ارائه یک مدل قطعه‌بندی نرم برای مدلسازی کشش زمانی در سیستمهای بازشناسی واج

فربد رزازی[1]، ابوالقاسم صیادیان[2]*

۱- دانشجوی دکترا، دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر

۲- دانشیار، دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر

***تهران، صندوق پستی ۴۴۱۳ – ۱۵۸۷۵**

eeas35@aut.ac.ir

(دریافت مقاله: بهمن ۱۳۸۱، پذیرش مقاله: بهمن ۱۳۸۲)

چکیده – یکی از پیش فرضهای اولیه مدل مارکوف پنهان، اعمال توزیع آماری هندسی به کشش زمانی بردارهای هر حالت مدل مارکوف پنهان است که به وضوح با طبیعت مسأله سازگار نیست. مدلهای قطعه‌ای اتفاقی و به طور خاص مدل مارکوف پنهان قطعه‌ای تا حدی این مشکل را به ازای پیچیدگی بیشتر در مراحل آموزش و بازشناسی حل می‌کند. در مقابل، مدلهای شبکه عصبی و مدل ادراک گفتار دانشگاه ام آی تی از این مدلسازی چشم‌پوشی کرده‌اند. در این مقاله، شیوه‌ای جدید برای مدلسازی اثر کشش زمانی در هر حالت پیشنهاد شده که ایده اصلی آن در نظر گرفتن اثر قطعات (حالات) آکوستیکی همسایه در تخمین پارامترهای آماری هر قطعه آکوستیکی است. این مدلسازی نه تنها به مقاوم شدن الگوریتم در مقابل تعیین غیردقیق مرز قطعات انجامیده، بلکه تبدیل تدریجی قطعات را نیز در طول زمان مدلسازی کرده است. این ایده به صورت تحلیلی فرمولبندی شده و بر روی بانک اطلاعاتی تیمیت به ازای دسته‌های مختلف واج آزموده شده است. همچنین روش پیشنهادی با استفاده از سه الگوریتم بازشناسی، آزموده و بهینه سازی شده و نتایج به دست آمده در شرایط مختلف مقایسه شده است. نتایج به دست آمده حاکی از کاهش قابل توجه زمان بازشناسی است، در حالی که دقت بازشناسی نیز افزایشی جزیی نشان می‌دهد که نشانگر انطباق بهتر مدل با طبیعت گفتار نسبت به مدلهای دیگر است.

کلید واژگان: بازشناسی گفتار، مدلسازی کشش زمانی، مدل مارکوف، مدل مارکوف پنهان قطعه ای، مدل قطعه ای اتفاقی، قطعه‌بندی نرم.

۱- مقدمه

مدلسازی آکوستیکی گفتار، هسته اصلی سیستمهای بازشناسی گفتار به شمار می‌رود. این مدلسازی باید توانایی بازنمایی مناسب مشخصات مختلف واحد گفتاری را داشته باشد و تفاوت میان افراد گوینده، شیوه بیان، سرعت ادای گفتار، لهجه و شرایط محیطی را تحمل کند. تنوع بسیار زیاد گویش یک واحد زبانی، مدلهای آماری آموزش پذیر را مدل غالب و موفق در مدلسازی آکوستیکی گفتار معرفی کرده است. مدلهای آماری مورد استفاده به طور عمده بر بازنمایی مناسب توزیع آماری بردارهای ویژگی و توصیف تغییرات بردارها در طول واحد گفتاری و همچنین تحمل کشش زمانی قطعات آکوستیکی تأکید داشته‌اند.

اگرچه معیار نهایی در ارزیابی توانایی مدل آکوستیکی،

قدرت مدل در طبقه‌بندی واحدهای زبانی است، اما از نظر روش به کار رفته در مدلسازی پدیده‌های آکوستیکی، مدلهای موجود به دو شاخه اصلی از نظر شیوه مدلسازی گفتار قابل تفکیک هستند. گروه اول با نگاهی تعمیمی و با افزایش درجه آزادی مدلهای کلاسیک و تحلیلی (نظیر مدل مارکوف پنهان[1] [۱]) سعی در ایجاد قابلیت انعطاف پذیری بیشتر در مدل داشته‌اند تا مدل پیشنهادی، با دقت بیشتری به بردارهای ویژگی آموزشی منطبق شود. نمونه این ایده‌های تعمیمی را می‌توان در مدلهای قطعه‌ای اتفاقی با مسیر مشروط[2] [۲] و مدل مارکوف پنهان قطعه‌ای[3] [۳-۶] یافت.

گروه دوم ایده‌های مدلسازی با تمرکز بر بخشهای با اهمیت‌تر مدلسازی آکوستیکی و بدون پایبندی به مدلهای کلاسیک، تعداد پارامترها را به حداقل رسانده و دقت مدلسازی را در پدیده‌های تفکیک کننده‌تر در بازشناسی گفتار متمرکز می‌کنند. از این دسته می‌توان به مدل شبکه‌های عصبی[۷-۹] و مدل ادراک گفتاری دانشگاه ام آی تی[4] [۱۰-۱۳] در طبقه‌بندی واج‌ها نام برد. اهمیت استفاده از این گروه مدلها هنگامی برجسته‌تر می‌شود که به محدود بودن بانکهای اطلاعاتی و کاهش دقت تخمین پارامترها با افزایش تعداد پارامترها توجه شود؛ عموماً در بازشناسی گفتار، مسأله محدود بودن بانکهای اطلاعاتی آموزشی، مسأله‌ای حیاتی محسوب می‌شود.

در این مقاله، روشی برای مدلسازی آکوستیکی پدیده کشش زمانی گفتار ارائه می‌شود که مبنای آن، در نظر گرفتن اثر قطعات همسایه در توزیع بردارهای هر قطعه است. در مدلسازی آکوستیکی پدیده کشش زمانی، مدل پیشنهادی این مقاله را می‌توان در گروه دوم مدلهای آکوستیکی دسته‌بندی کرد. در واقع این مدل تلاش دارد تا میزان تمرکز خود را به دو پدیده عمده تخمین توزیع هر قطعه آکوستیکی و تحمل کشش زمانی قطعات آکوستیکی، به صورتی عادلانه و منطبق بر اهمیت نسبی آنها در بازشناسی گفتار تقسیم کند. بنابراین حتی المقدور سعی شده است که با ساده‌سازی مدل و کاهش تعداد پارامترها ، حداکثر کارایی حاصل آید.

ساختار این مقاله به این صورت است که ابتدا در بخش دوم، پس از بررسی روشهای کلاسیک مدلسازی اثر کشش زمانی در مدلهای موجود، اثر مدل مارکوف مرتبه اول به عنوان شیوهٔ غالب در مدلسازی طول قطعات آکوستیکی بررسی و نقد می‌شود. الگوریتم پیشنهادی در بخش سوم ارائه و یک مدل تحلیلی برای استفادهٔ از این مدل پیشنهاد می‌شود. سیستم پیاده‌سازی شده و نتایج آزمون این روش بر روی بانک اطلاعاتی تیمیت[5] در بخش چهارم آورده شده. بخش پنجم روش بازشناسی نرم را به عنوان روش بهینه بازشناسی برای مدلسازی نرم ارائه می‌دهد و بخش ششم به مسألهٔ بهینه سازی میزان همپوشانی قطعات به عنوان مهمترین پارامتر مدلسازی نرم می‌پردازد. در نهایت در بخش هفتم، مقاله جمع بندی شده است.

۲- مروری اجمالی بر روشهای به کار رفته در مدلسازی کشش زمانی گفتار

۲-۱- ابعاد مختلف اثر کشش زمانی

مدلهای آماری مورد استفاده در بازشناسی گفتار، علاوه بر تحمل تنوع آماری - که مشخصه هر مدل آماری به شمار می‌رود - باید بتوانند اثر کشش زمانی را نیز تحمل کنند. این اثر که مشخصه خاص گفتار است، از این واقعیت نشأت می‌گیرد که طول زمانی گفتاری با محتوای کلامی ثابت، به راحتی تغییر می‌کند.

کشش زمانی به دو صورت در مدلسازی گفتار تاثیر می‌گذارد. اول آنکه مشخصات آماری نمونه‌های زمانی یک دنباله مشاهده از یک الگو، در طول زمان تغییر

1. Hidden Markov Model (HMM)
2. Constrained Mean Trajectory Stochastic Segment Models
3. Segmental HMM
4. Speech Understanding Model of MIT (SUMMIT)
5. TIMIT Database

می‌کند. به عنوان مثال، توزیع آماری ویژگیهای استخراج شده از فریمهای ابتدایی واحد زبانی با توزیع آماری بردارهای ویژگی فریمهای پایانی واحد زبانی متفاوت است. این اثر معمولاً با قطعه‌بندی واحد زبانی به چند قطعه[1] (یا حالت[2]) و تخمین توزیعی برای هر قطعه مدلسازی می‌شود. نحوه قطعه‌بندی و چگونگی تخمین توزیع هر قطعه، از مسائل اصلی برخورد با این مسأله به شمار می‌رود.

