بازشناسی اعداد پیوسته از طریق خط تلفن

Telephony Continuous Digit Recognition

ارائه دهنده: امین فاضل **afazel@ce.sharif.edu**

استاد راهنما: دکتر محمد تقی منظوری **manzuri@sharif.edu**

استاد مشاور: دکتر حسین ثامتی **sameti@sharif.ed**

چکیده

سیستمهای بازشناسی گفتار امروزی در شرایط گفتار تمیز وآزمایشگاهی توانسته‌اند به جوابهای خوبی نایل‌آیند با این وجود هنگامیکه سیگنال گفتار به وسیله شبکه تلفن خراب می‌شود درصد دقت و صحت بازشناسی کاهش چشمگیری پیدا می‌کند، همچنین با تغییر تعداد واژگان مورد بازشناسی تغییراتی در سیستم بازشناسی گفتار باید صورت گیرد تا بتوان نتایج قابل قبولی دریافت کرد. در این مقاله به بررسی برخی از روشهای بازشناسی مقاوم به نویز به خصوص نویز خط تلفن پرداخته می‌شود و سپس تغییرات لازم در سیستم موجود برای بازشناسی اعداد پیوسته بیان می‌شود.

کلیدواژه ها: بازشناسی گفتار پیوسته، بازشناسی اعداد، بازشناسی از طریق خط تلفن، بازشناسی مقاوم گفتار

۱- مقدمه

در یک سیستم بازشناسی گفتار پارامترهای مختلفی تعیین کننده درجه توانایی سیستم بازشناسی می‌باشند ولی از یک دید کلی می‌توان آنها را به دسته‌های زیر تقسیم کرد[۵].

بازشناسی کلمات گسسته[۱] وابسته به گوینده و مستقل از گوینده. در این سطح از بازشناسی یک دسته از برنامه‌های کاربردی به نام "دستور و کنترل" قرار دارند که این سیستمها توانایی بازشناسی یک دستور یک کلمه‌ای از یک دادگان با تعداد لغت کم را دارند. یکی از مسایل مهم در این گونه سیستمها حساسیت آنها به نویز محیط می‌باشد به گونه‌ای که این سیستمها توانایی کار در محیطهای نویزی را داشته باشند.

بازشناسی کلمات متصل[۲] وابسته به گوینده و مستقل از گوینده. در این سطح از بازشناسی سیستمهایی قرار می‌گیرند که توانایی خوبی در بازشناسی کلمات گسسته را داشته باشند و همچنین توانایی آن را نیز داشته باشد که اگر این کلمات پشت سر هم نیز گفته شوند سیتم توانایی بازشناسی آنها را داشته باشد. در این سطح از تکنولوژی بازشناسی گفتار دسته‌ای از برنامه‌های کاربردی قرار دارند که مبنای آنها بازشناسی رشته‌ای از اعداد یا رشته‌ای از اعداد و حروف با یکدیگر می‌باشند و کاربردهای عملی مانند سیستمهای شماره‌گیری صوتی[۳]، تأیید هویت کارتهای اعتباری[۴] و سفارش گاتالوگ[۵] دارند.

بازشناسی پیوسته گفتار[۶] یا بازشناسی گفتار روان[۷] وابسته به گوینده و مستقل از گوینده. در این سطح معمولاً دادگان با واژگان بزرگ قرار دارند و سیتمهای عملی مانند دیکته برای کامپیوتر وجود دارد.

سیتمهای فهم گفتار که توانایی که توانایی فهم پیامهای خاص موجود در گفتار را علاوه بر تشخیص کلمات گفتار را دارا می‌باشند.

سیتمهای مکالمات فی‌البداهه که توانایی تشخیص گفتار به صورت دقیق و همچنین فهم تمام گفتار را دارا می‌باشند. از کاربردهای عملی این سیستمها می‌توان به خلاصه کردن مکالمات اشاره کرد.

پروژه حاضر در واقع بازشناسی پیوسته می‌باشد ولی همچنان که در دسته‌بندی‌های فوق بیان شد این گونه سیستمها معمولاً در محدودهٔ لغات زیاد مورد استفاده قرار می‌گیرند و اگر بخواهد در سطح دادگان با تعداد لغات کم سیستمی طراحی شود معمولاً از سیتمهای با کلمات متصل استفاده می‌شود و با این کار گوینده را مجبور به رعایت پارهای از محدودیت‌ها می‌کنند. در این پروژه بازشناسی پیوسته در سطح تعداد لغات کم (اعداد بیان شده به صورت طبیعی و پیوسته) می‌باشد و همچنین برای آنکه پروژه جنبه عملی نیز پیدا کند سیستم باید توانایی بازشناسی در حضور خط تلفن را داشته باشد و فرض می‌شود که فقط نویز حاصل از خط تلفن بر روی گفتار وجود داشته باشد.

در این مقاله سعی می‌شود تفاوت بین بازشناسی خودکار گفتار تلفنی با گفتار بدون نویز که در محیطی قابل کنترل تهیه شده‌است بیان گردد، همچنین تفاوتهای بازشناسی گفتار با دادگان بزرگ و دادگان کوچک نیز بررسی گردد. بخش دوم مقاله به معرفی سیستم بازشناسی گفتار موجود می‌پردازد. در بخش سوم به معرفی دادگان مورد استفاده پرداخته خواهد شد. در بخش

---

[1] *Isolated Word Recognition*
[2] *Connected Word Recognition*
[3] *Voice Dialing*
[4] *Credit Card Authorization*
[5] *Catalog Ordering*
[6] *Continuous Speech Recohnition*
[7] *Fluent Speech Recognition*

چهارم شبکه خط تلفن مورد بررسی قرار می‌گیرد. بخش پنجم برخی از روشهای موجود برای بازشناسی مقاوم گفتار را معرفی‌می‌کند. بخش ششم هم مسألهٔ بازشناسی گفتار را از دید تعداد واژگان مورد بررسی قرار می‌دهد و در بخش هفتم زمانبندی ادامه پروژه ارائه خواهد شد.

