دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

پایان نامه کارشناسی ارشد
گرایش هوش مصنوعی

بازشناسی اعداد پیوسته از طریق خط تلفن

نگارنده:
امین فاضل دهکردی

استاد راهنما:
دکتر محمد تقی منظوری

دی ۱۳۸۳

دانشگاه صنعتی شریف

دانشکدهٔ مهندسی کامپیوتر

رساله کارشناسی ارشد

عنوان: بازشناسی اعداد پیوسته از طریق خط تلفن

نگارش: امین فاضل دهکردی

کمیتهٔ ممتحنین:

استاد راهنما: دکتر محمد تقی منظوری

امضاء

استاد مشاور: دکتر حسین ثامتی

امضاء

استاد مدعو: دکتر فرشاد الماس‌گنج

امضاء

تاریخ: ۱۳۸۳/۱۰/۲۶

تقدیم به:

پدر و مادر عزیزم که اسوهٔ ایثار و ازخودگذشتگی‌اند.

سپاسگزاری

ابتدا بر خود واجب می‌دانم تا از پدر و مادرم که کمال شکیبایی و تـلاش‌انـد، تقـدیر و تـشکر نمایم.

همچنین نهایت سپاسگزاری را دارم خدمت تمامی معلمان دوران تحصیلم به خصوص اسـاتید ارجمندم جناب آقای دکتر ثامتی و آقای دکتر منظوری که با صبوری فراوان مرا در انجـام ایـن پایان‌نامه راهنمایی فرمودند.

در پایان از همهٔ اعضای گروه پژوهشی بازشناسی گفتار دانشگاه صنعتی شریف سپاسگزارم.

چکیده

یکی از کاربردهای مهم سیستم‌های بازشناسی گفتار استفاده از آن‌ها در بازشناسی اعداد تلفنی می‌باشد. سیستم‌های بازشناسی گفتار در مواجه با گفتار تمیز و با کیفیت بالا توانسته‌اند به نتایج قابل قبولی دست یابند ولی هنگامی که سیگنال گفتار به وسیلهٔ شبکهٔ تلفن خراب می‌شود، دقت بازشناسی به شدت کاهش می‌یابد. روش‌های مختلفی در حوزه‌های استخراج ویژگی‌های مقاوم، مقاوم‌سازی مدل و بهبود گفتار وجود دارند که با به‌کارگیری آن‌ها خطای بازشناسی به میزان قابل توجهی کاهش می‌یابد. هدف از این پایان‌نامه، طراحی و پیاده‌سازی یک سیستم بازشناسی گفتار پیوسته برای بازشناسی اعداد از پشت خطِ تلفن می‌باشد. برای افزایش هرچه بیشتر دقت در محیط تلفنی، بسیاری از الگوریتم‌های حوزهٔ استخراج ویژگی‌های مقاوم مورد مطالعه قرار گرفتند و نحوهٔ استفاده از آن‌ها در سیستم بازشناسی موجود بررسی شدند. روش‌های CMS، PLP، RASTA، RCC، PCA و همچنین ترکیبی از آن‌ها پیاده سازی شدند و مورد آزمایش قرار گرفتند. همچنین در حوزهٔ استخراج ویژگی‌ها مقاوم روشی جدید بر اساس فیزیولوژی گوش انسان پیشنهاد و مورد آزمایش قرار گرفت. نتایج حاصل از این روش بیانگر کارایی آن در شرایط نویزی می‌باشد. از آنجایی که بهترین واحد آوایی برای آموزش یک سیستم بازشناسی گفتار پیوسته مبتنی بر HMM برای بازشناسی اعداد، کلمه (عدد) می‌باشد، در سیستم بازشناسی گفتار پایه تغییراتی صورت پذیرفت تا توانایی آموزش بر مبنای واحد آوایی کلمه را داشته باشد. همچنین به علت اینکه در زبان فارسی دادگان تلفنی معتبر زیادی برای آموزش (مبتنی بر کلمه) سیستم بازشناسی گفتار در دسترس نبود و یا دادگان در این زمینه در حد آزمایشگاهی بودند، دادگان تلفنی اعداد متصل و پیوسته زبان فارسی (FCCDigits) طراحی و ضبط گردید. هدف اصلی از تهیهٔ این دادگان فراهم کردن داده‌های آموزشی برای یک سیستم بازشناسی گفتار پیوسته مبتنی بر واحد کلمه برای بازشناسی اعداد فارسی از طریق خط تلفن است و هدف نهایی، استفاده از آن در محدودهٔ وسیعی از طرح‌های پژوهشی و تجاری می‌باشد. نتایج بدست آمده بر روی دادگان FCCDigits بیانگر دقت ۹۷ درصد برای بازشناسی اعداد متصل و ۹۵ درصد برای بازشناسی اعداد پیوسته می‌باشد.

کلید واژه‌ها: بازشناسی اعداد تلفنی (Telephony Digit Recognition)، بازشناسی گفتار پیوسته (Continuous Speech Recognition)، بازشناسی مقاوم گفتار(Robust Speech Recognition)، ویژگی‌های مقاوم (Robust Features)، مدل مخفی مارکوف (Hidden Markov Model).

فهرست مطالب

فهرست شکل‌ها و الگوریتم‌ها

فهرست جدول‌ها

فصل اول

مقدمه‌ای بر بازشناسی گفتار

۱-۱- پیش‌گفتار

برای اغلب انسان‌ها صحبت طبیعی‌ترین و کارآمدترین ابزار تبادل اطلاعات است. هدف نهایی در بازشناسی گفتار به وجود آوردن ماشین‌هایی است که بتوانند مانند انسان بشنوند، بفهمند و پاسخ مناسب نشان دهند. کاربرد چنین ماشین‌هایی می‌تواند بسیار زیاد باشد و در بسیاری از موارد، زندگی راحت‌تری را برای نوع بشر به همراه داشته باشد ولی متأسفانه با توجه به قابلیت‌ها و توانایی‌های بشر با آغاز هزارهٔ سوم میلادی دستیابی به این هدف کماکان به صورت رویا و دور از دسترس می‌باشد. اما بشر هیچگاه در مقابل رویاهای خود ساکت و بی‌تفاوت نمانده است و مانند موارد بسیاری از این قبیل ابتدا آن‌ها را در داستان‌ها و فیلم‌های تخیلی بروز داده است و همواره سعی داشته که با اعمال محدودیت‌های ساده کننده به آرزوهای خود بطور مجانبی نزدیک گردد. اگر چه در حال حاضر سیستم‌های زیادی برای بازشناسی گفتار وجود دارد ولی همگی این دستاوردها بگونه‌ای دسته‌ای از محدودیت‌های ساده کننده را یدک می‌کشند که حذف هر یک از این محدودیت‌ها می‌تواند به طور قابل ملاحظه‌ای بر پیچیدگی این سیستم‌ها بیفزاید. تنها مدل کاملاً موفق از سیستم بازشناسی گفتار، سیستم شنوایی انسان می‌باشد که هنوز اسرارآمیز و ناشناخته جلوه می‌کند. اولین سیستم تولید کنندهٔ گفتار مصنوعی در سالِ ۱۹۳۶ ساخته شد و چند سال بعد یعنی در سالِ ۱۹۴۰ اولین تلاش‌ها برای ساخت سیستم بازشناسی گفتار آغاز شد وتا اکنون این تلاش‌ها ادامه دارند.

۱-۲- تاریخچهٔ مختصری از بازشناسی گفتار

تحقیقات در زمینهٔ بازشناسی گفتار قدمتی در حدود پنج دهه دارد. اولین کوشش در این زمینه به دههٔ ۵۰ میلادی باز می‌گردد. درسال ۱۹۵۲ در آزمایشگاه Bell سیستمی جهت بازشناسی ارقام گسسته برای یک گوینده ساخته شد که بر مبنای اندازه‌گیری فرکانس‌های تشدید در حوزهٔ حروف صدادار هر رقم کار می‌کرد.

به طور مستقل در سال ۱۹۵۶ در آزمایشگاه RCA، سعی در بازشناسی ۱۰ هجای مختلف برای یک گوینده شد که این ده هجا از ۱۰ کلمهٔ تک هجایی انتخاب شده بودند. این سیستم بر مبنای اندازه‌گیری‌های طیفی، به وسیلهٔ یک بانک فیلتر آنالوگ، در حوزهٔ حروف صدادار هر کلمه کار می‌کرد. در سال ۱۹۵۹ درکالج انگلستان، Fry و Denes سعی در ساخت سیستمی برای شناسایی واج‌ها در سطح ۴ واج صدادار و ۹ واج بی‌صدا نمودند. آن‌ها از یک تحلیل‌کنندهٔ طیفی[1] و یک انطباق‌دهندهٔ الگوی طیفی استفاده نمودند. جنبهٔ جدید این تحقیق استفاده از اطلاعات آماری در مورد رشته‌های واجی مجاز زبان انگلیسی بود. آن‌ها این روش را برای بهبود دقت در بازشناسی واج‌ها برای کلمات دو یا چند واجی به کار بردند.

در سال‌های دههٔ ۶۰ در دانشگاه کیوتو سخت افزار شناسایی واج ساخته شد که در آن یک سیستم تقطیع گفتار، ورودی بخش تشخیص را فراهم می‌کرد. در تلاشی دیگر، ژاپنی‌ها سیستم سخت‌افزاری شناسایی ارقام را در آزمایشگاه NEC ساختند.

دستهٔ دیگری از پروژه‌های تحقیقاتی در اواخر دههٔ ۶۰ در آزمایشگاه RCA آغاز شد که نتیجهٔ آن، ارائهٔ مجموعه‌ای از روش‌های نرمال‌سازی زمانی[2] ابتدایی در ارتباط با غیر یکنواختی مقیاس در گفتار بود. این روش‌ها برای تشخیص مطمئن محل ابتدا و انتهای گفتار به کار می‌رفت. در همان زمان‌ها در شوروی سابق، روشی بر مبنای برنامه‌سازی پویا[3] ارائه گردید که بعدها انطباق زمانی پویا نامیده شد.

آخرین دستاوردهای قابل ذکر دههٔ ۶۰، تحقیقات Reddy در زمینهٔ بازشناسی گفتار پیوسته در تعقیب دینامیک واج‌ها بود. تحقیقات وی بعدها به تحقیقات طولانی و دنباله‌داری در دانشگاه CMU تبدیل گشت به طوری که تا امروز این دانشگاه، پیشرو ارائهٔ سیستم‌های بازشناسی گفتار پیوسته است.

---

[1] *Spectral Analyzer*
[2] *Time Normalization*
[3] *Dynamic Programming*

در دهۀ ۱۹۷۰، Itakura ایدۀ LPC[1]، که در ساخت رمزکننده‌های با نرخ بیت پایین موفق بود را به حوزه بازشناسی گفتار کشاند و معیار فاصلۀ مناسبی نیز برای آن ابداع نمود. دیگر رویداد مهم این دهه، آغاز تحقیقات موفق و دنباله‌دار IBM در زمینۀ بازشناسی گفتار با مجموعۀ کلمات بزرگ بود. در آزمایشگاه Bell نیز محققان آزمایش‌هایی را بر روی مستقل از گوینده کردن بازشناسی گفتار آغاز کردند.

همان طور که بازشناسی کلمات گسسته حوزۀ تمرکز تحقیقات در دهۀ ۷۰ بود، در دهۀ ۸۰ شناسایی کلمات متصل محققان زیادی را به خود مشغول داشت. هدف، ایجاد سیستم‌هایی برای تشخیص رشته‌ای از کلمات بود که به صورت مؤکد، با کمی مکث ادا می‌شوند. از جملۀ این سیستم‌ها DP بود که توسط Sakoe در NEC پیاده‌سازی گردید.

تحقیقات در دهۀ ۸۰ با انتقال از روش‌های تطبیق الگو به مدل‌های آماری به خصوص HMM همراه بود. اگر چه روش HMM قبلاً نیز شناخته شده بود ولی تا اواسط دهۀ ۸۰ که تئوری و روش‌های آن منتشر گشت، به کار گرفته نشد. تکنولوژی جدید دیگری که در اواخر این دهه وارد زمینۀ بازشناسی گفتار شد، شبکه‌های عصبی مصنوعی بود. شبکه‌های عصبی اولین بار در دهۀ ۵۰ معرفی شدند ولی به دلیل مشکلات عملی فراوان، زیاد مورد توجه قرار نگرفتند. اما در دهۀ ۸۰ فهم عمیق‌تر از توانایی‌ها و محدودیت‌های این تکنولوژی در رابطه با مسألۀ بازشناسی به دست آمد. همچنین در این دهه ایدۀ مدل‌سازی گفتار به صورت «وابسته به بافت»[2] با توجه به دوواجی‌ها و سه‌واجی‌ها مورد توجه قرار گرفت که در بازشناسی گفتار پیوسته منجر به نتایج خوبی گردید.

در دهۀ ۸۰ گام بزرگی در زمینۀ بازشناسی گفتار پیوسته با اندازۀ واژگان بزرگ توسط DARPA[3]برداشته شد و سیستمی با دقت بازشناسی بالا با مجموعه واژگان ۱۰۰۰کلمه‌ای پیاده‌سازی گردید.

---

[1] *Linear Predictive Coding*
[2] *Context Dependent*
[3] *Defense Advanced Research Project Agency*

سهم بزرگی از این دستاورد مرهون تلاش‌های محققان در آزمایشگاه‌های MIT Lincoln و CMU بود. در سال ۱۹۸۸ در دانشگاه CMU سیستم پایه‌ای SPHINX ساخته شد، که اولین سیستم نسبتاً موفق بازشناسی گفتار پیوسته مستقل از گوینده با مجموعه واژگان بزرگ بود که بهبود کارایی آن همچنان ادامه دارد.

در اوایل دهۀ ۹۰ روش‌های مختلفی برای ترکیب شبکه‌های عصبی و HMM برای غلبه بر نقاط ضعف آن‌ها مطرح گردید و همچنین روش‌های درج مدل زبانی در آن‌ها پیشرفت چشمگیری داشت. در این دهه مسألۀ بازشناسی گفتار در محیط‌های نویزی بیشتر مورد توجه قرار گرفت و روش‌هایی برای مقاوم‌سازی سیستم‌های بازشناسی در برابر نویز ارائه گردید.

اکثر تحقیقات در زمینۀ بازشناسی گفتار تا این دهه بر روی زبان انگلیسی بود. در دهۀ ۹۰ مسألۀ بازشناسی گفتار، بیشتر مورد توجه محققان غیرانگلیسی زبان قرار گرفت و به خصوص در اواخر این دهه تحقیقات جدی در زمینۀ بازشناسی گفتار پیوستۀ فارسی آغاز شد.

در دهۀ اخیر زمینه‌های تحقیق جدیدی بر روی محققان گشوده شده است از مهمترین این زمینه‌ها، بازشناسی گفتار محاوره‌ای می‌باشد که هدف آن بازشناسی گفتاری است که فی‌البداهه توسط گوینده بیان می‌شود و در آن کلمات معمولاً به طور ناقص و خطادار ادا می‌شوند. زمینۀ تحقیقی جدید دیگر، بازشناسی گفتار به صورت «شنودی_دیداری»[1] می‌باشد که در آن شکل و نحوۀ حرکت لب‌ها و دهان نیز در بازشناسی (به خصوص در حضور نویز) مورد توجه قرار می‌گیرد. همچنین در این دهه، مدل‌سازی وابسته به بافت با استفاده از اطلاعات آکوستیک وابسته به بافت گفتار مورد توجه قرار گرفته است.

[1] *Audio-Visual*

۱-۳- علوم مرتبط با بازشناسی گفتار

طبیعت چندرشته‌ای[1] بودن گفتار یکی از مشکلات عمده در بازشناسی گفتار می‌باشد. بنابراین تحقیق در زمینهٔ بازشناسی گفتار با علوم مختلف دیگری در ارتباط می‌باشد و کسی که می‌خواهد در این زمینه به تحقیق بپردازد می‌بایست تا حدودی با این علوم آشنایی داشته باشد.

- پردازش سیگنال: جهت استخراج ویژگی‌ها و الگوهای مناسب از سیگنال گفتار و یا آنالیز آن، آشنایی با پردازش سیگنال الزامی است.
- آواشناسی: ارتباط بین سیگنال گفتار و اندام گویایی انسان و همچنین چگونگی مکانیزم شنوایی انسان در این علم مورد بررسی قرار می‌گیرد.
- بازشناسی الگو: جهت پیاده‌سازی مجوعه‌ای از الگوریتم‌های لازم برای خوشه‌بندی[2] داده‌های آموزشی به عنوان الگوهای مرجع و مقایسهٔ الگوی داده‌های آزمایشی با این الگوهای مرجع به منظور پیدا کردن بهترین الگوی قابل انطباق، به این علم احتیاج است.
- تئوری اطلاعات و ارتباطات: ارتباط این مقوله از علم با بازشناسی گفتار در تخمین پارامترهای مدل‌های آماری، روش‌های مدرن کد و کدگشایی و محاسبه آنتروپی یک مدل زبانی می‌تواند باشد.
- زبان شناسی: صداشناسی[3] یا ارتباط بین صداها، دستور زبان[4] یا نحوهٔ شکل گرفتن جملات از کلمات زبان، معنا[5] یا نحوهٔ ترکیب کلمات برای شکل دادن پیام‌های با معنی و جنبهٔ عملی[6] یا مفهوم به دست آمده از معنی گفتار، ارتباط تنگاتنگی با زبان شناسی دارند.

---

[1] *Multidisciplinary*
[2] *Clustering*
[3] *Phonology*
[4] *Syntax*
[5] *Semantics*
[6] *Pragmatics*

- فیزیولوژی: این علم برای بیان و درک نحوهٔ عملکرد سیستم عصبی انسان در هنگام بیان و شنود گفتار به کار می‌آید.
- علوم کامپیوتر: جهت پیاده سازی کارآمد الگوریتم‌های مورد نیاز در سیستم بازشناسی گفتار، به علوم کامپیوتر نیاز است.
- روانشناسی: جهت فهم عامل‌های روانی که انسان در بیان، شنود و درک گفتار بکار می‌گیرد.

۱-۴- پیچیدگی مسألهٔ بازشناسی گفتار

پیچیدگی مسألهٔ بازشناسی گفتار در زمینه‌های مختلفی دسته‌بندی می‌شود و متعاقب هر یک از آن‌ها محدودیت‌هایی بر سیستم‌های بازشناسی گفتار اعمال می‌گردد. بیان این نکته ضروری است که حذف هر یک از این محدودیت‌ها می‌تواند به طور قابل ملاحظه‌ای بر پیچیدگی‌های این سیستم‌ها بیفزاید.

۱-۴-۱- میزان وابستگی یا استقلال از گوینده

هر اندازه تعداد گویندگانی که سیستم قرار است به آن‌ها پاسخ دهد زیادتر گردد بر پیچیدگی آن افزوده می‌شود. ویژگی‌های سیگنال گفتاری گویندگان مختلف با توجه به عواملی مانند سن، جنس، لهجه، لحن و سرعت ادای گفتار تغییر می‌کند. سیستم‌هایی که تنها به یک و یا چند گویندهٔ خاص پاسخ می‌گویند، وابسته به گوینده[1] و آن‌هایی که به تمام گویندگان یک زبان پاسخ می‌گویند مستقل از گوینده[2] نام دارند. اگر چه دقت بازشناسی در سیستم‌های وابسته به گوینده معمولاً بالاتر از سیستم‌های مستقل از گوینده است ولی

[1] *Speaker Dependent*
[2] *Speaker Independent*

لزوم آموزش مجدد این سیستم‌ها برای این که توسط گویندگان جدید مورد استفاده قرار گیرند، نقص بزرگی برای آن‌ها به حساب می‌آید.

۱-۴-۲- اندازهٔ مجموعهٔ واژگان[1]

حجم کلمات و عباراتی که سیستم باید تشخیص دهد نیز بر پیچیدگی مسألهٔ بازشناسی گفتار تأثیر دارد. تحقیقات نشان داده است که پیچیدگی یک سیستم بازشناسی با افزایش تعداد کلمات بصورت لگاریتمی زیاد می‌شود. سیستم‌های بازشناسی گفتار از لحاظ تعداد کلمات مورد بازشناسی به سه دستهٔ سیستم‌های بازشناسی با مجموعه واژگان کوچک[2] (زیر ۱۰۰ کلمه)، متوسط[3] (بین ۱۰۰ تا ۱۰۰۰ کلمه) و بزرگ[4] (بیش از ۱۰۰۰ کلمه) تقسیم می‌شوند.

۱-۴-۳- فی‌البداهه، پیوسته، متصل یا گسسته بودن گفتار مورد بازشناسی

شکل گفتار ورودی سیستم نیز می‌تواند بر میزان پیچیدگی بازشناسی گفتار تأثیر بگذارد. گوینده می‌تواند بین هر دو کلمهٔ متوالی کمی مکث کند، کلمات را به هم پیوسته ولی با تأکید ادا نماید، بطور پیوسته و طبیعی صحبت نماید و یا گفتار را بطور طبیعی به شکلی فوری و مکالمه‌ای بیان کند. به این ترتیب سیستم‌های بازشناسی از لحاظ گفتار ورودی به چهار دسته تقسیم می‌شوند:

[1] *Vocabulary Size*
[2] *Small Vocabulary*
[3] *Medium Vocabulary*
[4] *Large Vocabulary*

- سیستم بازشناسی کلمات مجزا[1]: در این سیستم‌ها، در مرحله‌ی بازشناسی، گوینده هنگام بیان جمله، کلمات را با فاصله از یکدیگر بیان می‌کند. مکث بین هر دو کلمه متوالی سبب می‌شود که مرز بین کلمات به راحتی شناسایی شود.
- سیستم‌های بازشناسی گفتار متصل[2]: در این سیستم‌ها، گوینده هر کلمه را با تأکید کافی بیان می‌کند و به دلیل وضوح الگوهای استرس، شناسایی مرزها آسانتر می‌شود.
- سیستم‌های بازشناسی گفتار پیوسته[3]: در این گونه سیستم‌ها، گوینده جملات را به طور طبیعی بیان می‌کند. از مهمترین مشکلاتی که در این نوع سیستم‌ها وجود دارد، می‌توان مشخص نبودن مرز بین کلمات متوالی و ادغام شدن ابتدای یک کلمه با انتهای کلمهٔ ما قبل خود را نام برد.
- سیستم‌های بازشناسی گفتار فی‌البداهه[4]: در این گونه سیستم‌ها، گوینده گفتاری با جملاتی ناقص، شروع‌های مجدد، خنده و سرفه بیان می‌کند که گفتار را از حالت سلیس بودن خارج می‌کند و اندازهٔ مجموعهٔ واژگان را عملاً نامحدود می‌کند.

۱-۴-۴- میزان محدودیت‌های زبانی[5]

در بازشناسی گفتار معمولاً محدودیت‌هایی در ارتباط با مدل زبانی ساخت جملات و عبارات از اجزاء کوچکتر مثل کلمه و آوا نیز اعمال می‌شود. محدودیت‌های زبانی را می‌توان با یک مدل از زبان بیان نمود. مدل زبانی یک زبان طبیعی شامل چهار جزء سمبل، دستور، معنا و جنبهٔ عملی می‌باشد. سمبل‌های زبان واحدهای طبیعی هستند که همهٔ پیغام‌ها از آن‌ها تشکیل شده‌اند که می‌توانند کلمه، هجا یا واج باشند. دستور زبان مرکب از محدودیت‌های واژگانی و نحوی است که بیانگر شکل گرفتن کلمات از واحدهای

[1] *Isolated Word Recognition*
[2] *Connected Speech Recognition*
[3] *Continuous Speech Recognition*
[4] *Spontaneous Speech Recognition*
[5] *Linguistic Constraint*

کوچکتر از کلمه و نیز شکل گرفتن جملات از کلمات می‌باشد. جنبهٔ معنایی نیز مرتبط به نحوهٔ ترکیب کلمات برای شکل دادن پیغام‌های با معنی است. در بالاترین سطح جنبهٔ عملی یک زبان جای دارد که بیانگر وابستگی ادا کردن و معنی کلمه به گوینده‌ها و محیط است. محدودیت‌های معنایی و جنبهٔ عملی به ندرت در سیستم‌های بازشناسی گفتار امروزی مورد استفاده قرار می‌گیرند ولی محدودیت‌های دستوری تقریباً در تمامی سیستم‌های بازشناسی گفتار پیوسته مورد استفاده قرار می‌گیرند. استفاده از مدل زبانی و نحوی در سیستم‌های بازشناسی گفتار می‌تواند به طور مؤثری دقت آن‌ها را افزایش دهد ولی این محدودیت‌ها به خصوص محدودیت‌های نحوی با توجه به زبان به دشواری قابل استفاده در سیستم‌های امروزی هستند.

۱-۴-۵- میزان تأثیر محیط بر سیستم بازشناسی گفتار

محیطی که سیستم بازشناسی در آن عمل می‌کند، بر کارایی بازشناسی مؤثر است. مسألهٔ بازشناسی گفتار با وجود اثرات مخربی چون نویز، پژواک، تداخل و اعوجاج که معمولاً از محیط، میکروفن و یا کانال انتقال ناشی می‌شوند بسیار متفاوت با مسألهٔ بازشناسی گفتار در محیطی آرام و بدون تخریب است. هرگاه گفتار مورد بازشناسی، با نویز تخریب گردد کارایی سیستم بازشناسی به شدت کاهش می‌یابد. در واقع یکی از دلایل مهم کاهش کارایی سیستم‌های بازشناسی در شرایط نویزی عدم تطابق بین محیط آموزش و آزمایش می‌باشد.

۱-۵- کاربردهای بازشناسی گفتار

بازشناسی گفتار می‌تواند کاربردهای متفاوتی داشته باشد [۱]. از جمله کاربردهای آن می‌توان به استفاده از آن در سیستم‌های تجاری و اداری جهت دستیابی به بانک‌های اطلاعاتی، ورود داده و کنترل و

مدیریت آن‌ها، دیکتهٔ خودکار، ترجمهٔ گفتار، کمک به معلولین جسمی اشاره کرد. کاربرد بازشناسی گفتار فقط به موارد بالا محدود نمی‌شود. از این سیستم‌ها می‌توان جهت کمک به ناشنوایان نیز بهره جست. با حل مسألهٔ بازشناسی گفتار و پیاده‌سازی سخت‌افزاری آن بصورت قابل‌حمل، فرد ناشنوا می‌تواند صورت متنی پیام را بر روی صفحهٔ نمایش ببیند که کمک در خور توجهی به این افراد خواهد کرد. کاربرد آن در مخابرات نیز می‌تواند زندگی بشر را آسان‌تر از گذشته کند. بعنوان مثال، در ساخت تلفن‌هایی با شماره گیر گفتاری و یا ساخت مراکز اطلاعات تلفن با اپراتور کامپیوتری از جمله کاربردهای آن در مخابرات است. سیستم‌های بازشناسی گفتار در صنعت نیز می‌توانند کاربرد داشته باشند. متداول‌ترین آنها، پیاده‌سازی سیستم‌های به اصطلاح «چشم آزاد، دست آزاد» در کارخانه‌هاست. در موارد دیگر از جمله ساخت اسباب‌بازی‌ها، کنترل بخش‌هایی از وسایل نقلیه نیز سیستم‌های بازشناسی گفتار می‌توانند به انسان‌ها خدمت کنند.

در اروپا سیستم‌های SPICOS و تعمیم‌های آن از شرکت فیلیپس در زمینهٔ دیکتهٔ اتوماتیک، سیستم‌های CSELT برای اطلاعات راه آهن اروپایی، سیستم‌های دانشگاه کمبریج انگلستان مثل ABBOT و LIMSI نمونه‌هایی موفق از نتایج تحقیقات در زمینهٔ بازشناسی گفتار می‌باشد. در ژاپن، سیستم‌های بازشناسی گفتار با تعداد کلمات زیاد برای کاربردهای تلفنی پیاده‌سازی شده است. در چین و تایوان سیستم‌های بازشناسی بر مبنای تشخیص هجا برای کاربردهای دیکتهٔ اتوماتیک با تعداد کلمات بسیار زیاد طراحی گردیده و در حال پیاده‌سازی است. در کانادا، تحقیقات قابل توجهی روی سیستم INRS برای بازشناسی ۸۶۰۰۰ کلمهٔ منفرد به انجام رسیده است. در ایالات متحده پروژه‌های گوناگونی در AT&T، IBM، CMU، MIT، SRI و دیگر مراکز تحقیقاتی، به نتایجی قابل توجهی دست یافته‌اند

۱-۶- هدف از این پایان‌نامه

هدف از این پایان‌نامه، طراحی و پیاده‌سازی یک سیستم بازشناسی گفتار پیوسته برای بازشناسی اعداد از پشت خطِ تلفن می‌باشد. در فصل ۲، مدل مخفی مارکوف به عنوان هستهٔ اصلی در سیستم بازشناسی اعداد معرفی شده‌است. فصل ۳، اجزای یک سیستم بازشناسی گفتار پیوسته را بیان کرده است. نویز و نحوهٔ تأثیر آن بر سیستم‌های بازشناسی گفتار و الگوریتم‌های مقابله با این مشکل، در فصل ۴ مورد بررسی قرار گرفته‌اند. در فصل۵، ویژگی‌های مقاوم به تفصیل بررسی می‌شوند. همچنین روشی جدید برای استخراج ویژگی‌ها در این فصل معرفی می‌شود. در فصل ۶ به نحوهٔ طراحی و تهیهٔ دادگان متصل و پیوسته زبان فارسی پرداخته می‌شود. همچنین سیستم بازشناسی اعداد تلفنی مبتنی بر کلمه و چگونگی آموزش آن به تفصیل در این فصل معرفی خواهد شد. در فصل۷، بستر مورد آزمایش و نتایج اولیه و نهایی گزارش شده است. فصل ۸ نیز به جمع‌بندی و ارائهٔ پیشنهادهایی برای ادامهٔ کار اختصاص دارد.

فصل دوم

مدل مخفی مارکوف

۲-۱- مقدمه

مدل مخفی مارکوف یک روش بسیار قدرتمند برای مدل‌سازی یک سری مشاهدات گسسته در زمان می‌باشد. این مدل نه تنها یک روش مؤثر برای ساخت بهینه و پارامتری مدل‌ها مهیا می‌کند بلکه می‌تواند از تکنیک‌های برنامه سازی پویا[1] نیز بهره‌مند گردد. نمونهٔ داده‌ها می‌توانند به صورت پیوسته یا گسسته در زمان پراکنده شده باشند و یا می‌توانند اسکالر و یا برداری باشند. در مدل مخفی مارکوف نمونهٔ داده‌ها به شکل فرآیندهای تصادفی پارامتری توصیف می‌شوند و این فرآیندهای تصادفی را می‌توان از روی یک چارچوب دقیق و تعریف شده تخمین زد. اصل تئوری مدل مخفی مارکوف به وسیلهٔ آقای Baum و همکارانش در یک سری از مقالات کلاسیک پایه‌ریزی شد[۲]. مدل مخفی مارکوف اکنون به عنوان یکی از روش‌های قدرتمند آماری برای مدل کردن سیگنال گفتار در آمده است. این مدل به صورت موفقیت آمیزی در رشته‌های بازشناسی خودکار گفتار، بهبود گفتار، مدل‌های زبانی آماری[2]، سنتز گفتار، فهم زبان، برچسب دهی ادات سخن[3] و ترجمهٔ ماشینی مورد استفاده قرار گرفته است[۱].

۲-۲- زنجیرهٔ مارکوف

زنجیرهٔ مارکوف با حداقل مقدار حافظه یک دسته از فرآیندهای تصادفی را مدل می‌کند. در این بخش زنجیرهٔ مارکوف گسسته مورد بررسی قرار می‌گیرد.

اگر $X = X_1, X_2, \cdots, X_n$ یک رشته از متغیرهای تصادفی از الفبای گسستهٔ متناهی $O = \{o_1, o_2, \cdots, o_M\}$ باشند آنگاه بر اساسِ قانون بیز خواهیم داشت:

[1] *Dynamic Programming*
[2] *Statistical Language Modeling*
[3] *Part-of-speech Tagging*

$$P(X_1, X_2, \cdots, X_n) = P(X_1)\prod_{i=2}^{n} P(X_i \mid X_1^{i-1}) \qquad (۱-۲)$$

که در رابطهٔ فوق $X_1^{i-1} = X_1, X_2, \cdots, X_{i-1}$ می‌باشد. متغیر تصادفی X زنجیرهٔ مـارکوف درجـهٔ اول نامیده می‌شود اگر:

$$P(X_i \mid X_1^{i-1}) = P(X_i \mid X_{i-1}) \qquad (۲-۲)$$

و بنابراین رابطهٔ (۲-۱) به صورت زیر بازنویسی می‌گردد:

$$P(X_1, X_2, \cdots, X_n) = P(X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1}) \qquad (۳-۲)$$

رابطهٔ (۲-۲) به عنوان فرض مارکوف شناخته می‌شود. این فرض مقدار خیلی کمی حافظه برای مـدل کردنِ رشته داده‌های پویا به کار می‌برد: احتمال هر متغیر تصادفی در یک زمان خاص فقط به مقدار متغیر در یک زمان قبل وابسته است. اگر زمان را در نظر نگیریم آنگاه زنجیرهٔ مارکوف مـی‌توانـد بـرای مـدل‌سـازی رویدادهای تغییر ناپذیر با زمان (ایستان[1]) استفاده گردد، یعنی:

$$P(X_i = s \mid X_{i-1} = s') = P(s \mid s') \qquad (۴-۲)$$

اگر X_i را به یک حالت مرتبط کنیم آنگاه زنجیرهٔ مارکوف را می‌توان با یـک ماشـن حالـت متنـاهی نمایش داد که گذر بین حالت‌های آن با تابع احتمال $P(s \mid s')$ مشخص می‌گردد. اگر نمایش ماشین حالـت متناهی را در نظر بگیریم، آنگاه فرض مارکوف موجود در رابطهٔ (۲-۲) به این صورت بیان می‌گردد: احتمال اینکه زنجیرهٔ مارکوف در حالت مشخصی در یک زمان خاص قرار بگیرد فقط بـه حالـت زنجیـرهٔ مـارکوف یک زمان قبل آن وابسته است.

یک زنجیرهٔ مارکوف با N حالت مجزای $\{1, \cdots, N\}$ را در نظر بگیرید، اگر حالتِ s در زمـان t بـا s_t نمایش داده شود آنگاه پارامترهای زنجیرهٔ مارکوف می‌توانند به صورت زیر تعریف گردند:

[1] *Stationary*

$$a_{ij} = P(s_t = j \mid s_{t-1} = i) \qquad 1 \le i, j \le N \tag{۵-۲}$$

$$\pi_i = P(s_1 = i) \qquad 1 \le i \le N \tag{۶-۲}$$

که a_{ij} احتمال گذر از حالت i به حالت j و π_i احتمال اولیه‌ای است که زنجیرهٔ مارکوف در حالت iام شروع شود. هر دو احتمال‌های گذر و اولیه دارای محدودیت‌های زیر می‌باشند:

$$\sum_{j=1}^{N} a_{ij} = 1; \qquad 1 \le i \le N \tag{۷-۲}$$

$$\sum_{j=1}^{N} \pi_i = 1 \tag{۸-۲}$$

زنجیرهٔ مارکوف توصیف شده در بالا مدل مارکوف قابل مشاهده نیز نامیده می‌شود زیرا خروجی هر فرآیند مجموعه‌ای از حالت‌ها در هر زمان t می‌باشند که هر حالت به یک رویداد قابل مشاهدهٔ X_i مربوط است. در واقع یک رابطهٔ یک به یک بین رشته رویدادهای قابل مشاهدهٔ X و رشته حالت‌های زنجیرهٔ مارکوف $S = s_1, s_2, \cdots, s_n$ وجود دارد.

