شورای پژوهشهای علمی کشور

کمیسیون اطلاع رسانی و فناوری اطلاعات

گزارش نهایی طرح ملی پردازش زبان فارسی

دانشگاه صنعتی امیر کبیر

فهرست مطالب

فهرست شکل ها

فهرست جداول

۱ بازشناسی گفتار گسسته و پیوستهٔ فارسی

۱ـ ۱ مقدمه

هدف در بخش بازشناسی گفتار گسسته و پیوستهٔ فارسی، دستیابی به سیستم‌های بازشناسی گفتارفارسی با توانائی بازشناسی گفتار گسسته در چارچوب دایرهٔ کلمات تعیین شده و نرخ بازشناسی قابل‌قبول، بر روی دادگان تعریف شده و نیز بازشناسی گفتار پیوستهٔ فارسی با دایرهٔ کلمات مشخص و درچارچوب دادگان اصلاح شدهٔ پیوسته و با نرخ بازشناسی موردنظر می‌باشد.

در هر یک از این دو راستا، اقدامات صورت گرفته شامل مطالعه، بررسی و پیاده‌سازی روش‌های‌مطرح در بازشناسی جهت دستیابی به نتایج مطروحه بوده است. تهیهٔ دادگان‌های گفتاری گسسته وپیوسته با مشخصات ازپیش تعیین شده، پیاده‌سازی روش‌های مبتنی برHMMهای با چگالی‌مشاهدات پیوسته و نیز شبکه‌های عصبی، ارزیابی و بهبود روش‌های پیاده‌سازی شده و احیاناً اصلاح یاپیشنهاداتی جهت اصلاح این روش‌ها می‌باشد.

۱ـ ۲ بازشناسی گفتار گسستهٔ فارسی

بازشناسی گفتار گسسته از جمله بخش‌های مهمّ در بحث بازشناسی گفتار می‌باشد که در کاربردهای‌مختلف ازجمله فرمان و کنترل موارد استفادهٔ فراوانی دارد.

۱ـ ۲ـ ۱ دایرهٔ کلمات و دادگان

دایرهٔ کلمات از جمله عوامل مهمّ در تأمین بازشناسی با کیفیت مناسب گفتار گسسته می‌باشد. هدف‌در این طرح پژوهشی دستیابی به بازشناسی در محدودهٔ کلمات صد کلمه بوده است. با این همه، از آنجاکه پیش از آغاز این طرح، یک دادگان گفتاری از کلمات گسسته (ارقام فارسی) با تعداد ده کلمه و بصورت ناوابسته به گوینده موجود بوده است، مرحلهٔ ابتدائی کار بر روی این دادگان به انجام رسید. لازم به اشاره‌است که این دادگان از مجموعه‌ای از دانشجویان در محدودهٔ سنی ۲۰ تا ۳۰ سال از هر دو جنسیت ضبط‌گردیده و مشتمل بر دو بخش آموزشی و آزمایشی بوده است. شرایط ضبط، محیط با نویز زمینهٔ کم بوده‌است.

دادگان دیگری برای دستیابی به اهداف این طرح پژوهشی طراحی و ضبط گردید. در طراحی این‌دادگان تلاش شد کلمات مورد استفاده، ضمن پوشش یک محدودهٔ صدکلمه‌ای، از ویژگی‌های خاصی‌برخوردار باشند. به همین جهت، مجموعهٔ کلمات شامل ۳۸ کلمهٔ اصلی تشکیل‌دهندهٔ کلیهٔ اعداد تاشش رقمی در زبان فارسی (صفر تا نوزده، مضارب ده، مضارب صد و مضارب هزار) بهمراه اسامی‌شهرهای مهمّ ایران می‌باشد. درعین حال ضبط دادگان بصورت مستقیم به داخل کامپیوتر از طریق یک‌میکروفون دینامیک مناسب(Sony F-VK98) و در محیط اداری (دفتر کار) با نویز محیط معمولی‌صورت گرفت. در مجموع از ۱۰۲ گوینده (نیمی مرد و نیمی زن)، هر یک از ۱۰۰ کلمهٔ تعریف شده، ۵ بارضبط گردیدند. البته باتوجه به ضبط گفتار در محیط دانشگاهی و از افرادی که دارای مدرک تحصیلی‌حداقل دیپلم متوسطه بوده‌اند، دسترسی به افراد با سنّ بالای ٤٠ سال عملاً مقدور نبوده و لذا سنّ افرادشرکت‌کننده در تهیهٔ این دادگان محدودهٔ وسیعی را شامل نمی‌شود [۱][۲].

پس از جمع‌آوری دادگان، اقدام به بررسی تک‌تک حدود ۵۱/۰۰۰ بیان گفتاری موجود گردید واشکالات مختلف موجود در آنها از قبیل صداهای اضافی ضبط شده، بریدگی کلمات در آغاز یا پایان،برخی لهجه‌ها و گویش‌های نامعمول، صدای دهان در آغاز ادای کلمات و ... که علیرغم پیش‌بینی‌های‌لازم در ضبط برخی کلمات وجود داشتند، مشخص و بیان‌های دارای مشکل حذف گردیدند [٤]. سپس‌دادگان به دو بخش آموزشی و آزمایشی تقسیم گردید. این دادگان به عنوان دادگان اصلی در این بخش ازطرح پژوهشی پردازش گفتار فارسی مورد استفاده واقع گردید.

درعین حال در بعضی مراحل، ازجمله در بازشناسی به کمک شبکه‌های عصبی نیز یک مجموعهٔ ۱۲کلمه‌ای از کلمات برای کنترل یک یونیت دندانپزشکی مورد استفاده قرار گرفت که بنابه عللی که بعداًاشاره خواهد گردید مورد استفادهٔ بیشتری نیافت [۲].

۱ـ ۲ـ ۲ سیستم‌های مورد استفاده

جهت دستیابی به بازشناسی کلمات گسسته، دو شیوه مورد توجه قرار گرفت که ذیلاً به این دو شیوه‌می‌پردازیم.

۱ـ ۲ـ ۲ـ ۱ مدل‌های مارکوف پنهان با چگالی مشاهدات پیوسته (CDHMMs)

مدل‌های مارکوف پنهان به عنوان یکی از شیوه‌های موفق، امروزه در کاربردهای بازشناسی گفتارپیوسته و گسسته کاربردهای فراوانی یافته‌اند. این مُدل‌ها باتوجه به توانائی بالائی که در مُدل‌نمودن‌ویژگی‌های گفتار و بخصوص ویژگی دینامیک گفتار دارند، مورد بررسی و استفادهٔ فراوانی در این زمینه‌قرار گرفته‌اند. مناسب‌ترین این مُدل‌ها جهت اینگونه مدل‌سازی، از نقطه‌نظر

چگالی احتمالات‌خروجی، انواع با چگالی مشاهدات پیوسته می‌باشند. این‌گونه ازHMM ها به دلیل دقت بالائی که درمُدل‌سازی ارائه می‌دهند مناسب تشخیص داده شده‌اند [۵][٦]. به همین دلیل، در این پیاده‌سازی نیز ازاینگونه مُدل‌های آماری استفاده گردید.

۱ـ۲ـ۲ـ۱ـ۱ توپولوژی مُدل‌ها

پیاده‌سازی اوّلیه بر روی دادگان گفتاری ده‌کلمه‌ای براساسCDHMMs دارای ٦ حالت صورت‌گرفته بود. در کاربرد ۱۰۰ کلمه‌ای، باتوجه به اینکه طول برخی کلمات طولانی‌تر بوده، ازCDHMMs با۸ حالت استفاده گردید. اینHMM ها از نوع چپ ـ راست و بدون انتقال جهشی درنظر گرفته شدند [٤].اشاره به این نکته نیز در اینجا خالی از اهمیت نیست که باتوجه به توانائیHMM ها در مُدل‌سازی‌دینامیک سیگنال گفتاری، مُدل‌نمودن سکوت ابتدا و انتهای کلمه نیز برعهدهٔ حالت‌های آغازین و پایانیHMM ها گذاشته شد و از آشکارسازی نقاط ابتدا و انتهای گفتار پرهیز گردید.

۱ـ۲ـ۲ـ۱ـ۲ استخراج ویژگی‌های طیفی

ویژگی‌های طیفی مورد استفاده در این بخش از طرح پژوهشی، ضرائب کپسترال LP انتخاب گردیدند.برای استخراج این ضرائب از سیگنال گفتاری رقمی‌شده، مراحل زیر به ترتیب به انجام رسید [۱][٤].

۱ـ پیش‌تأکید[1] . وظیفهٔ این بخش مسطح‌کردن طیف فرکانسی سیگنال گفتار برای پیشگیری از مشکلات‌احتمالی ناشی از محدودیت طول لغت دیجیتال در پردازش‌های بعدی می‌باشد.

۲ـ تقسیم به قاب‌ها[2] . سیگنال گفتار در این بخش به مجموعه‌ای از بلوک‌های متداخل با طول زمانی‌حدود ۲۰ تا ۲۵ میلی‌ثانیه و فاصلهٔ فریم ۱۰ تا ۱۵ میلی‌ثانیه تقسیم می‌گردد. هر قاب سپس مورد پردازش‌مستقل قرار گرفته و ویژگی‌های سیگنال گفتاری در داخل آن پایدار فرض می‌شود.

۳ـ اعمال پنجرهٔ همینگ[3] که باعث کاهش دامنهٔ نمونه‌های کناری قاب می‌گردد.

٤ـ استخراج ضرائب LPC[4].

۵ـ بدست‌آوردن ضرایب کپسترال[5] با استفاده از ضرائب LPC .

٦ـ اعمال لیفترجوانگ[6] جهت وزن‌دهی ضرائب کپسترال.

[1] Pre-emphasis
[2] Frame Blocking
[3] Hamming Window
[4] Linear Predictive Coding
[5] Cepstral Coefficients
[6] Juang Lifter

۷ـ بدست‌آوردن ضرائب دلتا و دلتا ـ دلتا برای ضرائب کپسترال و انرژی لگاریتمی.

بردارهای ویژگی بدست آمده از مراحل فوق به عنوان ویژگی‌های طیفی نمایندهٔ گفتار در مراحل‌بعدی پردازش مورد استفاده قرار می‌گیرند.

۱ـ۲ـ۲ـ۱ـ۳ آموزش مُدل‌ها

برای بدست آوردن یک سیستم بازشناسی گفتار مناسب، لازم است ابتدا مُدل‌ها به‌نحو مطلوبی مورد آموزش قرار گیرند. یکی از پراستفاده‌ترین روش ها در آموزش مدل‌هایHMM ، روشی مبتنی بر تکنیک‌درستنمائی بیشینه (ML[۷]) است، موسوم به Baum –Welch. این آلگوریتم یک روش تکراری است ورسیدن آن به یک ماکزیمم محلّی برای درستنمائی اثبات شده است [۷]. با استفاده از این آلگوریتم و درشرایط دنباله‌های مشاهدهٔ چندگانه، روابط باز تخمین برای پارامترهای اصلی در یکCDHMM به قرارزیر می‌باشند [۱][٦]:

$$\hat{c}_{jk} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k)}{\sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t^{(l)}(j,m)} \qquad (۱-۱)$$

$$\hat{\boldsymbol{\mu}}_{jk} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k)\cdot \boldsymbol{O}_t^{(l)}}{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k)} \qquad (۱-۲)$$

$$\hat{\boldsymbol{U}}_{jk} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k)\cdot(\boldsymbol{O}_t^{(l)}-\boldsymbol{\mu}_{jk})(\boldsymbol{O}_t^{(l)}-\boldsymbol{\mu}_{jk})^T}{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k)} \qquad (۱-۳)$$

در این روابط، c_{jk} وزن مخلوط kاُم از حالت jاُم، $\boldsymbol{\mu}_{jk}$ بردار میانگین آن و $\boldsymbol{U}_{jk}$ ماتریس کوواریانس آن می‌باشند و روابط برای $\boldsymbol{\mu}_{jk}$ و $\boldsymbol{U}_{jk}$ به صورت برداری (یا ماتریسی) نوشته شده‌اند. همچنین،$\gamma_t(j,k)$ احتمال بودن در مخلوط kام از حالت jام در لحظهٔ t و مشاهدهٔ $\boldsymbol{O}_t$ است.

پیش از اینکه مدل‌هایCDHMM بتوانند آموزش ببینند، لازم است مُدل‌های اوّلیه با مقادیر مناسبی‌مقداردهی شوند تا دستیابی به ماکزیمم محلّی بتواند مترادف با رسیدن به ماکزیمم کلّی تلقی شود. ازاین‌رو مقداردهی اوّلیهٔ پارامترها از اهمیت فوق‌العاده‌ای برخوردار است. برای این منظور در دو مرحله به‌شرح زیر عمل می‌شود:

الف: کلیهٔ مقادیر دنباله‌های مشاهدهٔ آموزشی موجود برای هر مُدل بطور یکنواخت به تعدادحالت‌های آن مُدل تقسیم شده و بر روی تمامی بردارهای بدست آمده برای هر حالت مقادیر

[7] Maximum Likelihood

بردارهای‌میانگین و واریانس بدست می‌آیند و به عنوان مقادیر ابتدائی برای پارامترهای آن حالت مورد استفاده‌قرار می‌گیرند.

ب: با استفاده از مقادیر ابتدائی فوق، تمامی دنباله‌های مشاهدهٔ موجود با استفاده از آلگوریتم ویتربی‌با حالات مدل مربوطه همردیف می‌شوند. سپس مقادیر همردیف شده بند الف فوق برای بدست‌آوردن پارامترهای هر حالت مورد استفاده قرار می‌گیرند. این عمل تا رسیدن به یک معیار همگرائی و یارسیدن به تعداد موردنظر از دفعات تکرار ادامه می‌یابد.

پارامترهای بدست آمده از مراحل فوق، سپس به عنوان پارامترهای اوّلیه در آلگوریتم Baum-Welch مورد استفاده قرار می‌گیرند.

درخصوص آموزش مدل‌های دارای چگالی‌های مشاهدهٔ مخلوط گوسین[8] ، اگرچه استفاده از روش‌فوق نیز جایز می‌باشد ولی ترجیح داده می‌شود که در مراحل بعدی نسبت به افزایش تعداد عناصرمخلوط اقدام نمود. بنابراین مراحل ابتدائی آموزش عموماً با چگالی‌های مشاهدهٔ تک‌گوسین صورت‌می‌پذیرد و سپس برای افزایش تعداد عناصر مخلوط از آلگوریتمی نظیر روش شکافت مخلوط[9] که دربحث بازشناسی گفتار پیوسته به تفصیل بیشتری مورد اشاره قرار خواهد گرفت، استفاده می‌شود.

نکتهٔ دیگری که در اینجا حائز اهمیت است، استفاده از نمایش درستنمائی بصورت لگاریتمی است که لازمهٔ پیاده‌سازی آلگوریتم‌هائی نظیرBaum-Welch می‌باشد چرا که مقادیردرستنمایی بعد از چند لحظهٔ زمانی به علّت ضرب‌شدن در مقادیر احتمالی بسیار کوچک، به سمت‌صفر میل خواهد نمود که عمل آموزش را غیرممکن می‌سازد. بهمین جهت لازم است که در پیاده‌سازی‌از نمایش لگاریتمی احتمالات و درستنمائی استفاده نمود. در این صورت ضرب‌ها هم تبدیل به جمع شده و بنابراین مشکل میل‌نمودن سریع مقادیر حاصلضرب به سمت صفر نیز از بین خواهد رفت.

۱ـ۲ـ۲ـ۱ـ٤ بازشناسی کلمات جدا

مرحلهٔ بازشناسی کلمات جدا با استفاده از HMMها عموماً به کمک آلگوریتم ویتربی[10] صورت‌می‌گیرد [۵][٦]. این آلگوریتم که براساس شیوهٔ برنامه‌نویسی پویا طراحی گردیده از مزایای چندی به‌شرح زیر برخوردار است:

ـ یافتن مسیر بهینهٔ حالت براساس اصل بهینگی Bellman .

ـ کاهش فراوان حجم محاسبات به دلیل استفاده از اصول برنامه‌نویسی پویا.

[8] Mixture-Gaussian Observation Densities
[9] Mixture Splitting
[10] Viterbi Algorithm

ـ یافتن همزمان احتمال دنبالهٔ مشاهدات به شرط وجود مدل($P(O|\lambda)$) که در بازشناسی گفتار کاربرددارد.

ـ عدم استفاده از جمع (برخلاف آلگوریتم جلورونده[11]) و درمقابل استفاده از بیشینه‌سازی که باعث‌ساده‌شدن پیاده‌سازی آن بصورت لگاریتمی می‌شود.

باتوجه به موارد فوق، آلگوریتم ویتربی یکی از روش‌های مناسب و مطلوب جهت انجام بازشناسی‌گفتار تلقی شده و لذا در این بخش از طرح پژوهشی جاری نیز مورد استفاده قرار گرفته است.

۱ـ ۲ـ ۲ـ ۲ شبکه‌های عصبی با تأخیر زمانی[12]

یکی دیگر از روش‌های مطرح در بازشناسی گفتار، استفاده از شبکه‌های عصبی با تأخیر زمانی(TDNN) می‌باشد. این نوع شبکه به عنوان گونه‌ای تغییریافته از[13] MLP جهت کاربرد در بازشناسی‌گفتار مطرح گردیده است. از آنجائی که شبکه‌های عصبی توانائی بالائی در طبقه‌بندی الگوهای استاتیک‌دارند، جهت اعمال آنها در این زمینه، نیاز به افزودن ویژگی توانائی برخورد با دینامیک الگوهای گفتاری‌به آنها می‌باشد. در TDNN ، اطلاعات چند فریم گفتاری، به همین منظور، بطور همزمان به سیستم‌اعمال می‌شود. در عمل ملاحظه گردیده است که توانائی نسبتاً قابل قبولی را می‌توان در این زمینه ازاین‌گونه شبکه‌ها انتظار داشت [۸].

۱ـ ۲ـ ۲ـ ۲ـ ۱ دادگان به‌کار رفته

برای این مرحله از پیاده‌سازی، از دادگانی که با استفاده از دوازده کلمهٔ کنترلی اشاره‌شدهٔ قبلی تشکیل‌گردیده بود استفاده گردید. این دادگان با استفاده از گفتار ۷۰ گوینده که به تساوی از مردان و زنان تشکیل‌شده بودند و با سه بار تکرار هر کلمه تشکیل گردیده بود. علاوه بر این یک بخش وابسته به گوینده نیز دراین دادگان پیش‌بینی شده بود تا بتوان توانائی سیستم بازشناسی را روی اینگونه کاربرد نیز آزمود. [۱]

۱ـ ۲ـ ۲ـ ۲ـ ۲ پیاده‌سازی

در مرحلهٔ پیاده‌سازی این سیستم نیز همانند سیستم مبنی برCDHMM ، نیاز به مراحل استخراج‌ویژگی‌های گفتاری، آموزش سیستم بازشناسی و بالاخره شبیه‌سازی و بازشناسی می‌باشد. موارد فوق برروی یک شبکهٔ عصبی از نوعTDNN پیاده‌سازی گردیده و شبکه با دو مجموعهٔ دادگان،

[11] Forward Algorithm
[12] Time Delay Neural Networks
[13] Multi-Layer Perceptron

یکی دادگان‌مبتنی بر مجموعهٔ ارقام اشاره شده در بخش ۱ـ ۲ـ ۱ و دیگری مجموعهٔ اشاره شده در بخش۱ـ ۲ـ ۲ـ ۱ آزمایش گردید [۲]. اگرچه نتایج بدست آمده از این مرحله نسبتاً مناسب می‌باشند، ولی‌کیفیت بالاتر سیستمCDHMM در بازشناسی گفتار گسسته کاملاً آشکار بوده است. بررسی دقیق‌ترنشان می‌دهد که شبکهTDNN ، علیرغم تأخیر زمانی ایجاد شده، قادر به دخالت دادن مناسب ومطلوب دینامیک گفتار در طبقه‌بندی نمی‌باشد. یکی از روش‌های پیشنهاد شده استفاده از نوعی پیچش‌زمانی به عنوان پردازش اوّلیه جهت انطباق مناسب زمانی سیگنال گفتار با الگوهای استفاده شده برای‌آموزش شبکه می‌باشد. از آنجا که روش‌هائی نظیر پیچش زمانی پویا(DTW[14]) خود به عنوان‌آلگوریتم‌های بازشناسی گفتار مستقیماً مورد استفاده می‌باشند، تصمیم گرفته شد تحقیق بیشتر در این‌زمینه متوقف شده و پیاده‌سازی تنها براساسCDHMMs ادامه یابد.

۱ـ ۲ـ ۳ مشکلات در بازشناسی و اصلاحات مربوطه

سیستم بازشناسی ایجاد شده براساس CDHMMs ، علیرغم توانائی‌های خوبی که از خود نشان‌می‌دهد، از دو مشکل اساسی رنج می‌برد: محدودیت سرعت و کاهش کیفیت در شرایط مختلف‌محیطی (نظیر وجود نویز زمینه). لذا دنبالهٔ کار در این بخش بر روی رفع این دو مشکل و یافتن‌راه‌حل‌هائی برای آنها متمرکز گردید.

۱ـ ۲ـ ۳ـ ۱ سرعت بازشناسی

مسئلهٔ سرعت در بازشناسی در مراحل مختلف کار قابل طرح می‌باشد. پیاده‌سازی بهینهٔ آلگوریتم‌هاموجب می‌گردد که با سرعت بالاتری به انجام برسند. چه در مرحلهٔ آموزش و چه در مرحلهٔ بازشناسی،این فاکتور دارای اهمیت زیادی می‌باشد. داشتن سرعت بالا در آموزش، درعین حال دارای اهمیت سرعت در بازشناسی نمی‌باشد چرا که آموزش عموماً در شرایطoff-line صورت می‌گیرد درحالیکه‌در کاربردهای واقعی، بازشناسی اغلبon-line می‌باشد. به همین جهت اهمیت سرعت در بازشناسی‌بیش از آموزش می‌باشد. البته باید توجه نمود که بهر صورت کندبودن آموزش خود باعث تأخیر درپیاده‌سازی موارد آزمایشی سیستم می‌گردد.

علاوه بر موارد فوق باید توجه نمود که چه در آموزش و چه در بازشناسی، از آنجا که نیاز به استخراج‌ویژگی‌های گفتاری می‌باشد، سرعت استخراج ویژگی‌ها دارای اهمیت می‌باشد. با این همه، در این‌خصوص نمی‌توان بهبودی چندانی، جز از طریق بهینه‌سازی برنامه‌نویسی و یا برخی تغییرات جزئی‌بدست آورد[۱]. درخصوص آلگوریتم آموزش نیز باتوجه به اینکه در این طرح پژوهشی،

[14] Dynamic Time Warping

آموزش‌بصورتoff-line صورت می‌گرفته است، افزایش سرعت اجرای الگوریتم چندان مورد توجه قرارنگرفته است.

آلگوریتم بازشناسی، باتوجه به چند مورد، نیاز به افزایش سرعت دارد. اوّل اینکه این آلگوریتم‌می‌تواند به صورتon-line مورد استفاده واقع شود. علاوه بر این، باتوجه به استفاده ازHMM های باچگالی پیوسته، عمدهٔ زمان صرف محاسبهٔ تابع چگالی احتمال مخلوط پیوسته (گوسین‌ها) می‌گردد که‌شامل محاسبهٔ مقادیر نمائی و ضرب‌های برداری می‌باشد. باتوجه به زمان‌گیربودن این محاسبات واینکه باید به‌ازاء هر بردار ورودی این کار تکرار گردد، زمان صرف‌شدهٔ کلّی به‌ازاء هر دنبالهٔ بردارهای‌ورودی (بیان گفتاری) قابل توجه می‌باشد. از آنجا که در بازشناسی با دایرهٔ کلمات ۱۰۰ کلمه، تعدادمُدل‌هائی که باید مورد بررسی قرار گیرند، زیاد است (۱۰۰ عدد)، این محاسبات باید ۱۰۰ بار تکرارگردد تا مُدل مناسب بتواند بررسی و یافت شود. بنابراین زمان زیادی باید صرف شود تا بازشناسی‌بتواند صورت گیرد.

شیوه‌های چندی برای بهبود سرعت در آلگوریتم بازشناسی به‌کار گرفته شده‌اند. آنچه در این طرح‌پژوهشی برای این منظور مورد توجه قرار گرفته از دو جنبه قابل طرح است. جنبهٔ اوّل بهبود پیاده‌سازی‌آلگوریتم جهت دستیابی به سرعت‌های بازشناسی قابل قبول می‌باشد. دوّمین جنبه از کار، توجه به‌روش‌های جستجوی بهبودیافته جهت کاهش فضای جستجو در هنگام کار با تعداد کلمات بالا می‌باشد.

در پیاده‌سازی آلگوریتم، توجه به ماتریس انتقال می‌تواند از جهت افزایش سرعت راهگشا باشد. توجه به غیرارگودیک[۱۵] بودنHMM و نیز محدودنمودن اختصاص فریم‌ها به حالت‌ها از نقطه‌نظر زمانی، باتوجه به محدودیت‌های طبیعی در گفتار، می‌توانند باعث افزایش نسبی سرعت اجرای الگوریتم شوند [۱][۲].

از جنبهٔ کاهش فضای جستجو، می‌توان بجای اقدام به اعمال تک‌تک مُدل‌ها (مثلاً ۱۰۰ مدل) به دادهٔورودی جهت محاسبهٔ درستنمائی مربوطه و سپس انتخاب بالاترین درستنمائی جهت بدست‌آوردن‌کاندیدای مناسب، به شیوهٔ دیگری عمل نمود [٤]. در شیوهٔ اخیر، اقدام به اعمال همزمان آلگوریتم‌ویتربی جهت همردیف‌سازی و بدست آوردن درستنمائی نهائی بین گفتار ورودی و تمامی ۱۰۰ مدل‌گفتاری موجود می‌گردد. این شیوه، اگرچه در ظاهر امر مانند شیوهٔ قبلی است، با این تفاوت که تمام‌محاسبات همزمان و بطور موازی صورت می‌گیرند، امّا دارای این ویژگی مطلوب است که می‌توان بااستفاده از تکنیک جستجوی شعاعی[۱٦]، در مقاطعی از عمل، فضای جستجو را با استفاده از یک شعاع‌جستجوی محدود که ازطریق اعمال یک سطح آستانهٔ نسبی (نسبت به

[15] Non-Ergodic
[16] Beam Search

درستنمائی بیشینه در هر مقطع)ایجاد می‌شود، کاهش داد. اینگونه جستجوها در موارد مختلف و با درجات موفقیت قابل قبولی درکاربردهای مختلف و بویژه کاربردهای بازشناسی گفتار مورد استفاده قرار گرفته‌اند [۹][۱۰]. به این‌ترتیب و با اعمال سطح آستانهٔ مناسب می‌توان ضمن داشتن نرخ خطای قابل قبول در بازشناسی،سرعت اجرای آلگوریتم بازشناسی را نیز افزایش داد.

۱ـ ۲ـ ۳ـ ۲ مقابله با اثرات شرایط مختلف محیطی

شرایط محیطی مختلف می‌توانند به شدت کیفیت یک سیستم بازشناسی گفتار را تحت تأثیر قراردهند. این شرایط شامل انواع نویزهای جمع‌شونده با سیگنال[۱۷] ناشی از منابع مختلف آلایندهٔ فضای‌صوتی می‌باشند. ازجملهٔ این موارد می‌توان به انواع نویز ناشی از کار دستگاههای مختلف برقی،موتورهای احتراقی و نظایر آنها، صدای سایر افراد که به عنوان نویز زمینه سیگنال گفتار را تحت تأثیر قرارمی‌دهد و یا هرگونه صدای دیگری نظیر صدای باد، ضربات و اصطکاک، جریان آب و نظائر آنها باشد. این منابع می‌توانند ازنظر طیف فرکانسی، انرژی سیگنال جمع‌شونده، و پایداری زمانی، کاملاًویژگی‌های متفاوتی داشته باشند.

منابع دیگری نیز که می‌توانند بر سیگنال گفتار موردنظر تأثیر بگذارند، شامل منابع نویز ضرب‌شونده[18] یا Convolutional است که عموماً مشخصات فرکانسی سیگنال گفتاری را تحت تأثیر قرار می‌دهند. این‌منابع عموماً شامل کانال‌های انتقال می‌باشند نظیر خط تلفن، میکروفون، ضبط گفتار، کانال رادیوئی وامثال آنها.

منابع نویز، تأثیر آنها بر کیفیت بازشناسی و روش‌های معمول در مقابله با آنها در [۲] مورد بررسی قرارگرفتند. یکی از روش‌هائی که درمقابله با نویز کارآئی خوبی از خود نشان داده است "معیار تصویروزن‌دهی شده[۱۹] " (WPM) نام دارد. در این روش، باتوجه به اثر نویز سفید بر روی بردارهای کپسترال یک‌معیار تصویر معرفی می‌شود که به کمک آن می‌توان برای هر بردار درستنمائی را به شکل جدیدی در هرحالت ازHMM محاسبه نمود [۱]. در بررسی‌های به‌عمل آمده مزایای چندی به شرح زیر برای روشWPM بدست آمد:

الف ـ این روش علاوه بر نویز سفید، برای نویزهای رنگی نیز تاحد زیادی مفید می‌باشد.

ب ـ دقّت بازشناسی در شرایط نویزی با استفاده از این روش افزایش قابل ملاحظه‌ای می‌یابد.

پ ـ آموزش مدل‌ها برای اعمال این روش دستخوش تغییر نشده وWPM تنها در مرحلهٔ بازشناسی‌قابل اعمال است.

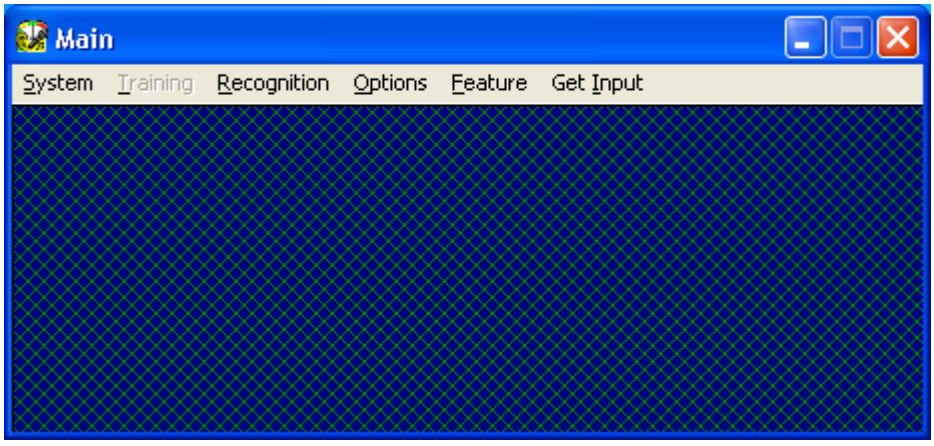
[17] Additive Noise
[18] Multiplicative
[19] Weighted Projection Measure

ت ـ هزینهٔ محاسباتی این روش بسیار ناچیز است.

پیاده‌سازی روشWPM مبیّن این نکته بوده است که در شرایط وجود نویز جمعی از نوع سفید، هرچه نسبت سیگنال به نویز کاهش یابد، کارآئی این روش بیشتر آشکار می‌گردد. به عنوان مثال، در شرایط سیگنال به نویزdB ۲۰، نرخ بازشناسی از ٤٤٪ به ۷۸٪ و در شرایط سیگنال به نویزdB ۱۰،این نرخ از ۹/۸٪ به ٤۸/٤٪ افزایش می‌یابد.

۱ـ ۲ـ ٤ واسط کاربر

برای آسان‌تر نمودن به‌کارگیری برنامه‌های تهیه شده برای بازشناسی گفتار گسستهٔ فارسی، اقدام به‌تهیهٔ یک واسط کاربرد برای این برنامه گردید. این برنامه که در محیط گرافیکWindows قابل اجراست،دارای محیط گرافیکی نشان داده شده در شکل ۱ـ ۱ می‌باشد.



شکل ۱ـ ۱ پنجرهٔ اصلی محیط گرافیکی.

این محیط گرافیکی دارای چند منو به شرح زیر می‌باشد:

الف ـ System : شامل دو بخشAbout وExit است که مورد اوّل اشارهٔ مختصری به ویژگی‌های‌این محیط گرافیکی داشته و مورد دوّم برای خروج از آن درنظر گرفته شده است.

ب ـ Training : این مسیر منو برای آموزش سیستم بازشناسی درنظر گرفته شده است و دارای دوگزینهٔ آموزش بهمراه استخراج ویژگی (Feature Extraction and Training) و آموزش تنها (Training only) می‌باشد.

پ ـ Recognition : این زیرمنو نیز خود دارای دو بخشBatch و Non-Batch می‌باشد. دربازشناسی بصورت Batch ، مجموعه‌ای از داده‌های گفتاری که قبلاً ضبط شده‌اند و یا از یک دادگان‌گفتاری استخراج گردیده‌اند، می‌توانند مورد استفاده قرار گرفته و آزمایش در مورد تمامی آنها

صورت‌گرفته و سپس نتایج اعلام گردد، درحالیکه در بازشناسی به صورت Non-Batch ، دادهٔ گفتاری مستقیماًاز میکروفون دریافت شده و بازشناسی می‌شود.

گزینهٔ Batch خود شامل دو گزینه یکی تحت عنوان استخراج ویژگی و بازشناسی (Feature Extraction and Recognition) و دیگری تحت عنوان بازشناسی تنها (Recognition only) می‌باشد.

ت ـ Options : دو گزینهٔ این بخش شامل تنظیم پارامترها (Parameter Setting) و غیرفعال کردن‌آموزش (Set Training off) می‌باشند. در منوی تنظیم پارامترها، کلیّهٔ پارامترهای کاری سیستم تنظیم‌می‌شوند و تمامی برنامه‌های سیستم براساس مقادیر تنظیم شده در این منو اجرا می‌شوند. پارامترهای‌تنظیم‌شده شامل تعداد حالت‌های مدل، تعداد عناصر مخلوط، تعداد مُدل‌ها، حداکثر زمان‌اختصاص‌یافته به یک بیان، حداکثر تعداد دفعات تکرار در آموزش ابتدائی، طول و میزان برهم‌افتادگی‌فریم‌ها (برحسب نمونه)، مرتبهٔ تحلیلLPC و ضرائب کپسترال، ابعاد بردار، مرتبهٔ رگرسیون در محاسبهٔ دلتا و مسیرها و اسامی فایل‌ها برای داده‌های آموزشی، ویژگی‌های آموزشی، داده‌های آزمایشی،ویژگی‌های آزمایشی و فایل مُدل‌ها می‌باشد. لازم به اشاره است که باتوجه به اینکه تمامی بخش‌های‌این برنامه از این پارامترها استفاده می‌نمایند، لازم است ابتدا این پارامترها تنظیم شده و سپس اقدام به‌استفاده از سایر ابزارها نمود. درعین حال، باتوجه به وابستگی بسیاری از این پارامترها به یکدیگر، لازم‌است تنظیم آنها متناسب با یکدیگر صورت پذیرد، چرا که در غیر اینصورت بخش‌های برنامه ممکن‌است دچار مشکل شوند. یک کلید نرم‌افزاری نیز در این منو برای برگشت مقادیر به مقادیر پیش‌فرض(Default Settings) پیش‌بینی گردیده است. گزینهٔ دوم زیرمنویOptions تنها برای غیرفعال کردن‌آموزش به‌کار می‌رود تا اشارهٔ سهو به گزینهٔ آموزش مدل‌ها موجب ازبین رفتن مُدل‌های موجود نشود.

ث ـ Feature : در این بخش تنها یک گزینه وجود دارد که مربوط به استخراج ویژگی‌ها می‌شود.این‌گزینه اقدام به استخراج ویژگی‌های سیگنال‌های گفتاری براساس مقادیر تنظیم شده در گزینهٔ تنظیم‌پارامترها در بخش قبل می‌نماید.

ج ـ Get Input : گزینه‌های این بخش جهت ضبط داده به برنامه اضافه شده‌اند و امکان جمع‌آوری‌دادهٔ گفتاری، بدون نیاز به برنامه‌های خارجی را فراهم می‌آورند. به این ترتیب امکان ضبط داده‌های‌گفتاری آموزشی یا آزمایشی با استفاده از نرم‌افزارGoldwave که از داخل برنامه قابل دسترسی است‌فراهم می‌گردد.

۱ـ ۳ بازشناسی گفتار پیوستهٔ فارسی

یکی از مهمّ‌ترین بخش‌های تعریف شده در این طرح پژوهشی، بخش بازشناسی گفتار پیوستهٔ فارسی‌می‌باشد. بازشناسی گفتار پیوسته، باتوجه به سابقهٔ کمی که در زبان فارسی دارد. از اهمیت بالائی‌برخوردار می‌باشد چرا که با شیوهٔ معمول مکالمه در زبان سروکار دارد. در این بخش از گزارش به بررسی‌موارد مختلفی که در این ارتباط در طرح پژوهشی حاضر مورد بررسی قرار گرفته است می‌پردازیم.