۲- معرفی سیستم موجود

در این بخش، در مورد سیستم بازشناسی موجود و پارامترهای مختلف آن توضیح مختصری داده خواهد شد. این سیستم، برای مدل‌سازی واحدهای آوایی و انجام بازشناسی گفتار پیوسته، از سیستم بازشناسی گفتار پیوسته که در [۱] شرح داده شده، استفاده گردیده است. سیستم بازشناسی مذکور، مدل‌سازی واحدهای آوایی را با استفاده از مدل مخفی مارکوف[1] انجام می‌دهد. مدلهای مخفی مارکوف به کاربرده شده از نوع چگالی پیوسته و با تلفیق‌های گوسی می‌باشند و پرش بین حالتهای آن فقط به صورت چپ به راست مجاز می‌باشد. توپولوژی مدلهای مربوط به هر واحد آوایی را در این سیستم می‌توان، جداگانه تعیین کرد. تعیین توپولوژی شامل تعیین تعداد حالت، تعداد تلفیق در هر حالت و تعیین پرشهای مجاز بین حالتها می‌باشد. با استفاده از این مدلها سیستم مذکور تطبیق الگو را برای بازشناسی سیگنال گفتار ورودی انجام می‌دهد و برای این منظور از روشهای مختلف جستجو برای یافتن بهترین دنبالهٔ واحدهای آوایی متناظر با دنبالهٔ آکوستیکی ورودی بهره می‌گیرد.

زبان مورد بازشناسی، مستقل یا وابسته بودن نسبت به گوینده و اندازهٔ مجموعهٔ واژگان مورد بازشناسی در این سیستم، همگی به دادگان گفتار مورد استفاده و مقدار و نوع داده‌های آموزشی بستگی دارند. با استفاده از دادگان گفتاری بزرگ و آموزش مدل‌ها با دادهٔ آموزشی کافی و متنوع از گوینده‌های مختلف، می‌توان این سیستم را برای بازشناسی مستقل از گوینده و با واژگان بزرگ به کار برد.

واحد آوایی به کار برده شده، واج می‌باشد یعنی به ازای هر واج زبان یک مدل مخفی مارکوف آموزش داده می‌شود، بنابراین ورودی سیستم یک سیگنال گفتار و خروجی سیستم، بهترین دنبالهٔ واجی یافت شده متناظر با دنبالهٔ آکوستیک ورودی می‌باشد.

بردارهای ویژگی استخراج شده از فریم های گفتار، شامل ۱۲ ضریب مل-کپستروم همراه با مشتقات زمانی اول و دوم آنها می‌باشد. مدل‌های مخفی مارکوف با استفاده از الگوریتم *segmental k-means* آموزش داده شده‌اند و توپولوژی آنها برای تمام واج ها یکسان در نظر گرفته شده است. در مرحلهٔ تطبیق الگو نیز از الگوریتم جستجوی «ویتربی شعاعی»[2] برای یافتن بهترین دنبالهٔ واجی استفاده شده است.

۳- دادگان گفتار

در این قسمت به معرفی دادگان موجود مورد استفاده در این زمینه پرداخته می‌شود. از آنجایی که کارایی سیستمهای بازشناسی گفتار بسیار به دادگان آموزشی وابسته است، انتخاب یک دادگان خوب در این زمینه می‌تواند نتایج بسیار مطلوبی هنگام استفاده به شکل عملی داشته باشد.

---

[1] *Hidden Markov Model*
[2] *viterbi*

۳-۱- دادگان فارس‌دات تلفنی

این دادگان توسط مرکزپردازش هوشمند علایم در سال ۱۳۸۰تهیه شده است[۶]. این دادگان در سطح واج تقطیع شده است و دارای تقریباً هشت ساعت گفتار از ۶۴ گوینده می‌باشد. این دادگان هم در بر دارندهٔ گفتار فی‌البداهه و هم گفتار رسمی می‌باشد. از آنجایی که این دادگان در بر دارندهٔ گفتار فی‌البداهه نیز می‌باشد هنگام برچسب گذاری از برچسبهای غیرزبانی نیز استفاده شده‌است تا بتوان آن قسمتهایی از گفتار را که به زبان مربوط نمی‌باشند نیز برچسب گذاری کرد.

گفتار فی‌البداهه این دادگان شامل سلام و احوالپرسی، نام و نام‌خانوادگی، میزان تحصیلات، سن، نام شهر یا استان محل زندگی، وضعیت اتاقی که در آن مستقر هستند، توضیح در مورد شهری که در آن مستقر هستند و یک خاطره تلخ یا شیرین می‌باشد و دادگان گفتار رسمی آن نیز شامل اعداد (صفر، یک، ...، بیست، سی،...، صد، دویست، پانصد، پونصد، هزار، میلیون و میلیارد)، ایام هفته، ماههای سال، حروف الفبا، پنجاه کلمه پر بسامد زبان فارسی به طور تصادفی، دو جمله ثابت، چهار جمله تصادفی ازدادگان فارس‌دات و هجاهای باز زبان فارسی.

از آنجایی که این دادگان شامل اعداد پیوسته زبان فارسی نمی‌باشد نمی‌توان از آن برای آزمایش اعداد پیوسته فارسی استفاده کرد به همین علت نسخه‌های مختلفی از این دادگان تهیه گردید تا بتوان آزمایشهایی جهت روشهای مختلف مقاوم به نویز به وسیله آنها انجام داد.

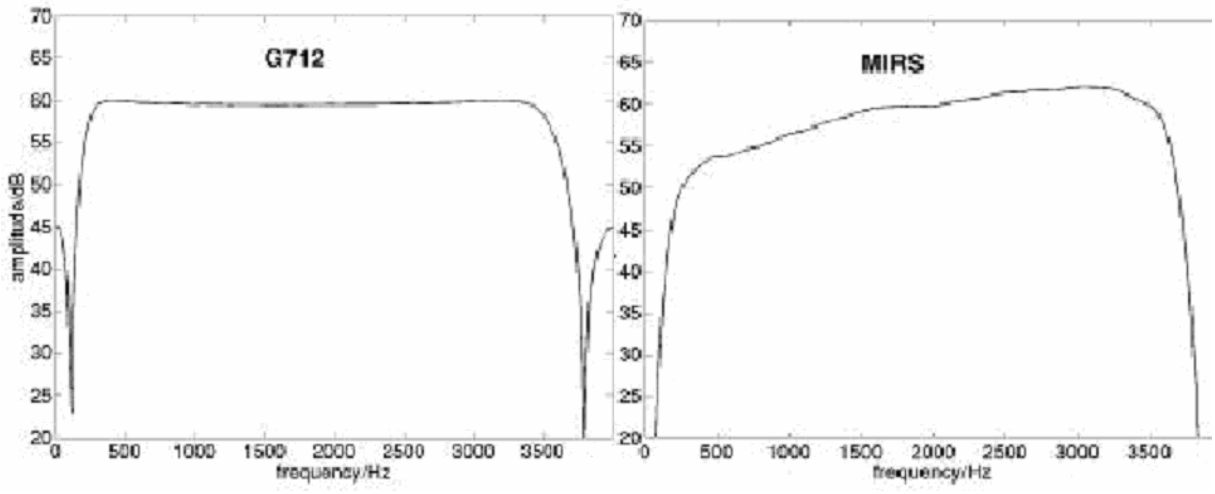
۳-۲- دادگان پروژه *AURORA 2*

این دادگان برای بررسی میزان دقت سیستمهای بازشناسی صحبت در محیطهای نویزی تهیه شده است و مورد استفادهٔ بیشتر از آن برای بررسی الگوریتمهای استخراج ویژگی مقاوم به نویز می‌باشد[۷].