اثر دوم، تغییر طول زمانی هر قطعه آکوستیکی است که از یک مشخصات آماری پیروی می‌کنند. به عبارت دیگر در صورتی که بخشی از واحد گفتاری در زمان کشیده شود، با توجه به اینکه طول فریمهایی که برای استخراج ویژگی به کار می‌روند ثابت است، تعداد بردارهای ویژگی قطعه تغییر می‌کند و این پدیده، موجب تغییر فاحش تابع درستنمایی قطعه، به ازای کششهای مختلف زمانی یک الگوی خاص می‌شود. با توجه به اینکه اکثر مدلسازیهای گفتار با هدف بازشناسی مقایسه‌ای انجام می‌شود، مقدار مطلق تابع درستنمایی ارزشمند نیست و در بسیاری از مدلسازیها، از اثر پدیده تکرار بردارهای ویژگی تقریباً یکسان در یک قطعه آکوستیکی برای محاسبه تابع درستنمایی چشم‌پوشی شده است. اما در مواردی که مقادیر توابع درستنمایی با هم مقایسه نمی‌شود (مانند کاربردهای تصدیق گفتاری) این مدلسازی ارزشمند است. حتی در موارد مقایسه‌ای نیز عدم توجه به این پدیده، موجب افزایش خطای بازشناسی می‌شود. پالایش داده‌ها، مدلسازی همبستگی آماری بردارهای هر قطعه، استفاده از فریمهای با طول متغیر و مدلهای پیچیده قطعه‌بندی اتفاقی[3]، برای مقابله با این اثر پیشنهاد شده اند.

این مقاله به بررسی اثر اول کشش زمانی یعنی مدلسازی قطعه‌بندی بهینه واحد زبانی به قطعاتی با توزیعهای ثابت در بازشناسی واج می‌پردازد. در مسأله خاص بازشناسی واج، هر یک از مدلهای مورد استفاده به شیوه‌های مختص خود با این اثر برخورد کرده‌اند. در مدل مارکوف پنهان، مدلسازی کشش زمانی با استفاده صریح از اتصال حالات با زنجیره مارکوف درجه یک انجام شده است. هدف از استفاده از این مدل، مدلسازی این واقعیت است که در دنباله گفتاری، مشخصات آماری بردارهای ویژگی برای زمانی ثابت می‌ماند و پس از آن، بردارهای ویژگی از توزیع دیگری پیروی می‌کنند. این مدت زمان بسته به کشش زمانی واحد زبانی متغیر است و مدل آماری باید بتواند این تغییر را تحمل کند. زمان سپری شده در هر حالت مدل مارکوف پنهان، توسط مدل مارکوف مرتبه اول اتصال قطعات (حالات) مدلسازی می‌شود. توزیع آماری این زمان به صورت ضمنی به شکل توزیع هندسی در مدل مارکوف مدلسازی شده است[۲،۱] . مشخصه این مدلسازی، استفاده از قطعه‌هایی با طول متغیر است که توسط برنامه‌نویسی پویا (الگوریتم ویتربی) انتخاب می‌شود و در محاسبه تابع درستنمایی استفاده می‌شود. این روش اگرچه به صورتی قطعه‌بندی واحد گفتاری را مدل می‌کند، اما برای تغییر تدریجی مشخصات آماری از یک قطعه به قطعه دیگر چاره‌ای نیندیشیده است.

استفاده از مدل مارکوف پنهان، محدودیتهای دیگری نیز در مدلسازی دارد. در نظر نگرفتن وابستگی بردارهای ویژگی در مدلسازی هر توزیع - که برطرف ساختن آن به روشهای مدلسازی قطعه‌ای اتفاقی[4] منجر می‌شود - عدم آموزش مقایسه‌ای - که برطرف ساختن آن به روشهای آموزش تفکیک کننده[5] منجر می‌شود - مدلسازی بر مبنای مدل مخلوطهای گاوسی[6] با مخلوطهای محدود - که برطرف ساختن آن مبنای استفاده از شبکه‌های عصبی و همچنین مدلهای غیرخطی مدلسازی توزیع هر حالت شده است - برخی از مشکلات است که هیچ یک از این مسائل، موضوع این مقاله نیستند.

1. Segment
2. State
3. Stochastic Segment Model
4. Stochastic Segment Modeling7
5. Discriminative
6. Gaussian Mixture Model (GMM)

روش کلاسیک رفع محدودیت مدل مارکوف در زمینه توزیع طول زمان حضور در هر حالت، استفاده از مدل مارکوف پنهان قطعه‌ای[1] است. در این مدل، زمان حضور در هر حالت مدل مارکوف پنهان به طور صریح مدلسازی و پارامترهای این توزیع زمانی نیز مستقیماً از داده‌های تجربی استخراج می‌شود. در هنگام خروج از هر حالت، گذر به حالتهای دیگر با یک مدل احتمالی ساده مدلسازی شده است. مزیت اصلی این تعمیم مدل مارکوف پنهان، بالا بردن نرخ بازشناسی در حدود ٪۳-۰ در موارد مختلف است و عیب اصلی آن افزایش تعداد پارامترهای سیستم است که موجب افزایش زمان آموزش، کاهش دقت تخمین پارامترها با بانک اطلاعاتی محدود و افزایش پیچیدگی سیستم و امکان به دام افتادن در حداقلهای محلی و همچنین افزایش زمان بازشناسی می‌شود [۳-۶].

روش دیگر، قطعه‌بندی با طول ثابت است که در مدلسازی بر مبنای شبکه عصبی پرسپترون چندلایه[2] مورد توجه قرار گرفته است[۷-۹]. در این مدل، هر واحد زبانی به چند قطعه با طول از پیش تعیین شده (و معمولاً مساوی) تقسیم می‌شود. سپس توسط شبکه عصبی، یک توزیع غیر خطی به هر قطعه منطبق می‌گردد. این رویکرد - که بیشتر برای بازشناسی و طبقه‌بندی واج‌ها استفاده شده - بیشتر به علت انطباق آسانتر آن با نحوه کارکرد این نوع از شبکه عصبی مورد نظر بوده است تا اینکه مستقیماً مدلی برای کشش زمانی محسوب شود. نتایج ارائه شده نیز همگی با استفاده از مدل شبکه عصبی به ازای هر توزیع ارائه شده‌اند.

مزیت استفاده از این مدل، سادگی و کاهش تعداد پارامترهای سیستم است و در مقابل، عیب عمده‌ای که موجب شده این روش کمتر در شیوه‌های غیر از شبکه عصبی مورد توجه باشد، عدم مدلسازی کشش زمانی در هر قطعه واحد زبانی است، در حالی که این پدیده در گفتار واقعی وجود دارد و مستقیماً نتایج بازشناسی را تحت تاثیر قرار می‌دهد.

مدل دیگری که در جهت بهبود مدلسازی کشش زمانی گام برمی‌دارد، مدل ادراک گفتاری دانشگاه ام آی تی است که مبنای بسیاری از سیستمهای عملی امروزی نظیر ژوپیتر[3] و وویجر[4] است.[۱۰-۱۶] در این مدل عملاً اثر کشش زمانی در سطح واج مدل نشده و توزیع بردارهای هر واج توسط یک یا چند مخلوط گاوسی مدل شده است. در این مدل، کشش زمانی، به پیچیدگی شیوه قطعه‌بندی دنباله گفتاری، به دنباله واج‌ها و انتخاب مناسب واحدهای زبانی سپرده شده و در واقع پیچیدگی سیستم به بخش قطعه‌بندی مناسب و جستجو در درخت قطعه‌ها منتقل شده است [۱۷]. اگرچه این روش ساختار هر واج را با یک قطعه آکوستیکی مدل می‌کند، اما فرایند جستجو در درخت قطعات (دندوگرام[5]) موجب محدودیت در دقت و سرعت سیستم بازشناسی شده است.

۲-۲- اثر مدل مارکوف مرتبه اول

انگیزه اصلی برای بررسی مدلسازی کشش زمانی، آزمونی است که گزارش مختصر آن در این بخش ارائه می‌شود. هدف از این آزمون، بررسی نقش مدلسازی مارکوف مرتبه یک در مدلسازی مارکوف پنهان در بازشناسی بوده است. برای این منظور سیستم بازشناسی با هفت واج مصوت (معادل هفت واج مصوت فارسی) با مشخصات جدول ۱ طراحی و پیاده‌سازی و آموزش داده شده است.

اثر مدلسازی توالی قطعات با مدل مارکوف در این سیستم بررسی شده است. نتایج به دست آمده برای بازشناسی مجموعه آزمون بانک اطلاعاتی تیمیت در جدول ۲ آورده شده است.