۱ـ ۳ـ ۱ دادگان و دایرهٔ لغات

۱ـ ۳ـ ۱ـ ۱ فارس‌دات

پیش از آنکه بتوان هرگونه تحقیقی بر روی بازشناسی گفتار پیوسته صورت دارد. نیاز به دادگان‌مناسب برای این منظور می‌باشد. متأسفانه در حال حاضر دادگان مطلوب و مناسبی برای این منظور درزبان فارسی وجود ندارد. دادگان شناخته شدهٔ موجود برای گفتار پیوستهٔ فارسی، دادگان فارس‌دات‌می‌باشد که دارای محدودهٔ کلماتی در حدود بیش از ۱۱۰۰ کلمه می‌باشد [۱۲]. این دادگان دارای جملاتی طراحی شده برای نیل به تعادل آکوستیک ـ فونتیک می‌باشد. بنابراین جملات ضبط شده ازمجموعه‌های گفتاری طبیعی (نظیر متن روزنامه‌ها، اخبار رسانه‌ها، مقالات، مکالمات روزمرّه و غیره)بدست نیامده‌اند و بنابراین برای بسیاری کاربردها در بحث بازشناسی گفتار پیوسته، ممکن است دادگان‌مطلوبی نباشند. با این‌حال باتوجه به منحصربه‌فرد بودن این دادگان، اقدام به استفاده از آن جهت‌پیاده‌سازی بازشناسی گفتار پیوستهٔ فارسی گردید.

این دادگان از حدود ٦۰۰۰ بیان (جمله) فارسی تشکیل شده که این جملات توسط ۳۰۰ گوینده (هرگوینده ۲۰ جمله) ادا گردیده‌اند. دو جمله از ۲۰ جملهٔ اداشده توسط هر گوینده، جملات خاصی هستندکه بین تمام گویندگان مشترک می‌باشند. درمجموع حدود ٤۰۰ جملهٔ مختلف در دادگان موجودمی‌باشد.

موارد چندی در این دادگان وجود داشتند که برای دستیابی به یک دادگان پایه برای بازشناسی گفتارپیوستهٔ فارسی می‌باید حلّ می‌شدند. برای این کار ابتدا ۳۰ واج پایه مورد استفاده در زبان فارسی‌مشخص گردیدند [۱]. سپس سیگنال‌های کلیهٔ بیان‌های گفتاری در این دادگان از طریق محیط نرم‌افزاری‌دادگان در سه فایل بزرگ استخراج گردیدند و در مرحلهٔ بعد تک‌تک بیان‌ها از این فایل‌ها استخراج و هریک به‌تنهائی ذخیره گردیدند. از میان سیگنال‌های ذخیره شده، گفتار گویندگان تهرانی انتخاب گردید(۱٤۷ گوینده). سپس فرکانس نمونه‌برداری سیگنال‌ها از٤٤/۱ KHz به١٦ KHz کاهش

یافت[۲۰]. پس ازآن تک‌تک فایل‌های گفتاری (۲۹٤۰ مورد) مورد بررسی دقیق قرار گرفته و موارد دارای اشکال (ازقبیل‌لهجه، مقطّع‌بودن، کش‌داربودن، غلط اداکردن برخی کلمات، مصنوعی اداکردن کلمات و اشکالات‌گفتاری نظیر نوک‌زبانی صحبت‌کردن و امثال آن) از مجموعه حذف گردیدند. درنتیجه، در مجموع ۲۷۰۷جمله از ۱۳۷ گوینده در دادگان باقی ماند. در این مجموعه حدود ٦۰٪ جملات (۱٦۱۵ جمله) توسط‌گویندگان مرد و بقیه (۱۰۹۲ جمله) توسط گویندگان زن ادا گردیده‌اند.

تقسیم داده‌های موجود به دو بخش آزمایش و آموزش بصورت تصادفی و با تخصیص حدوداً یک‌گوینده از هر سه نفر به مجموعهٔ داده‌های آزمایشی و بقیه به مجموعهٔ داده‌های آموزشی صورت‌پذیرفت. بنابراین، در مجموع ۱۸۱٤ جملهٔ آموزشی و ۸۹۳ جملهٔ آزمایشی برای کاربرد بازشناسی گفتارحاصل گردید.

برای تمامی مجموعهٔ آموزشی فوق، برچسب‌های واجی جملات با استفاده از تعاریف جملات‌بدست آمده و بصورت فایل‌هائی ذخیره گردیدند. برای جملات آزمایشی، برچسب‌های کلمه‌ای‌استخراج شده و ذخیره شدند، چرا که این‌گونه برچسب‌ها در کار بازشناسی برای بررسی نتایج و بدست‌آوردن دقّت کار موردنیاز می‌باشند. علاوه بر اینها، همچنانکه بعداً اشاره خواهد گردید برای انجام‌مناسب آموزش، نیاز به مُدل‌های اولیه‌ای برای هر یک از واحدهای پایه می‌باشد. از آنجائی که اینگونه‌مدل‌های اوّلیه نیازمند به برچسب‌های دارای مرزهای زمانی واج‌ها می‌باشند (برای آموزش اوّلیه)، نیازبه استخراج اینگونه برچسب‌ها برای تعداد محدودی از جملات آموزشی بوده است. به همین منظور،برای ۱۱۹ جمله از جملات آموزشی، برچسب‌های واجی دارای مرزهای زمانی از دادگان استخراج وپس از اصلاحات و تغییرات لازم در فرمت آنها، تک‌تک و به صورت فایل‌های جدا (به‌ازاء هر بیان‌جمله‌ای) ذخیره گردیدند.

دادگانی که به این ترتیب بدست آمد، سپس به عنوان دادگان مورد استفاده در کلیهٔ مراحل آموزش وتست سیستم‌های بازشناسی گفتار پیوستهٔ پیاده شده مورد استفاده قرار گرفت.

۱ـ ۳ـ ۱ـ ۲ دادگان اخبار

باتوجه به محدودیت‌های موجود در دادگان فارس‌دات ازقبیل محدودیت دایرهٔ لغات، طراحی‌براساس تعادل آکوستیک ـ فونتیک که موجب غیرطبیعی بودن پراکندگی اصوات و کلمات در دادگان‌می‌گردد، عدم برخورداری از بخش‌های وابسته به گوینده و تطبیق گوینده و نظائر آنها، تصمیم به‌جمع‌آوری دادگانی، برای برآورده‌کردن این منظورها، گرفته شد. در این راستا، باتوجه به هزینهٔ بسیاربالای پروژه‌های جمع‌آوری دادگان، تصمیم به استفاده از گفتار رادیوئی گرفته شد. دادگان

[20] Downsample

«اخبار» متشکل از حدود ۲۷ ساعت دادهٔ گفتاری جمع‌آوری شده از بخش‌های مختلف خبری رادیو (شبکه‌های مختلف صدای جمهوری اسلامی ایران) می‌باشد. این دادگان تا حدود زیادی پوشش کلمه‌ای مناسبی ازدایره لغات مطرح شده در مسائل اجتماعی، سیاسی، اقتصادی، فرهنگی و نظائر آنها که در بخش‌های‌خبری مطرح می‌گردند داشته و از این‌نظر دارای تنوع مناسب است. تمامی گفتار ضبط شده به صورت‌متن خوانده شده می‌باشد که با استفاده از یک گیرندهٔ رادیوئی با کیفیت مناسب (SONY ST-H3600) و ازموج FM ، ابتدا بر روی نوارهای کاست از نوع فلزی(METAL) ضبط و سپس با ارتباط الکتریکی‌مستقیم، ازطریق کارت صوتی به کامپیوتر منتقل گردیده‌اند.

برای ایجاد برچسب‌های لازم برای این گفتار، ابتدا تمامی ۲۷ ساعت گفتار (بجز بخش‌های مربوط به‌گزارش‌های تلفنی یا اخبار ورزشی و سیاسی خارجی که از اسامی غیرمعمول در فارسی استفاده‌می‌کنند) به فارسی تایپ گردیدند. برای این منظور از برنامهٔ تایپ ویژه‌ای استفاده گردید که خروجی آن‌به‌راحتی بتواند برای پردازش‌های بعدی مورد استفاده قرار گیرد. سپس تمامی فایل‌های خروجی توسط‌برنامه‌ای ویژه خوانده شده و براساس اصول دستوری و مفهومی و در شرایط با نظارت به رشتهٔ واجی‌قابل استفاده برای کاربردهای بازشناسی تبدیل گردیدند. متأسفانه به علت محدودیت منابع مالی،امکانات و زمان، تکمیل و بررسی دقیق و دسته‌بندی فایل‌های دادگان مقدور نگردید و برای اینکه این‌دادگان بتواند در کاربردهای بازشناسی مورد استفاده قرار گیرد، نیاز به کار بیشتری وجود دارد.

۱ـ ۳ـ ۲ سیستم‌های به‌کار گرفته شده

در بازشناسی گفتار پیوسته، امروزه دو شیوه در جهان بیشتر مورد توجه می‌باشند. یکی استفاده ازHMM ها و دیگری استفاده از سیستم‌های ترکیبی مبتنی برHMM و شبکه‌های عصبی. در این طرح‌پژوهشی هر دو شیوه مورد توجه و پیاده‌سازی قرار گرفته‌اند که در اینجا به شرح هر دو شیوه می‌پردازیم.

۱ـ ۳ـ ۲ـ ۱ مدل‌های مارکوف پنهان با چگالی مشاهدات پیوسته (CDHMMs)

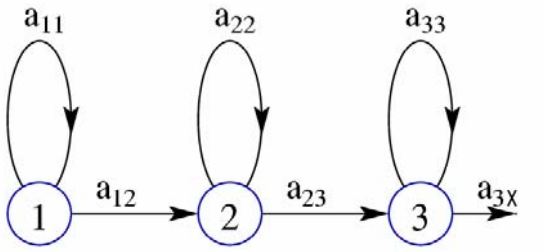
مدل‌های مارکوف پنهان، همچنانکه در مورد بازشناسی گفتار گسسته نیز ذکر گردید، پراستفاده‌ترین‌شیوه برای بازشناسی گفتار محسوب می‌شوند.HMMهای دارای چگالی مشاهدات پیوسته(CDHMMs) بخاطر دقت بالاتر در مُدل‌سازی گفتار، در این میان، مورد استقبال فراوانی قرار گرفته‌اند. ازهمین‌رو، در بازشناسی گفتار پیوسته در این طرح پژوهشی نیز، به عنوان اوّلین شیوهٔ بازشناسی موردتوجه و بررسی قرار گرفته‌اند. در ادامه به بررسی کلّی شیوه‌های مطالعه و پیاده‌سازی شده در این روندمی‌پردازیم.

۱ـ۳ـ۲ـ۱ـ۱ استخراج ویژگی‌ها

روش‌های استخراج ویژگی‌ها در بازشناسی گفتار پیوسته نیز، شباهت فراوان به شیوه‌های استخراج‌ویژگی گفتار گسسته دارند. در این طرح پژوهشی نیز برای پیاده‌سازی سیستم بازشناسی اقدام به‌استخراج ویژگی‌های کپسترال مبتنی برLPC گردیده است. در این راستا مؤلفه‌های انرژی، دلتا و دلتا ـدلتا، حسب مورد به ضرائب فوق اضافه گردیده‌اند. درعین حال، در برخی پیاده‌سازی‌ها نیز اقدام به‌استخراج ضرائبMFCC و استفاده از آنها، بهمراه ضرائب انرژی، دلتا و دلتا ـ دلتا گردیده است.

۱ـ۳ـ۲ـ۱ـ۲ مُدل‌های مورد استفاده

مُدل‌های استفاده شده در این موارد،CDHMM با توپولوژی نشان داده شده در شکل ۱ـ۲ می‌باشند. لازم به اشاره است که اینگونه مُدل‌ها برای مُدل‌سازی واج‌های پایهٔ زبان فارسی به کار گرفته‌شده‌اند که بنابراین در مجموع بالغ بر ۳۰ مدل می‌گردند. برای مُدل‌سازی سکوت و فاصلهٔ بین کلمات از مُدل‌های مشابهی با امکان جهش از حالت ورودی به حالت خروجی استفاده گردیده است. لازم به اشاره‌است که چگالی احتمال خروجی در هر یک از حالت‌ها می‌تواند یک تک ـ گوسی یا یک مخلوط ازگوسین‌ها باشد.



شکل ۱ـ۲ توپولوژی استفاده شده برای مدل‌سازی واج‌ها (مدل چپ ـ راست Bakis).

۱ـ۳ـ۲ـ۱ـ۳ آموزش سیستم [۳][۱۳]

برای آموزش سیستم مبتنی بر CDHMMs ، مقداردهی اوّلیهٔ پارامترها از اهمیت فوق‌العاده‌ای برخوردار است. از آنجا که روش استفاده شده برای آموزش، مبتنی بر آلگوریتم Baum-Welch است، لازم است ابتدا مدل‌ها آموزش مناسبی دیده تا سپس بتوانند به عنوان مدل‌های آغازین برای آموزش نهایی مورد استفاده قرار گیرند.

پیش از ادامهٔ بحث توضیح این نکته نیز لازم است که در آموزش CDHMMs برای مدل‌سازی واج‌ها در این طرح پژوهشی، علاوه بر آلگوریتم Baum-Welch از آلگوریتم معروف Segmental K-

means نیز استفاده گردید[۲]، ولی به دلیل توانائی بالاتر و اثبات‌شدهٔ آلگوریتمBaum-Welch ، در اینجا تنها به اشاره به این روند آموزش بسنده می‌کنیم.

در آموزش مُدل‌های واجی، در مرحلهٔ مقداردهی اوّلیه، همانند روندی که در مورد بازشناسی گفتارگسسته اشاره شد، ابتدا با استفاده از بخشی از دادگان آموزشی که دارای برچسب‌های زمانی می‌باشند،بردارهای ویژگی مربوط به هر واج از فایل ویژگی‌های گفتار جدا شده و مجموعهٔ بردارهای ویژگی‌بدست آمده برای هر واج بین ۳ حالت مُدل مربوطه به تساوی (تاحد ممکن) تقسیم و مقادیر بردارهای‌میانگین و واریانس برای یک تک گوسین به‌ازاء هر حالت محاسبه می‌گردند. در مرحلهٔ بعد برای‌بدست‌آوردن تخمین اوّلیهٔ مناسب‌تر، اقدام به اجرای آلگوریتم ویتربی برای بهبود تخصیص بردارها به‌حالت‌ها در مدل‌های براساس واج‌ها می‌گردد. دنبالهٔ حالت‌های بدست آمده از هر مرحلهٔ اجرای‌آلگوریتم ویتربی برای اصلاح مرحلهٔ مقداردهی فوق استفاده می‌گردد و بدین ترتیب پارامترهای‌جدیدی برای مُدل‌ها بدست می‌آیند. این آلگوریتم می‌تواند بصورت تکرارشونده اجرا شود. در این‌صورت تکرار آن تا رسیدن به مرحله‌ای از همگرائی می‌تواند ادامه یابد. این معیار همگرائی می‌تواندبراساس تغییرات در میزان درستنمائی میانگین بدست آمده از هر مرحله تعریف شده و یا صرفاً به‌اجرای تعداد مشخصی از دفعات تکرار بسنده گردد.

تا این مرحله، تنها تخمین اوّلیه‌ای از پارامترهای چگالی احتمال مشاهدات حالت‌ها برای هر مُدل‌بدست آمده است و پارامترهای انتقال حالت مورد توجه قرار نگرفته‌اند. بنابراین در مرحلهٔ بعد این‌پارامترها نیز مورد توجه قرار می‌گیرند. این مرحله شامل آموزشBaum-Welch تک‌تک مُدل‌های پایه‌می‌باشد. روابط مربوط به آموزش Baum-Welch در [۳]و [۱۳] ارائه گردیده‌اند. اجرای تکراری این‌آلگوریتم مدل‌های مناسبی برای تک‌تک واج‌های موردنظر ایجاد می‌نماید که در مرحلهٔ نهایی آموزش‌قابل استفاده می‌باشند. لازم به اشاره است که این مرحله نیاز به برچسب‌های زمان‌بندی شدهٔ واجی‌دارد. همانطور که پیش از این اشاره گردید، تعداد فایل‌های آماده شده از این دست در دادگان تهیه شده‌محدود است و لذا چنین آموزشی بسیار محدود خواهد بود.

آخرین مرحله و مرحلهٔ اصلی آموزش، Embedded Baum-Welch می‌باشد. در این مرحله تمامی‌جملات آموزشی مورد استفاده واقع می‌شوند و نیازی به برچسب‌های زمانی نمی‌باشد و تنهابرچسب‌های واجی جملات مورد استفاده واقع می‌شود. بنابراین در این مرحله لازم است مدل‌های‌ترکیبی[۲۱] (FSN) به‌ازاء هر جملهٔ آموزشی ایجاد شده و مقادیر مربوطه (صورت و مخرج روابط بازتخمین) محاسبه و جداگانه و به‌ازاء هر مُدل ذخیره شوند. این کار باید برای تمامی جملات

[21] Finite State Network

آموزشی‌انجام شده و سپس مقادیر نهائی بدست آمده برای صورت و مخرج روابط پارامترهای مُدل هر واج‌جهت محاسبات نهائی مورد استفاده قرار می‌گیرند [۱۳].

مُدل‌های بدست آمده از این مرحلهٔ آموزشی، مُدل‌های نهائی قابل استفاده در بازشناسی گفتارپیوسته می‌باشند. البته مسلّم است که مرحلهٔ فوق نیز باید بصورت تکراری انجام شود تا نتایج تاحدّامکان به سمت یک ماکزیمم محلّی میل نمایند.

۱ـ۳ـ۲ـ۱ـ٤ بازشناسی (دکودینگ) گفتار پیوسته

برای بازشناسی گفتار پیوسته، همانند گفتار گسسته، مناسب‌ترین روش آلگوریتم ویتربی می‌باشد. بااین همه، پیاده‌سازی آلگوریتم ویتربی برای بازشناسی گفتار پیوسته چندان ساده نمی‌باشد. درمقایسه باگفتار گسسته، در صورتی که بخواهیم روش مشابهی را در اینجا به‌کار ببریم، نیاز به ایجاد تمامی مُدل‌های‌جملات ممکن با استفاده از مُدل‌های واجی و سپس اعمال مقادیر گفتار جدید به تمامی این مُدل‌هابرای بدست آوردن بالاترین درستنمائی می‌باشد. باتوجه به تعداد فوق‌العاده زیاد جملات ممکن، این‌کار در عمل غیرممکن می‌نماید.

برای عملی ساختن این مورد، کار به‌این شرح پی گرفته می‌شود[۱۰][٤][۱۳]. برای جستجو به‌دنبال مناسب‌ترین جمله، بجای ساختن تمامی جملات ممکن در ابتدای کار، از ابتدا تنها مُدل تمامی‌کلماتی که می‌توانند در مرحلهٔ اوّل و در ابتدای جمله قرار بگیرند ساخته می‌شود. از آنجا که دایرهٔ کلمات‌دادگان محدود و حدود ۱۱۰۰ کلمه می‌باشد، انجام این کار در وهلهٔ اوّل مقدور است. سپس با استفاده ازیک نشانه[۲۲] که به ابتدای تمام مُدل‌ها منتقل می‌شود کار ادامه می‌یابد. این نشانه حامل مقادیردرستنمائی مسیر تا آن لحظه و نیز مسیر طی شده می‌باشد. در لحظات زمانی بعدی این نشانه به‌حالت‌های بعدی منتقل شده و درعین حال در صورت وقوع همزمان دو یا چند نشانه در یک حالت، وتنها نشانهٔ دارای درستنمائی بالاتر نگهداشته شده و سایر نشانه‌ها دور ریخته می‌شوند. به این ترتیب درهنگام رسیدن به انتهای هر کلمه، هر نشانه مجدداً در مقابلش به تعداد کلمات دادگان مسیر خواهدداشت. برای کاهش مسیرهای جستجو، می‌توان در اینجا از یک روش نظیر جستجوی شعاعی[۲۳] استفاده‌نمود. چنین روشی با استفاده از هَرَس‌کردن[۲٤]، با اعمال یک سطح آستانه در مقاطع زمانی مختلف، تنهانشانه‌های دارای بالاترین درستنمائی را حفظ نموده و سایر نشانه‌ها را حذف می‌نماید.

[22] Token
[23] Beam Search
[24] Pruning

آلگوریتم فوق، علیرغم زیر بهینه‌بودن، باعث کاهش فراوان در فضای جستجو، تنها با ازدست دادن‌جزئی دقت بازشناسی می‌شود. علاوه براین، در این آلگوریتم امکان اعمال یک مدل زبان در مقاطع‌کلمات وجود دارد که به‌نوبهٔ خود می‌تواند به مراتب فضای جستجو را کاهش داده و دقت بازشناسی راافزایش دهد.

۱ـ۳ـ۲ـ۱ـ۵ استفاده از چگالی مشاهدات مخلوط

علیرغم اینکه استفاده از چگالی مشاهدات مخلوط از نخستین گام آموزش مقدور می‌باشد،مدل‌سازی با چگالی مشاهدات تک گوسین مزایایی ایجاد می‌نماید که عموماً ترجیح داده می‌شود مُدل‌سازی اوّلیه به این صورت انجام شود. بنابراین لازم خواهد بود که پس از ساخته‌شدن سیستم‌تک‌گوسین، اقدام به افزایش تعداد عناصر مخلوط به تعداد دلخواه شود. برای این منظور در این طرح‌پژوهشی از آلگوریتمی تحت عنوان شکافت مخلوط استفاده گردیده است [٤][۱۳].

در این آلگوریتم، تلاش در هر مرحلهٔ افزودن تعداد عناصر چگالی مشاهدات به‌اندازهٔ یک واحد است.برای این منظور ابتدا عنصر دارای بالاترین وزن در مخلوط حاضر یافت شده و به دو عنصر همانند با وزن‌نصف وزن عنصر قبلی و با بردار میانگینی تغییریافته به‌اندازهٔ ۰/۲ انحراف معیار قبلی تقسیم می‌شود.پس از آنکه چگالی‌های مشاهدات تمام حالت‌های سیستم به این صورت افزایش عضو داده شدند،تمامی پارامترهای سیستم مجدداً تحت آموزش قرار می‌گیرند. بدیهی است که در این بحث فرض براستفاده از تعداد عناصر مخلوط یکسان برای چگالی‌های مشاهدات تمامی حالت‌های تمامی مدل‌های‌سیستم است. علاوه براین، آشکار است که پیاده‌سازی این آلگوریتم نیاز به توانائی آموزش چگالی‌های مخلوط در برنامهٔ آموزش embedded دارد. همچنین برای استفاده از این‌گونه مدل‌ها، برنامهٔ بازشناسی نیز باید به همین ترتیب تغییر داده شود.

مراحل فوق برای هر واحد افزایش تعداد عناصر چگالی مخلوط باید تکرار شوند. تعداد عناصرمطلوب می‌تواند با افزایش تعداد عناصر و ملاحظهٔ نتیجهٔ بازشناسی تا زمانی که پدیدهٔ undertraining رخ نداده باشد، بدست آید.

۱ـ۳ـ۲ـ۱ـ٦ همردیف‌سازی زمانی [۳][۱٤]

همردیف‌سازی زمانی گفتار پیوسته، در این طرح پژوهشی، پیش از پیاده‌سازی بازشناسی گفتارپیوسته و به عنوان مقدمه‌ای بر آن و همچنین برای مقدورساختن بدست آوردن یک هم‌ردیفی بین‌حالات مدل‌ها و نمونه‌های گفتار صورت گرفت. کاربردهای فراوانی در پردازش گفتار برای این کارمتصوّر است. ازجمله تعیین خودکار برچسب‌های زمانی آواها در گفتار پیوسته که به عنوان مثال

درهمین کاربرد، یعنی بازشناسی گفتار، پیش از این استفاده از آن مورد اشاره قرار گرفت، یا بررسی دقیق‌ترارتباط حالات درHMM ها با بخش‌های مختلف آوا برای بررسی‌های آکوستیک ـ فونتیکی.

مدل‌های به‌کار رفته در این کاربرد نیز از نوعHMM های با چگالی مشاهدات پیوسته و شبیه به‌مدل‌های اشاره شده در بخش ۱ـ ۳ـ ۲ـ ۱ـ ۲ می‌باشند. همچنین از ویژگی‌های کپسترال بدست آمده از LPC بهمراه ضرائب انرژی در کنار مقادیر دلتا و دلتا ـ دلتای آنها با مرتبهٔ ۱۲ در این کاربرد استفاده شده است. برای آموزش مدل‌ها، آلگوریتم K-means Clustering مورد استفاده قرار گرفته است که ازنظر پیاده‌سازی در این کاربرد، ساده‌تر از آلگوریتم Baum-Welch می‌باشد [۳]. برای همردیف‌سازی نیزآلگوریتم ویتربی مورد استفاده قرار گرفته است.

۱ـ۳ـ۲ـ۱ـ۷ مُدل‌سازی وابسته به متن[۲۵]

عدم توجه به شرایط متنی در بازشناسی گفتار پیوسته می‌تواند محدودیتی در بهبود کیفیت سیستم‌بازشناسی ایجاد نماید. به همین جهت این امر از اهمیت نسبتاً بالائی برخوردار است. در این طرح‌پژوهشی، به این نکته نیز توجه گردیده است.

برای بررسی تأثیر شرایط متنی بر کیفیت بازشناسی گفتار پیوستهٔ فارسی، با استفاده از یک ابزارموجود، اقدام به پیاده‌سازی و بررسی نتایج در یک سیستم وابسته به متن گردید [۱۵]. به این منظور یک‌مجموعه مدل وابسته به متن از نوع سه‌آوائی داخل کلمه‌ای[۲۶] براساس تمام شرایط متنی داخل کلمه‌ای‌موجود در دادگان تشکیل داده شد. تعداد مُدل‌ها در این مجموعه به بیش از ۲۰۰۰ و تعداد حالت‌ها به ۷۰۰۰ بالغ گردید. بدیهی است که در این شرایط، باتوجه به محدودیت داده‌های آموزشی، بسیاری از مدل‌ها ازآموزش مطلوب برای یک سیستم ناوابسته به گوینده برخوردار نخواهند گردید.

برای حلّ این مشکل از شیوهٔ گره‌زدن پارامترها برای کاهش تعداد کلّ پارامترها استفاده گردید [۱٦].در این کاربرد، حالت‌های مدل‌ها برای این منظور درنظر گرفته شدند و یک آلگوریتم خوشه‌بندی برای‌گروه‌بندی آنها پیاده‌سازی و سپس حالت‌های قرار گرفته در هر گروه (خوشه) گره زده شدند. به این‌ترتیب تعداد کلّ حالت‌ها در سیستم به ۱۰۲٦ کاهش یافت. درنتیجه بازشناسی از کیفیت مطلوب‌تری‌برخوردار گردید.

[25] Context Dependent Modeling
[26] Word-internal triphone

علاوه براین، در این پیاده‌سازی از شیوهٔ آموزش MAP[۲۷] با مقادیر پیشینهٔ[۲۸] اصلاح شده نیز استفاده‌گردید و ملاحظه شد که تخمینMAP در این کاربرد نتایج نسبتاً مناسب‌تری ارائه می‌نماید [۱۵].

۱ـ۳ـ۲ـ۱ـ۸ مُدل‌سازی هجا

از آنجا که زبان فارسی از ساختار هجائی ساده‌تری در مقایسه با بسیاری زبان‌های اروپائی نظیرانگلیسی برخوردار است، تصمیم گرفته شد که بر روی مدل‌سازی هجا در زبان فارسی نیز بررسی‌های لازم صورت پذیرد. به همین منظور و باتوجه به محدودیت هجا در زبان فارسی به سه نوع CV ، CVC وCVCC و نیز اینکه تعداد واکه‌ها در زبان فارسی محدود می‌باشد که خود باعث کاهش بیشتر تنوع در انواع هجا می‌گردد، مدل‌سازی براساس هجاهای زبان فارسی صورت پذیرفت [۳][۱۶]. استخراج هجاها از دادگان تحت بررسی نشان داد که در حدود ۸۴۵ هجا در این دادگان وجود دارد (که البته بسیار کمتر از کل هجاهای موجود در زبان فارسی که بالغ بر ۴۰۰۰ تخمین زده می‌شود، می‌باشد). مدل‌سازی براساس هجا، هرچند در عمل موفقیت‌آمیز می‌نمود، امّا بالابودن تعداد پارامترهای سیستم، امکان‌مدل‌سازی براساس چگالی‌های مخلوط را فراهم نمی‌نمود. به همین دلیل به شیوه‌هائی برای کاهش‌پارامترهای سیستم دست یازیده شد که از آن میان می‌توان به مدل‌سازی با تعداد حالات کاهش یافته ونیز گره‌زدن حالت‌ها اشاره نمود. شیوهٔ اخیر در عمل مفید واقع گردید و کلّ تعداد حالت‌های سیستم را ازبیش از ۷۶۰۰ حالت به ۱۰۱۸ حالت کاهش داد و بنابراین باعث افزایش دقت بازشناسی، چه در شرایط تک ـ گوسین و چه در شرایط استفاده از چگالی مخلوط گوسین گردید [۱۶] .

۱ـ ۳ـ ۲ـ ۲ مدل‌های ترکیبی HMM و شبکهٔ عصبی[۲۹]

یکی از روش‌هائی که در سال‌های اخیر برای بازشناسی گفتار پیوسته مورد توجه نسبی قرار گرفته‌است، استفاده از ترکیبHMM و شبکهٔ عصبی می‌باشد. علیرغم توانائی بالای نشان داده شده از سوی‌شبکه‌های عصبی در کاربردهائی نظیر طبقه‌بندی الگوها، این شبکه‌ها نتوانسته‌اند در طبقه‌بندی الگوهای‌گفتاری، به خوبی روش‌هائی نظیرDTW وHMM عمل نمایند. عمدهٔ این ضعف ناشی از دینامیک‌گفتار تلقی می‌شود که شبکهٔ عصبی در برخورد با آن ضعف نشان می‌دهد.

با این همه، در سیستم‌های مبتنی برHMM نیز ضعف‌هائی وجود دارد که درصورت رفع آنها انتظارنتایج بهتری را می‌توان داشت. مشکلات ناشی از استفاده از تخمینML در آموزش مدل‌ها،

[27] Maximum *a Posteriori*
[28] Prior parameters
[29] Hybrid HMM/ANN modeling

درجه اوّل‌فرض نمودن مُدل مارکوف، فرض ناوابستگی بردارهای ورودی به یکدیگر و بالاخره جایگزین‌نمودنpdf در هر حالت با یک مخلوط گوسین (یا نظایر آن) ازجملهٔ محدودیت‌ها در مدل‌سازی به کمکHMM تلقی می‌شوند. درمقابل، بررسی‌های انجام شده نشان داده است که در صورت استفاده از شبکهٔعصبی مصنوعی، می‌توان برخی از این ضعف‌ها را تاحدّ زیادی رفع نمود. ازجمله شبکه‌های عصبی‌قادرند به‌خوبی وابستگی بین بردارهای ورودی را آموخته و نیز برای تقریب تابع چگالی احتمال‌مشاهدات حالت مورد استفاده واقع شوند. علاوه براین، در صورت استفاده از سخت‌افزارهای موازی،شبکه‌های عصبی از مزایای فراوانی برخوردارند. یکی دیگر از مزایای شبکه‌های عصبی آن است که‌علیرغم فاز آموزش نسبتاً کُند، در هنگام کار از سرعت بالائی برخوردارند.

باتوجه به موارد فوق، در این طرح پژوهشی، برای بررسی عملکرد سیستم‌های ترکیبی، اقدام به‌پیاده‌سازی یک سیستم مبتنی بر ترکیبHMM و شبکه‌های عصبی گردید [۳][٤]. این پیاده‌سازی‌براساس ۳۲ مدل مبتنی بر واج‌های زبان فارسی که در بخش ۱ـ ۳ـ ۲ـ ۱ـ ۲ مورد اشاره قرار گرفتند،صورت گرفته است. ضرائب بردارهای ویژگی نیز همانند موارد استفاده شده در بخش ۱ـ ۳ـ ۲ـ ۱می‌باشند. مُدل‌های استفاده شده، برای امکان‌پذیر شدن آموزش مناسب، از نوع HMM های تک‌حالته‌می‌باشند که تخمین تابع چگالی احتمال مشاهدات آنها برعهدهٔ یک شبکهٔ عصبیMLP می‌باشد.شبکه دارای ۳۲ خروجی به‌ازاء ۳۲ واج (یا سکوت) مورد نظر است. بردار ورودی شبکه نیز شامل۱۲+c فریم از پارامترهای ویژگی ورودی می‌باشد. در عمل مقدارc از ۰ تا ۳ برای یافتن بهترین شرایط‌تغییر داده شد، که مقدار ۳ = c یعنی داشتن پارامترهای ۷ فریم در ورودی مطلوب تشخیص داده شدچرا که افزایش بیش از این مقدار موجب پیچیدگی بیش ازحدّ شبکهٔ عصبی می‌گردد.

پیش از اعمال پارامترهای ورودی به شبکه، لازم است که این پارامترها نرمالیزه شوند که بهترین‌شرایط کاری در صورت نرمالیزه‌کردن این بردارها با میانگین صفر و واریانس ۰/۵ بدست آمد. شبکهٔعصبی مورد استفاده از نوعMLP دولایه با لایهٔ مخفی دارای ۱۵۰۰ تا ۳۰۰۰ نرون و با تابعtanhمی‌باشد.

آموزش شبکهٔ عصبی با انتخاب نرخ آموزش بزرگ آغاز شده و با بررسی منحنیCross-validation ودر صورت نوسانی‌شدن آن درحین آموزش میزان این نرخ کاهش می‌یابد. این عمل مرتباً و در طول‌مرحلهٔ آموزش تا رسیدن به نقطهٔ مطلوب ادامه می‌یابد. لازم به اشاره است که در اینجا نیز از حدود۱۸۰۰ جملهٔ آموزشی دادگان فارس دات برای آموزش سیستم استفاده گردید، لکن به علّت نیاز سیستم‌به برچسب‌های زمانی در هنگام آموزش، اجباراً تمامی این جملات (البته به صورت خودکار و نه‌بصورت دستی) برچسب زمانی زده شدند. علاوه براین، برای امکان‌پذیرشدن بهره‌گیری از

ویژگی‌مطلوبHMM در مدل‌نمودن دینامیک گفتار، مقادیر احتمال انتقال حالت درHMMها بصورت آماری‌تخمین زده شده و مقداردهی شدند.

برای بازشناسی در این سیستم نیز، نظیر سیستم مبتنی برHMM (بخش ۱ـ ۳ـ ۲ـ ۱)، از روش‌جستجوی شعاعی و انتقال نشانه استفاده گردیده است که به دلیل اینکه این مورد مفصلاً در بخش۱ـ ۳ـ ۲ـ ۱ـ ٤ توضیح داده شد، در اینجا از اشارهٔ مجدد به آن خودداری می‌گردد. لازم به اشاره است که‌در این کاربرد، برای بهبود شرایط بازشناسی، از شعاع جستجوی متغیر استفاده گردیده است. در این‌صورت، در ابتدا شعاع جستجو بزرگ انتخاب می‌شود چرا که هرگونه اشتباه در کلمهٔ اوّل قابل بازگشت‌نبوده و باعث آسیب کلّی در بازشناسی می‌گردد. سپس و با پیشرفت مرحلهٔ بازشناسی به‌تدریج شعاع‌جستجو و بنابراین فضای جستجو کاهش داده می‌شود.

۱ـ ٤ پیاده‌سازی‌ها و بررسی نتایج

تمامی موارد برشمرده شده در بخش‌های مختلف فوق مورد بررسی و آزمایش تجربی قرار گرفتند. دراینجا تنها به نتایج مهمّی که در این بررسی‌ها بدست آمده است، اشاره می‌گردد. لازم به تذکر است که‌نتایج هر قسمت در گزارش‌های مرحله‌ای این طرح پژوهشی مورد اشاره و بررسی دقیق‌تر قرار گرفته‌است (مراجع [۱] تا [٤]).

در بخش بازشناسی گفتار گسستهٔ فارسی، سیستم مبتنی برCDHMM به نتایج مناسب‌تری دست‌یافت. این نتایج، با استفاده از دادگان ناوابسته به گویندهٔ ارقام فارسی (ده کلمه‌ای) برای سیستم دارای‌چگالی احتمال مخلوط پنج عضوی، حدود ۹٤٪ می‌باشد. علاوه بر این، دقت سیستم در شرایط وجودنویز سفید با نسبتdB ۱۵ N = S تنها حدود ۱۸/۸٪ می‌باشد که با اعمال آلگوریتمWPM این میزان‌به ٦۷/٦٪ افزایش می‌یابد [۲]. به این ترتیب سیستم بازشناسی از دقت نسبی مطلوب برخوردار است وکارآئیWPM در شرایط وجود نویز سفید (و البته سایر انواع نویز ـ که نتایج مربوطه در [۲] ارائه شده‌است) آشکار گردیده است.

گسترش سیستم بازشناسی گفتار گسسته به دادگان صدکلمه‌ای، براساس مُدل‌هائی با ۸ حالت و ٦عنصر مخلوط به‌ازاء هر حالت به نتیجه‌ای معادل ۸٦/٤٪ دقت در بازشناسی منجر گردید [٤]. این نتیجهٔبازشناسی باتوجه به ده برابر شدن فضای جستجو نسبت به شرایط قبل، قابل قبول می‌باشد. اعمال‌آلگوریتمی برای کاهش فضای جستجو و افزایش سرعت نیز مفید بوده و باعث گردیده سرعت سیستم‌به بیش از دو برابر، درازاء کاهش تنها چند درصد از دقت بازشناسی، افزایش یابد.