دادگان گفتار اصلی آن دادگان *TIDigits* می‌باشد که تشکیل شده‌است از اعداد متصل که به وسیله گوینده‌های انگلیسی زبان آمریکایی گفته شده‌اند. در این دادگان ابتدا *TIDigits* به ۸ کیلو هرتز زیرنمونه‌برداری[1] شده است و پس از آن یک فیلتر دیگر جهت شبیه‌سازی محیط خط تلفن بر روی دادگان اعمال گردیده است. این فیلتر یکی از دو فیلتر *G.712* و*MIRS* می‌باشد که در شکل(۱) نشان داده شده‌اند. گفتار حاصل از اعمال فیلترهای فوق در این دادگان گفتار تمیز نامیده می‌شوند. پس از آن نویزهایی با نسبت سیگنال به نویز[2]های مختلف به دادگان اضافه گردیده است، این نویزها عبارتند از نویز حاصل از قطار برون‌شهری، نویز همهمه[3]، نویز ماشین، نویز نمایشگاه، نویز رستوران، نویز خیابان، نویز فرودگاه و نویز ایستگاه قطار.

---

1 Downsample
2 Signal to Noise Ratio (SNR)
3 Crowd of people(babble)



شکل(۱)

پاسخ فرکانسی فیلترهای G.712 و MIRS

دادگان فوق دارای دو نوع داده آموزشی می‌باشد:

- آموزش به وسیله دادگان تمیز
- آموزش به وسیله دادگان تمیز و نویزی[1]

در نوع اول دادگان آموزش، ۸۴۴۰ جمله جهت آموزش استفاده شده است که به وسیلهٔ فیلتر *G.712* فیلتر شده‌اند. در دومین دادگان آموزش نیز همین ۸۴۴۰ جمله به بیست زیر مجموعهٔ ۴۲۲ جمله‌ای دسته‌بندی شده‌اند که این بیست زیر مجموعه در بر گیرندهٔ چهار نویز ماشین، قطار برون شهری، همهمه و نمایشگاه در پنج نسبت سیگنال به نویز مختلف ۵ ، ۱۰ ، ۱۵ و ۲۰ دسی‌بل می‌باشند. نویزها و گفتار این دسته آموزش نیز به وسیلهٔ فیلتر *G.712* فیلتر شده‌اند.

دادگان آزمایش *A* شامل ۴۰۰۴ جمله می‌باشد که به چهار زیر مجموعه ۱۰۰۱ جمله‌ای تقسیم‌بندی شده‌اند. به هر یک از این زیر مجموعه‌ها نویزهای ماشین، قطار برون شهری، همهمه و نمایشگاه اضافه شده‌اند و هر دسته نویزی حاصل در نسبتهای مختلف سیگنال به نویز ۵–، ۰، ۵، ۱۰، ۱۵و ۲۰ دسی‌بل وجود دارند. همچنین به عنوان حالت هفتم، نوع بدون نویز یکی از زیرمجموعه‌های فوق به‌عنوان حالت تمیز در این دادهٔ آزمایش موجود می‌باشد. در اینجا نیز از فیلتر *G.712* هم برای داده‌های گفتار و هم برای داده‌های نویز قبل از اضافه کردن آنها به یکدیگر استفاده می‌شود.

دادگان آزمایش *B* نیز دقیقاً شبیه به روش فوق به‌دست آمده‌اند با این تفاوت که در اینجا از نویزهای رستوران، خیابان، فرودگاه و نویز ایستگاه قطار استفاده شده است. در این حالت عدم انطباق بین دادگان آموزش و آزمایش حتی برای حالت آموزش به وسیلهٔ دادگان تمیز و نویزی هم وجود دارد.

در دادگان آزمایش *C* دو زیر مجموعه از چهار زیر مجموعهٔ ۱۰۰۱ جمله‌ای وجود دارند. در اینجا داده‌های گفتار و نویز به وسیلهٔ فیلتر *MIRS*، فیلتر شده‌اند و سپس نویز با نسبت سیگنال به نویزهای ۵–، ۰، ۵، ۱۰، ۱۵و۲۰ به گفتار اضافه شده است. در اینجا نیز حالت تمیز یعنی بدون حضور نویز جمع‌شونده نیز وجود دارد. نویزهای قطار برون شهری و خیابان به‌عنوان نویز جمع‌شونده انتخاب شده‌اند. این دسته از دادگان آزمایش برای نشان‌دادن اثر نویز کانال‌های متفاوت بر روی سیستم تشخیص گفتار، تهیه شده است.

---

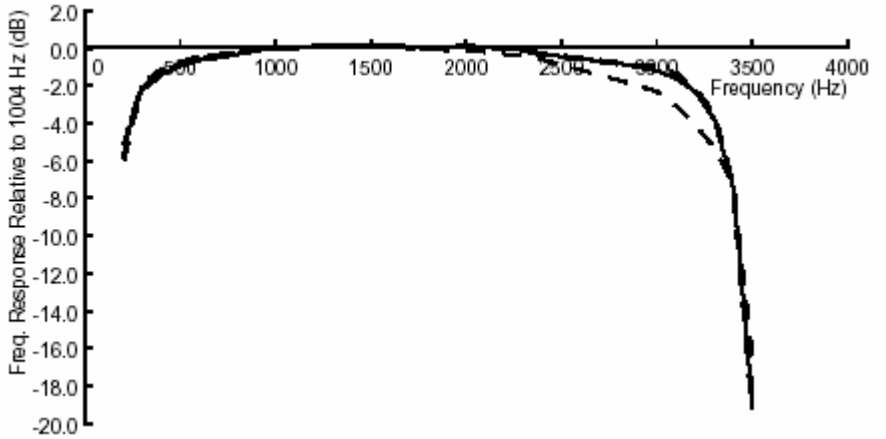
1 Multi Condition Training

۴- بررسی شبکه خط تلفن

در این قسمت شبکه تلفن با توجه به پاسخ فرکانسی، نسبت سیگنال به نویز و غیر خطی بودن آن مورد بررسی قرار می‌گیرد[۴]. اطلاعات موجود در اینجا بیشتر در مورد خط تلفن می‌باشد یعنی ویژگیهای مربوط به میکرفن تلفن در نظر گرفته نشده است.

