عنوان زیرپروژه:

تحلیل ابعاد دادگان گفتاری و امکان سنجی تهیه موتورهای بازشناسی گفتار زبان فارسی

فهرست مطالب

عنوان **شماره صفحه**

1 مقدمه

امروزه کاربردهای پردازش گفتار فراگیرتر شده اند. نیاز به داشتن سیستمهای اتوماتیک که همانند انسان با گفتار ارتباط داشته باشند روز به روز محسوس تر می گردد. از شاخه های پردازش گفتار می توان به کدینگ، بهبود کیفیت گفتار، بازشناسی گفتار (گسسته یا پیوسته)، بازشناسی گوینده (تعیین و تایید گوینده) اشاره کرد. برای آموزش سیستمهای گفتاری نیاز به داشتن دادگان آموزشی می باشد که با توجه به کاربرد بایستی دارای ویژگیهای خاصی باشند. در این ارتباط برای زبان انگلیسی دادگان مناسبی از قبیل TIMIT و Aurora طراحی شده است. برای زبان فارسی در سالهای اخیر گامهای مطلوبی برداشته شده است اما هنوز خلا داشتن یک پایگاه داده مناسب، استاندارد و قابل اطمینان حس می شود. برای طراحی دادگان مناسب جهت هر کاربرد پردازش گفتار بایستی نیاز آن کاربرد را برای دادگان آموزشی به خوبی مطالعه کنیم. در این مجموعه تلاش شده تا با مطالعه اجمالی روی برخی از کاربردهای پردازش گفتار به شاخص های مهمی که باید در طراحی پایگاه داده استاندارد دیده شود، اشاره شود. به همین منظور چارچوب گزارش بفرم زیر تنظیم گردیده است:

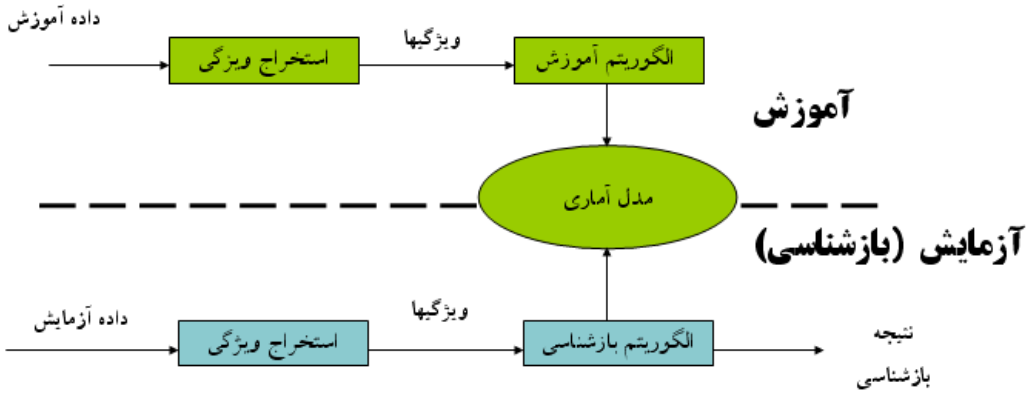
در فصل دوم مقدمه ای بر سیستم تشخیص گفتار خواهیم داشت. سپس در ادامه به مباحث سیگنالینگ، پیش پردازش، نمایش و خلاصه سازی اطلاعات صوتی خواهیم پرداخت. فصل چهارم پرکاربردترین الگوریتم یادگیری و مدل کردن در پردازش گفتار را معرفی می کند، در واقع به معرفی مدل مخفی مارکوف و کاربرد آن در بازشناسی گفتار می پردازد. ملاحظات عملی در استفاده از HMM ها نیز در فصل پنجم آورده شده است. روشهای جستجو، مؤلفه‌های فضای جستجو در سیستمهای بازشناسی گفتار و الگوریتم‌های جستجو برای بازشناسی گفتار نیز مباحث دیگری هستند که در فصول هفتم تا نهم بررسی شده اند. مقوله دیگری در پردازش گفتار، بازشناسی گفتار در محیطهای نویزی است که تحت عنوان مقاوم سازی بازشناسی گفتار در فصل دهم نگاه اجمالی شده است. تطبیق گوینده در سیستم های بازشناسی گوینده در فصل یازدهم آورده شده است. تشخیص هویت گوینده نیز بعنوان کاربرد دیگری از پردازش گفتار در فصل یازدهم مورد توجه قرار گرفته است. فصل دوازدهم حاوی اطلاعاتی در مورد سنتز گفتار می باشد. دادگان های گفتاری مهمترین بخش این گزارش می باشد که به شاخصهای دادگان گفتاری می پردازد، برای مشاهده این مقال به بخش سیزدهم مراجعه کنید. در انتها نیز جمع بندی کلی آورده شده است.

2 مقدمه ای بر سیستم تشخیص گفتار

2-1 مقدمه

سیگنال گفتار را می‌توان به صورت یک پوش طیفی که به آرامی تغییر می‌یابد، در نظر گرفت. این پوش طیفی بوسیله انسان دریافت می‌گردد و به دنباله‌ای از کلمات و معانی آنها ترجمه می‌گردد. سیستم‌های بازشناسی خودکار گفتار نیز بطور مشابه تلاش می‌کنند که این پوش طیفی را به دنباله‌ای از کلمات تبدیل نمایند. مشکلات فراوانی مانند تغییرات پوش طیفی گفتار در این راه وجود دارد. تغییرات پوش طیفی گفتار به دلایلی نظیر جنسیت گوینده، نحوه بیان و تأکید در گفتار گوینده و نیز تغییر محیط آکوستیکی که سیستم بازشناسی در آن عمل می‌کند، بوجود می‌آید. طراحی سیستم بازشناسی خودکاری که بتواند با توانایی انسان در مقابله با این تغییرات برابری نماید، هنوز یک مسأله و چالش اساسی محسوب می‌شود. بسیاری از سیستم‌های بازشناسی گفتار از روش‌های آماری برای مقابله با دسته‌ای از تغییرات پوش طیفی استفاده می‌کنند. کارایی این سیستم‌های آماری تا به امروز به طور قابل توجهی افزایش یافته است، به طوریکه برای یک کتاب لغت نامحدود مستقل از گوینده صحت بازشناسی بیش از ۹۰% بدست آمده است. به علاوه، برای کتاب لغت‌هایی با اندازه محدود، صحت بازشناسی بیش از ۹۵% نیز قابل دسترسی است. با این درصد کارایی، این سیستم‌ها قابل بهره‌برداری به نظر می‌رسند، ولی باید توجه کرد که اکثر این سیستمها در محیط‌هایی یکسان و ساکت (بدون حضور نویز) مورد آموزش و آزمایش قرار گرفته‌اند. این در حالی است که در شرایط عملی، سکوت به ندرت وجود دارد و محیط‌های آکوستیکیِ آزمایش و آموزش نیز عموماً با یکدیگر متفاوتند.

در یک سیستم بازشناسی گفتار، پس از تقسیم سیگنال‌ها و نمونه‌های گفتاری به دو دسته دادگان آموزش و آزمایش، سیگنال ورودی به فواصل زمانی ۲۵ تا ۳۰ میلی ثانیه (با در نظر گرفتن درصدی همپوشانی) قاب‌بندی می‌شود. سپس در مرحله استخراج ویژگی، از هر قاب ویژگی‌های گفتاری (نظیر ضرایب مل کپستروم یا ضرایب پیشگویی خطی ادراکی) استخراج می‌گردند. در مرحله آموزش، از این ویژگی‌ها برای آموزش مدل‌های آماری بازشناسی نظیر مدل مخفی مارکف (HMM) و شبکه عصبی مصنوعی (ANN) استفاده می‌شود. در مرحله آزمایش یا بازشناسی، ویژگی‌ها از طریق یک الگوریتم بازشناسی نظیر ویتربی با این مدل آماری مقایسه می‌شوند. این مراحل و پردازش‌ها در شکل ۱ قابل مشاهده‌اند.



شکل ۱ **بخش‌ها و مراحل پردازش در یک سیستم بازشناسی گفتار**

یکی از مراحل اصلی در روند بازشناسی گفتار فوق، مرحله استخراج ویژگی و تولید پارامترهایی از سیگنال گفتار است که علاوه بر کاهش حجم دادهٔ ورودی به بازشناس گفتار، خصوصیات برجسته‌ای را تعیین کند که واحدهای مختلف گفتاری نظیر واج‌ها، هجاها یا کلمات را از یکدیگر متمایز نماید. در این راستا اکثر سیستم‌های بازشناسی از ویژگی‌های مل کپستروم استفاده می‌نمایند که نحوه عملکرد گوش و سیستم شنوایی انسان را شبیه‌سازی می‌کند. با این وجود، بکارگیری خصوصیات شنوایی و سیستم تولید گفتار در انسان، لزوماً راه‌حل بهینه‌ای برای استخراج ویژگی در بازشناسی گفتار نیست. زیرا این ویژگی‌ها برای ایجاد تمایز میان واحدهای گفتاری بهینه نیستند و علاوه بر این، نیازمند روش‌های تکمیلی برای هنجارسازی گویندگان و مقابله با نویز محیط هستند.

برای رفع این نقیصه و ایجاد تمایز میان واحدهای گفتاری و همچنین برای مقابله با نویز راه‌حل‌های متعددی پیشنهاد شده‌اند. برای مقابله با نویز سه روش کلی وجود دارد:

۱. حذف نویز از سیگنال گفتار با روش‌هایی نظیر تفاضل طیف و بکاربردن فیلترهای حوزه زمان یا فرکانس نظیر فیلتر وینر.

۲. جبران تأثیر نویز بر ویژگی‌ها که با اعمال تبدیل بر ویژگی‌ها سعی در حذف نویز از ویژگی‌ها دارد. ازجمله روش‌های شناخته‌شده این دسته می‌توان به تفاضل میانگین ضرایب کپسترال و یا هنجارسازی ضرایب کپسترال بر حسب میانگین و واریانس اشاره نمود. روش‌های SDCN، SPLICE نیز در این گروه جای می‌گیرند.

۳. روش‌های تطبیق مدل تلاش می‌کنند که بجای اصلاح سیگنال یا پارامترها، مدل آکوستیک محیط یا مدل گوینده را در مرحله بازشناسی اصلاح نمایند. مزیت این روش‌ها آن است که در آنها داده‌های مشاهده شده تغییر نمی‌کنند و هیچ نوع فرض یا تصمیم‌گیری قبلی درباره سیگنال گفتار ضروری نیست. برخی از روش‌های مبتنی بر مدل عبارتند از: تصویر وزن‌دار، ترکیب موازی مدل‌ها (PMC) و بازگشت خطی با بیشترین درستنمایی (MLLR). روش MLLR در مسائل مقاوم سازی نسبت به گوینده و به عبارتی تطبیق گوینده بیشتر بکار گرفته شده است.

2-2 پارامترهای بازشناسی گفتار

پارامترهای مختلفی در یک سیستم بازشناسی گفتار موثر هستند. این پارامترها ،تعیین کننده درجه پیچیدگی سیستم می‌باشند. این پارامترها عبارتند از: وابسته و مستقل بودن از گوینده، بازشناسی کلمات مجزا و گفتار پیوسته، اندازه کتاب لغت، محدودیت‌های زبانی، گفتار مکالمه‌ای و شرایط محیطی که بازشناسی در آن انجام می‌گیرد. در این زیربخش این پارامترها به اختصار مورد بررسی قرار می‌گیرند.

۲-۲-۱ وابسته یا مستقل از گوینده

یک سیستم وابسته به گوینده در تعریف فقط برای استفاده یک گوینده طراحی می شود، درحالیکه یک سیستم مستقل از گوینده برای استفاده هر گوینده‌ای طراحی می‌گردد. بطور معمول سیستم‌های وابسته به گوینده دقیق‌تر از یک سیستم مستقل از گوینده هستند و نتایج بهتری را ارائه می‌دهند. ایراد عمده سیستم وابسته به گوینده این است که هر بار برای بازشناسی گوینده جدید نیاز به آموزش دارد . سیستم میانه سیستم‌های وابسته و مستقل از گوینده، سیستم چند گوینده است که برای تعداد گوینده‌های ثابت و کم بکار می‌رود.

۲-۲-۲ گفتار مجزا/ متصل/ پیوسته

در بازشناسی کلمات مجزا[1]، هر کلمه بصورت جداگانه و واضح بیان می شود و سیستم بازشناسی با هر کلمه بطور مستقل سرکار دارد. در بازشناسی کلمات متصل[2]، دنباله‌ای از کلمات برای بازشناسی مورد توجه قرار می گیرند، ولی کلمات جمله باید بطور مجزا و با فواصل زمانی سکوت از هم جدا شوند. در بازشناسی گفتار پیوسته[3]، کلمات با مکثهای از پیش تعریف شده از یکدیگر جدا نمی شوند و تلفظ لغات نیز تحت تأثیر آثار هم ادایی[4] قرار می‌گیرد. بنابراین واضح است که نسبت به دو مورد قبل مشکل‌تر است.

۲-۲-۳ اندازه کتاب لغت

تعداد کلمات موجود در کتاب لغت عامل مهمی در تشخیص کارایی یک سیستم بازشناس گفتار است. در سیستم‌های کتاب لغت کوچک (کمتر از ۱۰۰ لغت) معمولاً می‌توان به دقتی در حدود ۱۰۰% (حتی در سیستم‌های مستقل از گوینده) دست یافت. سیستم‌های با دایره لغات کوچک در کاربردهایی نظیر تشخیص کارت اعتباری و تشخیص شماره تلفن کاربرد دارند. به هر حال دقت بازشناس به لغات موجود در کتاب لغت نیز وابسته می‌باشد. اگر کلمات مشابه و گیج‌کننده باشند، رسیدن به دقت ۱۰۰% حتی برای کتاب لغت‌های بسیار کوچک، دشوار است.

۲-۲-۴ محدودیت‌های زبانی[5]

محدودیت‌های زبانی را می‌توان با یک مدل از زبان بیان نمود. مدل زبانی یک زبان طبیعی مرکب از چهار جزء می‌باشد. نمادها، دستور زبان[6]، معنا[7] و جنبه عملی[8]. نمادهای زبان واحدهای طبیعی هستند که همه پیغام‌ها از آنهـــا تشکیل می‌گردند و بیانگر کلمات یا واحدهای کوچکتر از کلمه، نظیر هجاها و واجها هستند. دستور زبان، مرکب از محدودیت‌های واژگانی[9] و نحوی[10] است که بیانگر شکل گرفتن کلمات از

---

[1] Isolated Word Recognition
[2] Connected Word Recognition
[3] Continues Speech Recognition
[4] Co-Articulation
[5] Linguistic constraints
[6] Grammar
[7] Semantic
[8] Pragmatic
[9] Lexical

واحدهای کوچکتر از کلمه و نیز شکل گرفتن جملات از کلمات می‌باشند. جنبه معنایی نیز مرتبط به نحوه ترکیب کلمات برای شکل دادن پیغام‌های با معنی است. به عنوان مثال جمله «اسب صحبت می کند» از لحاظ نحوی صحیح ولی از لحاظ معنایی نادرست است. در بالاترین سطح، جنبه عملی یک زبان جای دارد که بیانگر وابستگی ادا کردن و معنی کلمه به گوینده‌ها و محیط است. محدودیت‌های معنایی و جنبه عملی به ندرت در سیستم‌های بازشناسی گفتار استفاده می‌شوند. چرا که این محدودیت‌ها را به دشواری می‌توان به صورت فرمول بیان کرد. ولی محدودیت‌های دستوری تقریباً در تمامی سیستم‌های بازشناسی گفتار پیوسته به صورت محدودیت‌های واژگانی و نحوی مورد استفاده قرار می‌گیرند و تعداد جملات مجاز برای بازشناسی را کاهش می‌دهند.

۲-۲-۵ گفتار مکالمه‌ای[11]

گفتار بطور طبیعی به شکلی فوری و مکالمه‌ای بیان می‌شود که تشخیص آن بوسیله ماشین‌ها بسیار دشوار است. در چنین گفتاری، جملات ناقص، شروع‌های مجدد، خنده‌های بلند و سرفه کردن وجود دارد که گفتار را از حالت روان بودن خارج نموده و کتاب لغت را عملاً نامحدود می‌سازد.

۲-۲-۶ محیط[12]

محیطی که سیستم بازشناسی در آن عمل می‌کند، بر کارایی بازشناسی مؤثر است. شرایط نامناسبی همچون نویز محیط، ضعف میکروفن و اثرات کانال انتقال ممکن است کارایی را به میزان قابل توجهی کاهش دهد. در حالت کلی چنانچه یک سیستم بازشناسی برای یک محیط عاری از نویز طراحی شده باشد، بکار بردن آن، در شرایطی نامناسب و نویزی، بدون انجام اصلاحات، کارایی را به شدت کاهش خواهد داد.

2-3 اجزای یک سیستم بازشناسی

یک سیستم بازشناسی گفتار شامل اجزای مختلفی می‌شود. در ادامه این فصل به بررسی اجزای یک سیستم بازشناسی گفتار، از مرحله نمونه‌برداری تا مرحله پردازش زبان می‌پردازیم. بسته به نوع

---

[10] Syntactic
[11] Spontaneous speech
[12] Environment

بازشناسی برای گفتار مجزا، متصل یا پیوسته، ممکن است برخی از این اجزاء در سیستم موجود نباشند یا با جزئیات بیشتری در آن سیستم درنظر گرفته شده باشند.

۲-۳-۱ نمونه برداری از سیگنال صوتی

برای استفاده از الگوریتم‌های پردازش سیگنال گسسته، باید موج پیوسته سیگنال گفتار ورودی را به شکل گسسته تبدیل نمود. به این منظور باید از موج گفتار ورودی نمونه‌برداری کرد. بنابر قضیه نایکوئیست، فرکانس نمونه برداری موج پیوسته باید حداقل دو برابر بزرگترین مؤلفه فرکانسی موجود در موج پیوسته باشد. در کاربردهای عملی، موج گفتار پیوسته عموماً از خروجی یک میکروفن و یا از خط تلفن دریافت می‌شود. بنابراین در بسیاری موارد می‌توان پهنای باند خطوط انتقال تلفن را به عنوان معیاری برای تعیین فرکانس نمونه‌برداری در نظر گرفت.

یکی دیگر از مسائل مهم نمونه‌برداری از سیگنال پیوسته که دقت نمونه برداری را تعیین می‌کند، تعداد بیت‌های مورد استفاده در هر نمونه است. در مرحله نمونه برداری، اندازه هر نمونه به یکی از L سطح ممکن نسبت داده می‌شود. این امر سبب اضافه شدن خطا به اطلاعات می‌شود که به آن خطای چندی‌کردن اطلاق می‌گردد.

۲-۳-۲ استخراج ویژگی از سیگنال گفتار

پس از قطعه‌بندی به اجزای سازنده گفتار، باید ویژگی‌هایی را از سیگنال گفتار استخراج نمود تا از آنها در مرحله تطبیق الگو استفاده شود. سیگنال گفتار ویژگی‌های زیادی دارد که عموماً با طیف لحظه‌ای سیگنال گفتار یا شکل مجرای گفتار و... مرتبط می‌باشند. پردازش این همه ویژگی برای کاربردی بخصوص، همانند بازشناسی گفتار، کاری منطقی و عملی نخواهد بود. بدین منظور تبدیل‌هایی روی سیگنال گفتار انجام می‌شود تا بتوان ویژگی یا ویژگی‌های مفید را استخراج نمود. استخراج ویژگی به دو دلیل انجام می‌گیرد. اول آنکه سبب تمرکز روی اطلاعات موجود در سیگنال می‌شود و این امر منجر به بهبود میزان شباهت و عدم شباهت میان کلاس‌های مختلف می‌شود. ثانیاً داده‌ها را به نحو قابل ملاحظه‌ای کاهش داده، محاسبات به میزان زیادی کم می‌شود.

به منظور استخراج بردارهای ویژگی باید یک سری پردازش‌ها روی سیگنال انجام شود. این پردازش‌ها

عبارتند از: قاب‌بندی[13]، پیش‌تأکید کردن[14]، اعمال پنجره، تبدیل فوریه زمان کوتاه و....

۲-۳-۳ تطبیق الگو

برای ایجاد بانکِ قواره‌های مرجع در سیستمی که برای بازشناسی گفتار یک گوینده تعلیم می‌بینـــــد، از هر الگوی تلفظ شده توسط آن گوینده، یک یا چند قواره بدست آمده و ذخیره می‌شود. الگوی ذخیره شده می‌تواند یک مدل آماری بیانگر مشخصات آن الگو باشد. در هر حال همواره لازم است که میان مجموعه بردارهای ورودی که هر یک حاوی اطلاعات طیفی اخذ شده از بخشی کوتاه از سیگنال گفتار ورودی هستند و الگوی مرجع، یک تطبیق الگو و مسیریابی زمانی[15] انجام گردد تا بتوان تغییرات سرعت بیان عبارت را نیز در سیستم منظور کرد. به این ترتیب امکان محاسبـــه یک معیار شباهت مابین کلمه ورودی و قواره‌های ذخیره شده در سیستم فراهم می‌گردد.

در بعضی از سیستم‌های بازشناسی گفتار بر مبنای کلمه، تطبیق ورودی با قواره‌های مرجع با استفاده از یک روش بهینه‌سازی مشهور به برنامه ریزی پویا[16] انجام می‌گردد. حل مسأله مسیریابی زمانی با روش برنامه‌ریزی پویا به پیچش زمانی پویا[17] با علامت اختصاری DTW معروف است. در این روش برای هر کلمه یا عبارت، یک قواره در حافظه نگهداری می‌شود و هنگام تشخیص الگوی ورودی، میزان انطباق آن با تمامی قواره‌ها بررسی می‌گردد. این کار برای تعداد لغات زیاد، نیازمند حجم حافظه و محاسبات بسیار زیادی است. از طرف دیگر، این روش در بازشناسی گفتار پیوسته کارایی خود را از دست می‌دهد، چرا که مرز دقیق کلمات مشخص نیست. به علاوه این روش برای گویندگان متعدد نیز دارای کارایی نیست، زیرا نمی‌تواند تغییرات آکوستیک بین گوینده‌های مختلف را بخوبی مدل کند.

رویکرد دیگر آن است که از هر الگوی لغت‌نامه، یک مدل آماری ساخته شده و هر الگوی لغت‌نامه که مدل آماریِ آن بیشترین میزان شباهت (احتمال وقوع) را به الگوی ورودی مشاهده داشته باشد، به عنوان الگوی ورودی بازشناخته گردد. معمول‌ترین مدل مورد استفاده دراین مورد، مدل مخفی مارکوف[18] نام دارد. در مدل مخفی مارکوف، هر واحد از لغت‌نامه توسط مجموعه‌ای از حالت‌ها به همراه احتمالات انتقال از حالتی به حالت دیگر نمایش داده می‌شود. در بسیاری از سیستم‌های بازشناسی مبتنی بر مدل مخفی مارکوف، مدل بر پایه واج‌ها تولید شده و سپس در مرحله نهایی بازشناسی، اطلاعات مربوط به

[13] Framing
[14] Pre-Emphasizing
[15] Time Alignment
[16] Dynamic Programming
[17] Dynamic Time Wrapping
[18] Hidden Markov Model

واژگان با اطلاعات واجی ترکیب می‌گردد تا کلمات مورد شناسایی قرار گیرند. مرحله آموزش مدل مخفی مارکوف نسبت به DTW پیچیده‌تر است و به داده‌های آموزشی بیشتری نیاز دارد.

یکی دیگر از شیوه‌های تطبیق الگو استفاده از شبکه‌های عصبی مصنوعی است. شبکه‌های عصبی امکانات خوبی را در جهت پردازش‌های موازی و تطبیق‌یابی در اختیار قرار می‌دهند. در بعضی از تحقیقات انجام شده، روش‌های شبکه عصبی با روش‌های دیگری نظیر مدل مخفی مارکوف ترکیب شده‌اند. با توجه به اینکه سیستم تشخیص مورد استفاده در این پایان‌نامه از مدل مخفی مارکوف استفاده می‌کند، این مدل به طور جداگانه، به اختصار شرح داده می‌شود.

۲-۳-۴ پردازش زبان

صرفنظر از آنکه واحد پایه تشخیص گفتار چیست (کلمه، هجا، آوا یا واج)، برای تعیین چگونگی ادغام این واحدها از نظر ترتیب، متن و معنا، از محدودیت‌های زبان استفاده می‌شود. چنانکه قبلاً گفته شد اجزای مدل یک زبان عبارتند از: نمادها، دستور، معنا و جنبه عملی. این اجزاء به عنوان محدودیت‌های زبان استفاده می‌شوند تا در بازشناسی گفتار، تمامی اطلاعات موجود در ارتباط گفتاری در نظر گرفته‌شود.

4-2 انواع مدل سازی در ASR

بیشتر سیستمهای ASR شامل سه قسمت عمده می باشند:

- signal processing front-end

این واحد سیگنال گفتار را به دنباله ای از بردارهای ویژگی تبدیل می کند. استخراج بردارهای ویژگی به منظور کلاس بندی انجام می گیرد. و موجب کاهش نرخ اطلاعات برای پردازش نسبت به سیگنال اصلی می شود.

- acoustic modeling
- language modeling

سیستم های بازشناسی یا تشخیص خودکار گفتار برای تبدیل گفتار به متن مورد استفاده قرار می‌گیرند. یک سیستم بازشناسی گفتار قابل تجزیه به بلوکهای مجزای عملیاتی است و به هر بلوک می‌توان یک مجموعه ورودی و یک مجموعه خروجی نسبت داد. ازجمله می توان به موارد زیر اشاره نمود: پردازش سیگنال، ایجاد و مدیریت پایگاه داده گفتار، محاسبات آماری و حسابی، ارزیابی و پیاده‌سازی انواع گرامرها، الگوریتمهای جستجوی کارا، حذف نویز، بهبود گفتار و غیره که باید به نحو مطلوب پیاده سازی گردند.

روشهای مختلفی برای بازشناسی گفتار وجود دارد. الگوریتمهای اصلی بازشناسی گفتار بر پایه روشهای زیر می‌باشند:

۱- مدل مخفی مارکوف[19]. ۲- پیچش زمانی پویا[20]. ۳- شبکه عصبی[21].

سیستمهای مبتنی بر مدل مخفی مارکوف موفقیت خوبی در پردازش گفتار دارند چرا که روش مدل سازی مارکوف، در عین اینکه دقت بالای بازشناسی را حفظ می‌کند قابلیت مقاوم بودن را برای سیگنالهای گفتار فراهم می‌کند. مدلهای مخفی مارکوف را بر اساس تابع چگالی آنها به دو دسته مدل پیوسته و مدل گسسته تقسیم می‌کنند که مدل پیوسته دقت بازشناسی بهتری نسبت به مدل گسسته[22]، دارد.

در سیستم‌های بازشناسی گفتار که بر اساس مدل مخفی مارکوف می‌باشند، دو قسمت بیشترین زمان بازشناسی را به خود اختصاص می‌دهند: یکی «عملیات جستجو» است که بهترین کلمه منطبق را از روی مقایسه گفتار ورودی با مدلهای گفتاری مرجع پیدا می‌کند و دیگری قسمت مربوط به «محاسبه احتمال خروجی» در مدل پیوسته مارکوف می‌باشد. معمولاً از الگوریتم «ویتربی» برای انجام عملیات جستجو در سیستم‌های بازشناسی گفتار استفاده می‌شود.

در سیستم ASR، بازشناسی یعنی جستجو و برگرداندن دنباله حالتی که با دنباله مشاهده داده شده بیشترین مطابقت را داشته باشد. دقت بازشناسی به تعداد موارد آموزشی مدل و ظرفیت مدل برای مشاهده در مرحله آموزش بستگی دارد.

هر چند روشهایی مثل برنامه‌ریزی پویا و شبکه‌های عصبی نیز در زمینه بازشناسی گفتار فراوان استفاده شده اند ولی تکنولوژی ASR مبتنی بر مدل مخفی مارکوف استفاده بیشتری داشته ومدل مخفی مارکوف یکی از مدلهای آماری بسیار قوی در توصیف رفتار پیچیده فونهای گفتاری است.

۲-۴-۱ مدل مخفی مارکوف

مدل مخفی مارکوف ابزاری بسیار قدرتمند برای مدل کردن فرآیندهای تصادفی می‌باشد و کاربردهای زیادی در پردازش گفتار دارد. مدل مخفی مارکوف یک فرایند تصادفی دولایه است که در آن فرآیند تصادفی اصلی پنهان است و با استفاده از مجموعه‌ای دیگر از فرآیندهای تصادفی که سری مشاهدات را

[19] hidden Markov model (HHM)
[20] dynamic time wrapping (DTW)
[21] neural network
[22] discrete hidden Markov model (DHMM)

ایجاد می‌نمایند، آشکار می‌شود.

عناصر لازم برای تعریف مدل مخفی مارکوف، عبارتند از:

N، تعداد حالات

M، تعداد نمادهای مشاهده شده در هر حالت

A، ماتریس احتمالات انتقال بین حالت i به j (a_{ij})

B، توزیع احتمال مشاهده نماد k در حالت j $(b_j(k))$

π، توزیع احتمالات اولیه برای حالت i (π_i)

اگر سیگنال به یک دنباله $X = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}$ از بردارهای ویژگی‌ها تبدیل شود، مدل مخفی مارکوف با یک روال به صورت زیر بردارهای ویژگی را تولید می‌کند.

در زمان t=1 مدل با احتمال π_{q_1} در حالت q_1 قرار می‌گیرد.

اگر در زمان t مدل در حالت q_t باشد، در این حالت با احتمال $b_{q_t}(x_t)$ بردار ویژگی x_t را تولید خواهد کرد.