در بازشناسی گفتار پیوسته، سیستم مبتنی برCDHMM به شکل ارائه شده پیاده‌سازی گردید. درمرحلهٔ بازشناسی، با اعمال مدل زبان «زوج ـ کلمه» و آزمایش بر روی نزدیک به ۹۰۰ جملهٔ آزمایشی‌دادگان، با استفاده از سطح آستانهٔ ۱۰۰ در جستجوی شعاعی، نرخ بازشناسی در حدود ۸۳٪ برای‌سیستم دارای یک عنصر مخلوط و ۹۵/۱٪ برای سیستم دارای ٦ عنصر مخلوط بدست آمده است [٤].لازم به تذکّر است که بدنبال اصلاحات، تغییرات و تلاش‌هائی که در جهت بهبود کیفیت کاری این‌سیستم صورت گرفته، نرخ بازشناسی آن تاحدودی اصلاح گردیده به‌نحوی که در آخرین آزمایش‌های صورت پذیرفته، دقّت بازشناسی[30] این سیستم با استفاده از مدل زبان زوج کلمه (Word Pair)، پهنای‌شعاع ۱۲۵ و با چگالی احتمال مشاهدهٔ تک ـ گوسین بر روی کلّ مجموعهٔ آزمایشی دادگان، ۸۸/۱٪ ونرخ بازشناسی[31] آن ۹۲/۵٪ بوده است.

برای سیستم طراحی شده برای همردیف سازی زمانی، نتایج کار در مرجع [۳] به تفصیل ارائه‌گردیده‌اند. تنها به ذکر این بخش از نتایج اکتفا می‌نمائیم که برای ۷٦/۸ درصد از واج‌ها، اختلاف زمانی‌طول واج‌ها، ناشی از همردیف‌سازی زمانی و مقادیر برچسب زدهٔ دستی کوچکتر یا مساوی ۳ فریم‌می‌باشد.

مُدل‌سازی‌های وابسته به متن و براساس هجا در زبان فارسی نیز به‌نوبهٔ خود توانسته‌اند بهبودی‌مناسبی نسبت به سیستم پایهٔ مدل‌شده براساس واج‌ها (ناوابسته به متن) ایجاد نمایند. در این‌پیاده‌سازی‌ها، مدل‌سازی براساس هجا توانسته دقت بازشناسی کلمه را که در سیستم بازشناسی‌ناوابسته به متن براساس واج‌ها، حدود ٤٦/۹٪ بوده، تا ٦۸/۳٪ افزایش دهد. لازم به تذکر است که این‌نرخ‌ها مربوط به استفاده از سیستم تک گوسین و بدون اعمال مُدل زبان است. علّت عدم اعمال مُدل‌زبان در این شرایط، افزایش خطای سیستم جهت امکان‌پذیرنمودن بررسی بهبودی حاصل از اعمال‌مُدل جدید می‌باشد. دقّت همین سیستم، با افزایش مقدار عناصر گوسین در چگالی احتمال مشاهدات،تا ۷۸/۱٪ (به‌ازاء ۵ عنصر مخلوط) افزایش یافته است [٤][١٦].

در هنگام مُدل‌سازی وابسته به متن، دقت سیستم در شرایط مشابه فوق تا حدود ٦۷/۸٪ افزایش‌داشته است که بسیار نزدیک به نتیجهٔ حاصل از مدل‌سازی هجا می‌باشد. علاوه براین، چنانچه از تخمینMAP بهبود یافته به این منظور استفاده نمائیم. دقّت کار تا ۷۲/۱٪ به‌ازاء چگالی احتمال تک ـ گوسین و۸۱/۷٪ به‌ازاء چگالی احتمال مخلوط دارای پنج عنصر در شرایط اعمال نکردن مُدل زبان افزایش‌می‌یابد.

[30] Recognition Accuracy
[31] Recognition Rate

در پیاده‌سازی صورت گرفته با استفاده از مدل‌های ترکیبیHMM و شبکهٔ عصبی. بالاترین دقّت‌بازشناسی کلمه بدست آمده در شرایط استفاده از مُدل زبان زوج کلمه در حدود ۷۵٪ بوده است.

بررسی کلّی بر روی نتایج بدست آمدهٔ فوق نشان‌دهندهٔ موارد زیر است:

۱ـ در بازشناسی گفتار گسسته،HMMهای با چگالی مشاهدات پیوسته توانائی قابل قبولی از خود نشان‌داده‌اند و نتایج بدست آمده از استفاده از این نوعHMM ها درحدّ قابل قبول است. بدیهی است درصورت آموزش این نوع HMM ها با داده‌های آموزشی متنوع‌تر و بیشتر، انتظار نتایج بهتری از اینگونه‌سیستم‌ها می‌رود.

۲ـ در بازشناسی گفتار پیوسته، همCDHMM و هم ترکیبHMM و شبکه‌های عصبی نتایج قابل‌قبولی داشته‌اند. CDHMM ها توانسته‌اند با ارائه توانائی مناسب، به دقت‌های نسبتاً بالا در بازشناسی‌گفتار پیوسته دست یابند. درعین حال نشان داده شده که در صورت استفاده از مدل‌سازی‌های پیشرفته،نتایج باز هم بهبودی بیشتری می‌یابند و بنابراین استفاده از مدل‌سازی وابسته به متن یا استفاده ازمدل‌های براساس هجا یا نیم‌هجا توصیه می‌گردد.

درعین حال، ترکیبHMM ها و شبکه‌های عصبی نتایج درخور توجهی داشته‌اند. اگرچه این نتایج‌درحد نتایج بدست آمده ازCDHMM ها نبوده ولی نشان‌دهندهٔ این واقعیتند که اینگونه سیستم‌ها نیز ازتوانائی بالائی برخوردار بوده و در صورت توجه بیشتر ممکن است بتوان نتایج بهتری نیز از آنها بدست‌آورد.

بالاخره اینکه، علیرغم دستیابی به حدّ قابل قبولی از توانائی در بازشناسی گفتار پیوسته بر روی یک‌دادگان محدود با دایرهٔ کلمات متوسط، برای داشتن کیفیت‌های مطلوب‌تر و نیز شرایط آرمانی‌تر لازم‌است در سیستم‌های آتی موارد زیر حتماً مورد توجه و پیاده‌سازی قرار گیرند:

ـ مدل‌سازی وابسته به متن درون کلمه‌ای.

ـ مدل‌سازی وابسته به متن بین‌کلمه‌ای.

ـ دادگان با دایرهٔ لغات بیشتر و بدست آمده از شرایط طبیعی‌تر.

ـ مدل‌سازی آماری زبان به‌نحو مناسب.

ـ کار بیشتر بر روی روش‌های جستجو و دکودینگ.

ـ تطبیق گوینده و تطبیق با میکروفون و شرایط محیطی.

۲ سنتز گفتار فارسی

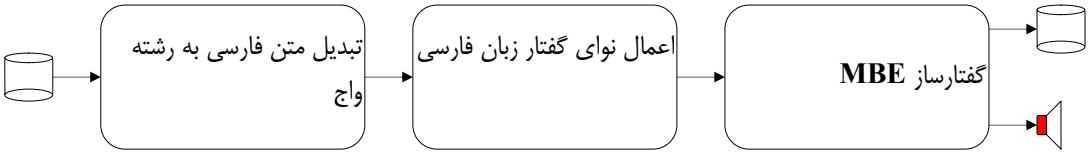
۲-۱ مقدمه

در این بخش از گزارش، ضمن بررسی کلی بخش سنتز گفتار فارسی در این طرح ملی، به مرحله لینک نهایی بین بخشهای گوناگون پروژه سنتز گفتار و فراهم کردن برنامه‌ای در محیط Windows 98 که قادر به خواندن متنهای فارسی باشد، پرداخته ایم.

این پروژه از بخشهای اصلی ذیل تشکیل یافته است:

۱. **تبدیل متن فارسی به رشته واج**
۲. **نوای گفتار**
۳. **گفتارساز**

در شکل ۲-۱ شمای کلی پروژه و چگونگی ارتباط بخشهای آن مشاهده می‌گردد:



شکل ۲- ۱ شمای کلی پروژه گفتارساز فارسی

پیش از انجام این مرحله از پروژه بخشهای فوق به صورت برنامه‌هایی در محیط DOS و یا حالت Console در محیط Windows تهیه شده بود که هر بخش از رابطهای ساده‌ای در ورودی و خروجی استفاده می‌کرد. در این پروژه ابتدا ماژولهای مورد نظر برای کار در محیط ویژوال آماده شده و سپس با یکدیگر لینک شدند. خاطر نشان می‌شود که در زمان تهیه این گزارش به دلیل آماده نبودن قسمت تحلیل نحوی برخی از قسمتهای دیگر که وابسته به اطلاعات خروجی این ماژول بودند فعال نمی‌باشند. به طور نمونه می‌توان به قسمت آهنگ جمله در ماژول نوای گفتار اشاره کرد که نیاز مستقیم به تحلیل نحوی جمله دارد.

۲-۲ نرم افزار نهایی

۲-۲-۱ لایه کاربری نرم افزار

پنجره اصلی نرم افزار که در شکل ۲-۲ دیده می‌شود از کاربری بسیار ساده‌ای برخودار است. همانگونه که مشاهده می‌گردد ورودی گفتارساز به دو صورت تایپ دستی و یا استفاده از یک پرونده متن انجام می‌گیرد. به این معنی که کاربر می‌تواند یک متن فارسی را در پنجره ورود متن بنویسد و یا پس از زدن تکمه Open Text و باز شدن دیالوگ آن یک پرونده متن فارسی را انتخاب نماید که در این صورت متن پرونده در پنجره ورود متن نوشته می‌شود و کاربر می‌توان آن را ویرایش نماید. سپس برای خوانده شدن متن وارد شده باید تکمه Say زده شود.

لازم به ذکر است که این نسخه از نرم افزار در سیستم عامل Windows 98 نوشته شده است. البته در نسخه‌های بالاتر Windows مانند ME ، 2000 و XP کار می‌کند ولی در دونسخه اخیر برخی از حروف متن ورودی به درستی مشاهده نمی‌شود.

شکل ۲-۲ ***پنجره رابط کاربر نرم افزار***

۲-۲-۲ لایه ارتباطی ماژولهای نرم افزار

از آنجا که ماژولهای اصلی این نرم افزار توسط گروههای مختلفی طراحی و پیاده‌سازی شده است، مناسب به نظر می‌رسید که برای برقراری ارتباط بین این ماژولها از زبانی ساده‌تر از زبانهای

برنامه نویسی و ساختارهای پیچیده آن استفاده شود. به همین جهت برای برقرار ارتباط بین ماژولهای کلی از متنهای ساده استفاده شده است. هر ماژول وظیفه دارد که متن ورودی را به ساختارهای مورد نیاز خود تبدیل کرده و برای تحویل خروجی نیز متن مورد نیاز ماژول بعدی را بسازد.

اگرچه در وهله اول اینگونه به نظر می‌رسد که این روش قدری کار برنامه نویسی هر ماژول را افزایش می‌دهد، اما حسن آن این است که سهولت فراوانی در یک کار گروهی ایجاد می‌کند. اولین نتیجه استفاده از این روش آزادی عمل برنامه نویسان برای پیاده‌سازی هر ماژول می‌باشد. در واقع هر برنامه نویس مجبور است که برای آزمودن ماژول خود از چنین روشی استفاده نماید. در ادامه خلاصه‌ای از اطلاعات ورودی و خروجی هر ماژول آورده شده است:

متن ورودی

هوا صاف است.

خروجی ماژول تبدیل متن به واج یا TTP (شامل تحلیل نحوی)

h@va	n(اسم)	m(مسند)
saf	j(صفت)	n(مسند الیه)
?@st	w(بن ماضی ربطی)	r(فعل ربط)

خروجی ماژول نوای گفتار

واجنوشته هجا	فرکانس گام	ضریب دیرش	ضریب دامنه
h@	133	1.0	1.0
va	150	1.0	1.0
saf	145	1.2	1.1
?@st	120	1.2	1.0

شایان ذکر است که اگر چه از این روش برای پیاده‌سازی و توسعه هر یک از ماژولها استفاده شده است اما در نرم افزار نهایی به علت یکسان بودن ساختارهای استفاده شده در بین برخی ماژولها این لایه ارتباطی به لایه ساختارها و توابع کلان نرم افزاری تنزل یافته است که این خللی در اصل کار ایجاد نمی‌کند.

در برخی از ارتباطهای نیز شکل این متن تا اندازه‌ای تغییر کرده است. به عنوان نمونه ورودی ماژول نوای گفتار برای هر جمله در حال حاضر رشته‌ای به شکل ذیل می‌باشد:

هوا h@va n m / **صاف** saf j n / **است** ?@st w r .

که در آن برای هر واژه به ترتیب متن فارسی، رشته واج، نقش صرفی و نقش نحوی آن آورده شده است. جدول علائم اختصاری مربوط به نقشهای صرفی و نحوی در انتهای این گزارش ارائه گردیده است.

۲-۳ ماژول تبدیل متن به واج (TTP)

هدف از پیاده‌سازی ماژول TTP تبدیل متن فارسی به رشته واج معادل آن است. این کار در بسیاری از زبانها به سهولت انجام می‌شود. اما متأسفانه در زبان فارسی به دلیل نوع نگارش مشکلات فراوانی در راه این تبدیل وجود دارد.

۲-۳-۱ مشکلات موجود در TTP فارسی

۲-۳-۱-۱ عدم فاصله[۳۲] گذاری صحیح بین کلمات

در برخی زبانها مانند انگلیسی چون حروف گسسته نوشته می‌شوند، تقید زیادی به فاصله گذاری بین کلمات وجود دارد. به این معنی که بین حروف یک واژه هیچ فاصله‌ای گذاشته نمی‌شود. اما در نگارش فارسی از آنجا که حروف یک واژه به صورت چسبیده به هم نوشته می‌شوند، هنگام تایپ متون فارسی فاصله گذاری لزوماً رعایت نمی‌شود. بنابراین قدم اول در راه تبدیل متن فارسی به رشته واج تقطیع صحیح واژه‌ها از یکدیگر است.

۲-۳-۱-۲ نوشته نشدن حرکات

در نگارش فارسی صداهای کوتاه نوشته نمی‌شوند. تشخیص و جایگزینی این صداها در واژگان فارسی یکی از جدی ترین مشکلات این ماژول می‌باشد.

۲-۳-۱-۳ تشخیص کسره اضافه

تلفظ برخی از واژه‌ها در یک جمله با تلفظ آنها به صورت جداگانه تفاوت دارد. یکی از مهمترین این تفاوتها در زبان فارسی اضافه شدن یک کسره به آخر واژه است که در بین مضاف و مضاف الیه

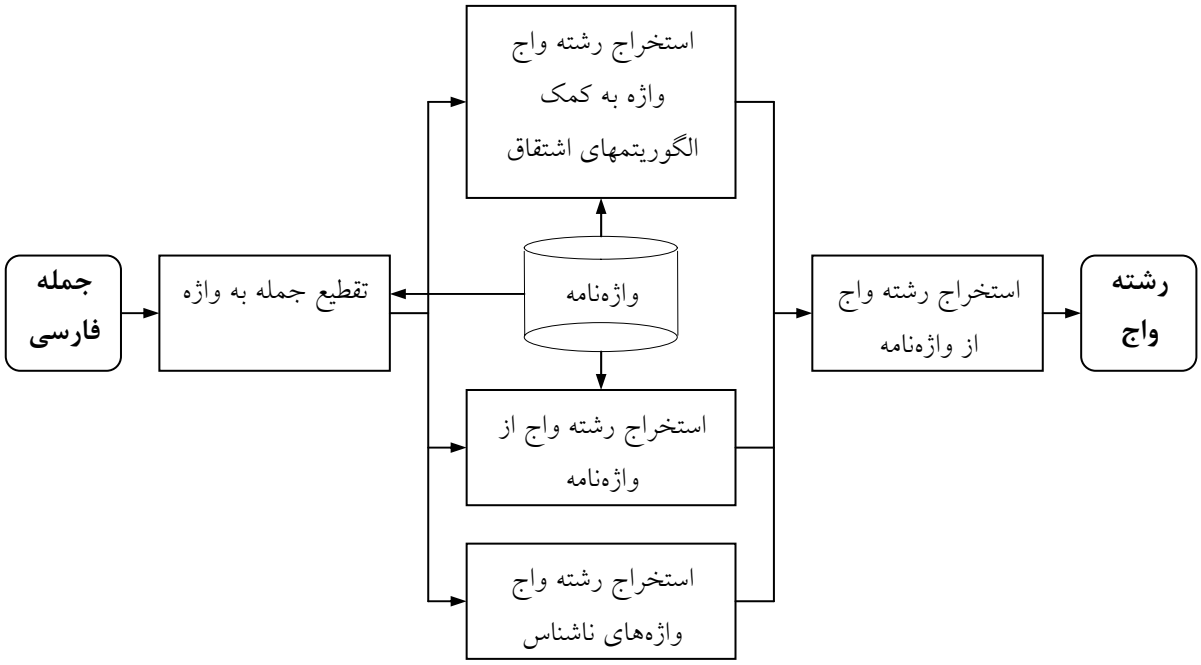
[32] Space

و یا موصوف و صفت ظاهر می‌گردد. به طور طبیعی این حرکت به علت کوتاه بودن در فارسی نوشته نمی‌شود. تلفظ نشدن این کسره در جمله‌های فارسی باعث منقطع و نامفهوم شدن آنها می‌شود.

در این پروژه اساس راه حل این مشکل بر استفاده از واژه‌نامه‌ای قرار داده شد که در آن نگارش واژه‌ها و رشته واج معادل آنها درج شده است. برای پیشگیری از افزایش بی رویه حجم واژه‌نامه مورد نیاز، روندهایی به منظور استخراج رشته واج معادل واژه‌ها از بن آنها طراحی و پیاده‌سازی گردیده است. با تمام پیش بینی های انجام شده باز هم بعضی از واژه‌ها به صورت ناشناس باقی می‌مانند که روندی نیز برای تلفظ این دسته از واژه‌ها ساخته شده است.

۲-۳-۲ بلوکهای ماژول TTP

شمای کلی این ماژول در شکل ۲-۳ مشاهده می‌گردد.



شکل ۲-۳ شمای کلی ماژول تبدیل متن به واج

۲-۳-۲-۱ بلوک تقطیع جمله به واژه‌ها

در این قسمت فرض اصلی بر این است که جمله فاصله کم ندارد و به عبارت دیگر هیچ دو واژه جداگانه‌ای به هم نچسبیده‌اند. اما فاصله اضافه مجاز است و بنابراین اگر بین اجزای یک واژه فاصله اضافی وارد شده باشد، این بلوک قادر به تشخیص آن است. این بلوک جمله دریافتی را به کمک فاصله‌ها و توسط شناسایی واژه‌ها تقطیع می‌کند. به طور معمول یک جمله فارسی به روشهای

گوناگونی تقطیع می‌شود. پس از تقطیع، نوع صرفی واژه به کمک واژه‌نامه قابل استخراج می‌باشد. با توجه به چگونگی قرار گرفتن واژه‌ها در کنار یکدیگر، تقطیعهای گوناگون امتیازدهی می‌شوند و در پایان تقطیعی که بالاترین امتیاز را دارد به عنوان تقطیع نهایی جمله برگزیده می‌شود.

۲-۳-۲-۲ بلوک استخراج رشته واج معادل واژه

پس از آنکه جمله به واژه‌های تشکیل دهنده‌اش تقطیع شد باید رشته واج معادل هر واژه تعیین گردد. این کار به کمک سه قسمت ذیل انجام می‌پذیرد.

۲-۳-۲-۲-۱ استخراج رشته واج متناظر واژه به کمک واژه‌نامه

در این بخش با رجوع به یک واژه‌نامه ۱۰/۰۰۰ کلمه‌ای، واژه مورد نظر جستجو می‌شود. چنانچه این واژه در واژه‌نامه موجود باشد رشته واج معادل و نوع صرفی آن به طور مستقیم استخراج می‌شود. نکته حائز اهمیت در این قسمت، سرعت دسترسی به واژه‌نامه می‌باشد. با توجه به اینکه مراجعه به واژه‌نامه به جز جستجوی مستقیم یک واژه در روتینهای تحلیل واژه‌های مشتق و تقطیع جمله به طور پیاپی مورد استفاده قرار می‌گیرد، سرعت پاسخگویی این قسمت بسیار مهم است. در حال حاضر با استفاده از الگوریتمها و روشهای مناسب این سرعت در حد بسیار خوبی است.

۲-۳-۲-۲-۲ تحلیل واژه‌های مشتق

واژه‌نامه‌ای که برای این پروژه تهیه شده است شامل ۱۰/۰۰۰ واژه می‌باشد که قدری محدودتر از مجموعه لغات رایج زبان فارسی می‌باشد. هر چند می‌توان این واژه‌نامه را گسترش داد، اما این گسترش نیازمند صرف هزینه مالی و زمانی زیادی می‌باشد. به علاوه سرعت دسترسی به واژه‌نامه نیز با افزایش حجم آن تا حدی کاهش می‌یابد. راه حل استفاده شده در این پروژه تحلیل و بررسی واژه‌های مشتق و به طور خاص فعلهای مشتق می‌باشد.

فعلهای مشتق شامل بخش پیشوند، اصلی و پسوند می‌باشد. در قسمت تحلیل فعل کلیه زمانها و صیغه‌های فعلهای فارسی تنها به شرطی که بن ماضی و مضارع آنها داخل واژه‌نامه موجود باشد شناسایی شده و رشته واج معادل آنها استخراج می‌شود. اطلاعات استخراج شده در این قسمت در ماژول نوای گفتار نیز استفاده می‌شود.

پیشوند و پسوندهای رایج در داخل جدولی گنجانده شده‌اند و نرم افزار با شناسایی آنها و به شرط موجود بودن قسمت اصلی در واژه‌نامه، رشته واج معادل واژه را تولید می‌نماید. بخش تحلیل پیشوند و پسوند حتی در شرایطی که چند پیشوند و پسوند به یک بدنه اصلی اضافه شده باشند قادر به تحلیل واژه می‌باشد.

۲-۳-۲-۲-۳ کلمات ناشناس

با آنکه بخش تحلیل واژه‌های مشتق باعث می‌گردد که نقش مؤثر واژه‌نامه چندین برابر شود با این حال باز هم واژه‌هایی وجود دارند که شناسایی نمی‌شوند. این بخش وظیفه تعیین رشته واج معادل این واژه‌ها را بر عهده دارد. هر چند نتیجه این بخش صد در صد صحیح نیست اما باید در نظر داشت که نتیجه این بخش در مقابل عدم قرائت واژه می‌باشد که به طور قطع نتیجه‌ای مطلوب تر است.

۲-۳-۲-۳ بلوک تشخیص کسره اضافه

پس از آنکه رشته واج معادل واژه‌ها به صورت مستقل تعیین شد برای تعیین رشته واج معادل جمله باید کسره‌های اضافه موجود در جمله شناسایی شده و به واژه‌های مربوط اضافه شوند. این کار به کمک نوع صرفی واژه‌ها که در قسمت تحلیل واژه‌ها استخراج شده‌اند، انجام می‌پذیرد. هر چند که اعمال دقیق کسره اضافه نیازمند تحلیل نحوی و حتی تحلیل معنایی می‌باشد، اما روش موجود که تنها مبتنی بر تحلیل صرفی می‌باشد نیز نتیجه خوب و قابل قبولی را ارائه داده است.

۲-۳-۳ محصولات جنبی

در کنار نرم افزارهای نوشته شده برای تبدیل متن به واج، دو محصول جانبی نیز تهیه شده است. این محصولات عبارتند از واژه‌نامه و بانک جملات زبان فارسی.

۲-۳-۳-۱ واژه‌نامه فونتیک

این واژه‌نامه شامل جدول ۱۰/۰۰۰ کلمه‌ای از واژه‌های پرکاربرد که پایه و اساس واژه‌های مشتق نیز قرار می‌گیرند تشکیل شده است. برای هر کلمه علاوه بر نگارش، رشته واج معادل و نوع صرفی آن نیز آورده شده است.

۲-۳-۳-۲ بانک جملات زبان فارسی

آزمایش صحت فرضیات و الگوریتمهای طراحی شده نیازمند یک بانک جمله‌های فارسی می‌باشد. این بانک با حجم ۱۰۰۰ جمله ایجاد گردید. هر جمله در این بانک شامل قسمتهای نگارش، نوع صرفی، رشته واج معادل برای تمام واژه‌های تقطیع شده آن جمله و نیز شامل نوع جمله می‌باشد. هرچند که این بانک برای استفاده در این پروژه طراحی شده است اما منبع بسیار مناسب و مفیدی برای استفاده در سایر تحقیقات زبان فارسی را فراهم می‌آورد

۲-٤ ماژول نوای گفتار

یکی از مهمترین عواملی که موجب طبیعی شدن گفتار مصنوعی می‌شود، نوای گفتار است. نوای گفتار یا عروض عبارتست از مجموعه مشخصات آوایی گفتار در یک زبان که حالت و روح بیان گفتار را مشخص می‌سازد. اینگونه پدیده‌ها، زبر زنجیری[۳۳] نامیده می‌شوند. به احتمال زیاد تا کنون در برنامه‌های تلویزیونی و یا سینمایی به گویش ماشینها و یا روباتها برخورده‌اید. تفاوت اصلی این نوع گویش با گویش رایج در زبان فارسی، عدم وجود نوای گفتار مناسب در آنست.

معمولاً نوای گفتار در دو بخش جداگانه مورد بررسی قرار می‌گیرد:

- تکیه[۳٤] در واژه **چه بطور منفرد و چه در جمله که توجه اصلی در این بخش به بررسی تکیه واژه‌ها بسته به نوع و نقش دستوری آنهاست.**
- آهنگ[۳۵] جمله **که توجه اصلی در این بخش به بررسی آهنگ جملات بسته به نوع آنهاست (خبری، پرسشی، امری و . . .).**

بنابراین براحتی می‌توان دریافت که بررسی دستوری جمله‌های فارسی یکی از مقدمات ضروری در ساختن گفتار فارسی به وسیله رایانه است و انجام این کار حداقل در حد چند پروژه کارشناسی است.

نوای گفتار نقش بسیار مهمی در موارد زیر دارد:

- **قابلیت فهم[۳٦] گفتار**
- **انتقال مفهوم صحیح خصوصاً در گفتار محاوره‌ای**
- **بیان حالت روحی گوینده**
- **آفرینش لهجه‌های گوناگون**

نوای گفتار شامل سه مشخصه اصلی در دستگاه گفتار است :

- **زیروبمی[۳۷]**
- **دیرش[۳۸]**
- **بلندی[۳۹]**

[33] Suprasegmental
[34] Accent or Stress
[35] Intonation
[36] Intelligibility
[37] Pitch
[38] Duration
[39] Loudness

در نوشتارهای فارسی نامهای گوناگونی را برای موارد یاد شده آورده اند ولی ما برای رعایت سادگی همین نامها را بکار می‌بریم. لازم به یادآوری است که بلندی متناسب با انرژی موج صوتی و بنابراین متناسب با مجذور دامنه[40] آن می‌باشد.

اکنون روشن شد که بررسی پدیده‌های زبر زنجیری زبان فارسی تا چه حد در ساختن گفتار فارسی بوسیله رایانه مؤثر است.

۲-٤-۱ میزان تأثیر نوای گفتار در زبان

میزان تأثیر نوای گفتار در زبانهای گوناگون متفاوت است، در برخی زبانها تنها مفهوم جمله را عوض می‌کند ولی در برخی دیگر حتی موجب تغییر واجها نیز می‌شود. روشن است که انسان در بسیاری از موارد بدون تغییر واژه‌ها، تنها با لحن گفتار منظور خود را بیان می‌کند. بسیاری از علامتگذاریهای نوشتاری به ویژه در نوشتن مکالمات (دیالوگ) به‌همین منظور انجام می‌گیرد. بطور کلی می‌توان دو نقش اساسی را برای نوای گفتار در نظر گرفت :

- **تمایز دهندگی[41] : ایجاد تفاوت معنایی (مثل گویا به معنی ناطق و گویا به معنی شاید)**
- **تباین دهندگی[42]: تفکیک واژه‌ها در جمله از یکدیگر**

زبان فارسی از جمله زبانهایی است که نوای گفتار در آن بیشتر نقش تمایز دهندگی دارد و نقش تباین دهندگی آن به جهاتی کمتر است. این نقش خصوصاً در گفتار محاوره أی بسیار مهم است و در بسیاری از موارد نوای گفتار بار ناشی از حذف برخی واژه‌ها در محاوره را بر دوش می‌کشد. در برخی زبانها نقش تمایز دهندگی از حد جمله و واژه نیز فراتر رفته و تغییر واجها را موجب می‌شود، مانند بعضی زبانهای آسیای شرقی. البته در این نوشتار موضوع اصلی در مورد گفتار با لهجه تهرانی و به گونه رسمی آن، مانند آنچه که در بخشهای خبری صدا و سیما بیان می‌شود، بوده است ولی این چیزی از اهمیت موضوع نمی‌کاهد.

۲-٤-۲ تحقیقات انجام شده تا کنون

در کشورهای پیشرفته دهها سال است که در زمینه نوای گفتار تحقیق و بررسی انجام می‌گیرد و نتایج آن در درسهای آکادمیک مطرح می‌گردد. در کشور ما نیز بیشتر تحقیقات در این زمینه به پیش از انقلاب بر می‌گردد و جز چند سال اخیر، توجه چندانی به این موضوع نشده است. در ذیل فهرستی از عناوین محققین و زمینه‌های تحقیقاتی آنها در زبان فارسی آورده شده است:

[40] Amplitude
[41] Oppositional
[42] Contrastive

۱. **دکتر تقی وحیدیان : نوای گفتار (تکیه، مکث، آهنگ)**
۲. **دکتر خانلری : نوای گفتار**
۳. **دکتر ساسان سپنتا : نوای گفتار**
٤. **دکتر یدالله ثمره : آواشناسی**

۲-٤-۳ روشهای مدلسازی نوای گفتار

برای مدلسازی نوای گفتار روشهای گوناگونی مطرح شده است که می‌توان به صورت کلان آنها را به روشهای ذیل تقسیم بندی کرد:

- **روشهای قاعده‌مدار**[٤٣]
- **روشهای داده‌مدار**[٤٤]
- **روشهای تلفیقی که از هر دو روش یادشده استفاده می‌کنند.**

۲-٤-۳-۱ روشهای قاعده‌مدار

این روشها بر اساس تحلیلهای دستوری شامل صرفی و نحوی و نیز تحلیلهای معنایی شکل می‌گیرند. به وسیله اینگونه روشها که معمولاً بر پایه زبانشناسی بنا شده‌اند، می‌توان یک متن واجنویسی شده را علامتگذاری کرد به صورتی که یک خواننده آشنا با علائم بتواند آنرا با آهنگ و تکیه صحیح بخواند. این علامتگذاریها، در واقع نوعی نت نویسی می‌باشند. در روند گفتارسازی از این علامتگذاریها برای تنظیم پارامترهای مورد نظر استفاده می‌شود.

حسن این روشها عبارتست از:

۱. **نزدیکی این روش به مبانی نظری در زبانشناسی**
۲. **قابل فهم بودن قواعد**
۳. **حجم کم اطلاعاتی که برای ذخیره کردن قواعد در رایانه لازم است**

معایب این روشها نیز عبارتند از:

۱. **نیاز به تحلیلهای دستوری و معنایی**
۲. **دشواری استخراج و تنظیم قواعد**
۳. **کامل و شامل نبودن قواعد معمول**

از جمله این روشها می‌توان به سیستم قوانین درختی اشاره کرد [۱۷].

[43] Rule Based
[44] Data Based

۲-٤-۳-۲ روشهای داده‌مدار

این روشها بطور عمده پس از بکارگیری رایانه در علوم گوناگون، رایج شدند و بر اساس مدلها و شبکه‌های قابل آموزش طراحی و پیاده‌سازی می‌شوند. به عنوان مثال می‌توان از شبکه‌های عصبی در این زمینه نام برد. اساس اینگونه روشها بررسی ماشینی اطلاعات موجود در محیط و تنظیم تدریجی پارامترهای مدل بر پایه آن می‌باشد. به عنوان مثال شکل منحنی F0 در یک هجا معمولاً به عنوان یک پارامتر در نظر گرفته می‌شود. این بدان معنی است که حالات محدودی را برای شکل تغییرات F0 در یک هجا فرض کنیم که بتوان آن شکل را بصورت کمّی معرفی کرد. به این تنظیم تدریجی پارامترها "آموزش" می‌گویند. برای آموزش مدلها در اینگونه روشها، می‌بایست از یک دادگان مناسب و با حجم کافی برخوردار بود. در هنگام آموزش سیستم، هیچگونه تحلیل نظری و هوشمندانه توسط انسان صورت نمی‌گیرد. آنچه که رخ می‌دهد نوعی کسب تجربه توسط سیستم می‌باشد.

مزایای این روش عبارتست از:

۱. **برداشت اطلاعات از گفتارهای طبیعی و در نتیجه طبیعیتر بودن تغییرات پارامترها**
۲. **بهتر شدن کیفیت کارکرد با افزایش میزان آموزش**
۳. **عدم احتیاج حتمی به تحلیلهای دستوری و زبانشناسی**

معایب این روش نیز عبارتست از:

۱. **نادیده گرفتن مشخصات زبان از دیدگاه نظری**
۲. **احتیاج به حجم بالای دادگان جهت آموزش**
۳. **پیچیدگی در انتخاب پارامترهای مناسب برای مدل**

از جمله این روشها می‌توان به شبکه عصبی اشاره کرد [۱۷].

۲-٤-۳-۳ روشهای تلفیقی

در این روشها مزایای هر دو نوع روش گذشته، استفاده می‌شود و معایب آنها حذف می‌گردد. در واقع این مدلها بر اساس ترکیبی از اطلاعات ورودی ذیل آموزش می‌بینند:

۱. **پارامترهای فیزیکی گفتار ضبط شده**
۲. **واجنوشته گفتار ضبط شده که از نظر نوایی بطور کامل علامتگذاری شده است**
۳. **تحلیلهای دستوری و معنایی گفتار ضبط شده**

البته در این مدل لازم نیست که تمامی موارد یاد شده در بندهای فوق مورد استفاده قرار گیرند. در مدل داده‌مدار تنها از اطلاعات بند ۱ استفاده می‌شود و پارامترهای فیزیکی مورد نیاز از گفتار ضبط شده استخراج می‌گردد. بدیهی است که اطلاعات فیزیکی نمی‌توانند کمبود اطلاعات

زبانشناسی را جبران کنند. این نوع مدلها از دیدگاه نظری به زبانشناسی و از دیدگاه عملی به گفتار طبیعی نزدیک هستند. از این نوع مدلها می‌توان به مدل معرفی شده در [۱۸] اشاره کرد که در آن از بندهای ۱ و ۲ از موارد یادشده استفاده شده است.

۲-٤-٤ پیاده‌سازی نوای گفتار در گروه سنتز

در گروه سنتز دانشکده برق دانشگاه صنعتی امیرکبیر، فعالیتهایی در زمینه مدلسازی نوای گفتار انجام شده است. این فعالیتها بیشتر در راستای روش مدلسازی قاعده‌مدار می‌باشد [۱۹]. در مدل مورد نظر دو پدیده نوایی "آهنگ در جمله" و "تکیه در واژه" بصورت مستقل مورد بررسی قرار گرفته و اثر آنها با یکدیگر جمع شده است.

۲-٤-٤-۱ پیاده‌سازی تکیه

در این پروژه قواعد تکیه مربوط به موارد ذیل آورده شده است:

۱. **اسم**
۲. **صفت**
۳. **عدد**
٤. **ضمیر**
٥. **انواع فعل**
٦. **قید**
۷. **حرف اضافه**
۸. **حرف ربط**
۹. **اصوات**
۱۰. **پیشوند و پسوند**

در ادامه نمونه‌ای از روش استفاده از جدول قواعد در مورد برخورد با یک فعل ماضی نقلی مشاهده می‌گردد. نقطه ارتباط هر جدول کلی به جزیی با حروف درشت نشان داده شده است.

```
#define WORDS    {VAJEGAN   , 0,   0,NULL}
#define PARTS    {VANDHA    , 0,   0,NULL}

static Accent_Rule VAJEH_Rules[]=
   {
   {ESM          ,-1, 400,ESM_Rules},
   {SEFAT        ,-1, 400,SEFAT_Rules},
   {GHEYD        ,-1, 400,GHEYD_Rules},
   {ZAMIR        ,-1, 400,ZAMIR_Rules},
```

```
   {ADAD         , 1, 400,ADAD_Rules},
   {HARF         , 0, 400,HARF_Rules},
   {FEL          , 1, 400,FEL_Rules},
   NULL
   };

static Accent_Rule FEL_Rules[]=
   {
   {MANFI        , 1, 400,NULL},
   {MAZI         ,-1, 400,MAZI_Rules},
   {MOZARE , 1, 400,MOZARE_Rules},
   {AYANDEH      ,-1, 400,AYANDEH_Rules},
   NULL
   };

static Accent_Rule MAZI_Rules[]=
   {
   {EKHBARI      ,-1, 400,NULL},
   {SADEH        ,-1, 400,NULL},
   {ESTEMRARI    , 1, 400,NULL},
   {ELTEZAMI     ,-1, 400,ELTEZAMI_Rules},
   {NAQLI        ,-2, 400,NAQLI_Rules},
   {BAEED        ,-1, 400,BAEED_Rules},
   {ABAD         ,-1, 400,ABAD_Rules},
   {MOSTAMER     , 1, 400,MOSTAMER_Rules},
   NULL
   };

static Accent_Rule NAQLI_Rules[]=
   {
   {"۳"          ,-1, 400,NULL},
   {"3"          ,-1, 400,NULL},
   WORDS,
   {" است "      , 0, 400,NULL},
   NULL
   };
```

در نمونه یاد شده و در جدول NAQLI_Rules به دو مورد قابل توجه بر می‌خوریم:

۱. رشته "۳" یا "3" معرف سوم شخص مفرد در نظر گرفته شده است. در مورد فعل ماضی نقلی در تمام شخصها تکیه روی هجای ماقبل آخر است جز در مورد سوم شخص که فعل به صورت دو جزء جدا از هم که دومی آن واژه "است" می‌باشد در می‌آید. در این حالت تکیه فعل بر روی هجای آخری قرار می‌گیرد.

