٔ آموزشی موجود ارتباط دارد. نتایج به‌دست آمده، کاملاً بر این مسأله صحه می‌گذارد. در شرایط عدم استفاده از مدل زبان (به اصطلاح بدون گرامر)، که نتایج آن در اینجا ارائه شده است، دقت بازشناسی، اگرچه با استفاده از مدلهای سه آوایی (گره نخورده) افزایش یافته، اما این نتیجه هنوز جای بهبود دارد.

جدول ۳ نتایج بازشناسی در حالت بدون گرامر.

مدل	تعداد حالت‌ها	تعداد تقریبی پارامترها	دقت بازشناسی
تک آوائی	۹۶	۵۰۰۰	٪۴۶/۲
سه آوائی بدون گره زدن	۷۲۸۷	۳۷۶۰۰۰	٪۶۰/۱
سه آوائی گره خورده براساس داده	۱۴۴۰	۸۴۱۰۰	٪۶۷/۸
سه آوائی گره خورده با کمک درخت تصمیم گیری	۱۴۴۲	۸۴۲۰۰	٪۷۱/۱

استفاده از گره زدن موجب شده که تعداد کل حالتها (و بنابراین به همان نسبت تعداد پارامترها) در سیستم کاهش یابد. این کاهش در هر دو مورد (گره زدن بر اساس داده و گره زدن درختی) با توجه به شرایط اعمال شده، بیش از ۸۰٪ بوده که کاهش قابل تأملی در تعداد پارامترهای سیستم است. همین کاهش موجب شده که دقت بازشناسی در هر دو مورد بهبود قابل توجهی پیدا کند.

بررسی و مقایسهٔ نتایج حاصل از دو مجموعه مدل گره زده شده بر اساس داده و گره زده شدهٔ درختی، نشان دهندهٔ بهبود جزیی در نتایج حاصل از شیوهٔ اخیر است. با این همه، مزیت عمدهٔ این شیوه، توانائی مدلسازی سه آوایی‌های مشاهده نشده در مجموعهٔ آموزسی است. این مطلب برای شیوهٔ گره زدن بر اساس داده، معضل تلقی می‌شود.

لازم است توضیح داده شود که این نتایج همگی براساس HMMهای دارای سه حالت به ازای هر مدل آوایی بدست آمده‌اند. همچنین در تمامی‌این موارد از چگالی مشاهدات تک ـ گوسین استفاده شده است. نتایج مربوط به مخلوطهای چند گوسین در اینجا ارائه نشده اگر چه این کار به راحتی قابل اعمال است و به‌طور طبیعی منجر به بهبودی قابل توجهی افزون بر موارد مطرح شده می‌گردد [۶]. نکته دیگری که باید در اینجا اشاره شود، علت ارائه نتایج در شرایط عدم استفاده از مدل زبان است. با توجه به پائین بودن ضریب سرگشتگی[2] گرامر در دادگان فارس دات ـ که حتی برای مدل زبان ساده ای نظیر زوج کلمه[3] در حدود ۳ است ـ استفاده از این مدل زبان موجب افزایش زیادی در نرخ بازشناسی گفتار، در تمامی‌شرایط مدلسازی آکوستیک می‌شود. این امر موجب غیر واقعی شدن نتایج و مشکل شدن بررسی نتایج به دست آمده از اعمال الگوریتم‌های مورد نظر می‌شود. به عنوان مثال، نتایج بازشناسی در این حالت برای سه آوایی‌های گره خورده بر اساس داده به

1. Undertraining

2. Grammar Perplexity

3. Word pair

۹۴/۱٪ و برای سه آوایی‌های گره خورده به کمک درخت تصمیم گیری به ۹۵/۲٪ در حالت استفاده از مدلهای تک - گوسین می‌رسد.

٦- نتیجه‌گیری و پیشنهادها

استفاده از مدلهای وابسته به متن در بازشناسی گفتار، در مقایسه با مدلهای نابسته به متن، به نتایج بهتری منتهی می‌شود. در مدلسازی وابسته به متن، تعداد پارامترها در سیستم بشدت افزایش می‌یابد. بنابراین، برای رسیدن به نتیجه بهتر باید تمهیداتی برای کم کردن تعداد پارامترهای آزاد اندیشید به نحوی که منجر به افزایش دقت بازشناسی شود. گره زدن حالتها با کاهش پارامترهای سیستم، موجب بهبود آموزش مدلها و در نتیجه افزایش دقت بازشناسی در سیستم می‌شود. استفاده از درخت تصمیم گیری مبتنی بر دانسته‌های آواشناسی، علاوه بر اینکه به نتیجهٔ بهتری منتهی می‌شود، امکان تخمین مدلهای دیده نشده را نیز می‌دهد.