۲-۳- مدل مخفی مارکوف

در زنجیرهٔ مارکوف، هر حالت به صورت مشخص به یک حالت قابل مشاهده مربوط می‌شود یعنی اینکه خروجی هر حالت به صورت تصادفی نمی‌باشد. یک توسعهٔ طبیعی برای زنجیرهٔ مارکوف می‌تواند فرآیند غیر مشخص برای هر یک از حالت‌های آن باشد به طوری که هر حالت آن یک دنباله از نمادهای قابل مشاهده را تولید می‌کند. بنابراین مشاهده در این مدل یک تابع تصادفی از حالت می‌باشد. این مدل جدید به نام مدل مخفی مارکوف نامیده می‌شود و می‌توان به آن به صورت یک فرآیند تصادفی دو گانه که در بر دارندهٔ یک فرآیند تصادفی (رشتهٔ حالت‌ها) غیر قابل مشاهده است، نگاه کرد. این فرآیند می‌تواند به

صورت احتمالی با فرآیندهای تصادفی قابل مشاهده مرتبط شود به طوری که رشته‌ای از ویژگی‌های قابل مشاهده را تولید کند.

مدل مخفی مارکف در واقع یک زنجیرهٔ مارکف است که مشاهدهٔ خروجی آن یک متغیر تصادفی مانند X است که با توجه به تابع احتمالی خروجی مربوط به هر حالت، تولید شده‌است. این بدان معناست که دیگر یک رابطهٔ یک به یک بین رشتهٔ مشاهدات و رشته حالت‌ها جود ندارد بنابراین نمی‌توان به طور قطع رشتهٔ حالت‌ها را برای یک رشتهٔ مشاهدات داده شده، مشخص کرد، یعنی اینکه رشته حالت‌ها قابل مشاهده نیستند و بنابراین مخفی می‌باشند. این دلیل قرار گرفتن "مخفی" قبل از "مدل مارکف" می‌باشد. با وجود این که حالت‌های HMM مخفی می‌باشند ولی معمولاً در بر دارندهٔ اطلاعات برجسته‌ای دربارهٔ دادهایی که مدل می‌شوند، باشد. با توجه به توضیحات فوق، مدل مخفی مارکف به وسیلهٔ پارامترهای زیر تعریف می‌شود:

- $O = \{o_1, o_2, \ldots, o_M\}$: الفبای مشاهدات خروجی. سمبل‌های مشاهدات خروجی فیزیکی سیستمی است که مدل می‌گردد.
- $\Omega = \{1,2,\ldots,N\}$: مجموعهٔ حالات که مشخص کنندهٔ فضای حالت می‌باشند. در این جا S_t بیان کنندهٔ حالت در زمان t می‌باشند.
- $A\{a_{ij}\}$: ماتریس احتمالاتی گذر بین حالت ها که a_{ij} احتمال رفتن از حالت iام به jام را بیان می‌کند. یعنی: $a_{ij} = P(s_t = j \mid s_{t-1} = i)$
- $B = \{b_i(k)\}$: ماتریس احتمالاتی خروجی است بطوری که $b_i(k)$ احتمال رخداد سمبل O_k را هنگامی که در حالت jام است نشان می دهد. اگر $X = X_1, X_2, \ldots, X_t, \ldots$ خروجی مشاهده شده مدل مخفی مارکف باشد، آنگاه رشته حالت‌های $S = s_1, s_2, \ldots, s_t, \ldots$ قابل

مشاهده نخواهند بود (مخفی) و $b_i(k)$ را می توان به صورت $b_i(k) = P(X_t = O_k \mid S_t = i)$ نوشت.

- $\Pi = \{\pi_i\}$: توزیع حالت اولیه است به طوری‌که $1 \le i \le N$ $\pi_i = P(s_o = i)$
- از آنجایی که a_{ij}، $b_{ij}(k)$ و π_i همگی توابع احتمالاتی می‌باشند، بنابراین باید خواص زیر را ارضا کنند:

$$a_{ij} \ge 0 \,,\, b_i(k) \ge 0 \,,\, \pi_i \ge 0 \qquad \forall all\, i, j, k \tag{۹-۲}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{۱۰-۲}$$

$$\sum_{k=1}^{M} b_i(k) = 1 \tag{۱۱-۲}$$

$$\sum_{i=1}^{N} \pi_i = 1 \tag{۱۲-۲}$$

در مجموع یک تعریف کامل مدل مخفی مارکوف در بردارندهٔ دو پارامتر با اندازهٔ ثابت M و N رشته مجموعهٔ ماتریس احتمالات A و B و π می باشند. M و N بیان‌کنندهٔ تعداد کل حالت‌ها و اندازهٔ الفبای مشاهده می‌باشند. بنابراین یک مدل مخفی مارکف به صورت $\Phi = (A, B, \pi)$ بیان می شد که مشخص کنندهٔ مجموعه پارامترهای مدل مخفی مارکف می‌باشد. در مدل مخفی مارکف درجهٔ اول بیان شده در بالا دو فرض وجود داد. اولین آن‌ها فرض مارکوف برای رنجیرهٔ مارکف می‌باشد.

$$P(s_t \mid s_1^{t-1}) = P(s_t \mid s_{t-1}) \tag{۱۳-۲}$$

که s_1^{t-1} بیان کنندهٔ رشته حالت‌های $S_1, S_2, \ldots, S_{t-1}$ می‌باشد. دومین فرض غیر وابسته بودن خروجی می‌باشد.

$$P(X_t \mid X_1^{t-1}, s_1^t) = P(X_t \mid s_t) \tag{۱۴-۲}$$

که X_1^{t-1} بیان کنندهٔ رشتهٔ خروجی $X_1, X_2, \ldots, X_{t-1}$ می باشد. فرض غیروابسته بودن خروجی بیان می‌دارد که احتمال اینکه یک سمبل خاص در زمان t مشاهده شود فقط بـه حالـت s_t بـستگی دارد و از مـشاهدات قبلی مستقل است.

ممکن است تصور شود که این فرض‌ها حافظهٔ مدل مخفی مارکف درجهٔ اول را محـدود مـی‌کنـد و باعث نا کارایی این مدل می‌گردد. در عمل این فرض‌ها ارزیابی، رمزگشایی و یادگیری مدل را بدون این که تاثیر قابل توجهی بر روی توانایی های مدل بگذارند، ممکن و کارا می کند و همچنین این فـرض‌هـا تعـداد زیاد پارامترهایی که باید تخمین زده شوند را کم می کند.

با تعریف فوق برای مدل مخفی مارکوف، قبل از این که بتوان از آن در کـاربرد هـای عملـی اسـتفاده کرد، سه مسألهٔ اساسی باید پاسخ داده شوند.

۱- مسألهٔ ازریابی – اگر مدل Φ و رشته‌ای از مشاهدات $X = (X_1, X_2, \ldots, X_T)$ ، داده شده باشند، آنگاه احتمال $P(X \mid \Phi)$، یعنی احتمال مدل که مشاهدات را تولید کند چقدر است؟

۲-مسألهٔ رمز گشایی – اگر مدل Φ و رشـته‌ای از مـشاهدات $X=(X_1, X_2, .., X_T)$ داده شـده باشـند، آنگاه احتمال محتمل‌ترین رشته حالت های $S = (s_0, s_1, \ldots, s_T)$ در مدل که این مـشاهدات را تولیـد کنـد، چیست؟

۳-مسألهٔ آموزش – اگر مدل Φ و یک مجموعه از مشاهدات داده شده باشند، به چه شکل مـی‌تـوان پارامترهای مدل Φ را تنظیم کرد که احتمال مشترک یا شباهت[1] $\prod_X P(X \mid \Phi)$ ماکزیمم گردد؟

اگر بتوانیم مسألهٔ ارزیابی را حل کنیم آنگاه خواهیم توانست که راه حلی برای این که تـا چـه انـدازه مدل مخفی مارکف با رشته‌ای از مشاهدات منطبق است را پیدا کنیم. بنابراین می‌توانیم مدل مخفی مارکف را برای بازشناسی الگو استفاده کنیم زیرا که شباهت $P(X \mid \Phi)$ می تواند برای محاسبهٔ احتمال $P(\Phi \mid X)$ بـه

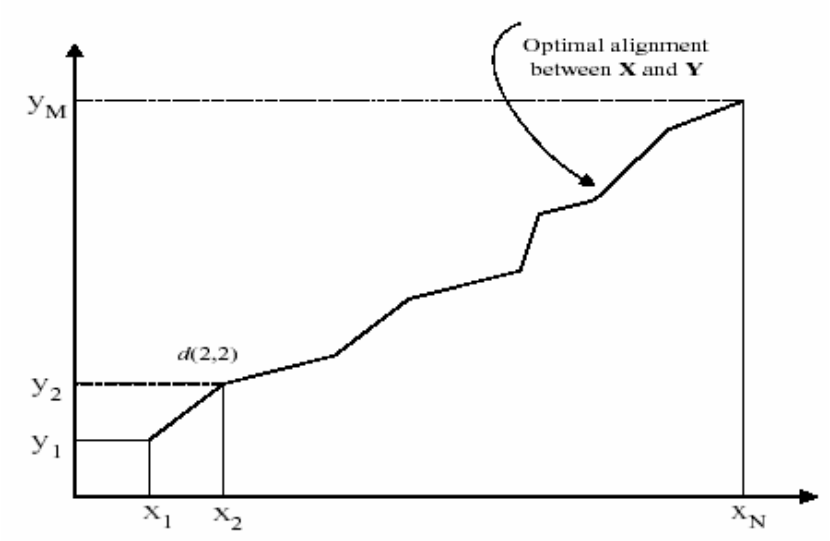
[1] *Likelihood*

کار رود و مدل مخفی مارکفی با بیشترین احتمال برای الگوی مورد نظر رشته مشاهدات را مشخص کرد. اگر مسألهٔ دوم را بتوانیم حل کنیم آنگاه می‌توانیم بهترین رشته حالت‌های منطبق شونده را برای رشتهٔ مشاهدات پیدا کنیم و یا به تعبیری دیگر می‌توانیم بخش پنهان مدل مخفی مارکف را شناسایی کنیم. و اگر بتوانیم مسألهٔ یادگیری را حل کنیم می‌توانیم به صورت خودکار پارامترهای مدل Φ را از روی داده‌های آموزشی تخمین بزنیم. برای پیاده سازی راه حل‌های کارا برای هر یک از مسائل فوق نیاز به استفاده از تکنیک‌های برنامه سازی پویا می‌باشد که در ادامه به صورت اجمالی مورد بحث قرار خواهند گرفت.

۲-۳-۱- برنامه‌سازی پویا و DTW

برنامه‌سازی پویا که همچنین به عنوان DTW در بازشناسی گفتار شناخته می‌شود[۳] به صورت وسیعی برای پیدا کردن شباهت دو الگوی گفتار به کار می‌رود. در این سیستم مبتنی بر الگو، هر الگوی گفتار از برداری از رشته‌های گفتار تشکیل شده است.

برای هر جفت (i, j) فاصلهٔ $d(i, j)$ بین دو بردار گفتار x_i و y_j وجود دارد. برای پیدا کردن بهترین مسیر بین نقطهٔ آغازین (i, j)، نقطهٔ انتهایی (N, M) از چپ به راست نیاز به محاسبهٔ فاصلهٔ بهینهٔ $D(N, M)$ می‌باشد. می‌توان تمامی مسیرهای بین $(1,1)$ تا (N, M) را بررسی کرد و کوچکترین فاصله را تشخیص داد.



شکل ۲-۱- مقایسه دو الگوی گفتار X و Y

همانطور که در شکل ۲-۱ مشخص است M حرکت برای هر قدم از چپ بـه راسـت وجـود دارد و تمام مسیرهای ممکن از $(1,1)$ تا (N,M) به صورت نمایی خواهد بود. اصول برنامه‌سـازی پویـا بـه طـور چشمگیری مقدار محاسبات را کم کند با کنار گذاشتن رشته‌هایی کـه نمـی‌تواننـد اصـولاً بهینـه باشـند و از آنجایی که مسیر بهینه در هر مرحله باید بر این مسیر در مرحلهٔ قبلـی مبتنـی باشـد بنـابراین فاصـلهٔ مینـیمم $D(i,j)$ به صورت رابطهٔ (۲-۱۵) معرفی می‌شود:

$$D(\mathrm{i},\mathrm{j}) = \min_{\mathrm{k}}\left[\mathrm{D}(\mathrm{i}-1,\mathrm{k}) + \mathrm{d}(\mathrm{k},\mathrm{j})\right] \qquad (۱۵-۲)$$

رابطهٔ فوق نشان می‌دهد که فقط کافی است که بهترین حرکت را برای هر جفت نگـهداری کنـیم در حالی که تعداد M حرکت وجود دارد. الگوریتم بازگشتی این امکان را مـی‌دهـد کـه مـسیر بهینـه جـستجو ازسمت چپ به راست شکل گیرد.

بنابراین حل زیر مسأله‌های $D(i-1,\mathrm{k})$ منجر به حل مسألهٔ کلی $D(i,j)$ می‌گردد و هنگامی کـه بـه سمت جلو حرکت می‌کنیم تطبیق بین x_i و y_j را پیدا می‌کنیم و زیرنویس آنرا در عنـصر جـدول $B(i,j)$

ذخیره می‌کنیم و در نهایت مسیر بهینه با یک الگوریتم Back-tracking مشخص می‌گردد. این الگوریتم در الگوریتم ۲-۱ آمده است.

مزیت استفاده از برنامه‌نویسی پویا در این است که هنگامی‌که یک زیر مسأله حل شد؛ نتیجه آن می‌تواند ذخیره گردد و دیگر احتیاجی به دوباره محاسبه کردن آن نمی‌باشد. پیاده‌سازی الگوریتم‌های بازشناسی گفتار مبتنی برDTW ساده و تا حدودی برای تعداد واژگان کم مؤثر است. برنامه‌سازی پویا می‌تواند الگوهای مختلف بیان کلمات برای گویندگان مختلف را در زمان و همچنین تکرار کلمات یک گوینده را یاد بگیرد ولی نمی‌تواند یک الگوی میانگین برای تعداد زیاد نمونه‌های موجود در یک دادگان فرا بگیرد. از آنجایی که عموماً به نمونه‌های آموزشی مختلف برای مدل‌کردن تغییرات مختلف موجود در گفتار نیاز است، DTW نمی‌تواند خوب عمل کند.

بازشناسی گفتار مبتنی بر DTW از نظر پیاده‌سازی بسیار ساده می‌باشد و تا حدودی برای بازشناسی گفتار با دامنه لغات کوچک مؤثر می‌باشد. با این وجود DTW حاوی قابلیتی برای استخراج یک الگوی میانگین برای همهٔ الگوهای موجود در مقدار زیادی از نمونه‌های آموزشی نمی‌باشد. بنابراین مدل مخفی مارکف روشی مؤثر برای پردازش گفتار می‌باشد.

الگوریتم ۲-۱- الگوریتم برنامه‌نویسی پویا

مرحلهٔ ۱: مقدار دهی اولیه

قرار بده

$D(1, 1) = d(1,1), \quad B(1,1) = 1$

برای jهای بین ۲ تا M قرار بده

$D(1, j) = \infty$

مرحلهٔ ۲: تکرار

- برای iهای بین ۲ تا M مراحل زیر را انجام بده

- برای jهای بین ۱ تا M محاسبه کن

$$D(i, j) = \min_{1 \le p \le M} \left[D(i-1, p) + d(p, j) \right]$$

$$B(i, j) = \arg\min_{1 \le p \le M} \left[D(i-1, p) + d(p, j) \right]$$

مرحلهٔ ۳: Backtracking و خاتمه

فاصلهٔ مینیمم برابر با $D(N,M)$ *است و مسیر بهینه* $(s_1, s_2, \ldots, s_N)$ *می‌باشد به طوری که* $s_N = M$ *و* $s_t = B(i+1, s_{t+1})$.

۲-۳-۲- ارزیابی مدل مخفی مارکوف – الگوریتم جلورو

یک روش مؤثر برای محاسبهٔ احتمال (شباهت) $P(X \mid \Phi)$ رشته مشاهدات $X = (X_1, X_2, \ldots, X_T)$، و مدل مخفی مارکوف Φ جمع‌کردن احتمال همهٔ رشته حالت‌های ممکن می‌باشد:

$$P(X \mid \Phi) = \sum_{\text{all S}} P(S \mid \Phi) P(X \mid S, \Phi) \qquad (۱۶-۲)$$

به بیان دیگر، برای محاسبهٔ $P(X \mid \Phi)$ ابتدا باید تمامی رشته حالت‌های ممکن S با طول T که رشته مشاهدات X را تولید می‌کنند، محاسبه کرد. احتمال هر مسیر S برابر با حاصلضرب احتمال رشته حالت‌ها (فاکتور اول) و احتمال خروجی مشترک (فاکتور دوم) در طول مسیر می‌باشد.

برای یک رشته حالت مشخص $S = (S_1, S_2, \ldots, S_T)$ ، که S_1 حالت شروع آن باشد، احتمال رابطهٔ (۲-۱۶) می‌تواند با اعمال فرض مارکوف به صورت زیر بازنویسی گردد:

$$P(X \mid \Phi) = P(s_1 \mid \Phi)\prod_{t=2}^{T} P(s_t \mid s_{t-1}) = \pi_{S_1} a_{s_1 s_2} \ldots a_{s_{T-1} s_T} = a_{s_0 s_1} a_{s_1 s_2} \ldots a_{s_{T-1} s_T} \quad (۱۷-۲)$$

که برای سادگی $a_{s_0 s_1}$ با π_{s_1} نشان داده شده است. برای رشته حالت S، احتمال خروجی مشترک در طول مسیر می‌تواند با اعمال فرض خروجی غیروابسته به صورت زیر بازنویسی گردد:

$$P(X \mid S, \Phi) = P(X_1^T \mid S_1^T, \Phi) = \prod_{t=1}^{T} P(X_t \mid s_t, \Phi) \quad (۱۸-۲)$$
$$= b_{s_1}(X_1) b_{s_2}(X_2) \cdots b_{s_T}(X_T)$$

با جایگزین کردن رابطه‌های (۲-۱۷) و (۲-۱۸) در رابطه (۲-۱۶) خواهیم داشت:

$$P(X \mid \Phi) = \sum_{\text{all S}} P(S \mid \Phi) = P(X \mid S, \Phi) \quad (۱۹-۲)$$
$$= \sum_{\text{all S}} \mathrm{a}_{s_0 s_1} b_{s_1}(X_1) \mathrm{a}_{s_1 s_2} b_{s_2}(X_2) \cdots \mathrm{a}_{s_{T-1} s_T} b_{s_T}(X_T)$$

رابطه (۲-۱۹) می‌تواند به صورت زیر تفسیر گردد. ابتدا همهٔ رشته حالت‌های ممکن با طول $T+1$ را محاسبه می‌کنیم. برای هر رشته حالت داده شده، با احتمال π_{s_1} یا $a_{s_0 s_1}$ از حالت اولیه شروع می‌کنیم. با احتمال $a_{s_{t-1} s_t}$ یک گذر از s_{t-1} به s_t می‌کنیم و مشاهده X_t را با احتمال $b_{s_t}(X_t)$ ایجاد می‌کنیم تا اینکه به آخرین گذر برسیم.

با وجود اینکه محاسبهٔ مستقیم رابطهٔ (۲-۱۹) با تعبیر فوق نیاز به محاسبهٔ $O(N^T)$ عدد از رشته حالت‌های ممکن دارد که نیاز به الگوریتمی از درجهٔ نمایی دارد، خوشبختانه الگوریتم بسیار بهینه‌ای برای محاسبهٔ رابطهٔ (۲-۱۹) وجود دارد. نکته این الگوریتم در نگهداری نتایج میانی و استفاده از نتیجهٔ این محاسبات برای کم‌کردن هزینه محاسباتی می‌باشد که الگوریتم جلورو نامیده می‌شود.

طبق فرض‌های مدل مخفی مارکوف، محاسبهٔ $P(s_t \mid s_t, \Phi) P(X_t \mid s_t, \Phi)$ فقط نیاز به s_{t-1} ، s_t و X_t دارد. بنابراین امکان محاسبهٔ شباهت با احتمال $P(X \mid \Phi)$ با اعمال الگوریتم بازگشتی بر روی t وجود دارد. احتمال جلورو را به صورت زیر تعریف می‌کنیم:

$$\alpha_t(i) = P(X_1^t, s_t = i \mid \Phi) \qquad (۲۰-۲)$$

$\alpha_t(i)$ احتمال آن است که مدل مخفی مارکوف در حالت iام باشند و زیررشتهٔ مشاهدات X_1^t را تولید کرده باشد. به شرطی که پارامترهای مدل معلوم باشند $\alpha_t(i)$ را می‌توان همانطور که در الگوریتم ۲-۲ نشان داده شده است محاسبه کرد.

$\alpha_t(i)$ می‌تواند به صورت استقرایی طبق الگوریتم ۲-۲ محاسبه شود. در بسیاری از مسأله‌های گفتار، نیاز به دانستن مدل مخفی مارکف در بعضی از حالت‌های خروجیِ مشخص می‌باشد و بنابراین می‌توان بیان داشت که $P(X \mid \Phi) = \alpha_T(S_F)$ که در آن S_F حالت انتهایی می‌باشد. به راحتی می‌توان نشان داد که الگوریتم پیشرو به جای این که از مرتبهٔ نمایی باشد از مرتبه $O(N^2T)$ است. این بدان علت است که می‌توان از احتمال‌های محاسبه شده میانی به طور کامل استفاده نمود.

الگوریتم ۲-۲- الگوریتم جلورو

مرحلهٔ ۱: مقدار دهی اولیه

قرار بده

$$\alpha_t(i) = \pi_t b_t(X_1) \qquad 1 \le i \le \mathrm{N}$$

مرحلهٔ ۲: استقراء

$$\alpha_t(j) = \left[\sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_{t-1}(i)\mathrm{a_{ij}}\right] \mathrm{b_j(X_t)} \qquad 2 \le t \le T; 1 \le i \le \mathrm{N}$$

مرحلهٔ ۳: خاتمه

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

و در حالت آخر $P(X \mid \Phi) = \alpha_T(s_F)$

۲-۳-۳- رمز گشایی مدل مخفی مارکوف - الگوریتم ویتربی

الگوریتم جلورو معرفی شده در بخش قبل با استفاده از احتمال همهٔ مسیرهای ممکن احتمال این را محاسبه می‌کند که مدل مخفی مارکوف رشته‌ای از مشاهدات را تولید کند و بنابراین بهترین مسیر (یا رشته حالت‌ها) را بدست نمی‌دهد. در خیلی از کاربردها نیاز به پیداکردن چنین مسیری است. از آنجایی که رشته حالت‌ها در HMM مخفی می‌باشند بهترین ملاک برای پیدا کردن چنین مسیری، یافتن محتمل‌ترین رشته حالت‌هایی می‌باشد که رشته مشاهدات را تولید کند. به بیانی دیگر، در جستجوی رشته حالات S هستیم که احتمال $P(S,X\mid\Phi)$ را ماکزیمم کند. این مسأله بسیار شبیه به یافتن مسیر بهینه در برنامه‌سازی پویا می‌باشد و بنابراین برای یافتن بهترین حالت‌های HMM از تکنیک مبنی بر برنامه‌سازی پویا استفاده می‌شود که ویتربی نامیده می‌شود[۴]. در عمل می‌توان روشی مشابه برای ارزیابی HMMها به کار برد که حل تقریبی و نزدیک به آنچه الگوریتم جلورو بدست می‌آورد را محاسبه کند.

الگوریتم ویتربی می‌تواند به عنوان الگوریتم برنامه سازی پویای اعمال شده بر مدل مخفی مارکف در نظر گرفته شود و یا اینکه می‌توان به آن به صورت الگوریتم جلوروی تغییر یافته نگریست. الگوریتم ویتربی به جای جمع کردن احتمال مسیرهای متفاوت که یک حالت مقصد مشخص منتهی می‌شوند، بهترین مسیر را در نظر می‌گیرد و آن را به خاطر می‌آورد. احتمال بهترین مسیر به صورت زیر تعریف می‌شود:

$$V_t(i)=P\left(X_1^t,S_1^{t-1},s_t=1\mid\Phi\right) \qquad (۲۱-۲)$$

که $V_t(i)$ احتمال محتمل‌ترین (با بیشترین شباهت) رشته حالت‌های زمان t می‌باشد که مشاهدات X_1^t را که در حالت iام خاتمه می‌یابند، را تولید کند. زیر برنامهٔ مبتنی بر استقراء برای الگوریتم ویتربی در الگوریتم ۲-۳ آمده است. الگوریتم ویتربی نیز از مرتبهٔ $O(N^2T)$ می‌باشد.

مرحلهٔ ۱: مقدار دهی اولیه

قرار بده

$$V_1(i) = \pi_i b_i(X_1) \qquad 1 \le i \le N$$
$$b_1(i) = 0$$

مرحلهٔ ۲: استقراء

$$V_t(j) = \underset{1 \le i \le N}{Max}\left(V_{t-1}(i)a_{ij}\right)b_j(X_t) \qquad 2 \le t \le T; 1 \le i \le \mathrm{N}$$
$$b_t(j) = \underset{1 \le i \le N}{Arg\max}\left(V_{t-1}(i)a_{ij}\right) \qquad 2 \le t \le T; 1 \le i \le \mathrm{N}$$

مرحلهٔ ۳: خاتمه

$$\text{The Best Score} = \underset{1 \le i \le N}{Max}\left(V_t(i)\right)$$
$$s_T^* = \underset{1 \le i \le N}{Arg\max}\left(B_T(i)\right)$$

مرحلهٔ ۴: Backtracking

$$s_t^* = B_{t+1}(s_{t+1}^*) \qquad t = T, T-1, T-2, \cdots, 1$$

و بهترین رشته $S^* = (s_1^*, s_2^*, \ldots, s_T^*)$ می‌باشد.

۲-۳-۴- تخمین مدل مخفی مارکوف – الگوریتم بام‌ـولش

بسیار حائز اهمیت است که پارامترهای مدل $\Phi = (A, B, \pi)$ طـوری تخمـین زده شـونده کـه رشـته مشاهدات را به طور دقیق توصیف کنند. این سخت‌ترین مسأله در بین سه مسألهٔ اساسی مدل مخفی مارکف می‌باشد زیرا هیچگونه راه حل تحلیلی برای ماکزیمم کردن احتمال مشترک داده‌های آموزشی وجـود نـدارد. به جای آن این مسأله به وسیله الگوریتم تکراری بام–ولش قابل حل است کـه بـه‌نـام الگوریتـم جلـورو–عقب‌رو نیز شناخته می‌شود. قبل از تو صیف الگوریتم بام–ولش، یک پارامتر جدید معرفی می‌گردد. مشابه با احتمال جلورو، احتمال عقب‌رو نیز می‌تواند به صورت زیر تعریف گردد:

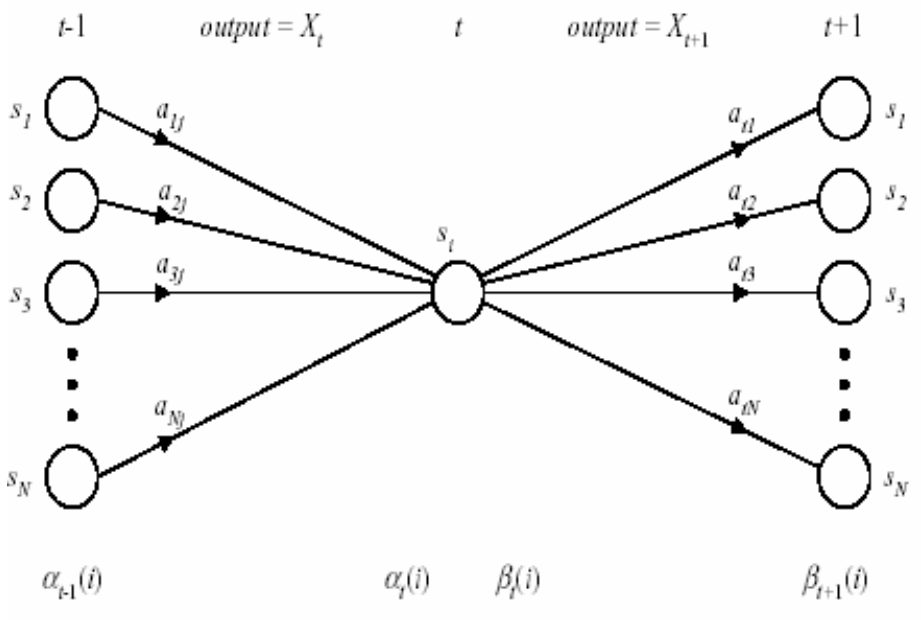
$$\beta_t(i) = P\left(X_{t+1}^T \mid s_t = i, \Phi\right) \qquad (۲۲-۲)$$

که $\beta_t(i)$ احتمال تولید مشاهدات میانی X_{t+1}^T (از $t+1$ تا انتها) می‌باشد که مـدل مخفـی مـارکف در حالت i و در زمان t باشد. $\beta_t(i)$ می‌تواند به صورت استقرایی به روش زیر محاسبه گردد.

$$\beta_T(i) = 1/N \qquad i \le i \le N \tag{۲۳-۲}$$

$$\beta_t(i) = \left[\sum_{j=1}^{N} a_{ij} b_j(X_{t+1}) \beta_{t+1}(j)\right] \qquad t = T-1,\ldots,1;\ 1 \le i \le N \tag{۲۴-۲}$$

رابطه بین α و β های کنار هم به صورت شکل ۲-۲ می تواند نمایش داده شود. α به صورت بازگشتی از سمت چپ به راست و β به شکل بازگشتی از سمت راست به چپ قابل محاسبه است.



شکل ۲-۲- رابطهٔ بین α_{t-1}، α_t و β_t ، β_{t+1} در الگوریتم جلورو-عقب‌رو

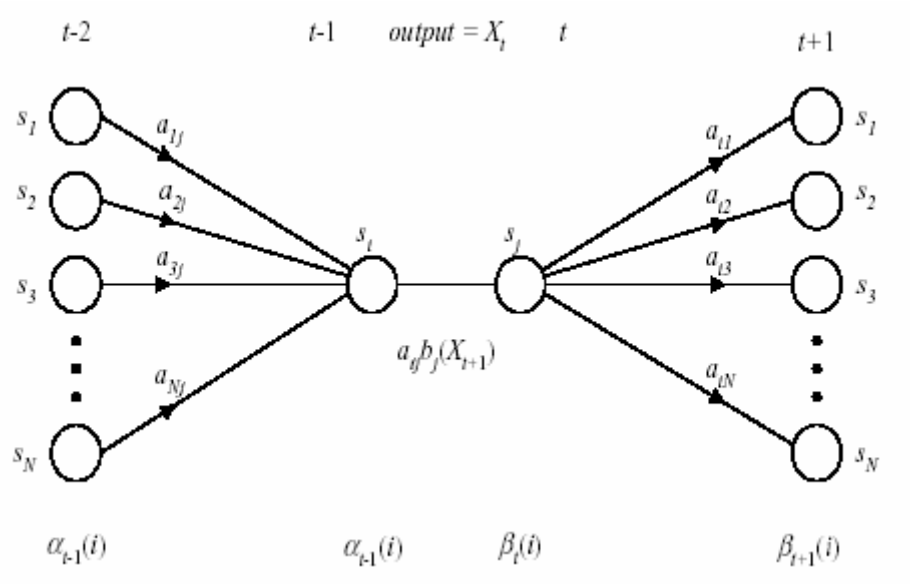
حال $\gamma_t(i,j)$ که احتمال گذر از حالت i به حالت jدر زمان t در یک مدل و رشتهٔ مشاهدات داده شده می‌باشد، تعریف می‌شود.

$$\gamma_t(i,j) = P\left(s_{t-1} = i, s_t = j \mid X_1^T, \Phi\right) \tag{۲۵-۲}$$

$$= \frac{P\left(S_{t-1} = i, S_t = j, X_1^T \mid \Phi\right)}{P(X_1^T \mid \Phi)}$$

$$= \frac{\alpha_{t-1}(i) a_{ij} b_j(X_t) \beta_t(j)}{\sum_{k=1}^{N} \alpha_T(k)}$$

رابطهٔ (۲-۲۵) می تواند به وسیلهٔ شکل ۲-۳ نمایش داده شود.



شکل ۲-۳- عمل‌های لازم برای محاسبهٔ $\gamma_t(i,j)$ (احتمال گذر از حالت i به حالت j در زمان t)

حال به طور تکراری بردار پارامترهای مدل مخفی مارکف $\Phi=(A,B,\pi)$ با ماکزیمم کردن شباهت $P(X|\Phi)$ برای هر تکرار، بهتر می‌شوند. $\hat{\Phi}$ نشان دهندهٔ بردار پارامتر جدید حاصل از بردار پارامتر Φ می‌باشد. پروسهٔ ماکزیمم کردن معادل ماکزیمم کردن تابع Q زیر می‌باشد:

$$Q(\Phi,\hat{\Phi})=\sum_{all\,s}\frac{P(X,S|\Phi)}{P(X|\Phi)}\log P(X,S|\Phi) \qquad (۲۶-۲)$$

که $P(X,S|\Phi)$ و $\log P(X,S|\tilde{\Phi})$ می توانند به صورت زیر بیان گردند:

$$P(X,S|\Phi)=\prod_{t=1}^{T}a_{s_{t-1}s_t}b_{s_t}(X_t) \qquad (۲۷-۲)$$

$$\log P(X,S|\Phi)=\sum_{t=1}^{T}\log a_{s_{t-1}s_t}+\sum_{t+1}^{T}\log b_{s_t}(X_t) \qquad (۲۸-۲)$$

و رابطهٔ (۲-۲۶) می تواند به صورت زیر بازنویسی گردد:

$$Q(\Phi,\hat{\Phi})=Q_{a_j}(\Phi,\hat{a}_i)+Q_{b_j}(\Phi,\hat{b}_j) \qquad (۲۹-۲)$$

که

$$Q_{a_j}(\Phi,\hat{a}_i)=\sum_i\sum_j\sum_t\frac{P(X,S_{t-1}=i,s_t=j|\Phi)}{P(X|\Phi)}\log\hat{a}_{ij} \qquad (۳۰-۲)$$

$$Q_{a_j}\left(\Phi,\hat{a}_j\right)=\sum_j\sum_k\sum_{t\in X_t=o_k}\frac{P\left(X,S_t=j\mid\Phi\right)}{P\left(X\mid\Phi\right)}\log\hat{b}_j(k) \qquad (۳۱-۲)$$

از آن جایی که تابع Q به سه رابطهٔ غیر وابسته تقسیم شدند، پروسه ماکزیمم کردن $Q\left(\Phi,\hat{\Phi}\right)$ می‌تواند با ماکزیمم کردن هر یک از روابط به طور جداگانه و با توجه به محدودیت‌های احتمال انجام شود:

$$\sum_{j=1`}^{N}a_{ij}=1 \qquad \forall all\, i \qquad (۳۲-۲)$$

$$\sum_{k=1}^{M}b_j(k)=1 \qquad \forall all\, j \qquad (۳۳-۲)$$

به علاوه تمام روابط واقع در رابطه‌های (۲-۳۰) و (۲-۳۱) به صورت زیر می‌باشند:

$$F(x)=\sum_i\gamma_i\log x_i \qquad (۳۴-۲)$$

که $\sum_i x_i=1$.