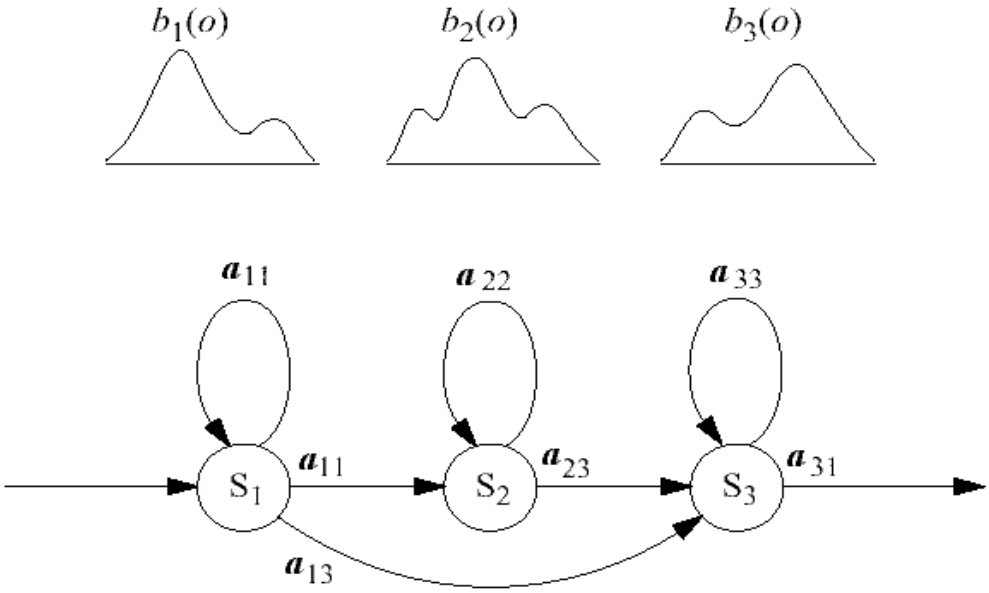
اگر در زمان t در حالت q_t باشد، در این حالت با احتمال $a_{q_t q_{t+1}}$ به حالت q_{t+1} در زمان t+1 خواهد رفت.

با توجه به روال فوق، احتمال تولید رشته بردار X توسط مدل، با فرض طی کردن رشته حالات $Q = \{q_1, q_2, \ldots, q_T\}$ از رابطه (۱) بدست می‌آید:

$$P(X \mid Q, l) = p_{q1} b_{q1}(x_1) a_{q1q2} \ldots a_{q_t q_{t+1}} b_{q_{t+1}}(x_{t+1}) \ldots b_{qT}(x_T) \qquad (۱)$$

رابطه (۱) را می‌توان چنین بیان کرد که مدل مخفی مارکوف λ با پارامترهای π_i و a_{ij} و$b_i(x)$، با فرض طی شدن رشته حالات Q، رشته بردار X را با احتمال P(X|Q, λ) تولید می‌کند.

با توجه به $b_i(x)$، دو نوع مختلف از مدل مخفی مارکوف خواهیم داشت. در حالت پیوسته، $b_i(x)$ می‌تواند به صورت مجموع چند تابع چگالی احتمال گوسی تعریف شود.شکل ۲ یک مدل مخفی مارکوف پیوسته سه حالته را نشان می‌دهد.



شکل ۲ **مثالی از مدل مخفی مارکوف سه حالته**

تابع چگالی احتمال گوسی به صورت زیر تعریف می‌شود:

$$p(x) = N(x, m, \Sigma) = \frac{1}{\sqrt{\det(\Sigma)(2p)^d}} \exp[\frac{-(x-m)^T \Sigma^{-1}(x-m)}{2}] \qquad (۲)$$

که در آن m بردار میانگین با طول d و Σ ماتریس کواریانس با ابعاد d×d است.

مدل مخفی مارکوف دو مسأله اساسی دارد:

مسأله آموزش: مجموعه مشاهده شده $X = \{x_1, x_2, ..., x_t, ..., x_T\}$ مفروض است، پارامترهای مدل چگونه تخمین زده‌شود که $P(X|\lambda)$ بیشینه شود؟ در مدل پیوسته روابط بازتخمین به صورت زیر بدست می‌آید:

$$\overline{c}_{jk} = \frac{\sum_{t=1}^{T} g_t(j,k)}{\sum_{t=1}^{T}\sum_{k=1}^{M} g_t(j,k)} \qquad (۳)$$

$$\overline{m}_{ik} = \frac{\sum_{t=1}^{T} g_t(j,k).X_t}{\sum_{t=1}^{T} g_t(j,k)} \qquad (۴)$$

$$\overline{\Sigma}_{i,k} = \frac{\sum_{t=1}^{T} g_t(j,k)(X_t - m_{jk})(X_t - m_{jk})^T}{\sum_{t=1}^{T} g_t(j,k)} \qquad (۵)$$

در روابط فوق $g_t(j,k)$ احتمال بودن در حالت j در زمان t است با شرط اینکه kامین مخلوط مبین $\mathbf{X}_t$ باشد.

مسأله بازشناسی: مجموعه مشاهده شده $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ..., \mathbf{x}_T\}$ مفروض است. هدف محاسبه P(X| λ) با داشتن پارامترهای مدل است.

$$p(X \mid l) = \sum_{allQ} p(X \mid Q, l) P(Q \mid l) \qquad (۶)$$

برای حل این احتمال از رویه‌های پیشرو و پسرو استفاده می‌شود. در رویه پیشرو متغیر پیشرو به صورت زیر تعریف می‌شود:

$$a_t(i) = p(X_1, ...X_T, q_T = i \mid l) \qquad (۷)$$

آنگاه با استفاده از استقرا، احتمال P(X| λ) به صورت زیر محاسبه می‌گردد:

$$a_1(j) = p_j b_j(X_1) \quad , \quad 1 \le j \le N$$

$$a_{t+1}(j) = [\sum_{i=1}^{N} a_t(i) a_{ij}] b_j(X_{t+1}) \quad , \quad 1 \le t \le T-1, \quad 1 \le j \le N$$

$$p(X \mid l) = \sum_{i=1}^{N} a_T(i)$$

در رویه پسرو متغیر اتفاقی پسرو به صورت زیر تعریف می‌شود:

$$b_t(i) = p(X_{t+1}, X_{t+2} ... X_T \mid q_T = i, l) \qquad (۸)$$

پس احتمال P(X| λ) با استفاده از استقرا به شکل زیر محاسبه می‌گردد:

$$b_T(i) = 1 \quad , \quad 1 \le i \le N \qquad (۹)$$

$$b_t(i) = \sum_{j=1}^{N} a_{ij} b_j(x_{t+1}) b_{t+1}(j) \quad , \quad 1 \le i \le N, \quad t = T-1, T-2, \mathbf{L}, 1 \qquad (۱۰)$$

$$p(X \mid l) = \sum_{j=1}^{N} p_{\ j} b_j(x_1) b_1(j) \qquad (۱۱)$$

شرح کامل مدل مخفی مارکوف در مراجع موجود است.

5-2 جمع بندی

در این بخش درباره پارامترهای تعیین کننده پیچیدگی که شامل اندازه کتاب لغت، محدودیت‌های زبانی، گفتار مکالمه‌ای و ... می‌باشند، بحث شد. همچنین اجزای مختلف یک سیستم بازشناسی از مرحله نمونه‌برداری و استخراج ویژگی تا مرحله تطبیق الگو و پردازش زبان مورد بررسی قرار گرفت. و در نهایت یک سیستم بازشناسی گفتار به عنوان سیستم پایه معرفی شد.

3 پیش پردازش، نمایش و خلاصه سازی اطلاعات صوتی

3-1 مقدمه

در کاربردهایی که به پردازش صوت نیاز دارند (مانند کد کردن، سنتر و بازشناسی) به نمایش خاصی از اطلاعات گفتار نیاز است. نیاز اصلی برای بازشناسی گفتار استخراج ویژگیهایی از صوت می‌باشد که تشخیص واحدهای پایه مورد نظر در بازشناسی را ممکن می سازد و مجموعه اطلاعات گفتار را به گونه ای خلاصه سازی می نماید که کار تشخیص به سادگی و با دقت لازم انجام گیرد.

3-2 تبدیل صوت از آنالوگ به دیجیتال

به منظور پردازش دیجیتالی صوت، سیگنال صوتی باید از آنالوگ به دیجیتال تبدیل گردد که این تبدیل نیاز به تجهیزات سخت افزاری دارد: یک میکروفن که یک سیگنال صوتی p(t) را به یک سیگنال الکتریکی $x_c(t)$ تبدیل می نماید سپس یک نمونه بردار با فواصل زمانی T_c (با فرکانس $f_c=1/T_c$)، مقادیر و ولتاژهای $x_c(nT_c)=x[n]$ را ارایه می نماید. نهایتاً یک مبدل آنالوگ به دیجیتال (A/D)، مقادیر x[n] , n=0,1,2,N-1 را به مقادیر مشخص، تبدیل می‌کند (معمولاً ۱۶ بیت).

نمونه بردار و مبدل A/D، اغلب در کارتهای صوتی کامپیوترها قرار داده می شوند. با توجه به قانون نایکوئیست، اگر فرکانس نمونه برداری بیشتر از دو برابر پهنای باند فرکانسی گفتار اصلی باشد، سیگنال نمونه برداری شده همان اطلاعات سیگنال پیوسته اصلی را خواهد داشت.

به عبارت دیگر سیگنال گفتار از نظر زمان و دامنه گسسته شده و می تواند در کامپیوتر ذخیره و پردازش شود. در این حالت ما یک دنباله N نمونه ای را که به یک نقطه واحد در فضای برداری N بعدی منطبق می شود خواهیم داشت. یک مرحله از فرآیند برای بازشناسی فونم‌ها تبدیل این مقادیر N تایی به L مقدار با ارزش می باشد که N>L بوده و به پیچیدگی محاسباتی کمتری نیاز دارد. این مرحله با محدودیتهایی انجام می شود بطوریکه برای هرقاب یک بردار ویژه‌ای اختصاص داده شود.

بعد از این فرآیند، الگوریتم بازشناسی، یک دسته بند ساده است که مشخص می کند که کدام مقادیر فضای L بعدی به هر فونم یا هر واحد پایه بازشناسی اختصاص داده می شوند. هر نقطه فضای L بعدی از شکل موج N نمونه‌ای به عنوان یک فونم مطابق با آن مقدار کلاس بندی می‌شود.

3-3 مشخصات فیزیکی سیگنال گفتار

صدا با تولید کننده هایی تولید می شود که در طول تولید یک فونم در مدت زمان کوتاهی در حالت ثابت[۲۳] هستند و با یک حرکت تلفظی دیگر به یک حالت ثابت دیگر می‌رسند عموماً سیگنالهای صوتی در 80-200ms با وضعیت ثابتی از دستگاه گویایی تولید می‌شوند. لذا در تحلیل گفتار قابها را حدود 30ms در نظر می‌گیرندکه در اینصورت می توان سیگنال صوتی را ایستان فرض نمود.

3-4 استخراج ویژگی

از قسمتهای ایستان سیگنال گفتار باید ویژگیهایی را استخراج کرد تا از آنها در مرحله بازشناسی استفاده نمود. سیگنال گفتار اطلاعات زیادی دارد که عموماً با طیف لحظه‌ای آن یا شکل مجرای گفتار و ... مرتبط می‌باشند. پردازش این همه ویژگی ضرورت ندارد بدین منظور تبدیلاتی روی سیگنال گفتار انجام می‌شود و ضرایبی استخراج می‌شوند که به محاسبات کمتری نیاز دارند و در عین حال اطلاعات لازم را در بردارند.

یکی از ضرایبی که از سیگنال گفتار استخراج می شود ضرایب MFCC[۲۴] می‌باشد هر چند روشهای پردازش دیگری نیز وجود دارد ولی این روش دقت کافی و پیچیدگی محاسبات کمتری در مقایسه با دیگر روشها دارد.

3-5 پیش پرداز سیگنال

اغلب پیک به‌دست آمده ازجهاز صوتی با فرکانس بالا، دامنه کوچکتری نسبت به فرمانتهای با فرکانس پایین دارد. لذا یک پیش تاکید برای فرکانسهای بالا به منظور یکنواخت شدن سیگنال لازم است. در حقیقت این فیلتر اثرات طیفی حنجره (دو قطب) و لبها (یک صفر) را حذف می‌کند. همچنین تغییرات ناگهانی موجود در سیگنال را که بر اثر نویزهای شدید محیط بوجود می آید، حذف می‌کند و باعث یکنواخت شدن سیگنال می‌گردد. اینگونه پردازش اغلب با فیلتر کردن سیگنال گفتار با یک فیلتر FIR مرتبه اول بدست می آید.

[23] stable
[24] mel frequency cepstral coefficient

پیش پردازهای دیگری نظیر حذف نویز و حذف سکوت نیز در سیستم های ASR مهم هستند. در سیستمِ ASR مبتنی بر HMM اگر سکوتهای طولانی از گفتار حذف نشوند کارایی بطور قابل ملاحظه کاهش پیدا می‌کند لذا یک تشخیص دهنده گفتار مناسب لازم است. اگر نسبت سیگنال به نویز تغییرات قابل ملاحظه‌ای نداشته باشد تشخیص دهنده‌های ساده مبتنی برمعیار انرژی،کارایی خوبی نشان می دهند.

3-6 پنجره بندی

روشهای معمول برای ارزیابی طیفی در صورتی که سیگنال ایستان (یعنی سیگنالی که خصوصیات آماری آن در زمان متغیر نباشد) باشد کارایی خوبی نشان می دهند. برای صوت این فرض فقط در فواصل زمانی کوتاه برقرار است یعنی یک تحلیل زمان کوتاه انجام می‌شود که سیگنال را به دنباله های متوالی پنجره بندی شده تبدیل می‌نماید و به طور جداگانه پردازش می‌شوند.

در سیستم‌های ASR پنجره ای که فراوان مورد استفاده قرار گرفته، پنجره همینگ می‌باشد که پاسخ ضربه آن به صورت زیر است:

$$W(n) = 0.54 - 0.46Cos\frac{2pn}{N-1} \qquad 0 \le n \le N-1 \qquad (۱۲)$$

3-7 پردازش بانک فیلتر

تحلیل طیف ویژگیهایی از سیگنال گفتار را آشکار می کند که اساساً با ترکیب جهاز صوتی مرتبط است. برای این منظور ویژگیهای طیفی گفتار را عموماً از خروجی بانکهای فیلتر که هر کدام در یک رنج فرکانسی تعریف می شوند بدست می آورند. معمولا یک مجموعه فیلتر ۲۴ تایی میان گذر برای شبیه سازی پردازش گوش انسان بکار برده می شود. فیلترها اغلب در محور فرکانس به شکل غیر یکسان توزیع می شوند. به عنوان یک قانون قسمتی از طیف که زیر۱KHz است با بانکهای فیلتر بیشتری پردازش می شود، چرا که آن قسمت اطلاعات بیشتری را از جهاز صوتی مانند فرمانت اول دربر دارد. پاسخ فرکانسی بانکهای فیلتر، پردازش ادراکی انجام شده در گوش را شبیه سازی می کنند و این فیلترینگ وزن دهی ادراکی نامیده می شود.

درسیستم های ASR، مقیاس ادراکی که بیشترین کاربرد را داشته، مقیاس mel می‌باشد. فرکانس مرکزی هر بانک فیلتر Mel قبل از ۱kHz به طور یکنواخت توزیع می شود و بعد از ۱kHz یک مقیاس لگاریتمی را دنبال می کند.

3-8 ضرایب کپسترال، دلتا، دلتا دلتا و انرژی

ضرایب کپسترال معمولا با یک ضریب انرژی e_t همراه هستند که لگاریتم انرژی قاب یا پنجره می‌باشد. این پارامتر مفید است چرا که فونم‌های مختلف انرژی‌های مختلفی دارند. پارامترهای کپسترال و انرژی، اطلاعات دینامیکی سیگنال گفتار را به حساب نمی آورند از طرفی این اطلاعات گاهاً برای ASR ها مفید هستند. لذا می توان دیفرانسیل مرتبه اول و دوم را برای بدست آوردن آن اطلاعات بکار برد که حاوی اطلاعات دینامیک و اطلاعات انتقال میان حالات مختلف گویش هستند.

چنانچه t شمارهٔ قاب و i شماره ضریب باشد، ضرایب دلتا کپسترال از روابط زیر بدست می‌آیند:

$$d_t(i)=\sum_{t=1}^{N}\left[\frac{t\,(C_{t+t}(i)-C_{t-t}(i))}{2\sum_{t=1}^{N}t^2}\right] N\le t\le T-N \qquad (۱۳)$$

که در این رابطه T تعداد کل قابها و N تعیین کننده طول پنجره‌ای است که روی آن مشتق گرفته می‌شود و C_t ضریب کپسترال در زمان t می باشد که از فرمول (۱۴) محاسبه می شود.

(۱۴)

در قابهای ابتدایی و انتهایی برای محاسبهٔ مشتقات ضرایب از روابط زیر استفاده می‌شود

$$d_t(i)=C_{t+1}(i)-C_t(i) \qquad t<N \qquad (۱۵)$$

$$d_t(i)=C_t(i)-C_{t-1}(i) \qquad t\ge T-N \qquad (۱۶)$$

مقدار T را می‌توان بطور اختیاری تعیین کرد. ولی بطور معمول مقادیر ۲ یا ۳ مناسب‌ترین مقدار برای آن هستند. گاه درمخرج رابطهٔ (۱۳) از یک مقدار ثابت نیز استفاده می‌شود.

برای محاسبهٔ ضرایب دلتا دلتا کپسترال، ضرایب دلتا کپسترال را باید در روابط ۲، ۴ و ۵ قرار داد. ضریب لگاریتم انرژی برای یک قاب از رابطهٔ (۱۷) بدست می‌آید. این ضریب باید برای همهٔ قابهای موجود در سیگنال گفتار بدست آید.

$$E = \log\left(\sum_{n=1}^{M} S^2(n)\right) \qquad (۱۷)$$

پس از بدست آمدن ضرایب انرژی، می‌توان مشتقات آن را بدست آورد. ضرایب دلتا انرژی را می توان از رابطهٔ (۱۸) محاسبه نمود.

$$\Delta E_t = \sum_{t=1}^{N}\left[\frac{t\left(E_{t+t} - E_{t-t}\right)}{2\sum_{t=1}^{N} t^2}\right] \qquad N \le t < T - N \qquad (۱۸)$$

در رابطهٔ فوق E_t ضریب لگاریتم انرژی برای قاب t، T تعداد کل قابهای موجود در سیگنال گفتار، و N طول پنجره‌ای است که روی آن مشتق گرفته می‌شود.

در قابهای ابتدایی و انتهایی، برای محاسبهٔ دلتا انرژی از روابط (۱۹) و (۲۰) استفاده می شود:

$$\Delta E_t = E_{t+1} - E_t \qquad t < N \qquad (۱۹)$$

$$\Delta E_t = E_t - E_{t-1} \qquad t \ge T - N \qquad (۲۰)$$

برای محاسبهٔ ضرایب دلتا دلتا انرژی باید ضرایب دلتا انرژی را در روابط بالا قرار داد.

4 مدل مخفی مارکوف و کاربرد آن در بازشناسی گفتار

4-1 مقدمه

همان طورکه گفته شد یکی از روشهای مدلسازی آماری مورد استفاده در گفتار مدل مخفی مارکوف می باشد. مدل کردن سیگنالهای گفتار، برای تشخیص خودکار گفتار، بهبود گفتار، سنتز گفتار و ترجمه ماشینی به کار رفته و توانایی مناسبی از خود نشان داده است.

4-2 فرآیندهای تصادفی

یک فرآیند تصادفی مجموعه‌ای از متغیرهای تصادفی $\{X(t), t \in T\}$ است که برای هر $t \in T$، $X(t)$ یک متغیر تصادفی می‌باشد.

4-3 زنجیره مارکوف

۴-۳-۱ زنجیره مارکوف زمان گسسته

در زنجیره مارکوف زمان گسسته، $X = X_1, X_2, ..., X_n$ یک دنباله از متغیرهای تصادفی از یک الفبای گسسته محدود $O=\{O_1, O_2, ..., O_n\}$ می‌باشد که در آن توزیع احتمال شرطی هر وضعیت آینده X_{n+1} بطوریکه وضعیتهای قبلی $X_0, X_1, ..., X_{n-1}$ و وضعیت فعلی سیستم X_n باشد مستقل از وضعیتهای قبلی است و فقط بستگی به وضعیت فعلی آن دارد یعنی به صورت رابطه‌ای خواهیم داشت:

$$P(X_1, X_2, ..., X_n) = P(X_1) \prod_{t=2}^{n} P(X_t \mid X_1^{t-1}) \qquad (۲۱)$$

که $X_1^{t-1} = X_1, X_2, ..., X_{t-1}$ و

$$P(X_t \mid X_1^{t-1}) = P(X_t \mid X_{t-1}) \qquad (۲۲)$$

و معادله (۲۱) به صورت زیر در می‌آید:

$$P(X_1, X_2, ..., X_n) = P(X_1)\prod_{t=2}^{n} P(X_t|X_{t-1}) \quad (۲۳)$$

چنین فرایند تصادفی زنجیره مارکوف و معادله (۲۳) فرض مارکوف نامیده می‌شود.

زنجیره مارکوف برای مدل کردن رخدادهای ایستان بکار می‌رود یعنی احتمال انتقال وضعیت 's به وضعیت s مستقل از زمان t می‌باشد. به چنین زنجیره‌های مارکوف، زنجیره مارکوف همگن گفته می‌شود که در آن:

$$P(X_t = s|X_{t-1=s'}) = P(s|s') \quad (۲۴)$$

اگر X_t را به یک حالت نسبت دهیم زنجیر مارکوف را با یک فرآیند حالت متناهی می‌توان نشان داد بطوریکه گذار بین حالتها با تابع احتمال $P(s|s')$ مشخص می‌شود. اگر مجموعه حالتها زنجیره مارکوف را به $\{1,...,N\}$ و حالت زمان t را با s_t نشان دهیم پارامترهای مدل مارکوف بصورت زیر خواهد بود:

$$a_{ij} = P(s_t = j|s_{t-1} = i) \qquad , 1 \le i, j \le N \quad (۲۵)$$

$$p_i = P(s_1 = i) \qquad ,1 \le i \le N \quad (۲۶)$$

a_{ij} احتمال انتقال از حالت i به حالت j است و p_i احتمال اولیه بودن زنجیره مارکوف در حالت i می‌باشد و با داشتن a_{ij}ها یک ماتریس انتقال وضعیت تعریف می‌شود. هر دو احتمال محدودیتهایی دارند:

$$\sum_{j=1}^{N} a_{ij} = 1 \quad , 1 \le i \le N \quad (۲۷)$$

$$\sum_{j=1}^{N} p_j = 1 \quad (۲۸)$$

۲-۳-۴ زنجیره‌های مارکوف زمان پیوسته

در زنجیره‌های مارکوف زمان پیوسته مدت زمانی که سیستم در یک وضعیت باقی می‌ماند خود یک متغیر تصادفی نمایی است در صورتیکه در زنجیره‌های مارکوف زمان گسسته هر انتقال یک مرحله است و زمان ماندن در یک وضعیت مطرح نیست که در حالت کلی بصورت زیر تعریف می‌شود:

فرآیند تصادفی با زمان پیوسته $\{N(t), t \ge 0\}$ که مقادیر غیر منفی و صحیح را اختیار می‌کند را در نظر

بگیرید فرآیند $\{N(t), t \geq 0\}$ یک زنجیره مارکوف با زمان پیوسته است اگر برای تمام $S, t \geq 0$ و مقادیر غیرمنفی i,j، $X(u)$ و $0 \leq u \leq s$ داشته باشیم:

$$P\{X(t+s) = j | X(s) = i \ , X(u) = x(u), 0 \leq u \leq s\} = P\{X(t+s) = j | X(s) = i\} \quad (۲۹)$$

4-4 تعریف مدل مخفی مارکوف

در زنجیره‌های مارکوف هر حالت بطور مشخص با یک رخدادِ قابل مشاهده متناظر است یعنی خروجی تصادفی نیست ولی در بعضی زنجیره‌های مارکوف، مشاهده یک تابع احتمالی از حالت است که مدل مخفی مارکوف نامیده می‌شود و از دو لایه تشکیل می‌شود، یک لایه فرآیند تصادفی قابل مشاهده و لایه زیرین، یک لایه فرآیند تصادفی است که مستقیماً قابل مشاهده نمی‌باشد.

پس یک مدل مخفی مارکوف یعنی زنجیره مارکوفی که در آن خروجی مشاهده شده از یک متغیر تصادفی X با تابع احتمال خروجی حالتها مرتبط است.

در این مدل برای یک دنباله مشاهده شده، نمی‌توان یک دنباله حالت که تناظر یک به یک با دنباله مشاهدات داشته باشد مشخص کرد یعنی دنباله حالتها، قابل مشاهده نیست و مخفی است برای همین اصطلاحاً مدل مخفی مارکوف نامیده می‌شود.

۴-۴-۱ اجزاء مدل مخفی مارکوف

یک مدل مخفی مارکوف با اجزاء زیر مشخص می‌شود:

یک الفبای خروجی مشاهده شونده یعنی سمبلهای خروجی فیزیکی سیستم مدل شونده

$$O = \{O_1, O_2, ..., O_m\}$$

مجموعه حالتها: $\Omega = \{1, 2, ..., N\}$

بطوریکه S_t حالت زمان t را نشان می‌دهد.

ماتریس احتمال انتقال: $A = \{a_{ij}\}$

که a_{ij} یعنی احتمال انتقال از حالت i به حالت j:

$$a_{ij} = P(s_t = j | s_{t-1} = i) \quad (۳۰)$$

ماتریس احتمال خروجی $B = \{b_t(k)\}$

که $b_t(k)$ احتمال تولید سمبل O_k در حالت i می‌باشد.

اگر $X = X_1, X_2, ..., X_t, ...$ خروجی مشاهده شده باشد دنباله حالتهای $S = S_1, S_2, ..., S_t, ...$ قابل مشاهده و مشخص نیست و $b_t(k)$ را می‌توان بصورت زیر نوشت:

$$b_t(k) = P(X_t = O_k \mid s_t = i) \qquad (۳۱)$$

توزیع حالت اولیه $p = \{p_i\}$

بطوریکه

$$p_i = P(s_0 = i) \qquad 1 \le i \le N \qquad (۳۲)$$

از آنجائیکه a_{ij} و $p_i, b_{ij}(k)$ مقادیر احتمال هستند باید خصوصیات زیر را داشته باشند:

$$a_{ij} \ge 0 \quad , b_t(k) \ge 0, p_i \ge 0 \ \forall all\ i, j, k \qquad (۳۳)$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad (۳۴)$$

$$\sum_{k=1}^{M} b_t(k) = 1 \qquad (۳۵)$$

$$\sum_{i=1}^{N} p_i = 1 \qquad (۳۶)$$

HMM دو پارامتر با مقادیر ثابت N, M دارد که N تعداد حالتها و M اندازه الفبای قابل مشاهده را نشان می‌دهند، همچنین مجموعه الفبای قابل مشاهده O، و سه مجموعه یا ماتریس احتمالات، که در حالت کلی یک HMM را بصورت زیر نمایش می‌دهیم:

$$\Phi = (A, B, p) \qquad (۳۷)$$

در مدل مخفی مارکوف مرتبه اول، دو فرض داریم که اولی فرض زنجیره مارکوف می‌باشد:

$$P(s_t \mid s_1^{t-1}) = P(s_t \mid s_{t-1}) \qquad (۳۸)$$

که s_1^{t-1} دنباله حالت $s_1, s_2, ..., s_{t-1}$ می‌باشد.

و فرض دوم استقلال خروجی است:

$$P(X_t \mid X_1^{t-1}, s_1^t) = P(X_t \mid s_t) \qquad (۳۹)$$

که X_1^{t-1} دنباله خروجی $X_1, X_2, ..., X_{t-1}$ را نشان می‌دهد.

شرط استقلال خروجی یعنی اینکه احتمال تولید یک سمبل در زمان t فقط به حالت s_t بستگی دارد و از خروجی قبلی، مستقل است.

این فرضها، بدون اینکه تاثیر چندانی در ظرفیت مدل وحافظه آن داشته باشند باعث می‌شوند که ارزیابی، کدگشایی و آموزش مدل ممکن شود و همچنین تعداد پارامترهایی که باید تخمین زده شوند کم شود.

۴-۴-۲ سه مسأله اساسی مدل مخفی مارکوف

سه مسأله اساسی که در رابطه با مدل HMM قابل طرح است به شرح زیر می باشد:

مسأله ارزیابی: با یک مدل Φ و دنباله مشاهده $X = (X_1, X_2, ..., X_T)$ داده شده، احتمال $P(X \mid \Phi)$ چیست؟ یعنی احتمال اینکه مدل، مشاهده فوق را تولید نماید.

مسأله کدگشایی: با یک مدل Φ مفروض و دنباله مشاهده $X = (X_1, X_2, ..., X_T)$، بهینه‌ترین دنباله حالات $S = (S_0, S_1, ..., S_T)$ در مدل که مشاهده فوق را تولید نماید کدام است؟

مسأله آموزش: با یک مدل Φ داده شده و یک مجموعه مشاهدات، چطور ما می‌توانیم پارامترهای مدل $\overline{\Phi}$ را برای بیشینه کردن احتمال توأم $\prod_x P(X \mid \Phi)$ تنظیم نماییم.

۴-۴-۳ نحوه ارزیابی HMM – الگوریتم پیشرو

برای محاسبه احتمال $P(X \mid \Phi)$ برای دنباله مشاهده، $X = (X_1, X_2, ..., X_T)$ و مدل مفروض Φ به‌صورت زیر عمل می‌شود:

$$P(X \mid \Phi) = \sum_{alls} P(X \mid S, \Phi) P(s \mid \Phi) \qquad (۴۰)$$

برای یک دنباله حالت $S = (S_0, S_1, ..., S_T)$ می‌توان نوشت:

$$P(S \mid \Phi) = P(S_1 \mid \Phi) \prod_{t=2}^{T} P(s_t \mid s_{t-1}, \Phi) = p_{s1}\, a_{s1s2}, ... a_{sT-1,sT} \qquad (۴۱)$$

همچنین داریم:

$$P(X \mid S,\Phi) = P(X_1^T \mid S_1^T,\Phi) = \prod_{t=1}^{T} P(X_t \mid s_t,\Phi) = b_{s1}(X_1)b_{s2}(X_2)...b_{ST}(X_T) \quad (۴۲)$$

با جایگذاری در معادله (۴۰) خواهیم داشت:

$$P(X \mid \Phi) = \sum_{alls} a_{s_o s_1} b_{s_1}(X_1) a_{s_1 s_2} b_{s_2}(X_2)...a_{s_{T-1} s_T} b_{s_T}(X_T) \quad (۴۳)$$

برای محاسبه معادله (۴۳) پیچیدگی $O(N^T)$ خواهد بود که برای TوN های بزرگ پیچیدگی نمایی بزرگی بدست می‌آید. برای حل این مشکل از الگوریتمی معروف به الگوریتم پیشرو استفاده می‌شود و ایده آن به این صورت است که نتایج میانی را برای استفاده در مراحل بعدی محاسباتِ دنباله حالات ذخیره می‌نماییم.