۲. اگر قسمت تحلیل نحوی برنامه واژه "است" را به عنوان جزیی از فعل ماضی نقلی معرفی نماید، با مراجعه به قاعده مورد نظر در انتهای همین جدول، این واژه بدون تکیه در نظر گرفته می‌شود. در غیر اینصورت این فعل نیز مانند بقیه افعال تکیه می‌گیرد.

۲-٤-٤-۲ پیاده‌سازی آهنگ

در این پروژه قواعد مربوط به انواع جملات ذیل آورده شده است:

۱. **خبری ساده یا معمولی**
۲. **پرسشی ساده**
۳. **پرسشی با واژه پرسش**
٤. **تعجبی سئوالی**
٥. **تعجبی**
٦. **تأسف آمیز**

برای پیاده‌سازی آهنگ نیز از جداول قواعد استفاده شده است. در ذیل نمونه‌ای از این جداول برای یک جمله معمولی (Normal) مشاهده می‌گردد:

```
static Intonation_Rule Normal[]=
{
     {SINGLE    ,PITCH, 0.00, 0.50, B, E, -200},
     {SENTENCE  ,PITCH, 0.00, 0.50, B, M, 400},
     {SENTENCE  ,PITCH, 0.50, 1.00, N, E, 400},
     {SENTENCE  ,PITCH, 0.00, 0.25, N, E, -600},
     {SENTENCE  ,PITCH, 0.00, 0.25, E, E, -400},
     NULL
};
```

مقادیری که برای این پارامترها در نظر گرفته شده اند به این مفهوم می‌باشد:

- **B و E مشخص کننده شماره واژه‌های ابتدا و انتها هستند.**
- **C و D بترتیب یکی از B بیشتر و یکی از E کمتر هستند.**
- **M شماره واژه مرکزی که حاوی هجای تکیه بر است را تعیین می‌کند.**
- **N و L بترتیب یکی از M بیشتر و یکی از آن کمتر را معین می‌کنند.**

البته آنچه که آورده شد در هنگامی بکار برده می‌شود که بخواهیم قواعد را بر روی جمله اعمال نماییم. ساختار Intonation_Rule چنین محدودیتی را ایجاد نمی‌کند. همانطور که مشاهده می‌شود برخی از قواعد موجود در مورد واژه هاست و نه جمله. اکنون قواعد جدول یادشده:

۱. **اگر جمله ورودی تنها دارای یک واژه است فاز ۰٪ تا ۵۰٪ از منحنی پایه با دامنه ٤۰- هرتز از ابتدا تا انتهای واژه اعمال می‌گردد.**
۲. **فاز ۰٪ تا ۵۰٪ از منحنی پایه با دامنه ٤۰ هرتز از ابتدای واژه اول جمله تا انتهای واژه مرکزی اعمال می‌گردد.**

۳. فاز ۰٪ تا ۲۵٪ از منحنی پایه با دامنه ٦۰- هرتز از ابتدای واژه بعد مرکزی جمله تا انتهای جمله اعمال می‌گردد.

٤. فاز ۰٪ تا ۲۵٪ از منحنی پایه با دامنه ٤۰- هرتز بر روی واژه انتهایی جمله اعمال می‌گردد.

۲-٤-٤-۳ یک نمونه عملی از بکارگیری نوای گفتار

در شکل ۲-٤ نمونه‌ای از خروجی این مدل در مورد جمله سئوالی "آن مرد آمد؟" مشاهده می‌گردد. در این شکل به ترتیب مقدار F0 برای حالات گوناگون از وجود تکیه و آهنگ آورده شده و در نهایت با شکل منحنی F0 در یک گفتار طبیعی مقایسه شده است.

۲-۵ ماژول گفتارساز MBE

پس از تبدیل متن تایپ شده فارسی به رشته فونمها و استخراج پارامترهای نوای گفتاری[٤٥] آن شامل تغییرات مناسب گام[٤٦] ، انرژی و دیرش[٤٧] واحدهای گفتاری نوبت به تبدیل رشته فونمها به سیگنال گفتار پیوسته است که به این مرحله [٤٨] PTS یا سنتز گفته می شود. یک سنتز کننده مطلوب باید دارای مشخصات زیر باشد:

۱. صدای تولید شده توسط آن تا حد ممکن طبیعی باشد.

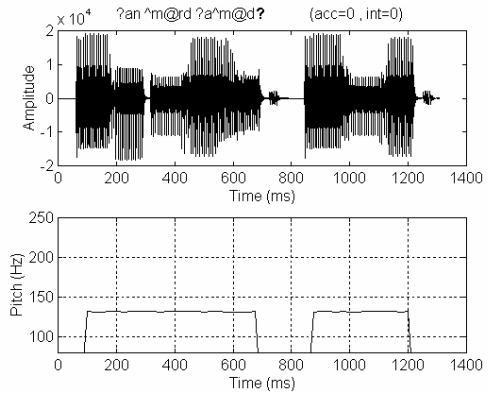
۲. تغییرات آهنگ کلام بوسیلهٔ آن ممکن باشد.

۳. کمترین حجم حافظه را جهت ذخیره دادگان پایهٔ خود اشغال کند.

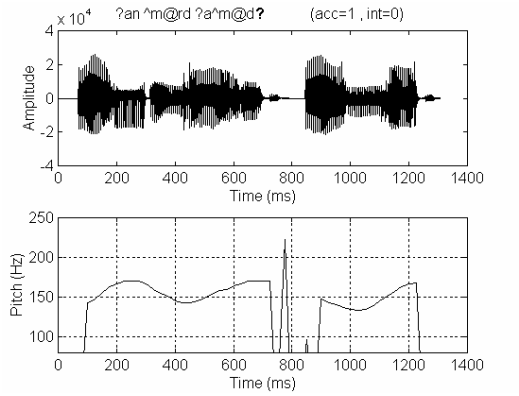
از اوایل دههٔ ۵۰ میلادی روی دو روش سنتز بطور موازی کار شده است: در روش اول که قاعده‌مند[٤٩] بوده و اولین بار توسط Fant ارائه گردید [۲۰]، برای ساختن واحدهای گفتاری از مدلهای کاملا"ریاضی استفاده می‌شود.

در روش دوم که بر اساس بهم چسباندن[٥۰] قطعات گفتار بوده و اولین بار توسط Harris ارائه گردید [۲۱]، واحدهای گفتار از قبل ذخیره شده و هنگام سنتز با تغییرات مناسبی به یکدیگر متصل می‌شوند. در هر دو روش از مدل منبع / فیلتر[٥۱] استفاده می‌شود. در این مدل یک فیلتر خطی که کانال صوتی آدمی را مدل می‌کند بوسیله یک یا چند منبع که دخالت تارهای صوتی را مدل می‌کند تحریک می‌شود، که تفاوت دو روش فوق در بدست آوردن پارامترهای کانال و تحریک می‌باشد.

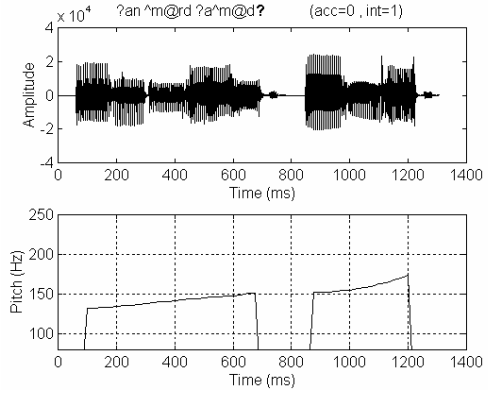
[45] Prosody
[46] Pitch
[47] Duration
[48] Phoneme To Speech
[49] Rule-Based
[50] Concatenation
[51] Source/Filter



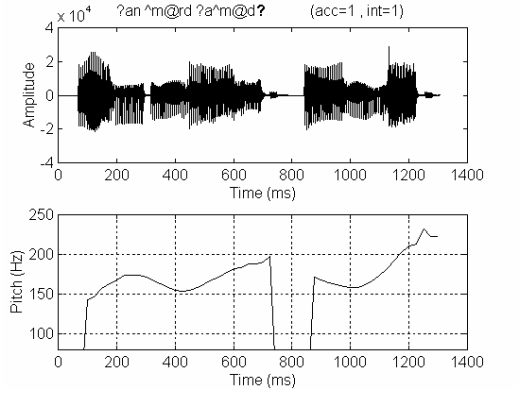
الف) مصنوعی بدون تکیه و آهنگ



ب) مصنوعی با تنها تکیه



ج) مصنوعی با تنها آهنگ



د) مصنوعی با تکیه و آهنگ

شکل ۲- ٤ تفکیک مراحل اعمال نوای گفتار

در روش قاعده‌مند، این پارامترها با قواعدی که در ارتباط با فرمنت هر فونم و میان فونمهاست استخراج می‌شود [۲۲]. در روش اتصال قطعات گفتاری، که واحدهای ذخیره شده گفتار بهم متصل می‌شوند، مدل مذکور برای بهبود سیگنال تولید شده از نظر یکنواختی در گام و انرژی و تغییر آهنگ کلام بکارمی‌رود.

روش قاعده‌مند سنتز سالیان درازی به عنوان تنها روش سنتز با کیفیت بالا مطرح بوده است، اما با ابداع روشهای همزمانی با گام PSOLA (Pitch Synchronous Overlap Add) که بطور جامع اولین بار توسط Moulines و Charpentier معرفی شدند [۲۳]، روش بهم چسباندن واحدهای گفتار نیز بعنوان روش سنتز با کیفیت بالا خصوصا" برای زبانهای اروپایی بکار گرفته می‌شود. بهمین منظور ما در قسمت تبدیل رشته فونم به صوت از این روش استفاده کردیم.

۲-۵-۱ کلیات روش بهم چسباندن واحدهای گفتار

در سیستمهایی که سنتز بر اساس اتصال واحدهای گفتار می‌باشد دو مشکل باید حل شود:

۱. **نوع واحدهای ذخیره شده گفتار چه باشند؟**
۲. **پردازش سیگنالهای ذخیره شده بمنظور تغییردر آهنگ ذاتی آن واحدها چگونه باشد؟**

۲-۵-۱-۱ انتخاب نوع واحدهای ذخیره شده

انتخاب واحدهای ذخیره شده اولیه نقش مهمی در کیفیت سیگنال تولید شده بازی می‌کند، زیرا ناپیوستگی در مرز واحدهای ذخیره شده در هنگام اتصال باید به کمترین مقدار برسد. این واحدهای ذخیره شده بسته به نوع کاربرد مورد نظر می‌تواند به چند دسته کلمه، هجا[۵۲]، واج[۵۳]، واجگونه و دو واج[۵٤] تقسیم شوند که ما در این طرح از واحد دوواج استفاده کردیم.

۲-۵-۱-۲ انتخاب روش پردازش سیگنال

مشکلاتی که در هنگام چسباندن واحدهای گفتار بیکدیگر بوجود می آیند ناشی از عدم تطابق فاز، دامنه و فرکانس پایه آنها می باشد که از عدم تطابق فاز و فرکانس پایه در دادگان ذخیره شده بندرت می‌توان اجتناب کرد و با یکی از روشهای پردازش سیگنال که در زیر بیان می گردد، باید بر این مشکلات غلبه نمود.

۲-۵-۱-۲-۱ روش[۵۵] LPC

این روش بصورت وسیعی برای مدل کردن و کد کردن سیگنالهای صوتی بکار میرود که در مرحله دی کد کردن و سنتز، تغییر آهنگ امکان پذیر خواهد بود زیرا پارامترهای کانال و تحریک بطور مجزا در اختیار بوده و قابل تغییر هستند.

در مدل LPC ساده سازی سیگنال تحریک بصورت تحریک متناوب یا نویزی باعث ماشینی شدن صدا[۵٦] می‌شود که این مشکل بوسیله انتخاب سیگنال تحریک بصورت ترکیبی[۵۷] قابل حل است [۲٤]. همچنین استفاده از تحریکهای پارامتریک نظیرMPLPC [۲۵] وCELP [۲٦] یا استفاده از تحریک غیر پارامتریک RELP [۲۷] کیفیت سیگنال سنتز شده را بهبود خواهد داد.

[52] Syllable
[53] Phoneme
[54] Diphone
[55] Linear Prediction Coding
[56] Metallic or Fuzzy sound
[57] Mixed Sources

۲-۵-۱-۲-۲ روش PSOLA[58]

در این روش ابتدا سیگنالهای ذخیره شده در پنجره هایی ضرب میشوند که اولا" همزمان با گام بوده و ثانیا" بیش از ۵۰٪ همپوشانی دارند. سپس تغییرات گام، دیرش و انرژی این فریمها با استفاده از روشی مناسب اعمال می‌شود که این روشها به سه دسته ذیل تقسیم می‌شوند.

۲-۵-۱-۲-۱ روش TD-PSOLA[59]

در این روش تغییرات گام و دیرش در حوزه زمان انجام میپذیرد و بنابراین به پردازش اضافی احتیاجی نخواهد بود. البته یافتن نشانگرهای گام[60] در این روش الزامی است که برای این کار از روش یافتن زمانهای بسته شدن حنجره[61] توسط تغییر در مشخصات ایستان طیفی زمان کوتاه فریمها استفاده می‌شود [۲۸].

برای تغییردادن گام و دیرش از روشهای حذف-تکرار فریمها و تغییر میزان همپوشانی استفاده می‌شود [۲۳]. البته تغییر گام برای فاکتورهای ۰/۵ تا ۲ امکان پذیر است زیرا هنگام کندکردن قسمتهای بی واک با فاکتورهای برابر با دو و بیشتر، تکرار سیگنالهای‌کوتاه مدت بی واک بطور نامنظم، یک همبستگی[62] کوتاه مدت را در سیگنال سنتز شده ایجاد می‌کندکه بعنوان نویز شبه پریودیک[63] معرفی می‌شود.

۲-۵-۱-۲-۲ روش LP-PSOLA[64]

در این روش پارامترهای LPC در هر فریم استخراج می‌شود که با آنها می‌توان تغییرات آهنگ را گسترده تر اعمال نمود و در ضمن تغییرات پوش طیف را آهسته و پیوسته نمود.

۲-۵-۱-۲-۳ روش FD-PSOLA[65]

در روش TD-PSOLA که از روش حذف-تکرار برای تغییر آهنگ کلام استفاده می‌شود تکرار فریمهای بی واک بطور نامنظم، نویز شبه پریودیک تولید می‌کند که از کیفیت سنتزکننده خواهد کاست. برای رفع اشکال فوق روش فرکانسی همزمان با گام پیشنهاد و بکار گرفته شد [۲۳].

در این روش ابتدا بوسیله STFT[66] سیگنال هر فریم به حوزه فرکانس برده شده و تغییرات مورد نظر آهنگ کلام روی آن اعمال خواهد شد. از بهترین روشهای فرکانسی همزمان با گام روش موسوم

---

58 Pitch Synchronous Overlap Add
59 Time Domain Pitch Synchronous Overlap Add
60 Pitch mark
61 Glottal closure instant
62 Correlation
63 Tonal noise
64 Linear Prediction Pitch Synchronous Overlap Add
65 Frequency Domain Pitch Synchronous Overlap Add
66 Short Time Fourier Transform

به MBR-PSOLA[٦٧] است که توسط Dutoit و Leich در سال ۹۳ میلادی معرفی شد [۲۹] که در آن از روش کدکردن MBE[٦٨] منظور استخراج پارامترهای کانال و تحریک مختلط هر فریم استفاده می‌شود [۳۰]. یکی از مزایای این روش آن است که فرکانس پایه بعنوان پارامتری مستقل بطور اتوماتیک در اختیار بوده و در نتیجه تغییر فرکانس پایه، امری آسان خواهد شد. همزمان با تنظیم فرکانس پایه، فاز نیز اصلاح می‌شود و در آخر با استفاده از درونیابی خطی در حوزه زمان، مشکل عدم تطابق دامنه نیز رفع خواهد شد. در این روش میان قسمتهای واکدار واحدهای گفتاری که به یکدیگر متصل میشوند، پیوستگی طیفی ایجاد نموده و این امر در بالارفتن کیفیت گفتار نقش بسزایی را داراست.

۲-۵-۲ گفتارساز فارسی با روش MBR-PSOLA

در کدکننده MBE از تکنیک آنالیز بوسیله سنتز[٦٩] و تقسیم هر فریم به باندهایی متناسب با گام آن فریم استفاده می‌شود. پس از استخراج پارامترها از روش مدکور تغییرات گام، دیرش و انرژی ممکن خواهد شد. می‌توان انرژی را با ضرب کردن هر فریم در مقداری مناسب و دیرش را با تکرار فریمهای پایدار واکدار تغییر داد. لازم به ذکر است که در آنالیز MBE می‌توان فریمهای واکدار و پایدار را با استفاده از معیارهای خطایی که در آن وجود دارد تشخیص داد.

با توجه به اینکه در آنالیز MBE گام هر فریم محاسبه می‌شود می‌توان آنرا براحتی تغییر داد بدون آنکه پوش طیف فریم دستخوش تغییر گردد. البته در تغییر دادن گام بایستی به پیوستگی فاز بین فریمها و طول باندهای جدید توجه نمود. اما این امر بسادگی امکان پذیر نیست و اشکالی که در تغییر گام بوجود می آید نه در داخل یک فریم بلکه در همپوشانی فریمهای مجاور پیش خواهد آمد. یعنی هنگام سنتز فریمهای واکدار با در نظر گرفتن $L = 4P_{SYN}$ داخل هر فریم ٤ پریود از شکل متناوب واکدار را خواهیم داشت که پیکهای آن در هر قسمت از فریم می‌توانند باشند که در صورت تغییر گام بدون توجه به موقعیت این پیکها در هر فریم، ممکن است فاصله بین آنها هنگام انجام OLA در کناره فریمهای مجاور با فاصله آنها در خود فریم تفاوت پیداکند و در نتیجه تناوب کلی از میان رود. برای از میان بردن مشکل فوق روشی که استفاده شده آن است که موقعیت پیکهای زمانی نسبت به مبدا زمانی هر فریم ثابت گدارده می‌شود. به عبارت دیگر نشانگرهای گام در هر فریم موقعیت ثابتی پیدا میکنند. این کار معادل بکار بردن فاز ثابت برای همه فریمهاست[٧٠] یعنی برای همه فریمها فاز مقداری ثابت قرار داده شود.

[67] Multiband Reharmonization PSOLA
[68] Multiband Excitation
[69] Analysis by Synthesis
[70] Phase reset strategy

اگر سیگنال را بصورت مجموعی از سینوسی ها در نظر بگیریم (تعبیر طیفی سیگنال) فاز هر یک، تاخیر آن را نسبت به مبدا زمانی نشان میدهد. طی تحقیقاتی که انجام گرفته است نشان داده شده است که اگر فاز هارمونیکها را در کل باند ثابت در نظر بگیریم منجر به صدای مصنوعی بصورت Metalic sound خواهد شد. به همین دلیل فاز بصورت تصادفی انتخاب می‌شود که در کار ما این طیف فاز از واکه ای دلخواه مثلاً (a) در فریمی که فرکانس پایه آن مساوی با فرکانس پایه سنتز است انتخاب شده است.

با تعویض فاز یک واکه با مقداری ثابت شکل سیگنال زمانی واکه کلاً عوض می‌شود اما دوره تناوب آن مساوی با دوره تناوب سنتز خواهد شد. با آزمایشات شنیداری ثابت شده است که آنچه در واکه ها مهم است، دوره تناوب آنهاست که گوش با دقت زیادی آنرا دنبال میکند و به آن حساس است اما به شکل موج آنها حساسیت چندانی ندارد فلدا تعویض فاز واکه ها تأثیر چندانی در کیفیت آنها نمی گذارد.

نتیجه مهم ثابت نگهداشتن فاز باندهای فرکانس پائین برای فریمهای سنتز شده با گام ثابت آن است که درونیابی خطی پوش طیف معادل با درونیابی مستقیم زمانی خواهد شدکه با آن مشکل سوم یعنی عدم تطابق دامنه طیفی بین فریمهای مجاور در سنتز کننده متن به گفتار حل خواهد شد. در روش بکاررفته عمق درونیابی خطی درحوزه زمان در قسمت پایدار و واکدار هر دو واجی تا مرز انتقالی قسمت واکدار پیش میرود. درونیابی با عمق متغیر با استفاده از اطلاعات کناری است که بر اساس توان کلی و نسبت توان باندهای واکدار به باندهای بی واک (P_V , P_{UV}) در مرحلهٔ آنالیز MBE استخراج شده اند. این اطلاعات کناری که میزان و نوع درونیابی خطی بین واحدهای مجاور را معین میکنند عبارتند از: فریم پایدار واکدار (VSS)[۷۱] ، فریم پایداربی واک (UVSS)[۷۲]، فریم انتقالی واکدار(VT)[۷۳] و فریم انتقالی بی واک (UVT)[۷۴].

پس از آنکه اطلاعات کناری فوق در مرحله آنالیز بدست آمد، تعداد فریمهایی که از واحدهای مجاور در امر درونیابی خطی بکار میروند مشخص شده و می‌توان درونیابی خطی را در حوزه زمان هنگام سنتز OLA بکار برد. بدینمنظور پنجره های بکاررفته در OLA در ضریبی متناظر با عمق درونیابی ضرب میشوند. اگر S_N^L و S_0^R بترتیب آخرین پنجره های OLA از دوواجهای متصل شده باشند و M_L و M_R بتر تیب تعداد فریمهای بکاررفته در درونیابی از دوواحهای سمت چپ و راست

[71] Voiced Steady State
[72] Unvoiced Steady State
[73] Voiced Transient
[74] Unvoiced Transient

باشند آنگاه می‌توان نشان داد که ضریب پنجره i ام از سمت چپ به اندازهٔ Δ_i و ضریب پنجره j ام از سمت راست به اندازهٔ Δ_j تغییر خواهندکرد.

فرمولهای بکار رفته برای درونیابی خطی عبارتند از :

$$S^L_{N-i} = S^L_{N-i} + (S^R_0 - S^L_N)\frac{1}{2}(\frac{M_L - i - 0\cdot 5}{M_L}) \quad , \quad for \quad i = 0,\mathrm{L}\ , M_L - 1$$

$$S^R_j = S^R_j + (S^L_N - S^R_0)\frac{1}{2}(\frac{M_R - j - 0\cdot 5}{M_R}) \quad , \quad for \quad j = 0,\mathrm{L}\ , M_R - 1$$

۲-۵-۲-۱ روش تغییر گام صحبت در گفتار ساز MBR-PSOLA

برای تغییر گام صحبت در گفتار ساز MBR-PSOLA از این خاصیت استفاده می شود که مشخصات طیفی کانال و تحریک بطور مجزا در دسترس بوده و بدینترتیب تغییرگام بدون تغییر در پوش طیفی کل سیگنال امکانپذیر خواهد بود. این روش مبتنی بر تغییر فاز و فرکانس سیگنال تحریک متناوب است بدینترتیب که سیگنال تحریکی که در آنالیز فریمها بکار گرفته شده است را متناسب با تفاوت گام آنالیز و سنتز شیفت میدهیم بدون انکه در فاز اصلاح شده طیفی کانال تغییری صورت پذیرد.

لازم به ذکر است، سیگنال تحریک را از تبدیل فوریه گرفتن از نمونه های پنجره همینگ بدست می آوریم که این نمونه ها از فواصلی مساوی دوره تناوب گام انتخاب میشوند.

بعنوان مثال میخواهیم دوره تناوب گام را برای دو فریم متوالی، از مقدار ثابت قبلی که۷۰ بوده است (طول پنجره انالیز و سنتز ٤ برابر این مقدار ثابت بوده است) به مقدار جدید ٦٥ و ۷۵ تغییر دهیم (اعداد فوق فاصله دو گام متوالی در حوزه زمان میباشد که برای تبدیل آن به فرکانس باید ۱۱۰۰۰ را بر آنها تقسیم نمود). برای این کار نمونه های انتخاب شده در پنجره همینگ را که در ساختن سیگنال تحریک بکار میروند بگونه ای در فریمهای متوالی شیفت میدهیم که در سیگنال منتجه تناوب کلی حاصل شود. یعنی اگر فاصله اولین نمونه در پنجره های همینگ را نسبت به مبدا با DIF نشان دهیم آنگاه این مقدار در فریمهای متوالی بصورت زیر باید تغییر کند:

```
K=0
do  K+ +      while ( ( DIF + P * K - TW / 2 ) < 0 )
DIF + = P * K - TW / 2
```

که در آن P دوره تناوب گام سنتز است و TW طول زمانی پنجره ها میباشد.

در فرمول فوق شیفت نمونه ها بگونه ای انتخاب شده است که این نمونه ها در همپوشانی فریمها و در ناحیهٔ مشترک آنها روی هم قرار گیرند.

۲-۵-۲-۲ روش تغییر طول و درونیابی بین سیلابها

در برنامه ابتدائی سنتز، تغییر طول هر سیلاب را با تکرار یا حذف کردن فریم یا فریمهای آخر قسمت CV و فریم یا فریمهای اول قسمت VC اعمال کردیم. همانطوریکه می دانیم تکرار یک فریم واکدار بطور متناوب (بدون داشتن کوچکترین تغییری در طیف آن) باعث صدای زنگ دار می‌شود. برای کاهش این مشکل، تکرار فریمها در کل سیلاب بطور یکنواخت توزیع شد که برای این کار برای هر فریم یک عدد تکرار محاسبه شده و بر حسب آن فریم یا حذف شده و یا تکرار می‌شود.

۲-۵-۲-۳ روش تغییر انرژی

در برنامه ابتدائی سنتز، تغییر انرژی هم در سنتز فرکانسی و هم در سنتز زمانی اعمال میشد که این کار باعث تفاوت سیگنال سنتز شده و سیگنال اصلی میشد. تغییر انرژی بدینصورت اعمال می‌شود که برای هر فریم ضریب تغییر انرژی محاسبه شده و هنگام سنتز در سیگنال ضرب می‌شود. برای رفع اشکال مورد نظر ضریب تغییر انرژی فقط در قسمت سنتز زمانی اعمال می‌شود. در ضمن با درونیابی خطی ضرایب انرژی در فریمهای تکرار شده از ثابت ماندن پیوسته انرژی سیگنال اجتناب می‌شود.

۲-۵-۲-٤ اعمال قواعد ثابت آهنگین کردن گفتار وتغییر ساختار سیلاب در کلمات

در کلمات خاصی قواعدی را باید روی سیلابهای آن اعمال نمود تا روانی گفتار طبیعی در سنتز بدست آید. از جمله این قواعد به چند نمونه در زیر اشاره می‌شود:

قاعده همزه تسهیل شده : این قاعده هنگام چسبیدن همزه به حروف دیگر بکار میرود به این صورت که انرژی همزه کاهش میابد مثلا"در ترکیب "برده است" یا در کلمه "دعا" انرژی همزه بایدکم شود. بنابراین در هنگام سنتز قاعده مزبور را بکار میبریم یعنی هنگام چسبیدن سیلاب شامل همزه از طرف همزه به سیلاب دیگر، ضریب انرژی فریمهای همزه را به نصف کاهش میدهیم.

قاعده تشدید : افزایش طول و/یا انرژی قسمت بیواک دوواجی باعث تشدید آن می‌شود. در مورد بیواکهای صدادار نظیر "م" افزایش طول و در مورد ایستانها نظیر "ب" افزایش طول سکوت ابتدائی باعث مشدد شدن آنها می‌شود. بیواکهای غیر مصوت نظیر "س" را با افزایش انرژی آنها می‌توان مشدد نمود.

باکنترل سکوت بین کلمات و سیلابها در کیفیت سیگنال سنتز شده بهبود زیادی حاصل می‌شود که این کار با تنظیم فاصله بین کلمات و سیلابها ممکن است.

طول قسمت واکدار سیلاب CVCC از مجموع طول قسمت واکدار دوواجهای CV و VC بیشتر است. بنابراین در سنتز طول قسمت واکدار را در ۱/٤ ضرب میکنیم.

۲-۵-۲-۵ تغییر و اصلاح الگوریتم سنتز جملات و سیلابها

در ابتدا مشکلی که در سنتز جملات، خود را بوضوح نشان میداد آن بود که طول زمانی سیلابها خصوصا" CVC زیاد بوده و بین سیلابها درونیابی خطی اعمال نمیشد و این باعث میشد که در مرز سیلابهای سنتز شده پرشهایی بوجود اید. اشکال دیگری که وجود داشت آن بود که انرژی دوواجها بایکدیگر خیلی متفاوت بود که این مشکل در سنتز جملات خود را نشان می‌داد.

اشکالات فوق به دو مورد بر می گشت: مورد اول آنکه در ضبط و انتخاب دوواجها دقت کافی بکار نرفته بود و مورد دیگر آنکه آلگوریتم سنتز سیلابها و جملات و درونیابی خطی بین سیلابها کامل نشده بود.

۲-۵-۲-٦ اصلاح و بهینه سازی آلگوریتم آنالیز دوفونیهای ذخیره شده

برای رفع اشکال اول که با آلگوریتم آنالیز مرتبط بود، دادگان جدیدی تهیه شد و در آن ابتدا جملاتی (حدود ۷۰ جمله) که در آنها تمام دوواجهای مورد نیاز گنجانده شده بود توسط یک گوینده و در شرایط احساسی یکسان و با انرژی تقریبا یکسان گفته شد. سپس از طریق تقطیع این جملات که تقریبا" طبیعی بودند دوواجهای مورد نیاز بصورت زیر بدست آمدند:

دوواج CV (که قبلا بطور مستقل بیان شده و از سکوت C تا سکوت V گفته شده و ذخیره میشد و در نتیجه طول آن غیر طبیعی میشد) از قسمت ایستان C تا قسمت ایستان V جدا شد و چون از جمله استخراج شد طول آن تقریبا" با طول گفتار طبیعی برابر بود.

طی بررسیهای انجام شده، طول دوواج CV که در سیلاب CVC بکارمیرود تقریبا" نصف سیلاب CV میباشد بنابراین دوواج بکاررفته در سیلاب CVC را از نصف کردن قسمت ایستان واکدار در دوواج CV تنها استخراج کردیم.

سیلاب CV که پس از آن سکوت قرار دارد را نیز از تکرار چند فریم واکدار از قسمت آخر ایستان دوواج CV ساختیم که این فریمهای تکرار شده را در ضرایبی که باعث افتادن سیگنال میشد ضرب کردیم.که به این عمل اصطلاحا" برونیابی[۷۵] گفته می‌شود.

دوواج VC را از میان سیلابهای CVC جملات گفته شده جدا کردیم بطوریکه طول زمانی قسمت واکدار آن کوتاه بوده، هنگام ساختن سیلابهای CVC طول آنها از حالت طبیعی بیشتر نشود.

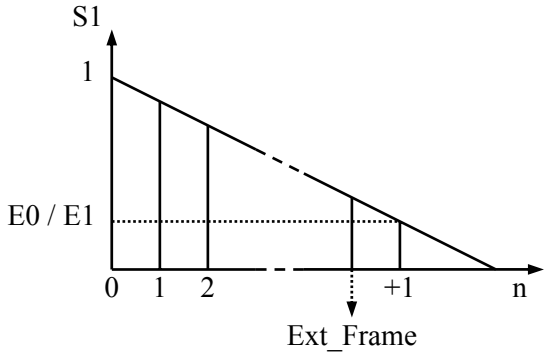
دوواج CC را از انتهای سیلابهای CVCC استخراج کردیم.

[75] Extrapolation

۲-۵-۲-۷ اصلاح و بهینه سازی آلگوریتم سنتز سیلابها و درونیابی خطی بین آنها

برای رفع اشکال دوم که با آلگوریتم سنتز مرتبط بود، در سنتز سیلاب CV متناسب با سیلابهای قبل و بعد آن در فریمهای ذخیره شده تغییراتی را اعمال نمودیم که در زیر حالات مختلف و تغییرات متناسب بیان می‌شود.

۱. اگر قبل از سیلاب **CV** سیلاب **CV** دیگری باشد بین قسمت واکدار سیلاب قبل و قسمت ایستان بیواک سیلاب فعلی برونیابی انجام میپذیرد بدینصورت که فریمهای واکدار**CV** قبلی به تعداد معینی(سه فریم) تکرار می‌شود و انرژی این فریمها بصورتی افت پیدا میکند که به حد انرژی قسمت بیواک دوواج بعدی برسد. لازم به ذکر است که انرژی قسمتهای ایستان**V** و **C** در دوواجهای**CV** توسط آنالیز **MBE** قبلا" در مرحله آنالیز بدست آمده و ذخیره شده اند. بمنظور بدست آوردن فرمول برونیابی مطابق شکل ۲ – ۵ عمل می‌کنیم.



شکل ۲– ۵ روش بدست آوردن فرمول برونیابی.

که در شکل فوق کمیتها عبارتند از :

S1 : ضریب هر فریم

E1 : انرژی آخرین فریم واکدار سیلاب قبل

E0 : انرژی اولین فریم بیواک سیلاب فعلی

n : تعداد فریمها

که در اینصورت خواهیم داشت:

$$Sl[n] = 1 + (E0 / E1 - 1) / (1 + Ext\_Frame) * n \qquad n = 1, 2, \ldots, Ext\_frame$$

Ext_frame : تعداد فریمهای ناحیه درونیابی

- اگر قبل از سیلاب **CV** سیلاب **CVC** یا **CVCC** باشد درونیابی با عمق کم (فقط روی یک فریم) از ابتدای **CV** انجام میپذیرد.

- اگر قبل از سیلاب **CV** سکوت باشد درونیابی با عمق بیشتر (حدود سه فریم) از ابتدای **CV** انجام میپذیرد.
- اگر بعد از سیلاب **CV** سکوت باشد چند فریم از قسمت واکدار انتهایی (حدود سه فریم) تکرار شده و برونیابی انجام می‌شود که افت فریمها بگونه ای است که بعد از تکرار فریمها به انرژی سکوت برسیم.
- حالت آخر آن است که بعد از سیلاب **CV** سیلاب دیگری باشد که همان حالت ۱ باید انجام شود.

۲. در مورد سیلاب **CVC** نیز حالات زیر اتفاق می افتد که عملکرد به یکی از چهار حالت ذیل بر می‌گردد:

- قبل از سیلاب **CVC** سیلاب **CV** باشد که همان حالت ۱ باید انجام شود.
- قبل از سیلاب **CVC** سیلاب **CVC** یا **CVCC** باشد که حالت ۲ باید انجام شود.
- قبل از سیلاب **CVC** سکوت باشد که حالت ۳ باید انجام شود.
- بعد از سیلاب **CVC** سیلاب دیگری باشد که حالت ۲ باید انجام شود.

۳. در مورد سیلاب **CVCC** باید گفت که طول قسمت واکدار بیشتر از **CVC** می‌شود.

۲-۵-۲-۸ تست واحدهای ذخیره شده و پیشنهاد اصلاح و در صورت لزوم افزایش واحدها

برای تست دادگان بدست آمده، با سنتز همه دوواجها بصورت سیلابهای CV و CVC اشکال دوواجهای ضبط شده را تشخیص دادیم. در ضمن با سنتز جملاتی که در آنها همه دوواجها گنجانیده شده بودند، اشکال دوواجهای ضبط شده را در جمله تشخیص داده و با ضبط و آنالیز مجدد آنها اشکالات برطرف گردید.

۲-۵-۲-۹ ادغام برنامه های آنالیز دوواجهای CV و VC و بهبود آنها

در ادامه کار یک برنامه جهت آنالیز همه دوواجهای CV و VC نوشته شد که در آن آرگومان ورودی فایل ضبط شده (که می‌تواند با پسوند raw. یا wav. باشد) همراه خود دوواجی است.

در ضمن روی مقادیر آستانه که در بدست آوردن گام و تخمین پارامترهای طیف بکار میرود کار شد و مقادیر آستانه ها بهینه گردید.

۲-۵-۲-۱۰ برنامه ادغام فایلهای پارامترها

در این برنامه پارامترهایی که در مرحله آنالیز در فایلهای مختلف ذخیره شده بودند تا آنالیز و کارکردن با آنها آسانتر باشد، در دو فایل پارامتر ادغام شدند. این کار اولاً سرعت سنتز را بالا برده ثانیاً تعداد فایلهای باز را هنگام سنتز کاهش داد.

۲-۵-۳ نتیجه گیری

طی این طرح، پس از پیاده سازی آنالیز و سنتز MBE و ذخیره سازی پارامترهای دوواجیها، بمنظور تولید گفتار مصنوعی قطعات گفتار بهم متصل شدند. در اتصال دوواجها به یکدیگر و در محل اتصال آنها با سه مشکل عدم تطابق فاز، گام و دامنه مواجه بودیم که با تغییر گام سنتز به مقدار دلخواه و تعویض فاز فریمهای واکدار با مقداری مشخص، دو مشکل اول از بین رفت و با اعمال درونیابی خطی با عمق متغیر در حوزه زمان مشکل سوم نیز حل شد.