برای رسیدن به کیفیت بهتر، می‌توان تعداد عناصر مخلوط استفاده شده در تابع چگالی احتمال حالتها را افزایش داد. در عین حال، استفاده از پایگاه داده گفتاری با دایره لغات وسیع‌تر و نمونه‌های گفتاری بیشتر و ضبط شده بر اساس جملات متنوع تر، در مقایسه با دادگان موجود، می‌تواند شرایط طبیعی تری را برای بررسی کیفیت سیستم بازشناسی ایجاد شده بر این اساس فراهم سازد.

استفاده از مدلهای زیر لغوی وابسته به متن بین کلمه‌ای نیز می‌تواند گام دیگری در هرچه دقیقتر کردن مدلسازی آکوستیک در بازشناسی گفتار پیوسته فارسی تلقی شود.

استفاده از مدلهای زیر لغوی وابسته به متن بین کلمه‌ای نیز می‌تواند گام دیگری در جهت دقیق‌تر کردن مدلسازی آکوستیک در بازشناسی گفتار پیوسته فارسی تلقی شود. اینگونه مدلسازی موجب می‌شود که تعداد مدلهای سه‌آوایی کلی در سیستم افزایش یافته و نیز پیاده‌سازی سیستم بازشناسی مبتنی بر کلمات بسیار سخت‌تر شود. علت این موضوع، عدم اطلاع از کلمهٔ ممکن بعدی در بررسی و ایجاد مدل برای هر کلمه است. به این ترتیب تعداد مدلهای کاندیدا در انتهای مدل ترکیبی هر کلمه بسیار زیاد خواهد بود.

این شیوهٔ مدلسازی، علاوه بر اینکه مشکلات فوق را ایجاد می‌کند، در گره‌زدن آواها نیز ـ در صورت استفاده از ساختار درختی مورد بحث ـ موجب بزرگتر شدن درخت تصمیم‌گیری خواهد شد که به نوبهٔ خود، ایجاد درخت تصمیم‌گیری را مشکلتر و پردازش لازم برای آن را بیشتر می‌کند. با این همه بهبودی نسبی حاصل از استفاده از این شیوهٔ مدلسازی، استفاده از آن را توجیه می‌کند.

۷- مراجع

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

[2] Lee, K.F.; "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition;" IEEE Trans. Acoust. Speech Sig. Proc., Vol. 38, pp. 599-609; 1990.

[3] Bellegarda, J.R.; Nahamoo, D.; "Tied Mixture Continuous Parameter Modeling for Speech Recognition;" IEEE Trans. Acoust. Speech Sig. Proc. Vol. 38, pp. 2033-2045; December 1990.

[4] Digalakis, V.; Murveit, H.; "Genones: Optimizing the Degree of Tying in a Large Vocabulary HMM-based Sppech Recognizer;" Proc. ICASSP-94; 1994.

[5] Odell, J.J.; "The Use of Context in Large Vocabulary Speech Recognition;" Ph.D. Thesis, Cambridge University, Engineering Department; 1995.

[٦] پرویزی، م؛ احدی، س.م؛ "بازشناسی گفتار فارسی پیوسته با دایره کلمات متوسط؛" مجموعه

مقالات کنفرانس مهندسی برق ایران، تهران ۱۳۸۰؛ ص ۱-۸.

[7] Young, S.J.; "The General Use of Tying in Phoneme-Based HMM Speech Recognisers;" Proc. ICASSP-93, Vol. 1, pp. 569-572, San Francisco; 1993.

[8] Young, S.J.; Odell, J.J.; Woodland, P.C.; "Tree-Based State Tying for High Accuracy Acoustic Modeling;" Proc. ARPA Workshop Human Language Technology, Princeton, NJ; March 1994.

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

[10] HTK: Hidden Markov Model Toolkit Ver. 1.4 Reference Manual, Cambridge University; 1992.

[11] Reichl, W.; Chou, W.; "Robust Decision Tree State Tying for Continuous Speech Recognition;" IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 5; September 2000; pp. 555-566

[12] Chou, W.; Reichl, W.; "High Resolution Decision Tree-based Acoustic Modeling beyond CART;" Proc. ICSLP-98, Sydney, Australia; pp. 2203-2206; 1998

[۱۳] ثمره، ی؛ *آواشناسی زبان فارسی: آواها و ساخت آوایی هجا*؛ چاپ پنجم، ویرایش دوم؛ تهران: مرکز نشر دانشگاهی، ۱۳۷۸.

[14] Bijankhan, M.; et al.; "FARSDAT - The Speech Database of Farsi Spoken Language;" Proc. 5th Australian International Conference on Speech Science and Technology (SST'94).