با استفاده ضرب کننده های لاگرانژ، اثبات می‌شود که تابع بالا بیشترین مقدار خود را موقعی بدست می آورد که:

$$x_i=\frac{\gamma_i}{\sum_i\gamma_i} \qquad (۳۵-۲)$$

به این ترتیب، تخمین مدل به صورت زیر خواهد بود:

$$\hat{a}_{ij}=\frac{\frac{1}{P(X\mid\Phi)}\sum_{t=1}^{T}P(X,s_{t-1}=i,s_t=j\mid\Phi)}{\frac{1}{P(X\mid\Phi)}\sum_{t=1}^{T}P(X,s_{t-1}=i\mid\Phi)}=\frac{\sum_{t=1}^{T}\gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{k=1}^{N}\gamma_t(i,k)} \qquad (۳۶-۲)$$

$$\hat{a}_{ij}(k)=\frac{\frac{1}{P(X\mid\Phi)}\sum_{t=1}^{T}P(X,s_t=j\mid\Phi)\delta(x_k,o_k)}{\frac{1}{P(X\mid\Phi)}\sum_{t=1}^{T}P(X,s_t=j\mid\Phi)}=\frac{\sum_{t\in X_t=O_k}\sum_i\gamma_t(i,j)}{\sum_{t=1}^{T}\sum_i\gamma_t(i,j)} \qquad (۳۷-۲)$$

با توجه دقیق به روابط تخمین مدل مخفی مارکف (۲-۳۶) و (۲-۳۷)، می توان مشاهده کرد که رابطهٔ (۲-۳۶) در واقع نسبت بین امید تعداد گذرها از حالت i به حالت jو امید تعداد گذرها از حالت i می باشد. در صورت رابطه تخمین احتمال خروجی(۲-۳۷)، امید تعداد دفعاتی است که داده مشاهدهٔ خارج شده از حالت jبا سمبل مشاهده O_k می‌باشد. مخرج آن امید تعداد دفعاتی است که داده مشاهده خارج شده از حالت jاست. الگوریتم جلورو–عقب‌رو (بام–ولش) تضمین می‌کند که در هر مرحلهٔ تکرار شباهت افزایش پیدا کند و بالاخره این شباهت به یک ماکزیمم محلی همگرا گردد. الگوریتم جلورو–عقب‌رو می‌تواند به صورت الگوریتم ۲-۴ بیان شود.

با وجود این که الگوریتم جلورو–عقب‌رو توصیف شده در بالا مبتنی بر یک رشته مشاهدهٔ آموزشی می‌باشد ولی می‌تواند با فرض غیروابسته بودن رشتهٔ مشاهدات، برای چندین رشته مشاهدات آموزشی عمومیت یابد. آموزش یک مدل مخفی مارکوف با M رشته داده معادل با پیدا کردن بردار پارامتر مدل مخفی مارکف Φ است که تابع زیر را ماکزیمم کند:

$$\prod_{i=1}^{M} P(X_i \mid \Phi) \qquad (۲-۳۸)$$

پروسه آموزش الگوریتم جلورو–عقب‌رو را بر روی هر یک از رشته‌های مشاهدات مستقل اجرا می‌کند تا روابط (۲-۳۶) و(۲-۳۷) را محاسبه کند. امیدهای موجود در مخرج این روابط می توانند به ترتیب به M رشته داده اضافه شوند. در انتها، همهٔ پارامترهای مدل (احتمالات) نرمالیزه می‌شوند تا جمع همگی آن‌ها برابر با یک گردد. کارهای بالا به انجام یک تکرار از تخمین بام–ولش منجر می‌شود و این تکرارها تا جایی ادامه پیدا می‌کنند تا همگرایی حاصل شود. این پروسه عملی و مناسب است زیرا که اجازه می‌دهد که یک مدل مخفی مارکف در یک سناریو بازشناسی گفتار با توجه به حجم زیاد داده‌های آموزشی، به خوبی

آموزش یابد. به عنوان مثال اگر $\gamma_t^m(i,j)$ بیان کنندهٔ $\gamma_t(i,j)$ برای رشته دادهٔ mام باشـد و T^m طـول آن را نشان دهد، رابطهٔ (۲-۳۶) می‌تواند به صورت زیر توسعه یابد:

$$\hat{a}_{ij} = \frac{\sum_{m=1}^{M}\sum_{t=1}^{T^M}\gamma_t^m(i,j)}{\sum_{m=1}^{M}\sum_{t=1}^{T^m}\sum_{k=1}^{N}\gamma_t^m(i,k)} \qquad (۲-۳۹)$$

الگوریتم ۲-۴- الگوریتم جلورو–عقب‌رو

مرحلهٔ ۱: مقدار دهی اولیه: یک تخمین اولیه برای Φ *انتخاب کن.*

مرحلهٔ ۲: براساس تابع کمکی $Q(\Phi,\hat{\Phi})$ *را محاسبه کن.*

مرحلهٔ ۳: برای ماکزیمم کردن تابع کمکی Q و با توجه به روابط (۲-۳۸) و (۲-۳۹)، $\hat{\Phi}$ را محاسبه کن.

مرحلهٔ ۴: تکرار: قرار بده $\Phi = \hat{\Phi}$ *و تا هنگام همگرایی از مرحلهٔ ۲ دوباره اجرا کن.*

فصل سوم

سیستم‌های بازشناسی گفتار

۳-۱- مقدمه

از آنجایی که اغلب رویکردها برای سیستم‌های بازشناسی گفتار رویکردهای فرآیندهای تصادفی می‌باشند و در این پروژه نیز از سیستم بازشناسی گفتار مبتنی بر HMM استفاده شده است در این فصل ابتدا قسمت‌های مختلف یک سیستم بازشناسی گفتار مبتنی بر HMM شرح داده می‌شوند و سپس در انتهای فصل به طور خلاصه به همهٔ تکنیک‌های بازشناسی گفتار اشاره می‌شود.

اگر چارچوب فرآیندهای تصادفی را در نظر بگیریم آنگاه یک مدل عمومی برای سیستم‌های بازشناسی گفتار را می‌توان به صورت زیر بیان کرد. اگر فرض کنیم W نشان دهندهٔ دنباله‌ای از کلمات و A نشان دهندهٔ دنباله‌ای از نمادهای آکوستیکی باشد آنگاه هدف از مسألهٔ بازشناسی گفتار اینست که با داشتن یک دنبالهٔ آکوستیکی مانند A، دنبالهٔ کلمات $\hat{W}$ را چنان پیدا کنیم که احتمال آن به شرط A ($P(W \mid A)$) ماکزیمم شود [۵]،[۶]:

$$\hat{W} = \arg\max_{W} P(W \mid A) \qquad (۱-۳)$$

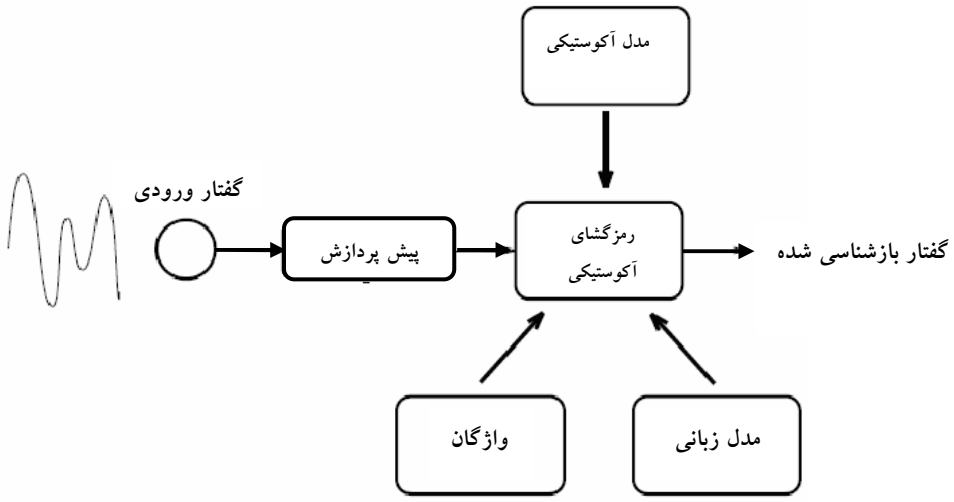
در رابطهٔ بالا سیستم بازشناسی گفتار مشروط بر مشاهدهٔ دنبالهٔ آکوستیکی A، محتمل‌ترین جمله از کلمات را به عنوان نتیجهٔ بازشناسی برمی‌گرداند. با استفاده از قانون بیز ، رابطهٔ بالا را می‌توان به صورت زیر بازنویسی کرد:

$$\hat{W} = \arg\max_{W} \frac{P(A \mid W)P(W)}{P(A)} = \arg\max_{W} P(A \mid W)P(W) \qquad (۲-۳)$$

در رابطهٔ بالا، $P(A|W)$ بیانگر احتمال مشاهدهٔ دنبالهٔ آکوستیک A، مشروط بر بیان جملهٔ W توسط گوینده می‌باشد. این احتمال در سیستم بازشناسی به وسیلهٔ یک رمزگشای آکوستیکی[1] و با استفاده از مدل آکوستیکی[2] گفتار محاسبه می‌شود. $P(W)$ نیز بیانگر احتمال وقوع جملهٔ W توسط گوینده می‌باشد.

[1] *Acoustic Decoder*
[2] *Acoustic Model*

این احتمال به وسیلهٔ یک رمزگشای زبانی[1] و با استفاده از مدل زبانی[2] و واژگان[3] محاسبه می‌گردد. در شکل ۳-۱ یک سیستم بازشناسی گفتار با سه منبع اصلی مدل آکوستیکی، مدل زبانی و درخت واژگان آمده است. همانطور که قبلاً نیز اشاره شد مدل آکوستیکی سیستم بازشناسی گفتار استفاده شده در این پایان‌نامه HMM می‌باشد که به طور کامل در فصل دوم شرح داده شد.



شکل ۳-۱- اجزای اصلی یک سیستم بازشناسی گفتار

واژگان یک نگاشت از آنچه مدل آکوستیکی مدل می‌کند به کلمات موجود در مدل زبانی و مجموعهٔ واژگان سیستم می‌باشد. در بازشناسی‌های با تعداد لغت کم، واژگان یک نگاشت یک به یک از مدل آکوستیکی به کلمات می‌باشد. در بازشناسی‌های با تعداد لغت زیاد ممکن است که مدل آکوستیکی واحدهای کوچکتر از کلمه را مدل کند که در این صورت واژگان چگونگی اتصالِ این واحدها را برای ایجاد کلمات نیز نشان می‌دهد. به عنوان مثال کلمهٔ *گفتار* ممکن است به شکل زیر به واحدهای آوایی واجی شکسته شود:

$$\text{گفتار} = \{ g\ o\ f / r \}$$

[1] *Linguistic Decoder*
[2] *Language Model*
[3] *Lexicon*

مدل زبانی در بر دارندهٔ این اطلاعات می‌باشد که چه رشته‌ای از کلمات می‌توانند پشت سرهم قرار بگیرند. این مدل با محدود کردن فضای جستجو به طور چشمگیری می‌تواند بار محاسباتی را کم کند و دقت بازشناسی را بالا ببرد. در یک سیستم بازشناسی گفتار مدل‌های زبانی مختلفی ممکن است مورد استفاده قرار بگیرند مانند گرامر دوکلمه‌ای که بیان می‌کند چه کلماتی می‌توانند پس از کلمهٔ جاری قرار بگیرند، یا مدل‌های آماری مانند N-grams که یک احتمال به رشته‌ای از کلمات نسبت می‌دهند.

مدل زبانی N-grams معمولاً در سیستم‌های بازشناسی مبتنی بر HMM مورد استفاده قرار می‌گیرد. احتمال یک رشته کلمهٔ خاص به صورت زیر بیان می‌شود:

$$P(W_1^n) = P(W_1)p(W_2 \mid W_1)\cdots p(W_n \mid W_1^{n-1}) \qquad (۳-۳)$$

که W_1^n رشته‌ای از کلمات از W_1 تا W_n می‌باشد. این امکان وجود ندارد که همهٔ آموزش داده شوند. یک فرض ساده این است که احتمال هر کلمه فقط به N-۱ کلمه قبل از خود وابسته باشد. در این صورت یک مدل زبانی N-gram خواهیم داشت که احتمال این مدل زبانی به صورت زیر بدست می‌آید:

$$P(W_1^n) \approx \prod_{k=1}^{n} p\left(W_k \mid W_{k-N+1}^{k-1}\right) \qquad (۴-۳)$$

که معمول‌ترین فرم‌های N-gram که استفاده می‌شوند trigram (N = ۳)، bigram (N = ۲) و unigram (N = ۱) می‌باشند.

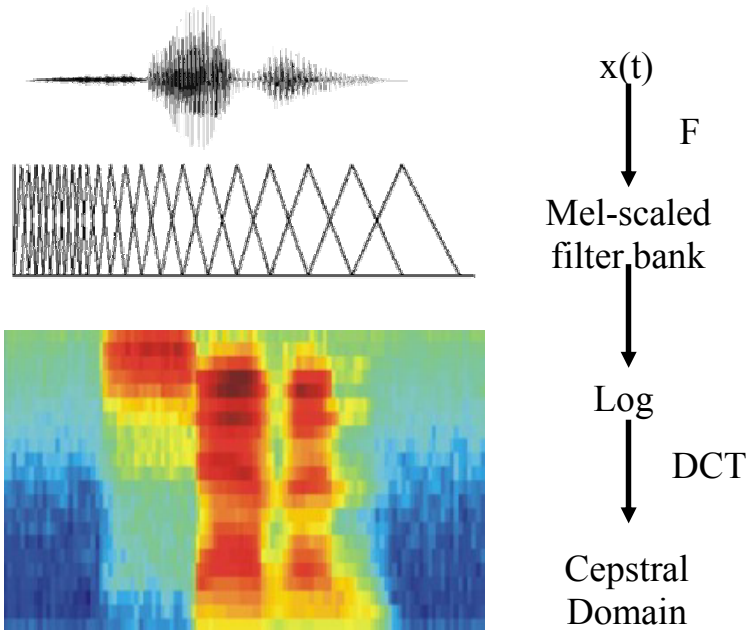
در بخش‌ها ۳-۲ تا ۳-۴ این فصل به طور مفصل به قسمت‌های پیش‌پردازش، مدل آکوستیکی و رمزگشای آکوستیکی از یک سیستم بازشناسی گفتار مبتنی بر HMM پرداخته می‌شود.

۳-۲- پیش‌پردازش - استخراج ویژگی‌های سیگنال گفتار

برای پردازش و آنالیز سیگنال گفتار، باید افزونگی‌های آن حذف گردند و در نتیجه محاسبات و پیچیدگی محاسباتی در سیستم‌های بازشناسی گفتار کاهش یابد. یکی از مسائلی که در مورد سیگنال گفتار وجود دارد، این است که این سیگنال با مشخصات مجرای تولید گفتار در طول زمان تغییر می‌کند و بنابراین باید سیگنال گفتار را که یک سیگنال غیرایستان است در یک بازه زمانی کوتاه ۲۰ الی ۴۰ میلی ثانیه‌ای تحلیل و آنالیز کرد. به همین دلیل احتیاج است که سیگنال گفتار را به فریم‌هایی که با هم همپوشانی دارند، تقسیم کنیم. در سیستم‌های بازشناسی گفتار هر فریم از سیگنال گفتار به یک بردار ویژگی تبدیل می‌شود، بدین وسیله افزونگی‌هایی که به آن نیاز نیست، از سیگنال گفتار حذف می‌شوند. به عنوان مثال یک فریم ۵۱۲ نمونه‌ای به یک بردار ویژگی ۳۹ تایی تبدیل می‌شود. پس از این کاهش بعد می‌توانیم این بردار را به یک سیستم بازشناسی گفتار بدهیم تا پردازش‌های لازم بر روی آن انجام گیرند. در ادامه به توضیح روش استخراج ضرایب کپسترال مبتنی بر معیار مل[1] به عنوان یکی از پرکاربردترین روش‌های استخراج ویژگی پرداخته می‌شود.

یکی از روش‌های استخراج ویژگی که بر اساس آنالیز بانک فیلتر و نیز ادراک گوش انسان از فرکانس کار می کند، استخراج ویژگی کپسترال مبتنی بر معیار مل است که ضرایب بدست آمده ضرایبMFCC نامیده می‌شوند.

[1] *Mel Frequency Cepstral Coefficients*

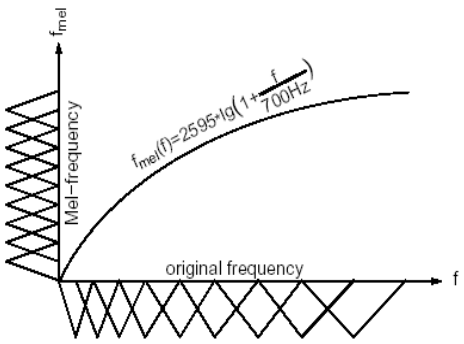


شکل ۳-۲- روش استخراج ویژگی‌های استاندارد MFCC

اگر واحد فرکانس بر اساس ادراک انسان را مل بنامیم رابطهٔ فرکانس در واحد مل با واحد فرکانس در واحد هرتز به صورت رابطهٔ زیر می‌باشد:

$$F_{mel} = 2595 \log_{10}\left(1+\frac{F_{HZ}}{700}\right) \qquad (۵-۳)$$

بر اساس این واقعیت که ادراک انسان از اصوات بر اساس هرتز نمی باشد، در روش استخراج ضرایب بر اساس مل، بانک فیلتر را بر روی محور فرکانس به صورت لگاریتم و بر اساس فرمرل (۳-۵) توزیع می‌کنند و شکل فیلترهایی که بر روی باند فرکانس توزیع می‌شوند مطابق شکل ۳-۳ است.



شکل ۳-۳- نحوهٔ تبدیل واحد فرکانس به واحد مل

مشاهده می‌شود که در فرکانس‌های بالا، پهنای باند فیلترها زیادتر می‌گردد و این بدان معناست که قدرت تفکیک فرکانسی و حساسیت نسبت به تغییر فرکانس در فرکانس‌های بالا در گوش انسان کمتر از این حساسیت و تفکیک در فرکانس‌های پایین است. شکل فیلترهایی که بر روی بانک فرکانسی توزیع می‌گردند، می تواند مثلثی، مستطیلی، کسینوسی یا ذوزنقه‌ای باشند.

۳-۳- مدل آکوستیکی – امتیازدهی ویژگی‌ها

برای این‌که گفتار به طور دقیق و مناسب مدل‌سازی گردد، لازم است گفتار به بخش‌های کوچک‌تری به نام «واحد آوایی» تقسیم گردد و هر بخش جداگانه مدل‌سازی شود. بسته به نوع و کاربرد سیستم بازشناسی یعنی پیوسته یا منفصل بودن بازشناسی، اندازهٔ واژگان، زبان مورد استفاده و ...، واحدهای آوایی مختلفی ممکن است به کار رود. واحدهای آوایی رایج عبارتند از: واج، هجا، دیاد (نیم‌هجا)، کلمه و واحدهای آوایی وابسته به بافت (مانند دوواجی, سه‌واجی، واجگونه و ...). در ادامه به عنوان نمونه کلمهٔ «گفتار» بر اساس واحدهای آوایی مختلف نمایش داده شده است:

واج: /g/ /o/ /f/ /t/ /â/ /r/
هجا: /gof/ /târ/
دیاد: /go/ /of/ /tâ/ /âr/
دوواجی: /$-g/ /g-o/ /o-f/ /f-t/ /t-â/ /â-r/

شکل ۳-۴- نمایش کلمهٔ «گفتار» بر اساس انواع واحدهای آوایی

مدل‌های مختلفی برای مدل‌سازی واحدهای آوایی گفتار می‌تواند مورد استفاده قرار گیرند که سیستم مورد استفاده در پروژه این کار را با کمک مدل مخفی مارکوف انجام می‌دهد. در بخش ۳-۵-۳ علاوه بر نحوهٔ این مدل‌سازی به کمک مدل مخفی مارکوف به انوع مدل‌های دیگر با توضیح مختصری راجع به هر یک اشاره خواهد شد.

۳-۴- رمزگشای آکوستیکی - رویهٔ جستجو

همان طور که قبلاً بیان شد، در بازشناسی گفتار هدف اینست که با داشتن یک دنبالهٔ آکوستیک مانند A، دنبالهٔ لغات $\hat{W}$ را چنان پیدا کنیم که احتمال $P(W \mid A)$ ماکزیمم شود؛ بنابراین مسألهٔ بازشناسی گفتار پیوسته، هم یک مسألهٔ بازشناسی الگو و هم یک مسألهٔ جستجو است.

در یک سیستم بازشناسی گفتار پیوسته، با در اختیار داشتن مدل‌های آکوستیک واحدهای آوایی، مسألهٔ رمزگشایی آکوستیکی[1] گفتار ورودی به یک مسألهٔ جستجو تبدیل می‌شود که هدف آن پیدا کردن دنباله‌ای از واحدهای آوایی است، به طوری که بهترین انطباق ممکن بین سیگنال ورودی و مدل‌های آکوستیک ایجاد شود.

در ضمن جستجو، بردارهای ویژگی گفتار ورودی، با مدل‌های مربوط به واحدهای آوایی، مقایسه شده و دنباله‌هایی از واحدهای آوایی به عنوان فرضیه شکل می‌گیرد. فرضیه‌ای که بیشترین امتیاز را داشته

[1] *acoustic decoding*

باشد، بهترین جواب رمزگشای آکوستیک خواهد بود. در صورت استفاده از HMM برای بازشناسی گفتار، حالت‌های HMM فضای جستجو را تشکیل می‌دهند.

در یک جستجوی کامل که همهٔ فرضیه‌های ممکن مورد بررسی قرار می‌گیرند، تعداد فرضیه‌ها با افزایش تعداد بردارهای ویژگی گفتار ورودی، به طور نمایی افزایش پیدا می‌کند. بنابراین انجام یک جستجوی کامل در سیستم‌های واقعی تقریباً غیرممکن است. به همین جهت، روش‌های جستجویی پیشنهاد شده است که قادرند به جای بررسی کل فضای جستجو، تنها با بررسی قسمتی از فضای جستجو، جواب خوبی بدهند [۵۲].

در بازشناسی گفتار پیوسته، روش‌های ترکیبی زیادی برای جستجو وجود دارد. مشهورترین روش‌های جستجو عبارتند از: جستجوی ویتربی[1] که بر مبنای الگوریتم ویتربی [۵]، [۶] عمل می‌کند؛ جستجوی «ویتربی شعاعی»[2] [۷] که شکل کاراتری از جستجوی ویتربی است و «جستجو بر مبنای پشته » [۸] که بر مبنای الگوریتم جستجوی*A عمل می‌کند. جستجوهای ویتربی و ویتربی شعاعی جزء جستجوهای پیمایش اول سطح[3] و جستجو بر مبنای پشته جزء جستجوهای پیمایش اول عمق[4] محسوب می‌شود.

۳-۵- تکنیک‌های بازشناسی گفتار

بطور کلی سه روش عمده برای بازشناسی گفتار وجود دارد که در ادامه به آن پرداخته می‌شود.

۳-۵-۱- روش ترکیبات آوایی

این روش بر پایهٔ نظریهٔ آواشناسی می‌باشد که بر اساس واحدهای آوایی مجزای موجود در زبان و واحدهای آوایی که به صورت گسترده به وسیلهٔ مجموعه‌ای از مشخصات سیگنال گفتار استخراج می‌گردند،

---

[1] *viterbi search*
[2] *beam viterbi*
[3] *Breath First Search*
[4] *Depth First Search*

مدل می‌شوند. مشخصات واحدهای آوایی تحت تأثیر واحدهای آوایی همجوار و گوینده به صورت خیلی زیاد تغییر می‌کنند. این روش با تقسیم کردن سیگنال گفتار به واحدهای مجزا که مشخصات آوایی سیگنال در آن واحدها یک یا چندین واحد آوایی را مشخص می‌کنند، شروع می‌شود و سپس به هر کدام از واحدهای متناظر با مشخصات آوایی، یک یا چندین برچسب واحد آوایی زده می‌شود. در نهایت با استفاده از دنبالهٔ برچسب‌های به دست آمده شروع به مشخص کردن یک کلمهٔ معتبر می‌نماید.

۳-۵-۲- روش هوش مصنوعی

ایدهٔ اصلی این روش برای بازشناسی گفتار به‌کارگیری اطلاعات از منابع گوناگون اطلاعاتی است. این منابع می‌توانند اطلاعات آوایی، اطلاعات واژگان، اطلاعات نحوی، اطلاعات معنایی و غیره باشد. روش‌های مختلفی برای وارد کردن منابع اطلاعاتی در یک بازشناسی کنندهٔ گفتار وجود دارد از جملهٔ آن‌ها می‌توان به روش بالا-به-پایین[1] یا پایین-به-بالا[2] اشاره کرد.

۳-۵-۳- روش بازشناسی الگو

در اکثر روش‌های از این دست، دو مرحله وجود دارد، مرحلهٔ اول آموزش است و مرحلهٔ دوم بازشناسی الگوها از طریق مقایسهٔ دادهٔ آزمون و الگوهاست. فرآیند آموزش باید بتواند مشخصات آوایی را از نسخه‌های گوناگونی که برای یک الگو در اختیار دارد، استخراج کند. متداول‌ترین الگوهایی که برای مدل‌سازی گفتار در مسألهٔ بازشناسی مورد استفاده قرار می‌گیرند، عبارتند از:

[1] *Top-Down*
[2] *Bottom-Up*

مدل انطباق زمانی پویا[1] (*DTW*)

نقطهٔ قوت تکنیک DTW همانطور که قبلاً نیز بیان شد در رفع عدم انطباق زمانی اجـزای واحـدهای بازشناسی، هنگام انطباق الگو بوسیلهٔ جابجایی پویای زمانی است. در این تکنیک برای هر واحـد بازشناسـی یک الگوی مرجع وجود دارد که هنگام بازشناسی ، میزان انطباق الگـوی ورودی بـا تمـام الگوهـای مرجـع محاسبه می‌شود. این کار در حجم لغات زیاد نیازمند حجم حافظه و محاسبات بسیار بالایی خواهـد بـود. از طرف دیگر در کاربرد برای بازشناسی گفتار پیوسته به لحاظ مشخص نبودن مرز کلمات به طور دقیق و نیـز در استفاده برای گویندگان متعدد، این تکنیک کارایی خود را از دست می‌دهد. به همـین دلیـل از ایـن مـدل بیشتر برای کاربردهایی که تعداد گویندگان محدود و تعداد لغات کمتر از ۱۰۰ کلمه است، استفاده می‌شود.

مدل مخفی مارکوف[2] (*HMM*)

HMM همانطوری که بیان شد یک مدل آماری از نوع اتفاقی دوگانه[3] است که هر واحد آوایـی را بـا مجموعه‌ای از حالت‌ها به همراه احتمالات انتقال از حالتی به حالت دیگر مدل مـی‌کنـد. توانـایی HMM در مدل‌سازی تغییرات زمانی سیگنال گفتار و داشتن پشتوانهٔ ریاضی قوی، آن را بـه عنـوان ابـزاری کارآمـد در بازشناسی گفتار مطرح می‌کند. در روش‌های آموزش کلاسیک HMM برای هر واحد آوایی بطـور جداگانـه با استفاده از نمونه‌های متناظر از مجموعهٔ آموزش، یک مدل ساخته می‌شود. اینکه برای آموزش یک مدل از واحدهای آوایی دیگر استفاده نمی‌شود یـک عیـب عمـده محـسوب مـی‌شـود، هرچنـد روش‌هـایی بـرای به‌کارگیری اطلاعات واحدهای آوایی دیگر وجود دارد ولی با این حال تا به امروز این روش مـدل‌سـازی از موفق‌ترین روش‌ها برای حل مسألهٔ بازشناسی گفتار پیوسته با تعداد کلمات زیاد و مـستقل از گوینـده بـوده است.

---

[1] *Dynamic Time Warping*
[2] *Hidden Markov Model*
[3] *Doubly Stochastic*

شبکهٔ عصبی مصنوعی[1] (*ANN*)

مدل دیگری که در بازشناسی گفتار از آن استفاده می‌شود شبکهٔ عصبی است که قدرت بالایی از نظر پردازش موازی، طبقه‌بندی و بازشناخت الگو دارد. یک مشکل عمدهٔ شبکهٔ عصبی در حجم محاسبات بالا و زمان طولانی یادگیری آن می‌باشد. در روش‌های آموزش شبکهٔ عصبی مصنوعی، امتیاز هر کلاس خاص (مدل) با توجه به امتیاز کلاس‌های دیگر تعیین می‌شود که نسبت به روش‌های کلاسیک آموزش HMM یک مزیت محسوب می‌شود.

مدل‌های ترکیبی[2]

هر کدام از مدل‌های فوق، نقاط ضعف و قوت ویژهٔ خود را دارند. با ترکیب مدل‌ها سعی می‌شود نقاط ضعف هر کدام پوشانده شود. به عنوان مثال شبکه‌های عصبی معمولاً طبقه‌بندی کننده‌های ایستایی هستند، در حالی که HMM دینامیک است و رفتار کوتاه مدت گفتار را به خوبی مدل می‌کند. در اوایل دههٔ ۹۰ ترکیب شبکه‌های عصبی و HMM با هدف گردآوری خواص خوب هر دو متداول گشت. همچنین ترکیب HMM و روش‌های فازی نیز به کار گرفته شده است.

[1] *Artificial Neural Network*
[2] *Hybrid Models*

فصل چهارم

بازشناسی گفتار در محیط‌های نویزی

۴-۱- مقدمه

امروزه، سیستم‌های بازشناسی گفتار به وفور در محیط‌های مختلف مورد استفاده قرار می‌گیرد. این سیستم‌ها، با وجود اینکه در شرایط آزمایشگاهی، خوب عمل می‌کنند ولی وقتی که سراغ شرایط واقعی می‌رویم، کارایی آنها به شدت کاهش می‌یابد.آزمایش‌های متعدد انجام شده، نشان می‌دهند که حتی سیستم‌هایی که در شرایط آزمایشگاهی، خیلی خوب عمل می‌نمایند، در برخی از حالت‌های غیر آزمایشگاهی، از دستیابی به حداقل کارایی مورد قبول نیز ناتوانند[۱۰].

محیط‌های واقعی، معمولاً دارای شرایط آکوستیکی غیر قابل کنترلی بوده و رفتارهایی کاملاً متمایز با محیط طراحی سیستم از خود نشان می‌دهند. این مسأله موجب ایجاد ناسازگاری بین شرایط آموزش و آزمایش گردیده و کاهش کارایی را به دنبال خواهد داشت.

در این فصل پس از ارائهٔ مدلی برای نویز به بررسی شبکه خط تلفن به عنوان عاملی مخرب برای سیستم‌های بازشناسی گفتار پرداخته می‌شود. سپس تأثیرات نویز در سیستم‌های بازشناسی گفتار بیان می‌گردد و در انتها انواع روش‌های مقابله با نویز ارائه می‌شوند.

۴-۲- مدلی عمومی برای نویز

منابع زیادی برای تغییر گفتار از حالت آزمایشگاهی وجود دارند که ما در یک دید کلی، همهٔ آن‌ها را منابع «نویز» می‌نامیم. معمولاً محیط در حالت آموزش، محیطی نسبتاً ایده‌آل و تا حد امکان، عاری از نویز می‌باشد، حال آنکه در مراحل آزمایش، انواع و اقسام نویزها بر عملکرد سیستم تأثیر می‌گذارند. برخی از این نویزها، حالت جمعی یا جمع شونده[1] و برخی دیگر حالت پیچشی[2] دارند. یک دسته از نویزهای مؤثر بر

[1] *Additive*
[2] *Convolutional*

سیستم‌های بازشناسی، نویزهایی کوتاه مدت و غیر ایستان می‌باشند. صدای عبور اتومبیل، بسته شدن درب و همچنین گفتار گوینده‌ای غیر از گویندهٔ اصلی در این دسته از نویزها قرار می‌گیرند. دستهٔ دیگر، نویزهای پیوسته در زمان همچون صدای کولر و فن می‌باشند. انعکاس و طنین گفتار در محیط بسته مانند اتاق، خانوادهٔ دیگری از نویزها را به نام نویزهای پیچشی شکل می‌دهند. مشکل‌ها و ضعف‌های سیستم انتقال و میکروفون و پهنای باند محدود کانال انتقال، هر یک موجب ایجاد نویزهای خاص خود می‌شوند البته در یک نگرش کلی‌تر می توان پهنای باند محدود را در خانوادهٔ نویزهای پیچشی قرار داد و بالاخره، باید از اثر لمبارد[1] یاد کرد. این اثر دارای این مفهوم است که در محیط‌های نویزی، خود گوینده هم حالت طبیعی حرف‌زدن را نداشته و تکیه‌ها و تاکیدهای متفاوت با حالت طبیعی، گفتار مورد نظر را ادا می‌کند[۱۱].

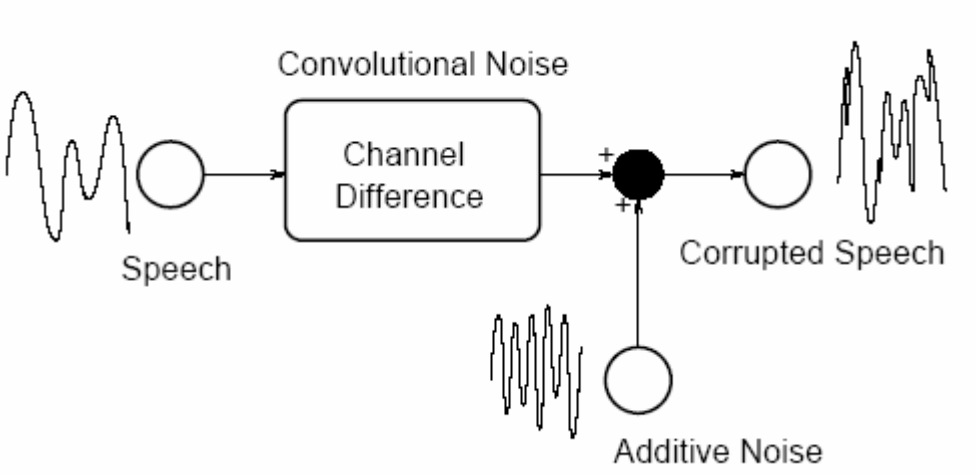
اگر همهٔ منابع نویز از یکدیگر مستقل فرض گردند، می توان گفت:

$$y(t) = \left[\left\{ \left(\left[s(t) \middle| \begin{matrix} \text{Stress} \\ \text{Lombard} \end{matrix} \right]_{n_1(t)} + n_1(t) \right) * h_{\text{mic}}(t) + n_2(t) \right\} * h_{\text{chan}}(t) \right] + n_3(t) \qquad (۱-۴)$$

که در این رابطه، *y(t)* خروجی گفتار که برای بازشناسی به سیستم داده می‌شود و *s(t)* گفتار بدون نویز و بدون اعوجاج که در محیط با نویز زمینهٔ $n_1(t)$ و تکیه و اثر لمبارد آن، بیان شده است. $n_1(t)$ نویز زمینه، $h_{\text{mic}}(t)$ پاسخ ضربهٔ میکروفون، $n_2(t)$ نویز خط انتقال، $h_{\text{chan}}(t)$ پاسخ ضربهٔ خط انتقال و $n_3(t)$ نویز گیرنده می‌باشد. با توجه به پیچیده بودن نحوهٔ اثر و در عین حال کم بودن تاثیر مسالهٔ لمبارد، از اثر این پدیده صرفنظر می‌شود. اگر اثر همهٔ نویزها جمعی در یک منبع و اثر همهٔ نویزهای پیچشی را در منبعی دیگر، خلاصه شوند، رابطهٔ (۴-۲) بدست می‌آید که *n(t)* و *h(t)* به ترتیب اثر کلی نویزجمع‌شونده و اثر کلی نویز پیچشی می‌باشند. رابطهٔ (۴-۲) یه صورت بلوک دیاگرام در شکل ۴-۱ آمده است.

$$y(t) = s(t) * h(t) + n(t) \qquad (۲-۴)$$

1 Lombard



شکل ۴-۱- مدل تأثیر نویز محیط بر گفتار

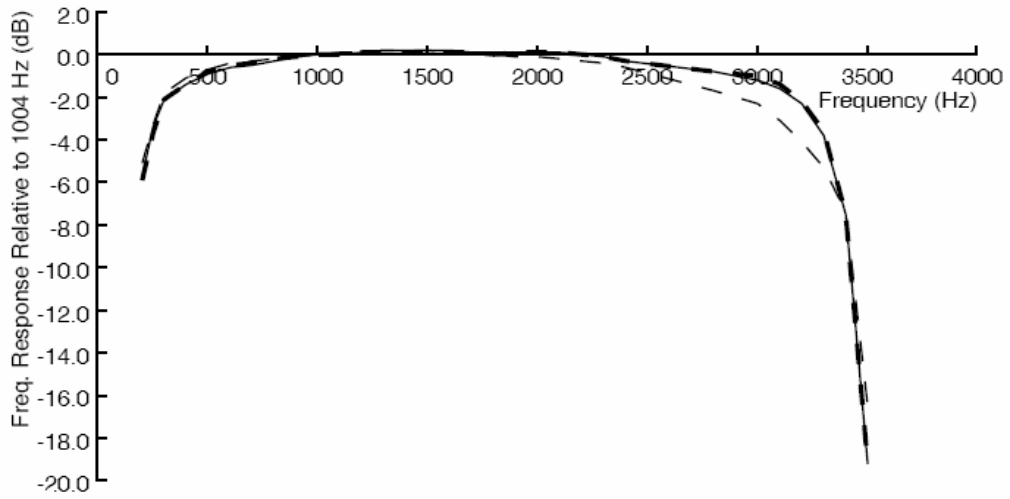
۴-۳- بررسی شبکهٔ خط تلفن

در این قسمت شبکه تلفن با توجه به پاسخ فرکانسی، نسبت سیگنال به نویز و غیر خطی بودن آن مورد بررسی قرار می‌گیرد[۱۲]. اطلاعات موجود در اینجا بیشتر در مورد خط تلفن می‌باشد یعنی ویژگی‌های مربوط به میکروفون تلفن در نظر گرفته نشده است.