احتمال پیشرو را به صورت زیر تعریف می‌کنیم:

$$a_t(i) = P(X_1^t, s_t = i \mid \Phi) \quad (۴۴)$$

یعنی احتمال بودن در حالت i در زمان t و تولید پاره مشاهده $X_1, X_2, X_3, ...X_t$.

$a_t(i)$ را می‌توان بصورت بازگشتی نوشت:

مرحله اول: مقداردهی اولیه

$$a_t(i) = P(X_1^t, s_t = i \mid \Phi) \quad (۴۵)$$

مرحله دوم: بازگشت

$$a_t(j) = \left[\sum_{i=1}^{N} a_{t-1}(i) a_{ij}\right] b_j(X_t) \quad 2 \le t \le T, 1 \le j \le N \quad (۴۶)$$

مرحله سوم: اتمام

$$P(X \mid \Phi) = \sum_{i=1}^{N} a_T(i) \quad (۴۷)$$

این الگوریتم دارای پیچیدگی $O(N^2T)$ خواهد بود.

۴-۴-۴ نحوه کدگشایی HMM – الگوریتم ویتربی:

از آنجائیکه دنباله حالات مخفی است روش ممکن و کاربردی، پیدا کردن دنباله حالاتی است که بیشترین احتمال را در تولید دنباله مشاهدات دارند.

به عبارت دیگر ما به دنبالِ دنباله حالات $S=(S_1,S_2,...,S_T)$ هستیم که $P(S,X|\Phi)$ را بیشینه نماید و همان چیزی است که در مسأله پیدا کردن مسیر بهینه در برنامه‌ریزی پویا استفاده می‌شود و الگوریتم ویتربی نام دارد. پیچیدگی الگوریتم ویتربی $O(N^2T)$ می‌باشد.

الگوریتم ویتربی را می‌توان تغییر یافته الگوریتم پیشرو قلمداد کرد که بجای جمع کردن احتمالات از مسیرهای متفاوت به یک حالت مقصد، مسیر بهینه را انتخاب و حفظ می‌کند. برای مشخص کردن احتمال مسیر بهینه:

$$V_t(i)=P(X_1^{t-1},S_1^{t-1},S_t=i|\Phi) \qquad (۴۸)$$

که $V_t(i)$ احتمال شبیه‌ترین دنباله حالت در زمان t می‌باشد که توسط مشاهده X_1^t (تا زمان t) بدست آمده و در حالت i پایان یافته است. رویه استقراء برای الگوریتم ویتربی بصورت زیر می‌باشد:

مرحله اول: مقداردهی اولیه

$$V_t(i)=p_i b_i(X_1) \quad 1\le i\le N$$
$$B_1(i)=0 \qquad (۴۹)$$

مرحله دوم: بازگشت

$$v_t(j)=\max_{1\le i\le N}\left[V_{t-1}(i)a_{ij}\right]b_j(X_t) \quad 2\le t\le T,1\le j\le N \qquad (۵۰)$$

$$B_t(j)=Arg\ \underset{1\le i\le N}{Max}\left[V_{t-1}(i)a_{ij}\right] \qquad (۵۱)$$

مرحله سوم: اتمام

$$\text{بهترین نمره}=\underset{1\le i\le N}{Max}[V_t(i)] \qquad (۵۲)$$

$$S_T^*=Arg\ \underset{1\le i\le N}{Max}[B_T(i)] \qquad (۵۳)$$

مرحله چهارم: مرحله پسرو

$$S_T^* = B_{t+1}(S_{t+1}^*) \qquad t = T-1, T-2, \ldots, 1 \tag{۵۴}$$

$S^* = (S_1^*, S_2^*, \ldots, S_T^*)$ بهترین دنباله می‌باشد.

۵-۴-۴ نحوه تخمین پارامترهای HMM- الگوریتم Baum – Welch:

تخمین پارامترهای مدل $\Phi = (A, B, p)$ برای توصیف دقیق‌تر دنباله‌های مشاهده شده خیلی مهم می‌باشد. راه‌حل مشخصی برای بیشینه کردن احتمال توأم در یک فرم بسته وجود ندارد ولی با روش تکراری الگوریتم Baum – Welch می‌توان مسأله را حل کرد که الگوریتم پیشرو- پسرو نامیده شده است.

معرفی احتمال پسرو:

$$b_t(i) = P(X_{t+1}^T \mid S_t = i, f) \tag{۵۵}$$

که $b_t(i)$ احتمال تولید X_{t+1}^T (از t+1 تا انتها) از زمان t و حالت i می‌باشد.

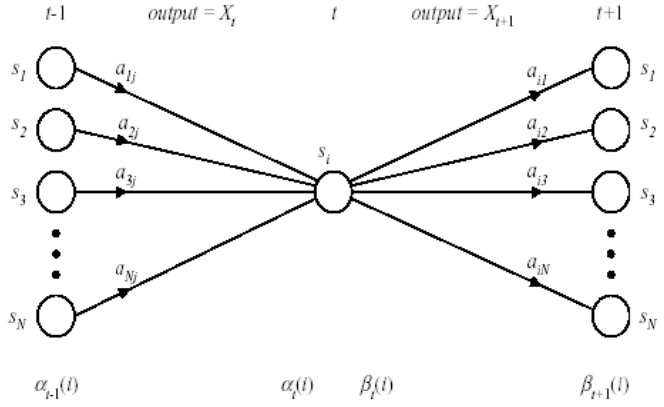
$b_t(i)$ را می‌توان بروش استقراء بدست آورد:

مقدار اولیه

$$b_T(i) = \frac{1}{N} \qquad 1 \le i \le N \tag{۵۶}$$

$$b_t(i) = \left[\sum a_{ij}\, b_j(X_{t+1}) b_{t+1}(j)\right] \quad t = T-1, \ldots, 1, \; 1 \le i \le N \tag{۵۷}$$

ارتباط همسایگی در $(b_t, b_{t+1}, a_{t-1}, a_t) b, a$ در شکل ۳ نشان داده شده است. a بطور بازگشتی از چپ براست و b از راست به چپ محاسبه می‌شوند.



شکل ۳ **ارتباط بین** $b_t, b_{t-1}, a_t, a_{t-1}$ **در الگوریتم پیشرو – پسرو**

برای شرح رویه تخمین پارامترهای مدل مخفی مارکوف $x_t(i,j)$ را بصورت زیر تعریف می‌کنیم:

$$x_t(i,j) = P(s_t = i,\ s_{t+1} = j \mid X_1^T, \Phi) \qquad (۵۸)$$

یعنی احتمال انتقال از حالت i به j برای مدل و دنباله مشاهده مفروض در زمان t.

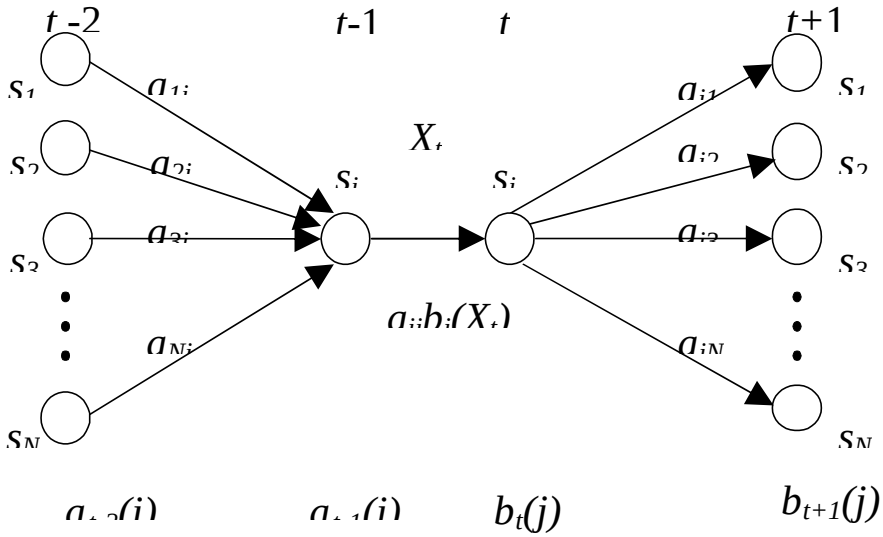
با استفاده از تعریف متغیرهای پیشرو و پسرو داریم:

$$x_t(i,j) = \frac{P(s_t = i, s_{t+1} = j, X_1^T \mid f)}{P(X_1^T \mid f)} = \frac{a_t(i)\, a_{ij}\, b_j(X_{t+1})\, b_{t+1}(j)}{\sum_{k=1}^{N} a_T(k)} \qquad (۵۹)$$

که

$$\sum_{k=1}^{N} a_T(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} a_t(i)\, a_{ij}\, b_j(X_{t+1})\, b_{t+1}(j) \qquad (۶۰)$$

رابطه (۶۰) را می‌توان به فرم شکل ۴ نشان داد:



شکل ۴ **نشان دادن عملیات لازم جهت محاسبه** $x_t(i,j)$

اکنون متغیر $g_t(i)$ را به صورت زیر تعریف می‌کنیم:

$$g_t(i) = P(S_t = i \mid X_1^T, \Phi) \qquad (۶۱)$$

یعنی احتمال بودن در حالت i در زمان t با دنباله مشاهده مفروض X_1^T وقتی مدل Φ است. با توجه به معادله (۶۱) و متغیرهای پیشرو و پسرو داریم:

$$g_t(i) = \frac{a_t(i)b_t(i)}{P(X_1^T \mid f)} = \frac{a_t(i)b_t(i)}{\sum_{i=1}^{N} a_t(i)b_t(i)} \qquad (۶۲)$$

خصوصیت زیر با توجه به عامل نرمالیزه (مخرج کسر) برای $g_t(i)$ برقرار است:

$$\sum_{i=1}^{N} g_t(i) = 1 \qquad (۶۳)$$

با توجه به تعریف $g_t(i)$ و $x(i)$ می‌توان آنها را اینگونه به هم مرتبط ساخت:

$$g_t(i) = \sum_{j=1}^{N} x_t(i,j) \qquad (۶۴)$$

با جمع بستن $g_t(i)$ بروی t، کمیتی حاصل می‌شود که می‌توان آن را به عنوان تعداد دفعات مشاهده حالت i یا تعداد گذرهای انجام شده از حالت i در نظر گرفت و جمع $x_t(i,j)$ بر روی t یعنی تعداد انتقال‌ها از حالت i به حالتj . با توجه به روابط بالا، روابط تخمین برای پارامترهای مدل بصورت زیر خواهد بود:

$$\bar{p}_i = g_1(i) \qquad (۶۵)$$

یعنی تعداد دفعات قرار داشتن در حالت i در زمان.t=1

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} x_t(i,j)}{\sum_{t=1}^{T-1} g_t(i)} \qquad (۶۶)$$

یعنی تعداد گذرها از حالت i به حالت j نسبت به تعداد گذرها از حالت i

$$\bar{b}_j(k) = \frac{\sum_{t=1, s=j, X_k=O_t} g_t(j)}{\sum_{t=1}^{T} g_t(j)} \qquad (۶۷)$$

یعنی تعداد دفعات بودن در حالت j و مشاهده سمبل X_k نسبت به دفعات بودن در حالت j.

مدل تخمین زده شده بصورت $\bar{\Phi} = (\bar{A}, \bar{B}, \bar{p})$ تعریف می‌شود که پارامترهای آن از طرف چپ معادلات (۶۵) تا (۶۷) بدست می‌آید.

به‌وسیله Baum ثابت شده است که یا مدل $\bar{\Phi}$ محتمل‌تر از مدل Φ می‌باشد، $P(X_1^T \mid \bar{\Phi}) > P(X_1^T \mid \Phi)$، که در نتیجه مدلی پیدا شده که با احتمال بیشتری دنباله مشاهده را تولید می‌نماید و یا اینکه مدل $\Phi = \bar{\Phi}$ می‌باشد. در حالت اول، محاسبات تخمین مجدد تا جایی تکرار می‌شود که یک شرط محدودیت برآورده شود که در نتیجه احتمال مشاهده رشته X_1^T در مدل بهبود می‌یابد و این رویه تخمین، یک تخمین احتمال بهینه یا ML از مدل مخفی مارکوف می‌باشد. محدودیتهای آماری که در تکرارهای تخمین باید مد نظر قرار گیرند بصورت زیر می‌باشند:

$$\sum_{i=1}^{N} \bar{p} = 1 \qquad (۶۸)$$

$$\prod_{i=1}^{N} \bar{a}_{ij} = 1 \qquad (۶۹)$$

$$\prod_{k=1}^{M} \bar{b}_j(k) = 1 \qquad (۷۰)$$

الگوریتم پیشرو – پسرو که باوم استفاده کرده است یک پیشرفت یکنواخت احتمال را در هر تکرار تضمین می‌کند و نهایتاً به یک احتمال محلی همگرا می‌شود.

4-5 مدل مخفی مارکوف پیوسته(CDHMM)

اگر مشاهدات در یک مجموعه متناهی نباشد بلکه از یک فضای پیوسته باشد توزیع خروجی گسته قبلی باید تغییر نماید. تفاوت بین HMM گسسته و پیوسته در توابع احتمال خروجی می‌باشد. در بازشناسی گفتار در HMM پیوسته چندی‌سازی یعنی نگاشت بردارهای مشاهده در فضای پیوسته به فضای گسسته انجام نمی‌گیرد و در نتیجه خطای ناشی از چندی‌سازی در نتایج اعوجاج ایجاد نمی‌شود.

۴-۵-۱ مدل مخفی مارکف با چگالی مخلوطی پیوسته

برای انتخاب یک تابع چگالی مشاهدات پیوسته، برخی محدودیتها باید برروی فرم تابع چگالی احتمال اعمال شود تا اطمینان حاصل شود که پارامترهای تابع چگالی احتمال می‌توانند به شکل مناسبی تخمین زده شوند. شکل عمومی تابع چگالی احتمال که می‌توان برای آن رویه تخمین مجدد را ارائه نمود مخلوطی متناهی بصورت زیر است:

$$b_j(X) = \sum_{K=1}^{M} C_{jk} \Pi(X, m_{jk}, \Sigma_{jk}) = \sum_{K=1}^{M} C_{jk} b_{jk}(X) \qquad 1 \le j \le N \qquad (۷۱)$$

که در آن N تعداد حالات، X بردار مدل شونده، M تعداد مولفه‌های مخلوط، C_{jk} ضریب k امین مخلوط در حالت j ام، Π یک چگالی احتمال بصورت لگاریتمی مقعر یا بیضوی متقارن با بردار میانگین m_{jk} و ماتریس کوواریانس Σ_{jk} برای k امین مخلوط در حالت j ام می‌باشد.

معمولاً Π از نوع توابع چگالی مخلوط گوسی چند متغیره انتخاب می‌شود زیرا این توابع می‌توانند هر تابع چگالی پیوسته را تخمین بزنند. در بعضی کاربردها از توابع چگالی احتمال دیگری نظیر چگالی مخلوط لاپلاسین نیز استفاده می‌شود.

تابع چگالی احتمال گوسی بصورت زیر تعریف می‌شود:

$$P(X) = N(X, m, \Sigma) = \frac{1}{\sqrt{\det \Sigma (2p)^d}} \exp\left[\frac{-(X-m)^T \Sigma^{-1} (X-m)}{2}\right] \qquad (۷۲)$$

این تابع چگالی احتمال d بعدی است که در آن بعد بردار ورودی (ویژگی) X است بردار میانگین d بعدی و ماتریس کوواریانس، ماتریسی متقارن با بعد $d \times d$ می‌باشد.

$$X = \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_d \end{bmatrix} \qquad m = E(X) = \begin{bmatrix} m_1 \\ . \\ . \\ . \\ m_d \end{bmatrix} \qquad (۷۳)$$

$$\Sigma = \begin{bmatrix} s_{11} & s_{21} & \dots & s_{d1} \\ s_{12} & s_{22} & \dots & s_{d2} \\ . & . & & . \\ . & . & & . \\ s_{1d} & s_{2d} & \dots & s_{dd} \end{bmatrix} = \begin{bmatrix} s_1^2 & s_{21} & \dots & s_{d1} \\ s_{12} & s_2^2 & \dots & \\ . & & & \\ s_{1d} & \dots & & s_d^2 \end{bmatrix} \qquad (۷۴)$$

ضرایب مخلوط C_{jk} در محدودیتهای آماری زیر باید قرار داشته باشد:

$$\sum_{k=1}^{M} C_{jk} = 1 \quad ، 1 \le j \le N \quad , \quad C_{jk} \ge . \qquad (۷۵)$$

بنابراین تابع چگالی احتمال با توجه به این رابطه بطور کامل نرمالیزه می شود یعنی:

$$\int_{-\infty}^{+\infty} b_j(x)\, dx = 1 \qquad (۷۶)$$

رابط تخمین برای C_{jk} و m_{jk} و Σ_{jk} بصورت زیر بدست می آید:

$$\overline{C}_{jk} = \frac{\sum_{t=1}^{T} g_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} g_t(j,k)} \qquad (۷۷)$$

یعنی تعداد دفعات قرار داشتن در حالت j با استفاده از مخلوط k ام نسبت به تعداد دفعاتی که سیستم در حالت j می باشد.

$$\overline{m}_{jk} = \frac{\sum_{t=1}^{T} g_t(j,k)\, o_t}{\sum_{t=1}^{T} g_t(j,k)} \qquad (۷۸)$$

هر ترم موجود در صورت را با مشاهده وزن دهی می‌کند و مقدار بخشی از بردار مشاهدات محسوب شده برای k امین جزء مخلوط را بدست می‌دهد.

$$\overline{\Sigma}_{jk} = \frac{\sum_{t=1}^{T} g_t(j,k)(o_t - m_{jk})(o_t - m_{jk})^T}{\sum_{t=1}^{T} g(j,k)} \qquad (۷۹)$$

که در رابطه فوق $g_t(j,k)$ احتمال بودن در حالت j در زمان t است بطوریکه k امین مخلوط مبین O_t باشد بعبارتی تعمیم $g_t(j)$ در حالت استفاده از یک مخلوط یا چگالی گسته است. رابطه $g_t(j,k)$ بصورت زیر است:

$$g_t(j,k) = \left[\frac{a_t(j)\, b_t(j)}{\sum_{j=1}^{N} a_t(j)\, b_t(j)}\right]\left[\frac{C_{jk}\Pi(o_t, m_{jk}, \Sigma_{jk}}{\sum_{k=1}^{M} C_{jk}\Pi(o_t, m_{jk}\, \Sigma_{jk})}\right] \qquad (۸۰)$$

رابطه تخمین احتمالات انتقال a_{ij} همانند روابط مورد استفاده آن در حالت چگالی گسسته می‌باشد.

۲-۵-۴ مقیاس‌گذاری

زمانیکه احتمالات پیشرو و پسرو را محاسبه می‌کنیم (در الگوریتم پیشرو و پسرو)، اگر طول دنباله مشاهده T به حد کافی بزرگ باشد مقادیر در الگوریتم به صفر میل می‌کنند یعنی برای Tهای خیلی بزرگ، مقادیر به حدی کوچک می‌شوند که از دقت هر ماشینی کمتر می‌شوند. برای حل این مسأله، می‌توان از یک سری ضرایب برای مقیاس‌گذاری استفاده کرد که در این صورت مقادیر در محدوده دقت ماشین قرار می‌گیرند و ضرایب را در انتهای محاسبات حذف می‌نماییم بدون اینکه در نتیجه بدست آمده تأثیری داشته باشد.

فرض کنید $a_t(i)$ در ضریب S_t زیر ضرب شود:

$$S_t = \frac{1}{\sum_i a_{\ t}(i)} \qquad (۸۱)$$

به‌طوریکه $\sum_t S_t a_t(i) = 1$ برای $1 \le t \le T$.

$b_t(i)$ را هم می‌توان در ضریب S_t برای $1 \le t \le T$ ضرب نمود. حالت بازگشتی را در محاسبه متغیرهای پیشرو و پسرو می‌توان در هر مرحله از زمان t با S_t مقیاس‌گذاری کرد. پس در زمان t ضریب کل مقیاس‌گذاری که به متغیر $a_t(i)$ پیشرو اعمال می‌شود به صورت زیر خواهد بود:

$$Scale_{\ a}(t) = \prod_{k=1}^{t} S_k \qquad (۸۲)$$

$$Scal\, e_b(t) = \prod_{k=t}^{T} S_k \qquad (۸۳)$$

و ضریب‌های مقیاس گذاری مجزا در الگوریتم بازگشتی پیشرو و پسرو درهم ضرب می‌شوند.

با فرض $a'_t(i)$ و $b'_t(i)$ و $x'_t(i,j)$ به عنوان متغیرهای مقیاس‌گذاری متناظر آنها داریم:

$$\sum_t a'_t(i) = Scale_a(T) \sum_i a_T(i) = Scale_{\ a}(T)\ P(X1\Phi) \qquad (۸۴)$$

و احتمال متوسط مقیاس‌گذاری شده $x'_t(i,j)$ را می‌توان به صورت زیر نشان داد:

$$x'_{\ t}(i,j) = (Scal\, e_a(t-1)\, a_{t-1}(i)\, a_{ij}\, b_j(X_t)\, b_{\ t}(j)^{Scal_{\ b}(t)} \Big/ Scal\, e_a(T) \sum_{i=1}^{N} a_T(i)) = x_t(i,j) \qquad (۸۵)$$

پس احتمالات متوسط را می‌توان همانند احتمالات مقیاس گذاری نشده به کار برد.

چرا که ضریب مقیاس گذاری حذف خواهد شد:

$$Scale_{\ a}(t)\, Scale_{\ b}(t) = \prod_{k=1}^{t} S_k \prod_{k=t+1}^{T} S_k = \prod_{t=1}^{T} S_k = Scale_{\ a}(T) \qquad (۸۶)$$

که از صورت و مخرج حذف خواهند شد.

بنابراین فرمول تخمین مجدد به همان صورت خواهد بود. بجز اینکه $P(X|\Phi)$ باید با فرمول زیر محاسبه شود:

$$P(X|\Phi) = \sum_i a'_T(i) / Scale_a(T) \qquad (۸۷)$$

در مواردیکه احتمال پاریز وجود ندارد مقیاس گذاری لازم نیست و $Scale_a$ را یک درنظر می‌گیریم.

یک روش دیگر معمول برای جلوگیری از پاریز استفاده از نمایش لگاریتمی برای همه احتمالات است. این عمل نه تنها باعث می‌شود که مقیاس گذاری غیرضروری باشد و پاریز رخ ندهد بلکه مقادیر صحیح بدست می‌آید و نیازی به عملگرهای ممیز شناور نیست که در عمل برای محاسباتی مانند ویتربی مناسب است.

برای ضرب دو عدد می‌توان لگاریتم آنها را با هم جمع کرد و تقسیم دو عدد نیز معادل تفریق لگاریتم آن دو عدد خواهد بود. فرض کنید در فرمول زیر $P_1 \geq P_2$ باشد:

$$\log_b (P_1 + P_2) = \log_b (b^{\log_b^{p1}} + b^{\log_b^{p2}}) = \log_b\ p1 + \log_b (1 + b^{\log_b^{p2} - \log_b^{p1}}) \quad (76)$$

اگر ترم دوم معادله بالا کوچک باشد جمع بسادگی برابر $\log_b^{p1}$ خواهد بود. برای بکارگیری این روش در تخمین مجدد، کلیه روابط باید بصورت لگاریتمی نوشته شود.

6-4 دنباله های چندین مشاهده ای

برای آموزش دنباله‌های مشاهده چندتایی با فرض استقلال دنباله‌ها از همان الگوریتم پیشرو – پسرو می‌توان استفاده کرد. مسأله اصلی در مدلهای چپ به راست آن است که نمی‌توان مدل را با یک دنباله مشاهده آموزش داد و پارامترهای آن را تخمین زد. بنابراین برای داشتن داده‌های کافی جهت تخمین مطمئن همه پارامترهای مدل، باید از دنباله‌های مشاهده چندتایی استفاده کرد و بنابراین باید روند تخمین مجدد اصلاح گردد.

مجموعه k مشاهده به صورت زیر مشخص است:

$$X = \left[X^{(1)}, X^{(2)}, ..., X^{(k)}\right] \quad (77)$$

بطوریکه $X^{(k)} = (X_1^{(K)}, X_2^{(k)}, ..., X_T^{(k)})$، k امین دنباله مشاهده است و هدف تنظیم پارامترهای Φ برای بیشینه کردن احتمال زیر است:

$$P(X|f) = \prod_{k=1}^{k} P(X^{(k)} \mid \Phi) = \prod_{k=1}^{k} P_k \quad (78)$$

از آنجائیکه معادلات تخمین مجدد براساس فرکانس‌های وقوع حوادث مختلف بنا نهاده شده است فرمولهای تخمین مجدد برای دنباله های چند مشاهده ای با جمع کردن فرکانسهای منفرد وقوع هر دنباله با یکدیگر بدست می‌آیند. بنابراین فرمولهای تخمین مجدد اصلاح شده برای $\overline{b}_j(l), \overline{a}_{ij}$ در حالت گسسته عبارتند از:

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{k} \frac{1}{p_k} \sum_{t=1}^{T_k-1} a_t^k(i)\, a_{ij}\, b_j(X_{t+1}^{(k)})\, b_{t+1}^k(j)}{\sum_{k=1}^{k} \frac{1}{p_k} \sum_{t=1}^{T_k-1} a_t^k(i) b_t^k(i)} \tag{79}$$

$$\bar{b}_j(l) = \frac{\sum_{k=1}^{k} \frac{1}{p_k} \sum_{t=1, S_t=O_l}^{T_k-1} a_t^k(i)\, b_t^k(j)}{\sum_{k=1}^{k} \frac{1}{p_k} \sum_{t=1}^{T_k-1} a_t^k(i) b_t^k(i)} \tag{80}$$

در مدل چپ به راست p_i هم تخمین زده نمی‌شود زیرا $p_1 = 1$ و بقیه p_i ها صفر هستند.

برای مقیاس کردن روابط، هر سری مشاهده دارای ضریب مقیاس خود است و ایده اصلی حذف عامل مقیاس‌بندی از هر ترم قبل از جمع کردن آنها می‌باشد. روابط تخمین مجدد بر حسب متغیرهای مقیاس شده بصورت زیر نوشته می‌شود:

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{k} \frac{1}{p_k} \sum_{t=1}^{T_K-1} Scale_a(t) a_t^k(i) a_{ij} b_j(O_t^k + 1) Scale_b(t)\, b_{t+1}^k(j)}{\sum_{k=1}^{k} \frac{1}{p_k} \sum_{t=1}^{T_K-1} Scale_a(t) a_t^k(i) Scale_b(t) b_t^k(i)} \tag{81}$$

و رابطه $\bar{b}_j(k)$ هم به همان صورت به دست می‌آید.

با در نظر گرفتن چگالی مشاهدات پیوسته، روابط تخمین مجدد برای مخلوطهای گوسی با در دست داشتن چندین رشته مشاهده به صورت زیر خواهد بود:

$$\bar{C}_{ij} = \frac{\sum_{k=1}^{k} \sum_{t=1}^{T_K} x_t^{(k)}(j,l)}{\sum_{k=1}^{k} \sum_{t=1}^{T_k} \sum_{l=1}^{M} x_t^{(k)}(j,l)} \tag{82}$$

$$\bar{m}_{jl} = \frac{\sum_{k=1}^{k} \sum_{t=1}^{T_k} x_t^{(k)}(j,l).O_t^{(k)}}{\sum_{k=1}^{k} \sum_{t=1}^{T_k} x_t^{(k)}(j,l)} \tag{83}$$

$$\bar{\Sigma}_{jl} = \frac{\sum_{k=1}^{k} \sum_{t=1}^{T_k} x_t^{(k)}(j,l)(O_t^{(k)} - m_{j,l})(O_t^{(k)} - m_{jl})^T}{\sum_{k=1}^{k} \sum_{t=1}^{T_k} x_t^{(k)}(j,l)} \tag{84}$$

که در روابط فوق $x_t^{(k)}(j,l)$ به صورت زیر تعریف می‌شود:

$$x_t^{(k)}(j,l)=\left[\frac{Scale_a(t)a_t^{(k)}(j)Scale_b(t)b_t^{(k)}(j)}{\sum_{j=1}^{N}Scale_a(t)a_t^{(k)}(j)Scale_b(t)b_t^{(k)}(j)}\right]\left[\frac{C_{jl}\Pi(O_t^{(k)},m_{jl},\Sigma_{jl})}{\sum_{m=1}^{M}C_{jm}\Pi(O_t^{(k)},m_{jm},\Sigma_{jm})}\right] \quad (85)$$

در روابط بالا k تعداد دنباله‌های آموزش، T_k طول دنباله و به عبارتی تعداد بردارهای مشاهدات دنباله k ام می‌باشد.

5 ملاحظات عملی در استفاده از HMM ها

5-1 تخمین‌های اولیه

می‌دانیم که الگوریتم تخمین مجدد در HMM ها باید برای تابع چگالی احتمال به یک بیشینه محلی برسد. با چه تخمینی ماکزیمم محلی برابرماکزیمم عمومی است؟

در HMM اگر احتمال، صفر فرض شود در حالت گسسته تا آخر صفرخواهد ماند. پس باید یک تخمین مستدل انتخاب گردد. تجربه نشان داده است که برای HMM گسسته می‌توان توزیع یکنواخت را بعنوان تخمین اولیه انتخاب کرد که برای کاربردهای گفتار مناسب است.