با توجه به تعداد دوواجیهای CV (۱۳۸=۲۳×٦) و متوسط حافظه مورد نیاز آنها (۹ کیلو بایت) و تعداد دوواجیهای VC (۱۳۸=۲۳×٦) و متوسط حافظه مورد نیاز آنها (۵ کیلو بایت) و تعداد دوواجیهای C (۲۳) و متوسط حافظه مورد نیاز آنها (۲کیلو بایت) حجم حافظه لازم برای ذخیره کردن دادگان بدون فشرده سازی حدود۲/٤ مگا بایت میباشد.

آزمایشات شنیداری مبین این نکته است که کیفیت این سنتز کننده بسیار بالا بوده و در مقایسه با سنتزکننده هایی نظیر KLATT که از مدلهای ریاضی برای تولید گفتار استفاده میکنند طبیعی تر گفتارسازی میکند. در ضمن بدلیل استفاده از حوزه فرکانس در تغییرات اعمال شده که دست ما را بسیار باز میکند این سنتزکننده نسبت به گفتارسازهای خانواده PSOLA ارجح میباشد.

۳ تصدیق هویت گوینده

۳-۱ مقدمه و هدف

امروزه شاهد فراگیر شدن فناوری پردازش خودکار گفتار در امور صنعتی، پزشکی، اداری و عمومی جامعه می باشیم. یکی از شاخه های مهم این فناوری تصدیق هویت گوینده می باشد. از سیستم های تصدیق هویت میتوان در امن نمودن دسترسی افراد مجاز به اماکن در مراکزی از قبیل بانکها و شرکتهای بیمه، مراکز کنترل پالایشگاه ها و نیروگاه ها، آزمایشگاه های تحقیقات استراتژیک و حتی بیمارستانها و نیز برای کنترل دسترسی به اطلاعات و خدمات از راه دور و نزدیک استفاده نمود و بدین ترتیب صدا را جانشین امضاء، اثر انگشت، کارت شناسائی، کارت مغناطیسی، رمز عبور و مانند آن نمود. ایده استفاده از صدا برای تصدیق هویت از آنجا ناشی میشود که صوت انسان از یک جهت دربردارنده خصوصیات یک گوینده از جهت ویژگیهای اندام گفتاری او و از طرف دیگر منعکس کننده خصوصیاتی رفتاری و اکتسابی چون میزان تحصیلات، موقعیت فرهنگی، اجتماعی، لهجه و مانند آن می باشد. با توجه به این نکات صدای انسان میتواند همانند یک کارت شناسایی که براحتی قابل جعل و تقلب نیست (جز با تقلید) و از طرفی همواره با آدمی همراه است به عنوان یک ابزار مناسب برای منظور تصدیق هویت بکار رود. مزیت دیگر صدا آن است که حصول و دسترسی بدان از طریق میکروفون و خصوصا" تلفن که در همه جا موجود است براحتی میسر بوده و با استفاده از شبکه های مخابراتی قابل انتقال میباشد. یعنی اینکه امکان تصدیق هویت توسط صدا از راه دور مثلا از طریق تلفن امکان پذیر میباشد. هدف از این بخش از طرح ملی، تحقیق و پژوهش در زمینه تصدیق هویت با استفاده از صدا از طریق تلفن می باشد. ماحصل این تحقیق میتواند برای طراحی سیستم های تصدیق هویت بکار رود. در تصدیق هویت، یک گوینده ابتدا خود را معرفی و سپس به بیان کلمه یا کلماتی میپردازد. وظیفه سیستم تصدیق هویت آن است که با مقایسه گفتار این گوینده با مدلی از صدای این فرد که قبلا در اختیار او گذاشته شده است تعیین کند که آیا این فرد همان کسی که ادعا میکند هست یا خیر.

در اینجا فرض بر آن گرفته ایم که گوینده هویت خود را توسط یک کد شناسائی ۷ رقمی اعلام می نماید. یک الگوریتم بازشناسی ارقام اقدام به بازشناسی ارقام کد شناسائی کاربر می نماید. با اعلام هویت گوینده، مدل مرجع آن گوینده استخراج و با مقایسه همان گفتار گوینده در بیان کد شناسائی وی با این مدل مرجع، عمل تصدیق هویت صورت می گیرد. انتخاب نوع ویژگی و روشهای مناسب

برای مدل کردن اکوستیکی ارقام و گویندگان، بررسی روشهای تصمیم گیری در رد یا قبول گویندگان در هنگام تصدیق هویت و در نهایت ارزیابی و تحلیل عملکرد روشهای پیشنهادی از اهداف این بخش از طرح می باشد. استفاده از بازشناسی ارقام در تصدیق هویت گوینده به این دلیل صورت میگیرد که اصولا اعلام هویت در تصدیق هویت گوینده میتواند به طرق مختلف مانند استفاده از صفحه کلید کامپیوتر یا تلفن، استفاده از کارت مغناطیسی و یا بیان آن بصورت شفاهی انجام شود. جهت انجام سهولت در تصدیق هویت از طریق تلفن فرض نموده ایم که گوینده هویت خود را از طریق بیان یک شماره شناسائی شخصی و بصورت شفاهی صورت میدهد.

بازشناسی گوینده شامل تعیین هویت و تصدیق هویت می باشد. تصدیق هویت گوینده میتواند مستقل از متن یا وابسته به متن باشد. در این طرح تصدیق هویت بصورت مستقل از متن و با استفاده از همان گفتار مربوط به بیان کد شناسائی شخصی گوینده انجام میشود. تصدیق هویت با مقایسه گفتار گوینده با مدل صدای او صورت میگیرد و نهایتا بااستفاده از روشهائی چون مقایسه اختلاف گفتار گوینده با یک سطح آستانه نسبت به رد یا قبول گوینده اقدام میشود. لذا در اینجا خطاهای رد یا قبول اشتباه گویندگان واقعی و دروغگو مطرح میشود که هدف آن است که حتی الامکان در سیستم های تصدیق هویت این خطاها کاهش یابند. در بخش های بعد ابتدا مروری بر کارهای تحقیقاتی انجام شده در زمینه بازشناسی ارقام فارسی و تصدیق هویت گوینده در ایران و جهان می پردازیم. بخش ۳ به خصوصیات گفتار تلفنی، بخش ٤ و ۵ به ترتیب به تشخیص گفتار از سکوت و استخراج ویژگی، بخش ٦ به روشهای مدل نمودن ارقام و گویندگان، بخش ۷ به ارزیابی روشهای بازشناسی ارقام و تصدیق هویت گوینده، بخش ۸ به تشخیص و تصحیح خطا، بخش ۹ به توضیح دادگان گفتاری مورد استفاده، بخش ۱۰ به بازشناسی کد شناسائی شخصی، بخش ۱۱ به تصدیق هویت گوینده و نهایتا بخش ۱۲ به نتیجه گیری اختصاص دارد.

۳-۲ مرور منابع علمی

اشاره شد که در سیستم تصدیق هویت گوینده ای که در این طرح بر روی آن تحقیقاتی صورت گرفته است، گوینده هویت خود را توسط یک کد شناسائی ۷ رقمی اعلام می نماید که می بایست بااستفاده از الگوریتم های بازشناسی ارقام نسبت به شناسائی ارقام این کد شناسائی اقدام گردد. در زمینه بازشناسی ارقام مجزای فارسی مستقل از گوینده از جمله فعالیتهائی که صورت گرفته می توان به موارد زیر اشاره کرد که بجز یک مورد همگی در محیط کنترل شده و با صداهای میکروفنی انجام شده اند. در سال ۱۳۷۱ فرامرز فکری و همکاران [۳۲] با استفاده از مدل پنهان مارکف گسسته یک سیستم بازشناسی ارقام مجزای فارسی را در محیط بدون نویز و میکروفنی پیاده سازی نمودند و

راندمان ۹۶/۰۹٪ را برای داده های آزمایشی بدست آوردند. در سال ۱۳۷۸ حسن بابابیک [۳۳] با استفاده از تلفیق مدل پنهان مارکف و شبکه عصبی راندمان ۹۸/۵٪ را بدست آورد. در سال ۱۳۷۷ شیوا رستم زاده و همکاران [۳٤] با استفاده از مدل پنهان مارکف پیوسته و با ۲۰۰ نمونه برای آموزش و ٤۰ نمونه برای تست در محیط کنترل شده و میکروفنی به راندمان ۹۹/۷۵٪ دست یافت. در سال ۱۳۷۸ سعید بابایی زاده و همکاران [۳۵] با استفاده از ترکیب شبکه های عصبی و مدل پنهان مارکف گسسته به راندمان ۹۸/۸٪ در محیط میکروفنی رسیده اند. در زمینه بازشناسی ارقام متصل فارسی بطور مستقل از گوینده می توان به کار آقای فرامرز فکری[۳۲] اشاره نمود که در محیط کنترل شده و با صداهای میکروفنی انجام شده است. متأسفانه هیچ نتیجه ای از راندمان سیستم پیاده سازی شده گزارش نشده است.

در زمینه تصدیق هویت گوینده نیز فعالیت های تحقیقاتی زیادی در سطح دنیا صورت گرفته است. در دهۀ ۱۹٤۰ بازشناسی گوینده را با استفاده از کمک گرفتن از حس بینائی در کنار حس شنوائی و با طیف نگاشت صوتی و بطور غیراتوماتیک انجام می دادند و بیشتر جنبه های پلیسی و قانونی داشته و به منظور شناخت مجرمین و یا رفع سوء ظن از متهمین مورد استفاده قرار می گرفت. در سال ۱۹٦٦ یک دادگاه قانونی برای اولین بار بازشناسی گوینده را براساس طیف نگاشت اصوات گفتار به رسمیت شناخت. در اوائل دهۀ ۱۹٦۰ بازشناسی خودکار گوینده (بوسیلۀ ماشین) بعنوان یک زمینۀ تحقیقاتی معرفی گردیده است. در روشهای اتوماتیک با دریافت صدای یک فرد از طریق میکروفون و رقمی نمودن آن اقدام به استخراج ویژگیهائی از گفتار می نمائیم که می توانند در متمایز نمودن این فرد از سایر گویندگان مفید واقع شود. به کمک این ویژگیها، در مرحله آموزش مدلی از گوینده ساخته شده و در مراحل بازشناسی با مقایسه صدای گوینده با این مدل و تعیین میزان شباهت بین آندو، عمل بازشناسی صورت می گیرد. ویژگیهای متعددی چون چگونگی تغییرات فرکانس ارتعاش تارهای صوتی که اصطلاحا" گام نامیده میشود، فرکانسهای تشدید مجرای گفتار یا باصطلاح فرمانتها، ضرائب طیف فوریه سیگنال گفتار، ضرائب پیشگوئی خطی و مشتقات آن چون ضرائب انعکاسی، ضرائب نسبت سطوح مقطع و ضرائب زوج خطوط طیفی، ضرائب کپستروم حاصل از آنالیز فوریه در معیار مل، ضرائب کپستروم حاصل از آنالیز پیشگوئی خطی و بسیاری ویژگیهای دیگر برای مشخص نمودن و بیان خصوصیات وابسته به گوینده موجود در گفتار مورد استفاده قرار گرفته و کارائی آنها ارزیابی شده است. این ویژگیها عموما سعی در مدل نمودن خصوصیات مجرای گفتار و یا خصوصیات سیگنال تحریک ناشی از ارتعاش تارهای صوتی گوینده می نمایند و بعضا به تنهائی یا ترکیبی از آنها برای بازشناسی گوینده بکار رفته اند. کارائی این ویژگیها می تواند در محیط های نویزی و در مواردی که صدا از طریق خطوط تلفنی و مخابراتی

منتقل میشود تحت الشعاع قرار گیرد[۳۶]. روشهائی چون لیفتر نمودن میانگذر، روش تفریق میانگین ضرایب کپسترال، روش تفاضل طیفی، استفاده از ویژگیهای کپسترال در حوزه مل، و نیز ویژگیهائی دیگری چون RASTA و PLP پیشنهاد شده اند که تا اندازه ای با تاثیرات نویز و محدودیتهای حاصل از خطوط مخابراتی چون محدودیت پهنای باند، اضافه شدن نویز و اکو و تاثیرات کانال انتقال ومیکروفون و مانند آن مقابله می نمایند. به منظور مدل نمودن گویندگان نیز روشهای متعددی پیشنهاد گردیده است که از آن جمله میتوان به روشهای در هم پیچیدن زمانی ، چندی سازی برداری، مدل مخفی مارکوف، روشهای آماری مرتبه دو، روشهای شبکه عصبی [۳۷] و سیستم های فازی[۳۸] اشاره نمود. خلاصه ای از فعالیتهای تحقیقاتی صورت گرفته در زمینه تصدیق هویت گوینده در جداول زیر آورده شده است. ستون دومِ جدول دادگان مورد استفاده، ستون سوم نوع بردارهای ویژگی، ستون چهارم روش بکار گرفته شده، ستون های بعد میزان داده های آموزش و تست و ستون آخر راندمان روش بکار رفته را نشان می‌دهد.

جدول۳-۱ بعضی از کارهای انجام شده در زمینه تصدیق هویت گوینده بصورت وابسته به متن

Author	DataBase	Feature	Technique	Test Set	Performance
[Doddington, 74]	50 Speakers	F0	Distance	mn	94.5%
[Li, 66]	20 Speakers	Spectrum	Distance	mn	90%
[Soong, 85]	100 Speakers	LPC	VQ	mn	98%
[Naik, 89]	20 Speakers	Filter Bank+ Energy	HMM	Sentence	97.7%
[Rosengerg, 91]	20 Speakers	LPCC + Δ LPCC	HMM	Digits 7	99%
[Naik, 86]	40 Speakers	PSC	DTW	Digits	< 99%
[Furui, 81]	50 Speakers	LPCC + Δ LPCC	DTW	2 s	< 98%
[Bernasconi, 90]	22 Speakers	LPCC + Δ LPCC	DTW	2-3 s	> 99.9%
[Homayounpour, 94]	55 Speakers	LPCC + Δ LPCC	DTW	Digits	98.1%

جدول۳-۲ بعضی از کارهای انجام شده در زمینه تصدیق هویت گوینده بصورت مستقل از متن

Author	DataBase	Feature	Technique	Train Set	Test Set	Performance
[Farell, 94]	TIMIT 38 Speakers	LPCC	VQ	7-13 s	0.2- 3.2 s	97.8%
[Reynolds, 94]	SWITCHBOARD 24 Speakers	MFCC	GMM	180 s	10 s	93%
[Carey, 91]	50 Speakers	11-Filter	HMM+N	50	Digits	95.5%

[Homayounpour, 95]	57 Speakers, Telephony	LPCC + Δ LPCC	LVQ3	32s	15	92%
[Homayounpour, 95]	57 Speakers, Telephony	LPCC + Δ LPCC	Order 2nd Statistics	32s	15	96%

در کشور ایران نیز در زمینه شناسایی گوینده کارهایی انجام گرفته است که به چند نمونه از آنها اشاره خواهیم کرد. در سال ۱۳۷۳ آقایان ذهابی و سپهری[۵۰] با استفاده از مدل پنهان مارکف، برای جمعیت ۵ نفری از گویندگان به ازای ٤۰ رقم برای آموزش و ۲۰ رقم برای بازشناسی سیستمی را پیاده سازی کردند. در سال ۱۳۷۳ آقایان مندولکانی و لطفی زاد [۵۱] با استفاده از تکنیک DTW برای یک جمعیت ۱۰ نفری و به ازای ۱۰ جمله برای آموزش و ۱۰ جمله برای آزمایش، به راندمان ۹۸٪ برای تعیین هویت گوینده دست یافته‌اند. در همین سال آقایان اصغری و عارف [۵۲]، با استفاده از کوانتیزاسیون برداری و بر روی یک جمیعت ۳۰ نفری از مردان به ازای ۳۰ عبارت کوتاه برای آموزش و ۳۰ عبارت کوتاه برای آزمایش، به راندمانی برابر ۹۹٪ برای تعیین هویت گوینده رسیده‌اند. باز هم در سال ۱۳۷۳ آقایان حدائق و لطفی‌زاد [۵۳]، با استفاده از تکنیک DTW و بر روی یک جمعیت ۱۰ نفری و به ازای ۱۰ تکرار از یک جملهٔ خاص برای آموزش و ۱۰ تکرار از همان جمله برای آزمایش، به راندمانی برابر ۱۰۰٪ برای تصدیق هویت گوینده دست یافته‌اند. در سال ۱۳۷٤، آقایان صیادیان و غفوری فرد [٥٤]، با استفاده از کوانتزاسیون برداری و بر روی یک جمعیت ۵۰ نفری از گویندگان، به ازای ۱۰ جمله برای آموزش و یک جمله برای آزمایش به کارآیی متوسط ۹۸/۰۳٪ برای تعیین هویت گوینده رسیده‌اند. باز هم در سال ۱۳۷٤، آقایان مقصودلو، نخعی و تبیانی [۵۵]، با استفاده از کوانتیراسیون برداری و بر روی جمعیت ۱۰ نفری از مردان، به ازای ۸ کُد پنج رقمی در دوره آموزش و کدهای سه رقمی در دوره آزمایش برای تصدیق هویت گوینده به راندمان ۹۹/۸۳٪ رسیده‌اند. در سال ۱۳۷۷، آقایان فیض‌آبادی و صدوقی [٥٦]، با استفاده از کوانتیزاسیون برداری و بر روی یک جمعیت ۳۰ نفری از گویندگان و به ازای ۲۰ جمله و ۲۰ رقم برای آموزش و یک جمله برای آزمایش، به راندمانی برابر ۱۰۰٪ برای تصدیق هویت گوینده رسیده‌اند. در سال ۱۳۷۹، آقایان علوی، غلامپور و نایبی، با استفاده از مدل پنهان مارکف و بر روی یک جمعیت ۲۰ نفری از گویندگان، به ازای ۱۰ عبارت در دورهٔ آموزش و ۱۰ عبارت در دوره تست به راندمان ۱۰۰٪ برای تصدیق هویت گوینده رسیده‌اند. در سال ۱۳۷۹ نیز آقایان صیادیان، بدیع، حکاک و بیک‌زاده [۵۷]، با استفاده از مدل مخلوط گاوسی(GMM) در سطح واج و یک مدل به ازای هر واج برای هر گوینده، بر روی یک جمعیت ٦۰ نفری (٤۰ مرد و ۲۰ زن) و به ازای ۱۰۰۰ جمله در دوره آموزش که بصورت دستی واج نگاری می‌شود و به ازای ۳ ثانیه گویش در دوره آزمایش به راندمان

۱۰۰٪ برای تعیین هویت گوینده رسیده‌اند. تمام کارهای انجام شده که در فوق ذکر شد همه در محیط میکروفنی انجام گرفته، تنها کاری که در محیط تلفنی در زبان فارسی برای شناسایی گوینده انجام شده، کار تحقیقاتی صورت گرفته در این طرح تحقیقات ملی است. همانطور که مشاهده می‌شود، در محیط تلفنی، در ایران کار زیادی انجام نگرفته است. زبان فارسی در زمینهٔ شناسایی گوینده در محیط های غیر کنترل شده و واقعی کار بیشتری را می‌طلبد.

۳-۳ خصوصیات خط تلفن و مکالمات تلفنی

پردازش سیگنالهای صوتی و گفتاری عبور داده شده از خط تلفن و نیز مکالمات تلفنی، بسیار متفاوت از پردازش در محیط های میکروفنی و بدون نویز است. پهنای باند خطوط تلفن محدود می باشد و بعنوان مثال محدودهٔ ۲۰۰ Hz تا ۳٤۰۰ Hz و حتی محدودتر از این هم درنظر گرفته می شود که بسیاری از اطلاعات مفید سیگنال گفتار را از بین می برد. این پدیده در سیستم های بازشناسی گوینده که اطلاعات گوینده در فرکانسهای بالا از اهمیت خاصی برای تمایز گویندگان برخوردار است، بیشتر اثر خود را نشان می دهد. بر روی خط تلفن پژواک وجود دارد. مشخصهٔ کانال تلفنی در باند عبور، یک مشخصهٔ هموار نیست و در فرکانسهای مختلف میزان تضعیف یا تقویت آن متفاوت است که این امر نیز کار بازشنایی را مشکل تر خواهد کرد. نکتهٔ بسیار مهمی که در مورد مکالمات تلفنی وجود دارد این است که گویندگان مختلف از دهنی های متفاوت در دستگاه تلفن خود استفاده می نمایند و پاسخ فرکانسی دهنی های مختلف ممکن است بسیار متفاوت از همدیگر و بسیار ناهموار باشد. تضعیف یا تقویت در مشخصهٔ فرکانسی یک دهنی، در باند تلفنی ممکن است تا 25 dB تغییر داشته باشد و بعنوان مثال دو دهنی یکی از نوع Electret و دیگری از نوع کربنی با هم مقایسه شده‌اند که مشخصهٔ فرکانسی آنها بسیار متفاوت از همدیگر و بسیار ناهموار است. علاوه بر مسائل فوق، اگر یک گوینده فقط از یک دهنی هم استفاده کند، در زمانهای متفاوت هیچ تضمینی وجود ندارد که مشخصهٔ کانال در تماسهای مختلف یکسان باشد. بر روی خط تلفن نویز نیز وجود دارد که لزوماً نویز جمع شونده نیست. در مکالمات تلفنی وضعیت قرار گرفتن دهان گوینده نسبت به دهنی، در مقایسه با مکالمات میکروفنی کنترل شده، تغییرات بیشتری دارد. در بعضی از دهنی ها، مثل دهنی کربنی، اعوجاجات هارمونیکی ایجاد میگردد و حتی پاسخ فرکانسی دهنی متغیر با زمان می‌باشد. نویزهای دیگری نیز بر روی خط تلفن وجود دارد که از آن جمله می توان به نویز آکوستیکی زمینه، نویز برق شهر، نویز CrossTalk ، نویز حاصل از ارتباطات مایکروویو و ... اشاره کرد. ملاحظه میشود که مسأله شناسایی گفتار و گوینده با مکالمات تلفنی بسیار متفاوت تر و دشوارتر از شناسایی گفتار و گوینده در محیط بدون نویز، میکروفنی و فقط با یک

میکروفن است. به علت وجود پدیده های فوق، باید برای سیستم های بازشناسی بر روی خط تلفن تمهیداتی بیندیشیم که در این مقاله برای کاهش اثرِ نویزهای جمع شونده و نیز برای جبران سازی[۷۶] مشخصهٔ کانالِ تلفنی راه حلی درنظر گرفته شده و در آزمایشات صورت گرفته برای بازشناسی ارقام فارسی و نیز تصدیق هویت گوینده نتایج آن ارائه گردیده است.

۳-٤ تشخیص گفتار از سکوت

در طرح پژوهشی حائز که اعلام کد شناسائی شخصی گوینده توسط ارقام مجزا یا متصل و نیز تصدیق هویت نیر بر اساس اطلاعات وابسته به گوینده موجود در ارقام صورت می گیرد نیاز به روشهائی برای جداسازی گفتار مربوط به ارقام از سکوت و تعیین محدوده ارقام ضروری می باشد. روشهای مختلفی برای این منظور در این طرح مورد استفاده قرار گرفت که از آن جمله میتوان به الگوریتم SAD و الگوریتم NP اشاره نمود. الگوریتم SAD بگونه ای پویا و بر مبنای انرژی فریم ها سعی در جدا سازی فعالیت های گفتاری از سکوت و طبقه بندی فریم ها به فریم های سکوت و گفتار دارد. این روش از تخمین لحظه ای SNR برای جدا سازی بخش های گفتاری از سکوت استفاده می نماید. توضیحات بیشتر این روش در مرجع [۱] آمده است. الگوریتم دیگر الگوریتم NP است که ادعا میشود قادر است درSNR= 0dB نیز بخوبی عمل نماید [۲]. در این روش از فریم های گفتار دارای همپوشانی تبدیل فوریه گرفته شده و پس از فیلتر نمودن مولفه های فرکانسی مربوط به برق شهر و فرکانسهای خارج از محدوده مفید گفتار، مولفه های طیف بصورت صعودی مرتب شده و سعی میشود به کمک طیف مرتب شده تخمینی از SNR بدست آید. یک ویژگی خوب این روش آن است که در آن SNR بدست آمده از گین و طول فریم مستقل می باشد. با مقایسه SNR بدست آمده با یک سطح آستانه که بتواند قادر به تفکیک سیگنال از سکوت باشد، براحتی میتوان سکوت زمینه را از سیگنال تشخیص داده و بدین ترتیب مرز بین کلمات با فرض آن که فاصله بین ارقام مورد نظر حاوی سکوت باشد به کمک این روش تعیین میگردد. برای توپضیحات بیشتر این روش به مرجع[۲]مراجعه کنید.

۳-۵ استخراج ویژگی

فرآیند استخراج ویژگی های مناسب یک گام اساسی و کلیدی در حل هر نوع مسئله تشخیص الگو می باشد. برای بازشناسی ارقام پارامترها یا بعبارتی ویژگی استخراج شده از سیگنال صحبت باید

[76] Channel Compensation

برای حاوی اطلاعات مفهومی گفتار و اینکه گوینده چه می گوید بوده و این اطلاعات برای گوینده های مختلف دارای تفاوت جزئی باشد و بالعکس برای تصدیق هویت گوینده، خصوصیات یا بعبارتی ویژگی استخراج شده از سیگنال صحبت میبایست برای گوینده مورد نظر دارای تغییرات جزئی بوده و در عین حال فاصله زیادی با سایر گویندگان داشته باشد. از ویژگی هایی که برای این منظور بکار رفته است می توان به تغییرات فرکانس گام، فرکانس فرمانتها، ضرائب سری فوریه، ضرائب خود همبستگی، ضرائب انعکاسی، ضرائب LPCC، ضرائب LFCC، ضرائب MFCC و بعضی از مشتقات آنها مثلΔMFCC، ΔΔMFCC، ΔLPCC و ΔΔLPCC اشاره کرد. در بازشناسی گفتار ضرائب MFCC کارائی خوبی از خود نشان داده اند. همچنین در بازشناسی گوینده (تصدیق و تعیین هویت گوینده) ضرائب MFCC و ضرائب LPCC و مشتقات آنها بیشترین استفاده را داشته اند. معمولا استخراج ویژگی شامل مراحلی است که از ان جمله میتوان به پیش تاکید، اعمال پنجره، حذف نویز توسط روشهای بهسازی گفتار مانند روش تفاضل طیفی، محاسبه ضرائب ویژگی، حذف خصوصیات میکروفون و کانال مخابراتی از ویژگیهای بدست آمده مانند روش تفاضل میانگین ضرائب کپسترال اشاره نمود.

۳-۶ مدل نمودن ارقام و گویندگان

در عمل روشهای متعددی برای مدل نمودن گفتار و گوینده بکار گرفته شده اند که از آن جمله میتوان به روش در هم پیچیدن زمانی، روش چندی سازی برداری، مدل مخفی مارکف، مدل مخلوط گوسی، روشهای آماری مرتبه دو و یا شبکه های عصبی از نوع SOM، LVQ، MLP یا هر شبکه مناسب دیگری اشاره نمود و اگر هم افزایش کارآیی مد نظر باشد میتوان یک مدل هیبرید از روشهای فوق بکار گرفت.

۳-۷ ارزیابی روشهای بازشناسی ارقام و تصدیق هویت گوینده

در بازشناسی ارقام کارائی روش مطرح شده بر اساس رابطه زیر صورت گرفته است:

$$Accuracy=100\%*(N-I-S-D)/N \qquad (۱-۳)$$

در این رابطهN مجموع تعداد کل نمونه های تست از کل ارقام، I تعدا کلمات داخل شده، D تعداد کلمات حذف شده، S تعداد کلمات جایگزین شده غلط می باشد.

در تصدیق هویت گوینده ارزیابی سیستم های تصدیق هویت و نحوه تعیین راندمان و کارائی آنها حائز اهمیت بالائی می باشد. برای ارزیابی هر سیستم لازم است ماهیت خطاها و اشتباهات موجود در آن سیستم بررسی گردد. در فناوری تصدیق هویت گوینده نیز این ضرورت بدیهی بنظر می‌رسد. همچنانکه در تعریف مسئله تصدیق هویت گوینده لازم است که میزان نزدیکی گفتار ورودی به مدل گوینده ادعا شده معین گردد. اصولا" در تصدیق هویت گوینده یک سطح آستانه جهت اخذ تصمیم در رابطه با پذیرش یا رد ادعای گوینده لازم است که می بایست آنرا در مرحله آموزش سیستم تعیین نمائیم.

در فناوری تصدیق هویت گوینده دو نوع خطا ممکن است رخ دهد که عبارتند از:

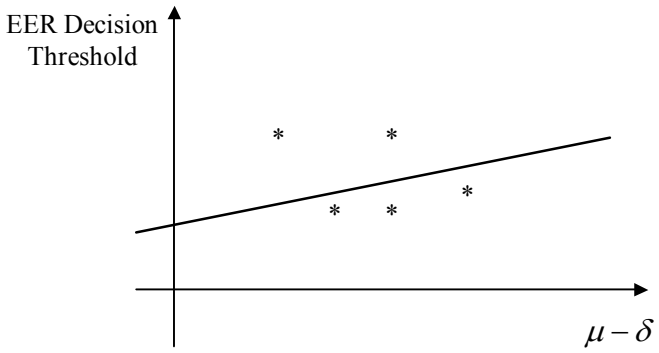
رد اشتباه یا باختصار FR: رد یک گوینده مجاز یا رد نادرست

پذیرش اشتباه یا باختصار FA: قبول یک گوینده غیر مجاز یا تصدیق نادرست

به منظور تصمیم گیری برای رد یا قبول گوینده معمولا فاصله یا شباهت بین گفتار ورودی گوینده و مدل وی با سطح آستانه مقایسه میشود. یک مقدار آستانه پائین باعث خطای رد اشتباه بالاتر و درعین‌حال خطای پذیرش اشتباه پائینر می‌شود. مسلما" بسته به نوع کاربرد مقدار این سطح آستانه متفاوت می‌باشد. روشهای مختلفی جهت پیدا نمودن سطح آستانه‌ای ارائه شده است. یک روش معمول برای انجام اینکار، روش نرخ خطای برابر و یا بطور مختصر EER می‌باشد. در این روش مقدار سطح آستانه محل تلاقی منحنی های درصد پذیرش اشتباه و درصد رد اشتباه بر حسب تغییرات سطح آستانه می باشد. یک اصلاحیه بر این روش این است که برای پیش‌بینی وضعیتهای آینده، بجای استفاده از EER، آن را در یک ضریب که معمولا بزرگتر از۱ است ضرب می‌کنند. اگر این ضریب ۱ باشد، مقدار نهایی همان EER خواهد بود. اگر این مقدار از ۱ بزرگتر باشد، سطح آستانه بطرف راست انتقال پیدا می‌کند که در اینصورت خطای قبول گوینده غیر مجاز زیاد ولی خطای رد گوینده مجاز کم می‌شود. این عمل برای مواردی که شرایط محیط در هنگام بکارگیری سیستم تغییرات زیادی می‌کند، مثلا روی خطوط تلفن بسیار مفید می‌باشد

در روش EER لازم است منحنی های تغییرات درصد پذیرش اشتباه و درصد رد اشتباه بر حسب تغییرات سطح آستانه برای بدست آوردن سطح آستانه تک تک گویندگان رسم شود و اگر به سیستم گوینده جدیدی اضافه گردد این روند برای بدست آوردن سطح آستانه تصمیم گیری وی نیز باید تکرار شود. در ادامه روشی ارائه می‌شود که در آن با توجه به اطلاعاتی که از تعداد محدودی گوینده آموزشی بدست می آوریم بتوانیم مقادیر سطوح آستانه را برای گویندگان جدید بدست آوریم. در این روش بازاء هر گوینده آموزشی، مدل وی با تکرارهائی از گفتار او و نیز تکرار هائی از گفتار سایر گوینده های آموزشی مقایسه و فواصل درون گوینده ای و بین گوینده ای وی بدست می آیند.

به کمک این فواصل EER و نیز تفاضل میانگین و انحراف استاندارد فواصل بین گوینده ای یعنی $\mu-\delta$ را بدست می آوریم. بازاء مقدار EER و $\mu-\delta$ هر گوینده، یک نقطه که توسط * مشخص شده در نمودار شکل زیر بدست می‌آید. این کار را برای کلیه گوینده های آموزشی صورت می دهیم.



شکل ۳- ۱ مقادیر آستانه تصمیم‌گیری EER برحسب میانگین منهای واریانس فواصل برون گوینده ای نظیر آنها.

حال با استفاده از روشهای عددی برازش خط، خط زیر را بگونه‌ای از بین نقاط فوق عبور می دهیم که خطای مربوط به نمایش این نقاط توسط این خط حداقل شود. این روش را روش برازش خط می نامیم:

$$thresh = c1*(\mu-\delta)+c2 \quad (۲-۳)$$

در واقع با استفاده از روشهای عددی ضرایب c1 و c2 بدست می‌آیند. با این روش برای بدست آوردن سطح آستانه تصمیم‌گیری یک گوینده نامشخص کافی است که ابتدا فواصل بین‌گویندگی او محاسبه شده و μ ، δ و در نتیجه $\mu-\delta$ این فواصل بدست آید. آنگاه با استفاده از رابطه فوق، مقدار آستانه تصمیم‌گیری برای آن گوینده بدست می‌آید. نکته قابل ذکر این است که می‌توان اصلاحیه‌ای که بر روش EER اعمال شد، برای این روش نیز بکار برد یعنی اینکه مقدار آستانه حاصل را در ضریبی ضرب نمود تا کارایی سیستم افزایش یابد.

همانگونه که قبلا گفته شد روش کلاسیک برای تصدیق هویت یک گوینده آن است که فاصله بین گفتار وی با مدل گفتار گوینده ادعا شده اندازه‌گیری و سپس با مقایسه این فاصله با یک سطح آستانه تصمیم گیری گوینده مورد نظر پذیرفته و یا رد شود. مزیت استفاده از چنین اندازه‌گیری جهت تنظیم سطح آستانه برای تصمیم‌گیری راجع به رد یا قبول هویت گوینده این است که این امتیازدهی انحراف و اختلاف واقعی بین مدل گوینده مجاز و گفتار ورودی را نشان می‌دهد. اما این امتیازدهی خام برای هم گوینده مجاز و هم گویندگان غیر مجاز بخاطر تغییرات درون‌گویندگی، تغییرات محیط

ضبط و یا تغییرات فونتیکی ممکن است تغییرات بسیار شدیدی نماید و بنابراین بصورت توزیعهایی با هم پوشانی زیاد در آیند. در این شرایط، نرخ خطای یکسان و یا EER زیاد شده و باعث می‌شود که کارایی تصدیق هویت گوینده کاهش یابد.

یک روش دیگر نگرش به مشکل فوق استفاده از سطح آستانه پویا می‌باشد که در آن از شباهت نرمال شده گروهی ۳۰ و یا امتیازدهی پراکندگی استفاده می‌شود. در این روش یک گروه از گویندگان که به گوینده مجاز بسیار نزدیک می‌باشند، بعنوان اعضای گروه وابسته به آن گوینده در نظر گرفته می‌شوند. تشکیل گروه را می‌توان به معنی ایجاد یک محیط محلی در فضای تمامی گویندگان دانست که به کمک آن اندازه‌های فاصله یا شباهت را تا اندازه‌ای نرمالیزه می نمائیم. ساده‌ترین شکل این نرمال سازی محاسبه تفاضل زیر می باشد:

A= فاصله بین گفتار ورودی و مدل گفتار گوینده مورد نظر

B= تابعی از فاصله بین گفتار ورودی و مدلهای گویندگان موجود در محیط محلی گوینده مورد نظر

A-B= فاصله نرمالیزه شده

این روش هم در سیستم مستقل از متن و هم وابسته به متن بکار گرفته شده است. برای باز هم بهتر شدن روند تصمیم گیری میتوان ترکیبی از دو روش سطح آستانه پویا (روش نرمال سازی گروهی) و روش معمول استفاده نمود. بدین صورت که قبل از عمل نرمال سازی گروهی ابتدا از یک سطح آستانه مطلق استفاده می‌شود که سطح آستانه با استفاده از روش EER محاسبه می‌گردد. ابتدا فاصله واقعی بین گفتار ورودی و گفتار مدل مرجع بدست آمده و اگر این مقدار از سطح آستانه کوچکتر بود، آنگاه از نرمالسازی گروهی جهت بدست آوردن فاصله جدید استفاده می‌شود و در غیر اینصورت گوینده گفتار ورودی سریعا رد هویت می‌شود. بعد از نرمالسازی گروهی، فاصله جدید با سطح آستانه دیگر جهت تصدیق و یا رد نهایی مقایسه می‌شود. برای توضیحات بیشتر روش نرمال سازی گروهی و سایر روشهای نرمال سازی امتیازات به به مرجع [۱] و [۳] مراجعه کنید.

برای ارزیابی سیستمهای تصدیق هویت گوینده لازم است که کارایی سیستم بصورتی بیان شود تا بتوان سیستمهای مختلف را تحت یک معیار واحد با هم مقایسه نمود. یک معیار برای این منظور متوسط حسابی یا متوسط هندسی نرخ خطای رد اشتباه بالاتر و نرخ خطای پذیرش اشتباه می باشد. در این طرح از روش میانگین حسابی استفاده شده است.

۳-۸ تشخیص و تصحیح خطای بازشناسی ارقام کد شناسائی شخصی

قبلا بیان نمودیم که اعلام هویت گوینده توسط بیان کد شناسائی هفت رقمی او صورت می گیرد. از آنجا که روشهای بازشناسی ارقام می توانند دارای خطا بوده و ارقام را اشتباهی شناسائی کنند لذا استفاده از روشهای تشخیص و تصحیح خطا بر روی ارقام بازشناسی شده میتواند به کاهش این نوع خطا منجر گردد. برای این منظور بررسیهائی صورت گرفته است که توضیحات بیشتر آن در مرجع [۲] آورده شده است.