نتایج فوق نشان می‌دهد که استفاده از مدل مارکوف

1. Segmental HMM
2. MultiLayer Perceptron (MLP)
3. Jupiter
4. Voyager
5. Dendogram

جدول ۱ پارامترهای HMM پیاده‌سازی شده

پارامتر	مقدار	پارامتر	مقدار
مدل	مارکوف پنهان با مخلوطهای پیوسته	مقادیر اولیه	K-means
تعداد مخلوطها	۴	الگوریتم بازشناسی	ویتربی(Viterbi)
تعداد حالتها	۳	بانک اطلاعاتی	تیمیت
روش آموزش	بام- ولش	حداقل واریانس	۰/۰۱
ویژگیها	MFCC + ΔMFCC + ΔLogE	پیش پردازش	فیلتر پیش‌تأکید
نوع قاب مورد استفاده	همینگ با طول ۱۶ m.s	همپوشانی قابها	٪۵۰

جدول ۲ نتایج آزمون اثر مدل مارکوف در بازشناسی

مجموعه / روش بازشناسی	مردان	زنان
ویتربی	۷۶/۵۷	۸۱/۲۵
حذف اثر مارکوف در قطعه‌بندی و تابع درستنمایی	۷۶/۹۹	۸۱/۲۰
حذف اثر مارکوف فقط در تابع درستنمایی	۷۶/۸۲	۸۱/۲۵

در بازشناسی تأثیر قابل توجهی در صحت بازشناسی ندارد و حتی در صورتی که این مدلسازی از محاسبه تابعدرستنمایی حذف شود - البته در حالی که قطعه‌بندی هر واج به قطعات آکوستیکی با محاسبه اثر مدل مارکوف باشد - نتیجه بازشناسی بهبود مختصری خواهد داشت. این نتیجه بیشتر از این نظر حائز اهمیت است که اثر قابل چشم‌پوشی مدلسازی مارکوف را در اتصال قطعات نشان می‌دهد. به عبارت دیگر اثر غالب در مدلسازی مارکوف پنهان، دقت تخمین توزیع قطعات آکوستیکی است که به صورت متوالی در واحد زبانی ظاهر می‌شوند و مدل آماری اتصال این قطعات تأثیر قابل توجهی بر بازشناسی نخواهد داشت.

۳- مدل قطعه‌بندی نرم

۳-۱- ایده اصلی

آنچه را در دو بخش قبلی بررسی شد، به صورت زیر می‌توان جمع بندی کرد:

الف- پیش فرضهای محدود کننده مدل مارکوف پنهان در مورد توزیع هندسی زمان حضور در قطعات (حالات) به مدلسازی غیر واقعی گفتاری می‌انجامد.

ب- جستجوی مرز دقیق قطعات به پیچیدگی مدلسازی می‌افزاید که به نوبه خود موجب افزایش زمان آموزش و بازشناسی می‌شود. همچنین افزایش پارامترهای مدلسازی سبب کاهش دقت تخمین پارامترها و در نتیجه محدودیت رشد نرخ بازشناسی می‌شود.

ج- قطعه‌بندی با طول ثابت، مدلسازی قطعات با طول متغیر را حذف می‌کند و در واقع این مدلسازی غیر واقعی، پیچیدگی و همچنین دقت مدلسازی را به بخشهای دیگری از مدل مانند تخمین پارامترهای توزیعها منتقل می‌کند.

د- در واقع مرز مشخصی نیز برای قطعات وجود ندارد و تغییر مشخصات آماری از یک قطعه به قطعات

مجاور به طور تدریجی انجام می‌شود.

روشی که در این مقاله پیشنهاد شده، استفاده از قطعه‌بندی ثابتی است که اثر متقابل هر قطعه در قطعات دیگر نیز در توزیع آن، در نظر گرفته شده است. به این ترتیب در مرحله آموزش، کشش زمانی قطعات با تغییر تدریجی هر قطعه به قطعه بعدی جبران و کم اثر شده است. قابلیتهای اولیه این مدلسازی به ازای واج‌های مصوت در طی مقاله کنفرانس [۱۸]، توسط نویسندگان، بررسی شده است. اما این مدلسازی به واج‌های مختلف تعمیم نیافته و در شرایط مختلف آزموده نشده است.

در شکل ۱ این نحوه قطعه‌بندی با قطعه‌بندی ثابت مقایسه و تفاوت آن نشان داده شده است. در مدل اول - که مبنای تمام مدلهای موجود در مدلسازی گفتار است - بردارهای هر قطعه به ازای گفتارهای مختلف جمع آوری و توزیع این بردارها توسط مدل مخلوطهای گاوسی مدلسازی می‌شود.

در مدل پیشنهادی، بردارهای همسایه هر قطعه نیز در محاسبه توزیع قطعه با وزن مناسب تأثیر داده می‌شود تا مدلسازی نسبت به عدم تطبیق دقیق مرزبندی قطعات در هنگام بازشناسی مقاوم باشد. مقدار این همپوشانی بسته به داده‌های آموزشی، قابل تنظیم است. واضح است که این همپوشانی به طول واحد زبانی محدود است.

ایده همپوشانی در فرایند قاب‌بندی و محاسبه بردارهای ویژگی با مقدار همپوشانی نوعی ۵۰٪ در بیشتر سیستمهای پردازش گفتار وجود دارد. این همپوشانی به از دست نرفتن اطلاعات فرکانسی در مرز قابها و دقت بیشتر طیف فرکانسی کمک می‌کند. در حالی که واحد واقعی اکوستیکی، قابهای همطول نیستند و استفاده از همپوشانی میان قطعات اکوستیکی با توزیع آماری یکسان، می‌تواند استفاده صحیح تری از همپوشانی قطعات به شمار رود.

در مرحله بازشناسی، با توجه به مقاوم شدن مدل در مقابل اثر کشش زمانی قطعات اکوستیکی و کاهش اثر جابه‌جایی مرزهای قطعات در داخل هر واحد زبانی و در نتیجه، کاهش اثر کشش زمانی در تابع درستنمایی، استفاده از قطعه‌بندی ثابت و سخت کافی خواهد بود. مزیت این روش، تغییر اساسی در سرعت بازشناسی است و حتی در صورتی که تغییر چشمگیری در نرخ بازشناسی حاصل نشود، عدم استفاده از قطعه‌بندی پویا، موجب کاهش زمان بازشناسی به یک سوم مقدار قبلی خواهد شد. در نتایج به دست آمده نشان داده می‌شود که در مجموع نه تنها کاهش نرخ بازشناسی مشاهده نمی‌شود، بلکه همراه با کاهش چشمگیر زمان بازشناسی افزایش نرخ بازشناسی، نیز قابل مشاهده خواهد بود.

مشکلی که این روش مدلسازی را تهدید می‌کند، کاهش دقت توزیع هر قطعه به علت همپوشانی قطعات مجاور است. هر چه این همپوشانی بیشتر باشد، توزیع قطعات بیشتر به یکدیگر نزدیک می‌شود و در واقع توزیع قطعات به یک توزیع عمومی برای واحد زبانی نزدیکتر می‌شود. این عامل، میزان همپوشانی قطعات را برای بهینه سازی نرخ بازشناسی محدود می‌سازد.

۳-۲- مدلسازی تحلیلی

مبنای مدلسازی مورد استفاده، تخمین پارامترهای توزیع آماری هر قطعه با استفاده از مدل GMM بوده است که احتمال هر بردار ویژگی را به صورت زیر مدل می‌کند:

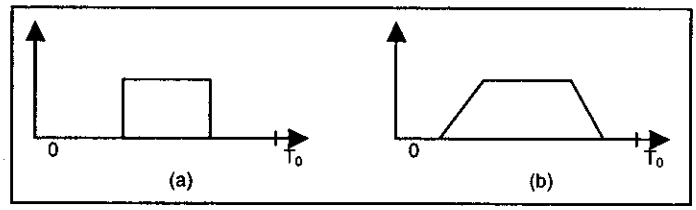
$$P(X_i \mid S_j) = \sum_{k=1}^{M} C_{jk} N(X_i; \mu_{jk}, \Sigma_{jk}) \qquad (1)$$

در مدل قطعه‌بندی نرم، در مرحله آموزش هر قطعه، قطعات همسایه نیز در این تخمین سهیم هستند در حالیکه در مرحله بازشناسی، فقط پارامترهای هر قطعه برای بازشناسی بردارهای آن قطعه استفاده شده‌اند.

۳-۲-۱- مرحله آموزش پارامترها

روش معمول برای تخمین پارامترهای مدل مخلوطهای گاوسی، استفاده از الگوریتم EM[1] به ازای بردارهای ویژگی هر قطعه است. با توجه به اینکه بردارهای مورد

1. Expectation-Maximization



شکل ۱ مقایسه مدل قطعه‌بندی سخت(a) و قطعه‌بندی نرم(b)

نظر در الگوریتم EM باید با وزن غیر یکسان اعمال شوند، ناگزیر این الگوریتم باید تصحیح شود. روش پیشنهادی آموزش بر مبنای این الگوریتم به صورت زیر است:

$$P_t(i) = \frac{\alpha_i N(Xt; \mu_i, \Sigma_i)}{\sum_{j=1}^{M} \alpha_j N(Xt; \mu_j, \Sigma_j)} \qquad (۲)$$

$$\alpha_i = \frac{\sum_{t=1}^{T_0} w_t . P_t(i)}{\sum_{t=1}^{T_0}\sum_{j=1}^{M} w_t . P_t(i)} \qquad (۳)$$

$$\mu_i = \frac{\sum_{t=1}^{T_0} w_t . P_t(i) . X_t}{\sum_{t=1}^{T_0} w_t . P_t(i)} \qquad (۴)$$

$$\Sigma_i = \frac{\sum_{t=1}^{T_0} w_t . P_t(i) . (X_t - \mu_i)(X_t - \mu_i)}{\sum_{t=1}^{T_0} w_t . P_t(i)} \qquad (۵)$$

که در آن، $P_t(i)$ احتمال مخلوط i ام به شرط مشاهده بردار X_t است.