اگر HMM های با چگالی مخلوط پیوسته استفاده شود تخمین اولیه مناسب ضروری است. چندین روش برای بدست آوردن اینگونه انتخابهای اولیه وجود دارد:

روش قطعه‌بندی یکنواخت[25]: بر اساس توزیع بردارهای مشاهده در میان حالات مدل مخفی مارکوف استوار است. در مشاهدات گسسته هر کدام از بردارهای مشاهده در یک حالت، یکی از M سمبل موجود در کتاب کد است. تعداد بردارهای موجود در یک رشته مشاهده بر تعداد حالات تقسیم می‌شود و به این ترتیب به هر حالت تعدادی از بردارهای مشاهده رشته مربوطه منتسب می‌شود. همین امر در مورد سایر رشته های مشاهده مربوط به همان نمونه نیز تکرار می‌شود. سپس تعداد در بردارهای مشاهده منتسب به حالت j ام از طریق جمع تعداد بردارهای منتسب به حالت j ام در تمامی رشته‌های مشاهده بدست می‌آید و مقادیر $b_j(k)$ با تقسیم تعداد بردارهای موجود با اندیس k در بردارهای منتسب به حالت j ام بر این مجموع بدست می‌آید. یعنی:

$\hat{b}_j(k) =$ تعداد بردارهای با اندیس k در حالت j تقسیم بر تعداد بردارهای موجود در حالت j

در مشاهدات پیوسته نیز، ابتدا بردارهای مشاهده منتسب به هر حالت j بدست می‌آید و سپس تمامی این بردارها با استفاده از یک الگوریتم خوشه‌بندی kmeans به M خوشه در حالت j ام تقسیم می‌شوند که M تعداد مخلوطها دریک حالت است و بنابراین هر خوشه بیانگر یکی از M مخلوط چگالی $b_j(O_t)$ است. سپس پارامترهای مدل در حالت j ام مطابق روابط زیر تخمین زده می‌شود.

[25] uniform segmentation

نسبت تعداد بردارهای موجود در خوشه m حالت j به تعداد بردارها در حالت j = $\hat{C}_{jm}$

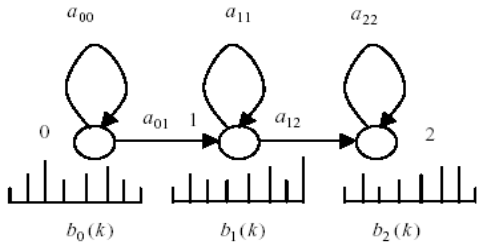
بردار میانگین بردارهای موجود در خوشه m ام حالت j ام = $\hat{m}_{jm}$

ماتریس کوواریانس بردارهای موجود در خوشه m ام حالت j ام = $\hat{\Sigma}_{im}$

به این ترتیب از این مقادیر بعنوان تخمین اولیه پارامترهای مدل مخفی مارکوف با چگالی گسسته یا پیوسته استفاده می‌شود تا بتوان تخمین نهایی مناسبی از این پارامترها به دست آورد.

2-5 توپولوژی مدل

گفتار یک سیگنال غیرایستان است و هر حالت HMM قابلیت نگهداری بعضی قطعات شبه ایستان از سیگنال گفتار را دارد. یک توپولوژی چپ به راست مانند شکل ۵ زیر یک انتخاب خوب برای مدل کردن



سیگنال گفتار می‌باشد که انتقال از هر حالت به خود همان حالت نیز ممکن است.

شکل ۵ **یک مدل مخفی مارکوف معمول برای مدل کردن واج، شامل سه حالت که هر حالت مرتبط با یک توزیع احتمال خروجی می‌باشد.**

زمانیکه یک قطعه گفتار شبه ایستان تولید می‌شود انتقال از چپ به راست نیز همان تولید گفتار را دنبال می‌کند یعنی برای مدل کردن سیگنالهایی که ویژگیهایشان در طول زمان بطور متوالی تغییر می‌کند مناسب است. در صورتیکه مدلهای کاملاً متصل برای این کاربردها خیلی مناسب نیستند. توزیع احتمال خروجی وابسته به حالت می‌تواند توزیع گسسته یا تابع چگالی مخلوط پیوسته باشد. در مدل چپ به راست دنباله‌حالتها از حالت ۱ شروع می‌شود یعنی

$$p_i = \begin{cases} 0 & i \neq 1 \\ 1 & i = 1 \end{cases} \qquad (۸۸)$$

و به حالت N ختم می‌شود و همچنین در ماتریس انتقال وضعیت $a_{ij} = 0 \quad j < i$ و تغییرات بزرگی در

اندیسها هم رخ نمی‌دهد یعنی اینکه پرشهای بیش از چند حالت (معمولاً ۲ حالت) مجاز نمی‌باشد یعنی:

$$a_{ij} = 0 \ , \ \ j > i + 2 \qquad (۸۹)$$

و برای وضعیت نهائی داریم:

$$a_{NN} = 1 \ \ , a_{Ni} = i \ , i < N \qquad (۹۰)$$

قابل ذکر است که تحمیل محدودیتهایی بر مدل چپ به راست تأثیری در روند تخمین مجدد ندارد. در این توپولوژی مهمترین چیز تعیین تعداد حالتها است. انتخاب توپولوژی مدل به داده‌ها آموزش در دسترس و نوع کاربرد بستگی دارد. اگر هر HMM برای نمایش یک واج بکار رود به سه تا پنج توزیع خروجی نیاز داریم. و برای کلمات، حالتهای بیشتری نیاز است. (با توجه به تلفظ و طول کلمه) و همچنین برای مدل کردن سکوت نیز 1 یا 2 حالت کفایت می‌کند. در عمل یک انتقال تهی (null) هم تعریف می‌شود که در مواقعی کاربرد دارد که بخواهیم HMM را بدون دیدن هیچ نمادی پیمایش کنیم.

3-5 ضوابط آموزش:یکنواخت کردن[26] پارامترها

اگر داده‌های آموزش محدود باشد، باعث می‌شود بعضی پارامترها بصورت ناقص آموزش داده شوند و یا آموزش داده نشوند و دسته‌بندی بر اساس مدلهای آموزش داده شده ضعیف، باعث افزایش میزان خطا خواهد شد. بعنوان مثال میدانیم محاسبه $b_j(k)$ نیازمند شمارش دفعات بودن در حالت j و مشاهده سمبل O_k بطور همزمان است. اگر دنباله آموزش آنقدر کوچک باشد که در آن چنین حادثه ای (مثلاً $S_t=j$ و $X_t=O_k$) رخ ندهد bj(k) برابر صفرخواهد بود و پس از تخمین مجدد هم صفرمی‌ماند و مدل حاصل برای هر دنباله مشاهده شامل ($S_t=j$ و $X_t=O_k$) احتمال صفر تولید خواهدکرد.

اولین راه حلی که به نظر می‌رسد افزایش مجموعه مشاهدات آموزشی است که همیشه ممکن نیست. دومین راه حل کاهش اندازه مدل مانند تعداد حالتها و تعداد سمبلها در هر حالت می‌باشد که همیشه ممکن نیست و اغلب دلایل فیزیکی مشخصی برای استفاده از یک مدل مفروض وجود دارد. راه حل سوم این است که می‌توان تخمین های یک مجموعه از پارامترها را با تخمینهای مجموعه دیگر از پارامترها درون یابی کرد. راه حل چهارم این است که می‌توان برای کم کردن تعداد پارامترها، آنها را به یکدیگر مرتبط نمود که در HMM شبه پیوسته از این تکنیک استفاده می‌شود. در اکثر این روشها باید از حد آستانه برای پارامترها استفاده کرد تا این اطمینان حاصل شود که تخمین هیچ یک از پارامترهای مدل از

[26] .smoothing

یک سطح کمتر نمی‌شود. مثلا برای یک مدل با مشاهدات گسسته داریم:

$$b_j(k)=\begin{cases} b_j(k) & , \quad b_j(k)\geq d_b \\ d_b & , \quad b_j(k)<d_b \end{cases} \qquad (۹۱)$$

و برای مدلی با توزیع پیوسته علاوه بر آنکه ماتریس کوواریانس هر مخلوط از نوع قطری انتخاب می‌شود تا تأثیر داده های آموزش ناکافی را خنثی نماید از شرط زیر نیز استفاده می‌شود: (مخصوصاً اگر همبستگی ضرایب ضعیف باشد مثل روش mfcc که ویژگیهای ناهمبسته دارد).

$$\Sigma_{jk}(r,r)=\begin{cases} \Sigma_{jk}(r,r) & , \quad \Sigma_{jk}(r,r)\geq d_I \\ d_I & \quad \Sigma_{jk}(r,r)<d_u \end{cases} \qquad (۹۲)$$

پس از اعمال شرایط فوق در روابط تخمین مجدد، باید تمامی پارامترهای باقی مانده طوری تغییر مقیاس داده شوند که چگالیها از محدودیتهای آماری لازم پیروی نمایند.

در عمل، در صورتیکه داده های آموزش ناکافی داشته باشیم درصد خطای بازشناسی گفتار بین ۵ تا ۲۰ درصد با روشهای مختلف یکنواخت کردن پارامترها، کم می‌شود.

6 روشهای جستجو

6-1 مقدمه

به هر روند جستجو می‌توان به صورت عبور از یک گراف جهت دار نگریست که در آن هر گره معرف یک وضعیت مسأله است و هر شاخه معرف رابطه بین وضعیت هایی است که توسط گره‌های دو سر شاخه ارائه می‌شود.

در طی جستجو باید مسیری در درون گراف یافت شود به طوریکه این مسیر از وضعیت آغازی شروع و در وضعیت‌های نهایی خاتمه یابد. گراف جستجو را می‌توان به طور کامل به کمک قواعدی که حرکتهای مجاز را تعریف می‌کنند، ساخت اما در عمل قسمت زیادی از گراف اصلاً ساخته نمی‌شود. اکثر برنامه‌های جستجو کار جالبی می‌کنند، آنها به جای آنکه اول به طور صریح گراف را بسازند وسپس جستجو کنند، گراف را به صورت غیر صریح توسط قواعد ارائه می‌کنند و به طور صریح فقط قسمتهایی از گراف را که مورد جستجو قرار می‌گیرند می‌سازند.

پیش از بحث در مورد تک تک روشها به دو نکته زیر که در همه تکنیکها بروز می‌کند توجه می‌کنیم:

- جهت اجرای جستجو
- استفاده از یک تابع هیوریستیک جهت پیش بردن و راهنمایی جستجو

6-2 استدلال جلو رو در مقابل عقب‌رو

هدف روال جستجو، کشف یک مسیر از میان فضای مسأله از یک وضعیت آغازی به وضعیت هدف است.چنین جستجویی می‌تواند در دو جهت حرکت کند:

- به طرف جلو، از وضعیت‌های آغازی
- به طرف عقب، از وضعیت‌های هدف

این دو روش قرینه یکدیگر هستند. فاکتورهای زیر ما را برای انتخاب جهت جستجو راهنمایی می‌کنند:

تعداد وضعیت‌های آغازی بیشتر است یا هدف؟ ما همیشه مایلیم که از تعداد کمتر وضعیت‌ها به طرف تعداد بیشتر برویم.

در کدام جهت فاکتور شاخه شاخه شدن بزرگتر است ؟ (این فاکتور تعداد متوسط گره‌هایی است که مستقیماً می‌توان از یک گره تک به آنها رسید) ما حرکت در جهتی که این فاکتور کوچکتر است را ترجیح می‌دهیم.

3-6 توابع هیوریستیک

یک هیوریستیک عبارت است از تکنیکی که با کشف راه حل‌هایی به مسائل کمک می‌کند اما تضمین نمی‌کند که به بیراهه منتهی نگردد. هیوریستیکهایی وجود دارند که موارد کاربرد عمومی و زیادی دارند ولی برخی از آنها فقط کاربردهای خاصی دارند. استراتژیهای زیر همگی همه منظوره هستند اما برای اینکه اینها در یک قلمرو به خصوص به درستی و خوب عمل کنند باید آنها را با هیوریستیکهایی که ویژه محسوب می‌شوند همراه کرد. یک روش برای این کار استفاده از تابع هیوریستیک است که هر وضعیت مسأله را بررسی می‌کند و میزان خوبی ومناسب بودن آن را تعیین می‌کند.

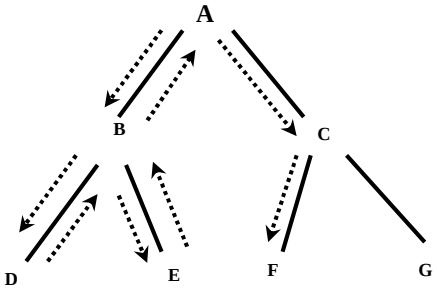
یک تابع هیوریستیک تابعی است که توصیف وضعیت مسأله را بر میزانهای مناسب بودن نگاشت می‌کند. جستجوی هیوریستیک ابزاری قوی برای حل مسائل مشکل است. استراتژیی که جهت کنترل اینگونه جستجو به کار می‌رود اهمیت به سزایی در تعیین میزان موفقیت این روش دارد. در زیر به برخی از این استراتژیهای کنترل اشاره خواهد شد.

- جستجو ی اول عمق[27]
- جستجوی اول پهنا[28]
- جستجوی اول بهترین[29]

۶-۳-۱ روش جستجوی اول عمق

در جستجوی اول عمق ابتدا هر مسیر احتمالی تا رسیدن به نتیجه (یا هدف) بررسی شده و سپس به مسیر دیگر پرداخته می‌شود. به منظور درک دقیق این شیوه جستجو، درخت زیر که در آنF هدف است را در نظر بگیرید:

[27] Depth first
[28] Breadth first
[29] Best first

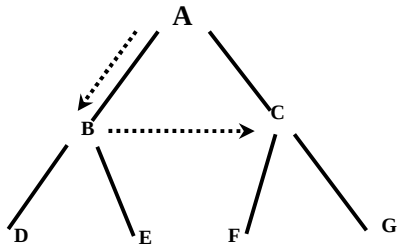


شکل ۶ **روند جستجو در شیوه اول عمق**

در یک جستجو به شیوه اول عمق این نمودار با ترتیب ABDBEBACF بررسی می‌شود. در این بررسی آنقدر به سمت چپ می‌رویم تا اینکه به یک گره انتهایی یا هدف برسیم. اگر به گره انتهایی رسیده باشیم، آنگاه یک گره به سمت عقب برگشته، به سمت راست می‌رویم و سپس تا هنگامی که یا به هدف و یا گره انتهایی دیگر برسیم به سمت چپ می‌رویم و این عمل تا هنگامیکه یا به هدف برسیم و یا اینکه آخرین گره موجود در فضای جستجو نیز بررسی شود ادامه می‌یابد. یک جستجو انجام شده به شیوه اول عمق قطعاً به جواب می‌رسد زیرا در بدترین حالت، این جستجو به یک جستجوی همگانی تبدیل می‌شود. در نمودار فوق اگر هدف G باشد یک جستجوی همگانی خواهیم داشت.

۶-۳-۲ روش جستجوی اول سطح

شیوه جستجوی اول سطح درست عکس شیوه جستجوی اول عمق است. در این روش تمام گره‌هایی که در یک سطح درخت هستند، پیش از بررسی گرهی در دیگر سطح مورد امتحان قرار می‌گیرند. در زیر این شیوه بین گره‌ها ارائه شده است (در اینجا هدف C است):



شکل ۷ **روند جستجو در شیوه اول عمق**

این نمایش نشان می‌دهد که جستجو ابتدا با گره‌های ABC برخورد می‌کند. همانند شیوه جستجوی اول عمق، روال جستجوی اول پهنا نیز در صورت وجود جواب تضمین می‌کند که آن را نهایتاً بیابد اما مشکلاتی هم دارد:

- احتیاج به حافظه زیادی دارد.
- احتیاج به کار زیادی دارد، به خصوص اگر مسیر حل طولانی باشد.

در مواردی که چندین راه حل طولانی وجود دارد بهتر است از همان روش اول عمق استفاده شود و نه از اول پهنا تا بتوان سرعت بیشتری به دست آورد.

6-4 ارزیابی روشهای جستجو

ارزیابی قابلیت اجرایی یک سیستم جستجو می‌تواند عملی بسیار پیچیده و دشوار باشد در واقع این ارزیابی بخش عمده‌ای از تحقیقات هوش مصنوعی را به خود اختصاص داده است. با این وجود در توصیف ما از این ارزیابی، دو اندازه‌گیری اولیه از اهمیت خاصی برخوردارند:

- یک جستجو با چه سرعتی یک راه حل را می‌یابد؟
- صحت این راه حل چقدر است ؟

انواع مختلفی از مسائل وجود دارند که در مورد آنها همه موضوعات در این نکته خلاصه می‌شود که یک راه حل (هر چه که می‌خواهد باشد) را با حداقل کوشش به دست آورید. در این نوع از مسائل، اندازه‌گیری مورد اول اهمیت دارد با این وجود در انواع دیگری از مسائل آنچه که مهم است آن است که راه حل یافت شده به راه حل بهینه تا حد ممکن نزدیکتر باشد. درک تفاوت بین یافتن یک راه حل «بهینه» و یافتن یک راه حل «خوب» مهم است. این تفاوت بر اساس این واقعیت به وجود می‌آید که یافتن یک راه حل بهینه اغلب اوقات مستلزم انجام یک « جستجوی همگانی » است زیرا این تنها شیوه‌ای است که به وسیله آن می‌توان تعیین کرد که آیا بهترین راه حل یافت شده است یا نه. با این وجود یافتن یک راه حل خوب به معنی یافتن راه حلی است که در داخل مجموعه راه‌حل‌ها قرار می‌گیرد و ممکن است یک راه حل بهتر از آن نیز وجود داشته باشد.

کلیه تکنیکهای جستجویی که ارائه شد در برخی مسائل مشکل، نسبت به یکدیگر دارای برتریهایی هستند. از این رو از نظر کلی به دشواری می‌توان گفت که یک شیوه جستجو «همواره» از شیوه‌ای دیگر بهتر است. با این وجود برخی از این تکنیکها به طور متوسط نسبت به ما بقی احتمال بهتر بودن بیشتری

دارند. همچنین خود شیوه‌ای که یک مسأله توسط آن تعریف می‌شود نیز در برخی موارد یک شیوه جستجوی مناسب را پیش روی ما قرار می‌دهد.

7 مؤلفه‌های فضای جستجو در سیستمهای بازشناسی گفتار

7-1 مقدمه

مدل آکوستیکی، مدل زبانی و درخت واژگان از اصلی‌ترین مؤلفه‌های تشکیل دهنده‌ی فضای جستجو در سیستمهای بازشناسی گفتار هستند که می‌توانند در محدود کردن جستجو و افزایش سرعت آن، نقش مهمی را ایفاء کنند.

7-2 مدل آکوستیکی

یک فرض کلیدی در پردازش تصادفی گفتار این است که سیگنال گفتار در محدوده‌های زمانی کوتاه، ایستان است. با این فرض بخش استخراج ویژگی، سیگنال گفتار پیوسته را به یک دنباله از بردارهای ویژگی تبدیل می‌کند. هدف این بخش از سیستم بازشناسی گفتار، ارائه تغییرات زمانی و طیفی سیگنال گفتار به صورت دنباله‌ای از بردارها می‌باشد.

بعد از استخراج دنباله‌ی بردارهای ویژگی X از سیگنال گفتار، لازم است که به کمک مدل‌های آکوستیکی مقدار احتمال یا امتیاز $P(X|W)$ محاسبه شود. در کاربردهایی که تعداد کلمات واژگان زیاد می‌باشد، محاسبه $P(X|W)$ برای تک تک کلمات موجود در بانک، غیرعملی می‌باشد. بنابراین دنباله کلمات به واحدهای صوتی پایه‌ای بنام واج شکسته می‌شوند.

برای هر واج یک HMM در نظر گرفته شده و به کمک داده‌های موجود در دادگان آموزشی، آموزش داده می‌شود. HMM یک ماشین حالت متناهی است که انتقال بین حالت‌های آن بر اساس یک توزیع مارکوف است و یک تابع چگالی احتمال، خروجی هر حالت آن را مدل می‌کند، این تابع نقش مهمی را در مدل کردن تغییرات طیفی سیگنال گفتار بازی می‌کند. بنا به پیچیدگی مساله‌ی بازشناسی، تابع چگالی احتمال به صورت پیوسته یا گسسته مدل می‌شود.

7-3 مدل زبانی

مدل زبانی، محدودیت‌هایی روی دنباله‌های کلمات مورد بازشناسی اعمال می‌کند و در واقع مکانیزمی

برای تخمین احتمال رخداد کلمات در یک دنباله از کلمات با فرض مشخص بودن دنباله کلمات پیرامون آن است. مدل زبانی به طور ضمنی دانش زبانی، دانش دامنه و هر گونه اطلاعات دیگری را در جهت کاهش فضای جستجو، جمع می‌کند. میزان محدودیتی که توسط مدل زبانی در یک سیستم بازشناسی ایجاد می‌شود، پیچیدگی آن مدل زبانی نامیده می‌شود و در واقع بیانگر میانگین فاکتور انشعاب است.

از آنجایی که احتمال بیان یک کلمه در هر مرحله از گفتار، در اکثر مواقع به کلمات بیان شده‌ی قبلی وابسته است، یک راه ساده و مؤثر برای به کارگیری مدل زبانی این است که هر دنباله‌ی M کلمه‌ای را توسط یک زنجیره‌ی مارکوف مرتبه n ام مدل کنیم که به اصطلاح N-gram نامیده می‌شود.

7-4 درخت واژگان

درخت واژگان، یکی از اساسی‌ترین اجزاء یک سیستم بازشناسی گفتار با واژگان زیاد می‌باشد. دو راه برای توصیف واجی واژگان در بازشناسی گفتار پیوسته وجود دارد: یکی ارائه خطی است که چندان مؤثر و کارا نیست چون از شباهت واجی بین کلمات سودی نمی‌برد و توصیف واجی هر واژه به صورت مستقل از بقیه ارائه می‌شود. پس فضای جستجو با افزایش تعداد واژگان به صورت خطی افزایش پیدا می‌کند. مثلاً اگر دو کلمه «برادر» و «برادران» که اولی پیشوند دومی می‌باشد، جزء واژگان باشد، توصیف واجی هر کدام به صورت مستقل ارائه می‌شود که در صورت زیاد بودن اینچنین کلماتی مقدار زیادی حافظه به هدر می رود.

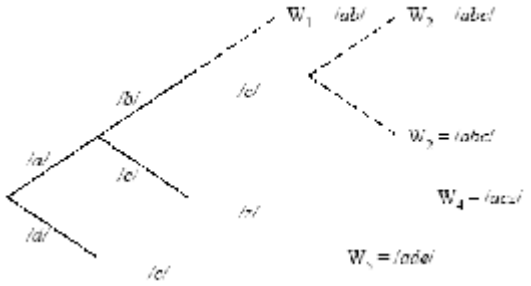
روش دوم ارائه توصیف واجی واژگان به صورت درختی می‌باشد که به اختصار درخت واژگان[30] یا درخت پیشوندی[31] نامیده می‌شود و در مواردی که تعداد واژگان زیاد باشد، متداول‌ترین روش محسوب می‌شود. جستجو بر مبنای درخت واژگان یک فاکتور اساسی برای ایجاد یک سیستم بازشناسی گفتار بلادرنگ با تعداد واژگان زیاد می‌باشد.

شکل ۸ نمونه‌ای از یک درخت واژگان را نشان می‌دهد. همان‌طور که دیده می‌شود واج‌های مشابه ابتدایی به اشتراک گذاشته شده‌اند. هر برگ این درخت متناظر با یک کلمه در واژگان می‌باشد. به دو دلیل زیر یک یال تهی، که به صورت خط‌چین نشان داده شده است، برای ارائه گره برگی هر کلمه به کار گرفته می‌شود:

وقتی توصیف واجی یک کلمه پیشوند کلمات دیگر می‌باشد. یال تهی به عنوان یک شاخه برای اتمام کلمه پیشوند عمل می‌کند.

[30] Lexical Tree
[31] Prefix Tree

وقتی در واژگان کلماتی با توصیف واجی یکسان وجود دارد، مانند دو کلمه "two" و "to". یالهای تهی برای اشاره به این کلمات استفاده می‌شود.



شکل ۸ **نمونه‌ای کوچک از یک درخت واژگان**

یک مزیت عمده استفاده از درخت واژگان این است که می‌تواند فضای جستجو را به طرز چشمگیری محدود کند.

8 الگوریتم‌های جستجو برای بازشناسی گفتار

8-1 مقدمه

در سیستم‌های بازشناسی گفتار، دیکدر، اساساً یک فرایند جستجو برای آشکار کردن دنباله کلماتی $\hat{W} = w_1 w_2 .. w_m$ است که بیشترین احتمال پسین[32]P(W|X) را برای دنباله مشاهدات داده شده $X = x_1 x_2 ... x_n$ داشته باشد.

$$\hat{W} = \arg\max_w P(W \mid X) = \arg\max_w \frac{P(W)P(X \mid W)}{P(X)} = \arg\max_w P(W)P(X \mid W) \qquad (۹۳)$$

یک راه بدیهی آن است که همه دنباله کلمه‌های ممکن را جستجو کرده و سپس کلمه با بیشترین مقدار $P(W)P(X \mid W)$ را انتخاب کنیم.

وقتی مدل‌های کلمات موجود باشد، بازشناسی گفتار به یک مسأله جستجو تبدیل می‌شود که هدف آن یافتن دنباله‌ای از مدل‌های کلمات است که به بهترین صورت، شکل موج ورودی را در برابر مدلهای لغات موجود بیان می‌کند. با توجه به اینکه تعداد کلمات و واج‌ها و همچنین مرزهای بین آنها در شکل موج ورودی مشخص نیست، استفاده از روش جستجویی مناسب که از عهده این الگوهای پویا با طول متغیر برآید، اهمیت زیادی دارد.

وقتی از HMM در سیستم بازشناسی گفتار استفاده می‌شود، حالتهای HMM ها فضای جستجو را تشکیل می‌دهند ولی علاوه بر این مؤلفه‌های دیگری هم در میزان اندازه فضای جستجو نقش دارند. در ادامه ابتدا با مؤلفه‌های فضای جستجو در بازشناسی گفتار آشنا می‌شویم و پس از آن روشهای جستجو در بازشناسی گفتار مطرح می‌شوند. اگرچه روشهایی که بیان خواهد شد بر مبنای HMM توضیح داده می‌شوند اما همه این روشها را می‌توان در سیستم‌هایی که از تکنیکهای مدل‌سازی دیگری استفاده کرده‌اند نیز به‌کار برد. درحقیقت بسیاری از این روشهای جستجو قبل از اینکه از HMM برای بازشناسی گفتار استفاده شود، کشف شده بودند.

[32] Maximum posterior probability

2-8 جستجوی ویتربی

وقتی از HMMها برای مدل آکوستیک استفاده می‌شود، از امتیاز[33] مدل آکوستیک طبق تعریف احتمال پیش‌رو[34] استفاده می‌شود. بنابراین همه دنباله حالت‌ها باید ملاحظه شوند:

$$P(X\ |W) = \sum_{\text{all possible } S_0^T} P(X, S_0^T \mid W) \qquad (۹۴)$$

که در رابطه فوق مجموع همه حالت‌های ممکن S برای یک دنباله کلمه مورد نظر W، حساب می‌شود.

الگوریتم پیش‌رو احتمال اینکه یک HMM، یک دنباله مشاهده را تولید کند، با جمع کردن احتمال‌های همه مسیرهای ممکن حساب می‌کند اما بهترین مسیر (یا دنباله حالت) را مشخص نمی‌کند، درحالیکه در بسیاری از کاربردها یافتن چنین مسیری مطلوب است. واقعیت این است که یافتن مسیر بهینه (دنباله حالت بهینه) اساس جستجو در بازشناسی گفتار پیوسته می‌باشد. از آنجا که دنباله حالت در ساختار HMM پنهان است (مشاهده نمی‌شود) و با توجه به اینکه هدف دیکدینگ آشکار کردن بهترین دنباله کلمه است، می‌توان مجموع را با بیشینه‌ای برای یافتن بهترین دنباله حالت تقریب زد. معیاری که اغلب استفاده می‌شود، یافتن دنباله حالتی است که هنگام تولید دنباله مشاهده بیشترین احتمال را به دست آورد. طبق قانون بیز، معادله (۹۴) به صورت زیر درمی‌آید:

$$\hat{W} = \arg\max_w P(W)P(X\ |W) \cong \arg\max_w \left\{P(W) \max_{S_0^T} P(X, S_0^T\ |W)\right\} \qquad (۹۵)$$

معادله (۹۵) اغلب با عنوان «تقریب ویتربی» شناخته می‌شود که می‌توان آن را چنین بیان کرد:

«محتمل‌ترین دنباله کلمه» با «محتمل‌ترین دنباله حالت» تقریب زده می‌شود. اگرچه اصولاً نتایج جستجو به روش احتمال پیش‌رو و روش تقریب ویتربی می‌توانند متفاوت باشند، اما در عمل این وضعیت به ندرت اتفاق می‌افتد.

به الگوریتم ویتربی می‌توان به عنوان یک الگوریتم برنامه‌ریزی پویا که به HMM اعمال می‌شود و یا یک الگوریتم پیش‌روی اصلاح شده نگاه کرد. الگوریتم ویتربی به جای جمع کردن احتمال‌های مسیرهای مختلفی که به یک حالت نهایی می‌رسند، بهترین مسیر(بهترین دنباله حالت) را برمی‌گزیند و به خاطر می‌سپارد. این الگوریتم دنباله حالتی $S = (s_1, s_2, ..., s_t)$ را جستجو می‌کند که مقدار $P(S, X \mid \Phi)$ را ماکسیمم سازد. این مسأله شباهت زیادی به مسأله یافتنِ مسیر بهینه در برنامه‌ریزی پویا دارد.

برای یافتن بهترین دنباله حالت $S = (s_1, s_2, ..., s_t)$ برای دنباله مشاهده داده شده

[33] score
[34] forward

$X = (x_1, x_2, ..., x_t)$ لازم است کمیت $d_t(i)$ را تعریف کنیم:

$$d_t(i) = \max_{s_1, s_2, ..., s_t} P\left[s_1, s_2, ..., s_{t-1}, s_t = i, x_1, x_2, ..., x_t \mid \Phi\right] \quad (۹۶)$$

$d_t(i)$در طول یک مسیر بالاترین امتیازی است که برای اولین t مشاهده که در حالتi پایان می‌یابد، به دست می‌آید. با این مقدمه داریم:

$$d_{t+1}(i) = \max_i \left[d_t(i) a_{ij}\right] b_j(x_{t+1}) \quad (۹۷)$$

برای بازیابی دنباله حالت، لازم است برای هر t و j آرگومانی که $d_t(i)$ را ماکسیمم می‌کند، نگه داریم. برای ذخیره آرگومان به آرایه‌ای $y_t(i)$ در الگوریتم نیاز است.