۴-۳-۱- پاسخ فرکانسی

پاسخ کانال یا پاسخ فرکانسی، یک اندازه‌گیری برای مقدار داده از بین رفته به علت باند فرکانسی در کانال‌های ارتباطی می‌باشد. شکل ۴-۲ پاسخ کانال در فاصله‌های کوتاه، متوسط و بلند را نشان می‌دهد. به طوری که فاصلهٔ کوتاه خطوط ارتباطی با طول بین صفر تا ۱۸۰ مایل، خطوط متوسط خطوطی با طول ۱۸۰ تا ۲۷۰ مایل و خطوط بلند، خطوطی با طول‌های بالاتر از ۲۷۰ مایل در نظر گرفته شده‌اند. طول‌های بیان

شده، فاصلهٔ نقطه به نقطه بین مکان‌های ارتباطی می‌باشد و فاصلهٔ واقعی که سیگنال طی می‌کند، می‌تواند متفاوت باشد.



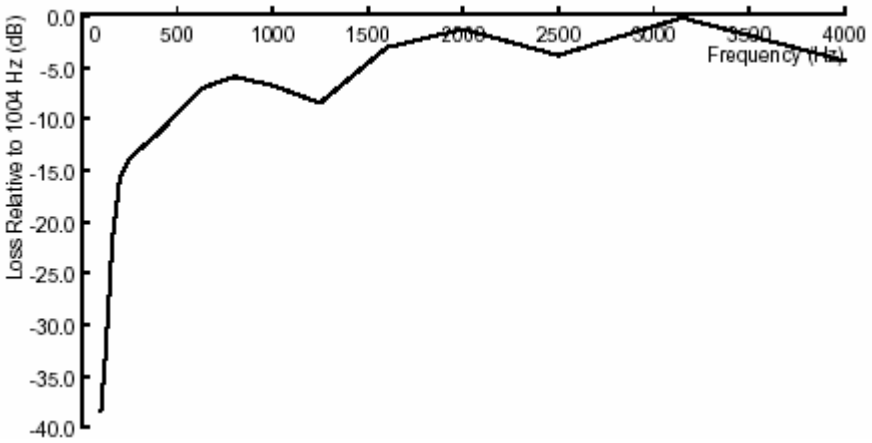
شکل ۴-۲- پاسخ کانال خط تلفن برای فاصله‌های کوتاه(خط بریده)، متوسط(خط بریده پررنگ) و بلند(خط پیوسته)[۱۲]

همانطوری که در شکل فوق مشاهده می‌گردد. بدترین پاسخ فرکانسی مربوط به فاصله‌های کم می‌باشد و این به علت آن است که تیم آقای Garey [۱۳] برای فاصله‌های کوتاه از واسطه‌های ارتباطی آنالوگ استفاده کرده‌اند. در حالی که برای فاصله‌های متوسط و بلند از اتصال‌دهنده‌های T1 دیجیتال استفاده شده است. هنگامیکه اثر میکروفون نیز در نظر گرفته شود پاسخ فرکانسی تقریباً مشابه پاسخ فرکانسی فوق است ولی یک تضعیف بیشتری در فرکانس‌های بالا خواهیم‌داشت.

۴-۳-۲- اثر میکروفون

باید در نظر داشت که در شکل‌های قبلی اثر میکروفون در نظر گرفته نشده است. میکروفون‌های جدید مانند میکروفون‌های خازنی پاسخ فرکانسی ثابت و تقریباً سطحی دارند با وجود این میکروفون‌های

کربنی[1] که هنوز نیز در شبکه‌های تلفن موجود می‌باشد دارای پاسخ فرکانسی غیر مسطح می‌باشند و همچنین اثرات غیر خطی بر روی سیگنال می‌گذارند. این اثرات غیر خطی شامل حساسیت‌های متفاوت نسبت به سطح سیگنال‌های متفاوت و پاسخ‌های متفاوت با توجه به جریان‌های قطبی می‌باشد. یک تابع نوعی برای یک میکروفون کربنی در شکل ۴-۳ نشان داده شده‌است.



شکل ۴-۳- حساسیت پاسخ فرکانسی برای یک میکروفون نوعی[۱۲]

۴-۳-۳- اثرات غیر خطی

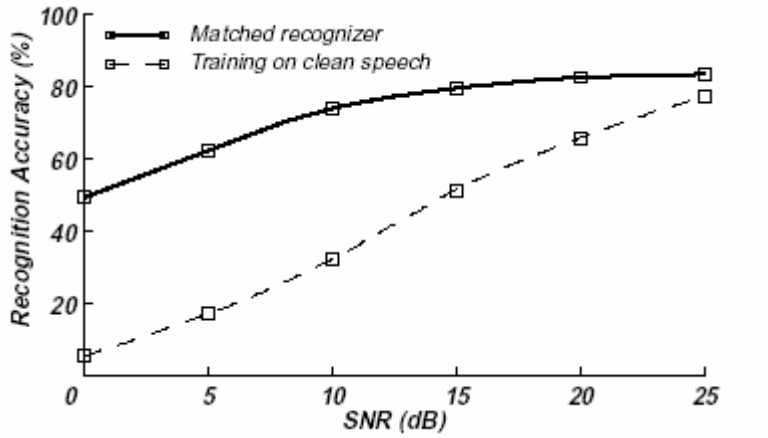
شبکه خط تلفن یک شبکه کاملاً خطی نمی‌باشد. این اثرات غیر خطی معمولاً به عنوان اعوجاج[2] در خط تلفن شناخته می‌شوند. با توجه به مقالات در این زمینه دو نوع اعوجاج در خط تلفن وجود دارد که عبارتند از اعوجاج هارمونیکی[3] و اعوجاج موجود در مدولاسین[4] اولین اعوجاج مربوط به وجود آمدن مؤلفهٔ فرکانس‌های ناخواسته در هارمونیک‌های سیگنال ورودی می‌باشد و دومین اعوجاج هنگامی رخ‌می‌دهد که دو یا چند مؤلفهٔ فرکانسی در کانال موجود باشند و بر یکدیگر اثر بگذارند.

[1] *Carbon Microphone*
[2] *Distortion*
[3] *Harmonic Distortion*
[4] *Modulation Distortion*

۴-۴- کاهش کارایی سیستم‌های بازشناسی گفتار در حضور نویز

هنگامی که سیستم‌های بازشناسی گفتار در محیط‌های نویزی قرار می‌گیرند کارایی آنها به شدت افت می‌کند. سیستم‌های بازشناسی گفتار بر این فرض عمل می‌کنند که نحوهٔ توزیع بردارهای ویژگی در داده‌های مرحلهٔ آزمون بسیار شبیه به بردارهای ویژگی در داده‌های مرحلهٔ آموزش باشد[۱۵].

اگر تشخیص دهنده با گفتاری آموزش داده شود که دقیقاً مشابه داده‌های زمان آزمون باشد، عدم تطابق می‌تواند برطرف شود. البته حتی در این صورت اثر نویز این است که تفاوت ذاتی بین نمونه‌های مختلف گفتار را افزایش می‌دهد و نتیجهٔ بازشناسی دقت کمتری نسبت به زمانی دارد که مراحل آموزش و آزمون هر دو تمیز باشند. شکل ۴-۴ کارایی یک سیستم بازشناسی را برای گفتاری که با نویز تخریب شده است، نشان می‌دهد. همانطور که دیده می‌شود، با افزایش نویز حتی با وجود یکسان بودن توزیع داده‌های مراحل آزمون و آموزش، دقت بازشناسی کاهش یافته است.



شکل ۴-۴- دقت بازشناسی به عنوان تابعی از نسبت سیگنال به نویز گفتار مورد بازشناسی[۱۵]

عواملی را که باعث تنزل کارایی یک سیستم بازشناسی گفتار می‌شوند، می‌توان به دو دستهٔ کلی عوامل مربوط به گوینده و عوامل محیطی تقسیم کرد.

عوامل مربوط به گوینده

در یک گویندهٔ خاص عواملی مانند تغییر حالت و سرعت گفتار، تغییر صدای گوینده در حضور نویز محیط (اثر لومبارد[1])، تأثیر منفی قابل توجهی در دقت سیستم دارد. همهٔ این عوامل باعث می‌شود یک نفر به سختی بتواند یک جملهٔ خاص را دوبار مثل هم بیان کند که این امر بازشناسی را مشکل می‌کند. از طرف دیگر، الگوی صحبت یک فرد می‌تواند کاملاً با الگوی صحبت فرد دیگر متفاوت باشد. وجود لهجه‌های مختلف، تفاوت در تلفظ، سرعت بیان گفتار، متفاوت بودن اندام تولید گفتار به عنوان مثال اندازه و شکل مجرای صوتی، تفاوت در سن، جنسیت و حتی تحصیلات بین گوینده‌های مختلف نیز سبب پیچیدگی بازشناسی و در نتیجه کاهش کارایی می‌شود. معمولاً برای مدل کردن واحدهای صوتی برای یک سیستم بازشناسی مستقل از گوینده، از گفتار چند صد گوینده استفاده می‌شود تا دقت بالایی برای گوینده‌هایی جدیدی که سیستم با آن‌ها آموزش ندیده است، داشته باشد. ولی با این حال خطای بازشناسی برای گویندهٔ با لهجهٔ متفاوت در حدود ۲ یا ۳ برابر افزایش می‌یابد [۱].

عوامل محیطی

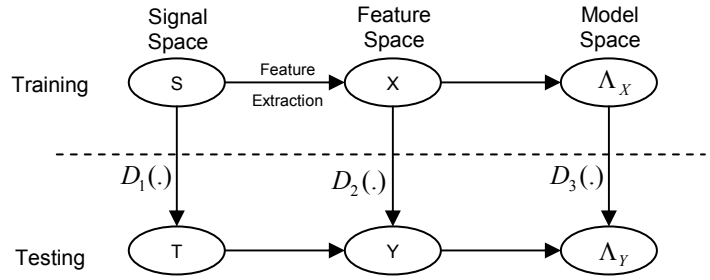
محیط پیرامون ما پر از صداهای گوناگون با بلندی‌های متفاوت است. وقتی با یک سیستم بازشناسی گفتار صحبت می‌کنیم ممکن است افرادی در حال حرف زدن باشند (نویز همهمه[2])، شیئی جابجا شود (نویز ضربه‌ای[3])، صدای دستگاهی شنیده شود یا طیف صدا به خاطر حرکت گوینده تغییر کند. در کاربردهایی مثل وسیلهٔ نقلیه، نویز پس‌زمینه به شدت در حال تغییر می‌باشد. از پارامترهای مهم دیگر می‌توان به نوع و محل قرار گرفتن میکروفون، فاصله دهان تا میکروفون، رسانهٔ انتقال، پهنای باند محدود، نویز میکروفون و A/D[4]،

[1] *Lombard Effect*
[2] *Babble Noise*
[3] *Spike Noise*
[4] *Analog to Digital Converter*

پژواک صدا[1] در محیط و همچنین نویزهایی که توسط خود گوینده تولید می‌شوند (مثل صدای بازدم)، اشاره کرد. به طور کلی چگونگی رویارویی با تغییرات محیط یکی از مهمترین و پیچیده‌ترین مسائل مطرح در بازشناسی گفتار به وسیلهٔ ماشین می‌باشد.

۴-۵- معرفی انواع روش‌های بازشناسی مقاوم گفتار

مسألهٔ تطبیق نویز را در فضاهای مختلف می توان مورد بررسی قرار داد. این فضاها در شکل ۴-۴ به نمایش درآمده‌است.



شکل۴-۵- فضاهای ممکن برای عدم انطباق بین داده‌های مرحلهٔ آموزش و مرحلهٔ آزمایش

فضای اول، خود سیگنال گفتار آموزش یا آزمایش می‌باشد. در فضای سیگنال، سیگنال مرحله آموزش،S و سیگنال مرحلهٔ آزمایش، T می‌باشد. یک تابع $D_1(.)$، عدم انطابق بین S و T را بیان می‌دارد. سیگنال‌ها در ادامه به بردارهای ویژگی تبدیل می‌شوند. در مرحلهٔ آموزش، بردار X و در مرحله آزمایش بردارY را داریم. عدم انطباق بین این دو با $D_2(.)$ بیان می‌شود. از آنجا که یکی از معمول‌ترین روش‌های بازشناسی، استفاده از مدل مخفی مارکوف می‌باشد، در ادامه سیستم از روی بردارهای ویژگی، مدل‌های λ_x و λ_y را بدست می‌آورد. عدم انطباق بین دو مدل اخیر، با تابع $D_3(.)$ بیان شده است[۱۴].

[1] *Reverberation*

با توجه به توضیحات فوق به طور کلی می‌توان روش‌های مقاوم نمودن سیستم‌های بازشناسی گفتار را در مقابل نویز به سه دستهٔ مهم زیر تقسیم بندی کرد[۱۲]:

- دسته اول از روش‌ها، آن‌هایی هستند که سیگنال گفتار را به ویژگی‌های ذاتاً مقاوم در مقابل نویز تبدیل می‌کنند. در این روش‌ها، بعد از استخراج ویژگی‌های سیگنال، پردازش اضافی دیگری مورد نیاز نمی‌باشد.
- در دستهٔ دوم، سیگنال گفتار عاری از نویز از روی سیگنال نویزی بدست می‌آید. سیگنال تمیز به‌دست آمده، در ادامه به یک سیستم بازشناسی معمولی داده می‌شود. در واقع در این دسته، از روش‌های بهبود گفتار برای بازشناسی بهره برده می‌شود.
- در روش‌های دستهٔ سوم، سعی بر آن است که مدل بازشناسی نسبت به نویز مقاوم شود. در این دسته، روش‌هایی مانند روش‌های بازگشت خطی با بیشترین شباهت[۱]، بیشترین احتمال پسین[۲] و ژاکوبین[۳] وجود دارند.

[1] *Maximum Likelihood Linear Regression (MLLR)*
[2] *Maximum A Posteriori (MAP)*
[3] *Jacobian*

فصل پنجم

ویژگی‌های مقاوم

۵-۱- مقدمه

همانطور که قبلاً بیان شد استفاده از ویژگی‌های مقاوم یکی از راه‌های بالا بردن دقت سیستم‌های بازشناسی گفتار می‌باشد. روش‌هایی که از ویژگی‌های مقاوم استفاده می‌کنند، قسمت استخراج ویژگی را در هر دو مرحلهٔ آموزش و آزمون تغییر می‌دهند. در این دسته از روش‌ها سعی می‌شود با توجه به تأثیر نویز بر خود سیگنال ویا ویژگی‌های آن، ویژگی‌هایی از سیگنال گفتار استخراج شود که نویز بر آن‌ها بی‌اثر یا کم‌اثر باشد. با این کار بردارهای ویژگی مشابهی برای نمونه‌های گفتار آموزشی و آزمایشی به دست آید و در نتیجه سیگنال گفتار به پارامترهایی ذاتاً مقاوم در مقابل نویز تبدیل می‌شود. از این رو در این روش‌ها، بعد از پارامتری کردن سیگنال، پردازش اضافی دیگری مورد نیاز نمی‌باشد [۱۶]. بدین منظور تأثیر محیط بر مشخصه‌های سیگنال گفتار را از طریق تأثیر آن‌ها بر خود سیگنال ویا ویژگی‌ها مشخص می‌نمایند. در بعضی از این روش‌ها به هر دو نوع سیگنال گفتار تمیز و نویزی نیاز می‌باشد.

در این فصل به ارائهٔ بعضی از پرکاربردترین ویژگی‌های مقاوم پرداخته می‌شود. در انتهای فصل نیز روشی نوین مبتنی بر فیزیولوژی گوش انسان ارائه خواهد شد.

۵-۲- نرمال کردن در حوزهٔ کپسترال

یکی از روش‌های مقاوم کردن ویژگی‌ها در سیستم‌های بازشناسی گفتار نرمال کردن ضرایب کپسترال می‌باشد. نرمال کردن این ضرایب می‌تواند به شکل‌های گوناگون انجام شود که در ادامه به بیان آن‌ها می‌پردازیم.

۵-۲-۱- نرمال کردن میانگین کپسترال[1] یا تفاضل میانگین کپسترال[2]

یکی از روش‌های قدیمی و معمول برای مقاوم کردن ضرایب کپسترال در مقابل نویز کانال و تغییرات کانال در سیستم‌های بازشناسی گفتار، تفاضل میانگین ضرایب کپسترال می‌باشد. در این روش، پس از تبدیل فریم‌های گفتار به ضرایب کپسترال، در یک پردازش اضافه، مقدار میانگین آن‌ها از تمام بردارهای کپسترال کم می‌شود. بردار ویژگی تفاضل میانگین ضرایب کپسترال از روی N بردار ضرایب کپسترال $\vec{C}_{y;i},\ i=1,\cdots,N$ و گفتار عبور کرده از کانال $y(n)$ با کم‌کردن میانگین تمام N بردار کپسترال از هر یک از بردارهای اصلی کپسترال $\vec{C}_{y,i}$ بدست می‌آید:

$$\vec{C}_{cms;i}=\vec{C}_{y;i}-\vec{C}_{y;avg},\ i=1,\cdots,N \qquad (۱-۵)$$

که $\vec{C}_{y;avg}=\frac{1}{N}\sum_{i=1}^{N}\vec{C}_{y;i}$. اصول این روش بر اساس رفتار ضرایب کپسترال در هنگام مواجه با نویز پیچشی است و فرض اصلی این روش این است که فیلتر کانال $h(n)$ در زمان ادای گفتار تغییرات چشمگیری نداشته باشد. یعنی اینکه $h(n)$ یک فیلتر خطی و تغییرناپذیر در زمان است. این فرض معمولاً برای گفتار ادا شده در طول چندین ثانیه قابل قبول است و در بسیاری از سیستم‌های بازشناسی گفتار معمول می‌باشد.

با توجه به اینکه نویز کانوالو شده در حوزهٔ زمان، مانند اعوجاج ایجاد شده توسط کانال، معادل نویز جمعی در حوزهٔ کپسترال می‌باشد، اگر سیگنال گفتار تمیز و قبل از عبور از کانال $h(n)$ با $s(n)$ و سیگنال عبور کرده از کانال با $y(n)$ بیان شود آنگاه:

$$y(n)=s(n)*h(n)\Leftrightarrow\vec{C}_y=\vec{C}_s+\vec{C}_h \qquad (۲-۵)$$

[1] *Cepstral Mean Normalization (CMN)*
[2] *Cepstral Mean Subtraction (CMS)*

که * عمل کانولوشن خطی و $\vec{C}_y$، $\vec{C}_s$ و $\vec{C}_h$ به ترتیب ضرایب کپسترال مرتبط را بیان می‌کنند. اگر از طرفین رابطه (۵-۲) میانگین گرفته شود خواهیم داشت:

$$\begin{aligned}\vec{C}_{y;avg} &= \frac{1}{N}\sum_{i=1}^{N}\vec{C}_{y;i} \\ &= \frac{1}{N}\sum_{i=1}^{N}\vec{C}_{s;i} + \frac{1}{N}\sum_{i=1}^{N}\vec{C}_{h;i} \\ &= \vec{C}_{s;avg} + \vec{C}_{h;avg}\end{aligned} \qquad (۵-۳)$$

از آنجایی که فرض شده است که کانـال در زمان بیان گفتـار تغییر نمی‌کند، مجموع $\frac{1}{N}\sum_{i=1}^{N}\vec{C}_{h;i}$ در رابطه (۵-۳) به صورت $\frac{1}{N}\sum_{i=1}^{N}\vec{C}_h$ یا فقط $\vec{C}_h$ که کپسترال کانال است، تغییر می‌کند. مجموع $\frac{1}{N}\sum_{i=1}^{N}\vec{C}_{s;i}$ به میانگین کپسترال سیگنال گفتار تمیز مربوط می‌شود. اگر تغییرات و اعوجاج صداها در $s(n)$ به گونه‌ای باشد که میانگین طیف گفتار نسبتاً مسطح باشد آنگاه بردار میانگین کپسترال مرتبط به سمت صفر میل خواهد کرد، یعنی:

$$\vec{C}_{s;avg} \Rightarrow \vec{0} \qquad (۵-۴)$$

میانگین کپسترال گفتار عبور کرده از کانال $y(n)$ که در رابطه (۵-۳) بیان شد، به صورت زیر تغییر خواهد کرد:

$$\begin{aligned}\vec{C}_{y;avg} &= \vec{C}_{s;avg} + \vec{C}_{h;avg} \\ &= \vec{0} + \vec{C}_h \\ &= \vec{C}_h\end{aligned} \qquad (۵-۵)$$

که همان کپسترال کانال می‌باشد. بنابراین کم کردن میانگین زمانی ضرایب کپسترال از هر یک از فریم‌ها که در رابطه (۵-۱) بیان شد معادل کم کردن کپسترال کانال می‌باشد. می‌توان رابطه (۵-۱) را به صورت زیر بازنویسی کرد:

$$\begin{aligned}\vec{C}_{cms;i} &= \vec{C}_{y;i} - \vec{C}_{y;avg} \\ &= (\vec{C}_{s;i} + \vec{C}_h) - \vec{C}_h \\ &= \vec{C}_{s;i}\end{aligned} \qquad (۶-۵)$$

بردار کپسترال CMS حاصل، $\vec{C}_{cms;i}$ ،دیگر به بردار $\vec{C}_h$ وابسته نیست و بنابراین تغییر ناپذیر در هر یک از کانال‌ها می‌باشد. روابط بالا همچنین با توجه به این نکته که کم کردن بردار $\vec{C}_h$ در حوزهٔ کپسترال معادل با دیکانولوشن با $h(n)$ در حوزهٔ زمان یا ضرب $\frac{1}{\left|H(e^{jw})\right|}$ در حوزهٔ فرکانس است، قابل محاسبه هستند.

از آنجایی که همیشه نمی‌توان انتظار داشت که بردار میانگین ضرایب کپسترال گفتار صفر شود، در بسیاری از سیستم‌ها می‌توان فرض کرد که بردار میانگین ضرایب کپسترال در سیگنال گفتار مقداری ثابت باشد، یعنی:

$$\vec{C}_{s;avg} = \vec{C} \qquad (۷-۵)$$

و بنابراین (۵-۵) را می‌توان به صورت زیر بازنویسی کرد:

$$\begin{aligned}\vec{C}_{y;avg} &= \vec{C}_{s;avg} + \vec{C}_{h;avg} \\ &= \vec{C} + \vec{C}_h\end{aligned} \qquad (۸-۵)$$

و رابطه (۵-۶) به صورت زیر تغییر می‌کند:

$$\begin{aligned}\vec{C}_{cms;i} &= \vec{C}_{y;i} - \vec{C}_{y;avg} \\ &= (\vec{C}_{s;i} + \vec{C}_h) - (\vec{C} + \vec{C}_h) \\ &= \vec{C}_{s;i} - \vec{C}\end{aligned} \qquad (۹-۵)$$

و بنابراین اگر سیستم در مرحله آموزش به جای، $\vec{C}_{s;i}$ با $\vec{C}_{s;i} - \vec{C}$ آموزش داده شود و در مرحله آزمایش از $\vec{C}_{cms;i}$ استفاده شود، در واقع اثر نویز از بین رفته است.

بردار جبران در این روش، تفاضل بردارهای ضرایب کپسترال مرحلهٔ آموزش و آزمون می‌باشد.

$$\vec{V}_{c;i} = \vec{C}_{s;i} - \vec{C}_{y;i} \qquad (۱۰-۵)$$

به این معنی که انجام عمل تفاضل میانگین ضرایب کپسترال معادل با آن است که بردار با داده‌های مرحلهٔ آموزش و آزمون جمع شود.

۵-۲-۲- نرمال کردن میانگین و واریانس[1]

می‌توان نرمال‌سازی را بدین صورت انجام داد که علاوه بر کم شدن میانگین کل بردارهای ویژگی از هر بردار ویژگی، حاصل را بر واریانس کل نیز تقسیم کرد. با این کار میانگین کل بردارها صفر و واریانس آن‌ها یک خواهد شد[۱۷]. در این حالت کم کردن میانگین همانند یک فیلتر بالاگذر خطی و تقسیم بر واریانس به کنترل بهره خودکار[2] تعبیر می‌شود.

۵-۳- ضرایب کپسترال ریشه‌ای[3]

فریم‌های گفتار می‌توانند به وسیلهٔ کانولوشن فیلتر مسیر صوتی با یک رشتهٔ تحریک کننده، که تشکیل شده است از پالس‌های متناوب برای گفتار صدادار و نویز برای گفتار بی‌صدا، مدل شوند. تابع انتقال مسیر صوتی به وسیلهٔ یک فیلتر تمام قطب مدل می‌شود. بعدها مدل تمام قطبی از گفتار نویزی در [۱۸] ارائه شد. واضح است که نویز در طیف سیگنال صفرهایی را به وجود می‌آورد. یکی از راه‌هایی که بتوان این سیستم را تحلیل کرد آن است که این سیستم به دامنهٔ خطی نگاشت شود به طوری که مولفه‌های تحریک کننده فیلتر گردند و فقط پاسخ مسیر صوتی باقی بماند. این کار به وسیلهٔ تبدیل‌های هوموورفیک مانند عملگر لگاریتم انجام می‌شود. Lim در [۱۹] تقریبی از دیکانوالو لگاریتمی به وسیلهٔ تابع ریشه بیان کرد به طوری که $\gamma(\cdot)$ و

[1] *Mean Variance Normalization (MVN)*
[2] *Automatic Gain Control (AGC)*
[3] *Root Cepstral Coefficients*

$(\cdot)^{1/\gamma}$ به جای عملگر لگاریتمی و عملگر نمایی مورد استفاده قرار می‌گیرند و کپسترال‌های لگاریتمی به عنوان حالت خاصی از کپسترال‌های ریشه‌ای بیان شدند. ضرایب کپسترال لگاریتمی در شرایط گفتار تمیز خیلی خوب عمل می‌نمایند ولی در هنگام نویزی شدن ورودی، عملکردشان چندان مطلوب نمی‌باشد. یکی ار راه‌حل‌هایی که برای استفاده از ضرایب کپسترال در محیط‌های نویزی به‌کار گرفته شده، استفاده از تابع ریشه بجای لگاریتم در محاسبه این ضرایب می‌باشد تا دیکانوالو هوموورفیک‌تری بدست آید. ضرایب RCC در مقایسه با ضرایب MFCC علاوه بر اینکه نسبت به نویز مقاوم‌تر می‌باشند در عمل سریعتر نیز محاسبه می‌گردند[۱۱]. برای محاسبهٔ ضرایب RCC ابتدا انرژی فیلتر بانک‌ها محاسبه می‌گردند:

$$e[j] = \sum_{k=0}^{N-1} w_j[k] \cdot \left|\widetilde{S}[k]\right|, \qquad \text{for j} = 1, 2, \ldots, \text{P} \tag{۱۱-۵}$$

در این رابطه $e[j]$ بیان کنندهٔ انرژی فیلتر jام، $w_j[k]$ وزن فیلتر مثلثی jام در kامین باند فرکانسی، $\left|\widetilde{S}[k]\right|$ اندازهٔ طیف حاصل از DFT در مقیاس مل و P تعداد فیلتر بانک‌ها است. در نهایت طبق رابطهٔ (۵-۱۲) تعداد m ضریب کپسترال ریشه‌ای استخراج می‌گردد:

$$RCC[i] = \sum_{\text{j}=1}^{P} \left\{ (e[j])^{\gamma} \cos\left(\frac{i(j-0.5)}{P} \pi \right) \right\}, \qquad \text{for i} = 1, 2, \ldots, \text{m} \tag{۱۲-۵}$$

که $0 < \gamma \leq 1$. γ ضریب ریشه است که برای فشرده کردن انرژی فیلتر بانک‌ها مورد استفاده قرار می‌گیرد.

۵-۴- وزن‌دهی ضرایب کپسترال یا اعمال لیفتر

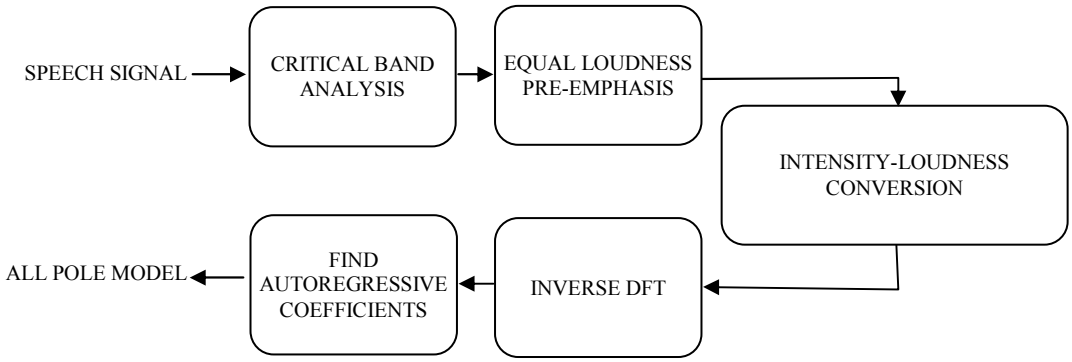
ضرایب بردار کپسترال نسبت به نویز حساسیت یکسانی ندارند[۴۹]. استفاده از لیفترها در بازشناسی مقاوم گفتار با این ایده انجام می‌شود که به ویژگی‌هایی که نسبت به نویز حساسیت بیشتری دارند، وزن

کمتری داده شود. جدای از مسأله نویز، لیفترها در بازشناسی گفتار نیز بکار رفته‌اند به این صورت که به ضرایبی که واریانس بیشتری دارند، وزن کمتری داده می‌شود.

با وجود اینکه ضرایب کپسترال دارای حساسیت یکسانی نسبت به نویز نیستند ولی استفاده از لیفتر در مورد سیستم‌های مبتنی بر مدل مخفی مارکف هیچ تأثیری در بازشناسی ندارد[۴۹].

۵-۵- پیشگویی خطی مبتنی بر درک انسان[1]

در این روش استخراج ویژگی هم از بانک فیلتر و هم از آنالیز پیشگویی خطی استفاده می‌شود. این روش از این ایده بهره می‌گیرد که اولاً قدرت تفکیک فرکانس و حساسیت نسبت به تغییر فرکانس در گوش انسان در فرکانس‌های مختلف متفاوت است و ثانیاً حساسیت گوش نسبت به شدت و انرژی در فرکانس‌های مختلف متفاوت است. بلوک دیاگرام استخراج ضرایب PLP در شکل ۵-۱ آمده است.



شکل ۵-۱- بلوک دیاگرام کلی استخراج ویژگی‌های PLP [۲۰]

در اولین قدم پس از اعمال پنجرهٔ طیف توان سیگنال گفتار به دست می‌آید. سپس یک سری از فیلترهای ذوزنقه‌ای یا مثلثی برروی باند فرکانسی که مطابق معیار بارک توزیع شده‌اند به طیف توان اعمال می‌شوند و انرژی طیف در باندهای مختلف بدست می‌آید. اگر تعداد فیلترها F باشد از این مرحله F مقدار

[1] *Perceptual Linear Prediction (PLP)*

به دست می‌آید که انرژی طیف توان در باندهای فرکانسی مختلف را نشان می‌دهند. این F عدد سپس تحت تاثیر منحنی برابری بلندی صوت[1] قرار می‌گیرند و وزن دهی می‌شوند. پس از این مرحله قانون ریشهٔ سوم شنوایی[2] به این F عدد اعمال شده و به توان ۰/۳۳ می‌رسند. سپس با داشتن این F عدد و اضافه کردن F عدد مشابه به صورت آینه‌ای، عکس تبدیل فوریه این رشته از اعداد یعنی 2F عدد به دست می‌آید و در حوزه زمان 2F عدد بدست می‌آید. با استفاده از این رشتهٔ زمانی این سیگنال را با مدل تمام قطب از مرتبهٔ P مدل می‌کنیم و تعداد P عدد از ضرایب پیشگویی خطی مبتنی بر درک انسان به دست می آید. پس از این مرحله P ضریب را به N ضریب کپسترال تبدیل می کنیم. به دست آوردن ضرایب پیشگویی خطی مبتنی بر دراک انسان و نیز ضرایب کپسترال مانند به دست آوردن ضرایب LPC می‌باشد.

همان طور که قلاً نیز ذکر شد، قدرت تفکیک فرکانس و حساسیت نسبت به تغییر فرکانس در فرکانس‌های مختلف در گوش متفاوت می‌باشد از این رو در آنالیز PLP ، فیلترها در بانک فیلتری مطابق با ادراک انسان بر روی توان طیف بدست آمده توزیع می‌شوند به طوریکه پهنای باند فیلترها در فرکانس‌های بالاتر، بیشتر می‌باشد. اگر فرکانس منطبق با ادراک انسان در آنالیز PLP را Ω بنامیم واحد این فرکانس بارک است و با فرکانس هرتز به صورت زیر مربوط می‌شود.

$$\Omega(\omega) = 6\ln\{(\omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{0.5}\} \qquad (۱۳-۵)$$

در روابطهٔ فوق ω فرکانس زاویه‌ای بر حسب رادیان بر ثانیه می‌باشد. اگر فرض کنیم که فرکانس نمونه برداری برابر ۱۰ کیلو هرتز باشد، آنگاه محدودهٔ Ω تقریباً از صفر تا ۱۷ بارک تغییر می‌کند و با فرض اینکه ۱۷ فیلتر را روی محور بارک بطور یکسان توزیع کنیم آنگاه پهنای باند هر فیلتر یک بارک می‌شود.