۳-۹ دادگان FARSDIGITS1

به منظور ارزیابی روشهای ارائه شده برای بازشناسی ارقام کد شناسائی شخصی و تصدیق هویت گوینده نیاز به یک دادگان تلفنی از ارقام داشتیم که به دلیل در اختیار نبودن این دادگان، اقدام به ضبط آن گردید. این پایگاه دادهٔ متشکل از ارقام مجزا و متصل فارسیِ صفر الی نُه است. این پایگاه داده تلفنی بوده و تقریباً دارای کیفیت SNR=8.8dB می‌باشد. در این پایگاهِ داده، گفتارِ ۱۰۰ گوینده با استفاده از یک اینترفیس تلفنی و یک کارت صدا با فرمت wave و با فرکانس ۱۱۰۲۵ هرتز و ١٦ بیت بازای هر نمونه و از روی خطوط عادیِ تلفنی و بصورت داخلِ شهری و یا برون شهری(راهِ دور) ضبط گردیده است. صدای ضبط شده از طریق خطوط تلفن با صدائی که مستقیما و از طریق یک میکروفون و در محیط عاری از نویز و سرو صدا ضبط میشود کاملا متفاوت است. خصوصیات صدای تلفنی باعث میشود که بازشناسی گفتار و گوینده به مراتب مشکلتر گردد.٦١ گوینده از ۱۰۰ گوینده مرد و ۳۹ نفر زن بوده‌اند که سن مردان از ۱۲ تا ٦١ سال و سن زنان از ١٤ تا ۵۲ سال بوده است. هر گوینده ارقام صفر تا نه را بصورت مجزا در یک الی دو جلسه با فاصلهٔ زمانی ۷ تا ۳۰ روز ضبط کرده است. از هر گوینده ۱۰ الی ١٦ تکرار از هر رقم در دست میباشد. برای لحاظ نمودن تأثیر کلمات ادا شده بر روی یکدیگر، مجددا" از تعداد ۳۰ نفر از گویندگان قبلی خواسته شده است که رشته های عددی ارقام را بصورت متصل بیان نمایند. در مجموع در این پایگاه داده برای هر رقم، تعداد ۱۲۲۲ نمونه ارقام مجزا و ٦٦٠ نمونه ارقام متصل ضبط شده است. در این دادگان برای هر فایل صدا، یک فایل برچسب تولید شده است که علاوه بر مشخصات کلی مربوط به ضبط صدای گوینده، اطلاعات مربوط به ابتدا و انتهای موقعیت ارقام نیز در آن وجود دارد. ابتدا و انتهای ارقام موجود در هر فایل با استفاده از اسپکتروگرام آن بصورت دستی تعیین گردیده و بعبارتی برچسب گذاری شده است .در این فایلها علاوه بر ارقام، محدوده نواحی سکوت بین ارقام نیز لحاظ گردیده است. مشخصات هر گوینده نیز شامل نام گوینده، سن، جنسیت، لهجه، تعداد جلسات ضبط صدا و یک کد هفت رقمی تصادفی در یک فایل دیگر ذخیره شده است. به منظور استفاده از این دادگان در

سیستم تصدیق هویت گوینده، به هر گوینده یک رشته ۷ رقمی تصادفی از ارقام به عنوان کد شناسائی شخصی او تخصیص داده شده است که با استفاده از فایلهای برچسب و فایلهای صدای متناظر آنها و با انتخاب و کنار هم گذاشتن ارقام مربوط به کد ۷ رقمی مورد نظر از فایلهای حاوی ارقام ۰ تا ۹، این کد ساخته می شود. یعنی مثل اینکه گوینده این کد را دقیقا به همان صورت بیان نموده باشد. به این ترتیب این امکان فراهم گردید که برای هر گوینده کد شناسائی او که توسط خود او بیان شده باشد و نیز کد شناسائی او که توسط سایر گوینده ها بیان شده باشد را تولید نمائیم.

۳-۱۰ بازشناسی کد شناسائی

قبل از این گفتیم که هر گوینده به منظور معرفی خود، کد شناسائی شخصی خود را که یک کد ۷ رقمی است بیان می نماید. یکی از وظایف این سیستم آن است که این کد شناسائی شخصی را بازشناسی نموده و ارقام آنرا تعیین نماید. لذا برای این منظور به یک سیستم بازشناسی ارقام فارسی . تا ۹ نیاز خواهیم داشت که در این بخش به بیان الگوریتم های بکار رفته برای این منظور پرداخته و نتایج حاصله را بیان خواهیم نمود. لازم بذکر است که بازشناسی ارقام در اینجا بصورت مستقل از گوینده صورت گرفته و فرض بر آن است که گوینده ارقام را بصورت مجزا و یا بصورت متصل و از طریق تلفن بیان می نماید. باز شناسی ارقام بیان شده شامل مراحل متعددی است که عمده ترین آنها تعیین محدوده ارقام، استخراج ویژگی ها و مدل نمودن آنها می باشد. بازشناسی ارقام مجزا در این پروژه توسط شبکه عصبی پیشگو و مدل مخفی مارکوف و بازشناسی ارقام متصل توسط مدل مخفی مارکوف بروش بازتخمین نهفته صورت گرفته است.

۳-۱۰-۱ بازشناسی ارقام کد شناسائی شخصی بصورت مجزا توسط شبکه عصبی پیشگو

در زیر به بیان مدل شبکه عصبی پیشگو در بازشناسی ارقام مجزای فارسی می پردازیم. شبکه عصبی پیشگو تلفیقی از شبکه عصبی چند لایه پرسپترونی و برنامه ریزی پویا بروش در هم پیچیدن زمانی می باشد. در ادامه این بخش پس از بیان چگونگی پیاده سازی شبکه عصبی پیشگو میزان کارائی آنرا با استفاده از دادگان ارقام تلفنی ضبط شده، مورد ارزیابی قرار می دهیم.

۳-۱۰-۱-۱ مدل شبکه عصبی پیشگو

در سالهای اخیر شبکه عصبی و بخصوص شبکه عصبی چند لایه پرسپترون در زمینه های مختلف پردازش اطلاعات بویژه بازشناسی گفتار مورد استفاده قرار گرفته است. از خصوصیات مهم

اینگونه شبکه ها می‌توان به الگوریتم یادگیری مؤثر بنام انتشار خطا به عقب و همچنین قابلیت آنها در نگاشت ورودی به خروجی در فضای مسئله اشاره نمود. شبکه عصبی چند لایه پرسپترون که در اینجا به اختصار شبکه عصبی نامیده می‌شود در اینجا به عنوان طبقه بندی کننده کلمات موجود در مجموعه کلمات یا لغتنامه که شامل ارقام ۰ الی ۹ می باشد بکار رفته است. روشی که در اینجا بررسی می‌گردد نگرش دیگری به مسئله بازشناسی ارقام با استفاده از شبکه های عصبی چندلایه است. در این روش از مدل پیشگوی عصبی استفاده می‌شود. در این مدل شبکه های عصبی چندلایه بعنوان پیشگویی کننده نمونه ها بکار گرفته می‌شوند. این مدل شامل یک دنباله شبکه عصبی چند لایه پرسپترون جهت پیشگویی غیرخطی هر کلاس می باشد. در واقع این مدل از ساختار زمانی گفتار جهت بازشناسی آن بهره می‌گیرد، چرا که ارتباط زمانی بین بردارهای ویژگی متوالی در گفتار از آنجا که شامل اطلاعات مهمی در بازشناسی گفتار است حائز اهمیت زیادی می باشد. در این روش تغییرات سرعت در بیان گفتار با بکارگیری الگوریتم برنامه‌ریزی پویا نرمالیزه می‌شود. برای توضیحات بیشتر در این زمینه مرجع [۱] مراجعه نمائید.

۳-۱۰-۱-۲ الگوریتم بازشناسی کلمات

در روش پیشگویی عصبی چند لایه پرسپترون، مدل هر لغت شامل دنباله ای از پیشگوهای عصبی می باشد. مسئله بازشناسی یک کلمه را می‌توان بصورت یافتن دنباله ای از پیشگو های عصبی روی صفحه محاسبات برنامه ریزی پویا تعریف نمود، بقسمی که مانده پیشگویی کل حداقل شود[۱]. در نهایت به کمک این روش، اختلاف کلی گفتار ورودی با مدل های لغات محاسبه و مدلی که کمترین اختلاف کلی با گفتار ورودی را داراست، بعنوان لغت بازشناسی شده اعلام میگردد.

۳-۱۰-۱-۳ الگوریتم آموزشی مدل کلمات

منظور از آموزش مدلها یافتن وزنهای پیشگوهای عصبی است بقسمی که مانده پیشگویی کل بازای مجموعه نمونه های آموزشی حداقل شود. تابع هدف بصورت متوسط مانده پیشگویی کل تمامی تکرارهای یک رقم تعریف می‌شود. الگوریتم آموزش برای بهینه سازی تابع هدف، ترکیبی از الگوریتم برنامه ریزی پویا و الگوریتم آموزش شبکه عصبی چند لایه یعنی انتشار خطا به عقب می باشد[۱].

۳-۱۰-۱-٤ استخراج ویژگی و آموزش مدل های ارقام

گفتار ۵۸ نفر از گویندگان (۳٦ نفر مذکر و ۲۲ نفر مونث) از دادگان ارقام فارسی FARSDIGITS1، برای انجام بررسیهای مربوط به بازشناسی ارقام به کمک روش شبکه عصبی

پیشگو استفاده شده است. این مجموعه ۵۸ گوینده ای به دو مجموعه یکی شامل ۵۰ گوینده برای آموزش و دیگری شامل ۸ گوینده برای ارزیابی تقسیم گردید. به منظور ساختن مدل ارقام، گفتار مربوط به بیان هر رقم به ۲۰ فریم مساوی تقسیم و از هر فریم پس از پیش تاکید و ضرب در پنجره همینگ، یک بردار ویژگی شامل ۱۲ پارامتر کپستروم بر مبنای معیار مل MFCC استخراج گردید. در هر بردار، ضرائب کپستروم به ۱/۱ برابر بزرگترین آنها نرمالیزه شدند.

همانگونه که قبلا اشاره شد در این اینجا مدل نمودن ارقام توسط مدل پیشگوی عصبی صورت گرفته است. تعداد پیشگوهای عصبی در مدل ارائه شده برای تمامی ارقام برابر ۱۰ انتخاب گردید. ساختار شبکه عصبی مورد استفاده در مدل پیشگوی فوق دارای سه لایه با ۱۲*۲ گره در لایه ورودی، ٤ گره در لایه مخفی و ۱۲ گره در لایه خروجی تعیین گردید. تعداد گره های لایه ورودی و خروجی به اندازه ای انتخاب شده اند که با قرار دادن بردارهای ویژگی خروجی دو فریم در ورودی شبکه بتوان بردار ویژگی فریم بعدی را پیشگوئی نمود. مدل هر رقم با استفاده از ویژگیهای بدست آمده از ۲ تکرار از هر گوینده که جمعا ۱۰۰ تکرار را شامل میشود آموزش داده شد. داده های آموزشی فوق طی ۱۰۰ تکرار در اختیار شبکه عصبی پیشگو قرار گرفت و بدین ترتیب مدل هر رقم بدست آمد.

۳-۱۰-۱-۵ نتایج آزمایشات

بعد از آموزش مدل کلیه ارقام ۰ تا ۹، اقدام به ارزیابی میزان کارائی روش ارائه شده در بازشناسی ارقام گردبد. راندمان بازشناسی برای نمونه‌های آموزشی ٪۹۵/٦ درصد بدست آمد. از طرفی طی آزمایش صورت گرفته بر روی داده های آزمایشی شامل ۱۰ تکرار از ارقام ۰ تا ۹ که توسط ۸ گوینده بیان شده بودند و سیستم هیچ آشنائی قبلی با آنها نداشت، راندمان ٪۸۳/۷ بدست آمد. طی آزمایشات دیگری اقدام به افزایش تعدا گره های لایه مخفی، تغییر تعداد تکرارها، برش زدن خروجی شبکه و تغییر پارامترهای شبکه عصبی شامل η و ضریب مومنتم α نمودیم. آزمایشات صورت گرفته گویای نکات زیر می باشد :

الف- با افزایش تعداد گره های لایه مخفی اگرچه خطای بازشناسی نمونه های آموزشی افزایش می‌یابد اما در عین حال خطای بازشناسی نمونه‌های آزمایشی کاهش می‌یابد. علت این امر تنظیم بیشتر وزنهای شبکه متناسب با نمونه‌های آموزشی و در عوض کاهش تعمیم پذیری آن می باشد. در مرجع [۱] نتایج آزمایش با تعداد لایه مخفی مختلف را آورده ایم. نتایج این آزمایش نشان داد که افزایش تعداد لایه ها از ٤ لایه به بعد موجب کاهش راندمان میگیرد.

ب- افزایش بیش از حد تعداد تکرار آموزش اگرچه باعث دقیقتر شدن پیشگوییها می‌شود اما بدلیل آموزش بیش از حد شبکه، تعمیم پذیری مدل کاهش می‌یابد. در مرجع [۱] نتایج بازشناسی با توجه به تعداد تکرار آموزش آمده است که گویای نکته فوق می باشد.

ج- جهت افزایش یادگیری سیستم و جلوگیری از حالت اشباع، خروجی گرهها را برش زدیم. میدانیم که وقتی تابع سیگموئید برای تابع فعالیت انتخاب شود داریم:

$$f(x) = 1/(1+\exp(-x)) \; ; \; f'(x) = f(x)\,[1- f(x)] \qquad (۳-۳)$$

بنابراین وقتی f(x) نزدیک 0 و یا 1 شود، مشتق آن بسمت صفر میل می‌کند در نتیجه یادگیری کاهش می‌یابد. برای جلوگیری از این مسئله مقدار تابع فعالیت گره های لایه مخفی به فاصله [۰/۱, ۰/۹] محدود شدند. نتایج این آزمایش نیز با توجه به برش و عدم برش خروجی گرهها در مرجع [۱] آمده است. این نتایج گویای آن است که برش یا عدم برش خروجی گره ها به نتایج تقریبا یکسان منجر می شود. این امر به این خاطر است که در شرایط آموزش با این داده های آموزشی، f(x) به 0 و یا 1 نزدیک نشده است.

د- برای بررسی نقش تعداد ویژگیهای استفاده شده، آزمایشات دیگری انجام شد. در این آزمایشات، تعداد ویژگیهای برابر ۱۰ و ۱۲ انتخاب شدند. نتایج حاصل در در مرجع [۳۲] آورده شده است و بیان می نماید که افزایش ویژگیها در افزایش راندمان بازشناسی مؤثر است.

ه- ارزیابی دیگری نیز به منظور بررسی نقش ویژگیها و مشتق اول آنها که گویای اطلاعات گذرا و دینامیک گفتار می باشند صورت گرفت. برای این منظور دو آزمایش انجام شد که در آنها تعداد ویژگیها مساوی اختیار شدند، با این تفاوت که در آزمایش اول از ۱۶ ضریب کپستروم و در آزمایش دوم ۸ ضریب کپستروم بعلاوه ۸ مشتق اول آنها استفاده گردید. این آزمایشات یک بار بازاء ۲۰۰ تکرار و بار دیگر بازاء ۵۰۰ تکرار صورت گرفتند. نتایج حاصله در در مرجع [۳۲] از این بررسی نشان می‌دهد که استفاده از مشتق اول ضرائب کپستروم یا بعبارتی استفاده از اطلاعات دینامیک موجود در گفتار موجب بهبود بازشناسی ارقام میگردد.

و- مسئله مهم دیگر پارامترهای ثابت یادگیری η و ضریب مومنتم α می‌باشد که در یادگیری موثر می‌باشند. مقادیر مختلفی را برای η و α بصورت سعی و خطا آزمایش کردیم و نهایتا مناسب ترین راندمان بازاء $\eta = 0.01$ و $\alpha = 0.6$ بدست آمد.

۳-۱۰-۲ بازشناسی کد شناسائی شخصی بصورت ارقام مجزا توسط مدل مخفی مارکوف

مدل پنهان مارکف ابزاری بسیار قوی برای پردازش گفتار و بطور کلی برای پردازش سلسله مشاهدات اتفاقی می باشد. در اینجا از مدل پنهان مارکف پیوسته با توابع چگالی مخلوط گاوسی نیز برای بازشناسی ارقام گسسته بر روی خط تلفن و برای جبران اثر کانال انتقال تلفنی از روش تفاضل میانگین در حوزه کپسترال CMS استفاده نمودیم. همچنین جهت حذف نویز جمع شونده موجود در خطوط تلفن از روش مشهور تفاضل طیفی و برای تشخیص مقاوم گفتار از سکوت زمینه از الگوریتمی استفاده گردید که در شرایط نویزی به خوبی کار می کند.

۳-۱۰-۲-۱ استخراج ویژگی

ویژگیهای استفاده شده در این قسمت، ویژگیهای کپسترال حاصل از بانک فیلتر بر اساس معیار مل یعنی MFCC و ویژگیهای کپسترال حاصل از آنالیز پیشگویی خطی یا LPCC می باشند. برای استخراج پارامترهای MFCC و LPCC، سیگنال گفتار به فریم های ۳۵ میلی ثانیه که شروع فریم ها با هم ۱۰ میلی ثانیه فاصله دارند تقسیم می شود. پس از این بر روی سیگنال هر فریم عمل پیش تأکید با $\alpha = 0.975$ انجام می شود و سپس پنجرهٔ همینگ اعمال می گردد. در آنالیز بانک فیلتر، ۱۸ فیلتر مثلثی که بر روی طیف فرکانسی بر اساس معیار مل توزیع شده اند استفاده می شود و سپس ۱۲ ضریب کپسترال استخراج میگردد. برای استخراج پارامترهای LPCC، پس از اعمال پنجرهٔ همینگ، آنالیز پیشگویی خطی بروش اتوکورولیشن با مرتبهٔ P=12 انجام می شود و سپس از ضرایب پیشگویی خطی بدست آمده، ۱۲ ضریب کپسترال استخراج می گردد. بر روی پارامترهای کپسترال LPCC و MFCC، یک لیفتر کاهنده در طرفین (لیفتر جوانگ) اعمال می شود و سپس پارامترهای کپسترال وزن دهی شده بدست می آیند. مشتق اول و دوم ضرایب کپسترال و نیز لگاریتم انرژی و مشتق اول و دوم لگاریتم انرژی نیز به ضرایب کپسترال اضافه می شوند. در نهایت بردارهای ویژگی، ۳۹ بعدی حاصل برای مدل نمودن ارقام بگونه ای که در بخش بعد توضیح داده خواهد شد مورد استفاده قرار گرفت.

۳-۱۰-۲-۲ آموزش مدل های پنهان مارکف

برای نزدیک شدن به ماکزیمم های سراسری، آموزش مدل های مارکف در دو مرحله به شرح زیر انجام گردید:

الف- مرحله اول:

در این مرحله برای ایجاد مدل اولیه هر رقم ابتدا به ماتریس گذر بین حالات مقادیر منطقی نسبت داده می شود. در اولین تکرار ابتدا کلیه بردارهای آموزشی حاصل از هر تکرار از رقم مورد نظر بطور مساوی بین حالات تقسیم شده و سپس مقادیر اولیهٔ ماتریس های کوواریانس و بردارهای میانگین محاسبه می شوند. در اینجا ماتریس های کوواریانس قطری هستند. با استفاده از این مدل اولیه و استفاده از الگوریتم ویتربی، بهترین تخصیص بردارها به حالات تعیین شده و سپس عملیات خوشه بندی توسط الگوریتم K-means انجام می شود و پارامترهای مدل تصحیح می گردند تا مدل جدیدی ایجاد شود. این کار تا رسیدن به همگرایی ادامه می یابد.

ب- مرحله دوم:

در این قسمت مدل حاصل از مرحله اول دریافت شده و بطور تکراری و با استفاده از فرمولهای تخمین بام-ولش، تا رسیدن به همگرایی، پارامترهای مدل تصحیح می شوند.

۳-۱۰-۲-۳ نتایج آزمایشات

در اینجا برای هر یک از ارقام صفر تا نه، مدلهای چپ به راست ٦ حالته که هر یک از حالت ها ١٦، ۳۲ یا ٦٤ مخلوط گاوسی در خود داشتند، در نظر گرفته شد. برای سکوت زمینه هم یک مدل یک حالته با ۳۲ تابع گاوسی در نظر گرفته شد. پایگاه صدای استفاده شده پایگاه تلفنی FARSDIGITS1 می‌باشد. گفتارهای تلفنی ضبط شده، با استفاده از روش تفاضل طیفی و روش تشخیص گفتار از سکوت ذکر شده بهسازی گردیدند. برای آموزش هر رقم، ٥٤٦ نمونه از هر رقم که مربوط به ۵۰ گویندهٔ اول پایگاه داده می باشند، استفاده شده است. در مرحله بازشناسی ٦٧٦ نمونه از هر رقم که مربوط به ۵۰ گوینده دوم پایگاه داده بوده و سیستم در مرحله آموزش گفتار آنها را تجربه نکرده است، استفاده گردید. بنابراین بازشناسی کاملاً مستقل از گوینده می باشد. ویژگیهای استفاده شده، ضرایب کپسترال، لگاریتم انرژی، مشتق اول و دوم ضرایب کپسترال و مشتق اول و دوم لگاریتم انرژی هستند. خلاصه نتایج بدست آمده آن است که ۳۲ مخلوط برای مدل کردن تابع چگالی نتایج بهتری نسبت به زمانی که تعداد مخلوطها ١٦ و ٦٤ هستند ارائه می دهد. درصد شناسایی بالای ٪۹۸/۵۷ بازای نمونه های آزمایشی برای پایگاه دادهٔ تلفنی با SNR=8.8dB گویای کارایی خوب روش بکار رفته در بازشناسی ارقام فارسی می باشد. همچنین در یک آزمایش به جای مدل کردن سکوت از جداسازی گفتار از سکوت استفاده شد و مشاهده گردید که راندمان سیستم حدود ٪۳ کاهش پیدا کرد. با توجه به این موضوع تصمیم گرفته شد که سکوت با استفاده از یک مدل مارکف یک حالته مدل شود. آزمایشات صورت گرفته همچنین نشان داد که کارایی پارامترهای حاصل از بانک فیلتر بهتر از پارامترهای حاصل از آنالیز پیشگویی خطی است. دلیل این برتری تأثیرپذیری ناچیز خروجی فیلترهای بانک فیلتر و نیز تأثیرپذیری شدید پارامترهای پیشگویی خطی از باقیمانده

نویز موجود در گفتار تلفنی یا نویز موزیکال ناشی از روش حذف نویز تفاضل طیفی و نیز توانایی بیشتر ضرایب کپسترال ناشی از خروجی بانک فیلتر با توزیع منطبق بر معیار مل که از خواص شنیداری گوش انسان الهام گرفته است در ارائه اطلاعات موجود در گفتار و در نتیجه تمایز بهتر اطلاعات مربوط به ارقام زبان فارسی، می باشد. در آزمایش دیگری که انجام شد برای جبران مشخصه کانال تلفنی از روش تفاضل میانگین در حوزه کپسترال استفاده گردید که کارایی سیستم را تا ۹۹/۱۰٪ افزایش داد.

۳-۱۰-۳ بازشناسی کد شناسائی شخصی بصورت ارقام متصل توسط مدل مخفی مارکوف

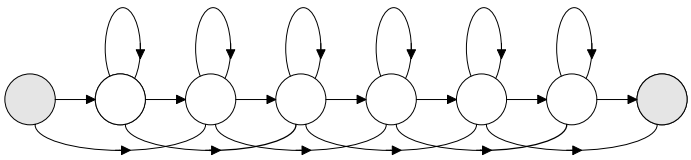
در این قسمت به بیان روش بازتخمین نهفته مدل مخفی مارکوف جهت بهبود راندمان سیستم های بازشناسی ارقام متصل فارسی می پردازیم. این روش موجب تخمین بهتر مقادیر پارامترهای مدل در حالت های مرزی مدل (حالت های ابتدا و انتها) و در نتیجه افزایش صحت بازشناسی می گردد. بازشناسی ارقام متصل نسبت به بازشناسی ارقام گسسته کاری مشکل تر و پیچیده تر می باشد. اثرگذاری ارقام بر روی همدیگر هنگام ادا شدن توسط گوینده ها و حتی حذف شدن واجهای ابتدایی و انتهایی هنگام چسبیدن ارقام به هم کار بازشناسی ارقام متصل را پیچیده تر می کند. در بازشناسی ارقام متصل توسط مدل مخفی مارکوف می توان سه راه در پیش گرفت: الف- برای بازشناسی ارقام متصل از مدل های بدست آمده برای بازشناسی ارقام گسسته استفاده کرد که داده های آموزشی آنها ارقام گسسته بوده است. ب- برای بازشناسی ارقام متصل از مدل هایی استفاده کرد که بر روی رشته ارقام متصل آموزش دیده اند و با استفاده از فرمولهای تخمین بام-ولش این مدل ها آموزش دیده اند. ج- برای بازشناسی ارقام متصل از مدل هایی استفاده کرد که داده های آموزشی آنها ارقام متصل است (مثل ب) و با فرمولهای بام-ولش بدست آمده اند با این تفاوت که پس از تخمین مدل ها با فرمولهای بام-ولش، از فرمولهای بازتخمین نهفته برای تخمین دوبارهٔ پارامترهای تمام مدل ها بطور موازی و افزایش کارایی آنها استفاده گردیده است. در اینجا از روند (ج) برای آموزش مدل ها استفاده گردیده است. طبیعی است که روند (ب) بهتر از روند (الف) می باشد. آزمایشها نشان می دهند که کارایی روش (ج) بهتر از روش (ب) بوده و سیستم را بهبود می بخشد.

۳-۱۰-۳-۱ پیش پردازش و استخراج ویژگی

گفتار به فریم های ۳۵ میلی ثانیه بگونه ای که فاصله بین فریم ها (شروع فریم ها) ۱۰ میلی ثانیه باشد تقسیم می گردد. پیش تأکید با ضریب $\alpha = 0.975$ اعمال شده و سپس پنجره همینگ اعمال می شود. آنالیز بانک فیلتر با معیار مل با تعداد فیلترهای مثلثی ۱۸ و تعداد ضرایب کپسترال ۱۲ انجام می گیرد. ضرایب کپسترال بدست آمده با یک پنجره کاهنده در طرفین (لیفتر جوانگ) وزن دهی می شوند. ضرایب کپسترال و لگاریتم انرژی بهمراه مشتق اول و دوم آنها به عنوان بردار ویژگی نهایی استفاده می گردند که رویهمرفته یک بردار ویژگی ۳۹ بعدی را تشکیل می دهند. برای بهسازی پارامترهای کپسترال بدست آمده، از روش تفاضل میانگین در حوزه کپسترال استفاده شده است. این عمل برای جبران اثر کانال انتقال بر روی پارامترهای کپسترال بکار می رود.

۳-۱۰-۳-۲ آموزش مدل های پنهان مارکوف برای ارقام متصل

آموزش مدل های پنهان مارکوف طی ٤ مرحله صورت می گیرد. الف- تقسیم یکسان بردارها بین حالات و محاسبه پارامترهای اولیه ب- تقسیم بهینه بردارها بین حالات با الگوریتم ویتربی و تصحیح پارامترهای مدل با استفاده از خوشه بندی K-means و تکرار این کار تا رسیدن به همگرایی ج- تصحیح پارامترهای مدل بطور تکراری با استفاده از فرمولهای تخمین بام-ولش تا رسیدن به همگرایی د- تصحیح پارامترهای مدل با استفاده از فرمولهای بازتخمین نهفته و تکرار این کار به دفعات لازم طوری که باعث آموزش بیش از حد نشود.



شکل ۳- ۲ مدل پنهان مارکوف برای ارقام متصل

مدل هر رقم را بصورت یک مدل شش حالته به اضافهٔ دو حالت ورودی و خروجی در نظر گرفته ایم. آموزش مدل ارقام بروش بازتخمین نهفته در مرجع [۵۸] آمده است.

۳-۱۰-۳-۳ آزمایشات و تحلیل نتایج

در این سیستم، برای آموزش مدل های ارقام در حالت متصل، از گفتار۳۰ نفر از گویندگان (مذکر و مؤنث) پایگاه صدای تلفنی FARSDIGITS1 استفاده شده است. در این پایگاه صدا، هر گوینده ۱۰۰

رشته دو رقمی متصل (تمام حالات ممکن) را بیان نموده است. برای آموزش مدل ها از تمام نمونه های آن رقم متعلق به ۱۵ گوینده اول استفاده گردید و مدل ها طبق مراحل الف، ب، ج، د آموزش داده شدند و برای بازشناسی از نمونه های رشته های ارقام ادا شده متعلق به ۱۵ گوینده دوم استفاده گردید. با توجه به تلفنی بودن پایگاه دادهٔ مورد نظر و برای مقابله با نویز از یک الگوریتم مقاوم برای تشخیص گفتار از سکوت و نیز از روش تقاضل طیفی برای بهسازی گفتار استفاده گردید. نتایج حاصل از آزمایشات در جدول زیر آمده است. این نتایج به ازای پارامترهای مبتنی بر بانک فیلتر توزیع شده بر اساس معیار مل، بدست آمده اند.

جدول ۳-۳ صحت بازشناسی ارقام متصل

نوع داده	آموزش بدون بازتخمین نهفته	تعداد دفعات تخمین مجدد در آموزش بروش بازتخمین نهفته					
		۱	۲	۳	٤	٥	٦
آموزشی	۹۰/۱	۹۰/۷	۹۱/۱	۹۱/۹	۹۱/۳	۹۰/۹	۹۰/۳
آزمایشی	۸۳/٥	۸۳/٥	۸۳/۷	۸۳/۷	۸۳/٦	۸۳/٥	۸۳/٥

همانگونه که در جدول ۳-۳ مشاهده می شود در صورت عدم استفاده از روش بازتخمین نهفته صحت بازشناسی ارقام متصل ۸۳/۱۵٪ می باشد. این کارایی در صورت استفاده از روش بازتخمین نهفته افزایش می یابد. بیشترین افزایش بازای ۳ بار تکرار تخمین بدست آمده است و بازاء تکرار بیشتر صحت بازشناسی کاهش می یابد.

۳-۱۱ تصدیق هویت گوینده

قبلا اشاره شد که در یک سیستم تصدیق هویت، پس از اعلام هویت توسط گوینده، میزان شباهت صدای این گوینده و مدل او تعیین و سپس با یک سطح آستانه مقایسه شده و نهایتا نسبت به رد یا قبول او اقدام میگردد. این بخش اختصاص دارد به ارائه روشهائی برای مدل کردن گویندگان که از روش شبکه عصبی، چندی سازی برداری و مدل مخفی مارکوف استفاده می نمایند. همچنین در این بخش به انواع روشهای محاسبه سطح آستانه تصمیم گیری، نرمالیزه کردن گروهی، محاسبه

خطاها و ارزیابی سیستم های تصدیق هویت گوینده خواهیم پرداخت و نتایج حاصل از آزمایشات صورت گرفته را ارائه خواهیم کرد.

۳-۱۱-۱ تصدیق هویت توسط تلفیق شبکه عصبی درخت برآمدگی[۷۷] و الگوریتم ژنتیکی

در زمینه بازشناسی الگو، الگوریتمهای مختلفی جهت خوشه‌بندی داده‌ها وجود دارد که عمدتا" در یک سطح داده‌ها را به خوشه‌های مختلف تقسیم بندی می‌نمایند. در اینجا روشی مورد استفاده قرار می گیرد که در آن خوشه‌بندی و طبقه‌بندی داده ها جهت ایجاد مدل هر گوینده بصورت سلسله مراتبی صورت می‌گیرد. تفاوت عمده این روش سلسله مراتبی با سایر الگوریتمهای خوشه‌بندی، سرعت سریع آن جهت یافتن خوشه دربرگیرنده داده ورودی می‌باشد. آموزش این روش بر اساس الگوریتمهای ژنتیکی و شبکه عصبی درخت برآمدگی بنا شده است. فاز آموزش این روش در واقع یافتن پارامترهای توزیع‌های نرمال این درخت می‌باشد. در این بررسی جهت یافتن پارامترهای بهینه از الگوریتمهای ژنتیکی استفاده شده است. بدلیل سلسله مراتبی بودن روش خوشه بندی بکار رفته سرعت دسترسی به خوشه ها سریع و از مرتبه $O(\log_2 n)$ می باشد.

۳-۱۱-۱-۱ شبکه درختی برآمدگی

یک درخت خوشه بندی از یک ریشه، تعدادی گره میانی و تعدادی برگ در آخرین سطح تشکیل میشود. متناظر هر گره میانی تابعی به شکل زیر اعمال میشود[۵۹]:

$$y = \prod_j \left[\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\left(\frac{(x_j-\mu_j)^2}{2\sigma_j^2} \right)} \right] \qquad (۳-٤)$$

در این تابع y حاصلضرب توابع نرمال با میانگین μ_j و واریانس σ_j متناظر عضو jام بردار ورودی می‌باشد. فضای تصمیم گیری چنین گره‌ای یک ابر بیضی می‌باشد که قسمتی از فضای ورودی گره را محصور می‌نماید. درخت برآمدگی یک ساختار مشابه درخت باینری فوق است که شرط زیر را نیز دارا می‌باشد[۶۰].

"مقدار تابع هر گره میانی باید در تمامی نقاط روی فضای ورودی از توابع متناظر شاخه‌های زیرین خود بزرگتر باشد."

دو گره فرزند گره ریشه باعث می‌شوند که منطقه محصور توسط گره ریشه به دو بخش تقسیم گردد. این عمل تقسیم بصورت بازگشتی برای هر گره تکرار می‌شود تا منطقه داده های ورودی به

[77] Bump-tree

زیر مناطق کوچکتر تقسیم گردند. در واقع هدف درخت برآمدگی این است که فضای ورودی بطور مکرر به مناطق کوچکتر تقسیم بندی شود تا اینکه نهایتا در سطح برگها عمل طبقه بندی داده‌ها بصورت جداپذیر خطی قابل انجام باشد. در سطح برگها، متناظر هر برگ یک شبکه عصبی خطی وجود دارد که وظیفه طبقه بندی نقاطی که در منطقه آن برگ قرار می‌گیرند را بر عهده دارد. وقتی داده ورودی X به شبکه درختی اعمال شود، شاخه های قویتر و در نهایت برگ فعالتر انتخاب می‌شود و در این گره است که عمل طبقه بندی با توجه به خروجی شبکه خطی صورت می‌گیرد. برای آموزش شبکه درختی برآمدگی اولین قدم تعیین ساختار درخت یعنی تعداد لایه‌ها، میانگین و واریانس گره‌های میانی می‌باشد. در مرحله دوم لازم است که شبکه های خطی متناظر هر برگ آموزش ببینند. روشهای مختلفی برای تعیین ساختار درخت وجود دارد که در اینجا از الگوریتمهای ژنتیک جهت یافتن ساختار بهینه استفاده شده است. روش دیگر خوشه‌بندی بازگشتی است که در آن داده‌های آموزشی با استفاده از الگوریتمهای خوشه بندی معمولی بطور بازگشتی به دو قسمت تقسیم می‌گردند و آنگاه بعد از هر تقسیم بندی عمل خوشه‌بندی با توجه به اینکه باید شامل توزیع‌های نرمال باشند، اصلاح می‌گردند. جهت آموزش شبکه‌های خطی از قانون پرسپترون و تنها در یک مرحله برای حداقل سازی خطای داده‌های آموزشی استفاده گردیده است.

گام اساسی در بکارگیری الگوریتمهای ژنتیکی یافتن کدینگ مناسب جهت کد کردن اطلاعات می‌باشد زیرا الگوریتمهای ژنتیکی نه بر پارامترها بلکه بر کد شده آنها اعمال می‌شوند. در واقع می‌توان گفت میزان کارایی الگوریتمهای ژنتیکی بستگی زیادی به نحوه کد کردن دارد. درخت شبکه عصبی بر آمدگی را میتوان بصورت بلاکی به گونه ای ارائه نمود که یافتن ساختار آن توسط الگوریتم ژنتیکی را میسر سازد [۱] هر بلاک معرف یک گره درخت و هر ژن (مقدار) در هر بلاک معرف یک پارامتر آن گره مثلا میانگین و یا واریانس می‌باشد. بلاک شماره صفر بیانگر ریشه درخت می‌باشد. فرزندان یک گره شماره n در بلاکهای شماره 2n+1 و 2n+2 قرار می‌گیرند. مسئله مهمی که در کدینگ شبکه درختی برآمدگی وجود دارد وجود وابستگی بین پارامترهای گره‌های مختلف می‌باشد. برای مثال با توجه به شرط اصلی این درختها باید پارامترهای گره‌های فرزند در فضای تعریف شده توسط گره پدر قرار گیرند. بنابراین پارامترهای گره‌های فرزند بایستی با توجه به گره‌های پدر انتخاب شوند. استفاده از یک کدینگ معمولی که در آن باید هر پارامتر مستقل از پارامترهای دیگر ارائه‌گردد، موجب می‌شود که با اعمال عملگرهای ژنتیکی (جهش و ترکیب) ساختارهایی بوجود آیند که شرط اصلی شبکه‌های درختی برآمدگی را دارا نباشند. این امر حتی با انتخاب نسل اول از شبکه‌های معتبر با گذشت زمان در نسلهای آتی بوقوع خواهد پیوست. برای رفع این مشکل روش کدینگ خاصی پیشنهاد شده است[۱] در کدینگ ارائه شده، هر ژن یک مقدار حقیقی در فاصله (۱+

و ۱-) می‌باشد. نحوه تبدیل یک کروموزم به یک ساختار درخت برآمدگی در مرجع [۱] آمده است. در این روش عملگر ترکیب به این صورت تعریف می‌شود که برای ترکیب دو ساختار یک گره بصورت تصادفی اختیار شده آنگاه زیرشاخه‌های متناظر آن گره در دو ساختار درختی با هم جابجا می‌شوند. در عمل جهش یک ساختار درختی یک گره بصورت تصادفی انتخاب شده و پارامترهای آن با احتمال کم با عدد تصادفی در فاصله (۰/۲+ و ۰/۲-) جمع می‌شود. بعد از اعمال عملگر جهش مقادیر پارامترها بین (۱+ و ۱-) برش می‌شوند تا کدینگ دچار مشکل نشود. نشان داده شده است که این روش کدینگ دارای خواص پیوستگی، هم ریختی، کامل بودن، بسته بودن و در برداشتن حداقل اطلاعات تکراری می‌باشد.