این مراحل تخمین آنقدر تکرار می‌شود تا تغییرات لگاریتمی تابع بهینه سازی از مقداری از پیش تعیین شده کمتر شود. در روابط بالا، وزن تأثیر بردارهای قطعات مجاور به صورت فراوانی کمتر یا مساوی یک برای این بردارها در نظر گرفته شده است. به عبارت دیگر فرض شده که در لحظه t، تعداد w_t بردار X_t برای مدلسازی مخلوطهای گاوسی حضور داشته است.

برای آن که پهنای هر یک از مخلوطهای گاوسی، بر اثر کمبود داده‌ها در مجموعه آموزشی از حدی کمتر نشود و مدل مخلوطهای گاوسی قدرت تعمیم خود را به داده‌های دیگر از دست ندهد، از روش هموارسازی معمول - که محدود کردن حداقل مقدار واریانس هر مؤلفه بردار است - استفاده شد. این مقدار، ۰/۰۱ واریانس هر مؤلفه در نظر گرفته شده است.

تابع بهینه سازی مورد استفاده، احتمال دنباله بردارهای X_t به شرط مدل فعلی است که با فرض استقلال بردارهای هر قطعه به صورت زیر قابل بازنویسی است:

$$P(X \mid \lambda) = \prod_{t=1}^{T_0} P(X_t)^{w_t} \qquad (۶)$$

در این تابع نیز، اثر پنجره به صورت فراوانی بردار X_t در لحظه t و به تعداد w_t مدلسازی شده است. برای مقادیر اولیه الگوریتم EM به ازای هر قطعه از الگوریتم K-means استفاده شده است. نتیجه این مقادیر اولیه تهیه یک کتاب کد توابع گاوسین با احتمالهای حضور متفاوت به ازای هر قطعه است.

۳-۲-۲- مرحله بازشناسی

در مرحله بازشناسی، پس از قطعه‌بندی هر واحد گفتاری، بردارهای هر قطعه با مدل مخلوطهای گاوسی قطعه سنجیده می‌شود. بنابراین با فرض استقلال بردارهای ویژگی در هر قطعه داریم:

$$j^* = ArgMax(LLF_j) = ArgMax(\prod_{t=1}^{T0} P(X_t \mid S_{t,j})) \qquad (۷)$$

شیوه قطعه‌بندی می‌تواند به صورت طول ثابت و از پیش تعیین شده باشد یا با استفاده از الگوریتم ویتربی، بهترین مرزبندی قطعات برای ایجاد بیشترین تابع درستنمایی جستجو شود. محاسبه احتمال هر بردار با فرض دانستن

قطعه مورد نظر، با استفاده از مدل مخلوطهای گاوسی انجام می‌شود. همچنین مدل غیرخطی کوانتیزاسیون برداری گاوسی، در بازشناسی مورد آزمون قرار گرفته است.

نحوه محاسبه تابع درستنمایی در حالت مدل مخلوطهای گاوسی به صورت زیر است:

$$P(X_t \mid S_j) = \sum_{k=1}^{M} C_{jk} N(X_t; \mu_{jk}, \Sigma_{jk}) \qquad (۸)$$

که در حالت کوانتیزاسیون برداری گاوسی به صورت زیر تغییر می‌یابد:

$$P(X_t \mid S_j) = ArgMax_k (C_{jk} N(X_t; \mu_{jk}, \Sigma_{jk})) \qquad (۹)$$

برای بهبود مدلسازی دنباله واج‌ها می‌توان از مدل زبانی (۱-گرام یا ۲-گرام) استفاده کرد، اما با توجه به اینکه هدف از این پژوهش، بررسی عملکرد مدل اکوستیکی بوده، اثر مدل زبانی در محاسبه نرخ بازشناسی در نظر گرفته نشده است. همچنین در این مقاله، همپوشانی به صورت نموداری ذوزنقه‌ای و به شکل خطی با حداکثر همپوشانی در نظر گرفته شده که در ادامه پژوهش باید بهینه سازی شود.

٤- پیاده‌سازی و تحلیل نتایج

۴-۱- پیاده‌سازی

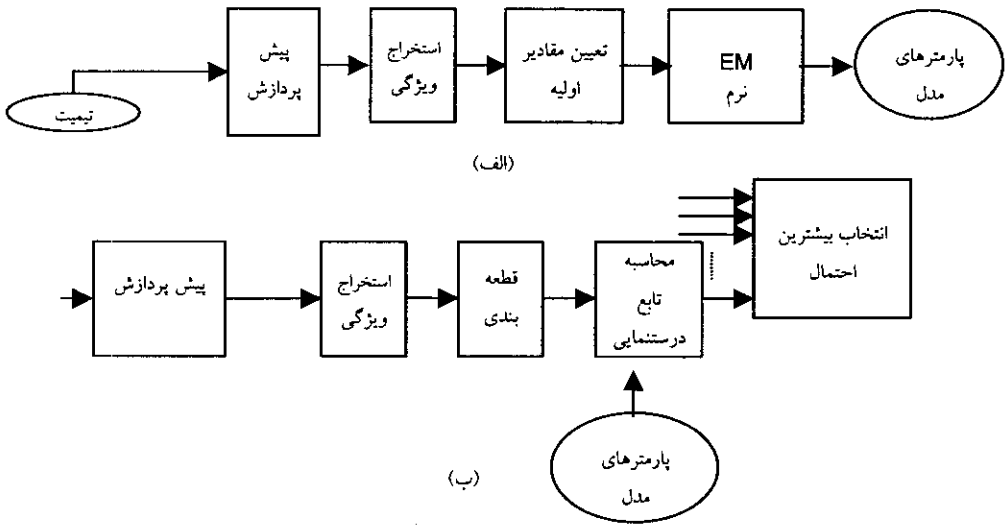
برای ارزیابی روش پیشنهادی، از بانک اطلاعاتی برچسب دار تیمیت استفاده شده است. این بانک اطلاعاتی که از لحاظ فونتیکی متعادل است، از گویش ۶۳۰ نفر تشکیل شده که ۷۰٪ آن را مردان و ۳۰٪ آن را زنان تشکیل می‌دهند. همچنین ۸ لهجه معمول امریکای شمالی با آمارگان منطبق بر زبان، مورد توجه و جمع آوری قرار گرفته است.

در کل این آزمایشها، طبقه‌بندی واج‌های دسته‌های مختلف واج‌های زبان انگلیسی برای آزمون مدل قطعه‌بندی نرم مورد نظر بوده است. برای ارزیابی سیستم قطعه‌بندی نرم، سیستم مدل مارکوف پنهان سه حالتی با چگالی احتمال پیوسته با چهار مخلوط به عنوان معیار مقایسه، پیاده‌سازی شده که مشخصات آن در جدول ۱ آورده شده است. برای آموزش سیستم، ابتدا تخمین اولیه مقادیر پارامترهای مدل مارکوف پنهان توسط الگوریتم K-means انجام و سپس با الگوریتم آموزش بام-ولش، پارامترهای مدل مارکوف پنهان تخمین زده شده است. برای حفظ توانایی تعمیم مدل به مشاهداتی که در مرحله آموزش حضور نداشته‌اند، حداقل واریانس هر مؤلفه بردار ویژگی در توزیع هر قطعه، برابر با ۰/۰۱ حداکثر آن مؤلفه در نظر گرفته شده است. در مرحله بازشناسی، از الگوریتم ویتربی برای جستجوی بهترین قطعه‌بندی (انتخاب مرز قطعات) و محاسبه تابع درستنمایی استفاده شده است.

شکل ۲ بلوک دیاگرام کلی سیستم آموزش و بازشناسی را با استفاده از مدلسازی نرم نشان می‌دهد. در مرحله آموزش، دنباله نمونه‌های آموزشی پس از استخراج ویژگی، به دنباله‌ای از بردارهای ویژگی تبدیل می‌شود و توزیع آماری هر قطعه توسط الگوریتم EM نرم با هشت مخلوط گاوسی تخمین زده شده و مشخصات این توزیعهای آماری به عنوان پارامترهای مدل در بانک اطلاعاتی ذخیره می‌شود. مقادیر اولیه پارامترهای مدل برای آغاز الگوریتم EM نرم، توسط تصحیح الگوریتم K-means برای قطعه‌بندی نرم به دست آمده است.

در مرحله بازشناسی، پس از استخراج دنباله بردارهای ویژگی، این دنباله قطعه‌بندی و تابع درستنمایی هر قطعه با فرض استقلال بردارهای ویژگی محاسبه می‌شود و مقایسه مقادیر این توابع درستنمایی به ازای مدلهای مختلف، منطبقترین واج بر دنباله ناشناس را آشکار می‌سازد.