سپس طیف توان توزیع شده حاصل با منحنی ماسک کنندهٔ $\Psi(\Omega)$، کانوالو می‌شود. اگر طیف توان فریم $P(\omega)$ باشد، آنگاه خروجی فیلتر i ام به صورت زیر خواهد بود:

[1] *Equal Loudness Curve*
[2] *Power Low of Hearing*

$$\theta(\Omega_i) = \sum_{\Omega=-13}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \qquad (۱۴-۵)$$

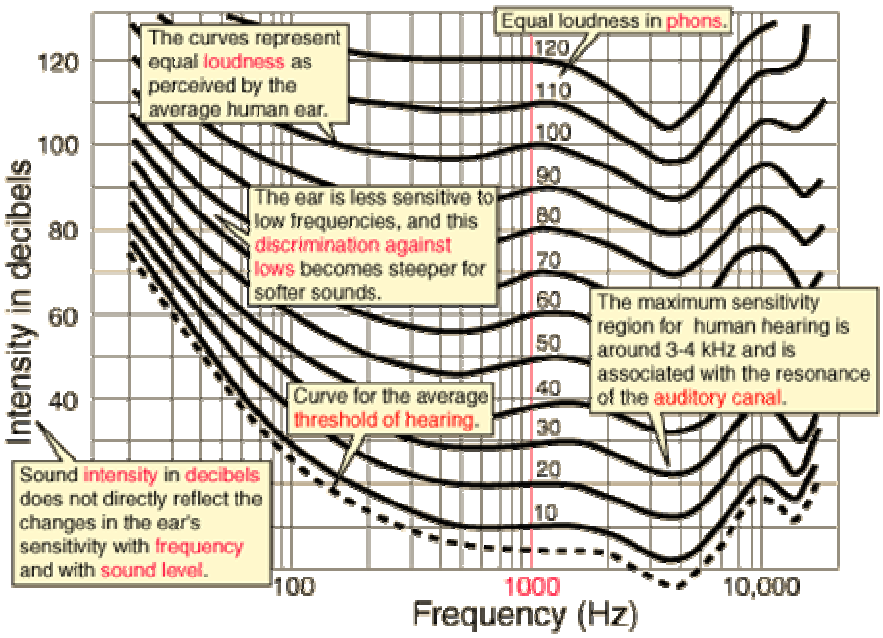
که در آن، فیلتر $\Psi(\Omega)$، به صورت رابطه زیر می‌باشد:

$$\Psi(\Omega) = \begin{cases} 0 & for \quad \Omega < 1.3 \\ 10^{2.5(\Omega+0.5)} & for \quad -1.3 \le \Omega \le -0.5 \\ 1 & for \quad -0.5 < \Omega < 0.5 \\ 10^{-(\Omega-0.5)} & for \quad 0.5 \le \Omega \le 2.5 \\ 0 & for \quad \Omega > 2.5 \end{cases} \qquad (۱۵-۵)$$

خروجی بدست آمده، سپس با استفاده از منحنی برابری بلندی صوت پیش تأکید می‌شود:

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \qquad (۱۶-۵)$$

همانطور که بیان شد اعمال این منحنی بدان علت است که حساسیت گوش انسان نسبت به شدت صوت در فرکانس‌های متخلف متفاوت است. حساسیت گوش انسان در فرکانس‌های مختلف در شکل ۵-۲ نشان داده شده است:



شکل ۵-۲- حساسیت گوش انسان در شدت صوت‌های مختلف(منحنی برابری بلندی صوت)

همان طور که مشاهده می‌گردد گوش انسان در محدودهٔ اطراف چهار کیلو هرتز بیشترین حساسیت را دارد. برای شبیه‌سازی عملکرد گوش انسان باید در جاهایی که حساسیت گوش انسان کم است، فیلتر ورودی را تضعیف کند. به عنوان مثال $E(\omega)$ می‌تواند طبق یکی از روابط زیر باشد:

$$E(\omega) = 1.151\sqrt{\frac{(\omega^2 + 144\times10^4)\omega^2}{(\omega^2 + 16\times10^4)(\omega^2 + 961\times10^4)}} \qquad (۱۷-۵)$$

$$E(\omega) = \frac{\omega^4 \times (\omega^2 + 56.8\times10^6)}{(\omega^2 + 6.3\times10^6)^2 \times (\omega^2 + 0.38\times10^9)} \qquad (۱۸-۵)$$

از آنجایی که در گوش انسان میزان احساس بلندی صوت با ریشه سوم انرژی آن متناسب است، برای تبدیل انرژی یا توان صوت، به بلندی صوت در گوش انسان انرژی به توان ۰/۳۳ رسانده می‌شود.

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \qquad (۱۹-۵)$$

پس از این مرحله همان طور که گفته شد، عکس تبدیل فوریهٔ رشته $\Phi_1,\ldots\Phi_f,\Phi_f,\ldots,\Phi_1$ محاسبه می‌گردد و یک رشتهٔ زمانی بدست می‌آید ودر انتها با استفاده از یک مدل تمام قطب مرتبهٔ P ضرایب پیشگویی خطی مبتنی بر درک انسان محاسبه می‌شوند.

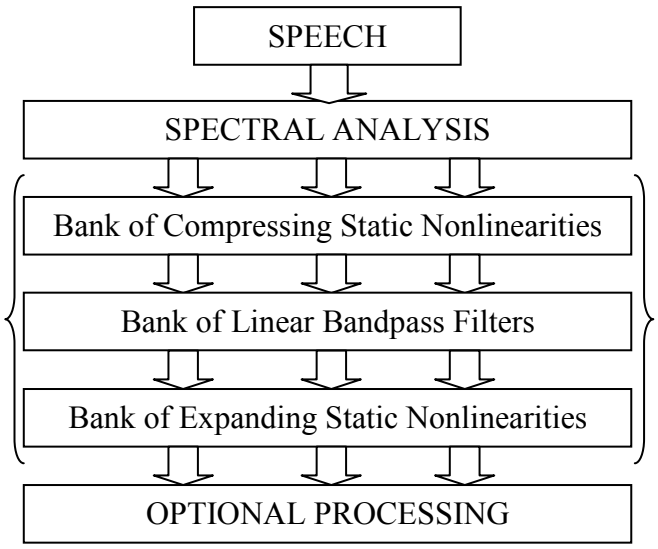
۵-۶- تحلیل طیف نسبی[1]

در آنالیز سیگنال گفتار عاملی که برای ما مهم است، علاوه بر مشخصات سیگنال تحریک، نحوهٔ تغییر شکل مجرای گفتار در طول زمان است. سیگنال گفتار علاوه بر اطلاعات زبان شناسی ممکن است شامل اطلاعات دیگری نیز باشد که برای سیستم‌های بازشناسی مفید نمی‌باشند. به عنوان مثال مشخصات یک نوع میکروفون یک جزء پیچشی در سیگنال گفتار است که در حوزهٔ طیف لگاریتمی یک عنصر جمع‌شونده با

[1] *RelAtive SpecTral Analysis(RASTA)*

طیف لگاریتمی سیگنال گفتار محسوب می‌شود و یا مشخصات طیفی یک کانال انتقال اغلب یا در طول زمان ثابت است و یا به کندی تغییر می‌کند و این مشخصه در شکل لگاریتمی، به صورت یک عنصر جمع‌شونده با تغییرات آهستهٔ زمانی در حوزهٔ طیف لگاریتمی ظاهر می‌گردد. تحلیل طیف نسبی با این هدف انجام می‌شود که تغییرات آهسته و یا شدیدی که خارج از محدودهٔ اطلاعات زبان شناسی گفتار قرار می‌گیرند را از سیگنال گفتار حذف کند. اعمال فیلتر RASTA که حوزهٔ عمل آن لگاریتم طیف می‌باشد، می‌تواند کارایی یک سیستم بازشناسی گفتار را هم در حضور نویز جمعی و هم در حضور نویز پیچشی افزایش دهد. فیلتر RASTA از جهاتی بر اساس عملکرد گوش انسان عمل می‌کند. گوش انسان به مؤلفه‌های فرکانسی که شدت آن‌ها در طول زمان با فرکانس حدود چهار کیلو هرتز تغییر می‌کند در مقایسه با مؤلفه‌هایی که شدت آن‌ها در طول زمان سریعتر یا آهسته‌تر از این مقدار تغییر می‌کند، حساسیت بیشتری دارد. این فیلتر نیز سعی می‌کند که این تغییرات آهسته یا شدیدی را که گوش انسان به آن‌ها حساسیت کمتری دارد حذف کند.

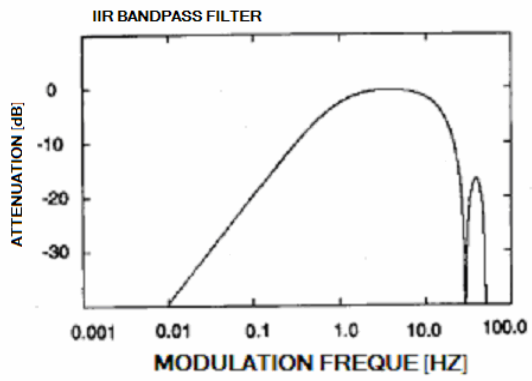
بلوک دیاگرام کلی تحلیل طیف نسبی در شکل ۵-۳ آمده است. همان‌طور که در این شکل نشان داده شده است ابتدا توان طیف سیگنال گفتاربه دست می‌آید و سپس اندازهٔ مولفه‌های بدست آمده با یک تبدیل غیرخطی فشرده کننده مانند لگاریتم فشرده می‌شوند، پس از آن دنباله زمانی اندازهٔ هر یک از مولفه‌ها در طول زمان، در حوزه فشرده شده، با فیلتر RASTA فیلتر می‌شوند و در انتها دنبالهٔ زمانی فیلتر شدهٔ هر یک مولفه‌ها تحت یک تبدیل غیر خطی منبسط کننده مانند تبدیل نمایی منبسط می‌گردند.



شکل ۵-۳- بلوک دیاگرام کلی تحلیل طیف نسبی [۲۱]

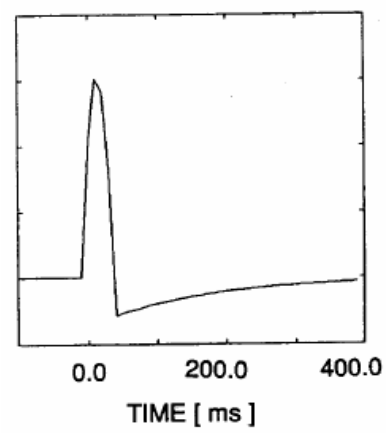
این عمل فشرده‌سازی طیف، فیلترکردن دنبالهٔ زمانی مولفه‌ها و سپس منبسط‌کردن طیف، تحلیل طیف نسبی نامیده می‌شود. پس از تحلیل طیف نسبی هر آنالیز دیگری مانند PLP یا MFCC می‌تواند انجام شود. اگر ضرایب PLP پس از تحلیل طیف نسبی به دست آیند، ضرایب حاصل را RASTA-PLP می‌نامند و به طریق مشابه ضرایب RASTA-MFCC نیز می توانند به دست آیند.

فیلتر RASTA یک فیلتر میان گذر است که می‌تواند پاسخ فرکانسی مانند شکل ۵-۴ داشته باشد.



شکل ۵-۴- پاسخ فرکانسی فیلتر میان‌گذر RASTA [۲۲]

فیلتر نشان داده شد، یک فیلتر IIR است که پاسخ ضربهٔ آن نیز در شکل ۵-۵ آمده است.



شکل ۵-۵- پاسخ ضربهٔ فیلتر میان‌گذر RASTA [۲۲]

اگر فیلتر RASTA مطابق شکل‌های ۵-۴ و ۵-۵ باشد دارای تابع تبدیلی به صورت زیر است:

$$H(z) = \frac{0.2 + 0.1Z^{-1} - 0.1Z^{-3} - 0.2Z^{-4}}{1 - 0.94Z^{-1}} \qquad (۲۰-۵)$$

تحلیل طیف نسبی را به دو صورت می‌توان استفاده کرد: یکی لگاریتمی یا Log RASTA و دیگری خطی-لگاریتمی یا Lin-Log RASTA (J-RASTA). تبدیل‌های فشرده کننده و منبسط کننده در Log RASTA بصورت $y = In\ x$ و $x = e^y$ می‌باشند و در J-RASTA این تبدیل‌ها به صورت $y = \ln(1 + Jx)$ و $x = \frac{e^y - 1}{J}$ می‌باشند. چون مقدار y فیلتر شده است دلیلی ندارد که حتماً x مثبت باشد. برای رفع این مشکل یک تبدیل معکوس به طور تقریبی و به صورت $x = \frac{e^y}{J}$ استفاده می‌گردد تا همیشه مقادیر x مثبت باشد. تبدیل J-RASTA به ازای مقادیر کوچک طیف خطی عمل می‌کند و به ازای مقادیر بزرگ طیف لگاریتمی عمل می‌نماید. مقدار بهینه J برای محیط‌های مختلف را می توان از رابطهٔ زیر به دست آورد:

$$J_{opt} = \frac{1}{C_{train}.E_{noise}} \qquad (۲۱-۵)$$

که C_{train} در رابطهٔ فوق به مجموعهٔ آموزشی بستگی دارد و برای محیط‌های مختلف با SNRهای مختلف ثابت است و E_{noise} انرژی میانگین نویز در آن محیط خاص با SNR خاص می‌باشد.

فیلتر RASTA دارای این مزیت نسبت به فیلترهای CMS است که پاسخ ضربهٔ آن با طول کمتری می‌باشد. بسیاری از نویزها و عوامل مزاحم ناشناخته در محیط ممکن است دارای تغییرات آهسته در طول زمان و یا تغییرات شدید در طول زمان باشند که فیلتر RASTA بر اساس سیستم شنوایی انسان این تغییرات را که در محدودهٔ مفید برای گوش انسان نیستند، حذف می‌کند.

فیلتر RASTA دارای مشخصات فاز هموار نیست و دنبالهٔ زمانی مولفه‌های فرکانسی با فرکانس‌های مدولاسیون مختلف را دچار تأخیر فازهای مختلف می‌کند. در [۲۳] نشان داده شده‌است که در یک کاربرد شناسایی ارقام متصل، اگر مشخصهٔ فاز فیلتر RASTA را هموار کنیم برای مدل‌های مستقل از متن کارایی چنین فیلتری بهتر از کارایی فیلتر RASTA است. این فیلتر که در تمام فرکانس‌های مدولاسیون دارای فاز صفر می‌باشد، فیلتر RASTA با تصحیح فاز[1] نام دارد. در واقع فیلتر RASTA با تصحیح فاز H_{RPC} برابر است با حاصلضرب فیلتر RASTA یا H_R در فیلتر تصحیح فاز یا H_{PC}. اگر فاز فیلتر H_R را Φ_R بنامیم آنگاه به دلیل اینکه

$$H_{RPC}(w) = H_R(w).H_{PC}(w) \qquad (۲۲-۵)$$

و از طرفی چون:

$$\left|H_{RPC}(w)\right| = 1 \qquad (۲۳-۵)$$

بنابراین:

$$\left|H_{RPC}(w)\right| = \left|H_R(w)\right| \qquad (۲۴-۵)$$

[1] *Phase Corrected RASTA*

$$\Phi_{PC}(w) = -\Phi_P(w) \qquad (۲۵-۵)$$

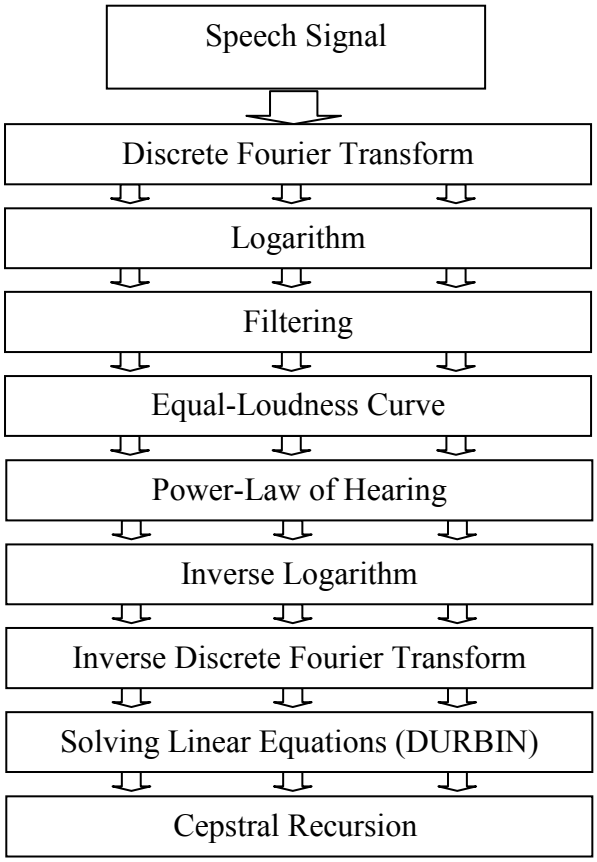
یعنی فیلتر H_{PC} در واقع یک فیلتر تمام گذرا است که مشخصهٔ فاز آن، همان مشخصهٔ فاز فیلتر RASTA ولی با علامت مخالف باشد و بنابراین تابع H_{PC} می‌تواند دارای فرمولی مشابه فرمول زیر باشد:

$$H_{PC}(z) = \frac{b_0 + b_1 z^{-1} + \cdots + b_q z^{-q}}{a_0 + a_1 z^{-1} + \cdots + a_p z^{-p}} \qquad (۲۶-۵)$$

۵-۷- ویژگی‌های RASTA-PLP

روش پیشگویی خطی مبتنی بر درک انسان بر پایه طیف زمان کوتاه گفتار می‌باشد. مانند دیگر تکنیک‌های مبتنی بر طیف زمان کوتاه، این روش نیز در مقابل اثرات کانال‌های مخابراتی بر مقادیر طیف زمان کوتاه، آسیب‌پذیر است. بکار بردن تحلیل طیف نسبی می‌تواند ضرایب PLP را در مقابل اثرات تخریبی کانال ارتباطی مقاوم کند.

RASTA مانند یک فیلتر میان‌گذر یا بالاگذر عمل می‌کند که شیب تندی در صفر دارد و در هر کانال فرکانسی، مقادیر ثابت یا با تغییرات کم را حذف می‌کند. از آنجا که در حوزه لگاریتم طیف اثر کانال به صورت مؤلفه‌های ثابت نمایان می‌شود، استفاده از این فیلتر می‌تواند اثرات کانال را کاهش دهد. شکل ۵-۶ مراحل محاسبه ضرایب RASTA-PLP را نشان می‌دهد. آخرین مرحله در بلوک دیاگرام نشان داده شده محاسبه ضرایب کپسترال از ضرایب RASTA-PLP است.



شکل ۵-۶- مراحل استخراج ضرایب RASTA- PLP [۲۴]

۵-۸- ارائه بهتر بردار ویژگی

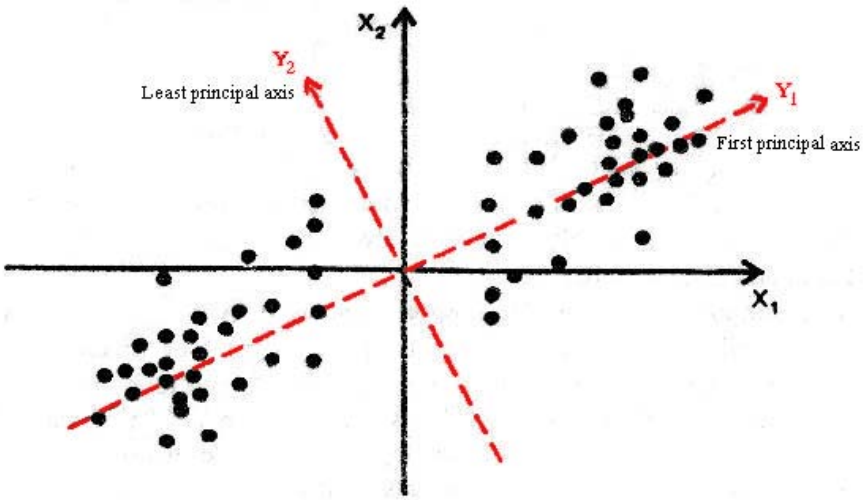
در یک سیستم بازشناسی هر چه عناصر بردار ویژگی که برای یک فریم به دست می‌آید از همدیگر مستقل‌تر باشند کار بازشناسی و تحلیل ویژگی‌ها آسانتر خواهد شد. این مورد زمانی که ماتریس کوواریانس بردار ویژگی را بخواهیم قطری در نظر بگیریم، مفیدتر خواهد بود. نشان داده شده که اگر بتوانیم تبدیلی اعمال کنیم که عناصر بردار ویژگی را تا حد ممکن مستقل یا غیرهمبسته کند، با توجه به کم شدن ابعاد بردار و دور ریخته شدن اطلاعات غیر ضروری در کم کردن اثر نویز مؤثر است. برای این کار روش معمول

تبدیل DCT[1] می‌باشد، تبدیل‌های مؤثرتر PCA[2] و LDA[3] می‌باشند[۲۵]، [۴۸]. در ادامه یکی از این تبدیل‌ها یعنی تبدیل PCA و استفاده از آن برای استخراج ویژگی‌های مقاوم مورد بررسی قرار می‌گیرد.

تحلیل اجزای اصلی یا PCA [۳۶]، یک روش آماری شناخته شده در کاهش تعداد ویژگی‌ها و کلاستر کردن می‌باشد. هدف روش این است که با اعمال تبدیل خطی و معکوس پذیر روی ویژگی‌ها، آن‌ها را به فضای جدیدی ببرد که در آن ویژگی‌های جدید دارای بیشترین تغییرات نسبت به همدیگر باشند. به عبارتی روش در جستجوی بردارهایی است که خطای نگاشت[4] به فضای جدید را مینیمم و واریانس داده‌های بعد از نگاشت را ماکزیمم نماید و به نوعی بردارهای فضای جدید را از همدیگر مستقل نماید. اگر بخواهیم با این روش بردار n بعدی X را به فضای جدید m بعدی Y ببریم ($m \le n$)، از تبدیلی به صورت $Y = T.X$ استفاده می‌کنیم که در آن T ماتریس تبدیل معکوس پذیر است که بر اساس اطلاعات آماری بردارهای ویژگی اولیه بدست می‌آید. این تبدیل از روش‌های آماری جهت تعیین میزان مهم بودن ویژگی‌ها استفاده می‌کند و اگر در آن میزان میانگین مربعات خطای[5] ناشی از کاهش تعداد ویژگی‌ها از n به m را E بگیریم، می‌توان نشان داد که مقدار E مینیمم است. درک شهودی کار این تبدیل در شکل ۵-۷ نشان داده شده است، در این شکل داده‌ها به صورت بردارهای دو بعدی تصادفی در فضای مختصات کارتزین پراکنده شده‌اند. واضح است که در این شکل با توجه به پراکندگی داده‌ها طرز قرار گرفتن بردارهای X_1 و X_2 طوری است که هیچکدام به درستی نمی‌توانند نقاط فضا را از هم جدا کنند اما اجزای اصلی[6] داده‌ها یا همان محورهای حاصل از تبدیل PCA این کار را خیلی بهتر می‌توانند انجام دهند. در شکل ۵-۷ محور Y_1 که معادل اولین جزء اصلی است به تنهایی کار جداسازی را توانسته انجام دهد.

---

[1] *Discrete Cosine Transform*
[2] *Principal Component Analysis*
[3] *Linear Discriminate Analysis*
[4] *Pøjection*
[5] *Mean Square Error (MSE)*
[6] *Principal Components*



شکل ۵-۷- فضای داده‌ها در مختصات کارتزین در مقایسه با محورهای حاصل از تبدیل PCA

علاوه بر PCA، این تبدیل دارای اسامی دیگری چون تبدیل بردارهای ویژه[۱]، تبدیل هتلینگ[۲] و KLT [3] نیز می‌باشد. از نظر تاریخی سابقه معرفی این تبدیل به سال ۱۹۰۱ برمی‌گردد که پیرسون[۴] آنرا معرفی کرد و از آن در زمینه زیستی استفاده نمود، در سال ۱۹۳۳ هتلینگ از آن در روانشناسی برای تبدیل متغیرهای گسسته به ضرایبی غیر همبسته استفاده کرد، کارهانن در سال ۱۹۴۷ از این ایده جهت تبدیل داده‌های پیوسته و بر اساس تئوری احتمال بهره گرفت و در سال ۱۹۴۸ لاف آنرا عمومیت داد. کاشمن در سال ۱۹۵۴ نشان داد که KLT دارای میانگین مربعات خطای مینیمم است. در سال ۱۹۵۶ تبدیل هتلینگ مجدداً توسط کرامر و ماتیوس مورد بررسی قرار گرفت.

برای بدست آوردن اجزای اصلی (PC) متغیر تصادفی X به صورت زیر عمل می‌کنیم: برای بردار n بعدی X، $X_i = [x_1 \ x_2 \ \ldots \ x_n]^T$ در صورت معلوم بودن تابع توزیع احتمال متغیر تصادفی X می‌توان میانگین آنرا به صورت $\mu_x = E[X]$ بدست آورد که در آن E بیانگر امید ریاضی متغیر تصادفی X است؛ اما

[1] *Eigenvectors*
[2] *Hitteling Transform*
[3] *Karhunen-Loeve (K-L) Transform*
[4] *Pearson*

در عمل معمولاً بایستی به صورت تقریبی و گسسته و با استفاده از رابطه $\mu_x = \frac{1}{M}\sum_{k=1}^{M} X_k$ مقدار میانگین را بدست آوریم، در این رابطه M تعداد بردارهای nتایی X می‌باشد. برای بردار X با مقدار میانگین μ_x ماتریس کواریانس به صورت $C_x = E[(X_k - \mu_x)(X_k - \mu_x)^T]$ خواهد بود که ماتریسی $n \times n$، حقیقی و متقارن است. مشابه بدست آوردن میانگین، بدست آوردن این ماتریس نیز در عمل بایستی به صورت تقریبی $C_x = \frac{1}{M}\sum_{k=1}^{M}[(X_k - \mu_x)(X_k - \mu_x)^T]$ محاسبه شود. هر کدام از عناصر σ_{ij} در ماتریس کواریانس نشان دهنده مقدار کواریانس (همبستگی) بین عناصر x_i و x_j در بردار X می‌باشد که در صورت غیرهمبسته[1] بودن x_i و x_j این مقدار صفر خواهد بود. برای این ماتریس مقادیر ویژه λ_i و بردارهای ویژه نرمال شده متناسب $\{\Phi_i\}$ را می‌توان با استفاده از رابطهٔ $\|\Phi_i\| = \sqrt{\Phi_i{}^T \Phi_i} = 1$ بدست آورد. حال بردار الگوی X می‌تواند بر روی زیرفضا یا زیرفضاهای پوشانده شده با بردار یا بردارهای ویژه نگاشته شود تا اجزای اصلی الگو را بدست آوریم. مثلاً برای بدست آوردن اولین (بزرگترین) جزء اصلی الگوی X می‌توان الگوی X را بر روی بزرگترین (مهم‌ترین) مقدار ویژه به صورت $X_i = \Phi_i{}^T X$ نگاشت کرد. سایر اجزای اصلی را نیز می‌توان به همان صورت بدست آورد که ترتیب اولویت آن‌ها از اول تا آخر با توجه به اهمیت بردارهای ویژه مشخص می‌شود و اهمیت بردارهای ویژه نیز با توجه به ترتیب بزرگ بودن مقادیر ویژه متناسب بدست می‌آیند. در شکل ۵-۷ محور Y_1 معادل اولین جزء اصلی (مهمترین جزء) و Y_2 معادل دومین و کم اهمیت‌ترین جزء اصلی است.

تبدیل کارهانن-لاف یا PCA از ایده اجزای اصلی استفاده می‌کند و یک تبدیل معکوس پذیر ایجاد می‌کند که در کارهایی مثل کاهش بعد و فشرده‌سازی از آن بتوان استفاده کرد. همانطور که بیان شد اجزای اصلی (PC) ترکیبی خطی از متغیرهای تصادفی‌اند که با توجه به واریانس آن‌ها خصوصیات خاصی دارند.

[1] *Uncorrelated*

جزء اصلی اول ترکیب خطی نرمال شده با بیشترین واریانس است و دومین جزء اصلی دارای دومین مقدار واریانس است و به همین ترتیب تا آخرین جزء اصلی که دارای کمترین واریانس است. در واقع اجزای اصلی با توجه به قدرتشان در تفاوت گذاشتن بین خوشه‌ها (مقدار واریانس) اولویت دهی می‌شوند. هسته اصلی این روش یک ماتریس تبدیل متعامد است. برای محاسبهٔ آن ابتدا برای کلیه بردارهای n بعدی X_k، مقدار میانگین هرکدام از n عنصر را در مسیر زمانی[1] با استفاده از رابطهٔ $\mu_x = \frac{1}{M}\sum_{k=1}^{M} X_k$ بدست می‌آوریم. حاصل کار بردار n تایی میانگین μ_x می‌باشد، که اگر هر کدام از M بردار X_k را بردار مشخصه‌ای در نظر بگیریم؛ عنصر اول بردار میانگین، برابر میانگین عناصر اول بردارهای X_k یا همان میانگین مشخصه اول خواهد بود و برای سایر عناصر بردار میانگین هم به ترتیب میانگین سایر عناصر بردارهای X_k خواهد بود. برای مرکزی کردن داده‌ها همهٔ مشخصه‌ها را از میانگین متناظر خودشان صبق رابطهٔ زیر کم می‌کنیم:

$$\bar{X}_k(i) = X_k(i) - \mu_x(i) \qquad ,i = 1,2,...,n \quad k = 1,2,...,M \tag{۲۷-۵}$$

حال ماتریس همبستگی $\bar{X}$ یا ماتریس کواریانس X را بدست می‌آوریم:

$$C_x = \frac{1}{M}\sum_{k=1}^{M}[\overline{X}_k \overline{X}_k^T] \tag{۲۸-۵}$$

که با توجه به داده‌ها و حقیقی بودن آن‌ها، ماتریس کواریانس حاصل، ماتریسی مربعی $n \times n$، حقیقی و متقارن است. مقادیر ویژه λ_i و بردارهای ویژه $\{\Phi_i\}$ ماتریس کواریانس C_x را بدست می‌آوریم. بردارهای ویژه بردارهایی n تایی‌اند و تعداد مقادیر ویژه و بردارهای ویژه نیز برابر n می‌باشد. بعلت حقیقی و متقارن بودن ماتریس کواریانس، مقادیر ویژه بدست آمده حقیقی و غیر صفر خواهند بود و بردارهای ویژه نیز متعامد می‌باشند و پیدا کردن n بردار ویژه متعامد همیشه ممکن است. میزان اهمیت هر کدام از بردارهای

[1] *Time Trajectory*

ویژه با توجه به بزرگی مقادیر ویژه متناسب بدست می‌آید چرا که مقدار مقادیر ویژه میزان واریانس داده‌ها را نشان می‌دهند و میزان توزیع واریانس در راستای هر کدام از محورهای جدید به صورت رابطهٔ (۵-۲۹) و متناظر با مقدار ویژه آن بردار است.

$$V = \frac{\lambda_k}{\sum_{i=1}^{n} \lambda_i} \qquad (۲۹-۵)$$

هر چه مقادیر ویژه بزرگتر باشند بردار ویژه متناظر دارای اهمیت بیشتری است و اجزای اصلی با اهمیت بالاتری را نتیجه می‌دهد. بدیهی است که برای هر مقدار ویژه و بردار ویژه متناظرش رابطهٔ $C_x\Phi_i = \lambda_i\Phi_i$ برقرار است. از روی مقادیر ویژه و بردارهای ویژه، ماتریس تبدیل $n \times n$ متعامد و معکوس‌پذیر T را بدست می‌آوریم. برای این کار سطر اول ماتریس T را بردار ویژه متناسب با بزرگترین مقدار ویژه می‌گیریم و سطر دوم آنرا بردار ویژه متناسب با دومین بزرگترین مقدار ویژه و به همین ترتیب تا n سطر ماتریس تبدیل از n بردار n تایی ایجاد شود:

$$T = [\Phi_1 \Phi_2 \cdots \Phi_n]^T \quad \lambda_1 \geq \lambda_2 \geq \cdots \lambda_n \qquad (۳۰-۵)$$

حال با داشتن این ماتریس تبدیل می‌توان بردارهای X_k و یا هر بردار تصادفی مثل Z را که همنوع و هم اندازه بردارهای X_k است، را به فضای جدید نگاشت و اجزای اصلی آنرا بدست آورد. برای این کار کافی است که ماتریس تبدیل T را در بردار مورد نظر ضرب کرد یعنی:

$$Y = T.(Z - \mu_z) \qquad (۳۱-۵)$$

که اجزای آن عبارت خواهند بود از:

$$Y_i = \Phi_i{}^T (Z_i - \mu_{zi}) \qquad (۳۲-۵)$$

تبدیل فوق را تبدیل هتلینگ می‌نامند. در این رابطه ماتریس تبدیل $T_{n\times n}$، در ماتریس ورودی $Z_{n\times p}$ (P تعداد بردارهای ویژگی اولیه) ضرب می‌شود تا فضای ورودی را به فضای جدیدی که از بردارهای

متعامد درست شده‌اند، نگاشت کند و ماتریس خروجی با همان اندازه را نتیجه دهد. یکی از کاربردهای اصلی این تبدیل در کار فشرده‌سازی و کاهش تعداد ویژگی‌های ورودی است که می‌تواند به جای n ویژگی‌ای که در هر کدام از بردارهای اولیه وجود دارند، فقط m ویژگی مهمتر ($m \leq n$) را حفظ کرد؛ که در ادامه به این مسأله و خصوصیات ویژگی‌های جدید بیشتر پرداخته می‌شود.

این تبدیل معکوس پذیر است، بدین معنی که می‌توان ویژگی‌های جدید را به فضای اولیه برگرداند و مسأله همیشه معکوس پذیر بودن این ماتریس را می‌توان ثابت کرد چون برای ماتریسی که سطرهای آن از بردارهای متعامد تشکیل شده‌اند می‌توان نوشت:

$$T^{-1} = T^{T} \qquad (۳۳-۵)$$

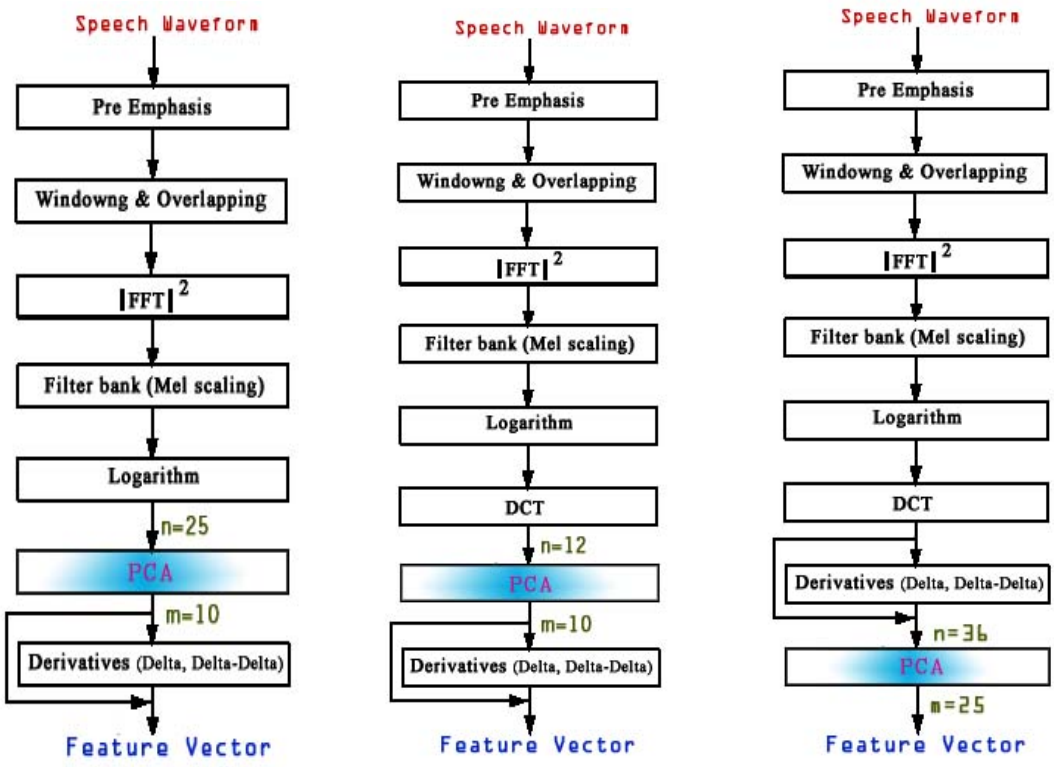
در این صورت می‌توان معکوس تبدیل هتلینگ را به صورت زیر بدست آورد:

$$Z = T^{T}.Y + \mu_z \qquad (۳۴-۵)$$

استفاده از این خاصیت در کاربردهایی مثل فشرده‌سازی خیلی حایز اهمیت است بدین صورت که چون سطرهای ماتریس تبدیل T به ترتیب از اهمیت آن‌ها کاسته می‌شود می‌توان به جای استفاده از کل ماتریس $n \times n$ از نمونه کاهش یافته آن استفاده کرد و سطرهای پایینی آنرا که دارای اهمیت کمتری‌اند حذف کنیم تا ماتریسی $m \times n$ ($m \leq n$) از آن باقی بماند و با نگاشت داده اصلی $n \times p$بعدی، داده‌های فضای جدید (نگاشت شده) به جای اندازه اولیه، $m \times p$ باشد. با این کار در واقع از اجزای با اهمیت کمتر داده‌ها صرفنظر شده است و داده کاهش بعد یافته و یا فشرده شده است.