۴-۱ پاسخ فرکانسی

پاسخ کانال یا پاسخ فرکانسی، یک اندازه‌گیری برای مقدار داده از بین رفته به علت باند فرکانسی در کانالهای ارتباطی می‌باشد. شکل(۲) پاسخ کانال در فاصله‌های کوتاه، متوسط و بلند را نشان می‌دهد. به طوری که فاصلهٔ کوتاه خطوط ارتباطی با طول بین صفر تا ۱۸۰ مایل، خطوط متوسط خطوطی با طول ۱۸۰ تا ۲۷۰ مایل و خطوط بلند، خطوطی با طولهای بالاتر از ۲۷۰ مایل در نظر گرفته شده‌اند. طولهای بیان شده، فاصلهٔ نقطه به نقطه بین مکانهای ارتباطی می‌باشد و فاصلهٔ واقعی که سیگنال طی می‌کند، می‌تواند متفاوت باشد.



شکل(۲)

پاسخ کانال خط تلفن برای فاصله‌های کوتاه، متوسط و بلند

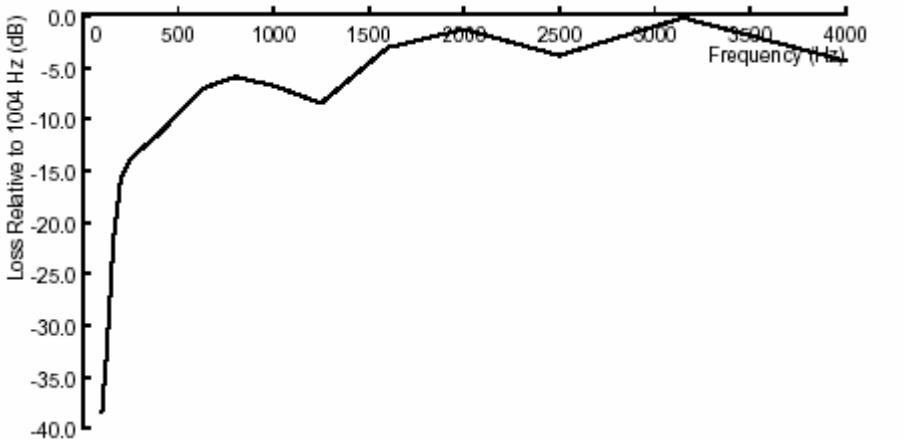
همانطوری که در شکل فوق مشاهده می‌گردد. بدترین پاسخ فرکانسی مربوط به فاصله‌های کم می‌باشد و این به علت آن است که تیم آقای *Garey* [۸] برای فاصله های کوتاه از واسطه‌های ارتباطی آنالوگ استفاده کرده‌اند. در حالی که برای فاصله‌های متوسط و بلند از اتصال‌دهنده‌های *T1* دیجیتال استفاده شده است. هنگامیکه اثر میکرفن نیز در نظر گرفته شود پاسخ فرکانسی تقریباً مشابه پاسخ فرکانسی فوق است ولی یک تضعیف بیشتری در فرکانسهای بالا خواهیم‌داشت.

۴-۲ اثر میکرفن

باید در نظر داشت که در شکلهای قبلی اثر میکرفن در نظر گرفته نشده است. میکرفن‌های جدید مانند میکرفن‌های خازنی پاسخ فرکانسی ثابت و تقریباً سطحی دارند با وجود این میکرفن‌های کربنی[1] که هنوز نیز در شبکه‌های تلفن موجود میباشد دارای پاسخ فرکانسی غیر مسطح می‌باشند و همچنین اثرات غیر خطی بر روی سیگنال می‌گذارند. این اثرات غیر خطی

1 Carbon Microphone

شامل حساسیت‌های متفاوت نسبت به سطح سیگنالهای متفاوت و پاسخ‌های متفاوت با توجه به جریانهای قطبی می‌باشد. یک تابع نوعی برای یک میکرفن کربنی در شکل(۳) نشان داده شده‌است.



شکل(۳)

حساسیت پاسخ فرکانسی برای یک میکرفن نوعی

۳-۴ اثرات غیر خطی

شبکه خط تلفن یک شبکه کاملاً خطی نمی‌باشد. این اثرات غیر خطی معمولاً به عنوان اعوجاج[1] در خط تلفن شناخته می‌شوند. با توجه به مقالات در این زمینه دو نوع اعوجاج در خط تلفن وجود دارد که عبارتند از اعوجاج هارمونیکی[2] و اعوجاج موجود در مدولاسین[3]. اولین اعوجاج مربوط به به وجود آمدن مؤلفهٔ فرکانسهای ناخواسته در هارمونیک‌های سیگنال ورودی می‌باشد و دومین اعوجاج هنگامی رخ می‌دهد که دو یا چند مؤلفهٔ فرکانسی در کانال موجود باشند و بر یکدیگر اثر بگذارند.

۵- بازشناسی مقاوم گفتار

امروزه، سیستم‌های بازشناسی گفتار به وفور در محیط‌های مختلف مورد استفاده قرار می‌گیرد. این سیستمها، با وجود اینکه در شرایط آزمایشگاهی، خوب عمل می‌کنند ولی وقتی که سراغ شرایط واقعی می‌رویم، کارایی آنها به شدت کاهش می‌یابد.آزمایش‌های متعدد انجام شده، نشان می‌دهند که حتی سیستم‌هایی که در شرائط آزمایشگاهی، خیلی خوب عمل می‌نمایند، در برخی از حالت‌های غیر آزمایشگاهی، از دستیابی به حداقل کارایی مورد قبول نیز ناتوانند[۹].

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

---

1 Distortion
2 Harmonic Distortion
3 Modulation Distortion

منابع زیادی برای تغییر گفتار از حالت آزمایشگاهی وجود دارند که ما در یک دید کلی، همهٔ آنها را منابع «نویز» می‌نامیم. معمولاً محیط در حالت آموزش، محیطی نسبتاً ایده‌آل و تا حد امکان، عاری از نویز می‌باشد، حال آنکه در مراحل آزمایش، انواع و اقسام نویزها بر عملکرد سیستم تأثیر می‌گذارند. برخی از این نویزها، حالت جمعی یا جمع شونده[1] و برخی دیگر حالت پیچشی[2] دارند. یک دسته از نویزهای مؤثر بر سیستم‌های بازشناسی، نویزهایی کوتاه مدت و غیر ایستان می‌باشند. صدای عبور اتومبیل، بسته شدن درب و همچنین گفتار گوینده‌ای غیر از گویندهٔ اصلی در این دسته از نویزها قرار می‌گیرند. دستهٔ دیگر، نویزهای پیوسته در زمان همچون صدای کولر و فن می‌باشند. انعکاس و طنین گفتار در محیط بسته مانند اتاق، خانوادهٔ دیگری از نویزها را به نام نویزهای پیچشی شکل می‌دهند. مشکلات و ضعف‌های سیستم انتقال و میکروفن و پهنای باند محدود کانال انتقال، هر یک موجب ایجاد نویزهای خاص خود می‌شوند البته در یک نگرش کلی‌تر می توان پهنای باند محدود را در خانوادهٔ نویزهای پیچشی قرار داد و بالاخره، باید از اثر لمبارد[3] یاد کرد. این اثر دارای این مفهوم است که در محیط‌های نویزی، خود گوینده هم حالت طبیعی حرف‌زدن را نداشته و تکیه‌ها و تاکیدهای متفاوت با حالت طبیعی، گفتار مورد نظر را ادا می‌کند[۲].