رویه استقراء برای الگوریتم ویتربی به صورت زیر می‌باشد:

مرحله اول: مقداردهی اولیه

$$d_t(i) = p_i b_i(x_1) \qquad 1 \le i \le N \ (states)$$
$$y_1(i) = 0 \quad (۹۸)$$

مرحله دوم: تکرار

$$d_t(j) = \max_{1 \le i \le N} \left[d_{t-1}(i) a_{ij}\right] b_j(x_t) \qquad 2 \le t \le T, \ 1 \le j \le N \quad (۹۹)$$

$$y_t(j) = Arg \ \underset{1 \le i \le N}{Max} \left[d_{t-1}(i) a_{ij}\right] \quad (۱۰۰)$$

مرحله سوم: اتمام

$$\text{بهترین امتیاز} = \underset{1 \le i \le N}{Max} \left[d_t(i)\right] \quad (۱۰۱)$$

$$S_T^* = Arg \ \underset{1 \le i \le N}{Max} \left[B_T(i)\right] \quad (۱۰۲)$$

مرحله چهارم: مرحله پسرو

$$S_T^* = y_{t+1}(S_{t+1}^*) \qquad t = T-1, \ T-2, \ \ldots, 1 \quad (۱۰۳)$$

بهترین دنباله می‌باشد $S^* = (S_1^*, S_2^*, ..., S_T^*)$

الگوریتم ویتربی در هر مرحله همه فرضیه‌ها را به طور موازی مورد پردازش قرار می‌دهد و به مرحله بعد

بسط می‌دهد. بنابراین یک جستجوی همزمان[35] محسوب می‌شودکه قبل از رفتن به زمان t+1 همه محاسبات مربوط به زمان t را انجام می‌دهد. برای زمان t، هر حالت با بهترین امتیاز از بین امتیازات همه حالات در زمان t-1 (به‌جای مجموع امتیازات همه مسیرهایی که وارد می‌شوند) به‌روز[36] می‌شود. به همین دلیل است که معمولاً به این روش «جستجوی ویتربی همزمان» می‌گویند. وقتی به روز رسانی انجام می‌شود، اشاره‌گرِ backtracking نیز ذخیره می‌شود تا حالت ورودی با بیشترین احتمال را به‌خاطر داشته باشیم. در پایان جستجو، بهترین دنباله حالت را می‌توان با ردگیری این اشاره‌گر پیدا کرد. در الگوریتم ویتربی چون همه فرضیه‌ها متعلق به یک دنباله‌ی ورودی هستند و از لحاظ زمانی با هم برابرند، می‌توان آنها را به طور مستقیم با هم مقایسه کرد. این الگوریتم راه حلی بهینه برای مسایل پیچش زمانی غیرخطی بین مدلهایHMM و مشاهدات آکوستیکی، تشخیص محدوده کلمه و تشخیص کلمه در بازشناسی گفتار پیوسته ارائه می‌کند.

3-8 جستجو در بازشناسی کلمات مجزا

در بازشناسی کلمات مجزا محدوده کلمات مشخص است. اگر مدلهای HMM کلمات موجود باشد، احتمال مدل آکوستیک $P(X|W)$ را می‌توان با الگوریتم پیش‌رو[37] حساب کرد. جستجو به یک مسأله ساده تشخیص الگو تبدیل می‌شود و کلمه‌ای $\hat{W}$ که بیشترین احتمالِ پیش‌رو را داشته باشد به عنوان کلمه بازشناسی شده انتخاب می‌شود. وقتی از مدل‌های زیرکلمه‌ای استفاده می‌شود، مدل‌های HMM کلمات با اتصال مدل‌های HMMِ زیرکلمه‌های متناظر بدست می‌آید.

4-8 جستجو در بازشناسی گفتار پیوسته

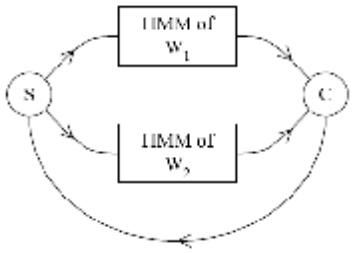
جستجو در بازشناسی گفتار پیوسته نسبتاً پیچیده است، حتی برای یک مجموعه لغات کوچک، چرا که الگوریتم جستجو بایستی احتمال شروع هر کلمه در هر فریم زمانی را درنظر بگیرد. در برخی از سیستمهای بازشناسی گفتار پیوسته، برای بازشناسی گفتار پیوسته یک روش دو مرحله‌ای درنظر گرفته می‌شد، ابتدا محدوده‌های کلمات فرض می‌شد و سپس از تکنیکهای انطباق الگو برای بازشناسی الگوهای قسمت شده استفاده می‌شد. اما این قسمت‌بندی و تعیین محدوده لغات چندان قابل اعتماد نبود.

در اینجا با مثال ساده‌ای (شکل ۹) نشان می‌دهیم که چگونه می‌توان روش‌های جستجوی کلمات مجزا

[35] Synchronous
[36] update
[37] forward

را به بازشناسی گفتار پیوسته گسترش داد. این سیستم تنها شامل دو کلمه می‌باشد و فرض می‌کنیم که مدل زبانی استفاده شده در اینجا یک تک‌گرام یکسان[38] باشد $P(W_1)=P(W_2)=1/2$.

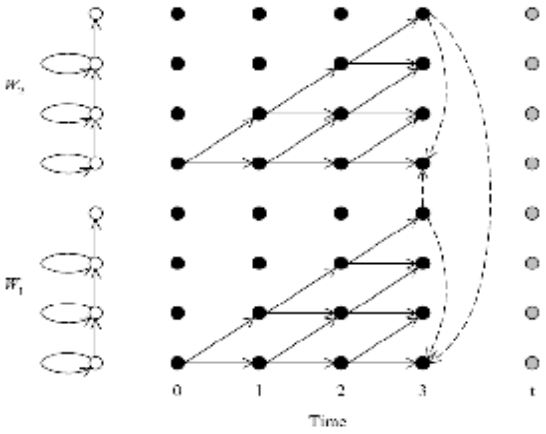
فراهم کردن ساختارهای زبان در همان چارچوبِ HMM اهمیت دارد. در شکل ۹ یک حالت شروع S و یک حالت جمع کننده[39] C اضافه کرده‌ایم. حالت شروع، یک انتقال پوچ[40] به حالت ابتدایی هریک از کلمات دارد که احتمال این انتقال‌ها براساس مدل زبانی مشخص می‌شود (در این مثال = ½). حالت خروجی هریک از کلمات نیز یک انتقال پوچ به حالت C دارد. همچنین حالت C یک انتقال پوچ به حالت شروع S دارد تا امکان تکرار وجود داشته باشد. مشابه با حالت جاسازی HMMهای فونم‌ها در HMMِ کلمه برای بازشناسی کلمات مجزا، می‌توان HMMهای کلمات W_1 و W_2 را در یک HMM جدید متناظر با ساختار شکل ۹ تعبیه نمود.



شکل ۹ **مثال ساده‌ای از عملیات بازشناسی گفتار پیوسته با دو کلمه W_1 و W_2. حالت S حالت شروع و حالت C یک جمع کننده حالت است تا پیوند کاملی بین لغات وجود داشته باشد.**

ساختار HMMِ ترکیبی که در شکل ۹ نشان داده شده است را می‌توان به عنوان یک شبکه آماری حالت نهایی[41] با احتمالات انتقال و توزیعات خروجی درنظر گرفت. الگوریتم جستجو اساساً یک انطباق بین مشاهدات X و مسیر (در اینجا یعنی دنباله حالات و انتقالها) در شبکه حالت نهایی تولید می‌کند. برخلاف بازشناسی گفتار مجزا، در بازشناسی گفتار پیوسته لازم است دنباله کلمه بهینه $\hat{W}$ مشخص شود. یک انتخاب بدیهی برای چنین کاری، الگوریتم ویتربی می‌باشد که دنباله حالت بهینه $\hat{S}$ به دنباله کلمه بهینه $\hat{W}$ متناظر می‌شود. شکل ۱۰ محاسبه شبکه ویتربی HMM را برای بازشناسی گفتار پیوسته دو کلمه شکل ۱۰ نشان می‌دهد.

[38] uniform unigram
[39] collector state
[40] null
[41] stochastic finite state



شکل ۱۰ **شبکه HMM برای مثالِ بازشناسی گفتار پیوسته در شکل ۸. وقتی حالت نهایی مدل یک کلمه ملاقات شد، یک کمان پوچ (که در شکل با خطوط منقطع مشخص شده است) از آن حالت به حالت ابتدایی همه کلمات پیوند می‌شود.**

5-8 نقش اشاره‌گر backtracking

در این‌جا لازم است نقش اشاره‌گر backtracking در جستجوی ویتربی برای بازشناسی گفتار پیوسته روشن شود. بطور کلی در بازشناسی گفتار ما به دنبال یافتن بهترین دنباله حالت نمی‌باشیم[-] در عوض ما فقط می‌خواهیم دنباله کلمه بهینه که با معادله ۹۳ مشخص می‌شود را پیدا کنیم. بنابراین ما فقط به‌خاطر یادآوری سابقه کلمه برای مسیر جاری از اشاره‌گر backtrack استفاده می‌کنیم و به این ترتیب دنباله کلمه بهینه در پایان جستجو قابل کشف می‌باشد. برای توضیح بیشتر، وقتی ما به حالت پایانی یک کلمه رسیدیم، یک نودِ سابقه[42] ایجاد می‌کنیم که شامل شناسه کلمه و شاخص زمان جاری[43] می‌باشد و این نود به اشاره‌گر فعلی backtracking ضمیمه می‌شود. سپس این اشاره‌گر به نود قبلی‌اش پاس می‌شود و به همین ترتیب پیش می‌رود تا دنباله مشخص شود. یک مزیت جانبی برای نگه‌داری این اشاره‌گر این است که دیگر ما نیازی به نگهداری کل شبکه در زمان جستجو نداریم بلکه فقط به فضایی برای نگهداری دو قسمت زمانی متوالی (ستونها) در محاسبه شبکه نیاز داریم (تکه زمانی قبلی و تکه زمانی جاری) چرا که همه اطلاعات backtracking اکنون در اشاره‌گر backtracking نگهداشته می‌شود. این ویژگی مزیت ممتازی برای پیاده‌سازی یک جستجوی همزمان ویتربی می‌باشد.

[-] البته با وجود اینکه ما به دنبال یافتن دنبالهٔ حالت‌هاي بهينه در ASR نيستيم اما آنها براي تعيين قطعه‌بندي فونتيك كاملاً مورد استفاده مي‌باشند و مي‌توانند اطلاعات مهمي براي توسعهٔ سيستم‌هاي ASR فراهم كنند.

[42] history
[43] time index

9 مقاوم سازی بازشناسی گفتار

9-1 مقدمه

دقت سیستمهای بازشناسی گفتار به هنگام عمل در محیطهای آکوستیکی مختلف که با یکدیگر ناسازگارند، به شدت کاهش می‌یابد. این کاهش را می‌توان ناشی از عدم تطبیق میان داده‌های گفتار آموزشی و داده‌های آزمایش در محیطهای آکوستیکی مختلف دانست. این عدم تطبیق می‌تواند ناشی از چند عامل باشد: آلودگی با نویز(جمع پذیر، انعکاسی[44] و کانال)، تغییر شیوه صحبت (سرعت بیان گفتار و اثر لومبارد[45])و یا تغییر گوینده (تغییر کیفیت صدا، فرکانس اصلی، جنسیت و لهجه گوینده). از این رو در سالهای اخیر روشهای متعددی برای مقاوم سازی بازشناسی گفتار و کاهش عدم تطبیق میان شرایط آموزش و آزمایش مطرح گردیده‌اند که از آن جمله می‌توان به بهبود سیگنال گفتار[46]، استفاده از ویژگیهای مقاوم[47] و نرمال سازی ویژگیها، استفاده از آرایه‌های میکروفنی، استفاده از خصوصیات شنوایی و گوش انسان و تطبیق و دگرگونی مدل گفتار اشاره کرد.

افزایش و بهبود دقت بازشناسی بوسیله روشهای فوق با محدودیتهایی مواجه است که بخشی از این محدودیتها به علت نارسایی مدل ریاضی است که برای مشخص کردن تخریب ویژگیهای آکوستیک بکار می‌روند. به این منظور باید مدلی ساده یافت که چگونگی تاثیر محیط را بر پارامترهای مشخصه سیستمهای بازشناسی گفتار و ورودی آنها نشان دهد.

روشهای پیشنهادی برای توصیف اثر تخریب محیط بر گفتار را می‌توان به دو دسته کلی تقسیم کرد: شیوه‌های مبتنی بر داده[48] و شیوه‌های مبتنی برمدل[49]. روشهای مبتنی بر داده سعی می‌کنند که تاثیر محیط بر مشخصه‌های گفتار را مشخص نمایند و خود سیگنال گفتار یا پارامترها و ویژگیها را بهبود بخشند و اصلاح نمایند که برخی از این روشها به این منظور نیازمند به وجود هر دو نوع سیگنال گفتار تمیز و نویزی هستند. شیوه‌های مبتنی برمدل تلاش می‌کنند که مدل ریاضی محیط را اصلاح نمایند و

---

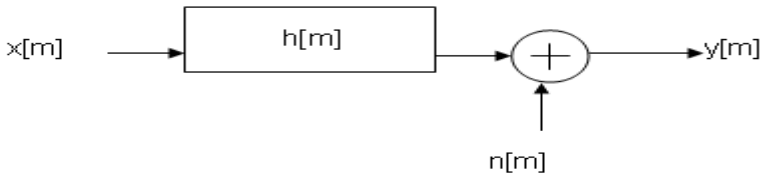
٤٤ Reverberation
٤٥ Lombard
٤٦ Speech enhancement
٤٧ Robust features
٤٨ Data-based methods
٤٩ Model-based methods

آن را به نحوی تغییر دهند که با محیط تطبیق یابد. در این شیوه داده‌های مشاهده شده تغییر نمی‌کنند و فرض و تغییری در مورد سیگنال گفتار لازم نیست.

در این بخش پس از مروری کلی بر مهمترین روشهای مبتنی بر داده و مبتنی بر مدل که برای مقابله با تاثیر نویز بردقت سیستمهای بازشناسی گفتار، در سالهای اخیر بکار رفته‌اند، کارآیی برخی شیوه‌های مبتنی بر داده با یکدیگر مقایسه می‌شود و مدلی از تاثیر محیط بر گفتار ارائه می‌گردد. سپس شیوه‌های مبتنی بر داده و روشهای مبتنی بر مدل بررسی می‌شوند و در نهایت آزمایشها و نتایج کلی ارائه می‌شوند.

2-9 مدلی از تاثیر محیط بر سیگنال گفتار

در بسیاری از کاربردهای بازشناسی گفتار، دو منبع مهم نویز وجود دارد: نویز جمع‌پذیر و نویز فیلتر خطی. از جمله نویزهای جمع پذیر می‌توان به تاثیر جریان هوا و صداهای پس زمینه موجود در محیط اشاره کرد. نویزهای ناشی از کانال انتقال همانند خط تلفن یا اثرات میکروفن و یا نویزهای ناشی از انعکاس و بازتابهای سطحی صدا از اجسام نیز از جمله نویزهای فیلتر خطی محسوب می‌شوند. نویز موجود بر روی خط تلفن نیز از این نوع است. اگر چه انواع دیگری از نویز نیز وجود دارد، اما اثر محیط را معمولا با این دو نوع نویز مدل می‌کنند.



شکل ۱۱ **مدل تاثیر نویز بر گفتار**

در حالتی که نویز به این دو نوع محدودباشد، می‌توان مدلی همانند شکل ۱۱ برای اثرتخریب محیط ارائه کرد [۱۳] . در این مدل سیگنال تمیز x[m] ابتدا از فیلتری با تابع تبدیل h[m] عبور می‌کند و سپس خروجی آن با نویز جمع‌پذیر n[m] که مستقل از x[m] است، جمع می‌گردد و سیگنال نویزی y[m] را شکل می‌دهد.رابطه زیر میان چگالی طیف توانهای سیگنال تمیز، سیگنال نویزی و نویز برقرار است:

$$P_y(w) = P_x(w)\ \left|H(w)\right|^2 + P_n(w) \qquad (۱۰۴)$$

این رابطه پس از اعمال لگاریتم در حوزه کپسترال به صورت زیر در می‌آید:

$$y = x + h + \log(1 - e^{n-x-h}) = x + h + s(x,n,h) \qquad (۱۰۵)$$

که در این رابطه x, n, h, y به ترتیب بیانگر $P_y(w), \left|H(w)\right|^2, P_n(w), P_x(w)$ می‌باشند. این رابطه را می‌توان به این صورت بازنویسی کرد:

$$y = x + f(x,n,h) \qquad (۱۰۶)$$

در این رابطه y را می‌توان بیانگر ویژگیها در محیط نویزی و x را نشانگر ویژگیها در محیط تمیز دانست و f(x,n,h) که تابع محیط نامیده می‌شود، اثر نویزهای جمع پذیر و فیلتر خطی را بر بردارهای ویژگی در محیط تمیز نشان می‌دهد. یکی از روشها در جبران تاثیر تخریب محیط آن است که با عمل در فضای ویژگیها و با در دست داشتن ویژگیهای نویزی y، تخمینی از ویژگیهای تمیز x بدست آید.

3-9 روشهای مبتنی بر داده

چنانکه گفته شد روشهای مبتنی بر داده تلاش می‌کنند که تاثیر محیط بر مشخصه‌های سیگنال گفتار را از طریق تاثیر آنها بر خود سیگنال و یا ویژگیها مشخص نمایند و این تاثیر را به نحوه‌ای بهبود بخشند و اصلاح نمایند که بتوان بازشناسی را تا حدامکان مستقل از تاثیر محیط انجام داد. این روشها را می‌توان به چند دسته کلی تقسیم کرد. اولین گروه روشهایی هستند که به صورت مستقیم بر روی سیگنال گفتار نویزی عمل می‌کنند و سیگنال گفتار تمیز را از سیگنال گفتار نویزی تخمین می‌زنند. ازجمله شیوه‌های این گروه، می‌توان به شیوه تفاضل طیف و فیلتر کردن گفتار نویزی اشاره کرد [۲۸،۱۳] .

گروه دوم روشهای استخراج ویژگیهای مقاوم هستند. در این روشها در اولین مرحله پردازش سیگنال گفتار تبدیلی بکار برده می‌شود تا بتوان بردارهای ویژگی مشابهی برای نمونه‌های گفتار آموزشی و آزمایشی بدست آورد و تفاوت آنها را کمتر نمود. از جمله این روشها می‌توان ازتفاضل میانگین ضرایب کپسترال[50] (CMS) ، [۱۳] آنالیز پیشگویی خطی و درکی گفتار [51] (PLP) [۲۶] و اعمال فیلتر

[50] Cepstral Mean Subtraction
[51] Perceptual Linear Predictive analysis of speech

RASTA بر آن [۲۲] و آنالیز ZCPA [52] [۱۵] نام برد.

روشهای دیگری نیز وجود دارند که تلاش می‌کنند از خصوصیات شنوایی و گوش انسان در برخورد با نویز محیط استفاده کنند.از جمله این روشها استفاده از تکنیک پوشش نویز[53] است که برای آن الگوریتمهای متفاوتی ارائه گردیده است[۱۳] .

قسمتهای بعدی این بخش به بررسی برخی از روشهای گروههای فوق که کاربرد بیشتری دارند و در کار حاضر مورد آزمایش قرار گرفته‌اند، می‌پردازد.

۹-۳-۱ تفاضل میانگین ضرایب کپسترال

تفاضل میانگین ضرایب کپسترال که در گروه ویژگیهای مقاوم جای می‌گیرد، یکی از ساده ترین و موثرترین الگوریتمهای موجود است که غالبا در بازشناسی گفتار با کتاب لغت وسیع بکارگرفته می‌شود. این الگوریتم متوسط بردارهای ویژگی کپسترال را برای یک نمونه از سیگنال گفتار محاسبه می‌کند و سپس این مقدار میانگین را از هر یک از بردارها کم می‌کند. این کار سبب می‌شود که تغییرات داده‌ها و ویژگیها کاهش یابد و نوعی نرمال سازی انجام گیرد. از این رو این شیوه را نرمال سازی ضرایب کپسترال با استفاده میانگین آنها[54] (CMN) نیز می‌نامند. اگر یک نویز از نوع ضرب شونده فرکانسی[55] نامتغیر با زمان موجود باشد، این مجموعه بردارهای ویژگی از تغییرات نویز تاثیر نمی‌پذیرند. در مورد نویز جمع‌پذیر نیز کم کردن این میانگین به مقاوم‌سازی سیستم کمک می‌کند. اما اگر هر دو نوع نویز مذکور به همراه هم وجود داشته باشند، به سختی می‌توان رفتار سیستم را پیش‌بینی کرد [۱۸،۱۳] .

۹-۳-۲ نرمال سازی ضرایب کپسترال با استفاده از میانگین و واریانس

در این روش که در گروه ویژگیهای مقاوم جای می‌گیرد، با توجه به تاثیر نویز بر میانگین و واریانس ضرایب کپسترال، نوعی نرمال‌سازی انجام می‌شود که در آن بردار ویژگی نه تنها از میانگین کل بردارهای ویژگی کم می‌شود، بلکه بر واریانس بردارهای ویژگی نیز تقسیم می‌شود. در این حالت کم کردن

---

۵۲ Zero Crossing with Peak Amplitude
۵۳ Noise Masking
۵۴ Cepstral Mean Normalization
۵۵ Convloutive

میانگین همانند یک فیلتر کردن بالاگذر خطی و تقسیم بر واریانس همچون کنترل بهره خودکار[56] تعبیر می‌شود [۱۶،۱۷] .

برای این کار ابتدا یک بافر با امکان ذخیره‌سازی N بردار کپسترال در نظر گرفته می‌شود ودر هر لحظه ضرایب نرمال شده با توجه به بردارهای ویژگی که در داخل بافر قرار گرفته‌اند، محاسبه می‌شوند. هنگامی که نیمی از بافر با بردارهای کپسترال پرشود،ضرایب نرمال شده محاسبه می‌شوند و اولین بردار بافر نرمال سازی می‌گردد و بردار جدید در بافر جایگزین می‌شود و پس از آن بردار دوم بافر نرمال سازی می‌گردد و این روند تا پر شدن بافر با بردارهای نرمال شده ادامه می‌یابد و سپس بافر خالی شده و N بردار بعدی در بافر قرار می‌گیرند [۱۷].

آزمایشهای عملی نشان داده است که طول بافر(N) باید به نحوی انتخاب شود که در حدود ۴۰ میلی ثانیه از گفتار را پوشش دهد. به این ترتیب کارآیی مناسب به دست می‌آید. علاوه بر اینکه اگر طول بافراز حدی بیشتر انتخاب شود، درصد بازشناسی از حد مشخصی فراتر نخواهد رفت و اصطلاحا منحنی بازشناسی گفتار اشباع می‌شود [۱۶] .

۹-۳-۳ آنالیز پیشگویی خطی و درکی گفتار و فیلتر RASTA

آنالیز پیشگویی خطی و درکی گفتار که به آنالیزPLP معروف است، از این ایده بهره می‌گیرد که قدرت تفکیک فرکانسی وحساسیت نسبت به تغییر فرکانس، در گوش انسان در فرکانسهای مختلف، یکسان نیست و حساسیت گوش نسبت به شدت و انرژی صوت نیز در فرکانسهای مختلف متفاوت است [۲۶]. در روش PLP ابتدا طیف توان زمان کوتاه سیگنال گفتار با استفاده از تبدیل فوریه محاسبه می‌گردد و سپس طیف میان باندهای بحرانی[57] تقسیم می‌شود. این تقسیم بر اساس معیار بارک[58] صورت می‌گیرد. پس از این مرحله، عمل کانولوشن طیف توان توزیع شده در باندهای بحرانی، با تابعی موسوم به تابع پوشش باندهای بحرانی[59] صورت می گیرد. این تابع مدلی از منحنی نامتقارن پوشش گوش در هر یک از باندهای بحرانی را ارائه می کند. حاصل این مرحله در تابع برابرسازی بلندی[60] ضرب می شود. این تابع تقریبی از نحوه حساسیت گوش نسبت به فرکانسهای مختلف را مدل می کند و حساسیت شنوایی را در

---

٥٦ Automatic Gain Control

١٠ Critical bands

١١ Bark

١١ Crtitical-band masking curve

١٣ Equal-loudness

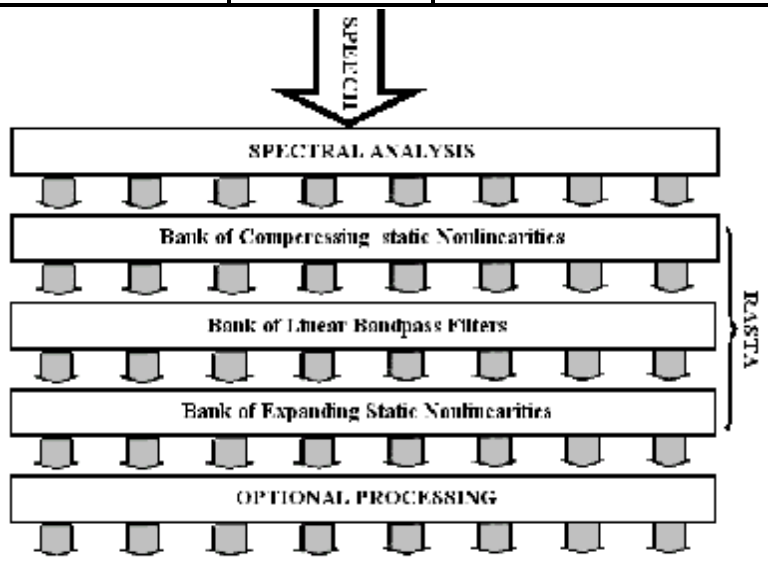
سطح 40 dB نیز شبیه سازی می نماید. مرحله بعدی در عملیات PLP، اعمال قانون ریشه سوم شنوایی می‌باشـــد. این قانون بیان می دارد که در گوش انسان میزان احساس بلندی صوت با ریشه سوم انرژی آن متناسب است یعنی در گوش انسان برای تبدیل انرژی یا توان صوت به بلندی صوت، انرژی صوت به توان ۰/۳۳ رسانده می‌شود. در مرحله آخر پس از اعمال تبدیل فوریه معکوس برحاصل، با اعمال یک مدل تمام قطب، ضرایب PLP محاسبه می شوند [۲۶] .

روش PLP بر اساس محاسبه طیف زمان کوتاه گفتار صورت می گیرد و بنابراین. مانند دیگر تکنیکهای مبتنی بر طیف زمان کوتاه، به دلیل تاثیر کانال‌های مخابراتی بر مقادیر طیف زمان کوتاه، آسیب‌پذیر است. بنابراین ضرایب PLPویژگیهای مقاومی محسوب نمی شوند. بنابراین باید ضرایب PLP را در مقابل اثرات تخریبی کانال ارتباطی مقاوم نمود. بکارگیری فیلتر RASTA از جمله شیوه هایی است که این مقاوم سازی را انجام می دهد. نام این فیلتر از واژه طیف نسبی[61] گرفته شده است [۲۲] .

آنالیز طیف نسبی با این هدف انجام می شود که تغییرات آهسته یا تندی که در خارج از محدوده زبان شناسی سیگنال گفتار قرار می گیرند، از این سیگنال حذف گردند. در واقع فیلترRASTA براساس عملکرد گوش انسان عمل می‌کند. گوش انسان نسبت به مؤلفه‌های فرکانسی که شدت آنها در طول زمان با فرکانس حدود 4 kHz تغییر می‌کنند، در مقایسه با مؤلفه‌هایی که شدت آنها در طول زمان سریعتر یا آهسته‌تر از 4 kHz تغییر می‌کنند، حساسیت بیشتری دارد.

مراحل کلی آنالیز طیف نسبی مطابق شکل ۱۲ است. در عملیات RASTA ابتدا یک تبدیل غیرخطی فشرده کننده بر طیف توان گفتار اعمال می شود. سپس یک فیلتر میان گذر بر طیف توان تبدیل یافته اثر می‌کند. در نهایت یک عمل تبدیل منبسط کننده که معکوس عمل فشرده سازی است، بر نتیجه اعمال می‌شود. مجموع عملیات فشرده سازی طیف، اعمال فیلتر و منبسط سازی طیف، آنالیز طیف نسبی یا RASTA نامیده می‌شود[۵].

[۱۵] Relative Spectra

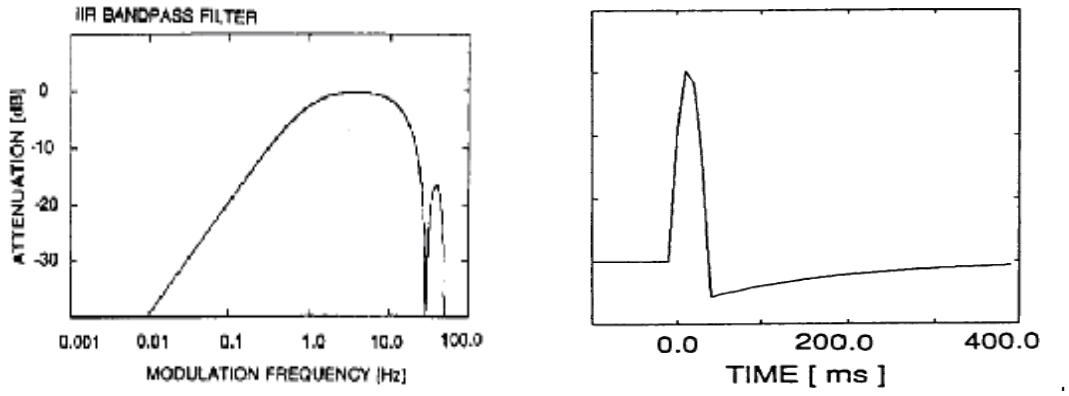


شکل ۱۲ **مراحل کلی آنالیز طیف نسبی**

پس از آنالیزطیف نسبی، هر آنالیز دیگری مانند PLP می تواند انجام شود. اگر ضرایب PLP پس از آنالیز RASTA بدست آید، ضرایب حاصل ویژگیهای مقاومی هستند که ضرایب RASTA-PLP نامیده می شوند.