برای استفاده از PCA در یک سیستم بازشناسی گفتار که دارای دو مرحلهٔ آموزش و استفاده و یا آزمون است، بایستی ماتریس تبدیلی را استخراج کنیم که در هر دو مرحله از آن استفاده کنیم. برای این کار ماتریس تبدیل را از روی کل داده‌های آموزش می‌سازیم. بدین صورت که به داده‌های آموزش به دید تعدادی بردار تصادفی نگریسته می‌شود. ابتدا روی مسیر زمانی مولفه‌های بردارهای آموزش با توجه به محل

قرار گرفتن و استفاده شدن تبدیل، میانگین گرفته و ماتریس کواریانس را به صورتی که بیان شد استخراج می‌کنیم و سپس برای این ماتریس مقادیر ویژه و بردارهای ویژه و نهایتاً ماتریس تبدیل را بدست می‌آوریم. در بکارگیری PCA به عنوان قسمتی از پیش پردازنده سیستم بازشناسی گفتار، می‌توان آنرا در جاهای مختلفی از این سیستم گذاشت. از جمله این موارد سه انتخابی است که در شکل ۵-۸ برای سیستمی که مبنای استخراج ویژگی‌های آن MFCC است، نشان داده شده است. در قسمت الف این شکل، PCA به عنوان آخرین بلاک پیش پردازنده برای کاهش تعداد ویژگی‌ها و مستقل کردن ویژگی‌ها استفاده شده است. در قسمت ب PCA بعد از بلاک DCT به عنوان انتخابی دیگر برای محل قرارگیری PCA، قرار گرفته است، می‌توان قرارگرفتن PCA در این قسمت را علاوه بر انتظار کاهش تعداد، به دید مستقل‌تر کردن ویژگی‌ها نیز نگریست. نهایتاً مطابق آنچه که در قسمت ج این شکل نشان داده شده است، می‌توان PCA را به عنوان جانشین بلاک DCT جهت کاهش تعداد ویژگی‌ها و به نوعی جدا سازی ویژگی‌های خروجی بانک فیلتر استفاده کرد.



ج ب الف

شکل ۵-۸- استفاده از PCA در محل‌های مختلف سیستم استخراج ویژگی‌های MFCC

۵-۹- روش‌های برگرفته از اندام تولیدی گفتار

در این مقوله سعی می‌شود نحوه اثر اندام گویایی بر سیگنال گفتار شناخته شود و با نرمال کردن این اثرات، تغییرات حاصل از گوینده‌های مختلف بر سیگنال گفتار کاهش یابد. به عنوان مثال طول و شکل مجرای صوتی در اشخاص مختلف متفاوت می‌باشد و این باعث تغییر در سیگنال گفتار می‌شود. اگر سیستم بازشناسی، مستقل از گوینده باشد، نباید این عامل اثر منفی داشته باشد. بدین منظور الگوریتم‌های VTN[1] را می‌توان اعمال نمود[۲۶].

[1] *Vocal Tract Normalization (VTN)*

فیلتر پیش تأکید[1] نیز برای خنثی کردن اثرات طیفی حنجره (دو قطب) ولب‌ها (یک صفر) به کار می‌رود. این فیلتر همچنین تغییرات ناگهانی موجود در سیگنال را که بر اثر نویزهای شدید محیط به وجود می‌آید حذف می‌کند و باعث یکنواخت شدن سیگنال می‌شود. پیاده‌سازی این فیلتر در زمان به صورت زیر می‌باشد:

$$y_p(t) = y(t) - \alpha \times y(t-1) \qquad (۳۵-۵)$$

که در اینجا α ضریب پیش‌تأکید می‌باشد و مقدار آن به صورت $0.9 \le \alpha \le 1$ انتخاب می‌شود.

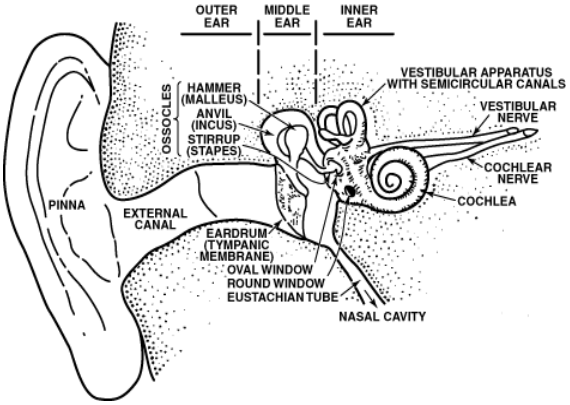
۵-۱۰- استخراج بازنماییِ مقاوم مبتنی بر مدل فیزیولوژی گوش انسان

در این بخش یک روش جدید با الهام گرفتن از گوش انسان و مدل‌کردن حرکت‌های مکانیکی غشای پایه[2] و سلول‌های مژکدار[3] برای دریافت فرکانس‌های مختلف صدا بر اساس یافته‌های فیزیولوژی ارائه می‌شود. در ابتدا توضیح مختصری از فیزیولوژی گوش انسان ارئه می‌شود و سپس مدل مبتنی بر آن بیان می‌گردد.

۵-۱۰-۱- فیزیولوژی سیستم شنوایی انسان

گوش اندام حسی تبدیل کنندهٔ صدا است. گوش انسان از سه بخش گوش بیرونی، گوش میانی و گوش درونی تشکیل شده است (شکل ۵-۹). گوش بیرونی امواج صوتی را به پردهٔ صماخ، که در حد فاصل گوش بیرونی و گوش میانی است، راه می‌دهد. امواج صوتی پردهٔ صماخ را به ارتعاش در می‌آورند. ارتعاش‌های پردهٔ صماخ از طریق سه استخوان کوچک موجود در گوش میانی به گوش درونی انتقال می‌یابند[۲۷].

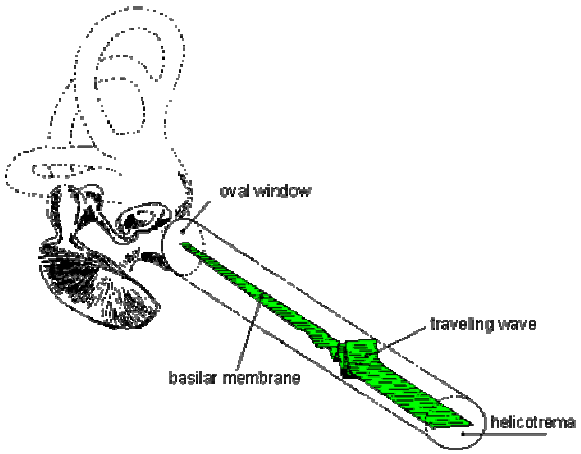
[1] *Pre-emphasis filter*
[2] *Basilar Membrane*
[3] *Hair Cells*



شکل ۵-۹- ساختار کلی گوش

گوش درونی حفرهٔ پر از مایعی است که حلزونی گوش نامیده می‌شود. حلزونی گوش انسان دارای یک ساختار مخروطی شکل است که فنروار در اطراف یک هستهٔ استخوانی پیچیده شده است. نیروی مکانیکی تقویت شده که از طرف گوش میانی به گوش داخلی فرستاده می‌شود بلافاصله توسط حلزونی گوش به فشار هیدرولیکی تبدیل می‌شود.این فشار هیدرولیکی، حرکت حاصل را به مجراء حلزونی گوش[۱] و اندام کورتی[۲] میرساند. در سراسر طول حلزون، نوعی غشاء امتداد یافته است که غشای پایه نامیده می‌شود. این غشاء پردهٔ خم‌پذیری است که هرگاه صدایی به پردهٔ صماخ برخورد کند آن را به ارتعاش در می‌آورد. ارتعاش غشای پایه به صورت جابجایی موج‌مانندی است که آن غشاء را تغییر شکل می‌دهد.

[1] *Cochlea duct*
[2] *Organ of Corti*

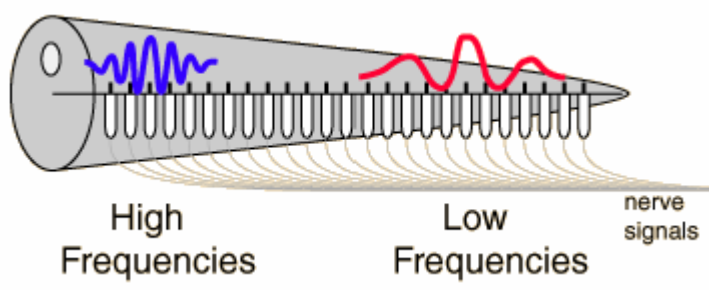


شکل ۵-۱۰- حرکت موج جابجایی در در طول غشای پایه

هنگامی که موج جابجایی در طول غشای پایه پیش می‌رود (شکل ۵-۹)، بر دامنهٔ آن افزوده می‌شود تا اینکه در نقطه‌ای به اوج برسد. اما وقتی که موج جابجایی غشای پایه از نقطهٔ اوج جلوتر می‌رود، دامنهٔ آن کاهش می‌یابد. محل اوج دامنهٔ موج جابجایی در غشای پایه، بستگی به فرکانس صوت دارد. اصوات با فرکانس بالا سبب می‌شوند که موج جابجایی در قاعدهٔ حلزون به اوج خود برسند در حالیکه اصوات با فرکانس پایین سبب می‌شوند که موج جابجایی در رأس حلزون به اوج خود برسد. بنابراین، طرح مکانیکی غشای پایه چنان است که به اصوات دارای فرکانس‌های مختلف با تغییر دادن شکل خود پاسخ می‌دهد[۲۷]. تغییر شکل غشای پایه را سلول‌های مژکدار به پتانسیل کار تبدیل می‌کنند. وقتی که غشای پایه مرتعش می‌شود و به اوج تغییر شکل خود می‌رسد، سلول‌های مژکدار نسبت به جسم آن سلول خم می‌شوند و از این طریق عصب شنوایی، پتانسیل کار تولید می‌کند که به مغز منتقل می‌شود و اطلاعات مربوط به فرکانس صوت را به مغز می‌رساند.

۵-۱۰-۲- مدل‌سازی حرکت‌های مکانیکی غشای پایه و سلول‌های مژکدار

از آنجایی که غشای پایه و سلول‌های مژکدار با همکاری یکدیگر فرکانس‌های مختلف صدا را به انرژی مکانیکی تبدیل می‌کنند با تقریب خوبی می‌توان هر قسمت از آن را به صورت گسسته به شکل یک دیاپازون در نظر گرفت. به این دیاپازون‌ها در واقع می‌توان به شکل سیستم‌های نوسانی میرا که توسط یک نیروی خارجی تحریک می‌گردند، نگاه کرد.



شکل ۵-۱۱- مدل کردن حرکت‌های مکانیکی غشاء پایه و سلول‌های مژکدار با دیاپازون

مدل هر یک از این سیستم‌های نوسانی میرا را می‌توان با رابطه ۵-۳۶ نمایش داد[۲۸]:

$$\ddot{x}+\frac{b}{m}\dot{x}+{\omega_0}^2 x=\frac{F}{m}\cos\omega t \qquad (۳۶-۵)$$

که در این رابطه b نشان‌دهندهٔ فاکتور میرایی، ω_0 فرکانس طبیعی هر کدام از این دیاپازون‌ها و F نیروی خارجی است که توسط صوت با فرکانس ω بر هر یک ار آن‌ها اعمال می‌شود. از آنجایی که این مدل‌سازی باید به صورت گسسته انجام شود، مثلاً می‌توان به تعداد ۲۵۶ عدد از این مدل‌ها برابر با ۲۵۶ باند فرکانسی در نظر گرفت. از آنجایی که هر یک از این مدل‌ها باید به فرکانس خاصی حساس باشند و در آن فرکانس حالت تشدید داشته باشند، برای هر یک از آن‌ها یک باند فرکانسی خاص در نظر گرفته شده است. این باند فرکانسی تعیین کننده مقدار b بر اساس رابطه زیر می‌باشند،

$$b=\frac{m\omega_0}{Q} \qquad (۳۷-۵)$$

که در رابطهٔ فوق Q فاکتوری است که تعیین کنندهٔ میزان انرژی از دست رفته است و همچنین k را به عنوان ضریبی تعریف می‌کنیم که به وسیلهٔ رابطهٔ $k = m\omega_0^2$ به جنس هر یک از دیاپازون‌ها وابسته است.

برای استفاده از این مدل در استخراج بازنمایی‌ها باید به ازای تعدادی نمونهٔ صوت شتاب هریک از دیاپازون‌ها به دست آید و از روی این شتاب و میزان انحراف آن‌ها انرژی برای تک‌تک آن‌ها محاسبه گردد. این انرژی به ازای تک‌تک دیاپازون‌ها می‌تواند به عنوان بردار بازنمایی برای سیستم بازشناسی گفتار مورد استفاده قرار بگیرد.

شتاب هر یک از این مدل‌ها را بر اساس سرعت و مکان قبلی و همچنین نیروی اعمال شدهٔ جدید که از طرف نمونه‌های صوت بر آن‌ها وارد می‌شوند، طبق رابطه زیر بدست می‌آید

$$a = \frac{F - bv_{pr} - kx_{pr}}{m} \qquad (۳۸-۵)$$

از روی این شتاب سرعت و مکان جدید به دست می‌آید.

$$\begin{bmatrix} x_{new} \\ v_{new} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t & 0 \\ 0 & 1 & \Delta t \end{bmatrix} \begin{bmatrix} x_{pr} \\ v_{pr} \\ a \end{bmatrix} \qquad (۳۹-۵)$$

در انتها انرژی از طریق رابطهٔ $E = \frac{1}{2}mv^2 + \frac{1}{2}kx^2$ به دست می‌آید و از آن برای بازنمایی‌های سیستم بازشناسی گفتار استفاده می‌شود. بردار بازنمایی نهایی که سیستم بازشناسی گفتار از آن استفاده خواهد کرد عبارت است از مقدار انرژی بدست آمده توسط رابطهٔ فوق که به ازای تک تک دیاپازون‌ها محاسبه شده است. همانطوری که بیان شد هر یک از این دیاپازون‌ها دارای مشخصه‌های خاص خود می‌باشند که از مدل کردن گسسته غشای پایه و سلول‌های مژکدار بدست آمده‌اند.

فصل ششم

طراحی و تهیهٔ دادگان تلفنی اعداد متصل و پیوسته زبان فارسی و استفاده از آن در سیستم بازشناسی اعداد تلفنی

۶-۱- مقدمه

در این فصل نحوهٔ طراحی و تهیهٔ دادگان اعداد زبان فارسی برای یک سیستم بازشناسی گفتار پیوسته بیان خواهد شد. در انتها نیز سیستم بازشناسی گفتار پیوسته مورد استفاده و نحوهٔ آموزش آن برای استفاده از این دادگان بیان خواهد شد. همانطور که در ادامه نیز بیان خواهد شد، سیستم مزبور باید توانایی استفاده از این دادگانِ مبتنی بر کلمه که در آن فقط برچسب کلمات آمده است و هیچگونه تقطیعی در سطح واج یا کلمه در آن صورت نگرفته است را داشته باشد.

۶-۲- طراحی و تهیهٔ دادگان تلفنی اعداد متصل و پیوسته زبان فارسی

برای آموزش سیستم‌های بازشناسی گفتار به یک دادگان معتبر نیاز است به طوری که سیستم بتواند از روی دادگان آموزشی تمامی واحدهای آوایی و گذارهای بین آن‌ها را فرا بگیرد. هدف اصلی از تهیهٔ این دادگان فراهم کردن داده‌های آموزشی برای یک سیستم بازشناسی گفتار پیوسته مبتنی بر واحد کلمه برای بازشناسی اعداد از طریق خط تلفن است و هدف نهایی، استفاده از آن در محدودهٔ وسیعی از طرح‌های پژوهشی و تجاری می‌باشد. برای زبان فارسی دادگان تلفنی معتبر زیادی در دسترس نیست. دو دادگان تلفنی OGI multi-language [۲۹] و Call Friend Farsi [۳۰] برای زبان فارسی وجود دارند که جهت بازشناسی خودکار زبان مورد استفاده قرار می‌گیرند و تنها دادگان تلفنی آماده شده جهت بازشناسی گفتار دادگان TFarsdat [۳۱] می‌باشد. این دادگان در سطح واج تقطیع شده است و معمولاً برای سیستم‌های بازشناسی گفتار که واحد آوایی آن‌ها واج می‌باشد مورد استفاده قرار می‌گیرد. از آنجایی که در این دادگان فرکانس تکرار هر کلمه به خصوص اعداد کم می‌باشد از آن نمی‌توان برای سیستم‌های بازشناسی اعداد مبتنی بر کلمه استفاده کرد.

این دادگان که FCCDigits[1] نامیده شده، بـه منظـور اسـتفاده در سیـستم‌هـای بازشناسـی گفتـار، در پژوهشکدهٔ پردازش هوشمند علایم ضبط گردیده است. کار مطالعه و طراحی این دادگان در حدود ۶ ماه به طول انجامیده است. در زیر مشخصات کلی این دادگان آمده است.

۶-۲-۱- توصیف گویندگان

گویندگان با توجه به سن، جنس و میزان تحصیلات انتخاب شده‌اند. تعـداد کـل گوینـدگانی کـه در جمع‌آوری قسمت اعداد متصل این دادگان همکاری کرده‌اند برابر با ۱۱۰ نفر می‌باشند کـه ۱۰۰ نفـر از ایـن افراد در قسمت اعداد پیوسته نیز همکاری نموده‌اند. در واقع قسمت اعداد متصل این دادگان مشتمل بر ۱۱۰ گوینده و قسمت اعداد پیوسته شامل ۱۰۰ گوینده است. این دادگان فقط در بردارنـدهٔ گوینـدگانی بـا لهجـهٔ تهرانی است که از نقاط مختلف شهر تهران از طریق شبکهٔ تلفن با مرکز جمع‌آوری دادگان تماس گرفته‌انـد و رشته‌های عددی از پیش تعیین شده را بیان کرده‌اند. نسبت تعداد گویندگان مرد به زن در کل این دادگـان تقریباً برابر با ۲ به ۱ می‌باشد که بیشترین آن‌ها در محدودهٔ سنی بین ۲۵ تا ۴۰ سال انتخاب شده‌انـد. تعـداد زنان و مردان در هر محدودهٔ سنی و تعداد کل آنها برای اعـداد متـصل و پیوسـته بـه صـورت جداگانـه در جدول‌های ۶-۱ و ۶-۲ آمده اند. نمودار نشان داده شده در شکل‌هـای ۶-۱ و ۶-۲ نیـز بیـان کننـدهٔ توزیـع مناسب افراد در محدوده‌های سنی مختلف می‌باشند.

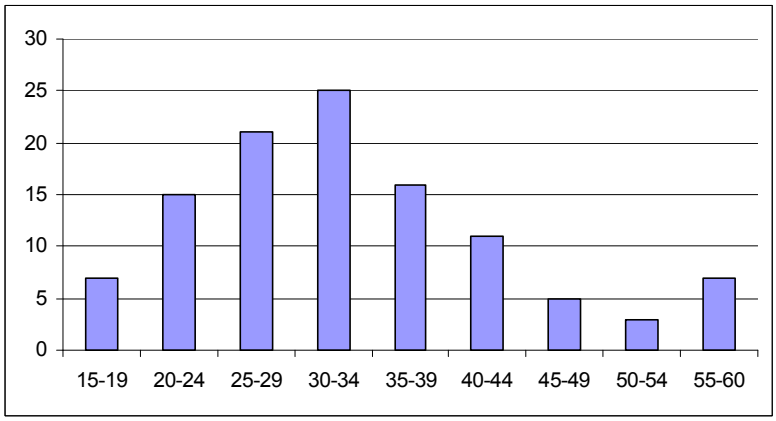
[1] *Farsi Connected and Continuous Digits*

جدول ۶-۱- نحوه توزیع گویندگان اعداد متصل دادگان با توجه به محدوده سنی آن‌ها

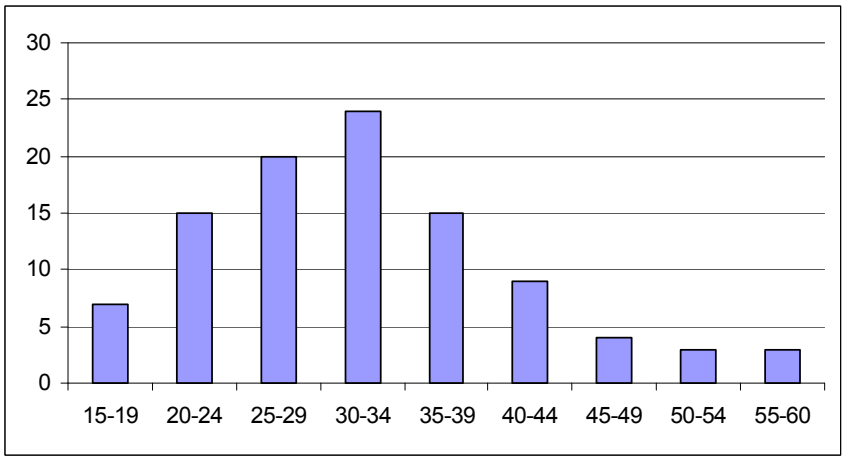
محدودهٔ سنی	مرد	زن	تعداد کل	درصد
۱۹ -۱۵	۳	۴	۷	٪۶.۳۶
۲۴ -۲۰	۸	۷	۱۵	٪۱۳.۴۶
۲۹ -۲۵	۱۲	۹	۲۱	٪۱۹.۰۹
۳۴ -۳۰	۱۶	۹	۲۵	٪۲۲.۷۳
۳۹ -۳۵	۱۱	۵	۱۶	٪۱۴.۵۵
۴۴ -۴۰	۶	۵	۱۱	٪۱۰
۴۹ -۴۵	۲	۳	۵	٪۴.۵۵
۵۴ -۵۰	۳	۰	۳	٪۲.۷۳
۶۰ -۵۵	۶	۱	۷	٪۶.۳۶

جدول ۶-۲-: نحوه توزیع گویندگان اعداد پیوسته دادگان با توجه به محدوده سنی آن‌ها

محدودهٔ سنی	مرد	زن	تعداد کل	درصد
۱۹ -۱۵	۳	۴	۷	٪۶.۳۶
۲۴ -۲۰	۸	۷	۱۵	٪۱۳.۶۴
۲۹ -۲۵	۱۱	۹	۲۰	٪۱۸.۱۸
۳۴ -۳۰	۱۵	۹	۲۴	٪۲۱.۸۲
۳۹ -۳۵	۱۰	۵	۱۵	٪۱۳.۶۴
۴۴ -۴۰	۶	۳	۹	٪۸.۱۸
۴۹ -۴۵	۱	۳	۴	٪۳.۶۴
۵۴ -۵۰	۳	۰	۳	٪۲.۷۳
۶۰ -۵۵	۳	۰	۳	٪۲.۷۳



شکل ۶-۱- نحوه توزیع گویندگان اعداد متصل دادگان با توجه به محدوده سنی آن‌ها



شکل ۶-۲- نحوه توزیع گویندگان اعداد پیوسته دادگان با توجه به محدوده سنی آن‌ها

۶-۲-۲- واژگان

گفتاری که از گویندگان در این دادگان جمع‌آوری شده است، رشته‌های عددی می‌باشد. اعدادی که گویندگان در قسمت اعداد متصل این دادگان بیان کردند عبارتند از "صفر"، "یک"، "دو"، "سه"، "چهار"، "پنج"، "شش"، "هفت"، "هشت"و "نه". در این قسمت از دادگان هر گوینده ۷۰ رشته از این اعداد را به طور متصل و به صورت ۲۰ رشتهٔ یک رقمی (هر کدام از ارقام "صفر" تا "نه" دو بار ادا شده اند)، ۱۰ رشتهٔ دو رقمی، ۱۰ رشتهٔ سه رقمی، ۱۰ رشتهٔ چهار رقمی، ۱۰ رشتهٔ پنج رقمی و ۱۰ رشتهٔ هفت رقمی بیان کرده

است. بنابراین هر گوینده ۲۳۰ عدد متصل و ۱۶۰ حالت گذر بین عددی را بیان کرده است. از الگوریتم زیر برای تولید لیست یکتایی از رشته‌های عددی متصل، که هر گوینده بیان می‌کند، استفاده شده است.

مرحلهٔ ۱: یک آرایهٔ هفتاد عنصری تولید کن. این آرایه در بر دارندهٔ طول رشته‌های عددی متصل است. به بیست عنصر این آرایه، که به‌طور تصادفی انتخاب شده‌اند، مقدار ۱ و به هر ده عنصر آرایه، که آن‌ها نیز به‌طور تصادفی انتخاب شده‌اند، مقادیر ۲، ۳، ۴، ۵ و ۷ نسبت بده. این آرایه بیان کنندهٔ مکان هر رشتهٔ عددی یک، دو، سه، چهار، پنج و هفت رقمی در لیستی که هر گوینده بیان می‌کند، می‌باشد.

مرحلهٔ ۲: برای رشته اعداد یک رقمی، از یک لیست بیست عنصری حاوی ارقام ۰ تا ۹، که هر یک دو بار تکرار شده‌اند، یک رقم را به طور تصادفی و بدون جایگزینی انتخاب کن. این ارقام به ترتیبی که آرایه هفتاد عنصری برای مکان رشته‌های عددی یک رقمی تعیین کرده است در لیست نهایی مربوط به گوینده قرار می‌گیرند.

مرحلهٔ ۳: در هر یک از ۵۰ مکان باقی‌مانده (مربوط به رشته‌های عددی با طول دو رقم و بیشتر) برای تعیین اولین رقم رشته‌های عددی متصل، از یک لیست ده عنصری شامل ارقام ۰ تا ۹، یک رقم را به صورت تصادفی و بدون جایگزینی انتخاب کن. اگر عدد انتخابی اول به عنوان مثال ۳ باشد یعنی این‌که رقم اول اولین رشته‌های عددی متصل دو، سه، چهار، پنج و هفت رقمی برابر ۳ خواهد بود.

مرحلهٔ ۴: برای تعیین ارقام بعدی در رشته‌های عددی متصل با طول دو رقم و بیشتر، به صورت تصادفی و بدون جایگزینی از لیست‌های گذار مربوط به ارقام قبلی یک رقم را انتخاب کن. ده لیست گذار مربوط به هر رقم وجود دارد که شامل ارقام ۰ تا ۹ است و بیان کنندهٔ رقم بعدی هر رقم می‌باشد. پس از انتخاب یک رقم از لیست گذار، رقم مذکور از این لیست حذف می‌گردد (این کار تضمین می‌کند که در رشته‌های با طول مختلف، گذارهای بین عددی مشابه وجود نداشته باشد).

مرحلهٔ ۵: به ازای هر گوینده، مراحل ۱ تا ۴ را تکرار کن.

الگوریتم فوق فقط نحوهٔ تولید اعداد متصل این دادگان را بیان می‌کند. در قسمت اعداد پیوسته دادگان هر گوینده کلمات زیر را دو بار بیان می‌کند:

"ده"، "یازده"، "دوازده"، "سیزده"، "چهارده"، "پانزده"،"شانزده"، "هفده"، "هجده"، "نوزده"، "بیست"، "سی"، "چهل"، "پنجاه"، "شصت"، "هفتاد"، "هشتاد"، "نود"، "صد"، "دویست"، "سیصد"، "چهارصد"، "پانصد"، "ششصد"، "هفتصد"، "هشتصد"، "نهصد"، "هزار"، "میلیون" و "میلیارد".

پس از آن به طور تصادفی چهار عدد دو رقمی پیوسته، شش عدد سه رقمی پیوسته، چهار عدد چهار رقمی پیوسته، چهار عدد پنج رقمی پیوسته، یک عدد شش رقمی پیوسته، پنج عدد هفت رقمی پیوسته، پنج عدد هشت رقمی پیوسته، یک عدد نه رقمی پیوسته، شش عدد ده رقمی پیوسته، شش عدد یازده رقمی پیوسته و دو عدد دوازده رقمی پیوسته توسط هر گوینده بیان می‌شود. به این ترتیب تقریباً تمام گذرهای بین عددی در دادگان اعداد پیوسته نیز بیان گردیده است. با توجه به مطالب گفته شده، در قسمت پیوسته دادگان هر گوینده در مجموع ۱۱۴ رشته عددی پیوسته را بیان کرده است.

۶-۲-۳- جمع‌آوری دادگان

برای جمع آوری دادگان ابتدا لیست رشته‌های عددی مربوط به هر گوینده به آنها داده شد سپس از طریق تماس تلفنی که از پژوهشکدهٔ پردازش هوشمند علایم با آن‌ها گرفته شد، از آن‌ها خواسته شد که رشته‌های عددی را بیان کنند. از گویندگان خواسته شده است که تا حد امکان در اتاقی بدون نویز رشته‌های عددی را بیان کنند. همچنین اعداد تا حد امکان به صورت پیوسته و بدون مکث ادا شده‌اند. صدای هر گوینده از طریق رابطی که از گوشی تلفن به ورودی کارت صدای کامپیوتر متصل بوده ضبط گردیده است.

نرخ نمونه برداری برابر با ۱۱,۰۲۵ کیلوهرتز و هر نمونه شانزده بیتی می‌باشد. برای رقمی کردن نمونه‌ها از کارت صدای Creative PC64 استفاده شده است.

برای هر گوینده یک فایل برچسب وجود دارد که برچسب رشته‌های عددی را به ترتیبی که گوینده ادا کرده، در خود جای داده است. برچسب هر رشتهٔ عددی شامل دنبالهٔ ارقام موجود در آن رشته می‌باشد.

۶-۲-۴- تصدیق صحت دادگان و آمایش آن

برای اطمینان از انطباق دادگان جمع‌آوری شده با لیست‌های آمده شده جهت هر گوینده، از گروهی از شنوندگان استفاده گردید. آنان هر یک از فایل‌های صوتی دادگان را با لیست‌های از پیش آماده شده مقایسه می‌کردند و در صورت مشاهدهٔ عدم تطابق بین لیست‌ها و گفتار ضبط شده، اشکال موجود گزارش می‌شد تا آن قسمت از دادگان دوباره ضبط گردد.

همچنین آزمایش‌هایی توسط گروهی از شنوندگان خبره برای تعیین کیفیت دادگان و همچنین قابل فهم بودن آن صورت پذیرفت تا در صورت وجود اشکال، کار تهیهٔ آن قسمت از دادگان دوباره صورت گیرد. در انتها و پس از جمع‌آوری کامل دادگان نیز فایل‌های صوتی دوباره مورد بازبینی قرار گرفتند. در این مرحله نیز آن دسته از فایل‌های صوتی که دارای اشکال بودند از دادگان حذف گردیدند. در جدول ۶-۳ مشخصات کلی این دادگان همراه با مقایسهٔ آن با سایر دادگان‌های فارسی موجود آمده است.

جدول ۶-۳- مشخصات کلی دادگان تلفنی اعداد متصل و پیوسته زبان فارسی و

مقایسه آن با سایر دادگان‌های تلفنی موجود زبان فارسی

نام دادگان	کاربرد	تعداد گویندگان	اندازه کل دادگان (ساعت)	اندازه دادگان برچسب دهی شده (ساعت)
TFarsdat	بازشناسی گفتار	۶۴	۸	۸
OGI Multilanguage	بازشناسی زبان	۲۰۵۲ (همه زبان‌ها)	۳۸.۵	۸
Call Friend FARSI	بازشناسی زبان	۶۰	۳۰	-
FCCDigits	بازشناسی گفتار	۱۱۰	۱۲	۱۲

۶-۲-۵- نحوهٔ برچسب دهی دادگان

در این دادگان فقط دنباله‌ای از رشته اعداد بیان شده به عنوان برچسب وجود دارد و هیچگونه تقطیعی در سطح واج یا کلمه وجود ندارد. تنها نکته‌ای که باید به آن توجه کرد آن است که این برچسب دهی باید به نحوی صورت می‌گرفت که علاوه بر راحتی توانایی آن را نیز داشته باشد که به طور خودکار قابل استفاده باشد. در مورد اعداد متصل هیچگونه مشکلی وجود ندارد و برای برچسب رشته اعداد بیان شده به صورت عددی پشت سر هم قرار گرفته‌اند. به عنوان مثال برچسب "۱۰۳۴" بیان کنندهٔ رشتهٔ "یک صفر سه چهار" می‌باشد. در مورد اعداد پیوسته نیز از اعداد به عنوان برچسب استفاده شده است و در اینجا در بین هر سه عدد کاما گذاشته شده است. مثلاً "یک‌هزار و یک‌صد و سه" به صورت "۱,۱۰۳" برچسب گذاری شده است. برای اینکه در اینگونه برچسب‌گذاری بین گفتاری مانند "صد" و "یک‌صد"،"هزار" و "یک‌هزار"،"میلیون" و "یک‌میلیون"،"میلیارد" و "یک‌میلیارد" تفاوت وجود داشته باشد، برای "یک‌صد"، "یک‌هزار"، "یک‌میلیون" و "یک‌میلیارد" از خود اعداد ولی برای "صد"، "هزار"، "میلیون" و "میلیارد" به ترتیب از برچسب‌های W، X، Y و Z استفاده شده است. به

عنوان مثال "یک‌میلیارد و صد میلیون و هزار و یک‌صد و سه" به صورت "۱,W۰۰,۰۰X,۱۰۳" و "ده‌میلیارد و یک‌صد میلیون و یک‌هزار و صد و سه" به شکل "۱۰,۱۰۰,۰۰۱,W۰۳" برچسب گذاری شده‌اند. نحوهٔ برچسب‌دهی بیان شده کاملاً گویا است و به راحتی می‌توان از آن به صورت خودکار در سیستم‌های بازشناسی استفاده کرد.