مشخصات مراحل پیش پردازش و استخراج ویژگی در جدول ۱ ارائه شده است. نمونه‌های ورودی پیش از استخراج ویژگی و تقسیم به قابهای ۱۶ میلی ثانیه‌ای، از فیلتر پیش تأکید عبور می‌کنند تا قله‌های کم دامنه در فرکانسهای بالا با دقت بیشتری بازنمایی شود. ویژگیهای مورد استفاده در این سیستم، ۱۲ ضریب کپستروم در



شکل ۲ بلوک دیاگرام سیستم الف - آموزش ب - بازشناسی

مقیاس مل[1] همراه با مشتق اول آنها و مشتق لگاریتم انرژی هر قاب است که در مجموع برداری ۲۵ عنصری را به ازای هر قاب تشکیل می‌دهند. میزان همپوشانی قابهای مجاور در مرحله استخراج ویژگی، ۵۰٪ در نظر گرفته شده و از پنجره همینگ در حوزه زمان برای جلوگیری از کاهش دقت طیف فرکانسی هر قاب استفاده شده است.

نرمالیزاسیون بردارهای ویژگی با فرض استقلال درایه‌های بردارها به صورتی انجام شده که میانگین و واریانس داده‌ها به ترتیب به صفر و یک منتقل شود و تمام تستها با هشت مخلوط انجام شده است. تمام آموزشهای بر روی تمام مجموعه‌های آموزشی تیمیت و آزمون سیستم با استفاده از کل مجموعه آزمون تیمیت انجام شده است.

در مرحله بازشناسی، الگوریتم‌های کوانتیزاسیون برداری گاوسی و ترکیب مخلوطهای گاوسی به صورت قطعات هم طول و با طول ثابت و همچنین الگوریتم ویتربی برای قطعه‌بندی پویا با قطعات با طول متغیر به کار رفته و با یکدیگر مقایسه شده است.

۴-۲- تحلیل نتایج

اولین آزمون انجام شده، مقایسه رفتار مدل قطعه‌بندی نرم با مدل مارکوف پنهان است. برای این منظور دو سیستم جداگانه بازشناسی واج‌ها برای مردان و زنان آموزش داده شده و واج‌های مجموعه تست با سه روش معرفی شده، بازشناسی شده‌اند. خلاصه این نتایج به ازای دسته‌های واجی مختلف در جدولهای ۳ و ۴ خلاصه شده است. نتایج به دست آمده برای قطعه‌بندی نرم و مقایسه آن با مدل مارکوف پنهان نشان می‌دهد که بهبودی در حدود یک تا پنج درصد در متوسط نرخ بازشناسی نسبت به مدل مارکوف پنهان استاندارد مشاهده می‌شود، اما آنچه بیشتر به چشم می‌آید، کاهش چشمگیر زمان بازشناسی است که در کاربردهای بی‌درنگ می‌تواند قابل توجه باشد.

در واج‌های مصوت، در مجموعه صدای زنان، روش قطعه‌بندی نرم تقریباً معادل مدل مارکوف پنهان رفتار کرده است، اما در مجموعه صدای مردان، بهبود قابل ملاحظه‌ای مشاهده می‌شود. در واج‌های شبه مصوت، این وضعیت معکوس است و بهبود قابل ملاحظه‌ای در مجموعه صدای زنان مشاهده می‌شود. اما نکته ارزشمند این است که در هیچ

1. Mel

یک از دسته‌های واجی، کاهش معنی داری در نرخ بازشناسی نسبت به مدل مارکوف پنهان وجود ندارد.

در جدول ۴ زمان بازشناسی دسته واج‌های مختلف به ثانیه داده شده است. واضح است که با توجه به تعداد و طول مختلف واج‌های دسته‌های واجی مختلف، زمان بازشناسی دسته‌های واجی متفاوت قابل مقایسه نیست، اما در مقایسه روشهای مدلسازی قطعه‌بندی نرم و مدل مارکوف پنهان، کاهش قابل ملاحظه‌ای در زمان بازشناسی نسبت به مدل مارکوف پنهان دیده می‌شود.

در هر دو سیستم مربوط به مردان و زنان به ازای دو الگوریتم بازشناسی کوانتیزاسیون برداری گاوسی و مدل مخلوطهای گاوسی، زمان بازشناسی با نسبت تقریبی ۳ بهبود نشان می‌دهد، در حالی که استفاده از الگوریتم ویتربی این مزیت را حذف می‌کند. در واقع ویتربی تکرار همان روش بازشناسی مدل مارکوف پنهان است با این تفاوت که تخمین پارامترهای توزیع هر قطعه به شکل دیگری انجام شده است.

جدول ۳ نرخ بازشناسی به ازای دسته‌های واجی مختلف

	مجموعه	مصوتها		شبه مصوتها		صامتها	
جنسیت	روش بازشناسی	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون
مردان	HMM	۷۲/۴۲	۷۵/۵۹	۸۲/۷۲	۸۰/۵۰	۶۸/۷۸	۶۷/۴۶
	VQ گاوسی	۷۷/۸۴	۷۸/۷۲	۸۲/۹۴	۸۱/۱۳	۷۰/۱۲	۶۸/۷۹
	GMM	۷۷/۸۹	۷۸/۷۶	۸۲/۰۲	۸۱/۱۳	۷۰/۲۴	۶۸/۵۹
	ویتربی	۷۷/۴۸	۷۸/۳۵	۸۲/۶۸	۸۱/۲۷	۶۸/۵۸	۶۷/۶۱
زنان	HMM	۷۹/۲۴	۷۸/۰۵	۷۶/۱۱	۷۴/۲۲	۶۸/۲۷	۶۵/۷۸
	VQ گاوسی	۷۹/۶۷	۷۷/۹۵	۸۲/۲۸	۸۱/۴۰	۶۹/۵۶	۶۶/۱۰
	GMM	۷۹/۶۳	۷۷/۷۹	۸۲/۳۸	۸۱/۶۲	۶۹/۶۶	۶۶/۴۰
	ویتربی	۷۹/۳۷	۷۸/۵۶	۸۱/۴۵	۸۰/۴۶	۶۷/۶۷	۶۴/۴۸

جدول ۴ زمان بازشناسی به ازای دسته‌های واجی مختلف

	مجموعه	مصوتها		شبه مصوتها		صامتها	
جنسیت	روش بازشناسی	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون
مردان	HMM	۴۲۱۰	۱۲۷۱	۱۹۹۴	۶۱۲	۲۰۰۸	۵۹۶
	VQ گاوسی	۱۲۶۹	۳۸۲	۶۴۸	۲۱۵	۶۵۶	۱۹۰
	GMM	۱۳۹۸	۴۱۶	۷۱۱	۲۲۹	۶۹۳	۲۰۴
	ویتربی	۴۵۲۱	۱۳۵۰	۲۱۷۶	۷۲۵	۲۱۱۰	۶۲۷
زنان	HMM	۱۷۴۸	۵۶۳	۷۸۶	۲۹۵	۸۵۷	۲۹۷
	VQ گاوسی	۵۱۱	۱۷۱	۲۵۶	۹۲	۲۸۲	۹۵
	GMM	۵۶۲	۱۸۸	۲۷۸	۱۰۰	۲۹۹	۱۰۲
	ویتربی	۱۸۲۶	۶۱۵	۸۷۴	۳۱۴	۹۲۷	۳۱۹

استفاده از ویتربی به عنوان روش بازشناسی مدلسازی قطعه‌بندی نرم با این امید انجام شده که قطعه‌بندی پویا و بهینه، موجب بهبود مرزبندی قطعات و واقعی‌تر شدن مقدار تابع درستنمایی شود. در حالی که در بعضی از دسته واج‌ها (مانند صامتها) این نحوه قطعه‌بندی بیشتر موجب افزایش تابع درستنمایی واج‌های رقیب شده است تا واج صحیح . در نتیجه مجموعاً استفاده از ویتربی موجب کاهش نرخ بازشناسی در این دسته واج‌ها شده است. به نظر می آید در واج‌های صامت، تغییر بین قطعات آنچنان تدریجی است که مرزبندی صریح قطعات به هر صورت موجب کاهش نرخ بازشناسی می‌شود.

مقایسه روشهای مختلف بازشناسی در قطعه‌بندی نرم نشان می‌دهد که در بیشتر تستها به ترتیب ویتربی، مدل مخلوطهای گاوسی و کوانتیزاسیون برداری گاوسی، احتمال بازشناسی بهتری داشته‌اند که با توجه به این که الگوریتم دوم مزیت کاهش زمان بازشناسی را نیز حفظ می‌کند، روش مناسبتری به نظر می‌رسد.

رفتار واج‌های مختلف در مدل قطعه‌بندی نرم یکسان نیست. به عنوان نمونه، رفتار هفت واج مصوت پرکاربرد زبان انگلیسی - که معادل هفت مصوت فارسی هستند - در یک سیستم بازشناسی جداگانه بررسی شده است. جدول ۵ و ۶ نتیجه بازشناسی این واج‌ها را نشان می دهد.