فیلتر بکار گرفته شده در آنالیز طیف نسبی می‌تواند شکلهای مختلفی داشته باشد. شکل ۱۳ پاسخ فرکانسی و پاسخ ضربه یک فیلترRASTA را نشان می‌دهد[۲۲]. تبدیل فشرده ساز در آنالیز RASTA می تواند به صورت $y = \ln(n)$ باشد. در این حالت فیلتر در حوزه لگاریتم طیف عمل می‌کند و آنالیز RASTA را log-RASTA می‌نامند.

هنگامی که عملیات RASTA در حوزه لگاریتم صورت گیرد، نویز فیلتر خطی یا تاثیر کانال که در حوزه لگاریتم طیف به صورت مولفه‌های طیفی جمع‌پذیر ظاهر می‌شود، به طور مؤثر حذف می‌شود. اما در مورد نویز جمع‌پذیر غیروابسته به سیگنال، پس از اعمال تبدیل غیرخطی لگاریتم، نویز وابسته به سیگنال می‌شود و به طور مؤثر با آنالیز RASTA حذف نخواهد شد. جهت حل این مسأله، استفاده از تبدیل فشرده‌ساز $y = \ln (1+Jx)$ به جای $y = \ln(x)$ پیشنهاد شده است که در آن J یک مقدار ثابت مثبت وابسته به سیگنال است. در صورتی که Jبسیار کوچکتر از یک اختیار شود،تبدیل حاصله یک تبدیل شبه خطی است.



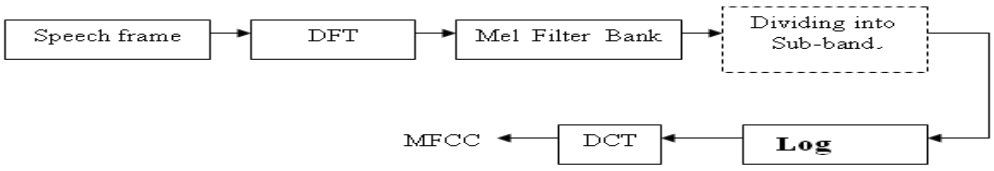
(الف) پاسخ ضربه **(ب) پاسخ فرکانسی**

شکل ۱۳ **فیلتر RASTA**.

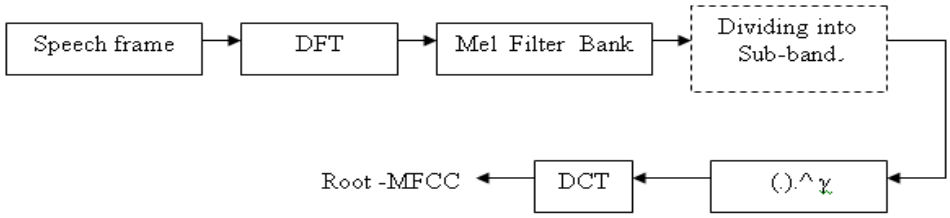
درصورتی که مقدارJ بسیار بزرگتر ازیک باشد، تبدیل شبه لگاریتمی خواهد بود. عملیات منبسط کننده متناظر با فشرده ساز به صورت $x = (e^{y}-1)/J$ می‌باشد. به آنالیز RASTA با چنین فشرده ساز و منبسط کننده‌ای lin–log–RASTA گفته می‌شود و گاه آن را J-RASTA نیز می نامند[۲۲]. فیلتر J-RASTA کارآیی یک سیستم بازشناس را هم در حضور نویز جمع شونده و هم در حضور نویز فیلتر خطی، افزایش می دهد[۱۳].

۹-۳-۴ ضرایب کپسترال ریشه ای

استفاده از دیگر انواع توابع فشرده سازی همانند تابع ریشه بجای تابع لگاریتم در استخراج ضرایب مل کپستروم، روش دیگری برای مقابله با حضور نویز به ویزه در حضور نویز جمع پذیر است [۲۳]. به این ترتیب میزان فشرده سازی انرژی زیرباندهای مل در مقایسه با یکدیگر و در حضور نویز ،تغییر می کند و موثرتر از تابع لگاریتم می‌گردد. تفاوت ویژگیهای مل کپسترال ریشه ای و ویژگیهای مل کپسترال عادی در شکل ۱۴ قابل رویت است.



(الف) روش قراردادی استخراج ضرایب مل کپسترال **(MFCC)**



(ب) ضرایب کپسترال ریشه ای **(RMFCC)** جاییکه۱ > γ > ۱-

شکل ۱۴ **روشهای استخراج MFCC و RMFCC**

۹-۳-۵ ضرایب خود همبستگی فاز[۶۲] (PAC)

تبدیل فوریه دنباله خودهمبستگی یک سیگنال در حوزه زمان، معادل با طیف توان آن سیگنال در حوزه فرکانس است. دنباله خود همبستگی را می توان به صورت یک ضرب داخلی مابین بردارهای گفتاری نیز در نظر گرفت. اخیرا معیار دیگری برای خود همبستگی بنام خود همبستگی فاز (PAC) مطرح شده است که از زاویه میان بردارهای گفتار به عنوان معیار همبستگی استفاده می کند[۹]. ایده استفاده از زاویه بر این اساس شکل گرفته است که در یک ضرب داخلی بین بردارها، زاویه کمتر از اندازه از نویز تاثیر می پذیرد[۲۷]. ضرایب خود همبستگی فاز را می توان به صورت زیر محاسبه کرد:

$$P[k] = q_k = Arc\cos\left(\frac{R[k]}{|x|^2}\right) \qquad (۱۰۷)$$

در این رابطه، R[k] و $|x|^2$ به ترتیب بیانگر ضرایب خود همبستگی عادی و انرژی یک قاب از سیگنال هستند. طیف فرکانسی محاسبه شده از ضرایب PAC بنام طیف خود همبستگی فاز یا طیف PAC نامیده می شود. مشابه با ویژگیهایی که از طیف عادی استخراج می شوند، ویژگیها می توانند از طیف خود همبستگی فاز نیز استخراج شوند. ویژگیهای MFCC که از طیف خود همبستگی فاز استخراج می شوند، بنام PAC-MFCC شناخته می شوند. نتایج آزمایشها در[۶] و[۹] نشان داده اند که ویژگیهای PAC-MFCC در حضور نویز مقاوم هستند، اما کارآیی بازشناسی گفتار تمیز را کاهش می دهند.

---

۵۵ Phase AutoCorrelation

۹-۳-۶ ویژگیهای مبتنی بر تاخیر گروه[۶۳]

تابع تاخیر گروه به صورت منفی مشتق فاز سیگنال صحبت تعریف می شود. این تابع را می توان اینگونه محاسبه کرد [۸،۱۰،۲۴] :

$$t_p(w) = -\frac{X_R(w)Y_R(w) + X_I(w)Y_I(w)}{|X(w)|^2} \qquad (۱۰۸)$$

در این رابطه اندیسهای I و R بیانگر بخشهای موهومی و حقیقی هستند و X(ω) و Y(ω)به ترتیب نشانگر تبدیل فوریه دنباله های x(n) و n.x(n) هستند. در [۸]، تابع اصلاح شده ای بر اساس رابطه (۱۰۶) برای محاسبه تاخیر گروه پیشنهاد شده است که مخرج این رابطه را حذف کرده است. زیرا مخرج این رابطه سبب ایجاد پرشهای ناگهانی[64] در مقدار تابع تاخیر گروه می شود که مطلوب نیست. رابطه جدید موسوم به طیف ضربی[65] است که به صورت زیر تعریف می شود:

$$Q(w) = |X(w)|^2 t_p(w) = X_R(w)Y_R(w) + X_I(w)Y_I(w) \qquad (۱۰۹)$$

در کار حاضر، ما وِیژگیهای MFCC را از این طیف ضربی استخراج می کنیم. آزمایشها نشان داده است که این ویژگیها نسبت به حضور نویز جمع پذیر برای بازشناسی کلمات و واجها مقاوم است[۸،۵]. این ویژگیها را در کار حاضر با نام GDF مشخص می نماییم.

4-9 روشهای مبتنی بر مدل

چنانکه گفته شد شیوه‌های مبتنی بر مدل تلاش می‌کنند که بجای اصلاح سیگنال یا پارامترها، مدل آکوستیک محیط را در مرحله بازشناسی اصلاح نمایند. مزیت این روشها آن است که در آنها داده‌های مشاهده شده تغییر نمی‌کنند و هیچ نوع فرض یا تصمیم‌گیری قبلی درباره سیگنال گفتار ضروری نیست و از این رو نیازمند پایگاه داده استریو نیستند. برخی از روشهای مبتنی بر مدل عبارتند از: تجزیه مدل مخفی مارکوف[66] [۲۵]،ترکیب موازی مدلها[67] (PMC) [۱۸] و بازگشت خطی با بیشترین شباهت[68]

---

٥٦ Group Delay
٥٧ Spikes
٥٨ Product Spectrum
٦٦ HMM decomposition
٦٧ Parallel Model Combination
٦٨ Maximum Likelihood regression

(MLLR) [۱۴].

در تجزیه مدل مخفی مارکوف، گفتار تمیز و نویز به صورت جداگانه و به عنوان دنباله‌ای از حالات مدل مخفی مارکوف مدل می‌شوند. مدلهای گفتار و نویز در یک مدل ادغام می‌شوند و به این ترتیب امکان انتقال از حالات گفتار به حالات نویز و بر عکس به وجود می‌آید. به این ترتیب، به هنگام بازشناسی الگوریتم ویتربی محتمل‌ترین مسیر را دریک فضای گسترده جستجو می‌کند. این شیوه نه تنها برای سیگنال و نویز، بلکه برای بازشناسی دو سیگنال همزمان با یکدیگر نیز کاربرد دارد[۲۵] . ایراد این شیوه، هزینه بالای محاسباتی جستجو در فضای گسترده است.

شیوه ترکیب موازی مدلها تلاش می‌کند تا تاثیر محیط بر توزیع کپستروم گفتار تمیز را در مدل محیط منعکس‌نماید و با فرض وجود دانش کامل از نویز، بردار میانگین و ماتریس کوواریانس توزیعهای گاوسی مدل مخفی مارکوف را به نحوی تغییردهد که به توزیعهای ایده‌آل کپستروم گفتار نویزی نزدیک شوند. چندین شیوه مختلف برای تغییر بردارهای میانگین و ماتریسهای کوواریانس در الگوریتم ترکیب موازی مدلها موجود است [۱۸] .

به هر حال،در تمامی فرمهای الگوریتم PMC، به دانش قبلی از نویز و بردارهای کانال نیاز است و نمونه‌های مجزا از نویز باید وجود داشته باشند تا بتوان پارامترهای PMC را تخمین زد [۱۹] .

روش MLLR در اصل برای تطبیق گوینده طراحی شده است، اما برای جبران اثر محیط و نویز نیز می‌تواند بکار رود. این الگوریتم بردارهای میانگین و ماتریسهای کوواریانس گفتار تمیزمدل شده بوسیله مدل مخفی مارکوف را تغییر می‌دهد و برای این تغییر از داده‌های آموزشی کمی استفاده می‌کند. این روش در نهایت یک ماتریس تبدیل می‌یابد که احتمال مشاهده بردار کپستروم نویزی در مدل را بیشینه کند[۱۴] .

روش MLLR از یک مدل صریح از محیط استفاده نمی‌کند، بلکه تنها فرض می‌کند که بردارهای میانگین توزیع کپستروم گفتار تمیز، تحت تاثیر محیط دوران یافته و انتقال می‌یابند [۱۹] .

۹-۴-۱ معیار تصویردهی وزن دار

تئوری معیار تصویردهی وزن دار*(WPM)* بر اساس این مشاهده *Mansour* و *Juang*استوار است که اندازه بردارهای ویژگی کپسترال در حضور نویز جمع پذیر سفید کاهش می یابد[۲۷] . طبق این خصوصیت در [۲۷] یک معیار محاسباتی بر پایه عمل تصویر کردن معرفی شد که کارآیی بازشناسی گفتار بوسیله روش *DTW* را درحضور نویز به طرز قابل توجهی بهبود بخشید. *Carlson* و *Celemnt* این معیار تصویردهی را گسترش دادند وآن را در یک سیستم بازشناسی مبتنی بر مدل مخفی مارکف با چگالی پیوسته

(CDHMM) بکار گرفتند. آنها یک عامل مقیاس را در توزیع حالات *CDHMM* و به عبارت دیگر در توزیع گاوسی احتمالات شباهت دخالت دادند تا به این وسیله کاهش اندازه بردار کپسترال جبران گردد. ویژگیهای مورد استفاده آنان ضرایب مل کپستروم*(MFCC)* بود. نتایج در [۲۱،۵،۷،۲۹] نشان داده است که که بکار بردن این معیار تصویردهی وزن دار ، نرخ بازشناسی کلمات مجزا را به طرز قابل توجهی در حضور انواع مختلف نویز از جمله نویز سفید و نویز رنگی بهبود می بخشد. رابطه جبران تعریف شده آنها برای توزیع گاوسی را می توان به صورت زیر بیان کرد:

$$b_{j,i}(c_t) = N(c_t, l_{j,i,t}, m_{j,i}, C_{j,i}) = \frac{\exp\left(-\frac{1}{2}(c_t - l_{j,i,t} m_{j,i})^T C_{j,i}^{-1}(c_t - l_{j,i,t} m_{j,i}) \right)}{(2p)^{\frac{n}{2}} |C_{j,i}|^{\frac{1}{2}}} \qquad (۱۱۰)$$

که پارامتر های موجود در آن این گونه تعریف می شوند:

c_t: بردار مشاهده کپسترال برای قاب t

n: بعد بردار مشاهده

$m_{j,i}$: بردار میانگین مخلوط گاوسی j ام در حالت i

$C_{j,i}$: ماتریس کوواریانس مخلوط گاوسی j ام در حالت i

$l_{j,i,t}$: عامل مقیاس برای قاب t در مخلوط گاوسی j ام در حالت i

$b_{j,i}(c_t)$: احتمال تولید شده برای بردار مشاهده c_t توسط مخلوط گاوسی j ام در حالت i

با گرفتن لگاریتم از رابطه (۱۰۵) و محاسبه لگاریتم احتمال شباهت، می توان به رابطه ای برای الگوریتم ویتربی دست یافت که بیانگر معیار تطبیق میان بردار مشاهده و بردارهای میانگین مخلوطهای گاوسی است. این رابطه را می توان اینگونه نوشت:

$$\log b_{j,i}(c_t) = (c_t - l_{j,i,t} m_{j,i})^T C_{j,i}^{-1}(c_t - l_{j,i,t} m_{j,i}) + \log |C_{j,i}| + N \log(2p) \qquad (۱۱۱)$$

با گرفتن مشتق از رابطه (۱۱۱) نسبت به $l_{j,i,t}$ می توان مقدار بهینه ای برای $l_{j,i,t}$ بدست آورد که به واسطه آن احتمال مشاهده $b_{ji}(c_t)$ بیشینه شود. این مقدار بهینه به صورت زیر بدست می آید:

$$l_{j,i,t} = \frac{c_t^T C_{j,i}^{-1} m_{j,i}}{m_{j,i}^T C_{j,i}^{-1} m_{j,i}} \qquad (۱۱۲)$$

با جایگذاری این مقدار برای $l_{j,i,t}$ در رابطه (۱۱۱)، مقداری برای لگاریتم احتمال شباهت بدست می آید

که *WPM* نامیده می شود. رابطه (۱۱۱) و (۱۱۲) برای تمامی مخلوطهای گاوسی موجود در یک حالت مدل مخفی مارکف بکار می روند و احتمال شباهت نهایی با استفاده از مخلوطهای گاوسی تطبیق یافته با معیار تصویردهی وزن دار محاسبه می شود.

5-9 شیوه‌هایی برای تطبیق با گوینده جدید

یکی از دلایل عدم تطبیق میان داده‌های گفتار آموزشی و داده‌های آزمایش، تغییر گوینده یا به عبارتی شناسایی کلام یک گوینده جدید است. برای آنکه سیستم بتواند با گوینده جدید تطبیق یابد، از شیوه‌های تطبیق گوینده استفاده می‌شود تا بتوان با در دست داشتن کمترین نمونه‌های گفتاری گوینده، گفتار او را بازشناسی کرد. در این راستا، شیوه‌های متعددی مطرح شده‌اند، همانند: نرمال سازی طول جهاز صوتی[69] (VTN)، خوشه‌بندی گوینده[70]، روش MLLR و انواع اصلاح شده آن، شیوه تخمین بیشینه پارامترهای پسین [71] (MAP) و انواع تغییر یافته آن [۱۲،۱۴].

شیوه نرمال سازی جهاز صوتی بر اساس این ایده استوار است که طول جهاز صوتی در گویندگان مختلف متفاوت است و این امر بر فرکانسهای فرمانت تاثیر می‌گذارد و سبب تغییراتی می‌شود و بنابراین باید داده‌ها بر اساس یک مقیاس فرکانسی خطی نرمال سازی شوند تا تاثیر تغییر طول جهاز صوتی در فرکانسهای فرمانت به حساب آورده شود و به این ترتیب به نوعی سیستم با تغییر گوینده تطبیق یابد [۱۲،۱۴] .

در خوشه‌بندی گوینده، تلاش می‌شود که گروهی از گویندگان یافت شوند گه از لحاظ آکوستیکی به گوینده جدید نزدیکترند و از ویژگیهای این گویندگان برای تخمین و اصلاح مجدد پارامترهای مدل گفتار استفاده می‌شود. ساده‌ترین شیوه این خوشه بندی بر اساس جنسیت گوینده است و توجه به پارامترهایی همانند فرکانس اصلی که در دو جنس متفاوت است [۱۴] .

شیوه MLLR نیز چنانچه گفته شد، عمل می‌کند اما این بار ماتریس تبدیل را چنان می‌یابد که احتمال مشاهده بردارهای کپستروم مربوط به گوینده جدید را بیشینه کند.

شیوه MAP نیز یک تخمین پسین برای پارامترهای مدل در نظر می‌گیرد و با استفاده از این تخمین در مرحله آموزش مدل مخفی مارکوف تغییری ایجاد می‌کند تا بتواند با داده‌های آموزشی کم مدل را با

---

۶۹ Vocal Tract Normalization
۷۰ Speaker Clustring
۷۱ Maximum A Posterirori

گوینده جدید تطبیق دهد [۱۲،۱۴] .

علاوه بر ابزارهای پردازش و بازشناسی گفتار که برای سیستم مورد نظر پیاده سازی شده اند، ابزارهای آماده دیگری نیز برای اهداف کاربردی و تحقیقاتی در مراکز تحقیقاتی جهان تهیه شده اند که استفاده از آنها برای همگان آزاد است و اطلاعات و منابع مربوط به آنها به صورت آزاد بر روی شبکه جهانی موجود است و برای کاربران آنها لیستهای پستی موجود است. از جمله این ابزارها که کار بر روی آن بیش از یک دهه است ادامه دارد، HTK [72]است. HTK ابزاری برای ساخت مدل مخفی مارکف و پردازش سیگنال گفتار بوسیله آن است که در دانشگاه کمبریج انگلستان تولید شده و توسعه یافته است. علاوه براین امکانات دیگر پردازش گفتار نظیر استخراج ویژگیها،گرامر زبان و تطبیق گوینده نیز در آن وجود دارد. از شیوه های تطبیق گوینده، روشهای MAP و MLLR در HTK پیاده سازی گردیده اند[۱۲] .

در این گزارش ابتدا مقدمه ای در مورد سیستمهای بازشناسی گفتار و روشهای متداول آن ارائه شد. سپس بحث پیش پردازش و بررسی ویژگیهای بکار رفته در بازشناسی بررسی شد. در ادامه مدل مخفی مارکوف بطور کامل تشریح شد و آنرا به عنوان روشی برای بازشناسی تحلیل کردیم. در فصل بعدی مسائل مهم در پیاده سازی عملی سیستم بازشناسی گفتار دیده شدند. در ادامه نیز روشهای جستجوی معمولی، مؤلفه های فضای جستجو و روشهای جستجوی ویتربی بکار رفته در سیستم بازشناسی گفتار ارائه شدند. در فصل آخر متدهای مختلف مقاوم سازی بازشناسی گفتار و نتایج بدست آمده از این متدها آورده شد.

سیستم بازشناسی گفتار به عنوان یکی از سرویسهای بکار رفته در پروژه حاضر از اهمیت خاصی برخوردار است که بدلیل بحثهای پردازش سیگنال به صورت گزارش مستقل آورده شد.

[72] Hidden Markov Model Toolkit

10 تطبیق گوینده

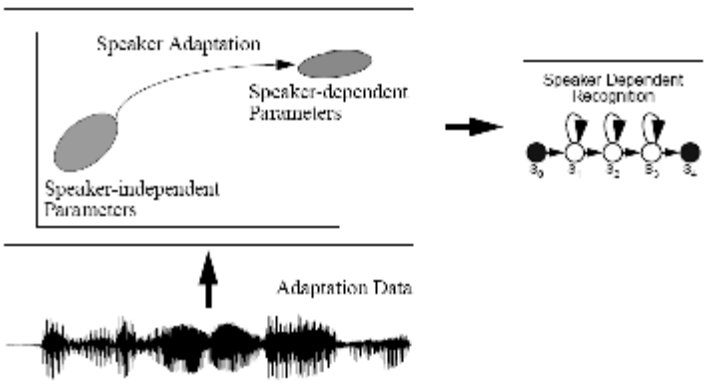
10-1 مقدمه

مبحث تطبیق گوینده هم برای سیستمهای بازشناسی گفتار و هم برای سیستمهای سنتز مطرح است. تطبیق گوینده در سیستمهای بازشناسی گفتار به منظور افزایش راندمان بازشناسی و در سیستمهای سنتز برای تغییر و تطبیق صدای سنتز شده با صدای فرد مورد نظر صورت می‌گیرد.

10-2 تطبیق گوینده برای سیستمهای بازشناسی گفتار

سیستمهای بازشناسی گفتار می توانند از نوع وابسته به گوینده (SD[۷۳]) یا مستقل از گوینده (SI[۷۴]) باشند. سیستم‌های وابسته به گوینده معمولا دارای راندمان بالاتری برای گوینده‌ای که بازاء داده های گفتاری او آموزش دیده‌اند می‌باشند. چنانچه سیستم بازشناسی موجود از نوع مستقل از گوینده باشد با استفاده از روشهای تطبیق گوینده و به کمک داده های گفتاری یک گوینده خاص سیستم را به سیستم وابسته به گوینده تبدیل و راندمان آنرا افزایش داد.

ایده اصلی تطبیق گوینده استفاده از مقدار کمی اطلاعات تطبیقی به منظور تغییر سیستم بازشناسی است به طوریکه تا حد ممکن بتواند اطلاعات مربوط به گوینده جدید را مدل کند. شکل ۱۵ عمل تطبیق گوینده جدید با مدلهای موجود در سیستم تشخیص با استفاده از داده های تطبیقی کم را نشان می دهد.



[73] Speaker Dependent
[74] Speaker Independent

شکل ۱۵ **نمایش سطح بالای عمل تطبیق گوینده**

با توجه به اینکه سیستم بازشناسی استفاده شده مبتنی بر HMM است روشهای تطبیق گوینده مبتنی بر HMM مورد بررسی قرار می‌گیرند. تکنیکهای تطبیق گوینده برای سیستمهای بازشناسی مبتنی بر HMM[۷۵] به دو دسته اصلی تقسیم شوند. دسته اول شامل تکنیکهایی هستند که گفتار ورودی گوینده جدید را به فضای برداری مشترک با گفتارهای آموزشی تبدیل می کنند. این تکنیکها با نام روشهای نگاشت طیفی[۷۶] شناخته می‌شود. دسته دوّم شامل روشهایی هستند که با تغییر پارامترهای مدل باعث تطبیق بهتر آن مدلها با ویژگیهای داده های تطبیقی می شوند. این تکنیکها را تکنیکهای نگاشت مدل[۷۷] می‌نامند.

10-3 انواع تطبیق ها

دو موضوعی که هنگام بحث در مورد روشهای نگاشت مدل باید مشخص شود، مد آموزش[۷۸] (با نظارت[۷۹] در مقابل بی‌نظارت[۸۰]) و مد تطبیق (افزایشی[۸۱] در مقابل دسته ای[۸۲]) است. در مد آموزشی با نظارت، سیستم تشخیص، آوانویسی درست کلمات را در اختیار داشته و تنها گفتار کاربر را با آوانویسی های موجود مقایسه و هم تراز[۸۳] می کند .در تطبیق بدون نظارت تشخیص دهنده آوانویسی درست کلمات را در اختیار نداشته و در نتیجه ممکن است گفتار کاربر را با یک آوای غلط مقایسه کند و به عبارتی تشخیص سیستم ممکن است با خطا همراه باشد. در نتیجه، مد با نظارت اگر قابل دسترسی باشد به مد بی‌نظارت ترجیح داده می شود.

10-4 روش نگاشت طیفی

ایده اصلی روش نگاشت طیفی افزایش یافته سیستم بازشناسی گفتار بوسیله نگاشت بردار ویژگی بدست آمده از گوینده جدید است. در این روش ویژگیهای گوینده را با یک تبدیل مناسب به ویژگی گوینده مدل آموزشی نگاشت می‌کنند.

---

[75] HMM-based recognition systems
[76] spectral mapping techniques
[77] model mapping approaches
[78] training mode
[79] supervise
[80] unsupervise
[81] incremental
[82] batch
[83] align

$$y_n = P(x_n) = A^T x_n \qquad (۱۱۳)$$

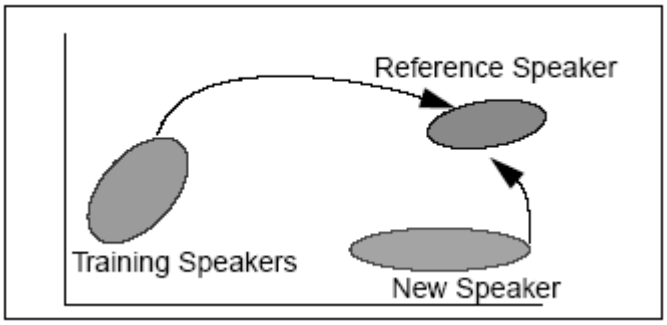
که P(.) تبدیلی است که بر روی ویژگی گوینده x_n، ورودی اعمال می‌شود. A^T ماتریس تبدیل و y_n نتیجه تبدیل یافته بردار ویژگی ورودی است. ماتریس تبدیل بصورتی تخمین زده می‌شود تا مقدار MSE[۸۴] بین ویژگی بدست آمده y_n و بردار ویژگی گوینده مرجع x_r مینیمم گردد.

$$D = E[(x_r - y_n)^T.(x_r - y_n)] = E[(x_r - A^T x_n)^T.(x_r - A^T x_n)] \qquad (۱۱۴)$$

نتیجه حاصل بصورت زیر است:

$$\min_A(D) \Rightarrow A = \left(E[x_n x_n^T]^{-1}\right).E[x_n x_r^T] \qquad (۱۱۵)$$

عمل نگاشت طوری طراحی می‌شود که اختلاف میان مجموعه بردار مرجع و مجموعه بردار نگاشت شده حداقل باشد. این اختلافات به واسطه اختلافات طیفی میان سیستمهای تولید گفتار گوینده ها است. این روش در شکل ۱۶ نمایش داده شده است و معمولاً با نام روش نرمال کردن گوینده شناخته می‌شود.



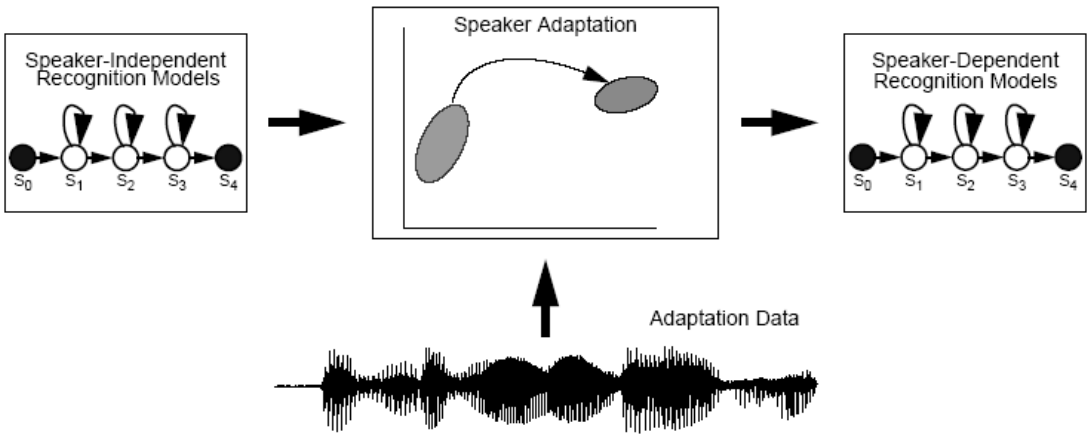
شکل ۱۶ **نگاشت بردارهای ویژگی گوینده های آموزشی و گوینده جدید به یک فضای مشترک**

10-5 روش نگاشت مدل

هدف تکنیکهای نگاشت طیفی بهبود تطبیق میان گوینده مرجع و گوینده جدید است. اما این روشها به صورت واضح سعی نمی کنند تا دقت مدلها را برای گوینده جدید افزایش دهند. به عبارت دیگر این روشها به طور کامل از مزایای داده‌های تطبیقی استفاده نمی کنند. به همین دلیل از دسته دیگری از روشها با نام روشهای نگاشت مدل استفاده می شود که به جای نگاشت تمام گوینده ها به یک فضا،

[84] Mean-Squared Error

پارامترهای مدل را با بهترین ارائه گوینده جدید تنظیم می‌کند. بلوک دیاگرام کلی این روشها در شکل ۱۷ دیده می‌شود.