۶-۳- سیستم بازشناسی اعداد تلفنی

سیستم بازشناسی اعداد تلفنی مبتنی بر سیستم بازشناسی گفتار بیان شده در [۵۱]، [۵۲] می‌باشد. این سیستم مدل‌سازی واحدهای آوایی را با استفاده از مدل مخفی مارکوف انجام می‌دهد. مدل‌های مخفی مارکوف به کاربرده شده از نوع چگالی پیوسته و با تلفیق‌های گوسی می‌باشند و پرش بین حالت‌های آن فقط به صورت چپ به راست مجاز می‌باشد. توپولوژی مدل‌های مربوط به هر واحد آوایی را در این سیستم می‌توان، جداگانه تعیین کرد. تعیین توپولوژی شامل تعیین تعداد حالت، تعداد تلفیق در هر حالت و تعیین پرش‌های مجاز بین حالت‌ها می‌باشد. با استفاده از این مدل‌ها سیستم مذکور تطبیق الگو را برای بازشناسی سیگنال گفتار ورودی انجام می‌دهد و برای این منظور از روش‌های مختلف جستجو برای یافتن بهترین دنبالهٔ واحدهای آوایی متناظر با دنبالهٔ آکوستیکی ورودی بهره می‌گیرد. زبان مورد بازشناسی، مستقل یا وابسته بودن نسبت به گوینده و اندازهٔ مجموعهٔ واژگان مورد بازشناسی در این سیستم، همگی به دادگان گفتار مورد استفاده و مقدار و نوع داده‌های آموزشی بستگی دارند. با آموزش مدل‌ها با دادهٔ آموزشی کافی و متنوع از گوینده‌های مختلف، می‌توان این سیستم را برای بازشناسی مستقل از گوینده به کار برد. واحد آوایی به کار برده شده در این سیستم، کلمه می‌باشد یعنی به ازای هر کلمه یک مدل مخفی مارکوف آموزش داده می‌شود، بنابراین ورودی سیستم یک سیگنال گفتار و خروجی سیستم، بهترین دنبالهٔ عددی یافت شده

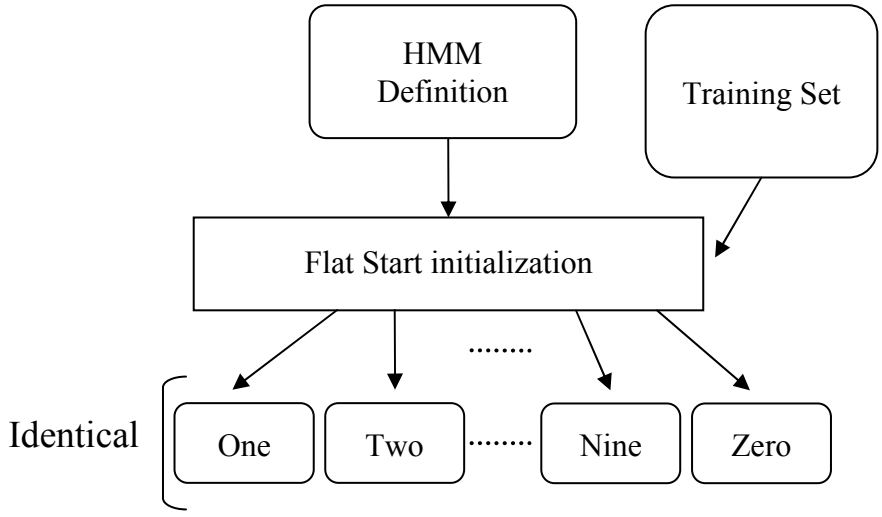
متناظر با دنبالهٔ آکوستیک ورودی می‌باشد. مدل‌های مخفی مارکوف با استفاده از الگوریتم Segmental k-means آموزش داده شده‌اند و توپولوژی آن‌ها برای تمام واحدهای آوایی یکسان در نظر گرفته شده است. در مرحلهٔ تطبیق الگو نیز از الگوریتم جستجوی ویتربی شعاعی برای یافتن بهترین دنبالهٔ کلمات استفاده شده است. همچنین برای حذف DC احتمالی موجود در فایل صوتی ورودی، تابعی به شکل زیر در نظر گرفته شده است که تا حد امکان بتواند آن را حذف کند:

$$y[n] = x[n] - x[n-1] + 0.999 * y[n-1] \qquad (۲-۶)$$

که $x[n]$ ورودی این تابع و $y[n]$ خروجی آن می‌باشد.

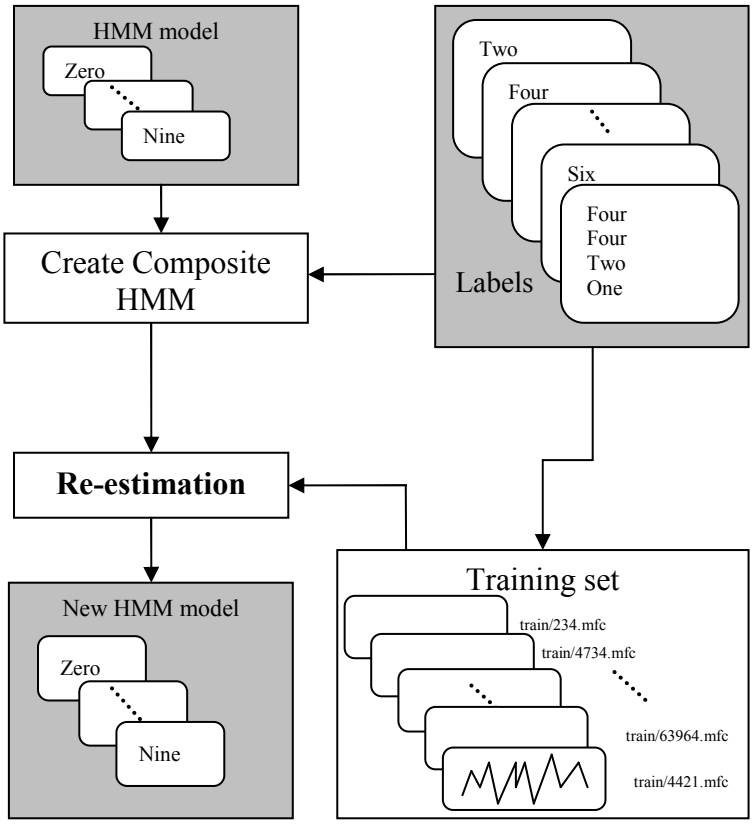
تغییراتی در این سیستم به وجود آمد تا توانایی استفاده از واحدهای آوایی کلمه (عدد) و همچنین دادگانی از رشته‌های عددی بدون تقطیع را داشته باشد. در ادامه به این تغیرات پرداخته می‌شود.

برای آموزش سیستم، در قسمت مقداردهی اولیه از روش Flat Start استفاده شده است [۳۴]. در این روش به حالت‌های همهٔ مدل‌ها یک میانگین و واریانس برابر، به عنوان مقدار اولیه، نسبت داده می‌شود. بلوک‌دیگرام این روش در شکل ۶-۳ آمده است.



شکل ۶-۳- بلوک دیاگرام مقدار دهی اولیه در سیستم بازشناسی اعداد تلفنی

برای تخمین دوبارهٔ مدل‌های به‌دست آمده، با استفاده از همهٔ داده‌های آموزشی، به طور همزمان همهٔ مدل‌ها به‌روز می‌شوند. در ابتدا مدل‌های مخفی مارکوف اولیه که به روش فوق به‌دست آمده‌اند، خوانده می‌شوند. همان‌طور که بیان شد هر یک از فایل‌های صوتی رشته‌های عددی دارای برچسبی معادل با آن چیزی که گوینده بیان کرده است، می‌باشد. با استفاده از این برچسب‌ها، مدل مخفی مارکوف مرکبی ساخته می‌شود که در بر گیرنده کل رشته عددی بیان شده توسط گوینده می‌باشد. این مدل مرکب از پشت سرهم قرار دادن مدل مخفی مارکوف هر یک از کلمات موجود در رشته عددی ساخته می‌شود. سپس الگوریتم جلورو_عقب‌رو [۵] بر روی هر یک از این مدل‌های مرکب اجرا می‌شود و مقادیر متغیرهای جلورو و عقب‌رو محاسبه می‌گردند. هنگامی‌که همهٔ فایل‌های موجود در مجموعه آموزش پردازش شدند، پارامترهای تخمین جدید با روش بام-ولش [۶]، از روی متغیرهای جلورو و عقب‌رو و مدل‌های اولیه محاسبه می‌شوند. مراحل به‌دست آوردن تخمین جدید در شکل ۶-۴ آمده است.



شکل ۶-۴- بلوک دیاگرام مراحل به‌دست آوردن تخمین جدید در سیستم بازشناسی اعداد تلفنی

فصل هفتم

آزمایش‌ها و نتایج بدست‌آمده

۷-۱- مقدمه

در این فصل، ابتدا در مورد بستر آزمایش‌های انجام شده و پارامترهای مختلف سیستم بازشناسی گفتار توضیح مختصری داده خواهد شد، سپس به آزمایش‌های گوناگونی که برای مقاوم‌سازی مبتنی بر ویژگی در این سیستم بازشناسی برای مقابله با نویز خط تلفن و همچنین آزمایش‌هایی که بر روی دادگان فارسی آماده شده صورت گرفته می‌پردازیم و نتایج حاصل از آن را بررسی می‌کنیم.

۷-۲- بستر انجام آزمایش‌ها

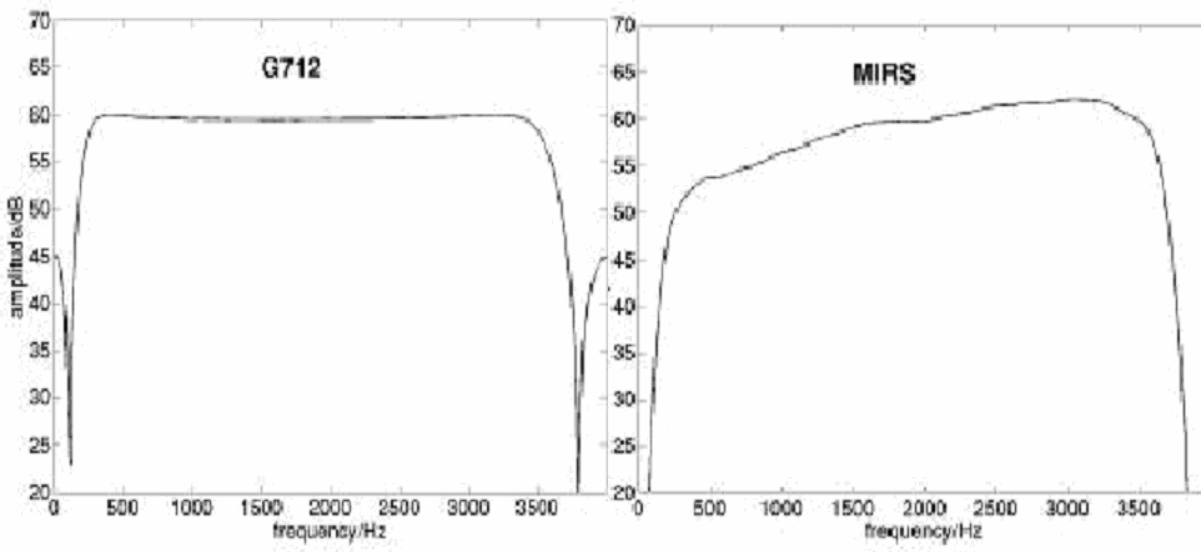
برای انجام آزمایش‌ها، از سیستم بازشناسی اعداد تلفنی شرح داده شده در فصل قبل و دادگان گفتاری FCCDigits، دادگان فارس‌دات تلفنی[۳۱] و AURORA 2 [۳۲] استفاده گردیده است.

دادگان فارس‌دات تلفنی توسط پژوهشکدهٔ پردازش هوشمند علایم در سال ۱۳۸۰تهیه شده است. این دادگان در سطح واج تقطیع شده است و دارای تقریباً هشت ساعت گفتار از ۶۴ گوینده می‌باشد. این دادگان هم در بر دارندهٔ گفتار فی‌البداهه و هم گفتار رسمی می‌باشد. از آنجایی که این دادگان در بر دارندهٔ گفتار فی‌البداهه نیز می‌باشد هنگام برچسب گذاری از برچسب‌های غیرزبانی نیز استفاده شده‌است تا بتوان آن قسمت‌هایی از گفتار را که به زبان مربوط نمی‌باشند نیز برچسب گذاری کرد. گفتار فی‌البداهه این دادگان شامل سلام و احوالپرسی، نام و نام‌خانوادگی، میزان تحصیلات، سن، نام شهر یا استان محل زندگی، وضعیت اتاقی که در آن مستقر هستند، توضیح در مورد شهری که در آن مستقر هستند و یک خاطره تلخ یا شیرین می‌باشد و دادگان گفتار رسمی آن نیز شامل اعداد (صفر، یک، ...، بیست، سی،...، صد، دویست، پانصد، پونصد، هزار، میلیون و میلیارد)، ایام هفته، ماه‌های سال، حروف الفبا، پنجاه کلمه پر بسامد زبان

فارسی به طور تصادفی، دو جمله ثابت، چهار جمله تصادفی ازدادگان فارس‌دات [۳۳] و هجاهای باز زبان فارسی.

دادگان AURORA 2 برای بررسی میزان دقت سیستم‌های بازشناسی صحبت در محیط‌های نویزی تهیه شده است و برای بررسی کارایی بسیاری از الگوریتم‌های استخراج ویژگی مقاوم در شرایط نویزی مختلف مورد استفاده قرار می‌گیرد. این دادگان از دادگان گفتاری TIDigits استفاده می‌کند که از اعداد متصل که به وسیله گوینده‌های انگلیسی زبان آمریکایی بیان شده‌اند، تشکیل شده است. در این دادگان ابتدا TIDigits به ۸ کیلو هرتز زیرنمونه‌برداری[1] شده است و پس از آن فیلتری جهت شبیه‌سازی خط تلفن بر روی دادگان اعمال گردیده است. دو نوع فیلتر به نام‌های G.712 و MIRS که در شکل ۷-۱ آمده‌اند، برای انجام کار شبیه سازی استفاده شده است. گفتار حاصل از اعمال فیلترهای فوق در این دادگان گفتار تمیز دادگان را تشکیل می‌دهند. سپس روی این دادگان تمیز نویزهایی با نسبت سیگنال به نویزهای مختلف بطور جداگانه به دادگان اضافه گردیده است تا مجموعه‌هایی با نسبت سیگنال به نویزهای ۵-، ۰، ۵، ۱۰، ۱۵و ۲۰ دسی‌بل را بسازند. نویزهای مورد استفاده عبارتنداز نویز همهمه(babble)، رستوران(restaurant)، نمایشگاه(exhibition)، ماشین(car)، خیابان(street)، فرودگاه(airport)، مترو(subway) و نویز ایستگاه قطار(train-station).

[1] *Sub Sample*



شکل ۷-۱- پاسخ فرکانسی فیلترهای G.712 و MIRS [۳۲]

در این دادگان مجموعه‌های آموزش و آزمایش بطور جداگانه مشخص شده‌اند. داده‌های آموزشی این دادگان خود بر دو نوع دادگان آموزشی تمیز و دادگان آموزشی ترکیبی (شامل تمیز و نویزی) می‌باشد. این دو مجموعه آموزشی هر دو به وسیلهٔ فیلتر G.712 فیلتر شده‌اند و شامل ۸۴۴۰ رشته عددی می‌باشند. اما در دادگان آموزشی نوع اول، چون داده‌ها با هیچ نوع نویزی جمع نشده‌اند آنرا به عنوان مجموعه تمیز می‌شناسند. در دومین دادگان آموزشی همان ۸۴۴۰ رشته عددی به بیست زیر مجموعهٔ ۴۲۲ تایی دسته‌بندی شده‌اند که این بیست زیر مجموعه در بر گیرندهٔ چهار نوع نویز مترو، همهمه، ماشین و نمایشگاه در چهار نسبت سیگنال به نویز مختلف ۵ ، ۱۰ ، ۱۵ و ۲۰ دسی‌بل و چهار زیر مجموعهٔ بدون نویز می‌باشند.

دادگان آزمایش نیز خود شامل سه مجموعه A، B و C است. دادگان آزمایشی A شامل ۴۰۰۴ جمله فیلتر شده با فیلتر G.712 می‌باشد که به چهار زیر مجموعه ۱۰۰۱ جمله‌ای تقسیم‌بندی شده‌اند. علاوه بر سیگنال‌های فیلتر شده این زیرمجموعه‌ها، به هر یک از این زیر مجموعه‌ها نویزهای مترو، همهمه، ماشین و نمایشگاه اضافه شده‌اند که برای هر دسته نویزی هم مجموعه‌هایی در نسبت‌های مختلف سیگنال به نویز ۵-، ۰، ۵، ۱۰، ۱۵و ۲۰ دسی‌بل وجود دارند تا با هم هفت حالت مختلف را تشکیل دهند. دادگان آزمایش

B نیز دقیقاً شبیه به روش فوق به‌دست آمده‌اند با این تفاوت که در اینجا از نویزهای رستوران، خیابان، فرودگاه و نویز ایستگاه قطار استفاده شده است. در این حالت عدم انطباق بین دادگان آموزش و آزمایش حتی برای حالت آموزش به وسیلهٔ دادگان آموزشی ترکیبی هم وجود دارد، چرا که دادگان آموزشی ترکیبی فقط شامل نویزهای مجموعه آزمایش A است. در دادگان آزمایش C دو زیر مجموعه از چهار زیر مجموعهٔ ۱۰۰۱ جمله‌ای وجود دارند. در این مجموعه داده‌های گفتار و نویز به وسیلهٔ فیلتر MIRS، فیلتر شده‌اند و سپس نویزهای مترو و خیابان با نسبت سیگنال به نویزهای ۵-، ۰، ۵، ۱۰، ۱۵و۲۰ به گفتار اضافه شده است. در اینجا نیز حالت تمیز، یعنی بدون حضور نویز جمع‌شونده وجود دارد. این دسته از دادگان آزمایش برای نشان‌دادن اثر نویز کانال‌های متفاوت بر روی سیستم تشخیص گفتار، تهیه شده است.

نتیجهٔ خروجی با مقایسهٔ دنبالهٔ کلمات حاصل از بازشناسی (که با خطا همراه می‌باشد) و دنبالهٔ کلمات صحیح و محاسبهٔ تعداد خطاهای درج[1]، حذف[2] و جایگزینی[3]، با استفاده از برنامه‌نویسی پویا[4] به دست می‌آیند. با توجه به انواع خطاها، درصد دقت و درصد صحت بازشناسی به ترتیب با روابط زیر محاسبه می‌شوند:

$$Accuracy\ rate = \frac{T-(D+I+S)}{T} \qquad (--۷)$$

$$Correctness\ rate = \frac{T-(D+S)}{T} \qquad (--۷)$$

در روابط بالا، I، D و S به ترتیب تعداد خطاهای درج، حذف و جایگزینی و T تعداد کل کلمات دنبالهٔ صحیح می‌باشد.

[1] *Insertion*
[2] *Deletion*
[3] *Substitution*
[4] *Dynamic Programming*

۷-۳- نتایج اولیه

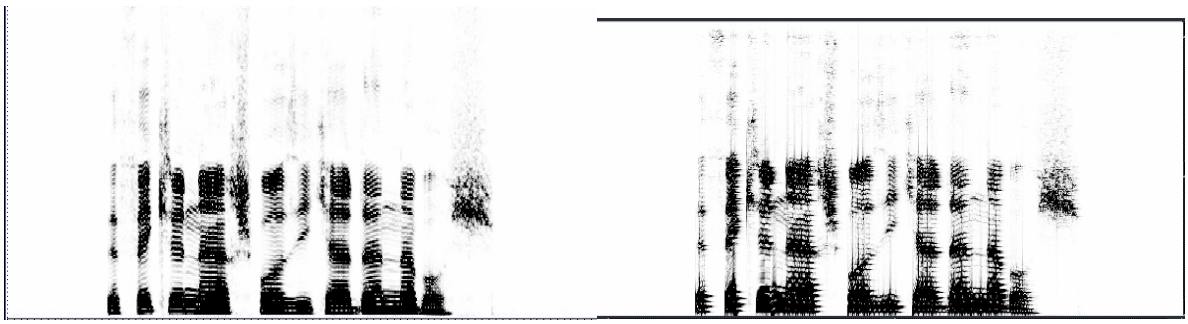
هدف از این سری از آزمایش‌ها امکان‌سنجی استفاده از دادگان فارس‌دات تلفنی برای بدست آوردن نتایج قابل قبول برای یک سیستم بازشناسی اعداد تلفنی بود. همانطور که بیان گردید این دادگان در سطح واج تقطیع شده است و معمولاً برای سیستم‌های بازشناسی گفتار که واحد آوایی آن‌ها واج می‌باشد مورد استفاده قرار می‌گیرد. از آنجایی که در این دادگان فرکانس تکرار هر کلمه، به خصوص اعداد، کم می‌باشد از آن نمی‌توان برای سیستم‌های بازشناسی مبتنی بر کلمه استفاده کرد. نتایج موجود در جدول ۷-۱ گواهی بر این ادعا است. در اینجا به صورت جداگانه آزمایش‌هایی برای اعداد متصل و پیوسته صورت گرفت. در آزمایش اول سیستم بازشناسی گفتار بر اساس مدل آوایی اعداد متصل و با دادگان فارس‌دات تلفنی آموزش داده شد و برای این‌که بتوان مقایسهٔ معنی‌داری بین این آزمایش و آزمایش‌های بعدی صورت بگیرد، می‌بایستی مجموعهٔ آزمون از خارج از این دادگان انتخاب شد. بدین منظور ۱۰۲ رشتهٔ عددی متصل (در کل شامل ۴۰۶ عدد) از طریق خط تلفن ضبط گردید و سیستم بازشناسی با آن‌ها آزمایش شد. در آزمایش دوم سیستم برای بازشناسی اعداد پیوسته آموزش داده شد و دادگان آزمون نیز اعداد پیوسته انتخاب گردیدند که خارج از دادگان آموزش به طور جداگانه ضبط گردیدند.

جدول ۷-۱- دقت بازشناسی حاصل از آموزش سیستم با دادگان TFarsdat برای بازشناسی اعداد متصل و پیوسته

نوع اعداد	کلمه		دقت بازشناسی رشته‌های عددی
	دقت	صحت	
اعداد متصل	٪۹۱,۸۷	٪۹۲,۶۱	٪۷۰,۵۹
اعداد پیوسته	٪۸۹,۵۶	٪۸۹,۹۳	٪۶۸,۹۲

۷-۴- نتایج حاصل از مقایسه روش پیشنهادی مبتنی بر فیزیولوژی گوش با MFCC

قبل از اینکه مدل پیشنهادی به عنوان روشی برای استخراج ویژگی در سیستم بازشناسی گفتار مورد استفاده قرار گیرد مقایسه‌ای بین تبدیل یافتهٔ یک سیگنال گفتار با این روش و طیف آن در شکل ۷-۲ آمده است. همانطوری که این شکل نشان می‌دهد تفاوت اندکی بین این دو تصویر وجود دارد و بیان کنندهٔ این نکته است که این روش علاوه بر کاربرد به عنوان استخراج ویژگی توانایی این را هم دارد که از آن به جای FFT یا تبدیل موجک[1] در کاربردهای مختلف تبدیل سیگنال بتوان استفاده کرد.



شکل ۷-۲- مقایسه تبدیل حاصل از روش پیشنهادی با طیف سیگنال گفتار

شکل سمت راست نشان‌دهندهٔ تبدیل یک سیگنال گفتار با مدل پیشنهادی است
و شکل سمت چپ طیف همان سیگنال گفتار را نشان می‌دهد.

برای نشان دادن تأثیر این بازنمایی آزمایش‌هایی برای مقایسه با بازنمایی‌های MFCC طراحی شد. برای آموزش سیستم بازشناسی اعداد از دادگان اعداد متصل زبان انگلیسی که شامل ۱۸ گوینده که هر گوینده ۷۷ رشته عددی را بیان می‌کند، استفاده شد. هنگام آزمایش هم از ۲۰۰ رشته عددی که به وسیلهٔ گوینده‌هایی غیر از گوینده‌های داده‌ای آموزش بیان شده‌اند، استفاده شد. دادگان آزمون به چهار دستهٔ ۵۰تایی تقسیم شدند و به هر دسته چهار نوع نویز مختلف با SNRهای متفاوت اضافه شد. در

[1] *Wavelet*

جدول ۷-۲ خلاصهٔ نتایج حاصل از آزمایش‌ها آمده است. همانطور که در این جدول مشخص است در اغلب موارد آزمایش بهبودهایی در نتیجه حاصل از بازشناسی به دست آمد.

جدول ۷-۲- نتایج حاصل از مقایسه بازنمایی‌های MFCC با بازنمایی‌های مبتنی بر مدل فیزیولوژیک

Word Error Rate								
Noise	**Feature**	**20 db**	**15 db**	**10 db**	**5 db**	**0 db**	**-5 db**	**Overall**
Car	MFCC	13.29	24.05	53.16	80.38	93.67	93.67	59.7
	Bio Model	13.29	23.42	55.06	74.05	82.28	93.67	56.96
Exhibition	MFCC	37.75	48.34	59.6	77.48	90.73	91.39	67.55
	Bio Model	33.77	36.42	64.24	78.15	86.09	89.4	64.68
Babble	MFCC	26.95	44.31	72.46	88.02	94.61	99.4	70.96
	Bio Model	23.95	34.73	63.47	76.05	83.23	88.62	61.68
Subway	MFCC	46.75	60.36	71.01	88.17	89.94	90.53	74.46
	Bio Model	47.93	57.4	72.78	83.43	87.57	91.12	73.37

۷-۵- نتایج حاصل از بکارگیری ویژگی‌های مقاوم در بازشناسی تلفنی اعداد انگلیسی

در این بخش روش‌های مختلف استخراج ویژگی‌های مقاوم به همراه نتایج حاصل از پیاده سازی آنها روی سیستم بازشناسی اعداد انگلیسی بررسی می‌شوند. برای انجام آزمایش‌ها از دادگان گفتاری AURORA 2 استفاده شده است که همانطور که گفته شد شامل دادگان تمیز و نویزی با اثر کانال و نویزهای جمع‌شونده مختلف در نسبت سیگنال به نویزهای مختلف می‌باشد. در این آزمایش‌ها هر عدد زبان انگلیسی با یک HMM دارای ۱۶ حالت و ۳ تلفیق گوسی مدل شد و از آن‌جا که ممکن است گوینده در فاصلهٔ ادای هر رقم کمی مکث کرده باشد، یک مدل سکوت نیز آموزش داده شده تا مکث‌های گوینده در حین صحبت را پوشش دهد. همچنین MFCC به عنوان روش مبنای استخراج ویژگی استفاده شده است که درصد دقت بازشناسی سیستم با آن برای مجموعه‌های آزمایش A و B در دادگان، برای نویزهای مختلف و در نسبت سیگنال به نویزهای متفاوت به صورت جدول ۷-۳ آورده شده است.

جدول ۷-۳- نتایج بازشناسی با MFCC (مبنا)

SNR ▸ / NOISE ▾	CLEAN	20	15	10	5	0	-5	AVG
SUBWAY	98.9	97.1	93.5	78.7	52.2	26.0	11.2	65.4
BABBLE	99.0	90.2	73.8	49.4	26.8	9.3	1.6	50.0
CAR	99.0	97.4	90.0	67.0	34.1	14.5	9.4	58.8
EXHIBITION	99.2	96.4	92.0	75.7	44.8	18.1	9.6	62.3
RESTAURANT	98.9	90.0	76.2	54.8	31.0	11.0	3.5	52.2
STREET	99.0	95.7	88.5	67.1	38.5	17.8	10.5	59.6
AIRPORT	99.0	90.6	77.0	53.9	30.3	14.4	8.2	53.3
TRAIN STATION	99.2	94.7	83.7	60.3	27.9	11.6	8.5	55.1

بکارگیری روش میانگین ضرایب کپسترال روی دادگان آزمایش مورد استفاده در این بخش نتایج جدول ۷-۴ را به همراه داشته است. اعمال این روش، به غیر از کاهش اندک نتایج در نویزهای مترو و نمایشگاه، بهبود قابل توجهی را برای سایر نویزها به دنبال داشته است.

جدول ۷-۴- نتایج بازشناسی با MFCC_MN یا CMS

SNR ▸ / NOISE ▾	CLEAN	20	15	10	5	0	-5	AVG
SUBWAY	99.0	96.3	91.6	77.7	51.0	24.7	14.0	64.9
BABBLE	99.2	97.6	93.7	83.5	58.2	25.2	12.2	67.1
CAR	99.0	97.1	93.0	78.1	46.7	21.1	12.9	64.0
EXHIBITION	99.2	95.3	89.5	74.5	46.2	20.3	10.5	62.2
RESTAURANT	99.0	97.5	94.4	85.2	63.1	30.7	12.6	68.9
STREET	99.2	97.2	93.3	81.9	56.0	25.4	12.7	66.5
AIRPORT	99.0	97.3	94.9	86.9	63.3	31.7	15.4	69.8
TRAIN STATION	99.2	97.8	94.0	82.9	56.2	25.4	12.7	66.9

از آنجا که روش ضرایب کپسترال حاصل از پیشگویی خطی مبتنی بر درک انسان بر پایهٔ طیف زمان کوتاه گفتار می‌باشد، مشابه سایر تکنیک‌های مبتنی بر طیف زمان کوتاه، در مقابل اثرات کانال‌های مخابراتی آسیب‌پذیر است که نتایج بکارگیری آن‌هم این مسأله را نشان می‌دهد. نتایج حاصل از بکارگیری آن هم در جدول ۷-۵ نشان داده شده است. استفاده از این روش منجر به کاهش نتیجه نسبت به حالت مبنا در همه شرایط نویزی شده است. یک راه بهبود کارایی این روش در حضور نویز کانال، اعمال یک فیلتر بالاگذر یا میان‌گذر است که منجر به روش تحلیل طیف نسبی می‌شود و نتایج آن در ادامه آمده‌اند.

جدول ۷–۵– نتایج بازشناسی با PLP

SNR ▸ / NOISE ▾	CLEAN	20	15	10	5	0	-5	AVG
SUBWAY	98.8	97.3	91.7	75.7	48.1	25.6	12.6	64.3
BABBLE	99.1	89.3	70.7	46.0	20.7	4.7	-0.3	47.2
CAR	98.9	95.4	81.2	54.3	25.8	10.3	7.3	53.3
EXHIBITION	99.3	95.9	87.5	65.1	30.8	10.6	6.7	56.5
RESTAURANT	98.8	88.5	71.3	47.4	22.3	3.2	-2.0	47.1
STREET	99.1	95.4	85.8	62.4	34.0	14.1	7.0	56.8
AIRPORT	98.9	90.0	73.6	48.1	23.3	8.5	2.9	49.3
TRAIN STATION	99.3	92.8	77.1	50.9	22.7	8.0	7.0	51.1

نتایج استفاده از روش RASTA-PLP در جدول ۷–۶ آورده شده است. این روش برای نویزهای مترو، نمایشگاه و ماشین نتایج را نسبت به حالت مبنا پایین آورده است.

جدول ۷–۶– نتایج بازشناسی با RASTA-PLP

SNR ▸ / NOISE ▾	CLEAN	20	15	10	5	0	-5	AVG
SUBWAY	98.7	95.6	89.8	70.4	39.2	20.0	12.2	60.8
BABBLE	98.9	96.4	91.9	75.5	41.8	19.5	11.2	62.2
CAR	98.8	96.4	89.7	62.7	29.2	17.5	9.7	57.7
EXHIBITION	99.1	95.7	89.0	70.4	34.3	16.0	8.3	59.0
RESTAURANT	98.7	96.3	92.1	79.4	49.6	23.6	11.7	64.5
STREET	98.9	96.3	89.8	71.1	40.9	19.3	10.6	61.0
AIRPORT	98.8	96.6	92.5	79.1	47.7	24.5	13.4	64.6
TRAIN STATION	99.1	96.9	90.9	72.4	39.6	20.2	10.9	61.4

همانطور که در فصل ۵ بیان شد عملکرد روش J-RASTA به شدت به مقدار J وابسته است و می‌بایست برای هر سطح نویز، مقدار بهینهٔ پارامتر J را بدست آورد. نتایج اعمال این روش به صورت نشان داده شده در جدول ۷–۷ می‌باشد که منجر به بهبود خوبی برای برخی نویزها و به طور مشخص برای نویز ماشین شده است، روی نویز همهمه تقریباً بدون تاثیر بوده و نتایج نویزهای نمایشگاه و رستوران را پایین آورده است. مقدار J بکار گرفته شده در اینجا $1.6*10^{-6}$ بوده است.

جدول ۷-۷- نتایج بازشناسی با J-RASTA

SNR ▸ / NOISE ▾	CLEAN	20	15	10	5	0	-5	AVG
SUBWAY	98.6	96.1	90.4	77.5	60.2	37.6	16.2	68.1
BABBLE	98.8	86.4	71.1	52.3	31.1	9.4	-6.8	48.9
CAR	98.8	97.3	94.7	86.9	70.0	41.6	15.0	72.0
EXHIBITION	99.1	90.0	79.9	63.4	40.4	15.7	0.6	55.6
RESTAURANT	98.6	82.2	66.2	49.7	28.5	10.0	-5.9	47.0
STREET	98.8	94.1	88.0	74.6	55.2	30.4	12.7	64.8
AIRPORT	98.8	89.4	80.0	63.1	43.2	22.6	4.2	57.3
TRAIN-STATION	99.1	93.2	88.5	78.4	58.3	33.8	12.3	66.2

یکی دیگر از راه‌حل‌هایی که برای استفاده از ضرایب کپسترال در محیط‌های نویزی به‌کار گرفته شد روش ضرایب کپسترال ریشه‌ای بود. همانطور که بیان شد انجام این کار در کنار بهبود کارایی سیستم، منجر به بهبود سرعت استخراج ویژگی‌ها می‌شود. نتایج بازشناسی حاصل با استفاده از این روش در جدول ۷-۸ آورده شده است که در صورت اعمال تفاضل میانگین روی این ضرایب، نتایج به صورت جدول ۷-۹ در می‌آیند. اعمال RCC به تنهایی، مشابه حالت قبل، برای نویزهای مترو و نمایشگاه منجر به کاهش نتیجه شده است و برای سایر موارد هم بهبود نسبتاً اندکی را به همراه داشته است ولی وقتی با روش تفاضل میانگین ضرایب ترکیب شود منجر به بهبود قابل توجهی در همهٔ شرایط نویزی شده است.

جدول ۷-۸- نتایج بازشناسی با RCC

SNR ▸ / NOISE ▾	CLEAN	20	15	10	5	0	-5	AVG
SUBWAY	99.1	94.2	81.8	56.7	32.4	16.4	9.6	55.7
BABBLE	99.0	95.2	84.0	60.0	29.2	6.7	-0.9	53.3
CAR	98.8	97.8	94.0	77.1	41.3	14.1	5.3	61.2
EXHIBITION	99.1	94.9	87.8	67.1	34.2	13.5	7.6	57.7
RESTAURANT	99.1	93.8	82.3	59.6	29.0	6.4	-1.1	52.7
STREET	99.0	96.2	90.1	71.8	42.7	20.0	9.6	61.3
AIRPORT	98.8	94.8	85.1	63.6	32.6	9.5	-1.8	54.6
TRAIN-STATION	99.1	95.1	87.5	70.0	36.3	10.3	0.9	57.0

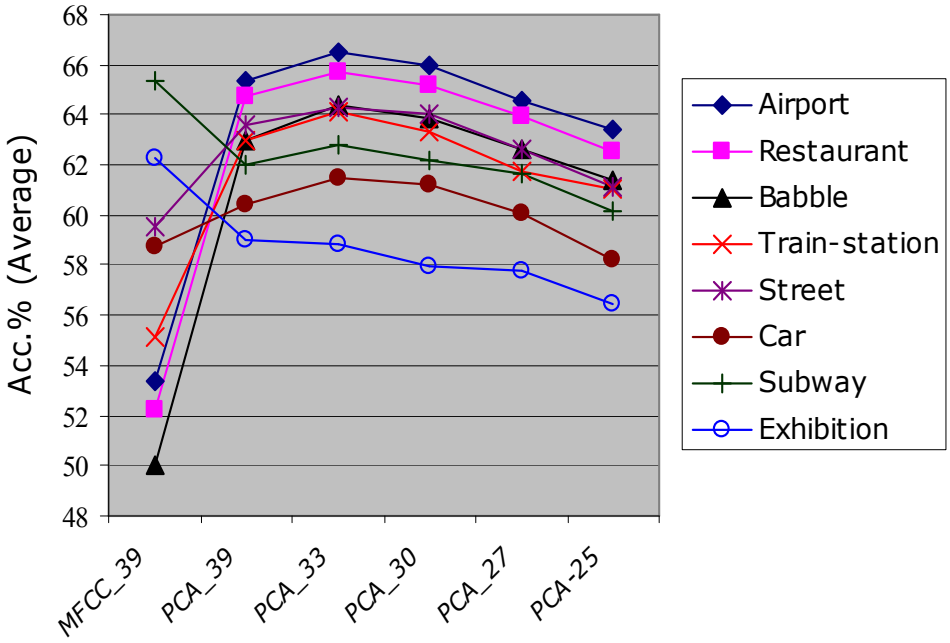
جدول ۷-۹- نتایج بازشناسی با RCC_MN

SNR ▸ / NOISE ▾	CLEAN	20	15	10	5	0	-5	AVG
SUBWAY	99.1	97.0	93.8	83.1	61.3	30.2	12.8	68.2
BABBLE	99.0	97.8	92.8	81.4	56.1	26.0	10.9	66.3
CAR	99.0	98.0	95.5	86.9	64.1	26.8	10.5	68.7
EXHIBITION	99.2	95.7	90.9	78.4	51.0	19.8	9.4	63.5
RESTAURANT	99.1	97.3	91.8	79.8	57.2	28.3	11.6	66.5
STREET	99.0	97.1	93.8	84.0	61.6	30.1	11.9	68.2
AIRPORT	99.0	98.0	95.1	86.6	64.9	33.9	13.5	70.1
TRAIN-STATION	99.2	97.4	94.0	84.9	62.9	29.4	11.7	68.5

در آزمایش‌های انجام شده برای ارزیابی نتایج حاصل از PCA این نتیجه حاصل شد که استفاده از میانگین ضرایب کپسترال نتایجی به مراتب بهتر از اعمال PCA به تنهایی خواهد داشت. همچنین ذکر این نکته ضروری است که نتایج بیان شده در این قسمت، از اعمال این روش به عنوان آخرین بلوک مراحل استخراج ویژگی استفاده شده است. نتایج کار برای نویزهای مختلف به صورت میانگین دقت روی نسبت سیگنال به نویزهای مختلف در شکل ۷-۳ به همراه نتیجه مبنای آنها آورده شده است. این نتایج نشان می‌دهند که تعداد بهینه در اینجا ۳۳ ویژگی است. این کاهش تعداد علاوه بر بهبود دقت، بعلت کاهش محاسبات مراحل بعد، باعث افزایش سرعت سیستم نیز می‌شود نتایج کامل اعمال این تبدیل در حالت استفاده از ۳۳ ویژگی در جدول ۷-۱۱ آورده شده است. مشابه اغلب روش‌های مورد بررسی، در اینجا نیز، برای نویزهای مترو و نمایشگاه نتایج کاهش یافته است ولی در سایر موارد بهبودهای قابل توجهی بدست آمده است.