در هر برگ درخت بر آمدگی یک پرسپترون ساده وجود دارد. آموزش شبکه‌های خطی متناظر برگها در مرجع [۱]آمده است.

۳-۱۱-۱-۲ تصدیق هویت گوینده توسط تلفیق درخت برآمدگی و الگوریتم ژنتیکی

در این بخش به تشریح چگونگی تلفیق ترکیب درخت برآمدگی و الگوریتم ژنتیکی برای انجام تصدیق هویت گوینده و توصیف نتایج حاصل از این روش می پردازیم. برای منظور آموزش و ارزیابی روش ارائه شده از دادگان FARSDIGITS1 استفاده نمودیم. از گفتار ۵۸ گوینده (۳٦ نفر مرد و ۲۲ نفر زن) این دادگان استفاده گردید. مجموعه گفتارهای هر گوینده به دوبخش ٦ تایی برای آموزش و ٤ تایی برای تست تقسیم شد. برای آموزش مدل هر گوینده علاوه بر گفتار وی از اطلاعات گفتار گویندگان همجنس او نیز استفاده شد. بدین منظور دو-سوم گویندگان همجنس وی بصورت اتفاقی انتخاب و همراه با اطلاعات وی جهت آموزش سیستم بکار گرفته شدند. در مرحله آزمایش، ٤ تکرار باقیمانده هر گوینده و همچنین یک-سوم گوینده های همجنس باقیمانده و همین تعداد گوینده غیر همجنس مورد استفاده قرار گرفتند. برای آموزش مدل یک گوینده و نیز انجام تستهای تصدیق هویت به کمک این مدل، از گفتار مربوط به بیان کد هفت رقمی توسط این گوینده و سایر گویندگان استفاده گردید.

سیگنال گفتار ورودی مربوط به بیان کد شناسائی گوینده به فریم های ۳۰ میلی‌ثانیه‌ای که ۲۰ میلی ثانیه همپوشانی دارند تقسیم شده و هر فریم تحت آنالیز LPC مرتبه ۱۲ قرار گرفت و از آن ۱۲ ویژگی LPCC استخراج گردید. هر گوینده توسط یک شبکه درختی ۵ لایه کامل شامل ۳۱ گره مدل گردید. تعداد لایه های فوق بدین دلیل انتخاب شدند که اولا تعداد لایه های بیشتر موجب می‌گردید که بدلیل محدودیت داده های آموزشی، مدل درختی هر گوینده شامل برگهائی تهی از داده آموزشی گردیده و در نتیجه درختهای ناقص ایجاد شود، ثانیا" تعداد لایه های کمتر پوشش ضعیفتری از اطلاعات یادگیری در فضای داده های آموزشی ایجاد می‌نمود.

گفتیم که آموزش درخت برآمدگی با استفاده از الگوریتم های ژنتیکی صورت گرفته است. مشخصات الگوریتم ژنتیکی مورد استفاده در این تحقیق عبارت است از:

تعداد تکرار (نسل) : ۵۰ تکرار

جمعیت هر نسل: ده درخت

نرخ ترکیب: ۷۰٪

نرخ جهش: ۱۰٪

تابع معیار: معکوس میانگین مربعات خطاهای شبکه‌های خطی

بدین ترتیب با استفاده از درخت برآمدگی و الگوریتم ژنتیکی با خصوصیات گفته شده مدلهای گویندگان ساخته شد. در مرحله تست هویت یک گوینده، بردار ویژگی حاصل از هر فریم گفتار ورودی به درخت اعمال شده و پس از پیمایش از ریشه به برگ در بر گیرنده آن، خروجی متناظر با آن برگ محاسبه گردید. این خروجی برای کلیه فریم های گفتار کد شناسائی ۷ رقمی گوینده محاسبه و میانگین‌گیری شد. این میانگین در واقع بیانگر میزان اختلاف بین مدل و گفتار ورودی می‌باشد. حال با در اختیار داشتن میانگین فوق و با توجه به آنچه که در بخش تصمیم گیری تصدیق هویت گوینده بیان نمودیم، با مقایسه این میانگین با سطح آستانه تصمیم گیری گوینده مورد نظر میتوان نسبت به رد یا قبول وی اقدام نمود.

برای مقایسه میزان کارائی روش فوق با کارائی روشهای متداول در تصدیق هویت گوینده، آزمایش دیگری صورت دادیم. در این آزمایش، از تکنیک چندی سازی برداریk- means برای مدل نمودن گویندگان استفاده کردیم. هر گوینده با استفاده از ۳۲و ٦٤ خوشه مدل گردید. در این روش به منظور تصدیق هویت یک گوینده گفتار وی با مدل گوینده ادعا شده مقایسه و فاصله بین بردار ویژگی هر فریم با مرکز ثقل نزدیکترین خوشه در مدل فوق بدست آمد و نهایتا میانگین فواصل بدست آمده بازاء کلیه فریمها در گفتار تست محاسبه گردید. این میانگین در واقع میزان اختلاف بین مدل گوینده ادعا شده و گفتار ورودی می‌باشد. با مقایسه این میانگین با سطح آستانه تصمیم گیری، میتوان نسبت به رد یا قبول گوینده تصمیم گیری نمود. نکته مهمی که در تصدیق هویت گوینده وجود دارد تعیین مناسب سطح آستانه تصمیم‌گیری می‌باشد. در آزمایشات فوق از روش نرخ خطای برابر و یا EER استفاده شده که این خطا نظیر محل تلاقی منحنی تغییرات خطاهای FR و FA می باشد. همانطور که در بخش بررسی خطاهای تصدیق هویت بیان نمودیم به منظور در نظر گرفتن تغییرات در محیط ضبط صدای گوینده در هنگام تست و همچنین تغییرات تدریجی صدای گوینده با گذشت زمان، در عمل بجای استفاده مستقیم از EER آنرا در یک ضریب بزرگتر ولی نردیک به یک ضرب می نمایند. بازاء مقادیر مختلفی برای این ضریب نتایج حاصل از تصدیق هویت با استفاده از درخت برآمدگی و

نیز روش چندی کردن برداری با تعداد ۳۲ و ٦٤ برای حجم کتابچه کد، در مرجع [۱] آورده شده است. در این نتایج مشاهده می‌شود که راندمان روش شبکه عصبی درختی در بهترین حالت بازاء ضریب ۱/۱ برابر ٪۹۳/۹ و راندمان روش چندی سازی برداری بازاء همین ضریب ٪۹٦/٦و ٪۹۷/۸ بترتیب بازاء ۳۲ و ٦٤ خوشه بدست می آید. افزایش تعداد خوشه‌ها از ۳۲ خوشه به ٦٤ خوشه تغییر قابل توجهی در نتایج حاصل ایجاد نمی‌نماید.

در آزمایش دیگر بجای استفاده از EER از روش یافتن خط برازش $y = c1*(m-d) + c2$ جهت تعیین آستانه تصمیم‌گیری استفاده شد. نتایج حاصل از تصدیق هویت گوینده با استفاده از درخت برآمدگی و چندی سازی برداری در جدول در مرجع [۱] آورده شده است. نتایج نشان میدهد که راندمان روش شبکه عصبی درختی در بهترین حالت بازاء ضریب ۱/۲ برابر ٪۹۱/۰ و راندمان روش چندی سازی برداری بازاء همین ضریب ٪۹٦/۰ و ٪۹٦/۹ بترتیب بازاء ۳۲ و ٦٤ خوشه بدست می آید. در آزمایش دیگری سعی شد تصمیم گیری نسبت به رد یا قبول گوینده با استفاده از روش نرمالیزه نمودن گروهی صورت گیرد. برای این منظور ۵ گوینده که در فاز آموزش نزدیکترین فاصله را با هر گوینده داشتند بعنوان گروه آن گوینده انتخاب شدند. برای تعیین فاصله نهایی یک گفتار با مدل هر گوینده، از فواصل گفتار ورودی با تک تک گویندگان هم گروه گوینده مرجع میانگین‌گیری شد و این میانگین بعنوان فاصله نهایی در نظر گرفته شد. با توجه به این فرضها، نتایج تصدیق هویت با استفاده از روش نرمالیزه نمودن گروهی برای روش درخت برآمدگی و روشهای تعیین سطح آستانه نرخ خطای برابر و برازش خط در مرجع [۱] آورده شده است. طبق نتایج حاصله با استفاده از نرمال سازی گروهی و روش تعیین سطح خطای نرخ برابر در بهترین حالت بازاء ضریب ۱/۱ کارایی برابر ٪۷۱/٦ بدست می آید. همچنین مشاهده گردید که با استفاده از نرمال سازی نتایج گروهی و روش تعیین سطح آستانه بکمک برازش خط، در بهترین حالت بازاء ضریب ۰/۹راندمان ٪٦۷/۳ بدست می‌آید. این نتایج نشان میدهد که استفاده از روش نرخ خطای برابر یا EER در مقایسه با روش برازش خط راندمان بالاتری را نتیجه می‌دهد. علت این موضوع با توجه به توضیحات مرجع [۱] بدین صورت قابل توضیح باشد که سطح آستانه بدست آمده در روش برازش خط مناسبترین سطح آستانه تصمیم گیری را برای آنکه خطای تصدیق هویت حداقل باشد، نتیجه نمی دهد.

۳-۱۱-۲ تصدیق هویت گوینده توسط سیستم هیبرید متشکل از مدل پنهان مارکُف و مدل مخلوط گاوسی

در قسمت قبل به بیان نحوه استفاده از ترکیب شبکه عصبی درخت برآمدگی و الگوریتم های ژنتیکی برای تصدیق هویت گوینده پرداختیم. مدل‌های پنهان مارکف HMM و مدل‌های مخلوط

گاوسیGMM [۷۸] کاربرد زیادی در تعیین هویت و تصدیق هویت گوینده دارند. در این قسمت از گزارش، با استفاده از تکنیک آمیختن داده ها[۷۹]، دو روش مدل مخفی مارکوف HMM و نیز مدل مخلوط گوسی GMM، را با هم موازی نموده و یک سیستم هیبرید را برای کاربرد در تعیین و تصدیق هویت گوینده بر روی خط تلفن تشکیل داده ایم. آزمایشها نشان می‌دهند که مدل هیبریدِ HMM⊕GMM در بازشناسی گوینده از هر یک از سیستم‌های HMM و GMM بهتر کار می‌کند. همچنین از آنجا که تاکید بیشتر این پروژه انجام تصدیق هویت گوینده از طریق خطوط تلفنی است لذا از تکنیک هائی برای افزایش کارائی در مقابل نویز و سایر تاثیرات خطوط تلفن استفاده نموده ایم. برای مقابله با نویز جمع‌شونده از روش تفاضل طیفی و نیز از معیارِ تصویر وزن‌دهی شده[۸۰]، و برای جبران‌سازی اثرِ کانالِ تلفن از روش تفاضل میانگین در حوزه کپسترال استفاده شده است که هر سه روش باعث بهبود سیستم بازشناسی گوینده شده‌اند.

۳-۱۱-۲-۱ پیش پردازش و استخراج ویژگی

برای مقابله با نویزِ جمعی موجود بر روی مکالمات از روش مشهور تفاضلِ طیفی [2] استفاده شده است. در این روش برای تخمین طیفِ نویز به الگوریتمی برای تشخیصِ گفتار از سکوت بصورت مقاوم نیاز داشتیم که از الگوریتم NP استفاده کردیم. ویژگیهای استفاده شده ویژگیهای مبتنی بر بانک فیلتر هستند که به طریق زیر بدست می آیند: فریم‌بندی سیگنالِ صحبت به فریم‌های ۳۵ میلی ثانیه که فاصلهٔ شروع هر دو فریمِ مجاور ۱۰ میلی ثانیه است، اعمال پیش تأکید (α=0.975)، اِعمال پنجرهٔ همینگ ، به‌کار بردن ۱۸ فیلتر مثلثی که بر اساس معیار *Mel* بر روی طیف فوریهٔ سیگنال توزیع شده‌اند ، استخراج ۱۲ ضریب کپسترال با اعمال یک لیفتر کاهنده در طرفین (لیفتر جوانگ) ، بدست آوردن مشتقات اول و دوم ضرایب کپسترال ، و در نهایت اعمال روش تفاضل میانگین در حوزهٔ کپسترال که در بخش بعدی توضیح داده خواهد شد.

۳-۱۱-۲-۲ آموزش مدلهای گویندگان

آموزش مدل پنهان مارکُف طی مراحل زیر صورت می‌گیرد :

الف- تقسیم یکسان بردارهای ویژگی بین حالات مدل و به دست آوردن تخمین اولیه برای پارامترهای مدل با استفاده از خوشه‌بندی .

[78] Gaussian Mixture Model
[79] Data Fusion
[80] WPM: Weighted Projection Measure

ب- تقسیم بهینهٔ بردارهای ویژگی بین حالات مدل با استفاده از الگوریتم ویتربی و خوشه‌بندی بردارها توسط الگوریتم *k-means* و به دنبال آن تخمین پارامترهای مدل و تکرار این کار تا همگرایی.

ج- تصحیح پارامترهای مدل با استفاده از فرمولهای تخمین بام- ولش و تکرار این کار تا رسیدن به همگرایی .

آموزش مدلهای مخلوط گاوسی نیز بصورت زیر است :

الف- تخمین اولیه برای میانگین‌ها ، واریانس‌ها و اوزانِ توابع گاوسی با استفاده از خوشه‌بندی .

ب- تصحیح پارامترهای مدل با استفاده از فرمولهای تخمین بام- ولش و تکرار این کار تا رسیدن به همگرایی .

به ازای هر گوینده یک مدل مخلوط گاوسی با ٦٤ تابع گاوسی در مدل و ۱۰ مدل پنهان مارکُف به ازای هر کدام از ارقام صفر تا نه درنظر گرفته شدند. یعنی به طور کلی برای ۱۰۰ گوینده ، ۱۰۰ مدل مخلوط گاوسی و ۱۰۰۰ مدل پنهان مارکُف درنظر گرفته شد. هر یک از مدلهای پنهان مارکُف دارای ٦ حالت و ۵ تابع گاوسی در هر حالت می‌باشد .

۳-۱۱-۲-۳ نحوه ساختن سیستم هیبرید

اگر فرض کنیم که O رشتهٔ مشاهدات (بردارهای ویژگی) باشد ، آنگاه احتمال تولید مشاهدات توسط مدل HMM را با احتمال تولید مشاهدات توسط مدل GMM به نحوهٔ زیر آمیخته می‌کنیم:

$$P_{hyb}(O|\lambda_i)=\alpha.P_n(O|\lambda_{i,GMM})+(1-\alpha).P_n(O|\lambda_{i,HMM}) \qquad (۵-۳)$$

که منظور از P_n، احتمال نرمالیزه شده می‌باشد که چنین به دست می‌آید :

$$P_n(O|\lambda_i)=P(O|\lambda_i)-\max_{j\neq i} P(O|\lambda_j) \qquad (٦-۳)$$

که منظور از λ_i گویندهٔ ادعا شده و λ_j گوینده‌ای غیر از گویندهٔ ادعا شده می‌باشد .

۳-۱۱-۲-٤ معیار تصویر وزن‌دهی شده

یکی از روشهایی که برای مقابله با نویز جمع‌شونده با پهنای باند گسترده[81] پیشنهاد شده است، استفاده از معیار تصویر وزن‌دهی شده یا *WPM* می‌باشد [٦۱، ٦۲، ٦۳]. آزمایشات نشان داده است که نویز سفید (یا نویزِ با پهنای باند گسترده) به صورت جمع‌شونده، بر اندازه یا طول بردارهای کپسترال تأثیر می‌گذارد، ولی جهتِ بردارها نسبت به نویزِ جمعی مقاوم‌تر است. برای محاسبهٔ فاصله بین دو بردار کپسترال نیز پیشنهاد شده است که به جای محاسبهٔ فاصلهٔ بین دو بردار، از یک معیار فاصلهٔ

[81] Additive Broadband Noise

مقاوم‌تر که به عبارتی دیگر زاویهٔ بین دو بردار را درنظر می‌گیرد، استفاده شود. یک تابع گاوسی برای محاسبهٔ احتمال در مدل پنهان مارکُف یا مدل مخلوط گاوسی چنین است :

$$N(o_t,\mu_i,\Sigma_i)=(2\pi)^{-\frac{n}{2}}.\left|\sum_i\right|^{-\frac{1}{2}}.\exp(-\frac{1}{2}(o_t-\mu_i)^T\Sigma_i^{-1}(o_t-\mu_i)) \qquad (۷-۳)$$

در این تابع گاوسی ، می‌توان فاصلهٔ اقلیدسی وزن‌دار را به عنوان معیاری برای فاصله به شکل زیر درنظر گرفت :

$$(o_t-\mu_i)^T\Sigma_i^{-1}(o_t-\mu_i)\, d_{WED}(o_t,\mu_i)= \qquad (۸-۳)$$

معیار فاصلهٔ اقلیدسی وزن‌دار و مبتنی بر تصویر و یا همان WPM به صورت زیر تعریف می‌گردد:

$$(o_t-\lambda\mu_i)^T\Sigma_i^{-1}(o_t-\lambda\mu_i)\, d_{WPM}(o_t,\mu_i)= \qquad (۹-۳)$$

مقدار بهینه برای λ باید طوری تعیین شود که بدون تغییر جهت در بردارهای o_t و μ_i ، فاصلهٔ وزن‌دار بین بردار o_t و $\lambda\mu_i$ می‌نیمم شود . مقدار بهینه برای λ از رابطهٔ زیر به دست می‌آید :

$$\lambda=\frac{o_t^T\Sigma_i^{-1}\mu_i}{\mu_i^T\Sigma_i^{-1}\mu_i} \qquad (۱۰-۳)$$

آزمایشات نشان داده است که با حضور نویز سفید جمع‌شونده با پهنای باندِ گسترده [۸۲] و حتی نویزهای رنگیِ جمع‌شونده با پهنای باند گسترده[۸۳] ، معیار WPM باعث ارتقاء کارآیی سیستم‌های بازشناسی می‌گردد. لازم به ذکر است که معیار WPM فقط در هنگام بازشناسی و تست سیستم اعمال می‌شود و در هنگام آموزش به هیچ تمهیدی نیاز ندارد. یکی از روشهایی که می‌تواند برای مقابله با نویزهای جمع‌شوندهٔ موجود بر روی خط تلفن بکار رود، همین معیار WPM است که در این طرح برای سیستم‌های بازشناسی گوینده به کار گرفته شده است .

۳-۱۱-۲-۵ آزمایشات

آزمایشات متعددی با استفاده از دادگان FARSDIGITS1 صورت گرفت. نصف تعداد تکرار های هر رقم توسط هر گوینده برای آموزش مدلهای گویندگان و نصف دیگر برای تست بکار گرفته شد. در دورهٔ تست، گوینده یکبار ارقام صفر تا نُه را برای شناخته شدن خود بیان می‌کند (ده رقم)، خطای تصدیق هویت گوینده بصورت $Error(\%)=\frac{FA+FR}{2}\times 100$ درنظر گرفته شده است.

آزمایش الف : در این آزمایش کارآیی سیستم هیبرید نسبت به هر یک از سیستم‌های HMM و GMM و به ازای ویژگیهای مختلف مقایسه می‌شود. نتایج این آزمایش در مرجع [۳] آمده است . کارآیی به صورت درصد صحت در تعیین هویت و درصد خطا در تصدیق هویت درنظر گرفته شده

[82] Additive Broadband White Noise
[83] Additive Broadband Colored Noise

است. یادآوری می‌شود که نتایج سیستم هیبرید بدون اعمال CMS و WPM می‌باشد. α_V و α_I مقادیر بهینهٔ α برای تصدیق و تعیین هویت گوینده می‌باشند. نتایج حاصله نشان میدهد که بازای سیستم‌های تک ماجولی (غیر هیبرید) ، بهترین کارآیی برای تصدیق هویت متعلق به HMM و به ازای پارامترهای MFCC+ΔMFCC است و بهترین کارآیی برای تعیین هویت متعلق به GMM و به ازای پارامترهای MFCC+ΔMFCC+ΔΔMFCC است. اضافه کردن مشتق اول ضرایب یعنی ΔMFCC کارآیی HMM را بالا می‌برد (به علت مدل کردن دینامیک محلی مجرای گفتار که مدل مارکف با تعداد حالات محدود برابر شش حالت احتمالاً قادر به لحاظ کردن آن نیست) ولی اضافه کردن ضرایب ΔΔMFCC (مشتق دوم) ، کارآیی HMM را پایین می‌آورد. این پدیده شاید به این دلیل باشد که با اضافه کردن مشتق دوم ضرایب به سمت مدل کردن دینامیک سراسری گفتار نزدیک می‌شویم که خود مدل پنهان مارکف با ماتریس گذر بین حالات آن را بهتر مدل می‌کند و اضافه کردن مشتق دوم ضرایب سودی ندارد ولی اضافه کردن مشتق دوم ضرایب در مدل مخلوط گاوسی که فاقد احتمالات گذر بین حالات است، کارآیی GMM را افزایش می‌دهد.

آزمایش ب : در این آزمایش اثر معیار تصویر وزن‌دهی شده بر روی یک سیستم بازشناسی گوینده مبتنی بر GMM بررسی می‌شود. با استفاده از پارامترهای MFCC و مشتق اول و دوم آنها و با ٦٤ تابع گاوسی، نتایج برای تصدیق و تعیین هویت گوینده بدون اعمال WPM و با اعمال WPM، در مرجع [۳] آمده است. در آنجا ملاحظه می‌شود که در هر دو حالت تصدیق و تعیین هویت گوینده اعمال WPM باعث پیشرفت کارآیی سیستم گردیده است. با اعمال WPM نرخ تصدیق هویت به میزان ۰۳/ ٪ کاهش می یابد.

آزمایش ج : در این آزمایش اثر تفاضل میانگین در حوزهٔ کپسترال یا CMS بر روی یک سیستم تصدیق هویت گوینده بررسی گردید. ملاحظه می‌شود که CMS باعث کاهش اثر کانال انتقال تلفنی بر روی پارامترهای کپسترال و در نتیجه کاهش خطای تصدیق هویت گوینده می‌گردد که در مرجع [۳] آورده شده است. ذکر این نکته لازم است که از WPM در این آزمایش استفاده نشده است. اعمال CMS خطا را به میزان ۲٤/ ٪ کاهش میدهد.

آزمایش د : در این آزمایش یک بررسی بر روی تعداد افراد جمعیت انجام می‌گیرد و راندمان سیستم برای تصدیق و تعیین هویت گوینده به ازای جمعیت‌های ۲۰، ٤۰، ۷۰ و ۱۰۰ نفری اندازه‌گیری می‌شود. سیستم پایه برای این آزمایش، مدل مخلوط گاوسی با ٦٤ تابع گاوسی وبا استفاده از معیار تصویر وزن‌دهی شده می‌باشد. نتایج این آزمایش در مرجع [۳] درج گردیده است. مشاهده می‌گردد که با افزایش تعداد گویندگان نرخ خطا در تعیین هویت گوینده بصورت خطی و بسیار سریعتر از نرخ خطا در تصدیق هویت گوینده رشد میکند. این میزان خطا برای تصدیق هویت ۰/۳۷۳ با

افزایش گویندگان از ۲۰ نفر به ۱۰۰ نفر بوده و با افزایش گویندگان به بیش از ۲۰ نفر تقریبا ثابت می باشد.

۳-۱۱-۳ مقایسهٔ چند روشِ نرمالیزاسیونِ امتیازات[۸۴] در سطحِ گویش و در سطحِ فریم برای افزایشِ کارآییِ تصدیق هویت گوینده بر روی خط تلفن

یکی از مسائل مهم در سیستم‌های بازشناسی گوینده، جنبهٔ تصمیم‌گیری است. در سیستم های تصدیق و تعیین هویت گوینده اگر امتیازات خام[۸۵] نرمالیزه گردند، میزان تمایز یک گوینده از گویندگان دیگر افزایش یافته و کارآیی سیستم بالا خواهد رفت. روشهای نرمالیزاسیون را می توان به روشهای نرمالیزاسیون امتیازات در سطح گویش، روشهای نرمالیزاسیون امتیازات در سطح فریم و روشهای نرمالیزاسیون امتیازات هم در سطح گویش و هم در سطح فریم تقسیم کرد. در بعضی سیستم ها نیز امتیازات در سطح فریم را برای بالابردن کارآیی سیستم وزن‌دهی می‌کنند. یکی از جنبه های دیگر در سیستم های بازشناسی گوینده، تعداد گویندگان است. نتایج بخش های قبل نشان داد که نرخ خطای تعیین هویت با زیاد شدن تعداد اعضای جمعیت بصورت خطی افزایش می یابد. در عوض نرخ خطای تصدیق هویت برای تعداد جمعیت بیش از ۲۰ نفر تقریباً ثابت می ماند و اضافه نمی گردد. در این بخش، چند روش نرمالیزاسیون امتیازات در سطح گویش و در سطح فریم و نیز روش وزن‌دهی امتیازات مدل برای افزایش کارآیی سیستم‌های تصدیق و تعیین هویت گوینده مورد ارزیابی قرار گرفته است. نشان داده شده که این روشها باعث افزایش تمایز بین گویندگان و در نتیجه کاهش خطا در سیستم‌های تصدیق و تعیین هویت گوینده می‌گردند.

۳-۱۱-۳-۱ روشهای نرمالیزاسیون امتیازات [٦٤]

معیار Bayes برای طبقه بندی یک مسألهٔ دو کلاسه را می توان بصورت زیر بکار برد :

$$\text{if } P(q_1|X) \geq P(q_2|X) \text{ then } X \in q_1 \text{ else } X \in q_2 \qquad (۱۱-۳)$$

اگر فرض کنیم X رشتهٔ مشاهدات یا رشتهٔ بردارهای ویژگی بصورت $X=\{x_1,x_2, \dots ,x_t, \dots , x_T\}$ و q_1 کلاس خودگوینده و q_2 کلاس گویندگان دیگر باشد آنگاه :

$$\frac{P(q_1).P(X \mid q_1)}{P(X)} \geq \frac{P(q_2).P(X \mid q_2)}{P(X)} \qquad (۱۲-۳)$$

$$\text{if } \frac{P(X \mid q_1)}{P(X \mid q_2)} \geq \left(\frac{P(q_2)}{P(q_1)} = Thr \right) \text{ then } X \in q_1 \text{ else } X \in q_2 \qquad (۱۳-۳)$$

[84] Score Normalization
[85] Raw Score

Thr سطح آستانهٔ تصمیم‌گیری می‌باشد. اگر λ_i و λ_c مدل گویندهٔ مدعی[۸۶] یا ادعا شده [۸۷] و λ_j طوری که j≠i ، مدل گویندگان دیگر باشد و نیز $\lambda_{\bar{c}}$ مدل گویندگان غیر از گویندهٔ ادعا شده باشد و S تعداد کل گویندگان باشد، آنگاه داریم :

$$\text{if } \frac{P(X \mid \lambda_c)}{P(X \mid \lambda_{\bar{c}})} \geq Thr \text{ then } X \in c \text{ else } X \in \bar{c} \qquad (۱۴-۳)$$

این احتمال را احتمال نرمالیزه شده یا نسبت احتمالات می نامند. اگر مبنای تصمیم گیری بصورت زیر باشد، احتمال نرمالیزه نشده یا خام مورد نظر است :

$$\text{if } P(X|\lambda_c) \geq Thr \text{ then } X \in c \text{ else } X \in \bar{c} \qquad (۱۵-۳)$$

بعضی اوقات از لگاریتم نسبت احتمالات[۸۸] استفاده می شود :

$$Log(\frac{P(X \mid \lambda_c)}{P(X \mid \lambda_{\bar{c}})}) = Log\ P(X|\lambda_c) - Log\ P(X|\lambda_{\bar{c}}) \qquad (۱۶-۳)$$

مدل $\lambda_{\bar{c}}$ را مدل ضد گوینده[۸۹] نیز می نامند. $P(X|\lambda_{\bar{c}})$ را می توان بصورت زیر بدست آورد :

$$P(X|\lambda_{\bar{c}}) = \frac{1}{S-1}\sum_{j \neq i} P(X \mid \lambda_j) = \sum_{j \neq i} P(\lambda_j).P(X \mid \lambda_j) \qquad (۱۷-۳)$$

فرمول بالا با این فرض نوشته شده است که احتمال پسین مدلها یعنی $P(\lambda_j)$ ها مساوی و برابر با $\frac{1}{S}$ باشند، بنابراین می‌توان نوشت:

$$Log(\frac{P(X|\lambda_c)}{P(X|\lambda_{\bar{c}})}) = \text{Log } P(X|\lambda_i) - Log(\frac{1}{S-1}\sum_{j \neq i} P(X|\lambda_j)) \qquad (۱۸-۳)$$

در فرمول فوق برای محاسبهٔ احتمال روی سایر گویندگان غیر از گوینده ادعا شده از میانگین استفاده گردیده است. بطور کلی و بصورت تقریبی می توان برای محاسبهٔ احتمال نرمالیزه شده در سطح گویش از فرمول کلی زیر بهره گرفت :

$$Log(\frac{P(X|\lambda_c)}{P(X|\lambda_{\bar{c}})}) = \text{Log } P(X|\lambda_i) - Log(\underset{j \neq i}{Stat}\ \{P(X|\lambda_j)\}) \qquad (۱۹-۳)$$

که منظور از *Stat* انجام یک عملیات آماری مانند میانگین و ... بر روی امتیازات حاصل از مدل‌های گوینده‌های دیگر می باشد. برای محاسبهٔ احتمال نرمالیزه‌شده بجای Stat می‌توان از آماره‌های[۹۰] زیر استفاده کرد:

الف- احتمال پسین[۹۱] [٦٤]

[86] Claimant Speaker
[87] Claimed Speaker
[88] Log-likelihood Ratio (LLR)
[89] Anti-Speaker
[90] Statistics
[91] Posterior Probability

در محاسبه احتمال نرمالیزه شده به این روش در واقع از فرمول محاسبهٔ احتمال پسین برای گویندهٔ i استفاده می کنیم :

$$P_n(X|\lambda_i)=\text{Log } P(X|\lambda_i)-\text{Log}\left(\frac{1}{S}\sum_{j=1}^{S} P(X|\lambda_j)\right) \qquad (۲۰-۳)$$

همانطور که مشاهده می‌شود برای محاسبهٔ $P(X|\lambda_{\bar{c}})$ در طرف راست معادله، احتمال $P(X|\lambda_i)$ نیز در $\sum$ در نظر گرفته می‌شود. استفاده از این فرمول برای محاسبهٔ احتمال نرمالیزه‌شده باعث می‌شود که اگر گویندهٔ مدعی، دروغگو باشد، بازای یکی از λ_j ها در قسمت $\sum$ از فرمول، که در حقیقت مدل خود گویندهٔ دروغگو است، $P(X|\lambda_j)$ زیاد شده و در کل $P_n(X|\lambda_i)$ کاهش یابد و احتمال قبول شدن او کمتر گردد.

ب- میانگین [۶۵، ۶۶]

فرمول محاسبه احتمال نرمالیزه شده به این روش به شکل زیر می باشد :

$$P_n(X|\lambda)=\text{Log } P(X|\lambda_i)-\text{Log}\left(\frac{1}{S-1}\sum_{j\neq i} P(X|\lambda_j)\right) \qquad (۲۱-۳)$$

تفاوت این فرمول با فرمول محاسبه به روش احتمال پسین این است که $P(X|\lambda_i)$ یعنی احتمال به ازای مدل گوینده ادعا شده در قسمت $\sum$ از فرمول، لحاظ نمی‌گردد. محاسبهٔ احتمال نرمالیزه‌شده به این روش نیز به دلیل مشابه با احتمال پسین موجب کاهش خطا می‌گردد.

ج- ماکزیمم [۶۶]

احتمال نرمالیزه‌شده در این روش به طریقه زیر محاسبه می گردد :

$$P_n(X|\lambda_i)=\text{Log } P(X|\lambda_i)-\text{Log}\left(\underset{j\neq i}{Max}\ P(X|\lambda_j)\right) \qquad (۲۲-۳)$$

استفاده از این آماره در واقع باز هم احتمال قبول شدن فرد دروغگو را کم می‌کند، زیرا ماکزیمم امتیاز فرد مدعی دروغگو روی مدلهای غیر از مدل ادعا شده (که اتفاقاً مدل حقیقی فرد دروغگو نیز یکی از آنهاست) زیاد بوده و امتیاز نرمالیزه شده او کم است. این نحوهٔ محاسبهٔ احتمال نرمالیزه شده همچنین به گوینده مدعی راستگو که ماکزیمم امتیاز او بر روی مدلهای دیگر معمولاً کم است، کمک می‌کند و احتمال نرمالیزه شده شخص مدعی راستگو، در سطح بالایی باقی می‌ماند.

د- می نیمم [۶۵]

احتمال نرمالیزه شده را در این روش بطریقه زیر محاسبه می کنیم:

$$P_n(X|\lambda_i)=\text{Log } P(X|\lambda_i)-\text{Log}\left(\underset{j\neq i}{Max}\ P(X|\lambda_j)\right) \qquad (۲۳-۳)$$

این روش از این ایده بهره می گیرد که می نیمم امتیاز فرد مدعی دروغگو بر روی مدلهای غیرادعا شده (که مدل خود او نیز در میان آنهاست) زیاد است، ولی می‌نیمم امتیاز فرد مدعی راستگو بر روی

مدلهای غیرادعا شده معمولاً کم است و بنابراین این روش، مدعی دروغگو را سرکوب کرده و به مدعی راستگو کمک می‌کند.

ه- شبیه‌ترین M گوینده [٦٧]

اگر امتیازات بردارهای ویژگی روی مدل‌های غیر ادعا شده را به ترتیب نزولی مرتب کنیم و M عدد از این امتیازات را که بیش از دیگران هستند برداریم، آنگاه احتمال نرمالیزه‌شده چنین است :

$$P_n(X|\lambda_i)=\text{Log } P(X|\lambda_i)-\text{Log}\left(\frac{1}{M}\sum_{j=1, j\neq i}^{M} P(X|\lambda_j)\right) \qquad (٢٤-٣)$$

که در آن :

$$P(X|\lambda_j) \geq P(X|\lambda_{j+1}) \geq \ldots \qquad (٢٥-٣)$$

و- نرمالیزاسیون گروهی[92] [٦٦]

در این روش که نرمالیزاسیون گروهی نام دارد، گروهی از گویندگان از میان گویندگان غیر از گوینده ادعا شده که به گوینده ادعا شده بیشتر شبیه هستند در دوره آموزش تعیین می شوند تعداد گویندگان این گروه، C می باشد و احتمال نرمالیزه شده به این طریق محاسبه می گردد :

$$P_n(X|\lambda_i)=\text{Log } P(X|\lambda_i)-\text{Log}\left(\frac{1}{C}\sum_{j\in Cohort(i)\ ,\ j\neq i} P(X|\lambda_j)\right) \qquad (٢٦-٣)$$

شباهت گویندهٔ i و گوینده j در دورهٔ آموزش به این طریق بدست می‌آید که امتیازات نمونه‌های آموزشی گویندهٔ i ام بر روی مدل گویندهٔ j ام و نیز امتیازات نمونه‌های آموزشی گویندهٔ j ام بر روی مدل i ام محاسبه شده و میانگین این دو عدد، میزان شباهت دو گویندهٔ i و j را نشان می‌دهد. بدیهی است که در روشهایی مثل کوانتیزاسیون برداری مقدار فاصلهٔ بدست آمده یا اعوجاج بدست آمده میزان عدم شباهت دو گوینده را نشان می‌دهد. مزیت بزرگی که نرمالیزاسیون گروهی نسبت به پنج روش قبلی دارد، این است که امتیاز مشاهدات به ازای تمام مدلهای گویندگان غیر مدعی محاسبه نمی‌شود و فقط به ازای گروهی از آنها محاسبه می گردد. قابل ذکر است گویندگانی که می توانند در یک گروه، شبیه به گوینده i ام قرار گیرند، لزومی ندارد که از همان جنسیت باشند.