مقایسه رفتار واج‌های مصوت مختلف نشان می‌دهد که افزایش نرخ بازشناسی در مدلسازی قطعه‌بندی نرم نسبت به مدل مارکوف پنهان استاندارد در تمام واج‌ها یکسان نیست و گاهی کاهش نرخ بازشناسی نیز دیده می‌شود. این پدیده با توجه به مقایسه یا بودن عملیات بازشناسی و همچنین با توجه به این واقعیت قابل توجیه است که رفتار واج‌های مختلف - بسته به اینکه توزیع آماری قطعات آنها در طول زمان چقدر متغیر است - تغییرهای متفاوتی را در نرخ بازشناسی نشان می‌دهند. اگرچه در مجموع این مدلسازی موجب بهبود نرخ بازشناسی متوسط واج‌ها شده است.

جدول ۵ نتایج بازشناسی برای مجموعه زنان

مدل	بازشناسی الگوریتم	ow	ux	uh	Iy	eh	ae	aa	متوسط
HMM	ویتربی	۷۵/۹۴	۶۳/۴۵	۷۲/۳۴	۸۵/۹۵	۶۶/۲۰	۷۴/۴۵	۸۵/۶۱	۷۸/۰۵
قطعه‌بندی نرم	VQ گاوسی	۷۴/۸۷	۵۶/۵۵	۶۸/۰۹	۸۷/۳۱	۶۶/۹۰	۷۶/۶۴	۸۳/۴۵	۷۷/۹۵
	GMM	۷۵/۴۰	۵۵/۸۶	۶۸/۰۹	۸۷/۶۱	۶۶/۲۰	۷۶/۰۱	۸۳/۰۹	۷۷/۷۹
	ویتربی	۷۸/۰۷	۵۷/۲۴	۷۰/۲۱	۸۶/۵۶	۶۴/۸۱	۸۰/۰۶	۸۴/۸۹	۷۸/۵۶

جدول ۶ نتایج بازشناسی برای مجموعه مردان

مدل	الگوریتم بازشناسی	ow	ux	uh	iy	eh	ae	aa	متوسط
HMM	ویتربی	۸۶/۲۹	۸۵/۱۶	۵۴/۷۰	۷۸/۰۸	۷۹/۱۳	۵۲/۴۲	۸۶/۲۶	۷۵/۵۹
قطعه‌بندی نرم	VQ گاوسی	۷۵/۸۱	۵۷/۸۱	۵۹/۸۳	۸۴/۵۷	۶۰/۹۵	۷۶/۰۸	۹۰/۷۸	۷۸/۷۲
	GMM	۸۷/۴۷	۷۵/۴۸	۶۰/۶۸	۸۴/۵۷	۶۰/۸۲	۷۶/۴۷	۹۰/۴۵	۷۸/۷۶
	ویتربی	۸۷/۷۱	۷۷/۴۲	۵۹/۸۳	۸۲/۵۲	۶۴/۷۹	۷۳/۵۹	۸۹/۷۸	۷۸/۳۵

در قدم بعدی سعی شده تا سیستمی پیاده‌سازی شود که بدون توجه به جنسیت، واج‌های مورد نظر را بازشناسی کند. برای این منظور کلیه مدلهای آموزش داده شده در آزمون اول در کنار یکدیگر قرار گرفت و سیستمی با تعداد مدل دو برابر واج‌ها به دست آمد که هر دو مدل آن، به یک واج اشاره می‌کند. این سیستم نسبت به صدای مردان و زنان مقاوم خواهد بود. نمونه نتایج استفاده از این سیستم به ازای مجموعه آزمون زنان و مردان برای واج‌های مصوت معادل فارسی در جدولهای ۷ و ۸ گردآوری شده است.

نتایج به دست آمده نشان می‌دهد که این استراتژی، کاهش قابل توجهی را در نرخ بازشناسی نسبت به حالت وابسته به جنسیت ایجاد نمی‌کند؛ در حالی که در هنگام آموزش، پارامترهای مدل با سرعت بیشتری تخمین زده می‌شوند و با توجه به مدلسازی جداگانه مردان و زنان دقت آن از سیستم یکپارچه‌ای برای مدلسازی کل افراد بیشتر است.

۵- بازشناسی نرم در مقابل بازشناسی سخت

در بخش سوم، پس از معرفی قطعه‌بندی نرم تصریح شد که استفاده از قطعه‌بندی ثابت برای بازشناسی مناسب است و نیازی به قطعه‌بندی پویا وجود ندارد و نتایج آزمونها نیز این مسأله را تأیید می‌کند. در آزمونهای انجام شده نیز از مرزهای ثابت و سخت در قطعه‌بندی قطعات استفاده شد. در حالی که مرحله آموزش، در واقع انجام متوالی محاسبه تابع درستنمایی و بازتخمین پارامترهای توزیعهای هر قطعه است و در مرحله محاسبه تابع درستنمایی، این تابع به صورت وزن دار و با در نظرگرفتن اثر قطعات همسایه محاسبه می‌شود.

بنابراین فرایند آموزش به بهینه سازی مدلی می‌انجامد که در محاسبه تابع درستنمایی آن، تمام بردارهای ویژگی با وزن مناسب نقش دارند و انتظار می‌رود که استفاده ازاین روشِ محاسبه تابع درستنمایی در مرحله بازشناسی، به علت انطباق بهتر با شیوه آموزش، به نرخ بازشناسی بالاتری منجر شود. این الگوریتم بازشناسی - که بازشناسی نرم نامگذاری شده - در بازشناسی مجموعه‌های مختلف آزموده شده و نتایج آن در جدولهای ۹ و ۱۰ آمده است. نتایج به دست آمده نشان می‌دهد که اگرچه بهبود اندکی در اغلب موارد بازشناسی حاصل شده، اما به علت استفاده مکرر از هر بردار ویژگی در محاسبه تابع درستنمایی قطعات مختلف، زمان بازشناسی تقریباً سه برابر شده است. به این ترتیب در عمل مزیت کاهش زمان بازشناسی نسبت به مدل مارکوف پنهان از میان رفته و در مقابل بهبود اندکی نسبت به بازشناسی سخت به دست آمده است.

جدول ۷ نتایج بازشناسی برای مجموعه زنان (سیستم مستقل از جنسیت)

بازشناسی الگوریتم	ow	ux	Uh	iy	eh	ae	aa	متوسط	زمان بازشناسی
VQ گاوسی	۶۷/۹۱	۵۳/۷۹	۴۸/۹۳	۸۹/۲۷	۶۹/۶۹	۶۸/۲۲	۸۷/۰۵	۷۶/۸۰	۳۹۴
GMM	۶۸/۴۵	۵۳/۱۰	۴۸/۹۴	۸۹/۲۷	۶۸/۶۴	۶۸/۸۵	۸۶/۶۹	۷۶/۷۰	۴۲۰
ویتربی	۶۸/۹۸	۵۳/۷۹	۵۳/۱۹	۸۹/۳۷	۷۲/۵۲	۶۹/۷۸	۸۵/۲۵	۷۷/۲۷	۱۳۵۱

جدول ۸ نتایج بازشناسی برای مجموعه مردان (سیستم مستقل از جنسیت)

بازشناسی الگوریتم	ow	ux	uh	iy	eh	ae	aa	متوسط	زمانبازشناسی
VQ گاوسی	۸۵/۸۲	۷۴/۱۹	۶۴/۱۰	۸۱/۵۴	۶۳/۱۲	۷۰/۸۵	۸۹/۴۵	۷۶/۹۳	۸۵۵
GMM	۸۵/۵۸	۷۴/۱۹	۶۴/۹۶	۸۱/۶۱	۶۳/۳۸	۷۲/۱۶	۸۹/۲۸	۷۷/۲۰	۹۴۹
ویتربی	۸۵/۳۵	۷۳/۵۵	۶۴/۹۶	۸۰/۴۱	۶۴/۵۳	۷۱/۳۷	۸۸/۲۷	۷۶/۶۸	۲۹۶۳

جدول ۹ مقایسه نتایج بازشناسی‌های نرم و سخت با روش بازشناسی کوانتیزاسیون برداری گاوسی

ردیف بالا: نرخ بازشناسی، ردیف پایین: زمان بازشناسی بر حسب ثانیه

	مجموعه	مصوتها		شبه مصوتها		صامتها	
جنسیت	نوع بازشناسی	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون
مردان	سخت	۷۷/۸۴	۷۸/۷۲	۸۲/۹۴	۸۱/۱۳	۷۰/۱۲	۶۸/۷۹
		۱۲۶۹	۳۸۲	۶۴۸	۲۱۵	۶۵۶	۱۹۰
	نرم	۷۸/۸۸	۷۹/۸۳	۸۲/۹۴	۸۱/۱۳	۶۹/۷۱	۶۸/۹۳
		۴۰۵۲	۱۱۹۷	۱۹۴۳	۶۴۹	۱۸۸۸	۵۶۸
زنان	سخت	۷۹/۶۷	۷۷/۹۵	۸۲/۲۸	۸۱/۴۰	۶۹/۵۶	۶۶/۱۰
		۵۱۱	۱۷۱	۲۵۶	۹۲	۲۸۲	۹۵
	نرم	۸۰/۲۶	۷۸/۹۸	۸۲/۲۶	۸۱/۹۰	۶۹/۷۰	۶۶/۴۹
		۱۶۱۲	۵۴۱	۷۷۶	۲۸۰	۸۴۱	۲۸۸