شکل ۱۷ **روش نگاشت مدل برای تطبیق مدلهای مخفی مارکوف**

اکثر روشها، پارامترهای تابع چگالی HMM را بوسیله ماکزیمم کردن احتمال یک مدل (l) برای یک مجموعه مشاهدات (O) تخمین می‌زند.

$$P(l \mid O) = \frac{P(O \mid l) P_p(l)}{P(O)} \qquad (۱۱۶)$$

و با فرض اینکه l از مقدار توزیع قبلی $P_p(l)$ بدست آمده است.

از روشهای معروف و شناخته شده که برای تطبیق مدل مخفی مارکوف با گوینده موردنظر استفاده می‌شوند می‌توان به موارد زیر اشاره نمود:

۱) MAP[۸۵]

۲) MLLR[۸۶]

۳) CAT[۸۷]

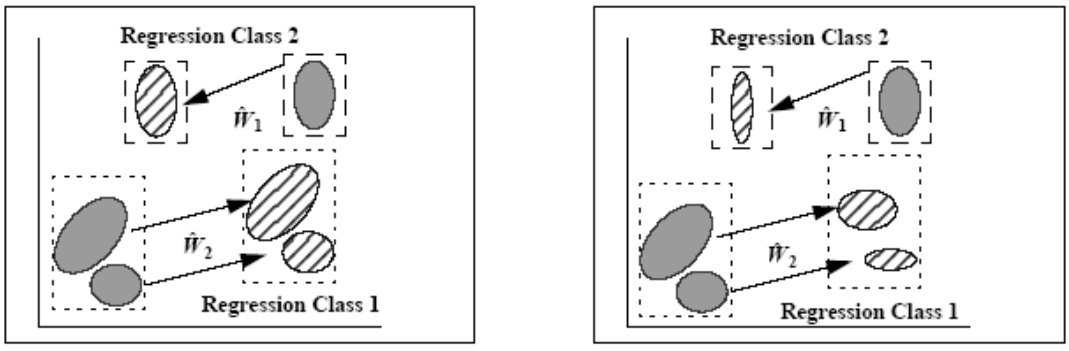
در ادامه روش MLLR شرح داده می‌شود.

[85] Maximum A Posteriori
[86] Maximum Likelihood Linear Regression
[87] Cluster Adaptive Training

6-10 تطبیق مدل با روش MLLR

در روش MLLR، مجموعه‌ای از تبدیلات محاسبه می‌شوند تا عدم شباهت بین مدل اولیه و داده تطبیقی را کاهش دهند. MLLR یک روش تطبیق مدل است که در آن پارامترهای میانگین و واریانس برای مدل مخفی مارکوف دارای چگالی گوسی بوسیله تبدیلهای خطی تخمین زده می‌شوند. نتیجه این تبدیلات این است که میانگین‌های اولیه شیفت و واریانس اولیه به کلی تغییر می‌کند.

MLLR قادر است تبدیلات تطبیقی قوی[88] بسازد که حتی برای مدلهایی که داده تطبیقی برای آنها وجود ندارد نیز با استفاده از تبدیل مشترک[89] عمل تطبیق را انجام دهد. این مساله به ما کمک می کند تا مشکل داده های تطبیقی محدود را برطرف کنیم. با وجود مقدار کمی داده، تنها یک تبدیل سراسری[90]، می تواند برای تمام مدلها به کار رود. هر چه داده بیشتری در دسترس باشد، تبدیلات دقیقتر و بیشتری می تواند به کار گرفته شود. در شکل ۴ اساس روش MLLR را آمده است که یک مدل مستقل از گوینده (بیضی توپر) را گرفته و از یک تبدیل استفاده می کند تا فضای مدل را به سمت یک به مدل وابسته به گوینده (بیضی راه راه) انتقال دهد. معمولا تنها تطبیق، روی میانگین انجام می شود، به خاطر اینکه فرض می شود اختلاف اصلی میان گوینده ها به خاطر موقعیت میانگین آواها در فضای صوتی است. تطبیق روی کواریانس معمولاً کمتر استفاده می شود، چون تاثیر کمتری نسبت به تطبیق میانگین دارد. در شکل ۱۸ می بینیم که تطبیق میانگین، موقعیت میانگین مدل ها را در فضا جابجا می کند، در حالیکه تطبیق کواریانس شکل توزیع را تغییر می دهد.



شکل ۱۸ **تطبیق مدل مستقل از گوینده به مدل وابسته به گوینده با استفاده از روش MLLR**

[88] robust adaptation transforms
[89] transform sharing
[90] global transform

یک ماتریس تبدیل برای یافتن مقدار جدید میانگین استفاده می‌شود تا میانگین جدید بوسیله این تبدیل تخمین زده شود:

$$\hat{m} = Wx \qquad (۱۱۷)$$

که W یک ماتریس $n\times(n+1)$ و n بعد داده است. x نیز بردار میانگین گسترش یافته است که یک مقدار w به آن اضافه شده است.

$$x = [w, m_1, m_2, ..., m_n]^T \qquad (۱۱۸)$$

می توان W را بصورت یک ماتریس A و یک بردار بایاس b بصورت $W = [b \;\; A]$ تجزیه نمود. سپس ماتریس تبدیل W با حل مسأله ماکزیمم کردن بوسیله روش ماکزیمم کردن امید ریاضی (EM)[۹۱] بدست می‌آید. اگر $g_q(t)$ را احتمال تولید o_t در حالت q و در زمان t، با دنباله مشاهدات O و مدل l باشد، ماتریس W_q با حل معادله زیر بدست می‌آید:

$$\sum_{i=1}^{T}\sum_{r=1}^{R} g_{qr}(t) U_{qr}^{-1} W_q x_{qr} x'_{qr} \qquad (۱۱۹)$$

که در آن x' ترانهاده ماتریس میانگین و U واریانس چگالی گوسی می‌باشند.

در ادامه تخمین ماتریس تبدیل برای واریانس فقط برای سیستمهای با کوواریانس قطری انجام می پذیرد. کوواریانس گوسی بصورت زیر تبدیل می یابد.

$$\hat{\Sigma}_m = B_m^T H_m B_m \qquad (۱۲۰)$$

که در آن H_m تبدیل خطی ای است که باید تخمین زده شود و B_m معکوس ضریب چولسکی[۹۲] Σ_m^{-1} است. بنابراین:

$$\Sigma_m^{-1} = C_m C_m^T \qquad (۱۲۱)$$

$$B_m = C_m^{-1} \qquad (۱۲۲)$$

بعد از بازنویسی تایع کمکی، ماتریس تبدیل H_m بصورت زیر تخمین زده می شود:

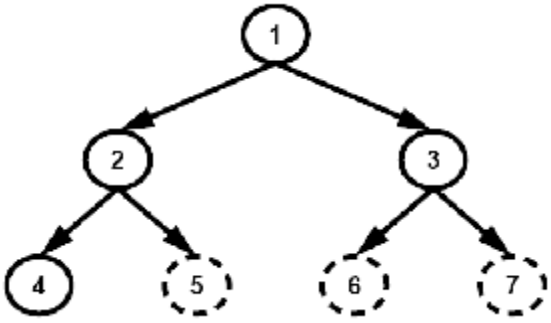
[91] Expectation Maximisation
[92] Choleski

$$H_m = \frac{\sum_{r=1}^{R_c} C_{m_r}^T \left[L_{m_r}(t)(o(t)-m_{m_r})(o(t)-m_{m_r})^T \right] C_{m_r}}{L_{m_r}(t)} \qquad (۱۲۳)$$

10-7 تبدیل اشتراک

به طور ایده ال، ما می خواهیم برای هر مدل گوسی در یک سیستم بر پایه HMM، از یک تبدیل تطبیقی استفاده کنیم تا تمام اختلافات میان مدل مستقل از گوینده و مدل وابسته به گوینده را به طور دقیق پیدا کنیم. هر چند، این کار در عمل نیاز به داده های تطبیقی زیادی دارد تا به طور دقیق، مدلهای مناسب را تخمین بزند. به همین خاطر، معمولا یک فرم از تبدیل اشتراک به کار گرفته می شود که مجموعه مدلهای گوسی که با هم کار می‌کنند را به وسیله یک تبدیل یکسان، تطبیق می دهد. در این طرح، مدلهایی که داده های تطبیقی کمی دارند و اصلا داده تطبیقی ندارند، توسط داده های تطبیقی مدلهای مشابه تطبیق داده می شوند.

یک روش معمول برای تبدیل ادغام[۹۳] استفاده از یک درخت برگشت باینری[۹۴] است که در شکل ۱۹ نیز نمایش داده شده است. همانطور که در شکل دیده می شود برگهای ۴، ۵، ۶ و ۷ کلاسهای پایه هستند. دایره های خط چین گره هایی را نشان می دهند که داده آموزشی کافی در اختیار ندارند و بنابراین این گره ها با گره های سطح بالاتر شریک می شوند. دایره‌های یک پارچه گره هایی هستند که داده آموزشی کافی برای تخمین انتقال دارند.



شکل ۱۹ **یک درخت بازگشت باینری برای ادغام اجزای تطبیق شده**

[93] transform pooling
[94] binary regression tree

یک الگوریتم ادغام نیاز به تعیین اجزایی دارد که در هر گره با هم ادغام شده اند. در اغلب موارد از یک الگوریتم جدا کننده مرکز ثقل[۹۵] ، برای جدا کردن اجزای یک گره به دو مجموعه از اجزا استفاده می شود. این عمل طوری انجام می شود که به ما اطمینان می دهد، اجزایی که به یکدیگر نزدیکترند، در فضای احتمال در یک گره با هم ادغام شوند.

[95] Centroid splitting algorithm

11 بازشناسی هویت از طریق گفتار

11-1 مقدمه:

بازشناسی گفتار یکی از مهمترین مباحث تحقیقات گفتاری است که در طول پنج دهه گذشته، یعنی از زمانی که چنین تحقیقاتی در اوایل دهه ۵۰ میلادی آغاز شد، پیشرفتهای زیادی داشته است.

فرایند بازشناسی گوینده[۹۶] که در اینجا به شاخه عمومی مسائلی از قبیل تعیین هویت گوینده[۹۷]، تصدیق گوینده[۹۸] و طبقه بندی گوینده اطلاق می شود، اولین بار توسط آتال[۹۹] در این زمینه مطرح شد. این اولین قدم در معرفی این زمینه تحقیقاتی بود که البته در ابتدا از موفقیت کمی برخوردار بود.

تعیین هویت گوینده به این صورت تعریف می شود که از میان N مدل گوینده مرجع، آن مدل گوینده ای که نزدیکترین و بیشترین شباهت را به گوینده نا مشخصی ورودی دارد پیدا می کند. از آنجایی که الگوهای گفتار با تک تک مدلهای مرجع مقایسه می شوند و همچنین از آنجایی که برای هر تصمیم گیری نادرست احتمال معینی برای هر مقایسه وجود دارد، بنابراین واضح است که احتمال تصمیم گیری کلی تابعی از N بوده و در نتیجه هر چه تعداد مدل ها بیشتر باشد احتمال خطا یا تصمیم گیری نادرست در تعیین هویت گوینده بیشتر می شود.

مسئله تصدیق گوینده به این صورت مطرح می شود که گفتار یک هویت گوینده نامشخص و مدل گوینده ای که وی ادعا می نماید داده شده است و بایستی مشخص شود که آیا گفتار این گوینده به اندازه کافی به مدل گوینده ادا شده شباهت دارد یا نه. بنابراین در این حالت تعداد مقایسه یکی است و معمولاً مستقل از تعداد جمعیت مدل مرجع است.

در یک تقسیم بندی، بسته به اینکه مدلی که برای شناسایی مورد آزمون قرار می گیرد جزو مدلهای مجموعه باشد یا نباشد، دو تعریف زیر ارائه می گردد:

---

[۹۶] Speaker Recognition

[۹۷] Speaker Identification

[۹۸] Speaker Verification

[۹۹] Atal

۱. تعیین هویت در مجموعه بسته: که در این حالت مدل مورد آزمون جزو N مدل موجود در شبکه می باشد.

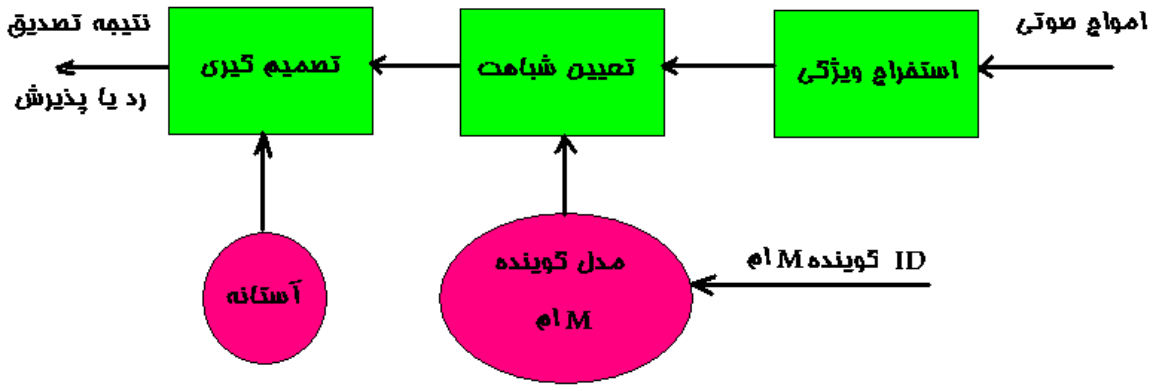
۲. تعیین هویت در مجموعه باز: که در این حالت گفتار ورودی نامشخص متعلق به هیچ کدام از اعضای مدلهای گویندگان مرجع موجود در سیستم نمی باشد.

در سیستم های تعیین هویت در مجموعه باز نتایج خروجی N+1 حالت می باشد که N تعداد مدلهای مرجع است. در واقع در این نوع سیستم، ترکیبی از دو نوع سیستم تعیین هویت مجموعه بسته و تصدیق هویت می باشد. دراین سیستم ابتدا نزدیکترین مدل به گفتار ورودی پیدا شده و سپس صحت تعلق گفتار ورودی به آن مدل مورد بررسی قرار می گیرد. بنابراین کارایی این سیستم نسبت به تعیین هویت مجموعه بسته کمتر است.

تصدیق هویت گوینده راحتی و اطمینان بیشتری را برای بسیاری از فعالیتهای روزمره آدمی که نیاز به امنیت دارد ایجاد می نماید. این تکنولوژی بدلیل اینکهاز خصوصیات ذاتی موجود در صدای انسان استفاده می کند، در مقابله با مسئله تقلید و کلاه برداری بسیار مقاوم بوده و از لحاظ ویژگی بازشناسی هویت از راه دور نسبت به برخی روشها از جمله اثر انگشت کارایی بسیار بیشتری دارد.

2-11 روشهای پیاده سازی سیستم های تصدیق گوینده:

شمای کلی یک سیستم تصدیق گوینده در شکل زیر آمده است:



شکل ۲۰ **شمای کلی یک سیستم تصدیق گوینده**

سیگنال گفتار گوینده به عنوان ورودی به سیستم انتقال داده می شود. سپس پیش پردازش و نهایتاً استخراج ویژگی ها بر روی آن اعمال می گردد و به این ترتیب به یک فضای جدید منتقل می گردد. ویژگی های استخراج شده از گفتار گوینده با مدل مرجع او مقایسه می شود. عمل مقایسه به این صورت

انجام میگیرد که میزان شباهت یا اختلاف این ویژگی ها با مدل مرجع به دست می آید. میزان شباهت به دست آمده با میزان آستانه مقایسه می شود و با توجه به این مقایسه خروجی سیستم مبنی بر تعلق و یا عدم تعلق گفتار ورودی به گوینده ادعا شده مشخص می شود.

قبل از معرفی روشهای پیاده سازی سیستم، به بررسی نکاتی در طراحی سیستم های بازشناسی (اعم از تعیین هویت و یا تصدیق هویت) می پردازیم:

۱- انتخاب ویژگی ها: برای استخراج ویژگی های گوینده بایستی چند عامل مد نظر قرار گیرد؛ از جمله: این ویژگی ها نبایستی به نویز محیط انتقال مانند کانال های مخابراتی حساس باشند، همچنین این ویژگی ها نبایستی به حالات روانی و فشار های محیط که گوینده در آن در حال صحبت کردن است وابسته باشد. در کل روش بهینه ای برای استخراج ویژگی ها وجود ندارد و معمولاً از طریق تجربه این عمل صورت می گیرد.

لازم به ذکر است که در بازشناسی گوینده از ویژگی های دینامیک که بستگی زیادی به حالات گوینده دارد استفاده می شود، در صورتی که در روشهای دیگر خصوصیات فیزیکی استاتیک و پایدار مورد بررسی قرار می گیرند. بنابراین یک سری محدودیتهای ذاتی در مورد استفاده از این سیگنالها وجود دارد. برای درک این محدودیتها بایستی اطلاعات تمایزدهنده گوینده ها و نحوه قرار گرفتن آنها در سیگنال گفتار مورد بررسی قرار گیرد. سیگنال گفتار از طریق حرکت اندامهای تولید گفتار بوجود می آید و توسط حنجره و سیستم عصبی کنترل می گردد. بنابراین دو منع اطلاعات گوینده در سیگنال صحبت وجود دارد یکی مربوط به خصوصیات فیزیکی و ساختاری مجرای گفتار و دیگری اطلاعات کنترلی از مغز و ماهیچه ها اندام گویایی است. این اطلاعات همراه با اطلاعاتی مربوط به هنگام حرکت دادن مفصل های اندامهای تولید گفتار، وارد سیگنال صحبت می شود. در کل اطلاعات سیگنال گفتار را به دو دسته سطح بالا (مانند: لحن، محتوای گفتار و استیل گفتار یعنی طریقه استفاده گرامری و نحوی از کلمات) و سطح پایین (مانند: خصوصیات سیگنال گفتار از قبیل دامنه طیف، فرکانی گام واکدار، فرکانس فرمانت، پهنای باند و خصوصیات تناوبی گفتار واکدار) تقسیم بندی می کنند. اطلاعات سطح بالا عملاً در بازشناسی گوینده توسط انسان کاربرد داشته و در عوض سیستم های بازشناسی گوینده اتوماتیک از ویژگی های سطح پایین سیگنال استفاده می شود.

۲- مدل های گویندگان: در سیستم های بازشناسی گوینده، مدل هر گوینده شامل خصوصیات آماری او است. معمولاً در مدل هی گویندگان از یکی از دو مدل پارامتری (مانند مدل گوسی) و غیر پارامتری (مانند مدل مراکز خوشه ها یا حالات چند گانه) استفاده می شود.

۳- طول گفتار آموزشی: که با افزایش این زمان نتایج مطلوب تری حاصل می گردد.

۴- انتخاب گفتار مناسب: پیشنهاد می گردد سکوت وبخشهای غیر گفتاری حذف گردد.همچنین

گفتارهای کم انرزی و بی واک در تمایز بین گوینده ها کم اثر بوده و پیشنهاد می گردد که از اطلاعات مدل حذف گردند.

۵- محیط نویزی: نویز پشت زمینه یک مشکل عمومی است و بهتر است برای تخمین آن از نویز پشت زمینه در زمان سکوت استفاده و با اعمال آن به مدل،تخمینی از مدل بدون نویز داشته باشیم.

۶- تنوع گوینده: اغلب سیگنال گفتار یک شخص در حالت های خوشحالی، ناراحتی و خستگی متفاوت و به این ترتیب کارایی سیستم را کاهش می دهند.بنابراین خصوصیات آماری یک فرد ثابت نیست.

12 سنتز گفتار (Speech Synthesis)

12-1 مقدمه

سنتز گفتار یک فناوری است که بوسیلهٔ آن متن به گفتار تبدیل می شود(TTS). یکی از کاربرد های سنتز گفتار تبدیل email و Fax به صدا و اخیراً تبدیل محتویات صفحات وب به صوت می باشد. نرم افزارهایی که در این زمینه طراحی شده اند ، بسیار زیاد هستند. با وجود آمدن گفتار مصنوعی نرم افزارهایی برای خواندن کتابها طراحی شده اند.

عروض(prosody)

عروض یکی از فاکتورهای اصلی برای بدست آوردن یک گفتار مصنوعی با کیفیت زیاد می باشد. مفهوم عروض ، زیر و بم کردن صدا و ریتم گفتارکه باعث تلفظ و برداشت مفهوم های مختلفی از گفتار می گردد، می باشد. همچنین اطلاعاتی راجع به وضعیت روانی سخنگو به ما می دهد. صدای هر شخص دارای یک فرکانس اصلی می باشد. که بر اساس آن تارهای صوتی به لرزه در می آیند. می توان گفت فرکانس اصلی یکی از فاکتورهای جداشوندهٔ صدا می باشد و آن قابل تغییر است. بطور ساده ، فرکانس اصلی بوسیلهٔ اصلاح شکل موج قابل دستکاری کردن می باشد. با استفاده از تکنیک پردازش سیگنال ، امواج صوتی می توانند از این تکنیک ها استفاده کنند.صدای مصنوعی کیفیت صدای طبیعی را ندارد چندین ایده را هنگام انجام کار می توان مد نظر قرار داد. یکی از مشکلات اساسی و لاینفک در این موضوع مسألهٔ کنترل وقفه های زمانی بوسیلهٔ سیستم کنترل امواج می باشدکه به منظور تولید یک صدای خوب و قابل فهم توسط انسان بر این مشکل بایستی غلبه گردد. هنگامیکه سنتز گفتار در یک سیستم بزرگ محاوره ای بکار گرفته شود، استفاده الگوهای عروضی گفتار اهمیت بحرانی داشته و استفاده نادرست از آن منجر به شکست خواهد شد.غلبه بر مشکلات عروضی گفتار بصورت دستکاری کردن شکل موج می باشد. بعضی از محققان با جلوگیری کردن از طولانی شدن متنها تلاش در کم کردن این مشکلات را دارند.

12-2 انواع متدهای تبدیل متن به گفتار

تکنیک های سنتز گفتار در سه مقوله زیر قابل بررسی می باشند :

۱- سنتز شمرده به شمرهٔ لغات (Articulatory Synthesis)

۲- سنتز فرمانت (Formant Synthesis)

۳- سنتز اتصالی (Concatinative Synthesis)

۱۲-۲-۱ سنتز شمرده به شمردهٔ لغات(Articulatory Synthsis)

در این روش از یک مدل کامپیوتری – مکانیکی برای تولید گفتار استفاده می شود. مثل حنجره که بوسیلهٔ ارتعاش تارهای صوتی باعث تولید صدا و تحریکات حلقی می شود. زبان بوسیلهٔ فشرده کردن هوا در محوطهٔ دهان باعث جدا کردن صداها از همدیگر می شود. مسیر عبور هوا بوسیلهٔ زبان با فشار بر روی سقف دهان یا دندانها تغییر می کند. لبها برای ایجاد هماهنگی و کمک به فرایند Articulation به زبان کمک می کنند. در حالت ایده آل سنتز کننده های شمرده به شمرده می توانند بوسیلهٔ ماهیچه های مصنوعی و کنترل آنها به شکل زبان ، لب و دهان شبیه سازی شوند. این کار مستلزم حل معادلات دیفرانسیل درجه سه وابسته به زمان می باشد. در هر صورت محاسبهٔ ایجاد گفتار در خروجی بسیار مشکل می باشد و نتیجهٔ آن در اکثر موارد تولید یک صدای غیر طبیعی است.

۱۲-۲-۲ سنتز فرمانت (Formant Synthesis)

سنتز فرمانت از یک سری از قوانین برای کنترل یک شکل موج که گفتار انسان را شبیه سازی می کند ، استفاده می کند. این روش باعث تولید یک گفتار ماشینی صحیح ، شبیه صدای روبات می باشد. اکثر نرم افزارهای TTS از این شکل سنتز برای تولید گفتار استفاده می کنند. روش سنتز فرمانت باعث تولید گفتاری قابل فهم و واضح می گردد ، اما صدای آن همانند صدای طبیعی نمی باشد. از مزایای این روش استفادهٔ معقول از حافظه و محاسبات کامپیوتری می باشد. سنتز فرمانت بطور گسترده برای پاسخگویی به کاربر مورد استفاده قرار می گیرد.

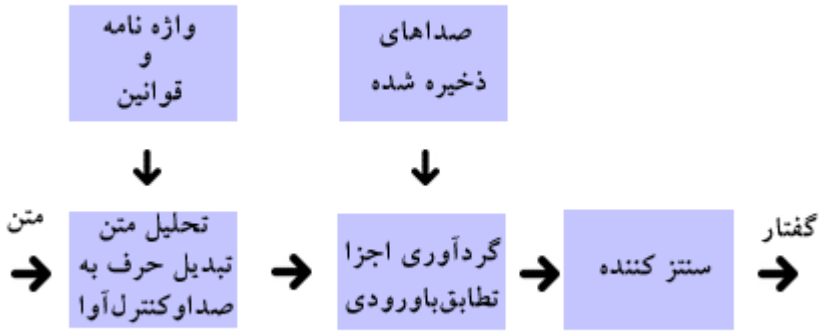
۱۲-۲-۳ سنتز اتصالی (Concatinative Synthesis)

در این رویکرد گفتار ذخیره شدهٔ طبیعی بصورت تکه تکه در کنار هم قرار می گیرند تا تولید یک گفتار خروجی کنند. به عنوان مثال ، برای اینکه به کاربر مقدار حساب بانکی اش را بیان کنیم باید لغات زیر را با اتصال به یکدیگر بیان کنیم.

"Your + total + account + balance + is + five + thousand + three + hundred + forty + nine + dollars + and + twenty + six + cents"

به علت اینکه فاکتورهای زیادی می توانند بر روی ویژگی های قطعات گفتاری در یک زبان طبیعی تاثیرگذار باشند در سیستمهای امروزی قطعات گفتاری زیادی در نوعها و قالبهای مختلف ذخیره شده است. همچنین برای زبانهای مختلف ، قطعات جداگانه ای ذخیره شده است. قطعه های گفتار می توانند به شکل موج خام و یا مجموعه ای از پارامترهای مشتق شده از انواع گوناگون موج ذخیره شوند. اگرچه

ذخیره قطعه ها بصورت مجموعه ای از پارامترها فضای کمتری را اشغال می کند، اما روند حرکت این تکنولوژی بر اساس ذخیرهٔ موج خام بوده است.هدف از این کار تولید صدایی شبیه به صدای انسان می باشد. هنگام سنتز یک جمله ، قطعه ها از بانک اطلاعات استخراج می شود و به هم متصل می گردند و به منظور داشتن فاصله زمانی مناسب و تلفظ مناسب اصلاح می شوند. یکی از کارهای مشکل و پیچیدهٔ این روش، شناخت و انتخاب قطعهٔ مناسب بر اساس ساختار جملهٔ گفتاری مورد نظر می باشد. اجزای گفتار مثل واجها شامل نیمهٔ اول یک صدا و نیمهٔ دوم صدای دیگر می باشند. بعضی از سنتز کننده های الحاقی که به آنها demisyllables گفته می شود از diphon برای مقیاس زمانی سیلابها استفاده می کنند. معنای diphon زمان گذرای بین واجها و داده ها و واجهای ضبط شده می باشد و معنای دمی سیلاب زمان گذرای نیم سیلاب می باشد(این زمان می تواند از زمان شروع تا وسط سیلاب و یا از وسط سیلاب تا پایان سیلاب در نظر گرفته شود) . در شکل ۲۱ بلوک دیاگرام TTS الصاقی نشان داده شده است.



شکل ۲۱ **فرایند سنتز گفتار**

بلوک اول ماژول تحلیل متن متن را بصورت کد اسکی دریافت کرده و آن را بصورت یک سری از سمبلهای تلفظ و عروض ، شامل فرکانس اصلی و فاصله های زمانی و دامنهٔ موج در می آورد. ماژول تحلیل متن از یک سری از ماژولهای جداگانه تشکیل شده است. متن ورودی در ابتدا تحلیل می شود و سمبلهای غیر از حروف الفبا و حروف اختصاری به لغات کامل تبدیل می شوند. برای مثال در جملهٔ :

Dr. Kay lives at 5201 Peak Dr.

کلمهٔ .Dr اول به کلمهٔ doctor تبدیل می شود و کلمهٔ دوم به کلمهٔ Drive و پس از آن ۵۲۰۱ به صورت پنج – صفر – صفر – یک تبدیل می شود. پس از آن پارسر سنتز کننده که وظیفهٔ آن تشخیص قطعه های گفتار برای هر لغت در جمله می باشد نشانه گذاری روی متن را انجام می دهد. ماژول بعدی دارای دو جزء اصلی است :

۱. جزء نرمال ساز متن

۲. جزء تجزیه کنندهٔ متن

قوانین نرمال کنندهٔ متن ، متن ورودی را می خواند و بصورت رشته ای از قطعات گفتاری نمایش داده و لغات مخفف را بصورت حرف حرف در می آورد. برای مثال جملهٔ St. John St. به جملهٔ Saint john Street تبدیل می شود. یا 1 oz به one ounce تبدیل می شود. یا $500 به پانصد دلار تبدیل می شود. کاربر می تواند بعضی از رفتارهای سیستم را با دستکاری بعضی از لغات در واژه نامه تغییر دهد. قوانین نرمالسازی ، زمان پایان جمله را تعیین می کنند. به مثال زیر که برای جملهٔ Dr. Smith is 43 آمده است توجه کنید:

Sentence : | sent |
Inp : | D | r | `.` | `` | S | m | i | t | h | `` | i | s | `` | 4 | ` | 3 | . |
Text : | d | o | c | t | o | r | `` | s | m | i | t | h | `` | i | s | `` | f | o | r | t | y | ` | t | h | r | e | e | . |

خطوط عمودی برای هماهنگ سازی رشتهٔ متن در نقاط مناسب مورد استفاده قرار می گیرد. نتیجهٔ حاصل از نرمال سازی جمله برای قوانین تجزیه کنندهٔ متن فرستاده می شود و در آنجا پردازش کلامی صورت می گیرد. قوانین تجزیهٔ متن عمل خود را روی متن نرمال شدهٔ قبل انجام می دهند و آن را به عبارتهای طبیعی – لغات – سیلابها و diphone ها می باشد. در شکل قبل دومین بلوکی که یونیتهای مختلف را برای ایجاد یک صدای طبیعی گردآوری می کند ، نشان داده شده است. این اطلاعات به یک سنتز کننده ، داده شده که برای کاربر تولید یک شکل موج گفتار می نماید.