جدول ۷-۱۰- نتایج بازشناسی با MFCC-PCA-MN با ۳۳ ویژگی

SNR / NOISE	CLEAN	20	15	10	5	0	-5	AVG
SUBWAY	98.8	95.8	90.1	74.1	45.9	21.7	13.3	62.8
BABBLE	98.8	97.1	91.6	79.3	50.5	21.8	11.8	64.4
CAR	98.6	96.4	916	73.6	41.2	18.6	10.5	61.5
EXHIBITION	98.8	94.0	85.3	68.4	38.8	17.0	9.3	58.8
RESTAURANT	98.8	96.4	92.4	81.2	54.9	24.8	11.3	65.7
STREET	98.8	96.7	90.9	77.3	51.4	22.9	11.9	64.3
AIRPORT	98.6	97.1	93.3	81.9	54.9	26.9	12.9	66.5
TRAIN-STATION	98.8	96.8	92.2	78.7	48.9	22.1	11.5	64.2



شکل ۷-۳- میانگین دقت سیستم در روش PCA-MN برای تعداد مختلف ویژگی‌ها و مقایسه آنها با روش مبنای MFCC

۷-۶- عملکرد ویژگی‌های مقاوم در بازشناسی تلفنی اعداد انگلیسی با تغییر کانال

نتایج ارایه شده در فوق برای ترکیب نویز کانال G.712 و نویزهای جمع‌شونده در نسبت سیگنال به نویزهای مختلف بود. برای بررسی کارایی این روش‌ها در شرایطی که کانال انتقال صحبت تغییر می‌کند از مجموعه آزمایش C که فقط دو نوع نویز مترو و خیابان را دارد و کانال شبیه سازی آن MIRS می‌باشد،

استفاده شده است. نتایج بازشناسی روی این مجموعه برای دو نوع نویز مترو و خیابان به صورت نشان داده شده در جدول‌های ۷-۱۱ و ۷-۱۲ می‌باشد. روش‌های J-RASTA و RCC-MN منجر به بهبود نتایج در هر دو حالت نویزی شده‌اند. به علاوه روش MFCC-MN یا CMS در نویز خیابان منجر بـه بـالا بـردن دقـت بازشناسی شده است. در سایر موارد اعمال روش‌ها باعث کاهش نتیجه شده است. استفاده از روش تفاضـل میانگین ضرایب چه به تنهایی و چه در کنار سایر روش‌ها بهبود قابل ملاحظه‌ای را به دنبال داشته است کـه می توان این مسأله را در تفاوت نتایج بکارگیری RCC و RCC-MN مشاهده کرد.

جدول ۷-۱۱- نتایج بازشناسی با روش‌های مختلف استخراج ویژگی برای نویز مترو در کانال MIRS

SNR / METHOD	CLEAN	20	15	10	5	0	-5	AVG
MFCC-39	99.1	93.5	86.8	73.9	51.3	25.4	11.8	63.1
RCC	99.1	91.8	79.8	54.1	26.1	11.0	8.1	52.9
RASTA-PLP	98.6	95.6	90.0	72.3	40.2	19.8	12.0	61.2
PCA MN-33	98.8	95.6	88.9	71.0	42.4	20.7	11.7	61.3
PLP	98.8	93.8	87.6	74.4	50.3	20.5	9.1	62.1
MFCC-MN	99.0	95.9	89.6	72.9	44.5	22.3	12.2	62.3
RCC-MN	99.1	96.1	91.5	78.7	55.6	27.2	11.6	65.7
J-RASTA	98.7	97.1	93.2	84.3	65.8	39.2	18.3	70.9

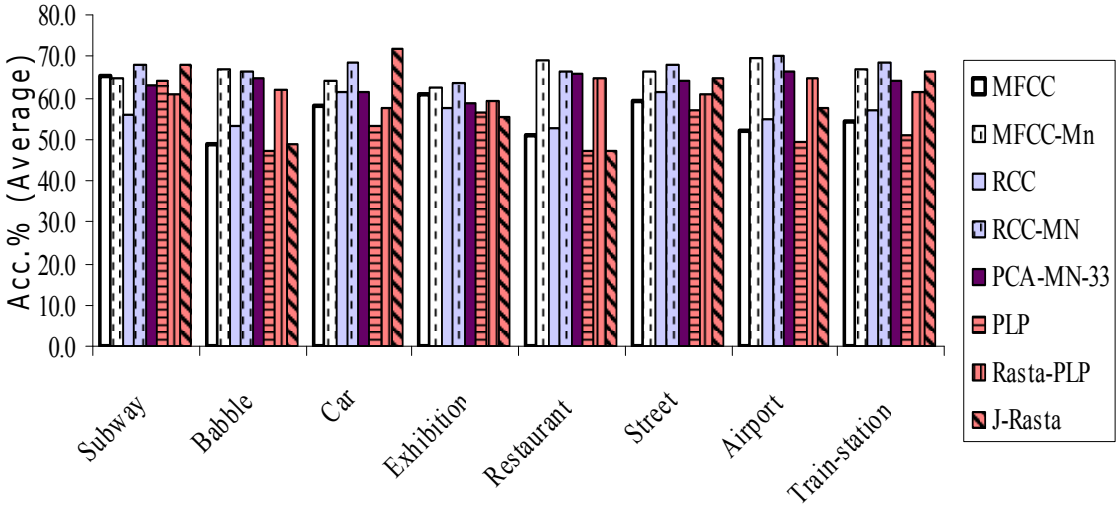
جدول۷-۱۲- نتایج بازشناسی با روش‌های مختلف استخراج ویژگی برای نویز خیابان در کانال MIRS

SNR / METHOD	CLEAN	20	15	10	5	0	-5	AVG
MFCC-39	99.0	95.1	88.9	74.4	49.2	22.9	11.2	63.0
RCC	99.0	94.0	87.0	72.4	49.4	24.0	11.5	62.5
RASTA-PLP	98.9	95.7	90.6	71.7	39.3	20.1	10.6	61.0
PCA MN-33	98.9	95.8	90.5	76.4	48.6	22.4	11.8	63.5
PLP	99.2	94.7	88.8	74.2	51.5	23.9	10.3	63.2
MFCC-MN	99.2	97.0	92.6	78.0	51.6	24.5	11.8	64.9
RCC-MN	99.0	96.6	92.3	80.2	57.4	28.2	13.4	66.7
J-RASTA	98.6	94.4	89.1	76.9	58.2	33.0	15.5	66.5

۷-۷- مقایسه و جمع بندی نتایج حاصل از اعمال انواع مختلف ویژگی‌های مقاوم

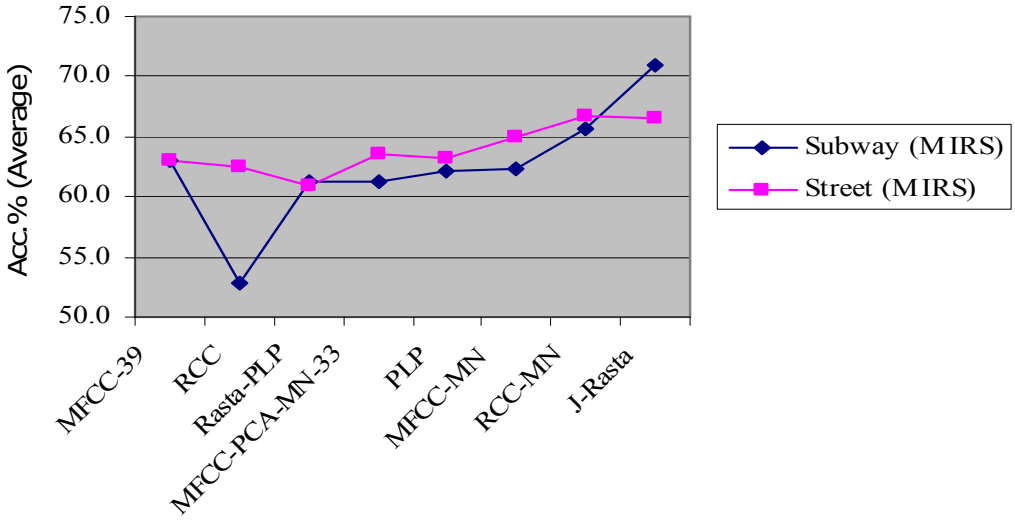
مسألهٔ مقاوم‌سازی سیستم‌های بازشناسی گفتار در این بخش با ارزیابی روش‌های مختلـف اسـتخراج ویژگی در شرایط مختلف نویزی مورد بررسی قرار داده شد. استفاده از تفاضل میانگین ضرایب کـلاً باعـث

بهبود کارایی سیستم بازشناسی در شرایط نویزی موجود شده است. سایر روش‌ها هم در برخی شرایط نویزی نسبت به برخی شرایط دیگر بهتر جواب داده اند. برای نویزهای مترو و نمایشگاه، استفاده از روش‌های مورد بررسی بجز در یکی- دو مورد آن‌هم بصورت جزیی، نتایج بازشناسی را پایین می‌آورد که علت این مسأله می‌تواند دارا بودن انرژی نسبتاً یکسان این دو نوع نویز در فرکانس‌های مختلف باشد در حالی که انرژی سایر نویزها بیشتر در فرکانس‌های پایین می‌باشد. مقایسه میانگین نرخ دقت بازشناسی روی نسبت سیگنال به نویزهای مختلف در شرایط نویزی مختلف، برای روش‌های مورد بررسی در این بخش بر روی مجموعه‌های آزمایش A و B به صورت نشان داده در شکل ۷-۴ می‌باشد.



شکل ۷-۴- مقایسه میانگین دقت بازشناسی (در نسبت سیگنال به نویزهای مختلف) روش‌های مختلف استخراج ویژگی برای شرایط نویزی متفاوت

بصورت مشابه، مقایسه کارایی این روش‌ها در شرایط تغییر کانال، با آزمایش روی مجموعه آزمایش C در شکل ۷-۵ آورده شده است. نتایج این مقایسه هم بیانگر توانایی خوب روش تفاضل میانگین ضرایب در حضور این نوع نویز است. در کنار این روش، روش J-RASTA هم به نسبت سایر روش‌ها موفق‌تر عمل کرده است.



شکل ۷–۵– مقایسه روش‌های مختلف استخراج ویژگی با تغییر کانال

۷–۸– نتایج حاصل از بازشناسی تلفنی اعداد فارسی

هدف از این آزمایش‌ها ارزیابی کارایی سیستم بازشناسی گفتار اعداد فارسی می‌باشد. در سری اول آزمایش‌های این بخش ابتدا دادگان اعداد متصل و پیوسته به دو مجموعهٔ آموزش و آزمایش تقسیم شدند. از بین ۱۱۰ گوینده موجود در دادگان ۱۳ گوینده شامل ۸ گویندهٔ مرد و ۵ گویندهٔ زن به عنوان گویندگان مجموعهٔ آزمایش به طور کاملاً تصادفی انتخاب گردیدند. مجموعهٔ آموزش اعداد متصل شامل ۷۲۱۶ رشتهٔ عددی می‌باشد که دارای ۲۴۰۹۶ کلمه (عدد) است. در مجموعهٔ آزمایش اعداد متصل نیز ۹۷۵ رشته عددی شامل ۳۲۶۳ کلمه (عدد) وجود دارد. از آنجایی که در فرکانس‌های بالاتر از ۸ کیلوهرتز عملاً هیچگونه داده‌ای وجود ندارد، فایل‌های این دادگان جهت انجام آزمایش‌ها به ۸ کیلوهرتز زیرنمونه برداری شده‌اند. نتایج حاصل از بازشناسی در جدول ۷–۱۳ آمده است. این جدول فقط در بر دارندهٔ نتایجی است که با ویژگی‌های RASTA-PLP بدست آمده‌اند. همان‌طور که در این جدول مشاهده می‌شود، سیستم بازشناسی

پس از آموزش با این دادگان توانسته است بر روی دادگان آزمایش به دقتی در حدود۹۷٪ دست یابـد. قابـل ذکر است که آزمایش سیستم فوق در محیط عملی و از پشت خط تلفن نیز گویای نتایج زیر است و سیستم در اکثر موارد توانایی بازشناسی صدای گوینده را دارا می‌باشد.

جدول ۷-۱۳- دقت بازشناسی حاصل از آموزش سیستم با دادگان FCCDigits برای بازشناسی اعداد متصل و پیوسته

مجموعهٔ مورد آزمایش	کلمه		دقت بازشناسی رشته‌های عددی
	دقت	صحت	
اعداد متصل	٪۹۷,۴	٪۹۷,۴۰	٪۹۱,۷۹
اعداد پیوسته	٪۹۵,۳	٪۹۶,۱۰	٪۹۰,۷۹

برای اینکه بتوان مقایسه‌ای بین نتایج حاصل از آموزش سیستم بـا ایـن دادگـان و دادگـان TFasrdat بدست آید مطابق آزمایش‌های بخش ۷-۳ آزمایش‌هایی طراحی شد فقـط بـا ایـن تفـاوت کـه در اینجـا از FCCDigits برای آموزش استفاده شد. نتایج حاصل از بازشناسی این رشـته‌هـای عـددی در جـدول ۷-۱۴ آمده است. همان‌طور که مشاهده می‌شود، با استفاده از دادگان طراحی شده، حدود ۸۰ درصد کاهش در نرخ خطای کلمه هم برای اعداد متصل و هم برای اعداد پیوسته نسبت به دادگان TFarsdat حاصل شده است.

جدول ۷-۱۴- دقت بازشناسی حاصل از آموزش سیستم با دادگان FCCDigits برای بازشناسی اعداد متصل و پیوسته جهت مقایسه با نتایج حاصل از دادگان TFarsdat

نوع اعداد	کلمه		دقت بازشناسی رشته‌های عددی
	دقت	صحت	
اعداد متصل	٪۹۸,۵۲	٪۹۹,۵۱	٪۹۴,۱۲
اعداد پیوسته	٪۹۷,۲۵	٪۹۸,۴۳	٪۹۳,۷۹

فصل هشتم

نتیجه‌گیری و ارائهٔ پیشنهادها برای ادامه کار

۸-۱- مقدمه

در این فصل به جمع‌بندی کارهای انجام شده در این پروژه پرداخته می‌شود و مراحل انجام آن به اجمال مرور می‌گردد. سپس یک نتیجه‌گیری کلی بیان می‌شود و در انتها نیز پیشنهادهایی جهت تکمیل این پروژه و ادامهٔ کار ارائه خواهد شد.

۸-۲- جمع‌بندی مراحل کار

هدف از این پایان‌نامه، طراحی یک سیستم واقعی بازشناسی اعداد تلفنی با به کارگیری انواع روش‌های استخراج ویژگی‌های مقاوم به نویز بود. سیستم بازشناسی موجود، مدل‌سازی گفتار را بر اساس کلمه (عدد) و با استفاده از HMM انجام می‌داد. در ابتدا سعی شد که با انجام آزمایش‌های منسجم تمام روشهای استخراج ویژگی بیان شده مورد ارزیابی قرار بگیرند و کارایی آنها در حضور نویز خط تلفن و همچنین نویزهای دیگر مورد بررسی قرار بگیرد. سپس با انجام تغیراتی در سیستم بازشناسی گفتار پیوسته پایه، این سیستم برای بازشناسی اعداد مهیا گردید.

در فصل اول مروری اجمالی بر تاریخچهٔ بازشناسی گفتار، علوم مرتبط با آن، پیچیدگی و ابعاد بازشناسی گفتار و کاربردهای مختلف آن ارائه شد. از آنجایی که مبنای سیستم بازشناسی مورد استفاده در این پروژه مدل مخفی مارکوف بود، در فصل دوم به طور اجمالی مدل مخفی مارکوف به عنوان یک روش بسیار قدرتمند برای مدل‌سازی توضیح داده شد. فصل سوم به معرفی قسمت‌های مختلف یک سیستم بازشناسی گفتار مبتنی بر مدل مخفی مارکوف اعم از قسمت پیش‌پردازش، نحوهٔ مدل کردن واحدهای آوایی و مدل آکوستیکی و رویه جستجو اختصاص یافت. همچنین در انتهای این فصل روش‌های دیگر بازشناسی گفتار مورد بررسی قرار گرفتند.

در فصل چهارم نویز نحوهٔ تأثیر آن بر سیستم بازشناسی بررسی شد و روش‌های کاهش اثر نویز بر سیستم بازشناسی دسته‌بندی شدند. روش‌های کم‌کردن اثر نویز به سه دسته تقسیم شدند. دستهٔ اول روش‌هایی هستند که ویژگی‌هایی را از سیگنال گفتار استخراج می‌کنند که نسبت به نویز حساسیت کمتری داشته باشند، دسته دیگر روش‌هایی هستند که بر پایهٔ تخمین گفتار تمیز از گفتار نویزی (حوزهٔ بهبود گفتار) عمل می‌کنند و دستهٔ سوم روش‌های مبتنی بر اصلاح مدل بازشناسی هستند.

در فصل پنجم مقاوم سازی بازشناسی از طریق ویژگی‌های مقاوم بررسی شد و چندین روش مطرح در این زمینه از جمله نرمال کردن در حوزهٔ کپسترال، ضرایب کپسترال ریشه‌ای، پیشگویی خطی مبتنی بر درک انسان، تحلیل طیف نسبی، تحلیل اجزای اصلی و ترکیبی از آن‌ها توضیح داده شد. در این فصل همچنین روشی نوین مبتنی بر فیزیولوژی گوش انسان ارائه شد.

در فصل ششم نحوهٔ طراحی و تهیهٔ یک دادگان تلفنی اعداد زبان فارسی که مشتمل بر ۱۱۰ گوینده با لهجهٔ تهرانی می‌باشد به تفصیل بیان گردید. همانطوری که در این فصل آمد این دادگان در بر دارندهٔ هم اعداد متصل و هم اعداد پیوسته می‌باشد. در این فصل همچنین تغییراتی که در سیستم بازشناسی گفتار پایه می‌بایست اعمال می‌گردید تا توانایی استفاده از این دادگان را داشته باشد، ارائه شد. پیش از این سیستم بازشناسی با استفاده از واحد آوایی واج سعی در بازشناسی اعداد داشت، ولی از آنجایی که پیچیدگی بازشناسی اعداد در سطح واژگان کم می‌باشد استفاده از واحد آوایی واج نمی‌توانست نتایج بسیار خوبی را بدست دهد. بنابراین علاوه بر تغییراتی که در سیستم برای توانایی آن در استفاده از واحد آوایی کلمه می‌بایست اعمال می‌گردید، دادگان مورد نیاز نیز باید طراحی و ضبط می‌شد.

در فصل هفتم بستر انجام آزمایش‌ها و نتایج حاصل از آن‌ها ارائه گردیدند. این آزمایش‌ها بیان کنندهٔ نتایج حاصل از به کارگیری ویژگی‌های مقاوم در محیط تلفنی و همچنین نتایج حاصل از بکارگیری سیستم

برای بازشناسی اعداد انگلیسی و فارسی می‌باشند. نتایج بیان شده و همچنین استفاده از این سیستم به شکل عملی بیان کنندهٔ این نکته است که می‌توان از آن در کاربردهای عملی استفاده کرد.

۸-۳- نتیجه‌گیری و پیشنهادها برای ادامهٔ کار

در این پایان‌نامه ویژگی‌های مقاوم متفاوتی برای مقابله با نویز تلفن مورد بررسی و آزمایش قرار گرفتند. از آنجایی که دامنهٔ این روش‌ها بسیار زیاد می‌باشد، باید ویژگی‌های مقاوم دیگر از جمله ویژگی‌های مبتی بر شنوایی[1] نیز مورد بررسی قرار بگیرند.

- افزایش تعداد گویندگان دادگان فعلی از مهمترین کارهایی است که باید انجام گردد. دادگان حاضر فقط در بر دارندهٔ لهجهٔ تهرانی است. برای داشتن یک دادگان کامل جهت سیستم‌های بازشناسی مستقل از گوینده نیاز به لهجه‌های متنوع زبان فارسی می‌باشد. همچنین تقطیع در سطح کلمه برای دادگان تلفنی اعداد پیوسته زبان فارسی می‌تواند تاثیر مثبتی در بهبود کارایی سیستم‌های بازشناسی گفتار داشته باشد.

- برای اینکه بتوان از دادگان ضبط شده برای بازشناسی اعداد در همهٔ محیط‌ها استفاده کرد به نظر می‌رسد که استفاده از روش‌های مبتنی بر مدل مانند روش ژاکوبین، MLLR و MAP می‌توانند مفید واقع شود. یکی از مزیت‌های استفاده از این روش‌ها عدم نیاز به دادگان بسیار زیاد می‌باشد و تنها با مقدار اندکی دادهٔ ضبط شده در محیط جدید می‌توان به نحو مؤثری به نتایج مورد نظر دست یافت. روش

[1] *Auditory based*

ژاکوبین در این سیستم‌ها که می‌توان در خلال صحبت نویز محیط جدید را به دست آورد و بـه سـرعت عمل تطبیق را انجام داد از مزیت بالایی برخوردار است.

- استفاده از یک آشکارساز فعال شدن صدا[1] به عنوان قسمتی از سیستم بازشناسی گفتار کمـک بـسیار شایانی در بالا بردن دقت بازشناسی خواهد داشت.

- در این پروژه از روش‌های حوزهٔ بهبود گفتار استفاده نشده است. در جاهایی که انتظار بازشناسـی در سیگنال به نویزهای کم وجود دارد، این روش‌ها می‌توانند مؤثر واقع شوند.

- یکی از کاربردهای جدید بازشناسی اعداد از طریق گوشی همـراه مـی‌باشـد. نحـوهٔ اسـتفاده از ایـن سیستم و اینکه تا چه حد این سیستم می‌تواند کارا باشد، به عنوان یکی دیگر از کارهای آینـده مـی‌توانـد واقع شود.

[1] *Voice Activity Detector*

منابع

[1] X. Huang, A. Acero & H.W. Hon, *Spoken Language Processing*, New Jersey, Prentice Hall, 2001.

[2] Baum, L.E. and J.A. Eagon, *"An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology"*, Bulletin of American Mathematical Society, Vol. 73, pp. 360-363, 1967.

[3] Sakoe, H. and S. Chiba, *"Dynamic Programming Algorithm Optimization for Spoken Word Recognition*," IEEE Transaction on Acoustics, Speech and Signal Processing, Vol. 26 No. 1, pp. 43-49, 1978.

[4] Viterbi, A.J., *"Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm,"* IEEE Transaction on Information Theory, Vol. 13, No. 2, pp. 260-269, 1967.

[5] L. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, New Jersey, Prentice Hall, 1993.

[6] L. Rabiner, *"A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition"*, In Proceedings of the IEEE, Vol. 77, No. 2, pp. 257-285, February 1989.

[7] N. Deshmukh, A. Ganapathiraju & J. Picone, *"Hierarchical Search for Large Vocabulary Conversational Speech Recognition"*, IEEE Signal Processing Magazine, Vol. 16, No. 5, pp. 84-107, September 1999.

[8] V. Goel, W. Byrne, *"Task Dependent Loss Functions in Speech Recognition: A* Search over Recognition Lattices"*, Proceedings ESCA Tutorial & Research Workshop for Accessing Information in Spoken Audio, Cambridge, UK, 1999.

[9] J.R. Deller, J.G. Proakis & J.H.L. Hansen, "*Discrete-Time Processing of Speech Signals*.", New York, Macmillan Publishing Company, 1993.

[10] C. H. Lee, *"On Stochastic Feature and Model Compensation Approaches to Robust Recognition"*, Speech Communication, Vol. 25, pp. 29-47, 1998.

[11] M. J. F. Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. Thesis, University. of Cambridge, 1995.

[12] Pedro J. Moreno , *Speech Recognition in Telephone Environments*, MS. Thesis, ECE Department, Carnegie Mellon University, December 1992.

[13] M. B. Carey,H. T Chen, A. Descloux,J. F. Ingle and K. I.Park, *"Analog Voice and Voiceband Data Transmission Performance Characterization of the Public Switched Network"*, Bell System Technical Journal, Vol. 63, pp. 2059-2119, November 1984.

[14] C. H. Lee, *"On Stochastic Feature and Model Compensation Approaches to Robust Recognition"*, Speech Communication, Vol. 25, pp. 29-47, 1998.

[15] B. R. Ramakrishnan, *Reconstruction of Incomplete Spectrograms for Robust Speech Recognition*, Ph.D. Thesis, Carnegie Mellon University, April 2000.

[16] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, 1990.

[17] K. Yao, E. Visser, O. W. Kwon, T. W. Lee, *"A Speech Processing Front-End with Eigenspace Normalization for Robust Speech Recognition in Noisy Automobile Environments*", In Proceedings of the EUROSPEECH, pp 9-12, 2003.

[18] D. Mansour and B. H. Juang, *"The Short-Time Modified Coherence Representation and Noisy Speech Recognition"*, IEEE Transaction on ASSP, Vol. 37, pp. 795-804, 1989.

[19] J. S. Lim, *"Spectral Root Homomorphic Deconvolution System*", IEEE Transaction on ASSP, Vol. 27, No 3, pp. 223-233, 1979.

[20] H. Hermansky, "*Perceptual Linear Predictive (PLP) Analysis for Speech"*, Journal of Acoustic. Society of America, pp. 1738-1752, 1990.

[21] H. Hermansky, N.Morgan, *"RASTA Processing of Speech"*, IEEE Transaction on Speech & Audio Processing, Vol.2, No. 4, pp. 587-589, October 1994.

[22] J. Koehler and et al, "*Integrated RASTA-PLP into speech recognition"*, In Proceedings of the IEEE ICASSP, Vol. 1, pp. 421-424, 1994.

[23] J. de Veth & L. Boves, *"Channel normalization techniques for automatic speech recognition over the telephone"*, Speech Communication, Vol. 25, pp. 149-164, Sep 1998.

[24] H. Hermansky, N.Morgan, A. Bayya, and P. Kohn, *"RASTA-PLP Speech AnalysisTechnique"*, In Proceedings of the IEEE ICASSP, Vol. 1, pp. 121-124, March 1992.

[25] R. Haeb-Umbach and H. Ney, *"Linear discriminant analysis for large vocabulary speech recognition"*, In Proceedings of the IEEE ICASSP, Vol. 1, pp. 13–16, 1992.

[26] A. Acero, R. M. Stern, *"Robust Speech recognition by Normalization of the Acoustic Space*", In Proceedings of the IEEE ICASSP, Toronto, Ontario, pp. 893-896, 1991.

[27] R. Rhoades and R. Pflanzer, *Human Physiology*, 3rd Edition, Saunders College Publishing, 1996.

[28] H. D. Young, R. A. Freedman, *University Physics with Modern Physics with Mastering Physics*, 11th Edition, Benjamin Cummings, 2004.

[29] Muthusamy, Y. K. et. al., *"The OGI Multi-Language Telephone Speech Corpus"*, Proceedings of the ICSLP. USA, pp. 895-898, 1992.

[30] LDC Homepage:
"http://WWW.ldc.upenn.edu/catalog/catalogEntry.jsp?catalogId=LDC96S50."

[31] Bijankhan, J. Sheykhzadegan, M. R. Roohani, R. Zarrintare,S. Z. Ghasemi and M. E. Ghasedi *"Tfarsdat - the Telephone Farsi Speech Database"*, In Proceedings of the EUROSPEECH, Geneva, Switzerland, pp. 1525-1528, 2003.

[32] H. G. Hirsch, D. Pearce, *"The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions"*, ISCA ITRW ASR2000, Paris, September 2000.

[33] Bijankhan, M. et. al., *"FARSDAT-The Farsi Spoken Language Database"*, In Proceedings of the 5th Australian International Conference on Speech Sciences and Technology. Vol. 2, pp. 826-829, 1994.

[34] S. Young, G. Everman, D. Kershaw, G Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book(for HTK Version 3.2)*, Cambridge University, 2002.

[35] K.Yao, E. Visser, O.W. Kwon & T.W. Lee, *"A Speech Processing Front-End with Eigenspace Normalization for Robust Speech Recognition in Noisy Automobile Environments"*, in Proc. Eurospeech 2003, pp 9-12.

[36] C.A. Ynoguti, F. Violaro, *"On the use of Principal Component Analysis over Mel Cepstral Coefficients"*, Telecomunicaçes, Revista do Instituto Nacional de Telecomunicaçes, ISSN 1516-2338, Vol. 05, no 2, pp. 13-17, December 2002.

[37] B. Babaali, H. Sameti, *"The Sharif Speaker Independent Large Vocabulary Speech Recognition System"*, In Proceedings of the 2nd Workshop on Information Technology & Its Disciplines, Kish, February 2004.

[38] S. Molau, M. Pitz & H. Ney, "*Histogram based Normalization in the Acoustic Feature Space*", ASRU Workshop on Automatic Speech Recognition & Understanding, 2001.

[39] A. Acero, R.M. Stern, *"Enviromental Robustness in Automatic Speech Recognition"*, In Proceedings of the ICASSP, Albuquerque, New Mexico, 1991.

[40] E.A. Wan, A.T. Nelson, *"Handbook of Neural Networks for Speech Processing"*, Boston, USA, 1998.

[41] S.F. Boll., "*Suppression of Acoustic Noise in Speech using Spectral Subtraction"*, IEEE Trans. Acoustics, Speech & Signal Processing, April 1979.

[42] J.L. Gauvain, C.H. Lee, *"Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains"*, IEEE Transaction on Speech & Audio Processing, Vol. 2, No. 2, pp. 291-298, April 1994.

[43] C. J. Leggetter, P.C. Woodland, *"Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs"*, Computer Speech & Language, Vol. 9, pp.171-186, 1995.

[44] S.V. Vaseghi, *Advanced Signal Processing and Noise Reduction*, John Wiley & Sons Publication Company, 2nd edition, 2000.

[45] A. Acero, R.M. Stern, *"Robust Speech Recognition by Normalization of the Acoustic Space"*, In Proceedings of the ICASSP, Toronto, Ontario, 1991.

[46] D.Y. Kim, C.K. UN, *"Probabilistic Vector Mapping with Trajectory Information for Noise Robust Speech Recognition"*, Electronic Letters, Vol. 32, pp. 1550-1551, August 1996.

[47] S. Sagayama, Y. Yamaguchi, S. Takahashi & J. Takahashi, *"Jacobian Approach to Fast Acoustic Model Adaptation"*, In Proceedings of the ICASSP, Munich, Germany, pp. 835-838, 1997.

[48] R. Haeb-Umbach, H. Ney, *"Linear Discriminant Analysis for Large Vocabulary Speech Recognition"*, In Proceedings of the IEEE on Acoustics, Speech & Signal Processing, Vol. 1, San Francisco, pp. 13–16, 1992.

[۴۹] رحمانی، محسن، *''اعمال روشهای بهبود گفتار به عنوان پیش پردازش، جهت بالابردن دقت بازشناسی گفتار فارسی''*، پایان نامه کارشناسی ارشد در رشتهٔ مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، ۱۳۸۰.

[۵۰] موثق، حامد، *''طراحی و پیاده‌سازی روش جستجوی بهینه برای بازشناسی گفتار پیوسته با مدل مخفی مارکوف''*، پایان‌نامهٔ کارشناسی ارشد، دانشکدهٔ مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۸۱.

[۵۱] بحرانی، محمد، *''به کارگیری ساختارهای وابسته به بافت در بازشناسی گفتار پیوسته مبتنی بر مدل مخفی مارکوف''* ، پایان‌نامهٔ کارشناسی ارشد، دانشکدهٔ مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۸۲.

[۵۲] باباعلی، باقر، *''بررسی روش‌های هرس‌کردن برای بهبود عملکرد یک سیستم بازشناسی گفتار پیوسته مبتنی بر مدل مخفی مارکوف و به‌کارگیری آن‌ها''*، پایان‌نامهٔ کارشناسی ارشد، دانشکدهٔ مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۸۲.

[۵۳] غلامپور، ایمان، *''بازشناسی مستقل از گوینده واجهای فارسی در صحبت پیوسته''*، پایان نامه دکترا، دانشکدهٔ مهندسی برق، دانشگاه صنعتی شریف، ۱۳۷۹.

Abstract

Telephony digits recognition is one of the most important applications of speech recognition systems. Speech recognition systems work very well dealing with high-quality speech, however, when the speech signal has been corrupted by the telephone network, recognition accuracies decrease dramatically. There are several ways in the categories of robust features, model-based techniques, and speech enhancement approaches to decrease significantly recognition error rates. The aim of this thesis is to design and implement a continuous speech recognition system for recognition of telephony digits. To improve the recognition accuracy in the telephony environment, many algorithms of robust features have been studied. CMS, PLP, RASTA, PCA, RCC methods and also combinations of them have been implemented and tested. In the category of robust features a new method based on human ear physiology suggested and tested. The result of this method indicates its capability in noisy environments. Because the best acoustic unit for training of a HMM-based continuous speech recognition for recognizing of digits is words (digits), some changes in the base speech recognition system applied in order that the system has the capability of training based on word acoustic units, and also because in Persian language there was not any telephony database for training (specially based on words) of the speech recognition systems, Farsi Connected and Continuous Digits (FCCDigits) database is designed, recorded, and transcribed. Recognition results show 97 percent accuracy on FCCDigits database for connected digits and 95 percent accuracy for continuous digits.

Keywords: Telephony Digit Recognition, Continuous Speech Recognition, Robust Speech Recognition, Robust Features, Hidden Markov Model.

Sharif University of Technology
Computer Engineering Department

M.Sc. Thesis

Telephony Continuous Digit Recognition

By:
Amin Fazel Dehkordi

Supervisor:
Dr. Mohammad Taghi Manzouri

January 2005