جدول ۱۰ مقایسه نتایج بازشناسی‌های نرم و سخت با روش بازشناسی مدل مخلوطهای گاوسی

ردیف بالا: نرخ بازشناسی، ردیف پایین: زمان بازشناسی بر حسب ثانیه

	مجموعه	مصوتها		شبه مصوتها		صامتها	
جنسیت	نوع بازشناسی	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون
مردان	سخت	۷۷/۸۹	۷۸/۷۶	۸۲/۰۲	۸۱/۱۳	۷۰/۲۴	۶۸/۵۹
		۱۳۹۸	۴۱۶	۷۱۱	۲۲۹	۶۹۳	۲۰۴
	نرم	۷۹/۰۱	۸۰/۰۵	۸۲/۹۸	۸۱/۱۶	۶۹/۷۸	۶۹/۰۳
		۴۲۶۲	۱۲۹۸	۲۰۹۶	۷۰۰	۲۰۵۳	۶۱۰
زنان	سخت	۷۹/۶۳	۷۷/۷۹	۸۲/۳۸	۸۱/۶۲	۶۹/۶۶	۶۶/۴۰
		۵۶۲	۱۸۸	۲۷۸	۱۰۰	۲۹۹	۱۰۲
	نرم	۸۰/۲۴	۷۹/۳۵	۸۲/۳۸	۸۱/۶۲	۶۹/۷۰	۶۶/۷۵
		۱۷۷۰	۵۱۰	۸۳۶	۳۰۲	۸۹۲	۳۰۹

نکته قابل توجه در مقایسه نتایج مدل مارکوف پنهان و این روش بازشناسی آن است که زمان بازشناسی، به سیستم معادل مدل مارکوف پنهان نزدیک شده، اما بهبودی در نرخ بازشناسی در شرایط مختلف نسبت به مدل مارکوف پنهان دیده می‌شود که قابل اعتماد به نظر می‌رسد. نکته قابل توجه دیگر این است که در هیچ یک از مجموعه‌های آزمون روش بازشناسی سخت بهتر عمل نکرده است. به عبارت دیگر روش بازشناسی نرم به ازای افزایش زمان بازشناسی، قدرت تعمیم بهتری به مدل بخشیده است تا مدل بتواند با آموختن مجموعه آموزشی، توصیف خوبی را از مجموعه آزمون ارائه دهد.

لازم است ذکر شود که در این مقاله در روش بازشناسی کوانتیزاسیون برداری گاوسی، از روشهای انتخاب تابع گاوسی[1] برای تسریع محاسبه تابع درستنمایی و انتخاب

1. Gaussian Selection

بهترین توزیع گاوسی استفاده نشده است.[۱۹-۲۲]. به کارگیری این روشها می‌تواند ضمن استفاده از مزایای مدلسازی قطعه‌بندی نرم، الگوریتم بازشناسی واج‌ها را تا چند برابر سریعتر کند، در حالی که هیچ تغییری در نرخ بازشناسی به دست آمده برای کوانتیزاسیون برداری گاوسی ایجاد نمی‌شود.

در مقایسه دو روش بازشناسی کوانتیزاسیون برداری گاوسی و مدل مخلوطهای گاوسی، دیده می‌شود که در بازشناسی نرم نیز مدل مخلوطهای گاوسی در اغلب موارد اندکی بهتر رفتار کرده و در مقابل، افزایش کمی نیز در زمان بازشناسی ایجاد کرده است. این مسأله در مورد واج‌های صامت به خوبی صدق نمی‌کند. با توجه به کوتاه بودن طول واج‌های صامت، مدلسازی مدت زمان حضور در هر قطعه داخلی یک صامت معنای ارزشمندی ندارد. این مسأله احتمالاً، دلیل عدم انطباق مناسب این مدل بر این دسته از واج‌ها است.

٦- بهینه سازی میزان همپوشانی قطعات

مهمترین پارامتری که در مدلسازی قطعه‌بندی نرم حضور دارد، میزان همپوشانی قطعات مجاور است. در صورت افزایش این همپوشانی، اگرچه اثر تغییر تدریجی از هر قطعه به قطعه دیگر و همچنین کشش زمانی قطعات، بهتر مدل می‌شود، اما در مقابل از محلی بودن توزیع قطعات در موضع خاصی از الگوی مدلسازی شده می‌کاهد. به عبارت دیگر، هر چه میزان همپوشانی افزایش یابد، مدلسازی از توالی توزیع‌های زیر قطعات به توزیعی عمومی برای کل واحد زبانی نزدیک می‌شود. در نتیجه این پارامتر نیازمند بهینه سازی است.

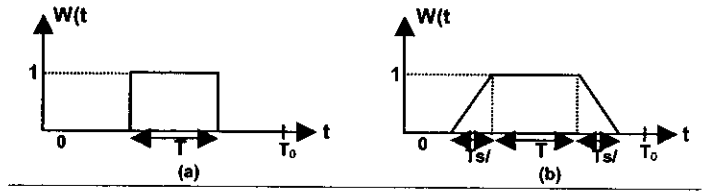
در فرایند بهینه سازی این پارامتر، با فرض خطی بودن همپوشانی قطعات و تقارن پنجره قطعات، میزان همپوشانی در مراحل آموزش و بازشناسی مصوتها و شبه مصوتها تغییر داده شده و نرخ بازشناسی در دو شیوه نرم و سخت با الگوریتم آموزش کوانتیزاسیون برداری گاوسی و بازشناسی مدل مخلوطهای گاوسی مقایسه شده است. میزان همپوشانی به صورت کمی زیر تعریف شده که نشان دهنده نسبت گستره همپوشانی قطعه نسبت به بخش مرکزی قطعه است:

$$OLC = \frac{L_1}{2L_2} \qquad (۱۰)$$

در این رابطه، L_1 طول منطقه قطعه با وزن غیر یک و L_2 طول منطقه قطعه با وزن یک است. شکل پنجره قطعه به ازای مقادیر OLC به برابر با صفر و ۰/۵در شکل ۳ نشان داده شده است. واضح است که همپوشانی صفر به قطعه‌بندی سخت می‌انجامد. همچنین در صورتی که همپوشانی از طول واحد زبانی از یک سو بیشتر باشد، فرض شده که عرض بخش همپوشانی در آن جهت به طول واحد زبانی محدود باشد به طوری که در مرز واحد زبانی وزن مورد استفاده به صفر برسد.

شبیه سازیهای انجام شده به ازای مقادیر مختلف همپوشانی قطعات، نشان دهنده مقدار بهینه همپوشانی در حدود ۰/۳ تا ۰/۷ است. در جدول ۱۱ مقادیر بهینه به دست آمده به ازای دسته‌های مختلف واجی و با دو روش بازشناسی نرم و سخت ارائه شده است. بدیهی است که در رویارویی با واج ناشناخته در مرحله بازشناسی، می‌توان از همپوشانی بهینه مدل هر دسته واج برای بازشناسی استفاده کرد. در مقادیر به دست آمده مقدار بهینه یکسان برای بازشناسی نرم و سخت دیده می‌شود. این مسأله در مورد صامتها مصداق ندارد که البته با طول کوتاه و ناکافی آنها برای مدلسازی کشش زمانی قابل توجیه است. نمونه رفتار نرخ بازشناسی به ازای مقادیر همپوشانی مختلف بازشناسی سخت و نرم در مورد شبه مصوتها به ازای مجموعه آزمون زنان در شکل ۴ آمده است. دیده می‌شود که رفتار بازشناسی نرم منحنی تک قله‌ای است؛ در حالی که نرخ بازشناسی در بازشناسی سخت به ازای مقادیر کمتر از حد بهینه تقریباً ثابت می‌ماند.