برای حل مسأله مقاوم سازی نسبت به نویز باید به دنبال دستیابی همزمان به دو هدف زیر باشیم:[۹]

- کارایی سیستم با تغیراتی چون کانال، محیط، گوینده در حد قابل قبولی باشد.
- تا جایی که امکان دارد، عملکرد سیستم بازشناسی و وابستگی آن به مجموعهٔ عظیمی از داده‌های آموزشی نجات‌یابد.

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

$$y(t)=\left[\left\{\left(\left[s(t)\Big|_{n_1(t)}^{\text{Stress}}_{\text{Lombard}}\right]+n_1(t)\right)*h_{\text{mic}}(t)+n_2(t)\right\}*h_{\text{chan}}(t)\right]+n_3(t)$$

که در این رابطه، $y(t)$ خروجی گفتار که برای بازشناسی به سیستم داده می‌شود و $s(t)$ گفتار بدون نویز و بدون اعوجاج که در محیط با نویز زمینهٔ $n_1(t)$ و تکیه و اثر لمبارد آن، بیان شده است. $n_1(t)$ نویز زمینه، $h_{\text{mic}}(t)$ پاسخ ضربهٔ میکروفن، $n_2(t)$ نویز خط انتقال، $h_{\text{chan}}(t)$ پاسخ ضربهٔ خط انتقال و $n_3(t)$ نویز گیرنده می‌باشد.

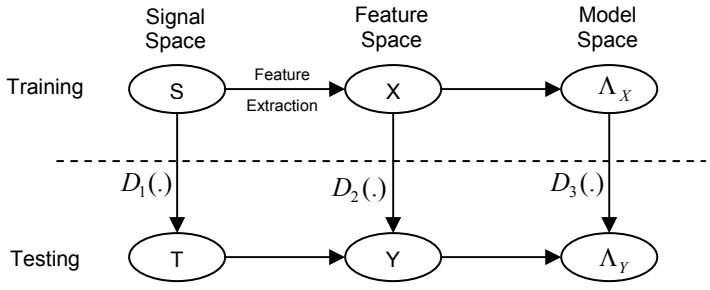
با توجه به پیچیده بودن نحوهٔ اثر و در عین حال کم بودن تاثیر مسالهٔ لمبارد، از اثر این پدیده صرفنظر می‌شود. اگر اثر همهٔ نویزها جمعی در یک منبع و اثر همهٔ نویزهای پیچشی را در منبعی دیگر، خلاصه شوند، رابطهٔ $y(t)=s(t)*h(t)+n(t)$ بدست می‌آید که $n(t)$ و $h(t)$ به ترتیب اثر کلی نویزجمع‌شونده و اثر کلی نویز پیچشی می‌باشند.

---

1 Additive
2 Convolutional
3 Lombard

۵-۱- تقسیم بندی روش‌های بازشناسی مقاوم گفتار

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



شکل(۴)

فضاهای ممکن برای عدم انطباق بین داده‌های مرحلهٔ آموزش و مرحلهٔ آزمایش

فضای اول، خود سیگنال گفتار آموزش یا آزمایش می‌باشد. در فضای سیگنال، سیگنال مرحله آموزش،S و سیگنال مرحلهٔ آزمایش، T می‌باشد. یک تابع $D_1(0)$، عدم انطابق بین S و T را بیان می‌دارد. سیگنالها در ادامه به بردارهای ویژگی تبدیل می‌شوند. در مرحلهٔ آموزش، بردار X و در مرحله آزمایش بردارY را داریم. عدم انطباق بین این دو با $D_2(0)$ بیان می‌شود. از آنجا که یکی از معمول‌ترین روش‌های بازشناسی، استفاده از مدل مخفی مارکف[1] می‌باشد، در ادامه سیستم از روی بردارهای ویژگی، مدل‌های λ_x و λ_y را بدست می‌آورد. عدم انطباق بین دو مدل اخیر، با تابع $D_3(0)$ بیان شده است[۹].

با توجه به توضیحات فوق به طور کلی می‌توان روش های مقاوم نمودن سیستم‌های بازشناسی گفتار را در مقابل نویز به سه دستهٔ مهم زیر تقسیم بندی کرد[۲]:

- دسته اول از روشها، آنهایی هستند که سیگنال گفتار را به ویژگی‌های ذاتاً مقاوم در مقابل نویز تبدیل می‌کنند. در این روش‌ها، بعد از استخراج ویژگی**هاي** سیگنال، پردازش اضافی دیگری مورد نیاز نمی‌باشد.
- در دستهٔ دوم، سیگنال گفتار عاری از نویز از روی سیگنال نویزی بدست می‌آید. سیگنال تمیز به‌دست آمده، در ادامه به یک سیستم بازشناسی معمولی داده می‌شود. در واقع در این دسته، از روشهای بهبود گفتار برای بازشناسی بهره برده می‌شود.
- در روش‌های دستهٔ سوم، سعی بر آن است که مدل بازشناسی نسبت به نویز مقاوم شود.

در ادامه به بررسی هر یک از این روش‌ها می‌پردازیم.

۵-۱-۱- ویژگی‌های مقاوم در مقابل نویز

هدف از این راه حل، انتخاب ویژگی‌هایی است که با استفاده از آنها بتوان اثر نویزهای جمعی و یا پیچشی را به حداقل رساند. به عبارت دیگر، ویژگی‌هایی را برای بیان سیگنال گفتار به کار می‌گیریم که بالاترین میزان مقاومت در مقابل انواع

---

1 Hiden Markov Model

نویزها را داشته باشند. برخی از این روشها (ویژگی‌ها) در مواجهه با نویز جمعی و برخی دیگر برای مقابله با نویز پیچشی، اعوجاج کانال و انواع دیگر نویز طراحی شده‌اند. در این میان روش‌هایی نیز هستند که قابلیت مقابله با هر دو نویز را دارند[۲]. این خانواده از روش‌ها در فضای بردار ویژگی فعالیت می‌نمایند و در ادامه به بررسی برخی از اعضای مهم این دسته پرداخته خواهد شد.