ز- نرمالیزاسیون گروهی هیبرید[93] [٦٥]

اگر می‌نیمم امتیازی (احتمال) را که گویش‌های گویندهٔ i ام در دورهٔ آموزش بر روی مدل خودِ گویندهٔ i ام یعنی λ_i کسب می‌کنند، $S^i_{\min}$ بنامیم، آنگاه احتمال نرمالیزه‌شده به این طریق بدست خواهد آمد :

$$\text{then} \quad S^i_{\min} \text{ if } \quad P(X|\lambda_i) > k$$

[92] Cohort Normalization
[93] Hybrid Threshold Cohort Normalization

$$P_n(X|\lambda_i)=\text{Log } P(X|\lambda_i)-\text{Log}(\frac{1}{C}\sum_{j\in Cohort(i)\ ,\ j\neq i} P(X|\lambda_j)) \qquad (۲۷-۳)$$

$$P_n(X|\lambda_i)= -\infty \qquad \text{Else}$$

که در این فرمول k می‌تواند مقداری بعنوان مثال در حدود 0.9 داشته باشد. میزان محاسبات این روش نیز مانند روش نرمالیزاسیون گروهی است.

ذکر این نکته لازم است که در تمامی روشهای فوق بطور تقریبی می‌توان به جای $Log(\frac{\sum P_j}{S-1})$ از $\frac{\sum Log\ P_j}{S-1}$ استفاده کرد که بجای میانگین حسابی، میانگین هندسی را محاسبه می‌نماید. در مرجع [۷۱] نشان داده شده که ایندو تقریباً یکسان و در برخی موارد، واسطهٔ هندسی جواب بهتری را داده است. همچنین می توان نشان داد که روشهای نرمالیزاسیونِ در سطح گویش، نرخ خطای تصدیق هویت گوینده را تحت تأثیر قرار می‌دهند ولی بر نرخ خطای تعیین هویت گوینده تأثیری ندارند.

۳-۱۱-۳-۲ روشهای نرمالیزاسیون امتیازات در سطح فریم [٦٧]

اگر $P_n(x_t|\lambda_i)$ احتمال نرمالیزه شدهٔ بردار x_t بر روی مدل گوینده i ام باشد، این احتمال را به طریق زیر حساب می کنیم :

$$P_n(x_t|\lambda_i)=\text{Log } P(x_t|\lambda_i)-\text{Log}(\underset{j}{Stat}\ \{P(x_t|\lambda_j)\}) \qquad (۲۸-۳)$$

که آمارهٔ $Stat$ می‌تواند هر یک از آماره‌های مطرح شده در قسمت قبل باشد. احتمال رشتهٔ بردارها در سطح گویش چنین بدست خواهد آمد :

$$P(X|\lambda_i)=\frac{1}{T}\sum_{t=1}^{T} P_n(x_t|\lambda_i) \qquad (۲۹-۳)$$

اگر بخواهیم $P(X|\lambda_i)$ را باز هم در سطح گویش نرمالیزه کنیم و بعنوان مثال آمارهٔ مورد نظر ما در سطح گویش آمارهٔ ماکزیمم باشد، چنین عمل می‌کنیم :

$$P_n(X|\lambda_i)=P(X|\lambda_j)-\underset{j\neq i}{MaxP}(X|\lambda_i) \qquad (۳۰-۳)$$

فرمول فوق با این فرض است که $P(X|\lambda_i)$ خود لگاریتم احتمال است نه احتمال. به این احتمال بدست آمده، احتمال نرمالیزه شده هم در سطح فریم و هم در سطح گویش می‌گویند.

۳-۱۱-۳-۳ وزن‌دهی امتیازاتِ مدل [٦٧]

اگر فرض کنیم $P(x_t|\lambda_i)$ احتمال تولید بردار x_t توسط مدل گوینده i اُم باشد، آنگاه احتمالات $P(x_t|\lambda_1)$ ، $P(x_t|\lambda_2)$ ، ... و $P(x_t|\lambda_S)$ را بصورت نزولی مرتب کرده و به هر مدل یک رتبه اختصاص می‌دهیم و برای مدل λ_i ، رتبه را r_i می‌نامیم. مدلی که بیشترین احتمال را تولید کند، دارای رتبهٔ ۱ و

مدلی که کمترین احتمال را تولید کند، دارای رتبهٔ S است (S تعداد گویندگان جمعیت است). امتیاز وزن‌دهی شدهٔ مدل از این طریق محاسبه می‌گردد :

$$P_w(x_t|\lambda_i)=w(r_i).P(x_t|\lambda_i) \qquad (۳۱-۳)$$

که در آن $w(r)$ ، یک تابع وزنِ کاهنده و بعنوان مثال بصورت زیر می‌باشد :

$$w(r)=\frac{S}{\alpha .r} \qquad (۳۲-۳)$$

مقدار α به عنوان نمونه می‌تواند یک باشد. در اینجا پس از وزن‌دهی امتیازات مدل در سطح فریم، عمل نرمالیزاسیون در سطح گویش نیز انجام می گیرد، بدین معنی که ابتدا وزن دهی امتیازات در سطح فریم مدل انجام می شود، احتمالات وزن‌دهی شدهٔ بردارها برای محاسبهٔ احتمال رشتهٔ بردار بر روی هم انباشته شده و سپس عمل نرمالیزاسیون بر روی احتمال در سطح گویش (رشتهٔ بردار) انجام می‌گیرد.

۳-۱۱-۳-٤ استخراج ویژگی

از ویژگیهای کپسترال مبتنی بر بانک فیلتر، به اضافهٔ مشتق اول و دوم ضرایب کپسترال استفاده گردید. سیگنال زمانی به فریم‌های ۳۵ *ms* که فاصلهٔ دو فریم متوالی ۱۰ *ms* بوده، تقسیم می‌گردد و پیش تأکید و پنجرهٔ همینگ اعمال می‌شود. قبل از این مراحل سیگنال گفتار توسط یک الگوریتم تشخیص مقاوم گفتار از سکوت NP و نیز روشِ تفاضلِ طیفی بهسازی گردید تا نویز جمعیِ موجود در سیگنال کمتر شود. پس از پیش تأکید و پنجرهٔ همینگ، طیف سیگنال بدست آمده و ۱۸ فیلتر مثلثی با توزیع بر اساس معیار *Mel* ، بر روی طیف اعمال شد و ۱۲ ضریب کپسترال استخراج گردید. با استفاده از یک لیفتر کاهنده در طرفین (لیفتر جوانگ)، ضرایب کپسترال و مشتقات اول و دوم آن وزن‌دهی و پس از این مرحله برای مقابله با اثرِ کانال انتقال، از تکنیکِ تفاضل میانگین در حوزهٔ کپسترال استفاده شدو میانگینِ بردارهای کپسترال در نواحی گفتار، محاسبه شده و این میانگین از بردارهای ویژگیِ اولیه کم گردید.

۳-۱۱-۳-۵ آموزش مدلهای مخلوط گاوسی

برای آموزش مدلهای مخلوط گاوسی، ابتدائاً با استفاده از خوشه‌بندی *k-means* ، یک مقدار اولیه برای میانگین، واریانس و اوزان خوشه‌ها بدست آمد. سپس در مرحلهٔ بعد با استفاده از فرمولهای تخمین بام-ولش پارامترهای مدل تخمین زده می‌شوند و این عمل تا رسیدن به شرط همگرایی ادامه یافت.

۳-۱۱-۳-٦ آزمایشات

پایگاه دادهٔ مورد استفاده پایگاه دادهٔ تلفنیِ FARSDIGITS1 می‌باشد. برای هر گوینده یک مدل مخلوط گاوسی با ٦٤ تابع گاوسی درنظر گرفته شده و به ازای ۱۰۰ گوینده، ۱۰۰ مدل مخلوط گاوسی آموزش داده می‌شود. در مرحلهٔ آزمایش هر یک از گویندگان، ارقامِ صفر تا نه را یکبار اَدا می‌کنند و راندمان سیستم اندازه‌گیری می‌شود.

آزمایش الف : در این آزمایش، هدف آن است که تأثیر روشهای نرمالیزاسیونِ در سطح گویش را بررسی کنیم. مرجع [٤] نتایج به‌دست آمده را به ازای روشهای مختلف نشان می‌دهد نتایج نشان داده اند که در بهترین حالت و به ازای یکبار بیانِ ارقام صفر تا نه توسط گوینده، خطای تصدیقِ هویتِ گوینده از ۳/۲۵٪ به ۰/٤٪ رسیده است که کاهش بسیار چشمگیری را نشان می‌دهد. همچنین می‌توان مشاهده نمود که حتی آمارهٔ می‌نیمم برای نرمالیزه کردنِ امتیازات نیز تا حدّی خطای تصدیق هویت را کاهش می‌دهد. آزمایشات نشان دهندهٔ این موضوع است که آمارهٔ ماکزیمم، کمترین خطا را داشته و نیز اینکه روشهای نرمالیزاسیون در سطح گویش تأثیری بر نرخ صحتِ تعیین هویت گوینده ندارند که با توجه به فرمولهای ارائه شده امری منطقی است.

آزمایش ب : در این آزمایش، هدف آن است که اثر نرمالیزاسیون امتیازات هم در سطحِ فریم و هم در سطحِ گویش بر روی نرخ صحت بازشناسی گوینده بررسی گردد. امتیازات مربوط به هر بردار ویژگی (در سطح فریم) به پنج روش از روشهای گفته نرمالیزه و سپس احتمالِ انباشته شده (در سطح گویش) با استفاده از آمارهٔ ماکزیمم نرمالیزه می‌گردد. نتایج حاصل از این آزمایش در مرجع [٤] درج گردیده است. از نتایج حاصله ملاحظه می‌گردد که در بهترین حالت، یعنی استفاده از آمارهٔ ماکزیمم، نرمالیزه کردن امتیازات در دو سطحِ فریم و گویش، کارآیی را نسبت به بهترین نتیجهٔ حاصل شده از نرمالیزاسیون امتیازات فقط در سطحِ گویش، ارتقاء می‌دهد. همچنین ملاحظه می‌گردد که نرمالیزاسیون امتیازات در سطح فریم نرخ صحت در تعیین هویت گوینده را بر خلاف روشهای نرمالیزاسیون در سطح گویش، تحتِ تأثیر قرار می‌دهد.

آزمایش ج : هدف از این آزمایش آن است که اثر وزن‌دهیِ امتیازاتِ مدل بر روی نرخ صحت تعیین هویت گوینده و نرخ خطای تصدیق هویت گوینده بررسی گردد. برای این منظور ابتدا امتیازاتِ مدلها در سطح فریم وزن‌دهی شده و سپس احتمال انباشته شده را در سطحِ گویش با استفاده از آمارهٔ ماکزیمم نرمالیزه می‌کنیم. نتایج به‌دست آمده در مرجع [٤] درج گردیده است. ملاحظه میشود که وزن‌دهیِ امتیازات در سطح فریم به میزان کمی خطای تصدیق هویت را کاهش داده است ولی بر نرخِ صحت در تعیین هویت گوینده تأثیری نداشته است.

۳-۱۲ نتیجه گیری

آزمایشات صورت گرفته جهت بازشناسی ارقام با تلفیق شبکه عصبی پیشگو و برنامه ریزی پویا نشان داد که افزایش بیش از حد تعداد گره ها در لایه مخفی و نیز ارائه بیش از حد داده های آموزشی به مدل در هنگام آموزش موجب گرایش بیش از حد مدل به داده های آموزشی و کاهش قدرت تعمیم پذیری آن و در نتیجه کاهش راندمان بازشناسی میگردد. در خصوص نوع ویژگیها و تعداد آنها نیز معلوم گردیدکه که افزایش تعداد ویژگیها تا حد معقول در بهبود راندمان مؤ ثر میباشد، ضمن آنکه اطلاعات دینامیک گفتار موجود در مشتق ضرائب کپسترال میتواند در بهبود کارائی تاثیر مثبت داشته باشد.

با استفاده از مدل پنهان مارکف پیوسته برای بازشناسی ارقام گسسته فارسی همراه با استفاده از روش تفاضل طیفی برای مقابله با نویز جمع شونده بر روی خط تلفن و روش تفاضل در حوزهٔ کپسترال برای جبران اثر کانال معلوم شد که بهترین کارایی به ازای پارامترهای کپسترال حاصل از بانک فیلتر با توزیع مل، لگاریتم انرژی، مشتق اول و دوم لگاریتم انرژی و ضرایب کپسترال بدست می آید. این میزان کارایی برای داده های آزمایشی ۹۹/۱۰٪ است. آزمایشها نشان داد که در محیط نویزی خط تلفن، پارامترهای حاصل از بانک فیلتر بهتر از پارامترهای حاصل از پیشگویی خطی عمل می کنند، خصوصا" آنکه مشخصه کانال تلفنی از روش تفاضل میانگین در حوزه کپسترال جبران گردد. همچنین مشخص گردید که مدل نمودن سکوت نتیجه بهتری را در مقایسه با حذف سکوت بدست میدهد. نتیجه بدست آمده برای پایگاه دادهٔ تلفنی با SNR=8.8dB گویای راندمان نسبتاً بالای استفاده از مدل مخفی مارکوف برای شناسایی ارقام می باشد.

روش بازتخمین نهفته به صحت بازشناسی بیشتری نسبت به روش معمول تخمین باوم–ولش برای بازشناسی ارقام تلفنی متصل فارسی منجر گردید. این روش با تخمین پارامترهای گذر بین حالات در حالت های مرزی (ابتدا و انتها) موجب افزایش راندمان سیستم بازشناسی ارقام متصل با رشته های دو رقمی بر روی خط تلفن به ازای داده های آزمایشی میگردد. آزمایشات انجام شده نشان دادند که اگر تعداد تکرار های بازتخمین نهفته کم یا بیش از حد باشد، کارایی سیستم افت خواهد کرد.

استفاده از روش درخت برآمدگی بعنوان یک روش سلسله مراتبی جهت خوشه‌بندی داده‌ها در تصدیق هویت گوینده از مزیت سرعت بازیابی نسبتا" زیاد در مقایسه با دیگر روشهای خوشه‌بندی برخوردار است. مزیت دیگر آن سازگاری جهت ادغام با روشهای ژنتیکی می‌باشد. اگرچه شبکه‌های عصبی چند لایه نیز این قابلیت را دارند اما کندی سرعت آنها باعث کاهش توجه به این قابلیت آنها می‌شود. لازم به ذکر است که شبکه‌های عصبی چند لایه برای آموزش از قاعده انتشار به عقب استفاده می‌کنند که در این الگوریتم لازم است تکرارهای بسیاری جهت یافتن پارامترهای بهینه

صورت گیرد، در صورتیکه همچنان که گفته شد در روش شبکه های عصبی درختی تنها در یک مرحله و با سرعتی بالا این امر میسر می‌باشد. نتایج حاصل از این بررسی بنوعی روشنگر نکات دیگری نیز میباشد که مهمترین آنها عبارتند از:

- سرعت بازیابی خوشه‌ها در این روش سریعتر از دیگر روشهای خوشه‌بندی می‌باشد.
- زمان آموزش این روش بدلیل بکارگیری الگوریتمهای ژنتیکی که عموما زمانبر می‌باشند نسبت به سایر روشها طولانی‌تر می‌باشد.
- از مقایسه آزمایش این روش با آزمایش چندی سازی برداری معمولی چنین برمی‌آید که کارایی این روش کمتر از روش چندی سازی برداری بروش k-means می‌باشد اما مزیت آن سرعت بسیار بالاتر این روش در مقایسه با روشهای معمول مثل روش k-means می‌باشد. در حالات کلی رتبه محاسبات لازم برای بازیابی خوشه در برگیرنده یک بردار ورودی در الگوریتمهای خوشه‌بندی معمولی $O(n)$ و در درخت برآمدگی $O(\log_2 n)$می‌باشد.

در خصوص تعیین سطح آستانه در تصدیق هویت گوینده نیز مشاهده شد که استفاده از روش نرخ خطای برابر یا EER در مقایسه با روش برازش خط راندمان بالاتری را نتیجه می‌دهد، لیکن روش نرخ خطای برابر نیاز به تعداد کافی و زیاد تستهای درون گویندگی با استفاده از داده های آموزشی دارد که این امر موجب می‌گردد که لازم شود گویندگان در تعداد دفعات بیشتری کد شناسایی شخصی خود را جهت آموزش سیستم بیان نمایند. از طرف دیگر روش تعیین سطح آستانه تصمیم گیری به روش برازش خط گرچه راندمان پایین تری را سبب می‌شوند، لیکن چون تعیین سطح آستانه به کمک آن تنها به آزمایشات برون گویندگی نیاز دارد، بنابراین این امکان فراهم می‌شود که گویندگان در تعداد دفعات کمتری کد شناسایی شخصی خود را بیان نمایند که این موضوع بدلیل سهولتی که برای گویندگان در مرحله آموزش فراهم می‌نماید در سیستم‌های واقعی تصدیق هویت گوینده از اهمیت بالایی برخودار می‌باشد.

آزمایشات استفاده از مدل هیبرید برای تصدیق و تعیین هویت گوینده متشکل از مدل پنهان مارکف و مدل مخلوط گاوسی نشان داد که مدل هیبرید از هر یک از مدلهای پنهان مارکف و مخلوط گاوسی کارآیی بیشتری دارد. همچنین برای مقابله با نویز جمع‌شوندهٔ موجود بر روی مکالمات تلفنی از روش تفاضل طیفی و برای کاهش اثر نویز جمع‌شونده بر روی بردارهای کپسترال از روش تصویر وزن‌دهی شده یا WPM استفاده گردید که راندمان سیستم بازشناسی گوینده را ارتقاء بخشید. برای مقابله با اثر کانال انتقال بر روی بردارهای کپسترال از روش تفاضل میانگین در حوزهٔ کپسترال یا CMS استفاده شد که این روش نیز راندان سیستم را افزایش داد.

آزمایشات نشان داد که نرمالیزاسیونِ امتیازات در هر دو سطحِ گویش و فریم در سیستم‌های تعیین و تصدیق هویت گوینده نتیجهٔ بهتری نسبت به نرمالیزاسیون فقط در سطح گویش می‌دهد. همچنین ملاحظه گردید که روش وزن‌دهیِ امتیازاتِ مدل در سطح فریم و قبل از نرمالیزاسیونِ امتیازات در سطح گویش، کارآیی سیستم را نسبت به حالت نرمالیزاسیون فقط در سطح گویش ارتقاء می‌دهد. که مخصوصاً به ازای تصدیق هویت گوینده این افزایش کارائی بسیار چشمگیری میباشد.

در خاتمه ابراز میدارد که آنچه در این گزارش ارائه گردید ماحصل فعالیتهای تحقیقاتی است که برای ایجاد یک سیستم تصدیق هویت از طریق تلفن صورت گرفته است. روشهای استفاده شده با هدف بهبود کارائی، سرعت و سهولت در انجام کار انتخاب شده و آزمایشاتی برای تعیین توانمندی این روشها صورت گرفته است. لزوما این روشها، بهترین روشها نبوده و البته این افق برای کلیه محققین در این زمینه باز است که روشها و تکنیکهای دیگر را مورد بررسی قرار داده و با تلاش خود کارائی سیستم های تصدیق هویت گوینده را بهبود بخشند و از این طریق به امکان استفاده کارآمد تر از این سیستم ها در کاربردهائی که برای آنها متصور است یاری نمایند.

مراجع

[۱] س. م. احدی، ح شیخ‌زاده و م. همایون‌پور، *گزارش پیشرفت کار (۲) ـ طرح پژوهشی ملّی پردازش گفتار فارسی*، دی ۷۷ ـ خرداد ۷۸.

[۲] س. م. احدی، ح شیخ‌زاده و م. همایون‌پور، *گزارش پیشرفت کار (۳) ـ طرح پژوهشی ملّی پردازش گفتار فارسی*، تیر ۷۸ ـ بهمن ۷۸.

[۳] س. م. احدی، ح شیخ‌زاده و م. همایون‌پور، *گزارش پیشرفت کار (٤) ـ طرح پژوهشی ملّی پردازش گفتار فارسی*، اسفند ۷۸ـ شهریور ۷۹.

[٤] س. م. احدی، ح شیخ‌زاده و م. همایون‌پور، *گزارش پیشرفت کار (۵) ـ طرح پژوهشی ملّی پردازش گفتار فارسی*، مهر ۷۹ـ فروردین ۸۰.

[5] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, Vol. 77, No.2, Feb. 1989.

[6] L.R. Rabiner, B-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[7] L.E. Baum and J.A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to....

[8] A.Waibel *et al.*, "Phoneme Recognition Using Time-Delay Neural Networks", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 37, No.3. March 1989.

[9] H. Ney and X. Aubert, "Dynamic Programming Search Strategies: From Digit Strings to Large Vocabulary Word Graphs", in *Automatic Speech and Speaker Recognition: Advanced Topics*, Ed. C-H. Lee *et al.*, pp. 385- 411, Boston: Kluwer Academic Publishers, 1996.

[10] S.J. Young, "A Review of Large-Vocabulary Continuous Speech Recognition", *IEEE Signal Processing Magazine*, Vol. 13, No.5, pp.45- 57, Sep. 1996.

[۱۱] م. ر. میرحسینی و س. م. احدی، "معیار تصویر وزن‌دهی شده برای بازشناسی مقاوم گفتار فارسی"، *مجموعه مقالات هشتمین کنفرانس مهندسی برق ایران*، اصفهان، ۱۳۷۹.

[۱۲] کتابچهٔ استفاده از دادگان گفتاری زبان فارسی (فارس دات)، مرکز تحقیقات پردازش هوشمندعلائم.

[۱۳] م. پرویزی و س. م. احدی، "بازشناسی گفتار پیوستهٔ فارسی با دایرهٔ کلمات متوسط"، *مجموعه مقالات نهمین کنفرانس مهندسی برق ایران*، تهران، ۱۳۸۰.

[۱٤] م. شیرزاد، س. م. احدی، "هم‌ردیف‌سازی زمانی گفتار پیوستهٔ فارسی"، *مجموعه مقالات هشتمین کنفرانس مهندسی برق ایران*، اصفهان، ۱۳۷۹.

[15] S.M. Ahadi, "Continuous Persian Speech Recognition via Improved-MAP Estimated Context-Dependent Modeling", in *Proc. ICSP- 2001*, Daejon, Korea.

[16] S.M. Ahadi, "Reduced Context Sensitivity in Persian Speech Recognition via Syllable Modeling", in *Proc. SST- 2000*, Canberra, Australia.

[۱۷] بهزاد نظری، "پیش بینی خصوصیات آوایی زبان فارسی جهت کاربرد در یک سیستم تبدیل متن به گفتار"، رساله دکتری، دانشگاه صنعتی شریف.

[18] Kenneth N. Ross, Mari Ostendorf, "A Dynamic System Model for Generating Fundamental Frequency for Speech Synthesis" in *IEEE Transactions on Speech and Audio Processing,* Vol. 7, No. 3, May 1999.

[۱۹] علی اشکوری، "تحقیق بر روی قواعد نوای گفتار در فارسی و اعمال آن در یک گفتارساز پارامتری"، پایان نامهٔ کارشناسی ارشد، دانشکده برق دانشگاه صنعتی امیرکبیر، ۱۳۷٦.

[20] G. Fant, "Speech Communication Research", *IVA* ,24(8), 1953.

[21] L.R. Harris, "A Study of Speech Production", *JASA* , 25, 1953.

[22] D.H. Klatt, "Review of Text-To-Speech Conversion For English", *JASA* , 82, 1987.

[23] E. Moulines and F. Charpentier, "Pitch-Synchronous waveform processing techniques for Text-to-Speech synthesis using diphones", *Speech Communication,* 9, 1990.

[24] J. Makhoul, "A mixed source model for Speech compression and synthesis", *JASA,* 64, 1978.

[25] B.S. Atal and J.R. Remde, "A new model of LPC excitation for production of natural-sounding speech at low bit rates", *ICASSP'82*.

[26] P. Kroon and B.S. Atal, "Strategies for improving the performance of CELP coders at low bit rates", *ICASSP'88*.

[27] K. Hakoda, K. Kabeya, T. Hirahara and K. Nagakura, "Japanese text-to-speech synthesizer based on residual excited speech synthesis", *ICASSP'86*.

[28] R. Di Francisco and E. Moulines, "Detection of the glottal closure by jumps in the statistical properties of the signal", *Eurospeech'89*.

[29] T. Dutoit and H. Leich, "MBR-PSOLA: Text-To-Speech synthesis based on MBE re-synthesis of the segments database", *Speech Communication*, 13, pp. 435-440, 1993.

[30] D.W. Griffin and J.S. Lim, "Multiband Excitation Vocoder", *IEEE Trans. Acoust., Speech, Signal Processing* , Vol. 36, August 1988.

[۳۱] طرح ملی بازشناسی گفتار گسته و پیوسته فارسی، گزارش فروردین ۱۳۷۷ تا دی ۱۳۷۷ ، ۱۳۷۷.

[۳۲] فرامرز فکری، محمدرضا نخعی، محمود تبیانی، شناسایی صحبت توسط کامپیوتر، دانشگاه صنعتی شریف، دانشکده مهندسی برق، پایان نامه کارشناسی ارشد، اسفندماه ۱۳۷۱.

[۳۳] حسن بابابیک، "بازشناسی گفتار با استفاده از تلفیق مدل مخفی مارکف و شبکه عصبی"، *هفتمین کنفرانس مهندسی برق ایران*، تهران، مرکز تحقیقات مخابرات ایران، اردیبهشت ۱۳۷۸.

[۳٤] شیوا رستم زاده، سید محمد احدی، حمید شیخ زاده نجار، "بازشناسی گفتار فارسی ناپیوسته بصورت ناوابسته به گوینده به کمک مدلهای پنهان مارکف با چگالی پیوسته"، *ششمین کنفرانس مهندسی برق ایران*، تهران، دانشگاه صنعتی خواجه نصیرالدین طوسی، دانشکده مهندسی برق، اردیبهشت ۱۳۷۷.

[۳۵] سعید بابایی زاده، ایمان غلامپور، کامبیز نایبی، "بهبود کارایی سیستم های بازشناسی گفتار گسسته با ترکیب شبکه های عصبی و مدل های مارکف پنهان"، *هفتمین کنفرانس مهندسی برق ایران*، تهران، مرکز تحقیقات مخابرات ایران، اردیبهشت ۱۳۷۸.

[36] L. F. Lamel, I. L. Gauvain, "Speaker Verification Over Telephone", *Speech communications,* 31 (2-3), pp. 141-154, 2000.

[37] C. P. Lim, et al., "Speech Recognition using Artificial Neural Networks" *Proceedings of First Conf. on web Information Engineering,* Vol. 1, pp. 419-423, 2000.

[38] W. Siew Chan, L.Chee Peng, R. Osman, "Text-Dependent speaker Recognition using the Fuzzy ARTMAP Neural Network", *Proceeding of Intelligent System Technology,* vol. 1, pp. 33-38, 2000.
[39] G. doddington, "A Method for Speaker Verification", *JASA*, pp. 49, 1974.
[40] K. P. Li, "Experimental Studies in speaker Verification Using an Adaptive System", *JASA*, pp. 40. 1966.
[41] F. K. S. soong, "A. E. Rosenberg, L. R. Rabiner, B. H. Juang, "A Vector Quantization approach to speaker Recognition", *ICASSP-85,* pp. 387-390, 1985.
[42] A. E. Rosenberg, C. H. Lee, S. Gokoen, "Connected Word Talker Verification Using Whole Hidden Markov Model", *ICASSP-91,* pp. 381-384.
[43] S. Furui, " Comparison of Speaker Recognition Methods Using Static Features and dynamic features" *IEEE Transaction on ASSP,* 29 (3), pp. 342-350, 1981.
[44] C. Bernasconi, " On Instantaneous and Transitional Spectral Information fpr Text-Dependent Speaker Verification. *Sppech communication,* 9 (2), pp. 129-130, 1990.
[45] M. M. Homayounpour, Comparison of Some Relevent Parametric Representations for Speaker Verification", *ESCA Workshop on speaker Verification, Idenrtification, and Recognition,* pp. 185-188, 1994.
[46] K. R. Farrell, R. Mammone, K. T. Assaleh, "Speaker Recognition Using Neural Tree Networks and Conventional Classifiers", *IEEE Trans. On Speech and Audio Processing,* 2(1), pp. 194-204,1994.
[47] D. Reynolds, "Speaker Identification and Verification using Gaussiau Mixture Speaker Models", *ESCA Workshop On Automatic Speaker Verification, Identification, and Recognition,* pp. 27-30, 1994.
[48] M. Carrey, *et al.,* "A speaker Verification system Using Alpha Nets", *ICASSP-91,* pp. 397-400, 1991.
[49] M. Mehdi Homayounpour, et al., "Neural Net Approaches to Speaker Verification: Comparison with Second Order Statistical Measures", *ICASSP-95,* 1995.
[۵۰] م. ر. ذهابی، ا. ا. سپهری، "استفاده از تصمیم گیرنده‌های باینری در مدل مخفی مارکُف برای شناسایی گوینده "، *دومین کنفرانس مهندسی برق ایران*، صص. ۳۶۱-۳۶۷، دانشگاه تربیت مدرس، تهران، ایران، ۱۳۷۳ .
[۵۱] م. مندولکانی، م. لطفی‌زاد، "تشخیص هویت گوینده توسط کامپیوتر"، *دومین کنفرانس مهندسی برق ایران*، صص. ۳۵۳-۳۶۰، دانشگاه تربیت مدرس، تهران، ایران، ۱۳۷۳.
[۵۲] ح. اصغری، م. ر. عارف، "بازشناسی گوینده با تحقق چندی‌کننده‌های برداری"، *دومین کنفرانس مهندسی برق ایران*، صص. ۳۳۵-۳۴۴، دانشگاه تربیت مدرس، تهران، ایران، ۱۳۷۳.
[۵۳] م. ص. حدائق، م. لطفی‌زاد، "تصدیق هویت گوینده توسط کامپیوتر "، *دومین کنفرانس مهندسی برق ایران*، صص. ۲۱۲-۲۲۱، دانشگاه تربیت مدرس، تهران، ایران، ۱۳۷۳.
[۵۴] ا. صیادیان، ح. غفوری‌فرد، "استفاده از تغییرات دینامیکی ضرایب LSPF جهت کاهش خطای سیستم‌های بازشناسی گوینده"، *سومین کنفرانس مهندسی برق ایران*، صص. ۲۰۷-۲۱۲، دانشگاه علم و صنعت ایران، تهران، ایران، ۱۳۷۴.
[۵۵] ح. مقصودلو، م. ر. نخعی، م. تبیانی، "سیستم تأیید هویت گوینده وابسته به متن با استفاده از روش کوانتیزاسیون برداری"، *سومین کنفرانس مهندسی برق ایران*، صص. ۱۵۵-۱۶۲، دانشگاه علم و صنعت ایران، تهران، ایران، ۱۳۷۴.

[٥٦] س. ذ. فیض‌آبادی، س. صدوقی، "سیستم تشخیص گوینده"، *ششمین کنفرانس مهندسی برق ایران*، صص. ٣٦٩-٣٧٢، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران، ١٣٧٧.

[٥٧] ا. صیادیان، ک. بدیع، م. حکاک، م. ر. بیک زاده، "ارائهٔ روش TSD-PGMM در بازشناسی گوینده مستقل از متن"، *هشتمین کنفرانس مهندسی برق ایران*، صص. ٣٧٦-٣٨٢، دانشگاه صنعتی اصفهان، اصفهان، ایران، ١٣٧٩.

[٥٨] محمدمهدی همایون‌پور، حسین بشیری، جهانشاه کبودیان، "استفاده از روش بازتخمین نهفته در مدل پنهان مارکوف برای بازشناسی ارقام متصل فارسی بر روی خط تلفن بطور مستقل از گوینده"، *نهمین کنفرانس برق ایران*، ١٣٨٠.

[59] B. V. Williams, R. T. J. Bostock, D. Bounds, and A. Harget, "Improving Classification Performance in the Bumptree Network by Optimizing Topology with a Genetic Algorithm", *First Conference on Genetic Algorithms*, I, pp. 490-495, 1994.

[60] S. M. Omahandro, "Bumptree for Efficient Function, Constraint, and Classification Learning", *Advances in Neural Information Processing System*, 3, Morgan Kaufmann, 1991.

[61] D. Mansour et al. , "A Family of Distortion Measures based upon Projection Operation for Robust Speech Recognition", *IEEE Trans. on ASSP*, Vol. 37, No. 11, Nov. 1989.

[62] B. A. Carlson et al. , "A Projection-based Likelihood Measure for Speech Recognition in Noise", *IEEE Trans. SAP*, Vol. 2, No. 1, Jan. 1994.

[٦٣] م. ر. میرحسینی، س. م. احدی، "معیار تصویر وزن‌دهی شده برای بازشناسی مقاوم گفتار فارسی"، *کنفرانس مهندسی برق ایران*، دانشگاه صنعتی اصفهان، اصفهان، ایران، ١٣٧٩.

[64] T. Matsui, S. Furui, "Similarity Normalization Method for Speaker Verification based on a Posteriori Probability", *ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*, pp. 59-62, Martigny, Switzerland, 1994 .

[65] F. Chen et al., "Hybrid Threshold Approach in Text-Independent Speaker Verification", *ICSLP-94*, pp. 1855-1858, Yokohama, Japan, 1994 .

[66] A. E. Rosenberg et al., "The Use of Cohort Normalized Scores for Speaker Verification", *ICSLP-92*, pp. 599-602, Banff, Canada, 1992 .

[67] K. P. Markov et al., "Text-Independent Speaker Recognition using Non-linear Frame Likelihood Transformation", *Speech Communication*, Vol. 24, pp. 193-209, 1998 .

[٦٨] ح. بشیری، شناسایی شمارهٔ شخصی گفتاری متصل بر روی خط تلفن با استفاده از مدل پنهان مارکُف، پایان‌نامهٔ کارشناسی ارشد، دانشکدهٔ مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، ١٣٧٩.

[69] J. Penack, D. Nelson, "The NP Speech Activity Detection Algorithm", *ICASSP-95*, Vol. 1, pp. 381-384, Detroit, USA, May 1995.

[70] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction", *IEEE trans. on ASSP*, Vol. ASSP-27, No. 2, pp. 113-120, Apr. 1979.

[71] T. Matsui, S. Furui, "Similarity Normalization Method for Speaker Verification based on a Posteriori Probability", ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification, pp. 59-62, Martigny, Switzerland, 1994 .

[٧٢] محمد مهدی همایونپور، امیر نجاری، "بازشناسی ارقام ناوابسته به گوینده با استفاده از مدل پیشگوی عصبی" ، *هفتمین کنفرانس مهندسی برق ایران*، صفحه ٨١-٧٥، ١٣٧٨.

[۷۳] محمد مهدی همایونپور، امیر نجاری، "تصدیق هویت گوینده توسط تلفیق شبکه های عصبی و الگوریتم ژنتیکی" ، *پنجمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران*، صفحه ۲٦٤-۲۵۷، ۱۳۷۸.

[۷٤] جهانشاه کبودیان، حسین بشیری، محمدمهدی همایونپور، "بازشناسی ارقام مجزای فارسی بر روی خط تلفن بطور مستقل از گوینده و با استفاده از مدل پنهان مارکف پیوسته"، *هفتمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران*، ۱۳۸۰.

[۷۵] محمدمهدی همایونپور، جهانشاه کبودیان، "تعیین و تصدیق هویت گوینده بر روی خط تلفن با استفاده از یک سیستم هیبرید متشکل از مدل پنهان مارکوف و مدل مخلوط گوسی"، *هفتمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران*، ۱۳۸۰.

[۷٦] محمدمهدی همایونپور، جهانشاه کبودیان، "مقایسهٔ چند روشِ نرمالیزاسیونِ امتیازات در سطحِ گویش و درسطحِ فریم برای افزایشِ کارآییِ سیستم‌های تصدیق و تعیین هویت گوینده بر روی خط تلفن"، *نهمین کنفرانس برق ایران*، ۱۳۸۰.

[۷۷] محمد مهدی همایونپور، امیر نجاری، "بازشناسی کد شناسائی شخصی و و تصدیق هویت گوینده به منظور کنترل دسترسی از راه دور توسط تلفن " ، *نشریه علمی امیرکبیر*.

پیوست الف – جداول نشانه های بخش سنتز گفتار

جدول نشانه های قراردادی واجنویسی

واکه‌ها	
نشانه	صدا یا حرف
@	ـَ
e	ـِ
o	ـُ
a	آ
i	او
u	ای

نیمه واکه‌ها	
v	و
y	ی

همخوانها	
نشانه	صدا یا حرف
?	ا ، ع
b	ب
p	پ
t	ت
s	ث ، س ، ص
j	ج
c	چ
h	ح ، ه
x	خ
d	د
z	ذ ، ز ، ض ، ظ
r	ر
#	ژ
$	ش
q	غ ، ق
f	ف
k	ک
g	گ
l	ل
m	م
n	ن

جدول نشانه‌های انواع صرفی

نوع	علامت
اسم	n
ضمیر ساده	e
ضمیر مرکب	f
ضمیر پرسشی	g
صفت	j
قید پرسش	b
قید زمان	h
قید مکان	i
قید تأکید (نفی)	k
قید تردید	l
قید حالت	*
قید مقدار	%
صوت تحسین	o
صوت تحسیر	p
صوت تعجب	q

نوع	علامت
صوت تنبیه	r
صوت تحذیر	t
وندها	s
عدد	#
حرف اضافه	a
حرف ربطی	c
حرف نشانه	d
مصدر	m
بن ماضی متعدی	v
بن ماضی لازم	u
بن مضارع متعدی	y
بن مضارع لازم	x
بن ماضی ربطی	w
بن مضارع ربطی	z
ناشناس	?

جدول نشانه‌های نقشهای نحوی

نقش	نشانه
اسم	e
فعل	f
فعل ربط	r
حرف اضافه	h
حرف ربط	H
فاعل	F
مفعول	m
ضمیر	z
صفت	s
قید	q
تمیز	t
مسند	m
مسند الیه	n
مضاف	M
مضاف الیه	N
عدد	#