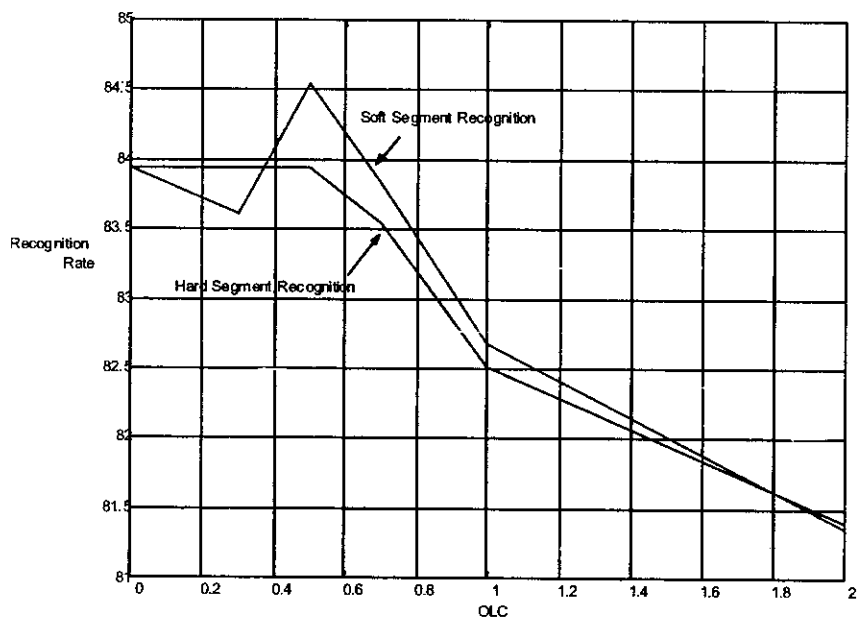
نکته قابل توجه دیگر، رفتار غیر یکسان بازشناسی



شکل ۳ مقایسه همپوشانی مختلف
(a)OLC=0 (b) OLC = 0.5

جدول ۱۱ مقدار بهینه همپوشانی در مدل قطعه‌بندی نرم

روش بازشناسی	مصوت	شبه مصوت	صامت
سخت	۰/۳	۰/۵	۰/۷
نرم	۰/۳	۰/۵	۰/۳



شکل ۴ نمودار تغییرات نرخ بازشناسی نسبت به میزان همپوشانی قطعات

واج‌های مختلف به ازای یک سیستم بازشناسی است. به نظر می‌رسد که واج‌هایی که توسط یک توزیع در کل واحد زبانی قابل توصیف هستند، همپوشانی بیشتر را ترجیح می‌دهند در حالی که واج‌هایی که در طول بیان از توزیعهای آماری مختلفی پیروی می‌کنند، به همپوشانی کمتری تمایل دارند. به طور کلی نمی‌توان روندی کاملاً نزولی یا صعودی را به ازای هر واج مشاهده کرد و این، راه را برای ارائه توجیهی ساده و همگانی مسدود می‌کند.

۷- نتیجه گیری

در این مقاله، روشی جدید برای مدلسازی کشش زمانی هر حالت واحد اکوستیکی گفتاری ارائه و آزموده شد. مدل ارائه شده به عنوان حالت خاص مدل قطعه‌ای اتفاقی قابل بررسی است، با این تفاوت که از مدلی معین و غیر آماری برای اتصال حالات واحد اکوستیکی استفاده می‌کند. از نظر دقت مدلسازی، مزیت اصلی این مدل، تحمل تبدیل تدریجی یک حالت به حالات مجاور می‌باشد در حالی که تقریباً در هیچیک از سایر مدلهای چنین رفتاری مورد توجه واقع نشده است یا اینکه پارامترهای زیاد مدل، مانع کارایی مدل شده است. از لحاظ سیستم بازشناسی گفتار، مزیت عمده آن، سرعت بالای بازشناسی و طبقه‌بندی واحدهای زبانی به علت عدم نیاز به استفاده از جستجوی پویا در یافتن مرزبندی بهینه در هر واحد گفتاری است.

استفاده مستقیم از قطعه‌بندی نرم در هنگام بازشناسی در اکثر مواقع، موجب افزایش جزئی در نرخ بازشناسی می‌شود که با انطباق بهینه سازی در مرحله آموزش و بازشناسی قابل توجیه است، اما در مقابل افزایش زمان بازشناسی نرم متناسب با میزان همپوشانی مشکلی محسوب می‌شود که باید در هنگام استفاده از این روش در کاربردهای بی‌درنگ مورد توجه قرار گیرد. بهینه سازی میزان همپوشانی برای حداکثر کردن نرخ بازشناسی، نشان می‌دهد که این روش در میزان همپوشانی نزدیک به ۳۰٪ به بالاترین نرخ بازشناسی می‌رسد، در حالی که زمان بازشناسی نیز نسبت به بازشناسی سخت تقریباً دو برابر شده است.

به نظر می‌رسد که روش قطعه‌بندی نرم در واج‌های مصوت و شبه مصوت بهتر عمل کرده است که اینها اتفاقاً واج‌هایی هستند که طول زمانی آنها بیشتر به مدلسازی نیاز دارد. در واج‌های صامت، با توجه به کوتاهی طول و تغییرات اندک آن، مدلسازی زمانی از ارزش کمتری برخوردار است.

مسائل باقیمانده در این مدلسازی، شکل پنجره مورد استفاده و تأثیر روشهای پالایش داده‌ها[۲۳] بر قطعه‌بندی نرم است که پژوهش در این زمینه‌ها ادامه دارد.

۸- منابع

[1] L.R. Rabiner; "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition"; Proceedings of The IEEE; Vol. 77, No. 2, Feb. 1989; PP. 257-286.

[2] M. Ostendorf; V. Digalakis; O. Kimball; "From HMM to Segment Models: A Unified View of Stochastic Modeling of Speech Recognition"; IEEE Transactions on Speech and Audio Processing; Vol. 4, No. 5, Sep. 1996; PP. 360-378.

[3] M.Russell; R.Moore; "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition"; Proceedings of IEEE ICASSP; 1985; PP.2376-2379.

[4] C. D. Mitchel; M. P. Harper; L. H. Jamieson; "Using Explicit Segmentation to Improve HMM Phone Recognition"; Proceedings of ICASSP; 1995; PP. 229-232.

[5] A. Anastasakos; R. Schwartz; H. Shu; "Duration Modeling in Large Vocabulary Speech Recognition"; Proceedings of ICASSP; 1995; PP. 628-631.

[6] D. Burshtein; "Robust Parametric Modeling of Durations in Hidden Markov Models"; IEEE Transactions on Speech and Audio Processing; Vol. 4, No. 3, May 1996; PP. 240-242.

[7] H. C. Leung; V.W. Zue; "Phonetic Classification Using Multi-Layer Perceptrons"; Proceedings of ICASSP; 1990; PP. 525-528.

[8] S. Zahorian; P. Silsbee; X. Wang "Phone Classification With Segmental Features

Interface for Weather Information"; IEEE Trans. Acous. Speech and Audio P rocess; Jan 2000; PP. 85-96.

[17] J. W. Chang; "Near Miss Modeling, A Segmental Based Approach to Speech Recognition"; PhD Thesis, MIT; June 1998.

[18] Razzazi F.; Sayyadian A.; "Soft Segment Modeling, A New Approach in Duration Modeling for Speech Recognition"; Proceedings of 9th Australian Conference of Speech Science and Technology, Melbourne; Dec. 2002.

[19] Knill K.M.; Gales M.J.F.; Young S.J.; "Use of Gaussian Selection in Large Vocabulary Continuous Speech Recognition Using HMMs"; Proceedings of ICSLP; 1996.

[20] Lee A.; Kawahara T.; Shikano K.; "Gaussian Mixture Selection Using Context-Independent HMM"; Proceedings of ICASSP; 2001.

[21] Lee A.; Kawahara T.; Shikano K.; "Gaussian Mixture Selection Using Context-Independent HMM"; Information Processing Society of Japan Journal; Aug 2002.

[22] Paul D.; "An Investigation of Gaussian Shortlists"; Proceeding of IEEE Workshop on Automatic Speech Recognition and Understanding; Sep. 1999.

[23] Razzazi F.; Sayyadian A.; "Data Refining HMM, A New Approach to HMM Based Speech Recognition System Improvement"; Proceedings of Forum Acusticum, Sevilla ; Aug 2002.

and a Binary-Pair Partitioned Neural Network Classifier"; Proceedings of ICASSP; 1997; PP. 1011-1014.

[9] S. Renals; R. Ruhwer; "Learning Phoneme Recognition Using Neural Networks"; Proceedings of ICASSP; 1989; PP. 413-416.

[10] V. Zue; J. Glass; M. Philips; S. Seneff; "Acoustic Segmentation and Phonetic Classification in the SUMMIT System"; ICASSP; 1989; PP 389-392.

[11] V. Zue; J. Glass; D. Goodine; M. Philips; S. Seneff; "The Summit Speech Recognition System: Phonological Modeling and Lexical Access"; ICASSP; 1990; PP 49-52.

[12] N. Strom; L. Hetherington; T.J. Hazen; E. Sandness; J. Glass; "Acoustic Modeling Improvements in a Segment-Based Speech Recognizer"; 1999.

[13] V. Zue; J. Glass; D. Goodine; H. Leung ; M. Philips; J. Polifroni; S. Seneff; "The Voyager Speech Understanding System: Preliminary Development and Evaluation"; ICASSP; 1990; PP. 73-76.

[14] V. Zue; J. Glass; D. Goodine; H. Leung; M. Philips; J. Polifroni; S. Seneff; "Integration of Speech Recognition and Natural Language Processing in the MIT Voyager System"; ICASSP; 1991; PP. 713-716.

[15] J. Glass; T. Hazen; L. Hetherington; "Real Time Telephone Based Speech Recognition in the JUPITER Domain"; 1999; PP. 61-64.

[16] V. Zue; S. Seneff; J. Glass; J. Polifroni; C. Pao; T. Hazen; L. Hetherington; "Jupiter: A Telephone-Based Conversational