۵-۱-۱-۱ نرمال کردن ضرائب کپسترال[1]

یک روش معمول، آن است که برای پایدارسازی ویژگیهای کپسترال در مقابل نویز جمعی، مقدار متوسط آنها را از مقدارشان کسر نمائیم، چرا که دیده شده که نویز جمعی فقط مقدار متوسط ویژگیها را تغییر می‌دهد[۲].

در این روش، پس از تبدیل فریمهای گفتار به ویژگیهای کپسترال، در یک پردازش اضافی، مقدار متوسط از ضرائب کپسترال کسر می‌گردد. این مقدار متوسط می‌تواند در زمان کوتاه و یا در جملهٔ محاسبه شده باشد. آزمایشهای انجام شده حکایت از قدرت این روش در مواجهه با نویز جمعی دارد.

۵-۱-۱-۲ آنالیز عمومی کپسترال

معمولی‌ترین عناصری که در بردارهای ویژگی مورد استفاده واقع می‌شود، ضرائب کپسترال می‌باشند. این ویژگیها در شرایط گفتار تمیز خیلی خوب عمل می‌نمایند ولی در هنگام نویزی شدن ورودی، عملکردشان چندان مطلوب نمی‌باشد. یکی ار راه حلهایی که برای استفاده از ضرایب کپسترال در محیط‌های نویزی به‌کار گرفته شده، استفاده از برخی توابع دیگر بجای لگاریتم در محاسبه این ضرائب می‌باشد. یکی از توابع مذکور، ریشه[2] می‌باشد[۲].

۵-۱-۱-۳ تحلیل طیف نسبی[3]

یکی از روش‌های مقابله با نویز در کانالهای متعدد با زمان با سرعت کم، *RASTA* می‌باشد[۹]. این پردازش، از این واقعیت استفاده می‌کند که نویز پیچشی تقریباً حالت ایستان دارد، بنابراین با اعمال یک فیلتر بالاگذر یا میان‌گذر می‌توان تقریباً اثر نویز پیچشی را حذف کرد.

حالت تعمیم یافته *RASTA* با نام *J-RASTA* برای مقابله با نویز جمعی و پیچشی توأم با هم طراحی شده است[۱۰]. در روش پایهٔ *RASTA*، از یک لگاریتم و عکس آن در مراحل پردازش استفاده می‌شود[۲] در حالی که در *J-RASTA* این مورد به صورت $y_i = \log(1 + Jx_i)$ اصلاح می‌شود که در آن x_i سیگنال ورودی است. عکس این تابع به صورت تقریبی به صورت $x_i = \frac{\exp(y_i)}{J}$ قابل حصول است. برای هر سطح نویز، می‌باید مقدار بهینهٔ J را یافت.

لازم به ذکر است که عملکرد روش *J-RASTA* به شدت به مقدار J وابسته است. اگر چه روشهایی برای یافتن مقدار بهینهٔ J وجود دارد ولی هنوز معلوم نیست که آیا این روشها در همهٔ محدودهٔ وسیع تغییرات میزان و نوع نویز به خوبی عمل می کنند یا خیر[۱۰].

---

1 Cepstral Mean Normalization (CMN)
2 Root
3 RelAtive SpecTrAL(RASTA)

۵-۱-۲- روشهای مبتنی بر بهبود گفتار[1]

به جای اعتماد به یک سری ویژگیهای مقاوم در مقابل نویز، راه دیگری که به نظر می‌رسد استفاده از یک سیستم بهبود گفتار برای تخمین سیگنال عاری از نویز از روی نمونه‌های نویزی می‌باشد. در حالت کلی رابطهٔ $\hat{S}(t) = H(o(t), M_s, M_n)$ بیانگر اصول کلی حاکم بر این روشهاست. در رابطه فوق $\hat{S}(t)$ سیگنال عاری از نویز، $o(t)$ بردار مشاهده و M_s و M_n مدلهای سیگنال تمیز و نویز می‌باشند. $H(.)$ نیز بیانگر تابع تخمین‌زننده می‌باشد. در بعضی از روشهای این دسته، تخمین نویز از روی نواحی سکوت، امری ضروری است.

۵-۱-۲-۱ روش تفاضل طیفی

در یک دید کلی باید گفت که در این روش هر فریم، یک طیف برای نویز زمینه تخمین زده شده و این طیف از طیف سیگنال نویزی کسر می‌گردد. در حالت کلی این روش، فرمول پایه دارای ویژگیهای متعددی است که با انتخاب مقدار خاص برای هر ویژگی، حالتهای ویژه تفاضل طیفی به دست می‌آید[۱۱].

فرض اصلی در تفاضل طیفی، این است که نویز، ایستان و با واریانس صفر بوده و هیچ نوع تغییری در تخمین طیف نویز و سیگنال گفتار رخ نمی‌دهد. از آنجا که این فرض چندان درست نمی‌باشد، می‌باید یک حد آستانه برای میزان افت طیف سیگنال گفتار تعیین نمائیم تا از منفی شدن طیف جلوگیری گردد. این مساله در نواحی کم انرژی گفتار موجب کاهش دقت تخمین می‌گردد.

علاوه بر این اگر ویژگیهای دینامیک وارد بردار ویژگی شوند، باید به صورت تفاضل ویژگیهای تخمین زده شدهٔ استاتیک در نظر گرفته شوند و اینجاست که خطاهای وارد در مسأله تخمین، تشدید می‌گردد.

۵-۱-۲-۲ فیلتر بهینهٔ احتمالی[2]

در این روش، تابع تبدیل H یک تابع قطعه به قطعه خطی تعریف می‌شود. این تبدیل از طریق کوانیتزه کردن فضای ویژگیها به یک مجموعه از نواحی مشخص و مینیمم‌سازی میانگین مربع خطا بدست می‌آید[۲].

۵-۱-۲-۳ نرمال کردن کپسترال وابسته به کتاب کد[3]

مشابه روش قبل، خانواده روشهای نرمال کردن کپسترال وابسته به کتاب کد در جهت یادگیری یک تبدیل میان فضای سیگنال نویزی به سیگنال عاری از نویز، تلاش می‌نمایند. شکل استاندارد *CDCN* قابلیت یادگیری رفتار یک محیط جدید بدون هیچگونه اطلاعات اضافی مانند داده‌های استریو را داراست. با توجه به این که در روش مستقیم پیاده‌سازی *CDCN*، حجم کار محاسباتی خیلی زیاد است، معمولاً ار تقریبهایی برای ساده‌کردن استفاده می‌شود[۱۲]. در حالت اولیه، این روش برای شرائط حضور نویز پیچشی به همراه نویز جمعی کم طراحی شده بود.