یکی از کارکردهای کلیدی سنتز گفتار ، گردآوری رشتهٔ صحیحی از صداها به همراه یک واژه نامهٔ تلفظ کمکی می باشد. بنابراین در جملهٔ Dr. Kay lives at 5201 Peak DR فعل live از جمله شناسایی می شود و همچنین در جملهٔ (to Mr Wright to give him the right direction Write a letter) سنتز کلمات wright و right و write به راحتی انجام می شود. اگر واژه نامهٔ لغات دچار ایراد شود قوانین کلی تبدیل حروف الفبا به صدا اعمال می شوند. در پایان بوسیلهٔ جملات نقطه گذاری شده و اطلاعات صوتی و نحوی موجود ، یک ماژول عروضی عملیات تولید یک عبارت را برای مقصد مورد نظر انجام می دهد.

برای مثال برای فرکانس اصلی ، فاصلهٔ زمانی بین واجها و دامنه توسط ماژول تعیین می شود. در سالهای اخیر TTS پیشرفتهای خوبی در زمینهٔ تولید صدای طبیعی داشته است که نتایج آن در پورتالهای صوتی و سرویس بانکها و آژانسهای مسافرتی شنیده می شود ، بطوریکه برای مشتری های این شرکتها تشخیص صدای کاربر واقعی و مصنوعی قابل تشخیص نمی باشد. سنتز یک بخش انتخاب شده بوسیله عوامل زیر قابل تاثیر است. یکی از مهمترین دیدگاهها به این موضوع افزایش قدرت پردازشی و ذخیره سازی سرور می باشد که بطور مستقیم بر روی سایز دستگاههای ذخیره سازی صوت تاثیر دارد.

امروزه سنتز اتصالی در بسیاری از موارد استفاده می شود. نقطه نظر دیگر نیز این است که تکنیک های کارای جستجو به ما اجازه می دهند که میلیونها بخش صوتی متفاوت را بصورت بلادرنگ برای بدست آمدن یک گفتار بهینه مورد جستجو قرار دهیم. درپایان اینکه ، می توانیم از برچسب زننده های اتوماتیک که باعث سرعت بخشی دسترسی بانک اطلاعات عروض و تلفظ می شوند ، استفاده کنیم.

3-12 ارزیابی سنتز گفتار

سه فاکتور کلیدی برای داشتن یک گفتار مصنوعی موفق وجود دارد.

۱-مفهوم بودن (Intelligibility)

درجه ایست که شنونده گفتار را متوجه می شود. این میزان بوسیلهٔ اندازه گیری تعداد واجهای تولید شده بوسیلهٔ یک نرم افزار در مقابل تعداد واجهایی که توسط کاربر قابل تشخیص است ، قیاس می شود.

۲-طبیعی بودن (Naturalness)

درجه ایست که شنونده گفتار سنتز شدهٔ مصنوعی را همانند گفتار طبیعی انسان می پندارد. اندازه گیری آن از لحاظ ذهنی بیشتر از قابلیت وضوح صدا می باشد ولی از نظر اهمیت در یک سطح قرار دارند.

۳-مطبوعیت(Pleasantness)

درجه ای است که بر اساس آن شنونده از صدا لذت می برد یا آن را قابل تحمل می داند. تعیین درجه های فوق بسته به نوع کاربران و نوع سرویس متفاوت است.

13 دادگان های گفتاری

13-1 مقدمه

استفاده از گفتار جهت برقراری ارتباط با ماشین‌ها در سال‌های اخیر مورد توجه پژوهشگران قرار گرفته است. امروزه با توجه به افزایش قدرت پردازشی پردازنده‌ها و حجم حافظه‌ها تمایل افراد به سمت ذخیره-ی اطلاعات گفتاری در رایانه‌ها بیشتر شده و این سبب شده است تا جستجو، تشخیص و پردازش‌های لازم مستقیما روی فایل‌های گفتاری ذخیره شده در رایانه اعمال شود.

در تهیه‌ی دادگان‌های صوتی قواعد خاصی مورد نظر قرار می‌گیرد از جمله پایگاه داده باید شامل کلمات و عبارات مناسبی باشد که معرف مجموعه گفتاری زبان مورد نظر باشد و از طرفی گویندگان باید از جنسیت، سن و سطح سوادهای مختلف جهت پوشش دادن مشخصه‌های گفتاری مختلف، انتخاب شوند.

در فصل حاضر مطالعه‌ی به نمونه هایی از پایگاه داده‌های گفتاری، خصوصیات آنها و کاربردشان اشاره می شود. برای تهیه دادگان گفتاری جدید خصوصیات دادگانهای موجود باید مطالعه و بررسی شده و بسته به کاربرد مورد نظر دادگان جدید طراحی گردند.

13-2 دادگان گفتاری فارس‌دات

دادگان فارس‌دات یک بانک اطلاعاتی از گفتار فارسی است که توسط پژوهشکده هوشمند علائم ارائه گردیده است. این دادگان حاوی ۳۸۶ جمله (ساخته شده با ۱۰۰۰ کلمه شامل تمامی دو واجها) است که توسط ۳۰۰ گوینده و با ۱۰ لهجه مختلف بیان شده‌اند. هر گوینده ۲۰ جمله را ادا کرده است. در مجموع ۶۰۰۰ جمله متمایز در فارس‌دات موجود است. همچنین ۱۱۰۰۰ کلمه مجزا نیز توسط ۱۰۰ گوینده بیان شده است. کلیه واجهای موجود در جملات این بانک اطلاعاتی دارای برچسب زمانی هستند. این دادگان می تواند در سیستم‌های بازشناسی گفتار، بازشناسی گوینده و لهجه بکار گرفته شود.

13-3 دادگان گفتاری فارس‌دات بزرگ

دادگان فارس‌دات بزرگ یک بانک اطلاعاتی از گفتار فارسی است که توسط پژوهشکده هوشمند علائم

ارائه گردیده است. این دادگان حاوی ۱۴۰ ساعت گفتار است که توسط ۳۰۰ گوینده و با ۱۰ لهجه مختلف بیان شده‌اند. هر گوینده ۴۰۰۰ کلمه را ادا کرده است. کلیه سیگنالها موجود در این بانک اطلاعاتی دارای برچسب در سطح کلمه هستند.

مناسب جهت کاربردهایی از قبیل بازشناسی گفتار پیوستهٔ فارسی مستقل از گوینده با واژگان بزرگ، بازشناسی گوینده مستقل از متن و تحقیقات در زمینهٔ واج‌شناسی آزمایشگاهی می باشد.

13-4 دادگان گفتاری فارس‌دات تلفنی (مونولوگ)

دادگان فارس‌دات تلفنی یک بانک اطلاعاتی از گفتار فارسی است که توسط پژوهشکده هوشمند علائم ارائه گردیده است. این دادگان توسط ۶۴ گوینده و با لهجه‌های مختلف بیان شده‌اند. این دادگان شامل ۷ ساعت سیگنا گفتاری است. کلیه سیگنالها موجود در این بانک اطلاعاتی دارای برچسب در سطح واج هستند. برای کاربردهایی از قبیل سیستم‌های رایانه‌ای تلفنی در بازشناسی گفتار و بازشناسی گوینده و بازشناسی زبان فارسی مناسب می باشد.

13-5 دادگان گفتاری فارس‌دات تلفنی بزرگ (محاوره ای)

دادگان فارس‌دات تلفنی بزرگ یک بانک اطلاعاتی از گفتار فارسی است که توسط پژوهشکده هوشمند علائم ارائه گردیده است. این دادگان حاوی ۲۵۲ مکالمه تلفنی دو طرفه پیرامون ۱۳ موضوع مختلف (شامل موضوعات فرهنگی، سیاسی، اجتماعی، اقتصادی، علمی، آموزشی، هنری، خانواده، ورزشی، صنعتی و فناوری، کشاورزی و دامداری، عمران و خدمات) است که توسط ۲۰۰ گوینده و با ۱۰ لهجه مختلف بیان شده‌اند. بخشی از سیگنالها موجود در این بانک اطلاعاتی دارای سه نوع برچسب زبانی در سطح کلمه، شامل برچسب‌واجی، آوایی و نوشتاری هستند. همچنین عناصر غیرزبانی از قبیل نوفهٔ خط تلفن، نوفهٔ محیطی، صدای تنفس، صدای برخورد لب‌ها صدای شک و تردید و کلماتی که ناقص تولید می‌شوند، برچسب‌دهی شده اند.

مناسب جهت کاربردهایی از قبیل بازشناسی گفتار پیوستهٔ تلفنی مستقل از گوینده با واژگان بزرگ، بازشناسی گویندهٔ تلفنی مستقل از متن، بازشناسی گفتار مکالمه‌ای روی خط تلفن به منظور خودکارسازی مراکز تلفنی، تحقیقات واج‌شناسی آزمایشگاهی و تجزیه و تحلیل کلام (گفتمان) می باشد.

13-6 دادگان اعداد و ارقام منفصل و پیوستهٔ فارسی

دادگان اعداد متصل و پیوسته فارسی یک بانک اطلاعاتی از اعداد فارسی است که توسط پژوهشکده هوشمند علائم ارائه گردیده است. تعداد گویندگان اعداد پیوسته ۱۰۰ نفر و تعداد گویندگان اعداد متصل ۱۱۰ نفر با لهجه تهرانی می باشد. این بانک اطلاعاتی دارای برچسب در سطح عدد هستند.

13-7 دادگان گفتاری TIMIT

دادگان TIMIT یک بانک اطلاعاتی از گفتار پیوسته انگلیسی است که توسط شرکت TI[100] و دانشگاه MIT[101] تهیه شده است و اداره استاندارد آمریکا (NIST[102]) آن را تأیید کرده است. این دادگان حاوی ۶۳۰۰ جمله است که توسط ۶۳۰ گوینده و با ۸ لهجه معمول آمریکای شمالی بیان شده‌اند. ۷۰% گویندگان مرد و ۳۰% آنها زن هستند. هر گوینده ۱۰ جمله را ادا کرده است که ۲ جمله از این ۱۰ جمله توسط سایر گویندگان نیز ادا شده است. در مجموع ۲۴۳۲ جمله متمایز در TIMIT موجود است که شامل ۲ جمله مشترک میان تمامی گویندگان، ۴۵۰ جمله مشترک میان گروههای ۷ نفری گویندگان و ۱۸۹۰ جمله تک گوینده است. کلیه کلمات و واجهای موجود در جملات این بانک اطلاعاتی دارای برچسب زمانی هستند. بانک اطلاعاتی TIMIT به دو بخش آموزش و آزمون تقسیم شده است که بخش آموزش شامل ۴۶۲ گوینده و بخش آزمون شامل ۱۶۸ گوینده است و گویندگان و جملات ادا شده در هر یک از این دو بخش با یکدیگر متفاوتند.

دادگان TIMIT عاری از نویز است و معمولاً برای ارزیابی نرخ بازشناسی واجها در بازشناسی گفتار پیوسته مورد استفاده قرار می‌گیرد. هر چند که با وجود برچسبهای زمانی برای کلمات و واجها، برای ارزیابی نرخ بازشناسی کلمات مجزا نیز قابل استفاده است. برای کاربرد این دادگان در ارزیابی روش‌های بازشناسی گفتار در حضور نویز، باید نویز را بطور مصنوعی به این دادگان اضافه نمود.

13-8 دادگان گفتاری TIDIGITS

دادگان TIDIGITS یک بانک اطلاعاتی گفتار انگلیسی است که در شرکت TI تهیه شده است. هدف از تهیه این دادگان طراحی و ارزیابی الگوریتمهای بازشناسی مستقل از گوینده برای دنباله متصل اعداد[103] بوده است. این دادگان حاوی گفتار ۳۲۶ گوینده متشکل از ۱۱۱ مرد، ۱۱۴ زن، ۵۰ پسر بچه و ۵۱ دختر

[100] Texas Instruments
[101] Massachusetts Institute of Technology
[102] National Institute for Standards and Technologies
[103] Connected Digit Sequences

بچه است که هر یک از گویندگان ۷۷ دنباله متصل عددی را بیان کرده اند. دنباله‌های عددی بیان شده با استفاده از ۱۱ رقم زیر ساخته شده‌اند:

zero, oh, one, two, three, four, five, six, seven, eight, nine.

این دنباله‌های عددی بیان شده برای هر گوینده را می‌توان این گونه تقسیم کرد: ۲۲ بار بیان اعداد مجزا (۲ بار بیان برای هر عدد) و ۱۱ مرتبه بیان برای هر یک از رشته های متشکل از ۲، ۳، ۴، ۵ و ۷ عدد که در مجموع ۷۷ دنباله مذکور را برای هر گوینده شکل می‌دهند. تمامی این دنباله‌ها به دو زیر مجموعه دادگان آموزش و دادگان آزمایش تقسیم شده‌اند.

دادگان TIDIGITS نیز عاری از نویز است و برای ارزیابی روشهای بازشناسی گفتار در حضور نویز باید به آن بطور مصنوعی نویز اضافه نمود.

9-13 دادگان گفتاری AURORA2

دادگان گفتاری AURORA2 برای ارزیابی دقت بازشناسی سیستم‌های بازشناسی گفتار در محیط‌های نویزی و به ویژه برای ارزیابی روش‌های جبران ویژگی در شرایط نویزی متفاوت مورد استفاده قرار می-گیرد. این دادگان با استفاده از گفتار بزرگسالان دادگان گفتاری TIDIGIT ساخته شده است، به این ترتیب که ابتدا نرخ نمونه‌برداری به ۸ کیلو هرتز کاهش یافته است و سپس فیلتری برای شبیه‌سازی تأثیر کانال انتقال تلفنی بر روی دادگان اعمال گردیده است. برای شبیه‌سازی اثر خط تلفن، دو نوع فیلتر استاندارد مورد استفاده قرار گرفته‌اند که با نامهای G.712 و MIRS شناخته می‌شوند.

گفتار حاصل از اعمال فیلترهای فوق، دادگان تمیز گفتاری را شکل می‌دهند. سپس به این دادگان تمیز نویزهای جمع‌پذیری با نسبتهای سیگنال به نویز ۲۰، ۱۵، ۱۰، ۵، ۰ و ۵- دسی‌بل اضافه می‌شوند. نویزهای جمع‌پذیر بکار رفته عبارتند از: نویزهای مترو[104]، همهمه[105]، ماشین[106]، نمایشگاه[107]، رستوران[108]،خیابان[109]، فرودگاه[110] و ایستگاه قطار[111].

در دادگان AURORA2 دو مجموعه مجزای آموزش و آزمایش مشخص شده‌اند. داده‌های آموزش خود

[104] Subway
[105] Babble
[106] Car
[107] Exhibition
[108] Restaurant
[109] Street
[110] Airport
[111] Train Station

به دو گروه تقسیم می‌شوند: داده های آموزش تمیز و داده‌های آموزش مختلط (هم تمیز و هم نویزی) است. این دو مجموعه آموزش بوسیله فیلتر G.712 فیلتر شده‌اند و شامل ۸۴۴۰ جمله می‌باشند. در دادگان آموزش مختلط ۸۴۴۰ جمله مذکور به ۲۰ زیرمجموعه ۴۲۲ جمله‌ای دسته‌بندی شده‌اند. این زیرمجموعهها در برگیرنده ۴ نوع نویز مترو، همهمه، ماشین و نمایشگاه با چهار نسبت سیگنال به نویز ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل و همچنین چهار زیرمجموعه بدون نویز می‌باشند.

دادگان آزمایش نیز به سه مجموعه تقسیم می‌شوند که با نامهای A، B و C شناخته می‌شوند. دادگان آزمایش A شامل ۴۰۰۴ جمله فیلتر شده با فیلتر G.712 می‌باشند که به چهار زیرمجموعه ۱۰۰۱ جمله‌ای تقسیم شده‌اند. علاوه بر زیرمجموعه‌های جداگانه تمیز مذکور، به هر یک از آنها نیز نویزهای مترو، همهمه، ماشین و نمایشگاه با نسبتهای سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل افزوده گردیده‌اند. دادگان آموزش B نیز مشابه با دادگان آموزش A ساختاردهی شده‌اند با این تفاوت که در آنها از نویزهای جمع‌پذیر رستوران، خیابان، فرودگاه و ایستگاه قطار استفاده شده است. برای مجموعه B عدم تطبیق میان داده های آموزش و آزمایش حتی برای دادگان آموزش مختلط نیز وجود دارد، زیرا دادگان آموزشی مختلط تنها شامل نویزهایی است که به مجموعه A اضافه شده‌اند. دادگان آزمایش C شامل دو زیرمجموعه از چهار مجموعه ۱۰۰۱ جمله‌ای است. در این مجموعه داده‌های گفتار و نویز بوسیله فیلتر MIRS فیلتر شده‌اند و سپس نویزهای مترو و خیابان با نسبتهای سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵ و ۲۰ دسی‌بل به گفتار افزوده شده‌اند. در این مجموعه نیز داده‌های تمیز بطور جداگانه وجود دارند. به این ترتیب مجموعه C برای آزمایش اثر کانالهای متفاوت مناسب است.

14 نتیجه گیری

در این گزارش به بررسی اجمالی پردازش گفتار و معرفی پایگاه های گفتار فارسی موجود پرداختیم. چنانکه دیده شد، هنوز خلا داشتن پایگاه داده گفتاری در اکثر کاربردهای پردازش گفتار دیده می شود.

در ادامه ویژگیهای مهمی که در طراحی پایگاه داده گفتار باید مدنظر قرار بگیرد آورده شده است.

۱- تعداد زبان

تعداد زبانهایی که در پایگاه داده ی گفتاری وجود دارد، مانند فارسی، انگلیسی.

۲- تعداد گویندگان

تعداد گویندگان نیز از اهمیت خاص خود برخوردار است.

۳- ویژگی های گویندگان

الف- یک زبانه بودن یا چند زبانه بودن

زبان مادری گوینده فارسی است و یا گوینده ای است که زبان فارسی بعنوان زبان دوم آن محسوب می شود. چراکه زبان مادری برای گوینده حل شده است و هیچ تردیدی در مورد آن ندارد، این امر باعث می شود که ادای جملات با فونوتیک درست صورت پذیرد.

ب- پراکندگی سنی

وجود داده گفتاری از هر گروه سنی می تواند به عملکرد بهتر سیستمهای گفتاری کمک کند.

ج- لهجه (تعداد گویش)

بسته به کاربرد می توان تنوع لهجه مانند خراسانی، یزدی، اصفهانی را در ساخت پایگاه داده ی گفتاری لحاظ نمود و یا می توان فقط از گویش کتابی استفاده نمود.

د- تجربه گوینده

داشتن تجربه قبلی ضبط صدا می تواند باعث کاهش لرزش صدا ناشی از استرس و

اضطراب گردد. همچنین از تبق زدن نیز جلوگیری می شود. در این زمینه می توان از روزنامه نگاران، گویندگان اخبار، بازیگران زن و مرد، راویان و یا خوانندگان بهره جست.

- حالتهای رفتاری

سیگنال گفتار گوینده در مدهای مختلف رفتاری دارای ویژگیهای متفاوت خواهد بود. عصبانیت، ناراحتی، خوشحالی و ... نمونه هایی از حالات گوینده می باشد که برای ضبط گوینده نبایستی در یکی از این حالات بسر ببرد.

۴- مدهای مکالمه

الف- قالب مکالمه

مکلمه می تواند براساس متن از قبل تنظیم شده باشد و یا مکالمه فی البداهه در شرایط طبیعی باشد.

ب- کیفیت صدا

صدا باید خوشایند، سازگار و دارای کیفیت یکسانی برای تمامی گویندگان در تمامی جلسات ضبط صدا باشد.

ج- رسائی صدا

باید برای ضبط از گویندگانی استفاده شود که دارای رسائی صدا هستند. صداهای گنگ، مبهم و ضعیف باعث کاهش کیفیت سیگنالهای گفتاری خواهد شد.

۵- ویژگیهای سیگنالینگ

الف- فرکانس نمونه برداری

فرکانس نمونه برداری می تواند از 8kHz تا 96kHz متغیر باشد.

ب- اندازه هر نمونه

تعداد بیت مورد نیاز برای نمایش هر نمونه که غالباً ۸ یا ۱۶ بیت می باشد.

۶- ابزار و شرایط ضبط صدا

الف- نوع میکروفون

میکروفون می تواند بصورت رومیزی و یا بصورت هدست باشد. همچنین میکروفون هایی وجود دارد که مقداری از نویز را حذف می کنند که این امر ممکن است باعث ایجاد تخریب شیگنال گفتار شود. لذا در استفاده از این نوع میکروفونها بایستی دقت لازم را داشته باشیم.

ب- محیط نویزی یا بدون نویز (نرخ سیگنال به نویز)

دادگان گفتار را می توان در یک اتاق آکوستیک (استودیو ضبط صدا) انجام داد تا دادگان ضبط شده حاوی نویز نباشند. در این حالت در صورت نیاز به دادگان نویزی بطور مصنوعی به دادگان تمیز ضبط شده نویز اضافه می کنیم. همچنین دادگان را می توان در شرایط مختلف نویزی ضبط نمود که در این حالت امتیاز داشتن شرایط طبیعی نویزی را خواهیم داشت.

۷- ویژگی های پایگاه از لحاظ فونتیک

الف- گفتار پیوسته یا گسسته

ب- تعداد جملات / کلمات / فونمها و پراکندگی مناسب آنها

ج- سطح برچسب زنی و دقت آن

د- تنوع موضوع (مانند سیاسی، ورزشی و هنری)

در بالا به ویژگیهای یک پایگاه داده گفتار استاندارد اشاره شد. در ادامه به ویژگیهایی که پایگاه داده برای هر کاربرد بایستی داشته باشد اشاره خواهیم کرد.

۱- بازشناسی گفتار گسسته و پیوسته:

- داشتن دادگان آموزشی و تست بطور مجزا
- داشتن دادگان زیاد جهت آموزش و ارزیابی درست سیستم بازشناسی
- داشتن تنوع گویش، جنسیت، رده های سنی
- داشتن دو بخش دادگان تمیز و نویزی جهت ارزیابی سیستم بازشناسی در شرایط نویزی. این امر می تواند روشهای آموزش multi-condition و clean را فراهم کند.
- برچسب زنی دقیق در سطح واج، سه واج، آوا و کلمه یا برچسب دنباله واجها یا کلمات

- انتخاب مناسب متون بمنظور داشتن دادگان مناسب از هر نوع واج، سه واج یا کلمه
- داشتن فرکانس نمونه برداری و bit rate مناسب

۲- بازشناسی گفتار مستقل از گوینده

- تمامی موارد گفته شده برای بازشناسی گفتار گسسته و پیوسته
- داشتن گویندگان زیاد بمنظور داشتن تنوع زیاد در هنگام آموزش مدلها

۲- تطبیق گوینده

- تمامی موارد گفته شده برای بازشناسی گفتار گسسته و پیوسته
- داشتن دادگان به اندازه کافی از هر گوینده هم در بخش دادگان آموزش و هم در بخش دادگان تست.

۳- بازشناسی گوینده

- تمامی موارد گفته شده برای بازشناسی گفتار گسسته و پیوسته
- داشتن گویندگان زیاد هم در بخش آموزش و هم بخش تست. بهتر است که کسر بیشتری از گویندگان دو بخش آموزش و تست همپوشانی داشته باشند.
- برای دو روش کلی بازشناسی گوینده وابسته به متن و مستقل از متن تمهیدات لازم دیده شود.

۴- سنتز گفتار

- داشتن دادگان زیاد جهت بکارگیری از آن در سیستم بازشناسی
- برچسب زنی دقیق در سطح واج، سه واج، آوا و کلمه یا برچسب دنباله واجها یا کلمات
- انتخاب مناسب متون بمنظور داشتن دادگان مناسب از هر نوع واج، سه واج یا کلمه
- تنوع لهجه،گوینده و رده سنی انتخابی است (ضروری نمی باشد.)

۵- تشخیص کلمات کلیدی

- برای این کاربرد می توان از دادگان تهیه شده برای بازشناسی گفتار بهره جست، بشرطی که بخوبی در سطح واج و کلمه برچسب زنی دقیق داشته باشند.

در این مجموعه تلاش شد تا در ابتدا به معرفی اجمالی شاخه های پردازش گفتار پرداخته شود، سپس نگاهی به ویژگیهای دادگان گفتاری داشته، در انتها نیز بسته با کاربرد پردازش گفتار ویژگی دادگان

بررسی گردید.

15 منابع و مآخذ

[1] L. R. Rabiner. *"A tutorial on hidden Markov models and selected applications in speech recognition*". Proceeding of the IEEE.

[2] X. D. Huang, Y. Ariki, and M. Jack. *"Hidden Markov Models for Speech Recognition*". Edinburgh University Press, 1990

[3] Lalit R. Bahl, Fred Jelinek, and R. L. Mercer. *"A maximum likelihood approach to continuous speech recognition*". IEEE Transactions on Pattern Analysis and Machine Intelligence, 5(2):179-190.

[4] Kao, Yu-Hung, "*N-best search for continuous speech recognition using viterbi pruning for non-output differentiation states*", Texas Instruments Incorporated; April 16, 2002.

[5] Babak Nasersharif, Ahmad Akbari, Mohammad Mehdi Homayounpour, "Application of HMM adaptation and robust features to sub-band speech recognition in noise", *Proceeding of 11'th international CSI computer conference,* Vol. 2, pp. 215-219, Tehran, Iran, January 2006.

[6] Babak Nasersharif, Ahmad Akbari, "Sub-band weighted projection measure for robust sub-band speech recognition", *Proceeding of EUROSPEECH,* pp. 945-948, Lisbon, 2005.

[7] Babak Nasersharif, Ahmad Akbari, "Improved HMM entropy for robust sub-band speech recognition", *Proceeding of 12'Th European Signal Processing Conferences (EUSIPCO),* Turkey, 2005.

[8] Zhu, D., Paliwal, K.,"Product of power spectrum and group delay function for speech recognition", *IEEE International Conference on Acoustic, Speech, and Signal processing,* vol. 1, pp. 125-128, 2004.

[9] Ikbal, S., Misra, H., Bourlard, H., "Phase autocorrelation derived robust speech features", *IEEE International Conference on Acoustic, Speech, and Signal processing,* vol. 2, pp. 133-136, 2003.

[10] Murthy, H.A., Gadde, V., "The modified group delay function and its application to phoneme recognition", *IEEE International Conference on Acoustic, Speech, and Signal processing,* vol. 1, pp. 68-71, 2003.

[11] Tyagi, V. , McCown, I., Misra, H., Bourlard, H., "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR", *IEEE workshop on Automatic Speech Recognition & Understanding,* pp. 399-404, 2003.

[12] S. Young, G. Evermann, D. Kershaw, G. Moore, J. J. Odell, D. Ollason, V. Valtchev, P. C. Woodland, "The HTK Book, Version 3.1 ", Cambridge University Engineering Department, 2002.

[13] X. Huang, A. Acero, H. Hon, " Spoken language processing: a guide to theory,

algorithm and system development", Prentice Hall, 2001.

[14] P.C.Woodland, “Speaker Adaptation for Continuous Density HMMs: A Review”, Cambridge University, 2001.

[15] D.Kim, S.Lee, R.M.Kil ,”Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environment” , IEEE Trans. Speech & Audio Processing ,January 1999.

[16] J.Hakkien, J.Sountasta,R.Hariharan,M.Vasilache,K.Launilla,”Improved Feature Vector Normalization for noise Robust Connected Speech Recognition”, EuroSpeech’99 , vol.6 , pp.1833-2836,1999

[17] O.Vikki,K.Launilla,”Cepstral Domain Segmental Feature Vector Normalization”,Speech Communication, No. 25, pp.133-147, 1998.

[18] M.J.F.Gales, “Model-based Techniques for Noise Robust Speech Recognition”,Phd thesis,Cambridge University,1996.

[19] P.J.Moreno, “Speech Recognition in Noisy Environment”, Phd Thesis, Carnegie Mellon University,1996.

[20] P.C.Woodland,M.J.F.Gales,D.Pye,”Improving Enviromental Robustness in Large Vocabulary Speech Recognition”, Proc. ICASSP'96, Vol. 1, pp. 65-68, Atlanta,1996.

[21] B.A. Carlson, M.A. Clements, "A projection-based likelihood measure for speech recognition in noise", *IEEE Trans. on Speechand Audio Processing*, vol. 2, no. 1, pp. 97-102, January 1994.

[22] H.Hermansky , N.Morgan, “RASTA Processing of Speech “, IEEE Trans. Speech and Audio Processing, Vol.2, No. 4 ,pp. 587-589, October 1994.

[23] Alexandre, P., Lockwood, P., " Root cepstral analysis: A unified view. Application to speech processing in car noise environments", *Speech Communication*, Vol. 12, Iss. 3, pp. 277-288, July 1993.

[24] Murthy, H.A, Yegnanarayana, B. "Formant Extraction from group delay function", *Speech Communication*, vol. 10, pp.209-221, 1991.

[25] A.P.Varga, R.K.Moore,” Hidden Markov Model Decomposition of Speech and Noise”, ICASSP’90 , pp. 845-848, 1990.

[26] H.Hermansky, “ Perceptual Linear Preductive Analysis of Speech” , Journal of Acoustical Society of America, pp. 1738-1752, April 1990.

[27] D. Mansour, B. Juang, "A family of distortion measure based upon projection operation for robust speech recognition", *IEEETrans. on Acoustic, Speech and signal processing*, vol. 37, pp.1659-1671, 1989.

[28] S.F.Boll,”Suprression of Acoustic Noise in Speech using Spectral Subtraction”, IEEE Trans. Acoustics,Speech & Signal Processing, April 1979.

[۲۹] بابک ناصرشریف، احمد اکبری، محمد مهدی همایونپور، "ترکیب روشهای مبنتی بر مدل و پردازش چندباندی گفتار برای مقاوم سازی بازشناسی گفتار نسبت به نویز"، سیزدهمین کنفرانس مهندسی برق ایران، دانشگاه زنجان، صفحات ۲۰۹-۲۱۴، اردیبهشت ۱۳۸۴.