---

1 Speech Enhacement
2 Probabilistic Optimal Filtering
3 Codebook Dependent Cepstral Normalization

۵-۱-۳ روشهای مبتنی بر مقاوم سازی مدلها

روش های قبلی، معمولاً در مرحله پارامتری کردن سیگنال گفتار وارد عمل می‌شوند. یک جایگزین برای آن روشها، اعمال تغییر در مدلهای اکوستیک می‌باشد. این روش این مزیت را به دنبال دارد که هیچگونه فرض یا تصمیم‌گیری در مورد گفتار را نیاز ندارد و داده‌های مشاهده در آن تغییر نمی‌کنند[۲].

در این روشها، مدل نویز و کانال انتقال، به طور مستقیم وارد بحث می‌شوند[۹]. نکتهٔ قابل ذکر دیگر آنکه این روشها در فضای مدلها فعالیت می‌نمایند.

۵-۱-۳-۱ انطباق برگشت خطی[1]

وظیفهٔ سازگار نمودن مدلها با یک محیط جدید را می‌توان به صورت کار تطبیق گوینده در نظرگرفت. در کار تطبیق گوینده یک سری از مدلهای مستقل از گوینده، برای یک گویندهٔ خاص یا گروهی از گویندگان خاص تنظیم می‌گردند.

یکی از روشهای متداول تطبیق گوینده *MLLR*[2] می‌باشد. در این روش، اگر فقط متوسطها در نظر گرفته‌شوند، رابطهٔ روبرو $\tilde{\mu}_s = A_s \xi_s$ نحوهٔ دستیابی به بردار متوسط سیگنال نویزی را می‌دهد. در این رابطه $\xi_s = [w(\mu)^T]^T$ و A_s تبدیلی است که ویژگیهای سیگنال تمیزرا به سیگنال نویزی تبدیل می‌نماید. ξ_s، بردار تعمیم یافتهٔ متوسطهاست که W در آن بایاس می‌باشد. در حالت بایاس صفر، $W=0$ و در غیر این صورت $W=1$ است[۲].

مشابه این رابطه را برای واریانسها هم می‌توان بدست‌آورد. یکی از اشکالات این روش، آن است که ماتریس A_s بدست آمده در واقع یک تناظر بین فریمها و حالتها را بیان می‌دارد. در حالت آموزش بدون ناظر که برچسب فریمها مشخص نیست و نیز در حالت آموزش با ناظر، هنگامی که مقدار اولیه‌ها چندان قوی انتخاب نشده است، تبدیلهای بدست آمده، چندان قوی نبوده و لذا منجر به ضغف عملکرد سیستم هم خواهند شد.

۵-۱-۳-۲ ماسک زدن بر روی نویز[3]

در این روش، هرگاه که در یک فریم، انرژی به زیر انرژی نویز بیفتد، خروجی انرژی فیلتر بانک را با یک سطح مناسب از نویز عوض می‌کنیم. در این صورت، اطلاعات نهفته در بخش های شدیداً نویزی گفتار حذف می‌شود. این روش در سیستمهای بازشناسی مبتنی بر مدل مخفی مارکف منجر به بهبود قابل توجه در عملکرد سیستم گردیده است[۹]. همانگونه که پیداست فرض اساسی در این بحث آن است که یکی از نویز یا گفتار بر دیگری اثر چیره کننده‌ای دارد و لذا هنگامی که انرژی این دو در اطراف هم باشند، این روش دچار مشکل خواهند شد.

روش ماسک زدن بر روی نویز با استفاده از الگوریتم‌هایی نظیر *Klatt*، *Bridle* و یا *Holmes* پیاده‌سازی می‌گردد[۲]. هر یک از این الگوریتمها در شرائط خاص مزایا و معایب مربوط به خود را دارا می‌باشند.

---

1 Linear Regression Adaptation
2 Maximum Likelihood Linear Regression
3 Noise Masking

۵-۱-۳-۳ تجزیهٔ گفتار و نویز[1]

می توان نشان داد که تابع همانندی[2] به کار رونده برای تشخیص حالت از روی سیگنال نویزی را می توان به صورت حاصل جمع تابعی از سیگنال تمیز و تابعی از نویز بیان نمود[۲].

مدلهای مخفی مارکف برای هر یک از دو سیگنال نویز و گفتار از روی داده‌های آموزشی به دست می‌آیند. در مرحلهٔ بازشناسی از یک الگوریتم ویتربی[3] در فضای مرکب سیگنال و نویز استفاده می‌شود و حالت بهینه به دست می‌آید. لازم به ذکر است در این روش، خروجی حالتها، توسط یک چگالی غیر گوسی مدل می‌گردد.

۵-۱-۳-۴ استفاده از فیلتر وینر[4]

معمول‌ترین سیستمهای بازشناسی گفتار، از ویژگیهای کپسترال استفاده می‌کنند. چرا که این ویژگیها، یکی از فشرده‌ترین شکلها را در بیان گفتار داشته و در ضمن ماتریس کوواریانس آنها قطری است و این، خود می تواند موجب سادگی کارها گردد[۲]. روش تجزیهٔ گفتار و نویز و بسیاری از روشهای ماسک زدن بر روی نویز بر اساس پارامترهای لگاریتم طیفی[5] طراحی شده‌اند. با توجه به استفاده از پارامترهای کپسترال در خیلی از سیستم‌های امروزی، اگر بتوان روشی برای تطبیق نویز با پارامترهای کپسترال طراحی نمود، کار ارزنده‌ای صورت گرفته است. یکی از روشهای مفید در این زمینه، فیلتر وینر می‌باشد که در سیستمهای بر اساس مدل مخفی مارکف و بر اساس انطباق زمانی پویا[6] بکار گرفته شده است[۲]و[۱۱].

۵-۱-۳-۵ روش های ترکیب موازی مدلها[7]

در این خانواده از روشها، به جای تغییر نمونه‌های آموزشی و آموزش مجدد سیستم، به طور مستقیم پارامترهای مدل برای حالت حضور نویز، جبران‌سازی می‌شود. این روشها، بر این فرض اساسی استوار است که مدلهای سیستم تمیز به خوبی بیانگر نمونه‌های آموزشی داده‌های تمیز می‌باشند[۲]و[۱۳].

جبران‌سازی از طریق ترکیب مدل گفتار تمیز با مدل نویز طبق یک تابع عدم انطباق[8] صورت می‌پذیرد. این تابع در واقع بیانگر نحوهٔ تأثیر نویز بر روی سیگنال گفتار می‌باشد.

یکی از مشکلاتی که این روش با آن روبروست، این است که در خیلی از مواقع، مدل خوبی از توزیع چگالی گفتار نویزی وجود ندارد و بنابراین باید دنبال راه حلهایی برای حل تقریبی و سریع مسأله بود.

---

1 Signal and Noise Decomposition (SND)
2 Likelihood
3 Viterbi
4 Wiener
5 Log.Spectral
6 Dynamic Time Warping (DTW)
7 Predictive Model based Compensation(PMC)
8 Mismatch Function

۶- ابعاد مسأله از نظر تعداد واژگان

سیستمهای بازشناسی مرسوم خیلی به مورد استفاده وابسته‌اند. کارایی این سیستمها هنگامیکه در دامنه‌ای آزماش می‌شوند که با دامنهٔ آموزش آنها متفاوت است به شکل چشمگیری افت می‌کند. این افت و کاهش در کارای سیستم از آن جهت به وجود می‌آید که این سیستمها برای کاری که مورد استفاده آنها می‌باشد بهینه شده‌اند. که این موجب می‌گردد نتوانند با محیط جدید تطبیق پیدا کنند.

۶-۱ سیستمهای مبتنی بر واحدهای آوایی وابسته به کلمه[1]

از نظر تئوری بهترین روش مدل‌سازی وابسته به بافت در سطح آواها، مدل‌سازی آواها به صورت وابسته به کلمه است، یعنی هر واج بسته به این که در چه کلمه‌ای قرار دارد، به صورت یک واحد آوایی جداگانه در نظر گرفته شود و برای آن مدل جداگانه‌ای تشکیل گردد؛ ولی این روش علاوه بر تعداد بسیار زیاد کلاسهایی که ایجاد می‌کند دو مشکل عمده دیگر نیز دارد، اول این که برای واجهای واقع در دو انتهای کلمه در گفتار پیوسته مدل‌سازی مناسبی صورت نمی‌گیرد چون این واجها تحت تاثیر کلمات قبلی یا بعدی نیز قرار می‌گیرند و کاملا وابسته به کلمه مورد نظر نیستند؛ دوم اینکه با تغیر مجموعه کلمات سیستم بازشناسی یا افزایش تعداد آنها، مدل‌سازی وابسته به بافت باید مجدداً انجام شود و این بسیار نامطلوب است. البته این روش برای مدل‌سازی واجهای تشکیل دهنده کلمات کوچک مانند "وَ"، "یا"، "اَز"، "به" بسیار موثر می‌باشد زیرا در کلمات کوچک تغییرپذیری طیفی واجها بسیار بیشتر از کلمات معمولی است.

۶-۲ سیستمهای مبتنی بر کلمه[2]

هنگامیکه تعداد واژگان کم باشد بهترین روش برای مدلسازی که بسیاری از اطلاعات وابسته به بافت را نیز مدل مورد نظر در بر داشته باشد استفاده از کلمه به عنوان واحد آوایی می‌باشد. با این عمل میتوان به خوبی هم‌تولیدی را که در بازشناسی گفتار پیوسته وجود دارد، مدل کرد.

۷- مراحل و زمان بندی ادامه پروژه

در فرصت باقی مانده، ساختارهای گوناگون مقاوم سازی نسبت به نویز خط تلفن درسیستم بازشناسی موجود به کار گرفته می‌شوند. همچنین ایده‌های مطرح جهت توانایی سیستم برای بازشناسی اعداد بیان شده به صورت طبیعی پیاده‌سازی خواهند شد. در مرحله بعد این ساختارها ارزیابی می‌گردند. در این مرحله سعی بر این است که با مطرح کردن ایده‌های جدید و اعمال تغییراتی در این روشها، کارایی آنها را افزایش دهیم. سپس ایده‌ها وتغییرات مورد نظر پیاده‌سازی می‌شوند و در خاتمه گزارش انجام این پروژه تهیه خواهد شد.

---

1 Word-dependent-phone-model
2Word-based-model

زمان بندی:

۱- پیاده‌سازی روشهای انتخابی	۲ ماه
۲- ارزیابی نتایج اولیه و به‌کارگیری ایده‌های جدید	۲ ماه
۳- نهایی کردن روشها و پیاده‌سازی تغییرات	۲ ماه
۴- تهیه گزارش نهایی	۲ ماه
جمع	۸ ماه

منابع

[1] L. Rabiner and B. H. Juang, *"Fundamentals of speech recognition"*, Prentice Hall, New Jersey, 1993.

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

[3] Uday Jain ,"*Connected Digit Recognition over Long Distance Telephone Lines using the SPHINX-II System"*, MS. Thesis, ECE Dept., Carnegie Mellon Univ., May 1996.

[4] Pedro J. Moreno , *"Speech Recognition in Telephone Environments"*, MS. Thesis, ECE Dept., Carnegie Mellon Univ., December 1992.

[5] L. R. Rabiner, *"Applications of speech recohnition in the area of telecommunications"*, IEEE 1997.

[6] M. Bijankhan, J. Sheykhzadegan, M. R. Roohani, R. Zarrintare,S. Z. Ghasemi and M. E. Ghasedi *"Tfarsdat - the Telephone Farsi Speech Database"*, Euro Speech 2003.

[7] H. G. Hirsch and D. Pearce, *"The AURORA experimental framework for the performance of speech recognition systems under noisy conditions"*,ISCA ITRW, September 2000.

[8] M. B. Carey,H. T Chen, A. Descloux,J. F. Ingle and K. I.Park, *"Analog Voice and Voiceband Data Transmission Performance Characterization of the Public Switched Network"*, AT&T Bell Labs Tech. J.,Vol. 63,PP. 2059-2119,November 1984.

[9] C. H. Lee, *"On stochastic feature and model compecsation approaches to robust recognition"*, Speech Communication, Vol. 25, PP. 29-47, 1998.

[10] J. Koehler and et al, *"Integrated RASTA-PLP into speech recognition"*, In proceeding ICASSP'94, Vol. 1, PP. 421-424, 1994.

[11] S. Vaseghi and B. P. Milner, *"Noise compensation methods for hidden markov model speech recognition in adverse environments"*, IEEE Transaction on Speech and Audio, Vol. 5, No. 1, 1997.

[12] A. Acero and R. M. Stern, *"Environmental robustness in automatic speech recognition"*, In Proceeding ICASSP'90, PP. 849-852, 1990

[13] M. J. F. Gales, "*Predicative model based compensation schemes for robust speech recognition"*, Speech Communication, Vol 25, PP. 49-74, 1998.